ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ
ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΤΜΗΜΑΤΟΣ ΕΦΑΡΜΟΣΜΕΝΗΣ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΜΕΛΕΤΗ ΑΛΓΟΡΙΘΜΩΝ ΑΠΟΜΕΙΩΣΗΣ ΔΕΔΟΜΕΝΩΝ ΡΟΗΣ - ΠΕΙΡΑΜΑΤΙΚΗ ΜΕΛΕΤΗ ΜΕ ΧΡΗΣΗ ΤΟΥ MOA

Διπλωματική Εργασία

του

Γεωργίου Αλέξανδρου

Θεσσαλονίκη, 07/2020

ΜΕΛΕΤΗ ΑΛΓΟΡΙΘΜΩΝ ΑΠΟΜΕΙΩΣΗΣ ΔΕΔΟΜΕΝΩΝ ΡΟΗΣ - ΠΕΙΡΑΜΑΤΙΚΗ ΜΕΛΕΤΗ ΜΕ ΧΡΗΣΗ ΤΟΥ MOA

Γεωργίου Αλέξανδρος

Πτυχίο Μαθηματικών, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, 2015

Διπλωματική Εργασία

υποβαλλόμενη για τη μερική εκπλήρωση των απαιτήσεων του

ΜΕΤΑΠΤΥΧΙΑΚΟΥ ΤΙΤΛΟΥ ΣΠΟΥΔΩΝ ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΠΛΗΡΟΦΟΡΙΚΗ

Επιβλέπων Καθηγητής

Ευαγγελίδης Γεώργιος

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την ηη/μμ/εεεε

Ονοματεπώνυμο 1	Ονοματεπώνυμο 2	Ονοματεπώνυμο 3
.....	.....	.....

Γεωργίου Αλέξανδρος
.....

Περίληψη

Η ραγδαία αύξηση των δεδομένων ροής δημιούργησε την ανάγκη για μια νέα προσέγγιση στον τρόπο με τον οποίο εξάγεται η πληροφορία διότι η παραδοσιακή μηχανική μάθηση δεν είναι σε θέση να αντιμετωπίσει τα νέα εμπόδια που προκύπτουν από την φύση των δεδομένων ροής. Η μείωση των δεδομένων αποτελεί μια από τις κύριες τεχνικές στον τομέα της εύρεση γνώσης από δεδομένα ροής. Η παρούσα μελέτη επικεντρώνεται στη μείωση των δεδομένων με την μέθοδο της επιλογής στιγμιοτύπων η οποία αποτελεί ένα σημαντικό βήμα προεπεξεργασίας και εφαρμόζεται σε πολλά σενάρια μηχανικής μάθησης. Ο σκοπός της επιλογής στιγμιοτύπων είναι η μείωση των δεδομένων σε ένα πιο διαχειρίσιμο όγκο απομακρύνοντας στιγμιότυπα που αποτελούν θόρυβο ή πλεονασματικές τιμές βελτιώνοντας με αυτό τον τρόπο την υπολογιστική πολυπλοκότητα και την ακρίβεια του αλγόριθμου.

Στην παρούσα εργασία συνοψίζονται θεμελιώδεις έννοιες της εξόρυξης γνώσης από δεδομένα ροής, αναλύεται η προεπεξεργασία γνώσης σε αυτά και επικεντρωνόμαστε σε μια αναλυτική παρουσίαση τεσσάρων βασικών μεθόδων μείωσης δεδομένων με επιλογή στιγμιοτύπων. Για να εμπλουτίσουμε την έρευνα διενεργούνται αναλυτικά πειράματα με την χρήση του λογισμικού MOA για αυτές τις μεθόδους και παρουσιάζεται μια ανάλυση της προβλεπτικής τους ικανότητας, του βαθμού μείωσης που επιτυγχάνουν καθώς και του υπολογιστικού κόστους και των απαιτήσεων τους σε μνήμη. Τα κύρια συμπεράσματα στα οποία καταλήγει η έρευνα είναι ότι κάποιες μέθοδοι επιτυγχάνουν σημαντική μείωση στα δεδομένα εντούτοις όλες έχουν πολύ υψηλό υπολογιστικό κόστος κάτι που τις καθιστά σε αυτό το στάδιο ανίκανες να διαχειριστούν πραγματικά ρεύματα δεδομένων.

Λέξεις Κλειδιά:

Εξόρυξη γνώσης από δεδομένα ροής - Μείωση δεδομένων - Προεπεξεργασία δεδομένων - Επιλογή στιγμιοτύπων - λογισμικό MOA

Abstract

The rapid growth of data streams created a need for a new approach in the way information is extracted mainly due to the fact that traditional machine learning can not overcome the barriers that occur from the nature of data streams. Data reduction constitute on of the main techniques in the domain of knowledge discovery from data streams. This research focuses in data reduction through instance selection which is an important step of preprocessing that can be applied in many machine learning tasks. The purpose of instance selection is reduction of the data to a more manageable volume by removing noisy or redundant instances and thus improving the computational complexity and the accuracy of the algorithm.

In the present master's dissertation fundamental concepts of data stream mining are summarised, followed by an analyses of the four most important methods of data reduction with instance selection. To enrich our study we conduct thorough experiments for this methods with the use of MOA software and an analyses is presented regarding to their predictive performance, and reduction rates as well as their demands in computational time and memory usage. The main conclusions that can be drawn in this research is that some of the instance selection methods present a prominent reduction rate but all of the methods have a very high computational complexity nonetheless thus rendering them as incapable of managing real world data streams.

Keywords:

Data stream mining - Data reduction - Data preprocessing - Instance selection -MOA software

Πρόλογος – Ευχαριστίες

Με την ολοκλήρωση της μεταπτυχιακής εργασίας θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή κ. Γεώργιο Ευαγγελίδη για την συνεχή καθοδήγηση, την αμέριστη υποστήριξη καθώς και τις ουσιώδεις συμβουλές που μου παρείχε όλο αυτό το χρονικό διάστημα.

Επιπλέον θα επιθυμούσα να ευχαριστήσω όλο το τμήμα εφαρμοσμένης πληροφορικής του πανεπιστημίου Μακεδονίας για την μόρφωση και το υψηλό επίπεδο σπουδών που μου παρείχε βοηθώντας με να αποκτήσω πολύ χρήσιμα εφόδια στην επιστήμη της πληροφορικής η οποία καθορίζει κάθε πτυχή της σύγχρονης ανθρώπινης δραστηριότητας.

Περιεχόμενα

Κατάλογος Εικόνων

Κατάλογος Πινάκων

ΚΕΦΑΛΑΙΟ 1. Εισαγωγή

1.1 Πρόβλημα – Σημαντικότητα του θέματος

Η παρούσα διπλωματική εργασία διαπραγματεύεται το θέμα της μείωσης του όγκου σε δεδομένα ροής. Οι τεχνικές μείωσης όγκου δεδομένων χρησιμοποιούνται με σκοπό τον καθαρισμό των δεδομένων από θορυβώδη και περιττά στιγμιότυπα με στόχο την βελτίωση της απόδοσης των αλγόριθμων εξόρυξης γνώσης από δεδομένα ροής (data stream mining). Η προεπεξεργασία των δεδομένων αν και δεν είναι τόσο γνωστή όσο άλλα τμήματα της διαδικασίας εύρεσης γνώσης από δεδομένα παίζει καθοριστικό ρόλο στην προσπάθεια επίτευξης μοντέλων υψηλής ακρίβειας. Επί του παρόντος ο ρυθμός των παραγόμενων δεδομένων αυξάνει εκθετικά ακολουθώντας το φαινόμενο των Μεγαλών Δεδομένων. Τα σύγχρονα σύνολα δεδομένων αυξάνουν σε τρεις διαστάσεις οι οποίες είναι: τα γνωρίσματα, τα στιγμιότυπα και η πληθικότητα καθιστώντας την μείωση της πολυπλοκότητας ως αναγκαίο βήμα προεπεξεργασίας. Η φύση των δεδομένων ροής δημιουργεί την ανάγκη για μια διαφορετική προσέγγιση του θέματος της μείωσης δεδομένων σε σχέση με την κλασική εξόρυξη γνώσης σε στατικά δεδομένα. Μια από τις πιθανές λύσεις που προτείνονται για να αντιμετωπιστεί ο τεράστιος όγκος των δεδομένων είναι η μείωση των δεδομένων με την τεχνική της επιλογής στιγμιοτύπων (instance selection). Οι αλγόριθμοι μείωσης που έχουν προταθεί στην βιβλιογραφία δεν έχουν δοκιμαστεί σε μεγάλα δεδομένα ροής κάτι που αμφισβητεί την ικανότητα τους να λειτουργήσουν κάτω από πραγματικές συνθήκες δεδομένων ροής. Συνεπώς απαιτείται επιπλέον ανάπτυξη και βελτίωση των τεχνικών προεπεξεργασίας δεδομένων για το περιβάλλον των δεδομένων ροής.

1.2 Σκοπός – Στόχοι

Η παρούσα έρευνα αποσκοπεί σε μια σχολαστική απαρίθμηση, ανάλυση και σύγκριση των μεθόδων μείωσης που χρησιμοποιούν την τεχνική της επιλογής στιγμιοτύπων και οι οποίες θεωρούνται ως οι πλέον σύγχρονες και αποτελεσματικές. Οι στόχοι της συγκεκριμένης έρευνας είναι να συγκρίνει και να μελετήσει εκ νέου τις μεθόδους μείωσης με χρήση γεννητριών δεδομένων ροής που παρέχονται από το λογισμικό MOA έτσι ώστε να καταλήξει στο ποια μπορεί να θεωρηθεί ως η βέλτιστη. Παρόλο που υπάρχουν αντίστοιχες έρευνες που αφορούν τις μεθόδους μειώσεις σε δεδομένα ροής η συγκεκριμένη προσπαθεί να αναδείξει με πειραματικό τρόπο τα σημεία που χρήζουν περαιτέρω έρευνας και που θα αποτελέσουν τις μελλοντικές προκλήσεις που πρέπει να αντιμετωπιστούν από την ερευνητική κοινότητα.

1.3 Συνεισφορά

Η συγκεκριμένη εργασία εκτός από την ενδελεχή βιβλιογραφική έρευνα που προσφέρει στον κλάδο των δεδομένων ροής και ειδικότερα στις μεθόδους μείωσης με την επιλογή στιγμιοτύπων, πραγματοποιεί και μια αναλυτική περιγραφή και παρουσίαση του λογισμικού MOA και κάποιων συγκεκριμένων γεννητριών δεδομένων ροής.

Επιπρόσθετα παρατίθεται ένα ευρύ πειραματικό πλαίσιο με σκοπό να εμπλουτιστεί το πεδίο της έρευνας πάνω στις συγκεκριμένες μεθόδους με πειράματα που εκτελούνται πάνω σε διαφορετικά σύνολα δεδομένων από αυτά που υπάρχουν στην βιβλιογραφία. Η έρευνα αναλύει την απόδοση των μεθόδων σε πρόβλεψη, μείωση, χρόνο και μνήμη. Επιπρόσθετα μη παραμετρικοί έλεγχοι χρησιμοποιούνται για να ενισχύσουν τα τελικά συμπεράσματα.

1.4 Βασική Ορολογία

Ο κύριος κλάδος μέσα στον οποίο υπάγεται ολόκληρη η μελέτη της εργασίας ονομάζεται Data stream mining, το οποίο μεταφράζεται ως *Εξόρυξη γνώσης από δεδομένα ροής*. Το βασικό θέμα που αναλύεται είναι η *Προεπεξεργασία δεδομένων* (Data preproccesing) και συγκεκριμένα η *μείωση δεδομένων* (Data reduction). Η τεχνική μείωσης δεδομένων που επιλέγεται να αναπτυχθεί ονομάζεται instance selection και μεταφράζεται ως *επιλογή στιγμιοτύπων ή παραδειγμάτων,* ελλείψει ταυτόσημο όρου. Τέλος ο όρος *Μηχανική μάθηση* (Machine learning) χρησιμοποιείται παράλληλα με την εξόρυξη γνώσης αν και δεν πρόκειται για ταυτόσημες έννοιες.

1.5 Διάρθρωση της μελέτης

Η δομή της εργασίας είναι η ακόλουθη. Το πρώτο και παρών κεφάλαιο αποτελεί την εισαγωγή της εργασίας στο αντικείμενο της έρευνας όπου αναφέρονται τα κύρια σημεία στα οποία θα επικεντρωθεί η εργασία στη συνέχεια. Στο δεύτερο κεφάλαιο παρουσιάζεται το θεωρητικό υπόβαθρο. Αρχικά εξηγούνται οι θεμελιώδεις έννοιες που διέπουν την εξόρυξη γνώσης από δεδομένα ροής, έπειτα δίνεται έμφαση στο κομμάτι της μείωσης δεδομένων και τέλος γίνεται μια λεπτομερής ανάλυση στη μείωση δεδομένων με την τεχνική της επιλογής στιγμιοτύπων όπου και παρουσιάζονται οι αλγόριθμοι που θα μας απασχολήσουν στην συνέχεια. Το τρίτο κεφάλαιο περιγράφει την μεθοδολογία που ακολουθήθηκε και παρουσιάζει το πειραματικό πλαίσιο πάνω στις μεθόδους που αναλύθηκαν στο προηγούμενο κεφάλαιο. Τέλος στο τέταρτο και τελευταίο κεφάλαιο συνοψίζονται τα συμπεράσματα και οι πληροφορίες που μπορούν να εξαχθούν από την έρευνα και την πειραματική μελέτη που προηγήθηκε.

ΚΕΦΑΛΑΙΟ 2. Βιβλιογραφική Επισκόπηση – Θεωρητικό Υπόβαθρο

2.1 Εξόρυξη γνώσης από δεδομένα ροής

Οι σύγχρονες πρόοδοι στο hardware και στο software επιτρέπουν την συγκέντρωση και αποθήκευση δεδομένων τεράστιου όγκου. Όμως όπως είχε προβλέψει ο νόμος του Wirth (Wirth 1995) η πρόοδος στην υπολογιστική επεξεργασία δεν αυξάνεται με τον ίδιο ρυθμό. Συνεπώς το να διαχειριζόμαστε και να εξάγουμε χρήσιμες πληροφορίες μέσα από τεράστιους όγκους δεδομένων έχει αναχθεί σε μια μεγάλη πρόκληση της σύγχρονης επιστήμης της πληροφορικής λόγω των φυσικών περιορισμών των σύγχρονων υπολογιστών.

Επιπρόσθετα με τα δεδομένα που παράγονται σε γιγαντιαίες ποσότητες οι περισσότερες γεννήτριες τα παράγουν συνεχόμενα και με αυτό τον τρόπο υπάρχει η άνοδος των δεδομένων ροής. Εφαρμογές δεδομένων ροής περιλαμβάνουν συνεχόμενα παραγόμενα σετ δεδομένων που είναι πολύ μεγάλα για να αποθηκευτούν στην μνήμη και ως αποτέλεσμα αποθηκεύονται σε δευτερεύουσα μνήμη. Δεδομένου ότι το να υπάρχει τυχαία πρόσβαση στην δευτερεύουσα μνήμη για να ανακτηθούν δεδομένα είναι μια ακριβή διαδικασία ο μόνος εφικτός τρόπος για να υπάρχει πρόσβαση στα δεδομένα είναι καθώς αυτά καταφθάνουν. Η τεχνική αυτή ονομάζεται single pass processing (Guha 2009).

Η εξαγωγή χρήσιμης πληροφορίας από δεδομένα ροής είναι μια πρόκληση από μόνη της. Οι περισσότερες τεχνικές εξόρυξης γνώσης από δεδομένα υποθέτουν ότι υπάρχει ένα στατικό σετ δεδομένων στο οποίο η κατανομή πιθανότητας είναι σταθερή και έτσι μπορεί να αναλυθεί από αλγορίθμους πολλών βημάτων. Καμία από τις παραπάνω συνθήκες δεν μπορούν να επαληθευτούν σε ένα σενάριο όπου έχουμε ροή δεδομένων κυρίως λόγο του ότι η υποκείμενη κατανομή είναι πολύ πιθανόν να αλλάξει με την πάροδο του χρόνου φαινόμενο που ονομάζεται ως concept drift.

Το ρεύμα δεδομένων είναι μια εν δυνάμει ανεξάντλητη ακολουθία στιγμιοτύπων τα οποία καταφθάνουν με την πάροδο του χρόνου. Επομένως το ρεύμα δεδομένων επιβάλλει κάποιους συγκεκριμένους περιορισμούς στο σύστημα εκμάθησης που δεν μπορούν να ικανοποιηθούν από τους κανονικούς αλγόριθμους εκμάθησης. Παρακάτω αναφέρουμε τις κύριες διαφορές μεταξύ των στατικών δεδομένων και των δεδομένων ροής :

- Τα στιγμιότυπα δεν δίνονται εξαρχής αλλά γίνονται διαθέσιμα μέσω κάποιας ακολουθίας το ένα μετά το άλλο ή δίνονται με την μορφή κάποιων παρτίδων με δεδομένα (data chunks) καθώς το ρεύμα προχωράει.

- Τα στιγμιότυπα ενδέχεται να καταφθάνουν με ραγδαίο ρυθμό και με διαφορετικό χρονικό διάστημα μεταξύ τους.

- Τα ρεύματα έχουν εν δυνάμει άπειρο μέγεθος συνεπώς είναι αδύνατο να αποθηκευτούν όλα τα εισερχόμενα δεδομένα στην μνήμη

- Κάθε στιγμιότυπο μπορεί να είναι προσβάσιμο για ένα περιορισμένο αριθμό φορών (σε ειδικές περιπτώσεις μόνο μια φορά) και έπειτα απορρίπτεται για να περιοριστεί η χρήση της μνήμης και ο χώρος της αποθήκευσης.

- Τα στιγμιότυπα πρέπει να επεξεργαστούν μέσα σε ένα περιορισμένο χρονικό διάστημα για να προσφέρουν ανταπόκριση σε πραγματικό χρόνο και να αποφευχθεί η συγκέντρωση δεδομένων που περιμένουν να επεξεργαστούν.

- Η πρόσβαση στα πραγματική τιμή είναι περιορισμένη εξαιτίας του μεγάλου κόστους της αναζήτησης της ετικέτας για κάθε εισερχόμενο στιγμιότυπο.

- Η πρόσβαση στη πραγματική ετικέτα μπορεί να καθυστερήσει, σε πολλές περιπτώσεις οι πραγματικές ετικέτες είναι διαθέσιμες μετά από μεγάλη χρονική περίοδο όπως για παράδειγμα η έγκριση πιστωτικών καρτών μπορεί να πάρει δύο ή και τρία χρόνια.

- Τα στατιστικά χαρακτηριστικά των στιγμιοτύπων που καταφθάνουν από το ρεύμα δεδομένων είναι πιθανό να υπόκεινται σε αλλαγές με την πάροδο του χρόνου.

Συνεπώς εξαιτίας των ανωτέρω διαφορών η διαδικασία εκμάθησης από τα δεδομένα ροής διαφέρει σε σημαντικό βαθμό από την παραδοσιακή μηχανική μάθηση όπου τα δεδομένα αποθηκεύονται σε πεπερασμένα διαρκή αποθηκευτήρια. Οι κύριες ασυμφωνίες που πρέπει να αντιμετωπιστούν αφορούν την διαχείριση του τεράστιου όγκου των δεδομένων ροής, την ακολουθιακή φύση των δεδομένων, τους περιορισμούς στην ταχύτητα επεξεργασίας των δεδομένων και στο ότι τα δεδομένα δεν μπορούν να είναι προσβάσιμα πολλές φορές.

Μια από τις πιο διαδεδομένες και ευρέως μελετημένες εργασίες στην εξόρυξη γνώσης από δεδομένα ροής είναι η επιτηρούμενη κατηγοριοποίηση. Η κατηγοριοποίηση σε δεδομένα ροής είναι μια παραλλαγή των αυξητικών αλγορίθμων εκμάθησης η οποία πρέπει να είναι σε θέση να ικανοποιεί όλες τις παραπάνω απαιτήσεις που προκύπτουν από την φύση των δεδομένων ροής. Επιπρόσθετα οι κατηγοριοποιητές δεδομένων ροής πρέπει να είναι προσαρμόσιμοι καθώς εργάζονται πάνω σε δεδομένα που είναι δυναμικά και περιβάλλοντα που δεν είναι σταθερά. Για να εκπληρωθούν αυτές οι απαιτήσεις αναπτύχθηκαν νέες μέθοδοι που περιλαμβάνουν ειδική διαχείριση των δεδομένων, μηχανισμούς που ξεχνάνε, εντοπιστές αλλαγής που καταγράφουν την υποκείμενη αλλαγή στο ρεύμα καθώς και αποδοτικούς κατηγοριοποιητές και προσαρμόσιμους αλγορίθμους σύνολα που συνεχώς αντιδράνε στις αλλαγές του ρεύματος. Παρακάτω δίνεται ο ορισμός της κατηγοριοποίησης σε δεδομένα ροής καθώς και ένας ορισμός του ρεύματος δεδομένων σύμφωνα με τον Barddal (2019).

2.1.1 Data stream Classification - Ταξινόμηση δεδομένων ροής

Ταξινόμηση είναι η διαδικασία με την οποία κατανέμονται στιγμιότυπα από ένα σετ σε διακριτές κλάσεις σύμφωνα με σχέσεις ή ομοιότητες. Δοσμένου ενός σετ από πιθανές κλάσεις $Y = \{y_1, \ldots, y_c\}$ ένας κατηγοριοποιητής (classifier) χτίζει ένα μοντέλο που προβλέπει για κάθε στιγμιότυπο χωρίς ετικέτα $\vec{x}$ την αντίστοιχη κλάση y με μια ακρίβεια.

Ορισμός

Η εργασία ταξινόμησης ενός συνόλου δεδομένων ορίζεται σύμφωνα με τον Barddal (2019) ως εξής : Ένα σύνολο από n στιγμιότυπα εκπαίδευσης που είναι στην μορφή ($\vec{x}$, y) όπου $y \in Y$ είναι μια διακριτή ετικέτα κλάσης ενώ $\vec{x}_i$ είναι ένα d - διαστατό διάνυσμα από γνωρίσματα τα οποία ανήκουν σε ένα σύνολο X. Τα γνωρίσματα αυτά μπορεί να είναι κατηγορικά, αριθμητικά ή μεικτά. Ένας κατηγοριοποιητής παράγει από το αυτό το σύνολο εκπαίδευσης ένα μοντέλο $f : \vec{x} \rightarrow Y$ το οποίο χρησιμοποιείται για να κατηγοριοποιήσει όλα τα μελλοντικά στιγμιότυπα τα οποία έρχονται χωρίς ετικέτα.

Η κατηγοριοποίηση σε δεδομένα ροής ή αλλιώς η online κατηγοριοποίηση είναι μια παραλλαγή της παραδοσιακής κατηγοριοποίησης ενός συνόλου δεδομένων (batch classification). Η διαφορά μεταξύ των δύο προσεγγίσεων έχει να κάνει με το πως τα δεδομένα παρουσιάζονται στον αλγόριθμο εκμάθησης (learner). Στην περίπτωση του batch σχηματισμού ένα στατικό και εξ ολοκλήρου προσβάσιμο σύνολο δεδομένων παρέχεται στον αλγόριθμο εκμάθησης ο οποίος επιστρέφει ένα μοντέλο f με το οποίο προβλέπονται μελλοντικά οι κλάσεις των μελλοντικών στιγμιοτύπων. Αντίθετα στα περιβάλλοντα ροής τα στιγμιότυπα δεν είναι διαθέσιμα στον κατηγοριοποιητή για εκπαίδευση αλλά παρουσιάζονται με μια συχνότητα με την πάροδο του χρόνου και ο αλγόριθμος εκμάθησης πρέπει να ενημερώνει το μοντέλο του σύμφωνα με την άφιξη των στιγμιοτύπων από το ρεύμα (Bifet 2010).

Ορισμός

Έστω $S = \left[(\vec{x}, \vec{y})\right]_{t=0}^{\infty}$ ο ορισμός ενός ρεύματος δεδομένων (data stream) το οποίο παρέχει στιγμιότυπα (x^t, y^t) από τα οποία κάθε ένα καταφθάνει με μια χρονοσήμανση t. Το $\vec{x}^t$ είναι είναι ένα d -διάστατο διάνυσμα με γνωρίσματα το οποίο ανήκει στο σύνολο X και y^t είναι η πραγματική κλάση του $\vec{x}^t$ (Barddal 2019).

Στην παραδοσιακή μηχανική μάθηση οι περισσότερες από τις υπάρχουσες τεχνικές θεωρούν ότι υπάρχει ένα στατικό σύνολο δεδομένων το οποίο έχει παραχθεί από μια άγνωστη και στατική κατανομή πιθανοτήτων το οποίο μπορεί να αποθηκευτεί και να αναλυθεί σε διάφορα βήματα από έναν batch αλγόριθμο. Καμία από τις παραπάνω υποθέσεις δεν μπορούν να επιβεβαιωθούν σε ένα σενάριο όπου έχουμε δεδομένα ροής. Η ανάπτυξη αλγορίθμων σε αυτή την περίπτωση πρέπει να λάβει υπόψη πολλούς περιορισμούς (Bifet 2010). Πρώτα από όλα τα στιγμιότυπα συνεχώς γίνονται διαθέσιμα με την πάροδο του χρόνου και δεν υπάρχει έλεγχος ως προς την σειρά με την οποία καταφθάνουν ούτε πως αυτά πρέπει να επεξεργαστούν. Επιπρόσθετα τα ρεύματα είναι εν δυνάμει απεριόριστα επομένως τα στιγμιότυπα πρέπει να απορρίπτονται αμέσως μετά την επεξεργασία τους δεδομένου ότι υπάρχει περιορισμένος χώρος μνήμης. Λόγω της έμφυτης παροδικής διάστασης των δεδομένων ροής η υποκείμενη κατανομή των δεδομένων αναμένεται να αλλάξει δυναμικά με την πάροδο του χρόνου το οποίο σημαίνει ότι θα πρέπει να γίνουν αλλαγές και στο μοντέλο εκμάθησης. Το φαινόμενο αυτό ονομάζεται concept drift.

2.1.2 Υποθέσεις και περιορισμοί του data stream mining

Σύμφωνα με τον Barddal (2019) πρέπει να ληφθούν υπόψη οι εξής περιορισμοί :

- **Επεξεργασία ενός περάσματος (single pass processing)**

Οι κατηγοριοποιητές πρέπει να είναι σε θέση να επεξεργαστούν στιγμιότυπα συνεχόμενα σύμφωνα με την άφιξη τους. Ένας κατηγοριοποιητής πρέπει να

επεξεργάζεται στιγμιότυπα αμέσως μόλις γίνουν διαθέσιμα και να τα απορρίψει αμέσως μετά. Παρόλο που δεν υπάρχει περιορισμός για προσωρινή αποθήκευση για ένα περιορισμένο χρονικό διάστημα αυτή η ενέργεια δεν πρέπει να θέσει σε κίνδυνο το χώρο μνήμης και τους περιορισμούς στον χρόνο επεξεργασίας.

- **Χώρος μνήμης**

Η αρχική μνήμη είναι πεπερασμένη και η χρήση της πρέπει να γίνεται με τον βέλτιστο τρόπο. Συνεπώς οι κατηγοριοποιητές και τα αντίστοιχα μοντέλα πρέπει να έχουν όρια σύμφωνα με το υπάρχουν διαθέσιμο hardware.

- **Χρόνος επεξεργασίας**

Ο χρόνος επεξεργασίας του κάθε στιγμιοτύπου που καταφθάνει δεν πρέπει να ξεπερνά την αναλογία με την οποία τα νέα στιγμιότυπα γίνονται διαθέσιμα. Αν ο χρόνο επεξεργασίας κάθε στιγμιοτύπου γίνεται ολοένα και μεγαλύτερος τότε τα στιγμιότυπα που καταφθάνουν θα απορρίπτονται ή θα συγκεντρώνονται για να έρθει η δικιά τους σειρά για επεξεργασία και αυτό θα έχει ως αποτέλεσμα το σύστημα να «κρασάρει». Επιπλέον αν ο αλγόριθμος δεν είναι σε θέση να επεξεργαστεί τα στιγμιότυπα σε πραγματικό χρόνο τότε δεν θα είναι σε θέση να προσαρμοστεί έγκυρα σε ένα ενδεχόμενο concept drift.

- **Διαθεσιμότητα ετικετών**

Υποθέτουμε ότι μετά την άφιξη ενός στιγμιοτύπου $\vec{x}^t$ η αντίστοιχη του ετικέτα y^t γίνεται διαθέσιμη για εκπαίδευση πριν την άφιξη του επόμενου στιγμιοτύπου. Αυτό είναι με διαφορά το πιο χρησιμοποιούμενο πλαίσιο για την ανάπτυξη learners δεδομένων ροής και λογισμικών όπως το MOA (massive online analysis)ή το SAMOA (scalable advanced massive online analysis). Παρόλα αυτά υπάρχουν και άλλες ρυθμίσεις για τα δεδομένα ροής όπως η ημι-επιβλεπόμενη (semi-supervised) ή μη-επιβλεπόμενη (unsupervised) μάθηση.

2.1.3 Προβλήματα που παρουσιάζονται στην εξόρυξη γνώσης σε δεδομένα ροής

Concept drift

Τα δεδομένα πάντα παράγονται από κάποια συνάρτηση. Τα συμβατικά συστήματα αλγορίθμων εξόρυξης γνώσης θεωρούν ότι κάθε σύνολο δεδομένων παράγεται από μια συγκεκριμένη στατική κρυμμένη συνάρτηση. Αυτό σημαίνει ότι μια ομοιόμορφη συνάρτηση χρησιμοποιείται για την εκπαίδευση και τον έλεγχο των δεδομένων. Αυτή η υπόθεση μπορεί να αποτύχει στην περίπτωση των δεδομένων ροής καθώς η συνάρτηση που παράγει τα στιγμιότυπα σε ένα χρονικό σημείο t μπορεί να μην είναι η ίδια με αυτή που παράγει στιγμιότυπα την χρονική στιγμή t+1. Αυτή η πιθανή αλλαγή στην βασική συνάρτηση παραγωγής δεδομένων ονομάζεται **concept drift** (Kulkarni & Ade 2014)**.** Με άλλα λόγια το concept drift μπορεί να θεωρηθεί με μια πιο αφηρημένη έννοια ως το εμπόδιο που δημιουργείται από ανεπαρκή, άγνωστα η μη παρατηρούμενα χαρακτηριστικά του συνόλου δεδομένων ένα γεγονός που ονομάζεται ως κρυμμένο περιεχόμενο. Εδώ το φαινόμενο αυτό που στηρίζει μια πραγματική και στατική εικόνα με την πάροδο του χρόνου για κάθε κλάση καλύπτεται τελείως από την όραση του μαθητευόμενου μοντέλου. Νέα δεδομένα παράγονται από μια κρυφή συνάρτηση και ο αλγόριθμος εκμάθησης δεν το γνωρίζει με αποτέλεσμα το φαινόμενο του concept drift να είναι απρόβλεπτο. Αν η συνάρτηση παραγωγής δεδομένων για τα concept drifts είχε αναγνωριστεί θα μπορούσε να δημιουργηθεί και ο κατάλληλος κατηγοριοποιητής για κάθε ένα concept drift και θα χρησιμοποιούνταν αντίστοιχα ο πιο κατάλληλος για όλα τα πρόσφατα δεδομένα. Με την έλλειψη αυτής της γνώσης όμως θα πρέπει να δημιουργηθεί ένας ενιαίος κατηγοριοποιητής που θα είναι σε θέση να αντιμετωπίζει αυτές τις αλλαγές με την πάροδο του χρόνου.

Πρώτος ορισμός του concept drift (Kulkarni & Ade 2014)

Το concept drift λαμβάνει χώρα σε ένα περιβάλλον το οποίο κοινώς αποκαλείται ως μη σταθερό περιβάλλον. Σε τέτοιες περιπτώσεις θεωρούμε ότι την χρονική

στιγμή t ο αλγόριθμος Α δέχεται ένα σύνολο από στιγμιότυπα με ετικέτα $\{X_0, \ldots, X_t\}$ όπου X_i είναι ένα ν-διάστατο διάνυσμα και κάθε στιγμιότυπο έχει μια αντίστοιχη ετικέτα της κλάσης y_j. Αν ένα στιγμιότυπο X_{t+1} χωρίς ετικέτα έρθει κατά τη χρονική στιγμή t+1 τότε ο αλγόριθμος αναμένεται να του παρέχει μια ετικέτα της κλάσης. Αυτό αποτελεί μια πρόβλεψη για την ετικέτα του συγκεκριμένου στιγμιοτύπου. Μόλις αυτό ολοκληρωθεί η πραγματική ετικέτα Y_{t+1} και ένα νέο στιγμιότυπο ελέγχου X_{t+2} παρουσιάζονται ώστε να γίνει ο έλεγχος του μοντέλου. Επιπλέον την κρυμμένη συνάρτηση f_h που παράγει το στιγμιότυπο στο χρόνο t την ονομάζουμε f_t. Το concept drift θεωρείται ότι εμφανίζεται από την στιγμή που η συνάρτηση παραγωγής δεδομένων f_h ξεκινάει να μεταβάλλεται με τον χρόνο. Υπάρχουν πολλοί τρόποι με τους οποίους μπορεί να συμβεί αυτή η αλλαγή και θα σχολιαστούν παρακάτω. Το θεώρημα του Bayes μπορεί να υπολογίσει την πιθανότητα το στιγμιότυπο X_{t+1} να ανήκει στην κλάση C_i ως εξής :

$$P(C_i \mid X_{t+1}) = \frac{P(C_i) \cdot P(X_{t+1} \mid C_i)}{P(X_{t+1})}$$

Επομένως αυτό το θεώρημα δίνει την σχέση μεταξύ των εκ των προτέρων και των δεσμευμένων πιθανοτήτων των στιγμιοτύπων και των κλάσεων τους. Το concept drift μπορεί να λάβει χώρα όσο αφορά και τις τρεις κύριες μεταβλητές του θεωρήματος Bayes. Δηλαδή μπορεί να μεταβάλει την πιθανότητα $P(C_i)$, την πιθανότητα $P(X_{t+1} \mid C_i)$ καθώς και την πιθανότητα $P(C \mid X_{t+1})$.

<u>Δεύτερος ορισμός του concept drift</u> (Barddal 2019)

Έστω η παρακάτω σχέση που συμβολίζει ένα concept C, ένα σύνολο εκ των προτέρων πιθανοτήτων των κλάσεων και την συνάρτηση πυκνότητας πιθανότητας (NGUYEN et al. 2012)

$$C = U_{y \in Y}\{(P[y], P[\overrightarrow{x} \mid y]\}$$

Δοθέντος ενός ρεύματος δεδομένων S στιγμιότυπα του $(\overrightarrow{x^t}, y^t)$ θα παράγονται από το τρέχων concept C^t. Αν για κάθε στιγμιότυπο t_i του S έχουμε ότι $C^{t_i} = C^{t_{i-1}}$ τότε το concept είναι σταθερό. Διαφορετικά αν μεταξύ δυο χρονικών περιόδων t_i και $t_j = t_i + \Delta$ ισχύει ότι $C^{t_i} \neq C^{t_j}$ τότε αυτό σηματοδοτεί την ύπαρξη ενός concept drift.

Ο λόγος που έχουμε αλλαγή (drift) δεν μπορεί να καθοριστεί ούτε να προβλεφθεί από τους συμβατικούς αλγορίθμους μάθησης διότι δεν είναι εφικτό να υποθέσουμε ότι αυτοί έχουν πρόσβαση σε δευτερεύουσες πηγές δεδομένων ή ότι το κόστος είναι τέτοιο που καθιστά την πρόσβαση σε αυτές τις πηγές απαγορευτική. Επομένως οι αλγόριθμοι κατηγοριοποίησης για δεδομένα ροής πρέπει να εντοπίσουν αυτές τις αλλαγές και να προσαρμοστούν αυτόματα και αυτόνομα. Αν η γεννήτρια δεδομένων δεν είναι σταθερή (όπως συμβαίνει βέβαια στις περισσότερες ρεαλιστικές εφαρμογές) τότε οι αλλαγές στο πλαίσιο έχουν επίπτωση στο concept που πρέπει να γίνει εκμάθηση. Επομένως ο εντοπισμός και η προσαρμογή σε αλλαγές στο concept είναι μια υποχρέωση που πρέπει να τηρούν οι νέοι αλγόριθμοι που ασχολούνται με την εξόρυξη γνώσης σε δεδομένα ροής. Η παρουσία αλλαγής μπορεί να επηρεάσει τις υπό μελέτη ιδιότητες των κλάσεων που το σύστημα εκμάθησης επιχειρεί να ανακαλύψει. Από ένα σημείο και μετά η μείωση της ποιότητας του χρησιμοποιούμενου μοντέλου εξαιτίας του concept drift μπορεί να είναι σε τέτοιο βαθμό που το μοντέλο πλέον να θεωρείται ακατάλληλο εργαλείο για πρόβλεψη. Επομένως οι μέθοδοι που διαχειρίζονται τις αλλαγές στα δεδομένα ροής θεωρούνται καίριας σημασίας σε αυτή την περιοχή της έρευνας. Επιπλέον εκτός από τον εντοπισμό των concept drift αναμένεται από αυτούς τους αλγόριθμους να είναι σε θέση να ξεχωρίζουν ένα concept drift από δεδομένα που αποτελούν θόρυβο (noisy data) ή έκτοπες τιμές (outliers).

Παρακάτω παρουσιάζεται εν συντομία μια ταξινομία του concept drift. Υπάρχουν δύο κύριες διαστάσεις που πρέπει να ληφθούν υπόψη όταν αναλύεται

η φύση των αλλαγών που λαμβάνουν χώρα στη τρέχουσα κατάσταση οποιουδήποτε ρεύματος δεδομένων.

Επίδραση στα μαθημένα όρια κατηγοριοποίησης

Εδώ διακρίνονται δυο είδη concept drift. Το πραγματικό concept drift επηρεάζει τα όρια απόφασης (δεσμευμένες πιθανότητες) και μπορεί να επηρεάσει και την συνάρτηση πυκνότητας πιθανότητας επομένως αποτελεί μια απειλή για το σύστημα εκμάθησης. Το εικονικό concept drift δεν επηρεάζει τα όρια απόφασης αλλά έχει επίδραση στην δεσμευμένη συνάρτηση πυκνότητας πιθανότητας επομένως δεν επηρεάζει το μοντέλο εκπαίδευσης που χρησιμοποιείται. Παρόλα αυτά θα πρέπει και αυτό να εντοπιστεί.

2.1.4 Τύποι αλλαγών στα δεδομένα

Τα concept drift εμφανίζονται με δύο τρόπους: ξαφνικά και σταδιακά. Για να καθοριστεί αν ένα drift εμφανίζεται ξαφνικά ή σταδιακά πρέπει να αναλυθεί το μέγεθος του παραθύρου της αλλαγής (drift window) W_{drift}. Σύμφωνα με τον Barddal (2019) αν υπολογίσουμε ότι ένα drift συμβαίνει μετά από το στιγμιότυπο $\vec{x}^t$ και ότι σταθεροποιείται μετά από το στιγμιότυπο $\vec{x}^{t+W_{drift}}$ αν ισχύει W_{drift} = 1 τότε η αλλαγή είναι ξαφνική διαφορετικά αν $W_{drift} > 1$ είναι σταδιακή. Μέσα στην ζώνη αλλαγής η πιθανότητα ένα στιγμιότυπο $\vec{x}^t$ να ανήκει σε ένα παλιό concept C_A ή σε ένα νέο C_B είναι συγκεκριμένο σε κάθε πρόβλημα. Ωστόσο πολλές από αυτές τις πιθανότητες μπορούν να συνδυαστούν μέσα από γνωστές συναρτήσεις κατανομής πιθανοτήτων. Ανεξάρτητα από την συνάρτηση που έχει επιλεχθεί πρέπει να εξασφαλίσουμε ότι $P[C_A] + P[C_B] = 1$ καθώς και ότι $P\left[\vec{x} \in C_A\right] = 1 - P\left[\vec{x} \in C_B\right]$. Τέλος ένα πολύ σημαντικό σημείο σε αυτές τις συναρτήσεις είναι η σύγκλιση αυτών των πιθανοτήτων δηλαδή όταν

$P\left[\vec{x} \in C_A\right] = P\left[\vec{x} \in C_B\right] = 0.5$ το οποίο είναι γνωστό ως η στιγμή αλλαγής (drift moment) και συμβολίζεται ως t_{drift}. Επιπλέον υπάρχει και το αυξητικό concept drift το οποίο έχει πολύ πιο αργό ρυθμό μεταβολών όπου η διαφορά μεταξύ των δύο κατανομών δεν είναι στατιστικά σημαντική.

Άλλες κατηγορίες αλλαγών

Τέλος μπορεί να παρουσιαστεί και το επονομαζόμενο επαναλαμβανόμενο concept drift το οποίο σημαίνει ότι ένα concept από τις k προηγούμενες επαναλήψεις μπορεί να επανεμφανιστεί και αυτό μπορεί να συμβεί μόνο μια φορά ή να συμβαίνει περιοδικά. Τα στίγματα γνωστά και ως έκτοπες τιμές (outliers) θα πρέπει να παραβλέπονται από τις μεθόδους καθώς οι αλλαγές που εκπροσωπούν είναι τυχαίες (Kuncheva 2008). Ακόμη ο θόρυβος αντιπροσωπεύει ασήμαντες διακυμάνσεις και θα πρέπει να παραληφθεί. Το μικτό concept drift είναι ένα υβριδικό φαινόμενο όπου περισσότερα από ενός τύπου concept drift μπορεί να εμφανιστούν κατά την διαδικασία της εξόρυξης γνώσης. Πρέπει να σημειωθεί ότι στα πραγματικά σενάρια ο τύπος της αλλαγής που θα παρουσιαστεί είναι άγνωστος εξαρχής και πρέπει να καθοριστεί κατά την διάρκεια της επεξεργασίας του ρεύματος δεδομένων.

Οι Minku, White και Yao (2009) πρότειναν το αυστηρό κριτήριο το οποίο επιτρέπει να διαχωριστεί μια αλλαγή μεταξύ τοπικής (local) και καθολικής (global). Η τοπική αλλαγή επηρεάζει μόνο μια μικρή περιοχή του χώρου των γνωρισμάτων ενώ η καθολική αλλαγή επηρεάζει ολόκληρο τον χώρο τον γνωρισμάτων. Το τι προκάλεσε την καθολική αλλαγή είναι πιο εύκολα εντοπίσιμο από την αιτία που προκάλεσε μια τοπική αλλαγή. Επιπρόσθετα μπορεί να παρουσιαστεί το λεγόμενο “feature drift ” (Barddal et al. 2016), όπου οι αλλαγές επηρεάζουν μόνο τα επιλεγμένα γνωρίσματα. Δυστυχώς σε εργασίες κατηγοριοποίησης με πραγματικά δεδομένα υπάρχει το ενδεχόμενο οι αλλαγές να εμφανιστούν ως μια μίξη των παραπάνω περιπτώσεων.

Μη ισορροπημένες κλάσεις

Οι μη ισορροπημένες κλάσεις παρουσιάζονται όταν ένα σύνολο δεδομένων δεν περιέχει κατά προσέγγιση τον ίδιο αριθμό παραδειγμάτων για κάθε κλάση κάτι το οποίο μπορεί να είναι δραστικό σε πολλές περιπτώσεις (Kubat, Holte & Matwin, 1998). Η ανισορροπία στην εξόρυξη γνώσης είναι ένα μείζον πρόβλημα και λεπτομερής έρευνα με πίνακες εκτίμησης και χρήση της τελευταίας λέξης της τεχνολογίας για να αντιμετωπιστεί το πρόβλημα έχει πραγματοποιηθεί και είναι διαθέσιμη από τους He και Garcia (2009). Μια απλή λύση σε αυτό το πρόβλημα είναι η υπό ή υπέρ δειγματοληψία για την κλάση πλειοψηφίας ή την κλάσης μειοψηφίας αντίστοιχα. Αυτή η προσέγγιση όμως έχει κάποιες παγίδες. Η υποδειγματοληψία απορρίπτει στιγμιότυπα από την κλάση της πλειοψηφίας χωρίς να λαμβάνει υπόψη της χρησιμότητα τους. Από την άλλη η υπερδειγματοληψία παράγει αντίγραφα από τα παραδείγματα της κλάσης μειοψηφίας. Υπάρχουν προφανώς πιο έξυπνες μέθοδοι όπως ο συμπυκνωμένος εγγύτερος γείτονας CNN ο οποίος απορρίπτει στιγμιότυπα από την κλάση πλειοψηφίας τα οποία είναι μακριά από τα όρια απόφασης. Άλλη μια προσέγγιση έγινε από τον Tomek (1976) ο οποίος προτείνει μια υποδειγματοληψία για την κλάση πλειοψηφίας καθορίζοντας μια απόσταση μεταξύ των παραδειγμάτων από τις διάφορες κλάσεις.

Μια πιο προηγμένη μέθοδος προτείνετε από τον αλγόριθμο SMOTE (Chawla et al. 2002) όπου κατασκευάζει τον χώρο γνωρισμάτων της κλάσης μειοψηφίας βάζοντας τακτικά τεχνητά στιγμιότυπα στο κομμάτι του τμήματος που συνδέει δύο παραδείγματα της μειοψηφίας. Ο SMOTE έχει αποδειχθεί ότι ενισχύει της ακρίβεια της κατηγοριοποίησης στην κλάση της μειοψηφίας. Επιπλέον ο SMOTEBoost (Chawla et al. 2003) συγχωνεύει τον SMOTE με τον AdaBoost.M2 για να βελτιώσει περαιτέρω την F-measure και ανάκληση. Ο αλγόριθμος BEV (Li 2007) εκπαιδεύει κατηγοριοποιητές εφαρμόζοντας την μέθοδο bagging σε όλα τα δεδομένα της κλάσης μειονότητας και υποσύνολα παίρνονται από την κλάση πλειοψηφίας για εκπαίδευση. Η εκμάθηση από μη ισορροπημένα δεδομένα είναι ένα θέμα στο οποίο απευθύνεται και ο αλγόριθμος με την ονομασία Learn++.UD (Muhlbaier, Topalis & Polikar, 2004) ο οποίος όμως δεν έχει την δυνατότητα να μάθει από κλάσεις που δημιουργούνται

από ένα νέο concept. Η εξελιγμένη μορφή του ο Learn++.UDNC είναι μια λύση στο πρόβλημα των μη ισορροπημένων δεδομένων καθώς και στην δυνατότητα που πρέπει να έχει ένας αλγόριθμος δεδομένων ροής να μαθαίνει από νέα concepts. Εναλλακτικά ο Learn++.UDNC επιλεκτικά ενσωματώνει νέες ικανότητες από διάφορους αλγορίθμους μέσα στην οικογένεια των Learn++. Για παράδειγμα μια προσέγγιση με βάρη στις κλάσεις, κανονικοποίηση προκαταρτικών μέτρων εμπιστοσύνης καθώς και την εισαγωγή μιας νέας συνάρτησης μεταφοράς η οποία είναι ικανή να ρίξει την μεροληψία ενός υποσυνόλου το οποίο είναι εκπαιδευμένο στην κλάση πλειοψηφίας. Ο DataBoost-IM αλγόριθμος ανακαλύπτει παραδείγματα της κλάσης πλειοψηφίας που είναι δύσκολο να ταξινομηθούν. Μέσω αυτών παράγει νέα τεχνητά δεδομένα (Guo & Viktor 2004).

Μη ισορροπημένες κλάσεις και concept drift

Ένα μεγάλο πλήθος εφαρμογών σε μη σταθερά περιβάλλοντα όπου οι πηγές δεδομένων τους παρουσιάζουν το φαινόμενο του concept drift επηρεάζονται ταυτόχρονα και από την μη ισορροπία κλάσεων. Χαρακτηριστικά παράδειγμα είναι τα δεδομένα που συγκεντρώνονται από την επιτήρηση του καιρού, την αναγνώριση ανεπιθύμητων ηλεκτρονικών μηνυμάτων, τον εντοπισμό απάτης σε πιστωτικές κάρτες και το σύστημα εντοπισμού ηλεκτρονικής επίθεσης. Οι προσεγγίσεις με σύνολα μοντέλων είναι αυτές που κατεξοχήν χρησιμοποιούνται για την αντιμετώπιση του φαινομένου του concept drift αλλά έχουν επιπλέον αποδειχθεί χρήσιμα και στην καταπολέμηση της μη ισορροπίας των κλάσεων (Gao et al. 2008).

Ένας αλγόριθμος κατάλληλος για μη σταθερά περιβάλλοντα από ασταθή δεδομένα έχει προταθεί από τον Gao et al. (2008). Ο αλγόριθμος αυτός που ονομάζεται UCB (uncorrelated bagging) βασίζεται σε ένα πλαίσιο πακέτου όπου εκπαιδεύει κατηγοριοποιητές σε ένα υποσύνολο από στιγμιότυπα της κλάσης πλειοψηφίας και από ένα συνδυασμό από όλα τα παραδείγματα των κλάσεων μειοψηφίας που έχουν παρατηρηθεί έως τότε. Υπάρχουν αρκετά μειονεκτήματα

σε αυτή την προσέγγιση εξαιτίας των έμμεσων υποθέσεων που πραγματοποιεί. Επιπρόσθετα δεν είναι αυξητικός αλγόριθμος ενός περάσματος γεγονός που τον κάνει εξαιρετικά δύσκολο στο να διαχειριστεί ένα απαιτητικό ρεύμα δεδομένων. Ο αλγόριθμος SERA των Chen και He (2009) στοχεύει σε μια μετρική ομοιότητας για να επιλέξει πρώην παραδείγματα των κλάσεων μειοψηφίας τα οποία είναι κοντά στο πιο πρόσφατο σύνολο δεδομένων. Ο SERA είναι λιγότερο επιρρεπής στις αλλαγές των κλάσεων μειοψηφίας από ότι ο UCB και μπορεί να εφαρμοστεί με δύο μεθόδους: Να δημιουργεί μόνο έναν κατηγοριοποιητή για κάθε σύνολο δεδομένων ή να εφαρμόζει την μέθοδο bagging με προτίμηση στα δεδομένα της μειοψηφίας. Επιπλέον ο SERA παραλείπει τα στιγμιότυπα από το παρών σύνολο εκπαίδευσης που δεν θεωρεί χρήσιμα αξιοποιώντας την μετρική απόστασης Mahalanobis. Τέλος ο SERA αλγόριθμος έχει αναβαθμιστεί έτσι ώστε να μπορεί να εφαρμόσει την τεχνική του συνόλου (ensemble). Δυστυχώς και αυτή η τεχνική δεν είναι αυστηρά ενός περάσματος καθώς έχει πρόσβαση και σε προηγούμενα δεδομένα. Συνεπώς και οι δύο προσεγγίσεις που αναφέρονται πιο πάνω λειτουργούν καλύτερα όταν η έννοια των δεδομένων της μειοψηφίας είναι σταθερή και τα προηγούμενα δεδομένα μπορούν να αποθηκευτούν για μεταγενέστερη χρήση. Οι Xioufis et al. (2011) δημιούργησαν μια προσέγγιση βασισμένη σε παράθυρα που εφαρμόζει έναν k-NN για ταξινόμηση με πολλές ετικέτες για δεδομένα που παρουσιάζουν τόσο το πρόβλημα του concept drift όσο και της αστάθειας των κλάσεων.

Οι Ditzler, Polikar και Chawla (2010) πρότειναν δύο τεχνικές βασισμένες σε σύνολα που μπορούν να μάθουν σε μια ευρύτατη περιοχή των concept drift σε μη σταθερά περιβάλλοντα και αντιμετωπίζοντας το πρόβλημα της αστάθειας των κλάσεων. Επιπρόσθετα οι τεχνικές τους αποφεύγουν τους δύο κύριους περιορισμούς των προηγούμενων μεθόδων που είναι η συγκέντρωση δεδομένων των κλάσεων μειοψηφίας και η επαναχρησιμοποίηση προηγούμενων δεδομένων. Οι μέθοδοι που προτείνονται από τους προαναφερόμενους ερευνητές ονομάζονται Learn++.CDS και Learn++.NIE και είναι ενός περάσματος που σημαίνει ότι δεν χρειάζεται η πρόσβαση σε προηγούμενα δεδομένα ούτε αποθηκεύουν δεδομένα της μειοψηφίας για να ισορροπήσουν την αστάθεια των κλάσεων. Συγκεκριμένα ο Learn++.CDS εφαρμόζει τον SMOTE για να ισορροπήσει τις κλάσεις και τα στιγμιότυπα τους ενώ ο Learn++.NIE εφαρμόζει

υποσύνολα αλγορίθμων με χρήση του bagging μαζί με εναλλακτικές μετρικές σφαλμάτων για να μάθει από ασταθή δεδομένα. Και οι δύο αλγόριθμοι είναι ικανοί να αποκτήσουν άμεσα γνώση και να αποθηκεύσουν παλιές πληροφορίες για το περιβάλλον και για τις ρυθμίσεις του κάτι το οποίο είναι πολύ χρήσιμο όταν έχουμε concepts που επαναλαμβάνονται. Συγκεκριμένα ο Learn++.CDS είναι ένας αλγόριθμος που αποτελεί ένωση του Learn++.NIE και του SMOTE που αντιμετωπίζουν το concept drift και την αστάθεια κλάσεων αντίστοιχα. Ο Learn++.NIE είναι ένας πιο ευφυής αλγόριθμος ο οποίος εφαρμόζει μια διαφορετική τεχνική που βασίζεται κυρίως σε δύο πυλώνες. Ο πρώτος εφαρμόζει υποσύνολα τεχνικών με την μέθοδο bagging για να ανακόψει την αστάθεια των κλάσεων χωρίς όμως να παράγει συνθετικά δεδομένα και χωρίς να αποθηκεύει δεδομένα των κλάσεων μειοψηφίας. Ο δεύτερος εφαρμόζει διάφορες μετρικές. Ο Learn++.NIE αποδίδει καλά αν και οι δύο ορισμοί των κλάσεων μειοψηφίας και πλειοψηφίας αλλάζουν και μια ισχυρή ισορροπία απαιτείται και στις δύο αυτές κλάσεις για να επιτευχθεί μια καλύτερη απόδοση (Ditzler & Polikar 2013).

Έλλειψη γνωρισμάτων

Η ομοιομορφία και η πληρότητα των δεδομένων είναι αναγκαία για κάθε σύστημα κατηγοριοποίησης. Ένας κατηγοριοποιητής απαιτεί ειδική εκπαίδευση για να αντιμετωπίσει το πρόβλημα της απουσίας γνωρισμάτων και δεν μπορεί να λειτουργήσει με παραδείγματα που έχουν χαμένα χαρακτηριστικά. Η απώλεια δεδομένων σε πραγματικές εφαρμογές αποτελεί ένα κοινό σενάριο που μπορεί να οφείλεται σε δυσλειτουργικούς αισθητήρες, λανθασμένη πληροφορία εικονοστοιχείου, κενές απαντήσεις σε ερωτήσεις που έγιναν σε κάποια έρευνα, κατεστραμμένο εξοπλισμό, ιατρικοί έλεγχοι κάτω από κάποιες ειδικές συνθήκες που δεν μπορούν να παρακολουθηθούν κ.ο.κ. Σε όλες αυτές τις περιπτώσεις αποτέλεσμα είναι η απώλεια χαρακτηριστικών που θα έπρεπε να έχουν τα δεδομένα. Επιπλέον τιμές γνωρισμάτων που είναι πέρα από τα επιτρεπτά όρια των δεδομένων που ενδεχομένως να δημιουργήθηκαν από κάποιο θόρυβο, απώλεια του σήματος, αλλά και καταστροφή δεδομένων μπορούν και αυτά με την σειρά τους να θεωρηθούν ως απώλεια γνωρισμάτων που θα αναμέναμε να

έχουν τα δεδομένα. Η απλούστερη μέθοδος για να αντιμετωπιστεί αυτή η κατάσταση είναι να απορριφθούν όλα τα στιγμιότυπα τα οποία τους λείπουν κάποια γνωρίσματα. Όταν μεγάλα ποσοστά των δεδομένων αντιμετωπίζουν το πρόβλημα της απώλειας γνωρισμάτων τότε το φιλτράρισμα ή η διαγραφή δεν μπορεί να θεωρηθεί ως η βέλτιστη λύση και ενδεχομένως να είναι και μη πρακτικές λύσεις. Μια πιο ρεαλιστική προσέγγιση είναι ο καταλογισμός όπου οι τιμές που λείπουν συμπληρώνονται με βάση κάποιον υπολογισμό όπως ο μέσος ή βρίσκοντας την αντίστοιχη τιμή των κ εγγύτερων γειτόνων κτλ. Για να εξαχθούν ασφαλείς εκτιμήσεις για τις κενές τιμές τα δεδομένα εκπαίδευσης θα πρέπει να περιέχουν έναν ικανοποιητικό αριθμό γνωρισμάτων (Kulkarni & Ade 2014).

Αυτές οι τεχνικές που χρησιμοποιούν την μέθοδο του καταλογισμού (imputation) τείνουν να έχουν μεροληπτικά αποτελέσματα. Η πολυωνυμική παλινδρόμηση είναι μια επιπλέον μέθοδος για την διαχείριση των κενών δεδομένων αλλά είναι εφαρμόσιμη μόνο σε ένα συγκεκριμένο είδος εφαρμογών. Προσεγγίσεις με καλές εγγυήσεις στην απόδοση των μοντέλων έχουν επίσης ανακαλυφθεί. Πολλές από αυτές τις προσεγγίσεις βασίζονται στην εκτίμηση του μοντέλου όπως η Bayesian εκτίμηση που υπολογίζει τις δεσμευμένες και εκ των προτέρων πιθανότητες ενώνοντας τον απόντα χώρο γνωρισμάτων. Αυτές οι μέθοδοι χρειάζονται επίσης αρκετή πυκνή κατανομή των δεδομένων ενώ η παραμετρική κατανομή δεδομένων θα πρέπει να είναι γνωστή εξαρχής σύμφωνα με τους Kulkarni και Ade (2014). Προφανώς τέτοια εκ των προτέρων γνώση είναι δύσκολο να αποκτηθεί στις περισσότερες περιπτώσεις. Ο αλγόριθμος EM (Expectation Maximization) των Dempster, Laird και Rubin (1977) θεωρητικά αποδείχθηκε ως μια επαναληπτική μέθοδος που είναι εύκολη στην κατασκευή της. Ο EM όμως πάσχει από δύο περιορισμούς. Πρώτον γίνεται αργή σύγκλιση αν μεγάλο μέρος των δεδομένων έχει κενές τιμές. Δεύτερον το βήμα της μεγιστοποίησης μπορεί να είναι αρκετά δύσκολο αν η μορφή της κατανομής δεδομένων δεν είναι γνωστή ή αν διαφορετικά παραδείγματα έχουν απώλεια διαφορετικών γνωρισμάτων. Συνεπώς ο EM σε τέτοιες περιπτώσεις δεν λειτουργεί καλά σε πραγματικά προβλήματα. Επιπρόσθετα κάτι που τον κάνει ακόμα πιο δύσχρηστο είναι η ανάγκη να υπάρχει εκ των προτέρων γνώση για την κατανομή κάτι το οποίο συνήθως δεν είναι διαθέσιμο.

Αν ο υπολογισμός της κατανομής είναι εσφαλμένος τότε θα οδηγήσει σε ασύμφωνα αποτελέσματα ενώ η μη διαθεσιμότητα αρκετών πυκνών δεδομένων μπορεί να δημιουργήσει απώλεια στην ακρίβεια του μοντέλου. Για να αντεπεξέλθουμε τέτοιες δυσκολίες προτάθηκαν πιο αυστηρές μέθοδοι όπως η χρήση Γκαουσιανών μικτών μοντέλων (GMM) ή μέθοδοι που βασίζονται σε νευρωνικά δίκτυα. Αλγόριθμοι που βασίζονται στο κοινό ασαφή min-max νευρωνικό δίκτυο ή το ARTMAP είναι ακόμα μερικά παραδείγματα. Μέθοδοι με σύνολα έχουν επίσης παρουσιαστεί όπως ο αλγόριθμος DECORATE των Melville και Mooney (2005) ο οποίος παράγει τεχνητά δεδομένα από ήδη υπάρχοντα δεδομένα (που παρουσιάζουν έλλειψη τιμών) αποδεικνύονται αρκετά ισχυρά μοντέλα για την αντιμετώπιση της έλλειψης γνωρισμάτων. Οι Juszczak και Duin (2004) πρότειναν την ένωση σε σύνολο ταξινομητών μιας κλάσης όπου κάθε ένας κατηγοριοποιητής θα εκπαιδεύεται σε ένα μόνο γνώρισμα. Αυτή η μέθοδος είναι ικανή να διαχειριστεί κάθε είδους συνδυασμό χαμένων δεδομένων. Επίσης μπορεί να είναι αρκετά ισχυρή όσο ένα χαρακτηριστικό είναι σε θέση να υπολογίσει τα όρια απόφασης. Αυτό όμως δεν είναι πάντα εφικτό. Οι Polikar et al.(2001) προτείνουν έναν νέο τρόπο που ανήκει στην οικογένεια των Learn++ αλγορίθμων και είναι γνωστός ως Learn++.MF. Ο αλγόριθμος αυτός παράγει επαρκή συχνότητα από κατηγοριοποιητές όπου κάθε ένας εκπαιδεύεται με ένα τυχαία επιλεγμένο υποσύνολο γνωρισμάτων.

Ένα στιγμιότυπο με χαμένες τιμές κατηγοριοποιείται με βάση την πλειοψηφία από τους κατηγοριοποιητές που δεν ανέπτυξαν τις χαμένες τιμές κατά την περίοδο της εκπαίδευσης τους. Επομένως αυτή η μέθοδος μπορεί να διαχωριστεί από άλλες μεθόδους με βάση αυτή την θεμελιώδη πλευρά. Ο Learn++.MF επιχειρεί να εξάγει την πιο άδικη πληροφορία που καθορίζεται από την παρουσία παλαιότερων δεδομένων με σκοπό να πάρει το περισσότερο κέρδος της αναπαραγωγής στο σύνολο των γνωρισμάτων. Με άλλα λόγια ο Learn++.MF δεν έχει καμία ανησυχία με την πρόβλεψη των χαμένων τιμών των δεδομένων. Συνεπώς ο Learn++.MF αποτρέπει πολλά από τα μειονεκτήματα της εκτίμησης που έχουν οι αλγόριθμοι που χρησιμοποιούν την μέθοδο του καταλογισμού. Σύμφωνα με τους Kulkarni και Ade (2014) αυτός είναι ένας καινοτόμος τρόπος για να αντιμετωπιστεί ένα σημαντικό κομμάτι χαμένων τιμών

από τα δεδομένα χωρίς την σταδιακή επιδείνωση της απόδοσης καθώς ο όγκος των δεδομένων αυξάνει.

2.1.5 Κύριοι τρόποι αντιμετώπισης του φαινομένου του concept drift

Όπως προαναφέρθηκε η διαχείριση του φαινομένου του concept drift είναι θέμα ύψιστης σημασίας για την εκμάθηση από δεδομένα ροής. Υπάρχουν τρεις κύριες κατευθύνσεις που μπορούμε να ακολουθήσουμε στην περίπτωση εμφάνισης ενός concept drift. Πρώτον η επανεκπαίδευση του συστήματος της κατηγοριοποίησης από την αρχή κάθε φορά που ένα νέο στιγμιότυπο ή μια νέα παρτίδα από στιγμιότυπα γίνεται διαθέσιμη. Δεύτερον ο εντοπισμός των αλλαγών και η επανεκπαίδευση του μοντέλου μόνο όταν ο βαθμός της αλλαγής θεωρείται αρκετά σημαντικός. Τρίτον η αξιοποίηση προσαρμοστικών μεθόδων εκμάθησης που είναι σε θέση να ακολουθήσουν τις αλλαγές που παρουσιάζει το ρεύμα δεδομένων. Προφανώς ο πρώτος τρόπος χαρακτηρίζεται από ένα υπολογιστικό κόστος που θεωρείται μη αποδεκτό άρα απομένουν οι δύο εναπομείναντες κατευθύνσεις που πρέπει να ληφθούν υπόψη.

Η διαχείριση δεδομένων που αλλάζουν κατανομή με την πάροδο του χρόνου απαιτεί στρατηγικές εντοπισμού και ποσοτικοποίησης της αλλαγής. Επίσης απαιτεί μηχανισμούς που ξεχνάνε τα ξεπερασμένα παραδείγματα καθώς και την αναθεώρηση του μοντέλου που χτίζεται. Δικαίως τέτοιες γενικές στρατηγικές υπάρχουν για να εντοπίζονται οι αλλαγές και να αποφασίζουν πότε τα παραδείγματα δεν είναι πλέον σχετικά δηλαδή πότε υπάρχει concept drift. Στην συνέχεια αναλύονται τέσσερις κύριες προσεγγίσεις για να αντιμετωπιστεί αποτελεσματικά το concept drift όταν αυτό εμφανίζεται στα δεδομένα ροής που προσπαθούμε να κατηγοριοποιήσουμε.

2.1.5.1 Στρατηγικές εντοπισμού αλλαγής

Τα παρακάτω διαφορετικά σενάρια αλλαγής στον τρόπο με τον οποίο καταφθάνουν τα δεδομένα έχουν αναγνωριστεί και αναπτυχθεί στην

βιβλιογραφία από τον Tsymbal (2004), τους Widmer και Kubat (1996) και τον Stanley (2003). Πρόκειται για τα :

- concept change το οποιο διαχωρίζεται σε concept drift και concept shift
- αλλαγή στην κατανομή ή στην δειγματοληψία.

Ο όρος έννοια (concept) αναφέρεται στην υπό εξέταση μεταβλητή την οποία προσπαθεί να προβλέψει το μοντέλο. Η αλλαγή στην έννοια (concept change) είναι η αλλαγή στην υποκείμενη μεταβλητή με την πάροδο του χρόνου. Ο όρος concept drift περιγράφει μια σταδιακή αλλαγή της έννοιας και ο όρος concept shift αναφέρεται στις περιπτώσεις όπου η αλλαγή μεταξύ δύο εννοιών είναι πιο απότομη.
Η αλλαγή στην κατανομή η οποία είναι γνωστή και ως αλλαγή δειγματοληψίας ή εικονικό concept drift αναφέρεται στην αλλαγή στην κατανομή δεδομένων. Ακόμα και αν το concept παραμένει το ίδιο η αλλαγή μπορεί συχνά να οδηγήσει στην αναθεώρηση του μοντέλου καθώς ο βαθμός σφάλματος του μοντέλου μπορεί να μην είναι πλέον αποδεκτός με την νέα κατανομή δεδομένων.

Κάποιοι ερευνητές όπως ο Stanley (2003) πρότειναν ότι από πρακτικής άποψης δεν είναι αναγκαίο να διαχωριστεί η αλλαγή στο concept από την αλλαγή στην κατανομή καθότι και στις δύο περιπτώσεις το μοντέλο θα χρειαστεί να αναθεωρηθεί.

Ο εντοπισμός αλλαγής δεν είναι εύκολη εργασία καθώς υπάρχει ένας θεμελιώδης περιορισμός. Η σχεδίαση ενός ανιχνευτή αλλαγής είναι ένα ρίσκο μεταξύ του εντοπισμού μιας πραγματικής αλλαγής και ενός λάθος συναγερμού (Gustafsson 2000).

Το τεστ CUSUM :

Το συσσωρευτικό άθροισμα (CUSUM αλγόριθμος) που προτάθηκε αρχικά από τον Page (1954) είναι σύμφωνα με τους Bifet και Kirkby (2009) ένας αλγόριθμος ανίχνευσης αλλαγής ο οποίος ενεργοποιεί έναν συναγερμό όταν ο μέσος των δεδομένων εισόδου είναι στατιστικά σημαντικά διαφορετικός από το μηδέν. Η εισακτέα τιμή ε_t του CUSUM μπορεί να είναι οποιοδήποτε φίλτρο για

παράδειγμα το προβλεπόμενο λάθος από ένα φίλτρο Kalman. Το CUSUM τεστ είναι το ακόλουθο :

$g_0 = 0$

$g_t = max(0, g_{t-1} + \varepsilon_t - \upsilon)$

Αν $g_t > h$ τότε ενεργοποίηση του συναγερμού και $g_t = 0$

Το τεστ CUSUM δεν απαιτεί μνήμη και η ακρίβεια του εξαρτάται από την επιλογή των u και h παραμέτρων.

Το τεστ Γεωμετρικού κινούμενου μέσου GMA

Το τεστ CUSUM είναι ένας κανόνας τερματισμού. Υπάρχουν και άλλοι κανόνες τερματισμού όπως το GMA (Geometric Moving Average) το οποίο προτάθηκε πρώτη φορά από τον Roberts (2000) και είναι το ακόλουθο :

$g_0 = 0$

$g_t = \lambda g_{t-1} + (1 - \lambda)\varepsilon_t$

Αν $g_t > h$ τότε ενεργοποίηση του συναγερμού και $g_t = 0$

Ο παράγοντας λ χρησιμοποιείται για να δώσει περισσότερο ή λιγότερο βάρος στα τελευταία δεδομένα που έχουν αφιχθεί. Το κατώφλι h χρησιμοποιείται για να συντονίσει την ευαισθησία και τον βαθμό λανθασμένου συναγερμού του ανιχνευτή.

Στατιστικά τεστ :

Οι δύο προηγούμενες είναι μέθοδοι που αντιμετωπίζουν αριθμητικές ακολουθίες. Για πιο σύνθετους πληθυσμούς χρειάζεται να χρησιμοποιήσουμε άλλες μεθόδους. Υπάρχουν κάποια στατιστικά τεστ που μπορούν να χρησιμοποιηθούν για να ανιχνευτούν αλλαγές. Ένα στατιστικό τεστ είναι μια διαδικασία με την οποία αποφασίζουμε κατά πόσο η μηδενική υπόθεση που αφορά κάποιο ποσοτικό γνώρισμα ενός πληθυσμού είναι αληθής ή ψευδής. Εξετάζουμε την υπόθεση εξάγοντας ένα τυχαίο δείγμα από τον πληθυσμό και υπολογίζοντας κάποια κατάλληλο στατιστικό μέγεθος από τα αντικείμενα του δείγματος αυτού. Αν κάνοντας αυτό λάβουμε μια τιμή του στατιστικού μεγέθους

που θα συνέβαινε σπανίως όταν η υπόθεση είναι αληθής τότε θα έχουμε λόγο να απορρίψουμε την μηδενική υπόθεση και να δεχθούμε την εναλλακτική.

Για να εντοπίσουμε την αλλαγή χρειάζεται η σύγκριση δύο πηγών από τα δεδομένα και να αποφασιστεί αν η μηδενική υπόθεση H_0 ότι τα δύο δείγματα ήρθαν από την ίδια κατανομή είναι αληθής (Bifet & Kirkby 2009). Έστω λοιπόν ότι έχουμε δυο εκτιμητές $\hat{\mu}_0$ και $\hat{\mu}_1$ με διακυμάνσειςσ_0^2 και σ_1^2. Αν δεν υπάρχει αλλαγή στα δεδομένα τότε αυτοί οι εκτιμητές θα είναι σταθεροί. Διαφορετικά ένα υποθετικό τεστ θα απορρίψει την H_0 και θα ανιχνευτεί η αλλαγή. Υπάρχουν πολλοί τρόποι κατασκευής τέτοιου υποθετικού τεστ με το απλούστερο να είναι η μελέτη της διαφοράς :

$\hat{\mu}_0 - \hat{\mu}_1 \in N(0,\sigma_0^2 + \sigma_1^2)$ κάτω από την μηδενική υπόθεση H_0 ή να πραγματοποιηθεί ένα τεστ χ^2, όπου $\frac{(\hat{\mu}_0 - \hat{\mu}_1)^2}{\sigma_0^2 + \sigma_1^2} \in \chi^2$ κάτω από την μηδενική υπόθεση H_0.

Το τεστ Kolmogorov- Smirnov είναι ακόμα ένα βασικό στατιστικό τεστ για την σύγκριση δύο πληθυσμών. Λαμβάνοντας δυο δείγματα ένα από κάθε πληθυσμό οι συναρτήσεις αθροιστικής κατανομής μπορούν να καθοριστούν και να σχεδιαστούν. Με αυτό τον τρόπο η μέγιστη τιμή της διαφοράς μεταξύ των γραφημάτων μπορεί να βρεθεί και να συγκριθεί με μια κρίσιμη τιμή. Αν η παρατηρούμενη τιμή υπερβαίνει την κρίσιμη αυτή τιμή τότε η υπόθεση H_0 απορρίπτεται και εντοπίζεται η αλλαγή. Δεν είναι εύκολο να εφαρμοστεί το Kolmogorov-Smirnov τεστ σε δεδομένα ροής. Οι Kifer, David και Gerhrke (2004) προτείνουν μια KS δομή για να εφαρμόσουν το συγκεκριμένο αλλά και όμοια τεστ στην περίπτωση των δεδομένων ροής.

Μέθοδος ανίχνευσης αλλαγής

Η μέθοδος ανίχνευσης αλλαγής (DDM) που προτάθηκε από τους Gama et al. (2004) ελέγχει τον αριθμό των σφαλμάτων που παράγονται από το μοντέλο εκμάθησης κατά την διάρκεια της πρόβλεψης. Συγκρίνει τα στατιστικά δύο παραθύρων. Το πρώτο περιέχει όλα τα δεδομένα και το δεύτερο περιέχει μόνο

τα δεδομένα από την αρχή έως ότου ο αριθμός των σφαλμάτων αυξηθεί. Αυτή η μέθοδος δεν αποθηκεύει αυτά τα παράθυρα στην μνήμη. Κρατάει μόνο τα στατιστικά και ένα παράθυρο με πρόσφατα δεδομένα.
Ο αριθμός των σφαλμάτων σε ένα δείγμα με n παραδείγματα μοντελοποιείται από μια διωνυμική κατανομή. Για κάθε σημείο i στην ακολουθία που γίνεται η δειγματοληψία ο βαθμός σφάλματος είναι η πιθανότητα της λάθος ταξινόμησης p_i με σταθερή απόκλιση που δίνεται από την σχέση $s_i = \sqrt{p_i(1-p_i)/i}$. Θεωρείται σύμφωνα με τους Bifet και Kirkby (2009) ότι ο βαθμός του σφάλματος p_i του αλγόριθμου εκμάθησης θα μειωθεί καθώς αυξάνεται ο αριθμός των παραδειγμάτων αν η κατανομή των παραδειγμάτων είναι σταθερή. Μια ενδεχόμενη σημαντική αύξηση στο σφάλμα του αλγορίθμου υποδηλώνει ότι η τάξη της κατανομής αλλάζει και επομένως το μοντέλο απόφασης θεωρείται πλέον ακατάλληλο. Επομένως αποθηκεύονται οι τιμές p_i και s_i όταν το άθροισμα $p_i + s_i$ φτάσει την ελάχιστη τιμή του κατά την επεξεργασία και όταν οι ακόλουθες συνθήκες ενεργοποιούνται :

- $p_i + s_i \geq p_{min} + 2 \cdot s_{min}$ για το επίπεδο προειδοποίησης. Πέρα από αυτό το επίπεδο τα παραδείγματα αποθηκεύονται αναμένοντας μια πιθανή αλλαγή στο πλαίσιο.

- $p_i + s_i \geq p_{min} + 3 \cdot s_{min}$ για το επίπεδο παράσυρσης(drift). Πέρα από αυτό το επίπεδο το concept drift θεωρείται ότι είναι αληθινό επομένως το μοντέλο που δημιουργήθηκε από τον αλγόριθμο εκμάθησης επαναρυθμίζεται και ένα νέο μοντέλο εκπαιδεύεται χρησιμοποιώντας τα παραδείγματα που αποθηκεύτηκαν από την στιγμή που ενεργοποιήθηκε το επίπεδο προειδοποίησης. Οι τιμές των p_i, s_i επαναρυθμίζονται και αυτές.

Αυτή η προσέγγιση έχει καλή συμπεριφορά στο να ανιχνεύει απότομες αλλαγές αλλά και παροδικές όταν αυτές δεν είναι πολύ αργές. Έχει όμως σημαντικές δυσκολίες όταν η αλλαγή στα δεδομένα είναι αργή διότι σε αυτή την περίπτωση τα παραδείγματα θα αποθηκευτούν για μεγάλο χρονικό διάστημα και επειδή το επίπεδο της παράσυρσης μπορεί να πάρει πολύ ώρα μέχρι να ενεργοποιηθεί θα

υπερβούμε το όριο της μνήμης που αποθηκεύονται δεδομένα (Bifet & Kirkby 2009). Οι Baena-Garcia et al. (2006) προτείνουν μια νέα μέθοδος την EDDM με σκοπό να βελτιωθεί η DDM. Η μέθοδος EDDM φαίνεται να είναι καλύτερη από την DDM για κάποια σύνολα δεδομένων αλλά χειρότερη για κάποια άλλα. Βασίζεται στην εκτιμώμενη κατανομή των αποστάσεων μεταξύ των λαθών ταξινόμησης. Η διαδικασία αλλαγής του μεγέθους του παραθύρου γίνεται με τις ίδιες ευρετικές μεθόδους.

Εκθετικός κινούμενος μέσος με βάρη EWMA

Ένας εκτιμητής είναι ένας αλγόριθμος που εκτιμά τα επιθυμητά στατιστικά στα δεδομένα εισόδου τα οποία μπορεί να αλλάξουν με την πάροδο του χρόνου. Ο απλούστερος αλγόριθμος εκτίμησης είναι ο γραμμικός εκτιμητής ο οποίος επιστρέφει τον μέσο των δεδομένων που περιέχονται στην μνήμη. Άλλα παραδείγματα αποδοτικών εκτιμητών είναι ο αυτοπαλινδρομικός, ο αυτοπαλινδρομικός κινούμενος μέσος και τα φίλτρα Kalman.

Ένας εκθετικά κινούμενος μέσος με βάρη (EWMA) εκτιμητής είναι ένας αλγόριθμος που ανανεώνει την εκτίμηση μια μεταβλητής συνδυάζοντας τις πιο πρόσφατες εκτιμήσεις της μεταβλητής με τον EWMA όλων των προηγούμενων μετρήσεων (Bifet & Kirkby 2009):

$$X_t = a z_t + (1 - a) X_{t-1} = X_{t-1} + a(z_t - X_{t-1})$$

όπου X_t είναι ο κινούμενος μέσος, z_t είναι η τελευταία μέτρηση και α είναι το βάρος που δίνεται από την τελευταία μέτρηση (που είναι μεταξύ 0 και 1). Η ιδέα είναι να παραχθεί μια εκτίμηση που δίνει περισσότερο βάρος σε πρόσφατες μετρήσεις με την υπόθεση ότι οι πρόσφατες μετρήσεις είναι πιο πιθανό να είναι σχετικές. Επιλέγοντας ένα επαρκή α είναι ένα δύσκολο πρόβλημα το οποίο δεν είναι ασήμαντο.

Επομένως ανακεφαλαιώνοντας οι εντοπιστές αλλαγής είναι εξωτερικά εργαλεία τα οποία χρησιμοποιούνται μαζί με το μοντέλο κατηγοριοποίησης. Μετράνε διάφορες ιδιότητες του ρεύματος δεδομένων, όπως την τυπική απόκλιση, το αναμενόμενο σφάλμα την κατανομή των στιγμιοτύπων ή την σταθερότητα

(Vallim & Mello 2014). Οποιαδήποτε αλλαγή σε αυτές τις ιδιότητες αποδίδεται στην ενδεχόμενη παρουσία αλλαγής και συνεπώς ενεργοποιείται η επίβλεψη της εξέλιξης του ρεύματος δεδομένων. Οι περισσότεροι εντοπιστές αλλαγής είναι ρυθμισμένοι να δουλεύουν σε δύο επίπεδα. Στο πρώτο ένα σήμα προειδοποίησης εκπέμπεται όταν η αλλαγή αρχίζει να λαμβάνει χώρα δίνοντας την πληροφορία στο σύστημα εκμάθησης ότι ένας νέος κατηγοριοποιητής θα πρέπει να αρχίσει να εκπαιδεύεται με βάση τα πιο πρόσφατα στιγμιότυπα. Στο δεύτερο ένα σήμα εντοπισμού πληροφορεί το σύστημα εκμάθησης ότι ο τρέχων βαθμός αλλαγής είναι σημαντικός και ότι ο παλαιότερος κατηγοριοποιητής θα πρέπει να αντικατασταθεί από έναν καινούργιο. Αυτή η λύση είναι γνωστή και ως σαφής χειρισμός αλλαγής. Τέλος σημαντική άνοδο και ενδιαφέρον για την ερευνητική κοινότητα παρουσιάζουν τα σύνολα εντοπιστών (Wozniak et al. 2016).

2.1.5.2 Παράθυρα (windowing)

Μια κοινή προσέγγιση για την διαχείριση των δεδομένων και την αντιμετώπιση του data drift είναι να διατηρήσουμε ένα προβλεπτικό μοντέλο σταθερό με ένα σετ από πρόσφατα στιγμιότυπα (Silva et al. 2013). Υπάρχουν τρεις κύριες κατηγορίες της τεχνικής windowing στην βιβλιογραφία και είναι οι εξής : sliding (συρόμενο), damped και landmark. Και στις τρεις περιπτώσεις η δυσκολία έγκειται στο γεγονός του να επιλεχθεί το κατάλληλο μέγεθος του σετ λόγω του διλήμματος πλαστικότητας - σταθερότητας (plasticity-stability dilemma). Ενώ τα μικρά παράθυρα (windows) αντικατοπτρίζουν την τρέχουσα κατανομή των δεδομένων και διασφαλίζουν γρήγορη προσαρμογή στο drift (πλαστικότητα - plasticity) συνήθως χειροτερεύουν την απόδοση του συστήματος σε σταθερές περιοχές. Αντίθετα μεγάλα παράθυρα δίνουν καλύτερη απόδοση σε σταθερές περιόδους (δηλαδή σε περιόδους όπου δεν αλλάζει η κατανομή με την οποία έρχονται τα δεδομένα) (σταθερότητα stability) ωστόσο έχουν αργή αντίδραση άρα και προσαρμογή σε περίπτωση drift (Gama 2010).

Sliding window (συρόμενο παράθυρο) **:**

Τα συρόμενα παράθυρα αποθηκεύουν στην μνήμη ένα προκαθορισμένο ή μεταβλητό σύνολο από πρόσφατα παραδείγματα. Στην περίπτωση που είναι

προκαθορισμένο το ποσό των παραδειγμάτων κάθε φορά που ένα νέο στιγμιότυπο καταφθάνει υπόκειται σε μια FIFO (πρώτο μέσα τελευταίο έξω) πολιτική δομής δεδομένων όπου το παλαιότερο στιγμιότυπο απορρίπτεται. Σε μεταβλητού μεγέθους παράθυρα ο αριθμός των στιγμιοτύπων σε αυτή την δομή δεδομένων μπορεί να αλλάξει με την πάροδο του χρόνου συνήθως σύμφωνα με τα αποτελέσματα του ανιχνευτή αλλαγής του ρυθμού με τον οποίο καταφθάνουν τα δεδομένα (δηλαδή ανιχνευτή concept drift event). Μια ευθείς ιδέα σύμφωνα με τον Barddal (2019) είναι να συρρικνώσουμε το παράθυρο όταν εντοπιστούν αλλαγές στα δεδομένα έτσι ώστε τα δεδομένα που είναι αποθηκευμένα στην μνήμη να αντικατοπτρίζουν το προηγούμενο concept και να διατηρούμε μεγάλα παράθυρα κατά την διάρκεια που έχουμε σταθερές περιοχές του ρεύματος (data stream).

Damping window :

Σε αντίθεση με το συρόμενο παράθυρο τα damping windows συνδέουν ένα βάρος σε κάθε δεδομένο το οποίο φθίνει με την πάροδο του χρόνου. Επομένως τα πιο πρόσφατα στιγμιότυπα έχουν ένα πιο μεγάλο βάρος από ότι τα παλαιότερα και αυτά τα βάρη φθίνουν με την πάροδο του χρόνου σύμφωνα με κάποια φθίνουσα συνάρτηση (Barddal 2019). Αυτή η τεχνική παραθύρου είναι ενδιαφέρουσα διότι τα βάρη μπορούν να χρησιμοποιηθούν ως ένδειξη για το κατά πόσο είναι σημαντικό ένα στιγμιότυπο στο τρέχον concept επομένως μπορεί να ληφθεί υπόψη κατά την πρόβλεψη.

Landmark window :

Τα landmark windows απαιτούν την επεξεργασία ενός ρεύματος (stream) με το να διαχειρίζονται ξένα μεταξύ τους σύνολα από δεδομένα ξεχωριστά από στιγμιότυπα που λέγονται landmarks (ορόσημα). Τα landmark στιγμιότυπά καθορίζονται σύμφωνα με τον χρόνο ή με τον αριθμό των στιγμιοτύπων που έχει δει ο αλγόριθμος από το προηγούμενο landmark ή σύμφωνα με τους περιορισμούς της μνήμης. Όλα τα στιγμιότυπα που ανήκουν στο ίδιο landmark αποθηκεύονται ή συνοψίζονται σε μια κοινή δομή δεδομένων η οποία χρησιμοποιείται έπειτα για να εκπαιδεύσει το μοντέλο. Όταν ένα νέο landmark εμφανιστεί όλα τα δεδομένα στο τωρινό παράθυρο απορρίπτονται και νέα

στιγμιότυπα ανακτώνται από το ρεύμα έως ότου ένα νέο ορόσημο (landmark) βρεθεί. Για άλλη μια φορά το πρόβλημα του να καθοριστεί το διάστημα μεταξύ των landmarks μας οδηγεί στο plasticity -stability dilemma (Barddal 2019).

Οι παρακάτω αλγόριθμοι βασίζονται σε μεθόδους windowing για να προσαρμοστούν αναλυτικά στο feature drift.

- Concept-adapting Very Fast Decision Tree (CVFDT)
- Heterogeneous Ensemble for Data Stream (HEFT-Stream)
- Hoeffding Adaptive Tree (HAT)
- Heuristic Updatable Weighted Random Subspaces (HUWRS)
- Adaptive Random Forest (ARF)

2.1.5.3 online μοντέλα εκμάθησης

Τα online μοντέλα εκμάθησης ανανεώνονται με κάθε νέο εισερχόμενο στιγμιότυπο επομένως προσαρμόζονται με τις αλλαγές στο ρεύμα δεδομένων μόλις αυτές εμφανιστούν. Αυτά τα μοντέλα πρέπει να ικανοποιούν ένα σύνολο από απαιτήσεις οι οποίες είναι οι ακόλουθες. Κάθε αντικείμενο πρέπει να επεξεργαστεί μόνο μια φορά κατά την διαδικασία της εκπαίδευσης, η υπολογιστική πολυπλοκότητα της διαχείρισης κάθε στιγμιοτύπου πρέπει να είναι όσο το δυνατόν πιο μικρή και η ακρίβεια του μοντέλου δεν πρέπει να είναι μικρότερη από αυτή ενός κατηγοριοποιητή που εκπαιδεύτηκε σε ένα σύνολο δεδομένων που συλλέχθηκε ως εκείνη την χρονική στιγμή. Η online μάθηση εκτελείται σε συνεχόμενους γύρους όπου σε κάθε γύρο δίνεται στον αλγόριθμο ένα παράδειγμα εκπαίδευσης το οποίο μπορεί να θεωρηθεί ως μια ερώτηση με την απάντηση του κρυμμένη. Ο αλγόριθμος κάνει μια πρόβλεψη στην ερώτηση αυτή και η απάντηση αποκαλύπτεται. Με βάση την διαφορά μεταξύ της πρόβλεψης του αλγορίθμου και της πραγματικής απάντησης υπάρχει μια

απώλεια. Ο στόχος λοιπόν του μοντέλου είναι να ελαχιστοποιήσει τη συνολική απώλεια μέσα από όλους τους γύρους. Πρέπει να σημειωθεί ότι ένα σύνολο από κλασικούς αλγόριθμους κατηγοριοποίησης μπορούν να δουλέψουν και σε online λειτουργία όπως τα νευρωνικά δίκτυα ή ο Naive Bayes. Ωστόσο υπάρχει πληθώρα μεθόδων που έχουν προσαρμοστεί έτσι ώστε να παρέχουν αποδοτική online λειτουργία (Czarnecki & Tabor 2015). Αυτές οι μέθοδοι επίσης προσφέρουν έμφυτη διαχείριση των concept drift.

2.1.5.4 Μέθοδοι συνόλων (ensembles)

Τα συστήματα συνόλων (ensemble) που ονομάζονται και ως συστήματα πολλαπλών ταξινομητών MCS (multiple classifier systems) έχουν σχεδιαστεί και έχουν αποδειχθεί ότι διεξάγουν με επιτυχία την διαδικασία της εκμάθησης σε μη σταθερά περιβάλλοντα (Kuncheva 2008). Η λειτουργία τους συνοψίζεται ως εξής: Με κάθε νέο σύνολο δεδομένων που καταφθάνει νέοι κατηγοριοποιητές προστίθενται στο σύνολο και πολλές φορές παλαιότεροι απομακρύνονται με σκοπό την δημιουργία ενός μοντέλου συνόλου το οποίο παρακολουθεί την εξέλιξη των δεδομένων. Με αυτό τον τρόπο τα σύνολα προσαρμόζονται εύκολα σε μη σταθερά δεδομένα ροής. Λόγω της τμηματικής τους δομής είναι ικανά να ενσωματώνουν νέα στοιχεία των δεδομένων εισάγοντας νέα στοιχεία στο σύνολο, ενημερώνοντας τα υπάρχοντα στοιχεία των κατηγοριοποιητών ή αλλάζοντας τα βάρη κατά την φάση της συγκέντρωσης. Οι μέθοδοι συνόλων κατηγοριοποιούνται κυρίως σε αυτούς που είναι τμηματικοί και στις online προσεγγίσεις.

Τα σύνολα που είναι βασισμένα σε τμήματα ουσιαστικά ανά περιόδους αξιολογούν τους επιμέρους κατηγοριοποιητές τους με βάση το νέο σετ δεδομένων και αντικαθιστούν τον κατηγοριοποιητή μέλος που είχε την χειρότερη επίδοση με ένα νέο υποψήφιο κατηγοριοποιητή. Επιπρόσθετα όλες οι προτεινόμενες μέθοδοι λειτουργούν με ένα προκαθορισμένο αριθμό κατηγοριοποιητών. Ένα γενικό σχέδιο ενός συνόλου που βασίζεται σε τμήματα παρουσιάζεται στον ακόλουθο ψευδοκώδικα :

Αλγόριθμος: Γενικευμένη μορφή block-based ensemble (Street & Kim 2001)

Είσοδος : S : ρεύμα δεδομένων με παραδείγματα χωρισμένα σε τμήματα μεγέθους d

k : αριθμός μελών του συνόλου

Q(): μετρική ποιότητας κατηγοριοποιητή

Έξοδος : ε : Σύνολο k κατηγοριοποιητών με βάρη

1: **for all** τμήματα $B_i \in S$ **do**

2: χτίσε και θέσε βάρος στον υποψήφιο κατηγοριοποιητή C_c με χρήση των $B_i, Q()$;

3: θέσε βάρος σε όλους τους κατηγοριοποιητές C_j του ε με χρήση των $B_i, Q()$;

4: **if** |ε| < k **then**

5: $\varepsilon \leftarrow \varepsilon \cup \{C_c\}$;

6: **else if** $\exists j : Q(C_c) > Q(C_j)$ **then**

7: αντικατέστησε το πιο αδύναμο μέλος του συνόλου με το C_c ;

8: **end if**

9: **end for**

Για κάθε τμήμα $B_i \in S$ τα βάρη των εκάστοτε κατηγοριοποιητών $C_j \in \varepsilon$ υπολογίζονται από μια μετρική ποιότητας Q() που εξαρτάται από κάθε αλγόριθμο ξεχωριστά. Για παράδειγμα στο σύνολο με βάρη ακρίβειας (AWE) η μετρική Q() είναι μια έκδοση του μέσου τετραγωνικού σφάλματος του κάθε κατηγοριοποιητή C_j του συνόλου που υπολογίζεται στο τρέχων τμήμα B_i και συγκρίνεται με το σφάλμα ενός τυχαίου κατηγοριοποιητή από το ίδιο τμήμα

(Wang et al. 2003). Επιπλέον με τον υπολογισμό με βάρη των αλγορίθμων του συνόλου, ένας υποψήφιος κατηγοριοποιητής C_c χτίζεται από το πρόσφατο μπλοκ B_i και προστίθεται στο σύνολο μόνο αν δεν υπερβαίνει το επιτρεπτό μέγεθος του. Αν το σύνολο είναι γεμάτο ο υποψήφιος κατηγοριοποιητής αντικαθιστά το πιο αδύναμο μέλος του συνόλου. Αξίζει να σημειωθεί ότι κάποιοι αλγόριθμοι όπως ο Learn++.NSE (Ditzler et al. 2015) δεν θέτουν όριο στον αριθμό των κατηγοριοποιητών που πρέπει να υπάρχουν μέσα στο σύνολο με σκοπό να αντιδράσουν σε επαναλαμβανόμενα concept drifts. Η πρόβλεψη ετικέτας για τα νέα παραδείγματα συνήθως βασίζεται σε μια ψήφο της πλειοψηφίας με βάρη των κατηγοριοποιητών που ανήκουν στο σύνολο. Τα περισσότερα σύνολα με βάση τα τμήματα εκμεταλλεύονται αλγορίθμους της παραδοσιακής μηχανικής μάθησης και τους έχουν ως κατηγοριοποιητές στο σύνολο τους. Αυτό δεν συμβαίνει για υβριδικούς αλγορίθμους όμως όπως το σύνολο αναβαθμισμένης ακρίβειας (AUE) που αναβαθμίζει τους κατηγοριοποιητές μετά από κάθε μπλοκ B_i (Brzezinski & Stefanowski 2014).

Η αρχή των online στατικών συνόλων προέρχεται από την έρευνα στον αλγόριθμο Winnow και στον Weighted Majority Algorithm (Littlestone & Warmuth 1994) ο οποίος συνδυάζει τις προβλέψεις αρκετών κατηγοριοποιητών με βάση την πλειοψηφία. Όταν το σύνολο ταξινομήσει λάθος κάποιο στιγμιότυπο τα βάρη των κατηγοριοποιητών που έσφαλαν μειώνονται με την χρήση μιας σταθεράς που έχει θέσει ο χρήστης. Το σύνολο DWM (Dynamic Weighted Majority) είναι μια επέκταση αυτής της ιδέας για δεδομένα ροής που παρουσιάζουν αλλαγή (Kolter & Maloof 2007). Το σύνολο αυτό περιέχει αυξητικούς κατηγοριοποιητές που παράγονται από τον ίδιο αλγόριθμο εκμάθησης. Όταν ένα νέο παράδειγμα είναι διαθέσιμο η τελική πρόβλεψη αποκτάται ως μια ψήφο με βάρη από όλους τους κατηγοριοποιητές του συνόλου. Τα βάρη όλων των κατηγοριοποιητών που ταξινομούν λάθος το παράδειγμα μειώνονται και σε αυτή την περίπτωση όπως και στον WMA. Ωστόσο ο DWM δημιουργεί και διαγράφει δυναμικά κατηγοριοποιητές όσον αφορά αλλαγές στην απόδοση τους. Αν η ολική πρόβλεψη του συνόλου είναι εσφαλμένη ένας νέος κατηγοριοποιητής προστίθεται στο σύνολο.

Ακόμα ένα σύνολο online αλγορίθμων περιλαμβάνει γενικεύσεις στατικών συνόλων. Τα πιο γνωστά τέτοια σύνολα είναι οι online εκδόσεις του bagging και

boosting (Oza & Russel 2001b). Στην περίπτωση του online bagging η κεντρική ιδέα είναι να προσαρμοστεί το βήμα της δειγματοληψίας bootstrap σε ένα σενάριο με ρεύματα δεδομένων. Αυτό υλοποιείται με την χρήση μοναδικών παραδειγμάτων πολλές φορές σύμφωνα με την κατανομή Poisson. Αυτή η πρόταση για την τυχαία ανανέωση των συνόλων εκπαίδευσης αποτέλεσε έμπνευση για την ανάπτυξη και άλλων μεθόδων όπως το Leveraging bagging, το online boosting ή το σύνολο DDD (Ditzler et al. 2015).

Υβριδικές μέθοδοι που συνδυάζουν ενεργό εντοπισμό, συρόμενα παράθυρα και σύνολα με κατηγοριοποιητές έχουν επίσης προταθεί όπως οι random forests με εντροπία (Abdulsalam, Skillicom & Martin, 2011) και η καινοτόμος ενσωμάτωση ενός φίλτρου Kalmar με ένα προσαρμόσιμο αλγόριθμο με συρόμενο παράθυρο που ονομάζεται ADWIN (Bifet et al. 2010).

Τέλος όσον αφορά τον αυξητικό αλγόριθμο Learn++.NSE της οικογένειας των Learn++ αυτός είναι σε θέση να παράγει έναν καινοτόμο κατηγοριοποιητή με κάθε νέο πακέτο δεδομένων που καταφθάνει. Επίσης είναι σε θέση να παρακολουθεί αρκετά περιβάλλοντα όπως το προοδευτικό, το γρήγορο, το απότομο αλλά και το κυκλικό drift. Είναι ικανός να αναγνωρίζει τους περισσότερο αλλά και τους λιγότερο σχετικούς κατηγοριοποιητές και να τους δίνει υψηλά και χαμηλά βάρη αντίστοιχα αλλά και να παρατηρεί πότε ένας κατηγοριοποιητής γίνεται πάλι σχετικός εάν το περιβάλλον ή οι ρυθμίσεις του έχουν κυκλική φύση. Ο αλγόριθμος αυτός τέλος είναι ικανός να απορροφά κλάσεις που προστέθηκαν ή αφαιρέθηκαν αλλά δεν έχει σχεδιαστεί για να διαχειρίζεται το πρόβλημα της μη ισορροπίας των κλάσεων (Kulkarni & Ade 2014).

2.2 Προετοιμασία δεδομένων (data preprocessing)

Τα πραγματικά δεδομένα στην αρχική τους μορφή παρουσιάζουν μεγάλη μεταβλητότητα και αστάθεια. Πολλές τιμές ενδέχεται να λείπουν, να υπάρχει ασυμφωνία μεταξύ διαφορετικών πηγών καθώς και να έχουν καταχωρηθεί εσφαλμένες τιμές. Για τον αναλυτή επομένως που θα αξιοποιήσει αυτά τα δεδομένα παρουσιάζονται πολλές προκλήσεις ως προς το πώς θα τα

αξιοποιήσει αποτελεσματικά. Για παράδειγμα ας λάβουμε υπόψη το σενάριο όπου χρειαζόμαστε να αξιολογήσουμε τα ενδιαφέροντα των καταναλωτών σύμφωνα με την δραστηριότητα τους στις ιστοσελίδες κοινωνικής δικτύωσης. Ο ερευνητής πρέπει αρχικά να καθορίσει το ποιες δραστηριότητες είναι χρήσιμες για την διαδικασία της εξόρυξης γνώσης. Αυτές οι δραστηριότητες μπορεί να σχετίζονται με τα ενδιαφέρονται που έχει καταχωρήσει ο ίδιος ο χρήστης, τα σχόλια που έχει κάνει καθώς ακόμη και το σύνολο των φίλων που έχει ο χρήστης και το ποια είναι τα ενδιαφέροντα αυτών. Όλες αυτές οι πληροφορίες παρουσιάζουν μεγάλη ποικιλομορφία και χρειάζεται να συλλεχθούν από διαφορετικές βάσεις δεδομένων μέσα στις ιστοσελίδες κοινωνικής δικτύωσης. Επιπρόσθετα κάποια δεδομένα δεν είναι διαθέσιμα κατευθείαν προς χρήση λόγω της αδόμητης μορφής τους. Συνεπώς χρήσιμα χαρακτηριστικά και πληροφορίες πρέπει να εξαχθούν από αυτές τις αδόμητες πηγές δεδομένων. Επομένως μια φάση όπου τα δεδομένα επεξεργάζονται και προετοιμάζονται πριν εισαχθούν στον κυρίως αλγόριθμο είναι επιβεβλημένη.

Η φάση της προετοιμασίας των δεδομένων είναι μια διαδικασία που αποτελείται από πολλά στάδια εκ των οποίων κάποια ή ακόμα και όλα μπορεί να χρησιμοποιηθούν σε μια εφαρμογή αν αυτό απαιτείται. Σύμφωνα με τον Aggarwal (2015, p. 27) τα βήματα που γίνονται είναι τα ακόλουθα :

Εξαγωγή γνωρίσματος και φορητότητα

Τα δεδομένα όπως συλλέγονται είναι συνήθως σε τέτοια μορφή όπου δεν είναι κατάλληλα για να ξεκινήσει η διαδικασία επεξεργασίας τους από κάποιον αλγόριθμο. Κάποια χαρακτηριστικά παραδείγματα είναι τα raw logs, τα αρχεία, τα δεδομένα που δεν έχουν δομή αλλά και άλλες μορφές ετερογενών δεδομένων. Σε τέτοιες περιπτώσεις είναι θεμιτό να εξαχθούν χρήσιμα γνωρίσματα από τα δεδομένα. Γενικά γνωρίσματα με καλή σημασιολογική ερμηνεία είναι πιο θεμιτά διότι απλοποιούν την ικανότητα του αναλυτή να κατανοεί ενδιάμεσα αποτελέσματα. Επιπρόσθετα είναι συνήθως καλύτερα συνδεδεμένα με τους στόχους της εκάστοτε εφαρμογής εξόρυξης γνώσης. Σε κάποιες περιπτώσεις όπου τα δεδομένα αποκτώνται από πολλαπλές πηγές πρέπει να ενσωματωθούν σε μια συγκεντρωτική βάση δεδομένων για να ακολουθήσει η διαδικασία της επεξεργασίας. Επίσης μερικοί αλγόριθμοι μπορεί

να λειτουργούν μόνο με μια συγκεκριμένη κατηγορία δεδομένων ενώ τα δεδομένα μπορεί να περιέχουν διάφορους τύπους δεδομένων. Σε αυτές τις περιπτώσεις λοιπόν η ευελιξία των τύπων δεδομένων γίνεται πολύ σημαντική διότι γνωρίσματα του ενός τύπου μετατρέπονται σε γνωρίσματα άλλου τύπου. Με αυτό τον τρόπο καταλήγουμε σε ένα πιο ομογενές σύνολο δεδομένων το οποίο μπορεί πλέον να επεξεργαστεί από τους υπάρχοντες αλγόριθμους (Aggarwal 2015).

Καθαρισμός δεδομένων

Στην φάση που γίνεται ο καθαρισμός απομακρύνονται από τα δεδομένα είσοδοι που λείπουν, που είναι λανθασμένοι ή που είναι ασυνεχείς με τα δεδομένα. Επιπλέον κάποιες είσοδοι που λείπουνε μπορούν να εκτιμηθούν από μια διαδικασία που είναι γνωστή ως καταλογισμός (imputation).

Μείωση δεδομένων, επιλογή και μετατροπή

Σε αυτή την φάση το μέγεθος των δεδομένων μειώνεται μέσω της επιλογής υποσυνόλου δεδομένων, της επιλογής υποσυνόλου γνωρισμάτων ή της μετατροπής των δεδομένων. Τα κέρδη που αποκομίζονται σε αυτή την φάση είναι διπλά. Πρώτον όταν το μέγεθος των δεδομένων μειώνεται οι αλγόριθμοι είναι γενικά πιο αποδοτικοί. Δεύτερον αν άσχετα γνωρίσματα ή άσχετα αρχεία απομακρυνθούν η ποιότητα της διαδικασίας εξόρυξης γνώσης αυξάνεται σημαντικά. Ο πρώτος στόχος σύμφωνα με τον Aggarwal (2015) επιτυγχάνεται μέσω της γενικής δειγματοληψίας και των τεχνικών της μείωσης των διαστάσεων. Για να επιτευχθεί ο δεύτερος στόχος πρέπει να εφαρμοστεί μια προσέγγιση που είναι εξειδικευμένη πάνω στην επιλογή γνωρισμάτων. Για παράδειγμα μια προσέγγιση επιλογής γνωρισμάτων που αποδίδει καλά στην συσταδοποίηση ενδέχεται να μην δουλεύει το ίδιο καλά και για την κατηγοριοποίηση. Η έρευνα των Ramírez-Gallego, S. et al. (2017) αναφέρει ότι η μείωση δεδομένων (García et al. 2016) είναι ένα σημαντικό βήμα προεπεξεργασίας των δεδομένων εξόρυξης, καθώς θέτει ως στόχο την απόκτηση γρήγορων και προσαρμόσιμων μοντέλων μεγάλης ακρίβειας τα οποία θα χαρακτηρίζονται ταυτόχρονα από χαμηλή υπολογιστική πολυπλοκότητα προκειμένου να ανταποκρίνονται γρήγορα στα εισερχόμενα αντικείμενα και σε

αλλαγές. Επομένως, η δραστική μείωση της πολυπλοκότητας των εισερχόμενων δεδομένων είναι ζωτικής σημασίας για την απόκτηση τέτοιων μοντέλων. Επιπλέον, λόγω της παρουσίας του concept drift ο αριθμός και η συνάφεια των χαρακτηριστικών και των περιπτώσεων μπορεί να αλλάζει εν καιρώ. Αυτό πρέπει επίσης να ληφθεί υπόψη όσο διατηρείται και ενημερώνεται ένα online μοντέλο. Επιπροσθέτως, η έρευνα αναφέρει παρακάτω τους βασικούς τομείς της προεπεξεργασίας δεδομένων για την μείωση της πολυπλοκότητάς τους.

2.2.1 Μείωση διαστάσεων

Υπάρχει ευρύ φάσμα τεχνικών στη βιβλιογραφία που στοχεύουν στη μείωση του αριθμού των χαρακτηριστικών, μεταξύ άλλων: Επιλογή Χαρακτηριστικών (Feature Selection - FS), Εξαγωγή Χαρακτηριστικών (Feature Extraction - FE) ή προβολή διατήρησης της τοπικότητας (locality preserving projection) (Shikkenawis & Mitra 2015) (Al-Shiha, Woo & Dlay, 2014) (Zhang et al. 2012). Το FS (Doquire & Verleysen 2012) εξαλείφει τα άσχετα ή περιττά χαρακτηριστικά-στήλες, ενώ η Εξαγωγή Χαρακτηριστικών (FE) δημιουργεί έναν απλούστερο χαρακτηριστικό χώρο μέσω μετασχηματισμών του αρχικού. Ο στόχος εδώ είναι να δώσουμε ένα ελάχιστο σύνολο χαρακτηριστικών έτσι ώστε η πιθανότητα διαδοχικής κατανομής των τάξεων να παραμένει όσο το δυνατόν πιο αμετάβλητη. Όπως το FS διατηρεί τα αρχικά χαρακτηριστικά, είναι πιο βολικό για την «ερμηνεία του μοντέλου» (model interpretation). Ανάλογα με την σχέση μεταξύ του επιλογέα και του αλγορίθμου πρόβλεψης, μπορούμε να ταξινομήσουμε τους αλγόριθμους FS σε τρεις κατηγορίες:

- τα φίλτρα, που λειτουργούν πριν από τη διαδικασία εκμάθησης (και άρα είναι ανεξάρτητα από αυτήν)
- τα «περιβλήματα» (wrappers), τα οποία χρησιμοποιούν τον καθορισμένο αλγόριθμο μάθησης για την αξιολόγηση υποομάδων χαρακτηριστικών
- τα ενσωματωμένα, όπου η αναζήτηση αποτελεί μέρος της ίδιας της μαθησιακής διαδικασίας.

Οι μέθοδοι περιτύλιξης τείνουν να είναι πιο ακριβείς από τα φίλτρα, αλλά περισσότερο πολύπλοκες. Οι ενσωματωμένες μέθοδοι είναι λιγότερο δαπανηρές από τα wrappers, αλλά απαιτούν άμεσες τροποποιήσεις της διαδικασίας εκμάθησης.

2.2.2 Μείωση περιπτώσεων

Η “Επιλογή Περίστασης (Instance Selection - IS)” ή “Δημιουργία Παραγώγου (Instance Generation - IG)” (López et al. 2014) αποσκοπεί στη μείωση του αριθμού των εκπαιδευτικών περιπτώσεων επιλέγοντας τα πιο αντιπροσωπευτικά παραδείγματα. Οι μέθοδοι IG μπορούν να δημιουργήσουν νέες περιπτώσεις για να γεμίσουν τα κενά στον ορισμό των εννοιών. Το IS διαφέρει από τη δειγματοληψία δεδομένου ότι η πρώτη κατηγοριοποιεί τις περιπτώσεις ανάλογα με το πρόβλημα, ενώ η δειγματοληψία είναι περισσότερο στοχαστική. Με βάση το είδος αναζήτησης που εφαρμόζεται από τους αλγόριθμους IS, μπορούν να ταξινομηθούν σε τρεις κατηγορίες που είναι: Η συμπύκνωση, η οποία αφαιρεί πλεονάζοντα σημεία μακριά από τα «σύνορα» (borders). Η έκδοση, η οποία αφαιρεί τα θορυβώδη σημεία κοντά στα όρια της κλάσης και το υβριδικό, το οποίο συνδυάζει τόσο τον θόρυβο όσο και την διαγραφή των περίσσιων σημείων.

2.2.3 Απλοποίηση χώρου λειτουργίας

Η «Διακριτοποίηση» (Ferreira & Figueiredo 2014) συνοψίζει ένα σύνολο συνεχών τιμών σε ένα πεπερασμένο σύνολο διακριτών διαστημάτων. Αυτή η διαδικασία επιστρέφει ονομαστικά χαρακτηριστικά που μπορούν να χρησιμοποιηθούν από οποιαδήποτε διαδικασία εξόρυξης. Αν και οι περισσότεροι από τους αλγόριθμους εξόρυξης λειτουργούν με συνεχή δεδομένα, πολλοί από αυτούς μπορούν να αντιμετωπίσουν μόνο ονομαστικά χαρακτηριστικά, ιδίως εκείνων που βασίζονται σε στατιστικά και ενημερωτικά μέτρα (π.χ. Naïve Bayes (NB)) (Lu et al. 2006). Άλλοι αλγόριθμοι, όπως οι κατηγοριοποιητές που βασίζονται στα δέντρα (Hu, Chen & Tang, 2009), δημιουργούν πιο ακριβή και

συμπαγή αποτελέσματα όταν χρησιμοποιούνται διακριτές τιμές. Οι καλοί διακριτικοί προσπαθούν να επιτύχουν την καλύτερη απόδοση πρόβλεψης που προέρχεται από διακριτά δεδομένα, ενώ παράλληλα μειώνουν όσο γίνεται το μεσοδιάστημα (Cano, Luna et al. 2016; Cano, Nguyen et al. 2016). Η έρευνα των (Ramírez-Gallego et al. 2017) διακρίνει δύο κύριες κατηγορίες, με βάση το πώς τα διαστήματα δημιουργούνται από τους διακριτοποιητές: Τις μεθόδους διαίρεσης, οι οποίες χωρίζουν το πιο υποσχόμενο διάστημα της κάθε επανάληψης σε δύο διαμερίσματα και τις μεθόδους συγχώνευσης, οι οποίες συγχωνεύουν τα δύο καλύτερα παρακείμενα διαστήματα σε κάθε επανάληψη.

Συνοψίζοντας η διαδικασία προεπεξεργασίας των δεδομένων είναι μια από τις πιο σημαντικές φάσεις μέσα στην διαδικασία της αναζήτησης γνώσης από δεδομένα. Παρόλο που είναι λιγότερο γνωστή από άλλα κυρίως μεταγενέστερα βήματα όπως η εξόρυξη γνώσης μπορεί πολύ συχνά να απαιτεί περισσότερη προσπάθεια μέσα στο πλαίσιο της ανάλυσης δεδομένων από οποιοδήποτε άλλη φάση.

Επί του παρόντος το πλήθος των παραγόμενων δεδομένων αυξάνει εκθετικά ακολουθώντας την άνοδο του φαινομένου που ονομάζεται Μεγάλα δεδομένα. Τα σύγχρονα σύνολα δεδομένων αυξάνουν σε τρεις διαστάσεις που αφορούν τα γνωρίσματα, το πλήθος των παραδειγμάτων και την πληθικότητα κάνοντας την μείωση της πολυπλοκότητας ένα αναγκαίο βήμα αν επρόκειτο να χρησιμοποιηθούν οι κλασικοί αλγόριθμοι. Οι τεχνικές μείωσης δεδομένων εκτελούν αυτή την απλούστευση επιλέγοντας και διαγράφοντας πλεονάζοντα και θορυβώδη γνωρίσματα και στιγμιότυπα ή διακριτοποιώντας σύνθετους συνεχόμενους χώρους των γνωρισμάτων. Με αυτό τον τρόπο διατηρείται η αρχική δομή και το νόημα των δεδομένων εισόδου αλλά ταυτόχρονα λαμβάνουμε και ένα πολύ πιο διαχειρίσιμο όγκο πληροφορίας. Τέλος η ταχύτερη εκπαίδευση, οι βελτιωμένες ικανότητες γενίκευσης καθώς και η βελτιωμένη κατανόηση και η ερμηνεία των αποτελεσμάτων είναι μερικά από τα επιπλέον κέρδη που λαμβάνουμε με την μείωση του όγκου των δεδομένων.

Με την άνοδο των Μεγάλων δεδομένων έρχεται όχι μόνο η αύξηση του όγκου τους αλλά και η ταχύτητα με την οποία αυτά καταφθάνουν. Σε πολλά προβλήματα που αναδύονται στον πραγματικό κόσμο δεν μπορούμε να κάνουμε την υπόθεση ότι θα έχουμε να διαχειριστούμε ένα στατικό σύνολο από

στιγμιότυπα. Αντιθέτως υπάρχει το ενδεχόμενο αυτά να καταφθάνουν συνεχόμενα οδηγώντας έτσι σε ένα εν δυνάμει ολοένα και αυξανόμενο σύνολο δεδομένων όπως αναφέραμε ότι συμβαίνει στα δεδομένα ροής όπου εκεί τα δεδομένα καταφθάνουν με μεγάλη ταχύτητα είτε ένα προς ένα είτε ως ολόκληρες παρτίδες. Η ανάγκη της συνεχόμενης αναβάθμισης του αλγορίθμου από τα δεδομένα που καταφθάνουν, η επεξεργασία αυτών μέσα σε αυστηρά χρονικά περιθώρια καθώς και η περιορισμένη μνήμη που είναι διαθέσιμη καθιστούν αδύνατη την ευθύ εφαρμογή των τεχνικών μείωσης δεδομένων που εφαρμόζονται στα στατικά δεδομένα καθώς οι περισσότερες από αυτές τις τεχνικές θεωρούν ότι ολόκληρο το σύνολο εκπαίδευσης είναι διαθέσιμο εξαρχής και οι ιδιότητες των δεδομένων δεν αλλάζουν με την πάροδο του χρόνου.

Παρόλο που η μείωση των δεδομένων είναι πολύ σημαντική στο τομέα των δεδομένων ροής δεν υπάρχουν πολλές προτάσεις στο συγκεκριμένο πεδίο σύμφωνα με την υπάρχουσα βιβλιογραφία. Οι περισσότερες μέθοδοι μείωσης δεδομένων είναι απλά αυξητικοί αλγόριθμοι που αρχικά κατασκευάστηκαν για να διαχειρίζονται πεπερασμένα σύνολα δεδομένων. Όπως προαναφέρθηκε η άμεση προσαρμογή των υπαρχόντων τεχνικών μείωσης δεδομένων στα δεδομένα ροής δεν μπορεί να υλοποιηθεί για τους εξής λόγους :

- Οι περισσότεροι αλγόριθμοι επιλογής στιγμιοτύπων στα στατικά δεδομένα απαιτούν πολλαπλά περάσματα των δεδομένων. Επιπλέον βασίζονται κυρίως σε αναζητήσεις των εγγύτερων γειτόνων κάτι που είναι χρονοβόρο. Επομένως οι τεχνικές αυτές είναι ανίκανες να χειριστούν υψηλής ταχύτητας δεδομένα ροής.

- Αντίθετα οι τεχνικές επιλογής γνωρισμάτων είναι εύκολα προσαρμόσιμες σε online σενάρια παρόλα αυτά παρουσιάζουν άλλα προβλήματα όπως την εξέλιξη του concept ή το δυναμικό (Masud et al. 2010) ή εναλλασσόμενο (Barddal et al. 2016) χώρο των γνωρισμάτων.

- Οι online επιβλεπόμενες μέθοδοι διακριτοποίησης παραμένουν ανεξερεύνητες. Οι περισσότερες από τις καθιερωμένες μεθόδους απαιτούν

πολλαπλές επαναλήψεις με έντονες προσαρμογές πριν εξάγουν μια πλήρως λειτουργική λύση σύμφωνα με τον Webb (2014)

2.3 Αλγόριθμοι Μείωσης Δεδομένων στις Ροές Δεδομένων

Σε σενάρια ροών, οι τεχνικές μείωσης απαιτούνται κατά προτίμηση να επεξεργάζονται στοιχεία online ή σε κατάσταση batch όσο γρήγορα γίνεται και χωρίς προηγούμενες υποθέσεις σχετικά με τη διανομή των δεδομένων.

2.3.1 Αλγόριθμοι Μείωσης Διαστημάτων

Πολλοί αλγόριθμοι FS για ροές δεδομένων έχουν προταθεί βιβλιογραφικά. Οι περισσότεροι από αυτούς είναι φυσικά «βαθμιαίοι» (incremental) αλγόριθμοι σχεδιασμένοι για επεξεργασία εκτός σύνδεσης (García et al. 2015), ενώ άλλοι είναι ειδικά σχεδιασμένοι να αντιμετωπίζουν τα δεδομένα ροής (Bolón-Canedo et al. 2015). Όλες οι μέθοδοι FS μπορούν να κατηγοριοποιηθούν σε τρεις ομάδες: φίλτρα, wrappers και υβρίδια. Σύμφωνα με το πότε προέκυψε η επιλογή: πριν και ανεξάρτητα από το βήμα μάθησης ή στενά συνδεδεμένο με αυτό.

Οι περισσότεροι από τους online επιλογείς που προτείνονται στη βιβλιογραφία είναι «σταδιακές» (incremental) προσαρμογές των φίλτρων χωρίς σύνδεση. Δεδομένου ότι αυτά τα φίλτρα βασίζονται σε αθροιστικές λειτουργίες (κυρίως βασισμένες σε πληροφορίες ή σε στατιστικά μέτρα), αυτά προσαρμόζονται εύκολα στο online περιβάλλον. Παρά το γεγονός ότι είναι απλά, τα ηλεκτρονικά φίλτρα φαίνεται να προσαρμόζονται καλά στα drifts και δεν χρειάζεται να χωνεύουν όλα τα δεδομένα ταυτόχρονα, όπως οι αντίστοιχες offline συσκευές. Επιπλέον, οι διαδικτυακές μέθοδοι συνήθως αντιμετωπίζουν προβλήματα από ροές που δεν μπορούν να αντιμετωπιστούν με μεθόδους εκτός σύνδεσης, όπως την άφιξη νέων χαρακτηριστικών ή κλάσεων.

Εστιάζοντας στο online FS, μπορούν να γίνουν και άλλες διακρίσεις ανάλογα με τις ιδιότητες των ροών. Ορισμένες μέθοδοι FS υποθέτουν ότι τα χαρακτηριστικά φτάνουν ένα-προς-ένα (χαρακτηριστικά ροής) όσο τα χαρακτηριστικά διανύσματα διατίθενται αρχικά (Wu et al. 2010) (Eskandari & Javidi 2016), ενώ άλλοι υποθέτουν ότι οι περιπτώσεις καταφθάνουν πάντα διαδοχικά και το σύνολο χαρακτηριστικών μπορεί να είναι υποκείμενο πιθανών αλλαγών (online FS) (Katakis, Tsoumakas & Vlahavas, 2005).

Μπορούν επίσης να προκύψουν και νέες τάξεις από ροές χωρίς προηγούμενη γνώση («εξέλιξη της ιδέας» (concept evolution)), απαιτώντας έναν πλήρη επαναπροσδιορισμό του χρησιμοποιούμενου μοντέλου. Στην εξόρυξη δεδομένων ροής, ο χαρακτηριστικός χώρος μπορεί επίσης να επηρεαστεί από αλλαγές στην διανομή των δεδομένων. Τα feature drifts εμφανίζονται κάθε φορά που η συνάφεια ενός δεδομένου μεταβάλλεται με την πάροδο του χρόνου, κάθε φορά που καταφθάνουν νέες περιπτώσεις στο σύστημα (Barddal et al. 2017). Όπως συμβαίνει και με άλλα concept drift, οι αλλαγές στη συνάφεια επιβάλλουν αλγόριθμους για την απόρριψη ή την προσαρμογή του μοντέλου που ήδη γνωρίζει, αφαιρώντας τα πιο άσχετα χαρακτηριστικά στο νέο σενάριο (Nguyen et al. 2012), όπως καθώς και τα πιο σχετικά (δυναμική FS). Όσο οι αλλαγές στην συνάφεια επηρεάζουν άμεσα τα «αποφασιστικά όρια» (decision boundaries), το feature drift μπορεί να θεωρηθεί ως ένας συγκεκριμένος τύπος πραγματικού concept drift.

Καθώς το σύνολο των επιλεγμένων χαρακτηριστικών εξελίσσεται με την πάροδο του χρόνου, είναι πιθανό πως ο χαρακτηριστικός χώρος σε δοκιμαστικές περιπτώσεις θα διαφέρει από την τρέχουσα επιλογή. Επομένως, όταν μια νέα περίπτωση κατηγοριοποιείται, χρειαζόμαστε να πραγματοποιήσουμε μία μετατροπή μεταξύ των χαρακτηριστικών χώρων για σκοπούς ομογενοποίησης (Masud et al. 2010). Οι τύποι μετατροπής που πρέπει να ληφθούν υπόψη είναι οι εξής:

- Lossy Fixed (Lossy-F): το ίδιο σύνολο χαρακτηριστικών χρησιμοποιείται για ολόκληρο το ρεύμα. Παράγεται από την πρώτη παρτίδα. Όλες οι ακόλουθες περιπτώσεις (εκπαίδευση και δοκιμή) θα αντιστοιχιστούν σε αυτό το σετ, με αποτέλεσμα την σαφή απώλεια μελλοντικών πληροφοριών.

- Lossy Local (Lossy-L): ένας διαφορετικός χαρακτηριστικός χώρος χρησιμοποιείται για κάθε μια παρτίδα εκπαίδευσης. Οι περιπτώσεις δοκιμής χαρτογραφούνται έτσι στον εκπαιδευτικό χώρο σε κάθε επανάληψη. Αυτή η μετατροπή είναι επίσης προβληματική διότι τα σχετικά χαρακτηριστικά της δοκιμής μπορεί να παραλειφθούν.
- Lossless Homogenizing (Lossless): Το Lossless είναι παρόμοιο με την προηγούμενη μετατροπή, με την διαφορά ότι ο χώρος χαρακτηριστικών στο σύνολο δοκιμών λαμβάνεται υπόψη εδώ. Υπάρχει ομοιογένεια μεταξύ χώρων, για παράδειγμα, με την ενοποίηση και των δύο διαστημάτων και την γέμιση μηδενικών σε οποιοδήποτε στοιχείο που μπορεί να λείπει στο άλλο σετ. Αυτή η μετατροπή έχει ως αποτέλεσμα τη χρήση όλων των τρεχουσών και προηγούμενων πληροφοριών. Υπό αυτήν την έννοια, μπορεί να θεωρηθεί και η προτιμότερη επιλογή.

Στα παρακάτω, η εργασία των (Ramírez-Gallego et al. 2017) επικεντρώνεται σε διαδικτυακές τεχνικές που επιτρέπουν την άφιξη νέων περιπτώσεων και λειτουργιών ταυτόχρονα, επειδή αντιπροσωπεύουν ένα σενάριο που υφίσταται σε προβλήματα του πραγματικού κόσμου. Στην συνέχεια αναφέρονται οι πιο συναφείς αλγόριθμοι σε αυτό το θέμα.

Το έργο των Katakis, Tsoumakas και Vlahavas (2005) εισαγάγει το πρόβλημα του δυναμικού χώρου χαρακτηριστικών σε ροές δεδομένων. Προτείνεται μία τεχνική που περιλαμβάνει μια μέθοδο ταξινόμησης χαρακτηριστικών (φίλτρο) για την επιλογή σχετικών χαρακτηριστικών. Καθώς η βαθμολογία σημαντικότητας για κάθε χαρακτηριστικό μπορεί να μετρηθεί χρησιμοποιώντας πολλές σωρευτικές λειτουργίες όπως η «Απόκτηση Πληροφορίας» (Information Gain - IG), χ^2 ή αμοιβαίες πληροφορίες, γίνεται να θεωρηθούν ως μια ευέλικτη λύση για την online κατάταξη χαρακτηριστικών σύμφωνα με την έρευνα.

Η έρευνα των (Carvalho & Cohen 2006) προτείνει την επιλογή «Προηγμένη Επιλογή Χαρακτηριστικών» (Extremal Feature Selection - EFS), μια online μέθοδο FS που χρησιμοποιεί τα βάρη που υπολογίζονται από έναν online κατηγοριοποιητή (Τροποποιημένη Ισορροπημένη Απόρριψη -Modified Balanced Window) για τη μέτρηση της συνάφειας των χαρακτηριστικών. Η βαθμολογία

υπολογίζεται ως η απόλυτη διαφορά μεταξύ των θετικών και των αρνητικών βαρών για κάθε χαρακτηριστικό.

Η έρευνα των (Masud et al. 2010) προσφέρει μια τεχνική ταξινόμησης συνεχούς ροής (DXMiner), η οποία χρησιμοποιεί το μέτρο βάρους απόκλισης για την κατάταξη χαρακτηριστικών κατά τη διάρκεια της φάσης ταξινόμησης. Επιπλέον, αναφέρεται ότι το DXMiner αντιμετωπίζει ομαλά το πρόβλημα των νέων τάξεων (concept evolution), δημιουργώντας ένα όριο απόφασης γύρω από τα δεδομένα εκπαίδευσης. Σε αντίθεση με τις προηγούμενες μεθόδους, το DXMiner χρησιμοποιεί μετατροπή δίχως απώλειες, η οποία είναι χρήσιμη για την ανίχνευση καινοτομίας. Για την ταξινόμηση χαρακτηριστικών στον χώρο δοκιμών, το DXMiner χρησιμοποιεί μία τεχνική χωρίς επιτήρηση (π.χ. την υψηλότερη συχνότητα στην παρτίδα) που επιλέγει χαρακτηριστικά πιο αντιπροσωπευτικά για τις εισερχόμενες έννοιες. Ας σημειωθεί ότι αυτό απαιτεί μία «κατά-παρτίδα ρύθμιση» (batch-mode) για τον υπολογισμό αυτών των στατιστικών στοιχείων.

Η έρευνα των Nguyen et al. (2012) σχεδίασε μια τεχνική συνόλου βασισμένη στο windowing για την ανίχνευση χαρακτηριστικών drifts. Ο αλγόριθμος βασίζεται σε ένα σύνολο ταξινομητών, όπου κάθε κατηγοριοποιητής έχει το δικό του σύνολο χαρακτηριστικών. Εάν ανιχνευτεί drift, το σύνολο ενημερώνεται μαζί με νέο κατηγοριοποιητή μαζί με ένα νέο υποσύνολο χαρακτηριστικών. Σε διαφορετική περίπτωση, κάθε κατηγοριοποιητής ενημερώνεται αναλόγως. Το Φίλτρο Βασισμένο στην Ταχεία Συσχέτιση (Fast Correlation-Based Filter FCBF) με βάση τη συμμετρική αβεβαιότητα είναι αυτό που αξιοποιείται εδώ. Το FCBF εφαρμόζει με ευεργετικό τρόπο μια τεχνική προς τα πίσω με διαδοχική στρατηγική αναζήτησης για την κατάργηση άσχετων και περιττών χαρακτηριστικών.

Η έρευνα των Gomes, et al. (2013) έχει ως πρόταση έναν αλγόριθμο για την εξόρυξη επαναλαμβανόμενων εννοιών. Εδώ υιοθετούν την ίδια λύση που προτείνεται στο (Katakis, Tsoumakas & Vlahavas, 2005). Ωστόσο, αντί να επιλέξουν έναν σταθερό αριθμό χαρακτηριστικών, προτείνουν να χρησιμοποιήσουν είτε ένα σταθερό όριο είτε ένα προσαρμοστικό ανάλογα με τα ποσοστά. Συγκρίνουν επίσης τα αποτελέσματα της χρήσης διαφορετικών μετατροπών χώρου (όπως Lossy-F, Lossy-L ή Lossless) (Masud et al. 2010).

Η έρευνα των (Wu et al. 2012) πρότεινε δύο προσεγγίσεις για τον χειρισμό των ρευμάτων και την αύξηση των όγκων χαρακτηριστικών με την πάροδο του χρόνου. Η μία ονομάζεται Online Επιλογή Λειτουργιών Ροής (Online Streaming Feature Selection - OSFS). Η Δεύτερη είναι η Γρήγορη Online Ροή Χαρακτηριστικών Επιλογών (Fast Online Streaming Feature Selection - Fast-OSFS). Βασίζονται σε δύο φάσεις ενός βέλτιστου συστήματος εντοπισμού υποσυνόλου: online ανάλυση της συνάφειας και στη συνέχεια πλεονασμός. Η συνάφεια που βασίζεται στην κλάση χρησιμοποιείται για την επιλογή ή απόρριψη μιας νέας λειτουργίας. Στη συνέχεια, ένα νέο και εκτεταμένο σύνολο χαρακτηριστικών αναλύεται για να ανιχνεύσει εάν υπάρχει ένα υποσύνολο χαρακτηριστικών το οποίο μπορεί να κάνει μία από τις χρησιμοποιούμενες λειτουργίες και την κλάση μεταβλητών υπό όρους ανεξάρτητες. Αν ναι, τότε ένα τέτοιο χαρακτηριστικό απορρίπτεται. Αυτό επιτρέπει τον έλεγχο της επέκτασης του χαρακτηριστικού χώρου. Στο Fast-OSFS η ανάλυση πλεονασμού χωρίζεται σε δύο μέρη. Πρώτον, ελέγχονται οι πλεονασμοί των νέων χαρακτηριστικών για να αποφασιστεί αν πρέπει να επιλεχθεί αυτή η δυνατότητα.
Μόνο αν το νέο χαρακτηριστικό συμπεριλήφθηκε, ο πλεονασμός των προηγούμενων χαρακτηριστικών αναλύεται. Αυτό οδηγεί σε μια σημαντική υπολογιστική επιτάχυνση αυτής της μεθόδου.

Στα έργα (Wang et al. 2014) (Wang et al. 2015) συστήνεται μια «αχόρταγη» online μέθοδος FS (που αποκαλείται OFS - Online FS) βασισμένη σε μια κλασσική τεχνική που κάνει μια ανταλλαγή μεταξύ της εξερεύνησης και της εκμετάλλευσης χαρακτηριστικών. Ο αλγόριθμος δαπανά ε προσεγγίσεις στην εξερεύνηση με τυχαία επιλογή των χαρακτηριστικών *N* από το σύνολο των χαρακτηριστικών, και τα υπόλοιπα βήματα για την εκμετάλλευση διαλέγοντας τα χαρακτηριστικά *N* για τα οποία ο γραμμικός κατηγοριοποιητής έχει μη μηδενικές τιμές. Σε αυτό το έργο, ούτε ένα χαρακτηριστικό drift δεν απευθύνεται ρητά ενώ παράλληλα δεν υπάρχει σύγκριση με κανένα προηγούμενο έργο.

Μια μέθοδος επιλογής διαδικτυακών χαρακτηριστικών βασισμένη στην «ανάλυση δομής ομάδας» (group structure analysis) προτάθηκε από την εργασία των (Wang et al. 2015). Αυτή η εργασία βασίστηκε στην υπόθεση ότι τα χαρακτηριστικά μπορεί να φτάσουν σε συγκεκριμένες ομάδες, όπως οι υφές, τα χρώματα κ.λ.π. Οι συγγραφείς πρότειναν τον αλγόριθμο Ομαδικής Επιλογής

Χαρακτηριστικών Online (Online Feature Group Selection - OFGS) που χρησιμοποίησε κριτήρια εντός της ομάδας και μεταξύ ομάδων. Το προηγούμενο κριτήριο χρησιμοποίησε φασματική ανάλυση για να επιλέξει διακριτικά χαρακτηριστικά σε κάθε ομάδα. Το τελευταίο εφάρμοσε παλινδρόμηση για να επιλέξει ένα βέλτιστο υποσύνολο από όλα τα προεπιλεγμένα χαρακτηριστικά. Αξίζει να σημειωθεί ότι ένα παρόμοιο πρόβλημα συζητήθηκε στο έργο των Li, Yang και Zhang (2015).

Εκτός από τους πιο σχετικούς αλγορίθμους που αναφέρθηκαν προηγουμένως υπάρχουν στη βιβλιογραφία αρκετές άλλες προτάσεις online και επιλογής χαρακτηριστικών ροής. Η έρευνα των (Ramírez-Gallego et al. 2017) τις αναφέρει παρακάτω εν συντομία.

Το έργο των (Yan et al. 2006) πρότεινε ταυτόχρονη εξαγωγή και επιλογή χαρακτηριστικών χρησιμοποιώντας αλγόριθμο ορθογώνιου κεντροειδούς. Η έρευνα των Tadeuchi et al. (2007) πρότεινε μια γρήγορη online επιλογή χαρακτηριστικών, η οποία θα χρησιμοποιεί φίλτρα για να δημιουργήσει διάφορα δυναμικά υποσύνολα και ένα wrapper για να επιλέξει το καλύτερο όλων αυτών. Οι συντάκτες εικάζουν ότι αυτή η λύση πρέπει να βρίσκεται σε θέση να χειριστεί την εμφάνιση ενός concept drift. Η έρευνα των (Cai et al. 2009) προτείνει την κανονικοποίηση με την χρήση του κανόνα l1 (l1-norm) για συνεχόμενη επιλογή μεταβλητών. Παρόμοια προσέγγιση χρησιμοποιήθηκε στην έρευνα των (Ooi & Ninomiya 2013) ωστόσο χρησιμοποιήθηκε μια κανονική παλινδρόμηση για αυτό το καθήκον. Οι έρευνες των (Fan & Bougouila 2012) (Fan & Bougouila 2014) παρουσίασαν ένα συνδυασμό ομαδοποίησης με βάση ένα μείγμα διεργασίας Dirichlet γενικευμένων κατανομών Dirichlet και της αυτόματης επιλογής επιτήρησης χωρίς εποπτεία σε σενάρια «σταδιακής» (incremental) εκμάθησης.

Το έργο των (Amayri & Bougouila 2013) προτείνει έναν παρόμοιο συνδυασμό ομαδικής ανακάλυψης και τη μείωση χαρακτηριστικών χρησιμοποιώντας πεπερασμένα μίγματα von Mises, ενώ η έρευνα των (Yao & Liu 2013) συνδυάζει την online επιλογή με την εκτίμηση της πυκνότητας. Ένα πρόβλημα της online επιλογής χαρακτηριστικών για τη μάθηση πολλαπλών εργασιών συζητήθηκε στο έργο των (Yang, Lyu & King 2013). Το ζήτημα της επεκτασιμότητας της εξεταζόμενης οικογένειας μοντέλων για μεγάλη εξόρυξη δεδομένων εξετάστηκε στην έρευνα των (Yu et al. 2014). Η έρευνα του (Roy 2015) εμβάθυνε στο πως

γίνεται να χρησιμοποιηθεί «σύνολο» (ensemble) νευρώνων του Kohonen για την επιλογή χαρακτηριστικών από ροές υψηλής διάστασης. Πρόσφατα, η έρευνα των (Yang et al. 2016) εισήγαγε μια παράλληλη μέθοδο χρησιμοποιώντας περιορισμένη μνήμη, ενώ η έρευνα των (Hammoodi, Stahl & Tennant, 2016) ανέφερε μια έννοια προσέγγισης ενός concept drift με την χρήση μόνο συγκεκριμένων χαρακτηριστικών. Παράταση της μεθόδου του OSFS χρησιμοποιώντας μία στο περίπου προσέγγιση των ροών δεδομένων γίνεται στην έρευνα των (Eskandari & Javidi 2016), ενώ ένας συνδυασμός διαδικτυακής διακριτοποίησης με χαρακτηριστική επιλογή για νευρικά δίκτυα απεικονίστηκε στο έργο των (Bolón-Canedo et al. 2016).

Ορισμένοι βλέπουν μία ακολουθία βίντεο ως μια ροή εικόνων. Ως αναλογία, στον συγκεκριμένο τομέα διερευνήθηκε η online δυνατότητα χειρισμού δεδομένων ώστε να μπορεί να αντέχεται η δυναμική εντόπιση αντικειμένων (dynamic object detection). Η έρευνα των (Yeh & Hsu 2009) σύστησε μια online εξαγωγή δεδομένων βασισμένη στο Boosting, όπου επιλέχθηκαν νέες λειτουργίες μία τη φορά για να αντισταθμιστούν οι αλλαγές στο παρασκήνιο. Η έρευνα των (Yang et al. 2016) περιέγραψε έναν ηλεκτρονικό μηχανισμό επιλογής χαρακτηριστικών διάκρισης Fisher για οπτική παρακολούθηση σε πραγματικό χρόνο. Τέλος, αξίζει να αναφερθεί το έργο των Yu, Ding και Wu (2016) όπου οι συγγραφείς υλοποίησαν πολλές δημοφιλείς μεθόδους επιλογής χαρακτηριστικών σε απευθείας σύνδεση και δημιούργησαν ένα πακέτο κώδικα ανοιχτού λογισμικού για το Matlab.

Εκτός από το FS, η μείωση των διαστάσεων μπορεί να επιτευχθεί μέσω μιας τεχνητής χαρτογράφησης μεταξύ του αρχικού χώρου των χαρακτηριστικών και ενός νέου χώρου με λιγότερες διαστάσεις. Οι τεχνικές εξαγωγών χαρακτηριστικών, αν και λιγότερο δημοφιλείς από αυτές των FS, έχουν δείξει την ικανότητά τους σε πολλά προγνωστικά προβλήματα. Μία από τις σημαντικότερες συνεισφορές εδώ είναι η Ανάλυση Κύριων Στοιχείων (Principal Component Analysis - PCA) από την έρευνα των (Jolliffe & Cadima 2016). Στην έρευνα των (Nie, Kotlowski & Warmuth, 2016), μελετώνται δύο εκδόσεις PCA εις βάθος. Ο στόχος της προηγούμενης εργασίας είναι η απόκτηση ενός ηλεκτρονικού μοντέλου με την μικρότερη διαφορά στις αθροιστικές ζημίες σε σχέση με τις καλύτερες offline εναλλακτικές λύσεις.

Μια νέα ανάλυση των θεωρητικών ιδιοτήτων του PCA ροής του Oja συζητήθηκε στην έρευνα των (Jain et al. 2016). Αν και είναι βέλτιστη, το διαδικτυακό PCA δεν είναι σε θέση να ενημερώσει τις προβολές σε λιγότερο από O (n^3) ανά επανάληψη (Hazan, Kale & Warmuth, 2010). Επομένως, στο μέλλον πρέπει να αναπτυχθούν αποδοτικότερες τεχνικές εάν θέλουμε μια πραγματική λύση ροής στην εξόρυξη χαρακτηριστικών. Μέχρι στιγμής αξίζει να αναφέρουμε τις εκδόσεις streaming του πυρήνα PCA που προτάθηκε από τους (Joseph, Tokumoto & Ozawa, 2016) και από τους (Ghashami, Perry & Phillips, 2016). Επιπλέον, το PCA εφαρμόστηκε επιτυχώς για την εντόπιση του concept drift σε μη στατικές ροές δεδομένων από τις (Kuncheva & Faithfull 2012), καθώς και από τους (Qahtan et al. 2015). Παρατηρείται επίσης ότι η εξαγωγή χαρακτηριστικών από τις ροές δεδομένων δεν είναι μόνο περιορισμένη για PCA και ότι άλλα έργα, αν και λίγα σε αριθμό, υπάρχουν. Οι (Allahyar & Yazdi 2014) περιέγραψαν τη διαδικτυακή ανάλυση διακριτικών στοιχείων για τον συνεχή υπολογισμό της ανάλυσης γραμμικών διακρίσεων. Η εργασία των (Sheikholeslami, Berberidis & Giannakis, 2015) προτείνει την εξαγωγή χαρακτηριστικών βάσει πυρήνα για ορυχεία ροής με περιορισμένους υπολογιστικούς πόρους. Η εργασία των (Li et al. 2015) εισήγαγε την κανονική ανάλυση συσχέτισης με την αβεβαιότητα που είναι κατάλληλη για την ταξινόμηση των ροών δεδομένων πολλών όψεων.

2.3.2 Αλγόριθμοι Μείωσης Δειγμάτων

Το lazy learning έχει χρησιμοποιηθεί ευρέως στην προβλεπτική αναλυτική (Cover & Hart 1967). Ωστόσο, οι βάσεις συμβάντων χειροτερεύουν φυσικά και μεγαλώνουν σε μέγεθος μακροπρόθεσμα. Σε ένα σενάριο ροής δεδομένων, τα διατηρούμενα παρελθοντικά συμβάντα που ανήκουν σε κάποια προηγούμενη έννοια μπορεί να υποβαθμίσουν την απόδοση της μάθησης εάν εμφανιστεί μια νέα έννοια. Ομοίως, νέες περιπτώσεις που αναγνωρίζονται ως μία νέα έννοια μπορεί να χαρακτηριστούν ως θόρυβος και να αφαιρεθούν από ένα σφάλμα του μηχανισμού IS, διότι διαφωνούν με παρελθοντικές έννοιες (Lu et al. 2016)

Μια ορισμένη «έκδοση» (edition) και «συμπύκνωση» (condensation) (García et al. 2015) θα πρέπει να λαμβάνει μέρος με βάση τις υποθέσεις των εκλεπτυσμένων διαδικασιών του IS, οι οποίες επιλέγουν εκείνες τις περιπτώσεις που αντικατοπτρίζουν καλύτερα την τρέχουσα κατάσταση της ροής δεδομένων. Ωστόσο, το μεγαλύτερο μέρος των παροντικών τεχνικών είναι σχεδιασμένες για στάσιμα περιβάλλοντα και αγνοούν το concept drift φαινόμενο. Η έρευνα των (Ramírez-Gallego et al. 2017) παρουσιάζει παρακάτω ένα υποσύνολο IS τεχνικών που επιλέγουν υποθέσεις από μια βάση αυξητικά ή κατά παρτίδες (Garcia et al. 2012)

Ο αλγόριθμος IB3 (Instance-Based Algorithm 3) (Aha, Kibler & Albert, 1991) είναι μία από τις πρώτες προσπάθειες αντιμετώπισης της μη στατικής φύσης των δεδομένων. Είναι βασισμένος στην ακρίβεια και τα μέτρα συχνότητας ανάκτησης. Μέσω μιας δοκιμής διαστήματος εμπιστοσύνης, το IB3 αποφασίζει εάν πρέπει μία υπόθεση να προστεθεί στην περίπτωση βάσης ή εάν πρέπει να περιμένει μέχρι η να επισημανθεί η ένθεσή του ως κατάλληλη. Η αφαίρεση των περιπτώσεων γίνεται όταν η ακρίβεια μιας περίπτωσης είναι κατώτερη (σε κάποιο βαθμό) από την συχνότητα της κλάσης της. Με τον IB3 κατά νου, αναβάλλεται η συμπερίληψη των παραδειγμάτων. Αυτό θα ήταν κατάλληλο μόνο για Βαθμιαίο concept drift.

Ο αλγόριθμος που βασίζεται στην Τεχνική Τοπικής Διαγραφής (Local Weighted Forgetting - LWF) (Salganicoff 1993) είναι μία τεχνική στάθμισης που βασίζεται στους k-πλησιέστερους γείτονες (kNN). Στον LWF, οι περιπτώσεις με βάρος κάτω ενός ορίου διαγράφονται. Ο αλγόριθμος LWF έχει επικριθεί λόγω της χαμηλότερης ασυμπτωτικής ταξινόμησης στα στατικά περιβάλλοντα και από το τις τάσεις του για overfitting (Klinkenberg 2004). Αυτή η μέθοδος έχει δείξει καλές επιδόσεις τόσο για τις σταδιακές όσο και για τις αιφνίδιες μετακινήσεις ιδεών.

Ο Salganicoff (1997) σχεδίασε τον αλγόριθμο «Πρόβλεψης Λάθους στα Πλαίσια Αλλαγών» (Prediction Error Context Switching - PECS), ο οποίος έχει σχεδιαστεί για να λειτουργεί τόσο σε δυναμικά όσο και σε στατικά περιβάλλοντα. Ο αλγόριθμος PECS βασίζεται στα ίδια μέτρα που χρησιμοποιεί ο IB3, υιοθετώντας επίσης την ίδια δοκιμή εμπιστοσύνης. Προκειμένου να εισαχθεί η χρονική διάσταση στις αποφάσεις του, το PECS θεωρεί μόνο τις πιο πρόσφατες προβλέψεις στους υπολογισμούς του. Επιπλέον το PECS προσθέτει αμέσως

νέες περιπτώσεις στη βάση για την επιτάχυνση της αργής διαδικασίας προσαρμογής. Το PECS απενεργοποιεί τις περιπτώσεις αντί να τις αφαιρεί μονίμως. Αυτές οι περιπτώσεις μπορούν να εισαχθούν εκ νέου εάν η συμβολή τους μπορεί να συμβάλει και πάλι στη βελτίωση της ακρίβειας. Η έρευνα των (Beringer & Hüllermeier 2007) διαφωνεί ότι το PECS διατηρεί υψηλές απαιτήσεις μνήμης και μια αργή διαδικασία αφαίρεσης, καθώς οι νέες περιπτώσεις διατηρούνται αμέσως μετά την άφιξή τους.

Ο αλγόριθμος Επαναληπτικού Φιλτραρίσματος Περιπτώσεων (Iterative Case Filtering - ICF) (Brighton & Mellish 2002) είναι μία τεχνική αφαίρεσης πλεονάσματος που διαγράφει τα σύνολα περιπτώσεων απόδοσης που είναι μικρότερα σε μέγεθος από το σύνολο προσβασιμότητας. Οι συγγραφείς συμπεριέλαβαν το επαναλαμβανόμενο Edited-NN (Tomek 1976) για την αφαίρεση του θορύβου γύρω από τα «σύνορα» (borders).

Παρόλο που υπάρχουν πιο σύνθετες προτάσεις στην βιβλιογραφία (Garcia et al. 2012) ο προηγούμενος κατάλογος περιλαμβάνει τις μεθόδους που έχουν αποτελέσει τα θεμέλια για περαιτέρω εξελίξεις όσον αφορά το concept drift στο IS (Lu et al. 2016). Η επόμενη λίστα ασχολείται με τις τεχνικές που ασχολούνται άμεσα και μόνο με το concept drift:

Οι (Delany et al. 2005) προτείνουν ένα μηχανισμό ελέγχου μετακίνησης με δύο επίπεδα, που ονομάζεται επεξεργασία βάσει ικανοτήτων (Competence-Based Editing - CBE). Στο πρώτο επίπεδο, ένα υβρίδιο δύο CBE μεθόδων: Απαλλαγή του Θορύβου Βασισμένο στην Ευθύνη (Blame Based Noise Removal - BBNR) και Συντηρητική Μείωση Πλεονάσματος (Conservative Redundancy Reduction - CRR). Το BBNR αποσκοπεί στη διαγραφή των περιπτώσεων των οποίων η απομάκρυνση δεν συνεπάγεται απώλεια κάλυψης, ενώ το CRR επιλέγει εσφαλμένες περιπτώσεις που έχουν τη μικρότερη κάλυψη. Να σημειωθεί ότι και οι δύο μέθοδοι έχουν σχεδιαστεί για στάσιμα περιβάλλοντα, κάτι που μπορεί να προκαλέσει ορισμένα προβλήματα όπως: την αφαίρεση των νέων concept όταν εμφανίζεται gradual drift ή το ξέχασμα μικρών ομάδων των περιπτώσεων όπου τα παραδείγματα καλύπτουν το ένα το άλλο, αλλά κατατάσσουν εσφαλμένα όλους τους γύρω γείτονες.

Το BBNR δεν διατηρεί το μοντέλο των ικανοτήτων της ενημερωμένο, μόνο ξαναχτίζει το μοντέλο στο δεύτερο επίπεδο. Ένα ξεπερασμένο μοντέλο ικανοτήτων μπορεί να οδηγήσει σε αντιφάσεις στην φάση αξιολόγησης, καθώς το μοντέλο δεν αντικατοπτρίζει με ακρίβεια την τρέχουσα έννοια.

Η Εκμάθηση Βασισμένη σε Υποδείγματα σε Ρευμάτων Δεδομένων (Instance-Based Learning on Data Streams - IBL-DS) (Beringer & Hüllermeier 2007) και τα IBLStreams (Shaker & Hüllermeier, 2012) παρουσιάζονται ως οι πρώτες λύσεις που θεωρούν τους παράγοντες χρόνου και χώρου για τον έλεγχο του σχήματος και του μεγέθους της «βάσης-υποθέσεων» (case-base). Και στους δύο αλγόριθμους, κάθε γείτονας εντός δοκιμαστικού εύρους αφαιρείται εάν η κατηγορία της νέας παρουσίας κυριαρχεί σε αυτό το εύρος.

Το IBL-DS εισάγει επίσης μια σαφή μέθοδο ανίχνευσης του drift αναπτυσσόμενη από τον Gama (Gama et al. 2004) που καθορίζει, με βάση το μέγεθος και τον χρόνο, πότε πρέπει να αφαιρείται ένας καθορισμένος αριθμός περιπτώσεων.

Ο αριθμός των αφαιρέσεων υπολογίζεται λαμβάνοντας υπόψη το ελάχιστο ποσοστό σφαλμάτων και το συνολικό σφάλμα των τελευταίων προβλέψεων. Και οι δύο αλγόριθμοι ελέγχουν το μέγεθος της «βάσης-υποθέσεων» αφαιρώντας τις παλαιότερες περιπτώσεις. Ωστόσο, η στρατηγική απομάκρυνσης βάσει χρόνου που έχουν εφαρμοστεί από αυτούς έχει επικριθεί επειδή κάποιες παλιές, αλλά ακόμα σχετικές περιπτώσεις μπορεί να εξαλειφθούν με αυτήν την διαδικασία.

Οι αλγόριθμοι FISH (Žliobaitė 2011) βασίζονται επίσης σε ένα συνδυασμό χρόνου και χώρου, όμως σε αυτή την περίπτωση, υπολογισμένες ως αποστάσεις. Η ιδέα πίσω από αυτούς τους αλγόριθμους είναι να επιλέγεται δυναμικά το πιο σχετικό παράδειγμα, το οποίο θα αξιοποιηθεί ως εκπαίδευση για το επόμενο μοντέλο. Προτείνονται τρεις διαφορετικές μορφές του FISH. Στο FISH1, το μέγεθος της εκπαίδευσης καθορίζεται στην αρχή. Το FISH2 επιλέγει το καλύτερο μέγεθος εκπαίδευσης. Το FISH3 επίσης σταθμίζει το χρόνο και το μέγεθος χρησιμοποιώντας ένα διαφορετικό βρόχο της διασταυρωμένης επικύρωσης. Το FISH2 θεωρείται ο ηγέτης της οικογένειας. Το FISH αντιπροσωπεύει μια χρονοβόρα επιλογή επειδή αποθηκεύει όλα τα παραδείγματα που βλέπουμε για να υπολογίσουμε τις αποστάσεις χώρου/ χρόνου.

Οι Zhao, Wang και Xu (2012) παρουσιάζουν έναν νέο πλησιέστερο αλγόριθμο για τη ροή δεδομένων, με βάση ένα τεχνητό ενδοκρινικό σύστημα που ονομάζεται AES (artificial endocrine system). Αυτό το σύστημα καταργεί την ανάγκη μιας πλήρους «βάσης υποθέσεων» (case-base), αντικαθιστώντας την με αναπαραστατικά κύτταρα, όπως και σε προηγούμενες εκδόσεις. Μια διαδικασία βασισμένη στην συμπύκνωση είναι επίσης ένα βασικό χαρακτηριστικό του AES. Ο αλγόριθμος διατηρεί μόνο τα οριακά K πρωτότυπα ή κύτταρα. Αυτά τα πρωτότυπα συνεχίζουν να κινούνται καθ 'όλη τη διαδικασία προκειμένου να προσαρμόσουν τους περιορισμούς των εννοιών στα εισερχόμενα drift.

Το COMPOSE (Dyer, Capo & Polikar, 2013) είναι ένα πλαίσιο βασισμένο σε γεωμετρία για ημι-εποπτευόμενη μάθηση και ενεργή μάθηση. Η ιδέα της COMPOSE είναι να δέχονται ετικέτες οι εισερχόμενες περιπτώσεις μέσω μιας ημι-εποπτευόμενης προσέγγισης. Στην συνέχεια, σκοπός είναι η δημιουργία και η επιλογή εκείνων των α-σχημάτων που βελτιστοποιούν καλύτερα το μοντέλο της τρέχουσας κατάστασης. Αυτή η επιλογή είναι, στην πραγματικότητα, μια διαδικασία διατήρησης που διατηρεί μόνο αυτούς τους αντιπροσώπους σχημάτων-πρωτοτύπων για την τωρινή κατάσταση. Το COMPOSE είναι κυρίως σχεδιασμένο για την αντιμετώπιση των gradual drift.

Το SimC (Mena-Torres & Aguilar-Ruiz 2014) στοχεύει στη δημιουργία ομάδων περιπτώσεων για κάθε κατηγορία έτσι ώστε η καθεμία να αντιπροσωπεύει μια διαφορετική περιοχή του χώρου. Τα θορυβώδη και τα παλιά παραδείγματα καταργούνται επιλέγοντας και απορρίπτοντας το λιγότερο σχετικό παράδειγμα στην παλαιότερη ομάδα. Καθώς εμφανίζεται το concept drift, ο αλγόριθμος δημιουργεί νέες ομάδες για να εκχωρήσει παραδείγματα που αντιπροσωπεύουν νέες έννοιες. Η συνάφεια στις ομάδες υπολογίζεται χρησιμοποιώντας τις αποστάσεις του χώρου και τις ηλικίες τους. Για μοναδικές περιπτώσεις, αξιοποιείται η ακρίβεια που χρησιμοποιεί τον «κοντινότερο κανόνα» (nearest rule).

Η έρευνα των (Lu et al. 2016) προτείνει μια τεχνική επεξεργασίας των case-base βασισμένη στη διατήρηση και την ενίσχυση (Smyth & Keane 1995). Η λύση αποτελείται από τρία στάδια: η πρώτη συγκρίνει την διανομή μεταξύ δύο παραθύρων προκειμένου να ανιχνευθεί εάν υπάρχει κάποιο drift ή όχι. Εκτός από την ανίχνευση του drift, αυτή η μέθοδος επίσης περιορίζει την περιοχή στην

οποία αλλάζει περισσότερο η κατανομή. Μετά από αυτό εφαρμόζεται η μέθοδος «Γρήγορης Αλλαγής Πλαισίου Βελτιωμένου Θορύβου» (Noise-Enhanced Fast Context Switch - NEFCS). Το NEFCS εξετάζει όλες τις νέες περιπτώσεις και καθορίζει εάν υπάρχει θόρυβος ή όχι (ενίσχυση). Ωστόσο, μόνο οι θορυβώδες περιπτώσεις που βρίσκονται εκτός των αναγνωρισμένων περιοχών αρμοδιότητας αφαιρούνται, επειδή μπορεί να αποτελούν μέρος καινοτόμων εννοιών. Η μέθοδος απομάκρυνσης πλεονασμού (Stepwise Redundancy Removal - SRR) στοχεύει στον έλεγχο του μεγέθους της θήκης (συντήρηση). Το SRR απομακρύνει περιττά παραδείγματα αναδρομικά έως ότου η κάλυψη των case-base αρχίσει να επιδεινώνεται.

Μπορούμε να σχεδιάσουμε τρεις μεγάλους τύπους επιλογής από τα προηγούμενα: βάσει ικανοτήτων, βάσει στάθμισης και βάσει της ακρίβειας. Οι μέθοδοι που βασίζονται σε ικανότητες (όπως CBE ή ICF) τείνουν να είναι πιο ακριβείς αλλά χρονοβόροι διότι απαιτούν συνεχείς ενημερώσεις του μοντέλου ικανότητας. Οι στρατηγικές επιλογής που βασίζονται στην απόσταση μπορεί να απαιτούν ακόμη περισσότερο χρόνο από τα μοντέλα που βασίζονται στην ικανότητα, όταν ο αριθμός των αποστάσεων ή / και των χαρακτηριστικών είναι μεγάλος. Οι μέθοδοι που βασίζονται στην ακρίβεια έχουν δυσκολίες στον εντοπισμό θορυβώδους παραδειγμάτων κατά τη διάρκεια μελλοντικών drifts. Τέλος, οι τεχνικές που βασίζονται στα χαρακτηριστικά τείνουν να υπερφορτώνουν τα δεδομένα και να είναι λιγότερο αποδοτικές από τους επιλογείς περιπτώσεων σύμφωνα με την έρευνα του (Klinkenberg 2004).

Ένα άλλο σχετικό θέμα που πρέπει να λαμβάνεται υπόψη κατά την εκλογή επιλογών περιπτώσεων είναι αν πρόκειται να εφαρμοστούν διεργασίες βελτίωσης ή / και συντήρησης ή όχι. Οι μέθοδοι που βασίζονται στις ικανότητες αποτελούνται συνήθως από δύο τεχνικές: μία για την απομάκρυνση του θορύβου και μία για το πλεόνασμα. Το πλεόνασμα συνήθως αγνοείται στις προαναφερόμενες τεχνικές καθώς οι περισσότερες από αυτές επιλέγουν περιπτώσεις σύμφωνα με τον εσφαλμένο αριθμό προβλέψεων που πράττεται από το καθένα. Αλγόριθμοι με βάση την απόσταση αφαιρούν απότομα τον πλεονασμό μέσω του συντελεστή χώρου στην φόρμουλα απόστασης.

2.3.3 Αλγόριθμοι Απλούστευσης Χαρακτηριστικών (διαστάσεων)

Οι αλγόριθμοι διακριτοποίησης για σενάρια ροής δεδομένων πρέπει επίσης να είναι σε θέση να χειριστούν την εμφάνιση των concept drifts. Ο ορισμός και ο αριθμός των «διαστημάτων» (intervals) διακριτοποίησης μπορεί να αλλάξει με την πάροδο του χρόνου, ως αποτέλεσμα της μετατόπισης των χαρακτηριστικών των δεδομένων. Ως εκ τούτου, είναι επιθυμητό τα διαστήματα διακριτοποίησης να είναι σε θέση να προσαρμόζονται ομαλά στην έννοια της μετατόπισης, χωρίς να επιβάλλεται αυξημένο υπολογιστικό κόστος όταν επανυπολογίζονται.

Η διακριτοποίηση ίσων συχνοτήτων (βάσει ιστογραμμάτων) μπορεί να θεωρηθεί ως μία από τις πρώτες τεχνικές αντιμετώπισης της σταδιακής διακριτοποίησης. Χρησιμοποιώντας ποσοτικά μεγέθη quantiles ως σημεία περικοπής, ο χαρακτηριστικός χώρος μπορεί να χωριστεί σε διαστήματα ίσων συχνοτήτων. Η εκτίμηση των ποσοτήτων σε ροές έχει μελετηθεί εις βάθος στην βιβλιογραφία σε προσεγγιστικές (Webb 2014) και ακριβείς (Gupta & Zane 2003) (Guha & McGregor 2009) μορφές. Μία από τις πιο εύχρηστες και αποτελεσματικές εναλλακτικές λύσεις διακριτοποίησης είναι ο Αλγόριθμος Σταδιακής Διακριτοποίησης (Incremental Discretization Algorithm - IDA) (Webb et al. 2014). Το IDA προσεγγίζει τα ποσοτικά μεγέθη μέσω της διατήρησης ενός δείγματος «δεξαμενής» (reservoir) της εισόδου του ρεύματος. Τα διαστήματα εδώ είναι δομημένα χρησιμοποιώντας σωρούς διαστημάτων, μία αποτελεσματική δομή δεδομένων που επιτρέπει την εισαγωγή και τη διαγραφή στοιχείων στο $O\left(log(n)\right)$ και την ανάκτηση του μέγιστου και του ελάχιστου (τα όρια διαστημάτων) σε συνεχή χρόνο. Όπως στις περισσότερες περιπτώσεις δεν είναι εφικτό να διατηρηθεί μια ολοκληρωτική καταγραφή όλων των δεδομένων, οι προσεγγιστικές λύσεις έχουν αποδειχθεί καταλληλότερες για την επεξεργασία ροών υψηλής απόδοσης από ότι οι ακριβείς λύσεις.

Άλλες τεχνικές που βασίζονται στη συχνότητα στηρίχθηκαν στη δημιουργία ορίων μεγέθους των bins. για να αντιμετωπίσουν την εξελισσόμενη διακριτοποίηση. Η έρευνα των (Lu, Zhang & Webb, 2006) παρουσίασε τον αλγόριθμο Σταδιακής Ευέλικτης Συχνότητας Διακριτοποίησης (Incremental

Flexible Frequency Discretization - IFFD). Το IFFD ορίζει ένα εύρος αντί για έναν αυστηρό αριθμό ποσοτικού μεγέθους. Αν φτάσει η συχνότητα των ενημερωμένων διαστημάτων στο μέγιστο και οι προκύπτουσες συχνότητες δεν είναι κάτω του ελάχιστου (για να αποφευχθεί μια υψηλή ταξινόμηση διακύμανσης), το IFFD διαιρεί το διάστημα σε δύο κατατμήσεις.

Ο διακριτικοποιητής ίσου πλάτους είναι μια άλλη μη επιτηρούμενη προσέγγιση που απαιτεί ως είσοδο μόνο το φάσμα χαρακτηριστικών και τον αριθμό των διαστημάτων διαίρεσης. Ωστόσο, το κύριο μειονέκτημα εδώ είναι ότι και οι δύο προσεγγίσεις απαιτούν ροή αρχείων που θα φτάνει σε τυχαία σειρά, κάτι που είναι αδύνατο σε πολλά προβλήματα μάθησης.

Μια άλλη σημαντική απαίτηση που πρέπει να ληφθεί υπόψη είναι ότι ορισμένοι «σταδιακοί» (incremental) αλγόριθμοι απαιτούν να διατηρηθεί το ίδιο σύνολο σημείων κοπής (αριθμός, δομή και νόημα) με την πάροδο του χρόνου (Webb 2014). Αυτή είναι η περίπτωση των πιο διακριτικών αλγορίθμων μάθησης. Εδώ προτείνεται η χρήση είτε ενός διακριτικού μέσου ίσου πλάτους είτε μιας ισοσταθμισμένης συχνότητας καθώς και τα δύο καθορίζουν τον αριθμό των bins εκ των προτέρων. Άλλοι στατικοί αλγόριθμοι (π.χ.: NB) δεν απαιτούν την συντήρηση των διαστημάτων κατά τη διάρκεια των επακόλουθων προβλεπτικών φάσεων, αλλά μόνο για εξοικονόμηση ορισμένων στατιστικών για το τρέχον βήμα διακριτοποίησης. Ωστόσο, οι ικανότητες εξακρίβωσης τέτοιων ταξινομητών εξακολουθούν να επηρεάζονται από τέτοιου είδους μετατοπίσεις σε ορισμούς, ειδικά αν είναι έντονες.

Σύμφωνα με τους Gama et al. (2004), ένα από τα κύρια προβλήματα των ανεπιτήρητων διακριτοποιητών είναι η ανάγκη ορισμού του αριθμού των διαστημάτων προκαταβολικά. Μια τέτοια απόφαση μπορεί να υποστηριχθεί από ορισμένους προκαθορισμένους κανόνες (π.χ., κανόνας Sturges) ή από διερευνητική διαδικασία ανάλυσης. Όμως, η διερευνητική ανάλυση δεν είναι πλέον δυνατή στους καιρούς μας όπου ο αριθμός των περιπτώσεων είναι πολύ μεγάλος και οι προκαθορισμένοι κανόνες έχουν δείξει ότι λειτουργούν μόνο με σύνολα δεδομένων μικρού μεγέθους. Ωστόσο, οι ανεπιτήρητοι διακριτοποιητές είναι σχεδιασμένοι για φυσικά περιβάλλοντα ροών εφόσον ο αριθμός των διαστημάτων παραμένει αμετάβλητος.

Οι περισσότερες εποπτευόμενες προσεγγίσεις τείνουν να πραγματοποιούν αρκετές συγχωνεύσεις και διαιρέσεις πριν αποκτήσουν κάποιο τελικό λειτουργικό σχέδιο. Απότομες αλλαγές στα χρονικά διαστήματα μπορεί να επηρεάσουν αρνητικά την online διαδικασία εκμάθησης. Ως εκ τούτου, οι μέθοδοι θα έπρεπε να επιδιώκουν την επίτευξη μεταβάσεων. Η έρευνα των (Ramírez-Gallego et al. 2017) παρουσιάζει έναν σύντομο κατάλογο επιτηρημένων προσεγγίσεων διακριτοποίησης.

Η έρευνα των (Gama et al. 2004) παρουσίασε τον αλγόριθμο Διαίρεσης Σταδιακής Διακριτοποίησης (Partition Incremental Discretization - PiD), που αποτελείται από δύο στρώματα. Το πρώτο συνοψίζει τα δεδομένα και δημιουργεί τα προκαταρκτικά χρονικά διαστήματα, τα οποία θα βελτιστοποιηθούν στο επόμενο στρώμα. Μία «επιπέδου πλάτους» (equal-width) στρατηγική μπορεί να χρησιμοποιηθεί για την προετοιμασία αυτού του βήματος. Στη συνέχεια, το πρώτο στρώμα ενημερώνεται με μια διαδικασία διαίρεσης όποτε ο αριθμός των στοιχείων ενός διαστήματος είναι μεγαλύτερος του προκαθορισμένου ορίου. Το δεύτερο στρώμα εκτελεί μια διαδικασία συγχώνευσης με την προηγούμενη φάση προκειμένου να αποδώσει το τελικό σχήμα διακριτοποίησης. Οποιοσδήποτε διακριτοποιητής μπορεί να χρησιμοποιηθεί στο δεύτερο στρώμα, δεδομένου ότι τα διαστήματα που δημιουργούνται στην προηγούμενη φάση χρησιμοποιούνται ως είσοδοι. Ο «Διακριτοποιητής Ελάχιστου Μήκους Περιγραφής» (Minimum Description Length Discretizer) χρησιμοποιείται ως αναφορά στο αρχικό χαρτί. Ωστόσο, υπάρχουν τρεις κύριοι λόγοι για την κριτική της προσέγγισης PiD. Πρώτον, δεν υπάρχει ακριβής αντιστοιχία μεταξύ του πρώτου και του δεύτερου στρώματος, το οποίο δημιουργεί ανακρίβειες που θα αλυσιδωθούν και θα αυξηθούν με την πάροδο του χρόνου. Δεύτερον, αν η διανομή των δεδομένων είναι εξαιρετικά λοξή, ο αριθμός των διαστημάτων που παράγονται θα αυξάνεται δραματικά, λόγω της υπερχείλισης της συχνότητας. Τέλος, η διαδικασία διαίρεσης μπορεί να γίνει ακόμη πιο ανακριβής αν εμφανιστούν πολλές επαναλήψεις μίας μόνο τιμής. Στην περίπτωση αυτή ένα τέτοιο σημείο κοπής μπορεί να δημιουργηθεί που χωρίζει περιπτώσεις με τις ίδιες τιμές χαρακτηριστικών σε δύο διαφορετικούς κάδους, οδηγώντας σε ασυνέπειες.

Στο Lehtinen, Saarela και Elomaa (2012) παρουσιάζεται μια ηλεκτρονική έκδοση του ChiMerge (OC), η οποία διατηρεί την $O\left(nlog(n)\right)$ πολυπλοκότητα χρόνου

που κατέχει ο αρχικός αλγόριθμος. Προκειμένου να διασφαλιστούν τα ίδια αποτελέσματα διακριτοποίησης, οι συγγραφείς εφαρμόζουν μια ηλεκτρονική προσέγγιση που βασίζεται σε «κυλιόμενα παράθυρα» (sliding windows). Χρησιμοποιούνται πολλές δομές δεδομένων για την εξομοίωση της ίδιας συμπεριφοράς που κατέχει η αρχική έκδοση. Πέρα της μεγάλης αποτελεσματικότητας που ισχυρίζονται οι συγγραφείς, αυτό που επιδεικνύεται από αυτήν την ηλεκτρονική έκδοση είναι μια μεγάλη αύξηση στην χρήση της μνήμης που προέρχεται από το σύνολο δομών δεδομένων. Το γεγονός αυτό μπορεί να είναι ανασταλτικός παράγοντας όσον αφορά κάποια σενάρια ροών δεδομένων με περιορισμένους υπολογιστικούς πόρους.

2.4 Επιλογή στιγμιοτύπων (instance selection)

Η επιλογή στιγμιοτύπων (instance selection) παίζει καίριο ρόλο στην προσπάθεια μείωσης των δεδομένων λόγω του γεγονότος ότι εκτελεί την συμπληρωματική διαδικασία της επιλογής γνωρισμάτων (feature selection). Παρόλο που είναι δυο ξεχωριστές διαδικασίες στις περισσότερες περιπτώσεις εφαρμόζονται από κοινού. Η αντιμετώπιση των τεράστιων όγκων των δεδομένων μπορεί να επιτευχθεί περιορίζοντας τα δεδομένα ως μια εναλλακτική για να αυξηθεί η αποτελεσματικότητα των αλγορίθμων εξόρυξης γνώσης. Η διαδικασία της επιλογής γνωρισμάτων επιτυγχάνει αυτό τον στόχο μέσω της μετακίνησης και διαγραφής άσχετων ή αχρείαστων χαρακτηριστικών. Με αυτό το σκεπτικό η μετακίνηση στιγμιοτύπων μπορεί να θεωρηθεί το ίδιο ή και περισσότερο ενδιαφέρουσα διαδικασία που έχει ως στόχο τον περιορισμό των δεδομένων σε συγκεκριμένες εφαρμογές (Liu & Motoda 2002).

Το σημαντικό πρόβλημα για τον περιορισμό των δεδομένων είναι η επιλογή και η αναγνώριση των πιο σχετικών δεδομένων από μια τεράστια πηγή στιγμιοτύπων και η προετοιμασία αυτών για να εισαχθούν σε έναν αλγόριθμο εξόρυξης γνώσης. Η επιλογή είναι συνώνυμο με την πίεση σε πολλά σενάρια όπως στους οργανισμούς, στις επιχειρήσεις ή στην φυσική εξέλιξη (Liu & Motoda 2002). Λαμβάνεται ως μια πραγματική ανάγκη στον κόσμο επομένως και στην

διαδικασία εξόρυξης γνώσης. Όπως είναι ευρέως γνωστό τα δεδομένα δεν είναι καθαρά όταν προσλαμβάνονται και κατά συνέπεια δεν είναι έτοιμα για να χρησιμοποιηθούν για εξόρυξη καθότι απουσιάζουν δεδομένα, υπάρχουν πολλά περιττά δεδομένα και σφάλματα προκύπτουν κατά την διαδικασία συλλογής και αποθήκευσης τους. Συνεπώς τα δεδομένα μπορεί να είναι τουλάχιστον κατά την αρχική τους μορφή αδύνατα να διαχειριστούν σωστά από μηχανισμούς εξόρυξης γνώσης.

Η επιλογή στιγμιοτύπων στοχεύει στο να επιλέξει ένα υποσύνολο από το σύνολο των δεδομένων για να επιτύχει τον αρχικό σκοπό μιας μεθόδου εξόρυξης γνώσης σαν να χρησιμοποιούσε όλο το σύνολο (Derrac, Garcia & Herrera, 2010). Ωστόσο σύμφωνα με τους Garcia, Luengo και Herera (2015) η μείωση των δεδομένων μέσω της επιλογής ενός υποσυνόλου δεδομένων δεν ανήκει πάντα στην κατηγορία της επιλογής στιγμιοτύπων. Συγκεκριμένα αντιστοιχούν την επιλογή στιγμιοτύπων με μια ευφυή διαδικασία κατηγοριοποίησης των στιγμιοτύπων σύμφωνα με το πόσο σχετικά είναι τα στιγμιότυπα ή πόσο θόρυβο έχουν και λαμβάνοντας πάντα υπόψη την διαδικασία εξόρυξης γνώσης που πρόκειται να εφαρμοστεί. Με αυτό τον τρόπο για παράδειγμα δεν θεωρούν την δειγματοληψία δεδομένων ως μια διαδικασία επιλογής στιγμιοτύπων διότι έχει μια πιο γενική χρήση και ο κύριος στόχος είναι η μείωση των δεδομένων με τυχαίο τρόπο για να ενισχυθεί η διαδικασία εκμάθησης. Παρόλα αυτά η δειγματοληψία δεδομένων (Domingo, Gavalda & Watanabe, 2000) ανήκει στην οικογένεια των μεθόδων μείωσης δεδομένων.

Το βέλτιστο αποτέλεσμα της επιλογής στιγμιοτύπων είναι ένα ελάχιστο υποσύνολο δεδομένων ανεξάρτητο του μοντέλου το οποίο μπορεί να επιτύχει το ίδιο έργο χωρίς να υπάρχει απώλεια στην επίδοση. Επομένως αν P είναι η απόδοση τότε πρέπει $P(DM_s) = P(DM_t)$, όπου DM είναι ο αλγόριθμος εξόρυξης γνώσης, s είναι ένα υποσύνολο στιγμιοτύπων που έχουν επιλεχθεί και t είναι ολόκληρο το σύνολο των στιγμιοτύπων εκπαίδευσης. Σύμφωνα με τους Liu και Motoda (2002) η επιλογή στιγμιοτύπων έχει τις ακόλουθες εξαιρετικές λειτουργίες :

- **ενεργοποίηση** : Η επιλογή στιγμιοτύπων κάνει το αδύνατο δυνατό δεδομένου ότι όταν ένα σύνολο δεδομένων είναι πολύ μεγάλο ένας αλγόριθμος εξόρυξης γνώσης ενδεχομένως να μην μπορεί να λειτουργήσει ή να μην μπορεί να αποδώσει αποτελεσματικά. Η διαδικασία της επιλογής στιγμιοτύπων όμως επιτρέπει στον αλγόριθμο να λειτουργήσει με πολλά δεδομένα.
- **εστίαση** : Τα δεδομένα σχηματίζονται από πολλές πληροφορίες από σχεδόν τα πάντα που αφορούν ένα πεδίο αλλά μια συμπαγής μέθοδος εξόρυξης γνώσης επικεντρώνεται μόνο σε ένα σημείο ενδιαφέροντος από όλο το πεδίο. Η διαδικασία της επιλογής στιγμιοτύπων επικεντρώνει τα δεδομένα στο συγκεκριμένο αυτό σημείο όπου επιθυμούμε να αντληθεί γνώση.
- **καθαρισμός** : Με την επιλογή σχετικών στιγμιοτύπων, περιττά καθώς και θορυβώδη στιγμιότυπα συνήθως απομακρύνονται βελτιώνοντας με αυτό τον τρόπο την ποιότητα των δεδομένων εισόδου και επομένως αναμένεται και η βελτίωση της απόδοσης του αλγορίθμου εξόρυξης γνώσης.

Στην παρούσα εργασία εστιάζουμε στην σημαντικότητα της επιλογής στιγμιοτύπων στην σύγχρονη εξόρυξη γνώσης και στα δεδομένα ροής καθώς είναι πολύ συχνό τα σύνολα δεδομένων να υπερβαίνουν το μέγεθος των δεδομένων που μπορούν να διαχειριστούν οι αλγόριθμοι. Επιπλέον ο τομέας της επιλογής στιγμιοτύπων προσελκύει όλο και περισσότερους ερευνητές που ασχολούνται με την μείωση δεδομένων. Η εμπειρία δείχνει ότι όταν ένας αλγόριθμος εφαρμόζεται σε ένα μειωμένο σύνολο δεδομένων επιτυγχάνει επαρκή και κατάλληλα αποτελέσματα αν επιλεχθεί η ορθή στρατηγική εξόρυξης γνώσης. Η εργασία αυτή προσανατολίζεται προς την διαδικασία της κατηγοριοποίησης συνεπώς επικεντρωνόμαστε σε μεθόδους που εφαρμόζουν την επιλογή στιγμιοτύπων και λαμβάνουν ένα υποσύνολο $S \subset T$ έτσι ώστε το σύνολο S να μην περιέχει αχρείαστα στιγμιότυπα και ταυτόχρονα να ισχύει ότι $Acc(S) \approx Acc(T)$ όπου $Acc(X)$ είναι η ακρίβεια της κατηγοριοποίησης που αποκτήθηκε χρησιμοποιώντας το X ως σύνολο εκπαίδευσης. Καθώς το σύνολο εκπαίδευσης μειώνεται, ο χρόνος που απαιτείται για την διαδικασία εκπαίδευσης μειώνεται και αυτός ταυτόχρονα ειδικά στις μεθόδους που είναι lazy learning ή instance-based

2.4.1 Επιλογή συνόλου εκπαίδευσης και επιλογή πρωτοτύπων

Αρχικά αρκετές προτάσεις είχαν γίνει για την επιλογή του πιο σχετικού συνόλου δεδομένων από το ευρύτερο σύνολο εκπαίδευσης με κυριότερο τον KNN αλγόριθμο. Αργότερα όταν ο όρος εκμάθηση βασισμένη σε στιγμιότυπα (Aha, Kibler & Albert, 1991) ή αλλιώς lazy learning πρωτοπαρουσιάστηκε για την συγκέντρωση όλων αυτών των μεθόδων που δεν εκτελούν μια φάση εκπαίδευσης κατά την εκμάθηση, ο όρος επιλογή πρωτότυπων προέκυψε από την βιβλιογραφία. Πλέον η οικογένεια των μεθόδων επιλογής στιγμιοτύπων περιλαμβάνει την επιλογή πρωτοτύπων που θεωρήθηκε ότι είναι λειτουργική και με άλλες μεθόδους εκμάθησης όπως τα δέντρα απόφασης ή τις μηχανές διανυσμάτων υποστήριξης. Ωστόσο δεν υπάρχει ισχυρή ένδειξη για να θεωρηθεί ότι η μέθοδος επιλογής στιγμιοτύπων είναι έγκυρη και μπορεί να εφαρμοστεί στον οποιοδήποτε αλγόριθμο εξόρυξης γνώσης μέσα στο ίδιο παράδειγμα εκπαίδευσης. Για τον λόγο αυτό οι Garcia, Luengo και Herrera (2015) διαχωρίζουν την επιλογή στιγμιοτύπων σε δύο κατηγορίες: την επιλογή πρωτοτύπων (PS) (Garcia et al. 2012) και επιλογή του συνόλου εκπαίδευσης (TS) (Cano, Herrera & Lozano, 2007).

Ορισμός instance selection (Garcia, Luengo & Herrera, 2015)

Έστω X_p είναι ένα στιγμιότυπο όπου $X_p = \left(X_{p1}, X_{p2}, \ldots, X_{pm}, X_{pc}\right)$, με το X_p να ανήκει σε μια κλάση c που δίνεται από το X_{pc} και ένας m-διάστατος χώρος στον οποίο το X_{pi} είναι η τιμή του i γνωρίσματος του p δείγματος. Έπειτα ας υποθέσουμε ότι υπάρχει ένα σύνολο εκπαίδευσης TR το οποίο αποτελείται από N στιγμιότυπα X_p και ένα σύνολο ελέγχου TS το οποίο αποτελείται από t στιγμιότυπα X_p. Έστω $S \subset TR$ είναι το υποσύνολο των επιλεγμένων δειγμάτων που προέκυψε από την εκτέλεση του αλγορίθμου επιλογής στιγμιοτύπων. Τότε κατηγοριοποιούμε ένα νέο σχέδιο από το σύνολο TS από έναν αλγόριθμο

εξόρυξης γνώσης που ενεργεί πάνω στο σύνολο S. Ολόκληρο το σύνολο ορίζεται να είναι το D και αποτελείται από την ένωση των TR και TS.

Σχετικά με τις PS μεθόδους αποτελούν μεθόδους επιλογής στιγμιοτύπων που αναμένεται να βρουν σύνολα εκπαίδευσης τέτοια ώστε να βελτιώσουν την ακρίβεια της κατηγοριοποίησης και να μειώσουν παράλληλα τον όγκο των δεδομένων αξιοποιώντας κατηγοριοποιητές που βασίζονται σε στιγμιότυπα με βάση κάποια ομοιότητα ή κάποια μετρική απόστασης. Πρόσφατα οι PS μέθοδοι έχουν αυξηθεί σε δημοτικότητα μέσα στο πεδίο της μείωσης δεδομένων. Διάφορες προσεγγίσεις για PS αλγόριθμους έχουν προταθεί στην βιβλιογραφία (Garcia et al. 2012).

Οι TSS μέθοδοι καθορίζονται με όμοιο τρόπο. Είναι γνωστές ως η εφαρμογή των μεθόδων επιλογής στιγμιοτύπων πάνω στο σύνολο εκπαίδευσης που χρησιμοποιείται για να κατασκευαστεί ένα προβλεπτικό μοντέλο. Επομένως οι TSS μέθοδοι μπορούν να αναπτυχθούν ως ένας τρόπος για να βελτιώσουν την συμπεριφορά των προβλεπτικών μοντέλων και συγκεκριμένα την ακρίβεια και την ερμηνευσιμότητα τους.

2.4.2 Ταξινομία

Οι μέθοδοι επιλογής στιγμιοτύπων συνήθως ταξινομούνται κάτω από τις εξής κατηγορίες: την κατεύθυνση της έρευνας, τον τύπο της επιλογής και την αξιολόγηση της έρευνας (Garcia et al.2012).

2.4.2.1 Κατεύθυνση της αναζήτησης

Η επιλογή στιγμιοτύπου μπορεί να θεωρηθεί σαν ένα πρόβλημα αναζήτησης. Με την χρήση μιας συγκεκριμένης μετρικής ο στόχος της μεθόδου αυτής είναι να βρει το πιο αντιπροσωπευτικό υποσύνολο στιγμιοτύπων για αυτή την μετρική (Cano, Herrera & Lozano, 2005b). Με βάση την κατεύθυνση της έρευνας η επιλογή στιγμιοτύπων σύμφωνα με τον Gonzalez (2018) μπορεί να χωριστεί σε πέντε ομάδες :

- **Αυξητικοί**: οι μέθοδοι αυτοί ξεκινάνε με ένα άδειο σύνολο δεδομένων και προσθέτουν στιγμιότυπα τα οποία ικανοποιούν ένα προκαθορισμένο κριτήριο. Το ελάττωμα τους είναι ότι είναι εξαιρετικά ευαίσθητες στην σειρά με την οποία εμφανίζονται τα στιγμιότυπα. Τα κύρια πλεονεκτήματα τους είναι ότι τα δεδομένα μπορούν να επεξεργαστούν άμα προέρχονται από ρεύμα δεδομένων, είναι ταχύτερες μέθοδοι και επιπλέον απαιτούν λιγότερο αποθηκευτικό χώρο.
- **Μειωτικοί** : Λειτουργούν κατά τον αντίθετο τρόπο από ότι η προαναφερόμενη ομάδα. Επομένως ξεκινάνε με ολόκληρο το σύνολο δεδομένων και απομακρύνουν στιγμιότυπα ακολουθώντας ένα προκαθορισμένο κριτήριο. Η σειρά με την οποία ελέγχονται τα στιγμιότυπα είναι και εδώ πολύ σημαντική αλλά όχι τόσο όσο στη προηγούμενη ομάδα. Το κύριο ελάττωμα είναι ότι ολόκληρο το σύνολο δεδομένων πρέπει να είναι διαθέσιμο στη μνήμη. Συνεπώς αυτές οι μέθοδοι δεν είναι κατάλληλες για δεδομένα ροής.
- **Παρτίδα (batch):** Τα στιγμιότυπα αναλύονται ανά παρτίδες επομένως προεπεξεργάζονται με επιτυχία και όσα είναι ακατάλληλα σημειώνονται για διαγραφή. Η διαγραφή όμως πραγματοποιείται μόνο προς το τέλος του αλγορίθμου. Το κύριο πλεονέκτημα τους είναι ότι διατηρούν μια συνολική εικόνα όλων των δεδομένων για κάθε χρονική στιγμή.
- **Μικτοί**: Μπορούν να χαρακτηριστούν ως μια μίξη μεταξύ των τριών προηγούμενων ομάδων. Ξεκινάνε με ένα προκαθορισμένο σύνολο και έπειτα στιγμιότυπα προστίθενται ή διαγράφονται σύμφωνα με ένα συγκεκριμένο κριτήριο.
- **Σταθεροί** : Αυτές οι μέθοδοι είναι μια υποκατηγορία των μικτών μεθόδων αλλά σε αυτή την περίπτωση ο τελικός αριθμός των στιγμιοτύπων είναι προκαθορισμένος ως παράμετρος εισόδου που μπαίνει από τον χρήστη.

2.4.2.2 Στρατηγική επιλογής

Το κλειδί αυτής της διαδικασίας είναι τα όρια κατηγοριοποίησης. Τα όρια απόφασης σχηματίζονται από στιγμιότυπα δύο ή και περισσότερων διαφορετικών κλάσεων οι οποίες είναι όμοιες. Κατά συνέπεια τα στιγμιότυπα

μπορεί να είναι είτε συνοριακά σημεία είτε κεντρικά σημεία (απομακρυσμένα από τα σύνορα) (Wilson & Martinez 1997). Τρείς ομάδες μεθόδων υπάρχουν στην στρατηγική επιλογής σύμφωνα με τον Gonzalez (2018) :

- **Αλγόριθμοι συμπύκνωσης:** Επιχειρούν να διατηρήσουν τα συνοριακά σημεία δηλαδή τα σημεία κοντά στα σύνορα απόφασης. Οι αλγόριθμοι αυτοί συνήθως επιτυγχάνουν μεγάλους ρυθμού μείωσης. Το κύριο μειονέκτημα τους όμως είναι ότι επηρεάζονται σημαντικά από σημεία που αποτελούν θόρυβο (Jankowski & Grochowski 2004).
- **Αλγόριθμοι συλλογής**: Οι αλγόριθμοι αυτοί εργάζονται κατά τον αντίθετο τρόπο από ότι οι αλγόριθμοι συμπύκνωσης αφαιρώντας τα συνοριακά σημεία. Διαγράφουν τα στιγμιότυπα τα οποία δεν είναι σε συμφωνία με τους εγγύτερους γείτονες τους. Επομένως δεν προσανατολίζονται στην μείωση των δεδομένων όσο στην μείωση των στιγμιοτύπων που αποτελούν θόρυβο. Ως συνέπεια ο όγκος μείωσης των δεδομένων από αυτούς τους αλγόριθμους είναι μικρότερος από ότι στους αλγόριθμους συμπύκνωσης.
- **Υβριδικοί αλγόριθμοι**: Οι υβριδικοί αλγόριθμοι όπως υποδηλώνει και το όνομα βρίσκονται μεταξύ των δυο ανωτέρω τεχνικών και ο στόχος τους είναι να βρούνε το μικρότερο και ταυτόχρονα το πιο ακριβές σύνολο παραδειγμάτων. Διαγράφουν και κεντρικά αλλά και συνοριακά στιγμιότυπα.

- **Αλγόριθμοι τάξης (rank methods):** Αυτοί οι αλγόριθμοι αποτελούν μια νέα προσέγγιση και δεν μπορούν να ενταχθούν σε καμία από τις παραπάνω κατηγορίες (Rico-Juan & Inesta 2012). Αυτό που κάνουν είναι να ταξινομούν τα στιγμιότυπα με βάση την σημαντικότητα τους δηλαδή με βάση την χρησιμότητα τους στην διαδικασία της κατηγοριοποίησης. Επομένως με αυτό τον τρόπο επιλέγεται ένα υποσύνολο με τα καλύτερα στιγμιότυπα (Valero-Mas et al. 2016).

2.4.2.3 Αναζήτηση αξιολόγησης

Οι μέθοδοι επιλογής στιγμιοτύπων μπορούν επίσης να ομαδοποιηθούν με βάση την στρατηγική που ακολουθούν για την επιλογή των στιγμιοτύπων. Σύμφωνα με τους Olvera-Lopez et al.(2009) οι μέθοδοι μπορούν να ομαδοποιηθούν σε δύο κατηγορίες οι οποίες είναι :

- **Μέθοδοι περιτύλιξης (wrapper)** : Η απόφαση για να επιλεχθεί ή να διαγραφεί ένα στιγμιότυπο αποκτάται από έναν κατηγοριοποιητή συνήθως έναν kNN.
- **Μέθοδοι φίλτρα (Filter)** : Η απόφαση για να επιλεχθεί ή να διαγραφεί ένα στιγμιότυπο λαμβάνεται με βάση κάποια ευρετική μέθοδο ή με κάποιον κανόνα και δεν βασίζεται σε κάποιον κατηγοριοποιητή.

2.4.3 Υπολογιστική πολυπλοκότητα

Ο στόχος των αλγορίθμων επιλογής στιγμιοτύπων είναι να μειώσουν το μέγεθος του συνόλου δεδομένων όπως ήδη έχει δηλωθεί. Η πολυπλοκότητα όμως των παραδοσιακών αλγορίθμων που χρησιμοποιούν αυτή την τεχνική είναι ένα από τα κυριότερα προβλήματα. Σύμφωνα με τους Jankowski και Grochowski (2004) η υπολογιστική πολυπλοκότητα της πλειοψηφίας αυτών των αλγορίθμων είναι τουλάχιστον log-linear. Ως συνέπεια μια πιθανή λύση για την διαχείριση συνεχώς αυξανόμενων συνόλων δεδομένων είναι η μείωση τους με την τεχνική της επιλογής στιγμιοτύπων. Δυστυχώς όμως οι μέθοδοι αυτοί επηρεάζονται από την υψηλή υπολογιστική πολυπλοκότητα που έχουν. Ως εκ τούτου κάποιες μέθοδοι προέκυψαν πρόσφατα με στόχο να αντιμετωπίσουν αυτό το πρόβλημα. Αυτές οι προσεγγίσεις αναφέρονται συνοπτικά παρακάτω.

2.4.4 Αυξητικές μέθοδοι (scaling-up approaches)

Κατά τα προηγούμενα χρόνια διάφορες προσεγγίσεις έγιναν με σκοπό να προσαρμοστούν οι υπάρχουσες μέθοδοι επιλογής στιγμιοτύπων για να αντιμετωπίζουν σύνολα δεδομένων με τεράστιο όγκο. Η πρώτη πρόταση ήταν η στρωματοποίηση που παρουσιάστηκε από τους Cano, Herrera και Lozano (2005a) για να ενισχύσει τις εξελισσόμενες μεθόδους επιλογής στιγμιοτύπων. Η ιδέα τους περιλαμβάνει των διαχωρισμό του αρχικού συνόλου δεδομένων σε υποσύνολα δεδομένων που θα είναι ξένα μεταξύ τους αλλά με την ίδια κατανομή κλάσης όπως το αρχικό σύνολο. Τα οφέλη του scaling-up μπορούν να συντονιστούν μέσω της διαμόρφωσης του μεγέθους του κάθε συνόλου. Επιπλέον η διαδικασία της στρωματοποίησης είναι κατάλληλη για την ενίσχυση οποιασδήποτε άλλης μεθόδου. Μια βελτιωμένη εκδοχή της παραπάνω μεθόδου παρουσιάστηκε από τους Haro-Garcia και Garcia-Pedrajas (2009). Η αρχή της διαδικασίας είναι η ίδια : ο διαχωρισμός ολόκληρου του συνόλου δεδομένων σε ανεξάρτητα υποσύνολα. Αφού η πρώτη παρτίδα συνόλων έχει επεξεργαστεί τα επιλεγμένα στιγμιότυπα από τον αλγόριθμο ενώνονται και η διαδικασία επαναλαμβάνεται εκ νέου.

Μια πιο καινοτόμα και ιδιαίτερη προσέγγιση κυρίως λόγο της επίδοσης της προτάθηκε από τους Garcia-Osorio et al. (2010). Η διαδικασία εκτελείται με έναν προκαθορισμένο αριθμό γύρων r. Σε κάθε γύρο μια διαίρεση χωρίζει το σύνολο δεδομένων σε ξένα υποσύνολα που ονομάζονται και κάδοι (bins). Όπως και στις προηγούμενες μεθόδους ένας αλγόριθμος επιλογής στιγμιοτύπων εφαρμόζεται σε κάθε κάδο. Ο αλγόριθμος ενημερώνει έναν πίνακα με ψήφους αυξάνοντας τους κατά ένα αν το στιγμιότυπο έχει επιλεχθεί. Αφού εκτελεστεί ένας προκαθορισμένος αριθμός γύρων ο πίνακας των ψήφων χρησιμοποιείται για να αποφασιστεί με την χρήση ενός κατωφλίου ποια στιγμιότυπα θα επιλεχθούν και ποια θα διαγραφούν.

Άλλες μέθοδοι όπως των Angiulli και Folino (2007) επικεντρώθηκαν στην ανάπτυξη μιας κατανεμημένης μεθόδου για τον υπολογισμό ενός συνέχους υποσυνόλου για πολύ μεγάλα σύνολα δεδομένων.

2.4.5 Σημαντικές μέθοδοι μείωσης στα στατικά δεδομένα

Στην βιβλιογραφία υπάρχει ένας μεγάλος αριθμός αλγορίθμων επιλογής στιγμιοτύπων ενώ αρκετοί προστίθενται κάθε χρόνο στην κατηγορία αυτών των μεθόδων. Οι πιο σημαντικές μέθοδοι επιλογής στιγμιοτύπων σύμφωνα με τους Garcia Luengo και Herrera (2016) είναι οι ακόλουθες: Ο συμπυκνωμένος εγγύτερος γείτονας CNN, ο αλλαγμένος εγγύτερος γείτονας ENN, η μέθοδος DROP (Decremental Reduction Optimization) και ο αλγόριθμος ICF (Iterative Case Filtering), ο οποίος θα παρουσιαστεί στην επόμενη παράγραφο.

2.4.5.1 Αλγόριθμος CNN

Ο αλγόριθμος του Hat (1968) με την ονομασία Condensed Nearest Neighbours CNN θεωρείται η πρώτη επίσημη πρόταση της τεχνικής επιλογής στιγμιοτύπων για τον εγγύτερο γείτονα. Η σκέψη της σταθερότητας όσον αφορά το σύνολο εκπαίδευσης είναι σημαντική σε αυτόν τον αλγόριθμο και καθορίζεται ως εξής : Δοθέντος ενός μη κενού συνόλου $X \neq \emptyset$ ένα υποσύνολο S του X ($S \subseteq X$) είναι σε συμφωνία όσον αφορά το X αν άμα χρησιμοποιηθεί το υποσύνολο S σαν σύνολο εκπαίδευσης ο κανόνας του εγγύτερου γείτονα μπορεί να ταξινομήσει σωστά όλα τα στιγμιότυπα του X. Ακολουθώντας τον ορισμό της συνέχειας αν θεωρήσουμε το σύνολο X σαν το σύνολο εκπαίδευσης τότε ένα συμπυκνωμένο υποσύνολο του θα πρέπει να έχει όλα τα χαρακτηριστικά του X και να είναι μικρότερο από αυτό.

Ο αλγόριθμος δείχνει την δομή της μεθόδου. Ξεκινά με ένα σύνολο δεδομένων S αρχικά μόνο με τυχαία επιλεγμένα στιγμιότυπα (εναλλακτικά το S θα μπορούσε να έχει ένα τυχαία επιλεγμένο στιγμιότυπο ανά κλάση). Έπειτα προσπαθεί να κατηγοριοποιήσει όλα τα στιγμιότυπα του X αξιοποιώντας αυτά του S σύμφωνα με τον κανόνα του εγγύτερου γείτονα. Αν αυτό είναι επιτυχές ο αλγόριθμος προχωρά με το επόμενο στιγμιότυπο διαφορετικά το λάθος κατηγοριοποιημένο στιγμιότυπο προστίθεται στο σύνολο S και η επιβεβαίωση

της σωστής κατηγοριοποίησης στο X ξεκινά από την αρχή. Τελικά θα τερματίσει επιστρέφοντας το S ως το επιλεγμένο σύνολο δεδομένων.

Αλγόριθμος : Condensed Nearest Neighbour CNN (Gonzalez 2018)

Είσοδος : σύνολο εκπαίδευσης $X = \{(x_1, y_1), \ldots, (x_n, y_n)\}$
Έξοδος : Το σύνολο των επιλεγμένων στιγμιοτύπων $S \subseteq X$

1: S = $\{x_1\}$
2: **foreach** $x \in X$ **do**
3: **if x** κατηγοριοποιείται λάθος με χρήση του S **then**
4: Πρόσθεσε το x στο S
5: Restart
6: **return** S

2.4.5.2 Αλγόριθμος ENN

Η πρώτη πρόταση για την αλλαγή του συνόλου δεδομένων παρουσιάστηκε από τον Wilson (1972) με την ονομασία Edited Nearest Neighbour (ENN). Είναι μια μέθοδος μείωσης επομένως ξεκινά με ολόκληρο το σύνολο X και κάθε στιγμιότυπο μετακινείται από το σύνολο αν δεν κατηγοριοποιείται ορθά από τον κανόνα του k εγγύτερου γείτονα. Ο αλγόριθμος είναι επίσης παραμετρικός διότι ο χρήσης πρέπει να θέσει την τιμή του k. Σύμφωνα με το αρχικό άρθρο το k παίρνει την τιμή 3.

Ο αλγόριθμος που ακολουθεί δείχνει τον ψευδοκώδικα του ENN. Η μέθοδος αυτή ουσιαστικά διαγράφει στιγμιότυπα τα οποία είναι θορυβώδη ή συνοριακά επιτυγχάνοντας έτσι πιο καθαρά σύνορα απόφασης. Επιπλέον τα κεντρικά στιγμιότυπα είναι ανεπηρέαστα από την διαδικασία αλλαγής. Ο στόχος αυτού του αλγόριθμου δεν είναι η μείωση του συνόλου δεδομένων αλλά η βελτίωση της ακρίβειας μέσω του επιλεγμένου υποσυνόλου. Ο αλγόριθμος αυτός

λόγω των ικανοτήτων του να καθαρίζει τα δεδομένα χρησιμοποιείται από πολλούς άλλους αλγόριθμους για φιλτράρισμα θορύβου.

Αλγόριθμος : Edited Nearest Neighbours ENN (Gonzalez 2018)

Είσοδος : σύνολο εκπαίδευσης $X = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, αριθμός k.
Έξοδος : Το σύνολο των επιλεγμένων στιγμιοτύπων $S \subseteq X$

1: S = X

2: **foreach** $x \in S$ **do**

3: **if** x κατηγοριοποιείται λάθος με χρήση του k-NN **then**

4: Μετακίνησε το x στο S

5: **return** S

2.4.5.3 Αλγόριθμος DROP

Η οικογένεια των DROP αλγορίθμων (Wilson & Martinez 2000) περιλαμβάνει μερικές από τις καλύτερες τεχνικές κατηγοριοποίησης που χρησιμοποιούν την επιλογή στιγμιοτύπων (Brighton & Mellish 2002). Το κριτήριο μέσω του οποίου μετακινούνται τα στιγμιότυπα βασίζεται σε δύο ιδέες. Στον συσχετισμό και στους εγγύτερους γείτονες. Η σχέση του συσχετισμού είναι αντίθετη από αυτή του εγγύτερου γείτονα. Ένα στιγμιότυπο p το οποίο έχει το q ως ένα από τους εγγύτερους γείτονες του ονομάζεται συνεργάτης του q. Το σύνολο των εγγύτερων γειτόνων ενός στιγμιοτύπου ονομάζεται γειτονία. Το σύνολο των συνεργατών για κάθε στιγμιότυπο είναι μια λίστα με όλα τα στιγμιότυπα τα οποία έχουν αυτό το συγκεκριμένο στιγμιότυπο στην γειτονιά τους.

Ο ψευδοκώδικας του DROP3 περιγράφεται παρακάτω. Ξεκινάει με ένα φίλτρο θορύβου όμοιο με του ENN και έπειτα από αυτό τα στιγμιότυπα ταξινομούνται με βάση την απόσταση τους από τον εγγύτερο εχθρό. Οι λίστες των εγγύτερων γειτόνων και των συνεργατών υπολογίζονται για κάθε στιγμιότυπο. Επειτα στον κύριο βρόγχο του αλγορίθμου για κάθε στιγμιότυπο x το σύνολο with περιέχει το πλήθος των συνεργατών του x οι οποίοι ταξινομούνται

σωστά όταν το x διατηρείται στο σύνολο δεδομένων ενώ το σύνολο without περιέχει το πλήθος των συνεργατών οι οποίοι ταξινομούνται σωστά όταν το x έχει απομακρυνθεί από το σύνολο δεδομένων. Αν το τελευταίο σύνολο είναι μεγαλύτερο ή ίσο από το προηγούμενο τότε το στιγμιότυπο x απομακρύνεται διότι η διαγραφή του δεν θα επηρεάσει την κατηγοριοποίηση των συνεργατών του. Σε περίπτωση που το x μετακινηθεί όλα τα στιγμιότυπα συνεργάτες θα πρέπει να ανανεώσουν την λίστα γειτόνων τους.

Αλγόριθμος: Decremental Reduction Optimization Procedure 3 DROP3 (Gonzalez 2018)

Είσοδος : σύνολο εκπαίδευσης $X = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, αριθμός k.
Έξοδος : Το σύνολο των επιλεγμένων στιγμιοτύπων $S \subseteq X$

1: Φίλτρο θορύβου : Μετακίνηση όλων των στιγμιοτύπων του X που έχουν
2: κατηγοριοποιηθεί λάθος από του k εγγύτερους γείτονες.
3: S = X
4: Ταξινόμηση του S σύμφωνα με την απόσταση από τον εγγύτερο εχθρό
5: **foreach** $x \in S$ **do**
6: Βρες τους k+1 εγγύτερους γείτονες του x στο S
7: Πρόσθεσε το x στη λίστα συνεργατών του κάθε γείτονα του
8: **foreach** $x \in S$ **do**
9: with = το σύνολο των συνεργατών του x που ταξινομήθηκαν σωστά
10: without = το σύνολο των συνεργατών του x που ταξινομήθηκαν σωστά χωρίς το x
11: **if** without ≥ with **then**
12: Διαγραφή του x από το S
13: **foreach** συνεργάτη **α** του **x do**
14: Απομάκρυνση του x από την λίστα εγγύτερων γειτόνων του α
15: Εύρεση ενός νέου εγγύτερου γείτονα για το α
16: Πρόσθεσε το α στη νέα λίστα συνεργατών των εγγύτερων γειτόνων του
17: **return** S

2.5 Μέθοδοι μείωσης με επιλογή στιγμιοτύπων στα δεδομένα ροής

Η επιλογή στιγμιοτύπων είναι μια από τις πιο αποτελεσματικές μεθόδους για να χειριστούμε δεδομένα με τεράστιο όγκο το οποίο σημαίνει ότι η εύρεση σχετικών δεδομένων ανάμεσα σε ένα τεράστιο πλήθος δεδομένων έχει ως αποτέλεσμα την μείωση των δεδομένων ή τον περιορισμό τους ενώ ταυτόχρονα διατηρείται η ουσία. Από την οπτική της εξόρυξης γνώσης ένα ιδανικό αποτέλεσμα άμα επιλεχθεί να εφαρμοστεί η επιλογή στιγμιοτύπων είναι το υποσύνολο των επιλεγμένων στιγμιοτύπων από το αρχικό σύνολο δεδομένων να προσφέρει τα ίδια αποτελέσματα ή συμπεράσματα σαν να χρησιμοποιήθηκε όλο το σύνολο. Η επιλογή στιγμιοτύπων δεν πρέπει να συγχέεται με την άλλη διάσημη τεχνική μείωσης δεδομένων που είναι η επιλογή χαρακτηριστικών η οποία αποσκοπεί στην επιλογή των πιο σχετικών γνωρισμάτων από όλο το σύνολο μειώνοντας κατά αυτό τον τρόπο τον αριθμό των διαστάσεων. Η επιλογή χαρακτηριστικών αποτελεί το επίκεντρο πολλών τεχνικών εξόρυξης γνώσης καθότι διευκολύνει την παραγωγή των δέντρων απόφασης, των κανόνων σύνδεσης και άλλων τεχνικών κατηγοριοποίησης (Galan, Liu & Torkkola, 2005). Από την άλλη η επιλογή στιγμιοτύπων επιχειρεί να μειώσει τα δεδομένα μειώνοντας τον αριθμό των παραδειγμάτων που ανήκουν στο αρχικό σύνολο δεδομένων. Και οι δύο τεχνικές τόσο η επιλογή στιγμιοτύπων όσο και η επιλογή χαρακτηριστικών αναζητούν στο να αποκτήσουν οφέλη και στις απαιτήσεις του χώρου αποθήκευσης αλλά και στην απόδοση του αλγορίθμου. Ωστόσο η επιλογή στιγμιοτύπων πολύ συχνά παραγκωνίζεται όπως φαίνεται από την βιβλιογραφία καθώς αρκετά περισσότερες έρευνες έχουν πραγματοποιηθεί πάνω στην επιλογή χαρακτηριστικών. Είναι γεγονός ότι η τελευταία τεχνική είναι η πλέον κατάλληλη για κάποιες εφαρμογές όπως για παράδειγμα η βιοπληροφορική (Galan, Liu & Torkkola, 2005) διότι ο αριθμός των γνωρισμάτων που απεικονίζεται από το M ξεπερνά κατά πολύ τον αριθμό των στιγμιοτύπων N, δηλαδή ισχύει ότι M >> N. Αυτό το πεδίο επομένως δεν μπορεί να κερδίσει πολλά από την επιλογή στιγμιοτύπων καθώς είναι συνηθισμένο να υπάρχουν κάτι λιγότερο από μερικές εκατοντάδες δείγματα τα οποία περιέχουν πληροφορία για χιλιάδες γονίδια που αποτελούν τα χαρακτηριστικά. Όμως το ακριβώς αντίθετο συμβαίνει για άλλες εφαρμογές όπως αυτές που περιλαμβάνουν δεδομένα ροής όπου υπάρχουν

πολλά οφέλη άμα χρησιμοποιηθεί η τεχνική μείωσης της επιλογής στιγμιοτύπων λόγο του ανεξάντλητου αριθμού παραδειγμάτων που παράγονται και που περνάνε κατά πολύ τον αριθμό τον γνωρισμάτων δηλαδή σε αυτή την περίπτωση ισχύει N >> M.

Κάποιες από τις μετρικές που μπορούν να χρησιμοποιηθούν για να μετρηθεί η απόδοση της τεχνικής της επιλογής στιγμιοτύπων είναι το επίπεδο συμπίεσης και η αναπαραγωγισιμότητα (Galan, Liu & Torkkola, 2005). Ορίζουμε τα δεδομένα ροής S και R ως συνεχόμενες ακολουθίες στοιχείων έτσι ώστε $S = (s_1, s_2, \ldots, s_n, \ldots)$ και $R = (r_1, r_2, \ldots, r_k, \ldots)$ όπου οι δείκτες 1,2,k,n αντιπροσωπεύουν την σειρά με την οποία τα στιγμιότυπα παράγονται και παρατηρούνται ενώ το S αντιπροσωπεύει το αρχικό ολόκληρο ρεύμα δεδομένων και το R αντιπροσωπεύει το τελικό ρεύμα που αναπαράχθηκε μετά την εφαρμογή της τεχνικής μείωσης επιλογής στιγμιοτύπων. Ορίζουμε το επίπεδο συμπίεσης C ως τον λόγο του πλήθους των στιγμιοτύπων που παρέμειναν από το αρχικό σύνολο δεδομένων μετά την μείωση δηλαδή το R προς το πλήθος των στιγμιοτύπων που αρχικά παρουσιάστηκαν και παρατηρήθηκαν στο ρεύμα S. Το κλάσμα αυτό το εκφράζουμε ως προς τις εκατό. Η αναπαραγωγισιμότητα αναφέρεται στο πόσο καλά το μειωμένο σύνολο δεδομένων αντιπροσωπεύει τα αρχικά δεδομένα ή ισοδύναμα στο κατά πόσο τα αρχικά δεδομένα μπορούν να αναπαραχθούν από το μειωμένο σύνολο δεδομένων. Η αναπαραγωγισιμότητα σύμφωνα με τους Galan, Liu και Torkkola (2005) μετράται στο κατά πόσο τα μειωμένα δεδομένα R αποκλίνουν από τα αρχικά S και τυπικά εκφράζεται υπό την μορφή απόλυτου αθροιστικού σφάλματος χρησιμοποιώντας την Ευκλείδεια απόσταση όπως φαίνεται στην σχέση :

$$D(S,R) = \sqrt{\sum_{i=1}^{n} \left(s_i - r_i\right)^2}.$$

Επομένως η εγγύτητα ή η ομοιότητα ενός ρεύματος S με την ακολουθία R που έχει αναπαραχθεί αντιπροσωπεύεται από το ρίζα των τετραγώνων της διαφοράς μεταξύ δύο ακολουθιών κάθε χρονική στιγμή για ένα μήκος n στιγμιοτύπων που έχουν αναπαραχθεί από τα δύο ρεύματα. Τόσο το επίπεδο συμπίεσης όσο και η αναπαραγωγισιμότητα είναι ανεξάρτητες μετρικές που σημαίνει ότι υπάρχει ένας

συσχετισμός μεταξύ τους. Συγκεκριμένα όσο μεγαλύτερη είναι η αναπαραγωγισιμότητα τόσο μικρότερο είναι το επίπεδο της συμπίεσης και αντιστρόφως. Συνεπώς ένας από τους σημαντικούς στόχους της τεχνικής της επιλογής στιγμιοτύπων είναι να επιτύχει την μέγιστη δυνατή συμπίεση με το ελάχιστο ποσό σφάλματος στην αναπαραγωγή.

Κατά μία έννοια η επιλογή στιγμιοτύπων είναι μια μορφή δειγματοληψίας (Galan, Liu & Torkkola, 2005). Παρόλο που μπορεί εύκολα να χαρακτηριστεί ως μια μορφή δειγματοληψίας σε τακτικά διαστήματα ή απλά τυχαία δειγματοληψία ο στόχος αυτή της τεχνικής είναι να αποδώσει καλύτερα από τις προαναφερόμενες μεθόδους. Τα δεδομένα ροής επιβάλλουν αρκετές προκλήσεις για την επιλογή στιγμιοτύπων εξαιτίας κάποιων έμφυτων χαρακτηριστικών τους όπως η μη ύπαρξη ορίων, οι μεγάλοι ρυθμοί παραγωγής παραδειγμάτων και τα προσωρινά στατιστικά τους κάτι που δεν παρουσιάζεται στα παραδοσιακά σύνολα δεδομένων. Το πιο σημαντικό είναι ότι στις παραδοσιακές βάσεις δεδομένων η διαδικασία της προεπεξεργασίας των δεδομένων μπορεί να εκτελεστεί offline και μπορεί να εκτελεστεί με πολλαπλά περάσματα του συνόλου αν αυτό κριθεί απαραίτητο για να επιτευχθεί η μέγιστη δυνατή επίδοση του αλγόριθμου. Στα δεδομένα ροής όμως η ανάλυση των δεδομένων πρέπει να γίνει online και οι μέθοδοι έχουν συνήθως μόνο μια ευκαιρία να εξετάσουν τα δεδομένα επομένως η απόδοση πρέπει να εξαχθεί με ένα πέρασμα των δεδομένων. Παρόλα αυτά πολλές τεχνικές που αναπτύχθηκαν για την ανάλυση σε παραδοσιακές βάσεις δεδομένων προσαρμόστηκαν στα δεδομένα ροής με την χρήση συρόμενων παραθύρων. Η τεχνική αυτή των συρόμενων παραθύρων βρίσκει εφαρμογή και στην προεπεξεργασία των δεδομένων με την επιλογή στιγμιοτύπων.

2.5.1 Η μέθοδος NEFCS - SRR

2.5.1.1 Εισαγωγή

Η εξελισσόμενη φύση καθώς και ο συγκεντρωτικός όγκος των δεδομένων του πραγματικού κόσμου δημιουργεί όπως έχουμε αναφέρει το φαινόμενο του concept drift. Συνεπώς η συλλογιστική που είναι γνωστή ως case-based reasoning (CBR) οφείλει να αλλάξει τις στρατηγικές της για να αντιμετωπίσει το πρόβλημα. Με βάση τα σύγχρονα θέματα που σχετίζονται με CBR τεχνικές που διαχειρίζονται το πρόβλημα του concept drift οι Lu et al. (2016) προτείνουν μια case-base τεχνική με δύο στάδια. Στο πρώτο στάδιο προτείνουν έναν αλγόριθμο με την ονομασία NEFCS (Noise-Enhanced Fast Context Switch) ο οποίος όπως υποδηλώνει και η ονομασία του αποσκοπεί στην απομάκρυνση των στιγμιοτύπων που αποτελούν θόρυβο αλλά σε ένα δυναμικό περιβάλλον. Στο δεύτερο στάδιο οι δημιουργοί αναπτύσσουν έναν καινοτόμο αλγόριθμο που καλείται SRR (Stepwise Redundancy Removal) ο οποίος μειώνει το μέγεθος του case-base απομακρύνοντας τα περιττά στιγμιότυπα ενώ ταυτόχρονα διατηρεί την case-base κάλυψη.

Η διαδικασία case-based Reasoning (CBR) που αναφέρεται παραπάνω είναι μια στρατηγική επίλυσης προβλημάτων η οποία χρησιμοποιεί προηγούμενη εμπειρία για να κατανοήσει και να επιλύσει νέα προβλήματα. Εν αντιθέσει με άλλες μεθόδους εκμάθησης οι οποίες αποθηκεύουν προηγούμενη εμπειρία ως γενικούς κανόνες και αντικείμενα τα συστήματα CBR αποθηκεύουν προηγούμενες εμπειρίες ως ατομικά επεισόδια επίλυσης προβλημάτων (Lopez et al. 2005) και καθυστερούν την γενίκευση έως ότου να έρθει η στιγμή της επίλυσης κάποιου προβλήματος. Επίσης υπάρχει και η έννοια case-base συντήρηση CBM η οποία αναφέρεται στη διαδικασία διόρθωσης των περιεχομένων ενός CBR συστήματος με στόχο την βελτίωση της ακρίβειας και της ικανότητας του (Wilson & Leake 2001). Οι μέθοδοι που εφαρμόζονται στην CBM περιλαμβάνουν την μείωση του μεγέθους ενός case base συστήματος ή του συνόλου εκπαίδευσης ενώ ταυτόχρονα επιχειρείται η διατήρηση ή ακόμα και η βελτίωση της γενικευμένης ακρίβειας (Delany & Cunningham 2004). Δυο είναι

οι κύριες περιοχές στην case-base συντήρηση που έχουν αναγνωριστεί και ερευνηθεί: 1) η ενίσχυση της ικανότητας του μοντέλου που στοχεύει στην μετακίνηση των περιπτώσεων που αποτελούν θόρυβο, συνεπώς ενισχύεται η ακρίβεια του κατηγοριοποιητή και 2) η συντήρηση της ικανότητας που αποσκοπεί στην μείωση των περιττών περιπτώσεων δηλαδή στην διαγραφή αχρείαστων περιπτώσεων που δεν συνεισφέρουν στην ικανότητα του κατηγοριοποιητή.

Με την υπάρχουσα βιβλιογραφία οι προσεγγίσεις που αφορούν τον χειρισμό του φαινομένου του concept drift μπορούν να διαιρεθούν σε τρεις βασικές κατηγορίες (Tsai, Lee & Yang, 2009): 1) Σε μεθόδους επιλογής στιγμιοτύπων (βασισμένες σε παράθυρα) όπου η κύρια ιδέα είναι η επιλογή των στιγμιοτύπων που είναι τα πιο σχετικά σύμφωνα με το σενάριο της δεδομένης χρονικής στιγμής. 2) στιγμιότυπα με βάρη όπου σε κάθε στιγμιότυπο ανατίθεται ένα βάρος το οποίο αντιπροσωπεύει την μειωμένη σχετικότητα των υπαρχόντων παραδειγμάτων εκπαίδευσης. Αυτά τα στιγμιότυπα μπορούν να λάβουν βάρη με βάση την παραμονή τους μέσα στο παράθυρο ή την σχετικότητα τους με βάση το τρέχων σενάριο. 3) Η εκμάθηση με σύνολα δηλαδή η εκμάθηση με πολλαπλά μοντέλα αναφέρεται ότι είναι η πιο δημοφιλής και η πιο επιτυχημένη προσέγγιση για την αντιμετώπιση δεδομένων που παρουσιάζουν το φαινόμενο του concept drift (Qu et al. 2009). Αξιοποιεί πολλαπλά μοντέλα με ψήφο (Bifet et al. 2009) ή με επιλογή των πιο σχετικών μοντέλων για να κατασκευάσει ένα αποτελεσματικό προβλεπτικό μοντέλο (Klinkenberg & Joachims 2000). Αυτό που αφορά τον αλγόριθμο που θα αναλυθεί παρακάτω είναι ότι τόσο αυτός όσο και άλλες στρατηγικές CBR που διαγράφουν τον θόρυβο και τις άσχετες ή περιττές περιπτώσεις είναι μια μορφή επιλογής στιγμιοτύπων.

Τα προβλήματα που προκύπτουν από το concept drift σε ένα CBR σύστημα είναι ότι παλιές περιπτώσεις μπορεί να μην αντιστοιχούν σε τωρινά concepts επομένως οι μέθοδοι case-base μπορεί να διατηρούν επιβλαβείς για την ακρίβεια περιπτώσεις. Επιπρόσθετα όταν τα όρια των κλάσεων αλλάζουν η πρωτότυπα concepts εμφανίζονται οι νέες περιπτώσεις που εκπροσωπούν τα πρωτότυπα αυτά concepts είναι πιθανόν να χειριστούν ως θόρυβος και να απομακρυνθούν από αλγορίθμους ενίσχυσης της ικανότητας επειδή είναι σε αντίθεση με παλαιότερα concepts. Αυτό μπορεί να καθυστερήσει ή και να εμποδίσει έναν case-base μοντέλο από το να μάθει νέα concepts. Τέλος η μείωση περιττών

περιπτώσεων αποτελεί μεγάλη πρόκληση σε πεδία με εξελισσόμενα concept drifts. Οι μέθοδοι συντήρησης της ικανότητας διατηρούν μόνο περιπτώσεις που βρίσκονται μέσα στα όρια αποφάσεων. Σύμφωνα με τους Lu et al.(2016) αυτό είναι ακατάλληλο ειδικά σε περιβάλλοντα που παρουσιάζουν concept drift για δύο κυρίως λόγους. Πρώτον καταστρέφεται η αρχική κατανομή των περιπτώσεων επηρεάζοντας έτσι τα αποτελέσματα οποιασδήποτε μεθόδου εντοπισμού αλλαγής που συγκρίνει άμεσα την κατανομή των περιπτώσεων. Δεύτερον κάνει έναν αλγόριθμο εκμάθησης πολύ ευαίσθητο σε θόρυβο επομένως λανθασμένα διατηρεί περιπτώσεις θορύβου ως πρωτότυπα concepts. Με αφετηρία όλες τις παραπάνω προκλήσεις και λαμβάνοντας υπόψη ότι δεν είναι δυνατόν η εκ των προτέρων γνώση της πιθανής παρουσίας ενός concept drift οι Lu et al.(2016) προτείνουν μια πρωτότυπη case-base μέθοδο αλλαγής η οποία απευθύνεται στην ενίσχυση της ικανότητας αλλά και στην διατήρηση της ικανότητας ενώ δουλεύει ικανοποιητικά τόσο σε στατικά όσο και σε μεταβαλλόμενα περιβάλλοντα.

2.5.1.2 Ανάλυση Εννοιών

Ικανότητα

Η ικανότητα (competence) είναι μια μέτρηση του πόσο καλά ένα σύστημα CBR εκπληρώνει τους στόχους του. Εφόσον το CBR είναι μια μεθοδολογία επίλυσης προβλημάτων η ικανότητα λαμβάνεται να είναι το ποσοστό των προβλημάτων όπου το σύστημα αυτό μπορεί να επιλύσει. Η πρώτη έρευνα για μετρήσεις ικανότητας διενεργήθηκε από του Smyth και Keane (1995) στην οποία η τοπική ικανότητα μιας περίπτωσης (case) χαρακτηριζόταν από την κάλυψη (coverage) και την προσβασιμότητα (reachability). Η κάλυψη μιας περίπτωσης είναι το σύνολο των στοχευμένων προβλημάτων τα οποία μπορεί να επιλύσει η περίπτωση αυτή. Η προσβασιμότητα ενός στοχευμένου προβλήματος είναι το σύνολο των περιπτώσεων που μπορούν να χρησιμοποιηθούν για να παρέχουν μια λύση για το συγκεκριμένο πρόβλημα. Εφόσον είναι αδύνατον να αριθμηθούν

όλα τα πιθανά μελλοντικά προβλήματα στην πράξη το case-base θεωρείται ότι είναι αντιπροσωπευτικό δείγμα της κατανομής των στοχευμένων προβλημάτων. Ως αποτέλεσμα η κάλυψη μιας περίπτωσης εκτιμάται από το σύνολο των περιπτώσεων που μπορούν να λυθούν από την ανάκτηση και την προσαρμογή του ενώ η προσβασιμότητα μιας περίπτωσης εκτιμάται ως το σύνολο των περιπτώσεων που μπορούν να επιφέρουν την επίλυση του.

Ενίσχυση ικανότητας

Η ενίσχυση της ικανότητας (competence enhancement) αποσκοπεί στην απομάκρυνση στιγμιοτύπων που αποτελούν θόρυβο με στόχο να αυξηθεί η ακρίβεια του κατηγοριοποιητή. Κάποιες έρευνες που επιλέγουν ένα σύνολο από κατάλληλες περιπτώσεις και καταλήγουν σε πιο βολικά όρια αποφάσεων μπορούν να κατηγοριοποιηθούν ως μια μορφή ενίσχυσης της ικανότητας.

Μια από τις πρώτες τεχνικές απομάκρυνσης θορύβου είναι η τεχνική του Wilson (1972). Στον αλγόριθμο που προτείνουν με την ονομασία ENN μια περίπτωση που κατηγοριοποιείται λάθος από τους εγγύτερους γείτονες διαγράφεται από την αρχική case-base μέθοδο ως θόρυβος. Για τον εντοπισμό και την διαγραφή μικρών συστάδων από περιπτώσεις που αποτελούν θόρυβο ο Tomek (1976) έκανε δυο τροποποιήσεις στον αλγόριθμο ENN. Η πρώτη ονομάζεται RENN και είναι ένας αλγόριθμος που εφαρμόζει επαναλαμβανόμενα τον ENN στο case-base έως ότου δεν υπάρχουν άλλες περιπτώσεις που μπορούν να διαγραφούν. Η δεύτερη ονομάζεται all k-NN και είναι ένας αλγόριθμος ο οποίος αυξάνει τον αριθμό των k γειτόνων μετά από κάθε κύκλο του RENN. Αυτές οι δύο τροποποιήσεις βελτιώνουν την ακρίβεια της ταξινόμησης ωστόσο υπάρχει ρίσκο να απομακρυνθούν ολόκληρες συστάδες ως θόρυβος ενώ στην ουσία αντιπροσωπεύουν γνήσια concept. Οι Ferri και Vidal (1992) ενσωμάτωσαν τον ENN με διασταυρωμένη επικύρωση η οποία χωρίζει τυχαία το σύνολο εκπαίδευσης σε n φακέλους και έπειτα επαναληπτικά απομακρύνει περιπτώσεις από κάθε φάκελο που δεν μπορούν να κατηγοριοποιηθούν σωστά από περιπτώσεις άλλων φακέλων. Τέλος οι Delany και Cunningham (2004) παρουσίασαν τον αλγόριθμο BBNR ο οποίος επιβλέπει όλες τις περιπτώσεις που

έχουν συνεισφέρει σε λανθασμένες ταξινομήσεις και μετακινούν μια περίπτωση αν η διαγραφή της δεν δημιουργεί απώλεια στην κάλυψη.

Διατήρηση ικανότητας

Η διατήρηση της ικανότητας αντιστοιχεί στην μείωση του πλεονάσματος η οποία αποσκοπεί στην απομάκρυνση των περιττών περιπτώσεων οι οποίες δεν συμβάλλουν στην ικανότητα της κατηγοριοποίησης. Η μέθοδος CNN του Hart (1968) είναι η πρώτη και πιο γνωστή μέθοδος μείωσης πλεονάσματος η οποία αυξητικά προσθέτει περιπτώσεις οι οποίες δεν μπορούν να ταξινομηθούν σωστά από το τρέχων case-base σε ένα άδειο case-base.Ο CNN κάνει πολλαπλά περάσματα από το αρχικό case-base έως ότου δεν μπορούν να γίνουν άλλες προσθήκες. Παρόλα αυτά ο CNN έχει δεχθεί κριτική ότι είναι ευαίσθητος στην σειρά με την οποία εξετάζει τις περιπτώσεις και στον θόρυβο. Για την αντιμετώπιση αυτών των προβλημάτων αρκετές τροποποιήσεις έγιναν όπως ο RNN (Gates 1972),ο SNN (Ritter et al. 1975), ο τροποποιημένος CNN (MCNN) (Devi & Murty 2002),ο γενικευμένος CNN (GCNN) (Chou, Kuo & Chang, 2006), ο ταχύς CNN (FCNN) (Angiulli 2007) και ο βελτιωμένος CNN (ICNN) (Hao et al. 2008).

2.5.1.3 Η μέθοδος CBR στην διαχείριση του concept drift

Η πρώτη απόπειρα να αντιμετωπιστεί το φαινόμενο του concept drift από μια case-based τεχνική υλοποιείται από τον αλγόριθμο IB3 (Aha, Kibler & Albert, 1991) ο οποίος καταγράφει την ακρίβεια κάθε περίπτωσης και την συχνότητα ανάκλησης. Ο IB3 αποτρέπει τις περιπτώσεις που αποτελούν θόρυβο αναβάλλοντας την εισαγωγή τους στο case-base έως ότου αυτές αποδειχθούν αξιόπιστες μέσω ενός τεστ βαθμού εμπιστοσύνης όπου ο αριθμός των επιτυχημένων κατηγοριοποιήσεων θεωρείται ότι κατανέμεται διωνυμικά. Μια περίπτωση που έχει συμπεριληφθεί στο case-base απομακρύνεται οριστικά αν η ακρίβεια της είναι σημαντικά πιο μικρή από παρατηρούμενη συχνότητα της κλάσης της. Ο IB3 έχει δεχθεί κριτική όσον αφορά την ικανότητα του να χειρίζεται μόνο σταδιακά concept drift και την ακριβή διαδικασία προσαρμογής του (Widmer & Kubat 1996). Ο αλγόριθμος LWF (Locally Weighted Forgetting) του

Salganicoff (1993) ο οποίος μειώνει τα βάρη των k εγγύτερων γειτόνων μια νέας περίπτωσης και απορρίπτει μια περίπτωση αν το βάρος της πέφτει κάτω από ένα κατώφλι θ, θεωρούνταν ότι ήταν ένας από τους καλύτερους προσαρμοστικούς αλγόριθμους εκμάθησης στην εποχή του. Ωστόσο ο LWF έχει χαμηλότερη ασυμπτωτική ταξινόμηση σε μη ασταθής συνθήκες. Ο Klinkenberg (2004) απέδειξε μέσω πειραμάτων ότι οι τεχνικές στιγμιοτύπων με βάρη τείνουν να υπερμοντελοποιούν τα δεδομένα με συνέπεια να αποδίδουν χειρότερα από άλλες ανάλογες τεχνικές επιλογής στιγμιοτύπων. Ο Salganicoff (1997) εισήγαγε τον αλγόριθμο PECS (Prediction Error Context Switching) ο οποίος επιτυγχάνει καλή απόδοση τόσο σε μεταβαλλόμενα όσο και σε στατικά περιβάλλοντα. Είναι όμοιος με τον IB3 υπό την έννοια ότι και αυτός καταγράφει την ακρίβεια της περίπτωσης και υιοθετεί το ίδιο τεστ με βαθμού εμπιστοσύνης για να καθορίσει τον θόρυβο παρόλο που ο PECS δεν κανονικοποιεί το κατώτερο όριο του βαθμού εμπιστοσύνης όσον αφορά την συνολικά παρατηρούμενη συχνότητα της κλάσης. Παρόλα αυτά ο PECS διαφέρει από τον IB3 με πολλαπλούς τρόπους. Πρώτον κάθε νέα παρατήρηση εισάγεται κατευθείαν στο case-base. Δεύτερον ο PECS υπολογίζει την ακρίβεια μια περίπτωσης βασιζόμενος μόνο στις τελευταίες l προβλέψεις του. Τρίτον αντί να διαγράφει οριστικά μια περίπτωση ως θόρυβος ο αλγόριθμος PECS την απενεργοποιεί και παρακολουθεί την ακρίβεια της με σκοπό την επανενεργοποίηση της. Τα πειράματα δείχνουν ότι ο PECS βελτιώνει την ευρωστία έναντι του IB3 κυρίως σε εργασίες που είναι μεταβαλλόμενες με τον χρόνο αλλά με το κόστος της ολοένα και μεγαλύτερης απαίτησης χώρου μνήμης. Ωστόσο ο PECS είχε αρχικά σχεδιαστεί με στόχο να βελτιώνει την απόδοση σε προβλήματα που περιέχουν concept drift, όπου ο θόρυβος μαζί με εισερχόμενες παρατηρήσεις δεν λαμβανόταν υπόψη. Ως αποτέλεσμα όλες οι παρατηρήσεις που αποτελούν θόρυβο κρατούνται αρχικά και μπορούν να απομακρυνθούν μόνο με αναστολή. Ως συνέπεια ο PECS έχει δεχθεί κριτική για αυτή του την τεχνική διότι απαιτείται ανεξάντλητη μνήμη καθώς οι περιπτώσεις δεν διαγράφονται πότε παρά μόνο απενεργοποιούνται. Οι Delany et al.(2005) προτείνουν έναν αλγόριθμο εκμάθησης δύο επιπέδων για την διαχείριση του concept drift. Στο πρώτο επίπεδο χρησιμοποιούν μια CBE μέθοδο αλλαγής που είναι μια υβριδική διότι συνδυάζει τις μεθόδους BBNR και CRR για να διαχειριστεί το case-base περιοδικά. Ειδικότερα ο BBNR αναλύει όλες τις περιπτώσεις που

έχουν συμβάλλει στην λανθασμένη κατηγοριοποίηση και απομακρύνει περιπτώσεις αν η διαγραφή τους δεν επιφέρει απώλεια στην κάλυψη. Η CRR επαναλαμβανόμενα επιλέγει μια περίπτωση με την μικρότερη κάλυψη η οποία δεν μπορεί να λυθεί σωστά. Οι δημιουργοί συγκρίναν την μέθοδο τους με ένα ολόκληρο case-base χωρίς διαχείριση, όπως και με ανανέωση βασισμένη σε παράθυρα και παρουσίασαν αρκετές βελτιώσεις. Ωστόσο ένα υβρίδιο δύο CBM μεθόδων που έχει σχεδιαστεί για στατικά περιβάλλοντα δεν εγγυάται αποτελεσματική εκμάθηση κάτω από μεταβαλλόμενα περιβάλλοντα. Στην χειρότερη περίπτωση πρωτότυπα concepts μπορεί να απορρίπτονται διαρκώς ειδικά υπό την παρουσία σταδιακού concept drift. Επιπρόσθετα ο BBNR έχει δυσκολίες στην απομάκρυνση μικρών ομάδων περιπτώσεων θορύβου. Για παράδειγμα δύο περιπτώσεις θορύβου που έχουν η μια την άλλη μέσα στο σύνολο κάλυψης τους μπορούν να ταξινομήσουν σωστά η μια την άλλη αλλά μπορούν να προκαλέσουν λανθασμένη ταξινόμηση για όλες τις άλλες κοντινές περιπτώσεις. Αυτές οι μικρές ομάδες θορύβου δεν μπορούν να απομακρυνθούν από τον αλγόριθμο BBNR ακόμα και αν εξακολουθούν να παρέχουν λανθασμένα αποτελέσματα ταξινόμησης. Αυτό το φαινόμενο μπορεί εύκολα να προκληθεί από ξεπερασμένες περιπτώσεις όταν λαμβάνει χώρα ένα concept drift. Τέλος ο BBNR παραμελεί το πρόβλημα του CBM. Ένα ασύμμετρο μοντέλο ικανότητας μπορεί να οδηγήσει στην λανθασμένη διατήρηση μια περίπτωσης που αποτελεί θόρυβο. Στο δεύτερο επίπεδο οι Delany et al. (2005) ξαναεπιλέγουν περιοδικά γνωρίσματα με σκοπό να ξαναχτίσουν εξολοκλήρου ένα CBR σύστημα. Το δεύτερο επίπεδο είναι πέραν του πεδίου μιας case-base μεθόδου αλλαγής αποτελεί όμως μια στρατηγική ανοικοδόμησης ενός μοντέλου που μπορεί να εφαρμοστεί σε οποιαδήποτε τεχνική επιλογής στιγμιοτύπων.

Οι Beringer και Hullermeier (2007) παρουσίασαν έναν αλγόριθμο εκμάθησης που βασίζεται στα στιγμιότυπα με την ονομασία IBL-DS ο οποίος αυτόνομα ελέγχει την σύσταση και το μέγεθος του case-base. Αυτός ο αλγόριθμος βασίζεται σε τρεις κανόνες τροποποίησης : 1) Όταν το μέγεθος του case-base ξεπερνά το όριο το παλαιότερο στιγμιότυπο απομακρύνεται, 2) Όταν το φαινόμενο του concept drift εντοπίζεται με την μέθοδο του Gama (Gama et al. 2004) ένας μεγάλος αριθμός στιγμιοτύπων θα διαγραφεί με χωρικά ομοιόμορφο αλλά προσωρινά μονομερή τρόπο. Ο αριθμός των περιπτώσεων που θα

απομακρυνθεί εξαρτάται από την απόκλιση μεταξύ του ελάχιστου βαθμού σφάλματος και του βαθμού σφάλματος των 20 πιο πρόσφατων ταξινομήσεων, 3) Για κάθε νέα περίπτωση που διατηρείται της οποίας η κλάση κυριαρχεί σε ένα εύρος, όλοι οι γείτονες σε ένα υποψήφιο εύρος που ανήκει σε μια διαφορετική κλάση θα διαγράφονται. Παρόλο που η μέθοδος IBL-DS είναι πληροφοριακή διότι αλλάζει την στρατηγική επιλογής στιγμιοτύπων της σε περίπτωση που εντοπιστεί κάποιο concept drift, η στρατηγική αυτή μπορεί να είναι δύσκολη να ενσωματωθεί σε κάποια πεδία προβλημάτων όπως το φιλτράρισμα ανεπιθύμητης αλληλογραφίας όπου όλα τα γνωρίσματα μπορεί να είναι δυαδικά. Τέλος η προσωρινή απομάκρυνση περιπτώσεων μπορεί να έχει ως αποτέλεσμα την απώλεια στην ικανότητα του case-base διότι μπορεί να διαγράφονται σπάνιες αλλά σωστές περιπτώσεις.

Οι δημιουργοί του αλγορίθμου NEFCS -SRR (Lu et al. 2016) υποστηρίζουν ότι υπάρχει ξεκάθαρη διαφορά μεταξύ των παραπάνω μεθόδων και του δικού τους αλγόριθμου διότι είναι μια case-base μέθοδο επιλογής στιγμιοτύπων που διαρκώς βελτιώνεται ή τουλάχιστον διατηρεί την αποτελεσματικότητα και την ακρίβεια ενός κατηγοριοποιητή σε άγνωστες καταστάσεις (concept drift) με την προϋπόθεση ότι παρέχονται στη μέθοδο νέες περιπτώσεις με ετικέτα.

2.5.1.4 Η νέα case-base μέθοδος αλλαγής

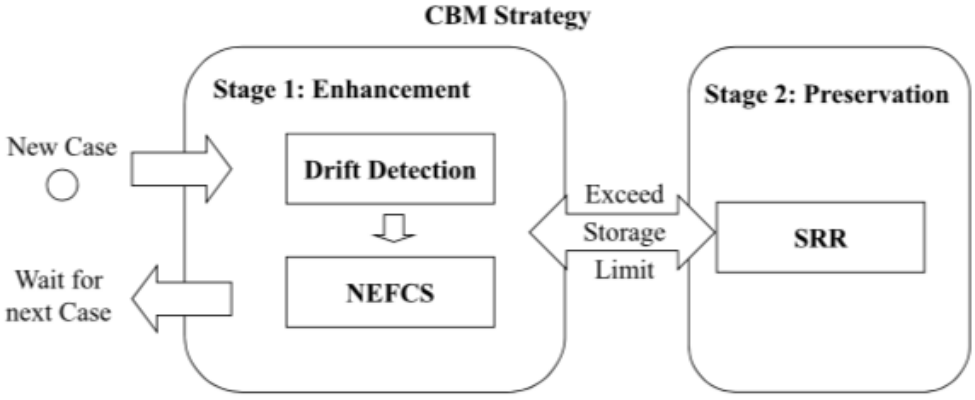
Οι παραδοσιακές case-base μέθοδοι αλλαγής είτε παραβλέπουν την εξελισσόμενη φύση πολλών πραγματικών σεναρίων κάνοντας έμμεσα υπόθεση ότι είναι σταθερά είτε είναι ανίκανες να αντιμετωπίσουν ταυτοχρόνως την ανάγκη για απόδοση αλλά και τα θέματα που αφορούν την ικανότητα (competence). Έχοντας ως κίνητρο τα παραπάνω προβλήματα οι Lu et al. (2016) κατασκεύασαν μια CBM διαδικασία που αποτελείται από δύο στάδια. Στο στάδιο 1 εκμάθησης (ενίσχυση) προτείνουν τον NEFCS αλγόριθμο που στοχεύει στην απομάκρυνση θορύβων σε δυναμικά περιβάλλοντα. Στο στάδιο 2 εκμάθησης (διατήρηση) προτείνουν τον αλγόριθμο SRR ο όποιος απομακρύνει τις περιττές περιπτώσεις με έναν επαναλαμβανόμενα ομοιόμορφο τρόπο ενώ διατηρεί την case base κάλυψη. Η τεχνική τους σχετίζεται τόσο με τη φιλοσοφία του CBM όσο

και με την κατηγοριοποίηση σε δεδομένα ροής που παρουσιάζουν concept drift. Θεωρούν ότι υπάρχει ένα case-based σύστημα που καταγράφει ένα ζωντανό ρεύμα νέων στιγμιοτύπων όπου νέα στιγμιότυπα με ετικέτα μπορούν να διατηρηθούν για μελλοντική χρήση. Καθώς μπορεί να παρουσιαστεί το concept drift και ο θόρυβος παραμένει οι δημιουργοί στόχευσαν στο να ενισχύσουν αυτό το σύστημα με την ικανότητα να μαθαίνει και να βελτιώνεται αναλόγως ενώ θα περιορίζει το case-base μέγεθος του αντί να το επιτρέπει να επεκτείνεται ανεξέλεγκτα. Η λύση τους αποτελείται από τρία τμήματα :

1. εντοπισμός αλλαγής βασισμένος στην ικανότητα (N. Lu, Zhang & J. Lu, 2014) ο οποίος συγκρίνει την κατανομή των περιπτώσεων μεταξύ δυο συρόμενων παραθύρων από πρόσφατες περιπτώσεις. Όταν ένα concept drift εντοπίζεται τότε εντοπίζεται παράλληλα και η περιοχή ικανότητας όπου η κατανομή παρουσιάζει την πιο σημαντική αλλαγή.

2. Ο αλγόριθμος NEFCS(Noise-Enhanced Fast Context Switch) ο οποίος ενισχύει τις δυνατότητες εκμάθησης του συστήματος υπό συνθήκες concept drift.

3. Η SRR (Stepwise Redundancy Removal) μέθοδος η οποία ελέγχει το μέγεθος του case-base συστήματος με σκοπό να αντιμετωπίσει το θέμα της απόδοσης.



Εικόνα 2.1 : Μια CBM διαδικασία δύο σταδίων (N. Lu et al. 2016)

Η διαδικασία φαίνεται αναλυτικά στην εικόνα 2.1. Για κάθε νέο στιγμιότυπο με ετικέτα εκτελείται από το Drift Detection εντοπισμός αλλαγής βασισμένος στην ικανότητα. Έπειτα ο NEFCS προσαρμόζει το case-base κάθε φορά που εμφανίζεται concept drift επομένως διαρκώς βελτιώνει την ακρίβεια κατηγοριοποίησης του συστήματος κατά την χρήση του. Τέλος όταν το case-based σύστημα υπερβαίνει το όριο χωρητικότητας του ο SRR ενεργοποιείται και συρρικνώνει το case-base ενώ διατηρεί την κατανομή των περιπτώσεων όσο το δυνατόν περισσότερο.

Εντοπισμός αλλαγής βασισμένος στην ικανότητα

Η CBM στρατηγική των Lu et al. (2016) παίρνει κίνητρο από την ιδέα ότι όταν δεν υπάρχει concept drift τότε παλιές περιπτώσεις μπορούν να βοηθήσουν στην αναγνώριση του θορύβου και να βελτιώσουν την ακρίβεια. Από την άλλη όταν η παρουσία του concept drift είναι εμφανής τότε τα νέα στιγμιότυπα είναι πιο αντιπροσωπευτικά του καινούργιου concept ενώ οι παλιές περιπτώσεις είναι ξεπερασμένες. Επομένως ένα μοντέλο εντοπισμού αλλαγής υιοθετείται από τους δημιουργούς με βάση τα παραπάνω.

Στην ουσία η συγκεκριμένη μέθοδος εντοπισμού αλλαγής διαιρεί το χώρο του προβλήματος σε μια ομάδα επικαλυπτόμενων σετ και έπειτα υπολογίζει την εμπειρική πιθανότητα πάνω σε μια περιοχή ικανότητας έστω A που αποτελεί υποσύνολο του $R^{CB}(S_i)$.Το $R^{CB}(S_i)$ είναι το σύνολο όλων των σχετικών σετ που περιέχουν τουλάχιστον μια περίπτωση $c \in S_i$,όπου $S_i \subseteq CB(i = 1,2)$ τα δείγματα της περίπτωσης. Ο υπολογισμός αυτός γίνεται μέσω του εμπειρικού βάρους (N.Lu, Zhang & J.Lu, 2014). Το εμπειρικό βάρος αποτελεί το άθροισμα των πυκνοτήτων όλων των σχετικών σετ $r \in R^{CB}(S_i)$ που ανήκουν στο υποσύνολο A. Όσο υψηλότερο είναι αυτό το βάρος τόσο μεγαλύτερο είναι το ποσοστό των περιπτώσεων στο S που υποστηρίζουν την επιλεγμένη περιοχή ικανότητας A. Παρακάτω δίνονται δύο ορισμοί και διατυπώνεται ένα θεώρημα μέσω των οποίων θα αναλυθεί το πως λειτουργεί ο εντοπισμός αλλαγής σε περιοχές ικανότητας.

Ορισμός 1 : Για δύο σύνολα περιπτώσεων $S_1, S_2 \subseteq CB$ και για ένα σχετικό σετ $r \in R^{CB}(S_1) \cup R^{CB}(S_2)$, με $R = \{r\}$ ορίζεται ως εμπειρική απόσταση του r με βάση το σχετικό σετ μεταξύ των S_1, S_2 να είναι η $d_r^{CB}(S_1, S_2) = \left| S_1^{CB}(R) - S_2^{CB}(R) \right|$ (Lu et al. 2016).

Θεώρημα : Με βάση τον ορισμό της εμπειρικής απόστασης με βάση το σχετικό σετ η εμπειρική απόσταση με βάση την ικανότητα μεταξύ του S_1 και του S_2 $d^{CB}(S_1, S_2) = sup_{A \in A} \left| S_1^{CB}(A) - S_2^{CB}(A) \right|$ μπορεί να υπολογιστεί ως εξής :

$$d^{CB}(S_1, S_2) = \sum_{r \in P} d_r^{CB}(S_1, S_2) \text{ ή } d^{CB}(S_1, S_2) = \sum_{r \in Q} d_r^{CB}(S_1, S_2)$$

όπου :

$P = \{r \in R^{CB}(S_1) \cup R^{CB}(S_2) : S_1^{CB}(R) - S_2^{CB}(R) > 0\}$

$Q = \{r \in R^{CB}(S_1) \cup R^{CB}(S_2) : S_1^{CB}(R) - S_2^{CB}(R) < 0\}$

$\sum_{r \in P} d_r^{CB}(S_1, S_2) = \sum_{r \in Q} d_r^{CB}(S_1, S_2)$

Σε σύγκριση με την εμπειρική απόσταση με βάση την ικανότητα η οποία καθορίζει την απόσταση μεταξύ δύο δειγματικών συνόλων, η εμπειρική απόσταση με βάση το σχετικό σετ καθορίζει την απόσταση μεταξύ δυο δειγματικών συνόλων σε ένα συγκεκριμένο σχετικό σύνολο. Το παραπάνω θεώρημα αποδεικνύει ότι η εμπειρική απόσταση με βάση την ικανότητα είναι το άθροισμα των εμπειρικών αποστάσεων με βάση το σχετικό σύνολο πάνω σε όλες τις διαιρέσεις (σχετικά σετ) οι οποίες είναι υπέρ του S_1 ή του S_2 (αναλόγως πιο έχει το μεγαλύτερο βάρος).

Ορισμός 2: Αν ένα σύνολο από σχετικά σύνολα $R_p \subseteq \{r \in R^{CB}(S_1) \cup R^{CB}(S_2) : d_r^{CB}(S_1, S_2) > 0\}$ ικανοποιεί τις τρεις παρακάτω συνθήκες :

(1) $d_r^{CB}(S_1, S_2) \geq d_{r'}^{CB}(S_1, S_2)$ για κάθε $r \in R_p$ και για κάθε $r' \notin R_p$

(2) $$\sum_{r \in R_p} d_r^{CB}(S_1, S_2) \geq p \sum_{r \in R^{CB}(S1) \cup R^{CB}(S_2)} d_r^{CB}(S_1, S_2)$$

(3) $$\sum_{r \in R_p} d_r^{CB}(S_1, S_2) - min\left(d_r^{CB}(S_1, S_2)\right) < p \sum_{r \in R^{CB}(S1) \cup R^{CB}(S_2)} d_r^{CB}(S_1, S_2)$$

όπου 0 ≤ p ≤ 1 τότε κάθε στοιχείο στο σύνολο R_p ονομάζεται top-p περιοχή ικανότητας (Lu et al. 2016).

Οι Lu et al.(2016) αποδεικνύουν ότι η μεγαλύτερη εμπειρική απόσταση με βάση το σχετικό σετ εμφανίζεται στις top-p περιοχές ικανότητας, επομένως συμπεριφέρονται σε αυτές τις περιοχές ως τις αναγνωρισμένες περιοχές ικανότητας που παρουσιάζουν concept drift στην πιο έντονη μορφή του. Τα αποτελέσματα στον εντοπισμό του concept drift μπορούν να βελτιωθούν σημαντικά αν πιο πολλά αντιπροσωπευτικά δείγματα χρησιμοποιούνται για τον εντοπισμό αλλαγής. Επιπρόσθετα οι Lu et al. (2016) επισημαίνουν ότι όσο μικρότερη είναι η p-τιμή τόσο περισσότερη εμπιστοσύνη θα υπάρχει στο ότι οι περιοχές που αναγνωρίστηκαν είναι περιοχές που παρουσιάζουν concept drift αν και λιγότερες περιοχές ικανότητας θα έχουν επιλεχθεί. Τέλος στην πράξη περιοχές που βρίσκονται κάτω από την επίδραση του concept drift μπορούν να αναγνωριστούν μεταφράζοντας τον χώρο ικανότητας σε χώρο γνωρισμάτων και επιλέγοντας τον χώρο των γνωρισμάτων όπυο υπάρχει μια τεράστια τάση θετικών σε ένα δειγματικό σύνολο και μια τεράστια τάση αρνητικών στο άλλο δειγματικό σύνολο.

Παρόλο που οποιαδήποτε τεχνική εντοπισμού μπορεί να υιοθετηθεί για να ολοκληρώσει αυτή την διεργασία οι δημιουργοί του αλγορίθμου επέλεξαν την παραπάνω μέθοδο εντοπισμού που είναι βασισμένη στην ικανότητα και η οποία χρησιμοποιεί το μοντέλο της ικανότητας ως μια τεχνική διαίρεσης του χώρου και συγκρίνει μέσω αυτής τις κατανομές δύο παραθύρων με περιπτώσεις. Ο κύριος λόγος που επιλέχθηκε επομένως η παραπάνω μέθοδος είναι ότι όχι μονό μπορεί να ειδοποιήσει για την παρουσία πιθανού concept drift αλλά μπορεί επιπλέον να

υποδείξει μια μικρή περιοχή του προβλήματος του χώρου η οποία ονομάζεται αναγνωρισμένη περιοχή ικανότητας. Αυτή η περιοχή είναι το πιο πιθανόν να επηρεαστεί από ένα concept drift. Αυτός ο τρόπος λοιπόν καθιστά την CBR μέθοδο πιο κατάλληλη για να αντιμετωπίσει την παρουσία τοπικού concept drift.

Η μέθοδος NEFCS

Η μέθοδος NEFCS παίρνει τα αποτελέσματα του εντοπισμού αλλαγής με βάση την ικανότητα ως είσοδο (είτε υπάρχει concepr drift και η περιοχή ικανότητας που παρουσιάζεται αυτό έχει εντοπιστεί) και στοχεύει στο να βελτιώνει συνεχώς την ικανότητα εκμάθησης. Η NEFCS αποτελείται από τρεις κύριες διαδικασίες :

- Τροποποιημένη μείωση θορύβου με βάση την ευθύνη M-BBNR
- αλλαγή περιεχομένου
- Ανανέωση μοντέλου ικανότητας

Παρακάτω δίνονται κάποιοι βασικοί ορισμοί για να γίνει κατανοητός ο τρόπος λειτουργίας της μεθόδου.

Ορισμός 3 : Για μια βάση περιπτώσεων $CB = \{c_1, c_2, \ldots, c_n\}$, δοσμένης μιας περίπτωσης $c \in CB$ το σύνολο κάλυψης του είναι το $CoverageSet(c) = \{c' \in CB : Solves(c, c')\}$, όπου το $Solves(c, c')$ σημαίνει ότι το c μπορεί να ανακτηθεί και να προσαρμοστεί για να λύσει την περίπτωση c' (Lu et al. 2016).

Ορισμός 4 : Το σύνολο ευθύνης ορίζεται να είναι το LiabilitySet(c) = $\{c' \in CB : Misclassifies(c', c)\}$, όπου $Misclassifies(c', c)$ σημαίνει ότι η περίπτωση c συμβάλει με κάποιο τρόπο στην λανθασμένη κατηγοριοποίηση της στοχευμένης περίπτωσης c' (Lu et al. 2016).

Ορισμός 5 : Η περίπτωση $c' \in CB$ μπορεί να λυθεί από το CB και αυτό δηλώνεται ως $Solves(CB, c')$, αν υπάρχει περίπτωση $c \in CB$ με $c' \neq c$ τέτοια ώστε $c' \in CoverageSet(c)$ (Lu et al. 2016).

Ορισμός 6 (Υποθετικός κανόνας BBNR) : Αν ένα concept drift έχει εντοπιστεί, μια νέα περίπτωση c θα διαγραφεί με ασφάλεια σύμφωνα με τον BBNR κανόνα ο οποίος είναι :

$|LiabilitySet(c)| > 0$ και $Solves\left(CB - \{c\}, c'\right)$ να ισχύει $\forall c' \in CoverageSet(c).$

μόνο όταν η περίπτωση c βρίσκεται εκτός κάθε αναγνωρισμένης περιοχής ικανότητας που πραγματοποιείται το concept drift (Lu et al. 2016).

- Τροποποιημένη μείωση θορύβου με βάση την ευθύνη M-BBNR

Η τροποποιημένη μείωση θορύβου M-BBNR εφαρμόζει τον υποθετικό κανόνα BBNR για να εξετάσει κάθε νέα περίπτωση c και να καθορίσει αν αυτή αποτελεί θόρυβο. Αν υπάρχει concept drift και η περίπτωση c βρίσκεται μέσα στην αναγνωρισμένη περιοχή ικανότητας όπου αυτό λαμβάνει χώρα τότε η c δεν θα διαγραφεί διότι μπορεί να εκπροσωπεί ένα πρωτότυπο σενάριο. Διαφορετικά η c θα απομακρυνθεί από το case-base (CB) αν ικανοποιεί τον BBNR υποθετικό κανόνα (ορισμός 6). Με αυτό τον τρόπο υπάρχει η δυνατότητα διαφοροποίησης μεταξύ των νέων στιγμιοτύπων και αποτρέπεται η πιθανότητα να συμπεριληφθεί ο θόρυβος. Αν η c απομακρυνθεί τότε η επόμενη διαδικασία (αλλαγή περιεχομένου) θα παραληφθεί διαφορετικά μόνο υπάρχοντες περιπτώσεις οι οποίες συγκρούονται με την c θα ελέγχονται κάτι το οποίο επιταχύνει την διαδικασία εκμάθησης σε νέα σενάρια. Μια περίπτωση c' που ανήκει στο CB θεωρούμε ότι είναι συγκρουόμενη περίπτωση από μια νέα περίπτωση c η οποία δεν ανήκει στο N αν ισχύει ότι $c \in LiabilitySet(c')$. Το σύνολο N αποτελεί το σύνολο των νέων περιπτώσεων οι οποίες διαγράφονται σύμφωνα με τον υποθετικό κανόνα BBNR. Ο υποθετικός BBNR κανόνας επιβάλει μια επιπρόσθετη συνθήκη στον BBNR κανόνα ο οποίος είναι ότι μια νέα περίπτωση c θα θεωρηθεί θόρυβος και θα απομακρυνθεί από τον BBNR κανόνα αν η c

βρίσκεται εκτός ενός αναγνωρισμένου πεδίου ικανότητας που έχει concept drift. Αυτό είναι ένα από τα πλεονεκτήματα της μεθόδου εντοπισμού αλλαγής βασισμένη στην ικανότητα η οποία μπορεί να αναγνωρίσει μικρές περιοχές ικανότητας οπού η αλλαγή είναι πιο έντονη. Οι Lu et al. (2016) τονίζουν ότι όταν εργάζονται με άλλες μεθόδους εντοπισμού αλλαγής οι οποίες δεν έχουν την ικανότητα να εντοπίσουν την αλλαγή τότε όλες οι νέες περιπτώσεις δεν απομακρύνονται κάτι το οποίο συμβαίνει στην περίπτωση που θέσουμε p = 0 στον ορισμό 2.

Για να αποτρέψει το περιττό κόστος ο M-BBNR εξετάζει μόνο τις περιπτώσεις που είναι σε σύγκρουση με την c αντί να ελέγχει ολόκληρο το CB όπως κάνει ο BBNR. Αυτό βελτιώνει κατά πολύ την αποδοτικότητα του σταδίου ενίσχυσης της ικανότητας διότι αυτές οι ανακτημένες περιπτώσεις έχουν ήδη αποκτηθεί κατά την διάρκεια επίλυσης του c επομένως καμία επιπλέον περίπτωση δεν χρειάζεται να ανακτηθεί. Επιπρόσθετα κάθε περίπτωση του τελευταίου παραθύρου θα εξαιρεθεί από αυτή την λίστα όταν υπάρχει concept drift αν αυτή η περίπτωση βρίσκεται μέσα στην αναγνωρισμένη περιοχή ικανότητας που έχει το concept drift. Υπάρχουν δύο κύριοι λόγοι για να εξαιρεθούν αυτές οι περιπτώσεις : 1) Υπάρχει ακόμα πιθανότητα η c να είναι θόρυβος και οι πιο πρόσφατες περιπτώσεις στην ίδια περιοχή βοηθάνε στο να μειωθεί αυτή η επίδραση. 2) Για να εμποδιστεί η διαγραφή άλλων πρόσφατα κρατημένων περιπτώσεων οι οποίες εκπροσωπούν πρωτότυπα σενάρια. Τέλος κάθε συγκρουόμενη περίπτωση θα εξετάζεται για θόρυβο με την χρήση του BBNR κανόνα και θα απομακρύνεται αν ικανοποιεί το κριτήριο. Παρακάτω παρατίθεται ο ψευδοκώδικας του BBNR αλγορίθμου στον οποίο οι Lu et al. (2016) έχουν κάνει μια μικρή τροποποίηση στην σειρά 5 για να εξασφαλίσουν το ότι περιπτώσεις που έχουν απομακρυνθεί δεν θα συμπεριληφθούν στην κάλυψη.

Αλγόριθμος: Blame-based noise reduction BBNR (N.Lu et al. 2016)

Τοπικές μεταβλητές :
aL : η αντικρουόμενη λίστα
CB : η βάση περιπτώσεων

CSet(C) : το σύνολο κάλυψης της c

| LSet(c) | : μέγεθος του συνόλου ευθύνης της c

1: ταξινόμησε την aL στη φθίνουσα σειρά του | LSet(c) |

2: **For each** x **in** aL

3: CB = CB - {x}

4: **For each** y **in** CSet(x)

5: **If** CB περιέχει το y

6: **If** y δεν επιλύεται από το CB

7: CB = CB +{x}

8: **break**

9: **end if**

10: **end if**

11: **end for**

12: **end for**

- αλλαγή περιεχομένου

Καθώς ο BBNR έχει δυσκολίες να απομακρύνει ξεπερασμένες περιπτώσεις η μέθοδος NEFCS ανατρέχει στην αποτελεσματικότητα της πληροφορίας για να αντιμετωπίσει το concept drift. Για να εντοπιστεί η αλλαγή στην ακρίβεια λόγο του concept drift χωρίς να υπάρχει τάση από μεγάλο όγκο παρατηρήσεων που προέρχονται από άλλο concept ένας κατάλογος αλλαγής ο οποίος αποθηκεύει τις τελευταίες l ανακτήσεις, κρατείται για κάθε περίπτωση ως μια μορφή της αποτελεσματικότητας της πληροφορίας. Ως αποτέλεσμα ο κατάλογος αλλαγής που κρατάει τις τελευταίες l προβλέψεις κάθε περιπτώσεις που μπορεί να ανακτηθεί θα πρέπει να ανανεώνεται για κάθε παρατηρούμενη νέα περίπτωση c, όπου ένα ανακτήσιμο στοιχείο c' στο c σημαίνει ότι $sim(c, c') \geq sim(c, c'_k)$ όπου c_k' είναι ο k εγγύτερος γείτονας της c.

Ο NEFCS υιοθετεί το ίδιο τεστ διαστήματος εμπιστοσύνης που χρησιμοποιείται και από τον IB3 (Aha, Kibler & Albert, 1991) και από τον PECS

(Salganicoff 1997) για να αλλάξει τις περιπτώσεις. Αυτό αποφέρει ένα διάστημα εμπιστοσύνης το οποίο υπολογίζεται σύμφωνα με τον τύπο :

$$confidence \quad interval = \frac{p_i + z^2 \pm z\sqrt{p_i(1-p_i)/n + z^2/4n^2}}{1 + z^2/n}$$

Αν το άνω όριο στην ακρίβεια μιας περίπτωσης c πέσει κάτω από το κατώφλι απενεργοποίησης p_{max}, αυτή απενεργοποιείται από την μελλοντική διαδικασία. Ωστόσο αυτό το ίδιο παράδειγμα μπορεί τελικά να μετακινηθεί πίσω στο CB αν το κατώτερο όριο της ακρίβειας του ανέβει πάνω από την συμφωνημένη πιθανότητα αποδοχής p_{min} αν το concept drift είναι κυκλικό. Στον ανωτέρω τύπο p_i είναι η υπολογισμένη ακρίβεια της περίπτωσης c_i και n είναι ο αριθμός των προσπαθειών ταξινόμησης του c_i. Το z είναι η σταθερά του διαστήματος εμπιστοσύνης.

- Ανανέωση μοντέλου ικανότητας

Αυτή η διαδικασία υπάρχει για να εξασφαλίσει ότι η μέθοδος M-BBNR αναφέρεται στο σωστό μοντέλο ικανότητας. Το μοντέλο ικανότητας ανανεώνεται για την προσθήκη περιπτώσεων και οι απομακρυσμένες περιπτώσεις θα διαγραφούν από το μοντέλο ικανότητας

Βαθμιαία απομάκρυνση πλεονάσματος - SRR

Οι Lu et al.(2016) υιοθετούν ένα όμοιο σχέδιο με τον CNN για να κατευθύνουν τον αλγόριθμο απομάκρυνσης περιττών παραδειγμάτων για διατήρηση της ικανότητας και αυτό το σχέδιο είναι η απόκτηση ενός υποσυνόλου του αρχικού CB το οποίο μπορεί να επιλύσει επιτυχώς όλες τις απομακρυσμένες περιπτώσεις. Παρακάτω εξηγείται ο SRR αλγόριθμος με την χρήση του k- NN κανόνα.

Μια μεγάλη διαφορά μεταξύ του SRR και άλλων μεθόδων διατήρησης της ικανότητας με βάση την διαγραφή είναι ότι ο SRR επαναλαμβανόμενα απομακρύνει περιττές περιπτώσεις με ένα τελείως ομοιόμορφο τρόπο κάτι το οποίο οι δημιουργοί του πιστεύουν ότι είναι πιο κατάλληλο για το πρόβλημα του

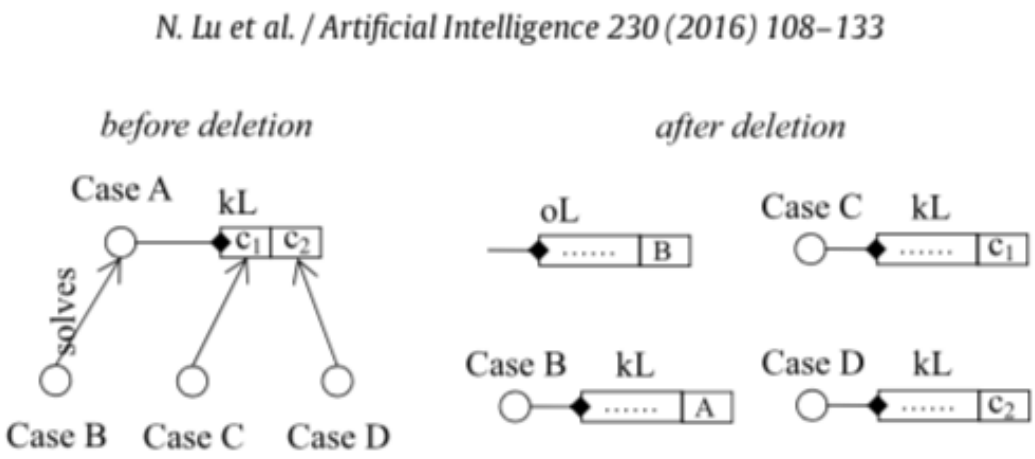
concept drift. Ένας παρόμοιος τρόπος για την διαχείριση ενός concept drift χρησιμοποιείται από τον IBL-DS (Beringer & Hüllermeier 2007), ο οποίος απομακρύνει περιπτώσεις με έναν χωρικά ομοιόμορφο τρόπο. Ωστόσο η ομοιόμορφη διαγραφή περιπτώσεων στον χώρο των γνωρισμάτων μπορεί να είναι δύσκολη να εφαρμοστεί σε πολυδιάστατα δεδομένα όπως για παράδειγμα το φιλτράρισμα ανεπιθύμητης αλληλογραφίας. Επιπρόσθετα ο IBL-DS παρουσιάζει το πρόβλημα της απώλειας της CB ικανότητας. Συνεπώς οι αρθρογράφοι υποστηρίζουν την ομοιόμορφη απομάκρυνση περιττών περιπτώσεων στον χώρο ικανότητας. Επιπλέον εκμεταλλευόμενοι τα μοντέλα ικανότητας η μέθοδος τους δεν χρειάζεται πολλαπλά περάσματα από όλο το CB κάτι το οποίο κερδίζει χρόνο.

Ο SRR ξεκινά ταξινομώντας τις περιπτώσεις σε φθίνουσα σειρά με βάση το σύνολο προσβασιμότητας τους. Ένα μεγάλο σύνολο προσβασιμότητας υποδηλώνει ότι η περίπτωση είναι μακριά από τα όρια της κλάσης (Brighton & Mellish 2002). Ο SRR εργάζεται με επαναλαμβανόμενο τρόπο. Καθώς τρέχει ο SRR διατηρεί τρεις δομές :

- μια λίστα διατήρησης -pL η οποία αποθηκεύει περιπτώσεις που δεν μπορούν να απομακρυνθούν.
- μια κλειδωμένη λίστα -oL η οποία αποτρέπει μια περίπτωση από το να διαγραφεί στην τρέχουσα επανάληψη και θα καθαριστεί στην αρχή της επόμενης επανάληψης.
- μια συνδεδεμένη λίστα -kL η οποία συνδέει αρκετές περιπτώσεις που έχουν απομακρυνθεί πιο πριν σε μια περίπτωση η οποία μπορεί να τις λύσει.

Η μέθοδος SRR λειτουργεί με επαναλαμβανόμενο τρόπο. Σε κάθε επανάληψη η SRR συνεχώς εξετάζει μια ξεκλείδωτη περίπτωση c με το μεγαλύτερο σύνολο προσβασιμότητας έως ότου όλες οι περιπτώσεις κλειδωθούν ή διατηρηθούν. Μια περίπτωση c θα απομακρυνθεί αν η c και όλες οι περιπτώσεις στην συνδεδεμένη λίστα kL της c μπορούν ακόμα να λυθούν σωστά χωρίς την περίπτωση c. Όταν η c απομακρύνεται οι κοντινότερες $(k+1)/2$ περιπτώσεις οι οποίες λύνουν το c κλειδώνονται και μπαίνουν στην συνδεδεμένη λίστα του c. Το $(k+1)/2$ επιλέγεται ως ο αριθμός των περιπτώσεων που θα κλειδωθούν διότι είναι ο

ελάχιστος αριθμός των περιπτώσεων που απαιτείται για να εξασφαλιστεί σωστή κατηγοριοποίηση για το c, συνεπώς δεν θα χαθεί καμία CB ικανότητα αν απορριφθεί η c. Επιπρόσθετα κάθε περίπτωση c_i' στη συνδεδεμένη λίστα kL της c θα προστεθεί στη kL των εγγύτερων $(k+1)/2$ περιπτώσεων οι οποίες λύνουν την c_i'.



Εικόνα 2.2 : Παράδειγμα διαγραφής περίπτωσης (N. Lu et al. 2016)

Η παραπάνω εικόνα δείχνει ένα παράδειγμα στο οποίο η περίπτωση A απομακρύνεται από τον κανόνα του εγγύτερου γείτονα. Αρχικά οι περιπτώσεις στον υπόλοιπο CB που μπορούν να λύσουν την A καθώς και οι περιπτώσεις c_1, c_2 στη συνδεδεμένη λίστα ανακτώνται. Ως αποτέλεσμα οι περιπτώσεις B,C και D ανακτώνται επίσης. Τέτοιες ανακτήσεις είναι άμεσες εξαιτίας του μοντέλου ικανότητας δηλαδή του ReachabilitySet. Επιτυχής ανάκτηση σημαίνει ότι η περίπτωση A και όλες οι περιπτώσεις με τις οποίες η A είναι συνδεδεμένη μπορούν ακόμα να λυθούν από το υπόλοιπό CB χωρίς την A. Επομένως η περίπτωση A μπορεί να απομακρυνθεί με ασφάλεια δηλαδή χωρίς απώλειας της ικανότητας και η περίπτωση B θα κλειδωθεί προσωρινά (θα προστεθεί δηλαδή στη oL λίστα). Έπειτα κάθε περίπτωση που είναι στη συνδεδεμένη λίστα της A θα προστεθεί στη αντίστοιχη kL λίστα της περίπτωσης που την επιλύει. Αποτυχία ανάκτησης οποιασδήποτε περίπτωσης εκ των B, C και D σημαίνει ότι η περίπτωση A δεν μπορεί να μετακινηθεί. Επομένως η A θα διατηρηθεί. Ο SRR θα προχωρήσει στην επόμενη επανάληψη όταν δεν υπάρχουν πλέον ξεκλείδωτες περιπτώσεις και θα σταματήσει αυτόματα όταν όλες οι περιπτώσεις

έχουν διατηρηθεί. Ωστόσο ο SRR μπορεί να σταματήσει αν πάσα στιγμή εφόσον εξασφαλίσει ότι όλες οι απομακρυσμένες περιπτώσεις μπορούν να λυθούν σωστά οποιαδήποτε στιγμή χωρίς απώλεια της κάλυψης. Παρακάτω παρουσιάζεται ο ψευδοκώδικας του αλγορίθμου SRR.

Αλγόριθμος: Stepwise redundancy removal algorithm SRR (N.Lu et al. 2016)

τοπικές μεταβλητές :
CB, η βάση περιπτώσεων
pL, η λίστα διατήρησης
oL, η κλειδωμένη λίστα
kL(c), η συνδεδεμένη λίστα της c
RSet(c), το σύνολο προσβασιμότητας της c ReachabilitySet(c)

```
1: ταξινόμηση του CB σε φθίνουσα σειρά με βάση το μέγεθος του RSet(c)
2: θέσε το oL να είναι κενό
3: While | CB-pL-oL | > 0
4:     θέσε το x ως την πρώτη περίπτωση στο CB-pL-oL
5:     If x και kL(x) μπορούν να λυθούν από το CB-{x}
6:         CB= CB-{x}
7:         For each y in kL(x)
8:             συνέδεσε το y σε όλες τις περιπτώσεις που το επιλύουν
9:         End for
10:        συνέδεσε το x σε όλες τις περιπτώσεις που το επιλύουν
11:        κλείδωσε τις περιπτώσεις που επιλύουν το x
12:    else
13:        pL = pL + x
14:    End if
15: End while
```

16: **If** CB υπερβαίνει το όριο μεγέθους και |oL| > 0

17: πήγαινε στο 2

18: **else**

19: **exit**

20: **End if**

Η λίστα διατήρησης pL διατηρεί τις σημαντικές περιπτώσεις η απομάκρυνση των οποίων θα προκαλέσει απώλεια στην ικανότητα δηλαδή θα οδηγήσει στην αποτυχία επίλυσης περιπτώσεων που έχουν απομακρυνθεί. Επιπρόσθετα η pL βοηθάει στην επιτάχυνση της διαδικασίας της μείωσης περιττών περιπτώσεων διότι οι περιπτώσεις μέσα στη λίστα αυτή θα εξαιρεθούν από τον έλεγχο στις μετέπειτα επαναλήψεις κάτι το οποίο βοηθάει στο να αποφευχθεί το κόστος της επαναλαμβανόμενης εξέτασης των ίδιων σημαντικών περιπτώσεων σε κάθε επανάληψη. Η κλειδωμένη λίστα oL εξασφαλίζει ότι οι περιττές περιπτώσεις θα διαγραφούν με έναν επιδέξια ομοιόμορφο τρόπο. Για κάθε περίπτωση που διαγράφεται μια λίστα από περιπτώσεις που την επιλύουν θα κλειδωθούν για να προστατευτούν από τον κίνδυνο διαγραφής τους κατά την τρέχουσα επανάληψη. Αυτό αποτρέπει σε μια συγκεκριμένη περιοχή περιπτώσεις να καθαριστούν γρήγορα. Η συνδεδεμένη λίστα kL είναι το κλειδί ώστε να εξασφαλιστεί ότι δεν υπάρχει απώλεια στην ικανότητα για οποιαδήποτε διαγραφή κατά την διάρκεια πολλαπλών επαναλήψεων. Αφού ο SRR απομακρύνει περιπτώσεις σταδιακά, η ύπαρξη ενός υποσυνόλου του CB που επιλύει σωστά όλες τις περιπτώσεις στην αρχή κάθε επανάληψης δεν εγγυάται ότι αυτό το υποσύνολο θα επιλύει όλες τις περιπτώσεις του αρχικού CB. Για παράδειγμα κατά την πρώτη επανάληψη η περίπτωση A απομακρύνεται και η περίπτωση B που επιλύει την A κλειδώνεται. Κατα την επόμενη επανάληψη αν η περίπτωση B απομακρυνθεί επειδή η C η οποία λύνει την B μπορεί να ανακτηθεί θα υπάρξει ρίσκο στο να μην μπορεί τελικά να λυθεί η A. Επομένως για κάθε περίπτωση c ο SRR διατηρεί την λίστα kL για να αποθηκεύει όλες τις περιπτώσεις για τις οποίες η c ανακτήθηκε για να λύσει. Για να μπορεί μια περίπτωση c να απομακρυνθεί πρέπει να εξασφαλίζεται ότι η c μπορεί να λυθεί σωστά καθώς και όλες οι περιπτώσεις στις οποίες η c συνδέεται από το υπόλοιπο CB.

Οι Lu et al.(2016) υποστηρίζουν ότι οι διαφορές μεταξύ του SRR και των υπαρχόντων μεθόδων διατήρησης της ικανότητας είναι τεράστιες. Αρχικά οι υπάρχοντες μέθοδοι παίρνουν την μέθοδο της πρόσθεσης μιας περίπτωσης ή της αφαίρεσης αυτής ενώ η SRR προσέγγιση είναι πιο υβριδική μεταξύ των δύο διότι η SRR μπορεί να προσθέσει μια περίπτωση και να την αποτρέψει από το να διαγραφεί μέσω της λίστας διατήρησης. Επιπλέον η SRR απομακρύνει ομοιόμορφα το πλεόνασμα μέσα από τον μηχανισμό κλειδώματος που διαθέτει. Αυτό όχι μόνο διευκολύνει την εξήγηση των περιπτώσεων αλλά επιπλέον κάνει την εκμάθηση πρωτότυπων concept πιο ομαλή χωρίς να αφήνει μεγάλο κενό στο χώρο γνωρισμάτων. Τρίτον για κάθε διαγραφή η SRR εγγυάται ότι δεν θα υπάρχει απώλεια κάλυψης για τις περιπτώσεις. Ως αποτέλεσμα η SRR μπορεί να σταματήσει οποιαδήποτε στιγμή όταν το όριο του μεγέθους εκπληρωθεί κάτι το οποίο δίνει σε αυτόν που λαμβάνει αποφάσεις πολύ περισσότερο έλεγχο στο μέγεθος της CB.

2.5.1.5 Συμπεράσματα

Η case-base διόρθωση είναι μια σημαντική πτυχή της case-base συντήρησης η οποία υιοθετεί μια προσέγγιση επιλογής στιγμιοτύπων για να διαχειριστεί το φαινόμενο του concept drift.

Στο άρθρο τους οι Lu et al.(2016) παρουσίασαν αρχικά μια μέθοδο ενίσχυσης ικανότητας που ονομάζεται NEFCS για να αποτρέψουν τον θόρυβο από το να συμπεριληφθεί και να προσαρμοστεί το CB άμεσα και σύμφωνα με το concept drift μέσω της ασφαλούς απομάκρυνσης ξεπερασμένων περιπτώσεων. Αντί να εμπιστεύεται όλες τις πιο πρόσφατες περιπτώσεις ο NEFCS ενσωματώνει μια μέθοδο εντοπισμού αλλαγής για να διαφοροποιήσει την συμπεριφορά σε νέες περιπτώσεις βασιζόμενος στο αν υπάρχει concept drift. Επιπρόσθετα όπως ένας online αλγόριθμος εκμάθησης ο NEFCS μπορεί να εκμεταλλευτεί άμεσα οποιαδήποτε ανάδραση του συστήματος επιτρέποντας έτσι μια άμεση αντίδραση σε ένα ενδεχόμενο concept drift. Εφόσον είναι σχεδόν αδύνατο να αυξηθεί το μέγεθος του CB χωρίς περιορισμό οι αρθρογράφοι προτείνουν μια μέθοδο διατήρησης της ικανότητας που ονομάζεται SRR για να

περιορίσουν τις απαιτήσεις σε μνήμη. Τα δύο κύρια χαρακτηριστικά της SRR είναι ότι : Πρώτον προσπαθεί να διατηρήσει την κατανομή των περιπτώσεων σε κάθε επανάληψη κάτι το οποίο είναι πολύ σημαντικό για προβλήματα με concept drift. Δεύτερον ως μια μέθοδος προσανατολισμένη στην ικανότητα η SRR διατηρεί την ικανότητα του CB δηλαδή εξασφαλίζει ότι κάθε στιγμιότυπο που απομακρύνθηκε μπορεί ακόμα να λυθεί από το CB που απομένει.

Στα πειράματα που διεξήγαγαν οι ερευνητές ανακάλυψαν τα ακόλουθα ενδιαφέροντα στοιχεία. Πρώτον όταν ένα απότομο concept drift λαμβάνει χώρα τότε ένας συγκεκριμένος βαθμός απομάκρυνσης πλεονάσματος μπορεί να βοηθήσει τον αλγόριθμο εκμάθησης στο να προσαρμοστεί πιο γρήγορα σε ένα πρωτότυπο concept. Δεύτερον πριν επιχειρηθεί να εντοπιστεί ένα concept drift ο αλγόριθμος NEFCS-SRR προσπαθεί να διατηρήσει τα προηγούμενα concept που έχει μάθει. Επομένως σε σύγκριση με τα απότομα concept drift οι δημιουργοί του πιστεύουν ότι είναι πιο κατάλληλος για τον εντοπισμό σταδιακών concept drift. Τρίτον στο πεδίο της ανεπιθύμητης αλληλογραφίας όπου το πεδίο του προβλήματος είναι πολύ αραιό μια μέθοδος επιλογής στιγμιοτύπων βασισμένη στην ικανότητα έχει ξεκάθαρο πλεονέκτημα από άλλες τεχνικές επιλογής στιγμιοτύπων. Τέταρτον παρόλο που κάθε μέθοδος συμπεριφέρεται πολύ διαφορετικά με βάση το σύνολο δεδομένων που δέχεται ως είσοδο ο NEFCS-SRR παρουσιάζει διαρκώς καλή συνολική ακρίβεια και επομένως οι δημιουργοί του υποστηρίζουν ότι είναι μια πιο γενική μέθοδος συγκριτικά με άλλες που παρουσιάζουν μεγάλες διαφορές στην ακρίβεια τους ανάλογα με το σύνολο δεδομένων που επεξεργάζονται κάθε φορά. Τέλος όσον αφορά την αποδοτικότητα όλες οι μέθοδοι με βάση την ικανότητα απαιτούν επιπλέον χρόνο για να διατηρήσουν τα μοντέλα ικανότητας. Αυτό κάνει τα μοντέλα αυτά ακατάλληλα για εφαρμογές που απαιτούν αποφάσεις σε πραγματικό χρόνο όπως ο εντοπισμός εισβολών στο διαδίκτυο όπου ένας εξαιρετικά μεγάλος όγκος πακέτων καταφθάνουν κάθε δευτερόλεπτο. Ωστόσο για το φιλτράρισμα ανεπιθύμητης αλληλογραφίας οι συνέπειες λίγο επιπλέον χρόνου είναι ασήμαντες.

Ανακεφαλαιώνοντας οι Lu et al.(2016) στην έρευνα τους καταλήγουν στο ότι η διαφοροποίηση των νέων περιπτώσεων σύμφωνα με τον εντοπισμό αλλαγής είναι καλύτερη από το να διατηρούνται ή να απορρίπτονται χωρίς

διάκριση όλες οι νέες περιπτώσεις. Επιπλέον μια ελεγχόμενη συντηρητική και βασισμένη στην ικανότητα μέθοδος είναι προτιμότερη από μια μη ελεγχόμενη ασύμμετρη μέθοδο που δεν βασίζεται στην ικανότητα για την διεξαγωγή διατήρησης της ικανότητας σε διεργασίες που είναι μεταβαλλόμενες με τον χρόνο.

2.5.2 Η μέθοδος ECUE (CBE)

2.5.2.1 Εισαγωγή

Το κόστος της ανεπιθύμητης αλληλογραφίας για τις εταιρίες παγκοσμίως ανέρχεται στο 20 δισεκατομμύρια δολάρια τον χρόνο και ο ρυθμός αύξησης του είναι της τάξης του 100% κάθε χρόνο σύμφωνα με τον J.Spira (2003). Επομένως γίνεται αντιληπτό ότι πρέπει να αντιμετωπιστεί κατάλληλα κάτι το οποίο παρουσιάζει πολλές δυσκολίες για ποικίλους λόγους. Μια από τις πιο προκλητικές διαστάσεις του προβλήματος είναι η δυναμική φύση της ανεπιθύμητης αλληλογραφίας επονομαζόμενης και ως spam. Επιπλέον ενώ τα φίλτρα προσαρμόζονται για να αντιμετωπίζουν τους σημερινούς τύπους των spam emails οι δημιουργοί των spam περιπλέκουν και μπερδεύουν τα φίλτρα με το να αποκρύπτουν τα emails τους έτσι ώστε να παρουσιάζονται ως έγκυρα. Αυτή η δυναμική φύση των spam email δημιουργεί απαιτήσεις για την ανανέωση των φίλτρων έτσι ώστε να μπορούν να είναι ικανά να αναγνωρίζουν την ανεπιθύμητη αλληλογραφία με την πάροδο του χρόνου.

Η πρόχειρη εκμάθηση (lazy learning) είναι καλή για δυναμικά μεταβαλλόμενα περιβάλλοντα. Με την πρόχειρη εκμάθηση η απόφαση στο πως να γενικεύσεις πέρα από τα δεδομένα εκπαίδευσης αναβάλλεται έως ότου να εξεταστεί κάθε νέο στιγμιότυπο. Σε σύγκριση με αυτό τα πιο σύνθετα συστήματα εκμάθησης καθορίζουν τον γενικευμένο μηχανισμό κατασκευάζοντας ένα μοντέλο βασιζόμενα στα δεδομένα εκπαίδευσης πρώτου δεχθούν νέα στιγμιότυπα. Οι Delany et al. (2005) παρουσίασαν μια νέα εφαρμογή πληροφοριών που βασίζεται σε προηγούμενες περιπτώσεις (CBR), μια πρόχειρη τεχνική μηχανικής μάθησης για το πρόβλημα της ανεπιθύμητης αλληλογραφίας με την ονομασία

ECUE (Email Classification Using Examples). Στην έρευνα τους επικεντρώθηκαν στο να αξιολογήσουν πως η ECUE μπορεί να βοηθήσει με το φαινόμενο του concept drift το οποίο είναι έμφυτο στην ανεπιθύμητη αλληλογραφία.

Η τεχνική CBR προσφέρει ένα πλήθος πλεονεκτημάτων στο τομέα του εντοπισμού της ανεπιθύμητης αλληλογραφίας. Τα spam είναι ασύνδετα πολλές φορές μεταξύ τους για παράδειγμα ένα spam το οποίο πουλάει φθηνές συνταγές για φάρμακα δεν έχει σχεδόν τίποτα κοινό με ένα spam το οποίο προσφέρει καλύτερες προσφορές σε δάνεια ακινήτων. Η κατηγοριοποίηση με βάση της περιπτώσεις λειτουργεί καλά για ασύνδετα σενάρια ενώ ο Naive Bayes μια τεχνική μηχανικής μάθησης η οποία είναι δημοφιλής για κατηγοριοποίηση κειμένων προσπαθεί να μάθει ένα ενοποιημένο σενάριο. Επιπρόσθετα υπάρχει μια φυσική ιεραρχία εκμάθησης στο CBR σύστημα όπου το απλούστερο επίπεδο εκμάθησης είναι απλά η ανανέωση του case-base με νέα στιγμιότυπα σπαμ ή κανονικών email. Το πλεονέκτημα του CBR σε αυτό το πρώτο επίπεδο εκμάθησης είναι ότι δεν απαιτεί να ξαναχτιστεί το μοντέλο όπως είναι αναγκαίο με άλλες λύσεις μηχανικής μάθησης στον τομέα του φιλτραρισμάτων email. Το δεύτερο επίπεδο της εκμάθησης είναι η επανεκπαίδευση του συστήματος με την επανεπιλογή γνωρισμάτων τα οποία μπορεί να είναι πιο προβλεπτικά για το spam. Αυτό το επίπεδο επανεκπαίδευσης μπορεί να εκτελεστεί περιστασιακά και να βασίζεται σε νέα δεδομένα εκπαίδευσης. Το υψηλότερο επίπεδο εκμάθησης το οποίο εκτελείται ακόμα πιο αραιά από ότι η επιλογή γνωρισμάτων επιτρέπει νέες τεχνικές εξαγωγής γνωρισμάτων να εισαχθούν στο σύστημα. Για παράδειγμα όταν συγκεκριμένα γνωρίσματα του πεδίου χρησιμοποιούνται στο σύστημα νέες τεχνικές εξαγωγής γνωρισμάτων θα επιτρέψουν νέα γνωρίσματα να συμπεριληφθούν. Στη ECUE μέθοδο οι δημιουργοί χρησιμοποιούν έναν αλγόριθμο ανάκτησης με βάση την ομοιότητα που βασίζεται στα δίκτυα ανάκτησης περιπτώσεων (CRN),(Lenz, Auriol & Manago, 1998) ο οποίος είναι μια δομή μνήμης που επιτρέπει την αποδοτική και ελαστική ανάκτηση περιπτώσεων. Το πλεονέκτημα της χρήσης του CRN για την εφαρμογή του δεύτερου και του τρίτου επιπέδου εκμάθησης που προαναφέρθηκαν είναι ότι μπορεί εύκολα να διαχειριστεί περιπτώσεις με νέα γνωρίσματα. Το γεγονός λοιπόν ότι αυτά τα γνωρίσματα μπορεί να λείπουν από τις παλαιότερες περιπτώσεις δεν αποτελεί πρόβλημα.

Η υπάρχουσα έρευνα για φιλτράρισμα της ανεπιθύμητης αλληλογραφίας με την βοήθεια της μηχανικής μάθησης επικεντρώνεται κυρίως στην χρήση του Naive Bayes (Androutsopoulos et al. 2000). Επιπλέον υπάρχουν εργασίες που χρησιμοποιούν μηχανές διανυσμάτων υποστήριξης (Androutsopoulos, Paliouras & Michelakis, 2004), (Drucker, Wu & Vapnik, 1999) καθώς και με λανθάνουσα σημασιολογική δεικτοδότηση (Gee 2003). Υπάρχει ακόμη έρευνα που χρησιμοποιεί κατηγοριοποιητές που βασίζονται στην μνήμη (Sakkis et al. 2004). Παρόλα αυτά όλες οι παραπάνω εργασίες δεν αντιμετωπίζουν το πρόβλημα του concept drift το οποίο είναι έμφυτο στα ανεπιθύμητα emails και οι αξιολογήσεις των ερευνών αυτών βασίζονται σε στατικά σύνολα δεδομένων. Μια τεχνική που χρησιμοποιείται για τον εντοπισμό concept drift είναι η εκμάθηση με σύνολα (ensemble learning) και το πεδίο που ασχολείται με το φιλτράρισμα ανεπιθύμητης βιβλιογραφίας την έχει χρησιμοποιήσει μαζί με την τεχνική boosting (Androutsopoulos, Paliouras & Michelakis, 2004), (Carreras & Marquez 2001) καθώς και σε συνδυασμό με τον Naive Bayes και με κατηγοριοποιητές βασισμένους στην μνήμη (Sakkis et al. 2001). Αυτές οι αξιολογήσεις όμως χρησιμοποιούν επίσης στατικά σύνολα δεδομένων και δεν επιχειρούν να εντοπίσουν το φαινόμενο του concept drift.

2.5.2.2 Η case-based μέθοδος ECUE

Σε έναν CBR αλγόριθμο εκμάθησης τα παραδείγματα στο σύνολο εκπαίδευσης παρουσιάζονται ως περιπτώσεις στην βάση περιπτώσεων (case-base). Στη μέθοδο ECUE κάθε email αποτελεί μια περίπτωση η οποία απεικονίζεται ως ένα διάνυσμα χαρακτηριστικών ή γνωρισμάτων. Οι περιπτώσεις διατάσσονται με δυαδικά γνωρίσματα που εφαρμόζονται ως τιμές Boolean. Αν το γνώρισμα υπάρχει στο email τότε η περίπτωση αναθέτει στο συγκεκριμένο γνώρισμα την τιμή αληθής διαφορετικά η τιμή του γνωρίσματος είναι ψευδής. Είναι πιο συνηθισμένο στην κατηγοριοποίηση κειμένων για λεξικά γνωρίσματα να κουβαλάνε πληροφορίες συχνότητας αλλά οι εκτιμήσεις των Delany et al. (2005) δείχνουν ότι δεν υπάρχει σημαντική διαφορά μεταξύ της χρήσης της συχνότητας της πληροφορίας και της μη χρήσης της. Μια δυαδική αναπαράσταση είναι καλύτερη για το συγκεκριμένο πεδίο λόγω της αξιοσημείωτης απόδοσης της χρήσης της συχνότητας πληροφορίας.

Επιλογή γνωρίσματος

Τα γνωρίσματα για κάθε περίπτωση αναγνωρίζονται με την χρήση μιας ποικιλίας γενικών λεξικών γνωρισμάτων πρωτίστως με τον αυτόματο χωρισμό σε λέξεις και προτάσεις (tokenisation) των email. Τα emails δεν αλλάζουν για να απομακρυνθούν οι HTML ετικέτες και έπειτα εφαρμόζονται οι τεχνικές no stop word, αποκοπής καταλήξεων (stemming) και λημματοποίησης (lemmatization). Έπειτα τα συνημμένα των email απομακρύνονται. Δεδομένού ότι τα σύνολα δεδομένων ήταν προσωπικά θεωρήθηκε ότι συγκεκριμένες επικεφαλίδες μπορεί να περιέχουν χρήσιμες πληροφορίες επομένως ένα υποσύνολο των πληροφοριών των επικεφαλίδων συμπεριλήφθηκε στην διαδικασία του tokenisation. Στο συγκεκριμένο στάδιο δεν συμπεριλήφθηκαν συγκεκριμένα γνωρίσματα του πεδίου παρόλο που προγενέστερες έρευνες από αυτή των Delany et al.(2005) έδειξαν ότι ενισχύεται η επίδοση των φίλτρων από την ένταξη τους. Ο αυτόματος χωρισμός λέξεων και προτάσεων σε 1000 emails καταλήγει σε ένα πολύ μεγάλο αριθμό γνωρισμάτων της τάξης των δεκάδων χιλιάδων. Επομένως η επιλογή γνωρισμάτων είναι αναγκαία για την μείωση των διαστάσεων του χώρου των γνωρισμάτων. Οι ερευνητές χρησιμοποίησαν το κέρδος πληροφορίας (IG) (Quinlan 1993) για να επιλέξουν τα πιο προβλεπτικά γνωρίσματα καθώς έχει αποδειχθεί ότι είναι μια τεχνική που αποδίδει στην επιθετική απομάκρυνση γνωρισμάτων για την κατηγοριοποίηση κειμένων (Yang & Pedersen 1997). Επιπλέον τα πειράματα που διεξήγαγαν οι Delany et al. (2005) μέσω της διασταυρωμένης επικύρωσης υπέδειξαν καλύτερη επίδοση στα 700 γνωρίσματα.

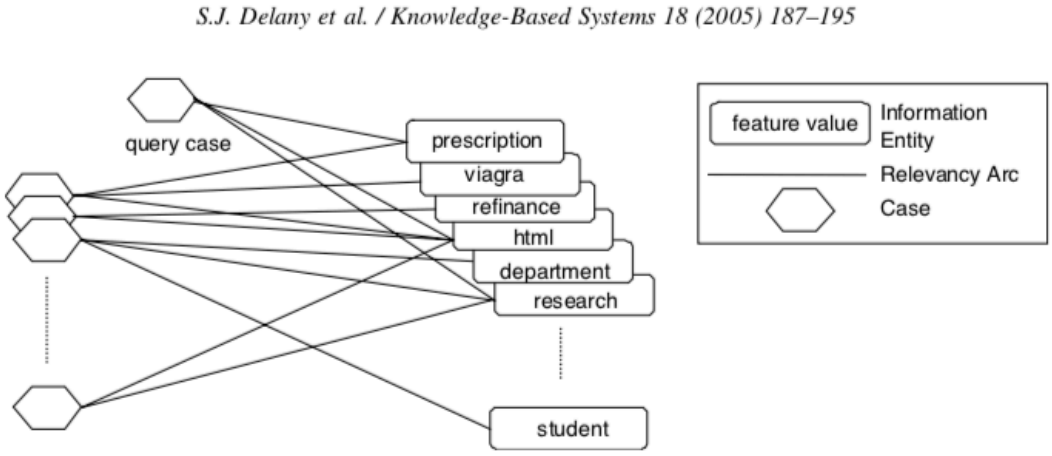
Ανάκτηση περίπτωσης

Το σύστημα χρησιμοποιεί έναν k-εγγύτερο γείτονα κατηγοριοποιητή για να ανακτήσει τις k πιο όμοιες περιπτώσεις με την στοχευμένη περίπτωση. Ο κλασικός k-NN αλγόριθμος υπολογίζει ατομικά την ομοιότητα κάθε περίπτωσης σε μια βάση περιπτώσεων (case-base) με αυτή της στοχευμένης περίπτωσης. Αυτή η προσέγγιση είναι αρκετά δυσλειτουργική σε πεδία όπου υπάρχει μείωση των τιμών των γνωρισμάτων ή υπάρχουν κενά γνωρίσματα στις περιπτώσεις. Επειδή η ανεπιθύμητη αλληλογραφία παρουσιάζει και τις δυο παραπάνω περιπτώσεις οι ερευνητές χρησιμοποιούν έναν εναλλακτικό αλγόριθμο

ανάκτησης με βάση την ομοιότητα που βασίζεται στην μέθοδο CRN η οποία διευκολύνει την αποδοτική και ελαστική ανάκτηση περιπτώσεων. Η μέθοδος CRN είναι ισοδύναμη για τις περιπτώσεις που ανακτά με τον k-NN αλγόριθμο αλλά ειδικά για αυτή την περίσταση είναι υπολογιστικά πιο αποτελεσματική.

Οι περιπτώσεις αποθηκεύονται μέσα στο CRN ως κόμβοι περιπτώσεων. Ένας δεύτερος τύπος κόμβων που ονομάζεται φορέας πληροφορίας (information entity IE) αντιπροσωπεύει ένα μοναδικό ζευγάρι τιμών γνωρισμάτων παράδειγμα 'viagra = true' υποδεικνύοντας την παρουσία της λέξης (γνώρισμα) 'viagra' μέσα στην περίπτωση. Οι κόμβοι περιπτώσεων είναι συνδεδεμένοι με τους κόμβους IE οι οποίοι τους εκπροσωπούν χρησιμοποιώντας τόξα συνάφειας. Κάθε τόξο συνάφειας μπορεί να έχει ένα βάρος το οποίο αντικατοπτρίζει την σημασία του IE κόμβου ή την σημασία του ζευγαριού τιμών γνωρισμάτων. Η CRN έχει ακόμη την έννοια της ομοιότητας τόξων μεταξύ όμοιων χαρακτηριστικών τιμών γνωρισμάτων αλλά αυτό δεν αξιοποιείται από την μέθοδο ECUE καθώς όλα τα γνωρίσματα είναι δυαδικά.

Μια στοχευμένη περίπτωση ενεργοποιεί την CRN με το να συνδέεται με τους IE κόμβους μέσω των τόξων σχετικότητας τα οποία εκπροσωπούν τις τιμές των γνωρισμάτων του. Αυτή η ενεργοποίηση επεκτείνεται από το δίκτυο στους κόμβους των περιπτώσεων κάθε ένας από τους οποίους συγκεντρώνει ένα σκορ κατάλληλο με την ομοιότητα του με την στοχευμένη περίπτωση. Οι κόμβοι περιπτώσεων με την μεγαλύτερη ενεργοποίηση είναι αυτοί που είναι πιο όμοιοι με την στοχευμένη περίπτωση.



Εικόνα 2.3 : Η μέθοδος CRN για φιλτράρισμα αλληλογραφίας (Delany et al. 2005)

Η εικόνα 2.3 απεικονίζει το CRN σύστημα για το φιλτράρισμα της ανεπιθύμητης αλληλογραφίας. Καθώς τα γνωρίσματα στην αναπαράσταση των περιπτώσεων είναι δυαδικά οι IE κόμβοι συμπεριλαμβάνονται μόνο για γνωρίσματα με τιμή αληθής. Τα τόξα σχετικότητας όλα λαμβάνουν βάρος με τιμή 1.

Ένα πολύ σοβαρό πρόβλημα για τα φίλτρα ανεπιθύμητης αλληλογραφίας είναι η παρουσία των εσφαλμένων θετικών (FP), δηλαδή όταν κανονικά emails κατηγοριοποιούνται εσφαλμένα ως ανεπιθύμητη αλληλογραφία. Ένα FP είναι αρκετά πιο σημαντικό σφάλμα από ένα FN (false negative) δηλαδή από ένα email το οποίο λανθασμένα έχει κατηγοριοποιηθεί ως κανονικό ενώ είναι ανεπιθύμητο. Για πολλούς ανθρώπους η περίπτωση των FP είναι μη αποδεκτή. Για αυτό τον λόγο ο κατηγοριοποιητής χρησιμοποίησε ομόφωνη ψηφοφορία για να καθορίσει αν η υπό εξέταση περίπτωση αποτελεί spam ή όχι. Για να αποφευχθεί η παρουσία των FP όλοι οι γείτονες της υπό εξέταση περίπτωσης πρέπει να έχουν κατηγοριοποιήσει ένα spam για να μπορέσει να κατηγοριοποιηθεί και αυτό ως spam.

Διαχείριση της βάσης περιπτώσεων

Η μέθοδος που προτείνουν οι Delany et al.(2005) για την διαχείριση του concept drift είναι η επιλογή στιγμιοτύπων. Με την πάροδο του χρόνου η βάση περιπτώσεων πρέπει να ενημερώνεται για να συμπεριλάβει νέους τύπους ανεπιθύμητων ηλεκτρονικών μηνυμάτων αλλά και μηνυμάτων που είναι κανονικά. Δεδομένου του όγκου των ηλεκτρονικών μηνυμάτων που μπορεί να λάβει ένας μόνο χρήστης μέσα σε μια εβδομάδα είναι εμφανές ότι υπάρχει ανάγκη της διαχείρισης του συνόλου εκπαίδευσης. Η προηγούμενη τους έρευνα στις τεχνικές διόρθωσης του case-base ανακάλυψε μια μέθοδο της διόρθωσης της βάσης περιπτώσεων που χρησιμοποιεί χαρακτηριστικά ικανότητας μιας case-base για να απομακρύνει θορυβώδεις περιπτώσεις καθώς και περιττές περιπτώσεις. Αυτή η μέθοδος ονομάζεται Competence-Bases editing (CBE).

Η CBE μέθοδος αρχικά κατασκευάζει ένα μοντέλο ικανότητας της βάσης περιπτώσεων αναγνωρίζοντας για κάθε περίπτωση τις ιδιότητες ικανότητας δηλαδή αυτές οι περιπτώσεις που συμβάλλουν στην σωστή κατηγοριοποίησης και αυτές που συμβάλλουν σε λάθος κατηγοριοποιήσεις. Έπειτα η τεχνική

αποτελείται από δυο στάδια επεξεργασίας. Στο πρώτο πραγματοποιείται η διαδικασία ενίσχυσης της ικανότητας με την απομάκρυνση των περιπτώσεων που αποτελούν θόρυβο χρησιμοποιώντας τις ιδιότητες ικανότητας που προαναφέρθηκαν. Το δεύτερο στάδιο δηλαδή το στάδιο διατήρησης της ικανότητας αναγνωρίζει και απομακρύνει τις περιττές περιπτώσεις (δηλαδή περιπτώσεις που βρίσκονται στο κέντρο μιας μεγάλης συστάδας περιπτώσεων στην ίδια κατηγορία). Το πλεονέκτημα της CBE μεθόδου όταν αυτή εφαρμόζεται στο πεδίο των ανεπιθύμητων ηλεκτρονικών μηνυμάτων σύμφωνα με τους Delany et al.(2005) είναι ότι καταλήγει σε ένα συντηρητικό κούρεμα του CB το οποίο οδηγεί σε μεγαλύτερες βάσεις περιπτώσεων αλλά με καλύτερη γενικευμένη ακρίβεια σε σύγκριση με άλλες τυπικές τεχνικές αλλαγής περιπτώσεων (Delany & Cunningham 2004).

2.5.2.3 Αξιολόγηση - Συμπεράσματα

Εφόσον τα εσφαλμένα θετικά FP είναι πιο σοβαρά από ότι τα FN οι Delany et al. (2005) κατέληξαν στο ότι η ακρίβεια (ή το σφάλμα) ως μέτρηση της απόδοσης δεν δίνουν την πλήρη εικόνα υπό την έννοια ότι δεν υπάρχει διαύγεια όσον αφορά το πλήθος των FP και των FN. Επομένως δύο φίλτρα με παρόμοια ακρίβεια μπορεί να έχουν πολύ διαφορετικά ποσοστά FP και FN. Προηγούμενες εργασίες στο συγκεκριμένο πεδίο χρησιμοποίησαν διάφορες μετρικές για να αναφέρουν την απόδοση. Η πιο κοινή μετρική απόδοσης είναι η ακρίβεια και η ανάκληση (Gee 2003). Οι Sakkis et al.(2004) εισήγαγαν μια ακρίβεια με βάρη η οποία ενσωματώνει μια μέτρηση του πόσο περισσότερο κόστος έχει ένα FP από ότι ένα FN. Παρόλο που αυτές οι μετρικές είναι χρήσιμες για σκοπούς σύγκρισης τα ποσοστά των FP και FN δεν είναι καθαρά οπότε η αληθινή αποτελεσματικότητα του κατηγοριοποιητή δεν είναι προφανής. Για τον λόγο αυτό επομένως οι Delany et al. (2005) χρησιμοποιούν τον δείκτη τον FP, τον δείκτη τον FN και τον μέσο δείκτη σφάλματος της κλάσης που είναι ο $Error = \frac{FPRate + FNRate}{2}$.

Ο αλγόριθμος Naive Bayes εμφανίζεται να είναι η επικρατέστερη επιλογή για το φιλτράρισμα ανεπιθύμητης αλληλογραφίας. Επομένως οι δημιουργοί του ECUE

συγκρίναν την μέθοδο του με τον Naive Bayes κατηγοριοποιητή. Πειράματα που διεξήχθησαν σε ένα αριθμό στατικών σύνολων δεδομένων δεν υπέδειξαν έναν αλγόριθμο από τους δύο να είναι καλύτερος από τον άλλον διαρκώς (Delany , Cunningham & Coyle, 2004). Περισσότερο ενδιαφέρον παρουσίασαν οι αξιολογήσεις που πραγματοποιήθηκαν για μια χρονική διάρκεια επιτρέποντας το σύστημα να ενημερώνει δυναμικά τα δεδομένα εκπαίδευσης του με παραδείγματα spam email αλλά και κανονικά τα οποία ταξινομήθηκαν λάθος. Ένα πλήθος πειραμάτων διεξήχθησαν αλλάζοντας από την μη ενημέρωση στα αρχικά CB δεδομένα εκπαίδευσης στην ενημέρωση μιας επεξεργασμένης case-base σε μηνιαία εβδομαδιαία και ημερήσια βάση με αυτά τα emails τα οποία είχαν κατηγοριοποιηθεί εσφαλμένα σε μια συγκεκριμένη χρονική περίοδο. Η αξιολόγηση έδειξε ότι η καλύτερη απόδοση παρουσιάστηκε όταν ενημερωνόταν η επεξεργασμένη CB σε ημερήσια βάση με όποιο email είχε κατηγοριοποιηθεί λάθος εκείνη την ημέρα. Τα ίδια πειράματα εκτελέστηκαν χρησιμοποιώντας τον Naive Bayes κατηγοριοποιητή σε ανεπεξέργαστα δεδομένα εκπαίδευσης. Τα δεδομένα εκπαίδευσης δεν μπορούν να επεξεργαστούν για έναν Naive Bayes κατηγοριοποιητή καθώς η τεχνική επεξεργασίας είναι για μεθόδους που βασίζονται στην ικανότητα και οι οποίες χρησιμοποιούν έναν k-NN κατηγοριοποιητή για να καθορίσουν την ικανότητα κάθε περίπτωσης στη CB και να αναλύσουν τις ιδιότητες ικανότητας των περιπτώσεων καθορίζοντας με αυτό τον τρόπο ποιες πρέπει να απομακρυνθούν. Λόγω της σημασίας των FP σε αυτό το πεδίο ο Naive Bayes κατηγοριοποιητής διαμορφώθηκε να είναι μεροληπτικός πέρα των FP. Παρόλο που ο Naive Bayes έχει καλύτερο ολικό δείκτη σφάλματος πάνω στα δεδομένα που χρησιμοποιήθηκαν χωρίς ενημέρωση το CBR σύστημα αποδίδει καλύτερα σε όλα τα σύνολα δεδομένων όταν ενημερώνονται δυναμικά τα δεδομένα για να μάθουν από email που έχουν ταξινομηθεί λάθος. Συνεπώς μπορεί να εξαχθεί το συμπέρασμα ότι η ημερήσια ενημέρωση των δεδομένων εκπαίδευσης με emails που ταξινομήθηκαν λάθος βελτιώνει την απόδοση του CBR συστήματος αλλά έχει μια γενικά επιβλαβή επίδραση στον Naive Bayes κατηγοριοποιητή. Ο Naive Bayes με την ημερήσια ενημέρωση βελτιώνει τον δείκτη των FP πιο σημαντικά από ότι ο ECUE αλλά η επιδείνωση του FN δείκτη έχει μια γενικά αρνητική επίδραση στην απόδοση. Αξίζει επίσης να σημειωθεί ότι η ενημέρωση ενός συστήματος που χρησιμοποιεί τον Naive Bayes με κάθε νέο

δεδομένο εκπαίδευσης απαιτεί μια ξεχωριστή διαδικασία εκμάθησης για τον επαναυπολογισμό των πιθανοτήτων όλων των γνωρισμάτων. Αντίθετα η ενημέρωση ενός CBR συστήματος όπως είναι ο ECUE αλγόριθμος με νέα δεδομένα εκπαίδευσης απαιτεί απλά οι νέες περιπτώσεις να προστεθούν στο CB.

Οι Delany et al.(2005) μέσω των πειραμάτων τους έδειξαν ότι η χρήση ενός πρόχειρου αλγορίθμου εκμάθησης μπορεί να διαχειριστεί το φαινόμενο του concept drift το οποίο είναι έμφυτο στα δεδομένα που αποτελούν ανεπιθύμητη αλληλογραφία. Η προσέγγιση τους είναι η ενημέρωση του CB στο τέλος κάθε ημέρας με περιπτώσεις που είχαν ταξινομηθεί λάθος από το σύστημα εκείνη την ημέρα και η περιοδική ανακατασκευή του CB χρησιμοποιώντας τις πιο πρόσφατες περιπτώσεις για την επανεπιλογή γνωρισμάτων. Αυτή η προσέγγιση αποδίδει καλύτερα από ότι η τυπική προσέγγιση που βασίζεται στα παράθυρα και στα δύο επίπεδα της ενημέρωσης του μοντέλου.

2.5.3 Η μέθοδος ICF

2.5.3.1 Εισαγωγή

Ο κατηγοριοποιητής που ονομάζεται εγγύτερος γείτονας είναι ένα απλό επιτηρούμενο σχέδιο εκμάθησης το οποίο κατηγοριοποιεί άγνωστα στιγμιότυπα βρίσκοντας το εγγύτερο στιγμιότυπο που έχει παρατηρηθεί , σημειώνοντας την κλάση του και προβλέποντας την κλάση των άγνωστων στιγμιοτύπων (Cover & Hart 1967). Όλοι οι αλγόριθμοι που αναπτύσσουν αυτό το σύστημα κατηγοριοποίησης παρουσιάζουν το ίδιο πρόβλημα: τα στιγμιότυπα τα οποία χρησιμοποιούνται για την εκπαίδευση του κατηγοριοποιητή αποθηκεύονται χωρίς διάκριση. Καμία διαδικασία επιλογής δεν εφαρμόζεται και ως αποτέλεσμα επιβλαβή και περιττά στιγμιότυπα αποθηκεύονται άσκοπα. Παραβλέποντας αυτό το πρόβλημα το σύστημα αυτής της κατηγοριοποίησης είναι απλό και πολύ αποτελεσματικό σε σύγκριση με άλλες μεθόδους όπως τα εμπροσθοτροφοδοτούμενα νευρωνικά δίκτυα ή τα δέντρα απόφασης (King, Feng & Sutherland, 1995).

Στο άρθρο τους οι Brighton και Mellish (2002) προσπαθούν να ελαττώσουν αυτό το πρόβλημα και ελέγχουν τα κριτήρια που χρησιμοποιούνται για την επιλεκτική αποθήκευση στιγμιοτύπων στο πρόβλημα της κατηγοριοποίησης. Το έργο τους ρίχνει βάρος στην απόκτηση διορατικότητας σχετικά με την δομή των προβλημάτων κατηγοριοποίησης γενικά. Θεωρώντας τα στιγμιότυπα ως διανύσματα γνωρισμάτων τότε μπορούμε να αντιληφθούμε τον χώρο των στιγμιοτύπων όπου κάθε στιγμιότυπο θα αποτελεί ένα σημείο. Επίσης η δομή των κλάσεων που δημιουργούνται από τα στιγμιότυπα μπορεί να είναι πολύ διαφορετική σε κάθε πρόβλημα το οποίο προφανώς δημιουργεί ασυνέχεια όταν εφαρμόζουμε ένα συγκεκριμένο σχέδιο επιλογής στιγμιοτύπων πάνω σε διαφορετικά προβλήματα. Οι αρθρογράφοι εξηγούν τις πιθανές δομές κλάσεων και πως αυτές μπορούν να ομαδοποιηθούν διότι όπως επισημαίνουν η γνώση της δομής των κλάσεων είναι ένα αναπόσπαστο τμήμα του σχεδιασμού και της ανάπτυξης ενός αλγορίθμου επιλογής στιγμιοτύπων. Με βάση την δομή του χώρου των στιγμιοτύπων παρουσιάζουν τον αλγόριθμο ICF

2.5.3.2 Διάτρηση της ικανότητας της κατηγοριοποίησης

Γενικά το πρόβλημα της επιλογής στιγμιοτύπων καθορίζεται ως η ανάγκη να εξαχθούν τα πιο χρήσιμα σύνολα στιγμιοτύπων από την βάση δεδομένων για την οποία γνωρίζουμε ότι περιέχει στιγμιότυπα τα οποία είναι περιττά ή επιβλαβή. Στο πλαίσιο της εκμάθησης με βάση τα στιγμιότυπα αναζητούμε να απορρίψουμε τις περιπτώσεις οι οποίες είναι περιττές ή επιβλαβείς για την διαδικασία της κατηγοριοποίησης. Στο πλαίσιο αυτό επομένως το πρόβλημα της επιλογής στιγμιοτύπων ανάγεται ουσιαστικά σε πρόβλημα διαγραφής στιγμιοτύπων. Στην περίπτωση που οι αποφάσεις διαγραφής δεν μειώνουν την ακρίβεια της κατηγοριοποίησης πρέπει να είναι ξεκάθαρο το είδος των αποφάσεων διαγραφής. Οι παρακάτω λόγοι είναι αυτοί για τους οποίους ένας κατηγοριοποιητής που χρησιμοποιεί τον k εγγύτερο γείτονα μπορεί να κατηγοριοποιήσει λανθασμένα ένα στιγμιότυπο :

- Όταν ο θόρυβος είναι παρών τοπικά για το υπό εξέταση στιγμιότυπο. Τα θορυβώδη στιγμιότυπα κερδίζουν με τον κανόνα της πλειοψηφίας και ως αποτέλεσμα προβλέπεται λανθασμένη κλάση για το στιγμιότυπο.

- Όταν το υπό εξέταση στιγμιότυπο καταλαμβάνει μια θέση κοντά στα σύνορα μεταξύ των κλάσεων όπου η διάκριση είναι δυσκολότερη εξαιτίας της παρουσίας πολλαπλών κλάσεων.

- Όταν η περιοχή που καθορίζει την κλάση ή τμήμα της κλάσης είναι τόσο μικρή που τα στιγμιότυπα που ανήκουν στην κλάση και περιβάλλουν το τμήμα κερδίζουν με την ψήφο της πλειοψηφίας. Αυτή η κατάσταση εξαρτάται από την τιμή του k.

- Όταν το πρόβλημα είναι άλυτο από έναν αλγόριθμο εκμάθησης βασισμένο στα στιγμιότυπα. Αυτό θα είναι εξαιτίας της φύσης της υποκείμενης συνάρτησης ή εξαιτίας του προβλήματος των αραιών δεδομένων.

Στο πλαίσιο της επιλογής στιγμιοτύπων μπορεί να αντιμετωπιστεί το πρώτο πρόβλημα και να επιχειρήσουμε να βελτιώσουμε την ακρίβεια απομακρύνοντας τον θόρυβο. Το τέταρτο πρόβλημα δεν μπορεί όμως να επιλυθεί διότι εκφράζει την ενδογενή δυσκολία του προβλήματος. Ωστόσο το δεύτερο και το τρίτο πρόβλημα καθορίζουν της απόφασης απομάκρυνσης. Η απομάκρυνση στιγμιοτύπων που είναι κοντά στα όρια δεν προτείνεται διότι αυτά τα στιγμιότυπα είναι σχετικά με την διάκριση μεταξύ των κλάσεων. Προσοχή χρειάζεται το 3ο πρόβλημα αλλά καθώς το k είναι τυπικά μικρό η εμφάνιση ενός τέτοιου προβλήματος είναι σπάνια. Ένα σημαντικό σημείο που σημειώνεται από τους Brighton και Mellish (2002) και το οποίο δεν υπήρχε έως τότε στην βιβλιογραφία είναι ότι μπορεί κάποιος να θέσει ένα θεωρητικό όριο στο πως ένας αλγόριθμος μείωσης εκτελείται. Στην πράξη επιλέγουν ένα τυχαίο δείγμα του προβλήματος της κατηγοριοποίησης και κρατάνε αυτά τα στιγμιότυπα για να ελέγξουν την ακρίβεια του εγγύτερου γείτονα. Έπειτα δοσμένης μια βάσης στιγμιοτύπων I δημιουργούν τα σύνολα ελέγχου και εκπαίδευσης. Κάνοντας τις ακόλουθες υποθέσεις :

1. | training | > | testing |
2. Υποθέτουμε τον κατηγοριοποιητή εγγύτερο γείτονα με k = 1.

Έπειτα μετά το φιλτράρισμα του συνόλου εκπαίδευσης το μέγιστο πλήθος στιγμιοτύπων στην εκπαίδευση που απαιτείται από τον κατηγοριοποιητή για να διατηρήσει την αρχική του ακρίβεια είναι στην ουσία όσο το πλήθος του συνόλου ελέγχου. Αυτό το αποτέλεσμα συνάγεται καθώς για κάθε στιγμιότυπο στο σύνολο ελέγχου από το οποίο εξάγουμε την ακρίβεια χρειαζόμαστε μόνο μια περίπτωση στο σύνολο εκπαίδευσης για να ταξινομήσουμε σωστά αυτό το ένα στιγμιότυπο ελέγχου. Αυτό το αποτέλεσμα μπορεί να χρησιμοποιηθεί σαν οδηγός για να ελέγχεται αν ο αλγόριθμος αποδίδει όπως θα έπρεπε. Το ελάχιστο πλήθος στιγμιοτύπων που απαιτούνται για να διατηρηθεί η ακρίβεια του κατηγοριοποιητή στο σύνολο ελέγχου μας δίνει μια εικόνα για το πόσο εύκολο είναι το πρόβλημα.

2.5.3.3 Η δομή του χώρου στιγμιοτύπων

Παραδοσιακά ο τρόπος με τον οποίο σημαντικά στιγμιότυπα αναγνωρίζονται σε έναν χώρο στιγμιοτύπων έχει θεωρηθεί ότι εφαρμόζεται σε όλα τα προβλήματα κατηγοριοποίησης. Ο στόχος είναι η εύρεση ενός αλγόριθμου ο οποίος μπορεί να εφαρμοστεί σε οποιοδήποτε πεδίο κάτι το οποίο αμφισβητούν οι Brighton και Mellish (2002) προτείνοντας δύο ευρείς κατηγορίες δομών κλάσεων οι οποίες όμως απαιτούν διαφορετική προσέγγιση.

Η πλειοψηφία των προβλημάτων ειδικά στο πεδίο της εξόρυξης γνώσης υπάγονται σε μια συγκεκριμένη κατηγορία. Αυτή η κατηγορία περιέχει χώρους στιγμιοτύπων που οι κλάσεις τους οριοθετούνται από ομογενείς περιοχές στιγμιοτύπων. Όμως υπάρχει και η κατηγορία προβλημάτων που αποτελείται από χώρους στιγμιοτύπων που δεν είναι ομογενείς. Στο παρελθόν ο χαρακτηρισμός ενός σημαντικού στιγμιοτύπου δεν εξαρτιόταν από το πρόβλημα κυρίως λόγο του ότι σπάνια παρουσιαζόντουσαν προβλήματα με μη ομογενείς δομές των κλάσεων. Σε ένα ομογενή χώρο στιγμιοτύπων οι ερευνητές συμφωνούν στο ότι τα πιο σημαντικά στιγμιότυπα είναι αυτά που βρίσκονται στα σύνορα μεταξύ των κλάσεων ενώ αυτά που εντοπίζονται στο εσωτερικό των κλάσεων είναι συνήθως περιττά και η απομάκρυνση τους δεν εμποδίζει τον

εγγύτερο γείτονα να κάνει διάκριση μεταξύ των κλάσεων. Στιγμιότυπα με μεγάλη χρησιμότητα μπορεί να αποτελούν συνοριακά σημεία αλλά αυτό δεν αποτελεί εγγύηση διότι ο τρόπος με τον οποίο αναγνωρίζουμε αυτά τα στιγμιότυπα δεν καθοδηγείται από την προϋπόθεση ότι τα συνοριακά σημεία είναι κρίσιμα. Το πρόβλημα με τις μεθόδους που βασίζονται στην χρησιμότητα είναι ότι απαιτείται γνώση για την προηγούμενη χρήση των στιγμιοτύπων. Όταν τα στιγμιότυπα δεν έχουν χρησιμοποιηθεί μπορεί να υπάρχει μια λανθασμένη μέτρηση της χρησιμότητας τους. Πράγματι τα συνοριακά σημεία είναι λιγότερο πιθανό να εξαιρεθούν από αυτό το πλαίσιο διότι τα εσωτερικά σημεία εξ ορισμού περιστοιχίζονται από περιπτώσεις της δικιάς τους κλάσης και επομένως έχουν μεγάλη πιθανότητα να προβλέψουν ορθά ένα στιγμιότυπο.

Οι Brighton και Mellish (2002) καθορίζουν μια μη ομογενή κλάση ως μια κλάση που ορίζεται από ένα σύνολο στιγμιοτύπων που δεν μοιράζονται την ίδια τοπικότητα. Σε αυτή την περίπτωση η έννοια του συνόρου δεν έχει νόημα. Επομένως σε αυτή την περίπτωση η διατήρηση μόνο πρωτότυπων στιγμιοτύπων είναι ο ασφαλέστερος δρόμος για την απομάκρυνση ενός πλήθους στιγμιοτύπων.

Συνοψίζοντας αποδεικνύεται ότι η φύση των κρίσιμων περιπτώσεων εξαρτάται από την δομή των κλάσεων. Στην πλειοψηφία των περιπτώσεων τα προβλήματα υπάγονται στην κατηγορία των ομογενών στιγμιοτύπων. Όμως υπάρχουν και άλλοι τύποι δομών στις κλάσεις. Δίνοντας προτίμηση στις ομογενείς δομές κλάσεων οι Brighton και Mellish (2002) υποστηρίζουν ότι τα πρωτότυπα μπορεί να είναι καλοί κατηγοριοποιητές διότι μπορούν να ταξινομήσουν πολλά στιγμιότυπα στον χώρο των στιγμιοτύπων. Ωστόσο δεν αποτελούν καλούς διακριτοποιητές.

2.5.3.4 Η μέθοδος ICF

Μια πρωτοποριακή προσέγγιση στη διατήρηση της ικανότητας είναι η πολιτική διαγραφής ίχνους των Smyth και Keane (1995) η οποία είναι ένα σύστημα φιλτραρίσματος το οποίο σχεδιάστηκε για χρήση μέσα στην δομή του CBR (Case-Based Reasoning). Η προσέγγιση αυτή αναφέρεται εδώ καθότι είναι σχετική με την μέθοδο ICF. Σε προηγούμενη έρευνα (Brighton 1997) οι ερευνητές

έδειξαν ότι κάποια από τα σενάρια που εισήχθησαν από τους Smyth και Keane (1995) μεταφέρονται στο απλούστερο πλαίσιο του αλγορίθμου ταξινόμησης εγγύτερος γείτονας. Η CBR ως γνωστόν αποτελεί μια προσέγγιση επίλυσης , συλλογιστικής και οργάνωσης διεργασιών με βάση προηγούμενες λύσεις (Kolodner 1993). Οι τεχνικές λεπτομέρειες είναι σχετικά ίδιες με την εκμάθηση με βάση τα στιγμιότυπα παρόλο που η ιδέα της προσαρμογής της περίπτωσης χρησιμοποιείται συνήθως ως μετρική ομοιότητας. Ένα CBR σύστημα σκοπεύει να επιλύσει μια νέα διεργασία προσαρμόζοντας λύσεις που έχουν αποθηκευτεί με τέτοιο τρόπο ώστε να μπορούν να εφαρμοστούν στο νέο πρόβλημα. Το περισσότερο από το έργο των Smyth και Keane (1995) βασίζεται στην έννοια της προσαρμογής των περιπτώσεων. Χρησιμοποιούν την ιδιότητα $Adaptable(c, c')$ για να ορίσουν το ότι η περίπτωση c μπορεί να προσαρμοστεί από την περίπτωση c'. Επομένως γενικά μπορούμε να πούμε ότι μπορούμε να διαγράψουμε μια περίπτωση για την οποία υπάρχουν πολλές άλλες περιπτώσεις που μπορούν να προσαρμοστούν σε αυτή. Σε προηγούμενη έρευνα (Brighton 1996) εισήχθηκε η έννοια του Local-Set μιας περίπτωσης c. Το Local-Set μιας περίπτωσης c ορίζεται να είναι το σύνολο των περιπτώσεων που περιέχονται στην μεγαλύτερη δυνατή υπερσφαίρα με κέντρο το c έτσι ώστε μονο περιπτώσεις της ίδιας κλάσης με της c να περιέχονται σε αυτήν.

Η πρωτοτυπία του έργου των Smyth και Keane (1995) προέρχεται από την ταξινομία που πρότειναν για τις ομάδες των περιπτώσεων. Καθορίζοντας τέσσερις κατηγορίες περιπτώσεων που αντικατοπτρίζουν την συνολική προσφορά στην ικανότητα που παρέχει κάθε περίπτωση καταφέρνουμε να κερδίσουμε μια εικόνα για την επίδραση που έχει η απομάκρυνση μιας περίπτωσης. Οι Brighton και Mellish (2002) καθορίζουν αυτές τις κατηγορίες με την βοήθεια δυο ιδιοτήτων που ονομάζονται κάλυψη και προσβασιμότητα. Αυτές οι ιδιότητες είναι σημαντικές καθώς η σχέση μεταξύ τους έχει χρησιμοποιηθεί σε σημαντικές έρευνες. Για μια βάση περιπτώσεων $CB = \{c_1, c_2, \ldots, c_n\}$ καθορίζονται οι κάλυψη και η προσβασιμότητα να είναι τα ακόλουθα σύνολα :

$$Coverage(c) = \{c' \in CB : Adaptable(c, c')\}$$
$$Reachable(c) = \{c' \in CB : Adaptable(c', c)\}$$

Με βάση αυτές τις δύο ιδιότητες μπορούν να καθοριστούν τέσσερα σύνολα με χρήση της θεωρίας συνόλων. Για παράδειγμα μια περίπτωση στο βασικό σύνολο καθορίζεται ως η περίπτωση που το σύνολο προσβασιμότητα της είναι κενό $Reachable(c) = \varnothing$.

Με βάση όλα τα παραπάνω η έρευνα των Brighton και Mellish (2002) διαφέρει από την διαγραφή ίχνους (Smyth & Keane 1995) μόνο στην αντικατάσταση της ιδιότητας Adaptable με αυτή του Local-Set. Αν μια περίπτωση c μπορεί να προσαρμοστεί σε μια περίπτωση c' εξαρτάται στο κατά πόσο η c είναι σχετική με την λύση της c'. Στην απλή μάθηση αυτό σημαίνει ότι η c είναι εγγύτερος γείτονας της c'. Ωστόσο δεν μπορούμε να υποθέσουμε ότι μια περίπτωση διαφορετικής κλάσης είναι σχετική με την λύση (σωστή πρόβλεψη) της c'. Επομένως οριοθετούμε την γειτονία της c' με την πρώτη περίπτωση που ανήκει σε διαφορετική κλάση. Με βάση αυτή την παράλληλη δομή βρίσκεται ότι η διαγραφή ίχνους αποδίδει καλά. Ακόμα πιο πολύ ενδιαφέρων παρουσιάζει το γεγονός ότι μια απλούστερη μέθοδος η οποία χρησιμοποιεί την local-set ιδιότητα και όχι την ταξινομία αποδίδει το ίδιο καλά. Με την διαγραφή με χρήση του local-set διαγράφονται περιπτώσεις με μεγάλα local-sets καθώς αυτές οι περιπτώσεις που έχουν τέτοια μεγάλα σετ εντοπίζονται στο εσωτερικό των περιοχών των κλάσεων. Το πρόβλημα έγκειται στο πόσες περιπτώσεις πρέπει να διαγραφούν. Οι Brighton και Mellish (2002) επέλεξαν να αξιοποιήσουν την μεθοδολογία των Smyth και Keane (1995) επιβάλλοντας ένα όριο το οποίο είναι προκαθορισμένη μεταβλητή. Αυτό έρχεται σε αντίθεση με άλλους αλγορίθμους οι οποίοι αποφασίζουν δυναμικά για το πότε θα σταματήσουν να απομακρύνουν περιπτώσεις.

Ο επαναλαμβανόμενος αλγόριθμος φιλτραρίσματος περιπτώσεων ICF (iterative case filtering) χρησιμοποιεί για τις περιπτώσεις τις ιδιότητες κάλυψης και προσβασιμότητας όταν μεταφέρει την πολιτική διαγραφής ίχνους που αναπτύχθηκε παραπάνω. Όπως ο επαναλαμβανόμενος αλγόριθμος του Wilson οι Brighton και Mellish (2002) εφαρμόζουν έναν κανόνα που αναγνωρίζει τις περιπτώσεις που θα πρέπει να διαγραφούν. Αυτές οι περιπτώσεις έπειτα απομακρύνονται και ο κανόνας εφαρμόζεται εκ νέου επαναληπτικά έως ότου καμία πλέον από τις περιπτώσεις δεν ικανοποιεί τις προϋποθέσεις του κανόνα. Ο

αλγόριθμος ICF χρησιμοποιεί τα σύνολα κάλυψης και προσβασιμότητας τα οποία μπορούν να παρομοιαστούν ως τα σύνολα γειτνίασης και συνεργατών (associates) που χρησιμοποιούνται από τους Wilson και Martinez (1997). Μια βασική διαφορά είναι ότι το σύνολο προσβασιμότητας δεν έχει προκαθορισμένο μέγεθος αλλά οριοθετείται από την εγγύτερη περίπτωση μιας διαφορετικής κλάσης. Αυτή η διαφορά είναι καίριας σημασίας καθώς ο αλγόριθμος ICF εξαρτάται από τα σχετικά μεγέθη αυτών των συνόλων. Ο κανόνας διαγραφής του ICF είναι απλός. Απομακρύνονται οι περιπτώσεις που το σύνολο προσβασιμότητας τους είναι μεγαλύτερο από το σύνολο κάλυψης τους. Μια πιο διαισθητική ανάγνωση του παραπάνω κανόνα είναι ότι είναι μια περίπτωση c απομακρύνεται όταν οι περιπτώσεις που μπορεί να λύσει είναι λιγότερες από αυτές που μπορούν να λύσουν την c. Αυτές οι περιπτώσεις θα είναι μακρύτερα από τα όρια της κλάσης καθώς τα σύνολα προσβασιμότητας τους θα είναι μεγάλα. Αφότου απομακρυνθούν αυτές οι περιπτώσεις ο χώρος των περιπτώσεων θα περιέχει ουσιαστικά πιο πυκνά σχήματα περιπτώσεων σε κάθε πλευρά των συνόρων των κλάσεων. Αυτό λοιπόν είναι το κριτήριο διαγραφής που χρησιμοποιεί ο αλγόριθμος και η διαδικασία προχωράει με τον επαναλαμβανόμενο υπολογισμό αυτών των ιδιοτήτων αφού έχει πραγματοποιηθεί το φιλτράρισμα. Συνήθως όλο και περισσότερες περιπτώσεις θα αρχίσουν να πληρούν τα κριτήρια καθώς η ελαχιστοποίηση προχωρά και τα σχήματα που περιβάλλουν τα όρια των κλάσεων θα γίνονται ολοένα και πιο στενά.

Ο αλγόριθμος αυτός όπως και η πλειοψηφία των αλγορίθμων που επικεντρώνονται στην απομάκρυνση των περιττών περιπτώσεων είναι πιθανόν να προστατεύουν τις περιπτώσεις που αποτελούν θόρυβο. Τα σύνολα προσβασιμότητας και κάλυψης μιας περίπτωσης που αποτελεί θόρυβο θα περιέχουν ένα μόνο στοιχείο (singleton). Αυτή η ιδιότητα προστατεύει τις περιπτώσεις αυτές από το να απομακρυνθούν. Για τον λόγο αυτό οι Brighton και Mellish (2002) εφαρμόζουν το φιλτράρισμα θορύβου που βασίζεται στην διαγραφή του Wilson και έχει υιοθετηθεί από τους Wilson και Martinez (1997).

The Iterative Case Filtering Algorithm: ICF (Brighton & Mellish 2002)

1: Εκτέλεσε την διαγραφή του Wilson
2: **for all** $x \in T$ **do**
3: **if** x κατηγοριοποιείται λάθος από τον k εγγύτερο γείτονα **then**
4: σημείωσε το x για απομάκρυνση
5: **for all** $x \in T$ **do**
6: **if** x σημειώθηκε για απομάκρυνση τότε $T = T - \{x\}$
7: επανάληψη έως ότου καμία άλλη περίπτωση δεν σημειώνεται
8: **repeat**
9: **for all** $x \in T$ **do**
10: υπολογισμός του reachable(x)
11: υπολογισμός του coverage(x)
12: progress = false
13: **for all** $x \in T$ **do**
14: **if** |reachable(x)| > |coverage(x)| **then**
15: σημείωσε το x για απομάκρυνση
16: progress = true
17: **for all** $x \in T$ **do**
18: **if** x σημειώθηκε για απομάκρυνση **then** $T = T - \{x\}$
19: until not progress
20: **return** T

Οι γραμμές 2 έως 6 του παραπάνω ψευδοκώδικα εκτελούν την διαγραφή του Wilson. Το υπόλοιπο του αλγορίθμου (7-20) επικεντρώνεται στην απομάκρυνση των περιττών περιπτώσεων με τον τρόπο που περιγράφεται πιο πάνω. Υπάρχει ένας έλεγχος (μεταβλητή progress) για να βεβαιωθεί ότι υπάρχει πρόοδος έπειτα

από κάθε επανάληψη. Ο αλγόριθμος είναι μειωτικός (decremental) όπως η οικογένεια των RT αλγορίθμων (Wilson & Martinez 1997) αλλά διαφέρει στο ότι περισσότερα από ένα περάσματα απαιτούνται για να ελαχιστοποιηθεί το σύνολο δεδομένων.

2.5.3.5 Αξιολόγηση - Συμπεράσματα

Η αξιολόγηση του ICF πραγματοποιήθηκε πάνω σε 30 σύνολα δεδομένων από το UCI αποθηκευτήριο (Blake & Merz 1998). Σύμφωνα με τους Brighton και Mellish (2002) η εξέταση του πλήθους των επαναλήψεων που κάνει ο αλγόριθμος πάνω σε ένα σύνολο δεδομένων και συγκεκριμένα το ποσοστό των περιπτώσεων που απομακρύνονται μετά από κάθε επανάληψη μας παρέχει μια σημαντική εικόνα για την φύση της δομής του χώρου των περιπτώσεων του συγκεκριμένου συνόλου δεδομένων. Για παράδειγμα στο σύνολο switzerland που εφαρμόστηκε ο αλγόριθμος κατά την αξιολόγηση του χρειάστηκαν 17 επαναλήψεις και σε κάθε μια από αυτές ένας μικρός αριθμός περιπτώσεων απομακρύνονταν κάθε φορά. Αυτή η παρατήρηση οδήγησε τους δημιουργούς του ICF στο συμπέρασμα ότι υπάρχει μια μεγάλη αναλογία συσχετιζόμενων περιοχών διότι για να περιοριστεί μια περιοχή ένα πλήθος από άλλες περιοχές έπρεπε να περιοριστούν πρώτα. Αντίθετα για το σύνολο δεδομένων zoo παρουσιάστηκε γρήγορη σύγκλιση και χρειάστηκαν μόνο δύο επαναλήψεις του ICF που οδήγησαν στην απομάκρυνση του 37% των περιπτώσεων στο πρώτο πέρασμα. Συνεπώς με βάση τα παραπάνω κατέληξαν στο συμπέρασμα ότι περιπτώσεις συνόλων για τα οποία ο αλγόριθμος ICF παρουσιάζει μικρό αριθμό επαναλήψεων και σε κάθε μια από αυτές απομακρύνει μεγάλο αριθμό περιπτώσεων τότε η δομή του χώρου των περιπτώσεων σε αυτά τα σύνολα είναι τέτοια που περιέχει μικρή αλληλεξάρτηση μεταξύ των περιοχών. Οι πιο προβληματικές δομές περιπτώσεων χαρακτηρίζονται από μεγάλο αριθμό επαναλήψεων που έχουν ως αποτέλεσμα λίγες μόνο περιπτώσεις να απομακρύνονται σε κάθε μια από αυτές.

Επιπρόσθετα στα πειράματα που έγιναν από τους Brighton και Mellish (2002) οι αλγόριθμοι ICF και RT3 πέτυχαν κατά προσέγγιση 80% μείωση πάνω στα 30 σύνολα που εφαρμόστηκαν. Επίσης διατηρήθηκε το 20% των

στιγμιοτύπων για έλεγχο κάτι το οποίο θεωρητικά σημαίνει ότι μόνο το 20% του συνόλου εκπαίδευσης απαιτείται για να επιτευχθεί η διατήρηση της ικανότητας. Επίσης ανακάλυψαν ότι τα πεδία τα οποία υποφέρουν από υποβάθμιση της ικανότητας ως αποτέλεσμα της εφαρμογής του ICF και του RT3 είναι ακριβώς αυτά για τα οποία η υποβάθμιση εμφανίζεται ως αποτέλεσμα της απομάκρυνσης του θορύβου. Αυτό υποδεικνύει ότι η απομάκρυνση του θορύβου είναι πολλές φορές επιβλαβής και ο ICF ως αποτέλεσμα υποφέρει τις επιπτώσεις. Αυτό το αποτέλεσμα υποστηρίζει τα συμπεράσματα των Daelemans, Bosch και Zavrel (1999) που ισχυρίζονται ότι το φιλτράρισμα φυσικών γλωσσών είναι λάθος λόγω του πλήθους των κλάσεων που αποτελούν εξαίρεση. Στα σύνολα δεδομένων που χρησιμοποιήθηκαν από τους αρθρογράφους οι κλάσεις εξαιρέσεις θα εμφανίστηκαν ως θόρυβος και επομένως θα απομακρύνθηκαν. Ωστόσο οι Daelemans, Bosch και Zavrel (1999) δεν χρησιμοποιούν κριτήρια φιλτραρίσματος τα οποία εξασφαλίζουν επαρκώς την διατήρηση των συνοριακών περιπτώσεων άρα το μόνο συμπέρασμα που εξάγουν οι Brighton και Mellish (2002) είναι ότι η απομάκρυνση του θορύβου είναι λανθασμένη όταν το σύνολο δεδομένων περιέχει πολλές κλάσεις που αποτελούν εξαίρεση. Τα αποτελέσματα τους καταλήγουν στο ότι σε κάποια προβλήματα ο θόρυβος δεν μπορεί να διακριθεί από τις κλάσεις που αποτελούν εξαίρεση.

Συνοψίζοντας οι Brighton και Mellish (2002) καταλήγουν στο ότι διαφορετικά πεδία μπορεί να έχουν δραστικά διαφορετικές δομές κλάσεων τις οποίες ταξινομούν σε ομογενείς και μη ομογενείς. Το κύριο σημείο στο οποίο επικεντρώνονται είναι στο ότι τα συστήματα μείωσης δεδομένων θεωρούνταν ως γενικές λύσεις στο πρόβλημα της επιλογής στιγμιοτύπων. Οι παρατηρήσεις τους στο πως αυτά τα συστήματα δουλεύουν και το πόσο καλά αποδίδουν σε διαφορετικά προβλήματα υποδεικνύουν ότι η επιτυχία ενός συστήματος μείωσης δεδομένων εξαρτάται σε μεγάλο βαθμό από την δομή του χώρου των στιγμιοτύπων. Επίσης υποστηρίζουν ότι ένα κριτήριο επιλογής δεν είναι αρκετό για υψηλή απόδοση πάνω στο πεδίο. Αυτό το επιχείρημα ενισχύεται όταν μεταβούμε σε πιο σύνθετα και μεγάλα σύνολα δεδομένων. Επιπλέον η δομή του συνόλου local-set που εισήγαγαν οι Brighton και Mellish είναι μια καλή μέτρηση για το πόσο ομογενείς είναι οι δομές κλάσεων στο χώρο των στιγμιοτύπων. Με τον υπολογισμό του μέσου μεγέθους των συνόλων local-set μπορούμε να έχουμε

μια εικόνα για το πόσο κοντά είναι μεταξύ τους τα στιγμιότυπα της ίδια κλάσης. Συνεπώς το ζήτημα είναι η ανάπτυξη αν επιθυμούμε να εξασφαλίσουμε επιτυχή επιλογή στιγμιοτύπων και το κλειδί για την ανάπτυξη είναι η σωστή κατανόηση για το πως δομούνται οι κλάσεις μέσα στον χώρο των στιγμιοτύπων.

2.5.4 Οι μέθοδοι FISH

2.5.4.1 Εισαγωγή

Η ύπαρξη του concept drift δημιουργεί μεγάλα προβλήματα στην κατασκευή επιτηρούμενων μοντέλων εκμάθησης για δεδομένα ροής. Η κατανομή των δεδομένων αλλάζει με την πάροδο του χρόνου και επομένως υπάρχει η ανάγκη για την δημιουργία προσαρμοστικών μοντέλων εκμάθησης. Στην επιτηρούμενη μάθηση η προσαρμοστικότητα αυτή μπορεί να επιτευχθεί είτε με την σχεδίαση συγκεκριμένων αλγορίθμων είτε με την χειραγώγηση του συνόλου εκπαίδευσης με τον χρόνο στο χώρο των στιγμιοτύπων ή και στο χώρο των γνωρισμάτων. Η χειραγώγηση του συνόλου εκπαίδευσης περιλαμβάνει την επιλογή στιγμιοτύπων , την χρήση βαρών στα στιγμιότυπα και την δυναμική επιλογή γνωρισμάτων (Wenerstrom & Carrier 2006). Οι στρατηγικές που αφορούν την διαχείριση του συνόλου εκπαίδευσης είναι μέθοδοι περιτύλιξης υπό την έννοια ότι μπορούν να χρησιμοποιηθούν για διαδικτυακή εκμάθηση σε διαφορετικούς τύπους κατηγοριοποιητών. Η διαδοχική επιλογή στιγμιοτύπων (εκπαίδευση παραθύρων) χρησιμοποιείται συνήθως σε ξαφνικά concept drift. Οι στρατηγικές εκπαίδευσης παραθύρων επιλέγουν τους εγγύτερους γείτονες σε κάθε χρονικό διάστημα για να σχηματίσουν το σύνολο εκπαίδευσης. Η επιλεκτική δειγματοληψία στον χώρο είναι ιδιαίτερα επικερδής όταν αναμένονται επαναλαμβανόμενα concept drifts. Σε αυτή την περίπτωση τα κοντινότερα στιγμιότυπα στον χώρο γνωρισμάτων επιλέγονται για να σχηματίσουν το σύνολο εκπαίδευσης.

Με βάση τα παραπάνω η Zliobaite (2011) στην έρευνα της παρουσίασε ένα σχέδιο που συνδυάζει τις αποστάσεις στον χρόνο και στον χώρο για την επιλογή του συνόλου εκπαίδευσης κάτω από την επίδραση του concept drift. Υποστηρίζει ότι χρειάζεται μια συνδυαστική οπτική για την επιλογή στιγμιοτύπων

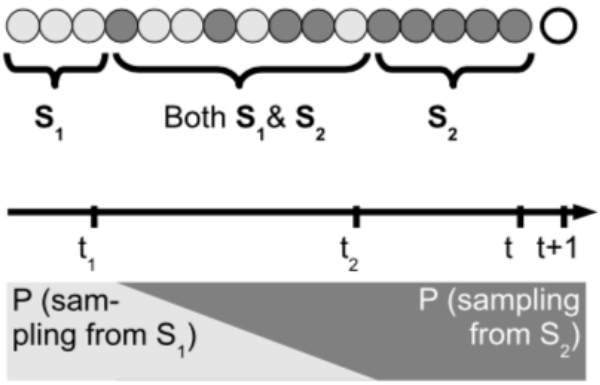
λόγο της σύνθετης φύσης των πραγματικών δεδομένων. Με την χρήση των εννοιών της ομοιότητας χρόνου και ομοιότητας χώρου ανέπτυξε μια οικογένεια μεθόδων για εκπαίδευση κατηγοριοποιητών που είναι ιδιαίτερα χρήσιμες όταν αναμένεται η παρουσία απότομων αλλαγών στα δεδομένα. Η επιλογή συνόλου εκπαίδευσης βασίζεται στην ομοιότητα του υπό εξέταση στιγμιοτύπου. Οι αποστάσεις στον χώρο και στον χρόνο συνδυάζονται γραμμικά. Οι μέθοδοι λοιπόν αυτοί καθορίζουν διαδικτυακά ένα βέλτιστο σύνολο εκπαίδευσης σε κάθε χρονικό βήμα χρησιμοποιώντας την διασταυρωμένη επικύρωση. Αποτελούν προσεγγίσεις περιτύλιξης πράγμα που σημαίνει ότι διαφορετικοί κατηγοριοποιητές μπορούν να τις εφαρμόσουν.

2.5.4.2 Ανάλυση του προβλήματος

Η σχετικότητα ενός αρχείου σύμφωνα με τα ενδιαφέροντα ενός αναγνώστη που διαβάζει διαδικτυακά νέα εξαρτάται από την ηλικία του αρχείου (απόσταση στον χρόνο) αλλά και από το περιεχόμενο (απόσταση στον χώρο). Εφόσον ο στόχος είναι η κατασκευή ενός κατηγοριοποιητή ο οποίος θα προσαρμόζεται σε ένα σταδιακό concept drift τότε για να επιτευχθεί αυτό πρέπει να επιλεχθούν τα πιο σχετικά ιστορικά στιγμιότυπα για να σχηματιστεί ένα σύνολο εκπαίδευσης. Στο παράδειγμα των διαδικτυακών νέων για κάθε νέο εισερχόμενο αρχείο (που δεν έχει ετικέτα) αναζητάμε όμοια αρχεία μέσα στο ιστορικό αρχείο. Η ομοιότητα μεταξύ δυο αντικειμένων στην εκμάθηση με βάση τα στιγμιότυπα (Aha & Kibler 1991) καθορίζεται ως μια συνάρτηση απόστασης στον χώρο. Αν το πεδίο δεν είναι στατικό τότε η απόσταση με βάση τον χρόνο είναι εξίσου σχετική.

Ας υποθέσουμε μια διεργασία διαδικτυακής κατηγοριοποίησης. Ένα στιγμιότυπο $X \in R^p$ λαμβάνεται μια χρονική στιγμή και η αντίστοιχη διακριτή ετικέτα της κλάσης έστω y είναι άγνωστη. Στον χρόνο t+1 ο σκοπός είναι η πρόβλεψη της ετικέτας της κλάσης y_{t+1} για το επιλεγμένο στιγμιότυπα X_{t+1}. Επιτρέπεται να επανεκπαιδευτεί ο κατηγοριοποιητής σε κάθε χρονικό βήμα αν αυτό απαιτείται. Κάθε επιλεγμένο ή και όλα τα δεδομένα του ιστορικού $X_1, X_2, \ldots, X_t$ με αντίστοιχες ετικέτες $y_1, y_2, \ldots, y_t$ μπορούν να χρησιμοποιηθούν ως σύνολο εκπαίδευσης για έναν κατηγοριοποιητή τη χρονική στιγμή t+1. Την χρονική στιγμή t+2 έπειτα από την απόφαση κατηγοριοποίησης η

πραγματική ετικέτα λαμβάνεται και μπορεί πλέον το X_{t+1} να προστεθεί στο σύνολο εκπαίδευσης και να προχωρήσει η διαδικασία στο X_{t+2}.



Εικόνα 2.4 : Σταδιακό concept drift (Zliobaite 2011)

Θεωρούμε το σταδιακό concept drift που φαίνεται στην εικόνα 2.4. Μέχρι την χρονική στιγμή t_1 η πηγή S_1 που παράγει τα δεδομένα είναι ενεργή. Σημειώνουμε ότι η πηγή δεν είναι η ίδια με την ετικέτα της κλάσης. Κάθε πηγή μπορεί να παράγει ένα στιγμιότυπο από οποιαδήποτε κλάση. Από την χρονική στιγμή $t_2 + 1$ και έπειτα η πηγή S_1 αντικαθίσταται πλήρως από την S_2. Στο χρονικό ενδιάμεσο $(t_1 + 1, t_2)$ και οι δύο πηγές είναι ενεργές και ένα στιγμιότυπο μπορεί να έρχεται από οποιαδήποτε από τις δύο πηγές με μια εκ των προτέρων πιθανότητα. Το κύριο πρόβλημα είναι ότι ο σχεδιαστής του μοντέλου δεν γνωρίζει πότε η πηγή θα αλλάξει οριστικά. Ο στόχος είναι να δοθεί μια ετικέτα κάποιας κλάσης στο στιγμιότυπο X_{t+1}. Αναμένεται ότι το concept drift είναι ενεργό επομένως αρκετές πηγές ενδεχόμενων να ήταν ενεργές κατά την χρονική στιγμή t+1. Επομένως για να χτιστεί ένα κατηγοριοποιητής που θα επιτυγχάνει καλή ακρίβεια είναι λογικό να επιθυμούμε το σύνολο εκπαίδευσης να προέρχεται από την ίδια ή τουλάχιστον όσο το δυνατόν κοντά στην πηγή του X_{t+1} στιγμιοτύπου. Η Zliobaite (2011) προτείνει ότι μπορούμε να βρούμε πόσο όμοιο είναι το X_{t+1} με τα στιγμιότυπα που έχουν εξεταστεί έως τότε ακόμα και αν η ετικέτα του συγκεκριμένου στιγμιοτύπου δεν είναι γνωστή. Επομένως στοχεύει να επιλέξει ένα σύνολο εκπαίδευσης το οποίο αποτελείται από στιγμιότυπα τα οποία είναι όμοια με το επιλεγμένο στιγμιότυπο.

2.5.4.3 Ομοιότητα σε χρόνο και χώρο για επιλογή συνόλου εκπαίδευσης

Η ιδέα της ομοιότητας στον χώρο και στον χρόνο ορίζεται σύμφωνα με την Zliobaite (2011) ως εξής: Έστω η ομοιότητα σε χώρο και χρόνο μεταξύ ενός επιλεγμένου στιγμιοτύπου X_j και ενός στιγμιοτύπου που έχει εξεταστεί X_i να είναι μια συνάρτηση απόστασης της μορφής :

$$D(X_i, X_j) = f\left(d_{ij}^{(S)}, d_{ij}^{(T)}\right) \quad (1)$$

όπου $d_{ij}^{(S)}$ είναι η απόσταση μεταξύ των δύο στιγμιοτύπων στον χώρο και $d_{ij}^{(T)}$ είναι η απόσταση μεταξύ των δύο στιγμιοτύπων στον χρόνο. Όσο μικρότερη είναι η απόσταση τόσο πιο όμοια είναι τα δύο στιγμιότυπα. Η απόσταση στον χρόνο μεταξύ των στιγμιοτύπων X_i , X_j στην περίπτωση των ίσων χρονικών διαστημάτων ορίζεται να είναι η συνάρτηση :

$$d_{ij}^{(T)} = f\left(|i-j|\right) \quad (2)$$

Διαφορετικές συναρτήσεις απόστασης μπορούν να επιλεχθούν με βάση την γνώση της δομής και της οπτικής επιθεώρησης των δεδομένων. Για παράδειγμα μια εκθετική συνάρτηση της μορφής $d_{ij}^{(T)} = e^{|i-j|}$ μπορεί να χρησιμοποιηθεί με σκοπό να δώσει έμφαση στις πιο πρόσφατες χρονικές στιγμές. Στην έρευνα της η Zliobaite χρησιμοποιεί την γραμμική απόσταση η οποία είναι η λιγότερο σύνθετη επιλογή . Επομένως εφαρμόζει την σχέση $d_{ij}^{(T)} = \left|i-j\right|$. Τα χρονικά διαστήματα μπορεί να είναι άνισα όπως για παράδειγμα στις τιμές των μετοχών οι οποίες καταγράφονται μόνο σε εργάσιμες ημέρες της εβδομάδος επομένως υπάρχει ένα κενό τριών ημερών μεταξύ της τιμής της Παρασκευής και της τιμής της Δευτέρας. Σε αυτή την περίπτωση χρησιμοποιείται η συνάρτηση $d_{ij}^{(T)} = \left|T(i)-T(j)\right|$ όπου Τ είναι η συνάρτηση που αντιστοιχεί τους δείκτες σε πραγματικές χρονικές τιμές. Όσον αφορά τον χώρο τώρα η απόσταση μπορεί να έχει διάφορες μετρικές (City-block , Ευκλείδεια). Επίσης μπορεί να χρησιμοποιηθεί και μια μετρική που δημιουργεί ο σχεδιαστής του αλγορίθμου. Οι

όροι ομοιότητα και απόσταση είναι αντιστρόφως ανάλογοι δηλαδή όσο μεγαλύτερη είναι η απόσταση τόσο μικρότερη η ομοιότητα και αντίστροφα. Η ομοιότητα χρησιμοποιείται ως γενική έννοια σαν όρος ενώ ο όρος απόσταση χρησιμοποιείται όταν αναφερόμαστε στην μετρική.

Ο σχεδιασμός μια συνάρτησης συνδυασμού D εξαρτάται από τις προσδοκίες που έχει ο σχεδιαστής με βάση τα δεδομένα. Η επιλογή των αναλογιών του χρόνου και του χώρου εξαρτώνται άμεσα από τους παρατηρούμενους τύπους αλλαγής αλλά και τις μελλοντικές προσδοκίες. Ο στόχος είναι η επιλογή του συνόλου εκπαίδευσης με τέτοιο τρόπο που να αντιπροσωπεύει επαρκώς τα τρέχοντα στιγμιότυπα. Ο γραμμικός συνδυασμός της μετρικής μεταξύ των στιγμιοτύπων X_i και X_j θα δίνεται τελικά από την σχέση :

$$D(X_i, X_j) = a_1 d_{ij}^{(S)} + a_2 d_{ij}^{(T)} \quad (3)$$

όπου οι a_1 και a_2 είναι σταθερές που εκφράζουν βάρη. Αν a_1 = 0 τότε ο σχεδιαστής επιλέγει ένα σύνολο εκπαίδευσης βασισμένο μόνο στην μετρική του χρόνου στρατηγική που ονομάζεται εκπαίδευση παραθύρου. Αν όμως $a_2 = 0$ τότε πρόκειται για μέθοδο επιλογής στιγμιοτύπων. Σε κάθε άλλη περίπτωση οι τιμές των σταθερών a_1 και a_2 μπορούν να καθοριστούν από τον σχεδιαστή με βάση το πεδίο της γνώσης ή την οπτική επιθεώρηση των δεδομένων ή ακόμα μπορεί να εξάγονται οι τιμές τους μέσα από ένα σετ επικύρωσης ή διαδικτυακά. Για λόγους ερμηνευσιμότητας η Zliobaite (2011) κανονικοποιεί τις αναλογίες του χώρου και χρόνου στις συναρτήσεις $d^{(S)}$ και $d^{(T)}$. Ρυθμίζει τις τιμές κάθε γνωρίσματος στο X να ανήκουν στο διάστημα $\left[0, \frac{1}{p}\right]$ με σκοπό να ισχύει ότι $d^{(S)} \in [0,1]$. Ομοίως θα ισχύει και $d^{(T)} \in [0,1]$. Για ένα μόνο σύνολο δεδομένων η κλιμάκωση αυτή δεν ειναι αναγκαία καθώς η αναλογία μπορεί να εντοπιστεί μέσω των a_1 και a_2 . Όμως με αυτό τον τρόπο οι μετρικές του χρόνου και του χώρου μπορούν να συγκριθούν σε διάφορα σύνολα δεδομένων. Για την επιλογή συνόλου εκπαίδευσης όταν συμβαίνει concept drift η Zliobaite (2011) αναφέρει ότι μας ενδιαφέρει η σχετική απόσταση. Επομένως για απλότητα τα a_1

και a_2 μπορούν να αντικατασταθούν από το $A = \frac{a_2}{a_1}$ δεδομένου ότι $a_1 \neq 0$.

Μας ενδιαφέρει επομένως ο βαθμός της απόστασης μεταξύ των στιγμιοτύπων που έχουν εξεταστεί και του επιλεγμένου στιγμιοτύπου X_{t+1} . Συνεπώς η σχέση (3) μπορεί να απλοποιηθεί στην παρακάτω :

$$D^*(X_i, X_{t+1}) = d^{(S)}_{i,t+1} + A\, d^{(T)}_{i,t+1} = D^*_i \quad (4)$$

Επομένως σύμφωνα με την Zliobaite (2011) καθορίζουμε την απόσταση στον χρόνο και χώρο D^* με σκοπό να βαθμολογήσουμε τα στιγμιότυπα που έχει εξετάσει η μέθοδος σύμφωνα με την απόσταση από το επιλεγμένο στιγμιότυπο X_{t+1} .Μια όμως ακόμα πολύ σημαντική απόφαση που πρέπει να παρθεί είναι το πόσα από τα όμοια στιγμιότυπα πρέπει να συμπεριληφθούν στο σύνολο εκπαίδευσης . Το μέγεθος του συνόλου εκπαίδευσης καθορίζεται με την εφαρμογή ενός κατωφλιού στη μετρική της απόστασης. Αφού η μετρική απόστασης D^* έχει καθοριστεί το μέγεθος του συνόλου εκπαίδευσης καθορίζεται με την μετακίνηση του κατωφλιού αυτού. Επομένως δοσμένου ενός στιγμιοτύπου X_{t+1} $, t = 1,...,t$ το οποίο δεν έχει ακόμα ετικέτα το στιγμιότυπο X_i επιλέγεται να εισέλθει στο σύνολο εκπαίδευσης αν $D^*(X_i, X_{t+1}) < h^D$, όπου h^D είναι το κατώφλι του συνόλου εκπαίδευσης. Το κατώφλι αυτό καθορίζεται από τον σχεδιαστή ή παράγεται από ένα σύνολο επικύρωσης.

2.5.4.4 Η οικογένεια μεθόδων FISH

Μετά την προσέγγιση που συνδυάζει την απόσταση σε χώρο και χρόνο η Zliobaite (2011) προτείνει την οικογένεια μεθόδων που ονομάζονται FISH (uni**F**ied **I**nstance **S**election algorit**H**m) η οποία ενσωματώνει τις ιδέες που παρουσιάστηκαν ανωτέρω. Τα στιγμιότυπα εκπαίδευσης επιλέγονται συστηματικά σε κάθε χρονικό βήμα. Οι μέθοδοι μπορούν να χρησιμοποιηθούν με διαφορετικούς βασικούς κατηγοριοποιητές. Συγκεκριμένα η οικογένεια αυτή περιέχει τρεις τροποποιήσεις τους αλγόριθμους FISH1, FISH2, FISH3. Στον

FISH1 το μέγεθος του συνόλου εκπαίδευσης είναι προκαθορισμένο και τοποθετείται εξαρχής. Η προέκταση του ο FISH2 λειτουργεί χρησιμοποιώντας μεταβλητό μέγεθος στο σύνολο εκπαίδευσης. Επίσης στον FISH2 οι αναλογίες του χώρου και του χρόνου (δηλαδή οι σταθερές a_1 και a_2 στην σχέση (3)) είναι προκαθορισμένες εξαρχής ως επιλογή του σχεδιαστή. Η επέκταση του FISH2 ο αλγόριθμος FISH3 δεν θέτει προκαθορισμένες τιμές στα a_1 και a_2 αλλά οι αναλογίες αυτές είναι εκπαιδεύσιμες διαδικτυακά. Η Zliobaite (2011) σημειώνει ότι ο αλγόριθμος FISH2 είναι ο κεντρικός μέσα στο σύνολο αυτών των μεθόδων. Υποστηρίζει ότι σε πολλές περιπτώσεις οι βέλτιστες αναλογίες του χρόνου και χώρου στην συνάρτηση απόστασης είναι εξαρτώμενες από το πεδίο και μπορούν να καθορίσουν εκτός δικτύου (με την χρήση για παράδειγμα ενός συνόλου επικύρωσης). Από την άλλη οι αλλαγές μπορεί να λάβουν μη ομοιόμορφες ταχύτητες επομένως το διαδικτυακά προσαρμόσιμο μέγεθος του συνόλου εκπαίδευσης μπορεί να είναι αναγκαίο.

FISH1

Ο ψευδοκώδικας που δίνεται παρακάτω για τον αλγόριθμο FISH1 περιέχει τα βήματα για την επιλογή συνόλου εκπαίδευσης για την λήψη αποφάσεων στον χρόνο t+1. Η μέθοδος κατατάσσει τα ιστορικά στιγμιότυπα χωρίς τις ετικέτες τους σύμφωνα με την απόσταση τους από το στιγμιότυπο που εξετάζεται και επιλέγει N στο πλήθος στιγμιότυπα αυτά που είναι τα πιο όμοια με σκοπό να δημιουργήσει το σύνολο ελέγχου. Δεδομένου ότι το μέγεθος του συνόλου εκπαίδευσης είναι προκαθορισμένο αν τύχει και είναι μικρό τότε τα στιγμιότυπα από μια μόνο κλάση μπορεί να καταλήξουν στο σύνολο εκπαίδευσης. Για να αποφευχθεί κάτι τέτοιο η Zliobaite (2011) προτείνει την επιλογή ενός στρωματοποιημένου συνόλου εκπαίδευσης. Αυτό σημαίνει ότι τα $\frac{N}{c}$ ποιο όμοια στιγμιότυπα επιλέγονται για κάθε κλάση , επομένως όλα μαζί σχηματίζουν το σύνολο εκπαίδευσης μεγέθους N .

Αλγόριθμος FISH1 (Zliobaite 2011)

Είσοδος

Δεδομένα : ιστορικά στιγμιότυπα $X^H = \left(X_1, \ldots, X_t\right)$ με ετικέτες y^H , στιγμιότυπο υπό εξέταση X_{t+1} χωρίς ετικέτα.

Παράμετροι : μέγεθος συνόλου εκπαίδευσης N , αναλογία χρόνου/χώρου A

Βασικός αλγόριθμος εκμάθησης : L

1: **for** i = 1:t

2: Υπολογισμός αποστάσεων χρόνου και χώρου D_i^*

3: Ταξινόμηση των αποστάσεων από το ελάχιστο στο μέγιστο $D_{z1}^* < D_{z2}^* < \ldots < D_{zt}^*$. Οι δείκτες $z1,...,zt$ καθορίζουν την μετάθεση $\left(X_1, \ldots, X_t\right) \rightarrow \left(X_{z1}, \ldots, X_{zt}\right)$.

4: Επιλογή των N στιγμιοτύπων με τις μικρότερες αποστάσεις D .

5: Παραγωγή του συνόλου των δεικτών $\{z1,...,zN\}$

Έξοδος

Οι δείκτες $I_t = \{z1,...,zN\}$ σχηματίζουν το σύνολο εκπαίδευσης $X_t^T = \left(X_{z1}, \ldots, X_{zN}\right)$.

FISH2

Ο FISH2 αποτελεί μια προέκταση του FISH1 και η κύρια διαφορά του είναι ότι το μέγεθος του συνόλου εκπαίδευσης βρίσκεται με διαδικτυακή εκμάθηση. Για να εφαρμοστεί ένα μεταβλητό μέγεθος στο σύνολο εκπαίδευσης η Zliobaite ενσωμάτωσε ιδέες που εμπνεύστηκαν από δυο μεθόδους παραθύρων. Η πρώτη αφορά την δυναμική ενσωμάτωση ταξινομητών για την διαχείριση του concept drift (Tsymbal et al. 2008) και η δεύτερη είναι εμπνευσμένη από το έργο των Klinkenberg και Joachims (2000) που αφορά τον εντοπισμό του concept drift με χρήση μηχανών διανυσμάτων υποστήριξης.

Ο αλγόριθμος ξεκινά με τον υπολογισμό των αποστάσεων στον χρόνο και στον χώρο μεταξύ του υπό εξέταση στιγμιοτύπου X_{t+1} και κάθε ιστορικού στιγμιοτύπου $X^H = \left(X_1, \ldots, X_t\right)$. Οι αποστάσεις του X_{t+1} προς κάθε ιστορικό στιγμιότυπο κατατάσσονται με βάση την απόσταση. Έπειτα με την χρήση διασταυρωμένης επικύρωσης αποφασίζεται πόσα από τα πιο όμοια στιγμιότυπα εκπαίδευσης πρέπει να επιλεχθούν . Για τον λόγο αυτό η Zliobaite (2011) αναφέρει ότι χτίζεται ένα σύνολο ταξινομητών $\left(L^1, L^2, \ldots, L^N\right)$ που χρησιμοποιούν διαφορετικού μεγέθους σύνολα εκπαίδευσης. Το σύνολο επικύρωσης σχηματίζεται με την χρήση k ιστορικών στιγμιοτύπων τα οποία βρέθηκαν να είναι τα πιο όμοια με το υπό εξέταση στιγμιότυπο X_{t+1}. Επιλέγεται το μέγεθος του συνόλου εκπαίδευσης να είναι το N^* το οποίο έδωσε την καλύτερη ακρίβεια στο σύνολο επικύρωσης. Η μέθοδος δουλεύει όμοια με την χρήση παραθύρων στην έρευνα των Klinkenberg και Joachims (2000). Οι τελευταίοι χρησιμοποιούν ακολουθιακά στιγμιότυπα με βάση τον χρόνο για να σχηματίσουν τα παράθυρα. Η μέθοδος διασταυρωμένης επικύρωσης «άφησε ένα έξω» χρειάζεται να εφαρμοστεί. Αυτό σημαίνει ότι η διαδικασία της επικύρωσης επαναλαμβάνεται k φορές για κάθε μέγεθος N συνόλου εκπαίδευσης που ελέγχεται. Κάθε φορά αφήνεται έξω ένα στιγμιότυπο επικύρωσης από το σύνολο εκπαίδευσης και έπειτα γίνεται έλεγχος με αυτό. Χωρίς την διασταυρωμένη επικύρωση το σύνολο εκπαίδευσης μεγέθους k είναι πιθανόν να δώσει την καλύτερη ακρίβεια διότι σε αυτή την περίπτωση το σύνολο εκπαίδευσης θα είναι ίσο με το σύνολο επικύρωσης. Το αποτέλεσμα της μεθόδου είναι ένα σύνολο δεικτών $I_t = \{z1,...,zN^*\}$ μεγέθους N^*. Οι δείκτες αυτοί υποδεικνύουν ποια από τα στιγμιότυπα που έχει δει ο αλγόριθμος θα επιλεχθούν για το σύνολο εκπαίδευσης $X_t^T = \left(X_{z1}, \ldots, X_{zN}\right)$. Χρησιμοποιώντας τα αρχικά στιγμιότυπα X_t^T ένας κατηγοριοποιητής $L_t^{N^*}$ εκπαιδεύεται για την τελική πρόβλεψη της ετικέτας y_{t+1} του υπό εξέταση στιγμιοτύπου X_{t+1} .

Αλγόριθμος FISH2 (Zliobaite 2011)

Είσοδος

Δεδομένα : X^H, y^H, X_{t+1}

Παράμετροι : μέγεθος γειτονιάς k , A

Βασικός αλγόριθμος εκμάθησης : L

1: **for** i = 1:t

2: Υπολογισμός αποστάσεων χρόνου και χώρου D_i^*

3: **for** N = k : βήμα : t επέλεξε το μέγεθος του συνόλου εκπαίδευσης

4: επέλεξε N στιγμιότυπα που έχουν τις μικρότερες αποστάσεις D

5: με χρήση διασταυρωμένης επικύρωσης χτίσε έναν κατηγοριοποιητή L^N χρησιμοποιώντας

6: τα στιγμιότυπα $\left(X_{z1}, \ldots, X_{zN}\right)$ ως σύνολο εκπαίδευσης.

7: έλεγξε το L^N στους k εγγύτερους γείτονες $\left(X_{z1}, \ldots, X_{zN}\right)$ και κατέγραψε το σφάλμα

8: ελέγχου e_N

9: Βρες τον κατηγοριοποιητή L^{N*} με το ελάχιστο σφάλμα όπου $N* = argmin_{N=k}^{t}(e_N)$.

10: Παραγωγή του συνόλου δεικτών $\{z1,...,zN*\}$.

Έξοδος

Οι δείκτες $I_t = \{z1,...,zN\}$ για τον σχηματισμό του συνόλου εκπαίδευσης X_t^T .

FISH3

Ο FISH3 αποτελεί επέκταση του FISH2. Συγκεκριμένα ο FISH2 χρησιμοποιεί μια προκαθορισμένη αναλογία A των αποστάσεων στον χρόνο και στον χώρο. Ο FISH3 μπορεί να μάθει την αναλογία αυτή διαδικτυακά χρησιμοποιώντας έναν επιπλέον βρόγχο διασταυρωμένης επικύρωσης. Αντί να καθορίζεται η αναλογία μεταξύ χρόνου και χώρου η Zliobaite (2011) επιχειρεί έναν αριθμό εναλλακτικών και επιλέγει τον αλγόριθμο εκμάθησης ο οποίος είναι ο πιο ακριβής στο σύνολο επικύρωσης , εφαρμόζει την ίδια αρχή που υπάρχει δηλαδή και στον FISH2.

Αλγόριθμος FISH3 (Zliobaite 2011)

Είσοδος

Δεδομένα : X^H, y^H, X_{t+1}

Παράμετροι : μέγεθος γειτονιάς k

Βασικός αλγόριθμος εκμάθησης : L

1: **For** j = 0 : βήμα : 1

2: $a_1 = j \quad , a_2 = 1 - a_1$ οι αναλογίες χώρου και χρόνου

3: **for** i=1:t

4: υπολόγισε τις αποστάσεις $D_i^j = a_1 d_{i,t+1}^{(S)} + a_2 d_{i,t+1}^{(T)}$

5: ταξινόμησε τις αποστάσεις με αύξουσα σειρά $D_{jz1}^j < D_{jz2}^j < \ldots < D_{jzt}^j$

6: for N = k : βήμα2 : t επέλεξε το μέγεθος του συνόλου εκπαίδευσης

7: επέλεξε N στιγμιότυπα αυτά με την μικρότερη απόσταση D^j

8: με χρήση διασταυρωμένης επικύρωσης χτίσε κατηγοριοποιητή L^{jN}

9: χρησιμοποιώντας τα στιγμιότυπα $\left(X_{jz1}, \ldots, X_{jzN}\right)$ ως σύνολο εκπαίδευσης

10: έλεγξε τον L^{jN} με τους k εγγύτερους γείτονες $\left(X_{jz1}, \ldots, X_{jzk}\right)$ και κατέγραψε το

11: σφάλμα ελέγχου e_N^j

12: Εύρεση του κατηγοριοποιητή με το ελάχιστο σφάλμα L^{jN*} όπου

13: $jN* = argmin_{j=0}^{1} min_{N=k}^{t}(e_N^j)$

14: Παραγωγή του συνόλου δεικτών $\{jz1,...,jzN*\}$.

Έξοδος

Οι δείκτες $I_t = \{jz1,...,jzN\}$ σχηματίζουν το συνόλο εκπαίδευσης X_t^T .

2.5.4.5 Αξιολόγηση - Συμπεράσματα

Στην πειραματική αξιολόγηση η Zliobaite (2011) έχοντας ως σκοπό να επιβεβαιώσει τις ιδιότητες των μεθόδων FISH εκτελεί εκτενή αριθμητικά πειράματα. Ο κύριος στόχος της είναι να αναδείξει το πλεονέκτημα που παρέχει ο συνδυασμός της απόστασης στον χώρο και τον χρόνο σε σύγκριση με την χρήση μόνο του κριτηρίου που αφορά τον χρόνο και τον χώρο. Εφαρμόζονται δυο όμοιες μέθοδοι και εκτελούνται παράλληλα με τον FISH σε έξι σύνολα δεδομένων. Με σκοπό να μειωθεί η προτίμηση σε κάποιο συγκεκριμένο βασικό αλγόριθμο επιλογής τα πειράματα εκτελούνται χρησιμοποιώντας τέσσερις διαφορετικούς βασικούς αλγόριθμους και δυο εναλλακτικές συναρτήσεις για τον υπολογισμό απόστασης στον χώρο.

Επιπλέον υπολογίζεται η απόδοση των μεθόδων FISH με βάση το σφάλμα ελέγχου και την πολυπλοκότητα. Με σκοπό να εντοπιστεί μια διαφορά που είναι στατιστικά σημαντική μεταξύ των βαθμών των σφαλμάτων των μεθόδων χρησιμοποιείται το τεστ McNemar το οποίο δεν απαιτεί την υπόθεση της ανεξαρτησίας και ίσης κατανομής των δεδομένων. Επιπλέον για τους αλγορίθμους FISH1, FISH2, FISH3, TSY χρησιμοποιείται η ευκλείδεια απόσταση του χώρου :

$$d^E\left(X_j, X_i\right) = \sqrt{\sum_{i=1}^{p} |\, x_j^{(i)} - x_i^{(i)} \,|^2}$$

όπου $x_j^{(i)}$ είναι το i γνώρισμα του στιγμιοτύπου X_j και p είναι η διάσταση. Τα γνωρίσματα έχουν αναχθεί στο διάστημα [0,1] πριν τον υπολογισμό της απόστασης στον χώρο. Επιπλέον ο αλγόριθμος FISH2 εξετάζεται και με την χρήση της απόστασης του συνημιτόνου (αντίστροφη ομοιότητα). ΄Οσον αφορά τον χρόνο χρησιμοποιείται η γραμμική απόσταση που ορίζεται από την σχέση (2). Επιπλέον οι αποστάσεις στον χρόνο και τον χώρο ανάγονται και αυτές στο διάστημα [0,1] δηλαδή $d^{(S)}, d^{(T)} \in [0,1]$ πριν τον υπολογισμό των αναλογιών $a_1 : a_2$.

Για την εξέταση των μεθόδων χρησιμοποιήθηκαν έξι σύνολα δεδομένων με ενδεχόμενο concept drift. Τρία εξ αυτών είναι αληθινά (Luxenbourg , Ozone ,

Electricity) ενώ τα υπόλοιπα είναι μεν αληθινά αλλά με εισαγόμενο τεχνητό concept drift (German, Vote2, Iono2). Όλα τα παραπάνω σύνολα δεδομένων έχουν δυαδική διεργασία ταξινόμησης.

Τα πειράματα στον FISH1 υλοποιήθηκαν με μεταβαλλόμενη την αναλογία του χρόνου και του χώρου στην μετρική της απόστασης. Η μεταβολή αυτή ήταν ελεγχόμενη δηλαδή καθορίστηκε το σύστημα εκτός της παραμέτρου που αποτελεί την αναλογία του χρόνου και του χώρου στη συνάρτηση της απόστασης. Ερευνήθηκε η επίδραση της αναλογίας αυτής στην ακρίβεια του αλγορίθμου και στα έξι σύνολα δεδομένων. Επιτράπηκε η αναλογία $a_1 : a_2$ να μεταβάλλεται από 0:1 έως 1:0 με βήμα 0,01. Ο βασικός αλγόριθμος ταξινόμησης που χρησιμοποιήθηκε ήταν ο NMC (Nearest Mean classifier). Τα αποτελέσματα που εξάγει η Zliobaite (2011) ακόμα και με την χρήση ενός προκαθορισμένου μεγέθους N συνόλου εκπαίδευσης υποδεικνύουν ότι η καλύτερη ακρίβεια επιτεύχθηκε με τον συνδυασμό αποστάσεων στον χώρο και στον χρόνο. Το ελάχιστο σφάλμα συνέκλινε αρκετά προς την απόσταση στον χώρο στην πλειοψηφία των συνόλων (δηλαδή όταν $a_1 > a_2$) κάτι το οποίο εξηγείται από την φύση των δεδομένων. Τα σύνολα δεδομένων επιλέχθηκαν έτσι ώστε να έχουν ετερογενή δομή στον χώρο και να έχουν επίσης προσωρινή διάταξη. Συνεπώς ακόμα και με την χρήση ενός προκαθορισμένου μεγέθους για το σύνολο εκπαίδευσης ο αλγόριθμος FISH1 αποδίδει καλύτερα από τον ALL σε πέντε από τα έξι σύνολα δεδομένων. Ο ALL είναι αλγόριθμος που χρησιμοποιεί όλα τα δεδομένα. Αν τα δεδομένα ήταν στατικά τότε ο αλγόριθμος ALL θα ήταν ο πιο ακριβής.

Ο αλγόριθμος FISH2 που θεωρείται ο κεντρικός στην οικογένεια των μεθόδων FISH συγκρίθηκε με τους KLI , TSY καθώς και τον ALL ο οποίος θεωρείται το επίπεδο αναφοράς. Ο έλεγχος πραγματοποιήθηκε με την χρήση τεσσάρων βασικών κατηγοριοποιητών και δύο εναλλακτικών μετρικών απόστασης στον χώρο. Συνεπώς εκτελέστηκαν από την Zliobaite (2011) 48 πειράματα για κάθε μέθοδο. Απο τα αποτελέσματα προκύπτει ότι ο FISH2 πετυχαίνει τον καλύτερο και με διαφορά βαθμό όταν χρησιμοποιεί ως βασικό κατηγοριοποιητή τον NMC ή δέντρα, ενώ είναι λίγο καλύτερος με τον PWC . Τέλος στην περίπτωση που βασικός κατηγοριοποιητής είναι ο kNN τότε η μέθοδος ALL αποδίδει καλύτερα από τον FISH2. Τα αποτελέσματα είναι

προφανώς υπέρ του FISH2 πρέπει να σημειωθεί όμως ότι κάποια από αυτά δεν παρουσιάζουν στατιστικά σημαντική διαφορά και επιπλέον πρέπει να υπολογιστεί το γεγονός ότι τα σύνολα που χρησιμοποιήθηκαν είναι μικρά σε μέγεθος . Επίσης το τεστ είναι μη παραμετρικό. Ο FISH2 είναι σχεδιασμένος να λειτουργεί εκεί όπου το concept drift δεν φαίνεται ξεκάθαρα δηλαδή σε περιπτώσεις που είναι σταδιακή ή επαναλαμβανόμενη η παρουσία του. Τα εύσημα για την απόδοση του FISH2 πρέπει να δοθούνε στην επιλογή του συνόλου εκπαίδευσης με βάση την ομοιότητα. Ο KLI χρησιμοποιεί μόνο την ομοιότητα στον χρόνο (παράθυρο εκπαίδευσης). Ο TSY χρησιμοποιεί μόνο την ομοιότητα στον χρόνο για την εκπαίδευση του κατηγοριοποιητή αλλά έπειτα χρησιμοποιεί την ομοιότητα στον χώρο για την επιλογή κατηγοριοποιητή. Επιπλέον ο FISH2 αποδίδει καλύτερα από τους KLI και TSY επειδή χρησιμοποιεί προσαρμοσμένο σύνολο επικύρωσης σε σύγκριση με τον KLI και μεταβλητό μέγεθος για το σύνολο εκπαίδευσης σε σύγκριση με τον TSY. Σε όλα τα πειράματα που αφορούσαν τον FISH2 χρησιμοποιήθηκαν ίσες αναλογίες μεταξύ χώρου και χρόνου στην συνάρτηση απόστασης $a_1 : a_2 = 1 : 1$ με σκοπό να υπάρχουν ομοιόμορφα και συγκρίσιμα αποτελέσματα μεταξύ όλων των συνόλων.

Η μέθοδος FISH3 χρησιμοποιεί μεταβλητό μέγεθος συνόλου εκπαίδευσης καθώς και μεταβλητές αναλογίες χρόνου και χώρου στην συνάρτηση της απόστασης , και στα δύο μεταβλητά μεγέθη γίνεται διαδικτυακή εκμάθηση για να βρεθούν οι τιμές τους. Στη μέθοδο FISH1 οι αναλογίες ήταν προκαθορισμένες πριν από κάθε τρέξιμο των πειραμάτων ενώ στον FISH3 υπάρχει μεταβλητή αναλογία για κάθε χρονικό βήμα. Η Zliobaite (2011) συγκρίνει τις ακρίβειες των τριών FISH μεθόδων χρησιμοποιώντας ως βασικό κατηγοριοποιητή τον NMC και την ευκλείδεια απόσταση στον χώρο. Χρησιμοποιούνται επίσης οι ίδιες προκαθορισμένες αναλογίες χρόνου και χώρου στη συνάρτηση απόστασης. Ο FISH3 αποδεικνύεται ότι έχει την καλύτερη ακρίβεια σε όλα τα σύνολα εκτός του Ozone το οποίο όμως έχει τεράστια ασυμμετρία κλάσεων. Στην σύγκριση συμπεριλήφθηκε και ο βασικός αλγόριθμος ALL για να επιβεβαιωθεί ότι οι μέθοδοι FISH ανταποκρίνονται στην παρουσία ενδεχόμενου concept drift. Η βελτίωση στην ακρίβεια που επιδεικνύει ο FISH3 σε σχέση με τον FISH2 μπορεί να θεωρηθεί αμελητέα. Για την ακρίβεια οι διαφορές μεταξύ τους είναι στατιστικά σημαντικές μόνο στα τρία από τα έξι σύνολα. Τέλος με βάση τα στατιστικά τεστ

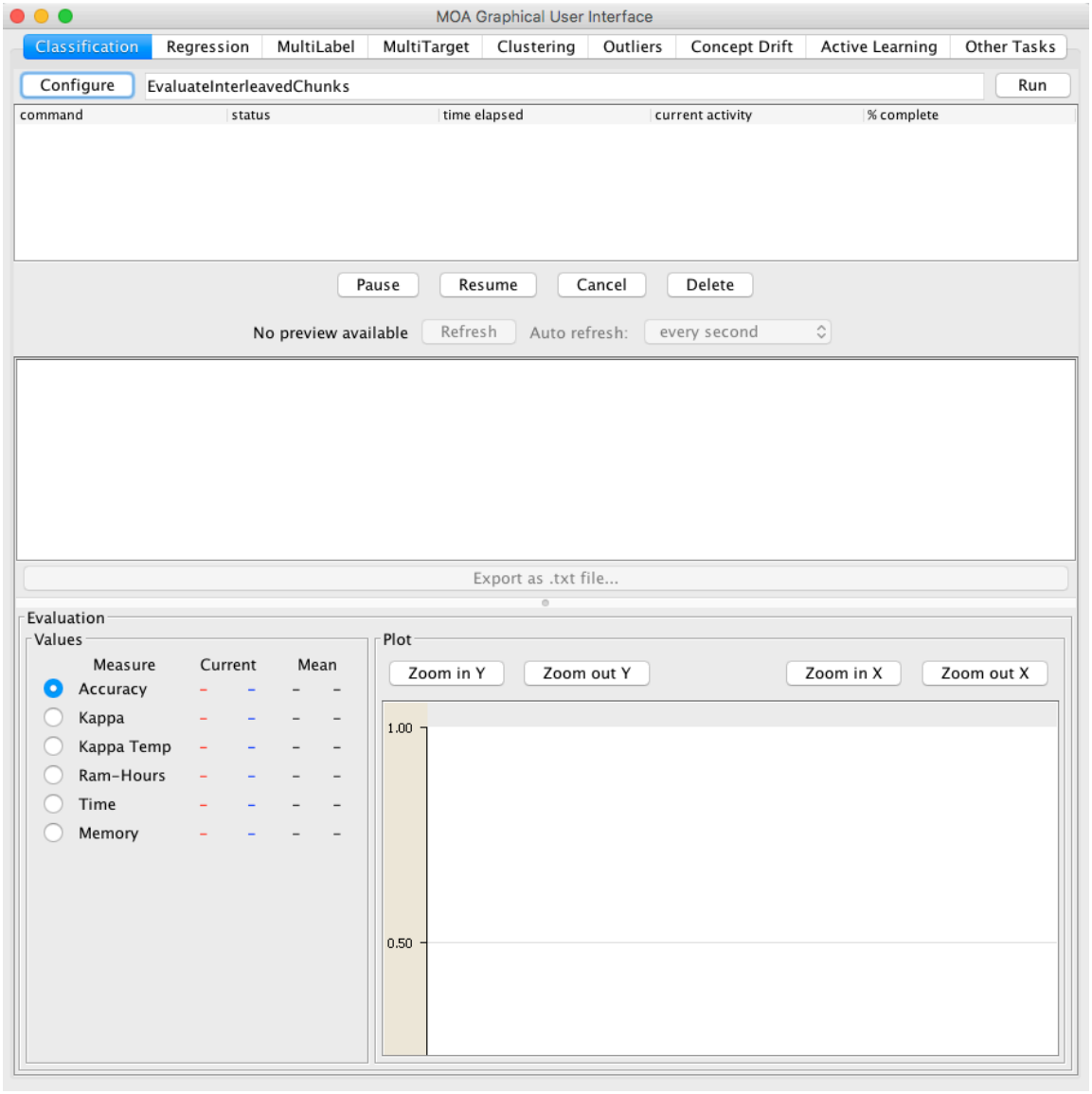
μπορεί να εξαχθεί το συμπέρασμα ότι αν το πεδίο επιτρέπει αυξημένη πολυπλοκότητα η χρήση μεταβλητού μεγέθους στο σύνολο εκπαίδευσης καθώς και μεταβλητών αναλογιών χρόνου και χώρου αξίζει να εφαρμοστούν σε σύνολα δεδομένων που παρουσιάζουν σταδιακό drift. Η έρευνα λοιπόν καταλήγει στο ότι οι μέθοδοι FISH πρέπει να θεωρηθούν ως μια επέκταση στις υπάρχουσες τεχνικές και στο ότι τονίζεται ότι οι σχέσεις χρόνου και χώρου δεν είναι διακριτές αλλά μπορούν να θεωρηθούν ως ένας συνεχόμενος χώρος.

ΚΕΦΑΛΑΙΟ 3. Μεθοδολογία

3.1 Το Λογισμικό MOA

Το MOA (Massive Online Analysis) είναι ένα περιβάλλον λογισμικού το οποίο επιτρέπει την ενσωμάτωση και εφαρμογή αλγορίθμων για την διεξαγωγή πειραμάτων μέσω τον οποίων γίνεται online εκμάθηση από εξελισσόμενα δεδομένα ροής. Το λογισμικό MOA εμπεριέχει μια συλλογή από offline και online μεθόδους όπως επίσης και εργαλεία για αξιολόγηση. Συγκεκριμένα εφαρμόζει τις τεχνικές boosting, bagging και τα δέντρα Hoeffding όλα με ή χωρίς Naive Bayes κατηγοριοποιητές στα φύλλα (Bifet et al. 2010)

Το MOA είναι συνδεδεμένο με το WEKA (Waikato Environment for Knowledge Analysis) το οποίο είναι ένα ανοιχτού κώδικα βραβευμένο λογισμικό το οποίο εμπεριέχει εφαρμογές από μιας ευρείας γκάμας στατικούς αλγόριθμους μηχανικής μάθησης. Το WEKA όπως και το MOA είναι γραμμένα στην προγραμματιστική γλώσσα Java. Τα κύρια οφέλη της Java είναι η φορητότητα της όπου οι εφαρμογές μπορούν να τρέξουν σε οποιαδήποτε πλατφόρμα εμπεριέχει μια εικονική μηχανή Java και οι ισχυρές και αναπτυγμένες βιβλιοθήκες που διαθέτει ως γλώσσα. Επιπλέον η χρήση της γλώσσας είναι διαδεδομένη και χαρακτηριστικά όπως η αυτόματη συλλογή απορριμμάτων βοηθάει στο να μειωθεί το φορτίο και τα σφάλματα του προγραμματιστή. Οι αλγόριθμοι γράφονται σε Java και υλοποιούνται ως κλάσεις οι οποίες έπειτα οργανώνονται σε πακέτα για να επικοινωνούν καλύτερα μεταξύ τους. Συνεπώς το MOA είναι υλοποιημένο σε πακέτα (packages) που ακολουθούν ιεραρχία καταλόγου. Κάθε φορά που η εικονική μηχανή της Java καλείται να εκτελέσει έναν αλγόριθμο, δημιουργεί ένα στιγμιότυπο της σχετικής κλάσης και κατανέμει τη μνήμη που χρειάζεται για την εκτέλεση.



Εικόνα 3.1 : Η γραφική διεπαφή χρήστη του MOA

3.1.1 Εξόρυξη γνώσης από δεδομένα ροής με το MOA

Μια ευρέως αποδεκτή υπόθεση που υπάρχει στην σύγχρονη κοινωνία είναι ότι είναι σημαντικό να καταγράφονται τα δεδομένα καθώς μπορεί να εμπεριέχουν σημαντικές πληροφορίες . Αυτό συμβαίνει σε όλες τις πλευρές της ζωής από τις αποδείξεις των σουπερ μάρκετ έως τις τηλεφωνικές διαλέξεις. Για να υποστηρίξουν αυτή την υπόθεση οι μηχανικοί και οι επιστήμονες παρήγαγαν ένα μεγάλο πλήθος ευρηματικών προγραμμάτων και συσκευών από προγράμματα επιβράβευσης έως συστήματα εντοπισμού αναγνώρισης αντικειμένων τα

λεγόμενα RFID tags. Εντούτοις δεν έχει δοθεί η ανάλογη σημασία στο πώς αυτή η τεράστια ποσότητα δεδομένων μπορεί να αναλυθεί .

Η μηχανική μάθηση το πεδίο της πληροφορικής το οποίο ασχολείται με το να βρεθούν τρόποι έτσι ώστε να εξάγεται αυτόματα πληροφορία από τα δεδομένα θεωρούνταν η λύση στο συγκεκριμένο πρόβλημα. Ιστορικά η μηχανική μάθηση είχε επικεντρωθεί στο να εκπαιδεύει τους αλγόριθμους της από μικρό αριθμό δεδομένων διότι αυτά ήταν περιορισμένα όταν το πεδίο αυτό άρχισε να αναπτύσσεται. Κάποιοι πολύ εκλεπτυσμένοι αλγόριθμοι κατασκευάστηκαν από την έρευνα η οποία μπορεί να κατασκευάσει μοντέλα με μεγάλη ακρίβεια αλλά από περιορισμένα παραδείγματα. Σε αυτές τις περιπτώσεις είναι κοινώς αποδεκτό ότι ολόκληρο το σύνολο εκπαίδευσης μπορεί να αποθηκευτεί στην μνήμη (Bifet & Kirkby 2009).

Προσφάτως η ανάγκη να επεξεργαστούν μεγαλύτερα σύνολα δεδομένων έδωσε κίνητρο στο να αναπτυχθεί το πεδίο της εξόρυξης γνώσης από δεδομένα (data mining) . Διάφοροι τρόποι ερευνούν τον τρόπο με τον οποίο θα μειώσουν τον υπολογιστικό χρόνο και την απαιτούμενη μνήμη που χρειάζεται για να επεξεργαστούν τεράστια σύνολα δεδομένων τα οποία όμως είναι στατικά . Αν τα δεδομένα δεν χωράνε στην μνήμη τότε ίσως είναι αναγκαίο να παρθεί ως δείγμα ένα μικρότερο σύνολο εκπαίδευσης. Εναλλακτικά ορισμένοι αλγόριθμοι μπορούν να καταφύγουν σε προσωρινή εξωτερική αποθήκευση των δεδομένων ή να επεξεργάζονται κάθε ένα υποσύνολο δεδομένων χωριστά κάθε φορά . Η βασική διαδικασία εκμάθησης διαχειρίζεται από τους επιστήμονες ουσιαστικά σαν μια αυξημένη έκδοση της κλασικής μηχανικής μάθησης όπου η μάθηση θεωρείται μια μοναδική ενδεχομένως και ακριβή υπολογιστικά διαδικασία όπου ένα σύνολο δεδομένων εκπαίδευσης χρησιμοποιείται για να εξάγει ένα τελικό στατικό μοντέλο (Bifet & Kirkby 2009).

Η προσέγγιση της εξόρυξης γνώσης από δεδομένα ναι μεν επιτρέπει την διαχείριση μεγαλύτερων συνόλων από δεδομένα από ότι η μηχανική μάθηση αλλά παρόλα αυτά δεν επιλύει το πρόβλημα της συνεχόμενης ροής δεδομένων (Bifet & Kirkby 2009). Τυπικά ένα μοντέλο το οποίο δημιουργήθηκε από πριν δεν μπορεί να ενημερωθεί όταν νέες πληροφορίες καταφθάνουν. Αντ' αυτού ολόκληρη η διαδικασία εκπαίδευσης πρέπει να επαναληφθεί εφόσον συμπεριληφθούν και τα νέα δεδομένα. Υπάρχουν περιπτώσεις όπου αυτός ο

περιορισμός είναι ανεπιθύμητος και ενδεχομένως να οδηγεί σε ανεπαρκή μοντέλο.

Η περίπτωση των δεδομένων ροής προέκυψε πρόσφατα εξαιτίας του προβλήματος του διαρκώς αυξανόμενου όγκου των δεδομένων. Οι αλγόριθμοι που έχουν γραφτεί για δεδομένα ροής είναι σε θέση να αντεπεξέλθουν με μεγέθη δεδομένων που είναι πολλές φορές μεγαλύτερα από την μνήμη και μπορούν να επεκταθούν στο να αντιμετωπίζουν δύσκολες πραγματικού χρόνου εφαρμογές οι οποίες δεν ήταν σε θέση να αντιμετωπιστούν από την μηχανική μάθηση ή από την κλασική εξόρυξη γνώσης . Η κύρια υπόθεση που γίνεται στην επεξεργασία δεδομένων ροής είναι ότι τα στιγμιότυπα εκπαίδευσης μπορούν να εξεταστούν για μικρό χρονικό διάστημα για μια μόνο φορά διότι καταφθάνουν σε ρεύματα με μεγάλη ταχύτητα και έπειτα να απορριφθούν για να δημιουργηθεί χώρος για τα επακόλουθα παραδείγματα (Bifet & Kirkby 2009). Ο αλγόριθμός που επεξεργάζεται το ρεύμα δεν έχει έλεγχο στην σειρά με την οποία θα εξετάσει τα παραδείγματα και πρέπει να ανανεώνει το μοντέλο του αυξητικά καθώς εξετάζει το κάθε παράδειγμα. Μια επιπλέον επιθυμητή ιδιότητα η αποκαλούμενη ανά πάσα στιγμή ιδιότητα (anytime property) απαιτεί το μοντέλο να είναι σε θέση να εφαρμοστεί σε οποιαδήποτε στιγμή μεταξύ των παραδειγμάτων εκπαίδευσης.

Η μελέτη αποκλειστικά των θεωρητικών πλεονεκτημάτων των αλγορίθμων είναι σίγουρα χρήσιμη και επιτρέπει νέες καινοτομίες αλλά οι απαιτήσεις των δεδομένων ροής αναγκάζουν αυτή την μελέτη να επεκταθεί και να επιβεβαιωθεί με πειραματικές μεθόδους. Το να δηλωθεί ότι ένας αλγόριθμος είναι κατάλληλος για κάποια σενάρια δεδομένων ροής προϋποθέτει ότι διαθέτει τις αναγκαίες πρακτικές ικανότητες. Προφανώς οι αμφιβολίες είναι ισχυρές αν αυτά που υποστηρίζεται ότι κάνει ένας αλγόριθμος δεν επιβεβαιώνονται από εμπειρικές αποδείξεις. Υπό αυτό το πρίσμα στη συνέχεια θα εξετάσουμε πειραματικά τους αλγόριθμους μείωσης δεδομένων που αναλύσαμε στο προηγούμενο κεφάλαιο

Ένα ακόμα σημαντικό κομμάτι είναι η αξιολόγηση των αλγορίθμων κατηγοριοποίησης σε δεδομένα ροής. Οι συγκεκριμένοι αλγόριθμοι απαιτούν κατάλληλες και ολοκληρωμένες μεθόδους εκτίμησης. Η εκτίμηση σύμφωνα με τους Bifet και Kirkby (2009) θα πρέπει να επιτρέπει στους χρήστες να είναι

βέβαιοι ότι συγκεκριμένα προβλήματα μπορούν να αντιμετωπιστούν, να ποσοτικοποιούν βελτιώσεις σε αλγόριθμους και να καθορίζουν ποιοι αλγόριθμοι είναι οι πιο κατάλληλοι για το πρόβλημα τους. Το MOA λαμβάνοντας υπόψη όλες τις παραπάνω παραμέτρους προτείνεται από τους Bifet και Kirkby (2009) ως το πλέον κατάλληλο εργαλείο.
Η μέτρηση της απόδοσης των αλγορίθμων κατηγοριοποίησης σε δεδομένα ροής είναι ένα τρισδιάστατο πρόβλημα το όποιο έχει ως παραμέτρους την υπολογιστική ταχύτητα, την μνήμη και την ακρίβεια. Δεν είναι δυνατόν να εφαρμόσουμε και ταυτοχρόνως να μετρήσουμε και τις τρεις παραμέτρους για αυτό τον λόγο στο MOA είναι αναγκαίο να ρυθμίσουμε το μέγεθος της μνήμης και έπειτα να καταγράψουμε τα αποτελέσματα των άλλων δύο παραμέτρων. Το λογισμικό MOA έχει αναπτυχθεί με σκοπό να παρέχει μια χρήσιμη εικόνα σχετικά με την απόδοση των αλγορίθμων κατηγοριοποίησης.

3.1.2 Υποθέσεις του MOA

Το λογισμικό MOA απασχολείται με το πρόβλημα της κατηγοριοποίησης η οποία είναι ίσως η πιο μελετημένη λειτουργία της μηχανικής μάθησης. Ο στόχος της κατηγοριοποίησης είναι να παράξει ένα μοντέλο το οποίο μπορεί να προβλέψει την κλάση παραδειγμάτων χωρίς ετικέτα. Το μοντέλο αυτό εκπαιδεύεται με παραδείγματα στα οποία η ετικέτα ή η κλάση τους είναι γνωστή. Για να αποσαφηνιστεί ο χώρος του προβλήματος το οποίο αντιμετωπίζεται οι Bifet και Kirkby (2009) κάνουν κάποιες υποθέσεις σχετικά με το τυπικό σενάριο εκμάθησης :

1. Τα δεδομένα θεωρείτε ότι έχουν ένα μικρό και καθορισμένο σύνολο από στήλες οι οποίες ονομάζονται γνωρίσματα ή χαρακτηριστικά . Πρέπει να είναι έως και κάποιες εκατοντάδες το πολύ τα γνωρίσματα αυτά.

2. Ο αριθμός των γραμμών ή αλλιώς τα παραδείγματα είναι πολύ μεγάλος και αριθμεί εκατομμύρια παραδείγματα στην μικρότερη των περιπτώσεων. Για την ακρίβεια οι αλγόριθμοι θα πρέπει να έχουν ικανότητα να επεξεργάζονται άπειρου πλήθους δεδομένα το οποίο ουσιαστικά σημαίνει ότι δεν υπερβούν

όρια της μνήμης διαφορετικά έχουν αποτύχει ανεξάρτητα από το πόσα παραδείγματα εκπαίδευσης έχουν επεξεργαστεί .

3. Τα δεδομένα έχουν έναν περιορισμένο αριθμό από πιθανές κλάσεις τυπικά λιγότερες από δέκα.

4. Το ποσό της μνήμης που είναι διαθέσιμο για έναν αλγόριθμο εκμάθησης εξαρτάται από την εφαρμογή. Το σύνολο των δεδομένων εκπαίδευσης θα είναι σημαντικά πιο μεγάλο από την διαθέσιμη μνήμη.

5. Θα πρέπει να υπάρχει ένα μικρό άνω όριο στον χρόνο που θα επιτραπεί στον αλγόριθμο να εκπαιδεύσει ή να κατηγοριοποιήσει ένα παράδειγμα. Αυτό επιτρέπει στους αλγόριθμους να αυξάνουν γραμμικά με τον αριθμό των παραδειγμάτων έτσι ώστε οι χρήστες να μπορούν να επεξεργαστούν Ν φορές περισσότερα από ένα δοσμένο σύνολο δεδομένων απλά περιμένοντας Ν φορές περισσότερο από ότι έχουν περιμένει.

6. Τα concepts ρευμάτων θεωρούνται ότι είναι στατικά ή αναπτυσσόμενα. Το concept drift συμβαίνει όταν το υποκείμενο concept που καθορίζει τον στόχο ο οποίος μαθαίνεται αρχίζει να αλλάζει με την πάροδο του χρόνου.

Οι πρώτες τρεις υποθέσεις δίνουν έμφαση στο ότι ο στόχος είναι να αυξήσουμε την ακρίβεια με το πλήθος των παραδειγμάτων. Πηγές δεδομένων που είναι μεγάλες σε άλλες παραμέτρους όπως ο αριθμός των γνωρισμάτων ή των πιθανών ετικετών δεν είναι ο σκοπός του πεδίου προβλήματος. Οι υποθέσεις 4 και 5 επισημαίνουν αυτό που χρειάζεται από μια λύση. Σχετικά με την υπόθεση 6 κάποιοι ερευνητές τονίζουν ότι το να αντιμετωπιστεί το φαινόμενο του concept drift είναι το πιο σημαντικό ζήτημα στην επεξεργασία δεδομένων ροής.

3.1.3 Απαιτήσεις που πρέπει να πληρούν οι αλγόριθμοι σε δεδομένα ροής

Οι συμβατικές μέθοδοι της μηχανικής μάθησης λειτουργούν υποθέτοντας ότι τα δεδομένα εκπαίδευσης είναι διαθέσιμα με την μορφή ενός ολόκληρου συνόλου

και κάθε παράδειγμα μπορεί να ανακτηθεί όπως είναι με ελάχιστο υπολογιστικό κόστος. Η εναλλακτική είναι να θεωρήσουμε τα δεδομένα εκπαίδευσης ως ένα ρεύμα δηλαδή μια εν δύναμη ατελείωτη ροή δεδομένων που καταφθάνει με τρόπο που δεν μπορεί να ελεγχθεί. Ένας αλγόριθμος ικανός να μάθει από ένα ρεύμα είναι εξ ορισμού ένας αλγόριθμος εξόρυξης γνώσης.

Η χρήση της κατηγοριοποίησης σε ένα σενάριο δεδομένων ροής προσφέρει αρκετά πλεονεκτήματα. Όχι μόνο αντιμετωπίζονται οι περιοριστικές υποθέσεις που γίνονται στις τεχνικές της βασικής μηχανικής μάθησης αλλά μπορούν να αντιμετωπιστούν και άλλες εφαρμογές ακόμα πιο απαιτητικές από ότι η εξόρυξη γνώσης από μεγάλες βάσεις δεδομένων. Ένα παράδειγμα μιας τέτοιας εφαρμογής είναι η παρακολούθηση ενός υψηλής ταχύτητας δικτύου δεδομένων όπου η ατελείωτη ροή δεδομένων είναι συντριπτική για να θεωρηθεί καν δυνατό να αποθηκευτούν τα δεδομένα και να ανακτηθούν στο μέλλον.

Ένας αλγόριθμος κατηγοριοποίησης πρέπει να ικανοποιεί κάποιες απαιτήσεις για να είναι σε θέση να λειτουργήσει με βάση τις υποθέσεις και να είναι κατάλληλος για να εκμάθηση από δεδομένα ροής. Οι απαιτήσεις σύμφωνα με τους Bifet και Kirkby (2009) είναι τέσσερις και περιγράφονται λεπτομερώς παρακάτω.

<u>1η απαίτηση</u> : Επεξεργασία ενός παραδείγματος κάθε φορά και εξέταση του μόνο μια φορά .

Το σημαντικό χαρακτηριστικό των δεδομένων ροής είναι ότι τα δεδομένα «ρέουνε» το ένα παράδειγμα μετά το άλλο. Δεν υπάρχει δυνατότητα για τυχαία πρόσβαση των δεδομένων που παρέχονται. Κάθε παράδειγμα πρέπει να γίνεται δεκτό καθώς καταφθάνει και με την σειρά που καταφθάνει. Μόλις εξεταστεί ή αγνοηθεί το παράδειγμα απορρίπτεται χωρίς να υπάρχει δυνατότητα επανάκτησης του.

Παρόλο που αυτή η απαίτηση υπάρχει στην είσοδο του αλγορίθμου δεν υπάρχει κανόνας που να απαγορεύει έναν αλγόριθμο να θυμάται παραδείγματα

εσωτερικά βραχυπρόθεσμα. Ένα παράδειγμα αυτής της περίπτωσης είναι ο αλγόριθμος ο οποίος αποθηκεύει ένα σύνολο παραδειγμάτων για χρήση από ένα συμβατικό σύστημα . Ενώ ο αλγόριθμος είναι ελεύθερος να ενεργήσει με αυτό τον τρόπο θα πρέπει να απορρίψει τα αποθηκευμένα παραδείγματα από κάποιο σημείο και μετά διότι πρέπει να τηρεί και την δεύτερη απαίτηση .

Ο κανόνας της μια φοράς της εξέτασης ενός παραδείγματος μπορεί να μην τηρηθεί σε περιπτώσεις όπου είναι πρακτικό να ξαναστείλουμε ολόκληρο το ρεύμα , το οποίο είναι ισοδύναμο με την δυνατότητα πολλαπλών σκαναρισμάτων της βάσης δεδομένων. Σε αυτή την περίπτωση ένας αλγόριθμος έχει την ευκαιρία κατά τα επαναλαμβανόμενα περάσματα να βελτιώσει το μοντέλο το οποίο έμαθε . Ωστόσο ένας αλγόριθμος γενικά που απαιτεί περισσότερα από ένα περάσματα για να λειτουργήσει δεν είναι ευέλικτος αρκετά για καθολική χρήση στα δεδομένα ροής.

2η απαίτηση : **Χρήση περιορισμένης μνήμης**

Το κυριότερο κίνητρο για την κατασκευή των μοντέλων που δουλεύουν πάνω σε δεδομένα ροής είναι ότι επιτρέπουν την επεξεργασία δεδομένων τα οποία είναι πολλές φορές μεγαλύτερα από την διαθέσιμη λειτουργική μνήμη. Ο κίνδυνος που προκύπτει από την επεξεργασία τόσο τεράστιων όγκων δεδομένων είναι ότι η μνήμη θα εξαντληθεί πολύ γρήγορα.

Η μνήμη που χρησιμοποιείται από έναν αλγόριθμο μπορεί να χωριστεί σε δύο κατηγορίες: στην μνήμη που χρησιμοποιείται για να αποθηκευτούν τα στατιστικά που είναι σε εξέλιξη και στην μνήμη που χρησιμοποιείται για να αποθηκευτεί το τρέχων μοντέλο. Για τον πιο αποδοτικό στη μνήμη αλγόριθμο αυτές οι δύο κατηγορίες θα είναι το ένα και το αυτό δηλαδή οι στατιστικές αποτελούν το μοντέλο το οποίο χρησιμοποιείται για πρόβλεψη .

Αυτός ο περιορισμός στην μνήμη είναι φυσικός και μπορεί να χαλαρώσει μόνο αν χρησιμοποιηθεί εξωτερικός χώρος αποθήκευσης όπου θα αποθηκεύονται προσωρινά τα παραδείγματα. Κάθε τέτοια χρήση εξωτερικού χώρου αποθήκευσης πρέπει να γίνει λαμβάνοντας υπόψη την τρίτη απαίτηση.

<u>3η απαίτηση</u> : Λειτουργία σε περιορισμένο χρόνο

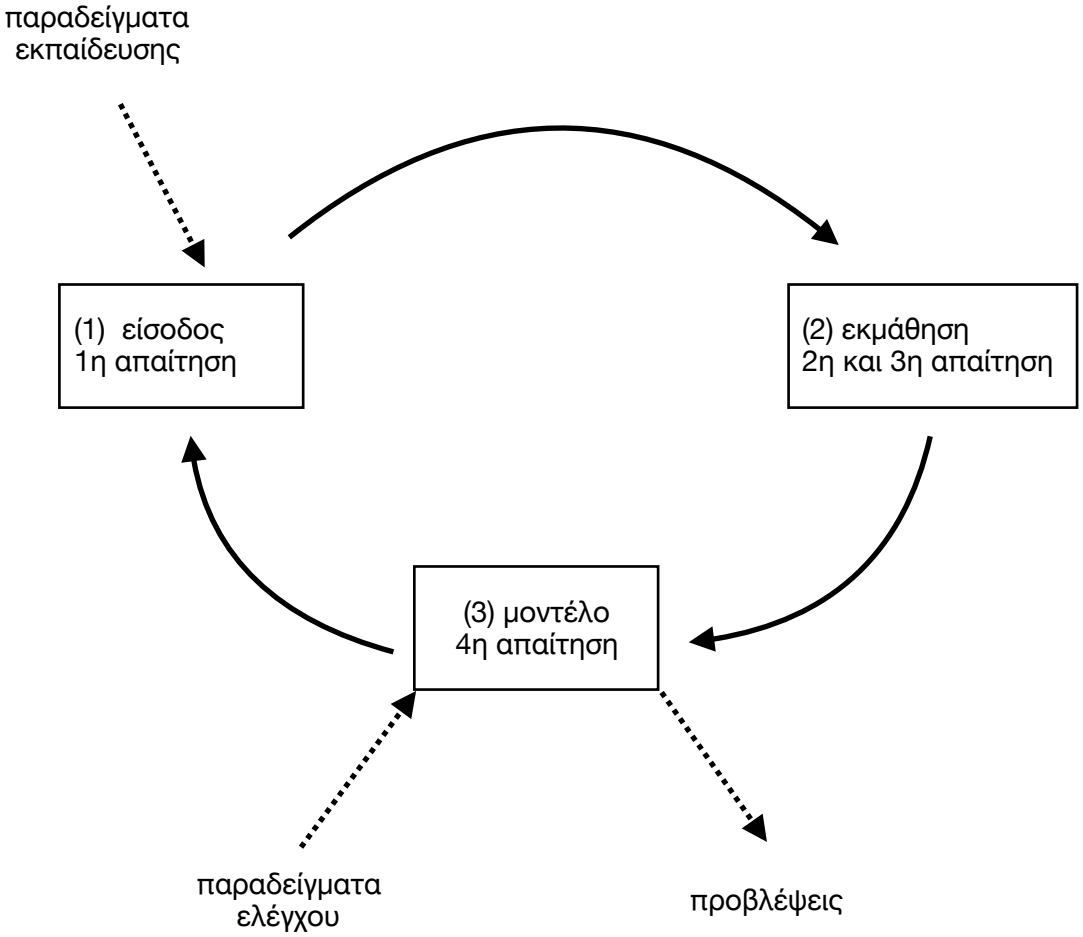
Για να εξετάσει ένας αλγόριθμος άνετα οποιοδήποτε πλήθος παραδειγμάτων η χρονική πολυπλοκότητα πρέπει να είναι γραμμική του αριθμού των παραδειγμάτων. Αυτό μπορεί να επιτευχθεί στο πλαίσιο της εξόρυξης γνώσης από δεδομένα ροής αν υπάρχει μια σταθερά κατά προτίμηση μικρή η οποία θα είναι το άνω χρονικό όριο της επεξεργασίας κάθε παραδείγματος.

Επιπλέον αν ένας αλγόριθμος πρέπει να είναι σε θέση να λειτουργεί σε πραγματικό χρόνο τότε πρέπει να επεξεργάζεται τα παραδείγματα το ίδιο γρήγορα αν όχι ταχύτερα από ότι καταφθάνουν. Σε περίπτωση που δεν το επιτυγχάνει αυτό τότε αναπόφευκτα υπάρχει απώλεια δεδομένων.

Ο αυστηρός συγχρονισμός δεν είναι καίριας σημασίας σε εφαρμογές με μικρότερες απαιτήσεις όπως όταν ο αλγόριθμος χρησιμοποιείται για να κατηγοριοποιήσει μια μεγάλη αλλά διαρκή πηγή δεδομένων. Ωστόσο όσο πιο αργός είναι ο αλγόριθμος τόσο λιγότερο χρήσιμος θα είναι για χρήστες οι οποίοι χρειάζονται αποτελέσματα μέσα σε ένα λογικό χρονικό πλαίσιο.

<u>4η απαίτηση</u> : Ο αλγόριθμος να είναι σε θέση να κάνει πρόβλεψη οποιαδήποτε στιγμή

Ένας ιδανικός αλγόριθμος θα πρέπει να είναι σε θέση να παράξει το βέλτιστο μοντέλο που μπορεί από τα δεδομένα που έχει εξετάσει αφού δει οποιοδήποτε πλήθος παραδειγμάτων. Στην πράξη είναι πολύ πιθανόν ότι θα υπάρχουν περίοδοι όπου το μοντέλο θα παραμένει σταθερό όπως όταν ένα batch-base αλγόριθμος αποθηκεύει το νέο σύνολο (batch). Η διαδικασία παραγωγής του μοντέλου θα πρέπει να είναι όσο πιο αποδοτική γίνεται , με το καλύτερο σενάριο να είναι να μην χρειάζεται μετάφραση. Αυτό σημαίνει ότι το τελικό μοντέλο διαχειρίζεται απευθείας στην μνήμη από τον αλγόριθμο καθώς επεξεργάζεται παραδείγματα από ότι να πρέπει να ξαναυπολογιστεί το μοντέλο με βάση τις τρέχουσες στατιστικές .



Εικόνα 3.2: Ο κύκλος κατηγοριοποίησης δεδομένων ροής (Bifet & Kirkby 2009)

Η εικόνα 3.2 αποτυπώνει την τυπική χρήση ενός αλγορίθμου κατηγοριοποίησης για δεδομένα ροής και πώς οι απαιτήσεις που έχουμε εφαρμόζονται. Το γενικό μοντέλο της κατηγοριοποίησης δεδομένων ροής ακολουθεί τα επόμενα τρία βήματα μέσω ενός επαναλαμβανόμενου κύκλου :

1. Ο αλγόριθμος δέχεται το επόμενο διαθέσιμο παράδειγμα από την ροή (απαίτηση 1)

2. Ο αλγόριθμος επεξεργάζεται το παράδειγμα και αναβαθμίζει τις δομές δεδομένων του. Αυτό το πραγματοποιεί χωρίς να υπερβαίνει τα όρια της μνήμης που υπάρχουν (απαίτηση 2) και όσο το δυνατόν ταχύτερα (απαίτηση 3)

3. Ο αλγόριθμος είναι έτοιμος να δεχθεί το επόμενο παράδειγμα. Οποιαδήποτε στιγμή αν απαιτηθεί ο αλγόριθμος είναι έτοιμος να παρέχει ένα μοντέλο το οποίο μπορεί να κάνει πρόβλεψη για την κλάση παραδειγμάτων που δεν έχουν εξεταστεί ακόμα (απαίτηση 3)

3.1.4 Διαδικασία αξιολόγησης αλγορίθμων σε δεδομένα ροής

Αυτή η παράγραφος υποθέτει ότι η καίρια μεταβλητή η οποία μετράτε από τις διαδικασίες εκτίμησης είναι η ακρίβεια του αλγορίθμου εκμάθησης. Η ακρίβεια ή ισοδύναμα το αντίθετο της δηλαδή το σφάλμα δεν είναι η μόνη μεταβλητή που πρέπει να μας απασχολεί αλλά είναι συνήθως η πιο συναφής. Η ακρίβεια τυπικά μετράται ως το ποσοστό των σωστά κατηγοριοποιημένων παραδειγμάτων που κατηγοριοποιεί ένα μοντέλο δοθέντος ενός συνόλου δεδομένων. Ο πιο εύστοχος αλγόριθμος είναι αυτός που κάνει τα λιγότερα λάθη όταν προβλέπει τις ετικέτες των παραδειγμάτων. Σε προβλήματα κατηγοριοποίησης η επίτευξη της μέγιστης δυνατής ακρίβειας είναι ο πιο άμεσος και προφανής στόχος. Η αξιόπιστη εκτίμηση για την ακρίβεια ενός αλγορίθμου επιτρέπει την σύγκριση διαφορετικών μοντέλων έτσι ώστε η καλύτερη διαθέσιμη μέθοδος για ένα δοσμένο πρόβλημα να μπορεί να καθοριστεί .

Είναι πολύ αισιόδοξο σενάριο να μετρηθεί η ακρίβεια που επιτυγχάνεται από ένα μοντέλο εκμάθησης στα ίδια δεδομένα που χρησιμοποιήθηκαν για την εκπαίδευση του διότι ακόμα και αν το μοντέλο πετύχει τέλεια ακρίβεια στα δεδομένα εκπαίδευσης κάτι τέτοιο δεν μπορεί συναχθεί για την ακρίβεια που θα επιτύχει σε δεδομένα που δεν έχει ακόμα δει. Για την εκτίμηση ενός αλγορίθμου εκμάθησης πρέπει να εξεταστεί η ικανότητα του να γενικεύει σε άγνωστα παραδείγματα που δεν έχει δει προηγουμένως. Ένα μοντέλο λέγεται ότι υπερμοντελοποιεί (overfit) τα δεδομένα αν δυσκολεύεται πολύ να ερμηνεύσει τα δεδομένα εκπαίδευσης τα οποία είναι συνήθως θόρυβος επομένως έχει χαμηλή προβλεπτική ικανότητα όταν προβλέπει την κλάση της ετικέτας από παραδείγματα που δεν έχει ακόμα δει. Μια από τις μεγαλύτερες προκλήσεις της μηχανικής μάθησης είναι να κατασκευάσει αλγορίθμους που μπορούν να αποφύγουν το πρόβλημα της υπερμοντελοποίησης (Bifet & Kirkby 2009).

Το περιβάλλον των δεδομένων ροής έχει διαφορετικές απαιτήσεις από αυτό των στατικών δεδομένων. Συγκεκριμένα η αξιολόγηση σε αλγόριθμους εκμάθησης σε στατικά δεδομένα επικεντρώνεται στην επαναχρησιμοποίηση των δεδομένων για να τα αξιοποιήσει στο μέγιστο δυνατό δεδομένου ότι έχει ένα μειωμένο αριθμό παραδειγμάτων. Στην περίπτωση των δεδομένων ροής υπάρχει αφθονία στα δεδομένα επομένως η επαναχρησιμοποίηση τους δεν είναι αναγκαία. Επομένως εφόσον υπάρχει τεράστιο πλήθος δεδομένων η γενικευμένη ακρίβεια μπορεί να μετρηθεί με την μέθοδο holdout χωρίς να υπάρχουν τα προβλήματα που παρουσιάζονται στο σενάριο των στατικών δεδομένων τα οποία και οδήγησαν στους ερευνητές στην αναζήτηση εναλλακτικών μεθόδων αξιολόγησης. Η σημαντική διαφορά είναι ότι ένα μεγάλο σύνολο από παραδείγματα μπορεί να αποθηκευτεί για λίγο για να χρησιμοποιηθεί για έλεγχο του μοντέλου και για μέτρηση της ακρίβειας χωρίς να υπάρχει έλλειμμα στον αλγόριθμο εκμάθησης από παραδείγματα εκπαίδευσης.

Η διαδικασία αξιολόγησης ενός αλγόριθμου εκμάθησης καθορίζει ποια παραδείγματα χρησιμοποιούνται για εκπαίδευση του αλγορίθμου και ποια χρησιμοποιούνται για να ελεγχθεί το αποτέλεσμα μοντέλου. Η διαδικασία που χρησιμοποιείται παραδοσιακά στους αλγορίθμους στατικών δεδομένων

εξαρτάται εν μέρη από το μέγεθος των δεδομένων. Μικρά σύνολα δεδομένων με λιγότερο από χίλια παραδείγματα είναι αρκετά στην μηχανική μάθηση σε στατικά δεδομένα διότι υπάρχουν μέθοδοι που διενεργούν την μέγιστη δυνατή χρήση των δεδομένων για αυτό και έχει εγκαθιδρυθεί ως διαδικασία αξιολόγησης η δέκα φορές επαναλαμβανόμενη διασταυρωμένη επικύρωση με δέκα φακέλους. Καθώς το μέγεθος των συνόλων αυξάνεται πρακτικοί χρονικοί περιορισμοί αποτρέπουν διαδικασίες που επαναλαμβάνουν την εκπαίδευση πολλές φορές. Είναι κοινώς αποδεκτό ότι με αρκετά πιο μεγάλες πηγές δεδομένων είναι αναγκαίο να μειωθεί ο αριθμός των επαναλήψεων ή των φακέλων για να επιτραπεί στο πείραμα να ολοκληρωθεί μέσα σε λογικό χρόνο. Σε περίπτωση που υπάρχουν δεδομένα εκατοντάδων χιλιάδων παραδειγμάτων στην εκμάθηση από στατικά δεδομένα τότε χρησιμοποιείται η μέθοδος holdout καθώς αυτή απαιτεί την λιγότερη υπολογιστική προσπάθεια. Μια δικαίωση για την χρήση αυτής της μεθόδου είναι ότι πέρα από τον χρόνο που είναι λιγότερος η αξιοπιστία που χάνεται λόγω της έλλειψης και άλλων επαναλήψεων ανταμείβεται από την αξιοπιστία που κερδίζεται από το μεγάλο πλήθος των παραδειγμάτων που έχουν συμπεριληφθεί στην διαδικασία.

Όταν αναλογιζόμαστε τι διαδικασία πρέπει να χρησιμοποιηθεί στο πεδίο των δεδομένων ροής μια από τις κύριες ανησυχίες που πρέπει να ληφθούν υπόψη είναι το πως θα χτιστεί μια εικόνα της ακρίβειας του μοντέλου με την πάροδο του χρόνου. Δύο κύριες προσεγγίσεις έχουν αναπτυχθεί, σύμφωνα με τους Bifet και Kirkby (2009, p.26) η πρώτη είναι μια φυσική επέκταση της μεθόδου holdout που χρησιμοποιείται στα στατικά δεδομένα και η δεύτερη είναι μια ενδιαφέρουσα εκμετάλλευση ιδιοτήτων που είναι μοναδικές στους αλγόριθμους δεδομένων ροής. Οι δύο μέθοδοι παρουσιάζονται παρακάτω.

Η μέθοδος Holdout

Όταν η εκμάθηση σε στατικά δεδομένα φτάνει σε ένα επίπεδο όπου η διασταυρωμένη επικύρωση είναι πολύ χρονοβόρα γίνεται αποδεκτό ότι είναι καλύτερα να μετρηθεί η απόδοση του αλγορίθμου με ένα holdout σύνολο. Αυτό είναι ιδιαίτερα χρήσιμο όταν ο διαχωρισμός μεταξύ του συνόλου ελέγχου και του συνόλου εκπαίδευσης έχει προκαθοριστεί έτσι ώστε τα αποτελέσματα από διαφορετικές έρευνες μπορούν να συγκριθούν άμεσα. Εξετάζοντας τα

προβλήματα δεδομένων ροής ως προβλήματα στατικών δεδομένων μεγάλης κλίμακας τότε γίνεται αποδεκτό ότι ένα holdout σύνολο είναι κατάλληλο για να χρησιμοποιηθεί και στην περίπτωση των δεδομένων ροής.

Για να παρακολουθείται η επίδοση ενός μοντέλου καθώς ο χρόνος περνά το μοντέλο μπορεί να εκτιμάται περιοδικά για παράδειγμα μετά από ένα εκατομμύριο παραδείγματα. Το να εξετάζουμε το μοντέλο πολύ συχνά έχει ως πιθανή συνέπεια να επιβραδυνθεί η διαδικασία εκτίμησης το οποίο εξαρτάται και από το μέγεθος του συνόλου ελέγχου.

Μια πιθανή πηγή για παραδείγματα για την μέθοδο holdout είναι νέα παραδείγματα από το ρεύμα τα οποία δεν έχουν χρησιμοποιηθεί για να εκπαιδευτεί ο αλγόριθμος εκμάθησης. Μια διαδικασία μπορεί να αναζητήσει μια παρτίδα από παραδείγματα από το ρεύμα για να χρησιμοποιηθούν ως παραδείγματα ελέγχου και αν απαιτείται η αποδοτική χρήση των παραδειγμάτων τότε μπορούν να δοθούν έπειτα στον αλγόριθμο για περαιτέρω εκπαίδευση όταν ο έλεγχος έχει ολοκληρωθεί. Αυτή η μέθοδος είναι προτιμότερη σε σενάρια που υπάρχει concept drift καθώς μετράει την ικανότητα ενός αλγορίθμου να προσαρμόζεται στην τελευταία τάση των δεδομένων.

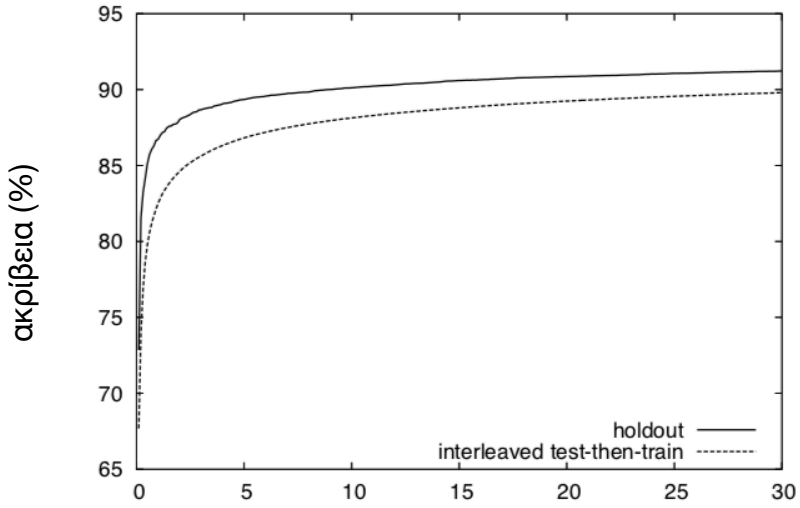
Στην περίπτωση που θεωρείται ότι δεν υπάρχει concept drift ένα στατικό hold out σύνολο θα είναι αρκετό διότι αποφεύγεται το πρόβλημα των πολλαπλών εκτιμήσεων μεταξύ πιθανών συνόλων ελέγχου. Θεωρώντας ότι το σύνολο ελέγχου είναι ανεξάρτητο και επαρκώς μεγάλο σε σχέση με την πολυπλοκότητα του concept θα παρέχει μια ακριβή μέτρηση της γενικευμένης ακρίβειας. Σύμφωνα με άλλες έρευνες τα σύνολα ελέγχου που είναι της τάξης των δεκάδων χιλιάδων παραδειγμάτων θεωρούνται επαρκή.

Η μέθοδος Prequential ή Interleaved Test-Then-Train

Στην εξόρυξη γνώσης από δεδομένα ροής η πλέον χρησιμοποιούμενη μέθοδος εκτίμησης είναι η prequential ή αλλιώς interleaved test - then -train . Η ιδέα της είναι πολύ απλή, χρησιμοποιεί κάθε στιγμιότυπο πρώτα για να ελέγξει το μοντέλο και έπειτα για να το εκπαιδεύσει και από αυτό η ακρίβεια του μοντέλου ανανεώνεται αυξητικά . Όταν σκόπιμα εκτελείτε η εκτίμηση σε αυτή την σειρά το μοντέλο πάντα εξετάζεται σε παραδείγματα τα οποία δεν έχει δει . Αυτό το σχέδιο έχει το πλεονέκτημα ότι δεν χρειάζεται κάποιο holdout σύνολο για έλεγχο

διότι κάνει μέγιστη χρήση των διαθέσιμων δεδομένων. Σύμφωνα με τους Bifet και Kirkby (2009, p.26) αυτή η μέθοδος εξασφαλίζει μια πιο ομαλή απεικόνιση της ακρίβειας με την πάροδο του χρόνου καθώς κάθε παράδειγμα θα γίνεται ολοένα και λιγότερο σημαντικό στον συνολικό μέσο. Τα μειονεκτήματα αυτής της προσέγγισης είναι ότι κάνει δύσκολο τον ακριβή διαχωρισμό και την μέτρηση των χρόνων εκπαίδευσης και ελέγχου. Ακόμη η αληθινή ακρίβεια που ένας αλγόριθμος είναι ικανός να πετύχει σε κάθε χρονική στιγμή είναι δυσδιάκριτη, οι αλγόριθμοι θα τιμωρούνται για τα αρχικά τους λάθη ανεξάρτητα από το επίπεδο της ακρίβειας που είναι ικανοί να καταφέρουν πιο μετά παρόλο που φυσικά αυτή η χαμηλή ακρίβεια από τα αρχικά λάθη θα εξαφανιστεί με την πάροδο του χρόνου καθώς νέα στιγμιότυπα εξετάζονται (Bifet & Kirkby 2009). Με αυτή την μέθοδο τα στατιστικά ενημερώνονται με κάθε παράδειγμα στο ρεύμα και μπορούν να καταγραφούν σε αυτό το επίπεδο λεπτομέρειας αν είναι επιθυμητό. Για λόγους αποδοτικότητας μια παράμετρος δειγματοληψίας μπορεί να χρησιμοποιηθεί για να μειωθούν οι απαιτήσεις αποθήκευσης των αποτελεσμάτων καταγράφοντας τα στατιστικά μόνο κατά περιόδους που επιλέγουμε εμείς όπως στην μέθοδο holdout.

Σύγκριση των δύο μεθόδων



Εικόνα 3.3: Καμπύλες εκμάθησης που κατασκευάστηκαν για τον ίδιο αλγόριθμο από τις δύο μεθόδους εκτίμησης (Bifet & Kirkby 2009)

Η εικόνα 3.2 είναι ένα παράδειγμα του πως οι καμπύλες εκμάθησης μπορούν να διαφέρουν μεταξύ των δύο μεθόδων εκτίμησης δεδομένου ότι έχουμε ένα συγκεκριμένο αλγόριθμο και μια συγκεκριμένη πηγή δεδομένων. Η μέθοδος holdout μετράει την άμεση ακρίβεια που έχει το μοντέλο σε ένα συγκεκριμένο σημείο χωρίς να έχει μνήμη των προηγούμενων επιδόσεων του μοντέλου. Κατά την διάρκεια των πρώτων εκατομμυρίων στιγμιοτύπων το γράφημα που απεικονίζει την ακρίβεια είναι απότομο. Αν το σύνολο ελέγχου ήταν μικρό άρα και αναξιόπιστο ή ο αλγόριθμος είναι ασταθής τότε οι διακυμάνσεις στην ακρίβεια είναι περισσότερο ευδιάκριτες. Η prequential μέθοδος σε αντίθεση με την holdout μετράει την μέση ακρίβεια που επετεύχθη από το μοντέλο σε ένα συγκεκριμένο σημείο, συνεπώς μετά από τριάντα εκατομμύρια παραδείγματα η γενικευμένη ακρίβεια έχει μετρηθεί σε κάθε ένα από τα 30 εκατομμύρια παραδείγματα από ότι τα ανεξάρτητα ένα εκατομμύριο παραδείγματα που χρησιμοποιήθηκαν από την μέθοδο holdout. Αυτό εξηγεί σύμφωνα με τους Bifet και Kirkby (2009, p.27) γιατί η prequential μέθοδος έχει πιο ομαλή καμπύλη. Επίσης εξηγεί γιατί η εκτίμηση της ακρίβειας είναι πιο χαμηλή διότι κατά τα αρχικά στάδια της εκμάθησης το μοντέλο ήταν λιγότερο ακριβές χαμηλώνοντας έτσι την μέση ακρίβεια. Η prequential μέθοδος κάνει την μέτρηση τόσο του χρόνου όσο και της ακρίβειας πιο δύσκολη. Μπορεί να βελτιωθεί αν χρησιμοποιηθεί μια τροποποίηση που εισάγει εκθετική μείωση. Η μέθοδος εκτίμησης holdout προσφέρει το καλύτερο αποτέλεσμα από τις δύο μεθόδους διότι η μέση ακρίβεια που θα λαμβάναμε από την prequential μέθοδο μπορεί να εκτιμηθεί βρίσκοντας τον μέσο της ακρίβειας από διαδοχικά δείγματα.

Ο παρακάτω αλγόριθμος των Bifet και Kirkby (2009) παρουσιάζει έναν ψευδο-κώδικα της διαδικασίας εκτίμησης που χρησιμοποιείται από το λογισμικό MOA. Η διαδικασία είναι όμοια με αυτή που χρησιμοποιείται από τους Domingos και Hulten (2000) των οποίων η έρευνα βρέθηκε να έχει πολύ αναλυτικές πρακτικές εκτίμησης .

Αλγόριθμος διαδικασίας αξιολόγησης (Bifet & Kirkby 2009)

Καθορισμός του m_{bound} το μέγιστο χώρο μνήμης που επιτρέπεται να χρησιμοποιήσει το μοντέλο .
Holdout n_{test} παραδείγματα για έλεγχο

while (περισσότερη εκτίμηση χρειάζεται) **do**
ξεκίνα τον χρονομετρητή για την εκπαίδευση
for i =1 to n_{train} **do**
πάρε το επόμενο παράδειγμα e_{train} από το το ρεύμα εκπαίδευσης
εκπαίδευσε και ενημέρωσε το μοντέλο βεβαιώνοντας ότι το m_{bound} τηρείται
end for
τέλος χρονομετρητή εκπαίδευσης και καταγραφή χρόνου εκπαίδευσης
έναρξη χρονομετρητή ελέγχου
for i =1 to n_{train} **do**
πάρε το επόμενο παράδειγμα e_{test} από το ρεύμα ελέγχου
έλεγχος του μοντέλου με το e_{test} και ενημέρωση ακρίβειας
end for
τέλος χρονομετρητή ελέγχου και καταγραφή χρόνου ελέγχου
καταγραφή στατιστικών του μοντέλου (ακρίβεια , μέγεθος κτλ)
end while

Ο αλγόριθμος προσφέρει ευελιξία σχετικά με το ποια στατιστικά επιλέγουμε. Η n_{train} παράμετρος καθορίζει πόσα παραδείγματα θα χρησιμοποιηθούν για εκπαίδευση πριν εφαρμοστεί η εκτίμηση στο σύνολο ελέγχου. Ένα σύνολο από n_{test} παραδείγματα κρατιέται στην άκρη για να χρησιμοποιηθεί για έλεγχο. Σε περίπτωση δεδομένων ροής χωρίς το φαινόμενο του concept drift αυτό το σύνολο μπορεί εύκολα να αποτελείται από τα πρώτα n_{test} σε πλήθος παραδείγματα που συλλέγονται από το ρεύμα. Για να αποκτηθούν αξιόπιστοι

χρόνοι εκτίμησης τα σύνολα n_{test} και n_{train} πρέπει να είναι αρκετά μεγάλα. Σε πραγματικές εφαρμογές ο χρονομετρητής καταγράφει τον χρόνο που επεξεργάζεται τον αλγόριθμο η CPU με στόχο να μειωθεί το πρόβλημα που δημιουργείται από ένα λειτουργικό σύστημα που διενεργεί και άλλες εργασίες. Στα περισσότερα πειράματα το σύνολο n_{test} είχε ρυθμιστεί να έχει ένα εκατομμύριο παραδείγματα το οποίο βοηθάει να μετρηθεί ο χρόνος αλλά επιπλέον εξασφαλίζει αξιοπιστία στις εκτιμήσεις που γίνονται για την ακρίβεια.

Το πλαίσιο έχει σχεδιαστεί για να ελέγχει έναν αλγόριθμο ο οποίος τείνει να αθροίζει πληροφορία με τον χρόνο επομένως θα χρειάζεται και περισσότερη μνήμη καθώς εκπαιδεύει περισσότερα παραδείγματα. Ο αλγόριθμος θα πρέπει να είναι σε θέση να περιορίσει την συνολική μνήμη που χρησιμοποιεί επομένως πρέπει να υπακούει στο όριο m_{bound} ανεξάρτητα από πόσα παραδείγματα εκπαιδεύει (Bifet & Kirkby 2009, p.29).

Μια από τις μεγαλύτερες προκλήσεις στην αξιολόγηση είναι η απόφαση του πότε πρέπει να σταματήσει η εκπαίδευση και να αρχίσει ο έλεγχος (Bifet & Kirkby 2009). Σε περιπτώσεις που υπάρχει λίγη μνήμη διαθέσιμη κάποιοι αλγόριθμοι θα φτάσουν σε σημείο όπου θα έχουν εξαντλήσει όλη την μνήμη και δεν θα μπορούν πλέον να λάβουν νέα πληροφορία. Σε αυτό το σημείο το πείραμα μπορεί να τερματιστεί καθώς τα αποτελέσματα δεν πρόκειται να αλλάξουν από αυτό το σημείο και έπειτα.

Ακόμα πιο προβληματική είναι η κατάσταση κατά την οποία ο χρόνος ή τα παραδείγματα εκπαίδευσης έχουν εξαντληθεί πριν παρατηρηθεί το τελικό επίπεδο απόδοσης του αλγορίθμου. Κάποιοι αλγόριθμοι είναι καλύτεροι σε ακρίβεια από άλλους βραχυπρόθεσμα ωστόσο καθώς αυξάνεται ο αριθμός των παραδειγμάτων που βλέπουν οι αλγόριθμοι η κατάσταση αυτή μπορεί να αλλάξει μακροπρόθεσμα. Το ποιος αλγόριθμος τελικά είναι ο καλύτερος αυτό εξαρτάται από την εφαρμογή. Αν επομένως υπάρχει χρονικός περιορισμός ή έλλειψη σε παραδείγματα τότε ο αλγόριθμος που επιτυγχάνει καλύτερη ακρίβεια με αυτές τις παραμέτρους είναι ο βέλτιστος (Bifet & Kirkby 2009).

Για να αποκλειστεί οποιαδήποτε επίδραση έχει η σειρά με την οποία καταφθάνουν τα δεδομένα στη διαδικασία εκμάθησης η διαδικασία αξιολόγησης πρέπει να επαναληφθεί αρκετές φορές κάθε φορά με διαφορετικό σύνολο δεδομένων εκπαίδευσης από το ίδιο πρόβλημα. Οι παρατηρήσεις που

μαζεύονται από κάθε επανάληψη μπορούν μετά να δώσουν έναν γενικό μέσο ως αποτέλεσμα. Ένα πλεονέκτημα αυτής της προσέγγισης είναι ότι η διακύμανση της συμπεριφοράς μπορεί επιπλέον να παρατηρηθεί. Στην ιδανική περίπτωση τα δεδομένα μεταξύ των επαναλήψεων θα είναι μοναδικά κάτι το οποίο είναι εφικτό με δεδομένα που παράγονται από συνθετικές γεννήτριες ή από πηγές με άφθονα δεδομένα. Αν υπάρχει έλλειψη σε δεδομένα τότε μια κίνηση είναι η επαναδιάταξη των δεδομένων εκπαίδευσης (Bifet & Kirkby 2009).

Μια ιδανική μέθοδος αξιολόγησης θα περίμενε έως ότου η ακρίβεια έπιανε ένα σταθερό επίπεδο και θα επαναλαμβανόταν πολλές φορές για να διασφαλιστεί η αξιοπιστία των αποτελεσμάτων. Δυστυχώς καμία από τις δυο παραπάνω κινήσεις δεν είναι εφικτή άμα λάβουμε υπόψη το μεγάλο όγκο του πειραματικού έργου που χρειάζεται.

Το ερώτημα του πότε ένας αλγόριθμος θεωρείται καλύτερος από έναν άλλον καθορίζεται από την εξέταση τον τελικών αποτελεσμάτων. Τα αποτελέσματα της ακρίβειας αναφέρονται ως ποσοστά με δυο δεκαδικά ψηφία και αν η τελική ακρίβεια μιας μεθόδου είναι καλύτερη από μια άλλη τότε θεωρείται ανώτερη. Η μέτρηση του τυπικού σφάλματος των αποτελεσμάτων μέσα από πολλαπλές επαναλήψεις θα επέτρεπε τα αποτελέσματα να αναλυθούν πιο επίσημα αλλά κάθε επιπλέον επανάληψη θα πολλαπλασίαζε τις χρονικές απαιτήσεις. Μια εναλλακτική και με μικρότερο κόστος πρακτική είναι να εξεταστούν οι διαφορές μεταξύ των αλγορίθμων χρησιμοποιώντας το έλεγχο McNemar όπως αυτό προτείνεται από τον Diettrich (1998).

3.1.5 Πειραματικές ρυθμίσεις σε εξελισσόμενα ρεύματα

Αυτή η παράγραφος προτείνει ένα νέο πειραματικό πλαίσιο για τα δεδομένα ροής μελετώντας το concept drift με την χρήση του MOA (Bifet & Kirkby 2009). Η πλειοψηφία της έρευνας που αφορά το concept drift στην εξόρυξη γνώσης από δεδομένα ροής έχει πραγματοποιηθεί με την χρήση κλασικών εφαρμογών εξόρυξης γνώσης όπως το λογισμικό WEKA (Witten & Frank 2005). Δεδομένου ότι το πλαίσιο των δεδομένων ροής έχει περιορισμούς τους οποίους δεν έχουν τα κλασικά περιβάλλοντα εξόρυξη γνώσης οι ερευνητές Bifet και Kirkby (2009)

πιστεύουν ότι το λογισμικό MOA είναι πιο κατάλληλο για να βελτιωθεί η αξιολόγηση αυτών των μεθόδων.

Στην εξόρυξη γνώσης από δεδομένα ροής ενδιαφερόμαστε για τρεις διαστάσεις: ακρίβεια, απαιτούμενος χώρος ή μνήμη υπολογιστή και τον χρόνο που απαιτείται για να μάθει ο αλγόριθμος από τα παραδείγματα εκπαίδευσης αλλά και να προβλέψει. Αυτές οι ιδιότητες μπορεί να είναι αλληλένδετες. Η προσαρμογή του χρόνου και του χώρου που χρησιμοποιεί ένας αλγόριθμος μπορεί να επηρεάσει την ακρίβεια του. Η αποθήκευση περισσότερης προ-επεξεργασμένης πληροφορίας όπως πίνακες μπορεί να επιτρέψει έναν αλγόριθμο να τρέξει γρηγορότερα αλλά με κόστος τον διαθέσιμο χώρο. Ένας αλγόριθμος μπορεί επίσης να τρέξει γρηγορότερα με το να επεξεργάζεται λιγότερη πληροφορία είτε σταματώντας νωρίς είτε αποθηκεύοντας λιγότερη πληροφορία συνεπώς έχει λιγότερα δεδομένα να επεξεργαστεί κατά αυτόν τον τρόπο. Επιπλέον όσο περισσότερο χρόνο έχει ένας αλγόριθμος τόσο πιο πιθανό είναι η ακρίβεια να αυξηθεί. Όσον αφορά τώρα τα εξελισσόμενα ρεύματα δεδομένων μας απασχολούν τα εξής: η εξέλιξη της ακρίβειας, η πιθανότητα των λανθασμένων συναγερμών, η πιθανότητα των αληθινού εντοπισμού και ο μέσος χρόνος καθυστέρησης στον εντοπισμό της αλλαγής (Bifet & Kirkby 2009).

Πολλές φορές οι μέθοδοι εκμάθησης δεν έχουν ενσωματωμένους εντοπιστές αλλαγής επομένως είναι δύσκολο να καθοριστεί η αναλογία μεταξύ λανθασμένων θετικών και λανθασμένων αρνητικών καθώς και ο μέσος χρόνος καθυστέρησης στον εντοπισμό της αλλαγής (Bifet & Kirkby 2009). Σε αυτές τις περιπτώσεις οι καμπύλες εκμάθησης είναι μια χρήσιμη εναλλακτική για την παρατήρηση της εξέλιξης της ακρίβειας σε μεταβαλλόμενα περιβάλλοντα.

Συνοψίζοντας οι κύριες ιδιότητες για μια ιδανική μέθοδος εκμάθησης από εξελισσόμενα δεδομένα ροής είναι οι εξής: υψηλή ακρίβεια και άμεση προσαρμογή στην αλλαγή, χαμηλό υπολογιστικό κόστος τόσο σε χώρο όσο και σε χρόνο, εγγύηση στην θεωρητική απόδοση του μοντέλου και ελάχιστο αριθμό παραμέτρων.

Πλαίσιο concept drift

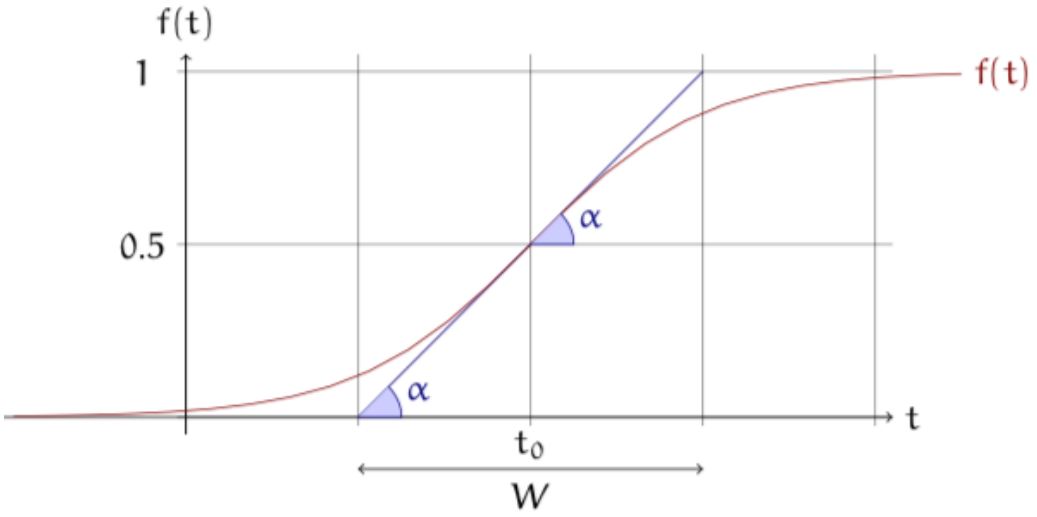
Ο στόχος των Bifet και Kirkby (2009, p.40) είναι η εισαγωγή τεχνητού drift σε γεννήτριες δεδομένων ροής με ευθύ τρόπο. Ένα πλαίσιο που είναι όμοιο με αυτό των Bifet και Kirkby (2009) έχει προταθεί από τους Narasimhamurthy και Kucheva (2007) οι οποίοι παρουσιάζουν ένα γενικό πλαίσιο για την παραγωγή δεδομένων προσομοιώνοντας περιβάλλοντα αλλαγής. Το πλαίσιο τους παρουσιάζει την στρατηγική παραγωγής των συνθετικών εξελισσόμενων δεδομένων STAGGER και Moving Hyperplane. Θεωρούν ένα σύνολο κ πηγών δεδομένων με γνωστές κατανομές. Καθώς αυτές οι κατανομές στις πηγές τους είναι σταθερές η κατανομή δεδομένων στον χρόνο t , $D^{\{t\}}$ καθορίζεται μέσω της $v_i(t)$, όπου η $v_i(t) \in [0,1]$ καθορίζει το μέγεθος της επιρροής της πηγής i το χρόνο t :

$$D^{\{t\}} = \{v_1(t), v_2(t), \ldots, v_k(t)\} \text{ , όπου } \sum_i v_i(t) = 1$$

Το πλαίσιο τους καλύπτει τόσο τα σταδιακά όσο και τα απότομα concept drifts. Η προσέγγιση όμως των Bifet και Kirkby (2009) είναι πιο συμπαγής. Ξεκινάνε αντιμετωπίζοντας ένα απλό σενάριο: ένα ρεύμα δεδομένων και δύο διαφορετικά concepts. Έπειτα μελετάται και η γενικότερη περίπτωση που περιλαμβάνει περισσότερα από ένα concept drift γεγονότα.

Θεωρώντας τα ρεύματα δεδομένων ως δεδομένα που παράγονται από καθαρές κατανομές οι συγγραφείς μοντελοποιούν ένα concept drift γεγονός ως ένα συνδυασμό με βάρη δύο καθαρών κατανομών που χαρακτηρίζουν τα στοχοποιημένα concepts πριν και μετά το drift. Στο πλαίσιο που παρουσιάζουν καθορίζουν την πιθανότητα κάθε στιγμιότυπο του ρεύματος να ανήκει στο νέο concept με την αλλαγή. Χρησιμοποιείται η σιγμοειδής συνάρτηση ως μια εκλεπτυσμένη αλλά και πρακτική λύση. Η σιγμοειδής συνάρτηση με τύπο $f(t) = \frac{1}{\left(1 + e^{-s(t-t_0)}\right)}$,

όπως φαίνεται και από το γράφημα της εικόνας 3.4 έχει παράγωγο στο σημείο t_0 ίση με $f'(t_0) = \frac{s}{4}$. Η εφαπτομένη υπό γωνία α έχει κλίση ίση με την παράγωγο αυτή άρα ισχύει ότι $\varepsilon\varphi a = \frac{s}{4}$. Επιπλέον παρατηρούμε ότι ισχύει $\varepsilon\varphi a = \frac{1}{W}$ και ότι επειδή s = 4• εφα θα είναι και $s = \frac{4}{W}$. Επομένως η παράμετρος s στην σιγμοειδή συνάρτηση δίνει το μήκος του W καθώς και την γωνία α. Σε αυτό το σιγμοειδή μοντέλο χρειάζεται να καθορίσουμε μόνο δύο παραμέτρους, την χρονική στιγμή t_0 που είναι το σημείο που ξεκινάει η αλλαγή καθώς και το W που εκφράζει το συνολικό διάστημα της αλλαγής.



Εικόνα 3.4: Μια σιγμοειδής συνάρτηση (Bifet & Kirkby 2009)

Παρατηρούμε ότι: $f(t_0 + \beta \cdot W) = 1 - f(t_0 - \beta \cdot W)$ και ότι οι $f(t_0 + \beta \cdot W)$, $f(t_0 - \beta \cdot W)$ είναι σταθερές τιμές που δεν εξαρτώνται από τις παραμέτρους t_0 και W .

Επομένως με βάση τους παραπάνω τύπους θα έχουμε ότι :

$$f\left(t_0 + \frac{W}{2}\right) = 1 - f(t_0 - \frac{w}{2}) = \frac{1}{1 + e^{-2}} = 88.08\ \%$$

$$f\left(t_0 + W\right) = 1 - f(t_0 - W) = \frac{1}{1 + e^{-4}} = 98.20\ \%$$

$$f\left(t_0 + 2W\right) = 1 - f(t_0 - 2W) = \frac{1}{1 + e^{-8}} = 99.97\ \%$$

Επομένως σύμφωνα με τα παραπάνω αποτελέσματα όσο μεγαλύτερη είναι η διάρκεια της αλλαγής τόσο υψηλότερη είναι η ακρίβεια.

Ορισμός (Bifet & Kirkby 2009)

Δοθέντων δύο ρευμάτων με δεδομένα a και b ορίζεται ως $c = a \oplus_{t_0}^{W} b$ το ρεύμα δεδομένων που κατασκευάζεται από την συνένωση των ρευμάτων a και β όπου t_0 είναι το σημείο αλλαγής και W είναι το μήκος της αλλαγής. Επιπλέον :

$$Pr\left[c(t) = a(t)\right] = \frac{e^{\frac{-4(t - t_0)}{W}}}{1 + e^{\frac{-4(t - t_0)}{W}}} \text{ και } Pr\left[c(t) = b(t)\right] = \frac{1}{1 + e^{\frac{-4(t - t_0)}{W}}}$$

όπου η πρώτη σχέση είναι η πιθανότητα το ρεύμα c να είναι ίσο με το ρεύμα a και η δεύτερη σχέση είναι η πιθανότητα το ρεύμα c να είναι ίσο με το ρεύμα b.

Επίσης αν a ≠ b ισχύουν οι ακόλουθες αλγεβρικές ιδιότητες:

- $a \oplus_{t_0}^{W} b \neq b \oplus_{t_0}^{W} a$
- $a \oplus_{t_0}^{W} a = a$
- $a \oplus_{0}^{0} b = b$
- $a \oplus_{t_0}^{W} \left(b \oplus_{t_0}^{W} c\right) \neq \left(a \oplus_{t_0}^{W} b\right) \oplus_{t_0}^{W} c$

- $a \oplus_{t_0}^{W} \left(b \oplus_{t_1}^{W} c \right) \approx \left(a \oplus_{t_0}^{W} b \right) \oplus_{t_1}^{W} c$, αν $t_0 < t_1$ και $W << |t_1 - t_0|$

Η δημιουργία ενός ρεύματος δεδομένων με πολλαπλές αλλαγές μπορεί να πραγματοποιηθεί σύμφωνα με τα παραπάνω ενώνοντας διαφορετικά ρεύματα δεδομένων με διαφορετικά concept drifts.

3.2 Επιλογή Δεδομένων ροής (Data Streams)

Για τους σκοπούς της έρευνας των αλγορίθμων κατηγοριοποίησης σε δεδομένα ροής υπάρχει έλλειψη κατάλληλων και δημόσια διαθέσιμων συνόλων δεδομένων που αφορούν προβλήματα του πραγματικού κόσμου. Το UCI (Asuncion & Newman 2007) καθώς και το KDD (Hettich & Bay 1999) αποτελούν αποθήκες αρχείων όπου διαθέτουν τα πιο κοινά δεδομένα ορόσημα για αλγορίθμους μηχανικής μάθησης. Εντούτοις πολλά από αυτά τα σύνολα δεδομένων δεν είναι κατάλληλα για την αξιολόγηση των αλγορίθμων κατηγοριοποίησης σε δεδομένα ροής. Το αρχείο KDD έχει αρκετά μεγάλα σύνολα δεδομένων αλλά δεν διαθέτει προβλήματα κατηγοριοποίησης με επαρκή παραδείγματα. Το σύνολο δεδομένων Forest Covertype είναι ένα από τα μεγαλύτερα και έχει λιγότερα από 600.000 παραδείγματα επομένως σύμφωνα με όσα αναπτύχθηκαν ανωτέρω γίνεται κατανοητό ότι τα δεδομένα αυτά δεν επαρκούν.

Για να παρουσιάσουν τα συστήματα τους αρκετοί ερευνητές χρησιμοποίησαν ιδιωτικά πραγματικά δεδομένα τα οποία δεν μπορούν να αναπαραχθούν από άλλους. Παράδειγμα αποτελεί το ίχνος δικτύου από το πανεπιστήμιο της Washington το οποίο χρησιμοποιήθηκε από τους Domingos και Hulten για να αξιολογήσουν το σύστημα VFDT (Domingos & Hulten 2000) καθώς και το σύνολο δεδομένων που αφορά απάτες με πιστωτική κάρτα που χρησιμοποιήθηκε από τους Wang et al.(2003) και τους Chu και Zaniolo (2004) .

Κατά κανόνα οι ερευνητές δημοσιεύουν αποτελέσματα τα οποία είναι βασισμένα πάνω σε συνθετικά παραγόμενα δεδομένα. Σε πολλές από αυτές τις περιπτώσεις οι συγγραφείς επινοούν πρωτότυπα μοτίβα παραγωγής συνθετικών δεδομένων με σκοπό να αξιολογήσουν τους αλγόριθμους που έχουν

κατασκευάσει. Παράδειγμα αποτελεί η γεννήτρια παραγωγής τυχαίων δέντρων η οποία χρησιμοποιήθηκε και αυτή από τους Domingos και Hulten (2000) για την αξιολόγηση του VFDT και οι ειδικής κατασκευής γεννήτριες που περιγράφονται από τους Oza και Russel (2001a), Street και Kim (2001) και Chu και Zaniolo (2004). Τα συνθετικά δεδομένα έχουν αρκετά πλεονεκτήματα, είναι ευκολότερα να αναπαραχθούν και υπάρχει μικρό κόστος όσον αφορά την αποθήκευση και την μετάδοση τους. Παρόλα αυτά τα πλεονεκτήματα υπάρχει έλλειψη εγκαθιδρυμένων και ευρέως χρησιμοποιούμενων συνθετικών δεδομένων ροής.

Για την πειραματική αξιολόγηση των μεθόδων μείωσης σε δεδομένα ροής που αναλύθηκαν στο προηγούμενο κεφάλαιο θα χρησιμοποιηθούν τέσσερις γεννήτριες δεδομένων ροής που είναι ενσωματωμένες στο λογισμικό MOA. Ο λόγος που επιλέγονται τα συγκεκριμένα σύνολα δεδομένων είναι η δυνατότητα που μας δίνεται από το λογισμικό να εισάγουμε εμείς το τεχνητό concept drift σε οποιοδήποτε σημείο του πειράματος επιθυμούμε. Επιπλέον επιλέγουμε και την διάρκεια που επιθυμούμε να έχει το drift. Με βάση αυτό το σκεπτικό θα εξεταστούν ως προς την απόδοση οι συγκεκριμένες μέθοδοι στα ίδια σύνολα δεδομένων όταν αυτά παρουσιάζουν drift σε ένα συγκεκριμένο σημείο και όταν αυτά διατηρούν την ίδια κατανομή (δηλαδή όταν δεν ενεργοποιούμε το τεχνητό drift). Οι τέσσερις γεννήτριες που επιλέχθηκαν παρουσιάζονται αναλυτικά παρακάτω.

Γεννήτρια STAGGER

Εισήχθησαν από τους Schlimmer και Granger (1986). Στο STAGGER μια περιγραφή του concept είναι ένα σύνολο από αριθμημένα με βάρη γνωρίσματα. Αυτό το σύνολο περιέχει τρία κατηγορικά γνωρίσματα τα οποία είναι το μέγεθος, το χρώμα και το σχήμα όπου το πρώτο παίρνει από τρεις τιμές μικρό μεσαίο και μεγάλο ενώ τα άλλα δύο γνωρίσματα παίρνουν τις τιμές κόκκινο και πράσινο και κυκλικό, μη-κυκλικό αντίστοιχα. Πριν από το πρώτο σημείο αλλαγής τα στιγμιότυπα παίρνουν την ετικέτα θετικό αν το χρώμα είναι κόκκινο και το μέγεθος είναι μικρό. Μετά από αυτό το σημείο και πριν από την δεύτερη αλλαγή

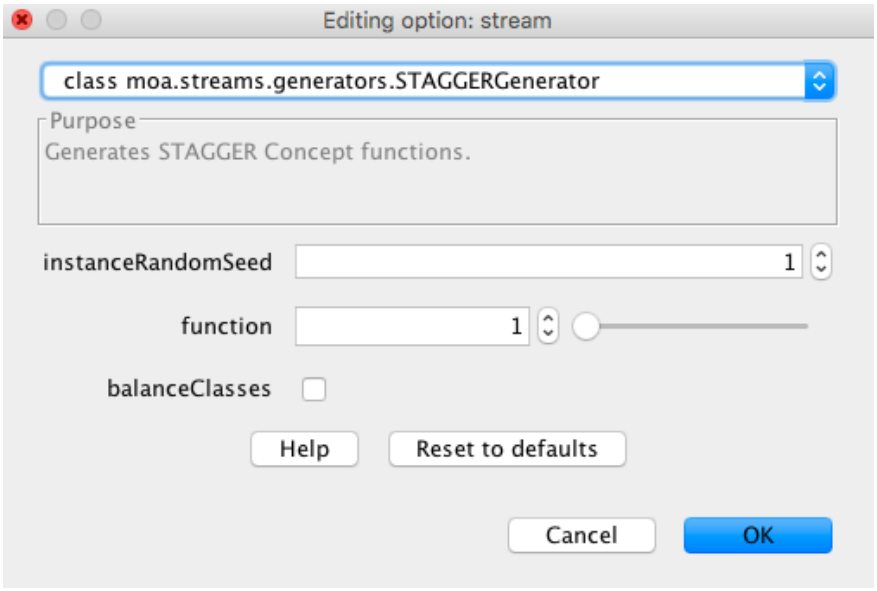
τα στιγμιότυπα κατηγοριοποιούνται ως θετικά αν το χρώμα είναι πράσινο ή το σχήμα κυκλικό και τελικά μετά το δεύτερο drift τα στιγμιότυπα κατηγοριοποιούνται θετικά μόνο αν το μέγεθος είναι μεσαίο ή μεγάλο. Επομένως με βάση τα παραπάνω τα concepts των STAGGER είναι οι ακόλουθες συναρτήσεις κατηγοριοποίησης :

1: Συνάρτηση η οποία επιστρέφει το 1 αν το μέγεθος είναι μικρό και το χρώμα κόκκινο.

2: Συνάρτηση η οποία επιστρέφει το 1 αν το χρώμα είναι πράσινο και το μέγεθος είναι κύκλος.

3: Συνάρτηση η οποία επιστρέφει το 1 αν το μέγεθος είναι μεσαίο η μεγάλο.

Ένα επιπλέον γνώρισμα είναι η πιθανότητα που δίνεται να εξισορροπηθούν οι κλάσεις το οποίο σημαίνει ότι η κατανομή των κλάσεων θα τείνει προς την ομοιόμορφη κατανομή.



Εικόνα 3.5 : Παράθυρο με τις παραμέτρους που αφορούν τα STAGGER data

Γεννήτρια SEA

Αυτό το σύνολο δεδομένων περιέχει απότομο concept drift και παρουσιάστηκε αρχικά από τους Street και KIM (2001). Τα δεδομένα παράγονται χρησιμοποιώντας τρία γνωρίσματα εκ των οποίων μόνο τα δύο πρώτα είναι σχετικά μεταξύ τους. Όλα τα γνωρίσματα λαμβάνουν τιμές από 0 έως 10 . Τα σημεία του συνόλου δεδομένων χωρίζονται σε 4 τμήματα με διαφορετικές αλλαγές το καθένα. Σε κάθε τμήμα η κατηγοριοποίηση πραγματοποιείται με την χρήση της σχέσης $f_1 + f_2 \leq \theta$, όπου f_1 και f_2 αντιπροσωπεύουν τα δύο πρώτα γνωρίσματα ενώ το θ είναι ένα κατώφλι. Οι πιο συχνές τιμές είναι : 9 , 8, 7 και 9.5 για τα δεδομένα των τμημάτων. Σύμφωνα με τον ορισμό στη σελίδα 148 η γεννήτρια SEA σύμφωνα με τους Bifet και Kirkby (2009) μπορεί να οριστεί ως εξής :

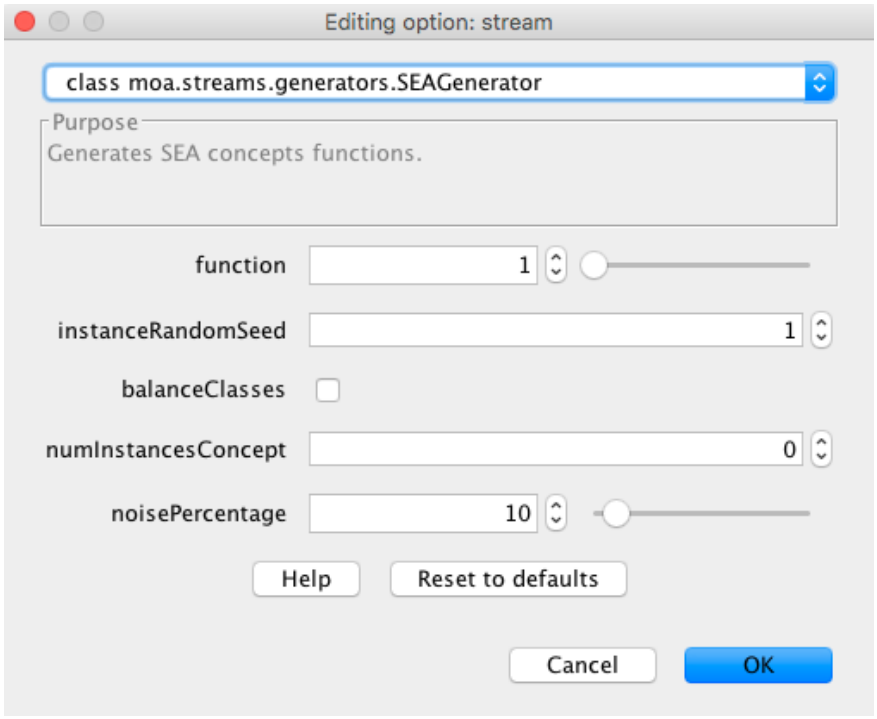
$$\left(\left(\left(SEA_9 \oplus_{t_0}^{W} SEA_8 \right) \oplus_{2t_0}^{W} SEA_7 \right) \oplus_{3t_0}^{W} SEA_{9.5} \right)$$

Συνεπώς με βάση τα παραπάνω η γεννήτρια θα κατηγοριοποιήσει ένα στιγμιότυπο με βάση τις δύο δυνατές ετικέτες. Πιο αναλυτικά οι συναρτήσεις θα είναι οι εξής:

- Συνάρτηση 1 : αν $f_1 + f_2 \leq 8$ διαφορετικά 1
- Συνάρτηση 2 : αν $f_1 + f_2 \leq 9$ διαφορετικά 1
- Συνάρτηση 3 : αν $f_1 + f_2 \leq 7$ διαφορετικά 1
- Συνάρτηση 4 : αν $f_1 + f_2 \leq 9.5$ διαφορετικά 1

Το concept drift μπορεί να εισαχθεί αλλάζοντας την συνάρτηση κατηγοριοποίησης. Το συγκεκριμένο ρεύμα δεδομένων έχει δυο επιπλέον παραμέτρους που πρέπει να ληφθούν υπόψη. Η πρώτη είναι η δυνατότητα ισορροπίας των κλάσεων έτσι ώστε η κατανομή των κλάσεων να τείνει προς την ομοιόμορφη και η δεύτερη είναι η δυνατότητα εισαγωγής θορύβου στο ρεύμα η

οποία με βάση κάποια πιθανότητα αλλάζει την επιλεγμένη ετικέτα για ένα στιγμιότυπο.

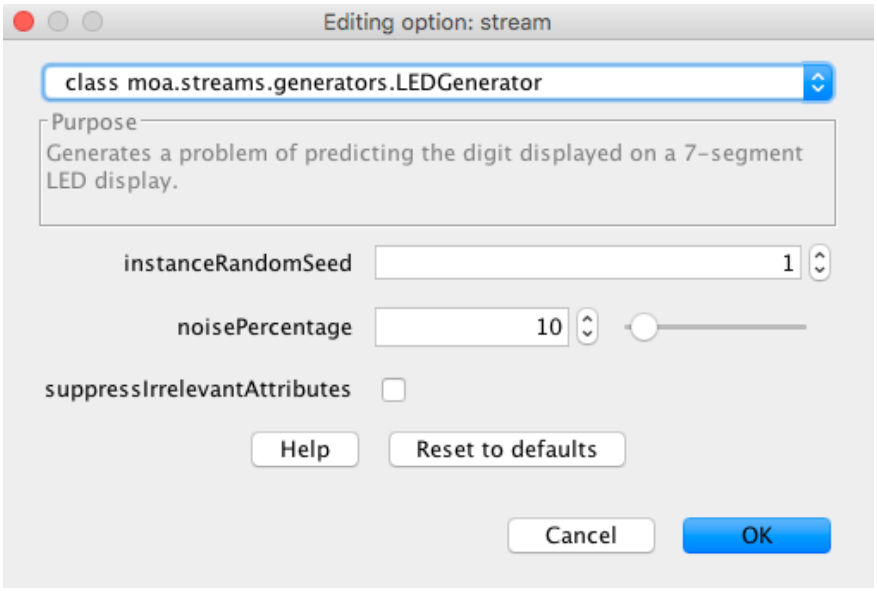


Εικόνα 3.6 : Παράθυρο με τις παραμέτρους που αφορούν τα SEA data

Γεννήτρια LED

Αυτή η πηγή δεδομένων προέρχεται από το βιβλίο Classification and Regression Trees των Breiman et al. (1984). Ο στόχος είναι να προβλεφθεί το ψηφίο που εμφανίζεται σε μια διάταξη LED εφτά τμημάτων , όπου κάθε γνώρισμα έχει 10% πιθανότητα να αναστραφεί. Η πηγή δεδομένων αυτή έχει βέλτιστη Bayes κατηγοριοποίηση της τάξης του 74%. Η συγκεκριμένη ρύθμιση της γεννήτριας που χρησιμοποιείται για πειράματα παράγει 24 δυαδικά γνωρίσματα 17 από τα οποία είναι άσχετα.

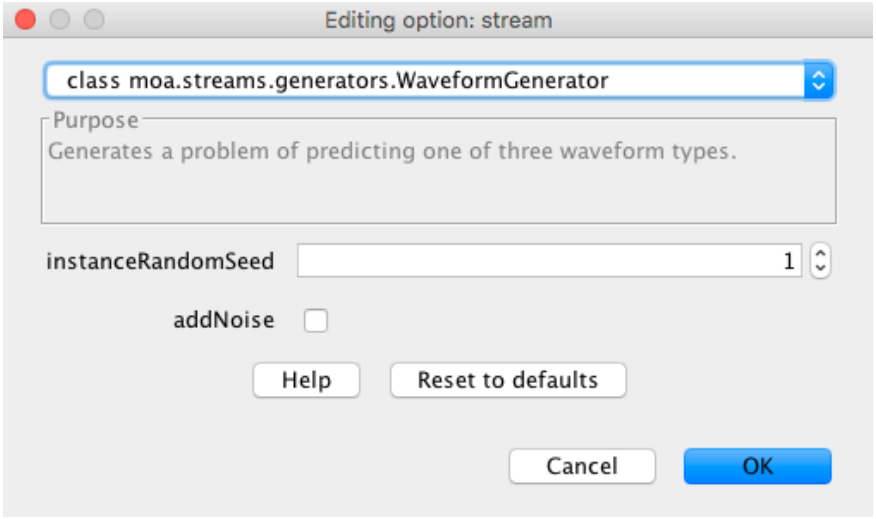
Οι παράμετροι που περιλαμβάνει η συγκεκριμένη γεννήτρια είναι τρεις και αφορούν τα εξής : Τον σπόρο (seed) για την τυχαία παραγωγή στιγμιοτύπων (κάτι που περιλαμβάνουν όλες οι γεννήτριες), την δυνατότητα εισαγωγής θορύβου (default υπάρχει 10% θόρυβος) και την δυνατότητα να μειωθούν τα δεδομένα ώστε να περιέχουν μόνο 7 σχετικά δυαδικά γνωρίσματα.



Εικόνα 3.7 : Παράθυρο με τις παραμέτρους που αφορούν τα LED data

Γεννήτρια WAVEFORM

Αυτή η γεννήτρια έχει ίδια προέλευση με την LED και δόθηκε από τον David Aha στο UCI αρχείο. Ο σκοπός της εργασίας εδώ είναι να διαφοροποιήσει τα δεδομένα μεταξύ τριών διαφορετικών κλάσεων του waveform κάθε μια από τις οποίες παράγεται από έναν συνδυασμό δύο ή και τριών βασικών κυμάτων. Η βέλτιστη Bayes κατηγοριοποίηση είναι γνωστό ότι έχει ποσοστό 86%. Υπάρχουν δυο εκδοχές του προβλήματος. Ο WAVE21 έχει 21 αριθμητικά γνωρίσματα τα οποία εμπεριέχουν θόρυβο. Ο WAVE40 εισάγει επιπλέον 19 άσχετα μεταξύ τους γνωρίσματα. Οι παράμετροι επομένως που υπάρχουν για αυτή την γεννήτρια είναι η εισαγωγή σπόρου για τυχαία παραγωγή στιγμιοτύπων και η δυνατότητα εισαγωγής θορύβου που αν ενεργοποιηθεί θα προστεθούν 19 επιπλέον άσχετα μεταξύ τους γνωρίσματα.



Εικόνα 3.8 : Παράθυρο με τις παραμέτρους που αφορούν τα Waveform data

3.3 Μέγεθος δεδομένων

Μια από τις μεγαλύτερες προκλήσεις στη συγκεκριμένη εργασία αποτέλεσε η εύρεση του κατάλληλου μεγέθους που πρέπει να έχουν τα δεδομένα ροής. Δεδομένου ότι η μελέτη πραγματοποιείται στο πεδίο των δεδομένων ροής θα ανέμενε κανείς οι αλγόριθμοι αυτοί να είναι σε θέση να διαχειριστούν εκατομμύρια δεδομένα μέσα σε σύντομο χρονικό διάστημα και με μικρό κόστος στην μνήμη έτσι ώστε να ικανοποιούνται οι απαιτήσεις που πρέπει να πληρούν οι αλγόριθμοι των δεδομένων ροής όπως αυτές αναφέρθηκαν στην σελίδα 131. Δυστυχώς κάτι τέτοιο δεν υλοποιείται με τις συγκεκριμένες μεθόδους. Στην έρευνα τους για την προεπεξεργασία δεδομένων στην εξόρυξη γνώσης με δεδομένα ροής οι Gallego et al. (2017) μελετώντας τις μεθόδους επιλογής στιγμιοτύπων σημειώνουν ότι αυτές είναι ακατάλληλες για μεσαίου μεγέθους σύνολα δεδομένων και για τον λόγο αυτό απορρίπτουν προβλήματα που εμπεριέχουν στιγμιότυπα άνω των εκατό χιλιάδων και επιλέγουν μόνο τεχνητά σύνολα που αριθμούν έως και δέκα χιλιάδες στιγμιότυπα. Για την επιλογή τους αυτή βασίζονται και σε προγενέστερες έρευνες και συγκεκριμένα στην έρευνα των Lu et al. (2016) η οποία παρουσιάζει τον αλγόριθμο μείωσης NEFCS-SRR και τον εξετάζει πειραματικά σε ένα πλήθος συνόλων. Συγκεκριμένα τον εξετάζουν σε πραγματικά σύνολα δεδομένων με μεγαλύτερο το Wine Quality το

οποίο αριθμεί 4998 στιγμιότυπα αλλά και σε συνθετικά δεδομένα όπως τα SEA και το περιστρεφόμενο υπερεπίπεδο στα οποία όμως το πλήθος των στιγμιοτύπων φτάνει τις δέκα χιλιάδες. Στο τελευταίο πείραμα τους εξετάζουν τον αλγόριθμο NEFCS-SRR σε φίλτρα ανεπιθύμητης αλληλογραφίας. Με βάση όλα αυτά καταλήγουν στο συμπεράσματα ότι όλες οι μέθοδοι που είναι βασισμένες στην ικανότητα (competence-based) απαιτούν πολύ χρόνο για να διατηρήσουν το μοντέλο ικανότητας. Συνεπώς κάτι τέτοιο καθιστά ένα competence-based μοντέλο ακατάλληλο για εφαρμογές που χρειάζονται ταχύτατους χρόνους εκτέλεσης για πολύ μεγάλο όγκο δεδομένων που καταφθάνει κάθε δευτερόλεπτο. Επισημαίνουν ότι οι μέθοδοι αυτές είναι κατάλληλες μόνο για μικρά σύνολα ή για φιλτράρισμα ανεπιθύμητης αλληλογραφίας όπου εκεί το κόστος σε χρόνο δεν είναι τόσο σημαντικό.

Σε αυτή την εργασία δοκιμάστηκαν αρχικά πειράματα με εκατό χιλιάδες στιγμιότυπα και επιβεβαιώθηκαν τα ευρήματα των ανωτέρω ερευνών. Οι αλγόριθμοι χρειαζόντουσαν τεράστιο χρονικό περιθώριο για να ολοκληρώσουν γεγονός που τους καθιστούσε ακατάλληλους για εξόρυξη γνώσης σε δεδομένα ροής. Για τον λόγο αυτό καταφύγαμε σε σύνολα δεδομένων που αριθμούν δέκα χιλιάδες στιγμιότυπα έτσι ώστε να μπορέσουμε να εξετάσουμε τους αλγόριθμους σε εύλογο χρονικό διάστημα. Ακόμα και σε αυτές τις περιπτώσεις όμως παρουσιάστηκαν προβλήματα που θα αναλυθούν παρακάτω. Επιπλέον τα ίδια πειράματα εκτελέστηκαν και με την παρουσία concept drift για να εξετάσουμε το πως συμπεριφέρονται οι μέθοδοι μείωσης σε αυτή την περίπτωση. Παρακάτω δίνεται ένας πίνακας που κάνει εμφανές το χρονικό πρόβλημα που παρουσιάζουν οι μέθοδοι αυτοί σε σύγκριση με τον απλό kNN.

Πίνακας 3.1: Χρόνοι μεθόδων στα STAGGER για 10000 instances

μέθοδος	Χρόνος (cpu sec)	Κόστος μνήμης (ram-Hours)
FISH	1863	0,00111
NEFCS-SRR	4589	0,05875
ICF	2248	0,00085
CBE	1279	0,00035
kNN	2	0,00000

3.4 Τεχνικά Χαρακτηριστικά Η/Υ - Λογισμικού

Προτού προχωρήσουμε στην παρουσίαση των αποτελεσμάτων των αλγορίθμων του MOA, θα αναφερθούμε στα τεχνικά χαρακτηριστικά του συστήματος (Η/Υ, έκδοση λογισμικού) το οποίο χρησιμοποιήθηκε για την εκτέλεση των πειραμάτων. Τα τεχνικά χαρακτηριστικά φαίνονται συνοπτικά στον παρακάτω πίνακα :

Πίνακας 3.2 : Τεχνικά χαρακτηριστικά συστήματος εκτέλεσης αλγορίθμων MOA

Μνήμη (RAM)	18 GB 800 MHz DDR2 FB-DIMM
Επεξεργαστής	2 x 2,8 GHz Quad-Core Intel Xeon
Τύπος συστήματος	Quad-Core Intel Xeon (64 bit)
Λειτουργικό σύστημα	macOS High Sierra version 10.13.6
έκδοση του MOA	MOA Release 2018.06

3.5 Εκτέλεση πειραμάτων για μέγεθος 10000 στιγμιοτύπων

3.5.1 Στόχοι

Οι στόχοι των πειραμάτων που θα πραγματοποιηθούν στα τέσσερα δεδομένα ροής που αναλύθηκαν ανωτέρω χρησιμοποιώντας τις τέσσερις μεθόδους μείωσης επιλογής στιγμιοτύπων είναι να διαπιστωθεί το πως επηρεάζουν την απόδοση της κατηγοριοποίησης. Για τον σκοπό αυτό θα συγκριθούν τα αποτελέσματα των μεθόδων στην περίπτωση που δεν υπάρχει κάποιο drift. Έπειτα θα εισαχθεί τεχνητό drift και τα αποτελέσματα θα συγκριθούν εκ νέου. Θα εξεταστεί η ακρίβεια καθώς και το κόστος σε χρόνο και σε μνήμη που απαιτούν οι συγκεκριμένοι αλγόριθμοι μείωσης δεδομένων. Έπειτα θα εξεταστεί το μέγεθος της μείωσης που επιτυγχάνει ο κάθε αλγόριθμος για τα 4 αυτά σύνολα δεδομένων.

3.5.2 Εκτέλεση πειραμάτων χωρίς την παρουσία concept drift

3.5.2.1 Δεδομένα STAGGER

Ξεκινάμε εξετάζοντας αρχικά το πως θα εφαρμοστούν οι τέσσερις μέθοδοι για τα STAGGER δεδομένα. Το κρίσιμο σημείο για την εφαρμογή των αλγορίθμων είναι η μεταβλητή p. Η μεταβλητή αυτή καθορίζει το μέγεθος του περιβάλλοντος δηλαδή το κάθε πότε θα ενεργοποιείται η μέθοδος μείωσης. Η default τιμή που δίνουν οι δημιουργοί είναι p = 500 η οποία όμως δεν αποδίδει ικανοποιητικά τόσο ως προς την ακρίβεια όσο και ως προς τον χρόνο. Επομένως για κάθε αλγόριθμο και σε κάθε σύνολο δεδομένων πρέπει να γίνει έρευνα για το ποια είναι η κατάλληλη τιμή της παραμέτρου p η οποία πρέπει να τεθεί ούτως ώστε να λάβουμε την καλύτερη δυνατή απόδοση στην κάθε περίπτωση. Οι άλλες παράμετροι οι οποίες είναι κοινές για όλα τα πειράματα είναι οι ακόλουθες :

i : μεταβλητή που εκφράζει το πλήθος των στιγμιοτύπων που έχουμε επιλέξει για το πείραμα.

c: μεταβλητή που εκφράζει το πλήθος των στιγμιοτύπων σε κάθε κομμάτι (chunk)

f : μεταβλητή που εκφράζει πόσα στιγμιότυπα υπάρχουν μεταξύ των δειγμάτων της διαδικασίας αξιολόγησης .

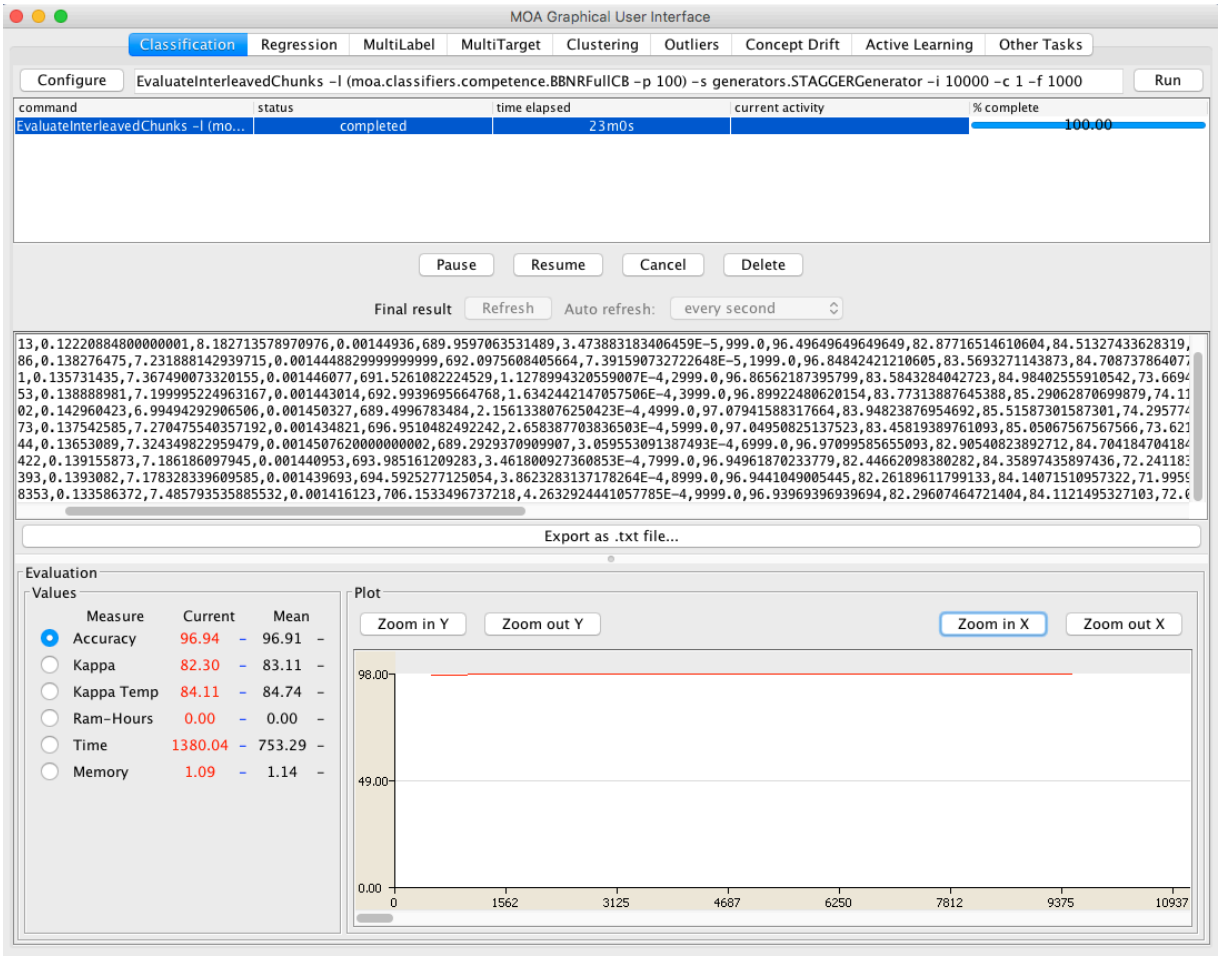
t: χρονικό όριο για εκπαίδευση/έλεγχο του μοντέλου (-1 : δεν υπάρχει όριο)

b: όριο στο μέγεθος που καταλαμβάνει το μοντέλο (-1: δεν υπάρχει όριο)

s : η γεννήτρια παραγωγής δεδομένων

l : ο αλγόριθμος που εφαρμόζεται

Οι περισσότερες από τις ανωτέρω μεταβλητές φαίνονται στην εικόνα 3.9 που ακολουθεί. Δίπλα από το πλαίσιο configure υπάρχει η γραμμή εντολών (η οποία μπορεί να εκτελεστεί και από την κονσόλα) και η οποία δείχνει να εκτελείται ένα πείραμα στα δεδομένα STAGGER για την μέθοδο CBE (BBNRFullCB). Όποιες από τις παραπάνω μεταβλητές δεν αναγράφονται στην γραμμή εντολών είναι διότι έχουν την τιμή default.



Εικόνα 3.9 : Πείραμα στα δεδομένα STAGGER για την μέθοδο CBE

Στον πίνακα 3.3 που ακολουθεί παρουσιάζονται τα αποτελέσματα ως προς την ακρίβεια και τον χρόνο (sec) από τα πειράματα που εκτελέστηκαν για τις τέσσερις μεθόδους μείωσης στα δεδομένα STAGGER χρησιμοποιώντας διαφορετικές τιμές για το p. Όπως προαναφέρθηκε το p υποδηλώνει το κάθε πότε ενεργοποιείται η μέθοδος μείωσης. Για παράδειγμα για την προκαθορισμένη τιμή των 500 στιγμιοτύπων σε σύνολο δέκα χιλιάδων

στιγμιοτύπων που θα εκτελεστεί το κάθε πείραμα η μέθοδος μείωσης θα ενεργοποιηθεί 19 φορές. Τα κενά σημεία του πίνακα υποδηλώνουν ότι τα πειράματα εκεί δεν εκτελέστηκαν διότι θα απαιτούσαν πολύ ώρα και υψηλό κόστος σε μνήμη για να ολοκληρωθούν.

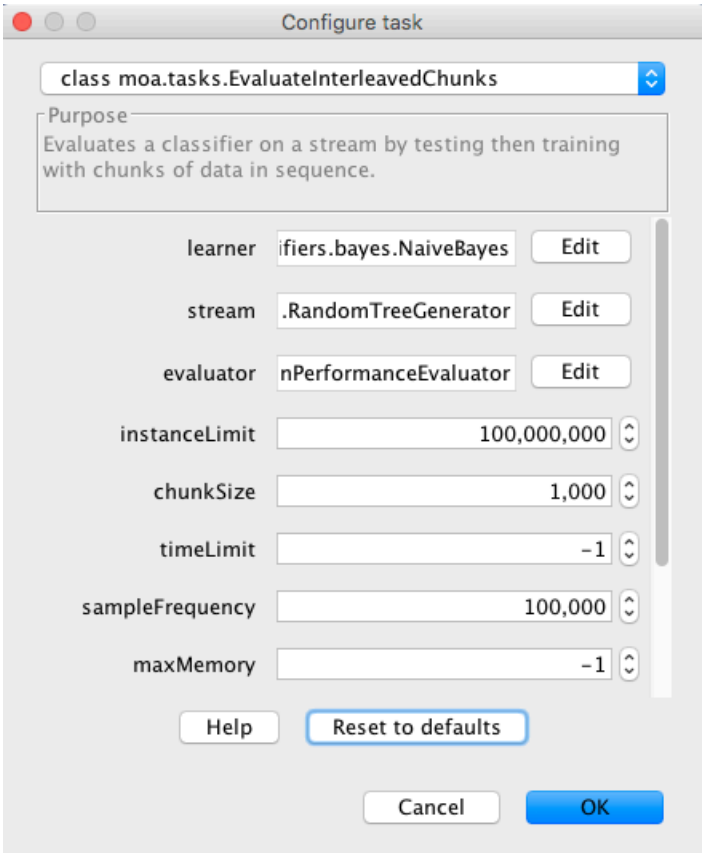
Πίνακας 3.3: Έρευνα για την τιμή του p στα δεδομένα STAGGER

	CBE		**NEFCS-SRR (s=100)**		**FISH**		**ICF**	
	accuracy	Time	accuracy	Time	accuracy	Time	accuracy	Time
p = 50	92,29	407	**99,91**	**409**	83,50	16354	-	
p = 100	**96,94**	**1279**	99,91	436	82,31	11215	-	
p = 500	99,29	29662	99,97	4589	81,81	1863	89,39	32929
p=1000	-		99,97	1576	82,46	750	89,69	13627
p=2000	-		99,96	2550	75,53	465	90,52	4600
p=5000	-		99,96	3301	**94,69**	**112**	**92,76**	**2248**

Σύμφωνα με τον ανωτέρω πίνακα η default τιμή (p=500) δεν αποδίδει τόσο καλά. Συγκεκριμένα παρουσιάζει πολύ μεγάλο κόστος σε χρόνο στην περίπτωση που εφαρμόζεται για τις μεθόδους CBE και ICF. Επιπλέον παρατηρούμε ότι οι μέθοδοι CBE και NEFCS-SRR είναι πιο γρήγοροι για μικρά p χωρίς να ρίχνουν αισθητά την ακρίβεια. Αντίθετα οι ICF και FISH είναι πιο γρήγοροι αλλά έχουν και υψηλότερη ακρίβεια για μεγάλα p και ειδικά για p = 5000 δηλαδή για την περίπτωση που η μέθοδος μείωσης ενεργοποιείται μόνο μια φορά για τα πρώτα 5000 στιγμιότυπα.

Παρόλα αυτά για να μπορέσουμε να έχουμε μια σωστή σύγκριση μεταξύ των αλγορίθμων θα εξετάσουμε την ακρίβεια τους καθώς και το κόστος σε χρόνο και μνήμη για την τιμή p=500, δηλαδή για την περίπτωση που κάθε μέθοδος μείωσης ενεργοποιήθηκε 19 φορές μέσα στο σύνολο των δέκα χιλιάδων στιγμιοτύπων των δεδομένων STAGGER. Τέλος αναφέρουμε ότι τα παραπάνω πειράματα εκτελέστηκαν για c=1, δηλαδή τα στιγμιότυπα κατέφθαναν ένα ένα. Αυτό ενδεχομένως έχει κόστος στην αρχική ακρίβεια του μοντέλου διότι το μοντέλο βλέπει αρχικά λίγα στιγμιότυπα. Παρόλα αυτά παρατηρείται ότι δεν βελτιώνει αρκετά την ακρίβεια και σε κάποιες περιπτώσεις οδηγεί σε πολύ πιο χρονοβόρα πειράματα.

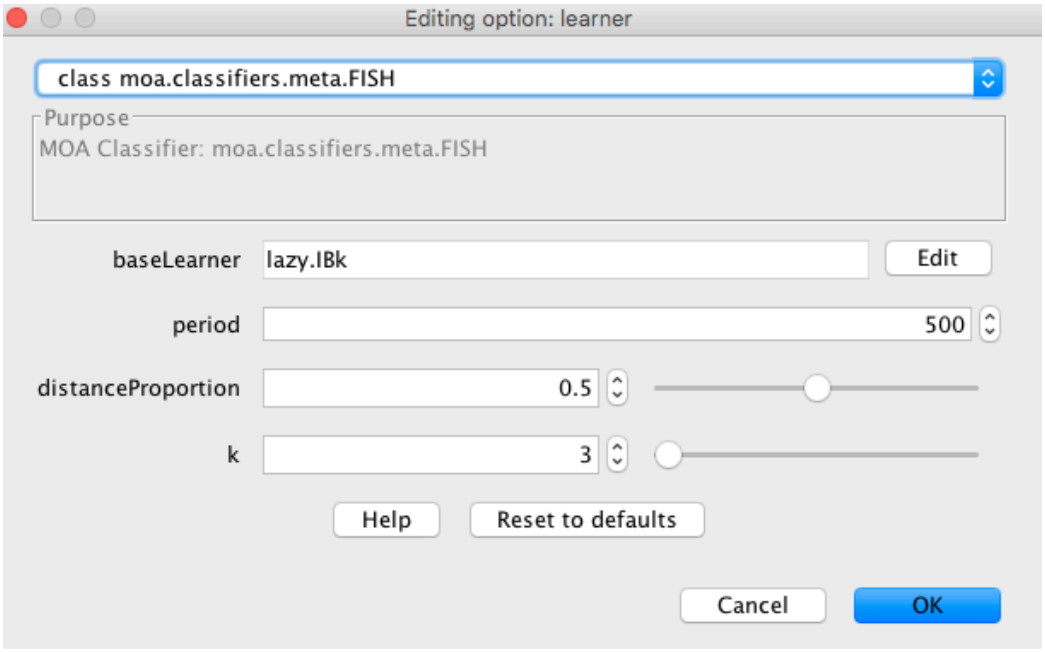
Παρακάτω παρουσιάζεται με την βοήθεια εικόνων ο τρόπος που εκτελείται ένα πείραμα με το λογισμικό MOA. Επιλέγοντας το πλαίσιο configure που φαίνεται στην εικόνα 3.9 καταλήγουμε στο παράθυρο που φαίνεται στην εικόνα 3.10 το οποίο περιέχει τις default τιμές που υπάρχουν από τους δημιουργούς του MOA.



Εικόνα 3.10 : Παράθυρο ρύθμισης του MOA

Για την εκτέλεση πειράματος στα δεδομένα STAGGER για την μέθοδο μείωσης FISH επιλέγουμε στο πλαίσιο του stream τα δεδομένα STAGGER. Στο πλαίσιο του learner επιλέγοντας τον αλγόριθμο FISH εμφανίζεται το παράθυρο της εικόνας 3.11. Σε αυτό το παράθυρο επιλέγουμε την προκαθορισμένη τιμή των 500 στιγμιοτύπων στο πλαίσιο δίπλα από την μεταβλητή period. Επιπλέον διατηρούμε την αναζήτηση του kNN να γίνεται με 3 εγγύτερους γείτονες και την αναλογία μεταξύ των μεταβλητών χρόνου και χώρου να είναι ένα προς ένα. Όσον αφορά το παράθυρο της εικόνας 3.10 ρυθμίζουμε το chunk size να είναι 1 καθώς και το instanceLimit να είναι 10,000. Επιθυμούμε να μας δίνει αναφορά

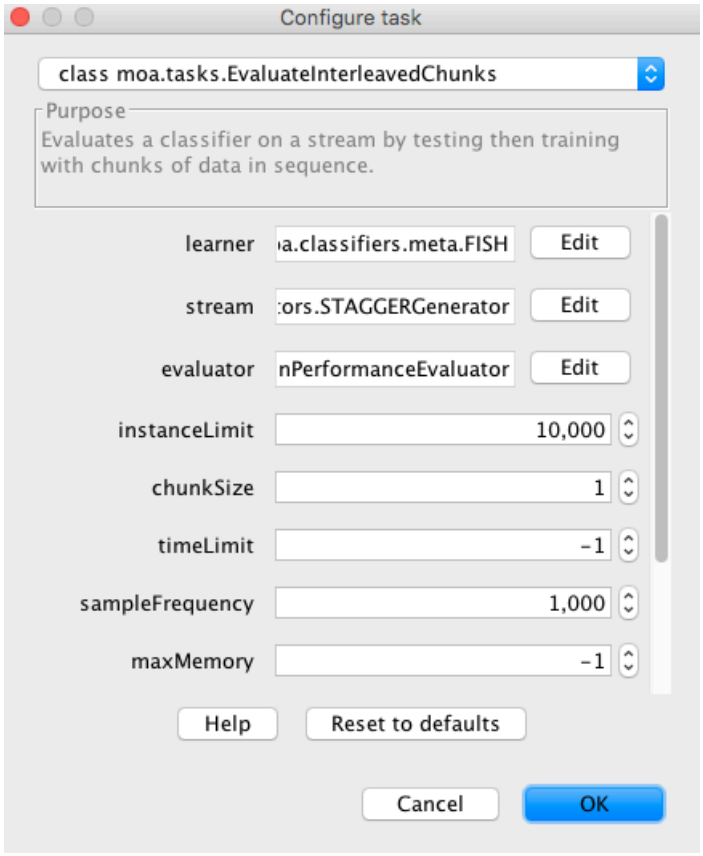
το λογισμικό ανά 1000 στιγμιότυπα για την απόδοση του learner επομένως επιλέγουμε sampleFrequency ίσο με 1000. Τέλος δεν θέτουμε περιορισμό ως προς τον χρόνο ή την μνήμη επομένως τα ορίσματα αυτά παίρνουν την τιμή -1.



Εικόνα 3.11 : Παράθυρο ρύθμισης των παραμέτρων του FISH

Με βάση τα παραπάνω καταλήγουμε στο τελικό παράθυρο της εικόνας 3.12. Πατώντας OK έχουνε εισαχθεί τα κατάλληλα ορίσματα για να εκτελεστεί το πείραμα που αφορά την μέθοδο FISH στα δεδομένα STAGGER. Έπειτα πατώντας την εντολή Run που βρίσκεται δίπλα από την γραμμή εντολών (εικόνα 3.9) το πείραμα τρέχει και μπορούμε να καταγράψουμε και να εξάγουμε τα αποτελέσματα του σε ένα αρχείο txt. Φυσικά τα πειράματα μπορούν να εκτελεστούν και χωρίς την χρήση της γραφικής διεπαφής μέσω της κονσόλας. Για να συμβεί αυτό πρέπει να αλλάξουμε directory στην τοποθεσία που είναι το αρχείο MOA jar και να εκτελέσουμε την παρακάτω εντολή:

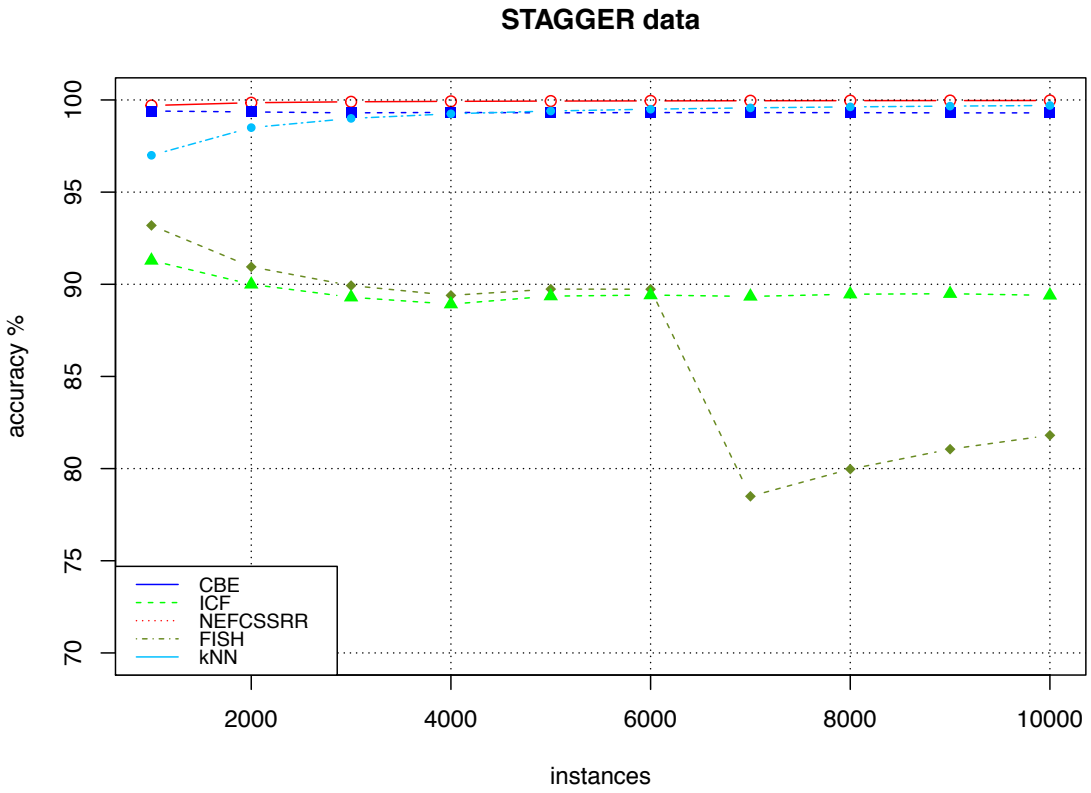
```
java -cp moa.jar moa.DoTask EvaluateInterleavedChunks -l
moa.classifiers.meta.FISH -s generators.STAGGERGenerator -i
10000 -c 1 -f 1000
```



Εικόνα 3.12 : Παράθυρο με ρυθμισμένες παραμέτρους

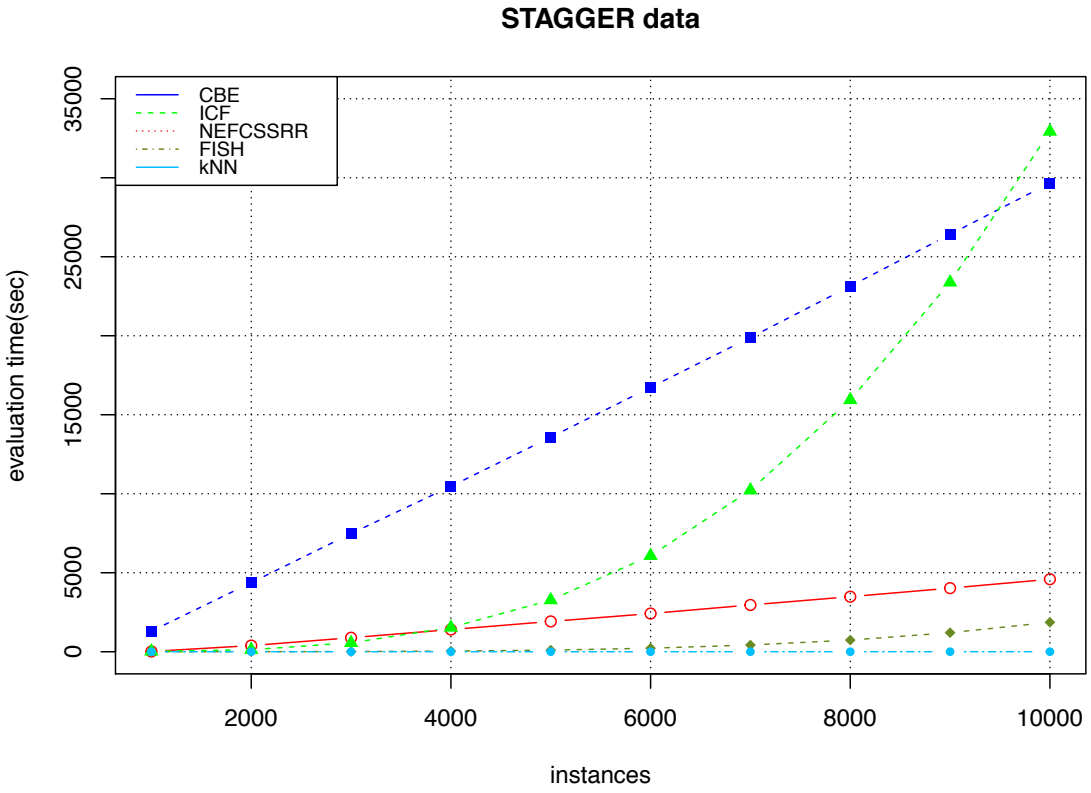
Με όμοιο τρόπο εκτελούνται και τα αντίστοιχα πειράματα για τις άλλες τρεις μεθόδους μείωσης πάνω στα δεδομένα STAGGER. Επιλέγουμε σε κάθε μέθοδο το p να ισούται με 500 και στην μέθοδο NEFCS-SRR επιλέγουμε το case base να έχει μέγεθος έως 100. Κάθε φορά που το σύνολο αυτό υπερβαίνει το όριο των 100 ενεργοποιείται ο αλγόριθμος SRR μείωσης πλεονάσματος που αναλύθηκε στη σελίδα 88.

Με βάση τις παραπάνω διαδικασίες καταλήγουμε στα ακόλουθα γραφήματα που μας δίνουν μια εικόνα για την απόδοση των αλγορίθμων μείωσης πάνω στα δεδομένα STAGGER αλλά και για τις απαιτήσεις τους σε χρόνο και μνήμη. Τα γραφήματα έγιναν με χρήση της γλώσσας προγραμματισμού R αντλώντας τα δεδομένα από τα αρχεία txt που μπορούμε να εξάγουμε από το λογισμικό MOA.



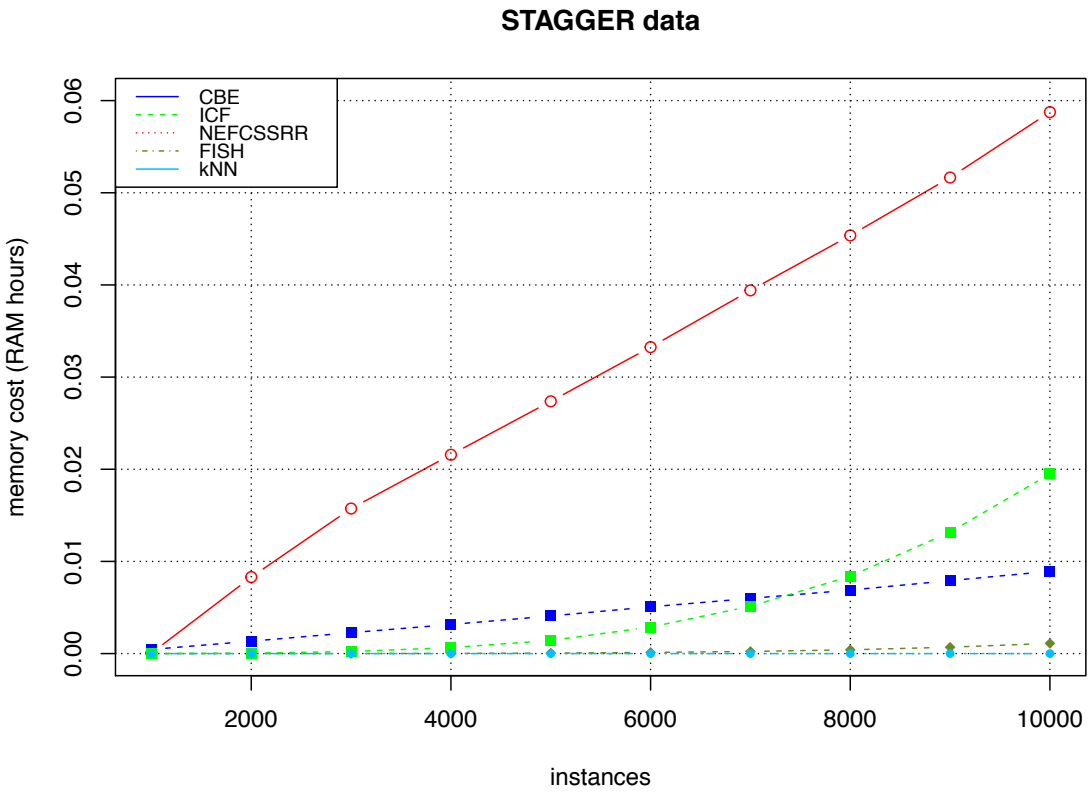
Εικόνα 3.13 : Ακρίβεια των μεθόδων στα δεδομένα STAGGER

Στο γράφημα της εικόνας 3.13 παρατηρούμε ότι οι μέθοδοι μείωσης επιτυγχάνουν υψηλή ακρίβεια στα δεδομένα STAGGER. Συγκεκριμένα οι μέθοδοι NEFCS-SRR και CBE επιτυγχάνουν την ίδια υψηλή ακρίβεια με αυτή που επιτυγχάνει ο βασικός αλγόριθμος kNN. Αντίθετα οι μέθοδοι ICF και FISH παρουσιάζουν χαμηλότερη ακρίβεια σε σχέση με τις άλλες δύο μεθόδους με την μέθοδο FISH μετά τα 6000 στιγμιότυπα να ρίχνει την ακρίβεια κάτω από 80% αν και δείχνει να παρουσιάζει ανάκαμψη από τα 7000 στιγμιότυπα και μετά. Μια τόσο απότομη πτώση στην ακρίβεια είναι συνήθως ένδειξη παρουσίας concept drift αν και στην συγκεκριμένη περίπτωση δεν έχουμε εισάγει κάποιο drift.



Εικόνα 3.14 : Χρόνος αξιολόγησης των μεθόδων στα δεδομένα STAGGER

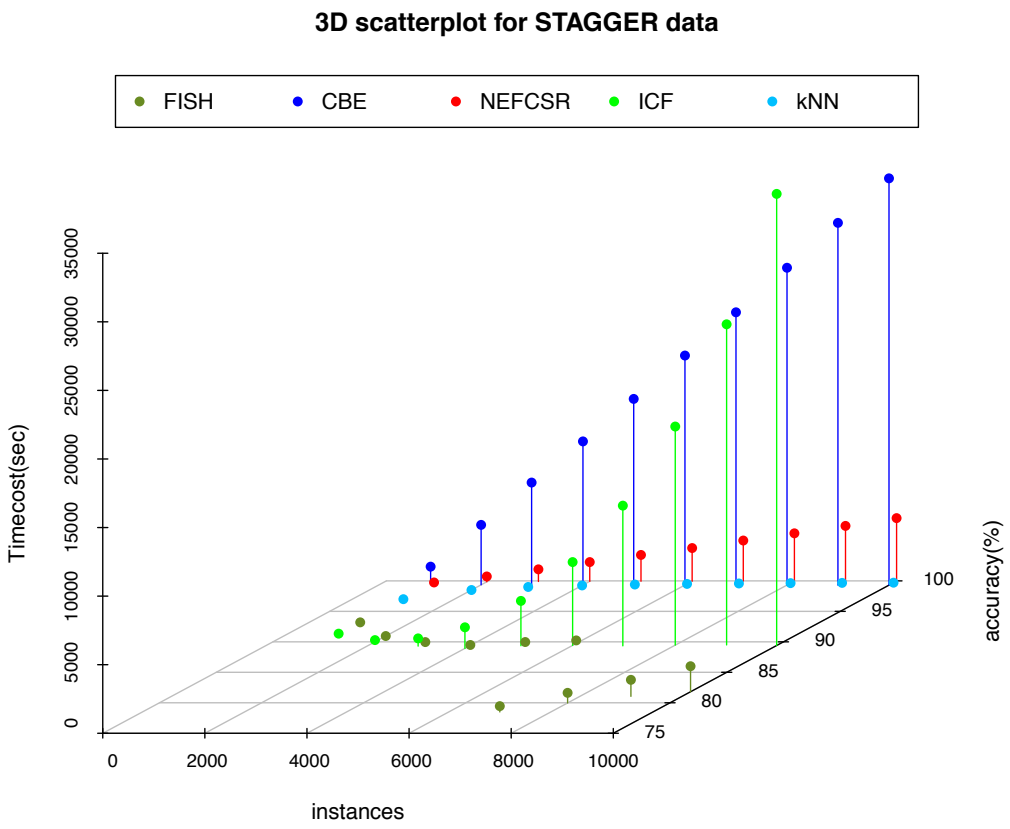
Στο γράφημα της εικόνας 3.14 γίνεται αντιληπτό ότι οι αλγόριθμοι CBE και ICF απαιτούν πολύ χρόνο για να ολοκληρώσουν τα πειράματα. Συγκεκριμένα ο αλγόριθμος CBE παρουσιάζει γραμμική αύξηση για κάθε 1000 στιγμιότυπα που εξετάζει ενώ ο ICF παρουσιάζει εκθετική αύξηση και ειδικά από τα 5000 στιγμιότυπα και μετά αυξάνει με μεγάλο ρυθμό την χρονική διάρκεια που χρειάζεται για να ολοκληρώσει. Αντίθετα οι αλγόριθμοι NEFCS-SRR και FISH χρειάζονται πολύ λιγότερο χρόνο για να ολοκληρώσουν με τον FISH να απαιτεί περίπου 2000 δευτερόλεπτα. Την καλύτερη επίδοση την επιτυγχάνει ο βασικός αλγόριθμος kNN ο οποίος χρειάζεται μόλις 2 δευτερόλεπτα για να κατηγοριοποιήσει τα δέκα χιλιάδες στιγμιότυπα.



Εικόνα 3.15 : Κόστος μνήμης των μεθόδων στα δεδομένα STAGGER

Όσον αφορά την χρήση της μνήμης αυτή υπολογίζεται σε RAM-Hours όπου 1 RAM-Hour ισούται με 1 GB της RAM το οποίο μοιράζεται ανά ώρα επεξεργασίας. Η μονάδα RAM-Hours χρησιμοποιείται ως μέτρηση αξιολόγησης των πόρων που χρειάζονται οι αλγόριθμοι δεδομένων ροής. Σύμφωνα με το γράφημα της εικόνας 3.15 υψηλές απαιτήσεις σε μνήμη έχει η μέθοδος NEFCS-SRR ενώ οι υπόλοιπες δεν απαιτούν πάνω από 0,02 RAM-Hours για να ολοκληρώσουν.

Στο γράφημα της εικόνας 3.16 που ακολουθεί παρουσιάζεται σε τρισδιάστατο σύστημα συντεταγμένων η πορεία της ακρίβειας των τεσσάρων μεθόδων (άξονας z) καθώς και το κόστος σε χρόνο (άξονας y) καθώς οι μέθοδοι αυτοί εξετάζουν τα στιγμιότυπα (άξονας x).



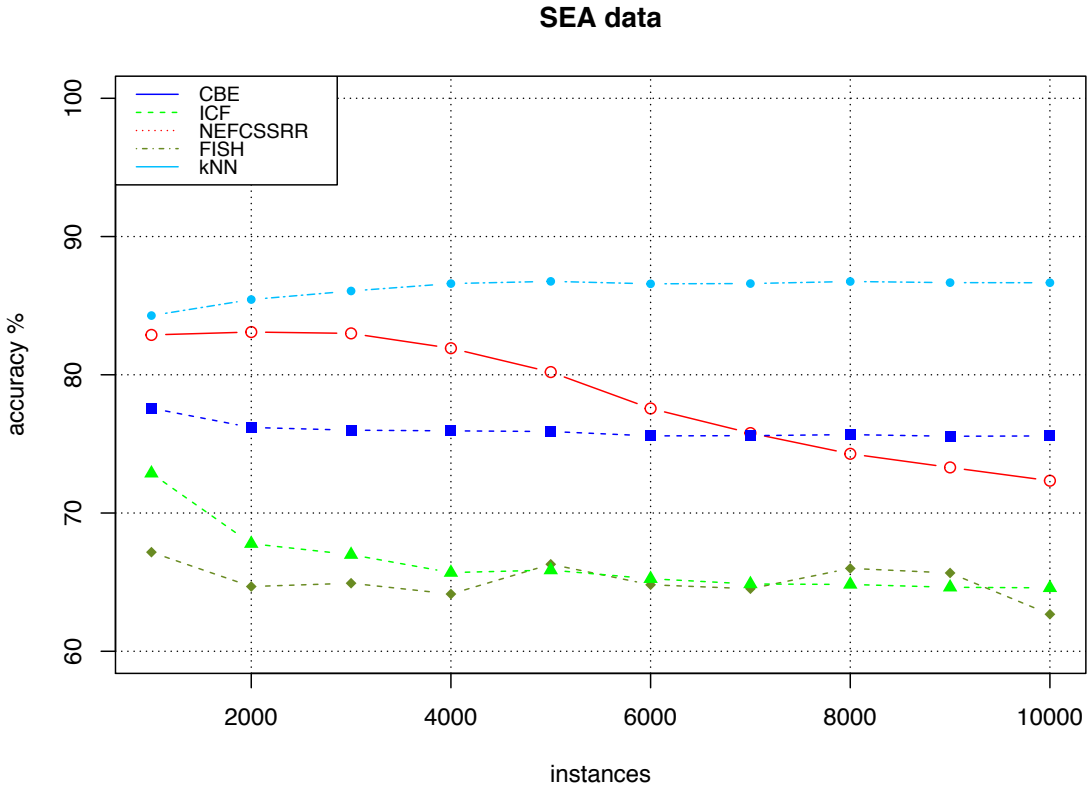
Εικόνα 3.16 : Διάγραμμα διασποράς στα δεδομένα STAGGER

3.5.2.2 Δεδομένα SEA

Στον πίνακα 3.4 εξετάζουμε την ακρίβεια που επιτυγχάνουν οι αλγόριθμοι στα δεδομένα SEA καθώς και τον χρόνο που απαιτείται για να ολοκληρωθούν τα πειράματα για τις διάφορες τιμές του p. Παρατηρούμε ότι στο συγκεκριμένο σύνολο δεδομένων η καλύτερη ακρίβεια επιτυγχάνεται για p = 5000 όπου οι αλγόριθμοι μείωσης εκτελούνται μόνο μια φορά στα πρώτα 5000 στιγμιότυπα. Επίσης παρατηρούμε όπως και στα STAGGER δεδομένα ότι ο αλγόριθμος CBE χάνει σε ακρίβεια για μικρά p αλλά είναι πολύ πιο γρήγορος εκεί, ενώ οι υπόλοιποι αλγόριθμοι επιτυγχάνουν την πιο γρήγορη απόδοση για p = 5000. Όμοια με τα δεδομένα STAGGER θα επιλέξουμε και εδώ να συγκρίνουμε τους αλγόριθμους για p = 500. Εξετάζουμε αρχικά την ακρίβεια των μεθόδων με την βοήθεια της γραφήματος της εικόνας 3.17.

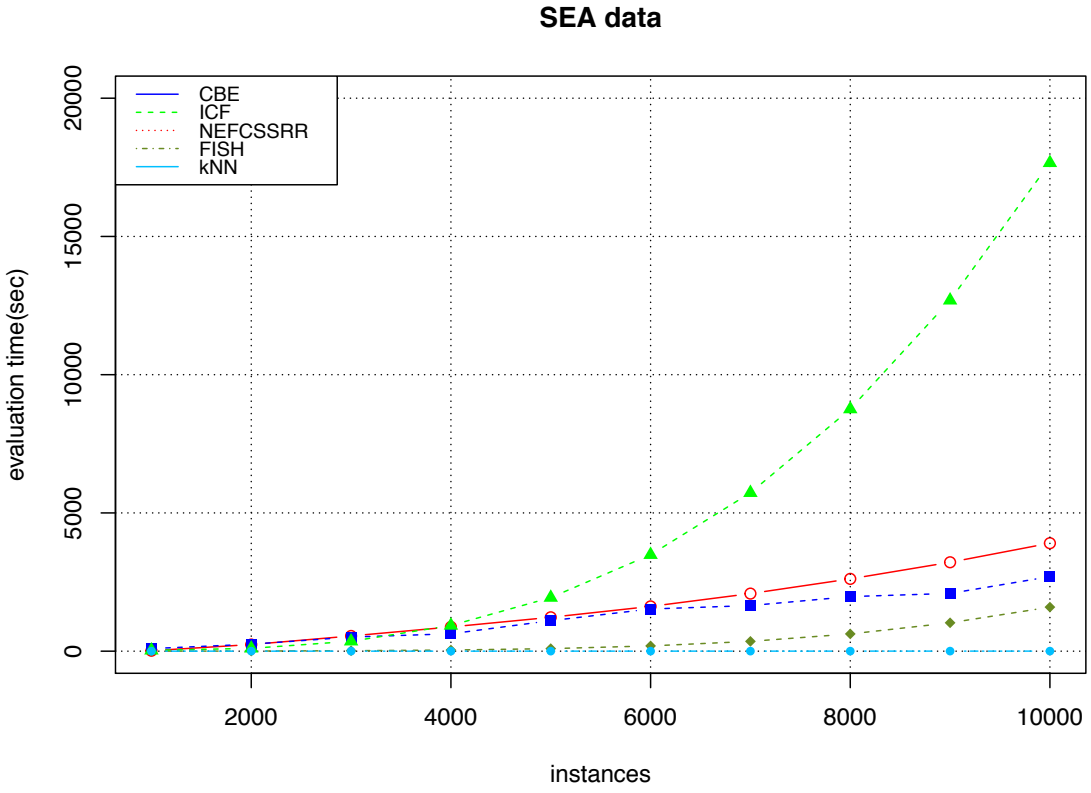
Πίνακας 3.4: Έρευνα για την τιμή του p στα δεδομένα SEA

	CBE		NEFCS-SRR (s=100)		FISH		ICF	
	accuracy	Time	accuracy	Time	accuracy	Time	accuracy	Time
p = 50	69,47	347	64,29	4930	59,43	19598	-	
p = 100	71,50	858	65,26	3459	60,62	10555	-	
p=500	75,58	2696	72,34	3905	62,68	1593	64,58	17661
p=1000	77,06	5581	70,55	3077	65,47	997	65,60	7998
p=2000	78,76	5234	76,24	3371	66,88	300	67,81	3203
p=5000	**79,88**	11297	**80,91**	2282	**80,95**	156	**74,28**	2918



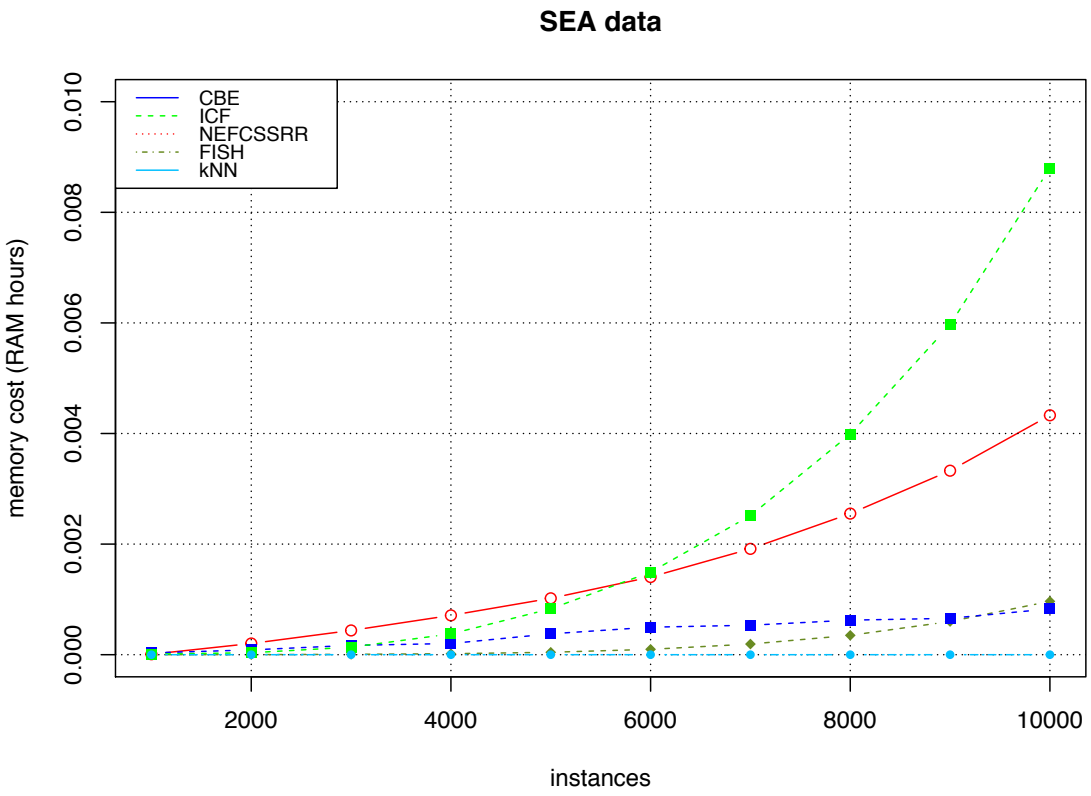
Εικόνα 3.17 : Διάγραμμα ακρίβειας στα δεδομένα SEA

Παρατηρούμε ότι ο αλγόριθμος kNN παρουσιάζει την υψηλότερη ακρίβεια. Την δεύτερη καλύτερη ακρίβεια επιτυγχάνει ο CBE ο οποίος σταθεροποιεί την απόδοση του περίπου στο 75% για όλη την διάρκεια του πειράματος. Στις άλλες τρεις μεθόδους η ακρίβεια μειώνεται σταδιακά με την μεγαλύτερη πτώση να παρουσιάζεται στην μέθοδο NEFCS-SRR.



Εικόνα 3.18 : Χρόνος αξιολόγησης των μεθόδων στα δεδομένα SEA

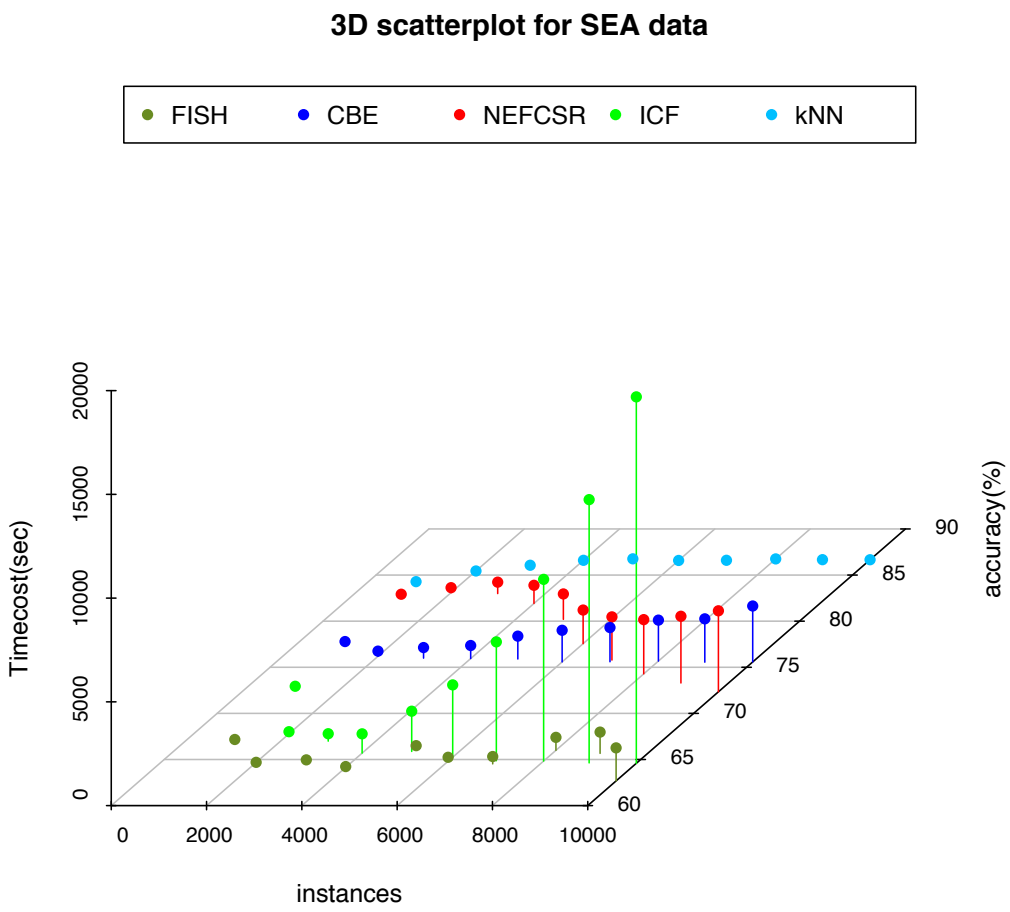
Όσον αφορά στην χρονική διάρκεια που χρειάζονται οι αλγόριθμοι για να ολοκληρώσουν παρατηρείται εκθετική αύξηση του χρόνου από την μέθοδο ICF η οποία αποδεικνύεται ιδιαίτερα χρονοβόρα και σε αυτή την περίπτωση. Οι υπόλοιποι αλγόριθμοι δεν απαιτούν πάνω από 5000 δευτερόλεπτα για να ολοκληρώσουν με τον ταχύτερο να είναι φυσικά ο kNN και να ανταγωνίζεται μόνο από τον FISH ο οποίος χρειάζεται περίπου 1500 δευτερόλεπτα.



Εικόνα 3.19 : Κόστος μνήμης των μεθόδων στα δεδομένα SEA

Σύμφωνα με την εικόνα 3.19 ο αλγόριθμος ICF έχει μεγάλο κόστος και στην μνήμη παρουσιάζοντας και εδώ εκθετική αύξηση. Δεύτερος σε κόστος μνήμης ανέρχεται ο NEFCS-SRR ενώ οι υπόλοιπες μέθοδοι διατηρούν χαμηλό κόστος.

Στην εικόνα 3.20 παρουσιάζεται το διάγραμμα διασποράς για τα δεδομένα SEA. Μέσω του διαγράμματος έχουμε μια τρισδιάστατη εικόνα για την απόδοση των αλγορίθμων μείωσης σε συνάρτηση με τις δύο βασικές μετρικές που εξετάζουμε αυτή της ακρίβειας και του χρόνου αξιολόγησης.



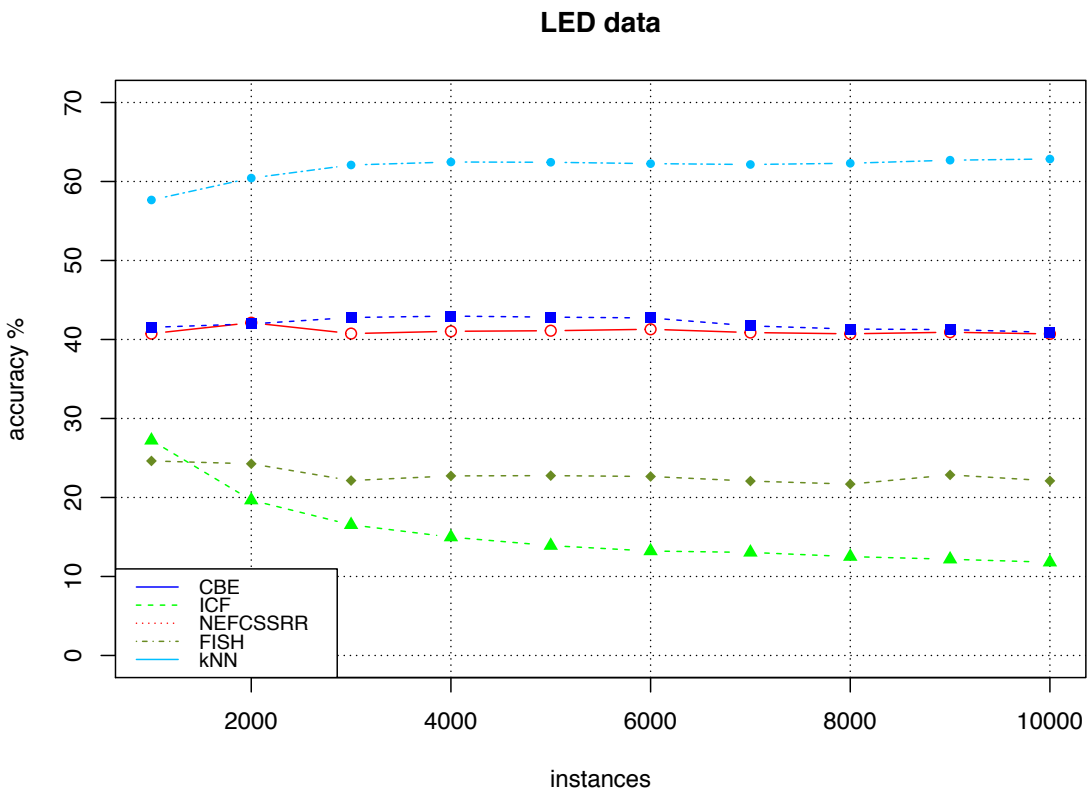
Εικόνα 3.20 : Διάγραμμα διασποράς στα δεδομένα SEA

3.5.2.3 Δεδομένα LED

Στον πίνακα 3.5 που ακολουθεί παρουσιάζονται τα αποτελέσματα των πειραμάτων που εκτελέστηκαν πάνω στα δεδομένα LED για τις διάφορες τιμές του p. Όπως και στην περίπτωση των SEA έτσι και εδώ η καλύτερη ακρίβεια παρατηρείται για p=5000 για όλες τις μεθόδους. Όμως στην περίπτωση του ICF και του CBE οι χρόνοι δεν είναι οι βέλτιστοι για αυτή την τιμή του p. Και εδώ όπως και στις προηγούμενες περιπτώσεις παρουσιάζουμε τα γραφήματα για p = 500.

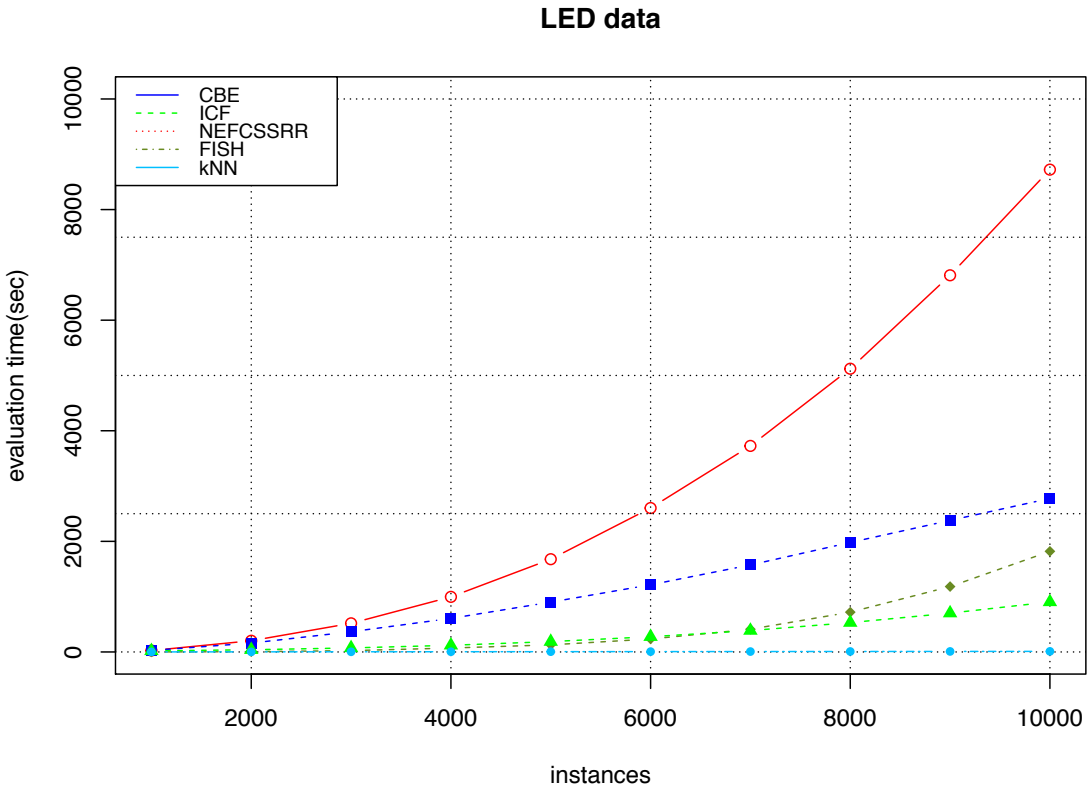
Πίνακας 3.5: Έρευνα για την τιμή του p στα δεδομένα LED

	CBE		NEFCS-SRR (s=100)		FISH		ICF	
	accuracy	Time	accuracy	Time	accuracy	Time	accuracy	Time
p = 50	34,93	601	39,63	40245	20,96	17623	-	
p = 100	36,24	1033	38,92	23031	20,15	13457	10,31	2670
p=500	40,90	2776	40,70	8722	22,10	1820	11,80	904
p=1000	43,56	3317	43,00	6678	24,74	811	13,45	746
p=2000	46,39	3188	45,08	5258	26,10	313	17,45	874
p=5000	**49,14**	2792	**48,86**	2579	**32,46**	157	**30,25**	2739



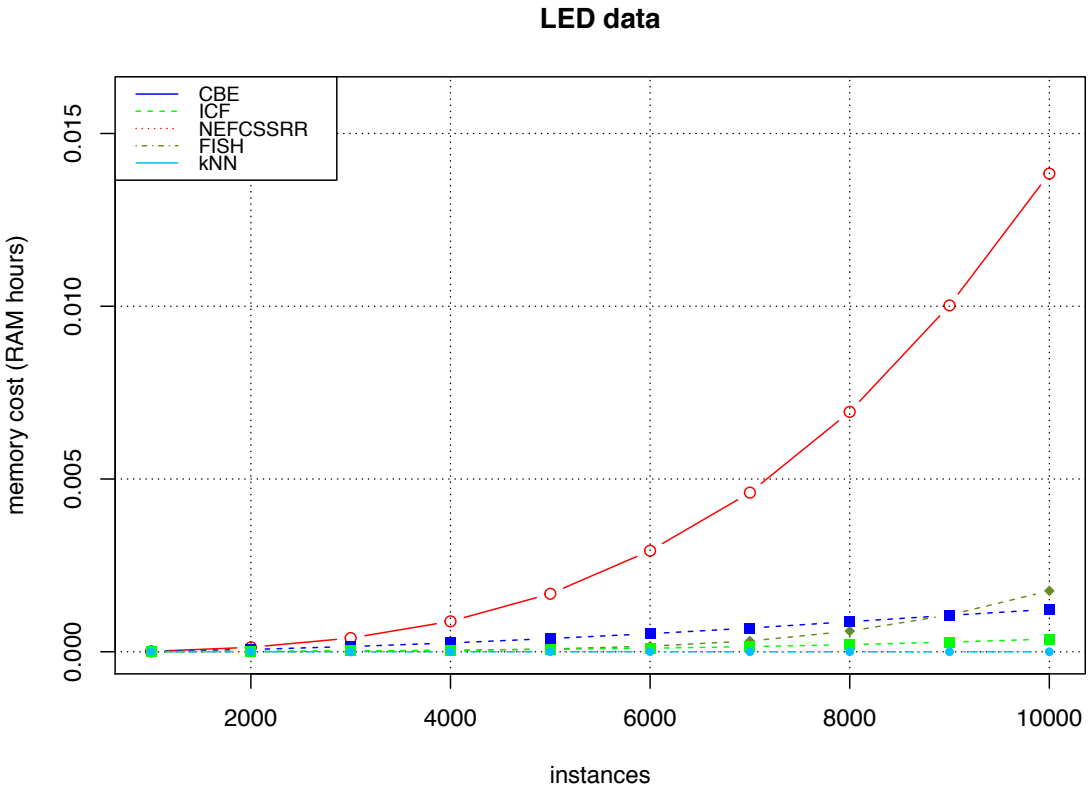
Εικόνα 3.21 : Διάγραμμα ακρίβειας στα δεδομένα LED

Παρατηρούμε από το γράφημα της εικόνας 3.21 ότι η ακρίβεια στα δεδομένα LED για τις μεθόδους μείωσης αλλά και για τον kNN διατηρείται σταθερή με εξαίρεση τον αλγόριθμο ICF που παρουσιάζει αισθητή μείωση της ακρίβειας και φθάνει στο πολύ χαμηλό ποσοστό 10%. Την καλύτερη επίδοση την σημειώνει ο kNN με 63%. Στην περίπτωση των δεδομένων ροής LED η ακρίβεια των μεθόδων έχει μειωθεί αισθητά σε σχέση με τα προηγούμενα δεδομένα. Πρέπει να σημειωθεί ότι υπάρχει προκαθορισμένη προσθήκη θορύβου 10% που συμβάλλει στα χαμηλά ποσοστά ακρίβειας που παρουσιάζουν οι αλγόριθμοι.



Εικόνα 3.22 : Χρόνος αξιολόγησης μεθόδων στα δεδομένα LED

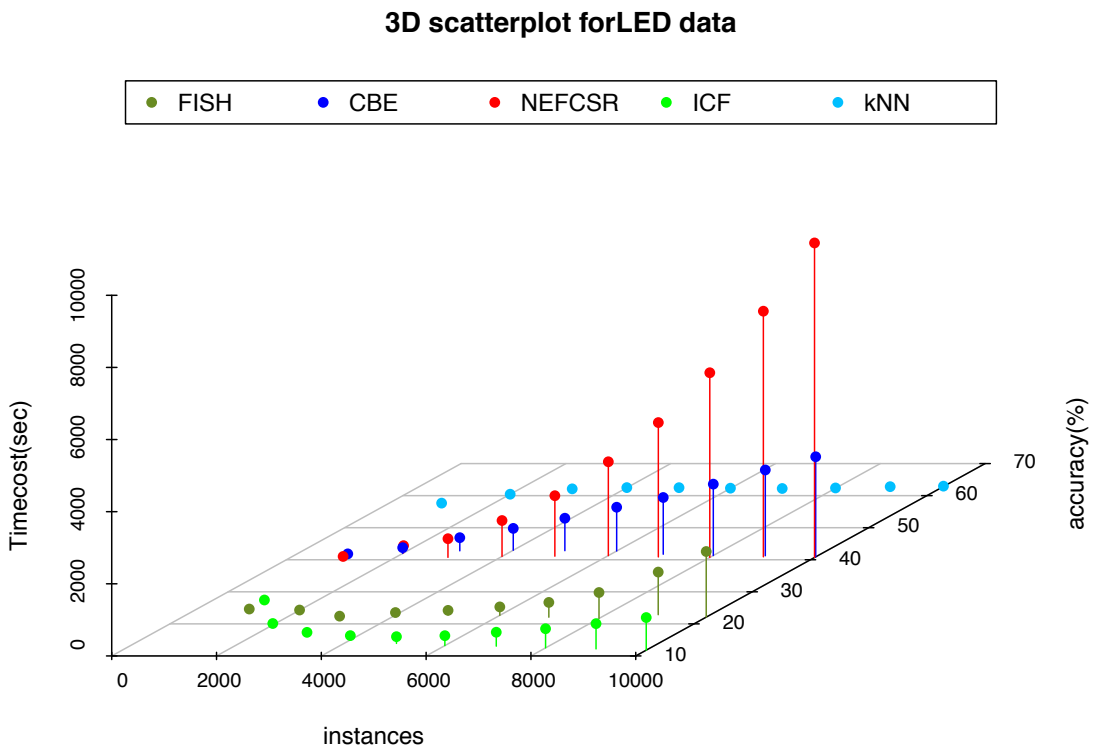
Σχετικά με τον χρόνο αξιολόγησης των μεθόδων αυτή την φορά μεγαλύτερο κόστος παρουσιάζει ο αλγόριθμος NEFCS-SRR ενώ γενικά όλοι οι αλγόριθμοι χρειάζονται λιγότερο χρόνο για να ολοκληρώσουν σε σχέση με τα αντίστοιχα πειράματα που έτρεξαν για τα δεδομένα STAGGER και SEA.



Εικόνα 3.23 : Κόστος μνήμης των μεθόδων στα δεδομένα LED

Το κόστος μνήμης σε RAM-Hours φαίνεται στο διάγραμμα της εικόνας 3.23 όπου η μέθοδος NEFCS-SRR που χρειάζεται και περισσότερο χρόνο για να ολοκληρώσει είναι αυτή που απαιτεί και μεγαλύτερο χώρο στην μνήμη. Όλες οι άλλες μέθοδοι παρουσιάζουν χαμηλές απαιτήσεις σε μνήμη.

Τέλος δίνεται το διάγραμμα διασποράς που δίνει μια καλύτερη και πιο άμεση εικόνα για το ποια μέθοδος απέδωσε καλύτερα στα δεδομένα ροής LED με βάση την ακρίβεια και τον χρόνο αξιολόγησης του μοντέλου.



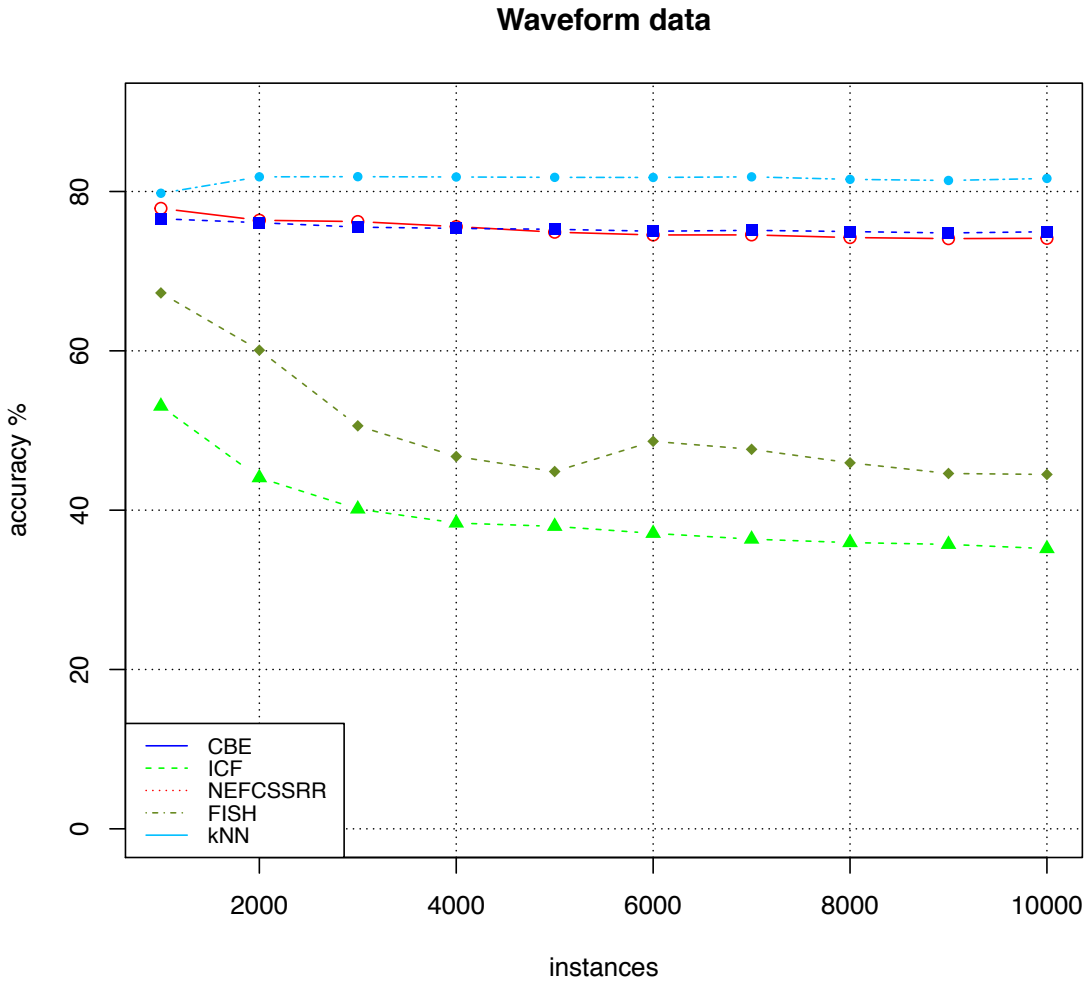
Εικόνα 3.24 : Διάγραμμα διασποράς στα δεδομένα LED

3.5.2.4 Δεδομένα WAVEFORM

Στα waveform δεδομένα ροής οι μέθοδοι μείωσης επιτυγχάνουν την καλύτερη ακρίβεια για p= 5000 δηλαδή όταν ενεργοποιούνται μόνο μια φορά. Στο ίδιο συμπέρασμα καταλήξαμε και σε όλα τα προηγούμενα δεδομένα ροής που εξετάστηκαν. Επομένως αυτό αποτελεί ισχυρή ένδειξη ότι οι αλγόριθμοι αυτοί αν και μειώνουν το σύνολο δεδομένων απομακρύνοντας τα περιττά και θορυβώδη στιγμιότυπα εντούτοις έχουν κόστος στην ακρίβεια του τελικού μοντέλου. Εκτελούμαι τα αντίστοιχα πειράματα στα δεδομένα ροής Waveform για p = 500.

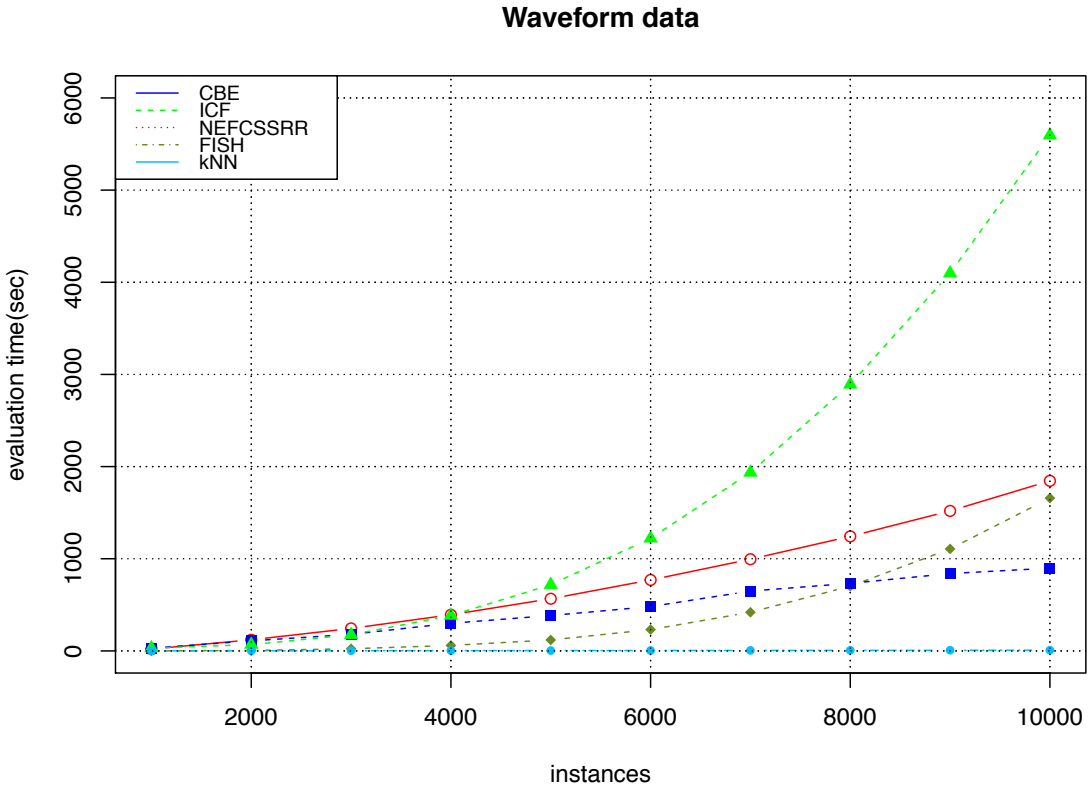
Πίνακας 3.6: Έρευνα για την τιμή του p στα δεδομένα WAVEFORM

	CBE		NEFCS-SRR (s=100)		FISH		ICF	
	accuracy	Time	accuracy	Time	accuracy	Time	accuracy	Time
p = 50	69,73	116	70,69	3847	42,19	17428	-	
p = 100	71,08	204	70,57	2451	44,59	9273	-	
p=500	74,94	898	74,12	1845	44,48	1658	35,17	5596
p=1000	75,46	1696	76,46	2544	48,60	800	37,55	2734
p=2000	76,91	3477	77,67	3462	45,77	404	42,28	1842
p=5000	**78,81**	4159	**79,93**	2420	**78,44**	138	**57,13**	3277



Εικόνα 3.25 : Διάγραμμα ακρίβειας στα δεδομένα Waveform

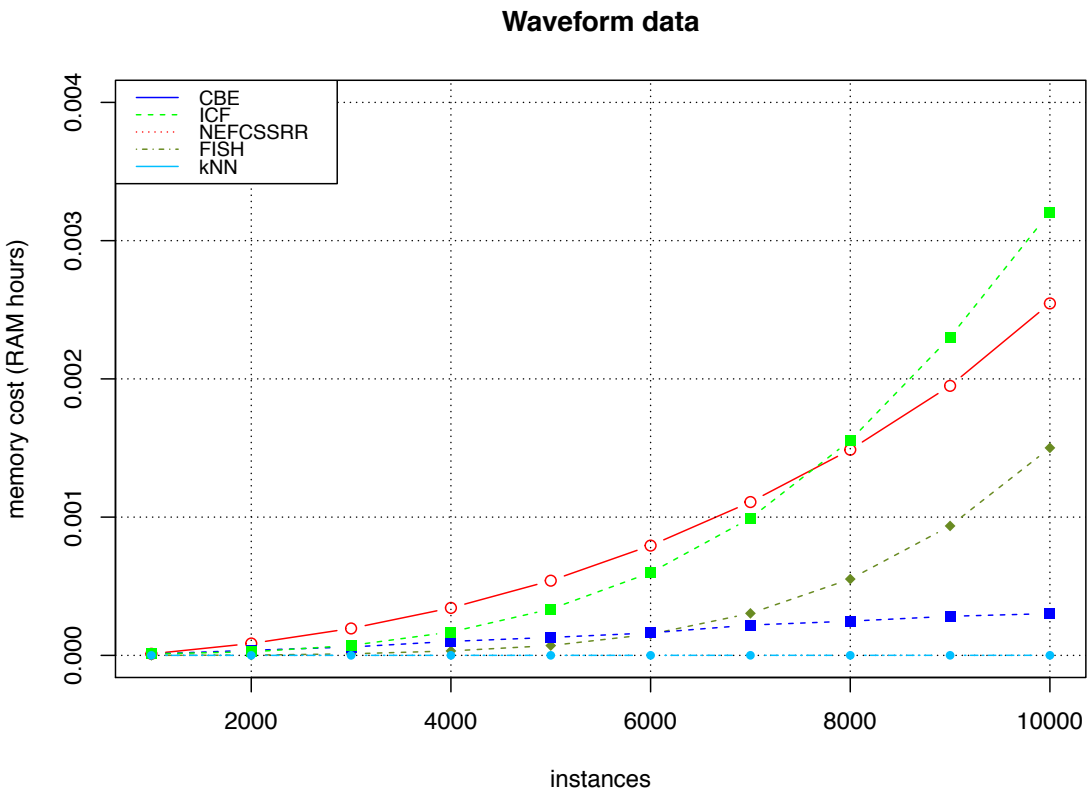
Και σε αυτή την περίπτωση παρατηρούμε ότι ο kNN επιτυγχάνει την υψηλότερη ακρίβεια και ακολουθείται από τις μεθόδους CBE και NEFCS-SRR. Οι μέθοδοι ICF και FISH αντιθέτως σημειώνουν μείωση στην ακρίβεια τους με την πάροδο του χρόνου με την μέθοδος ICF να σημειώνει ακρίβεια κάτω από 40%.



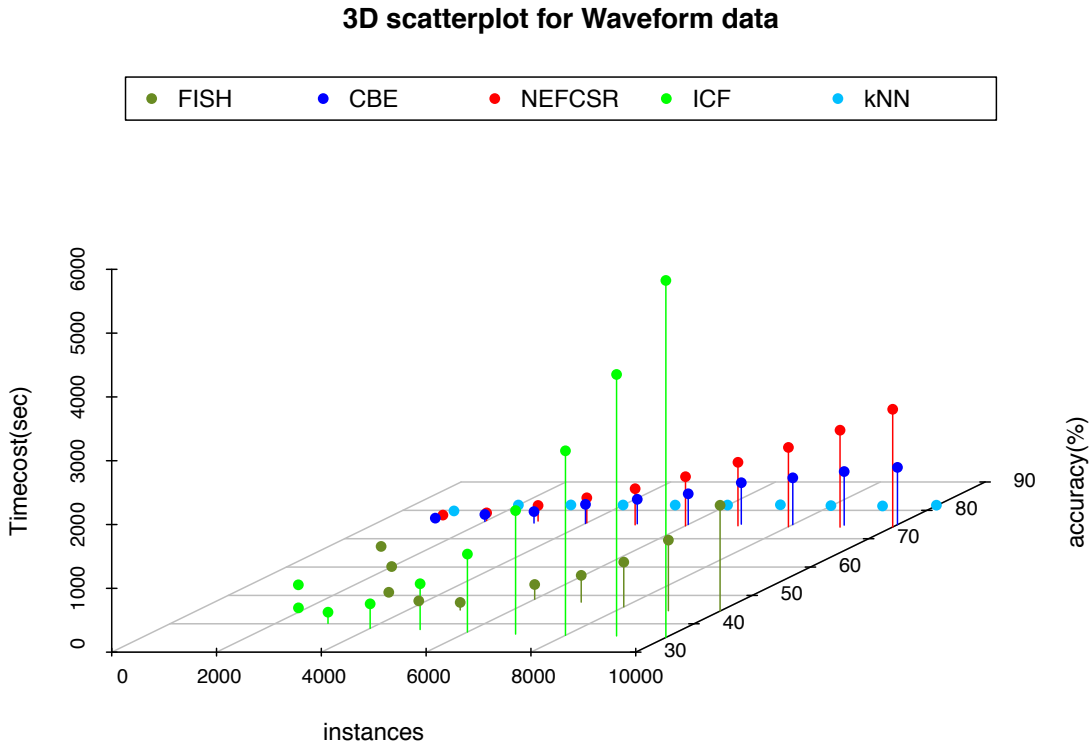
Εικόνα 3.26 : Χρόνος αξιολόγησης μεθόδων στα δεδομένα Waveform

Στην περίπτωση των δεδομένων ροής Waveform ο αλγόριθμος ICF δεν αποδίδει καλά ούτε ως προς τον χρόνο καθώς απαιτεί πάνω από 5000 δευτερόλεπτα για να ολοκληρώσει το πείραμα. Οι υπόλοιπες μέθοδοι χρειάζονται κάτω από 2000 δευτερόλεπτα με τον ταχύτερο αλγόριθμο να παραμένει ο απλός kNN.

Επιπρόσθετα εξετάζοντας την εικόνα 3.27 παρατηρούμε ότι ο αλγόριθμος ICF απαιτεί και την περισσότερη μνήμη σε σχέση με τις υπόλοιπες μεθόδους. Δεύτερος σε κόστος μνήμης έρχεται ο αλγόριθμος NEFCS-SRR ενώ οι αλγόριθμοι FISH και CBE απαιτούν πολύ λίγη μνήμη.



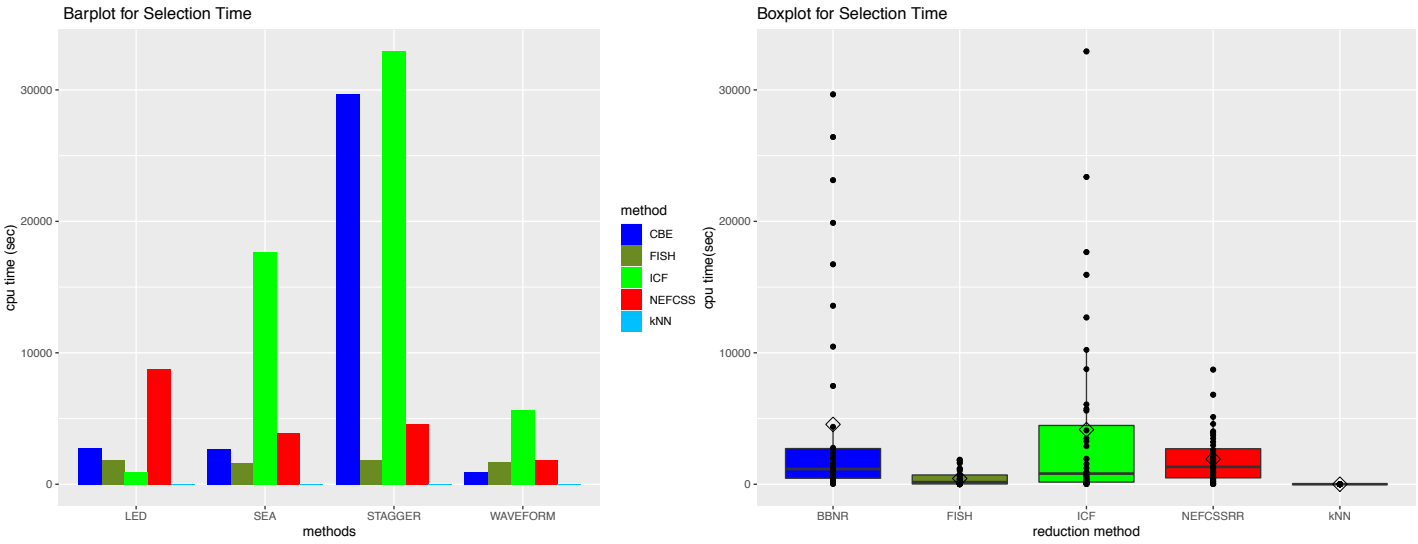
Εικόνα 3.27: Κόστος μνήμης των μεθόδων στα δεδομένα Waveform



Εικόνα 3.28 : Διάγραμμα διασποράς στα δεδομένα Waveform

3.5.2.5 Σύγκριση μεθόδων

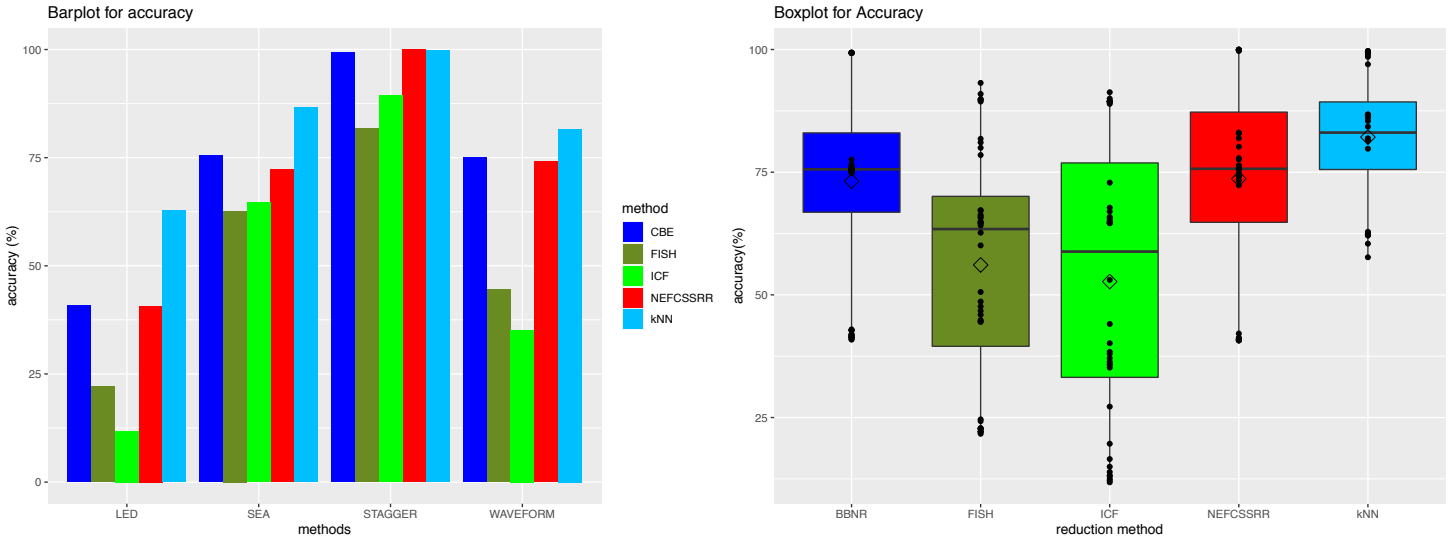
Στο τμήμα αυτό θα επιχειρήσουμε να συγκρίνουμε τις τέσσερις μεθόδους ως προς την απόδοση τους στα τέσσερα δεδομένα ροής. Συγκεκριμένα θα εξετάσουμε με την βοήθεια ραβδογραμμάτων αλλά και θηκογραμμάτων ποια ήταν η ακρίβεια που πέτυχαν οι μέθοδοι μείωσης αλλά και οι απαιτήσεις που είχαν σε χρόνο . Επιπλέον θα εξετάσουμε ποια ήταν η μείωση που επετεύχθη από κάθε μέθοδο στα τέσσερα αυτά σύνολα. Τέλος θα εξετάσουμε με χρήση μη παραμετρικών ελέγχων αν υπάρχει στατιστικά σημαντική διαφορά μεταξύ των μεθόδων.



Εικόνα 3.29 : Ραβδόγραμμα και θηκόγραμμα για τον χρόνο των μεθόδων

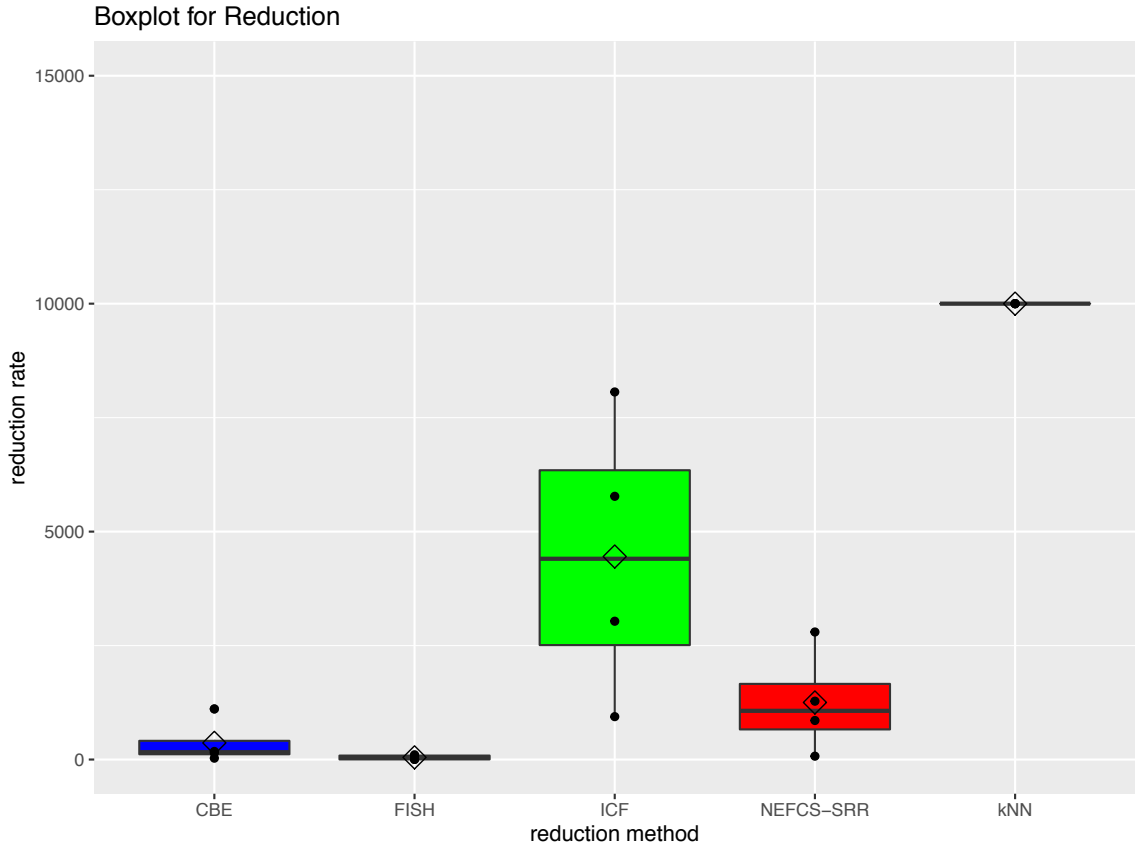
Με βάση τα διαγράμματα της εικόνας 3.29 παρατηρούμε ότι ο αλγόριθμος ICF είναι ιδιαίτερα χρονοβόρος στα δεδομένα SEA και STAGGER. Ο αλγόριθμος CBE με εξαίρεση τα δεδομένα STAGGER όπου παρουσιάζει πολύ υψηλή πολυπλοκότητα ως προς τον χρόνο αποδίδει ικανοποιητικά στα υπόλοιπα σύνολα δεδομένων. Εντούτοις αυτός που έχει σταθερά χαμηλές απαιτήσεις ως προς τον χρόνο είναι ο αλγόριθμος FISH ο οποίος έρχεται δεύτερος μετά τον kNN. Ο τελευταίος δεν εκτελεί κάποια μέθοδο μείωσης επομένως είναι αναμενόμενο να έχει υψηλή απόδοση ως προς τον χρόνο. Λαμβάνοντας

συνεπώς υπόψη και το θηκόγραμμα μπορούμε να συνάγουμε το συμπέρασμα ότι η πιο γρήγορη μέθοδος είναι η FISH ενώ αυτή που έχει την υψηλότερη πολυπλοκότητα ως προς τον χρόνο είναι η μέθοδος ICF.



Εικόνα 3.30 : Ραβδόγραμμα και θηκόγραμμα για την ακρίβεια των μεθόδων

Από τα γραφήματα της εικόνας 3.30 μπορούμε να συνάγουμε ότι ο kNN έχει την υψηλότερη ακρίβεια και ακολουθείται από τους NEFCS-SRR και CBE. Συγκεκριμένα από τα θηκογράμματα φαίνεται η διάμεσος της μεθόδου NEFCS-SRR να ισούται με 75% όσο και η διάμεσος της CBE. Επιπλέον η διάμεσος και των δύο μεθόδων παρουσιάζεται στο μέσο των ορθογωνίων γεγονός που υποδηλώνει ότι η κατανομή των αποτελεσμάτων της ακρίβειας που παρουσιάζουν οι δύο μέθοδοι είναι συμμετρική. Η μέθοδος μείωσης ICF φαίνεται να μην αποδίδει τόσο καλά ενώ η μέθοδος FISH αν και έχει χαμηλότερη απόδοση από τις NEFCS-SRR και CBE εντούτοις όπως προαναφέρθηκε αποδίδει καλύτερα από όλες ως προς την πολυπλοκότητα σε χρόνο.



Εικόνα 3.31: Θηκογράμματα μείωσης των μεθόδων

Αναφορικά με την μείωση που επιτυγχάνουν οι τέσσερις μέθοδοι από τα θηκογράμματα της εικόνας 3.31 παρατηρούμε ότι την μεγαλύτερη μείωση την επιτυγχάνει ο αλγόριθμος FISH. Συγκεκριμένα η διάμεσος του θηκογράμματος που αντιστοιχεί στον FISH είναι κάτω από 100 στιγμιότυπα γεγονός που υποδηλώνει ότι η μέθοδος αυτή πραγματοποιεί μείωση της τάξης του 99% και άνω. Ο εξαιρετικός βαθμός μείωσης που παρουσιάζει ο FISH εξηγείται από το γεγονός ότι επιλέγει τον kNN για κάθε νέο παράδειγμα. Αυτό όμως εξηγεί και την χαμηλή του απόδοση ως προς την ακρίβεια. Ο FISH ακολουθείται από τον CBE ο οποίος επίσης παρουσιάζει πολύ υψηλό ποσοστό μείωσης στα δεδομένα που εξέτασε ενώ και σε αυτή την περίπτωση ο ICF αποδίδει χειρότερα από τους άλλους αλγορίθμους έχοντας τιμή διαμέσου κοντά στα 5000 στιγμιότυπα.

Στον πίνακα 3.7 βλέπουμε τα αποτελέσματα (p-values) του μη παραμετρικού ελέγχου Wilcoxon που διενεργήθηκε κατά ζεύγη για τις αποδόσεις των τεσσάρων μεθόδων μείωσης και του kNN. Διενεργούμε τον συγκεκριμένο μη παραμετρικό έλεγχο διότι δεν μπορούμε να υποθέσουμε ότι τα δεδομένα μας

προέρχονται από κανονική κατανομή. Ο στατιστικός έλεγχος Wilcoxon όπως και όλοι οι στατιστικοί έλεγχοι υποθέτει ότι τα σφάλματα είναι ανεξάρτητα.

Δεδομένου ότι έχουμε μικρό δείγμα τιμών ο έλεγχος Wilcoxon θα δίνει τιμή p μεγαλύτερη του 0.05 σε κάθε περίπτωση ανεξαρτήτως του πόσο απέχει ο δειγματικός μέσος από τον υποθετικό μέσο. Εξετάζοντας όμως την κάθε δυάδα μεθόδων με βάση το αν ο μέσος της ακρίβειας είναι μεγαλύτερος ή μικρότερος μπορούμε να εξάγουμε πιο ασφαλή συμπεράσματα. Για τον σκοπό αυτό εκμεταλλευόμαστε το θηκόγραμμα της εικόνας 3.30. Για παράδειγμα με την βοήθεια του εξετάζουμε την μηδενική υπόθεση :

H_0 : η μέση ακρίβεια του kNN είναι μικρότερη από την μέση ακρίβεια του CBE

Από τον πίνακα 3.7 παρατηρούμε ότι αυτός ο στατιστικός έλεγχος δίνει τιμή p = 0,03 η οποία είναι μικρότερη από το επίπεδο σημαντικότητας a = 0,05 επομένως συνάγουμε ότι πρέπει να απορρίψουμε την μηδενική υπόθεση και να ενεργοποιήσουμε την εναλλακτική η οποία είναι ότι η μέση ακρίβεια του kNN είναι μεγαλύτερη από την μέση ακρίβεια του CBE κάτι το οποίο επαληθεύεται και στο θηκόγραμμα της εικόνας 3.30 . Με όμοιο τρόπο εργαζόμαστε και στα άλλα ζεύγη.

Πίνακας 3.7: Κατά ζεύγη συγκρίσεις με χρήση του Wilcoxon signed rank test

	CBE	FISH	ICF	NEFCS-SRR
FISH	0,03	-	-	-
ICF	0,03	0,209	-	-
NEFCS-SRR	0,089	0,03	0,03	-
kNN	0,03	0,03	0,03	0,053

Από τις τιμές που βλέπουμε στον πίνακα συμπεραίνουμε τα εξής:

- Η καλύτερη μέθοδος όσον αφορά την ακρίβεια φαίνεται να είναι ο kNN ο οποίος αποδίδει καλύτερα από τους CBE, FISH και ICF. Ωστόσο στη σύγκριση που υπάρχει μεταξύ του kNN και του NEFCS-SRR δεν μπορούμε να εξάγουμε το συμπέρασμα ότι ο kNN αποδίδει καλύτερα από τον NEFCS-SRR διότι η τιμή του p είναι 0,053. Φυσικά το ότι το p είναι μεγαλύτερο από

το επίπεδο σημαντικότητας a = 0,05 δεν μας επιτρέπει να υποθέσουμε την μηδενική υπόθεση ότι ο kNN έχει μικρότερη ακρίβεια από τον NEFCS-SRR. Απλώς δεν έχουμε πειστικά στοιχεία για να ισχυριστούμε με ασφάλεια την εναλλακτική υπόθεση του ότι ο kNN αποδίδει καλύτερα από τον NEFCS-SRR.

- Με την ίδια λογική δεν μπορούμε να συνάγουμε το συμπέρασμα ότι ο NEFCS-SRR έχει καλύτερη ακρίβεια από τον CBE ούτε το ότι ο ICF έχει καλύτερη ακρίβεια από τον FISH.

- Ο NEFCS-SRR μέσω του παραμετρικού ελέγχου φαίνεται ότι έχει καλύτερη επίδοση από τους FISH και ICF αλλά δεν μπορούμε να συνάγουμε ότι είναι καλύτερος του CBE ούτε του kNN.

- Ο FISH και ICF δεν αποδίδουν καλύτερα από καμία άλλη μέθοδο αλλά στη μεταξύ τους σύγκριση δεν μπορούμε να εξάγουμε με ασφάλεια ότι υπάρχει στατιστικά σημαντική διαφορά ως προς την ακρίβεια τους.

- Τέλος ο CBE είναι κατώτερος του kNN και επιπλέον δεν μπορούμε να συνάγουμε ότι έχει διαφορά ως προς την ακρίβεια με τον NEFCS-SRR. Παρόλα αυτά φαίνεται να αποδίδει καλύτερα από τους FISH και ICF.

Πίνακας 3.8: Friedman rank sum test

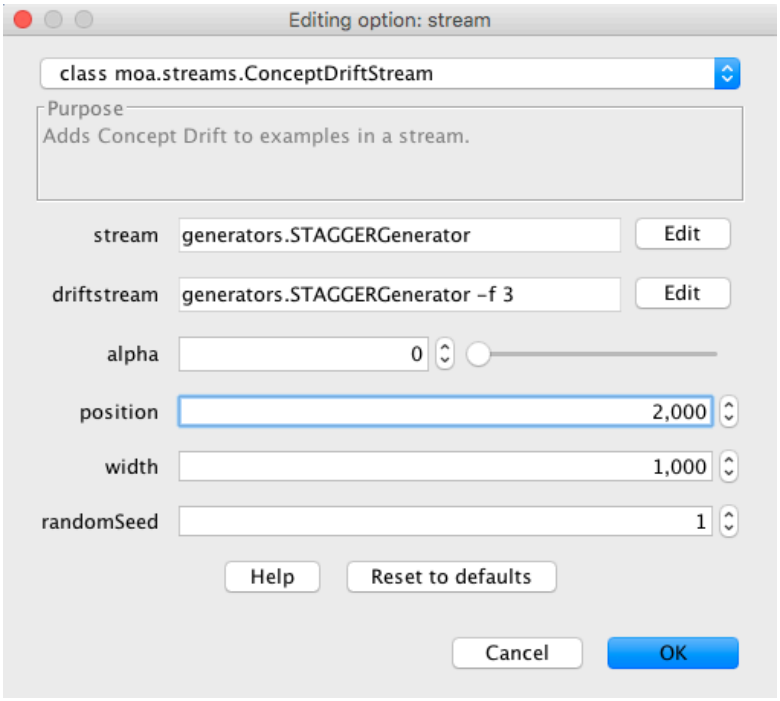
Friedman chi-squared	**17,12**
df	4
p-value	0,001832

Τέλος ο πίνακας 3.8 μας παρουσιάζει τα αποτελέσματα του μη παραμετρικού ελέγχου Friedman. Μέσω του ελέγχου αυτού επιβεβαιώνουμε ότι οι τέσσερις μέθοδοι έχουν στατιστικά σημαντικές διαφορές όσον αφορά την ακρίβεια τους.

3.5.3 Εκτέλεση πειραμάτων υπό την παρουσία concept drift

Σε αυτή την παράγραφο της μεθοδολογίας θα εξεταστούν εκ νέου οι τέσσερις μέθοδοι μείωσης στα ίδια δεδομένα και για ίδιο αριθμό εκτελέσεων (p =500) στα οποία δεδομένα όμως αυτή την φορά θα εισαχθεί τεχνητό concept drift. Στο σιγμοειδή μοντέλο του τεχνητού concept drift που εισάγουν οι δημιουργοί του MOA χρειάζεται να καθορίσουμε την χρονική στιγμή που ξεκινάει το drift καθώς και το πλάτος που αυτό θα έχει. Επιλέγουμε να εισάγουμε το concept drift στο στιγμιότυπο 2000. Δηλαδή η κατανομή θα αλλάζει σε αυτό το σημείο και επιπλέον το πλάτος της θα είναι W = 1000. Ο σκοπός της εισαγωγής του τεχνητού concept drift είναι να εξεταστούν οι αποδόσεις των μεθόδων μείωσης στην περίπτωση που το ρεύμα δεδομένων αλλάζει κατανομή. Θέλουμε δηλαδή να εξετάσουμε το πως συμπεριφέρονται οι μέθοδοι όταν υπάρχει το φαινόμενο του concept drift και να τις συγκρίνουμε με τις αντίστοιχες περιπτώσεις στα ίδια δεδομένα όπου δεν υπήρξε αλλαγή στην κατανομή με την οποία καταφθάνουν τα δεδομένα από τις γεννήτριες.

Το τεχνητό drift που εισάγουμε φαίνεται στην εικόνα 3.32. Χρησιμοποιείται η κλάση ConceptDriftStream η οποία είναι η κύρια κλάση για την προσομοίωση των concept drifts. Το πλαίσιο δίπλα από την παράμετρο stream δείχνει το τρέχων ρεύμα δεδομένων και το πλαίσιο δίπλα από την παράμετρο driftstream υποδηλώνει το νέο concept. Για να υπάρξει προσομοίωση ενός concept drift πρέπει να αλλαχθεί το concept και για να συμβεί αυτό χρειάζεται μια γεννήτρια η οποία έχει μια κατανομή η οποία μπορεί να ρυθμιστεί. Οι γεννήτριες STAGGER , SEA , LED , Waveform παρέχουν αυτή την δυνατότητα. Συγκεκριμένα τα STAGGER δεδομένα που εξετάζουμε παρακάτω έχουν τρεις συναρτήσεις που μπορούν να μεταβληθούν με σκοπό να αλλάξουν δραστικά το υποκείμενο concept. Στην περίπτωση που ακολουθεί αλλάζουμε από τα STAGGER με την συνάρτηση 1 στα STAGGER με την συνάρτηση 2. Με όμοιο τρόπο εισάγουμε το drift και στις υπόλοιπες γεννήτριες δεδομένων.



Εικόνα 3.32: Εισαγωγή τεχνητού concept drift στα δεδομένα STAGGER

3.5.3.1 Δεδομένα STAGGER

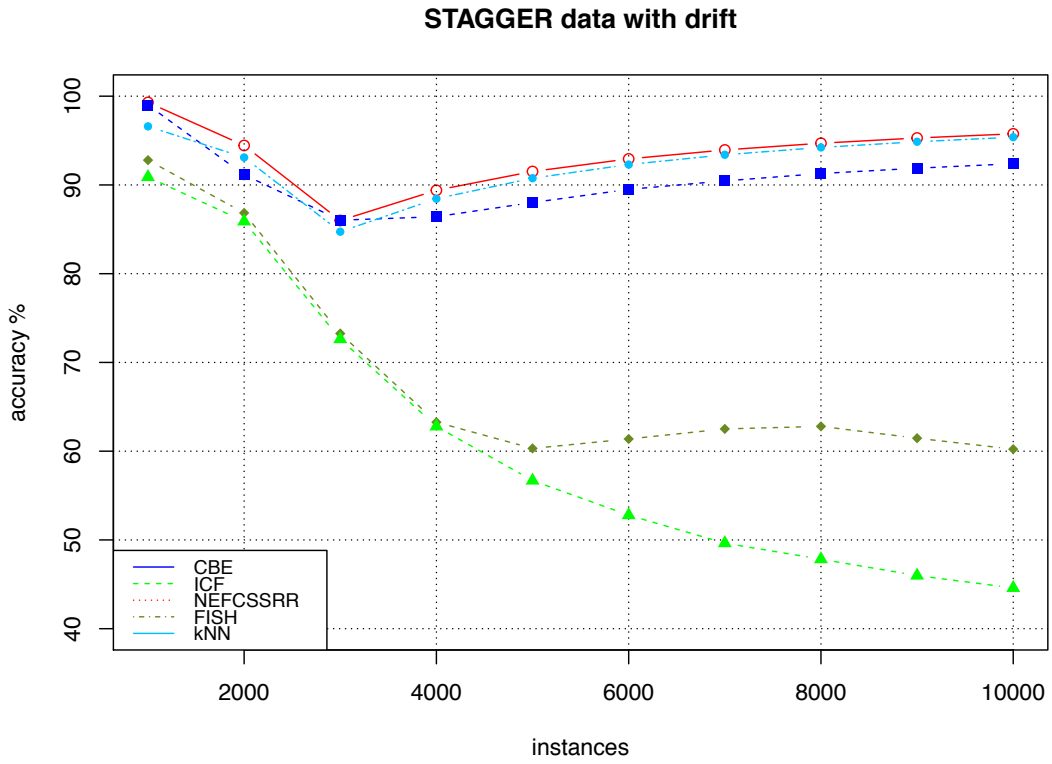
Ξεκινάμε να εξετάσουμε τις μεθόδους για την γεννήτρια δεδομένων STAGGER. Στον πίνακα που ακολουθεί παρατίθενται οι αποδόσεις των μεθόδων σε τρεις περιπτώσεις. Στην πρώτη περίπτωση η οποία έχει εξεταστεί στην προηγούμενη παράγραφο παρουσιάζονται οι αποδόσεις ως προς τον χρόνο και την ακρίβεια για κάθε μέθοδο χωριστά. Στην δεύτερη περίπτωση (δεύτερη στήλη) παρουσιάζονται οι αποδόσεις των μεθόδων όταν εμφανίζεται το concept drift με τα χαρακτηριστικά που περιγράψαμε παραπάνω. Στην τρίτη περίπτωση (τρίτη στήλη) παρατίθενται τα αποτελέσματα ως προς την ακρίβεια και τον χρόνο όταν κατά την διάρκεια του κάθε πειράματος υπήρξε ένα επαναλαμβανόμενο concept drift. Συγκεκριμένα το επαναλαμβανόμενο drift εμφανίζεται τρεις φορές, έχει πλάτος W = 500 και μεταξύ των drift υπάρχει απόσταση χιλίων στιγμιοτύπων. Το πρώτο drift εμφανίζεται για p = 1000 επομένως το επόμενο θα εμφανιστεί για p = 2500 κ.ο.κ. Τα αποτελέσματα των πειραμάτων παρουσιάζουν ιδιαίτερο ενδιαφέρον.

Πίνακας 3.9: Σύγκριση μεθόδων στα STAGGER

	no drift		one drift		recurrent drift	
	accuracy	Time	accuracy	Time	accuracy	Time
data stream	**STAGGER data**					
CBE	99,29	29662	92,38	19869	43,37	12461
NEFCS	99,97	4589	95,75	768	85,78	869
ICF	89,39	32928	44,59	10907	81,54	24094
FISH	81,81	1863	60,22	1545	81,68	1500
kNN	99,69	2	95,37	2	88,22	1,9

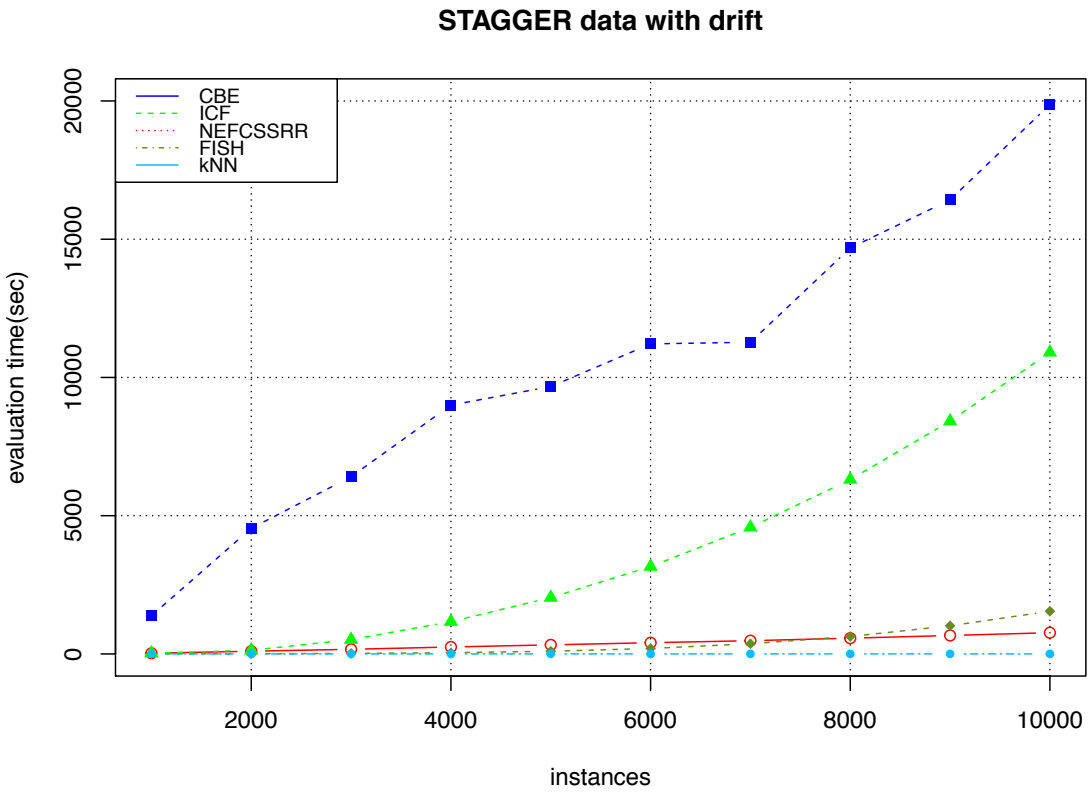
Στην περίπτωση που εμφανίζεται ένα drift παρατηρούμε ότι οι μέθοδοι CBE και NEFCS-SRR αντιδρούν πολύ καλά σε αυτό διότι αν και μειώνεται ελάχιστα η ακρίβεια εντούτοις μειώνεται και η υπολογιστική πολυπλοκότητα. Δεν ισχύει όμως το ίδιο για τις μεθόδους ICF και FISH οι οποίες αν και παρουσιάζουν βελτιωμένο χρόνο εντούτοις φαίνονται να επηρεάζονται σε σημαντικό βαθμό από την αλλαγή της κατανομής όσον αφορά την ακρίβεια τους. Παρατηρώντας την τρίτη στήλη δεν μπορούμε να εξάγουμε τα ίδια συμπεράσματα όταν το drift είναι επαναλαμβανόμενο. Σε αυτή την περίπτωση παρατηρούμε ότι ο αλγόριθμος CBE επηρεάζεται αρκετά ως προς την ακρίβεια η οποία βυθίζεται στο 43,37%. Ο αλγόριθμος NEFCS-SRR δείχνει να επηρεάζεται και αυτός αλλά όχι σε τέτοιο βαθμό καθώς εξακολουθεί να διατηρεί υψηλή ακρίβεια. Ενδιαφέρον παρουσιάζουν και οι μέθοδοι ICF και FISH οι οποίες έχουν καλύτερες αποδόσεις στην περίπτωση του επαναλαμβανόμενου concept drift από ότι στην περίπτωση του σταδιακού drift. Καλύτερη επίδοση από όλους εξακολουθεί να έχει ο απλός kNN, ενώ με βάση των πίνακα 3.9 για τα δεδομένα STAGGER μπορούμε να υποθέσουμε ότι η μέθοδος NEFCS-SRR κερδίζει τις υπόλοιπες και στις τρεις περιπτώσεις.

Ακολουθεί το γράφημα της εικόνας 3.33 όπου παρουσιάζεται η καμπύλη της ακρίβειας για κάθε μέθοδο χωριστά.



Εικόνα 3.33: Ακρίβεια των μεθόδων στα δεδομένα STAGGER με drift

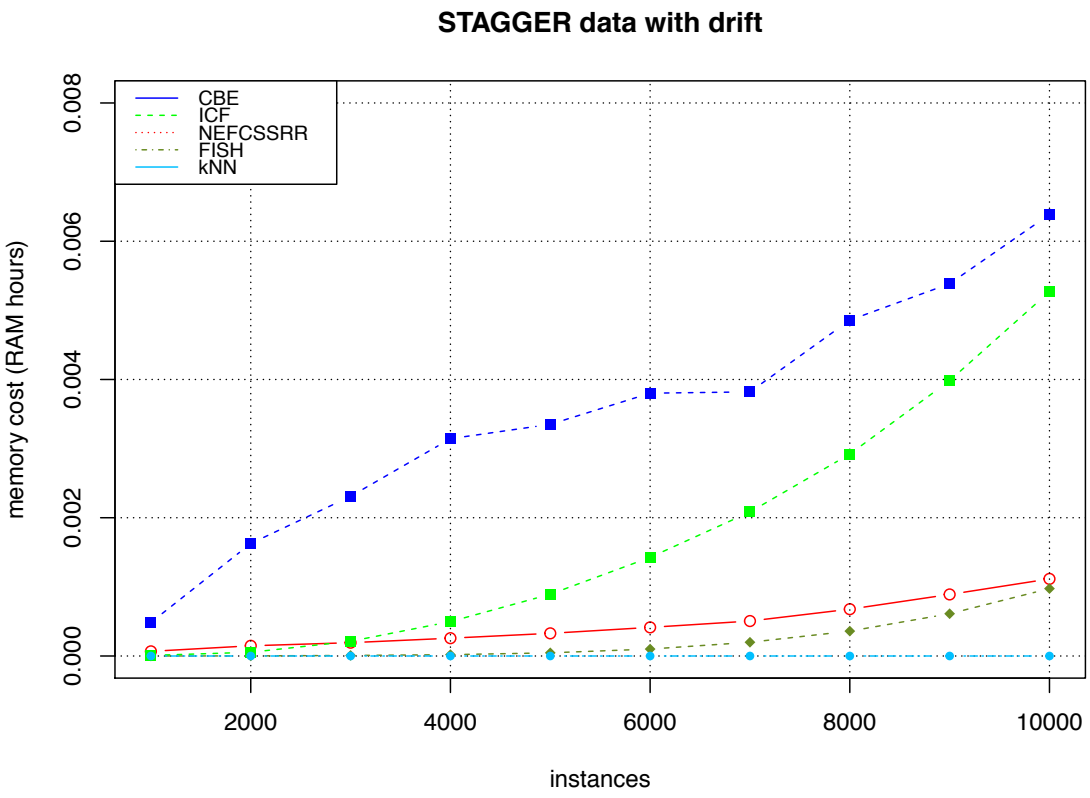
Όπως προαναφέρθηκε το drift εισάγεται στα 2000 στιγμιότυπα. Από τις καμπύλες της ακρίβειας της εικόνας 3.33 παρατηρούμε ότι όλες οι μέθοδοι επηρεάζονται από το concept drift. Παρόλα αυτά οι NEFCS-SRR και CBE φαίνεται να ανακάμπτουν γρήγορα μετά το τέλος του drift που σημειώνεται στα 3000 στιγμιότυπα εν αντιθέσει με τους ICF και FISH οι οποίοι δεν παρουσιάζουν βελτίωση. Συγκεκριμένα ο FISH φαίνεται να σταθεροποιείται στα 5000 στιγμιότυπα και μετά ενώ ο ICF διαρκώς ρίχνει την απόδοση του έως το τέλος.



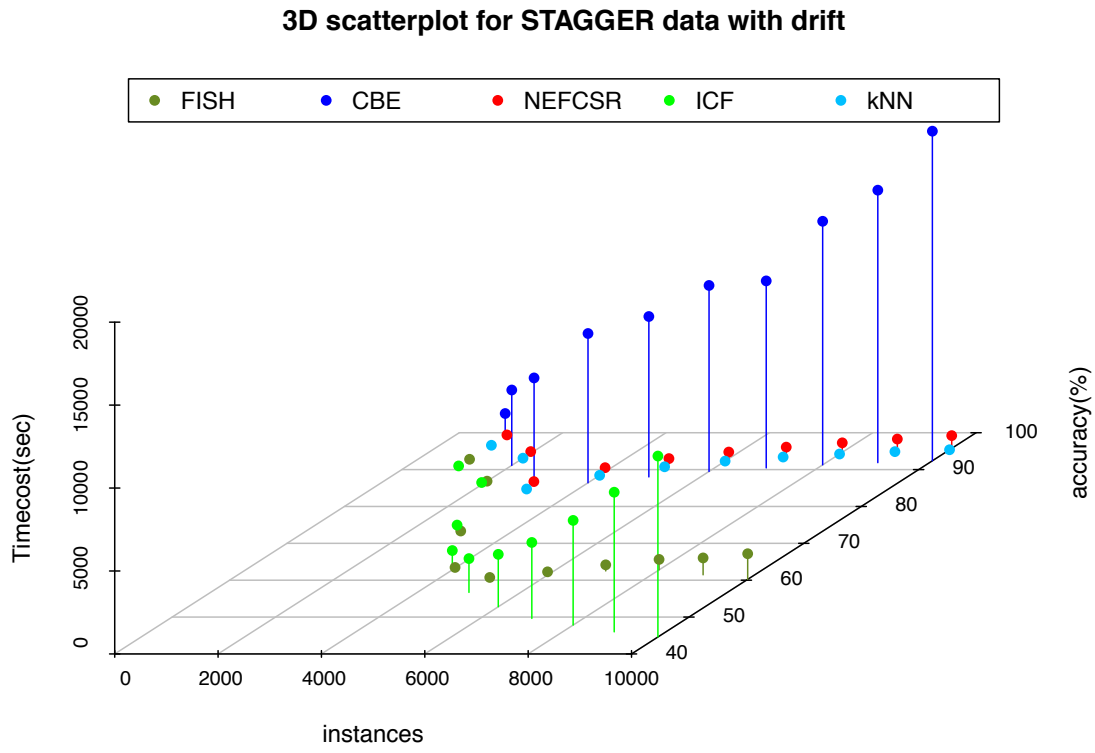
Εικόνα 3.34: Χρόνος αξιολόγησης των μεθόδων στα δεδομένα STAGGER με drift

Η μέθοδος CBE εξακολουθεί να είναι η πιο χρονοβόρα στα δεδομένα STAGGER όπως και στην περίπτωση που δεν υπήρχε το concept drift. Παρόλα αυτά όλες οι μέθοδοι σε αυτή την περίπτωση παρουσιάζουν βελτίωση στην χρονική τους πολυπλοκότητα σε σύγκριση με την περίπτωση της μη ύπαρξης drift με τον NEFCS-SRR να ξεχωρίζει ξανά για την επίδοση του. Ο ICF επίσης παρουσιάζει μεγάλη βελτίωση αλλά έχει μεγάλο κόστος ως προς την ακρίβεια.

Μια ακόμα σημαντική παράμετρος που δεν εξετάζεται στον πίνακα 3.9 είναι αυτή της απαίτησης σε μνήμη που έχουν οι μέθοδοι. Στην εικόνα 3.35 παρατηρούμε ότι οι NEFCS-SRR και FISH έχουν το χαμηλότερο κόστος μνήμης εξαιρουμένου φυσικά του kNN.



Εικόνα 3.35: Κόστος μνήμης των μεθόδων στα δεδομένα STAGGER με drift



Εικόνα 3.36: Διάγραμμα διασποράς στα δεδομένα STAGGER με drift

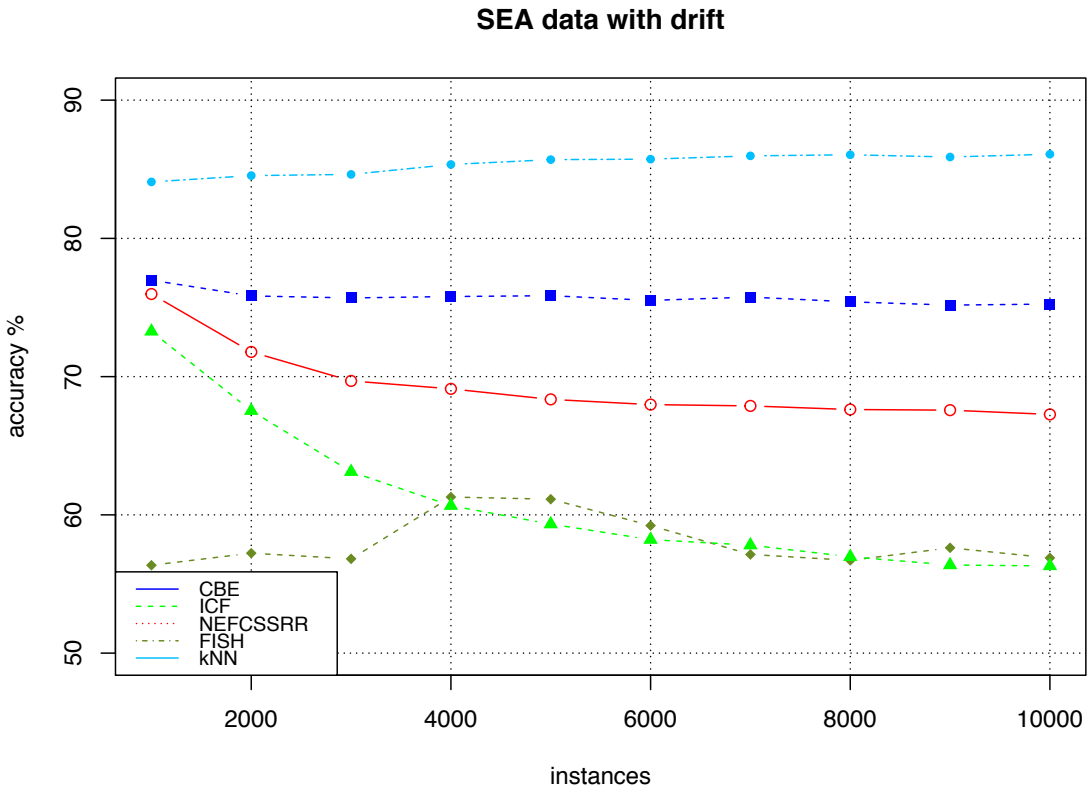
Το διάγραμμα διασποράς της εικόνας 3.36 το οποίο εξετάζει τις μεθόδους ως προς την χρονική πολυπλοκότητα και ως προς την ακρίβεια ταυτοχρόνως υποδεικνύει την μέθοδο NEFCS-SRR ως την καταλληλότερη για την περίπτωση των δεδομένων STAGGER με drift. Η μέθοδος αν και δεν μπορεί να συγκριθεί ως προς την χρονική παράμετρο με τον kNN εντούτοις καταφέρνει να παρουσιάσει την ίδια ακρίβεια με αυτόν.

3.5.3.2 Δεδομένα SEA

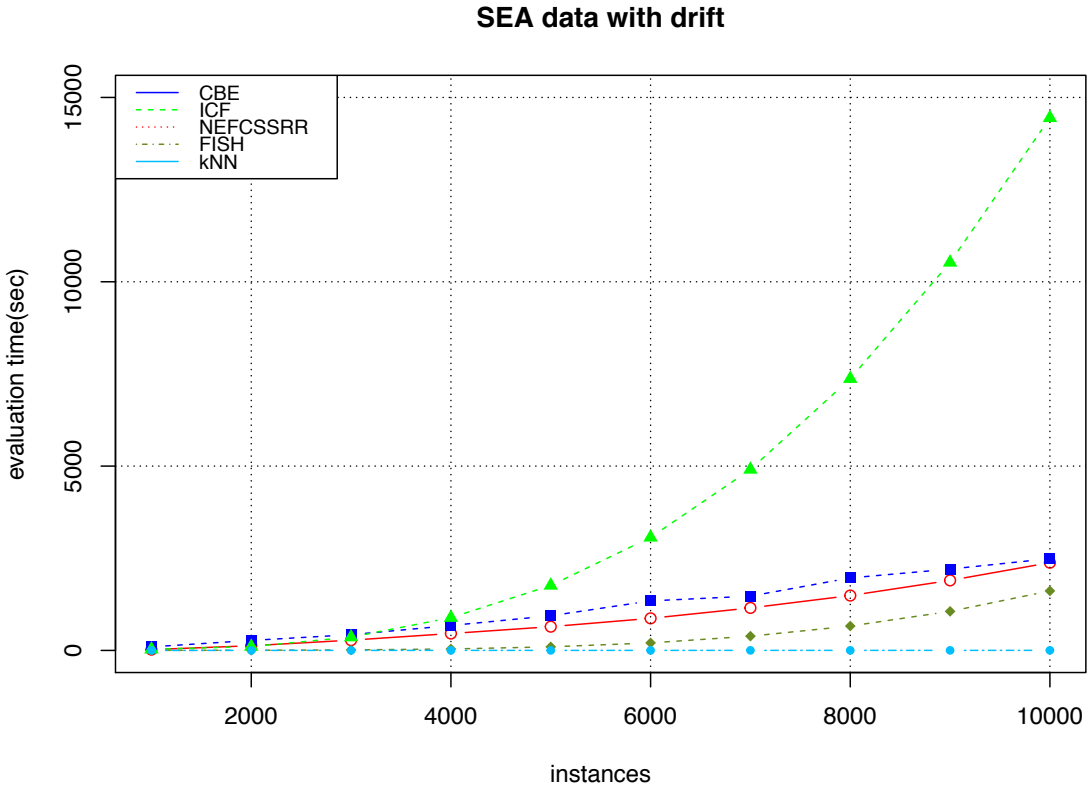
Πίνακας 3.10: Σύγκριση μεθόδων στα SEA

	no drift		drift		recurrent drift	
	accuracy	Time	accuracy	Time	accuracy	Time
data stream	**SEA data**					
CBE	75,58	2696	75,24	2490	74,18	3514
NEFCS	72,34	3905	67,26	2379	67,67	2557
ICF	64,58	17661	56,30	14446	62,93	16513
FISH	62,68	1593	56,88	1613	60,49	1613
kNN	86,65	1,9	86,08	1,7	85,05	1,9

Ο πίνακας 3.10 δείχνει τις επιδόσεις των τεσσάρων μεθόδων και του kNN στα δεδομένα SEA όταν αυτά δεν αλλάζουν κατανομή, όταν υπάρχει ένα drift και όταν το drift είναι επαναλαμβανόμενο. Στην περίπτωση των δεδομένων SEA ο αλγόριθμος CBE φαίνεται να διατηρεί την απόδοση του και στις τρεις περιπτώσεις χωρίς να παρουσιάζει βελτίωση ως προς την χρονική πολυπλοκότητα. Ο NEFCS-SRR παρουσιάζει μείωση στην ακρίβεια αλλά απαιτεί λιγότερο χρόνο για να ολοκληρώσει υπό την παρουσία του drift. Όσον αφορά τους FISH και ICF ρίχνουν και οι δύο την απόδοση τους. Με βάση τα παραπάνω ο αλγόριθμος CBE φαίνεται να είναι ο καλύτερος στα δεδομένα SEA διότι έχει καλύτερη ακρίβεια και την διατηρεί ακόμα και με την παρουσία ενός ή πολλών drifts.

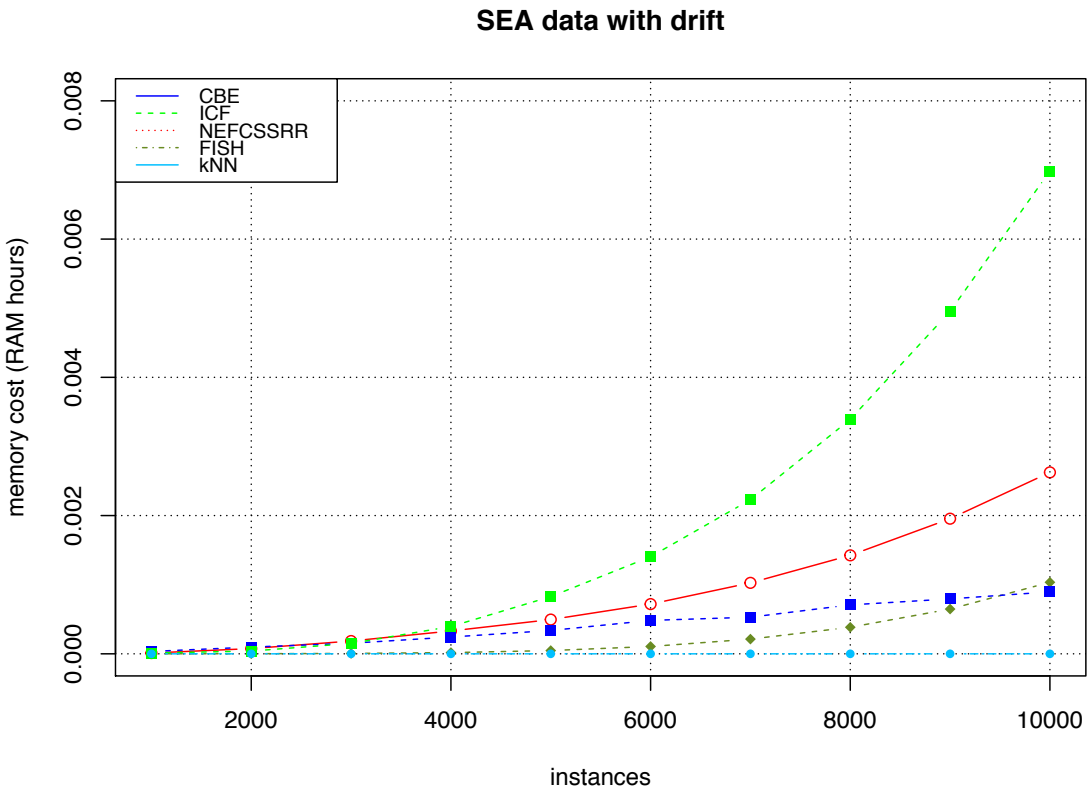


Εικόνα 3.37: Ακρίβεια των μεθόδων στα δεδομένα SEA με drift



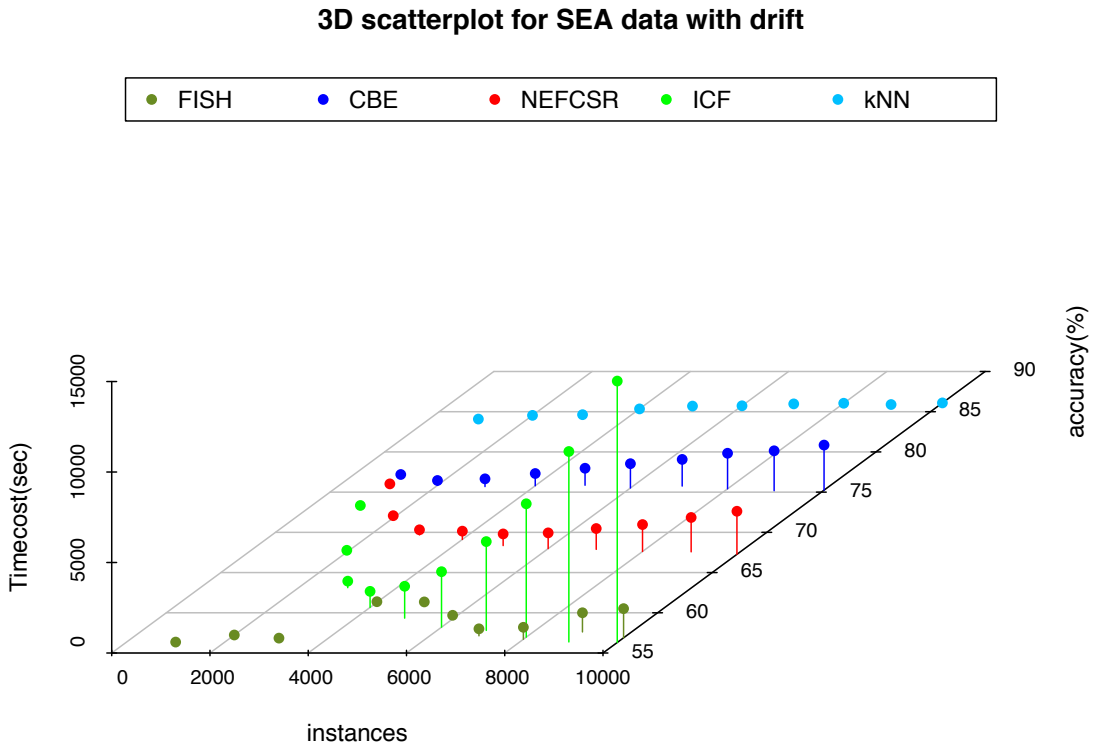
Εικόνα 3.38: Χρόνος αξιολόγησης των μεθόδων στα δεδομένα SEA με drift

Με βάση τα γραφήματα των εικόνων 3.37 και 3.38 παρατηρούμε ότι η μέθοδος CBE διατηρεί την ακρίβεια της σταθερή παρόλο που εμφανίζεται concept drift με κέντρο τα 2000 στιγμιότυπα. Επομένως φαίνεται να μην επηρεάζεται από το drift κάτι το οποίο δεν συμβαίνει για τους υπόλοιπους αλγόριθμους μείωσης. Όσον αφορά το κόστος σε χρόνο ο CBE ανταγωνίζεται τις μεθόδους NEFCS - SRR και FISH. Η μέθοδος που φαίνεται να έχει την χειρότερη επίδοση είναι ο ICF που απαιτεί πολύ χρόνο και επιτυγχάνει χαμηλή ακρίβεια κάτι το οποίο παρατηρήθηκε και στην περίπτωση που δεν υπήρξε drift. Συνεπώς δεν μπορούμε να συνάγουμε το συμπέρασμα ότι είναι η παρουσία του drift που τον οδηγεί να έχει χαμηλές επιδόσεις αλλά η φύση των δεδομένων SEA.



Εικόνα 3.39: Κόστος μνήμης των μεθόδων στα δεδομένα SEA με drift

Όσον αφορά την παράμετρο της μνήμης ο CBE φαίνεται και σε αυτή την περίπτωση να έχει την καλύτερη επίδοση συγκριτικά με όλες τις άλλες μεθόδους μείωσης.



Εικόνα 3.40: Διάγραμμα διασποράς στα δεδομένα SEA με drift

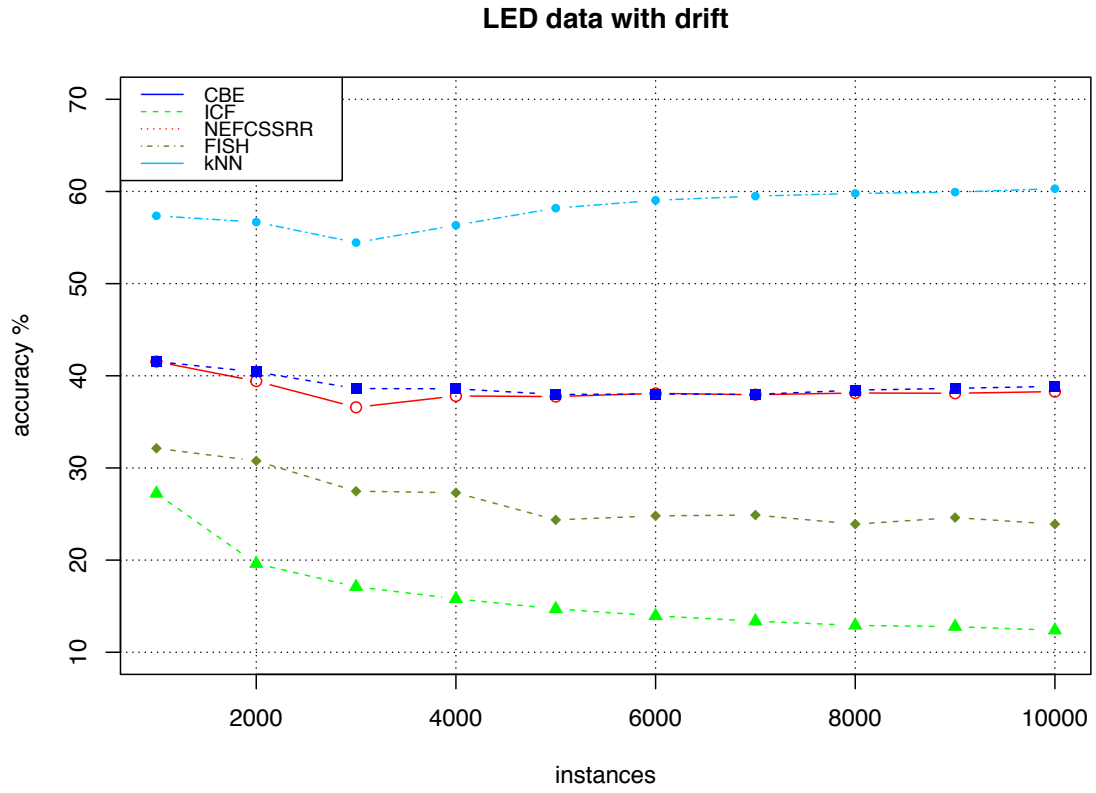
Το διάγραμμα διασποράς επιβεβαιώνει ότι ο CBE έχει την καλύτερη επίδοση ως προς τις δύο κύριες παραμέτρους του χρόνου και της ακρίβειας συγκριτικά πάντα με τις υπόλοιπες μεθόδους μείωσης καθώς ο kNN διαρκώς επιτυγχάνει καλύτερη ακρίβεια και σε λιγότερο χρόνο.

3.5.3.3 Δεδομένα LED

Ο πίνακας 3.11 που ακολουθεί παραθέτει τα αποτελέσματα των τεσσάρων μεθόδων και του kNN ως προς την ακρίβεια και τον χρόνο για τα δεδομένα LED όταν αυτά δεν παρουσιάζουν drift, όταν παρουσιάζουν ένα drift στα 5 από τα 7 τμήματα της LED διάταξης και όταν το drift είναι επαναλαμβανόμενο. Βάση αποτελεσμάτων η ακρίβεια μειώνεται για λίγο στους CBE και NEFCS-SRR χωρίς να υπάρχει ευδιάκριτη βελτίωση στον χρόνο. Αντιθέτως οι μέθοδοι ICF και FISH φαίνεται να μην επηρεάζονται καθόλου από την παρουσία του drift καθότι δεν μειώνεται η ακρίβεια τους αντιθέτως αυξάνεται οριακά.

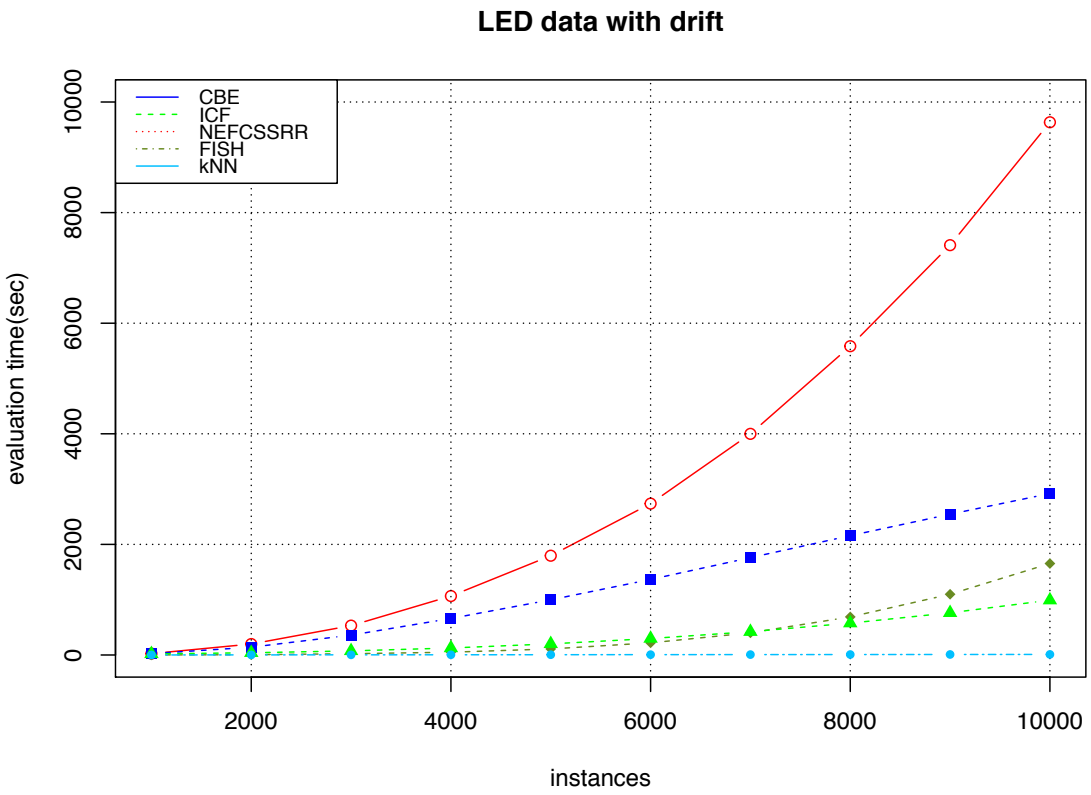
Πίνακας 3.11: Σύγκριση μεθόδων στα LED

	no drift		drift		recurrent drift	
	accuracy	Time	accuracy	Time	accuracy	Time
data stream	**LED data**					
CBE	40,90	2776	38,85	2919	36,99	3306
NEFCS	40,70	8722	38,28	9633	37,51	10589
ICF	11,80	904	12,39	991	12,26	970
FISH	22	1820	23,90	1654	21,04	2138
kNN	62,84	10	60,30	10	56,20	10

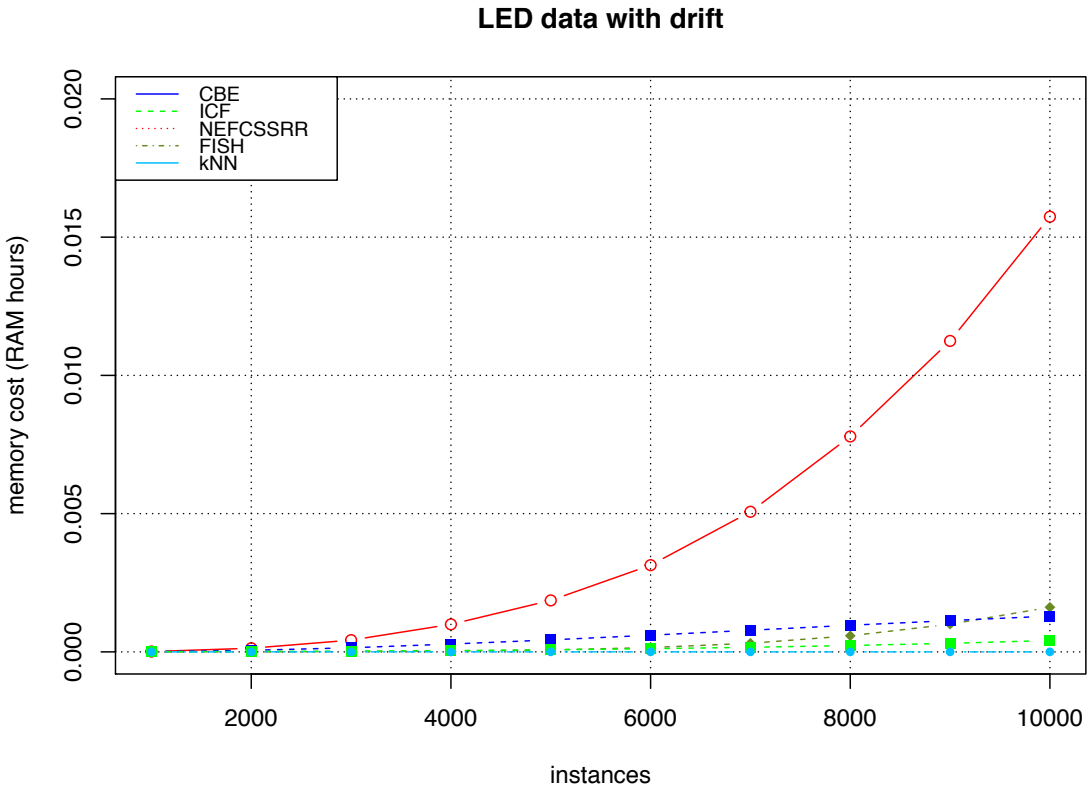


Εικόνα 3.41: Ακρίβεια των μεθόδων στα δεδομένα LED με drift

Οι μέθοδοι CBE και NEFC-SRR αντιδρούν καλύτερα στο drift εν αντιθέσει με τους ICF και FISH που φαίνεται η ακρίβεια τους να μειώνεται διαρκώς. Την καλύτερη επίδοση επιτυγχάνει ο kNN που αν και ρίχνει την ακρίβεια του στο διάστημα που εμφανίζεται το drift βελτιώνει έπειτα την απόδοση του.

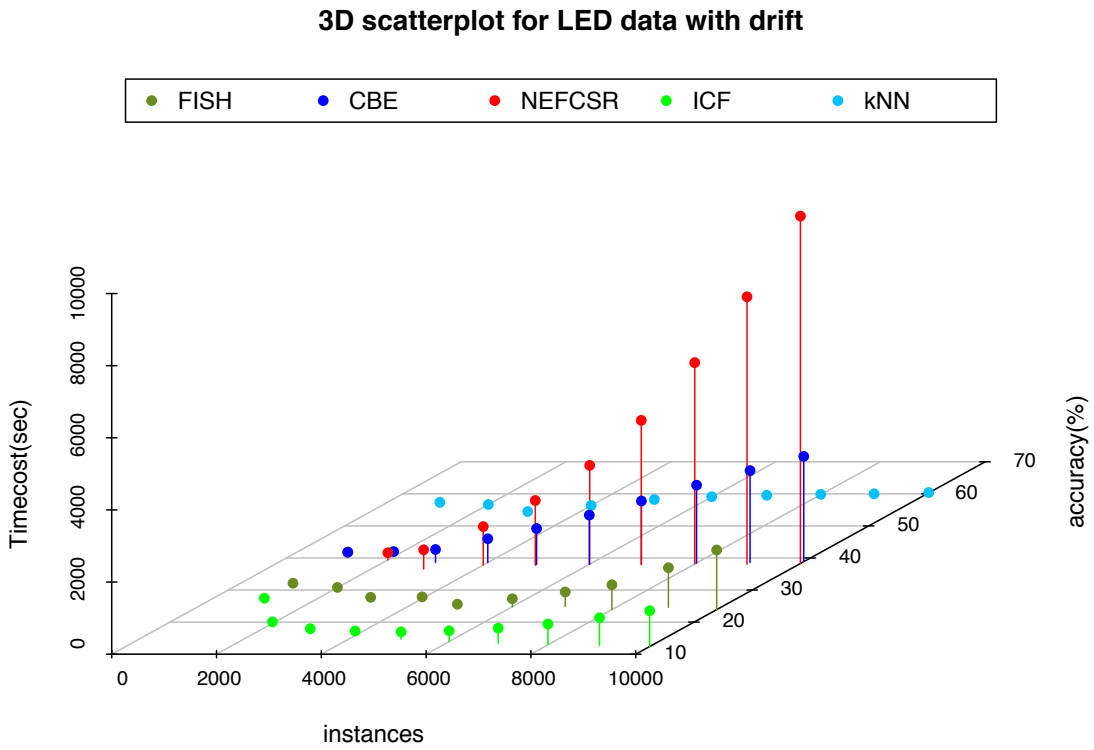


Εικόνα 3.42: Χρόνος αξιολόγησης των μεθόδων στα δεδομένα LED με drift



Εικόνα 3.43: Κόστος μνήμης των μεθόδων στα δεδομένα LED με drift

Από τις εικόνες 3.42 και 3.43 παρατηρούμε ότι η μέθοδος NEFCS-SRR έχει υψηλό κόστος σε μνήμη και χρόνο σε σχέση με τις υπόλοιπες μεθόδους. Η αμέσως επόμενη είναι η μέθοδος CBE η οποία όμως έχει πολύ χαμηλότερες απαιτήσεις. Λαμβάνοντας υπόψη το ότι οι δύο αυτές μέθοδοι επιτυγχάνουν την υψηλότερη ακρίβεια μπορούμε να υποθέσουμε ότι στα LED με drift η μέθοδος CBE φαίνεται να έχει πιο καλή απόδοση.



Εικόνα 3.44: Διάγραμμα διασποράς στα δεδομένα LED με drift

Το διάγραμμα διασποράς της εικόνας 3.44 φαίνεται να επαληθεύει την παραπάνω υπόθεση καθότι μας δείχνει τον CBE να επιτυγχάνει την ίδια ακρίβεια με τον NEFCS-SRR αλλά σε πολύ μικρότερο χρόνο. Ο αλγόριθμος kNN για άλλη μια φορά φαίνεται να ξεπερνά σε συνολική απόδοση όλες τις μεθόδους μείωσης.

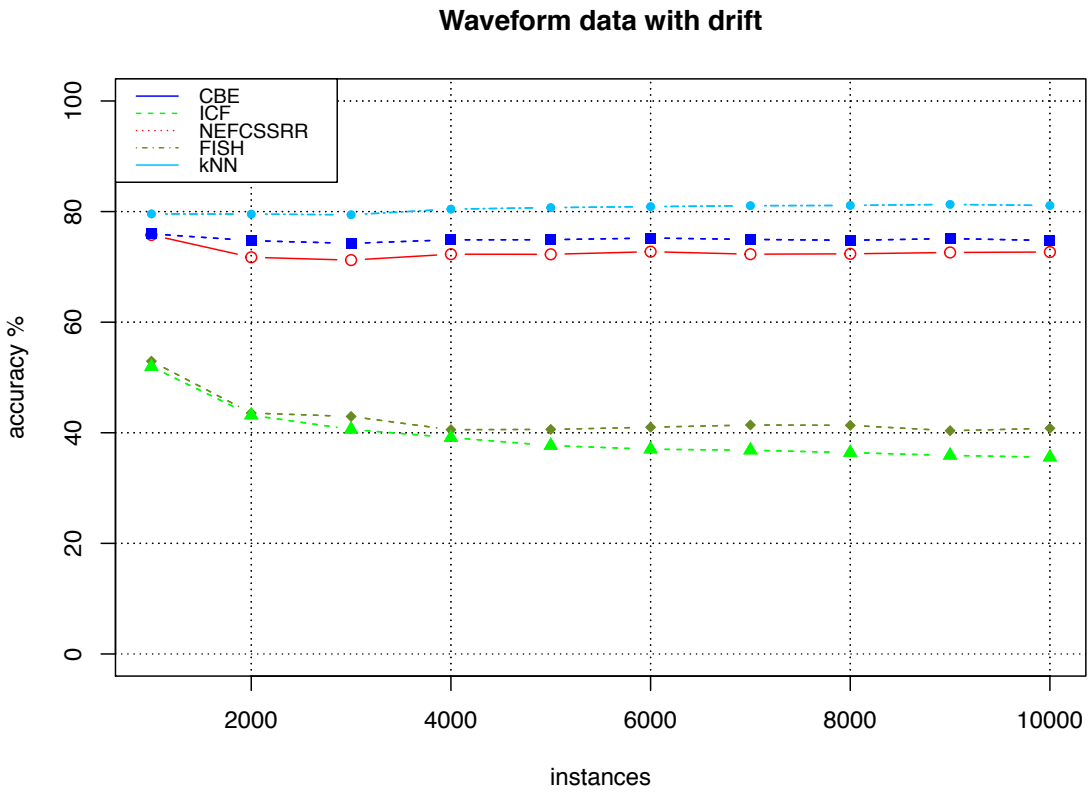
3.5.3.4 Δεδομένα Waveform

Πίνακας 3.12: Σύγκριση μεθόδων στα Waveform

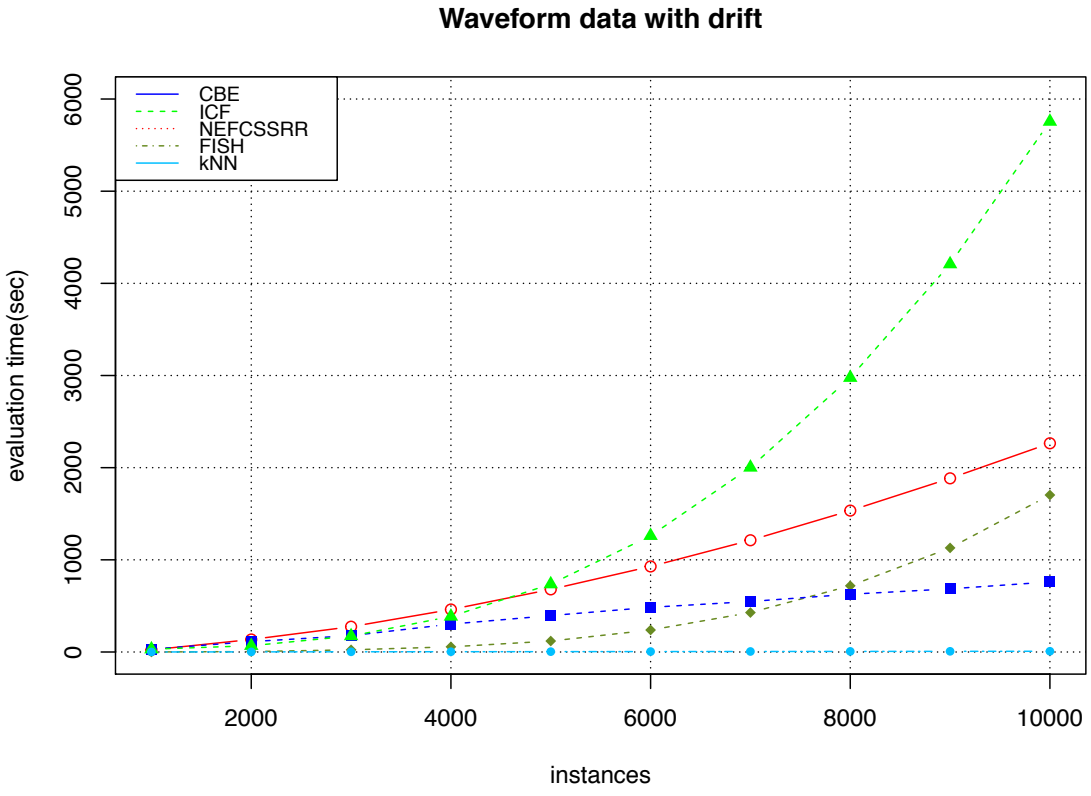
	no drift		drift		recurrent drift	
	accuracy	Time	accuracy	Time	accuracy	Time
data stream	**Waveform data**					
CBE	74,94	898	74,77	760	72,56	760
NEFCS	74,12	1845	72,68	2264	71,88	2241
ICF	35,17	5595	35,57	5755	35,38	5614
FISH	44,48	1658	40,81	1703	43,81	1979
kNN	81,63	8	81,09	8	80,19	8

Σύμφωνα με τον πίνακα 3.12 η μέθοδος που επιτυγχάνει την καλύτερη ακρίβεια και στις τρεις περιπτώσεις είναι η CBE. Επιπρόσθετα η CBE απαιτεί πολύ λιγότερο χρόνο για να ολοκληρώσει συγκριτικά με τις άλλες μεθόδους. Αξιοσημείωτο είναι το γεγονός ότι δεν ρίχνει καθόλου την ακρίβεια της στην περίπτωση που υπάρχει ένα drift ενώ στην περίπτωση του επαναλαμβανόμενου drift μειώνεται ελάχιστα απαιτώντας όμως και λιγότερο χρόνο για να ολοκληρώσει. Δεύτερη σε απόδοση είναι η μέθοδος NEFCS-SRR η οποία δείχνει να επηρεάζεται από την παρουσία του drift. Τέλος η ICF φαίνεται να διατηρεί την απόδοση της και στις τρεις περιπτώσεις ενώ η μέθοδος FISH αποδίδει καλύτερα στην περίπτωση που δεν υπάρχει drift. Το τεχνητό drift που εισήχθη περιλαμβάνει την εμφάνιση drift σε 20 από τα 40 αριθμητικά γνωρίσματα της γεννήτριας Waveform.

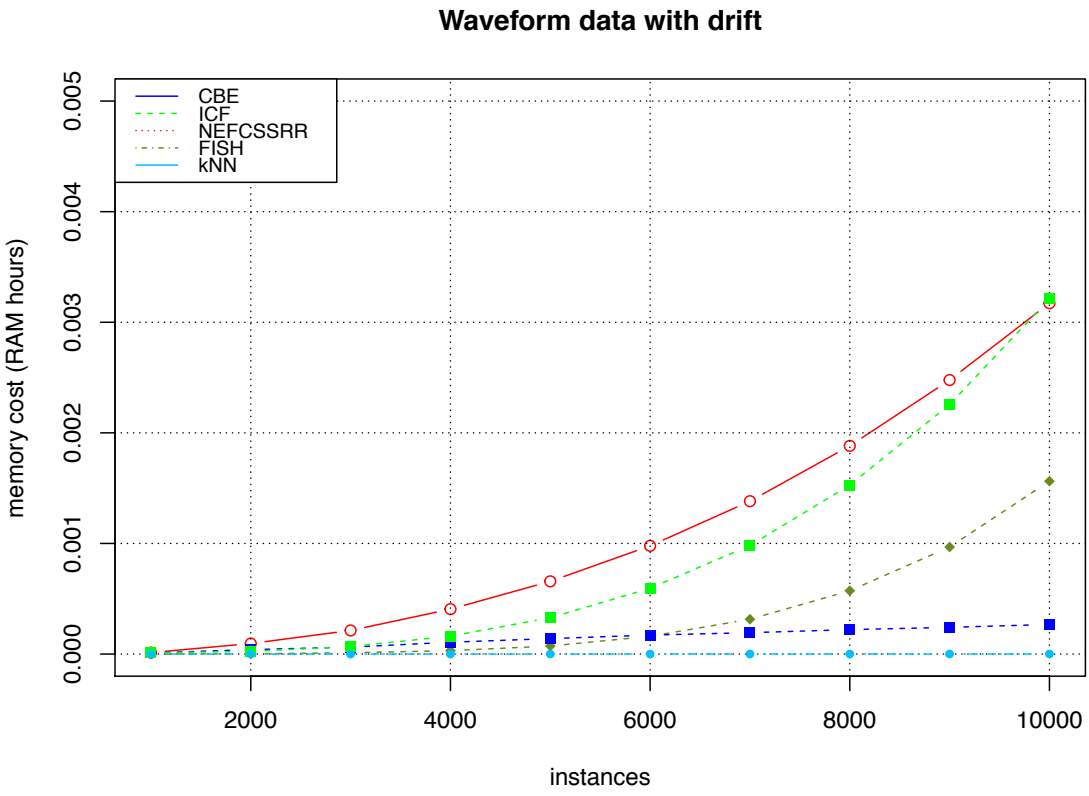
Στη συνέχεια παρουσιάζονται τα γραφήματα που συγκρίνουν τις επιδόσεις των μεθόδων ως προς την ακρίβεια, την χρονική πολυπλοκότητα και τις απαιτήσεις σε μνήμη. Μέσω των γραφημάτων μπορούμε να συνάγουμε το συμπέρασμα ότι και στα δεδομένα Waveform ο CBE φαίνεται να αποδίδει καλύτερα από τις υπόλοιπες μεθόδους μείωσης.



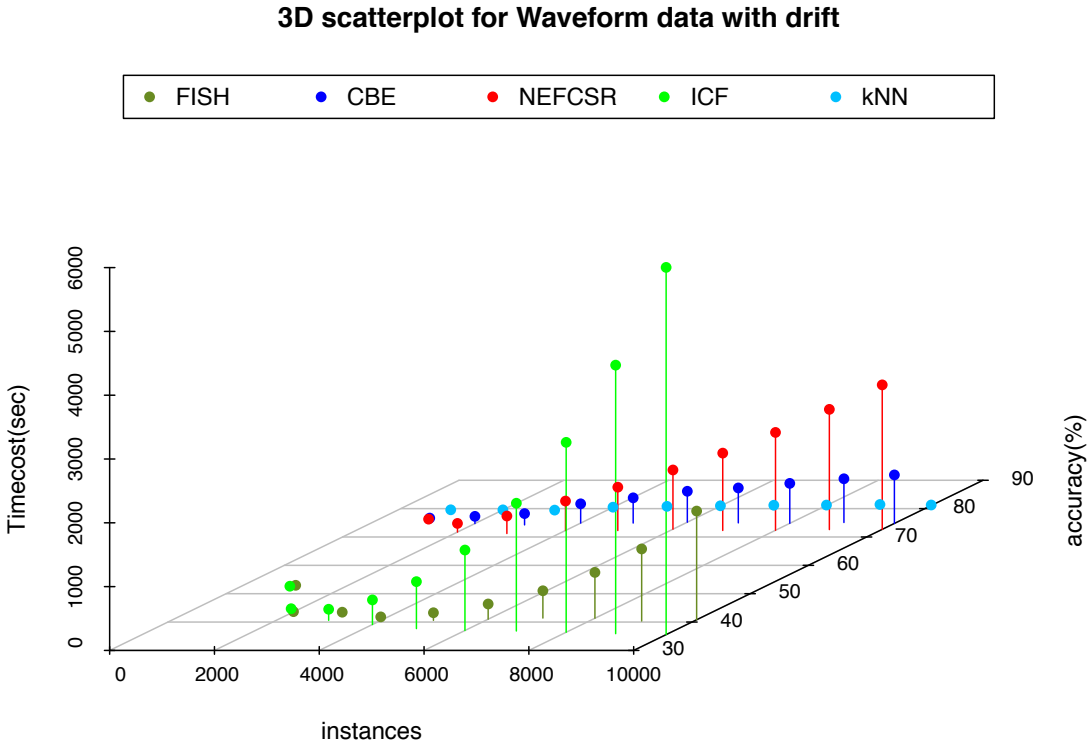
Εικόνα 3.45: Ακρίβεια των μεθόδων στα δεδομένα Waveform με drift



Εικόνα 3.46: Χρόνος αξιολόγησης των μεθόδων στα δεδομένα Waveform με drift



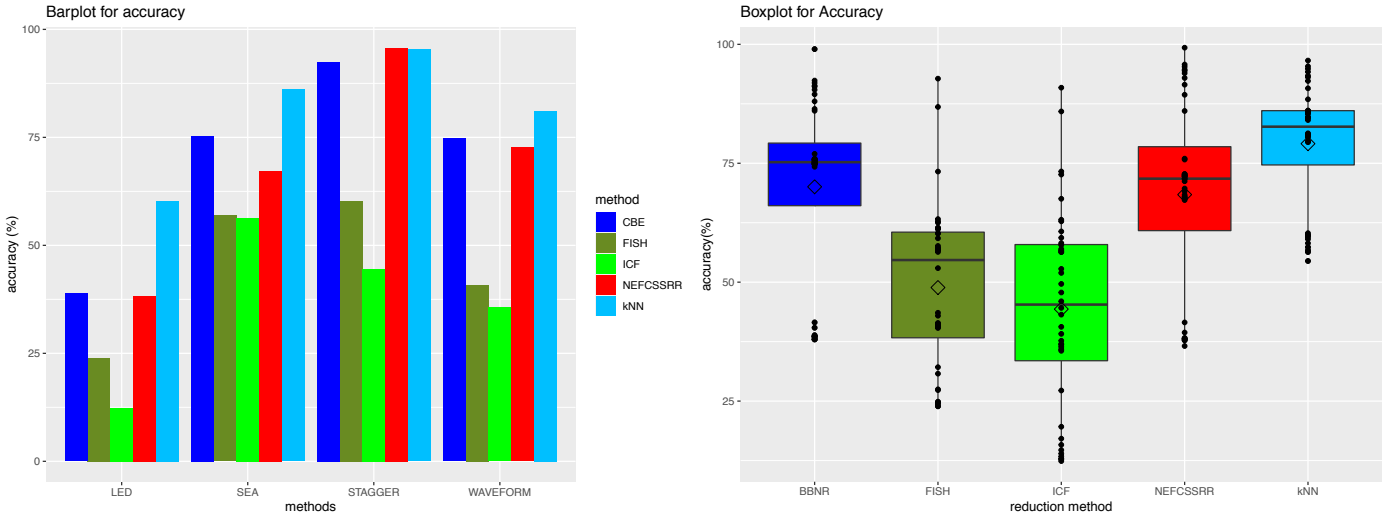
Εικόνα 3.47: Κόστος μνήμης των μεθόδων στα δεδομένα Waveform με drift



Εικόνα 3.48: Διάγραμμα διασποράς στα δεδομένα Waveform με drift

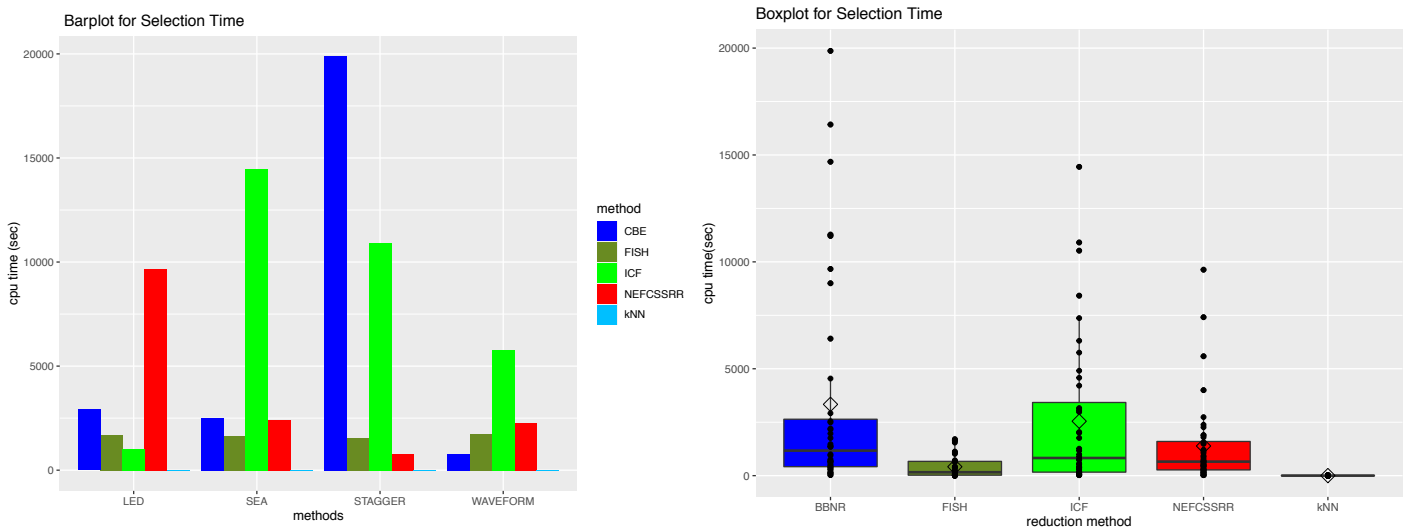
3.5.3.5 Σύγκριση μεθόδων

Θα επιχειρήσουμε και στην περίπτωση που υπάρχει drift να συγκρίνουμε τις μεθόδους με χρήση γραφημάτων και μη παραμετρικών ελέγχων έτσι ώστε να καταλήξουμε στο ποια μέθοδος αποδίδει καλύτερα σε δεδομένα ροής που εμφανίζουν concept drift.



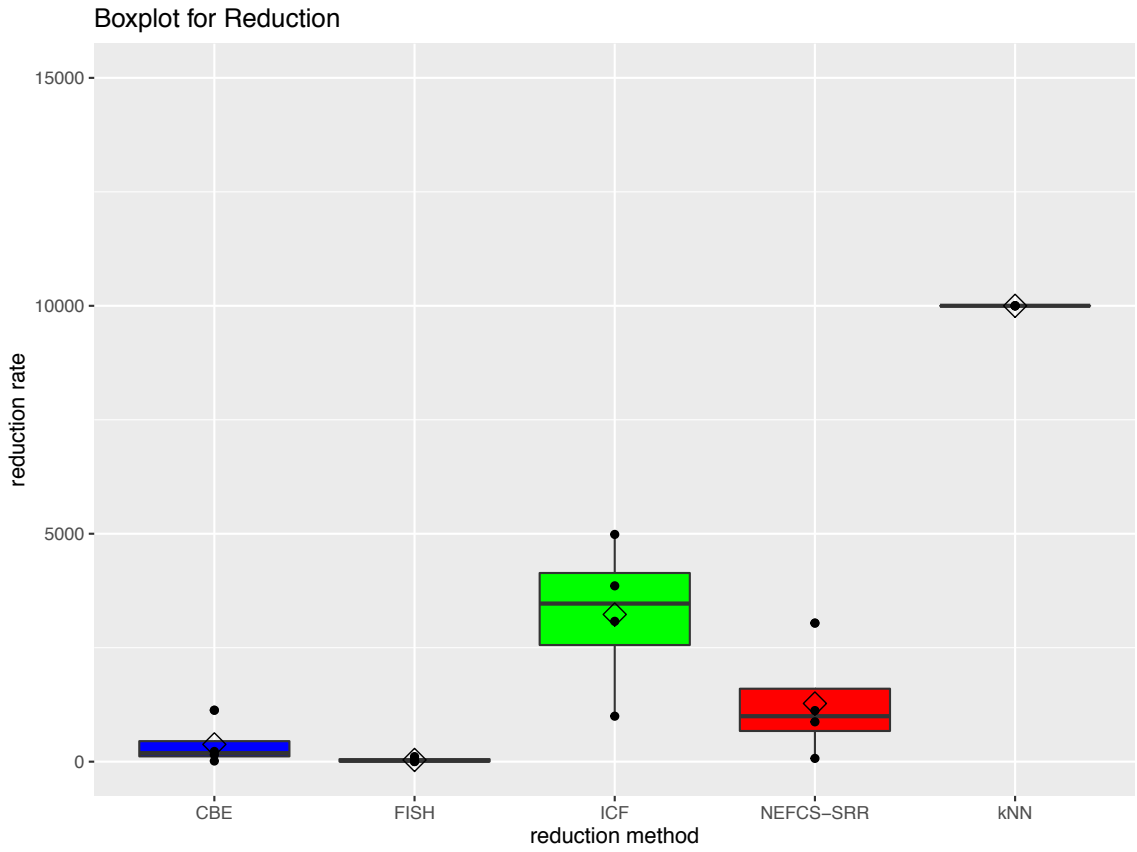
Εικόνα 3.49: Ραβδόγραμμα και θηκόγραμμα για την ακρίβεια των μεθόδων

Από το θηκόγραμμα της εικόνας 3.49 παρατηρούμε ότι ο kNN αποδίδει πολύ καλύτερα από τους υπόλοιπους αλγορίθμους στο σύνολο. Όσον αφορά το κάθε σύνολο δεδομένων ξεχωριστά έχει υψηλότερη ακρίβεια σε όλα εκτός από τα STAGGER όπου ο NEFCS-SRR φαίνεται να τον ξεπερνάει οριακά. Οι μέθοδοι NEFCS-SRR και CBE φαίνεται σύμφωνα με το θηκόγραμμα να επιτυγχάνουν περίπου την ίδια ακρίβεια συνολικά. Εξετάζοντας το ραβδόγραμμα παρατηρούμε ότι σε κάθε σύνολο δεδομένων ξεχωριστά οι αποδόσεις των δύο μεθόδων είναι περίπου οι ίδιες με εξαίρεση τα δεδομένα SEA όπου η CBE επιτυγχάνει καλύτερη ακρίβεια. Οι μέθοδοι FISH και ICF φαίνεται να αποδίδουν χειρότερα από τις δύο προηγούμενες μεθόδους.



Εικόνα 3.50: Ραβδόγραμμα και θηκόγραμμα για τον χρόνο των μεθόδων

Σχετικά με την χρονική πολυπλοκότητα παρατηρούμε ότι η μέθοδος FISH επιτυγχάνει την καλύτερη απόδοση συνολικά μετά τον kNN. Από το θηκόγραμμα βλέπουμε επίσης ότι μετά την μέθοδο FISH η NEFCS-SRR είναι αυτή που παρουσιάζει τον δεύτερο καλύτερο χρόνο ενώ πολύ κοντά είναι και η μέθοδος CBE. Από το ραβδόγραμμα παρατηρούμε ότι μόνο στα δεδομένα STAGGER η μέθοδος CBE αποδίδει πολύ χειρότερα από την NEFCS-SRR ενώ σε όλες τις άλλες περιπτώσεις η CBE φαίνεται να αποδίδει το ίδιο ή καλύτερα. Συνεπώς όσον αφορά την παράμετρο του χρόνου η μέθοδος που φαίνεται να ξεχωρίζει είναι η FISH.



Εικόνα 3.51: Θηκογράμματα μείωσης των μεθόδων όταν υπάρχει drift

Στα θηκογράμματα της εικόνας 3.51 παρατηρούμε ότι διατηρείται το ίδιο μέγεθος μείωσης με την περίπτωση που δεν υπήρχε concept drift στα δεδομένα ροής. Εξαίρεση αποτελεί η μέθοδος ICF που παρουσιάζει μεγαλύτερη μείωση από ότι πριν. Συγκεκριμένα η χειρότερη μείωση που επιτυγχάνει είναι περίπου 5000 στιγμιότυπα στα 10000 το οποίο σημειώνεται για τα δεδομένα LED. Η μέθοδος ICF στα δεδομένα χωρίς drift είχε παρουσιάσει πολύ μικρή μείωση (περίπου 2000 στιγμιότυπα) στα δεδομένα STAGGER γεγονός που ανέβαζε την διάμεσο του θηκογράμματος κοντά στα 5000 στιγμιότυπα.

Όσον αφορά τις άλλες μεθόδους παρατηρούμε ότι την μεγαλύτερη μείωση επιτυγχάνουν οι μέθοδοι CBE και FISH. Η μέθοδος FISH είναι γρηγορότερη μεταξύ των δυο μεθόδων αλλά παρουσιάζει πολύ χειρότερη ακρίβεια από την μέθοδο CBE (εικόνα 3.49). Τέλος ο NEFCS-SRR είναι εξίσου ενδιαφέρουσα επιλογή διότι επιτυγχάνει ικανοποιητική μείωση, είναι ακριβής και είναι ταχύτερος από τον CBE (εικόνα 3.50). Όσον αφορά την ακρίβεια όπως θα δούμε στην συνέχεια δεν μπορούμε να υποθέσουμε ότι ο CBE αποδίδει καλύτερα απο τον NEFCS-SRR.

Πίνακας 3.13: Κατά ζεύγη συγκρίσεις με χρήση του Wilcoxon signed rank test

	CBE	FISH	ICF	NEFCS-SRR
FISH	0,03	-	-	-
ICF	0,03	0,03	-	-
NEFCS-SRR	0,21	0,03	0,03	-
kNN	0,03	0,03	0,03	0,053

Ο πίνακας 3.13 μας δίνει τα αποτελέσματα (p-value) των κατά ζεύγη μη παραμετρικών ελέγχων που πραγματοποιήθηκαν με την βοήθεια του Wilcoxon signed rank test. Παρατηρούμε ότι όπως και στην περίπτωση που δεν είχαμε drift ότι ο kNN αποδίδει καλύτερα από όλες τις μεθόδους πλην της NEFCS-SRR όπου δεν μπορούμε να απορρίψουμε την μηδενική υπόθεση ότι ο kNN έχει χαμηλότερη ακρίβεια από τον NEFCS-SRR. Επιπλέον δεν μπορούμε να ισχυριστούμε ότι ο NEFCS-SRR έχει καλύτερη ακρίβεια από τον CBE. Η μόνη διαφορά σε σχέση με την περίπτωση που δεν υπήρχε drift παρατηρείται μεταξύ των μεθόδων ICF και FISH. Η μηδενική υπόθεση που κάνουμε μεταξύ αυτών είναι η εξής:

H_0 : η μέθοδος FISH έχει χαμηλότερη ακρίβεια από την μέθοδο ICF

Παρατηρούμε ότι η p-τιμή που δίνει ο μη παραμετρικός έλεγχος Wilcoxon είναι $p = 0,03 < 0,05$ επομένως απορρίπτεται η μηδενική υπόθεση και ενεργοποιείται η εναλλακτική γεγονός που μας επιτρέπει να υποθέσουμε ότι η μέθοδος FISH έχει καλύτερη ακρίβεια από την ICF στα δεδομένα όταν παρουσιάζεται το φαινόμενο drift κάτι το οποίο δεν μπορούσαμε να υποθέσουμε όταν δεν υπήρχε drift στα δεδομένα.

Πίνακας 3.14: Friedman rank sum test

Friedman chi-squared	18,08
df	4
p-value	0,00119

Τέλος εκτελώντας τον μη παραμετρικό έλεγχο Friedman απορρίπτουμε την μηδενική υπόθεση ότι οι μέθοδοι έχουν την ίδια ακρίβεια.

ΚΕΦΑΛΑΙΟ 4. Επίλογος

1.Σύνοψη και συμπεράσματα

Ολοκληρώνοντας την διπλωματική εργασία είναι σημαντικό να αναφερθούν κάποια χρήσιμα συμπεράσματα και αποτελέσματα που εξήχθησαν από τις πειραματικές μετρήσεις. Τα αποτελέσματα αυτά απαριθμούνται παρακάτω:

- Όσον αφορά την ακρίβεια και στην περίπτωση που εμφανίζεται drift αλλά και όταν δεν υπάρχει την καλύτερη επίδοση την επιτυγχάνει ο kNN. Σύμφωνα με τους μη παραμετρικούς ελέγχους όμως αλλά και με την βοήθεια των γραφημάτων μπορούμε να συνάγουμε το συμπέρασμα ότι οι μέθοδοι CBE και NEFCS-SRR αποδίδουν εξίσου ικανοποιητικά.

- Σχετικά με τον χρόνο που απαιτούνε οι μέθοδοι μείωσης και στις δύο περιπτώσεις φαίνεται να αποδίδουν καλύτερα οι FISH και NEFCS-SRR. Δεν συγκρίνουμε τις μεθόδους μείωσης με τον kNN όσον αφορά την παράμετρο του χρόνου διότι ο kNN δεν εκτελεί κάποια μείωση επομένως δεν χρειάζεται επιπλέον χρόνο.

- Οι αλγόριθμοι FISH, CBE και NEFCS-SRR επιτυγχάνουν πολύ μεγάλη μείωση του συνόλου δεδομένων. Λαμβάνοντας υπόψη και την παράμετρο της ακρίβειας μπορούμε να επιλέξουμε τον CBE ως την καλύτερη επιλογή. Βεβαίως και ο NEFCS-SRR εμφανίζεται ως μια ενδιαφέρουσα επιλογή καθότι τα αποτελέσματα του όσον αφορά την ακρίβεια και την μείωση είναι αρκετά όμοια με αυτά του CBE.

- Σχετικά με την αντίδραση των μεθόδων στην παρουσία του concept drift ξανά αυτοί που φαίνεται να αποδίδουν ικανοποιητικά ρίχνοντας λίγο την ακρίβεια τους είναι οι NEFCS-SRR και CBE. Οι FISH και ICF αντιδρούν ικανοποιητικά σε 3 από τα 4 σύνολα που παρουσιάζουν concept drift. Συγκεκριμένα στα STAGGER η ακρίβεια τους πέφτει δραματικά. Από τους

τρεις καλύτερους αλγορίθμους ο CBE φαίνεται ως η καλύτερη επιλογή όταν εμφανίζεται ένα σταδιακό drift.

- Το κύριο συμπέρασμα που μας απασχολεί είναι ότι φαίνεται οι αλγόριθμοι που βασίζονται στην ικανότητα (competence-based) να διατηρούν βάσεις περιπτώσεων οι οποίες είναι καθαρές από θορυβώδη ή περιττά στιγμιότυπα στον μέγιστο βαθμό. Ωστόσο όλες οι μέθοδοι που εξετάστηκαν χαρακτηρίζονται από πολύ υψηλό υπολογιστικό κόστος. Με βάση τις μετρήσεις μπορούμε με ασφάλεια να πούμε ότι όλες οι μέθοδοι επιλογής στιγμιοτύπων επιδεικνύουν μια αρνητική συμπεριφορά όσον αφορά τις απαιτήσεις τους σε χρόνο και μνήμη. Αυτό το γεγονός τους περιορίζει και τους αποτρέπει στην τωρινή τους κατάσταση να εφαρμοστούν σε ουσιαστικές εφαρμογές για εξόρυξη γνώσης από δεδομένα ροής όπως για παράδειγμα το network intrusion όπου ένας τεράστιος όγκος δεδομένων καταφθάνει κάθε δευτερόλεπτο.

Καταλήγοντας επισημαίνουμε ότι όταν επιλέγουμε μεθόδους προεπεξεργασίας για εξόρυξη γνώσης από δεδομένα ροής πρέπει να λάβουμε υπόψη όχι μόνο την ακρίβεια αλλά και το υπολογιστικό κόστος που δημιουργείται από κάθε μέθοδο ξεχωριστά. Όπως φάνηκε από την έρευνα όλες οι μέθοδοι ανεξαρτήτως απόδοσης παρουσίασαν απαιτήσεις σε CPU και μνήμη που χαρακτηρίζονται ως απαγορευτικές για την εξόρυξη γνώσης σε δεδομένα ροής.

2.Όρια και περιορισμοί της έρευνας

Η εργασία χρησιμοποίησε για την μεθοδολογία μόνο 4 σύνολα δεδομένων ροής. Ωστόσο αυτό έγινε με σκοπό να υπάρξει αναλυτική παρουσίαση της συμπεριφοράς των μεθόδων σε κάθε σύνολο χωριστά. Προφανώς η χρήση περισσότερων συνόλων θα έδινε μια ποιο γενική εικόνα για την συμπεριφορά και τις αποδόσεις των μεθόδων μείωσης. Επιπρόσθετα κατά την πειραματική μελέτη επιλέχθηκε τα στιγμιότυπα να καταφθάνουν ένα ένα και όχι ως πακέτα (batches). Αυτό έγινε διότι παρατηρήθηκε ακόμα μεγαλύτερη υπολογιστική πολυπλοκότητα σε κάποιες περιπτώσεις. Ακόμη για την σύγκριση των μεθόδων επιλέχθηκε η μείωση στιγμιοτύπων να ενεργοποιείται ανά 500 στιγμιότυπα. Αυτό

έγινε για δύο λόγους, πρώτον για να υπάρχει μια σωστή σύγκριση για τις μεθόδους εφόσον κάθε μέθοδος ενεργοποίησε τον μηχανισμό μείωσης το ίδιο πλήθος φορών. Δεύτερον αυτή η τιμή είναι αυτή που προτείνεται ως default από τους δημιουργούς των μεθόδων. Όπως όμως είδαμε οι μέθοδοι για αυτή την τιμή χάνουν πολύ σε ακρίβεια και απαιτούν πολύ χρόνο σε σχέση με άλλες περιπτώσεις που οι μείωση εφαρμόζεται πιο λίγες φορές.

3.Μελλοντικές Επεκτάσεις

Με βάση την συμπερασματολογία είναι φανερό ότι η κύρια πρόκληση που πρέπει να αντιμετωπιστεί είναι η υψηλή υπολογιστική πολυπλοκότητα που παρουσιάζουν οι μέθοδοι επιλογής στιγμιοτύπων. Υπάρχει ανάγκη οι υπάρχουσες μέθοδοι να βελτιωθούν δίνοντας έμφαση σε αυτές που είναι πιο πολλά υποσχόμενες όπως ο CBE και ο NEFCS-SRR έτσι ώστε να αποκτήσουν χαμηλές υπολογιστικές απαιτήσεις κάτι που θα τους επιτρέψει να εφαρμοστούν σε δεδομένα ροής υψηλών ταχυτήτων και για πολύ μεγάλο πλήθος δεδομένων. Είναι γεγονός ότι οι μέθοδοι που βασίζονται στην ικανότητα είναι προτιμότεροι από τις μεθόδους που απορρίπτουν αδιακρίτως όλες τις νέες περιπτώσεις. Η διατήρηση όμως νέων περιπτώσεων μεγαλώνει την βάση περιπτώσεων (case-base) και η σύγκριση με κάθε νέα περίπτωση αυξάνει το υπολογιστικό κόστος. Μια κατεύθυνση που θα μπορούσε να ακολουθήσει η έρευνα για να χαμηλώσει το υπολογιστικό κόστος των μεθόδων αυτών είναι να εφαρμοστεί μια πολιτική με βάρη όπου θα εξισορροπείται καλύτερα η απόφαση για το κατά πόσο μια περίπτωση αποτελεί θόρυβο ή νέο concept. Επιπρόσθετα υπάρχει η ανάγκη για επιπλέον έρευνα στον τρόπο αντιμετώπισης του concept drift από τις μεθόδους επιλογής στιγμιοτύπων. Ένας τρόπος που θα μπορούσε να επιτευχθεί αυτό είναι ο συνδυασμός των μεθόδων αυτών με εντοπιστές των drift. Τέλος σημασία πρέπει να δοθεί και στην δημιουργία μεθόδων προεπεξεργασίας για τις υπόλοιπες περιπτώσεις που υπάρχει πρόβλημα στα δεδομένα ροής όπως οι μη ισορροπημένες κλάσεις ή η έλλειψη γνωρισμάτων.

Βιβλιογραφία

Abdulsalam, H., Skillicorn, D. and Martin, P., 2011. Classification Using Streaming Random Forests, IEEE Transactions on Knowledge and Data Engineering, 23(1), pp.22-36.

Aggarwal, C. C., 2015. Data Mining. New York : Springer. pp. 27-29

Aha D.W, Kibler D. and Albert M.K., 1991. Instance-based learning algorithms. In: Machine Learning, pp. 37-66.

Allahyar, A. and Yazdi, H. S., 2014. Online discriminative component analysis feature extraction from stream data with domain knowledge. Intelligent Data Analysis, 18(5), pp. 927-951.

Al-Shiha, A. A. M., Woo, W. L. and Dlay, S. S., 2014. Multi-linear neighborhood preserving projection for face recognition. *Pattern Recognition*, *47*(2), pp. 544-555.

Amayri, O. and Bouguila, N., 2013. On online high-dimensional spherical data clustering and feature selection. Engineering Applications of Artificial Intelligence, 26(4), pp.1386-1398.

Androutsopoulos, I., Koutsias, J., Konstantinos, V., Chandrinos, V., Paliouras, G. & Spyropoulos, C., 2000. An evaluation of Naive Bayesian anti- spam filtering. In: G. Potamias, V. Moustakis, M. van Someren (Eds.), Proceedings of the ECML 2000 Workshop on Machine Learning in the New Information Age, pp. 9–17.

Androutsopoulos, I., Paliouras, G. and Michelakis, E., 2004. Learning to Filter Unsolicited Commercial E-Mail. [pdf] Athens: NCSR Demokritos. Available at: < http://nlp.cs.aueb.gr/pubs/TR2004_updated.pdf > [Accessed 15 July 2020].

Angiulli, F., 2007. Fast nearest neighbor condensation for large data sets classification, IEEE Trans. Knowl. Data Eng. 19 (11), pp.1450–1464.

Angiulli, F. and Folino, G., 2007. Distributed nearest neighbor-based condensation of very large data sets. *Knowledge and Data Engineering, IEEE Transactions on*, 19(12), pp.1593–1606.

Asuncion, A. and Newman, D. J., 2007. UCI Machine Learning Repository. [online] Available at:< https://archive.ics.uci.edu/ml/datasets.php>[Accessed 15 July 2020].

Baena-Garcia, M., Campo-Avila, J., Fidalgo, R., Bifet, A., Gavalda, R. and Morales-Bueno R., 2006. Early drift detection method. In: *Fourth International Workshop on Knowledge Discovery from Data Streams.*

Barddal, J. P., 2019. *Feature analysis in evolving data streams: Issues and algorithms.* Ph.D. thesis, Pontifícia Universidade Católica do Paraná (PUCPR), Brazil.

Barddal, J. P., Gomes, H. M., Enembreck, F. and Pfahringer, B., 2017. A survey on feature drift adaptation: Definition, benchmark, challenges and future directions. Journal of Systems and Software, 127, pp. 278-294.

Barddal, J. P., Gomes, H. M., Enembreck, F., Pfahringer, B. and Bifet, A. (2016, September). On dynamic feature weighting for feature drifting data streams. In: *Joint European conference on machine learning and knowledge discovery in databases* (pp. 129-144). Springer, Cham.

Beringer, J. and Hüllermeier, E., 2007. Efficient instance-based learning on data streams. Intelligent Data Analysis, 11(6), pp.627-650.

Bifet, A., 2010. *Adaptive Stream Mining: Pattern Learning and Mining from Evolving Data Streams. Frontiers in Artificial Intelligence and Applications,* 207, pp. 1-212.

Bifet, A., Frank, E., Holmes, G. and Pfahringer, B., 2010. Accurate ensembles for data streams: Combining restricted Hoeffding trees using stacking. In: 2nd Asian Conference on Machine Learning in Journal of Machine Learning Research, 13. Tokyo.

Bifet, A., Holmes, G. and Pfahringer, B. (2010, September). Leveraging bagging for evolving data streams. In: Joint European conference on machine learning and knowledge discovery in databases (pp. 135-150). Springer, Berlin, Heidelberg.

Bifet, A., Holmes, G. and Pfahringer, B., Kirkby, R., Gavalda, R., 2009. New ensemble methods for evolving data streams. In: *15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, Paris, France, pp. 139-149.

Bifet, A. and Kirkby, R., 2009. Data Stream Mining A Practical Approach. [pdf] Hamilton, New Zealand: University of Waikato. Available at: <https://sourceforge.net/projects/moa-datastream/files/documentation/StreamMining.pdf/download > [Accessed 20 June 2018].

Blake, C. and Merz, C., 1998. UCI repository of machine learning databases.

Bolon-Canedo, V., Fernández-Francos, D., Peteiro-Barral, D., Alonso-Betanzos, A., Guijarro-Berdiñas, B. and Sánchez-Maroño, N., 2016. A unified pipeline for online feature selection and classification. Expert Systems with Applications, 55, pp.532-545.

Bolón-Canedo, V., Sánchez-Maroño, N. and Alonso-Betanzos, A., 2015. Recent advances and emerging challenges of feature selection in the context of big data. Knowledge-Based Systems, 86, pp.33-45.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. J., 1984. *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA.

Brighton, H., 1996. Experiments in case-based learning. Undergraduate Dissertation, Department of Artificial Intelligence, University of Edinburgh, Scotland.

Brighton, H., 1997. Information filtering for lazy learning algorithms. Masters Thesis, Centre for Cognitive Science, University of Edinburgh, Scotland.

Brighton, H. and Mellish, C., 2002. Advances in instance selection for instance-based learning algorithms. Data mining and knowledge discovery, 6(2), pp.153-172.

Brzezinski, D. and Stefanowski, J. (2014, September). Prequential AUC for classifier evaluation and drift detection in evolving data streams. In: International Workshop on New Frontiers in Mining Complex Patterns (pp. 87-101). Springer, Cham.

Cai, Y., Sun, Y., Li, J. and Goodison, S. (2009, April). Online Feature Selection Algorithm with Bayesian ℓ 1 Regularization. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 401-413). Springer, Berlin, Heidelberg.

Cano, J. R., Herrera, F. and Lozano, M. (2005a).Stratification for scaling up evolutionary prototype selection. *Pattern Recognition Letters*,26(7), pp.953 – 963.

Cano, J.R., Herrera, F. and Lozano, M. (2005b). A Study on the Combination of Evolutionary Algorithms and Stratified Strategies for Training Set Selection in Data Mining, (pp. 271–284).Springer Berlin Heidelberg, Berlin, Heidelberg.

Cano, J.R., Herrera, F. and Lozano, M., 2007. Evolutionary stratified training set selection for extracting classification rules with trade off precision-interpretability. In: *Data & Knowledge Engineering,* 60(1), pp. 90-108.

Cano, A., Luna, J. M., Gibaja, E. L. and Ventura, S., 2016. LAIM discretization for multi-label data. Information Sciences, 330, pp.370-384.

Cano, A., Nguyen, D. T., Ventura, S. and Cios, K. J., 2016. ur-CAIM: improved CAIM discretization for unbalanced and balanced data. Soft Computing, 20(1), pp. 173-188.

Carreras, X. and Marquez, L., 2001. Boosting trees for anti-spam email filtering. In: Proceedings of 4th International Conference on Recent Advances in Natural Language Processing, Tzigov Chark, Bulgaria.

Carvalho, V.R. and Cohen, W.W. (2006, August). Single-pass online learning: Performance, voting schemes and online feature selection. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 548-553).

Chawla, N.V., Bowyer, L. O., Hall and Kegelmeyer, W. P., 2002. SMOTE: Synthetic Minority Oversampling Technique. Journal of Artificial Intelligence, 16, pp. 321-357.

Chawla, N.V., Lazarevic, A., Hall, L. O. and Bowyer, K. W., 2003. Improving Prediction of the Minority Class in Boosting. In: *7th European Conference on Principles and Practise of Knowledge Discovery in Databases (PKDD)* (pp.107-119). Dubrovnik, Croatia.

Chen, S. and He, H., 2009. Selectively recursive approach towards non stationary imbalanced stream data mining. In: *International Joint Conference on Neural Networks* (IJCNN 2009) (pp.522-529). Atlanta, GA.

Chou, C.-H., Kuo, B.-H. and Chang, F., 2006. The generalized condensed nearest neighbor rule as a data reduction method. In: 1*8th International Conference on Pattern Recognition, ICPR '06*, Hong Kong, China, Aug. 20–24, pp. 556–559.

Chu, F. and Zaniolo, C., 2004. Fast and light boosting for adaptive mining of data streams. In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 282–292

Cover, T. and Hart, P., 1967. Nearest neighbor pattern classification. IEEE transactions on information theory, 13(1), pp.21-27.

Czarnecki, W. M. and Tabor, J., 2015. Online extreme entropy machines for streams classification and active learning. In: *Proceedings of the 9th International Conference on Computer Recognition Systems CORES 2015*, (pp. 371–381).Wroclaw, Poland.

Daelemans, W., van den Bosch, A. and Zavrel, J., 1999. Forgetting exceptions is harmful in language learning. Machine Learning, 34(1/3), pp.11–41.

de Haro-García, A. and García-Pedrajas, N., 2009. A divide-and-conquer recursive approach for scaling up instance selection algorithms. *Data Mining and Knowledge Discovery*, 18(3), pp.392–418.

Delany, S.J. and Cunningham, P., 2004. An analysis of case-base editing in a spam filtering system. In: *7th European Conference on Case Based Reasoning, ECCBR'04*, Madrid, Spain, Aug. 30–Sep. 2, pp. 128–141.

Delany, S.J., Cunningham, P. and Coyle, L., 2004. An assessment of case-base reasoning for spam filtering. In: L. *McGinty, B. Crean (Eds.), Proceedings of 15th Artificial Intelligence and Cognitive Science Conference.*

Delany, S.J., Cunningham, P., Tsymbal, A. and Coyle, L. (2004, December). A case-based technique for tracking concept drift in spam filtering. In: International Conference on Innovative Techniques and Applications of Artificial Intelligence (pp. 3-16). Springer, London.

Delany, S.J., Cunningham, P., Tsymbal, A. and Coyle, L., 2005. A case-based technique for tracking concept drift in spam filtering, Knowl.-Based Syst. 18 (4–5), pp. 187–195.

Dempster, A. P., Laird, N. M. and Rubin D. B., 1977. Maximum Likelihood from Incomplete Data via the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological), 39(1), pp. 1-22.

Derrac, J., Garcia, S. and Herrera, F., 2010. Instance and feature selection based on cooperative coevolution with nearest neighbour rule. Pattern Recognition, 43, pp. 2082-2105.

Devi, V.S. and Murty, M.N., 2002. An incremental prototype set building technique, Pattern Recognit., 35 (2), pp. 505–513.

Diettrich, T.G., 1998. Approximate statistical test for comparing supervised classification learning algorithms. *Neural Computation*, 10(7), pp.1895–1923.

Ditzler, G. and Polikar, R., 2013. Incremental Learning of Concept Drift from Streaming Imbalanced Data. Knowledge and Data Engineering, IEEE Transactions on. 25., pp. 2283-2301.

Ditzler, G., Poolikar, R. and Chawla, N., 2010. An Incremental learning Algorithm for Non-stationary Environments and Class Imbalance. In: *20th International Conference on Pattern Recognition* (ICPR 2010) (pp.2997-3000). Istanbul,Turkey.

Ditzler, G., Roveri, M., Alippi C. and Polikar, R., 2015. Learning in nonstationary environments: A survey. Computational Intelligence Magazine, IEEE, 10(4), pp. 12-25.

Domingo, C., Gavalda, R. and Watanabe, O., 2000. Practical Algorithms for Online Selection. In: *Proceedings of the first International Conference on Discovery Science.*

Domingos, P. and Hulten, G. (2000, August). Mining high-speed data streams. In Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining , pp. 71-80.

Doquire, G. and Verleysen, M., 2012. Feature selection with missing data using mutual information estimators. Neurocomputing, 90, pp. 3-11.

Drucker, H.D., Wu, D. and Vapnik, V., 1999. Support vector machines for spam categorization, IEEE Transactions On Neural Networks, 10(5), pp.1048-1-54.

Dyer, K.B., Capo, R. and Polikar, R., 2013. Compose: A semisupervised learning framework for initially labeled nonstationary streaming data. IEEE transactions on neural networks and learning systems, 25(1), pp.12-26.

Eskandari, S. and Javidi, M. M., 2016. Online streaming feature selection using rough sets. International Journal of Approximate Reasoning, 69, pp.35-57.

Fan, W. and Bouguila, N. (2012, November). Online learning of a dirichlet process mixture of generalized dirichlet distributions for simultaneous clustering and localized feature selection. In Asian Conference on Machine Learning, pp. 113-128.

Fan, W. and Bouguila, N., 2014. Online variational learning of generalized Dirichlet mixture models with feature selection. Neurocomputing, 126, pp.166-179.

Ferreira, A. J. and Figueiredo, M. A., 2014. Incremental filter and wrapper approaches for feature discretization. Neurocomputing, 123, pp.60-74.

Ferri, F.J. and Vidal, E., 1992. Small sample size effects in the use of editing techniques. In: *11th IAPR International Conference on Pattern Recognition, ICPR '92*, Hague, Netherlands, Aug. 30–Sep. 3, pp. 607–610.

Galan, M, Liu, H and Torkkola, K., 2005. Intelligent instance selection of data streams for smart sensor applications. In: *KL Priddy (ed.), Proceedings of SPIE - The International Society for Optical Engineering*. vol. 5803, 14, pp. 108-119, Intelligent Computing: Theory and Applications III, Orlando, FL, United States, 3/28/05.

Gama, J., 2010. Knowledge Discovery from Data Streams. London: Chapman & Hall/ CRC.

Gama, J., Medas, P., Castillo, G. and Rodrigues, P., 2004. Learning with drift detection. In: *17th Brazilian Symposium on Artificial Intelligence, SBIA '04*, Sao Luis, Maranhao, Brazil, Sep. 29–Oct. 1, pp. 286–295.

Garcia, S., Derrac, J., Cano, J. and Herrera, F., 2012. Prototype selection for nearest neighbor classification: Taxonomy and empirical study. IEEE transactions on pattern analysis and machine intelligence, 34(3), pp.417-435.

García-Osorio, C., de Haro-García, A. and García-Pedrajas, N., 2010. Democratic instance selection: A linear complexity instance selection algorithm based on classifier ensemble concepts.*Artificial Intelligence*, 174(5-6), pp.410–441.

García, S., Luengo, J. and Herrera, F., 2015. *Data preprocessing in data mining*. Cham, Switzerland: Springer International Publishing, pp.195-243.

García, S., Luengo, J. and Herrera, F., 2016. Tutorial on practical tips of the most influential data preprocessing algorithms in data mining. *Knowledge-Based Systems*, *98*, pp.1-29.

Gao, J., Ding, B., Wei, F., Jiawei, H. and Yu, P. S., 2008. Classifying data streams with skewed class distributions and concept drifts. IEEE Internet Computing, 12(6), pp.37-49.

Gates, G.W., 1972. The reduced nearest neighbour rule, IEEE Trans. Inf. Theory, 18(3), pp.431–433.

Gee, K.R., 2003. Using Latent Semantic Indexing to Filter Spam, in: Proceedings of the 2003 ACM Symposium on Applied Computing (SAC), pp. 460-464.

Ghashami, M., Perry, D. J. and Phillips, J. (2016, May). Streaming kernel principal component analysis. In Artificial intelligence and statistics ,pp. 1365-1374.

Gomes, J.B., Gaber, M. M., Sousa, P. A. and Menasalvas, E., 2013. Mining recurring concepts in a dynamic feature space. IEEE Transactions on Neural Networks and Learning Systems, 25(1), pp.95-110.

Gonzalez, A.A., 2018. Estudio de Métodos de Selección de Instancias. Ph.D. Thesis. University of Burgos , Spain.

Guha, S., 2009. Tight results for clustering and summarising data streams. *In: Proceedings of the 12th International Conference on Database Theory* (pp. 268-275). New York,NY, USA: ACM.

Guha, S. and McGregor, A., 2009. Stream order and order statistics: Quantile estimation in random-order streams. SIAM Journal on Computing, 38(5), pp.2044-2059.

Guo, H. and Viktor, H., 2004. Learning from Imbalanced Data sets with Boosting and Data Generation: The DataBoost-IM Approach. ACM SIGKDD Explorations Newsletter, 6(1), pp.30-39

Gupta, A. and Zane, F. X. (2003, January). Counting inversions in lists. In Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms (pp. 253-254). Society for Industrial and Applied Mathematics.

Gustafsson, F., 2000. Adaptive Filtering and Change Detection. Wiley.

Hammoodi, M., Stahl, F. and Tennant, M. (2016, August). Towards online concept drift detection with feature selection for data stream classification. In Proceedings of the Twenty-second European Conference on Artificial Intelligence (pp. 1549-1550). IOS Press.

Hao, X., Zhang, C., Xu, H., Tao, X., Wang, S. and Hu, Y., 2008. An improved condensing algorithm, In: *7th IEEE/ACIS International Conference on Computer and Information Science, ICIS'08*, Portland, OR, USA, May 14–16, pp. 316–321.

Hart, P., 1968. The condensed nearest neighbor rule (corresp.). *Information Theory, IEEE Transactions on*, 14(3), pp.515 – 516.

Hazan, E., Kale, S. and Warmuth, M. K. (2010, June). On-line variance minimization in O (n2) per trial?. In COLT ,pp. 314-315.

He, H. and Garcia, E. A. (2009,September). Learning from Imbalanced Data. IEEE Transactions of Information Theory, 14(3), pp.1263-1284.

Hettich, S. and Bay, S. D., 1999. UCI KDD Archive .[online] Available at: <http://kdd.ics.uci.edu> [Accessed 15 July 2020].

Hu, H.W., Chen, Y.L. and Tang, K., 2009. A dynamic discretization approach for constructing decision trees with a continuous label. IEEE Transactions on knowledge and data engineering, 21(11), pp.1505-1514.

Jain, P., Jin, C., Kakade, S. M., Netrapalli, P. and Sidford, A. (2016, June). Streaming PCA: Matching matrix Bernstein and near-optimal finite sample guarantees for Oja's algorithm. In: Conference on learning theory ,pp. 1147-1164.

Jankowski, N. and Grochowski, M., 2004. Comparison of instances seletion algorithms I. algorithms survey.In Rutkowski, L., Siekmann, J., Tadeusiewicz, R., and Zadeh, L. A., editors, *Artificial Intelligence and Soft Computing - ICAISC 2004*, volume 3070 of *Lecture Notes in Computer Science*, pp.598–603. Springer Berlin Heidelberg.

Jolliffe, I.T. and Cadima, J., 2016. Principal Component Analysis: A Review and Recent Developments. Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences, 374.

Joseph, A.A., Tokumoto, T. and Ozawa, S., 2016. Online feature extraction based on accelerated kernel principal component analysis for data stream. Evolving Systems, 7(1), pp.15-27.

Juszczak, P. and Duin, R., 2004. Combining One-Class Classifiers to Classify Missing Data, pp.92-101.

Katakis, I., Tsoumakas, G. and Vlahavas, I. (2005, November). On the utility of incremental feature selection for the classification of textual data streams. In: Panhellenic Conference on Informatics (pp. 338-348). Springer, Berlin, Heidelberg.

Kifer, D., Ben-David, S. and Gerhrke, J., 2004. Detecting change in data streams. In: *Proc. 30th VLDB Conference*.Toronto, Canada.

King, R.D., Feng, C. and Sutherland, A., 1995. Statlog: Comparison of classification algorithms on large real-world problems. Applied Artificial Intelligence, *9*(3), pp. 289–333.

Kolodner, J.L., 1993. Case-Based Reasoning. San Mateo, CA: Morgan Kaufmann.

Kolter, J.Z. and Maloof, M. A., 2007. Dynamic weighted majority: An ensemble method for drifting concepts. The Journal of Machine Learning Research, 8, pp.2755–2790.

Klinkenberg, R., 2004. Learning drifting concepts: Example selection vs. example weighting. Intelligent data analysis, 8(3), pp.281-300.

Klinkenberg, R. and Joachims, T., 2000. Detecting concept drift with support vector machines. In: ICML ’00: Proceedings of the 17th international conference on Machine Learning, Morgan Kaufmann Publishers Inc., pp. 487–494.

Kubat, M., Holte, R. and Matwin, S., 1998. ‘Machine Learning for the Detection of Oil Spills in Satellite Radar Images’, *Machine Learning*, 30, pp.195-215.

Kulkarni, P. and Ade, R., 2014. Incremental Learning From Unbalanced Data with Concept Class, Concept Drift and Missing Features: A Review. *International Journal of Data Mining & Knowledge Management Process (IJDKP),* 4(6), pp. 17-25.

Kuncheva, L.I. (2008, July). Classifier ensembles for detecting concept change in streaming data: Overview and perspectives. In: 2nd Workshop SUEMA, pp. 5-10.

Kuncheva, L. I. and Faithfull, W. J., 2012. PCA feature extraction for change detection in multidimensional unlabeled data. IEEE transactions on neural networks and learning systems, 25(1), pp.69-80.

Lehtinen, P., Saarela, M. and Elomaa, T., 2012. Online chimerge algorithm. In Data Mining: Foundations and Intelligent Paradigms (pp. 199-216). Springer, Berlin, Heidelberg.

Lenz, M., Auriol, E. and Manago, M., 1998. Diagnosis and decision support, in *M. Bartsch-Sporl, S. Wess (Eds.), Case-Based Reasoning Technology: From Foundations to Applications, LNCS 104*, Springer, Berlin.

Li, C., 2007. Classifying imbalanced data using a bagging ensemble variation (BEV). *In: ACM Southeast Regional Conference* (pp.203-208).Winston-Salem,NC

Li, W. P., Yang, J. and Zhang, J. P., 2015. Uncertain canonical correlation analysis for multi-view feature extraction from uncertain data streams. Neurocomputing, 149, pp.1337-1347.

Littlestone, N. and Warmuth, M., 1994. The Weighted Majority Algorithm. Information and Computation, 108(2), pp. 212-261

Liu, H. and Motoda, H., 2002. On issues of instance selection. Data Mining and Knowledge Discovery, 6(2), pp.115-130.

López, V., Triguero, I., Carmona, C. J., García, S. and Herrera, F., 2014. Addressing imbalanced classification with instance generation techniques: IPADE-ID. Neurocomputing, 126, pp.15-28.

Lopez de Mantaras, R., McSherry, D., Bridge, D., Leake, D., Smyth, B., Craw. S., Faltings, B., Maher, M. L., Cox, M. T., Forbus, K., Keane, M., Aamodt, A. and

Watson, I., 2005. Retrieval reuse, revision and retention in case-based reasoning. Knowledge Engineering Revision, 20(3), pp.215-240.

Lu, N., Lu, J., Zhang, G. and De Mantaras, R. L., 2016. A concept drift-tolerant case-base editing technique. Artificial Intelligence, 230, pp.108-133.

Lu, J., Yang, Y. and Webb, G. I. (2006, August). Incremental discretization for naive-bayes classifier. In International Conference on Advanced Data Mining and Applications (pp. 223-238). Springer, Berlin, Heidelberg.

Lu, N., Zhang, G. and Lu, J., 2014. Concept drift detection via competence models, Artificial Intelligence,209, pp.11–28.

Masud, M. M., Chen, Q., Gao, J., Khan, L., Han, J. and Thuraisingham, B. (2010, September). Classification and novel class detection of data streams in a dynamic feature space. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 337-352). Springer, Berlin, Heidelberg.

Melville, P. and Mooney, R.J., 2005. Creating diversity in ensembles using artificial data. Information Fusion, 6, pp.99-111.

Mena-Torres, D. and Aguilar-Ruiz, J. S., 2014. A similarity-based approach for data stream classification. Expert Systems with Applications, 41(9), pp.4224-4234.

Minku, L. L., White, A. P. and Yao, X., 2009. The impact of diversity on online ensemble learning in the presence of concept drift. IEEE Transactions on knowledge and Data Engineering, 22(5), pp.730-742.

Muhlbaier, M., Topalis, A. and Polikar, R., 2004. Incremental learning from unbalanced data. In: *Proc. of Int. Joint Conference on Neural Networks (IJCNN 2004)* (pp. 1057-1062). Budapest, Hungary.

Narasimhamurthy, A. and Kucheva L. I., 2007. A framework for generating data to simulate changing environments. In: *AIAP'07*, pp. 384–389.

Nguyen, H. L., Woon, Y. K., Ng, W. K. and Wan, L. (2012, May). Heterogeneous ensemble for feature drifts in data streams. In: Pacific-Asia conference on knowledge discovery and data mining (pp. 1-12). Springer, Berlin, Heidelberg.

Nie, J., Kotłowski, W. and Warmuth, M. K., 2016. Online PCA with optimal regret. The Journal of Machine Learning Research, 17(1), pp.6022-6070.

Olvera-López, J. A., Martínez-Trinidad, F. J., Carrasco-Ochoa, J. A. and Kittler, J. (2009). Prototype selection based on sequential search.*Intelligent Data Analysis*, 13(4), pp.599–631.

Ooi, K. and Ninomiya, T., 2013. Efficient Online Feature Selection based on ℓ1-Regularized Logistic Regression. In: ICAART (2) ,pp. 277-282.

Oza, N. C. and Russel, S. (2001a). Experimental comparisons of online and batch versions of bagging and boosting. In: *International Conference on Knowledge Discovery and Data Mining*, pp. 359–364.

Oza, N. C. and Russel, S. (2001b). Online bagging and boosting. In: *Artificial Intelligence and Statistics*, pp. 105-112.

Page, E. S., 1954. Continuous inspection schemes. *Biometrika*, 41(1/2), pp.100-115.

Polikar, R., Upda, L., Upda, S. and Honavar, V., 2001. Learn++: an incremental learning algorithm for supervised neural networks. *IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews)*, 31(4), pp.497-508.

Qahtan, A. A., Alharbi, B., Wang, S. and Zhang, X. (2015, August). A pca-based change detection framework for multidimensional data streams: Change detection in multidimensional data streams. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 935-944.

Qu, W., Zhang, Y., Zhu, J., Qiu, Q., 2009. Mining multi-label concept-drifting data streams using dynamic classifier ensemble. In: 1st Asian Conference on Machine Learning, ACML ’09, Nanjing, China, pp.308-321.

Quinlan, J. R., 1993. C4.5 Programs for Machine Learning, Morgan Kaufmann Publishers, San Mateo, CA.

Ramírez-Gallego, S., Krawczyk, B., García, S., Woźniak, M. and Herrera, F., 2017. A survey on data preprocessing for data stream mining: Current status and future directions. Neurocomputing, 239, pp.39-57.

Rico-Juan, J. R. and Iñesta, J. M., 2012. New rank methods for reducing the size of the training set using the nearest neighbor rule. *Pattern Recognition Letters*, 33(5), pp.654 – 660.

Ritter, G.,Woodruff, H., Lowry, S. and Isenhour, T., 1975. An algorithm for a selective nearest neighbor decision rule, IEEE Trans. Inf. Theory, 21 (6), pp.665–669.

Roberts, S. W., 2000. Control chart tests based on geometric moving averages. *Technometrics*, 42(1), pp.97-101.

Roy, A. (2015, July). Automated online feature selection and learning from high-dimensional streaming data using an ensemble of Kohonen neurons. In: 2015 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.

Sakkis, G., Andoutsopoulos, I., Paliouras, G., Karkaletsis, V. and Spyropoulos, C., 2001. Stacking classifiers for anti-spam filtering of e-mail. In: *Lee, Harman (Eds.), Proceedings of the 6th Conference on Empirical Methods in Natural Language Processing*, pp.44–50.

Sakkis, G., Andoutsopoulos, I., Paliouras, G., Karkaletsis, V., Spyropoulos, C. and Stamatopoulos, P., 2004. A memory-based approach to anti-spam filtering for mailing lists, Information Retrieval ,6(1), pp.49–73.

Salganicoff, M. (1993, December). Density-adaptive learning and forgetting. In: Proceedings of the Tenth International Conference on Machine Learning, pp. 276-283.

Salganicoff, M., 1997. Tolerating concept and sampling shift in lazy learning using prediction error context switching. In: Lazy learning (pp. 133-155). Springer, Dordrecht.

Schlimmer, J. C. and Granger, R. H., 1986. Incremental learning from noisy data. *Machine Learning*, 1(3), pp. 317–354.

Shaker, A. and Hüllermeier, E., 2012. IBLStreams: a system for instance-based classification and regression on data streams. Evolving Systems, 3(4), pp. 235-249.

Sheikholeslami, F., Berberidis, D. and Giannakis, G. B. (2015, December). Kernel-based low-rank feature extraction on a budget for big data streams. In: 2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP), pp. 928-932.

Shikkenawis, G. and Mitra, S. K., 2015. 2D orthogonal locality preserving projection for image denoising. IEEE transactions on Image Processing, 25(1), pp.262-273.

Silva, J. A., Faria, E. R., Barros, R. C., Hruschka, E. R., Carvalho, A. C. and Gama, J. , 2013. Data Stream Clustering: A Survey. ACM Computing Surveys, 46(1), pp. 13:1-13-31.

Smyth, B. and Keane, M. T. (1995, August). Remembering to forget. In Proceedings of the 14th international joint conference on Artificial intelligence. Citeseer, (pp. 377-382.

Spira, J., 2003. Spam e-mail and its impact on it spending and productivity. Technical report, Basex Inc.

Stanley, K., 2003. Learning concept drift with a committee of decision trees. Technical Report AI Technical Report 03-302, Department of Computer Science, University of Texas at Austin, Trinity College.

Street, W. N. and Kim, Y., 2001. A streaming ensemble algorithm (SEA) for large-scale classification. In *International Conference on Knowledge Discovery and Data Mining*, pp. 377–382.

Tadeuchi, Y., Oshima, R., Nishida, K., Yamauchi, K. and Omori, T. (2007, October). Quick Online feature selection method for regression-A feature selection method

inspired by human behavior. In 2007 IEEE International Conference on Systems, Man and Cybernetics, pp. 1895-1900.

Tomek, I., 1976. Two Modifications of CNN. IEEE Transactions on Systems, Man and Cybernetics, Part A, 6(11), pp.769-772.

Tsai, C.J., Lee, C.I., Yang, W.P., 2009. Mining decision rules on data streams in the presence of concept drifts. Expert Systems with Applications, 36(2), pp. 1164-1178.

Tsymbal, A., 2004. The problem of concept drift: Definitions and related work. Technical Report TCD-CS-2004-15, Department of Computer Science, University of Dublin, Trinity College.

Tsympal, A., Pechenizkiy, M., Cunningham, P. and Puuronen, S., 2008. Dynamic integration of classifiers for handling concept drift, Information Fusion, 9(1), pp. 56–68.

Vallim, R. M. and De Mello, R. F., 2014. Proposal of a new stability concept to detect changes in unsupervised data streams. *Expert Systems with Applications, 41*(16), pp.7350–7360.

Valero-Mas, J. J., Calvo-Zaragoza, J., Rico-Juan, J. R., and Iñesta, J. M., 2016. An experimental study on rank methods for prototype selection.*Soft Computing*, pp. 1–13.

Wang, H., Fan W., Yu, P. S. and Han, J., 2003. Mining concept-drifting data streams using ensemble classifiers. In: *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* pp.226-235.

Wang, J., Wang, M., Li, P., Liu, L., Zhao, Z., Hu, X., and Wu, X., 2015. Online feature selection with group structure analysis. IEEE Transactions on Knowledge and Data Engineering, 27(11), pp.3029-3041.

Wang, J., Zhao, P., Hoi, S. C. and Jin, R., 2014. Online feature selection and its applications. IEEE Transactions on Knowledge and Data Engineering, 26(3), pp. 698-710.

Webb, G. I. (2014, December). Contrary to popular belief incremental discretization can be sound, computationally efficient and extremely useful for streaming data. In 2014 IEEE International Conference on Data Mining , pp.1031-1036.

Wenerstrom, B. and Giraud-Carrier, C., 2006. Temporal data mining in dynamic feature spaces. In: ICDM ’06: Proceedings of the 6th international conference on Data Mining, IEEE Computer Society, pp. 1141– 1145.

Widmer, G. and Kubat, M., 1996. Learning in the presence of concept drift and hidden contexts, Maching Learning, 23(1), pp.69–101.

Wilson, D. L., 1972. Asymptotic properties of nearest neighbor rules using edited data. *Systems, Man and Cybernetics, IEEE Transactions on*, SMC-2(3), pp.408–421.

Wilson, D., Leake, D.B., 2001. Maintaining case-based reasoners: dimensions and directions, Computational Intelligence, 17(2), pp.196–213.

Wilson, D. R. and Martinez, T. R., 1997. Instance pruning techniques. In: Machine Learning: Proceedings of the fourteenth international conference (ICML'97), pages 404–411. Morgan Kaufmann.

Wilson, D. R. and Martinez, T. R., 2000. Reduction techniques for instance-based learning algorithms. *Machine Learning*, 38(3), pp.257–286.

Wirth, N., 1995. A plea for lean software. *Computer*, 28(2), pp.64-68.

Witten, I. H. and Frank, E., 2005. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, San Francisco, 2nd Edition.

Wozniak, M., Ksieniewicz, P., Cyganek, B. and Walkowiak, K., 2016. Ensembles of Heterogeneous Concept Drift Detectors - Experimental Study. In: 15th IFIP International Conference on Computer Information Systems and Industrial Management (CISIM), Sep 2016, Vilnius, Lithuania., pp.538-549.

Wu, X., Yu, K., Ding, W., Wang, H. and Zhu, X., 2012. Online feature selection with streaming features. IEEE transactions on pattern analysis and machine intelligence, 35(5), pp.1178-1192.

Wu, X., Yu, K., Wang, H. and Ding, W., 2010. Online streaming feature selection.

Xioufis, E. S., Spiliopoulou, M., Tsoumakas, G. and Vlahavas, I., 2011. Dealing with concept drift and class imbalance in multi-label stream classification. In: *International Joint Conference on Artificial Intelligence (IJCAI 2011),* pp. 1583-1588.

Yan, J., Zhang, B., Liu, N., Yan, S., Cheng, Q., Fan, W., Yang, Q., Xi, W. and Chen, Z., 2006. Effective and efficient dimensionality reduction for large-scale and streaming data preprocessing. IEEE transactions on Knowledge and Data Engineering, 18(3), pp.320-333.

Yang, H., Fujimaki, R., Kusumura, Y. and Liu, J. (2016, August). Online feature selection: A limited-memory substitution algorithm and its asynchronous parallel variation. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining ,pp. 1945-1954.

Yao, Z. and Liu, W., 2013. Extracting robust distribution using adaptive Gaussian Mixture Model and online feature selection. Neurocomputing, 101, pp.258-274.

Yang, H., Lyu, M.R. and King, I., 2013. Efficient online learning for multitask feature selection. ACM Transactions on Knowledge Discovery from Data (TKDD), 7(2), pp.1-27.

Yang, Y. Y. and Pedersen, J. O., 1997. A comparative study on feature selection in text categorization, in: Proceedings of ICML-97, pp. 412–420.

Yeh, Y. J. and Hsu, C. T., 2009. Online selection of tracking features using AdaBoost. IEEE Transactions on Circuits and Systems for Video Technology, 19(3), pp. 442-446

Yu, K., Ding, W. and Wu, X., 2016. LOFS: a library of online streaming feature selection. Knowledge-Based Systems, 113, pp.1-3.

Yu, K., Wu, X., Ding, W. and Pei, J. (2014, December). Towards scalable and accurate online feature selection for big data. In: 2014 IEEE International Conference on Data Mining, pp. 660-669.

Zhao, L., Wang, L. and Xu, Q., 2012. Data stream classification with artificial endocrine system. Applied Intelligence, 37(3), pp.390-404.

Žliobaitė, I., 2011. Combining similarity in time and space for training set formation under concept drift. *Intelligent Data Analysis*, *15*(4), pp.589-611.