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

Επιβλέπων Καθηγητής: Ευαγγελίδης Γεώργιος

Γεωργίου Αλέξανδρος - 02/11/2020

# Περιεχόμενα

# Εισαγωγή

Οι σύγχρονες γεννήτριες δεδομένων παράγουν δεδομένα σε τεράστιες ποσότητες και με μεγάλη ταχύτητα. Αποτέλεσμα είναι η άνοδος των δεδομένων ροής. Η εξαγωγή χρήσιμης πληροφορίας από δεδομένα ροής αποτελεί πρόκληση διότι η φύση τους επιβάλλει περιορισμούς που δεν μπορούν να ικανοποιηθούν από τους κλασικούς αλγορίθμους εκμάθησης.

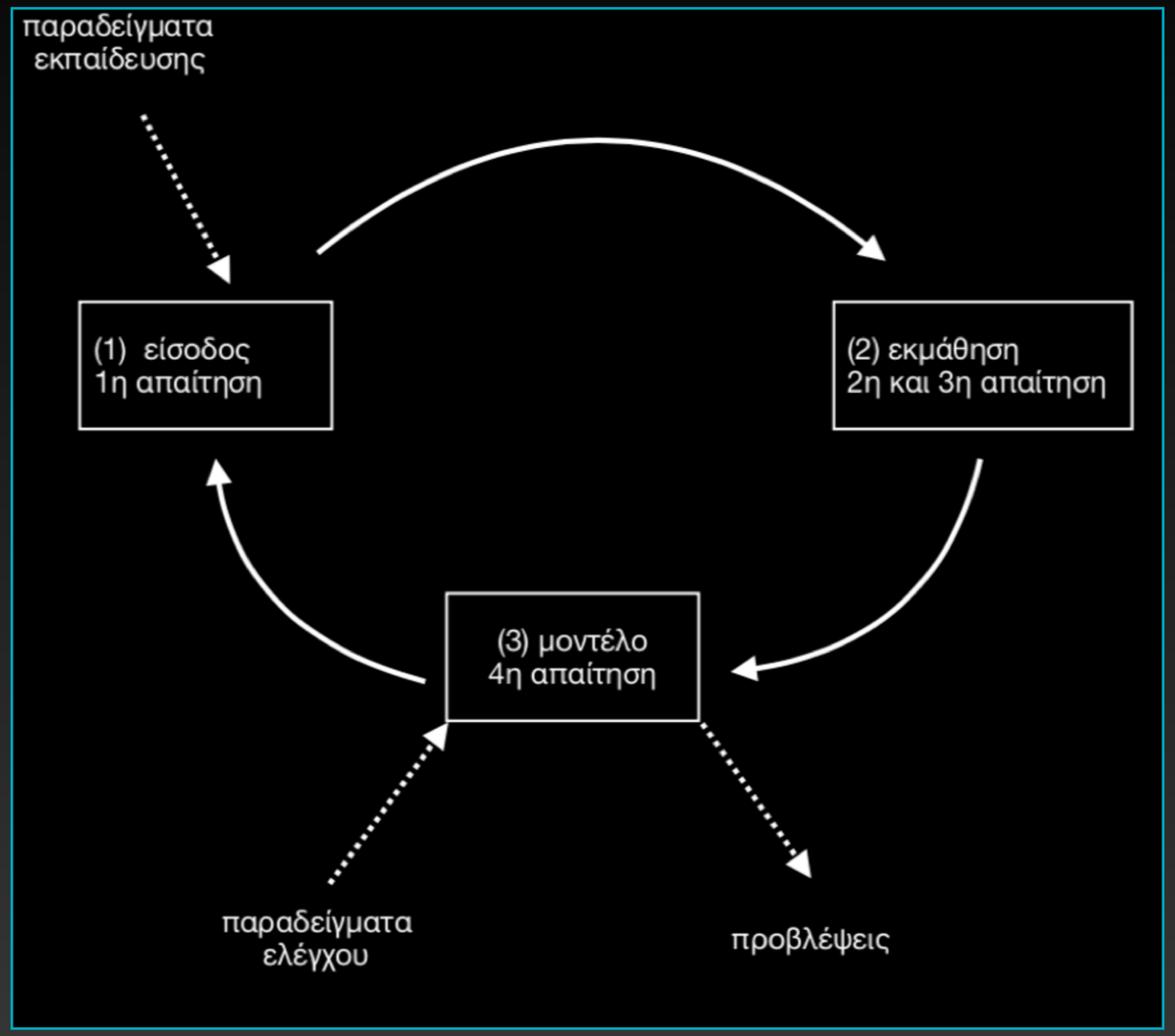
# Κύριες διαφορές στατικών δεδομένων και δεδομένων ροής

- Τα στιγμιότυπα δεν είναι διαθέσιμα εξαρχής
- Καταφθάνουν με ραγδαίο ρυθμό
- Τα δεδομένα ροής έχουν εν δύναμη άπειρο μέγεθος άρα δεν αποθηκεύονται σε μνήμη
- Κάθε στιγμιότυπο είναι συνήθως προσβάσιμο μόνο μια φορά
- Κάθε στιγμιότυπο επεξεργάζεται μεσα σε λίγο χρόνο
- Η πρόσβαση στην πραγματική τιμή είναι περιορισμένη
- Πιθανότητα αλλαγής στη βασική συνάρτηση παραγωγής δεδομένων (φαινόμενο concept drift).

# Απαιτήσεις που πρέπει να πληρούν οι αλγόριθμοι σε δεδομένα ροής

- Επεξεργασία ενός παραδείγματος κάθε φορά και εξέταση του μόνο μια φορά .
- Χρήση περιορισμένης μνήμης
- Λειτουργία σε περιορισμένο χρόνο
- Ο αλγόριθμος να είναι σε θέση να κάνει πρόβλεψη οποιαδήποτε στιγμή

# Κύκλος κατηγοριοποίησης για δεδομένα ροής




Bifet & Kirkby (2009)

# Προεπεξεργασία στα δεδομένα ροής

Η διαδικασία προεπεξεργασίας των δεδομένων είναι μια από τις πιο σημαντικές φάσεις στην διαδικασία αναζήτησης γνώσης από 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.

Οι τεχνικές μείωσης δεδομένων εκτελούν αυτή την απλούστευση επιλέγοντας και διαγράφοντας πλεονάζοντα και θορυβώδη γνωρίσματα και στιγμιότυπα ή διακριτοποιώντας σύνθετους συνεχόμενους χώρους των γνωρισμάτων.

## Η μείωση δεδομένων  είναι επιβεβλημένη καθώς θέτει ως στόχο

- Απόκτηση γρήγορων και προσαρμόσιμων μοντέλων
- Την βελτίωση της ακρίβειας
- Τα μοντέλα να έχουν χαμηλή υπολογιστική πολυπλοκότητα
- Την διαχείριση του φαινομένου του concept drift

# Κύριες κατευθύνσεις που επικεντρώνεται η έρευνα

- Μείωση διαστάσεων μέσω της επιλογής χαρακτηριστικών (FS)
- Μείωση περιπτώσεων μέσω της επιλογής στιγμιοτύπων (IS)
- Διακριτοποίηση μέσω της δημιουργίας διακριτών διαστημάτων

# Οι τεχνικές των παραπάνω κατηγοριών σε στατικά δεδομένα δεν εφαρμόζονται άμεσα στα δεδομένα ροής διότι

- Οι αλγόριθμοι επιλογής στιγμιοτύπων απαιτούν πολλαπλά περάσματα
- Οι αλγόριθμοι επιλογής γνωρισμάτων προσαρμόζονται σε online σενάρια αλλά αδυνατούν να αντιμετωπίσουν ένα ενδεχόμενο concept drift
- Οι μέθοδοι διακριτοποίησης  απαιτούν πολλαπλές επαναλήψεις με έντονες προσαρμογές.

# Data reduction με instance selection

Κύριος στόχος η επιλογή και η αναγνώριση των πιο σχετικών στιγμιοτύπων από μια τεράστια πηγή δεδομένων

Το βέλτιστο αποτέλεσμα της επιλογής στιγμιοτύπων είναι ένα ελάχιστο υποσύνολο δεδομένων το οποίο μπορεί να επιτύχει το ίδιο έργο χωρίς να υπάρχει απώλεια στην απόδοση. Δηλαδή:

$$P(DM_s) = P(DM_t)$$

# Λειτουργίες που έχει το instance selection

- Ενεργοποίηση. Επιτρέπει στον αλγόριθμο να λειτουργήσει με πολλά δεδομένα.
- Εστίαση. Η διαδικασία επιλογής στιγμιοτύπων επικεντρώνει τα δεδομένα στο συγκεκριμένο σημείο που επιθυμούμε να αντληθεί γνώση.
- Καθαρισμός. Απομακρύνονται θορυβώδη και περιττά στιγμιότυπα βελτιώνοντας έτσι την ποιότητα των δεδομένων.

# Η μέθοδος NEFCS-SRR

Μέθοδος που βασίζεται στην φιλοσοφία του CBM. Η CBM διαδικασία αναφέρεται στη διαδικασία διόρθωσης των περιεχομένων ενός CBR συστήματος και αναλύεται σε δύο επιμέρους στρατηγικές :

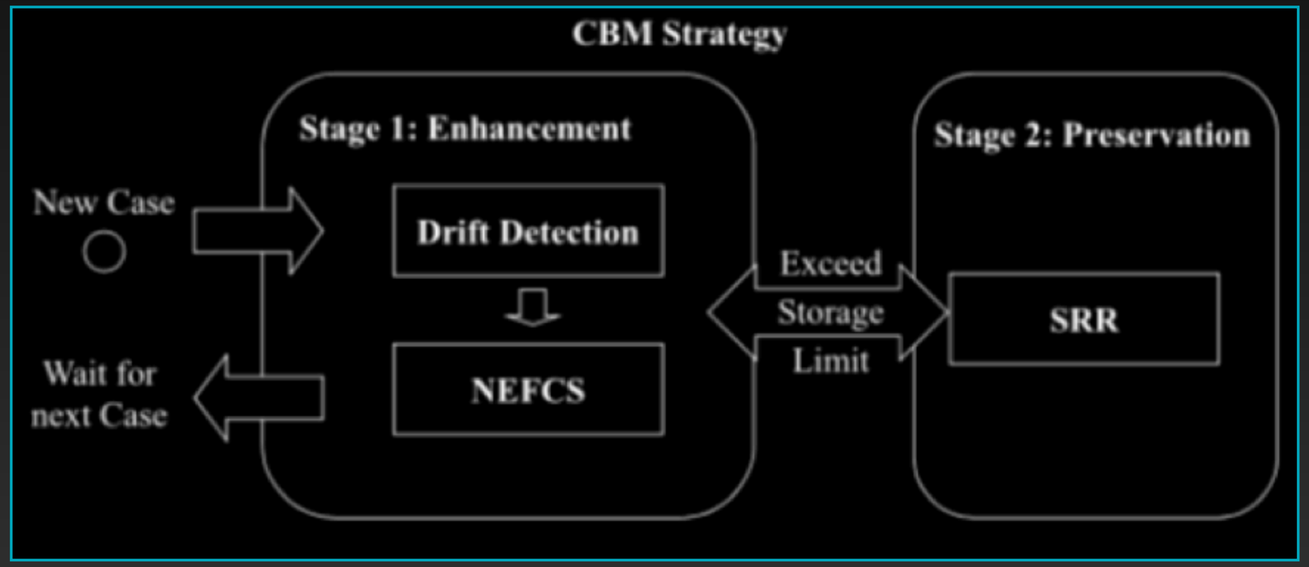
- Την ενίσχυση ικανότητας που στοχεύει στη μετακίνηση περιπτώσεων που αποτελούν θόρυβο με στόχο την βελτίωση της ακρίβειας του κατηγοριοποιητή.
- Την συντήρηση ικανότητας που αποσκοπεί στην μείωση περιττών περιπτώσεων που δεν συνεισφέρουν στην ικανότητα του κατηγοριοποιητή.

## Προβλήματα που προκύπτουν από το concept drift σε ένα CBR σύστημα:

- Διατήρηση επιβλαβών για την ακρίβεια περιπτώσεων που δεν αντιστοιχούν σε τωρινά concepts
- Νέες περιπτώσεις μετά την εμφάνιση του concept να χειριστούν ως θόρυβος και να απομακρυνθούν.

**Οι Lu et al.(2016)** προτείνουν μια πρωτότυπη **case-base** μέθοδο που αποτελείται από τρία στάδια:

- εντοπισμός αλλαγής βασισμένος στην ικανότητα.
- Ο αλγόριθμος **NEFCS** που στοχεύει στην απομάκρυνση θορύβων σε δυναμικά περιβάλλοντα (ενίσχυση).
- Η μέθοδος **SRR** η οποία απομακρύνει τις περιττές περιπτώσεις με έναν ομοιόμορφο τρόπο (διατήρηση).




Lu et al.(2016)

# Εντοπισμός αλλαγής βασισμένος στην ικανότητα

Η μέθοδος αυτή χρησιμοποιεί το μοντέλο της ικανότητας ως μια τεχνική διαίρεσης του χώρου και συγκρίνει μέσω αυτή τις κατανομές δυο παραθύρων με στιγμιότυπα. Με αυτό τον τρόπο ειδοποιεί για την παρουσία πιθανού concept drift αλλά και υποδεικνύει μια περιοχή που μπορεί να επηρεαστεί από αυτό.

# Μέθοδος NEFCS

Αποτελείται από τρεις διαδικασίες:

- Την τροποποιημένη μείωση θορύβου M-BBNR η οποία εφαρμόζει τον υποθετικό κανόνα BBNR για να εξετάσει αν κάθε νέα περίπτωση αποτελεί θόρυβο. Αν υπάρχει concept drift και η υπό εξέταση περίπτωση βρίσκεται μεσα στην περιοχή που λαμβάνει χώρα η αλλαγή τότε δεν θα διαγραφεί διότι μπορεί εκπροσωπεί ένα πρωτότυπο σενάριο. Διαφορετικά αν η περίπτωση ικανοποιεί τον υποθετικό κανόνα BBNR τότε απομακρύνεται από το cb

- Αλλαγή περιεχομένου. Αποθηκεύονται οι τελευταίες περιπτώσεις και διενεργείται ένα τεστ για την ακρίβεια της κάθε περίπτωσης. Αν δεν ικανοποιείται το κατώφλι τότε αυτή απενεργοποιείται από την μελλοντική διαδικασία. Υπάρχει όμως η δυνατότητα το παράδειγμα αυτό να μετακινηθεί πίσω στο cb

- Ανανέωση μοντέλου ικανότητας. Ανανεώνεται το μοντέλο ικανότητας με την προσθήκη περιπτώσεων αλλά και την οριστική διαγραφή των απομακρυσμένων περιπτώσεων.

# Βαθμιαία απομάκρυνση πλεονάσματος - SRR

Η μέθοδος αυτή χρησιμοποιεί λίστες μέσω των οποίων επιτυχγάνει τα εξής:

- Προσθέτει μια περίπτωση σε λίστα διατήρησης και την αποτρέπει να διαγραφεί.
- Απομακρύνει ομοιόμορφα το πλεόνασμα μέσα από τον μηχανισμό κλειδώματος που διαθέτει
- Για κάθε διαγραφή που γίνεται η SRR εγγυάται ότι δεν υπάρχει απώλεια κάλυψης για τις υπόλοιπες περιπτώσεις του CB.

# Η μέθοδος ECUE (CBE)

Η μέθοδος ECUE βασίζεται και αυτή σε προηγούμενες περιπτώσεις (CBR). Αποτελεί μια lazy learning τεχνική μηχανικής μάθησης με σκοπό την αντιμετώπιση του concept drift στα spam emails. Χωρίζεται σε τρία στάδια:

- Επιλογή γνωρίσματος
- Ανάκτηση περίπτωσης
- Διαχείριση της βάσης περιπτώσεων

# Επιλογή γνωρίσματος

Οι τεχνικές χωρισμού λέξεων οδηγούν σε ένα πολύ μεγάλο αριθμό γνωρισμάτων. Για τον λόγο αυτό χρησιμοποιείται η επιλογή στιγμιοτύπων με βάση το κέρδος πληροφορίας (IG) ούτως ώστε να επιλεχθούν τα πιο προβλεπτικά γνωρίσματα.

# Ανάκτηση περίπτωσης

Εδώ χρησιμοποιείται ένας εναλλακτικός αλγόριθμος ανάκτησης με βάση την ομοιότητα που βασίζεται στην μέθοδο CRN (δομή μνήμης που επιτρέπει την αποδοτική και ελαστική ανάκτηση περιπτώσεων). Οι περιπτώσεις αποθηκεύονται ως κόμβοι ενώ υπάρχει και ένας δεύτερος τύπος κόμβων που ονομάζεται information entity (IE). Οι κόμβοι είναι συνδεδεμένοι μεταξύ τους με τόξα.

Η υπό εξέταση περίπτωση ενεργοποιεί την CRN με το να συνδέεται με τους IE κόμβους. Αυτή η ενεργοποίηση επεκτείνεται μέσω του δικτύου στους κόμβους περιπτώσεων οι οποίοι συγκεντρώνουν ένα σκορ ανάλογα με την ομοιότητα τους με την στοχευμένη περίπτωση. Αυτοί που έχουν την μεγαλύτερη ενεργοποίηση είναι οι πιο όμοιοι με την στοχευμένη περίπτωση και επομένως ανακτώνται από το σύστημα.

# Διαχείριση της βάσης περιπτώσεων

Εφαρμόζεται η μέθοδος CBE η οποία αναγνωρίζει τις περιπτώσεις που συμβάλλουν στην σωστή κατηγοριοποίηση καθώς και αυτές που συμβάλλουν σε λάθος κατηγοριοποίηση. Επειτα:

1ο στάδιο: Απομακρύνονται από το case-base οι περιπτώσεις που αποτελούν θόρυβο ( competence - enhancement)

2ο στάδιο: Αναγνωρίζονται και απομακρύνονται οι περιττές περιπτώσεις (competence maintenance)

# Η μέθοδος ICF

Για ένα case-base $CB = \{c_1, c_2, \ldots, c_n\}$ ορίζονται τα σύνολα:

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

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

Οι Brighton και Mellish (2002) αντικαθιστούν την ιδιότητα Adaptable με αυτή του Local-Set με σκοπό την διαγραφή περιπτώσεων που έχουν μεγάλα loca-sets.

Ο αλγόριθμος ICF χρησιμοποιεί τα σύνολα κάλυψης και προσβασιμότητας.
Μια περίπτωση c διαγράφεται από τον ICF αν:
$|Reachable(c)| > |Coverage(c)|$. Η διαδικασία προχωράει με τον επαναλαμβανόμενο υπολογισμό αυτών των ιδιοτήτων.

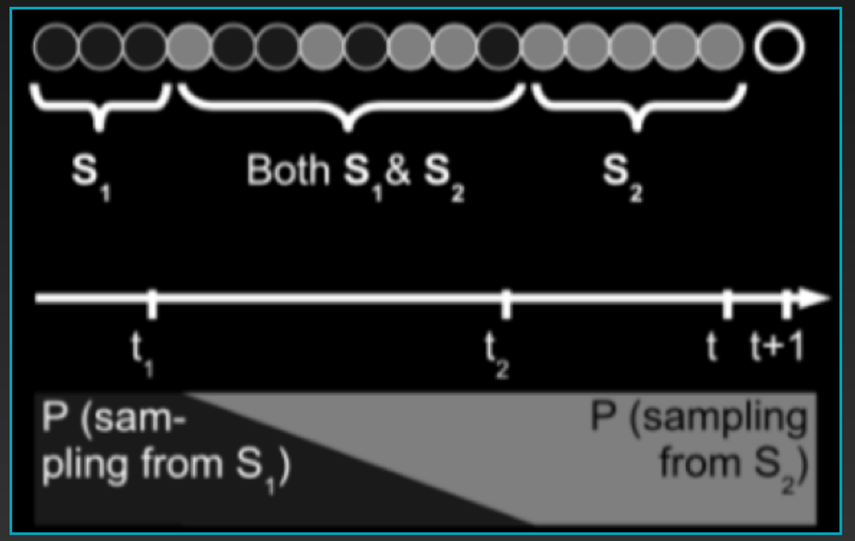
Το σύνολο προσβασιμότητας δεν έχει προκαθορισμένο μέγεθος κάτι το οποίο επηρεάζει τον αλγόριθμο ICF.

Ο αλγόριθμος ICF επικεντρώνεται στην απομάκρυνση των περιττών περιπτώσεων. Επομένως είναι πιθανόν να προστατεύει περιπτώσεις που αποτελούν θόρυβο. Για τον λόγο αυτό χρησιμοποιείται το φιλτράρισμα θορύβου που βασίζεται στη διαγραφή του Wilson και Martinez (1997).

# Οι μέθοδοι FISH

Στόχος είναι να δοθεί μια ετικέτα στο στιγμιότυπο $X_{t+1}$. Για να κατασκευαστεί ένας κατηγοριοποιητής με καλή ακρίβεια αναζητείται ένα σύνολο εκπαίδευσης το οποίο θα είναι όσο το δυνατόν πιο κοντά στην πηγή του $X_{t+1}$.

Η Zliobaite (2011) προτείνει ότι μπορούμε να βρούμε ποσο όμοιο είναι το με τα στιγμιότυπα που έχουν εξεταστεί έως τότε επιλέγοντας ένα σύνολο εκπαίδευσης το οποίο αποτελείται από στιγμιότυπα τα οποία είναι όμοια με το $X_{t+1}$.




Zliobaite (2011)

Χρησιμοποιείται η ιδέα της ομοιότητας στον χώρο και στον χρόνο σύμφωνα με την σχέση: $D(X_i, X_j) = a_1 d_{ij}^{(S)} + a_2 d_{ij}^{(T)}$. Η επόμενη απόφαση που πρέπει να παρθεί είναι πόσα από τα όμοια στιγμιότυπα που έχουν εξεταστεί πρέπει να συμπεριληφθούν στο σύνολο εκπαίδευσης. Το μέγεθος του συνόλου εκπαίδευσης καθορίζεται με ένα κατώφλι. Το στιγμιότυπο $X_i$ εισέρχεται στο σύνολο εκπαίδευσης αν $D^*(X_i, X_{t+1}) < h^D$, όπου το $h^D$ προκαθορίζεται από τον σχεδιαστή (FISH1) ή παράγεται από ένα σύνολο επικύρωσης (FISH2, FISH3).

**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 $(X_{z1}, \ldots, X_{zN})$ ως σύνολο εκπαίδευσης.

7: έλεγξε το $L^N$ στους k εγγύτερους γείτονες $(X_{z1}, \ldots, X_{zN})$ 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**

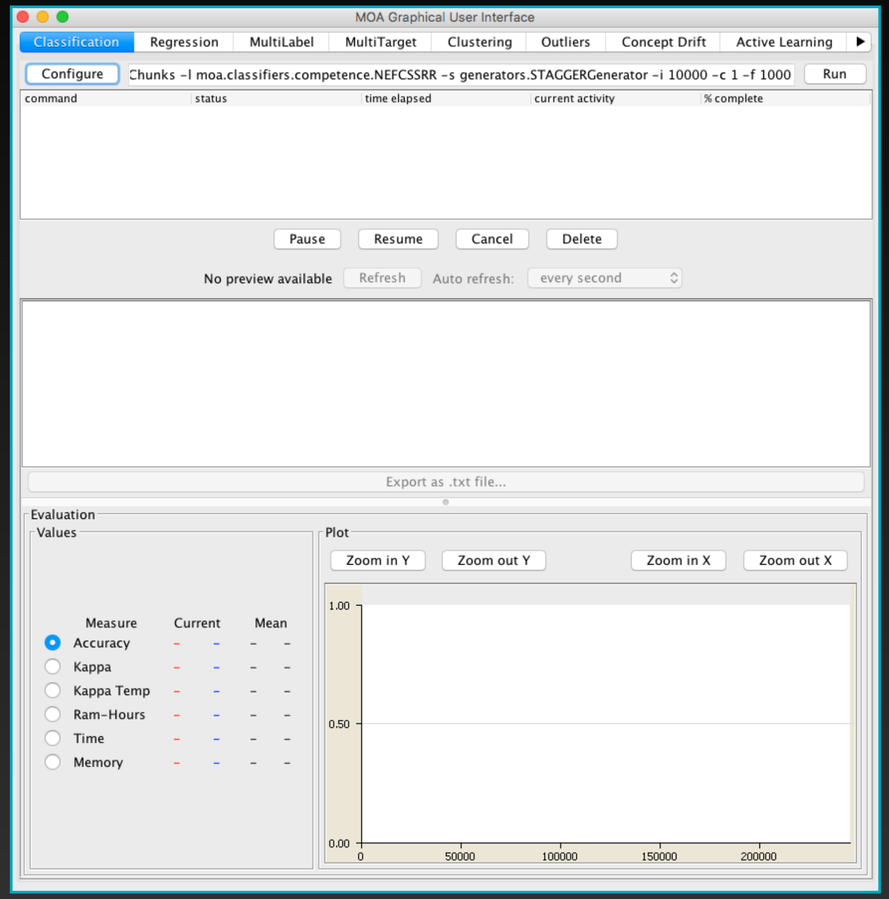
Οι δείκτες $I_t = \{z1,...,zN\}$ για τον σχηματισμό του συνόλου εκπαίδευσης $X_t^T$.

# Το λογισμικό MOA (Massive Online Analysis) και οι γεννήτριες παραγωγής δεδομένων

# Το λογισμικό MOA (Massive Online Analysis)

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.

Οι αλγόριθμοι γράφονται σε Java και υλοποιούνται ως κλάσεις οι οποίες έπειτα οργανώνονται σε πακέτα για να επικοινωνούν καλύτερα μεταξύ τους.



# Γεννήτρια STAGGER

Τρία κατηγορικά γνωρίσματα: μέγεθος, χρώμα σχήμα

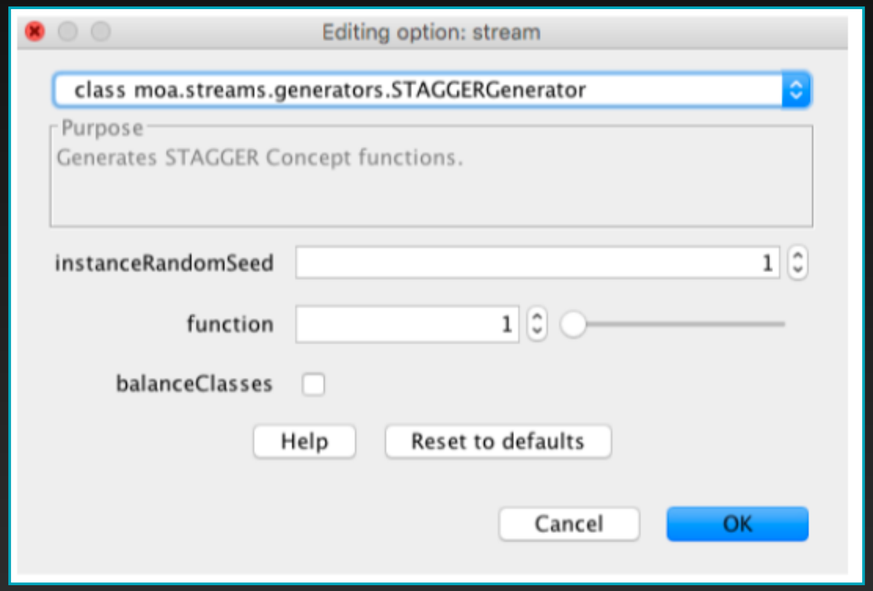
Μέγεθος → μικρό, μεσαίο, μεγάλο

χρώμα → Κόκκινο, πράσινο

Σχήμα → κυκλικό, μη κυκλικό

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.



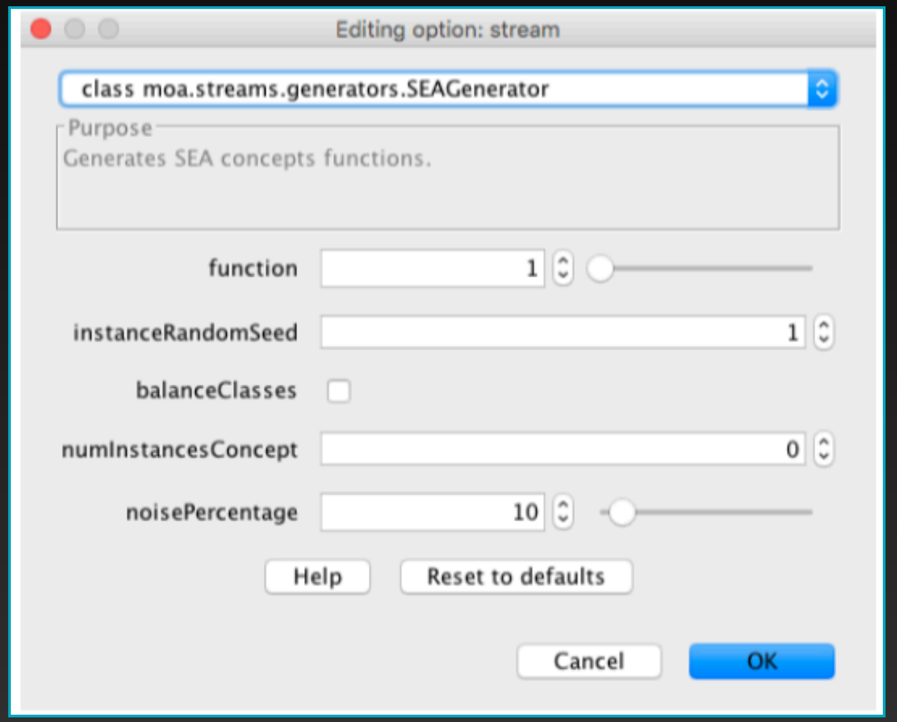
# Γεννήτρια SEA

Παράγει δεδομένα με τρία γνωρίσματα. Μονο τα δύο πρώτα είναι σχετικά μεταξύ τους.

Όλα τα γνωρίσματα λαμβάνουν τιμές στο [0,10]

Η γεννήτρια κατηγοριοποιεί ένα στιγμιότυπο με βάση τις εξής συναρτήσεις:

- 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

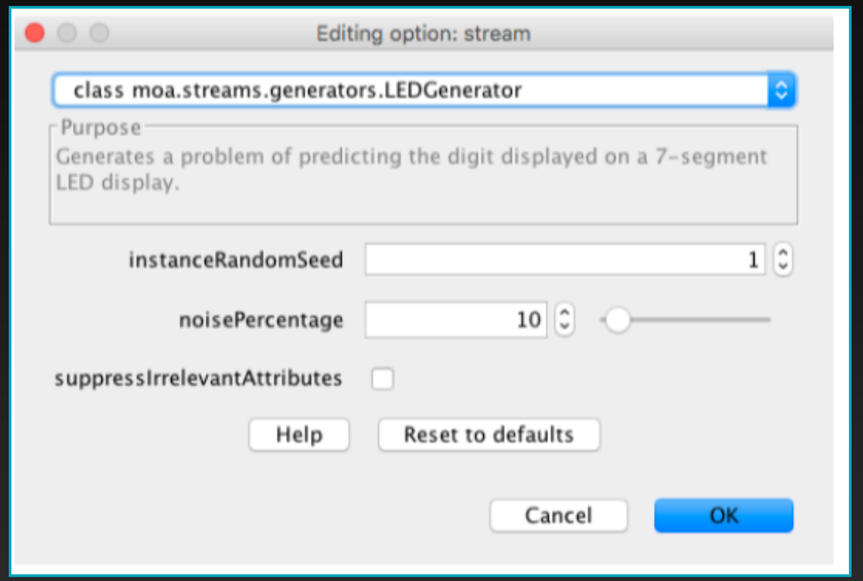


# Γεννήτρια LED

Στόχος είναι να προβλεφθεί το ψηφίο που εμφανίζεται σε μια διάταξη LED 7 τμημάτων.

Η συγκεκριμένη ρύθμιση της γεννήτριας παράγει 24 δυαδικά γνωρίσματα 17 από τα οποία είναι άσχετα.

Βέλτιστη Bayes κατηγοριοποίηση : 74%

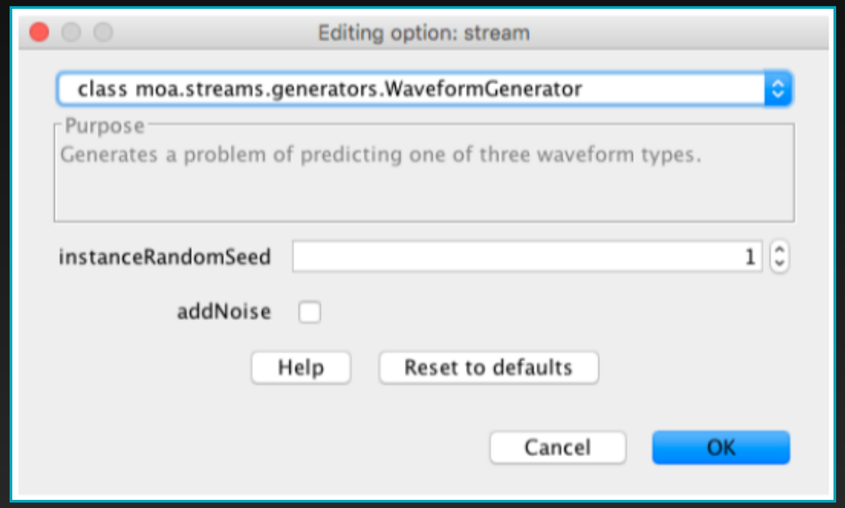


# Γεννήτρια LED

Στόχος η κατηγοριοποίηση των δεδομένων μεταξύ τριών διαφορετικών κλάσεων.

Βελτιστη Bayes κατηγοριοποίηση : 86%

Δυο εκδοχές: WAVE21 με 21 αριθμητικά γνωρίσματα και WAVE40 που εισάγει 19 άσχετα μεταξύ τους γνωρίσματα.



# Πειραματικές Ρυθμίσεις

# Μέγεθος δεδομένων

Η πειραματική μελέτη της εργασίας επιβεβαιώνει την έρευνα των Gallego et al. (2017) ότι οι μέθοδοι επιλογής στιγμιοτύπων είναι ακατάλληλες για μεσαίου ή μεγάλου μεγέθους σύνολα δεδομένων.

Καταλήγουν στο συμπέρασμα ότι οι competence based μέθοδοι απαιτούν πολύ χρόνο για να διατηρήσουν το μοντέλο ικανότητας.

Η πειραματική έρευνα επομένως αναγκάζεται να καταφύγει σε σύνολα δεδομένων που αριθμούν 10.000 στιγμιότυπα.

| 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 |

Χρόνοι μεθόδων στα STAGGER για 10000 instances

# Έρευνα για την τιμή του p

Το p ορίζεται ως η μεταβλητή που υποδηλώνει το κάθε πότε θα ενεργοποιείται η μέθοδος μείωσης.

Στα 10.000 στιγμιότυπα η προκαθορισμένη τιμή των p=500 θα ενεργοποιήσει σε κάθε πείραμα την μέθοδο μείωσης 19 φορές.

Στον πίνακα φαίνεται το trade off ανάμεσα στην ακρίβεια και το μέγεθος του μοντέλου.

Παρατηρούμε ότι η default τιμή (p=500) δεν αποδίδει τόσο καλά. Το ίδιο συμβαίνει και για τα υπόλοιπα σύνολα δεδομένων.

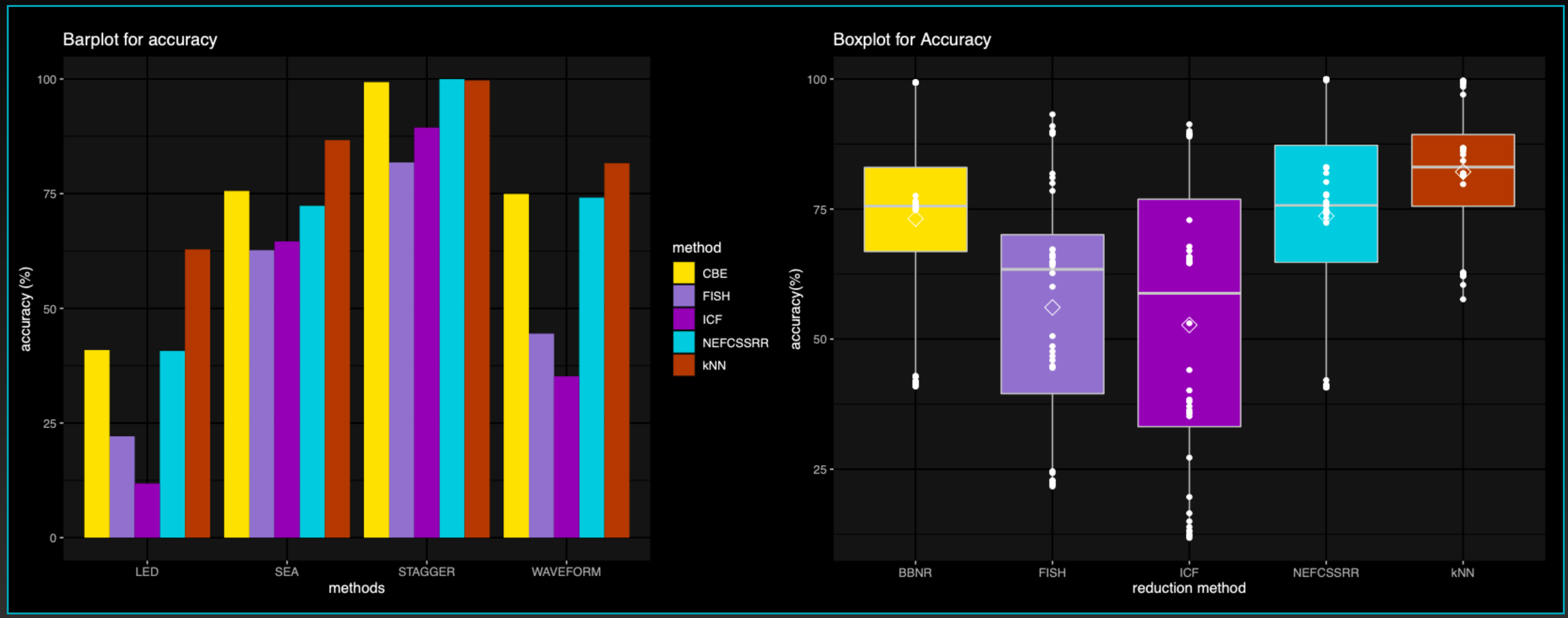
Ωστόσο για να υπάρξει σωστή σύγκριση μεταξύ των αλγορίθμων θα εξεταστεί η ακρίβεια τους καθώς και το κόστος σε χρόνο και μνήμη για την τιμή p=500.

| | 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** |

Έρευνα για την τιμή του p στα δεδομένα STAGGER

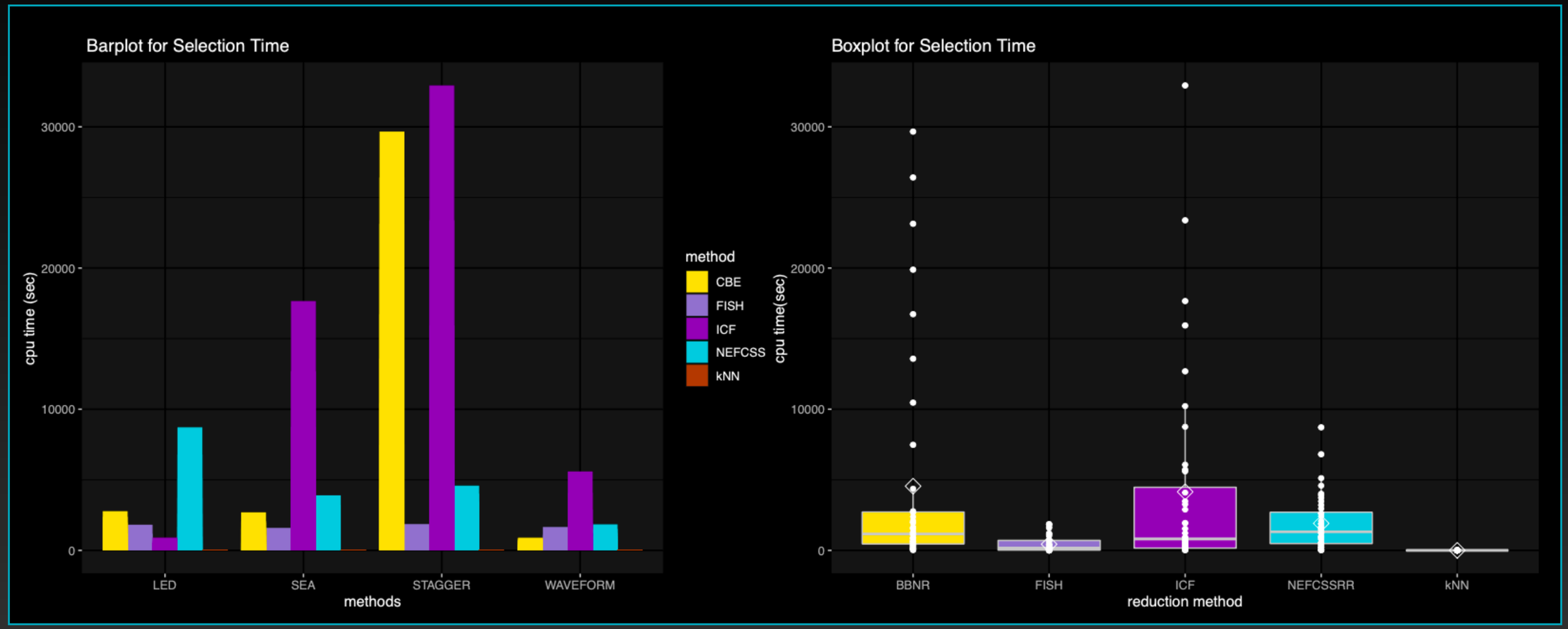
# Αποτελέσματα - Σύγκριση Μεθόδων

# Περίπτωση χωρίς concept drift



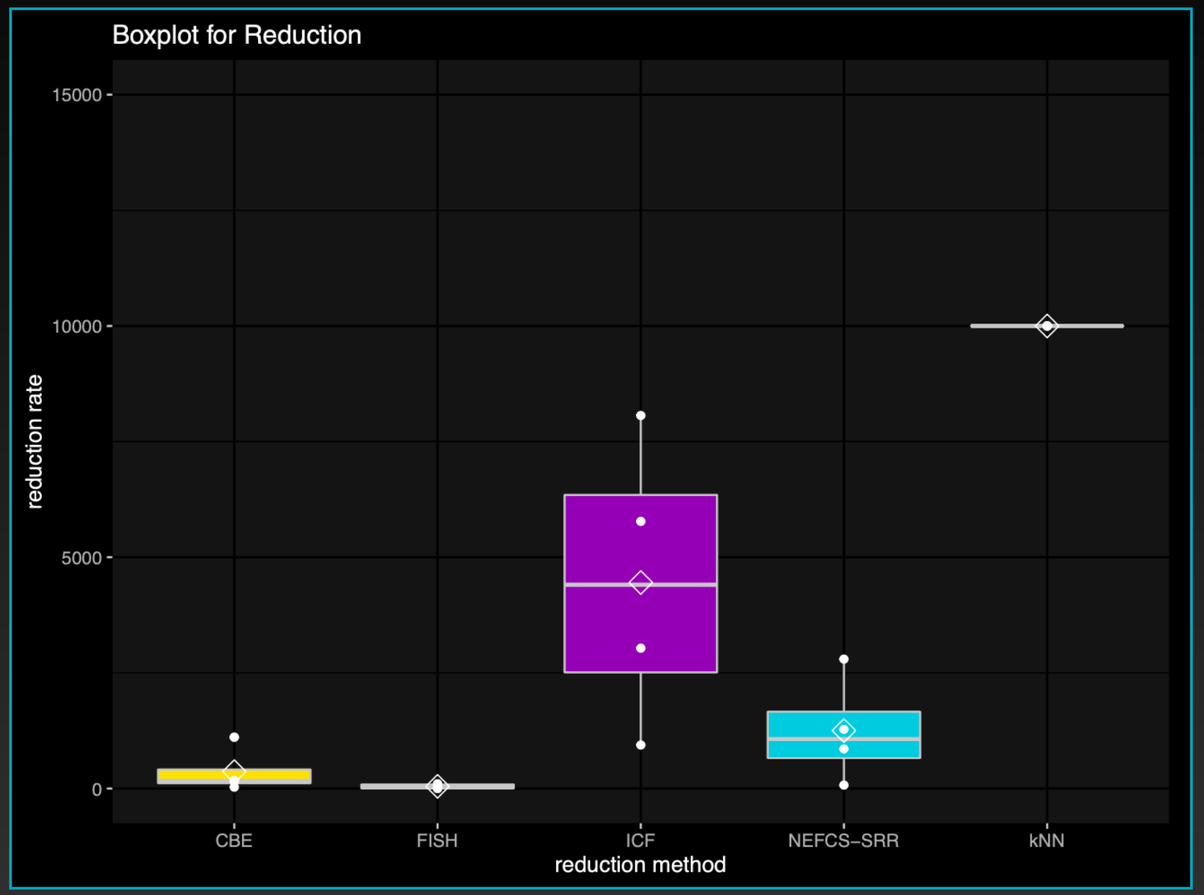

ο kNN έχει την υψηλότερη ακρίβεια και ακολουθείται από τους NEFCS-SRR και CBE.

# Περίπτωση χωρίς concept drift



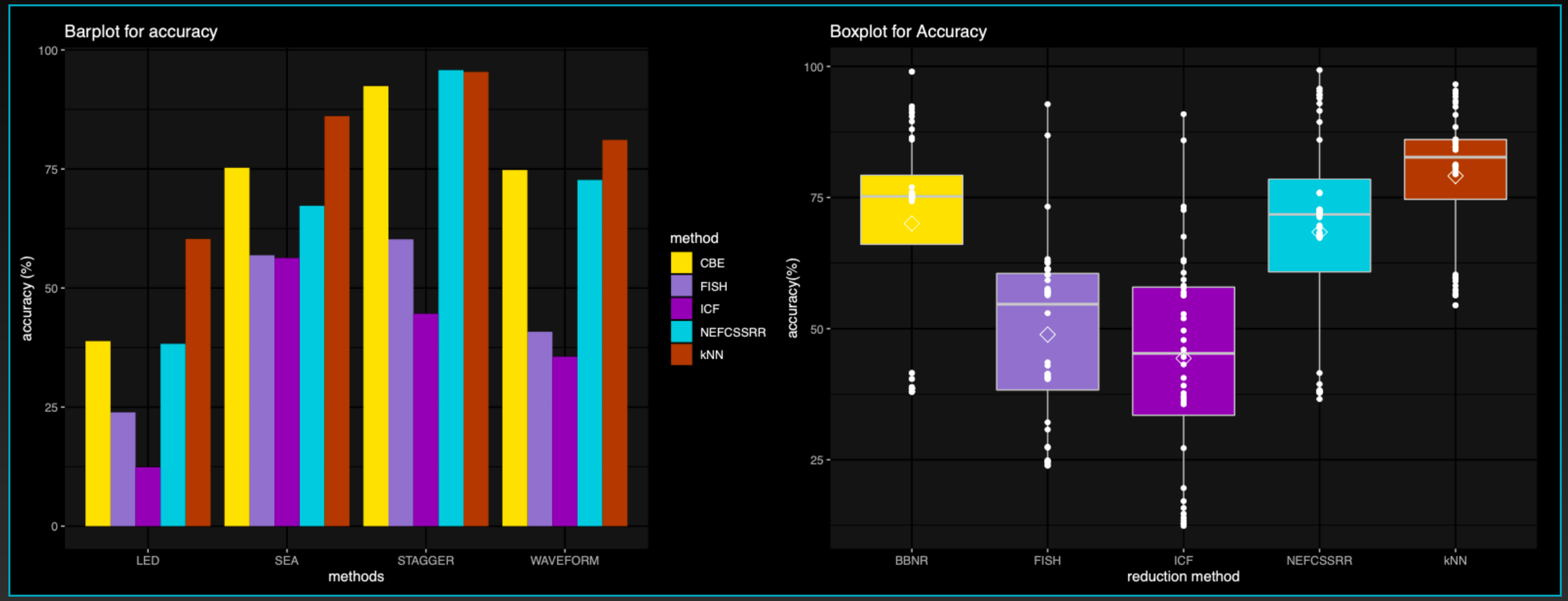

ο ICF είναι ιδιαίτερα χρονοβόρος κυρίως στα δεδομένα SEA και STAGGER. Ο αλγόριθμος που δείχνει να έχει χαμηλές απαιτήσεις σε χρόνο είναι ο FISH και ακολουθείται από τον NEFCS-SRR.

# Περίπτωση χωρίς concept drift



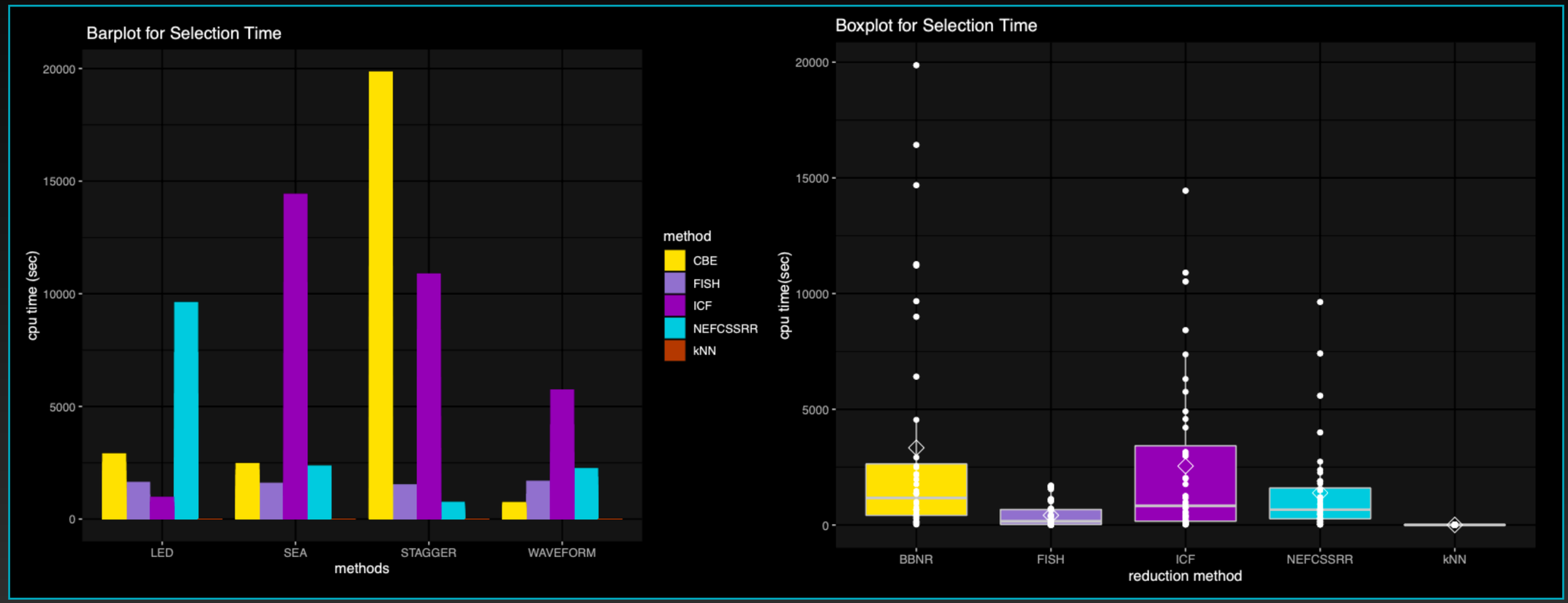

Η μείωση που επιτυγχάνουν οι τέσσερις μέθοδοι είναι αξιοσημείωτη ειδικά για τις μεθόδους FISH και CBE. Στην μέθοδο NEFCS-SRR καθορίζουμε εμείς το μέγεθος του case-base του.

# Περίπτωση με εισαγωγή τεχνητού concept drift



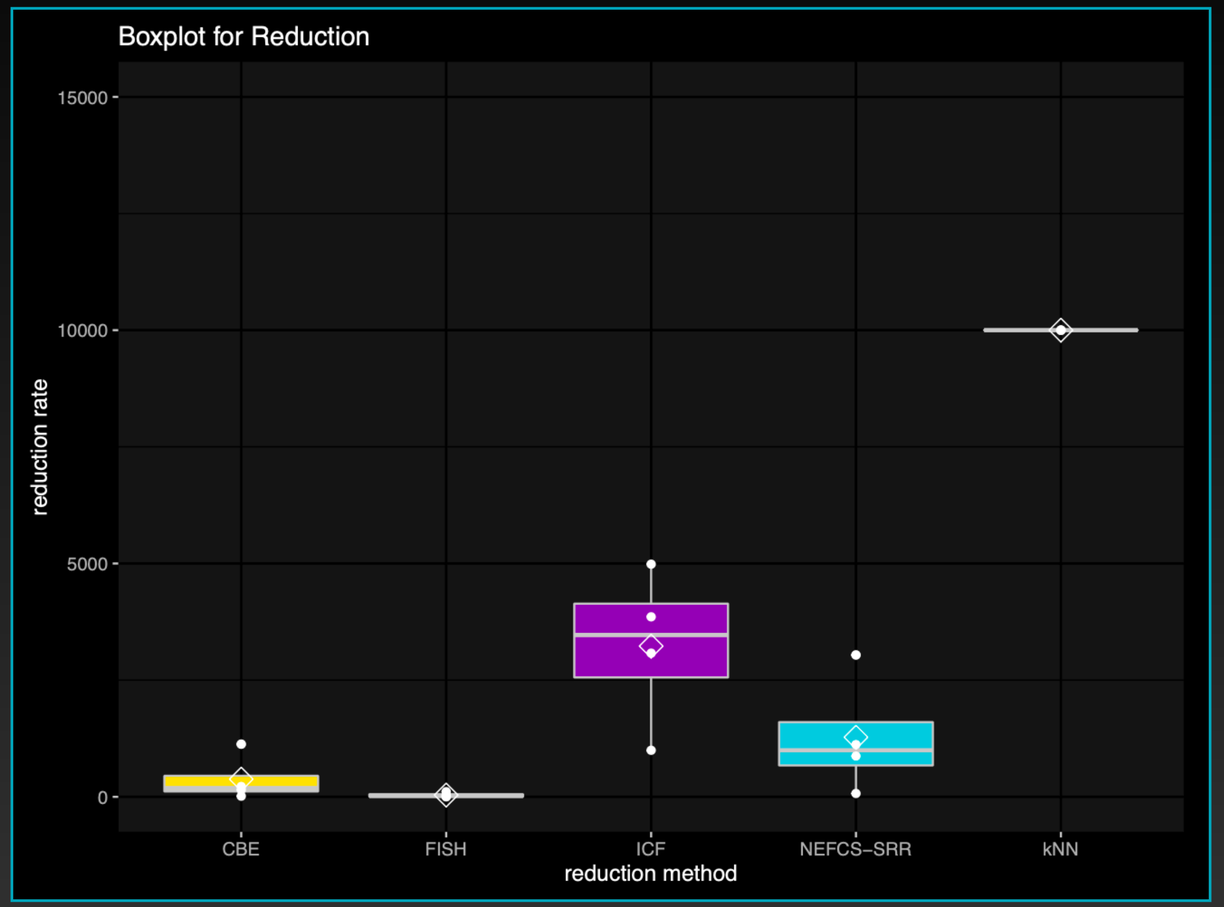

Εισάγουμε τεχνητό concept drift ρυθμίζοντας κατάλληλα την κατανομή. Στις γεννήτριες που εκτελούμε τα πειράματα μας δίνεται αυτή η δυνατότητα. Οι μέθοδοι NEFCS-SRR και CBE φαίνεται να αποδίδουν καλύτερα υπό την παρουσία concept drift.

# Περίπτωση με εισαγωγή τεχνητού concept drift




Σχετικά με την χρονική πολυπλοκότητα παρατηρούμε ότι η μέθοδος FISH επιτυγχάνει την καλύτερη απόδοση ενώ πολλά υποσχόμενη είναι και η NEFCS-SRR.

# Περίπτωση με εισαγωγή τεχνητού concept drift




Στα θηκογράμματα 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.

# Σύνοψη - Συμπεράσματα

Από την πειραματική έρευνα εξήχθησαν τα εξής σημαντικά συμπεράσματα:

- Σύμφωνα με τους μη παραμετρικούς ελέγχους αλλά και τα γραφήματα 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

- 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.

- 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

- 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.

“Σας ευχαριστώ”