ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ
ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΤΜΗΜΑΤΟΣ ΕΦΑΡΜΟΣΜΕΝΗΣ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΜΕΛΕΤΗ ΤΟΥ ΠΕΡΙΒΑΛΛΟΝΤΟΣ SCIKIT-MULTIFLOW ΓΙΑ ΤΗΝ ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ ΣΕ ΔΕΔΟΜΕΝΑ ΡΟΗΣ

Φοιτητής: Σταυρίδης Απόστολος

Σημαντικότητα του Προβλήματος

- **Επεξεργασία Δεδομένων:** Διαθέσιμα δεδομένα για πολλαπλή πρόσβαση
- **Συστήματα Βάσεων Δεδομένων:**
  - Χαρακτηριστικό παράδειγμα αποθηκευμένων δεδομένων
- **Ροές Δεδομένων:**
  - Συνεχή δεδομένα σε μια δυνητικά άπειρη ροή
  - Επεξεργάζονται από ένα σύστημα με περιορισμένους πόρους
    - Η κύρια μνήμη είναι μικρή
    - Μπορεί να περιέχει μόνο ένα μικρό τμήμα του stream
    - Τα δεδομένα απορρίπτονται αμέσως μετά την επεξεργασία
- **Διαδικτυακή μάθηση:**
  - Ενημερώνει το μοντέλο της μετά από κάθε εμφάνιση δεδομένων
  - Δεν απαιτεί πρόσβαση σε όλα τα δεδομένα του παρελθόντος

Θεωρητικό Υπόβαθρο

- **Μηχανική μάθηση:**
  - Ασχολείται με την κατασκευή αλγορίθμων που βασίζονται σε μια συλλογή παραδειγμάτων κάποιου φαινομένου
  - Τα παραδείγματα μπορούν να προέρχονται από έναν άλλο αλγόριθμο
  - Διαδικασία επίλυσης πρακτικού προβλήματος με
    - Συλλογή συνόλου δεδομένων
    - Αλγοριθμική δημιουργία στατιστικού μοντέλου βάσει του συνόλου δεδομένων

Θεωρητικό Υπόβαθρο

- **Διαδικτυακή Μηχανική μάθηση Μεγάλων Δεδομένων Ροής:**
  - **Data streams:**
    - Δημιουργούνται με πολύ υψηλούς ρυθμούς
    - Πρέπει να αναλυθούν γρήγορα για να φθάσουν σε ευαίσθητες πληροφορίες
    - Προέρχονται από μετρήσεις δικτύου, εγγραφές κλήσεων, επισκέψεις σε ιστοσελίδες, αναγνώσεις αισθητήρων κ.ο.κ.,
    - Φθάνουν συνεχώς σε πολλαπλές, ταχείες, χρονικά μεταβαλλόμενες, πιθανώς απρόβλεπτες και απεριόριστες ροές
    - Βασικός περιορισμός επεξεργασίας τους οι πόροι του συστήματος

Θεωρητικό Υπόβαθρο

- **Διαδικτυακή Μηχανική μάθηση Μεγάλων Δεδομένων Ροής:**
  - Τα γρήγορα δεδομένα σχετίζονται με τη γρήγορη σημασιολογία
  - Οι παραδοσιακοί αλγόριθμοι μάθησης batch κατασκευάζουν στατικά μοντέλα
  - Οι αλγόριθμοι μάθησης ροής πρέπει να δημιουργήσουν μοντέλα που εξελίσσονται με την πάροδο του χρόνου
  - Η επεξεργασία εξαρτάται από τη σειρά παραδειγμάτων που παράγονται από συνεχή, μη στατική ροή δεδομένων
  - Η μοντελοποίηση επηρεάζεται από ενδεχόμενες μετακινήσεις ιδεών ή από αλλαγές στη διανομή

Σκοπός Εργασίας

- Διερεύνηση και σύγκριση υλοποιημένων αλγορίθμων κατηγοριοποίησης του scikit-multiflow πάνω σε ροές δεδομένων.
- Ανάλυση αλγορίθμων απομείωσης δεδομένων σε πραγματικό χρόνο
- Αποτίμηση απόδοσης κατηγοριοποίησης με ή χωρίς απομείωση δεδομένων

Μεθοδολογία

- **Κριτική αποτίμηση διεθνών βιβλιογραφικών πηγών**
- **Ανάλυση:**
  - Βιβλιοθήκης λογισμικού scikit-multiflow,
  - Αρχιτεκτονικής scikit-multiflow
  - Παράδειγμα εργασίας ταξινόμησης με χρήση γεννήτριας SEA
  - Χρήση Προσαρμοστικής Παραθυροποίησης (ADWIN)

Βιβλιοθήκη Λογισμικού Scikit-Multiflow

Scikit-Multiflow:

- Βιβλιοθήκη μηχανικής μάθησης
- Για δεδομένα ροής πολλαπλών εξόδων/πολλαπλών ετικετών
- Γραμμένα σε γλώσσα Python
- Ελεύθερο & Ανοιχτού Κώδικα Λογισμικό
- Επεκτείνει τα υπάρχοντα εργαλεία για επιστημονικούς σκοπούς
- Αναπτύχθηκε με στόχο:
  - Τον εύκολο σχεδιασμό και εκτέλεση πειραμάτων
  - Επιτρέπει τη γρήγορη πρωτοτυποποίηση και πειραματισμό

Βιβλιοθήκη Λογισμικού Scikit-Multiflow

Scikit-Multiflow:

- Περιλαμβάνει:
  - Συλλογές υπερσύγχρονων μεθόδων για
    - Ταξινόμηση
    - Παλινδρόμηση
    - Ανίχνευση εννοιολογικής απόκλισης
    - Ανίχνευση ανωμαλιών
  - Γεννήτριες δεδομένων και αξιολογητών
- Σχεδιασμένο για αλληλεπίδραση με *NumPy* και *SciPy*
- Αναπτύχθηκε με στόχο:
    - Τον εύκολο σχεδιασμό και εκτέλεση πειραμάτων
    - Επιτρέπει τη γρήγορη πρωτοτυποποίηση και πειραματισμό

Αρχιτεκτονική Scikit-Multiflow

- **StreamModel:**
  - Βασική κλάση στο scikit-multiflow
  - Περιέχει αφηρημένες μεθόδους:
    - _fit : Εκπαιδεύει ένα μοντέλο κατά batch τρόπο
    - _ partial fit : Αυξανόμενα εκπαιδεύει ένα μοντέλο ροής
    - _predict : Προβλέπει την αξία του στόχου σε μεθόδους μάθησης υπό επίβλεψη
    - _predict_proba: Υπολογίζει πιθανότητες ανά τάξη σε προβλήματα ταξινόμησης
- Ένα **αντικείμενο StreamModel** αλληλεπιδρά με:
  - Ένα αντικείμενο Stream που παρέχει συνεχή ροή δεδομένων κατόπιν αιτήματος
  - Ένα αντικείμενο StreamEvaluator (προαιρετικά) που εκτελεί πολλαπλές εργασίες
    - ερωτά τη ροή δεδομένων
    - εκπαιδεύει και δοκιμάζει το μοντέλο στα εισερχόμενα δεδομένα
    - παρακολουθεί συνεχώς την απόδοση του μοντέλου

Παράδειγμα Εργασίας Ταξινόμησης

- **Χρησιμοποιείται η γεννήτρια SEA**
  - Γεννήτρια ροής δεδομένων
  - Δεν αποθηκεύει αλλά παράγει δεδομένα κατά ζήτηση
- **Κλάση SEAGenerator:**
  - Δημιουργία δεδομένων προβλήματος δυαδικής ταξινόμησης
  - Περιέχουν 3 αριθμητικά χαρακτηριστικά
  - Μόνο 2 σχετίζονται με την μάθηση
- **Σκοπός:**
  - Η εκπαίδευση ενός ταξινομητή Naive Bayes

Παράδειγμα Εργασίας Ταξινόμησης

- **Χρησιμοποιείται η γεννήτρια SEA:**
  - Γεννήτρια ροής δεδομένων
  - Δεν αποθηκεύει αλλά παράγει δεδομένα κατά ζήτηση
- **Κλάση SEAGenerator:**
  - Δημιουργία δεδομένων προβλήματος δυαδικής ταξινόμησης
  - Περιέχουν 3 αριθμητικά χαρακτηριστικά
  - Μόνο 2 σχετίζονται με την μάθηση
- **Σκοπός:**
  - Η εκπαίδευση ενός ταξινομητή Naive Bayes
  - Αξιολόγηση prequential για την παρακολούθηση της απόδοσης των μεθόδων μάθησης ροής

Παράδειγμα Εργασίας Ταξινόμησης

Παράδειγμα 1

- Όταν φτάνει ένα δείγμα δεδομένων (X, y):
  - **Γίνονται προβλέψεις** για το νέο δείγμα (X) για την αξιολόγηση της απόδοσης του μοντέλου
  - Το νέο δείγμα χρησιμοποιείται για να **εκπαιδεύσει το μοντέλο** και ενημερώνει την εσωτερική του κατάσταση.
- **Υλοποίηση βρόχου prequential:**
- **Ακρίβεια ταξινομητή Naive Bayes:** 93.95%

```
stream = SEAGenerator(random_state=1)
classifier = NaiveBayes()

n_samples = 0
correct_cnt = 0
max_samples = 2000

# Βρόχος αξιολόγησης μεθόδου prequential
while n_samples < max_samples and \
stream.has_more_samples():
  X, y = stream.next_sample()
  # Πρόβλεψη κλάσης για τα νέα δεδομένα
  y_pred = classifier.predict(X)
  if y[0] == y_pred[0]:
    correct_cnt += 1
  # Partially fit (εκπαίδευση) μοντέλου με νέα δεδομένα
  classifier.partial_fit(X, y)
  n_samples += 1

print('{} samples analyzed.'.format(n_samples))
print('Accuracy: {}'.format(correct_cnt / n_samples))

> 2000 samples analyzed.
> NaiveBayes classifier accuracy: 0.9395
```

Παράδειγμα Εργασίας Ταξινόμησης

Παράδειγμα 2

- Ο **βρόχος prequential μπορεί να τροποποιηθεί** για ανάγνωση από άλλες δομές δεδομένων (numpy.ndarray ή pandas.DataFrame)
- Για **εφαρμογές σε πραγματικό χρόνο με εξ' ολοκλήρου δεδομένα ροή** (buffers πρωτοκόλλου της Google), η κλάση Stream μπορεί να επεκταθεί
- Η **μέθοδος prequential** υλοποιείται στην **EvaluatePrequential**
- Χρήση **SGDClassifier** για συμβατότητα με αυξητικές μεθόδους scikit-learn

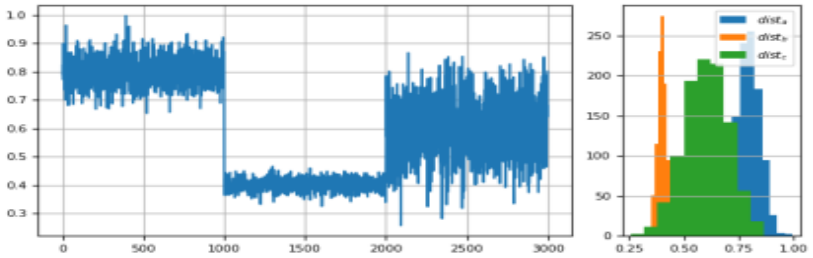
```
stream = SEAGenerator(random_state=1)
nb = NaiveBayes()
svm = SGDClassifier()
# Ρύθμιση της αξιολόγησης
metrics = ['accuracy', 'kappa',
           'running_time', 'model_size']
eval = EvaluatePrequential(show_plot=True,
                max_samples=20000,
                metrics=metrics)
# Εκτέλεση της αξιολόγησης
eval.evaluate(stream=stream, model=[nb, svm],
              model_names=['NB', 'SVM']);
```

Παράδειγμα Εργασίας Ταξινόμησης

- **Σύγκριση Παραδειγμάτων:**
  - **NaiveBayes:**
    - Παρουσιάζει καλύτερη απόδοση στην αρχή της ροής
    - Είναι γρηγορότερος
    - Δεσμεύει ελαφρώς περισσότερη μνήμη
  - **SGDClassifier:**
    - Τελικά ξεπερνά τον NaiveBayes
    - Είναι αργότερος
    - Έχει μικρότερο αποτύπωμα μνήμης.

Ανίχνευση Εννοιολογικής Απόκλισης

- Χρήση συνθετικής ροή δεδομένων τριών σετ 1000 δειγμάτων ροής που περιέχουν ακολουθία από κανονική κατανομή:
  - 1ο σετ ($\mu_\alpha = 0.8, \sigma_\alpha = 0.05$)
  - 2ο σετ ($\mu_b = 0.4, \sigma_b = 0.2$)
  - 3ο σετ ($\mu_c = 0.6, \sigma_c = 0.1$)



Ανίχνευση Εννοιολογικής Απόκλισης

- **Μέθοδος Ανίχνευσης Εννοιολογικής Απόκλισης:**
  Προσαρμοστικής Παραθυροποίησης (ADWIN)
- **Σκοπός:**
  Ανίχνευση απόκλισης μετά τα δείγματα 1000 και 2000 στη συνθετική ροή δεδομένων.

```
drift_detector = ADWIN()
 for i, val in enumerate(stream_int):
    drift_detector.add_element(val)
    if drift_detector.detected_change():
      print('Change detected at index {}'.format(i))

      drift_detector.reset()
```

> Change detected at index 1055
> Change detected at index 2079

Επίδραση Απόκλισης στη Μάθηση

- **Παράδειγμα σύγκρισης** δύο δημοφιλών μοντέλων ροής:
  - **HoeffdingTreeClassifier**
  - **HoeffdingAdaptiveTreeClassifier**
- Δεδομένα από αρχείο csv
- Χρήση της κλάσης FileStream
- Δεδομένα με 3 βαθμιαίες αποκλίσεις στα 5k, 10k, και 15k δείγματα
- **Βαθμιαία απόκλιση:** η παλιά έννοια σταδιακά αντικαθίσταται από μια νέα
- Υπάρχει μεταβατική περίοδος που οι δύο έννοιες είναι παρούσες

Επίδραση Απόκλισης στη Μάθηση

- **Πρώτα 5k δείγματα:** και οι δύο μέθοδοι συμπεριφέρονται με παρόμοιο τρόπο
- **Στο σημείο 5k:** εμφανίζεται η πρώτη απόκλιση με το HoeffdingAdaptiveTreeClassifier να ανακάμπτει γρηγορότερα
- **Στο σημείο 15k:** Παρατηρείται ίδια συμπεριφορά
- **Μετά την απόκλιση στα 10k:** το HoeffdingTreeClassifier είναι καλύτερο για μικρή περίοδο αλλά μετά γρήγορα ξεπερνιέται.
- **HoeffdingAdaptiveTreeClassifier:**
  - επιτυγχάνει καλύτερη απόδοση
  - απαιτεί λιγότερο χώρο στη μνήμη

Συμπεράσματα

- **Παράδειγμα εργασίας ταξινόμησης**
  - Χρησιμοποιήθηκε η γεννήτρια SEA
  - Ο ταξινομητής Naive Bayes πέτυχε ακρίβεια 93.95%
  - Σύγκριση ταξινομητών NaiveBayes και SGDClassifier:
    - Ο NaiveBayes παρουσίασε καλύτερη απόδοση στην αρχή της ροής (γρηγορότερος και δεσμεύει ελαφρώς περισσότερη μνήμη)
    - Ο SGDClassifier τελικά τον ξεπερνά (αργότερος με μικρότερο αποτύπωμα μνήμης)

Συμπεράσματα

- **Παράδειγμα επίδρασης απόκλισης στη μάθηση:**
  - Συγκρίθηκαν τα μοντέλα ροής:
    - **HoeffdingTreeClassifier**
    - **HoeffdingAdaptiveTreeClassifier** (λαμβάνει υπόψη την απόκλιση)
    - Το HoeffdingAdaptiveTreeClassifier πέτυχε καλύτερη απόδοση ενώ απαιτούσε λιγότερο χώρο στη μνήμη.
    - Εφαρμογή μηχανισμού αντικατάστασης κλάδου που ενεργοποιείται από το ADWIN
    - Λιγότερο περίπλοκη δομή δέντρου αντιπροσώπευσης των δεδομένων

Περιορισμοί Έρευνας

- Η εργασία περιορίστηκε σε δεδομένα που αναζητήθηκαν από μελέτες της διεθνούς βιβλιογραφίας
- Δεν υπήρξε παραγωγή πρωτογενών δεδομένων
- Μελλοντικά θα μπορούσε να πραγματοποιηθεί σύγκριση πρωτογενών αλγορίθμων κατηγοριοποίησης του scikit-multiflow πάνω σε ροές δεδομένων.