ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΕΥΦΥΕΙΣ ΤΕΧΝΟΛΟΓΙΕΣ ΔΙΑΔΙΚΤΥΟΥ - WEBINTELLIGENCE

Εξόρυξη γνώσης από ροές δεδομένων μέσω του λογισμικού MOA

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

Σταμπόλκου Γρηγορία

Επιβλέπων : Στέφανος Ουγιάρογλου
Επίκουρος Καθηγητής, ΔΙ.ΠΑ.Ε

Θεσσαλονίκη, Ιούνιος 2023

Η σελίδα αυτή είναι σκόπιμα λευκή.

ΔΙΕΘΝΕΣ ΠΑΝΕΠΙΣΤΗΜΙΟ ΤΗΣ ΕΛΛΑΔΟΣ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ ΕΥΦΥΕΙΣ ΤΕΧΝΟΛΟΓΙΕΣ ΔΙΑΔΙΚΤΥΟΥ – WEBINTELLIGENCE

Τίτλος Μεταπτυχιακής Διπλωματικής Εργασίας

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

Σταμπόλκου Γρηγορία

Επιβλέπων : Στέφανος Ουγιάρογλου
Επίκουρος Καθηγητής, ΔΙ.ΠΑ.Ε.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή στις Choose a date.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Όνομα Επώνυμο	Όνομα Επώνυμο	'Ονομα Επώνυμο
Choose an item.ΔΙ.ΠΑ.Ε.	Choose an item.ΔΙ.ΠΑ.Ε.	Choose an item.ΔΙ.ΠΑ.Ε.

Θεσσαλονίκη, Choose a date

(Υπογραφή)

.....

Click here to enter text.

Click here to enter text.

Περίληψη

Στη δεκαετία του 1990 η εξόρυξη δεδομένων ήταν μια συναρπαστική νέα ιδέα που ασχολούταν με την ανακάλυψη γνώσης σε κρυφά μοτίβα. Το 2010 οι άνθρωποι άρχισαν να μιλούν για μεγάλα δεδομένα και για δυνητικά άπειρες και γρήγορες ακολουθίες δεδομένων, που ονομάζονται ροές δεδομένων, οι οποίες παράγονται από δυναμικά μεταβαλλόμενα περιβάλλοντα. Υπάρχουν πολλά παραδείγματα τέτοιων εφαρμογών, όπως η δημιουργία συστημάτων πρόβλεψης που βασίζονται σε οικονομικά δεδομένα, ζήτηση ενέργειας, παρακολούθηση δικτύου ιστού και αισθητήρων, εντοπισμού κακόβουλου λογισμικού. Στα δυναμικά μεταβαλλόμενα περιβάλλοντα η συνάρτηση πυκνότητας πιθανότητας της διαδικασίας δημιουργίας δεδομένων αλλάζει με την πάροδο του χρόνου, προκαλώντας το φαινόμενο της εννοιολογικής απόκλισης. Καταρρίπτοντας τη θεμελιώδης υπόθεση που γίνεται από τις περισσότερες προσεγγίσεις της μηχανικής μάθησης ότι τα δεδομένα εκπαίδευσης και δοκιμών δημιουργούνται από την ίδια, αν και άγνωστη, σταθερή κατανομή πιθανοτήτων. Η μάθηση σε δυναμικά μεταβαλλόμενα περιβάλλοντα απαιτεί προσαρμοστικές ή εξελισσόμενες προσεγγίσεις που μπορούν να παρακολουθούν μια εννοιολογική απόκλιση, σε πραγματικό χρόνο, και την προσαρμογή ενός μοντέλου μάθησης σε αυτήν.

Η παρούσα μεταπτυχιακή διπλωματική εργασία, μελετά την κατηγοριοποίηση ροών δεδομένων στοχεύοντας στη διαχείριση της εννοιολογικής απόκλισης. Παρουσιάζονται οι τεχνικές αναγνώρισης εννοιολογικής απόκλισης ενός αλγόριθμου κατηγοριοποίησης και οι συνδυαστικοί αλγόριθμοι μάθησης (ensembles). Επιπλέον γίνεται μια συγκριτική μελέτη στο πειραματικό περιβάλλον του MOA, ένα δημοφιλές περιβάλλον ανοιχτού κώδικα για την εξόρυξη ροών δεδομένων.

Λέξεις Κλειδιά: " εννοιολογική απόκλιση, κατηγοριοποίηση, ροές δεδομένων, MOA "

Η σελίδα αυτή είναι σκόπιμα λευκή.

Abstract

In the 1990s, data mining was an exciting new idea that dealt with discovering knowledge in hidden patterns. In 2010, people started talking about big data and potentially infinite and fast sequences of data called data streams, which are generated by dynamically changing environments. There are many examples of such applications, such as creating prediction systems based on economic data, energy demand, web and sensor network monitoring, and malware detection. In dynamically changing environments, the probability density function of the data generation process changes over time, leading to the phenomenon of concept drift. By challenging the fundamental assumption made by most machine learning approaches that training and test data are created by the same, albeit unknown, stationary distribution of probabilities, learning in dynamically changing environments requires adaptive or evolving approaches that can track concept drift in real-time and adapt a learning model to it.

This master's thesis focuses on the classification of data streams, aiming to manage concept drift. Techniques for recognizing conceptual drift in a learner and combined learning algorithms are presented. Furthermore, a comparative study is conducted in the experimental environment of MOA, a popular open-source framework for data streams mining.

Keywords: "concept drift, classification, data streams, MOA"

Ευχαριστίες

Η παρούσα διπλωματική εργασία αποτελεί την ολοκλήρωση των σπουδών μου στο Μεταπτυχιακό Πρόγραμμα Σπουδών στις Ευφυείς Τεχνολογίες Διαδικτύου (MSc in Web Intelligence) του τμήματος Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του Διεθνούς Πανεπιστημίου της Ελλάδος.

Ιδιαίτερες ευχαριστίες θα ήθελα να δώσω στον επιβλέποντα καθηγητή της διπλωματικής μου εργασίας κ. Ουγιάρογλου Στέφανο, για την καθοδήγηση, την υπομονή του, τις πολύτιμες συμβουλές του, την άμεση βοήθεια όταν χρειάστηκε καθώς και τη στήριξη που μου προσέφερε καθ' όλη τη διάρκεια της εκπόνησης της εργασίας.

Η σελίδα αυτή είναι σκόπιμα λευκή.

Πίνακας περιεχομένων

1 Εισαγωγή 1

1.1 Ανάλυση σε πραγματικό χρόνο 1

1.2 Ροές δεδομένων 2

1.3 Κίνητρο και συνεισφορά 3

1.4 Οργάνωση κειμένου 5

2 Εξόρυξη ροών δεδομένων 6

2.1 Εξόρυξη ροών δεδομένων 6

2.2 Απαιτήσεις αλγορίθμων εξόρυξης ροών δεδομένων 7

2.3 Κατηγοριοποίηση και Ροές δεδομένων 8

2.4 Μετρικές απόδοσης 11

3 Εννοιολογική απόκλιση (Concept drift) 13

3.1 Εννοιολογική απόκλιση 13

3.2 Ειδή Εννοιολογικής απόκλισης 14

3.2.1 Ειδή Εννοιολογικής απόκλισης σχεικά με την ταχύτητα τους 15

3.3 Επιθυμητές ιδιότητες ενός συστήματος διαχείρισης Εννοιολογικής απόκλισης 17

3.4 Διαχείριση της Εννοιολογικής απόκλισης 17

3.5 Διαχείριση δεδομένων στη μνήμη 19

3.6 Μέθοδοι Προσαρμογής 20

3.7 Τεχνικές αναγνώρισης Εννοιολογικής απόκλισης 21

3.7.1 Εκτιμητές 22

3.7.2 Ανιχνευτές αλλαγών 24

4 Αλγόριθμοι κατηγοριοποίησης ροών δεδομένων 31

4.1 Βασικοί κατηγοριοποιητές 31

4.1.1 Κατηγοριοποιητής πλειοψηφίας 31

4.1.2 Κατηγοριοποιητής χωρίς αλλαγή 31

4.1.3 Naïve Bayes (NB) 31

4.2 Δέντρα αποφάσεων 32

4.2.1 Hoeffding Tree (HT) 32

4.2.2 Hoeffding Adaptive Tree (HAT) 35

4.3 SingleClassifierDrift 35

4.4 Συνδυαστικοί αλγόριθμοι μάθησης(ensembles) 35
4.4.1 Accuracy Weighted Ensemble (AWE) *36*
4.4.2 Accuracy Updated Ensemble (AUE) *37*
4.4.3 Dynamic Weighted Majority (DWM) *38*
4.4.4 Paired Learners *39*
4.4.5 Learn++.NSE *39*
4.4.6 Hoeffding option tree (HOT) *40*
4.4.7 Adaptive Hoeffding option tree(AdaHOT) *41*
4.4.8 Bagging *41*
4.4.9 Online Bagging (OzaBag) *42*
4.4.10 ADWIN Bagging (OzaBagAdwin) *42*
4.4.11 Leveraging Bagging *43*
4.4.12 Bagging using ASHTs of different sizes(OzaBagASHT) *43*
4.4.13 Online Boosting(OzaBoost) *44*
4.4.14 OCBoost *45*
5 Το περιβάλλον εκτέλεσης πειραμάτων MOA 46
5.1 Εισαγωγή στο περιββάλον του MOA 46
5.2 Συνθετικά Δεδομένα 50
5.3 Πραγματικά Δεδομένα 52
6 Πειραματική μελέτη σε συνθετικά δεδομένα 53
6.1 Εγκαθίδρυση πειραμάτων 53
6.1.1 Προσομοίωση Απότομων και Σταδιακών Εννοιολογικών Αποκλίσεων 53
6.1.2 Πειραματικά αποτελέσματα 55
6.1.3 Συζήτηση 56
6.1.3.1 Προσομοίωση Επαναλαμβανόμενων Εννοιολογικών Αποκλίσεων 58
6.2.1 Προσομοίωση Βαθμιαία Σταδιακών Εννοιολογικών Αποκλίσεων 60
6.2.2 Πειραματικά αποτελέσματα 60
6.2.3 Συζήτηση 61
6.3.1 Προσομοίωση Βαθμιαία Σταδιακών Εννοιολογικών Αποκλίσεων 63
6.3.2 Πειραματικά αποτελέσματα 64
6.3.3 Συζήτηση 64
7 **Πειραματική μελέτη σε πραγματικά δεδομένα** 65
7.1 Εγκαθίδρυση πειραμάτων 66
7.1.1 Πειραματικά αποτελέσματα (poker) *66*

7.1.2 Συζήτηση *67*

7.2 Εγκαθίδρυση πειραμάτων(elec) *68*

7.2.1 Πειραματικά αποτελέσματα (elec) *68*

7.2.2 Συζήτηση *69*

8 Συμπεράσματα και μελλοντική έρευνα **70**

9 Βιβλιογραφία **71**

Κεφάλαιο 1 –Εισαγωγή

1.1 Ανάλυση σε πραγματικό χρόνο

Στον ψηφιακό κόσμο, δημιουργείται καθημερινά ένας τεράστιος όγκος δεδομένων από όλα τα είδη συσκευών, σε διαφορετικές μορφές, από ανεξάρτητες ή συνδεδεμένες εφαρμογές[10]. Οι πρόσφατες εξελίξεις στην τεχνολογία του υλικού και του λογισμικού επιτρέπουν τη σύλληψη διαφορετικών μετρήσεων δεδομένων σε ένα ευρύ φάσμα πεδίων. Αυτές οι μετρήσεις παράγονται συνεχώς από πολλές πηγές και σε πολύ υψηλούς κυμαινόμενους ρυθμούς δεδομένων από μη σταθερές διανομές[52]. Παραδείγματα τέτοιων πηγών είναι η κίνηση στο Διαδίκτυο και στον Παγκόσμιο Ιστό, δεδομένα GPS, κλήσεις κινητών τηλεφώνων, ηλεκτρονική αλληλογραφία, δεδομένα δικτύων αισθητήρων, ροές κλικ πελατών κ.λπ[14]. Επιπλέον, η ταχεία επέκταση των δεδομένων επιταχύνεται από τη δραματική αύξηση της αποδοχής των εφαρμογών κοινωνικής δικτύωσης, οι οποίες επιτρέπουν στους χρήστες να δημιουργούν ελεύθερα περιεχόμενο και να αυξάνουν το ήδη τεράστιο μέγεθος του Ιστού[10].

Αν και η εξόρυξη δεδομένων έχει γίνει πλέον ένα αρκετά καλά εδραιωμένο πεδίο, αυτή η πλημμύρα των μεγάλων δεδομένων έχει ξεπεράσει την ικανότητά των παραδοσιακών Συστημάτων Διαχείρισης Βάσεων Δεδομένων να επεξεργάζονται, να αναλύουν, να αποθηκεύουν και να κατανοούν αυτά τα σύνολα δεδομένων. Το ψηφιακό σύμπαν το 2007 υπολογίστηκε [42] ότι είναι 281 exabytes ή 281 δισεκατομμύρια gigabytes, και υποστηρίχτηκε ότι μέχρι το 2011, το ψηφιακό σύμπαν θα έχει 10 φορές μεγαλύτερο μέγεθος. Το 2007, για πρώτη φορά, ο όγκος των πληροφοριών που δημιουργήθηκαν ή καταγράφηκαν ξεπέρασε τον διαθέσιμο χώρο αποθήκευσης. Σύμφωνα με πιο πρόσφατη έρευνα [43] μεταξύ του 2018 και του 2025, το μέγεθος των δεδομένων που παράγονται σε πραγματικό χρόνο παγκοσμίως αναμένεται να δεκαπλασιαστεί, από 5 zettabytes σε 51 zettabytes ή 51.000 δισεκατομμύρια gigabytes.

Για να αντιμετωπιστεί αυτός ο εκπληκτικά μεγάλος όγκος δεδομένων, χρειάζονται γρήγορες και αποτελεσματικές μέθοδοι που λειτουργούν σε πραγματικό χρόνο χρησιμοποιώντας ένα λογικό ποσό υπολογιστικών πόρων. Είναι σημαντικό για τους οργανισμούς όχι μόνο να λαμβάνουν απαντήσεις σε ερωτήματα αμέσως, αλλά να το κάνουν σύμφωνα με τα δεδομένα που μόλις έφτασαν[10].

1.2 Ροές δεδομένων

Οι ροές δεδομένων είναι μια αλγοριθμική αφαίρεση για την υποστήριξη αναλυτικών στοιχείων σε πραγματικό χρόνο. Είναι ακολουθίες στοιχείων, πιθανώς άπειρες, με κάθε στοιχείο να έχει μια χρονική σήμανση, και έτσι μια χρονική σειρά. Τα στοιχεία δεδομένων φτάνουν ένα προς ένα, ενώ είναι επιθυμητό να δημιουργηθούν και να διατηρηθούν μοντέλα μάθησης, των στοιχείων αυτών, σε πραγματικό χρόνο. Υπάρχουν δύο κύριες αλγοριθμικές προκλήσεις κατά τον χειρισμό των δεδομένων μιας ροής, η ροή είναι μεγάλη και γρήγορη και οφείλεται να εξαχθούν πληροφορίες σε πραγματικό χρόνο από αυτήν. Αυτό σημαίνει συνήθως ότι είναι αποδεκτές κατά προσέγγιση λύσεις με σκοπό τη χρήση λιγότερης μνήμης και του λιγότερου δυνατού χρόνου[10].

Επίσης, μια τρίτη μεγάλη πρόκληση είναι ότι τα δεδομένα μιας ροής δεδομένων ενδέχεται να εξελίσσονται ή να αλλάζουν με την πάροδο του χρόνου προκαλώντας το φαινόμενο της εννοιολογικής απόκλισης(concept drift). Αυτή η εξέλιξη της ροής δεδομένων μπορεί να επηρεάσει την απόδοση των αλγορίθμων εξόρυξης δεδομένων, καθώς τα αποτελέσματα μπορεί να γίνουν απαρχαιωμένα με την πάροδο του χρόνου. Επομένως, τα μοντέλα μάθησης πρέπει να προσαρμόζονται όταν υπάρχουν αλλαγές στα δεδομένα[47].

Η ακρίβεια, ο χρόνος και η μνήμη είναι οι τρεις κύριες διαστάσεις που χρειάζεται να ληφθούν υπόψη κατά τη διαδικασίας εξόρυξης ροών δεδομένων. Αποδεκτές μέθοδοι δημιουργίας μοντέλων είναι αυτές που επιτυγχάνουν τη μέγιστη ακρίβεια σε ελάχιστο χρόνο και χρήση χαμηλής συνολικής μνήμης. Αξιοσημείωτο είναι επίσης ότι ενώ τα δεδομένα φτάνουν με μεγάλη ταχύτητα, δεν μπορούν να αποθηκευτούν στην προσωρινή μνήμη, οπότε ο χρόνος επεξεργασίας ενός στοιχείου είναι εξίσου σημαντικός με τον συνολικό χρόνο, ο οποίος είναι αυτός που συνήθως λαμβάνεται υπόψη στη συμβατική εξόρυξη δεδομένων[10].

1.3 Κίνητρο και Συνεισφορά

Το πρόβλημα της κατηγοριοποίησης ροών δεδομένων είναι ένα από τα πιο ευρέως μελετημένα στο πλαίσιο της εξόρυξης ροών δεδομένων. Το πρόβλημα αυτό γίνεται πιο δύσκολο από την εξέλιξη ή αλλαγή της κατανομής μιας ροής δεδομένων. Επομένως, οφείλεται να σχεδιαστούν αποτελεσματικοί αλγόριθμοι, ώστε να προσαρμόζονται στο φαινόμενο μιας εννοιολογικής απόκλισης[47]. Έχουν προταθεί πολλές τεχνικές αναγνώρισης εννοιολογικής απόκλισης που παρακολουθούν και να αναλύουν τη φύση αυτών των αλλαγών με την πάροδο του χρόνου. Επιπλέον, έχει προταθεί ένας μεγάλος αριθμός συνδυαστικών αλγορίθμων μάθησης οι όποιες αναγνωρίζουν μια εννοιολογική απόκλιση είτε ακολουθώντας μια αντιδραστική προσέγγιση (AUE, DWE, AWE) είτε μια ενεργητική προσέγγιση (Bagging using ADWIN, Leveraging Bagging, ADWIN Bagging)

Στην παρούσα διπλωματική γίνεται μια προσπάθεια ερμηνείας του φαινομένου της εννοιολογικής απόκλισης και των ειδών της (απότομη, σταδιακή, βαθμιαία σταδιακή επαναλαμβανόμενη). Παρουσιάζονται οι τεχνικές αναγνώρισης εννοιολογικής απόκλισης ενός αλγόριθμου κατηγοριοποίησης και των συνδυαστικών αλγορίθμων μάθησης. Ακολουθεί μια συγκριτική πειραματική μελέτη με τη χρήση του λογισμικού MOA, ένα από τα πιο δημοφιλές εργαλεία ανοιχτού κώδικα, εξόρυξης ροών δεδομένων. Στόχος της παρούσας διπλωματικής είναι να αναδειχτούν οι καλύτερες τεχνικές για κάθε είδος εννοιολογικής απόκλισης.

Η συνεισφορά της διπλωματικής συνοψίζεται ως εξής:

1. Προσομοιώθηκαν 3 απότομες εννοιολογικές αποκλίσεις σε μια ροη δεδομένων και 3 σταδιακές διαφορετικού μεγέθους σε μια άλλη ροή, χρησιμοποιήθηκαν τα συνθετικά δεδομένα των γεννητριών SEA και AGRAWEL.
2. Δοκιμάστηκαν οι τεχνικές αναγνώρισης εννοιολογικής απόκλισης (DDM, EDDM και όσες παρουσιάστηκαν στην ενότητα 3.7) και οι συνδυαστικοί αλγόριθμοι μάθησης (ensembles) στο γραφικό περιβάλλον του MOA με τη μέθοδο κατηγοριοποίησης SingleClassifierDrift και βασικό μαθητή τον Naive Bayes.
3. Αξιολογήθηκε η επίδοση των τεχνικών αναγνώρισης και βρέθηκε ότι όλες οι μέθοδοι έχουν μια καλή απόδοση σε μια απότομη εννοιολογική απόκλιση και ότι η απόδοση τους έχει κάποιες διαφορές στην περίπτωση εμφάνισης σταδιακών εννοιολογικών αποκλίσεων, εξαρτάται από τη συχνότητα τους και το μέγεθος των παραδειγμάτων τους. Οι τεχνικές αυτές, σε πολλές περιπτώσεις έχουν μια καλύτερη απόδοση από τους αλγόριθμους συνδυαστικούς μάθησης(ensembles).

4. Προσομοιώθηκε μια ροή δεδομένων με βαθμιαίες σταδιακές αποκλίσεις και μια με πολύ λεπτές σταδιακές (ανεπαίσθητες), χρησιμοποιήθηκαν τα συνθετικά δεδομένα της γεννήτριας HYPERPLANE.
5. Δοκιμάστηκαν οι αλγόριθμοι συνδυαστικής μάθησης (AUE, AWE, DMW, Leveraging Bagging και όσες αναφέρονται στην ενότητα 4.2) στο γραφικό περιβάλλον του MOA.
6. Αξιολογήθηκε η επίδοση των αλγορίθμων και βρέθηκε ότι οι παθητικές μέθοδοι ,σε αρκετές περιπτώσεις, έχουν μια καλύτερη απόδοση από πολλές ενεργητικές μεθόδους
7. Προσομοιώθηκαν δυο ροές, μια ροή με γρήγορες βαθμιαίες σταδιακές εννοιολογικές αποκλίσεις και μια ροη μετρίας ταχύτητας.
8. Δοκιμάστηκαν οι αλγόριθμοι συνδυαστικής μάθησης (ενότητα 4.2) στο γραφικό περιβάλλον του MOA.
9. Αξιολογήθηκε η επίδοση των αλγορίθμων και βρήκαμε ότι οι ευεργετικές μέθοδοι έχουν, συνήθως, καλύτερη απόδοση σε μια γρήγορη ροή με εννοιολογικές αποκλίσεις.
10. Αξιολογήθηκαν όλοι οι αλγόριθμοι κατηγοριοποίησης που έχουν αναφερθεί σε πραγματικά δεδομένα του συνόλου poker και Elec. Δόθηκε ένα γενικό πλαίσιο των μετρικών αξιολόγησης και των συνθηκών που επηρεάζουν την απόδοση ενός μοντέλου.

1.4 Οργάνωση κειμένου

Η διπλωματική εργασία είναι δομημένη στα εξής κεφάλαια:

Στο Κεφάλαιο 2, γίνεται μια αναφορά στις απαιτήσεις που οφείλουν να πληρούν οι αλγόριθμοι εξόρυξης ροών δεδομένων, περιγράφεται η κατηγοριοποίηση ροών δεδομένων και οι τεχνικές αξιολόγησης της.

Στο Κεφάλαιο 3, γίνεται αναφορά στο φαινόμενο της εννοιολογικής απόκλισης, τα είδη της, στη προσαρμοστική μάθηση και στις τεχνικές αναγνώρισης εννοιολογικής απόκλισης, που έχουν προταθεί.

Στο Κεφάλαιο 4, παρουσιάζονται οι αλγόριθμοι κατηγοριοποίησης ενός αλγόριθμου κατηγοριοποίηση, καθώς και οι αλγόριθμοι συνδυαστικής μάθησης της κατηγοριοποίησης.

Στο Κεφάλαιο 5, παρουσιάζεται το γραφικό πειραματικό περιβάλλον του MOA, ένα από τα πιο δημοφιλής εργαλεία ανοιχτού κώδικα. Επιπλέον, παρουσιάζονται τα συνθετικά

δεδομένα και τα δεδομένα πραγματικού κόσμου που θα χρησιμοποιηθούν για πειραματικές μελέτες στα κεφάλαια 6 και 7.

Στο Κεφάλαιο 6, γίνεται μια πειραματική μελέτη σε συνθετικά δεδομένα που είναι ενσωματωμένα στο ΜΟΑ. Προσομοιώνονται τα είδη μιας εννοιολογικής απόκλισης (απότομη, σταδιακή, βαθμιαία σταδιακή και επαναλαμβανόμενη).

Στο Κεφάλαιο 7, γίνεται μια πειραματική μελέτη πάνω σε δεδομένα πραγματικού κόσμου, τα όποια περιέχουν εννοιολογικές αποκλίσεις.

Στο κεφάλαιο 8, εξάγονται κάποια συμπεράσματα.

Κεφάλαιο 2

2.1 Εξόρυξη ροών Δεδομένων

Οι κύριοι αλγόριθμοι στην εξόρυξη ροής δεδομένων είναι η κατηγοριοποίηση, η παλινδρόμηση, η συσταδοποίηση και η εξόρυξη κανόνων συσχέτισης.

Έχουμε υποθετικά μια ροή στοιχείων, τα οποία ονομάζονται επίσης πρότυπα ή παραδείγματα, που φτάνουν συνεχώς. Στη ρύθμιση της κατηγοριοποίησης μιας ροής απαιτείται να εκχωρήσουμε μια ετικέτα σε κάθε στοιχείο επιλέγοντας μέσα από ένα σύνολο ονομαστικών ετικετών, ως συνάρτηση των άλλων χαρακτηριστικών του αντικειμένου. Ένας κατηγοριοποιητής μπορεί να εκπαιδευτεί, εφόσον η σωστή ετικέτα για πολλά από τα παραδείγματα είναι αργότερα διαθέσιμη. Ένα παράδειγμα κατηγοριοποίησης είναι η επισήμανση των εισερχόμενων μηνυμάτων ηλεκτρονικού ταχυδρομείου ως ανεπιθύμητων ή μη. Η παλινδρόμηση είναι μια εργασία πρόβλεψης παρόμοια με την κατηγοριοποίηση, με τη διαφορά ότι η ετικέτα που πρέπει να προβλεφθεί είναι μια αριθμητική τιμή αντί για μια ονομαστική. Ένα παράδειγμα παλινδρόμησης είναι η πρόβλεψη της αξίας μιας μετοχής στο χρηματιστήριο για την επόμενη ημέρα.

Η κατηγοριοποίηση και η παλινδρόμηση χρειάζονται ένα σύνολο σωστά επισημασμένων παραδειγμάτων για να εκπαιδευτεί ένα μοντέλο, έτσι ώστε να μπορεί να χρησιμοποιηθεί για να προβλέψει τις ετικέτες των αθέατων παραδειγμάτων. Είναι τα κύρια παραδείγματα εποπτευόμενων μαθησιακών εργασιών. Όταν τα παραδείγματα δεν επισημαίνονται, μια ενδιαφέρουσα εργασία είναι να ομαδοποιηθούν σε ομοιογενής συστάδες. Η συσταδοποίηση μπορεί να χρησιμοποιηθεί, για παράδειγμα, όπως η λήψη προφίλ των χρηστών σε έναν ιστότοπο, αποτελώντας παράδειγμα μιας μη εποπτευόμενης μαθησιακής εργασίας.

Η συχνή εξόρυξη κανόνων συσχέτισης αναζητά τις πιο σχετικές συσχετίσεις μέσα στα παραδείγματα. Για παράδειγμα, σε ένα σύνολο δεδομένων σούπερ μάρκετ πωλήσεων, είναι δυνατόν να είναι γνωστό ποια είδη αγοράζονται μαζί και να δημιουργηθούν κανόνες

συσχέτισης, όπως για παράδειγμα, τις περισσότερες φορές οι πελάτες αγοράζουν τυρί, αγοράζουν επίσης κρασί[10].

2.2 Απαιτήσεις αλγορίθμων εξόρυξης ροών

Οι πιο σημαντικές απαιτήσεις για έναν αλγόριθμο εξόρυξης ροών είναι οι ίδιες για τα προγνωστικά μοντέλα, τα μοντέλα συσταδοποίησης και της εξόρυξης κανόνων συσχέτισης :

- **Απαίτηση 1: Σταδιακή επεξεργασία παραδειγμάτων.**

Το βασικό χαρακτηριστικό μιας ροής δεδομένων είναι ότι τα δεδομένα «ρέουν», το ένα παράδειγμα ακολουθεί το άλλο. Δεν υπάρχει δικαίωμα για τυχαία πρόσβαση στα δεδομένα που παρέχονται. Κάθε παράδειγμα πρέπει να γίνει αποδεκτό, καθώς φτάνει και με τη σειρά που φτάνει, ενώ, αφού επιθεωρηθεί ή αγνοηθεί, ένα παράδειγμα απορρίπτεται χωρίς δυνατότητα ανάκτησής του ξανά.

Αν και αυτή η απαίτηση υπάρχει στην είσοδο ενός αλγόριθμου, δεν υπάρχει κανόνας που να εμποδίζει έναν αλγόριθμο να θυμάται εσωτερικά παραδείγματα βραχυπρόθεσμα. Ένα παράδειγμα αυτού μπορεί να είναι ο αλγόριθμος που αποθηκεύει μια παρτίδα παραδειγμάτων για χρήση από ένα συμβατικό αλγόριθμο κατηγοριοποίησης. Ωστόσο, ο αλγόριθμος είναι ελεύθερος να λειτουργεί με αυτόν τον τρόπο, θα πρέπει να απορρίψει αποθηκευμένα παραδείγματα σε κάποιο σημείο εάν θέλει να συμμορφωθεί με την απαίτηση 2.

Ο κανόνας μιας επιθεώρησης μπορεί να χαλαρώσει μόνο σε περιπτώσεις όπου είναι πρακτικό να αποσταλεί εκ νέου ολόκληρη η ροή, που ισοδυναμεί με πολλαπλές σαρώσεις σε μια βάση δεδομένων. Σε αυτή την περίπτωση μπορεί να δοθεί μια ευκαιρία σε έναν αλγόριθμο κατά τη διάρκεια των επόμενων περασμάτων να βελτιώσει το μοντέλο που έχει εκπαιδεύσει. Ωστόσο, ένας αλγόριθμος που απαιτεί περισσότερα από ένα πέρασμα για να λειτουργήσει δεν είναι αρκετά ευέλικτος για καθολική εφαρμογή σε ροές δεδομένων.

- **Απαίτηση 2: Χρήση περιορισμένης ποσότητα μνήμης.**

Το κύριο κίνητρο για τη χρήση του μοντέλου ροής δεδομένων είναι ότι επιτρέπει την επεξεργασία δεδομένων που είναι πολλές φορές μεγαλύτερα από τη διαθέσιμη μνήμη εργασίας. Ο κίνδυνος με την επεξεργασία τόσο μεγάλων ποσοτήτων δεδομένων είναι ότι η μνήμη εξαντλείται εύκολα εάν δεν υπάρχει σκόπιμα καθορισμένο όριο στη χρήση της.

Η μνήμη που χρησιμοποιείται από έναν αλγόριθμο μπορεί να χωριστεί σε δύο κατηγορίες: τη μνήμη (προσωρινή) που χρησιμοποιείται για την αποθήκευση στατιστικών στοιχείων που εκτελούνται και τη μνήμη που χρησιμοποιείται για την αποθήκευση του τρέχοντος

μοντέλου. Ο αποδοτικότερος από άποψη μνήμης αλγόριθμος θα είναι ένα αυτός που τα τρέχοντα στατιστικά αποτελούν άμεσα το μοντέλο που χρησιμοποιείται για την πρόβλεψη.

Αυτός ο περιορισμός μνήμης είναι ένας φυσικός περιορισμός που μπορεί να χαλαρώσει μόνο εάν χρησιμοποιείται εξωτερικός χώρος αποθήκευσης, για παράδειγμα προσωρινά αρχεία. Οποιαδήποτε τέτοια λύση πρέπει να γίνει λαμβάνοντας υπόψη την απαίτηση 3.

- **Απαίτηση 3: Επεξεργασία ενός παραδείγματος σε περιορισμένο χρόνο.**

Προκειμένου ένας αλγόριθμος να μπορεί να κλιμακώνεται άνετα σε οποιονδήποτε αριθμό παραδειγμάτων, η πολυπλοκότητα του χρόνου εκτέλεσης πρέπει να είναι γραμμική ως προς τον αριθμό των παραδειγμάτων. Αυτό μπορεί να επιτευχθεί στη ρύθμιση ροής δεδομένων, εάν υπάρχει ένα σταθερό, κατά προτίμηση μικρό, ανώτερο όριο στην ποσότητα επεξεργασίας ανά παράδειγμα.

Επιπλέον, εάν ένας αλγόριθμος πρόκειται να είναι ικανός να λειτουργεί σε πραγματικό χρόνο, χρειάζεται να επεξεργαστεί τα παραδείγματα τόσο γρήγορα, αν όχι ταχύτερα από ο,τι φτάνουν. Αν δεν πραγματοποιηθεί αυτό σημαίνει αναπόφευκτα απώλεια δεδομένων.

Ο απόλυτος χρονισμός δεν είναι τόσο κρίσιμος σε λιγότερο απαιτητικές εφαρμογές, όπως όταν ο αλγόριθμος χρησιμοποιείται για την κατηγοριοποίηση μιας μεγάλης αλλά επίμονης πηγής δεδομένων. Ωστόσο, όσο πιο αργός είναι ο αλγόριθμος, τόσο μικρότερη αξία θα έχει για χρήστες που απαιτούν αποτελέσματα σε εύλογο χρονικό διάστημα.

- **Απαίτηση 4: Άμεση ανατροφοδότηση ανά πάσα στιγμή.**

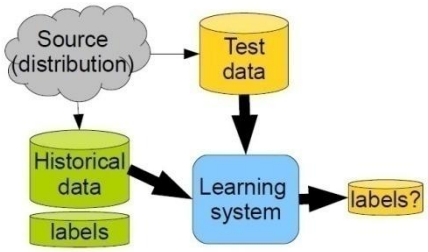
Ένας ιδανικός αλγόριθμος θα πρέπει να είναι ικανός να παράγει το καλύτερο μοντέλο εκπαίδευσης που μπορεί από τα δεδομένα που έχει παρατηρήσει, αφού δει οποιονδήποτε αριθμό παραδειγμάτων. Στην πράξη είναι πιθανό ότι θα υπάρξουν περίοδοι όπου το μοντέλο εκπαίδευσης παραμένει σταθερό, όπως όταν ένας αλγόριθμος που βασίζεται σε παρτίδες αποθηκεύει την επόμενη παρτίδα[10][12][13].

- **Απαίτηση 5: Προσαρμογή στις αλλαγές που προκύπτουν με την πάροδο του χρόνου**

Μια ροή δεδομένων μπορεί να αλλάζει καθώς εξελίσσεται με την πάροδο του χρόνου. Έτσι, δεδομένα από το παρελθόν μπορεί να γίνουν άσχετα (ή ακόμα και επιβλαβή) για την τρέχουσα φύση της ροής. Τα μοντέλα εκπαίδευσης οφείλουν ανά πάσα στιγμή να προσαρμόζονται στην αλλαγή [10][13].

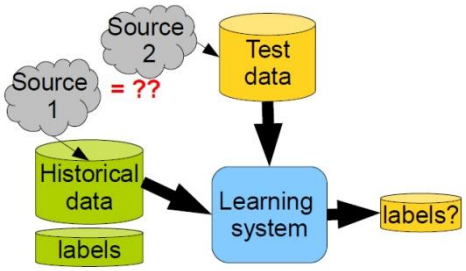
2.3 Κατηγοριοποίηση και Ροές δεδομένων

Στην παραδοσιακή κατηγοριοποίηση σε παρτίδες (batch) ή εκτός σύνδεσης, δίνεται σε έναν αλγόριθμο κατηγοριοποίησης ένα σύνολο παραδειγμάτων με ετικέτα. Ο αλγόριθμος δημιουργεί ένα μοντέλο κατηγοριοποίησης. Στη συνέχεια, το μοντέλο αναπτύσσεται, δηλαδή χρησιμοποιείται για την πρόβλεψη της ετικέτας για παραδείγματα χωρίς ετικέτα, που δεν έχει δει. Μια καλή και ευρέως χρησιμοποιούμενη μεθοδολογία είναι να χωριστεί το σύνολο δεδομένων, το οποίο είναι από την αρχή διαθέσιμο και μπορεί να αποθηκευτεί, σε δύο μέρη, στο σύνολο εκπαίδευσης και στο σύνολο δοκιμής, ή η χρήση της διασταυρούμενης επικύρωσης (cross validation), για να εξασφαλιστεί ότι το μοντέλο κατηγοριοποίησης είναι αρκετά ακριβές. Σε κάθε περίπτωση, υπάρχει μια πρώτη φάση εκπαίδευσης και δοκιμής, σαφώς χρονικά διαχωρισμένη από τη φάση πρόβλεψης.



Εικόνα 1 : Παραδοσιακή Κατηγοριοποίηση [27].

Δίνεται ένα πεπερασμένο σύνολο εκπαίδευσης D={(x,y)} , όπου y={y_1, y_2, …, y_k}, |D|=n, βρείτε μια συνάρτηση y=f(x) που μπορεί να προβλέψει την τιμή y για ένα αόρατο παράδειγμα x



Εικόνα 2: Κατηγοριοποίηση σε περιβάλλον Ροή [27].

Δίνεται μια άπειρη ακολουθία ζευγών της μορφής (x,y) όπου y={y_1, y_2, …, y_k}, βρείτε μια συνάρτηση y=f(x) που μπορεί να προβλέψει την τιμή y για ένα αόρατο παράδειγμα x.

Στο διαδικτυακό περιβάλλον, και ιδίως στη ροή δεδομένων, αυτός ο διαχωρισμός μεταξύ εκπαίδευσης, αξιολόγησης και δοκιμών είναι πολύ λιγότερο σαφής και παρεμβάλλεται. Συνεπώς, ο κατηγοριοποιητής οφείλει να κάνει προβλέψεις πριν εισαχθούν σε αυτόν όλα τα δεδομένα, αφού τα παραδείγματα μιας ροής μπορεί να είναι άπειρα και να μην τελειώσουν ποτέ. Επομένως, λαμβάνει τα δεδομένα των οποίων την ετικέτα προβλέπει, για να συνεχίσει να εκπαιδεύει το μοντέλο. Η διαδικασία αυτή, πρώτα δοκιμής και μετά εκπαίδευσης, που ακολουθεί ένας αλγόριθμος κατηγοριοποίησης, είναι γνωστή ως test-then-train. Αυτή η διαδικασία, εγείρει αλλά ζητήματα, όπως η συχνή αξιολόγηση του μοντέλου και η ανάγκη επανεκπαίδευσης του σε τακτά διαστήματα. Σε γενικές γραμμές, ένας κατηγοριοποιητής εξόρυξης ροών είναι έτοιμος να κάνει, ένα από τα ακόλουθα, ανά πάσα στιγμή:

- Λήψη ενός παραδείγματος χωρίς ετικέτα και πρόβλεψη της ετικέτας του με βάση το τρέχον μοντέλο του.

- Λήψη της ετικέτας για ένα παράδειγμα που έχει δει στο παρελθόν και χρήση της για την προσαρμογή του μοντέλου, δηλαδή για εκπαίδευση.

Ένας αλγόριθμος κατηγοριοποίησης εκτελεί τον ακόλουθο βρόγχο:

- Λήψη ενός παραδείγματος χωρίς ετικέτα *x*.
    - Πρόβλεψη της ετικέτας $\hat{y} = f(x)$ του *x*, όπου *f* είναι το τρέχον μοντέλο εκπαίδευσης.
    - Λήψη της πραγματικής ετικέτας *y* για το *x*.
    - Χρήση του ζεύγους (*x, y*) για την ενημέρωση (εκπαίδευση) *του μοντέλου f* και χρήση του ζεύγος ($\hat{y}$, *y*) για την ενημέρωση των στατιστικών στοιχείων (που είναι σχετικά με την απόδοση του κατηγοριοποιητή).
- Λήψη του επόμενου παραδείγματος[10].

Δεδομένου αυτού του βρόγχου, που ακολουθεί ένας αλγόριθμος κατηγοριοποίησης στην ενεργό μάθηση, προκύπτει το ακόλουθο ζήτημα. Πώς θα γίνει η αξιολόγηση ενός αλγορίθμου κατηγοριοποίησης; Η διαδικασία αξιολόγησης ενός αλγορίθμου μάθησης δίνει μια εκτίμηση του σφάλματος ενός κατηγοριοποιητή, καθορίζει ποια παραδείγματα χρησιμοποιούνται για την εκπαίδευση του αλγορίθμου και ποια χρησιμοποιούνται για τη δοκιμή της εξόδου του μοντέλου κατηγοριοποίησης από τον αλγόριθμο.

Το κύριο ζήτημα είναι να οικοδομηθεί μια ακριβή εικόνα της ακρίβειας, ενός κατηγοριοποιητή, με την πάροδο του χρόνου. Μια λύση είναι η λήψη παραδειγμάτων σε διαφορετικές χρονικές στιγμές, κατά τη διάρκεια της δημιουργίας του, για να βρεθεί το πως ποικίλει η ακρίβεια του. Προκύπτουν δύο κύριες προσεγγίσεις:

- **Holdout:** Η προσέγγιση αυτή, μετρά την απόδοση σε ένα τμήμα (holdout) της ροής. Είναι ιδιαίτερα χρήσιμη όταν η διαίρεση μεταξύ του συνόλου εκπαίδευσης και δοκιμών έχει προκαθοριστεί, έτσι ώστε τα αποτελέσματα από διαφορετικές μελέτες να μπορούν να συγκριθούν άμεσα. Ωστόσο, το holdout δίνει μια ακριβή εκτίμηση της τρέχουσας ακρίβειας ενός κατηγοριοποιητή μόνο εάν το σύνολο holdout έχει παρόμοια κατανομή με τα τρέχοντα δεδομένα, κάτι που είναι δύσκολο να διασφαλιστεί σε ένα δυναμικό περιβάλλον.
- **Παρεμβαλλόμενη (Interleaved) test-then-train ή προκαταρκτική**: Κάθε μεμονωμένο παράδειγμα χρησιμοποιείται για τη δοκιμή **(test)** του μοντέλου πριν χρησιμοποιηθεί για εκπαίδευση **(train)** και από αυτό η ακρίβεια μπορεί να ενημερωθεί σταδιακά. Όταν η αξιολόγηση εκτελείται σκόπιμα με αυτή τη σειρά, το μοντέλο δοκιμάζεται πάντα σε παραδείγματα που δεν έχει δει. Αυτό το σχήμα έχει

το πλεονέκτημα ότι δεν απαιτεί το διαχωρισμό ενός συνόλου holdout για τη δοκιμή, αξιοποιώντας στο έπακρο τα διαθέσιμα δεδομένα. Εξασφαλίζει επίσης μια ομαλή γραφική παράσταση ακρίβειας με την πάροδο του χρόνου, καθώς κάθε μεμονωμένο παράδειγμα θα γίνεται όλο και λιγότερο σημαντικό για τον συνολικό μέσο όρο. Στην παρεμβαλλόμενη αξιολόγηση, όλα τα παραδείγματα που ο αλγόριθμος έχει δει μέχρι στιγμής, λαμβάνονται υπόψη για τον υπολογισμό της ακρίβειας, αυτά που βρίσκονται σε ένα παράθυρο ορόσημο (Landmark window). Σε ένα παράθυρο ορόσημο έχει σημασία όλη η ιστορία της ροής κανένα παράδειγμα δεν απορρίπτεται και όλα έχουν το ίδιο βάρος. Αντιθέτως στην προκαταρκτική, μόνο τα πιο πρόσφατα, αυτά που βρίσκονται σε ένα συρόμενο παράθυρο (sliding windows) ή έναν παράγοντα αποσύνθεσης. Τα μεγέθη του συρόμενου παραθύρου και ο συντελεστής αποσύνθεσης είναι παράμετροι[12].

2.4 Μετρικές απόδοσης

2.4. 1 Μέτρα αξιολόγησης των επιδόσεων

Ακρίβεια (Accuracy): Ένα πρώτο μέτρο εκτίμησης απόδοσής είναι ο λόγος των ορθών κατηγοριοποιημένων παραδειγμάτων προς το σύνολο των παραδειγμάτων:

$$Accuracy = \frac{TN+TP}{TN+TP+FN+FP} \quad (1)$$

Στατιστική Kappa: Η στατιστική Kappa είναι ένα πιο ευαίσθητο μέτρο για την ποσοτικοποίηση της προγνωστικής απόδοσης των κατηγοριοποιητών μιας ροής. Κανονικοποιεί την πραγματική ακρίβεια ενός κατηγοριοποιητή p_0 με αυτή ενός κατηγοριοποιητή που προβλέπει τυχαία p_c:

$$k = \frac{p_0 - p_c}{1 - p_c} \quad (2)$$

όπου, p_o είναι η προκαταρκτική ακρίβεια του κατηγοριοποιητή, p_c η υποθετική πιθανότητα ενός κατηγοριοποιητή που προβλέπει τυχαία[25].

Στατιστική Kappa Temp: Η στατιστική Kappa Temp εξετάζει την παρουσία χρονικών εξαρτήσεων στις ροές δεδομένων :

$$k_{temp} = \frac{p_0 - p_c^{'}}{1 - p_c^{'}} \text{ (3)}$$

όπου, p_o είναι η προκαταρκτική ακρίβεια του κατηγοριοποιητή, $p_c^{'}$ ο είναι ο κατηγοριοποιητής χωρίς αλλαγή [32].

2.4. 2 Μέτρο κόστους

Το ζήτημα της ταυτόχρονης μέτρησης τριών διαστάσεων αξιολόγησης έχει οδηγήσει σε ένα άλλο σημαντικό ζήτημα στην εξόρυξη ροής δεδομένων, δηλαδή, την εκτίμηση ενός συνδυασμού του κόστους εκτέλεσης των διαδικασιών μάθησης και πρόβλεψης, όσον αφορά το χρόνο και τη μνήμη. Για παράδειγμα, υπάρχουν πολλές επιλογές κόστους ενοικίασης:

- Κόστος ανά ώρα χρήσης: Το Amazon Elastic Compute Cloud (Amazon EC2) είναι μια διαδικτυακή υπηρεσία που παρέχει υπολογιστική ικανότητα με δυνατότητα αλλαγής μεγέθους στο cloud. Το κόστος εξαρτάται από το χρόνο και το μέγεθος του ενοικιαζόμενου μηχανήματος (για παράδειγμα, μικρό παράδειγμα με 2 GB μνήμης RAM, μεγάλο με 8 GB ή πολύ μεγάλο με 16 GB).
- Κόστος ανά ώρα και μνήμη που χρησιμοποιείται: Το GoGrid είναι μια διαδικτυακή υπηρεσία παρόμοια με το Amazon EC2, αλλά χρεώνει με ώρες RAM. Κάθε GB μνήμης RAM που αναπτύσσεται για 1 ώρα ισούται εξ ορισμού με 1 ώρα μνήμης RAM.

Η χρήση των ωρών μνήμης RAM, όπως ορίζονται ανωτέρω, εισήχθη στο [47] ως ένα μέτρο αξιολόγησης των πόρων που χρησιμοποιούνται από αλγόριθμους ροής. Αν και προτείνεται για τη διαδικασία της εκπαίδευσης, μπορεί να εφαρμοστεί και στις άλλες εργασίες εξόρυξης δεδομένων[10].

Κεφάλαιο 3 - Εννοιολογική Απόκλιση

3.1 Εννοιολογική Απόκλιση

Μια ροή δεδομένων με την πάροδο του χρόνου ενδέχεται να αλλάζει ή να εξελίσσεται. Τι σημαίνει όμως ότι μια ροή δεδομένων αλλάζει ή εξελίσσεται; Είναι απίθανο να σημαίνει ότι τα στοιχεία που παρατηρούνται τη σημερινή ημέρα δεν είναι ακριβώς τα ίδια με αυτά που είχαν παρατηρηθεί την προηγούμενη ημέρα. Μια πιο λογική αντίληψη είναι ότι οι στατιστικές ιδιότητες των δεδομένων αλλάζουν με μεγαλύτερο ρυθμό από αυτό που μπορεί να αποδοθεί σε τυχαίες διακυμάνσεις, θόρυβο. Για να γίνει κατανοητή αυτή η ιδέα, βοηθά να γίνει η υπόθεση ότι τα δεδομένα είναι στην πραγματικότητα το αποτέλεσμα μιας τυχαίας διαδικασίας που κάθε φορά δημιουργεί ένα στοιχείο, σύμφωνα με μια κατανομή πιθανότητας που χρησιμοποιείται εκείνη την ακριβή στιγμή και που μπορεί να είναι ή να μην είναι η ίδια που χρησιμοποιείται σε οποιαδήποτε άλλη δεδομένη στιγμή. Δεν υπάρχει καμία αλλαγή, όταν αυτή η υποκείμενη κατανομή παραγωγής παραμένει σταθερή. Η αλλαγή πραγματοποιείται κάθε φορά που διαφέρει η κατανομή από το ένα βήμα στο επόμενο [10].

Η ανίχνευση αλλαγής είναι ένα πολύ σύνθετο πρόβλημα που έχει μελετηθεί από πολλούς ερευνητές και έχουν αποδοθεί πολλοί ορισμοί που δηλώνουν ξεκάθαρα το εύρος του ερευνητικού πεδίου.

- εννοιολογική αλλαγή (concept change) -αλλαγή στόχου
    - εννοιολογική απόκλιση (concept drift) - σταδιακή αλλαγή
    - εννοιολογική μετατόπιση (concept shift)- απότομη αλλαγή
- distribution or sampling change-αλλαγή κατανομής ή δειγματοληψίας

Ο όρος έννοια (concept) αποδίδεται στη μεταβλητή στόχο. Ο όρος εννοιολογική αλλαγή (concept change) αναφέρεται στην αλλαγή του στόχου με την πάροδο του χρόνου, **με τρόπο αυθαίρετο**. Ο όρος εννοιολογική απόκλιση (concept drift) περιγράφει μια σταδιακή

αλλαγή του στόχου. Ο όρος εννοιολογική μετατόπιση (concept shift) συμβαίνει όταν μια αλλαγή μεταξύ δύο εννοιών είναι πιο απότομη.

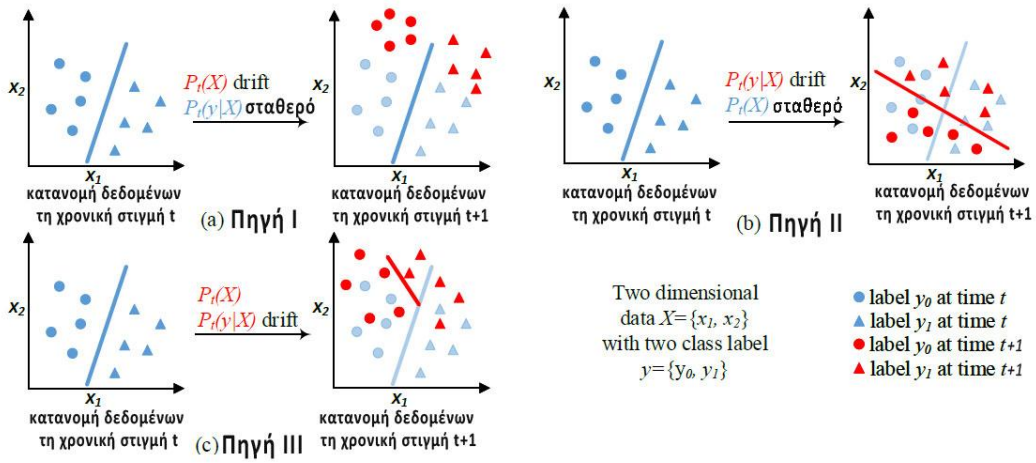
Η αλλαγή κατανομής, επίσης γνωστή ως αλλαγή δειγματοληψίας, αναφέρεται στην αλλαγή της κατανομής των δεδομένων. Ακόμα κι αν ο στόχος παραμένει ο ίδιος, μία αλλαγή στην κατανομή των δεδομένων μπορεί συχνά να οδηγήσει σε αναθεώρηση του τρέχοντος μοντέλου, καθώς το ποσοστό σφάλματος του μοντέλου μπορεί να μην είναι πλέον αποδεκτό για τη νέα κατανομή. Ορισμένοι συγγραφείς, όπως ο Stanley [23], έχουν προτείνει ότι, από πρακτική άποψη, δεν είναι απαραίτητο να γίνει διάκριση μεταξύ αλλαγής στόχου και αλλαγής δειγματοληψίας, καθώς το τρέχον μοντέλο πρέπει να αλλάξει και στις δύο περιπτώσεις. Η αντίληψη αυτή θα υπερασπιστεί σε αυτήν την εργασία, όπου και θα χρησιμοποιηθεί στο εξής ο όρος εννοιολογική απόκλιση (concept drift) για κάθε αλλαγή και μια διάκριση τους θα γίνεται ως προς τους τύπους που περιγράφονται στην ενότητα 3.2, όπως έχει συναντηθεί στις περισσότερες βιβλιογραφικές αναφορές[13].

Επίσημα, η εννοιολογική απόκλιση ορίζεται ως εξής: Δίνεται σε μια χρονική περίοδος [0, t], ένα σύνολο δειγμάτων, που συμβολίζεται ως $S_{0,t}$ ={ d_0, . . . , d_t } , όπου $d_i = (X_i, y_i)$ είναι μία παρατήρηση (ή μια παρουσία δεδομένων), Xi είναι το διάνυσμα χαρακτηριστικών, y_i είναι η ετικέτα και το $S_{0,t}$ ακολουθεί μια ορισμένη κατανομή $F_{0,t}(X, y)$. Η εννοιολογική απόκλιση εμφανίζεται στη χρονική σήμανση t+1, εάν $F_{0,t}(X, y) \neq F_{t+1,\infty}(X, y)$, που συμβολίζεται ως :

$$\exists t: P_t(X, y) \neq P_{t+1}(X, y) \quad (4)$$

3.2 Ειδή Εννοιολογικής απόκλισης

Από τον τύπο (4) συμπεραίνεται ότι η εννοιολογική απόκλιση που προκύπτει τη στιγμή t μπορεί να οριστεί ως η αλλαγή της κοινής πιθανότητας των X και y τη στιγμή t. Δεδομένου ότι η κοινή πιθανότητα $P_t(X,y)$ μπορεί να αποσυντεθεί σε δύο μέρη ως $P_t(X,y) = P_t(X)P_t(y|X)$, η εννοιολογική απόκλιση μπορεί να ενεργοποιηθεί από τρεις πηγές:

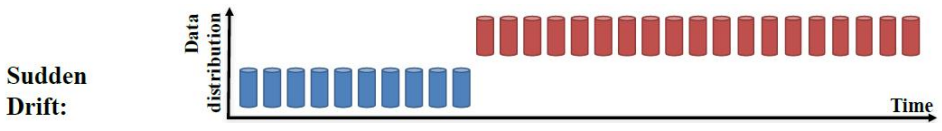


Εικόνα 3:Πήγες concept drift [26]

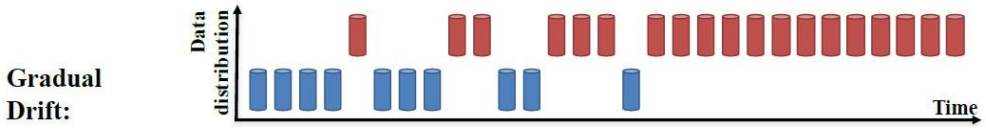
- Πηγή I: $P_t(X) \neq P_{t+1}(X)$ ενώ $P_t(y|X)=P_{t+1}(y|X)$, δηλαδή, το επίκεντρο της έρευνας είναι η αλλαγή του $P_t(X)$ ενώ το $P_t(y|X)$ παραμένει αμετάβλητο. Από την αλλαγή στο $P_t(X)$ η κατανομή εισόδου αλλάζει αλλά δεν επηρεάζεται το όριο απόφασης, έχει επίσης θεωρήθηκε ως **εικονική εννοιολογική απόκλιση**.
- Πηγή II: $P_t(y|X) \neq P_{t+1}(y|X)$ ενώ $P_t(X) = P_{t+1}(X)$ ενώ το $P_t(X)$ παραμένει αμετάβλητο. Αυτή η αλλαγή θα προκαλέσει αλλαγή των ορίων απόφασης και θα οδηγήσει σε μείωση της ακρίβειας μάθησης, η οποία ονομάζεται επίσης **πραγματική εννοιολογική απόκλιση**.
- Πηγή III: μείγμα Πηγής I και Πηγής II, συγκεκριμένα $P_t(X) \neq P_{t+1}(X)$ και $P_t(y|X) \neq P_{t+1}(y|X)$. Η εννοιολογική απόκλιση επικεντρώνεται στη αλλαγή τόσο του $P_t(y|X)$ όσο και του $P_t(X)$, αφού και οι δύο αλλαγές μεταφέρουν σημαντικές πληροφορίες για μαθησιακό περιβάλλον [26].

3.2.1 Ειδή Εννοιολογικής απόκλισης σχετικά με την ταχύτητας τους

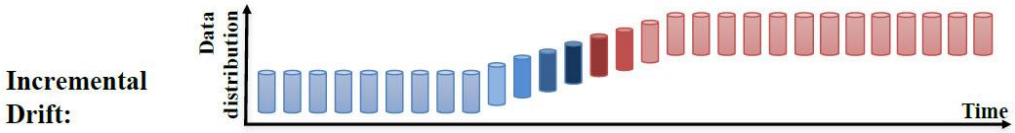
Επίσης, μια εννοιολογική απόκλιση ανάλογα με την ταχύτητα και την ευκρίνεια που δημιουργείτε, μπορεί να διακριθεί στους παρακάτω τύπους :



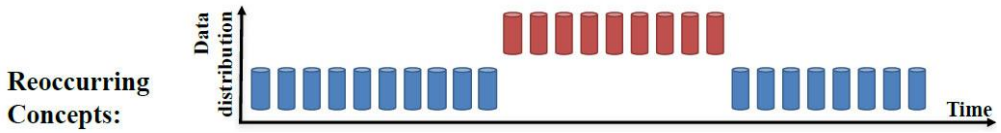
Εικόνα 4: Μια νέα εννοιολογική απόκλιση εμφανίζεται σε σύντομο χρονικό διάστημα [26].



Εικόνα 5: Μια νέα εννοιολογική απόκλιση αντικαθιστά σταδιακά μια παλιά σε μια χρονική περίοδο [26].



Εικόνα 6 : Μια παλιά εννοιολογική απόκλιση αλλάζει σταδιακά σε μια νέα κατά τη διάρκεια μιας χρονικής περιόδου [26].



Εικόνα 7: Μια παλιά εννοιολογική απόκλιση μπορεί να επαναληφθεί μετά από κάποιο χρονικό διάστημα[26].

Η έρευνα για την προσαρμογή σε μια εννοιολογική απόκλιση στους τύπους 1-3 επικεντρώνεται στο πώς ελαχιστοποιηθεί η μείωση της ακρίβειας ενός μοντέλου και στο πως επιτευχθεί ο ταχύτερος ρυθμός ανάκτησης της. Δηλαδή στο πόσο γρήγορα θα εντοπίσει ένα μοντέλο μάθησης μια εννοιολογική απόκλιση και θα προσαρμοστεί στην τρέχουσα κατανομή των δεδομένων που έχει αλλάξει. Αντίθετα, η ερεύνα για την προσαρμογή σε μια εννοιολογική απόκλιση τύπου 4 δίνει έμφαση στη χρήση ιστορικών αλλαγών, δηλαδή στο πώς να βρούμε τις καλύτερες αντίστοιχα ιστορικές αλλαγές και στο συντομότερο χρονικό διάστημα.

- **Ξαφνική (**Sudden) εννοιολογική απόκλιση συμβαίνει όταν η κατανομή έχει παραμείνει σταθερή για μεγάλο χρονικό διάστημα, στη συνέχεια αλλάζει απότομα σε λίγα βήματα.(εικόνα6)
- **Σταδιακή ή Βαθμιαία σταδιακή** (Gradual / Incremental) εννοιολογική απόκλιση συμβαίνει όταν, για μεγάλο χρονικό διάστημα, η κατανομή βιώνει σε κάθε βήμα μια μικροσκοπική, ελάχιστα αισθητή αλλαγή, αλλά αυτές οι συσσωρευμένες αλλαγές γίνονται σημαντικές με την πάροδο του χρόνου (εικόνα 7,8).
- **Επαναλαμβανόμενες εννοιολογικές αποκλίσεις** (Recurring) εμφανίζονται όταν οι κατανομές που έχουν εμφανιστεί στο παρελθόν τείνουν να επανεμφανίζονται με την πάροδο του χρόνου. Ένα παράδειγμα είναι η εποχικότητα, όπου οι καλοκαιρινές διανομές είναι παρόμοιες μεταξύ τους και διαφορετικές από τις χειμερινές διανομές(εικόνα9) [26].

- Επίσης η εννοιολογική απόκλιση μπορεί να είναι **καθολική ή μερική**, ανάλογα με το αν επηρεάζει όλο το χώρο του στοιχείου ή μόνο ένα μέρος του [10].

3.3 Επιθυμητές ιδιότητες ενός συστήματος διαχείρισης Εννοιολογικής απόκλισης

Η ενσωμάτωση της ανίχνευσης της εννοιολογικής απόκλισης στη διαδικασία εκπαίδευσης είναι ένα από τα πιο δύσκολα προβλήματα κατά την εκπαίδευση από ροές δεδομένων. Το κύριο πρόβλημα είναι να ανιχνευτούν τα σημεία αλλαγής, την ακριβή στιγμή, που συμβαίνει μια εννοιολογική απόκλιση. Σε πραγματικά προβλήματα μεταξύ δύο διαδοχικών ακολουθιών S_i και S_{i+1} θα μπορούσε να υπάρξει μια μεταβατική φάση όπου ορισμένα παραδείγματα **και των δύο κατανομών εμφανίζονται μικτά.** Ένα παράδειγμα που δημιουργείται από μια κατανομή F_{i+1} είναι ο θόρυβος για τη κατανομή F_i. Αυτή είναι μια άλλη δυσκολία που αντιμετωπίζουν οι αλγόριθμοι ανίχνευσης εννοιολογικής απόκλισης, οφείλουν να διαχωρίζουν τον θόρυβο από την εννοιολογική απόκλιση. Η διαφορά μεταξύ του θορύβου και των παραδειγμάτων μιας άλλης κατανομής είναι η επιμονή, χρειάζεται να υπάρχει ένα συνεπές σύνολο παραδειγμάτων της νέας κατανομής. Οι αλγόριθμοι για την ανίχνευση αλλαγών είναι απαραίτητο να συνδυάζουν την ανθεκτικότητα στο θόρυβο με την ευαισθησία σε μια εννοιολογική απόκλιση. Οι επιθυμητές ιδιότητες ενός μοντέλου ανίχνευσης εννοιολογικής απόκλισης είναι [14]:

- Εντοπισμός μιας εννοιολογικής απόκλισης (και προσαρμογή των μοντέλων, εάν χρειάζεται) το συντομότερο δυνατό.
- Ταυτόχρονα, να είναι ανθεκτικό στο θόρυβο και τις ακραίες τιμές.
    - Διάκριση θορύβου από μια εννοιολογική απόκλιση
- Λειτουργία σε λιγότερο από το χρόνο άφιξης και την υπογραμμική μνήμη (ιδανικά, κάποια σταθερή, προκαθορισμένη ποσότητα μνήμης).
- Αναγνώριση και αντίδραση σε επαναλαμβανόμενα πλαίσια [36].

3.4 Διαχείριση της Εννοιολογικής απόκλισης

Σε δυναμικά μεταβαλλόμενα ή μη σταθερά περιβάλλοντα, η κατανομή των δεδομένων μπορεί να αλλάξει με την πάροδο του χρόνου προκαλώντας το φαινόμενο της εννοιολογικής απόκλισης. Οι εννοιολογικές αποκλίσεις μπορούν να προσαρμοστούν

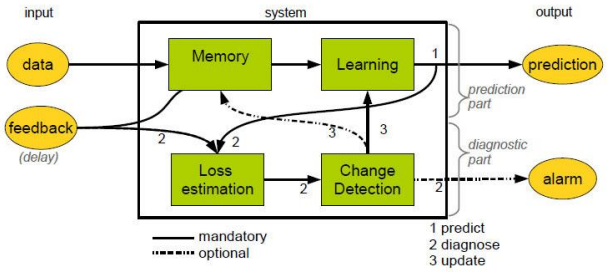
γρήγορα με την αποθήκευση περιγραφών των εννοιών, έτσι ώστε να μπορούν να επανεξεταστούν και να χρησιμοποιηθούν ξανά αργότερα. Ως εκ τούτου, απαιτείται προσαρμοστική μάθηση για την αντιμετώπιση των δεδομένων σε μη σταθερά περιβάλλοντα. Όταν εντοπιστεί μια εννοιολογική απόκλιση, το τρέχον μοντέλο πρέπει να ενημερωθεί για να διατηρηθεί η ακρίβεια του[39].

Η διαδικτυακή προσαρμοστική μάθηση ορίζεται επίσημα ως εξής. Ένα μοντέλο απόφασης είναι μια συνάρτηση L που αντιστοιχίζει τις μεταβλητές εισόδου στον επιθυμητό στόχο y = L(X). Ένας αλγόριθμος κατηγοριοποίησης καθορίζει τον τρόπο δημιουργίας ενός μοντέλου από ένα σύνολο παρουσιών δεδομένων.

Η διαδικτυακή διαδικασία προσαρμοστικής μάθησης είναι η ακόλουθη.

1. Πρόβλεψη. Όταν φτάσει το νέο παράδειγμα X_t, γίνεται μια πρόβλεψη $\hat{y}_t$ χρησιμοποιώντας το μοντέλο της ροής L_t.
2. Διάγνωση. Μετά από κάποιο χρονικό διάστημα γίνεται λήψη της αληθινής ετικέτας y_t και εκτιμείται η απώλεια ως $f(\hat{y}_t, y_t)$.
3. Ενημέρωση. Μπορεί να χρησιμοποιηθεί το παράδειγμα (X_t, y_t) για την ενημέρωση του μοντέλου L_{t+1}.

Ανάλογα με τους υπολογιστικούς πόρους, τα δεδομένα μπορεί να χρειαστεί να απορρίπτονται μόλις υποβληθούν σε επεξεργασία χρησιμοποιώντας την πιο πρόσφατη έκδοση του μοντέλου L_{t+1} = train $((X_t, y_t),L_t)$. Εναλλακτικά, ορισμένα από τα προηγούμενα δεδομένα ενδέχεται να παραμείνουν προσβάσιμα L_{t+1} = train $((X_i, y_i),.., (X_t,y_t),L_t)$. Υπάρχουν διάφοροι τρόποι χειρισμού δεδομένων στο διαδίκτυο. Μετά την ενημέρωση του μοντέλου έρχεται νέο παράδειγμα X_{t+1} και ο βρόχος πρόβλεψης-διάγνωσης-ενημέρωσης του μοντέλου ανατροφοδότησης συνεχίζεται απεριόριστα. Σε κάποια χρονικά βήματα σταθερότητας μπορεί να επιλεχτεί να διατηρηθεί το τρέχον μοντέλο $L_{t+1} = L_t$.



Εικόνα 8: 'Ενα γενικό σχήμα Διαδικτυακού αλγόριθμου προσαρμοστικής μάθησης

Η εικόνα 8 απεικονίζει ένα γενικό σχήμα για έναν διαδικτυακό αλγόριθμο προσαρμοστικής μάθησης. Η μονάδα μνήμης ορίζει πώς και ποια δεδομένα παρουσιάζονται στον αλγόριθμο μάθησης (μαθησιακή ενότητα). Η μονάδα εκτίμησης απωλειών παρακολουθεί την απόδοση

του αλγόριθμου εκπαίδευσης και στέλνει πληροφορίες στη μονάδα ανίχνευσης αλλαγών για ενημέρωση το μοντέλο εάν είναι απαραίτητο[11].

Έχουν προταθεί πολλές μέθοδοι στη Μηχανική Μάθηση για την αντιμετώπιση μιας εννοιολογικής απόκλισης. Όλες αυτές οι μέθοδοι υποθέτουν ότι τα πιο πρόσφατα παραδείγματα είναι τα πιο σχετικά. Γενικά, οι προσεγγίσεις για την αντιμετώπιση μιας εννοιολογικής απόκλισης μπορούν να αναλυθούν στις εξής διαστάσεις: διαχείριση δεδομένων στη μνήμη, μέθοδοι προσαρμογής κατά τη εκπαίδευσης, τεχνικές αναγνώρισης εννοιολογικών αποκλίσεων [14].

3.5 Διαχείριση δεδομένων στη μνήμη

Οι μέθοδοι διαχείρισης δεδομένων χαρακτηρίζουν τις πληροφορίες σχετικά με τα δεδομένα που είναι αποθηκευμένα στη μνήμη ώστε να διατηρηθεί ένα μοντέλο απόφασης συνεπές με την πραγματική κατάσταση της φύσης.

3.5.1 Πλήρης μνήμη

Μέθοδοι που αποθηκεύουν στη μνήμη επαρκή στατιστικά στοιχεία για όλα τα παραδείγματα. Τα παραδείγματα περιλαμβάνουν τη στάθμιση των παραδειγμάτων ανάλογα με την ηλικία τους. Τα σταθμισμένα παραδείγματα βασίζονται στην απλή ιδέα ότι η σημασία ενός παραδείγματος πρέπει να μειώνεται με το χρόνο.

3.5.2 Μερική μνήμη

Μέθοδοι που αποθηκεύουν στη μνήμη μόνο τα πιο πρόσφατα παραδείγματα. Τα παραδείγματα αποθηκεύονται σε μια first-in first-out (fifo) δομή δεδομένων. Ένα παράδειγμα υλοποίησης μιας μεθόδου επεξεργασίας fifo είναι να οριστεί ένα χρονικό παράθυρο στη ροή των εισερχόμενων παραδειγμάτων. Σε κάθε χρονικό βήμα, ο αλγόριθμος εκπαίδευσης προκαλεί ένα μοντέλο απόφασης χρησιμοποιώντας μόνο τα παραδείγματα που περιλαμβάνονται στο παράθυρο. Η βασική δυσκολία είναι πώς θα επιλεχτεί το κατάλληλο μέγεθος παραθύρου. Ένα παράθυρο μικρού μεγέθους, που αντικατοπτρίζει με ακρίβεια την τρέχουσα κατανομή, μπορεί να εξασφαλίσει γρήγορη προσαρμοστικότητα σε φάσεις με εννοιολογικές αποκλίσεις, αλλά σε πιο σταθερές φάσεις μπορεί να επηρεάσει την απόδοση του μαθητή. Ενώ ένα μεγάλο παράθυρο θα παρήγαγε καλά και σταθερά μαθησιακά αποτελέσματα σε περιόδους σταθερότητας, αλλά δεν μπορεί να αντιδράσει γρήγορα σε μια εννοιολογική απόκλιση.

- **Συρόμενα Παράθυρα Σταθερού Μεγέθους (Sliding window-fixed window size)**

  Αυτές οι μέθοδοι αποθηκεύουν στη μνήμη έναν σταθερό αριθμό από τα πιο πρόσφατα παραδείγματα. Όποτε είναι διαθέσιμο ένα νέο παράδειγμα, αποθηκεύεται στη μνήμη και το παλαιότερο απορρίπτεται. Αυτή είναι η απλούστερη μέθοδος για την αντιμετώπιση μιας εννοιολογικής απόκλισης και μπορεί να χρησιμοποιηθεί ως βάση για συγκρίσεις.

- **Παράθυρα προσαρμοστικού μεγέθους (adaptive window size)**.

  Το σύνολο των παραδειγμάτων στο παράθυρο είναι μεταβλητό συνήθως χρησιμοποιούνται σε συνδυασμό με ένα μοντέλο ανίχνευσης. Μειώνεται το μεγέθους του παραθύρου κάθε φορά που το μοντέλο ανίχνευσης σηματοδοτεί αντιμετώπιση μιας εννοιολογικής απόκλισης και αυξάνεται όταν η κατανομή παραμένει σταθερή.

Ένα μοντέλο ροής δεδομένων θα πρέπει να μπορεί να αντιδρά σε μια εννοιολογική απόκλιση ξεχνώντας παλιά δεδομένα, ενώ εκπαιδεύεται κάνοντας χρήση νέων παραδειγμάτων που αντικατοπτρίζουν την τρέχουσα φύση της ροής. Το μοντέλο διαχείρισης δεδομένων υποδεικνύει επίσης τον **μηχανισμό λήθης**. Τα Παραδείγματα στάθμισης αντιστοιχούν σε σταδιακή λήθη. Η συνάφεια των παλαιών πληροφοριών είναι όλο και λιγότερο σημαντική. Τα χρονικά παράθυρα (μερικής μνήμης) αντιστοιχούν σε απότομη λήθη. Τα παραδείγματα διαγράφονται από τη μνήμη. Φυσικά μπορούν να συνδυαστούν και οι δύο μηχανισμοί λήθης σταθμίζοντας τα παραδείγματα σε ένα χρονικό παράθυρο [11][14].

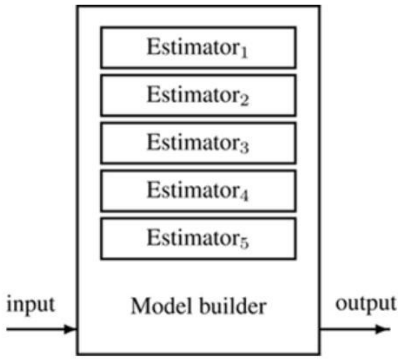
3.6 Μέθοδοι Προσαρμογής

Οι μέθοδοι προσαρμογής χαρακτηρίζουν τις αλλαγές στο αλγόριθμο κατηγοριοποίησης κατά τη διαδικασία εκπαίδευσης.

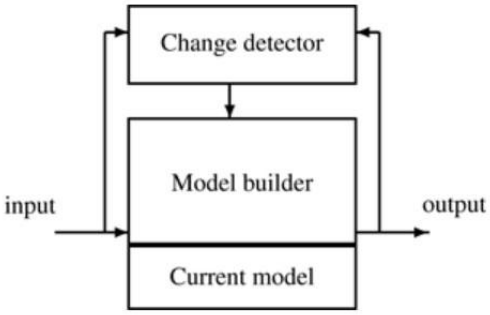
- **Τυφλές μέθοδοι:** Μέθοδοι που προσαρμόζουν τον μοντέλο εκπαίδευσης σε τακτά χρονικά διαστήματα χωρίς να εξετάζουν αν έχουν όντως συμβεί μια εννοιολογική απόκλιση. Τα παραδείγματα περιλαμβάνουν μεθόδους που σταθμίζουν τα παραδείγματα ανάλογα με την ηλικία τους και μεθόδους που χρησιμοποιούν χρονικά παράθυρα σταθερού μεγέθους.
- **Ενημερωμένες μέθοδοι:** Μέθοδοι που τροποποιούν το μοντέλο εκπαίδευσης μόνο αφού εντοπιστεί μια αλλαγή. Χρησιμοποιούνται σε συνδυασμό με ένα μοντέλο ανίχνευσης εννοιολογικής απόκλισης. Μια αναλυτική περιγραφή των μοντέλων αυτών παρουσιάζεται στην ενότητα 3.7.2[11][14].

3.7 Τεχνικές αναγνώρισης Εννοιολογικής απόκλισης

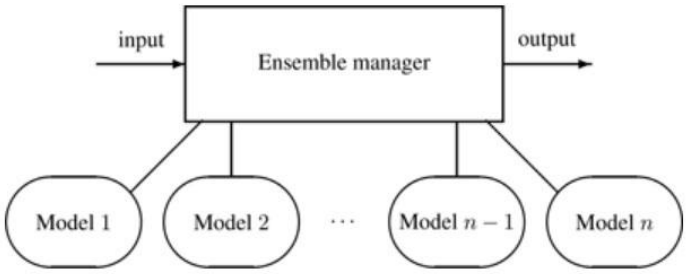
Οι τεχνικές για την διαχείριση εννοιολογικής απόκλισης μπορούν να ομαδοποιηθούν σε τρεις οικογένειες ή και σε συνδυασμό αυτών. Α) Στην πρώτη περίπτωση γίνεται η χρήση προσαρμοστικών εκτιμητών για την διατήρηση των σχετικών στατιστικών και έπειτα ενός αλγόριθμου. Ο αλγόριθμος σε συγχρονισμό με τα αυτά τα στατιστικά στοιχεία, δημιουργεί ένα μοντέλο Β) Η δημιουργία μοντέλων που προσαρμόζονται ή ανακατασκευάζονται όταν ένας ανιχνευτής αλλαγών υποδεικνύει ότι έχει μια εννοιολογική απόκλιση. Γ) Μέθοδοι συνόλου, οι οποίοι διατηρούν δυναμικούς πληθυσμούς μοντέλων [10].



Εικόνα 8 : Α) Προσαρμοστικοί εκτιμητές Και αλγόριθμος[10]



Εικόνα 9 : Β) Μοντέλα με ανιχνευτές εννοιολογικών αποκλίσεων[10]

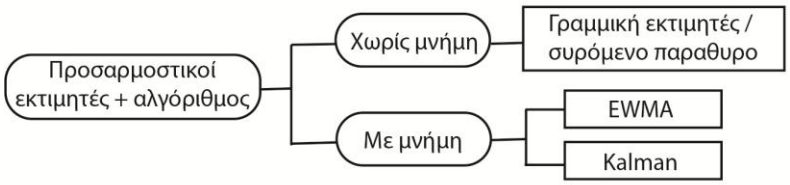


Εικόνα 10: Γ) Μέθοδοι συνόλου για διαχείριση εννοιολογικής απόκλισης [10]

Στις δύο ενότητες αναλύονται οι δύο πρώτες προσεγγίσεις. Η μέθοδοι συνόλου κατηγοριοποίησης αναλύονται στην ενότητα 4.4

3.7.1 Εκτιμητές

Ένας εκτιμητής είναι ένας αλγόριθμος που εκτιμά ένα ή περισσότερα στατιστικά στοιχεία σχετικά με τα δεδομένα εισόδου, τα οποία μπορεί να αλλάξουν με την πάροδο του χρόνου. Επισημαίνεται η περίπτωση που ένα τέτοιο στατιστικό στοιχείο είναι σχετικό με μια αναμενόμενη τιμή της τρέχουσας κατανομής των δεδομένων. Σε αυτή την περίπτωση πρόβλημα είναι ότι όταν προκύψει μια εννοιολογικής απόκλιση είναι δύσκολο να είμαστε σίγουροι ποια προηγούμενα στοιχεία της ροής εξακολουθούν να είναι αξιόπιστα, ως δείγματα της τρέχουσας κατανομής και ποια είναι ξεπερασμένα.



Εικόνα 11 : Εκτιμητές

Υπάρχουν δύο είδη εκτιμητών: εκείνοι που αποθηκεύουν ρητά ένα δείγμα της ροής δεδομένων (θα ονομάσουμε αυτόν τον χώρο αποθήκευσης "προσωρινή μνήμη") και εκτιμητές χωρίς μνήμη. Μεταξύ των πρώτων αναλύεται ο γραμμικός εκτιμητής πάνω από συρόμενα παράθυρα. Μεταξύ των τελευταίων, εξηγούμε το EWMA και το φίλτρο Kalman [10].

Συρόμενα Παράθυρα και Γραμμικοί Εκτιμητές

Ο απλούστερος αλγόριθμος εκτιμητή για την αναμενόμενη τιμή είναι ο γραμμικός εκτιμητής, ο οποίος απλώς επιστρέφει τον μέσο όρο των στοιχείων δεδομένων που περιέχονται στη μνήμη. Μια εύκολη υλοποίηση της μνήμης είναι ένα συρόμενο παράθυρο που αποθηκεύει τα πιο πρόσφατα στοιχεία W που ελήφθησαν. Τις περισσότερες φορές, το W είναι μια σταθερή παράμετρος του εκτιμητή. Οι πιο εξελιγμένες προσεγγίσεις μπορεί να αλλάξουν το W με την πάροδο του χρόνου, ίσως ως αντίδραση στη φύση των ίδιων των δεδομένων [10].

Εκθετικά σταθμισμένος κινητός μέσος όρος (EWMA)

Ο εκθετικά σταθμισμένος εκτιμητής κινητού μέσου όρου (EWMA) αλλάζει την εκτίμηση μιας μεταβλητής συνδυάζοντας την πιο πρόσφατη μέτρηση μιας μεταβλητής με την EWMA όλων των προηγούμενων μετρήσεων:

$$A_t = ax_t + (1-a)A_{t-1}\,, A_1 = x_1 (5)$$

όπου A_t είναι ο κινητός μέσος όρος κατά τη χρονική περίοδο t , x_t είναι η τελευταία μέτρηση και α ∈ (0, 1) είναι μια παράμετρος που αντικατοπτρίζει το βάρος που δίνεται στην

τελευταία μέτρηση. Συχνά ονομάζεται παράγοντας αποσύνθεσης ή εξασθένισης. Πράγματι, επεκτείνοντας την παραπάνω επανάληψη, μπορούμε να δούμε ότι η εκτίμηση στο χρόνο t είναι:

$$A_t = \sum_{i=2}^{t} \alpha(1-\alpha)^{\mathrm{t}-i}\mathrm{x_i} + (1-\alpha)^{\mathrm{t}-1}\mathrm{x_1} \text{ (6)}$$

έτσι το βάρος κάθε μέτρησης διασπάται εκθετικά γρήγορα με βάση το 1 − α. Οι μεγαλύτερες τιμές α συνεπάγονται ταχύτερη λήθη των παλαιών μετρήσεων και οι μικρότερες δίνουν μεγαλύτερη σημασία στην ιστορία[35][10].

Μονοδιάστατο φίλτρο Kalman

Το μονοδιάστατο φίλτρο Kalman [5] αντιμετωπίζει το πρόβλημα της εκτίμησης της κρυφής κατάστασης x∈R μιας ελεγχόμενης w διεργασίας διακριτού χρόνου που διέπεται από τη γραμμική εξίσωση της στοχαστικής διαφοράς:

$$x_t = x_{t-1} + w_{t-1} \text{ (7)}$$

όπου το x παρατηρείται έμμεσα μέσω μιας μέτρησης z ∈R που είναι

$$z_t = x_t + v_t \text{ (8)}$$

Εδώ w_t και v_t είναι τυχαίες μεταβλητές που αντιπροσωπεύουν τη διαδικασία και το θόρυβο μέτρησης, αντίστοιχα. Θεωρείται ότι είναι ανεξάρτητες μεταξύ τους και με κανονικές κατανομές πιθανότητας

$$w \sim N(0, Q), v \sim N(0, R) \text{ (9)}$$

Στο περιβάλλον μας, x_t είναι η αναμενόμενη τιμή στο χρόνο t κάποιας ιδιότητας των στοιχείων ροής, z_t είναι η αξία της ιδιότητας στο στοιχείο που παρατηρείται πραγματικά κατά το χρόνο t , και w_t , v_t είναι η τυχαία αλλαγή στο x_t και ο θόρυβος παρατήρησης στο z_t. Η εκτίμηση y_t της κατάστασης κατά το χρόνο t ενημερώνεται στο φίλτρο Kalman ως εξής, όπου P και K είναι βοηθητικές ποσότητες.

$$K_t \leftarrow P_{t-1} / (P_{t-1} + R),$$

$$y_t \leftarrow y_{t-1} + K_t(z_t - y_{t-1}),$$

$$P_t \leftarrow (1 - K_t)P_{t-1} + Q. \text{ (10)}$$

Η αποτελεσματικότητα του φίλτρου Kalman σε οποιαδήποτε συγκεκριμένη εφαρμογή εξαρτάται από την εγκυρότητα των στοχαστικών εξισώσεων, των gaussian υποθέσεων για το θόρυβο και από την ακρίβεια της εκτίμησης των διακυμάνσεων Q και R .

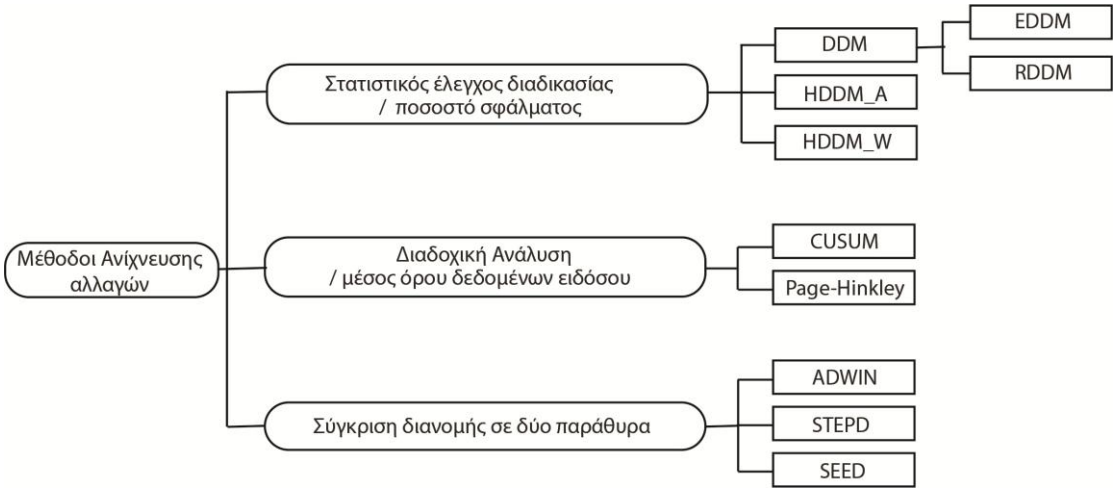
Αυτή είναι μια πολύ απλή έκδοση του φίλτρου Kalman. Γενικότερα, μπορεί να είναι πολυδιάστατο, όπου το x είναι ένα διάνυσμα και κάθε συστατικό του x είναι ένας γραμμικός συνδυασμός όλων των συνιστωσών του την προηγούμενη φορά, συν το θόρυβο.

Επιτρέπει επίσης μεταβλητές ελέγχου που μπορούμε να αλλάξουμε για να επηρεάσουμε την κατάσταση και επομένως να δημιουργήσουμε βρόχους ανάδρασης. Αυτός είναι στην πραγματικότητα ο αρχικός και κύριος σκοπός του φίλτρου Kalman. Μπορεί να επεκταθεί σε μη γραμμικές σχέσεις μέτρησης προς διαδικασία [5][10].

3.7.2 Ανιχνευτές αλλαγών

Η ανίχνευση εννοιολογικών αποκλίσεων στα δεδομένα είναι ένα θέμα μεγάλων διαστάσεων που έχει μελετηθεί ευρέως από πολλούς ερευνητές. Αυτή η ενότητα εξετάζει αλγόριθμους που επιτρέπουν την ανίχνευση μιας εννοιολογικής απόκλισης, γνωστοί ως ανιχνευτές αλλαγών. Προειδοποιούν τον βασικό μαθητή, ότι το μοντέλο πρέπει να ξαναχτιστεί ή να ενημερωθεί[49]. Παρακάτω παρουσιάζονται οι περισσότεροι μέθοδοι που έχουν ενταχθεί στο MOA.

.



Εικόνα 11: Μέθοδοι Ανίχνευσης Αλλαγών

3.7.2.1 Στατιστικός έλεγχος διαδικασίας

Ο στατιστικός έλεγχος διαδικασίας (SPC) χρησιμοποιείται για την παρακολούθηση της ποιότητας της μαθησιακής διαδικασίας ανιχνεύοντας την εξέλιξη του διαδικτυακού ποσοστού σφάλματος των βασικών μαθητών. Μια εννοιολογική απόκλιση θεωρείται ότι έχει συμβεί εάν η υποβάθμιση της απόδοσης του μοντέλου είναι στατιστικά σημαντική. Στην περίπτωση αυτή θα ενεργοποιηθεί μια διαδικασία αναβάθμισης (συναγερμός ενός concept drift) [49].

DDM: Μέθοδος ανίχνευσης ολίσθησης.

Η μέθοδος DDM που προτείνεται από τους Gama et al. [8] ελέγχει τον αριθμό των σφαλμάτων που παράγονται από το ένα μοντέλο κατηγοριοποίησης κατά την πρόβλεψη.

Συγκρίνει τα στατιστικά στοιχεία δύο παραθύρων, το πρώτο παράθυρο περιέχει όλα τα δεδομένα και το δεύτερο περιέχει μόνο τα δεδομένα από την αρχή μέχρι να αυξηθεί ο αριθμός των σφαλμάτων. Αυτή η μέθοδος δεν αποθηκεύει αυτά τα παράθυρα στη μνήμη, αλλά διατηρεί μόνο στατιστικά στοιχεία και ένα παράθυρο με πρόσφατα δεδομένα.

Ο αριθμός των σφαλμάτων σε ένα δείγμα n παραδειγμάτων μοντελοποιείται με μια διωνυμική κατανομή. Για κάθε σημείο i της ακολουθίας που γίνεται δειγματοληψία, το ποσοστό σφάλματος είναι η πιθανότητα εσφαλμένης κατηγοριοποίησης (p_i), με τυπική απόκλιση που δίνεται από τον τύπο $s_i = \sqrt{p_i(1-p_i)t}$ (11).

Υποθέτει (όπως μπορεί να υποστηριχθεί π.χ. στο μοντέλο εκπαίδευσης PAC [51]) ότι το ποσοστό σφάλματος του αλγορίθμου εκπαίδευσης (p_i) θα μειωθεί όσο αυξάνεται ο αριθμός των αναλυόμενων δειγμάτων, εφόσον η κατανομή των δεδομένων είναι σταθερή. Μια αύξηση στο σφάλμα του αλγορίθμου, που ξεπερνά ένα υπολογιζόμενο όριο, υποδηλώνει ότι η κατανομή κλάσης αλλάζει και αποτελεσματικά, το πραγματικό μοντέλο απόφασης υποτίθεται ότι είναι ακατάλληλο (το μοντέλο πρέπει να ξαναδημιουργηθεί).

Το κατώφλι ανίχνευσης υπολογίζεται σε συνάρτηση με δύο στατιστικές, που λαμβάνονται όταν το ($p_i + s_i$) είναι ελάχιστο. Το DDM αποθηκεύει τη μικρότερη τιμή p_{min} από τα ποσοστά σφάλματος που παρατηρούνται μέχρι το χρόνο t και την τυπική απόκλιση s_{min} σε αυτό το σημείο. Στη συνέχεια εκτελεί τους ακόλουθους ελέγχους:

- Αν $p_t + s_t \geq p_{min} + 2 \cdot s_{min}$, δηλώνεται προειδοποίηση. Από αυτό το σημείο και μετά, αποθηκεύονται νέα παραδείγματα εν αναμονή μιας πιθανής δήλωσης αλλαγής.
- Αν $p_t + s_t \geq p_{min} + 3 \cdot s_{min}$, δηλώνεται αλλαγή. Το μοντέλο που προκαλείται από τη μέθοδο εκπαίδευσης απορρίπτεται και δημιουργείται ένα νέο μοντέλο χρησιμοποιώντας τα παραδείγματα που έχουν αποθηκευτεί από τότε που εμφανίστηκε η προειδοποίηση. Οι τιμές για p_{min} και s_{min} επαναφέρονται επίσης.

Αυτή η προσέγγιση είναι καλή στον εντοπισμό απότομων και σταδιακών εννοιολογικών αποκλίσεων όταν η σταδιακή εννοιολογική απόκλιση δεν είναι πολύ αργή, αλλά έχει δυσκολίες όταν είναι αργά σταδιακή. Σε αυτήν την περίπτωση, τα παραδείγματα θα αποθηκευτούν για μεγάλο χρονικό διάστημα, η δήλωση αλλαγής μπορεί να πάρει πολύ χρόνο για να ενεργοποιηθεί και μπορεί να γίνει υπέρβαση της μνήμης παραδειγμάτων.

EDDM : Early Drift Detection Method

Η μέθοδος EDDM που προτείνεται από τους Baena-García et al.[7] είναι μια προσαρμογή της μεθόδου DDM. Στοχεύει στη βελτίωση του ρυθμού ανίχνευσης της σταδιακής εννοιολογικής απόκλισης του DDM, διατηρώντας παράλληλα καλή απόδοση σε μια απότομη εννοιολογική απόκλιση.

Ωστόσο, ο ίδιος μηχανισμός συναγερμού προειδοποίησης που χρησιμοποιείται από το DDM χρησιμοποιείται στο EDDM. Αντί να παρακολουθεί το ποσοστό σφάλματος, το EDDM παρακολουθεί την απόσταση μεταξύ δύο διαδοχικών σφαλμάτων. Κατά συνέπεια, όταν οι έννοιες των εισερχόμενων δεδομένων είναι σταθερές, η απόσταση μεταξύ των δύο διαδοχικών σφαλμάτων μειώνεται. Διαφορετικά, σηματοδοτούνται προειδοποιήσεις και εννοιολογικές αποκλίσεις.

RDDM: Reactive Drift Detection Method

Η μέθοδος RDDM [3] μετριάζει το πρόβλημα απώλειας απόδοσης της μεθόδου DDM, το οποίο οφείλεται στη μειωμένη ευαισθησία όταν μια έννοια (concept), έχει μεγάλο αριθμό μελών. Η μέθοδος RDDM βελτιώνει τη μέθοδο DDM αφαιρώντας περιοδικά παλιές παρουσίες δεδομένων μεγάλων concept. Οι συγγραφείς υποστήριξαν ότι το RDDM παρέχει υψηλότερη ή ίση συνολική ακρίβεια από το DDM και ανιχνεύει μια εννοιολογική απόκλιση νωρίτερα στις περισσότερες περιπτώσεις.

HDDM_A και HDDM_W

Οι αλγόριθμοι ανίχνευσης ολίσθησης που δημιουργήθηκαν σύμφωνα με τη μέθοδο ανίχνευσης ολίσθησης Hoeffding (HDDM) τροποποιούν το αρχικό DDM χρησιμοποιώντας την ανισότητα του Hoeffding .

Το HDDM_A είναι ένα από αυτούς τους αλγόριθμους και χρησιμοποιεί το κινητό μέσο όρο του παρατηρούμενου ποσοστού σφάλματος για την εκτίμηση του πραγματικού ποσοστού σφάλματος του κατηγοριοποιητή. Ο αλγόριθμος υπολογίζει το μέσο ποσοστό σφάλματος σε ένα συρόμενο παράθυρο σταθερού μεγέθους και το συγκρίνει με το όριο του Hoeffding. Εάν το παρατηρούμενο ποσοστό σφάλματος υπερβαίνει το όριο, ο αλγόριθμος σηματοδοτεί μια εννοιολογική απόκλιση.

Ο αλγόριθμος HDDM_W είναι παρόμοιος με τον αλγόριθμο HDDM_A αλλά χρησιμοποιεί τον εκθετικά σταθμισμένο μέσο όρο (EWMA) ως εκτιμητή σφάλματος. Ο αλγόριθμος εκχωρεί ένα βάρος σε κάθε σημείο δεδομένων με βάση την πρόσφατη έκδοσή του, δίνοντας στα πρόσφατα σημεία δεδομένων υψηλότερα βάρη. Ο σταθμισμένος μέσος όρος χρησιμοποιείται στη συνέχεια για την εκτίμηση του πραγματικού ποσοστού σφάλματος του κατηγοριοποιητή και στη συνέχεια σε σύγκριση με το όριο του Hoeffding. Το HDDM_W είναι αποτελεσματικό στην ανίχνευση τόσο απότομων όσο και σταδιακών παρασύρεται και ξεπερνά τις επιδόσεις των μεθόδων DDM και HDDM_A σε ορισμένα σενάρια.

Και στις δυο μεθόδους υπάρχει η επιλογή παρακολούθησης των αυξήσεων και μειώσεων των σφαλμάτων (δύο όψεων) ή μόνο προσαυξήσεων (μονόπλευρες)[44].

3.7.2.2 Τεχνικές παραθύρων

Οι τεχνικές παραθύρων είναι μέθοδοι ανίχνευσης εννοιολογικής απόκλισης που παρακολουθούν τις διανομές σε δύο διαφορετικά χρονικά παράθυρα. Ένα παράθυρο αναφοράς, που συνήθως συνοψίζει προηγούμενες πληροφορίες και ένα παράθυρο πάνω από τα πιο πρόσφατα παραδείγματα[49].

ADWIN :ADaptive sliding WINdow

Το ADWIN [34] [2] είναι ένας αλγόριθμος ανίχνευσης αλλαγών και εκτίμησης. Επιλύει, με έναν καλά καθορισμένο τρόπο, το πρόβλημα της παρακολούθησης του μέσου όρου μιας ροής bit ή αριθμών πραγματικής αξίας. Το ADWIN διατηρεί ένα παράθυρο μεταβλητού μήκους με τα παραδείγματα που εμφανίστηκαν πρόσφατα, με την ιδιότητα ότι το παράθυρο έχει το μέγιστο μήκος, όταν είναι στατιστικά συνεπής, η υπόθεση ότι "δεν υπήρξε καμία αλλαγή στη μέση τιμή μέσα στο παράθυρο".

Πιο συγκεκριμένα, ένα παλαιότερο τμήμα του παραθύρου απορρίπτεται εάν και μόνο εάν υπάρχουν αρκετές ενδείξεις ότι η μέση τιμή του διαφέρει από αυτή του υπόλοιπου παραθύρου. Αυτό παρουσιάζει συνέπειες: α) Η αλλαγή εντοπίζεται αξιόπιστα κάθε φορά που το παράθυρο συρρικνώνεται. β) Ανά πάσα στιγμή ο μέσος όρος του υπάρχοντος παραθύρου μπορεί να χρησιμοποιηθεί ως αξιόπιστη εκτίμηση του τρέχοντος μέσου όρου στη ροή (εκτός από μια πολύ μικρή ή πρόσφατη αλλαγή που δεν είναι ακόμη στατιστικά σημαντική).

Οι είσοδοι σε ένα ADWIN είναι μια τιμή εμπιστοσύνης $\delta \in (0, 1)$ και μια (πιθανώς άπειρη) ακολουθία πραγματικών τιμών $x_1 , x_2 , x_3 , ..., x_t , ...$ Η τιμή του x_t είναι διαθέσιμο μόνο την ώρα t .

Μια δοκιμή $T (W_0 , W_1 , \delta)$ δείχνει την παραμετροποίηση που χρειάζεται ο αλγόριθμος, ο οποίος συγκρίνει τους μέσους όρους δύο παραθύρων W_0 και W_1 και αποφασίζει αν είναι πιθανό να προέρχονται από την ίδια διανομή. Μια καλή δοκιμή πρέπει να πληροί τα ακόλουθα κριτήρια:

- Εάν δημιουργήθηκαν από την ίδια κατανομή (καμία αλλαγή), τότε με πιθανότητα τουλάχιστον $1 − \delta$ η δοκιμή λέει "καμία αλλαγή".
- Εάν W_0 και W_1 δημιουργήθηκαν από δύο διαφορετικές διανομές των οποίων ο μέσος όρος διαφέρει περισσότερο από κάποια ποσότητα ε (W_0 , W_1 , δ) τότε με πιθανότητα τουλάχιστον $1 − \delta$ η δοκιμή λέει "αλλαγή (εννοιολογική απόκλιση)".

Είναι σημαντικό να σημειωθεί ότι το ADWIN δεν διατηρεί το παράθυρο ρητά, αλλά το συμπιέζει χρησιμοποιώντας μια παραλλαγή της τεχνικής του εκθετικού ιστογράμματος [5]. Αυτό σημαίνει ότι διατηρεί ένα παράθυρο μήκους W χρησιμοποιώντας μόνο μνήμη

O(logW) και χρόνο επεξεργασίας O(logW) ανά στοιχείο, αντί για το O(W) που περιμένει κανείς από μια αρχική υλοποίηση.

STEPD Μέθοδος ανίχνευσης με χρήση στατιστικών δοκιμών

Η μέθοδος ανίχνευσης STEPD, αναπτύχθηκε από τους K. Nishida και K. Yamauchi[1], για την επίτευξη της γρήγορης και ακριβούς ανίχνευσης της εννοιολογικής απόκλισης. Η βασική αρχή της μεθόδου είναι να ληφθούν υπόψη δύο ακρίβειες ενός βασικού μαθητή, η πρόσφατη και η συνολική. Υπολογίζει και συγκρίνει την ακρίβεια στις W πιο πρόσφατες περιπτώσεις και τη συνολική ακρίβειά του από την αρχή της μαθησιακής διαδικασίας. Υποθέτει ότι οι δύο ακρίβειες είναι ίσες εάν ο στόχος (concept) παραμένει σταθερός και ότι μια σημαντική μείωση της πρόσφατης ακρίβειας υποδηλώνει αλλαγή (ανίχνευση εννοιολογικής απόκλισης). Η δοκιμή πραγματοποιείται με τον υπολογισμό του παρακάτω στατιστικού:

$$T_{(r_0,r_r,n_0,n_r)} = \frac{|r_0/n_0 - r_r/n_r| - 0.5(1/n_0 + 1/n_r)}{\sqrt{\hat{p}\ (1-\hat{p})(1/n_0 + 1/n_r)}} \quad (12)$$

και συγκρίνοντας την τιμή του με το εκατοστημόριο της τυπικής κανονικής κατανομής για να ληφθεί το παρατηρούμενο επίπεδο σημαντικότητας (P-value). Όπου r_0 είναι ο αριθμός των σωστών κατηγοριοποιήσεων μεταξύ των συνολικών παραδειγμάτων n_0, εκτός των πρόσφατων παραδειγμάτων W, r_r είναι ο αριθμός των σωστών κατηγοριοποιήσεων των παραδειγμάτων W(=n_r) και $\hat{p}$ = (r_o+r_r)/(n_o+n_r). Εάν η P-value, P, είναι μικρότερη από ένα επίπεδο σημασίας, τότε η μηδενική υπόθεση ($r_o/n_o = r_r/n_r$) απορρίπτεται και η εναλλακτική υπόθεση ($r_o/n_o > r_r/n_r$) γίνεται αποδεκτή, δηλαδή ένα concept drift έχει ανιχνευτεί. Το STEPD χρησιμοποιεί δύο επίπεδα σημασίας: α_w και α_d. Αποθηκεύει παραδείγματα στη βραχυπρόθεσμη μνήμη καθώς το P <α_w είναι ικανοποιημένο. Στη συνέχεια, αναδομεί τον κατηγοριοποιητή από τα αποθηκευμένα παραδείγματα και επαναφέρει όλες τις μεταβλητές εάν P<α_d. Αξίζει να σημειωθεί ότι η μέθοδος αρχίζει να ανιχνεύει μια εννοιολογική απόκλιση αφού ικανοποιήσει την υπόθεση $n_o + n_r \geq 2_W$ και τα αποθηκευμένα παραδείγματα αφαιρούνται εάν P ≥ α_w[1].

SEED: Detecting Volatility Shift in Data Streams

Ο ρυθμός αλλαγής (ή η ανίχνευση μεταβλητότητας) στη ροή συζητείται στο SEED Huang et al., 2014 [4]. Έχει ήδη συζητηθεί η εννοιολογική απόκλιση, αλλά πόσο συχνά προκύπτει η αλλαγή είναι επίσης μια σημαντική παράμετρος. Το SEED καθορίζει το σημείο κοπής για την ανίχνευση μιας εννοιολογικής απόκλισης. Βασίζεται σε ένα μηχανισμό συρόμενου παραθύρου (sliding windows), με μέγεθος μπλοκ=3, επιπρόσθετα το σημείο κοπής συμβολίζεται με το σύμβολο «j». Τα διαδοχικά μπλοκ ελέγχονται για να βρεθεί

ομοιογένεια. Εάν αυτή υπάρχει, τα δύο μπλοκ συγχωνεύονται (γνωστό ως συμπίεση μπλοκ). Σε αυτή τη διαδικασία, τα σημεία κοπής που έχουν χαμηλό δυναμικό αφαιρούνται. Στην ανίχνευση μεταβλητότητας, βρίσκει το διάστημα μεταξύ των σημείων κοπής. Εδώ, το σημείο αποκοπής σημαίνει ότι είναι ένα σημείο όπου συμβαίνει μια εννοιολογική απόκλιση.

Εάν υπάρχουν τρία διαστήματα, σημαίνει ότι δεν υπάρχει μεταβολή στη μεταβλητότητα, και εάν είναι έξι, υπάρχει αλλαγή στη μεταβλητότητα. Η μέθοδος SEED χρησιμοποιεί την ανισότητα Hoeffding με διόρθωση Bonferroni για να βρει το σημείο κοπής, η οποία έχει καλή απόδοση, όσον αφορά τον χρόνο εκτέλεσης και στις περισσότερες περιπτώσεις απαιτεί λιγότερη μνήμη από το ADWIN2 [4].

3.7.2.3 Μέθοδοι διαδοχικής ανάλυσης.

Δοκιμές CUSUM και Page-Hinkley

Η δοκιμή σωρευτικού αθροίσματος (CUSUM) [6] έχει σχεδιαστεί για να δίνει συναγερμό όταν ο μέσος όρος των δεδομένων εισόδου αποκλίνει σημαντικά από την προηγούμενη τιμή του. Στην απλούστερη μορφή της, η δοκιμή CUSUM έχει ως εξής: δεδομένης μιας ακολουθίας παρατηρήσεων $\{x_t\}_t$, ορίστε $z_t = (x_t - \mu)/\sigma$, όπου μ είναι η αναμενόμενη τιμή του x_t και σ είναι η τυπική τους απόκλιση σε "κανονικές" συνθήκες. Εάν μ και σ δεν είναι εκ των προτέρων γνωστά, υπολογίζονται από την ίδια την ακολουθία. Στη συνέχεια, το CUSUM υπολογίζει τους δείκτες και τον συναγερμό:

$$g_0 = 0,$$

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

Αν $g_t > h$, δήλωση αλλαγής και επαναφορά $g_t = 0$, και μ και σ. (13)

Το CUSUM είναι χωρίς μνήμη και χρησιμοποιεί συνεχή χρόνο επεξεργασίας ανά στοιχείο. Πώς η συμπεριφορά του εξαρτάται από τις παραμέτρους k και h είναι δύσκολο να αναλυθεί ακριβώς. Μια κατευθυντήρια γραμμή είναι να οριστεί το k στο ήμισυ της τιμής των μεταβολών που πρέπει να ανιχνευθούν (μετρούμενες σε τυπικές αποκλίσεις) και το h σε ln(1/δ) όπου δ είναι ο αποδεκτός ρυθμός ψευδούς συναγερμού. Οι τιμές στην περιοχή 3 έως 5 είναι τυπικές. Γενικά, η είσοδος z_t στο CUSUM μπορεί να είναι οποιοδήποτε υπολειπόμενο φίλτρο, για παράδειγμα, το σφάλμα πρόβλεψης ενός φίλτρου Kalman.

Μια παραλλαγή της δοκιμής CUSUM είναι η δοκιμή Page-Hinkley:

$$g_0 = 0,$$

$$g_t = g_{t-1} + z_t - k,$$

$$G_t = \min\{g_t, G_{t-1}\},$$

Αν $g_t - G_t > h$, δήλωση αλλαγής και επαναφορά $g_t = 0, G_t$, και μ και σ. (14)

Αυτή η μέθοδος ανίχνευσης αλλαγών λειτουργεί με τον υπολογισμό των παρατηρούμενων τιμών και του μέσου όρου τους μέχρι την τρέχουσα στιγμή. Το Page-Hinkley δεν σηματοδοτεί ζώνες προειδοποίησης, μόνο ανιχνεύσεις αλλαγών. Η μέθοδος λειτουργεί μέσω του τεστ Page-Hinkley. Σε γενικές γραμμές, θα ανιχνεύσει μια εννοιολογική απόκλιση εάν ο παρατηρούμενος μέσος όρος σε κάποια στιγμή είναι μεγαλύτερος από μια τιμή κατωφλίου λάμδα.

Αυτές οι μέθοδοι είναι μονόπλευρες με την έννοια ότι εγείρουν συναγερμό μόνο όταν αυξάνεται ο μέσος όρος[6][10].

Κεφάλαιο 4 - Αλγόριθμοι κατηγοριοποίησης ροών δεδομένων

4.1 Βασικοί κατηγοριοποιητές

4.1.1 Κατηγοριοποιητής πλειοψηφίας(Majority Class)

Ο αλγόριθμος αυτός είναι ένας από τους απλούστερους κατηγοριοποιητές, προβλέπει ότι η ετικέτα που αντιστοιχεί κλάση για κάθε ένα νέο εισερχόμενο παράδειγμα είναι η συχνότερη κλάση. Χρησιμοποιείται κυρίως ως γραμμή βάσης, αλλά και ως προεπιλεγμένος κατηγοριοποιητής στα φύλλα των δέντρων αποφάσεων [10].

4.1.2 Κατηγοριοποιητής χωρίς αλλαγή(No-change Classifier)

Ένας άλλος απλός κατηγοριοποιητής για ροές δεδομένων είναι ο κατηγοριοποιητής χωρίς αλλαγή, ο οποίος προβλέπει ότι η ετικέτα που αντιστοιχεί σε ένα νέο εισερχόμενο παράδειγμα θα είναι η πραγματική ετικέτα του προηγούμενου παραδείγματος. Όπως και ο κατηγοριοποιητής πλειοψηφίας, δεν απαιτεί τα χαρακτηριστικά του παραδείγματος, επομένως είναι πολύ εύκολο να εφαρμοστεί. Στην περίπτωση ανίχνευσης εισβολής όπου ακολουθούνται από μεγάλα περάσματα "χωρίς εισβολή" με συντομότερες περιόδους "εισβολής", αυτός ο κατηγοριοποιητής κάνει λάθη μόνο στις οριακές περιπτώσεις, προσαρμόζοντας γρήγορα ένα συνεπές μοτίβο των ετικετών[10].

4.1.3 Naïve Bayes (NB)

Βασίζεται στο θεώρημα του Bayes, εκτελεί την κλασική bayesian πρόβλεψη ενώ κάνει την αφελή υπόθεση ότι όλες οι είσοδοι είναι ανεξάρτητες. Ο Naïve Bayes είναι ένας αλγόριθμος κατηγοριοποίησης γνωστός για την απλότητα και το χαμηλό υπολογιστικό του κόστος. Είναι μια ειδική περίπτωση αλγορίθμου που δεν χρειάζεται προσαρμογή στις ροές

δεδομένων. Αυτό οφείλεται στο γεγονός ότι είναι απλή η σταδιακή εκπαίδευση του και δεν προσθέτει δομή στο μοντέλο, έτσι ώστε η χρήση της μνήμης να είναι μικρή και περιορισμένη. Δεδομένου n_c διαφορετικών κλάσεων, ο αλγόριθμος Naïve Bayes προβλέπει για κάθε παράδειγμα **I**, χωρίς ετικέτα, την κλάση C στην οποία ανήκει με μεγάλη ακρίβεια.

Το μοντέλο λειτουργεί ως εξής: Έστω $x_1, \ldots, x_k$ είναι k διακριτές ιδιότητες χαρακτηριστικών και υποθέτει ότι το x_i μπορεί να πάρει n_i διαφορετικές τιμές. Έστω C το χαρακτηριστικό κλάσης, το οποίο μπορεί να πάρει διαφορετικές τιμές n_C. Κατά τη λήψη ενός παραδείγματος χωρίς ετικέτα $I = (x_1 = v_1, \ldots, x_k = v_k)$, ο κατηγοριοποιητής Naïve Bayes υπολογίζει την "πιθανότητα" το I να ανήκει στην κατηγορία c ως:

$$\boldsymbol{Pr}[\boldsymbol{C} = \boldsymbol{c}|\boldsymbol{I}] \cong \prod_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{k}} \boldsymbol{Pr}[\boldsymbol{x_i} = \boldsymbol{v_i}|\boldsymbol{C} = \boldsymbol{c}]$$

$$= \boldsymbol{Pr}[\boldsymbol{C} = \boldsymbol{c}] \prod_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{k}} \frac{\boldsymbol{Pr}[\boldsymbol{x_i} = \boldsymbol{v_i}|\boldsymbol{C} = \boldsymbol{c}]}{\boldsymbol{Pr}[\boldsymbol{C} = \boldsymbol{c}]} \quad (15)$$

Οι τιμές $Pr[x_i = v_j \wedge C = c]$ και $Pr[C = c]$ υπολογίζονται από τα δεδομένα εκπαίδευσης. Έτσι, η περίληψη των δεδομένων εκπαίδευσης είναι απλώς ένας τρισδιάστατος πίνακας που αποθηκεύει για κάθε τριπλό (x_i , v_j , c) ένα πλήθος $N_{i,j,c}$ περιπτώσεων εκπαίδευσης με $x_i = v_j$, μαζί με έναν μονοδιάστατο πίνακα για μετράει το C = c. Αυτός ο αλγόριθμος είναι φυσικά αυξητικός, μόλις λάβει ένα νέο παράδειγμα (ή μια παρτίδα νέων παραδειγμάτων), απλώς αυξάνει τις σχετικές μετρήσεις. Οι προβλέψεις μπορούν να γίνουν ανά πάσα στιγμή από τις τρέχουσες μετρήσεις [10][38].

4.2 Δένδρα αποφάσεων

Τα δέντρα αποφάσεων είναι μια πολύ δημοφιλής τεχνική κατηγοριοποίησης, καθώς είναι πολύ εύκολο να ερμηνευθούν και να απεικονιστούν τα μοντέλα δέντρων. Σε ένα δέντρο αποφάσεων, κάθε εσωτερικός κόμβος αντιστοιχεί σε ένα χαρακτηριστικό που διαιρείται σε έναν κλάδο για κάθε τιμή χαρακτηριστικού και τα φύλλα αντιστοιχούν σε προγνωστικά κατηγοριοποίησης, συνήθως κατηγοριοποιητές κλάσης πλειοψηφίας. Παρακάτω περιγράφονται τα δέντρα αποφάσεων που είναι κατάλληλα για ροές δεδομένων.

4.2.1 Hoeffding Tree (HT)

Τα δέντρα Hoeffding εισήχθησαν από τους Domingos και Hulten το 2000 [9] . Αναφέρονται στην υλοποίησή τους ως VFDT, ένα αρκτικόλεξο για το Very Fast Decision Tree Learner. Ένα VFDT δέντρο είναι ένας σταδιακός αλγόριθμος επαγωγής ενός δέντρου αποφάσεων που ανά πάσα στιγμή είναι ικανός να μάθει από τεράστιες ροές δεδομένων, υποθέτοντας ότι τα

παραδείγματα που δημιουργούν την διανομή δεν αλλάζουν με την πάροδο του χρόνου. Η βασική ιδέα προέρχεται από την παρατήρηση ότι αρκεί να ληφθεί υπόψη μόνο ένα μικρό υποσύνολο των παραδειγμάτων εκπαίδευσης που διέρχονται από έναν κόμβο, για να επιλέξει, ο αλγόριθμος, το βέλτιστο χαρακτηριστικό διαχωρισμού για αυτό τον κόμβο. Αυτή η ιδέα υποστηρίζεται μαθηματικά από το όριο Hoeffding, το οποίο ποσοτικοποιεί τον αριθμό των παρατηρήσεων (παραδειγμάτων) που απαιτούνται για την εκτίμηση ορισμένων στατιστικών με μια προδιαγεγραμμένη ακρίβεια (στην περίπτωσή αυτή, την καλή ποιότητα ενός χαρακτηριστικού). Πιο συγκεκριμένα, το όριο Hoeffding δηλώνει ότι με πιθανότητα 1−δ, ο πραγματικός μέσος όρος μιας τυχαίας μεταβλητής εύρους R δεν θα διαφέρει από τον εκτιμώμενο μέσο όρο μετά n ανεξάρτητες παρατηρήσεις από:

$$\epsilon = \sqrt{\frac{R^2 \ln(1/\delta)}{2n}} \quad (16)$$

Στο VFDT ένα δέντρο απόφασης κατασκευάζεται αναδρομικά αντικαθιστώντας τα φύλλα με κόμβους απόφασης. Κάθε φύλλο αποθηκεύει τα επαρκή στατιστικά στοιχεία σχετικά με τις τιμές χαρακτηριστικών των εισερχόμενων παραδειγμάτων. Τα επαρκή στατιστικά στοιχεία είναι εκείνα που χρειάζονται μια ευρετική συνάρτηση αξιολόγησης που εκτιμά την αξία των δοκιμών διαχωρισμού με βάση τις τιμές των χαρακτηριστικών. Όταν ένα παράδειγμα είναι διαθέσιμο, διασχίζει το δέντρο από τη ρίζα σε ένα φύλλο, σε κάθε κόμβο αξιολογεί το κατάλληλο χαρακτηριστικό και ακολουθεί τον κλάδο τον οποίο η τιμή χαρακτηριστικού του αντιστοιχεί με το παράδειγμα. Όταν το παράδειγμα φτάσει σε ένα φύλλο, τα επαρκή στατιστικά στοιχεία ενημερώνονται. Στη συνέχεια, αξιολογείται κάθε πιθανή συνθήκη που βασίζεται στις τιμές χαρακτηριστικών. Εάν υπάρχει αρκετή στατιστική υποστήριξη υπέρ μιας δοκιμής έναντι όλων των άλλων, το φύλλο μετατρέπεται σε κόμβο απόφασης. Ο νέος κόμβος απόφασης θα έχει τόσα φύλλα απογόνους, όσες είναι οι πιθανές τιμές για την επιλεγμένη δοκιμή (επομένως αυτό το δέντρο δεν είναι απαραίτητα δυαδικό). Οι κόμβοι απόφασης διατηρούν μόνο τις πληροφορίες σχετικά με τη δοκιμή διαχωρισμού που είναι αποθηκευμένη σε αυτόν τον κόμβο.

Πιο συγκεκριμένα, αρχικά, γίνεται η χρήση του ορίου Hoeffding για να αποφασιστεί το μέγεθος του δείγματος που θα παρατηρηθεί έπειτα πραγματοποιείται η δοκιμή διαχωρισμού σε κάθε φύλλο. Έστω H() η συνάρτηση αξιολόγησης ενός χαρακτηριστικού. Για το κέρδος πληροφοριών, το εύρος R, της H( ) είναι $\log_2$(#classes). Ας υποτεθεί ότι μετά την παρατήρηση n παραδειγμάτων σε ένα δεδομένο φύλλο, το X_a είναι το χαρακτηριστικό με την υψηλότερη τιμή H( ), και το X_b το χαρακτηριστικό με τη δεύτερη υψηλότερη H(). Έστω $\overline{\Delta H} = \overline{H}(x_a) - \overline{H}(x_b)$ η διαφορά μεταξύ των δύο καλύτερων χαρακτηριστικών. Τότε, εάν ισχύει $\overline{\Delta H} > \epsilon$ δηλώνει με πιθανότητα 1-δ, ότι το X_a είναι πραγματικά το χαρακτηριστικό με την

υψηλότερη τιμή στη συνάρτηση αξιολόγησης στο σύμπαν, π.χ. ακόμα και αν ελεγχτούν άπειρα παραδείγματα. Σε αυτή την περίπτωση το φύλλο πρέπει να μετατραπεί σε κόμβο απόφασης που χωρίζεται στο X_a. Εάν $\overline{\Delta H} < \epsilon$, το μέγεθος του δείγματος δεν είναι αρκετό για να ληφθεί μια σταθερή απόφαση. Οφείλεται να επεκταθεί το δείγμα βλέποντας περισσότερα παραδείγματα. Καθώς το μέγεθος του δείγματος αυξάνεται, το ϵ μειώνεται και εάν υπάρχει ένα πληροφοριακό χαρακτηριστικό, θα προωθηθεί προς τα πάνω.

Η αξιολόγηση της συνάρτησης διαχωρισμού για κάθε παράδειγμα μπορεί να είναι πολύ δαπανηρή. Μάλιστα, αποδεικνύεται ότι δεν είναι αποτελεσματικό να υπολογίζετε το H( ) κάθε φορά που έρχεται ένα νέο παράδειγμα. Το VFDT υπολογίζει τη συνάρτηση αξιολόγησης χαρακτηριστικών H( ) μόνο όταν έχει παρατηρηθεί ένας ελάχιστος αριθμός παραδειγμάτων από την τελευταία αξιολόγηση. Αυτός ο ελάχιστος αριθμός παραδειγμάτων συνιστά μια παράμετρο που ορίζει ο χρήστης. Όταν δύο ή περισσότερα χαρακτηριστικά παρουσιάζουν συνεχώς πολύ παρόμοιες τιμές της συνάρτησης H( ), ακόμη και για μεγάλο αριθμό παραδειγμάτων, το όριο Hoeffding δεν θα αποφασίσει μεταξύ τους (αναφέρεται ως ισοπαλία). Αυτή η κατάσταση μπορεί να συμβεί ακόμη και για δύο ή περισσότερα εξίσου πληροφοριακά χαρακτηριστικά. Για την επίλυση αυτού του προβλήματος, το VFDT χρησιμοποιεί μια σταθερά **τ** που εισάγει ο χρήστης για την απορροή. Λαμβάνοντας υπόψη ότι το ϵ μειώνεται όταν το n αυξάνεται, εάν $\overline{\Delta H} < \epsilon < \tau$ τότε το φύλλο μετατρέπεται σε κόμβο απόφασης. Η δοκιμή διαχωρισμού βασίζεται στο καλύτερο χαρακτηριστικό.

Πολλές βελτιώσεις στον βασικό αλγόριθμο VFDT έχουν προταθεί. Μια διαφορετική προσέγγιση για την αντιμετώπιση των αριθμητικών ιδιοτήτων προτάθηκε από τους Gama et al. [19]. Χρησιμοποιούν δυαδικά δέντρα ως τρόπο δυναμικής διακριτοποίησης αριθμητικών τιμών. Η ίδια εργασία διερευνά επίσης τη χρήση ενός πρόσθετου κατηγοριοποιητή σε κόμβους φύλλων, δηλαδή του Naive Bayes.

Το VFDT δεν μπορεί να χειριστεί μια εννοιολογική απόκλιση, γιατί μόλις δημιουργηθεί ένας κόμβος, δεν μπορεί ποτέ να αλλάξει. Οι Hulten et al. [20] παρουσίασαν τον αλγόριθμο πολύ γρήγορου δέντρου αποφάσεων (CVFDT) ως επέκταση του VFDT για την αντιμετώπιση του της εννοιολογικής απόκλισης. Το CVFDT διατηρεί ένα μοντέλο που είναι συνεπές με τις παρουσίες που είναι αποθηκευμένες σε ένα συρόμενο παράθυρο σταθερού μεγέθους, για να προσδιορίσει ποιοι κόμβοι γερνούν και μπορεί να χρειάζονται ενημέρωση. Η μέθοδος αυτή δεν είναι διαθέσιμη στο MOA, αντιθέτως έχει ενταχθεί το παρακάτω προσαρμοστικό δέντρο[9][14].

4.2.3 Hoeffding Adaptive Tree (HAT)

Το Hoeffding Adaptive Tree προτάθηκε από Bifet A. & Gavalda R[29]. Το HAT χρησιμοποιεί τον ανιχνευτή και εκτιμητή αλλαγών ADWIN, για την παρακολούθηση της απόδοσης των κλαδιών στο δέντρο και για την αντικατάστασή τους με νέους κλάδους όταν η ακρίβειά τους μειώνεται εάν τα νέα κλαδιά είναι πιο ακριβή.

4.3 Κατηγοριοποιητής εννοιολογικής απόκλισης

SingleClassifierDrift

Ο κατηγοριοποιητής SingleClassifierDrift δημιουργήθηκε για το χειρισμό μιας εννοιολογικής απόκλισης σε ροές δεδομένων, με βάση τη χρήση ενός ανιχνευτή αλλαγών, παρακολουθώντας το ποσοστό σφάλματος του κατηγοριοποιητή. Όταν ο ανιχνευτής αλλαγών εγείρει ένα προειδοποιητικό σήμα, δημιουργείται ένας νέος κατηγοριοποιητής και όταν ανυψώνεται ένα σήμα αλλαγής, ο τρέχων κατηγοριοποιητής αντικαθίσταται από ένα νέο. Είναι η στρατηγική στη μέθοδο DDM των Gama et al. [8], που περιγράφεται στην ενότητα 3.7.2. Υπάρχουν διάφορες μέθοδοι ανίχνευσης αλλαγών, που έχουν είδη περιγραφεί στην ενότητα 3.7.2, που μπορούν να χρησιμοποιηθούν. Βασικός κατηγοριοποιητής στη μέθοδο αυτή μπορεί να είναι ο NB ή το HT δέντρο, με σκοπό να δημιουργηθεί μια ενημερωμένη μέθοδος μιας εννοιολογικής απόκλισης, βέβαια μπορεί να επιλεχτεί κάθε κατηγοριοποιητής που είναι διαθέσιμος στο MOA [38].

4.4 Συνδυαστικοί αλγόριθμοι μάθησης (Ensembles)

Η πρόβλεψη ενός συνδυαστικού αλγορίθμου μάθησης είναι ένας συνδυασμός προβλέψεων μικρότερων μοντέλων. Πρόκειται για αλγορίθμους μάθησης που εκτελούνται ανεξάρτητα ή και εξαρτημένα (πχ μέθοδος boosting), και οι μεμονωμένες προβλέψεις τους συνδυάζονται με κάποιο τρόπο (π.χ. μέσος όρος ή ψηφοφορία) για την εξαγωγή μιας τελική πρόβλεψης[37]. Τόσο σε σενάρια παρτίδας όσο και σε σενάρια ροής, οι συνδυαστικοί αλγόριθμοι μάθησης τείνουν να βελτιώνουν την ακρίβεια πρόβλεψης, με κόστος τη χρήση περισσότερων πόρων, χρόνου και μνήμης.

Στη ροή, οι συνδυαστικοί αλγόριθμοι μάθησης έχουν πρόσθετα πλεονεκτήματα έναντι των μεθόδων ενός αλγόριθμου κατηγοριοποίησης είναι εύκολο να κλιμακωθούν και εάν εξορύσσουν μια κατανεμημένη ροή, δεν απαιτούν τη συγκέντρωση των διαφορετικών ροών σε ένα αλγόριθμο. Επίσης, οι συνδυαστικοί αλγόριθμοι μάθησης μπορούν να προσαρμοστούν στις αλλαγές με το κλάδεμα των τμημάτων τους, που υπολειτουργούν και την προσθήκη νέων κατηγοριοποιητών[10].

4.4.1 Accuracy Weighted Ensemble (AWE)

Στο πλαίσιο της μηχανικής μάθησης ένα απλό σχήμα συνδυαστικού αλγορίθμου μάθησης(ensembles) αποτελείται από N ένα μοντέλο κατηγοριοποίησης. Για κάθε κλάση C εκπαιδεύει N μοντέλα και για μια δεδομένη είσοδο x παράγουν $C_1(x),\ldots,C_N(x)$, προβλέψεις. Έπειτα εφαρμόζεται κάποια συνάρτηση $f(C_1(x),\ldots,C_N(x))$, όπου $C_i(x)$ υποδηλώνει την πρόβλεψη της C_i για x. Τα μοντέλα κατηγοριοποίησης μπορεί να διορθωθούν από την αρχή, ίσως να εκπαιδευτούν offine από δεδομένα παρτίδας ή να εκπαιδευτούν online στην ίδια ροή.

Στο πλαίσιο της κατηγοριοποίησης, η συνάρτηση f μπορεί απλά να ψηφίζει, δηλαδή να παράγει την πιο δημοφιλή κλάση μεταξύ του $C_i(x)$. Στη σταθμισμένη ψηφοφορία, κάθε κατηγοριοποιητής C_i έχει βάρος w_i που επισυνάπτεται στην ψηφοφορία. Τα βάρη μπορούν να καθοριστούν για όλη τη διαδικασία, να καθοριστούν από έναν εμπειρογνώμονα ή να ποικίλλουν με την πάροδο του χρόνου [10].

Τα Σύνολα σταθμισμένης ακρίβειας (AWE) προτάθηκαν από τους Wang et al. [22]. Ένα σύνολο AWE χωρίζει μια εισερχόμενη ροή σε διαδοχικά κομμάτια $S_1, S_2, \ldots, S_n$ ίσου μεγέθους W (chunksize), με το S_n να είναι το πιο ενημερωμένο κομμάτι. Δημιουργεί έναν νέο κατηγοριοποιητή C_i για κάθε εισερχόμενο κομμάτι δεδομένων S_i και χρησιμοποιεί τα δεδομένα του S_i για να αξιολογήσει τα υπάρχοντα μέλη του συνόλου και να επιλέγουν οι καλύτεροι κατηγοριοποιητές.

Οι καλύτεροι βασικοί κατηγοριοποιητές επιλέγονται σύμφωνα με τα βάρη w_i που υπολογίζονται για κάθε κατηγοριοποιητή C_i. Οι κατηγοριοποιητές σταθμίζονται με βάση την αναμενόμενη ακρίβεια κατηγοριοποίησης στα δεδομένα δοκιμής. Για να σταθμιστούν σωστά οι κατηγοριοποιητές, οφείλεται να είναι γνωστή η πραγματική λειτουργία εκπαίδευσης, η οποία δεν είναι διαθέσιμη σε ένα περιβάλλον ροής. Επομένως, οι συγγραφείς του AWE προτείνουν τον υπολογισμό τιμών των βαρών των κατηγοριοποιητών υπολογίζοντας το ποσοστό σφάλματος στο πιο πρόσφατο κομμάτι δεδομένων S_n. Υποθέτοντας ότι η κατανομή των κλάσεων του κομματιού S_n, είναι πιο κοντά στην κατανομή των κλάσεων των δεδομένων της τρέχουσας δοκιμής.

Πιο συγκεκριμένα, υποθέτοντας ότι το κομμάτι των πρόσφατων δεδομένων S_n περιέχει παραδείγματα της μορφής (x,c), όπου c είναι η ετικέτα που αντιστοιχεί στην πραγματική κλάση του παραδείγματος. Το σφάλμα κατηγοριοποίησης του C_i για το παράδειγμα (x,c) είναι $1-f_c^i(x)$, όπου $f_c^i(x)$ είναι η πιθανότητα που δίνεται από τον κατηγοριοποιητή i ότι το x είναι ένα παράδειγμα της κλάσης c. Το μέσο τετραγωνικό σφάλμα (MSE) του C_i μπορεί να εκφραστεί ως:

$$MSE_i=\frac{1}{|S_n|}\sum_{(x,c)\in S_n}(1-f_c^i(x))^2 \quad (17)$$

Το βάρος του C_i είναι αντιστρόφως ανάλογο προς το MSE_i. Από την άλλη πλευρά, ένας κατηγοριοποιητής που προβλέπει τυχαία, την πιθανότητα το x να ανήκει στην κλάση c ισούται την κατανομή της κλάσης p(c), θα έχει μέσο τετραγωνικό σφάλμα:

$$MSE_r = p(c)(1 - p(c))^2 \quad (18)$$

Αν, για παράδειγμα, c ∈{0,1} και η κατανομή της κλάσης είναι ομοιόμορφη, το MSE_r =0,25. Δεδομένου ότι ένα τυχαίο μοντέλο δεν περιέχει χρήσιμη γνώση σχετικά με τα δεδομένα. Χρησιμοποιείται το ποσοστό σφάλματος του τυχαίου κατηγοριοποιητή ως κατώφλι για τη στάθμιση των κατηγοριοποιητών. Απορρίπτονται δηλαδή οι κατηγοριοποιητές που το σφάλμα τους είναι ίσο ή μεγαλύτερο από το MSE_r. Επιπλέον, για να γίνει πιο εύκολα ο υπολογισμός αυτός δίνεται σε κάθε C_i κατηγοριοποιητή ένα βάρος w_i [22]:

$$w_i = MSE_r - MSE_i \quad (19)$$

Τα πλεονεκτήματα αυτής της μεθόδου είναι η απλότητά της και το γεγονός ότι μπορεί να χρησιμοποιηθεί με βασικούς κατηγοριοποιητές παραδοσιακής μηχανικής μάθησης και να λειτουργήσει καλά τόσο σε σταθερές όσο και σε μη σταθερές ροές. Ένα μειονέκτημα της μεθόδου είναι ότι το μέγεθος του κομματιού πρέπει να προσδιοριστεί εξωτερικά, από έναν εμπειρογνώμονα στον τομέα, λαμβάνοντας υπόψη την καμπύλη μάθησης των βασικών κατηγοριοποιητών [10][14].

4.4.2 Accuracy Updated Ensemble (AUE)

Οι Brzezinski και Stefanowski [21] έχουν προτείνει τον αλγόριθμο Accuracy Updated Ensemble (AUE). Ο προτεινόμενος αλγόριθμος είναι εμπνευσμένος από τον αλγόριθμο AWE και τον μηχανισμό στάθμισής του, αλλά βελτιώνει τα ελαττώματα του. Η μέθοδος AUE δεν επιλέγει μόνο τους καλύτερους κατηγοριοποιητές, αλλά τους ενημερώνει σύμφωνα με την τρέχουσα κατανομή της ροής. Στην μέθοδο AWE τα κομμάτια της εισερχόμενης ροής επιτρέπεται να εκπαιδευτούν με αλγόριθμους παραδοσιακής κατηγοριοποίησης που χρησιμοποιούνται σε παρτίδες (όχι ειδικούς διαδικτυακούς) και αργότερα να προσαρμοστούν τα βάρη τους σύμφωνα με την τρέχουσα κατανομή. Στην μέθοδο AUE, οι συγγραφείς του στρέφονται στους διαδικτυακούς αλγόριθμους κατηγοριοποίησης. Αυτό επιτρέπει την ενημέρωση των βασικών κατηγοριοποιητών αντί να προσαρμόζονται μόνο το βάρος τους. Εάν δεν συμβεί αλλαγή μεταξύ μιας σειράς κομματιών, η κατηγοριοποίηση θα βελτιωθεί ακριβώς σαν να ήταν χτισμένη σε ένα μεγαλύτερο κομμάτι δεδομένων το οποίο είναι πιο κατάλληλο για περιόδους σταθερότητας. Ως αποτέλεσμα, μπορεί να μειωθεί το μέγεθος του κομματιού χωρίς να υπάρχει κίνδυνος μείωσης της ακρίβειας

κατηγοριοποίησης. Επιπλέον, διατηρεί τα βασικά στοιχεία του μηχανισμού στάθμισης του AWE και αποσβένει κατηγοριοποιητές εάν συμβεί μια ξαφνική εννοιολογική απόκλιση. Ο συνδυασμός επιλογής κατηγοριοποιητή και ενημέρωσης τους υποστηρίζεται ότι κάνει την μέθοδο AUE καλύτερη από την μέθοδο AWE σε περιόδους σταθερότητας ή σταδιακής εννοιολογικής απόκλισης, ενώ είναι τουλάχιστον εξίσου ακριβής όταν προκύψει μια ξαφνική εννοιολογική απόκλιση.

Ένα άλλο μειονέκτημα της μεθόδου AWE είναι η λειτουργία στάθμισής του. Επειδή ο αλγόριθμος έχει σχεδιαστεί για να αποδίδει καλά σε δεδομένα που είναι ευαίσθητα στο κόστος, το όριο MSE_r στην Εξίσωση 15 μειώνει τις κατηγοριοποιήσεις όλα τα μέλη του συνδυαστικού αλγορίθμου μάθησης(ensembles) δεν προβλέπουν την κλάση των εισερχόμενων παραδειγμάτων. Για να αποφευχθεί αυτό, στο AUE προτείνετε μια απλούστερη συνάρτηση στάθμισης:

$$w_i = \frac{1}{MSE_i + \varepsilon} \qquad (20)$$

Το MSE_i υπολογίζεται ακριβώς όπως στην Εξίσωση 17 και το ε είναι μια πολύ μικρή σταθερή τιμή, η οποία επιτρέπει τον υπολογισμό του βάρους σε σπάνιες περιπτώσεις όταν $MSE_i = 0$. Στη μέθοδο επιδιώκεται να ενημερωθούν τα μέλη των κατηγοριοποιητών σύμφωνα με την τρέχουσα κατανομή, διατηρώντας παράλληλα την ποικιλομορφία τους. Για να το πετύχει αυτό η μέθοδος AUE, ενημερώνει μόνο επιλεγμένους κατηγοριοποιητές. Πρώτα απ 'όλα, λαμβάνει υπόψη μόνο τα τρέχοντα μέλη του συνόλου - τους k κορυφαίους κατηγοριοποιητές. Στη συνέχεια, χρησιμοποιεί το MSE_r ως κατώφλι για να επιτρέψει την διαδικτυακή ενημέρωση μόνο των κλάσεων που προβλέπουν αρκετή ακρίβεια [21].

4.4.3 Dynamic Weighted Majority (DWM)

Ένα θεμελιώδες έργο, είναι το σύστημα που παρουσιάζεται από τους Kolter και Maloof (ICDM03, ICML05)[15]. Ο αλγόριθμος Dynamic Weighted Majority (DWM) είναι μια μέθοδος συνδυαστικού αλγορίθμου μάθησης (ensembles) για την παρακολούθηση της εννοιολογικής απόκλισης. Διατηρεί ένα σύνολο βασικών κατηγοριοποιητών, που αναφέρονται ως εμπειρογνώμονες(expert) και ο καθένας έχει ένας βάρος. Οι εμπειρογνώμονες μπορούν να δημιουργηθούν από τον ίδιο αλγόριθμο βασικής εκπαίδευσης, αλλά σε διαφορετικά χρονικά βήματα, ώστε να χρησιμοποιούν διαφορετικά σύνολα παραδειγμάτων εκπαίδευσης. Όποτε είναι διαθέσιμο ένα παράδειγμα, ζητείται η γνώμη κάθε εμπειρογνώμονα, να κάνει μια πρόβλεψη της κλάσης του. Η τελική πρόβλεψη προκύπτει ως σταθμισμένη ψήφος τους. Για κάθε υποψήφια κλάση, ο αλγόριθμος DWM αθροίζει τα βάρη όλων αυτών των εμπειρογνωμόνων που την προβλέπουν και έπειτα

προβλέπει την κλάση με το μεγαλύτερο βάρος. Ο αλγόριθμος DWM, δηλαδή πρώτα προβλέπει την κλάση κατηγοριοποίησης του κάθε παραδείγματος εκπαίδευσης και έπειτα δημιουργεί και διαγράφει δυναμικά εμπειρογνώμονες ως απάντηση σε αλλαγές στην απόδοση τους. Τα βάρη όλων αυτών των εμπειρογνωμόνων που κατηγοριοποίησαν εσφαλμένα το παράδειγμα μειώνονται κατά μια πολλαπλασιαστική σταθερά β. Εάν η συνολική πρόβλεψη ήταν λανθασμένη, ένας νέος εμπειρογνώμονας προστίθεται στο σύνολο με βάρος ίσο με το συνολικό βάρος του συνόλου. Αφαιρεί τον εμπειρογνώμονα με το χαμηλότερο βάρος πριν προσθέσει το νέο μέλος. Τέλος, όλοι οι εμπειρογνώμονες εκπαιδεύονται στο παράδειγμα[14][15].

4.4.4 Paired Learners

Οι ερευνητές που σχεδιάζουν αλγόριθμους για να αντιμετωπίσουν μια εννοιολογική απόκλιση έχουν από καιρό αναγνωρίσει τη σημασία της εξισορρόπησης της αντιδραστικότητας και της σταθερότητας. Η εννοιολογική απόκλιση αναφέρεται σε μια εργασία εκπαίδευσης στο διαδίκτυο στην οποία ο στόχος αλλάζει με την πάροδο του χρόνου. Για την αντιμετώπιση της εννοιολογικής απόκλισης, οι Bach S. H.& Maloof M. A. [18], συνδύασαν έναν σταθερό διαδικτυακό μαθητή με έναν αντιδραστικό. Ένας σταθερός μαθητής προβλέπει με βάση όλη την εμπειρία του, ενώ ένας αντιδραστικός μαθητής προβλέπει με βάση την εμπειρία του σε ένα σύντομο, πρόσφατο χρονικό διάστημα, ένα παράθυρο. Η Βασική ιδέα είναι να χρησιμοποιηθεί η αλληλεπίδραση μεταξύ αυτών των δύο μαθητών και οι διαφορές τους στην ακρίβεια, σε αυτό το παράθυρο, για να αντιμετωπιστεί μια εννοιολογική απόκλιση. Ο σταθερός μαθητής ξεπερνά τις επιδόσεις του αντιδραστικού μαθητή όταν αλλάζει ο στόχος (concept change), αλλά ο αντιδραστικός κατηγοριοποιητής υπερτερεί του σταθερού κατηγοριοποιητή στην περίοδο μετά την αλλαγή του στόχου. Πράγματι, όταν ο αντιδραστικός κατηγοριοποιητής ξεπερνά τις επιδόσεις του σταθερού κατηγοριοποιητή σε σύντομο χρονικό διάστημα, τότε η μέθοδος της εκπαίδευσης σε ζεύγη αντικαθιστά τη γνώση του σταθερού κατηγοριοποιητή με αυτή του αντιδραστικού κατηγοριοποιητή. Στη συνέχεια, το ζευγάρι συνεχίζει να εκπαιδεύεται[18].

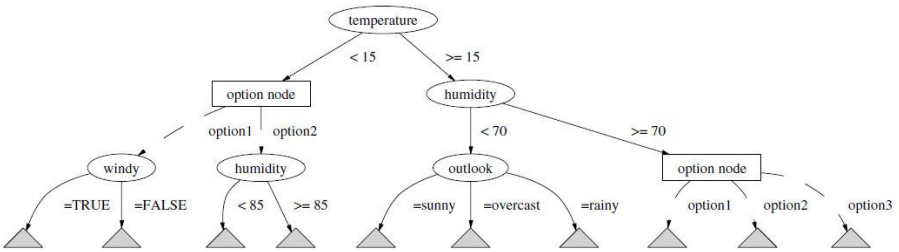
4.4.5 Learn++.NSE

Ο προτεινόμενος αλγόριθμος, Learn++.NSE, είναι μέλος της οικογένειας αλγορίθμων Learn++. Ο κοινός παρονομαστής σε όλους τους αλγόριθμους Learn++ είναι η δημιουργία ενός συνόλου κατηγοριοποιητών που εκπαιδεύονται σταδιακά (χωρίς πρόσβαση σε

προηγούμενα δεδομένα) στις εισερχόμενες παρτίδες (batch) δεδομένων. Έπειτα, οι αλγόριθμοι συνδυάζονται με κάποια μορφή ψηφοφορίας, με σταθμισμένη πλειοψηφία. Η μέθοδος Learn++.NSE είναι κατάλληλη για μη στάσιμα περιβάλλοντα, η καινοτομία της προσέγγισης είναι στον τρόπο καθορισμού του τρόπου των βαρών της ψήφου, με βάση τη χρονικά προσαρμοσμένη ακρίβεια, του κάθε κατηγοριοποιητή, σε τρέχοντα και προηγούμενα περιβάλλοντα. Αυτή η προσέγγιση επιτρέπει στον αλγόριθμο να αναγνωρίζει και να ενεργεί σε όλες τις εννοιολογικές αποκλίσεις (απότομες, σταδιακές, επαναλαμβανόμενες) αναλόγως στις υποκείμενες διανομές δεδομένων. Η μέθοδος Learn++.NSE είναι γενικά ανεξάρτητη από τον βασικό κατηγοριοποιητή, λειτουργεί καλύτερα εάν δεν χρησιμοποιείται κλάδεμα (ιδιαίτερα για επαναλαμβανόμενα περιβάλλοντα), και όπως ήταν αναμενόμενο, όσο πιο αργή είναι η εμφάνιση μιας εννοιολογικής απόκλισης, τόσο καλύτερα μπορεί να παρακολουθηθεί το περιβάλλον[28].

4.4.6 Hoeffding option tree (HOT)

Η μέθοδος Hoeffding option tree έχει τη μορφή ενός απλού δέντρου αποφάσεων Hoeffding Tree (HT), εκτός από την παρουσία πρόσθετων κόμβων επιλογής, που απεικονίζονται στην εικόνα (11) ως ορθογώνια. Σε αυτούς τους κόμβους θα εφαρμοστούν πολλαπλές δοκιμές, υπονοώντας ότι ένα παράδειγμα μπορεί να ταξιδέψει σε πολλαπλά μονοπάτια του δέντρου αποφάσεων και να φτάσει σε πολλαπλά φύλλα. Αποτελούν μια ενιαία δομή που αντιπροσωπεύει αποτελεσματικά πολλά δέντρα, ένα σύνολο δέντρων.



Εικόνα 14: Ενα Hoeffding Option Tree (HOT)

Ένα HOT περιέχει, εκτός από τους κόμβους που δοκιμάζουν ένα χαρακτηριστικό, πρόσθετους κόμβους επιλογής που δεν εφαρμόζουν καμία δοκιμή και απλώς διακλαδίζονται σε πολλά δευτερεύοντα δέντρα. Όταν ένα παράδειγμα φτάσει σε έναν κόμβο επιλογής, συνεχίζει να κατεβαίνει σε όλους τους απογόνους του κόμβου, φτάνοντας τελικά σε αρκετά φύλλα. Οι προβλέψεις της μεθόδου HOT γίνονται βάση της σταθμισμένης πλειοψηφίας των προβλέψεων όλων των φύλλων που εφαρμόζουν μια δοκιμή σε ένα παράδειγμα. Πειραματικές μελέτες απέδειξαν ότι αποδίδει καλύτερα από τη μη σταθμισμένη ψηφοφορία.

Οι πρόσθετοι κόμβοι επιλογών εισάγονται αν μετά τη λήψη ενός λογικού αριθμού παραδειγμάτων, η συνάρτηση αξιολόγησης χαρακτηριστικών του αλγορίθμου Hoeffding Tree εκτιμήσει ότι υπάρχουν πολλά παρόμοια χαρακτηριστικά (ισοπαλία). Η μέθοδος HOT δεν αποφασίζει αυθαίρετα πιο χαρακτηριστικό θα επικρατήσει, αντιθέτως προσθέτει όλα τα παρόμοια χαρακτηριστικά ως διαφορετικές επιλογές. Με αυτόν τον τρόπο, μετριάζονται οι επιπτώσεις της περιορισμένης προοπτικής και της αστάθειας, οδηγώντας τελικά σε ένα πιο ακριβές μοντέλο[40].

4.4.7 Adaptive Hoeffding option tree (AdaHOT)

Το Adaptive Hoeffding option tree για ροή δεδομένων με προσαρμοστική κατηγοριοποίηση Naive Bayes στα φύλλα. Ένα adaptive Hoeffding option tree είναι ένα Hoeffding option tree με την ακόλουθη βελτίωση, κάθε φύλλο αποθηκεύει μια εκτίμηση του τρέχοντος σφάλματος. Χρησιμοποιεί έναν εκτιμητή EWMA με α=0,2. Το βάρος κάθε κόμβου στην ψηφοφορία είναι η διαδικασία αντιστρόφως ανάλογη του τετράγωνου του σφάλματος. Υλοποιείται στο MOA ως AdaHoeffdingOptionTree.

4.4.8 Bagging

Η μέθοδος Bagging χρησιμοποιείται για τη μείωση της διακύμανσης. Εισήχθη από τον Breiman (1996). Στη μέθοδο αυτή επιλέγεται ένας βασικός αλγόριθμος εκπαίδευσης για να δημιουργηθεί ένα σύνολο M βασικών μοντέλων. Τα βασικά μοντέλα είναι δυνητικά διαφορετικά, επειδή το καθένα εκπαιδεύεται με ένα διαφορετικό bootstrap δείγμα μεγέθους N, που δημιουργείται με τη λήψη τυχαίων δειγμάτων με αντικατάσταση (επανατοποθέτηση) από το αρχικό σύνολο εκπαίδευσης. Το σύνολο εκπαίδευσης κάθε βασικού μοντέλου έχει το ίδιο μέγεθος με τα αρχικά δεδομένα, ορισμένα παραδείγματα δεν εμφανίζονται σε αυτό ενώ άλλα μπορεί να εμφανίζονται περισσότερες από μία φορές, περιέχει καθένα από τα αρχικά παραδείγματα εκπαίδευσης K φορές όπου το P(K = k) ακολουθεί μια διωνυμική κατανομή:

$$P(K = k) = \binom{n}{k} p^k (1-p)^{n-k} = \binom{n}{k} \frac{1}{n}^k \ (1 - \frac{1}{n})^{n-k} \quad (21)$$

Το μετά-μοντέλο που προκύπτει κάνει μια πρόβλεψη λαμβάνοντας την απλή πλειοψηφία των προβλέψεων των M κατηγοριοποιητών που δημιουργήθηκαν με αυτόν τον τρόπο. Η βασική ιδέα πίσω από τη μέθοδο bagging είναι ότι η ψηφοφορία μειώνει τη διακύμανση του δείγματος του βασικού αλγορίθμου, δηλαδή τη διαφορά μεταξύ κατηγοριοποιητών που εκπαιδεύονται από διαφορετικά δείγματα από την ίδια κατανομή πηγής. Στην

πραγματικότητα, λειτουργεί καλύτερα όσο υψηλότερη είναι η διακύμανση μεταξύ των κατηγοριοποιητών bootstrapped [33][10].

4.4.9 Online Bagging (OzaBag)

Η μέθοδος bagging φαίνεται να απαιτεί την πρόγνωση του μεγέθους του συνόλου εκπαίδευσης, το οποίο δεν είναι διαθέσιμο (ή χωρίς νόημα) στο διαδικτυακό πλαίσιο ροής. Επειδή, για κάθε βασικό μοντέλο, η δειγματοληψία με αντικατάσταση γίνεται με την εκτέλεση τυχαίων κληρώσεων σε ολόκληρο το σύνολο εκπαίδευσης.

Η διωνυμική κατανομή για μεγάλες τιμές του n τείνει σε μια κατανομή Poisson(1), όπου Poisson(1) = exp(−1)/k! (22). Χρησιμοποιώντας αυτό το γεγονός, ο Oza και ο Russell [33] πρότειναν τη διαδικτυακή μέθοδο Online Bagging, που αντί για δειγματοληψία με αντικατάσταση, δίνει σε κάθε παράδειγμα ένα βάρος σύμφωνα με τη κατανομή Poisson(1). Η διαδικτυακή έκδοση εκπαιδεύει διαδικτυακά (online) τα M βασικά μοντέλα. Προσομοιώνει τη διαδικασία bootstrap στέλνοντας K αντίγραφα κάθε νέου παραδείγματος για την ενημέρωση κάθε βασικού μοντέλου, όπου το K είναι μια κατάλληλη τυχαία μεταβλητή Poisson. Το απλό αυτό τέχνασμα αποφέρει μαθησιακή συμπεριφορά παρόμοια με αυτή του bagging, για παρτίδες (batch), αφαιρώντας την εξάρτηση από τον αριθμό των παραδειγμάτων. Έτσι σχεδιάστηκε ο αλγόριθμος αποθήκευσης για ροές δεδομένων ανοιχτού τύπου[14][33][10].

1: Initialize base models hm for all m 2 {1, 2, ...,M}
2: **for all** training examples **do**
3: **for** m = 1, 2, ...,M **do**
4: Set w = Poisson(1)
5: Update h_m with the current example with weight w
6: **anytime output:**
7: **return** hypothesis: $h_{fin}(x) = \arg max_{y \in Y} \sum_{t=1}^{T} I(h_t(x) = y)$

Πίνακας1:Oza and Russell's Online Bagging for M models

4.4.10 ADWIN Bagging (OzaBagAdwin)

Ένα πρόβλημα με την παραπάνω προσέγγιση είναι ότι δεν έχει σχεδιαστεί ειδικά για να αντιδρά σε μια εννοιολογική απόκλιση, εκτός εάν οι βασικοί κατηγοριοποιητές είναι οι ίδιοι εξαιρετικά προσαρμοστικοί. Η βασική ιδέα του ADWIN Bagging είναι να χρησιμοποιηθούν τα προσαρμοστικά δέντρα Hoeffding (HAT) αντί για τα μη προσαρμοστικά Hoeffding δέντρα (HT) ως βασικοί κατηγοριοποιητές για τη μέθοδο του συνόλου online bagging. Το ADWIN Bagging, που υλοποιήθηκε στο MOA ως OzaBagADWIN, βελτιώνει τη μέθοδο online bagging ως εξής, χρησιμοποιεί M παρουσίες του ADWIN για την παρακολούθηση του ποσοστού

σφάλματος των βασικών κατηγοριοποιητών. Όταν εντοπιστεί μια εννοιολογική απόκλιση, ο χειρότερος κατηγοριοποιητής του συνόλου αφαιρείται και προστίθεται ένας νέος κατηγοριοποιητής σε αυτό. Αυτή η στρατηγική ονομάζεται μερικές φορές "αντικαταστήστε τον ηττημένο"[16][10].

1: Initialize base models hm for all m 2 {1, 2, ...,M}
2: **for all** training examples **do**
3: **for** m = 1, 2, ...,M **do**
4: Set w = Poisson(1)
5: Update h_m with the current example with weight w
6: **if** ADWIN detects change in error of one of the classifiers **then**
7: Replace classifier with higher error with a new one
6: **anytime output:**
7: **return** hypothesis: $h_{fin}(x) = \arg max_{y \in Y} \sum_{t=1}^{T} I(h_t(x) = y)$

Πίνακας2:Adwin Bagging for M models

4.4.11 Leveraging Bagging

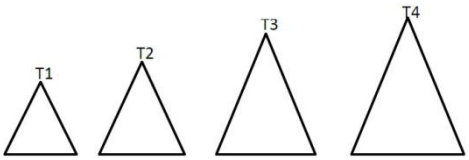
Στη μέθοδο Leveraging Bagging οδήγησε μια πειραματική παρατήρηση κατά τη χρήση της μεθόδου online bagging που έδειξε ότι η προσθήκη περισσότερης τυχαιότητας στην είσοδο βελτιώνει την απόδοση της. Δεδομένου ότι η προσθήκη τυχαιότητας αυξάνει την ποικιλομορφία ή τη διακύμανση των βασικών κατηγοριοποιητών.

Πρόσθετη τυχαιότητα σε μια μέθοδο μπορεί να εισαχθεί με δειγματοληψία με άλλες κατανομές εκτός της Poisson(1). Η μέθοδος Leveraging Bagging χρησιμοποιεί Poisson(> 1) για να σταθμίσει νέες παρουσίες για να κάνει διαδικτυακή δειγματοληψία. Εκτός από την προσθήκη τυχαιότητας στην είσοδο, η μέθοδος διαθέτει τη δυνατότητα επιλογής να προσθέσει και τυχαιοποίηση στην έξοδο του συνόλου χρησιμοποιώντας κωδικούς εξόδου.

Επίσης ενσωματώνει τη στρατηγική "αντικαταστήστε τον ηττημένο", του ADWIN Bagging. Όταν εντοπιστεί μια αλλαγή, αφαιρείται ο χειρότερος κατηγοριοποιητής και προστίθεται ένας νέος κατηγοριοποιητής. Υλοποιείται στο MOA ως LEVERAGINGBAG [17].

4.4.12 Bagging using ASHTs of different sizes (OzaBagASHT)

Η μέθοδος adaptive-size hoeffding tree (ASHT) [24], που εφαρμόζεται στο MOA ως OzaBagASHT, δημιουργεί ένα σύνολο hoeffding tree, διαφορετικού μεγέθους, με τους ακόλουθους περιορισμούς:



Εικόνα 15: Bagging με χρήση ενός σύνολου Hoeffding Tree διαφορετικού μεγέθους

- Κάθε δέντρο έχει μια σχετική τιμή που αντιπροσωπεύει τον μέγιστο αριθμό εσωτερικών κόμβων που μπορεί να δημιουργηθούν σε όλο το δέντρο.
- Όταν το μέγεθος του δέντρου υπερβαίνει τη μέγιστη τιμή μεγέθους, υπάρχουν δύο διαφορετικές επιλογές διαγραφής: α) Διαγράψτε τον παλαιότερο κόμβο, τη ρίζα και όλα τα παιδιά του εκτός από εκείνον όπου έγινε η διαίρεση. Μετά από αυτό, η ρίζα του παιδιού που δεν έχει διαγραφεί γίνεται η νέα ρίζα. Β) Διαγράψτε όλους τους κόμβους του δέντρου, δηλαδή επαναφέρετε το δέντρο στο κενό δέντρο.
- Μπορεί να προσαρμοστεί σε μια εννοιολογική απόκλιση.

Εάν υπάρχουν N δέντρα στο σύνολο, ένας απλός τρόπος για να εκχωρήσετε τα μέγιστα μεγέθη είναι να επιλέξετε δυνάμεις, για παράδειγμα, του 2: 2, 4, ... έως 2^k. Η διαίσθηση πίσω από αυτή τη μέθοδο είναι ότι τα μικρότερα δέντρα προσαρμόζονται πιο γρήγορα στις αλλαγές και τα μεγαλύτερα δέντρα αποδίδουν καλύτερα σε περιόδους με ελάχιστες ή καθόλου αλλαγές, απλώς και μόνο επειδή κατασκευάστηκαν χρησιμοποιώντας περισσότερα δεδομένα. Τα δέντρα που περιορίζονται σε μέγεθος s θα μηδενίζονται περίπου δύο φορές πιο συχνά από τα δέντρα που περιορίζονται σε μέγεθος 2s. Αυτό δημιουργεί ένα σύνολο διαφορετικών ταχυτήτων επαναφοράς, για μια μέθοδο συνόλου τέτοιων δέντρων και επομένως ένα υποσύνολο δέντρων που είναι μια καλή προσέγγιση για τον τρέχοντα ρυθμό αλλαγής.

Η δημιουργία του συνόλου μπορεί να βασίζεται στη πλειοψηφία μεταξύ όλων των δέντρων, αν και έχει διαπιστωθεί πειραματικά ότι είναι καλύτερο να τα σταθμίσουμε αντιστρόφως ανάλογα με το τετραγωνικό τους σφάλμα. Το σφάλμα κάθε δέντρου παρακολουθείται από έναν εκτιμητή EWMA. Είναι σημαντικό να σημειωθεί ότι οι επαναρυθμίσεις θα συμβαίνουν συνεχώς, ακόμη και για σταθερά σύνολα δεδομένων, αλλά αυτή η συμπεριφορά δεν πρέπει να βλάπτει την προγνωστική απόδοση του συνόλου[24].

4.4.13 Online Boosting (OzaBoost)

Ο Schapire (1990) πρότεινε αρχικά μια γενική μέθοδο για τη μετατροπή ενός αδύναμου μαθητή σε έναν που επιτυγχάνει, αυθαίρετα, υψηλή ακρίβεια. Ο αλγόριθμος που αναπτύχθηκε αρχικά βασίστηκε σε ένα θεωρητικό μοντέλο που είναι γνωστό ως μοντέλο ασθενούς εκπαίδευσης. Αυτό το μοντέλο υποθέτει ότι υπάρχουν αδύναμοι αλγόριθμοι

εκπαίδευσης που μπορούν να κάνουν ελαφρώς καλύτερα μια πρόβλεψη-εικασία από μια τυχαία, ανεξάρτητα από την υποκείμενη κατανομή πιθανότητας D που χρησιμοποιείται για τη δημιουργία των παραδειγμάτων. Το έργο του Schapire δείχνει πώς να ενισχύονται αυτοί οι αδύναμοι μαθητές ώστε να επιτύχουν αυθαίρετα υψηλή ακρίβεια[14].

Στη μέθοδο boosting, όπως και στη μέθοδο bagging, οι αλγόριθμοι συνδυάζουν πολλαπλά βασικά μοντέλα τα οποία εκπαιδεύονται με τα δείγματα της εισόδου για να επιτύχουν ένα χαμηλότερο σφάλμα κατηγοριοποίησης. Αλλά σε αντίθεση με τη μέθοδο bagging, τα μοντέλα δημιουργούνται διαδοχικά και όχι παράλληλα, με την κατασκευή κάθε νέου μοντέλου να εξαρτάται από την απόδοση των μοντέλων που κατασκευάστηκαν προηγουμένως. Η διαισθητική ιδέα της μεθόδου boosting είναι να δοθεί μεγαλύτερη βαρύτητα στα παραδείγματα που κατηγοριοποιούνται λανθασμένα από το τρέχον σύνολο κατηγοριοποιητών, έτσι ώστε ο επόμενος κατηγοριοποιητής στην ακολουθία να δίνει μεγαλύτερη προσοχή σε αυτά τα παραδείγματα [10].

Αυτή η διαδοχική φύση καθιστά την εφαρμογή της μεθόδου boosting πιο δύσκολη από την εφαρμογή της μεθόδου bagging σε ένα περιβάλλον ροής. Για διαδικτυακές ρυθμίσεις, οι Oza και Russell [33] πρότειναν το Online Boosting, μια διαδικτυακή μέθοδο που αντί να δημιουργεί νέα μοντέλα κάθε φορά που φτάνει ένα νέο παράδειγμα, ενημερώνει κάθε μοντέλο με ένα βάρος που υπολογίζεται ανάλογα με την απόδοση των προηγούμενων κατηγοριοποιητών. Στην πραγματικότητα στη μέθοδο αυτή διαιρείται το συνολικό βάρος του κάθε παραδείγματος σε δύο μισά. Το μισό του βάρους αποδίδεται στα σωστά κατηγοριοποιημένα παραδείγματα, και το άλλο μισό στα λανθασμένα παραδείγματα. Οι συγγραφείς χρησιμοποιούν την κατανομή Poisson για να αποφασιστεί η τυχαία πιθανότητα να χρησιμοποιηθεί ένα παράδειγμα για εκπαίδευση, μόνο που αυτή τη φορά η παράμετρος αλλάζει ανάλογα με το boosting βάρος του παραδείγματος, καθώς περνά από κάθε μοντέλο με τη σειρά. Υλοποιείται στο MOA ως OzaBoost [33] .

4.4.14 OCBoost -Online Coordinate Boosting

Οι Pelossof et al. πρότειναν την μέθοδο Online Coordinate Boosting, έναν νέο διαδικτυακό αλγόριθμο boosting για την προσαρμογή των βαρών ενός booster κατηγοριοποιητή, ο οποίος προσεγγίζει περισσότερο τον αλγόριθμο AdaBoost των Freund και Schapire [45]. Η διαδικασία ενημέρωσης των βαρών προκύπτει ελαχιστοποιώντας την απώλεια του AdaBoost όταν εμφανίζεται μια βαθμιαία σταθμισμένη εννοιολογική απόκλιση. Αυτή η μέθοδος ενίσχυσης μπορεί να οδηγήσει σε μια μορφή παρόμοια με τον αλγόριθμο των Oza και Russell [41].

Κεφάλαιο 5 – Το πειραματικό περιβάλλον του MOA

5.1 Εισαγωγή στο περιβάλλον του MOA

Το MOA (Massive Online Analysis) είναι από τα δημοφιλέστερα περιβάλλοντα ανοιχτού κώδικα για εξόρυξη ροών δεδομένων. Περιέχει ένα μεγάλο αριθμό αλγορίθμων μηχανικής μάθησης για την κατηγοριοποίηση, την παλινδρόμηση, την συσταδοποίηση , ανίχνευση ακραίων τιμών, ανίχνευση αλλαγών και συστήματα συστάσεων και επίσης πολλά εργαλεία αξιολόγησης. Ένα άλλο μεγάλο πλεονέκτημα του είναι ότι περιέχει ένα ικανοποιητικό αριθμό συνθετικών δεδομένων για πειραματισμό.

Το MOA είναι επέκταση του WEKA, μπορεί να χρησιμοποιηθεί μέσα από το περιβάλλον του WEKA ή και μόνο του όπως θα παρουσιαστεί παρακάτω, μπορεί να δανειστεί επίσης τους αλγόριθμους του WEKA. Είναι γραμμένο σε java, μπορεί να χρησιμοποιηθεί (ως αυτόνομο) με τρείς τρόπους μέσω του API της java, μέσω της γραμμής εντολών (cmd) και μέσω του γραφικού της περιβάλλοντος, όπως θα δούμε στην παρούσα εργασία.

- Βήμα 1 Λήψη του πλαισίου του MOA
- Βήμα 2 Φόρτωση του MOA.
- Βήμα 3 Ρύθμιση της εργασίας.
- Βήμα 4 Έξοδος αποτελεσμάτων, τα οποία μπορούν να αποθηκευτούν και να επεξεργαστούν.

Βήμα 1

Το MOA είναι διαθέσιμο στην διεύθυνση https://moa.cms.waikato.ac.nz , λήψη αρχείου, είναι ένα zip αρχείο, χρειάζεται αποσυμπίεση.

Βήμα 2

Ο πιο εύκολος τρόπος φόρτωσης είναι με διπλό κλικ στο αρχείο moa.bat για windows ή στο αρχείο moa.sh για Mac ή Linux. Τα αρχεία αυτά περιέχονται στο φάκελο bin. Εναλλακτικά μπορεί να χρησιμοποιηθεί η ακόλουθη εντολή.

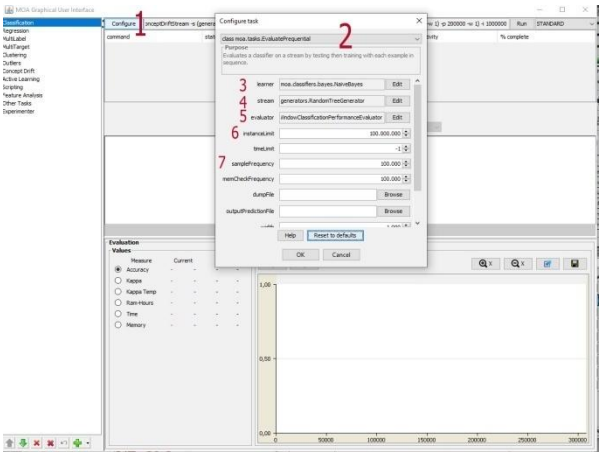
Εντολή φόρτωσης του MOA

java –Xmx16G -cp moa.jar -javaagent:sizeofag-1.0.4.jar moa.gui.GUI

Αυτή η εντολή καλεί το -cp moa.jar, java agent. Έχει δυο πλεονεκτήματα, επιτρέπει να οριστεί το μέγεθος της Head memory και την λήψη των μετρήσεων του κόστους RAM/hours.

Βήμα 3 - Ρύθμιση Εργασίας

Η ρύθμιση της εργασίας περιέχει 7 βασικές επιλογές και πολλές προαιρετικές. Παρακάτω παρουσιάζονται οι 7 βασικές επιλογές.



Εικόνα 16: Το γραφικό περιβάλλον του MOA

1) **Επιλογή ρύθμισης** (configure) (εικόνα 16 (1)) ανοίγει η καρτέλα configure Task εδώ επιλέγετε η κύρια εργασία, που θέλει να εκτελέσει ο χρήστης (εικόνα 16 (2)).
2) **Επιλογή της Κύριας Εργασίας στο MOA**. Στο πλαίσιο της κατηγοριοποίησης είναι η επιλογή της διαδικασίας αξιολόγησης του αλγορίθμου εκπαίδευσης. Η επιλογή αυτή καθορίζει ποια παραδείγματα θα χρησιμοποιηθούν για την εκπαίδευση του αλγορίθμου και ποια θα χρησιμοποιηθούν για τη δοκιμή του.
    - **Holdout**: Η μέθοδος εκτίμα την απόδοση σε ένα τμήμα (holdout) της ροής. Στο MOA, η υλοποίηση αυτής της **Holdout** απαιτεί από τον χρήστη να καθορίσει δύο παραμέτρους το μέγεθος του πρώτου παραθύρου, δηλαδή το σύνολο παραδειγμάτων που χρησιμοποιούνται για την δοκιμή και την συχνότητα των δοκιμών .
    - **EvaluateInterleavedTestThenTrain:** Κάθε παράδειγμα χρησιμοποιείται πρώτα για τη δοκιμή του μοντέλου και μετά για εκπαίδευση του, έπειτα η ακρίβεια μπορεί να ενημερωθεί σταδιακά. Έτσι, το μοντέλο δοκιμάζεται πάντα σε παραδείγματα που δεν έχει δει. Στο MOA, αυτό το σχήμα υλοποιείται χρησιμοποιώντας ένα μοντέλο παραθύρου ορόσημο (τα δεδομένα στη ροή λαμβάνονται υπόψη από την αρχή έως τώρα).

- **EvaluatePrequential:** Όπως το Interleaved test-then-train, αλλά, στο MOA, υλοποιεί την ιδέα ότι τα πιο πρόσφατα παραδείγματα είναι πιο σημαντικά, χρησιμοποιώντας ένα συρόμενο παράθυρο (sliding windows) ή έναν παράγοντα αποσύνθεσης. Τα μεγέθη του συρόμενου παραθύρου και ο συντελεστής αποσύνθεσης είναι παράμετροι**.** Μια επιπλέον επιλογή είναι η δυνατότητα χρήσης ενός προσαρμοστικού παραθύρου**.**

3) **Επιλογή Αλγόριθμου Εκπαίδευσης** (εικόνα 16(3)). Υπάρχει η δυνατότητα επιλογής ενός πλήθους αλγορίθμων κατηγοριοποίησης, έχουν αναλυθεί αρκετοί από αυτούς στην ενότητα 4.

4) **Επιλογή Ροής Δεδομένων** (εικόνα 16(4)). Το MOA έχει ενσωματωμένες πολλές ροές δεδομένων από συνθετικά δεδομένα, δίνει και την δυνατότητα χρήσης ροών από ARFF αρχεία.

   Επίσης δίνει την δυνατότητα ενσωμάτωσης της εννοιολογικής απόκλισης σε μια ροή δεδομένων ενώνοντας 2 ροές με παρόμοια χαρακτηριστικά μέσω της επιλογής ConceptDriftStream.

   - -s : Αρχική γεννήτρια
   - -d : Γεννήτρια μετά τη δημιουργία της εννοιολογικής απόκλισης
   - -p : Κεντρική θέση της εννοιολογικής απόκλισης
   - -w : Πλάτος της εννοιολογικής απόκλισης. Θέτοντας w=1 προσομοιώνεται μια απότομη εννοιολογική απόκλιση.

Εννοιολογικές αποκλίσεις μπορούν να δημιουργηθούν και με διαφορετικό τρόπο μέσω των συνθετικών δεδομένων Random RBF Generator ή HYPERPLANE μια περιγραφή τους ακολουθεί στην επόμενη ενότητα.

5) **Επιλογή Evaluator** (εικόνα 16(5))

Καθορίζει ποια παραδείγματα θα χρησιμοποιήσει ο αλγόριθμος εκπαίδευσης για να είναι συνεπής με την τρέχουσα φύση της ροής.

- **BasicClassificationPerformanceEvaluator**: Περιλαμβάνει παραδείγματα από όλη την ιστορία της ροής, εκτελεί σταδιακή αξιολόγηση. Χρησιμοποιείται σε συνδυασμό με την μέθοδο Interleaved test-then-train.
- **EWMAClassificationPerformanceEvaluator:** Αξιολογητής που ενημερώνει τα αποτελέσματα αξιολόγησης χρησιμοποιώντας έναν εκθετικό σταθμισμένο κινητό μέσο όρο. Χρησιμοποιείται με την μέθοδο Prequential, προσθέτοντας ένας μηχανισμό στάθμισης στα δεδομένα της ροής, δίνοντας ένα μεγαλύτερο βάρος στα πρόσφατα παραδείγματα, είναι κατάλληλο για ροές με εννοιολογικές αποκλίσεις.

- **AdwinClassificationPerfornaceEvaluation:** Αξιολογητής που ενημερώνει τα αποτελέσματα χρησιμοποιώντας ένα προσαρμοστικό παράθυρο μεταβλητού μεγέθους. Είναι κατάλληλος αξιολογητής για ροές με εννοιολογικές αποκλίσεις, χρησιμοποιείται με την μέθοδο Prequential.
- **WindowsClassificationPerfornaceEvaluation:** Αξιολογητής ενημερώνει τα αποτελέσματα χρησιμοποιώντας ένα συρόμενο παράθυρο (sliding window). Το μέγεθος του παραθύρου ορίζεται από τον χρήστη, εξ ορισμού είναι w=1000. Χρησιμοποιείται με την μέθοδο Prequential.

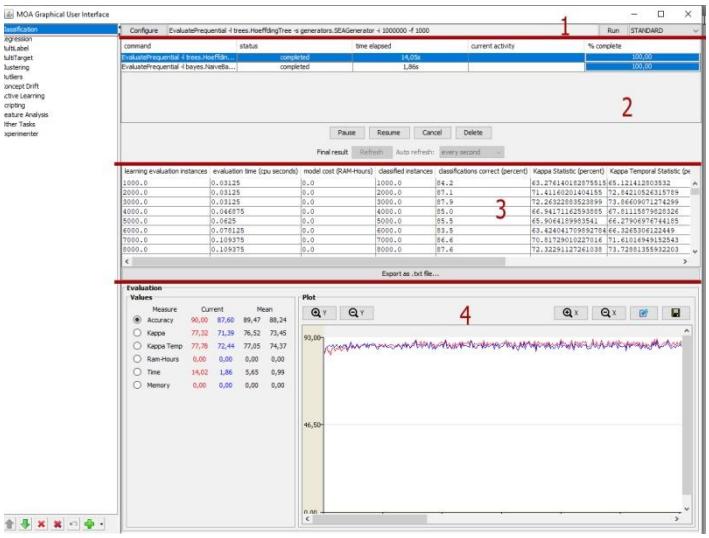
6) **Επιλογή μέγιστου αριθμών παραδειγμάτων** (εικόνα 16(6)). Καθορίζει το μέγιστο αριθμό παραδειγμάτων που θα χρησιμοποιηθούν για δοκιμή και εκπαίδευση
7) **Καθορισμός αριθμού παραδειγμάτων**(εικόνα 16(7)). παρουσίασης των μετρήσεων. Ανά πόσα παραδείγματα θα λαμβάνονται οι μετρήσεις. Έπειτα οκ και run για να τρέξει η εργασία. Το MOA έχει την δυνατότητα να εκτελεί πολλές εργασίες ταυτόχρονα.

Βήμα 4 έξοδος αποτελεσμάτων

Έστω ότι εκτελέσαμε τις εξής εργασίες:

- EvaluatePrequential -l bayes.NaiveBayes -s generators.SEAGenerator -i 1000000 -f 1000
- EvaluatePrequential -l trees.HoeffdingTree -s generators.SEAGenerator -i 1000000 -f 1000

Οι εντολές στη διεύθυνση γραμμής εργασιών, configure, μπορούν να αντιγραφούν και αποθηκευτούν και να επικολληθούν.



Εικόνα 17: Το γραφικό περιβάλλον του MOA

Μετά την εκτέλεση των εργασιών, το MOA έχει την μορφή της εικόνας 17. Θα μπορούσε να θεωρηθεί ότι το περιβάλλον του χωρίζεται σε 4 μέρη. Το μέρος 1 έχει ήδη συζητηθεί,

αφορά την ρυθμίσεις των εργασιών. Το μέρος 2 περιλαμβάνει μια λίστα των εργασιών, μπορούν να εκτελεστούν ταυτόχρονα πολλές εργασίες και να έχει πρόσβαση ο χρήστης, ανά πάσα στιγμή σε αυτές. Το μέρος 3 περιλαμβάνει ένα csv αρχείο με τις μετρήσεις του κάθε πειράματος. Το οποίο μπορεί να αποθηκευτεί και να επεξεργαστεί. Το μέρος 4 δίνει τις κυριότερες μετρήσεις της κάθε εργασίας που εκτελέστηκαν και μια γραφική απεικόνιση. Επιπλέον δίνει τη δυνατότητα σύγκρισης δύο εργασιών, με κόκκινο είναι η τρέχων εργασία και με μπλε η προηγούμενη[10][38][12].

5.2 Συνθετικά Δεδομένα

Τα συνθετικά δεδομένα έχουν πολλά πλεονεκτήματα, είναι ευκολότερο να αναπαραχθούν και υπάρχει μικρό κόστος όσον αφορά την αποθήκευση και τη μετάδοση τους. Στο MOA, τα δεδομένα έχουν συλλεχθεί από τις γεννήτριες που βρίσκονται συχνότερα στη βιβλιογραφία.

SEA Concepts Generator: Δημιουργείται χρησιμοποιώντας τρία χαρακτηριστικά, όπου μόνο τα δύο πρώτα είναι σχετικά. Και τα τρία χαρακτηριστικά παίρνουν τιμές μεταξύ 0 και 10. Τα σημεία του συνόλου δεδομένων χωρίζονται σε 4 μπλοκ με διαφορετικές έννοιες. Σε κάθε μπλοκ, η κατηγοριοποίηση γίνεται χρησιμοποιώντας $f1 + f2 \leq \theta$, όπου τα f1 και f2 αντιπροσωπεύουν τα δύο πρώτα χαρακτηριστικά και το θ είναι μια τιμή κατωφλίου. Οι πιο συχνές τιμές είναι 9, 8, 7 και 9,5 για τα μπλοκ δεδομένων.

STAGGER Concepts Generator (with abrupt concept drift): Δημιουργείται χρησιμοποιώντας τρία ονομαστικά χαρακτηριστικά: size = {small, medium, large}, color= {red, green} και shape={circular, non-circular}. Πριν από το πρώτο σημείο εμφάνισης μιας εννοιολογικής απόκλισης, οι περιπτώσεις χαρακτηρίζονται θετικές εάν (color = red) $\wedge$ (size = small). Μετά από αυτό το σημείο και πριν από τη δεύτερη εννοιολογική απόκλιση, οι περιπτώσεις κατηγοριοποιούνται θετικές εάν (color= green) v (shape = circular), και τέλος μετά από αυτό το δεύτερο σημείο μετατόπισης, οι περιπτώσεις κατηγοριοποιούνται θετικές μόνο εάν (size =medium) v (size = large).

LED (with gradual concept drift): Ο στόχος του συνόλου δεδομένων είναι να προβλέψει το ψηφίο που εμφανίζεται σε μια οθόνη LED, 7 τμημάτων, που το κάθε ψηφίο έχει 10% πιθανότητα να εμφανιστεί. Το σύνολο δεδομένων έχει 7 χαρακτηριστικά που σχετίζονται με την κλάση και 17 άσχετα. Μια εννοιολογική απόκλιση προσομοιώνεται με την ανταλλαγή σχετικών χαρακτηριστικών.

Waveform Generator: Αυτή η γεννήτρια μοιράζεται την προέλευσή της με το LED και δωρήθηκε επίσης από τον David Aha στο αποθετήριο UCI. Ο στόχος της εργασίας είναι η

διαφοροποίηση μεταξύ τριών διαφορετικών κατηγοριών κυματομορφής, καθεμία από τις οποίες δημιουργείται από συνδυασμό δύο ή τριών κυμάτων βάσης. Το βέλτιστο ποσοστό κατηγοριοποίησης Bayes είναι γνωστό ότι είναι 86%.

Υπάρχουν δύο εκδοχές του προβλήματος. Το WAVE21 έχει 21 αριθμητικά χαρακτηριστικά, τα οποία περιλαμβάνουν όλα θόρυβο. Το WAVE40 εισάγει επιπλέον 19 άσχετα χαρακτηριστικά[38].

Sine generator: Sine1 (με απότομη εννοιολογική απόκλιση): Αποτελείται από δύο χαρακτηριστικά x και y ομοιόμορφα κατανεμημένα στο [0,1]. Η συνάρτηση κατηγοριοποίησης είναι y = sin(x). Οι περιπτώσεις κατηγοριοποιούνται ως θετικές εάν βρίσκονται κάτω από την καμπύλη. διαφορετικά κατηγοριοποιούνται ως αρνητικά. Σε ένα σημείο εμφάνισης εννοιολογικής απόκλισης, οι ετικέτες κλάσεων αντιστρέφονται.

Sine2 (με απότομη εννοιολογική απόκλιση): Διαθέτει δύο χαρακτηριστικά των x και y που είναι ομοιόμορφα κατανεμημένα στο [0, 1]. Η συνάρτηση κατηγοριοποίησης είναι 0,5 + 0,3 * sin(3 * π * x). Οι περιπτώσεις κάτω από την καμπύλη κατηγοριοποιούνται ως θετικές ενώ οι άλλες περιπτώσεις κατηγοριοποιούνται ως αρνητικές. Σε ένα σημείο εμφάνισης εννοιολογικής απόκλισης, το σχήμα κατηγοριοποίησης αντιστρέφεται[53].

Random RBF Generator

Η γεννήτρια RBF (Radial Basis Function) λειτουργεί ως εξής: Δημιουργείται ένας σταθερός αριθμός τυχαίων κεντροειδών. Κάθε κέντρο έχει μια τυχαία θέση, με μια ενιαία τυπική απόκλιση, ετικέτα κλάσης και βάρος. Δημιουργούνται νέα παραδείγματα με την επιλογή ενός κέντρου τυχαία, λαμβάνοντας υπόψη τα βάρη, έτσι ώστε να είναι πιο πιθανό να επιλεγούν κέντρα με μεγαλύτερο βάρος. Επιλέγεται μια τυχαία κατεύθυνση για να αντισταθμίσει τις τιμές των χαρακτηριστικών από το κεντρικό σημείο. Το μήκος της εννοιολογικής απόκλισης λαμβάνεται τυχαία από μια κατανομή Gauss με τυπική απόκλιση που προσδιορίζεται από το επιλεγμένο κέντρο. Το επιλεγμένο κέντρο καθορίζει επίσης την ετικέτα κλάσης του παραδείγματος. Αυτό δημιουργεί αποτελεσματικά μια κανονικά κατανεμημένη υπερσφαίρα παραδειγμάτων που περιβάλλουν κάθε κεντρικό σημείο με ποικίλες πυκνότητες. Δημιουργούνται μόνο αριθμητικά χαρακτηριστικά. Το RRBFS αναφέρεται σε ένα απλό τυχαίο σύνολο δεδομένων RBF—100 κέντρα και δέκα χαρακτηριστικά. Το RRBFC είναι πιο σύνθετο—1000 κέντρα και 50 χαρακτηριστικά. Και τα δύο είναι προβλήματα δύο κλάσεων[38].

$$\sum_{i=1}^{d} w_i x_i \ = w_0 = \sum_{i=1}^{d} w_i \quad (23)$$

5.3 Δεδομένα Πραγματικού κόσμου

Σύνολο δεδομένων Poker-Hand: Αποτελείται από 1.000.000 περιπτώσεις και 11 χαρακτηριστικά. Κάθε εγγραφή του συνόλου δεδομένων Poker-Hand είναι ένα παράδειγμα ενός χεριού που αποτελείται από πέντε τραπουλόχαρτα που έχουν τραβηχτεί από μια τυπική τράπουλα των 52 χαρτιών. Κάθε φύλλο περιγράφεται χρησιμοποιώντας δύο χαρακτηριστικά (κοστούμι και κατάταξη), για συνολικά 10 προγνωστικά χαρακτηριστικά. Υπάρχει ένα χαρακτηριστικό Class που περιγράφει το "Πόκερ Χέρι". Η σειρά των φύλλων είναι σημαντική, γι' αυτό υπάρχουν 480 πιθανά χέρια Royal Flush αντί για 4.

Δεδομένα ELEC: Είναι ένα άλλο ευρέως χρησιμοποιούμενο σύνολο δεδομένων που περιγράφεται από τον M. Harries και αναλύεται από τον Gama. Αυτά τα δεδομένα συλλέχθηκαν από την αυστραλιανή αγορά ηλεκτρικής ενέργειας της Νέας Νότιας Ουαλίας. Στην αγορά αυτή, οι τιμές δεν είναι σταθερές και επηρεάζονται από τη ζήτηση και την προσφορά της αγοράς. Ρυθμίζονται κάθε πέντε λεπτά. Το σύνολο δεδομένων ELEC περιέχει 45.312 περιπτώσεις. Η ετικέτα κατηγορίας προσδιορίζει τη μεταβολή της τιμής σε σχέση με τον κινητό μέσο όρο των τελευταίων 24 ωρών [10][38][12].

Κεφάλαιο 6 - Πειραματική μελέτη σε συνθετικά δεδομένα

6.1 Εγκαθίδρυση Πειραμάτων

Για να μπορεί να γίνει η συγκριτική μελέτη των αποτελεσμάτων όλα τα πειράματα έγιναν στον ίδιο υπολογιστή: Zbook. Επεξεργαστής: Intel(R) Xeon(R) CPU E3-1505M v5 @ 2.80GHz 2.81 GHz. Μνήμη RAM: 16,0 GB. Λειτουργικό σύστημα windows 10 64 bit

6.1.1 Προσομοίωση Απότομων και Σταδιακών Εννοιολογικών Αποκλίσεων

Στο παρόν πείραμα χρησιμοποιήθηκαν τα συνθετικά δεδομένα των γεννητριών SEA και Agrawal για να γίνει προσομοίωση:

Μια ροής, από τα συνθετικά δεδομένα SEA, με 3 απότομες εννοιολογικές αποκλίσεις με πλάτος 1 έτσι ώστε δυο έννοιες να αλλάζουν ακριβώς στο επόμενο βήμα.

Ενδεικτική εντολή (αλγόριθμος NB) MOA

EvaluatePrequential -l bayes.NaiveBayes -s (ConceptDriftStream -s generators.SineGenerator -d (ConceptDriftStream -s (generators.SineGenerator -f 2) -d (ConceptDriftStream -s generators.SineGenerator -d (generators.SineGenerator -f 2) -p 50000 -w 1) -p 50000 -w 1) -p 50000 -w 1 -r 5) -e (WindowClassificationPerformanceEvaluator -o -p -r -f) -i 200000 -f 1000

1) Μια ροής, από τα συνθετικά δεδομένα SEA, με 3 σταδιακές εννοιολογικές αποκλίσεις, έτσι ώστε το μέγεθος κάθε έννοιας να είναι διαφορετικό το ένα από το άλλο.

Ενδεικτική εντολή (αλγόριθμος HAT) MOA

EvaluatePrequential -l trees.HoeffdingAdaptiveTree -s (ConceptDriftStream -s generators.SEAGenerator -d (ConceptDriftStream -s (generators.SEAGenerator -f 2) -d (ConceptDriftStream -s (generators.SEAGenerator -f 3) -d (generators.SEAGenerator -f 4) -p 50000 -w 10000) -p 50000 -w 8000) -p 50000 -w 6000 -r 5) -e (WindowClassificationPerformanceEvaluator -o -p -r -f) -i 200000 -f 1000

2) Μια ροής, από τα συνθετικά δεδομένα Agrawal, με 3 απότομες εννοιολογικές αποκλίσεις με πλάτος 1 έτσι ώστε δυο έννοιες να αλλάζουν ακριβώς στο επόμενο βήμα.

Ενδεικτική εντολή (αλγόριθμος NB-EDDM) MOA

EvaluatePrequential -l (drift.SingleClassifierDrift -d EDDM) -s (ConceptDriftStream -s (generators.AgrawalGenerator -f 5) -d (ConceptDriftStream -s (generators.AgrawalGenerator -f 3) -d (ConceptDriftStream -s (generators.AgrawalGenerator -f 4) -d (generators.AgrawalGenerator -f 6) -p 50000 -w 1) -p 50000 -w 1) -p 50000 -w 1 -r 5) -e (WindowClassificationPerformanceEvaluator -o -p -r -f) -i 200000 -f 1000

3) Μια ροής, από τα συνθετικά δεδομένα Agrawal, με 3 σταδιακές εννοιολογικές αποκλίσεις, έτσι ώστε το μέγεθος κάθε έννοιας να είναι διαφορετικό το ένα από το άλλο.

Ενδεικτική εντολή (αλγόριθμος AUE) MOA

EvaluatePrequential -l meta.AccuracyUpdatedEnsemble -s (ConceptDriftStream -s (generators.AgrawalGenerator -f 5) -d (ConceptDriftStream -s (generators.AgrawalGenerator -f 3) -d (ConceptDriftStream -s (generators.AgrawalGenerator -f 4) -d (generators.AgrawalGenerator -f 6) -p 50000 -w 10000) -p 50000 -w 8000) -p 50000 -w 6000 -r 5) -e (WindowClassificationPerformanceEvaluator -o -p -r -f) -i 200000 -f 1000

Περισσότερες πληροφορίες, στον πίνακα 3, για τη θέση της κάθε εννοιολογικής απόκλισης και το μέγεθος της.

Ροή	Αριθμός drift	Θέση	πλάτος	θέση	πλάτος	θέση	πλάτος
SEA (απότομη)	3	50000	1	100000	1	150000	1
SEA (σταδιακή)	3	50000	6000	100000	8000	150000	10000
Agrawal(απότομη)	3	50000	1	100000	1	150000	1
Agrawal(σταδιακή)	3	50000	6000	100000	8000	150000	10000

Πίνακας 3 Πληροφορίες εννοιολογικών αποκλίσεων ροών.

Τα δεδομένα όλων των ροών ανέρχονται σε 200.000 παραδείγματα. Η αξιολόγηση τους γίνεται με χρήση της μεθόδου EvaluatePrequential πάνω σε ένα συρόμενο παράθυρο (sliding windows) πλάτους w=1000. Οι μετρήσεις ορίζεται να επιστρέφονται κάθε 1000 παραδείγματα.

Στο πρώτο πείραμα των πινάκων 4 και 5 ο βασικός αλγόριθμος κατηγοριοποίησης είναι ο Naive bayes που σε συνδυασμό με συρόμενο παράθυρο είναι μια τυφλή μέθοδος προσαρμογής. Στα υπόλοιπα πειράματα γίνεται χρήση των ενημερωμένων μεθόδων προσαρμογής. Επιλέγετε ο κατηγοριοποιητής SingleClassifierDrift που χρησιμοποιεί έναν

ανιχνευτή αλλαγών (ενότητα 3.7.2). Ως βασικός αλγόριθμος κατηγοριοποίησης επιλέγεται ο Naive bayes.

Στο πρώτο πείραμα, των πινάκων 5 και 6, ο βασικός αλγόριθμος κατηγοριοποίησης είναι ο HT στο δεύτερο ο HAT και στα υπόλοιπα πειράματα οι αλγόριθμοι συνδυαστικής μάθησης (ensembles) της ενότητας 4.2 .

6.1.2 Πειραματικά αποτελέσματα

SEA - ΑΠΟΤΕΛΕΣΜΑΤΑ NB ΚΑΙ NB ΜΕ ΑΝΙΧΝΕΥΤΗ ΑΛΛΑΓΩΝ								
ΜΟΝΤ. ΜΑΘΗΣΗΣ	**ACCURACY**		**KAPPA**		**KAPPA TEMP**		**RAM HOURS**	
Εννοιολογική Απόκλιση	Απότομη Current/Mean	Σταδιακή Current/Mean	Απότομη Current/Mean	Σταδιακή Current/Mean	Απότομη Current/Mean	Σταδιακή Current/Mean	Απότομη	Σταδιακή
NB	81,80/83,80	80,80/83,75	62,38/64,05	60,37/63,57	62,70/65,67	60,57/65,24	1,499	1,183
NB-DDM	84,20/86,43	82,10/84,61	67,58/70,13	63,07/66,59	67,62/70,94	63,39/66,89	2,073	2,073
NB-EDDM	8(8,50/87,1	**88,60/**87,18	76,68/71,58	76,89/71,74	76,43/72,45	**76,69/**72,47	1,668	**1,499**
NB-RDDM	**87,60/88,33**	87,70/**87,33**	74,79/**73,91**	75,03/**71,84**	74,59/74,82	74,85/72,82	**8,384**	**7,540**
NB-HDDM_A	**88,60/**88,25	87,60/**87,34**	**76,89/**73,73	74,80/**71,96**	**76,64/**74,65	74,64/72,87	2,538	2,191
NB-HDDM_w	**88,60/88,33**	87,70/86,96	**76,89/**73,89	75,00/71,11	**76,64/**74,81	74,85/72,10	2,474	2,477
NB-SEED	**88,60/**88,26	**88,60/**87,18	**76,89/**73,77	**76,89/**71,44	**76,64/**74,68	**76,69/**72,47	2,787	3,930
NB-ADWIN	**88,60/**88,31	87,60/87,27	**76,89/**73,87	74,80/71,80	**76,64/**74,78	74,64/72,72	3,741	2,995
NB-STEPD	85,90/87,71	85,90/87,10	71,35/72,48	71,37/71,28	71,11/73,47	71,17/72,27	1,251	2,048
NB-CUSUM	**88,60/88,33**	**88,60/87,76**	**76,89/73,91**	**76,89/72,86**	**76,64/74,82**	**76,69/73,71**	2,039	2,039
NB-PH	**88,60/**88,20	86,10/87,53	**76,89/**73,64	75,43/72,37	**76,64/**74,55	75,26/73,25	2,199	2,431

Πίνακας 4 Αποτελέσματα απότομων και σταδιακών εννοιολογικών αποκλίσεων – SEA

AGRAWAL - ΑΠΟΤΕΛΕΣΜΑΤΑ NB ΚΑΙ NB ΜΕ ΑΝΙΧΝΕΥΤΗ ΑΛΛΑΓΩΝ								
ΜΟΝΤ. ΜΑΘΗΣΗΣ	**ACCURACY**		**KAPPA**		**KAPPA TEMP**		**RAM HOURS**	
Εννοιολογική Απόκλιση	Απότομη Current/Mean	Σταδιακή Current/Mean	Απότομη Current/Mean	Σταδιακή Current/Mean	Απότομη Current/Mean	Σταδιακή Current/Mean	Απότομη	Σταδιακή
NB	62,60/60,75	62,60/60,75	19,03/18,03	19,01/18,05	20,93/18,72	21,10/18,85	7,065	7,065
NB-DDM	**73,50/67,55**	73,40/67,11	43,53/32,07	43,34/31,18	43,97/32,78	43,88/31,94	3,772	3,897
NB-EDDM	73,30/64,65	**73,60/64,49**	43,22/26,48	43,93/26,00	43,55/26,79	44,30/26,62	2,881	**3,073**
NB-RDDM	72,60/68,15	73,20/67,15	41,36/33,12	42,92/31,34	42,07/34,04	43,46/32,04	**1,468**	**1,468**
NB-HDDM_A	**73,50/68,12**	**73,60/67,26**	43,59/33,07	43,88/31,66	43,97/33,96	44,30/32,25	4,225	4,227
NB-HDDM_w	72,60/67,50	72,30/66,31	39,98/31,73	40,40/29,49	41,01/32,70	41,56/30,33	4,642	4,780
NB-SEED	73,10/67,86	**73,70**/66,60	42,74/32,61	44,12/30,33	43,13/33,45	44,51/31,06	7,711	8,218
NB-ADWIN	73,40/68,14	73,60/67,04	43,35/33,13	43,82/31,20	43,76/34,02	44,30/31,80	6,454	6,447
NB-STEPD	69,80/67,08	70,00/66,24	34,62/30,78	35,04/29,18	36,15/31,83	36,71/30,17	4,168	4,170
NB-CUSUM	73,30/68,04	73,40/67,21	43,11/32,94	43,34/31,39	43,55/33,81	43,88/32,15	3,719	3,839
NB-PH	**73,60/68,09**	**73,60/67,17**	43,77/33,05	43,77/31,28	44,19/33,91	44,30/32,05	4,236	4,236

Πίνακας 5 Αποτελέσματα απότομων και σταδιακών εννοιολογικών αποκλίσεων – AGRAWAL

SEA ΑΠΟΤΕΛΕΣΜΑΤΑ HT / HAT/ ΣΥΝΟΛΑ								
ΜΟΝΤ. ΜΑΘΗΣΗΣ	**ACCURACY**		**KAPPA**		**KAPPA TEMP**		**RAM HOURS**	
Εννοιολογική Απόκλιση	Απότομη Current/Mean	Σταδιακή Current/Mean	Απότομη Current/Mean	Σταδιακή Current/Mean	Απότομη Current/Mean	Σταδιακή Current/Mean	Απότομη	Σταδιακή
HT	85,10/85,58	86,50/84,69	69,58/85,58	72,51/66,79	69,47/69,24	72,39/67,01	1,061	1,052
HAT	**89,00**/87,80	88,40/87,40	77,74/72,97	76,53/72,16	77,46/73,72	76,28/72,90	1,328	1,071
AdaHOT	87,70/85,07	86,70/84,80	75,07/69,18	72,92/67,08	74,80/70,09	72,80/67,49	6,090	4,375
OzaBag	86,20/85,80	85,80/84,96	71,86/68,75	71,02/67,33	71,72/69,72	70,96/67,73	2,113	2,158
OzaBagASHT	87,40/86,68	87,00/86,64	74,37/70,59	73,53/70,73	74,18/71,51	73,42/71,31	2,221	2,257
OzaBagAdwin	**88,00/87,38**	**88,80/87,54**	75,61/72,13	77,33/72,51	75,41/72,92	77,10/73,20	4,844	4,508
LeveragingBag	**89,10/**88,60	**89,00/88,44**	77,91/74,85	77,74/74,52	77,66/75,51	77,51/75,14	1,028	9,239

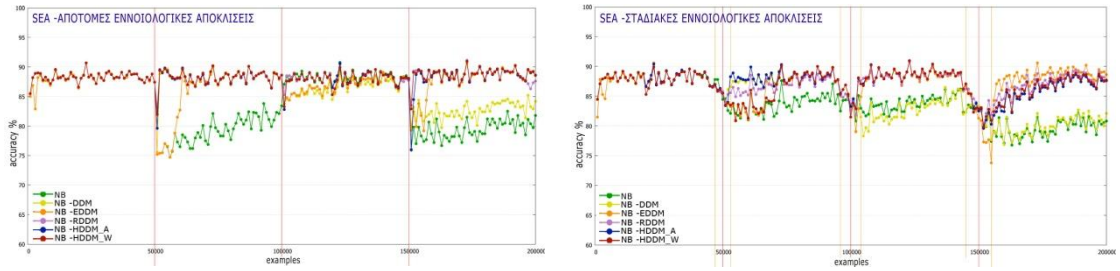
OzaBoost	**88,60/87,05**	88,40/86,47	76,88/71,46	76,48/70,35	76,64/72,24	76,28/70,94	2,464	2,509
OCBoost	**88,10**/87,40	88,00/87,09	75,92/72,10	75,54/71,61	75,61/72,80	75,46/72,25	2,871	2,933
AWE(10model)	87,50/87,46	87,20/86,98	74,68/72,19	74,03/71,17	74,39/72,98	73,82/72,04	1,544	1,532
AUE(10 model)	**88,60/88,13**	**89,20/87,69**	76,94/73,88	78,14/72,99	76,64/74,44	78,14/72,99	3,293	3,291
DWM	87,90/88,09	88,40/87,61	75,42/73,37	76,46/72,51	75,20/74,33	76,28/73,36	2,247	2,131
LearnNSE(S=10)	85,60/85,61	85,80/85,26	70,82/67,94	71,22/67,29	70,49/69,01	70,96/68,34	1,147	1,115
PairerdLearners	83,80/84,57	83,80/84,01	67,20/65,78	67,22/64,61	66,80/66,87	66,87/65,76	6,321	6,982

Πίνακας 6 Αποτελέσματα απότομων και σταδιακών εννοιολογικών αποκλίσεων – SEA

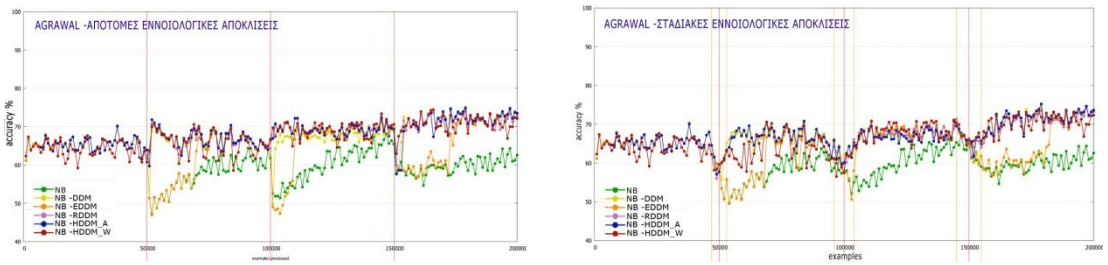
AGRAWAL- ΑΠΟΤΕΛΕΣΜΑΤΑ HT / HAT/ ΣΥΝΟΛΑ								
ΜΟΝΤ. ΜΑΘΗΣΗΣ	**ACCURACY**		**KAPPA**		**KAPPA TEMP**		**RAM HOURS**	
Εννοιολογική Απόκλιση	Απότομη Current/Mean	Σταδιακή Current/Mean	Απότομη Current/Mean	Σταδιακή Current/Mean	Απότομη Current/Mean	Σταδιακή Current/Mean	Απότομη	Σταδιακή
HT	67,90/64,78	67,20/64,41	31,08/28,17	29,04/27,24	32,14/27,04	30,80/26,41	5,291	4,766
HAT	**90,60/85,38**	**82,10/78,61**	80,37/70,32	64,71/56,72	80,13/69,51	62,24/55,53	2,604	2,344
AdaHOT	68,50/68,43	68,20/67,88	31,64/34,90	30,60/33,60	33,40/34,56	32,91/33,55	4,237	4,012
OzaBag	67,40/66,58	67,50/66,13	29,08/30,19	28,96/29,30	31,08/30,84	31,43/30,03	4,627	5,195
OzaBagASHT	77,70/70,87	75,70/69,97	51,27/38,62	46,52/36,83	52,85/39,72	48,73/37,96	4,537	4,466
OzaBagAdwin	**90,00/82,84**	**91,00/81,01**	79,14/64,40	81,42/60,99	78,86/64,21	81,01/60,48	8,345	8,136
LeveragingBag	81,20/78,24	81,20/75,23	60,89/53,90	60,97/44,47	60,25/54,60	60,34/48,48	2,443	2,506
OzaBoost	74,20/70,41	75,10/70,16	45,42/38,64	47,19/38,08	45,45/38,58	47,47/38,20	5,636	5,097
OCBoost	73,00/68,78	73,80/69,35	42,50/37,36	43,87/36,51	42,92/37,38	44,73/36,63	5,901	5,591
AWE(10model)	72,30/73,49	72,00/72,28	40,58/43,47	39,90/41,20	41,44/44,37	40,93/42,45	1,928	2,900
AUE(10 model)	**88,90/82,59**	**88,90/81,18**	77,26/63,86	77,14/60,87	76,53/63,66	76,58/60,84	8,056	8,648
DWM	71,20/67,15	71,40/66,34	38,01/30,78	37,64/29,26	39,11/31,98	39,66/30,39	2,867	2,618
LearnNSE(S=10)	70,40/64,74	69,40/64,09	35,98/25,76	34,00/24,57	37,42/27,02	35,44/25,78	2,800	2,772
PairerdLearners	73,60/68,02	73,80/67,27	43,77/32,90	44,19/31,51	44,19/33,76	44,73/32,27	1,916	2,973

Πίνακας 7 Αποτελέσματα απότομων και σταδιακών εννοιολογικών αποκλίσεων – AGRAWAL

6.1.3 Συζήτηση

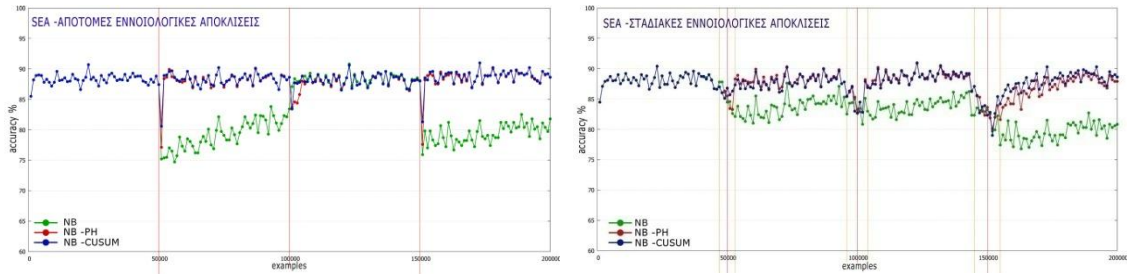


Εικόνα 18: SEA Απότομες και σταδιακές εννοιολογικές αποκλίσεις

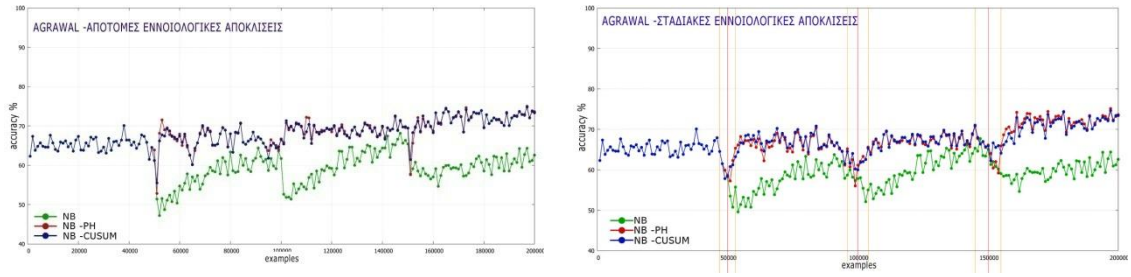


Εικόνα 19: AGRAWAL απότομες και σταδιακές εννοιολογικές αποκλίσεις

Η μέθοδος DDM είναι μια καλή επιλογή για την ανίχνευση απότομων εννοιολογικών αποκλίσεων και για σταδιακών, όταν δεν είναι πολύ αργά σταδιακές, όταν μια έννοια δεν περιέχει πολλά παραδείγματα. Η EDDM βελτιώνει την DDM στην ανίχνευση σταδιακών εννοιολογικών αποκλίσεων διατηρώντας μια εξίσου καλή ή και καλύτερη απόδοση στην ανίχνευση απότομων εννοιολογικών αποκλίσεων. Η RDDM αποδίδει εξίσου καλά τόσο στην ανίχνευση σταδιακών όσο και όσο απότομων εννοιολογικών αποκλίσεων, άλλα έχει ένα υψηλότερο κόστος RAM/hours. Οι μέθοδοι HDDM$_A$ και HDDM$_w$ αποδίδουν καλά και σε σταδιακές και σε απότομες εννοιολογικές αποκλίσεις διατηρώντας ένα χαμηλό κόστος RAM/Hours, έχουν μικρές διαφοροποιήσεις στην απόδοση τους.

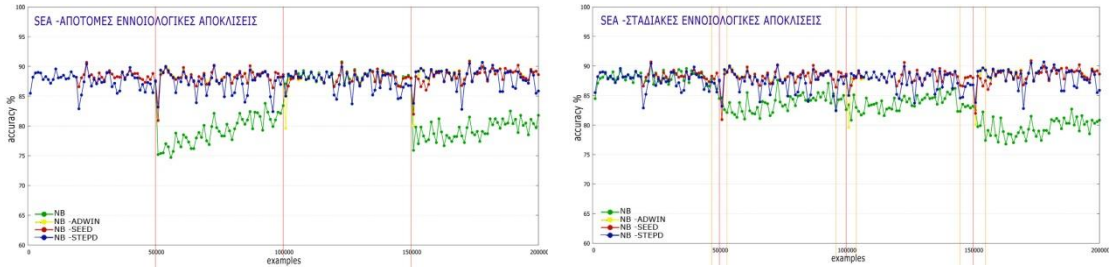


Εικόνα 20: SEA Απότομες και σταδιακές εννοιολογικές αποκλίσεις (PH-CUSUM)

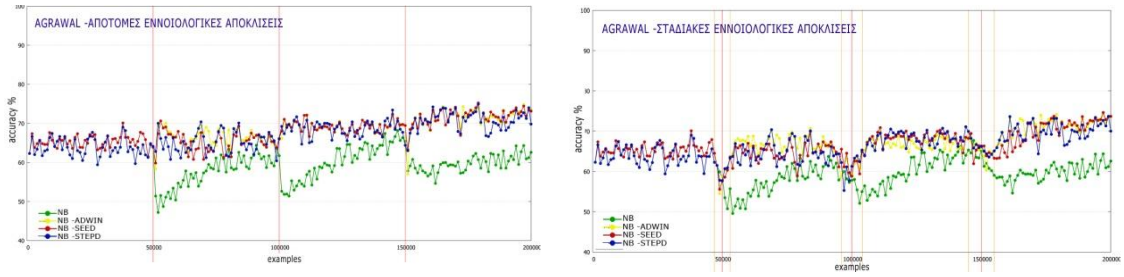


Εικόνα 21: AGRAWAL απότομες και σταδιακές εννοιολογικές αποκλίσεις (PH-CUSUM)

Οι μέθοδοι διαδοχικής ανάλυσης CUSUM και PH έχουν μια εξίσου καλή απόδοση στην αντιμετώπιση τόσο σταδιακών όσο και απότομων εννοιολογικών αποκλίσεων.



Εικόνα 22: SEA Απότομες και σταδιακές εννοιολογικές αποκλίσεις (ADWIN-SEED-STEPD)



Εικόνα 23: SEA Απότομες και σταδιακές εννοιολογικές αποκλίσεις (ADWIN-SEED-STEPD)

Οι τεχνικές παραθύρων είναι κατάλληλοι για την αντιμετώπιση τόσο σταδιακών όσο και απότομων εννοιολογικών αποκλίσεων. Τη χαμηλότερη απόδοση την έχει η μέθοδος STEPD, ίσως την καλύτερη απόδοση την έχει η μέθοδος SEED αλλά έχει υψηλότερο κόστος RAM/hours. Μια καλή επιλογή είναι η μέθοδος ADWIN, η οποία επιτυγχάνει μια καλή απόδοση διατηρώντας ένα λογικό κόστος RAM/hours και έχει θεωρητικές εγγυήσεις.

Οι μέθοδοι ανίχνευσης αλλαγών είναι μια καλή επιλογή για την αντιμετώπιση απότομων και σταδιακών εννοιολογικών αποκλίσεων. Οι τεχνικές αυτές φαίνεται να αποδίδουν καλύτερα από αρκετούς από τους συνδυαστικούς αλγόριθμους μάθησης (σύνολα), στην αντιμετώπιση αυτού του είδους των αλλαγών. Μια εξίσου καλή απόδοση ή και καλύτερη απόδοση πετυχαίνουν οι μέθοδοι OzaBagAdwin, OzaBagASHT, OzaBoost, LeveragingBag και AUE αλλά με υψηλότερο κόστος RAM/hours.

6.1.3.1 Προσομοίωση Απότομων και Σταδιακών Επαναλαμβανόμενων Εννοιολογικών Αποκλίσεων

Στο παρόν πείραμα χρησιμοποιήθηκαν τα συνθετικά δεδομένα των γεννητριών SINE για την δημιουργία

1. Μια ροής, από τα συνθετικά δεδομένα SINE, με 3 απότομες επαναλαμβανόμενες εννοιολογικές αποκλίσεις. Οι έννοιες επαναλαμβάνονται sine1-sine2-sine1-sine2.

Ενδεικτική εντολή (αλγόριθμος NB STEPD) MOA

EvaluatePrequential -l (drift.SingleClassifierDrift -d STEPD) -s (ConceptDriftStream -s generators.SineGenerator -d (ConceptDriftStream -s (generators.SineGenerator -f 2) -d (ConceptDriftStream -s generators.SineGenerator -d (generators.SineGenerator -f 2) -p 50000 -w 1) -p 50000 -w 1) -p 50000 -w 1 -r 5) -e (WindowClassificationPerformanceEvaluator -o -p -r -f) -i 200000 -f 1000

2. Μια ροής, από τα συνθετικά δεδομένα sine, με 3 σταδιακές εννοιολογικές αποκλίσεις, έτσι ώστε το μέγεθος κάθε έννοιας να είναι διαφορετικό το ένα από το άλλο. Οι έννοιες επαναλαμβάνονται sine1-sine2-sine1-sine2.

Ενδεικτική εντολή (αλγόριθμος OzaBag) MOA

EvaluatePrequential -l meta.OzaBag -s (ConceptDriftStream -s generators.SineGenerator -d (ConceptDriftStream -s (generators.SineGenerator -f 2) -d (ConceptDriftStream -s generators.SineGenerator -d (generators.SineGenerator -f 2) -p 50000 -w 10000) -p 50000 -w 8000) -p 50000 -w 6000 -r 5) -e (WindowClassificationPerformanceEvaluator -o -p -r -f) -i 200000 -f 1000

Ροή	Αριθμός drift	Θέση	πλάτος	θέση	πλάτος	θέση	πλάτος
SINE(abrupt CD)	3	50000	1	100000	1	150000	1
SINE(abrupt CD)	3	50000	6000	100000	8000	150000	10000

Πίνακας 8 Πληροφορίες εννοιολογικών αποκλίσεων ροών.

Τα δεδομένα όλων των ροών ανέρχονται σε 200.000 παραδείγματα. Η αξιολόγηση τους γίνεται με χρήση της μεθόδου EvaluatePrequential πάνω σε ένα συρόμενο παράθυρο (sliding windows) πλάτους w=1000. Οι μετρήσεις ορίζεται να επιστρέφονται κάθε 1000 παραδείγματα.

6.1.3.2 Πειραματικά αποτελέσματα

SINE STREAM (ΑΠΟΤΟΜΗ-ΣΤΑΔΙΑΚΗ)								
ΜΟΝΤ. ΜΑΘ.	**ACCURACY**		**KAPPA**		**KAPPA TEMP**		**RAM HOURS**	
Εννοιολογική Απόκλιση	Απότομη Current/Mean	Σταδιακή Current/Mean	Απότομη Current/Mean	Σταδιακή Current/Mean	Απότομη Current/Mean	Σταδιακή Current/Mean	Απότομη	Σταδιακή
NB	3,60/48,59	5,20/48,66	-92,01/-2,19	-89,18/-2,16	-95,14/-3,82	-90,74/-3,41	3,647	3,951
NB-DDM	91,80/92,43	**93,10**/90,08	83,44/84,69	86,08/80,06	83,40/84,68	86,12/80,03	2,416	2,531
NB-EDDM	92,10/91,50	92,40/89,91	84,04/82,95	84,67/78,36	84,01/82,71	84,71/78,36	1,785	**1,874**
NB-RDDM	**92,40/92,99**	**93,10/90,15**	84,63/85,83	86,08/80,18	84,62/85,83	86,12/80,15	9,776	**9,352**
NB-SEED	91,70/92,62	92,50/89,25	83,23/85,07	84,85/78,36	83,20/85,08	84,91/78,36	5,281	5,465
NB-ADWIN	91,80/92,55	92,80/89,83	83,44/84,94	85,48/79,54	83,40/84,95	85,51/79,51	4,745	4,589
NB-STEPD	**92,90/93,12**	**92,90/89,72**	**85,59/86,09**	85,60/79,30	85,63/86,09	85,71/79,28	2,985	4,062
NB-HDDM_A	91,80/92,54	**93,00/90,04**	83,44/84,92	85,88/79,96	83,40/84,93	85,92/79,93	2,782	3,032
NB-HDDM_w	**91,80/92,56**	92,30/89,91	83,44/84,96	84,46/79,70	83,40/84,97	84,51/79,68	2,975	3,114
NB-CUSUM	**91,80/92,56**	93,00/90,04	83,44/85,01	85,88/79,96	83,40/85,01	85,92/79,93	2,603	2,150
NB-PH	91,80/92,44	92,50/89,97	83,44/84,71	84,87/79,83	83,40/84,70	84,91/79,80	2,795	2,918
HT	97,00/78,60	95,50/68,08	93,96/56,77	90,93/36,05	93,93/56,77	90,95/35,85	1,636	1,353
HAT	99,40/98,02	99,20/93,41	98,79/96,01	98,39/86,77	98,79/95,97	98,39/86,73	3,739	1,759
OzaBag	92,20/86,15	98,90/75,53	98,39/72,29	97,78/51,26	98,38/71,94	97,79/50,93	2,509	2,405
OzaBagASHT	96,80/93,66	88,40/88,04	93,51/87,21	97,74/88,04	93,52/87,10	94,77/75,94	3,008	3,008
OzaBagAdwin	**99,50/95,89**	96,40/90,62	98,99/91,74	92,75/81,17	98,99/91,62	92,76/81,13	5,648	2,584
LeveragingBag	**99,70/97,03**	99,80/94,77	99,40/94,00	99,60/89,51	99,39/93,93	99,60/89,47	8,390	9,634
OzaBoost	98,90/97,73	97,90/91,62	97,78/95,40	95,77/83,16	97,77/95,36	95,77/83,13	3,095	3,149
OCBoost	**99,50/96,97**	98,80/91,74	98,99/94,02	97,58/83,40	98,99/93,78	97,59/83,36	3,094	3,229
AWE/10model	97,40/95,92	97,50/92,27	94,76/91,77	94,96/84,48	94,74/91,66	94,97/84,40	2,644	2,668
AUE/10 model	**99,90/97,80**	99,70/94,23	99,80/95,60	99,40/88,45	99,80/95,44	99,40/88,36	3,958	4,151
DWM	92,70/93,37	92,70/90,09	85,23/86,59	83,85/80,52	85,22/86,58	83,90/80,05	1,042	1,538
LearnNSE (S=10)	91,50/92,22	91,30/88,94	82,72/84,28	82,33/77,73	82,72/85,23	82,49/77,73	1,799	1,727
PairerdLearne	91,80/92,64	92,40/89,95	83,41/85,11	84,67/79,77	83,40/85,11	84,71/79,75	8,796	1,669

Πίνακας 9 Αποτελέσματα επαναλαμβανόμενων απότομων και σταδιακών εννοιολογικών αποκλίσεων – Sine

6.1.3.2 Συζήτηση

Στις επαναλαμβανόμενες έννοιες οι μέθοδοι συνόλου Bagging (LeveragingBag , OzaBagAdwin) και οι μέθοδοι Boosting(OzaBoost, OCBoost) έχουν καλύτερη απόδοση από τους ανιχνευτές αλλαγών. Επίσης καλή απόδοση από τις παθητικές μεθόδους συνόλου έχει η AUE και τα δέντρα HAT. Η μέθοδος STEPD έχει την καλύτερη απόδοση από όλους τους ανιχνευτές αλλαγών.

6.2.1 Προσομοίωση Βαθμιαίων Σταδιακών Εννοιολογικών Αποκλίσεων (incremental)

Στο παρόν πείραμα χρησιμοποιήθηκε η γεννήτρια HYPERPLANE για την προσομοίωση μιας ροής με βαθμιαία σταδιακές εννοιολογικές αποκλίσεις. Η γεννήτρια HYPERPLANE δημιουργεί εννοιολογικές αποκλίσεις μεταβάλλοντας τις τιμές των βαρών ενός περιστρεφόμενου υπερεπιπέδου, καθώς η ροή δεδομένων εξελίσσεται. Οι βασικές παράμετροι της γεννήτριας, που ελέγχουν το μέγεθος και την κατεύθυνση της εννοιολογικής απόκλισης είναι:

- -t : Μέγεθος της αλλαγής για κάθε παράδειγμα.
- -s : Ποσοστό πιθανότητας να αντιστραφεί η κατεύθυνση της αλλαγής.

1) Θέσαμε t=0,01 για την προσομοίωση μιας ροής με βαθμιαίες σταδιακές εννοιολογικές αποκλίσεις και

Ενδεικτική εντολή MOA (αλγόριθμος DWM)

EvaluatePrequential -l meta.DynamicWeightedMajority -s (generators.HyperplaneGenerator -k 10 -t 0.01) -i 100000 -f 1000

2) t=0,001 για την προσομοίωση μιας ροής δεδομένων με βαθμιαίες σταδιακές εννοιολογικές αποκλίσεις, με πολλές λεπτές, ανεπαίσθητες αλλαγές.

Ενδεικτική εντολή MOA (αλγόριθμος HT)

EvaluatePrequential -l trees.HoeffdingTree -s (generators.HyperplaneGenerator -k 10 -t 0.001) -i 100000 -f 1000

Τα δεδομένα όλης της ροής ανέρχονται σε 100.000 παραδείγματα. Η αξιολόγηση κάθε ροής γίνεται με χρήση της μεθόδου EvaluatePrequential

- α) πάνω σε ένα συρόμενο παράθυρο (sliding windows) σταθερού μεγέθους, ορίζουμε w=1000 και

- β) πάνω σε ένα παράθυρο προσαρμοστικού μεγέθους (adwin sliding windows). Οι μετρήσεις ορίζεται να επιστρέφονται κάθε 1000 παραδείγματα.

Στο πείραμα αυτό, δεν έχουμε γνώση για την ακριβή θέσης της δημιουργίας της κάθε εννοιολογικής απόκλισης. Η παρατήρηση, ότι η αλλαγή του μεγέθους ενός συρόμενου παραθύρου αλλάζει και τα πειραματικά αποτελέσματα, μας οδήγησε στο συμπέρασμα ότι θα είχαμε πιο αξιόπιστες μετρήσεις αν κάναμε τη χρήση ενός συρόμενου και ενός προσαρμοστικού παραθύρου.

6.2.2 Πειραματικά αποτελέσματα

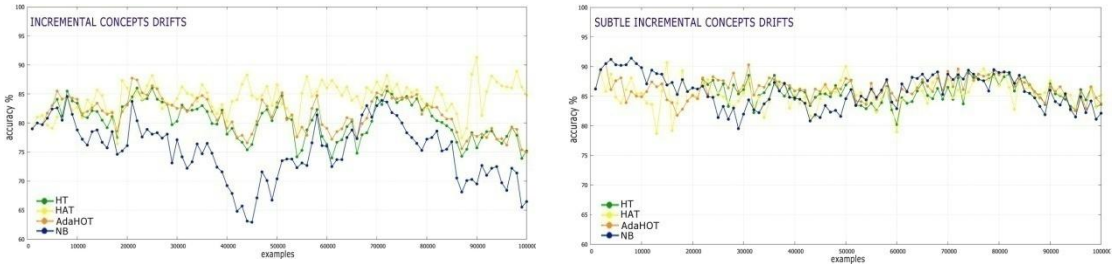
HYPERPLANE GENERATOR / t=0,01 βαθμιαία αυξητική εννοιολογική απόκλιση / t=0,001 λεπτή βαθμιαία sliding window								
MONT. ΜΑΘΗΣΗΣ	**ACCURACY (current/mean)**		**KAPPA (current/mean)**		**KAPPA temp (current/mean)**		**RAM HOURS**	
Εννοιολογική Απόκλιση	Βαθμιαία αυξητική	Λεπτά Βαθμιαία αυξητική	Βαθμιαία αυξητική	Λεπτά Βαθμιαία αυξητική	Βαθμιαία αυξητική	Λεπτά Βαθμιαία αυξητική	Βαθμιαία αυξητική	Λεπτά Βαθμιαία αυξητική
NB	66,50/75,27	82,10/85,83	32,94/50,51	64,16/71,64	31,07/50,53	63,47/71,56	5,843	4,944
HT	75,20/80,40	83,60/85,66	50,27/60,78	67,14/71,28	48,97/60,79	66,53/71,20	9,633	4,444
HAT	84,80/84,32	85,20/85,49	69,52/68,62	70,37/70,94	68,72/68,63	69,80/70,86	6,103	1,031
AdaHOT	75,00/81,44	83,70/86,38	49,56/62,86	67,35/72,72	48,56/62,86	66,73/72,63	6,206	4,094
OzaBag	76,00/81,13	84,50/86,66	51,87/62,23	68,96/73,29	50,62/62,24	68,37/73,21	3,414	1,981
OzaBagASHT	81,80/84,55	**86,60/88,17**	63,41/69,08	73,19/76,32	62,55/69,09	72,65/76,25	2,810	2,262
OzaBagAdwin	**86,40/**85,30	**87,50/87,80**	72,73/70,59	74,97/75,58	72,02/70,61	74,49/75,51	4,881	4,257
LeveragingBag	83,50/84,79	**87,40/86,20**	66,94/69,56	74,80/72,37	66,05/69,56	74,29/72,29	1,200	1,108
OzaBoost	82,40/84,21	85,80/86,53	64,73/68,40	71,57/73,04	63,79/68,41	71,02/72,96	2,602	2,162
OCBoost	81,80/84,15	86,00/86,83	63,56/62,25	71,99/73,63	62,55/63,68	71,43/73,55	2,883	2,820
AWE(10model)	85,90/86,11	**89,00/**89,41	**71,72/72,19**	**77,97/78,81**	**70,99/72,17**	**77,55/78,71**	1,681	1,518
AUE(10 model)	85,00/85,83	88,60/87,31	69,95/71,64	77,20/74,60	69,14/71,62	76,73/74,50	3,763	3,787
DWM	**89,50/**88,08	**89,50/**90,21	78,95/76,14	78,97/80,39	78,40/76,14	78,57/80,34	1,081	1,421
LearnNSE(S=10)	**87,90/**86,08	87,30/86,19	75,73/72,13	74,62/72,35	75,10/72,13	74,08/72,27	1,064	9,392
PairerdLearners	82,70/81,22	81,30/80,57	65,37/62,40	62,57/62,99	64,40/62,45	61,84/60,89	1,125	1,091

Πίνακας 10. Αποτελέσματα βαθμιαίων σταδιακών και βαθμιαία σταδιακών εννοιολογικών αποκλίσεων/sliding-window

HYPERPLANE GENERATOR / t=0,01 βαθμιαία αυξητική εννοιολογική απόκλιση / t=0,001 λεπτή βαθμιαία Adwin sliding window								
MONT. ΜΑΘΗΣΗΣ	**ACCURACY (current/mean)**		**KAPPA (current/mean)**		**KAPPA temp (current/mean)**		**RAM HOURS**	
Εννοιολογική Απόκλιση	Βαθμιαία αυξητική	Λεπτά Βαθμιαία αυξητική	Βαθμιαία αυξητική	Λεπτά Βαθμιαία αυξητική	Βαθμιαία αυξητική	Λεπτά Βαθμιαία αυξητική	Βαθμιαία αυξητική	Λεπτά Βαθμιαία αυξητική
NB	69,32/75,60	83,85/86,02	38,64/51,20	67,70/72,04	38,68/51,10	67,60/71,90	1,304	1,123
HT	77,53/80,63	85,35/85,72	55,05/61,27	70,71/71,44	55,08/61,18	70,62/71,28	1,578	7,371
HAT	**86,15/84,52**	85,33/85,61	72,29/69,04	70,65/71,22	72,31/68,95	70,56/71,04	6,759	7,671
AdaHOT	78,06/81,97	85,48/86,56	56,12/63,94	70,95/73,12	56,15/63,85	70,87/72,96	1,237	8,567
OzaBag	77,90/81,42	85,64/86,84	55,81/62,83	71,27/73,67	55,83/62,75	71,19/73,52	2,348	2,394
OzaBagASHT	83,32/84,62	**87,51/88,07**	66,63/69,25	75,01/76,14	66,65/69,16	74,94/75,99	2,584	2,584
OzaBagAdwin	**86,43/85,03**	**87,08/87,86**	72,86/70,06	74,17/75,71	72,87/69,97	74,09/75,57	4,701	4,953
LeveragingBag	85,57/84,46	**87,15/85,89**	71,14/68,92	74,31/71,77	71,16/68,83	74,23/71,60	9,803	1,149
OzaBoost	82,93/84,19	86,07/86,98	65,87/68,32	72,14/72,25	65,89/68,30	72,05/72,59	2,290	2,429
OzaBoostAdwin	71,50/75,74	78,74/77,61	43,00/51,47	57,49/55,21	43,04/51,36	57,35/54,95	2,233	1,129
AWE(10model)	87,11/85,96	**89,10/89,73**	74,22/71,92	78,20/79,47	74,24/71,82	78,14/79,34	1,665	1,659
AUE(10 model)	87,02/85,52	87,84/87,15	74,04/71,04	75,69/74,29	74,05/70,94	75,61/74,13	3,951	4,168
DWM	**88,57/87,60**	**88,83/89,99**	**77,14/75,20**	77,65/79,99	77,16/75,12	77,58/79,86	1,269	1,653
LearnNSE(S=10)	86,31/85,61	86,74/86,30	72,61/71,22	73,47/72,59	72,63/71,14	73,39/72,43	1,057	1,093
PairerdLearners	82,55/81,30	75,00/80,43	65,10/62,60	50,00/60,86	65,12/62,51	49,85/60,63	1,337	1,610

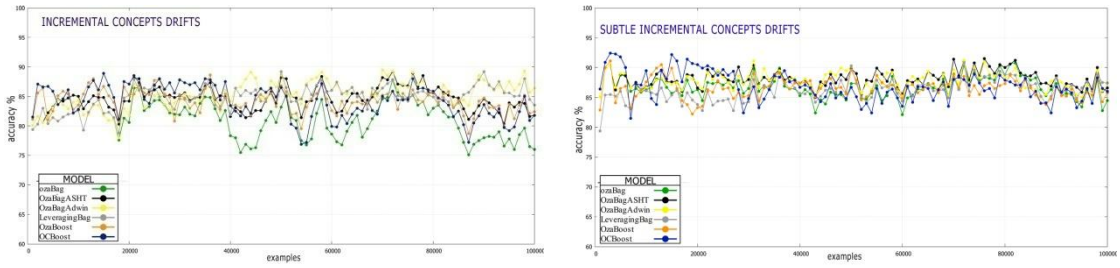
Πίνακας 11 Αποτελέσματα βαθμιαίων σταδιακών και βαθμιαία σταδιακών εννοιολογικών αποκλίσεων/adwin-sliding-window

6.2.3 Συζήτηση



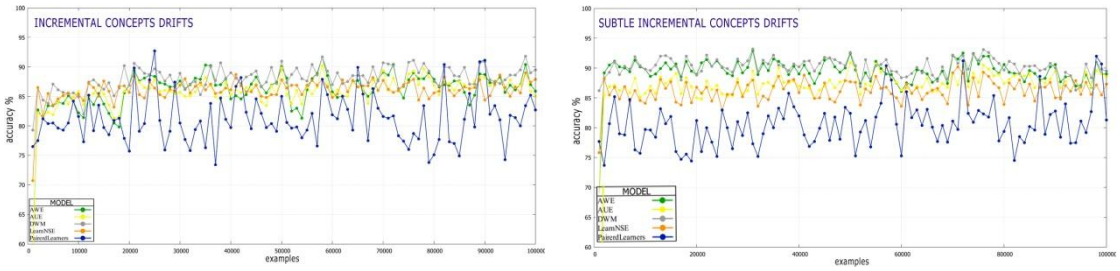
Εικόνα 24:ACCURACY HT / HAT/ADAHOT/ NB

Η μέθοδος HT δίνει σχεδόν πάντα καλύτερες μετρήσεις από την μέθοδο NB. Το σύνολο δέντρων επιλογής, Adaptive Hoeffding option tree (AdaHOT), δεν δίνει πολύ καλύτερα αποτελέσματα από ένα απλό HT δέντρο. Μια σημαντική βελτίωση στα αποτελέσματα πετυχαίνει το προσαρμοστικό δέντρο HAT. Όταν οι εννοιολογικές αποκλίσεις είναι ανεπαίσθητες η απόδοση των αλγορίθμων είναι παρόμοια.



Εικόνα 25:ACCURACY OzaBag/OzabagASHT/OzaBagAdwin/ Leveraging Bag/ OzaBoost/OcBoost

Από τους συνδυαστικούς αλγόριθμους μάθησης (σύνολα) Bagging την καλύτερη απόδοση έχει η μέθοδος OzaBagAdwin διατηρώντας ένα σχετικά καλό μέτρο κόστος, χαμηλότερο και από ένα απλό HT δέντρο. Στη ροή με λεπτές βαθμιαίες εννοιολογικές αποκλίσεις μια εξίσου καλή απόδοση έχουν και οι μέθοδοι OzaBagASHT και LeveragingBag. Από τους αλγόριθμους boosting η μέθοδος OCBoost έχει μια καλή απόδοση στη ροή με τις ανεπαίσθητες εννοιολογικές αποκλίσεις.



Εικόνα 26:ACCURACY AWE/AUE/DWM/learnNSE/PairedLearnes

Η μέθοδος AWE ξεπερνάει την μέθοδο AUE, ειδικά στη ροή με μικρές ανεπαίσθητες εννοιολογικές αποκλίσεις, διατηρώντας ένα χαμηλό μέτρο κόστους. Η μέθοδος DWM έχει μια πολύ απόδοση με μικρό κόστος και στις δύο ροές. Μια λιγότερο καλή αλλά ικανοποιητική απόδοση έχει η μέθοδος LearnNSE. Η μέθοδος PairerdLearners δεν έχει τόσο καλή απόδοση στη ροή με ανεπαίσθητες εννοιολογικές αποκλίσεις.

Στο πείραμα αυτό δεν παρουσιάστηκαν τα αποτελέσματα των ανιχνευτών αλλαγών επειδή τα αποτελέσματα τους έδειξαν ότι δεν είναι καλές πρακτικές για ανίχνευση βαθμιαίων σταδιακών αποκλίσεων. Οι μέθοδοι AWE και DMW φαίνεται να είναι δύο καλές επιλογές πετυχαίνουν μια πολύ καλή απόδοση διατηρώντας ένα χαμηλό υπολογιστικό κόστος και στις δύο ροές.

6.3.1 Προσομοίωση Βαθμιαίων Σταδιακών Εννοιολογικών Αποκλίσεων (incremental)

Χρησιμοποιήσουμε τη γεννήτρια RandomRBFGeneratorDrift για την προσομοίωση των βαθμιαίων σταδιακών εννοιολογικών αποκλίσεων. Η γεννήτρια αυτή, δημιουργεί μια τυχαία ροή, με τη χρήση μιας συνάρτησης ακτινικής βάσης με μετατόπιση. Μια εννοιολογική απόκλιση εισάγεται μετακινώντας τα κεντροειδή με σταθερή ταχύτητα. Η βασική παράμετρος που καθορίζει την ταχύτητα της ροής είναι:

- -s : Ταχύτητα αλλαγής κεντροειδών στο μοντέλο.

Στόχος αυτού του πειράματος είναι να αξιολογηθεί η απόδοση μιας ροής με εννοιολογικές αποκλίσεις ως προς ταχύτητα της.

- Θέτουμε s=0,001 για να γρήγορη ροή δεδομένων, με γρήγορες εννοιολογικές αποκλίσεις και

Ενδεικτική εντολή MOA (αλγόριθμος OzaBagAdwin)

EvaluatePrequential -l meta.OzaBagAdwin -s (generators.RandomRBFGeneratorDrift -s 1.0E-4) -e (WindowClassificationPerformanceEvaluator -o -p -r -f) -i 1000000 -f 10000

- s=0,0001 για να δημιουργηθεί μια ροή μέτριας ταχύτητας.

Ενδεικτική εντολή MOA (αλγόριθμος Levaring Bag)

EvaluatePrequential -l meta.LeveragingBag -s (generators.RandomRBFGeneratorDrift -s 1.0E-4) -e (WindowClassificationPerformanceEvaluator -o -p -r -f) -i 1000000 -f 10000

Τα δεδομένα όλης της ροής ανέρχονται σε 1.000.000 παραδείγματα. Αξιολογούμε τις μεθόδους NB,HT,HAT και τους συνδυαστικούς αλγόριθμους μάθησης χρησιμοποιώντας τη μέθοδο EvaluatePrequential πάνω σε ένα συρόμενο παράθυρο (sliding windows) σταθερού μεγέθους, ορίζουμε w=1000. Οι μετρήσεις ορίζεται να επιστρέφονται κάθε 10.000 παραδείγματα.

6.3.2 Πειραματικά αποτελέσματα

RandomRBFGeneratorDrift/ ΓΡΗΓΟΡΗ ΡΟΗ S=0,001/ΜΕΤΡΙΑΣ ΤΑΧΥΤΗΤΑΣ ΡΟΗ S=0,0001 **sliding windows** w=1000								
ΜΟΝΤ. ΜΑΘΗΣΗΣ	**ACCURACY (current/mean)**		**KAPPA(current/mean)**		**KAPPA temp(current/mean)**		**RAM HOURS**	
Εννοιολογική Απόκλιση	**Γρήγορη**	**Μέτρια**	**Γρήγορη**	**Μέτρια**	**Γρήγορη**	**Μέτρια**	**Γρήγορη**	**Μέτρια**
NB	49,30/53,50	51,80/53,71	-1,08/7,07	3,92/7,45	0/6,68	4,93/6,77	6,153	6,796
HT	53,70/57,33	67,10/69,29	7,41/14,65	34,17/38,54	8,68/14,02	35,11/38,16	4,864	7,559
HAT	71,30/62,64	77,10/79,61	42,73/25,24	54,17/59,17	43,39/24,74	54,89/58,37	1,943	1,734
AdaHOT	61,30/64,58	**81,60/81,19**	22,64/29,12	63,18/62,35	23,67/28,68	63,71/62,13	3,611	1,826
OzaBag	62,10/63,70	75,70/78,55	24,23/27,36	51,40/57,06	25,25/26,86	52,07/56,80	1,927	2,445
OzaBagASHT	71,80/69,15	**82,00/80,44**	46,69/38,25	63,99/60,84	44,38/37,86	64,50/60,64	2,059	2,298
OzaBagAdwin	71,50/67,82	**84,40/85,69**	43,09/35,56	68,79/71,35	43,79/35,19	69,23/71,20	3,464	4,635
LeveragingBag	**84,40/81,25**	**90,20/89,89**	**68,80/62,46**	80,39/79,76	**69,23/62,26**	80,67/79,65	1,017	1,319
OzaBoost	64,10/65,75	78,00/83,24	28,23/31,48	55,96/66,45	29,19/31,02	56,61/66,25	2,471	2,858
OCBoost	63,40/65,94	**83,00/83,97**	26,76/31,85	66,01/67,92	27,81/31,38	66,47/67,72	1,034	9,140
AWE(10model)	70,80/66,12	71,40/70,11	41,71/32,18	42,84/40,18	42,41/31,79	42,41/39,97	1,344	1,533
AUE(10 model)	**76,40/71,90**	**90,80/88,53**	**52,90/43,76**	**81,59/77,01**	**53,45/43,42**	81,85/76,92	4,113	4,728
DWM	70,60/67,87	75,80/70,41	41,31/35,65	51,58/40,75	42,01/35,34	52,27/40,44	1,306	2,128
LearnNSE(S=10)	72,80/67,87	70,50/68,55	45,65/35,55	41,01/37,01	46,35/35,21	41,81/36,71	1,054	1,047
PairerdLearners	69,80/65,64	68,20/65,32	39,64/91,17	36,41/30,50	40,43/30,82	37,28/30,21	1,789	1,727

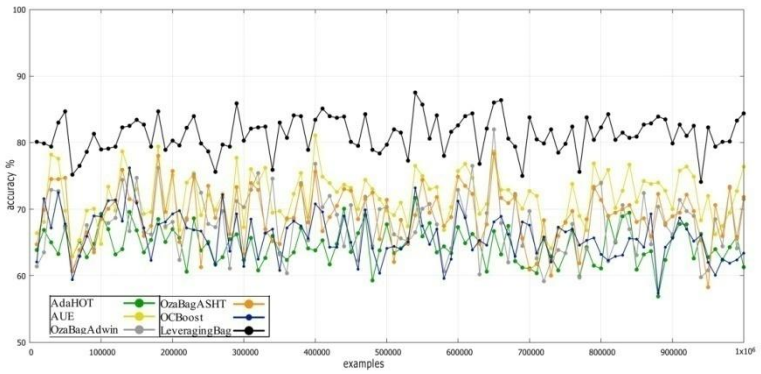
Πίνακας 12. Αποτελέσματα βαθμιαίων σταδιακών και βαθμιαία σταδιακών εννοιολογικών αποκλίσεων/sliding-window

6.3.3 Συζήτηση

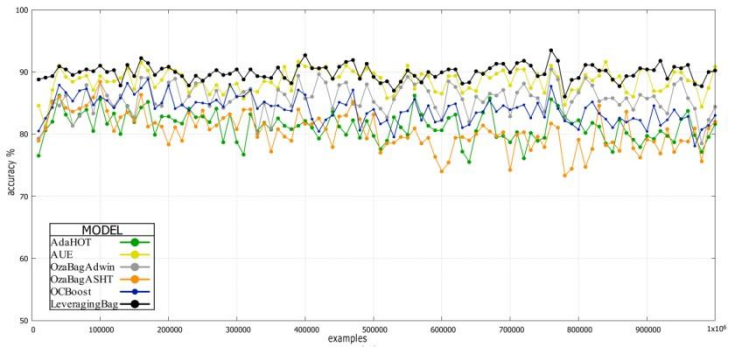
Η ταχύτητα των εννοιολογικών αποκλίσεων είναι ένα κρίσιμο σημείο. Οι αλγόριθμοι μάθησης έχουν καλύτερη απόδοση σε μια ροή δεδομένων, με εννοιολογικές αποκλίσεις μέτριας ταχύτητας. Στις γρήγορες εννοιολογικές αποκλίσεις (αλλάζουν οι έννοιες γρήγορα), ιδιαίτερα όταν είναι βαθμιαία σταδιακές, οι κατηγοριοποιητές έχουν μια δυσκολία προσαρμογής. Μια ουσιαστική μέθοδος κατηγοριοποίησης για ροή με γρήγορες εννοιολογικές αποκλίσεις είναι η Leveraring Bagging η δεύτερη καλύτερη μέθοδος είναι η AUE, μπορεί να χαρακτηριστεί η απόδοση της. Η μέθοδος Leveraging Bagging έχει ακρίβεια 84,40% και στατιστική kappa 68,80% και στατιστική Kappa Temp 69,23%. Η μέθοδος AUE

έχει ακρίβεια 76,40%,η στατιστική Kappa είναι 52,90% και η στατιστική Kappa Temp 53,45%.

Σε μια ροή με όπου οι εννοιολογικές αποκλίσεις δημιουργούνται με μέτρια σχετικά ταχύτητα καλή απόδοση έχουν οι συνδυαστικοί αλγόριθμοι μάθησης LeveragingBag, AUE, OzaBagAdwin, OCBoost, AdaHOT, AUE, OzaBagASHT.



Εικόνα 27:ACCURACY ΣΕ ΓΡΗΓΟΡΗ ΡΟΗ



Εικόνα 28:ACCURACY ΣΕ ΡΟΗ ΜΕΤΡΙΑΣ ΤΑΧΥΤΗΤΑΣ

Κεφάλαιο 7 - Πειραματική μελέτη σε πραγματικά δεδομένα

7.1 Εγκαθίδρυση Πειραμάτων (Poker)

Τα δεδομένα του συνόλου δεδομένων poker ανέρχονται σε 1.000.000 παραδείγματα. Αξιολογούμε τους αλγόριθμους NB, τους αλγορίθμους ανιχνευτών αλλαγών(3.7.2), HT, HAT και τους συνδυαστικούς αλγόριθμους μάθησης (σύνολα)(4.2) χρησιμοποιώντας τη μέθοδο EvaluatePrequential

- πάνω σε ένα συρόμενο παράθυρο (sliding windows) σταθερού μεγέθους, ορίζουμε w=1000 και
- πάνω σε ένα παράθυρο προσαρμοστικού μεγέθους (adwin sliding windows).

Οι μετρήσεις ορίζεται να επιστρέφονται κάθε 10.000 παραδείγματα.

7.1.1 Πειραματικά αποτελέσματα

	POKER / sliding window				POKER / adwin sliding window			
ΜΟΝΤΕΛΟ ΜΑΘΗΣΗΣ	**ACCURACY** (current/mean)	**KAPPA** (current/mean)	**KAPPA temp** (current/mean)	**RAM HOURS**	**ACCURACY** (current/mean)	**KAPPA** (current/mean)	**KAPPA temp** (current/mean)	**RAM HOURS**
NB	35,90/57,72	4,18/14,32	-38,44/-74,01	1,609	21,21/56,08	10,53/13,02	-88,30/-84,97	4,963
NB-DDM	46,70/62,17	6,42/20,41	-15,12/-56,12	4,671	43,82/61,73	21,63/18,05	-34,27/-60,77	1,271
NB-EDDM	81,40/77,50	65,82/51,17	59,83/9,60	2,396	77,42/77,73	58,10/42,87	46,03/7,11	8,765
NB-RDDM	**83,00/76,98**	69,49/48,34	63,28/7,30	8,940	81,68/76,68	66,16/36,12	56,21/2,43	3,305
NB-HDDM_A	**84,60/76,62**	71,77/49,23	66,74/6,02	3,096	81,28/75,91	65,29/37,52	55,25/-0,12	1,117
NB-HDDM_w	**84,40/77,29**	71,42/49,06	66,31/8,62	3,283	81,55/73,07	66,80/33,78	55,90/-12,75	1,241
NB-SEED	81,00/75,63	66,06/46,39	58,96/2,56	3,739	77,38/75,52	57,70/35,89	45,94/-1,16	1,363
NB-ADWIN	**85,10/74,34**	73,32/43,63	67,82/-2,33	3,847	80,81/73,49	63,63/32,15	54,14/-8,83	1,347
NB-STEPD	**83,50/77,54**	69,77/50,35	64,58/9,82	3,044	79,91/77,90	61,87/42,11	51,98/7,72	1,158
NB-CUSUM	81,00/72,81	66,76/41,22	58,96/-8,98	2,973	79,40/71,42	61,16/28,52	50,76/-17,89	1,066
NB-PH	70,80/70,31	51,60/36,31	36,93/-19,73	3,254	67,93/69,19	23,95/26,58	23,35/-26,06	1,130
HT	65,20/77,28	*37,33/49,15*	24,84/9,99	3,021	67,51/77,41	48,22/42,52	22,34/8,39	8,74
HAT	50,90/66,00	14,42/27,40	-6,05/-37,06	6,424	51,13/64,23	22,48/17,45	-16,79/-46,79	1,09
AdaHOT	71,40/81,39	46,09/58,01	38,23/26,37	1,738	67,51/77,59	48,22/43,19	22,34/9,39	8,698
OzaBag	80,20/83,68	60,39/62,04	57,24/34,78	1,636	77,64/82/91	56,12/54,27	46,56/29,81	2,316
OzaBagASHT	66,20/77,15	32,87/46,18	27,00/8,43	2,223	66,95/76,30	37,75/37,06	21,01/3,32	3,84
OzaBagAdwin	77,70/74,54	57,86/41,15	51,84/-1,99	2,946	72,07/73,26	40,57/26,09	33,26/-10,32	4,939
LeveragingBag	**99,10/88,19**	**98,46/72,54**	**98,06/50,80**	5,459	**98,74/87,20**	97,94/63,18	96,99/45,14	2,289

OzaBoost	**98,50/97,44**	97,44/73,28	96,76/52,75	2,387	**96,01/87,66**	93,49/67,58	90,47/48,27	2,996
OCBoost	72,00/79,87	36,64/51,76	39,52/19,52	2,989	67,36/78,64	17,77/43,72	21,99/12,59	3,722
AWE(10model)	56,20/59,83	*30,73/21,24*	5,40/-63,32	9,030	68,90/55,31	39,68/14,97	25,67/-86,58	1,121
AUE(10 model)	79,20/66,22	58,84/28,06	55,08/-38,19	2,736	79,75/60,33	60,43/19,70	51,62/-66,17	3,451
DWM	73,30/72,99	55,16/41,96	42,33/-9,17	1,482	71,71/72,20	40,52/28,31	32,40/-15,60	1,964
LearnNSE(S=10)	62,10/50,54	40,76/19,54	18,14/-64,40	7,989	70,33/57,92	49,73/14,18	29,09/-76,48	8,325
PairerdLearners	78,80/80,14	62,05/56,35	54,21/20,97	1,959	76,95/76,87	59,72/46,28	44,92/16,20	3,693

Πίνακας 13. Αποτελέσματα poker

7.1.2 Συζήτηση

Στο σύνολο αυτό των δεδομένων δεν είμαστε σε θέση να έχουμε ακριβή γνώση των ειδών των εννοιολογικών αποκλίσεων καθώς και την ακριβή τους θέση. Από την περιγραφή του συνόλου, που παρουσιάστηκε στην ενότητα 5, γίνεται αντιληπτό ότι οι εννοιολογικές αποκλίσεις θα συμβαίνουν επανειλημμένα και γρήγορα.

Η στατιστική kappa [50], που παίρνει τιμές 0-1, υποστηρίζει ότι ένας αλγόριθμος κατηγοριοποίησης επιτυγχάνει μια συμφωνία ως εξής:

- 0 = συμφωνία ισοδύναμη με την τύχη.
- 0,1 – 0,20 = ελαφρά συμφωνία.
- 0,21 – 0,40 = δίκαιη συμφωνία.
- 0,41 – 0,60 = μέτρια συμφωνία.
- 0,61 – 0,80 = ουσιαστική συμφωνία.
- 0,81 – 0,99 = σχεδόν τέλεια συμφωνία
- 1 = τέλεια συμφωνία [50].

Συμπεραίνουμε ότι οι αλγόριθμοι LeveragingBag και OzaBoost είναι οι αλγόριθμοι που επιτυγχάνουν μια τέλεια συμφωνία για αυτό το σύνολο. Οι αλγόριθμοι αυτοί επιτυγχάνουν 98,46% και 97,44% αντίστοιχα στατιστική Kappa. Επιπλέον επιτυγχάνουν μια πολύ καλή ακρίβεια 99,10% και 99,50% αντίστοιχα. Μια εξίσου καλή απόδοση έχουν στη στατιστική Kappa temp, δεν κατηγοριοποιούν δηλαδή χωρίς έλεγχο, με βάση την ετικέτα πρόβλεψης του προηγούμενου παραδείγματος. Ένα μειονέκτημα είναι ότι όχι έχουν κάποιο κόστος RAM/hours 5,459 και 2,387 αντίστοιχα. Από αυτό συμπεραίνουμε ότι ο OzaBoost είναι ένας ίσως ο καλύτερος αλγόριθμος για το σύνολο poker.

Οι ανιχνευτές αλλαγών ADWIN, HDDM_A, HDDM_w, STEPD, RDDM, EDDM, SEED, CUSUM, PH επιτυγχάνουν μια ουσιαστική συμφωνία. Αποδίδουν καλύτερα από τους περισσότερους αλγόριθμους συνδυαστικής μάθησης (σύνολα). Οι μέθοδοι συνόλου OzaBag, OzaBagAdwin, PairerdLearners, AUE, DWM, LearnNSE επιτυχαίνουν μια μέτρια συμφωνία. Μια δίκαια συμφωνία επιτυγχάνουν οι αλγόριθμοι HT,OzaBagASHT,AWE και OCBoost. Μια ελαφρά συμφωνία πετυχαίνει οι αλγόριθμοι HAT και Naïve Bayes.

7.2 Εγκαθίδρυση Πειραμάτων (Elec)

Τα δεδομένα του συνόλου δεδομένων elec ανέρχονται σε 45.312 παραδείγματα. Αξιολογούμε τους αλγόριθμους NB, τους αλγορίθμους ανιχνευτών αλλαγών(3.7.2), HT, HAT και τους συνδυαστικούς αλγόριθμους μάθησης (σύνολα) (4.2) τη μέθοδο EvaluatePrequential

- πάνω σε ένα συρόμενο παράθυρο (sliding windows) σταθερού μεγέθους, ορίζουμε w=1000 και
- πάνω σε ένα παράθυρο προσαρμοστικού μεγέθους (adwin sliding windows).

Οι μετρήσεις ορίζεται να επιστρέφονται κάθε 1.000 παραδείγματα.

7.2.1 Πειραματικά αποτελέσματα

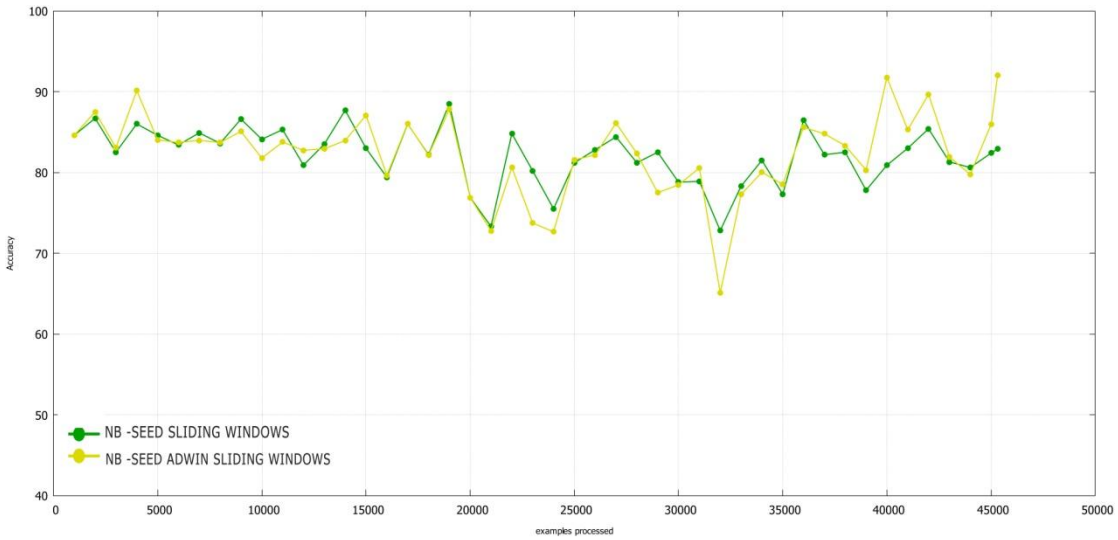
	ELEC/sliding window				ELEC/adwin sliding window			
ΜΟΝΤΕΛΟ ΜΑΘΗΣΗΣ	**ACCURACY (current/mean)**	**KAPPA (current/mean)**	**KAPPA temp (current/mean)**	**RAM HOURS**	**ACCURACY (current/mean)**	**KAPPA (current/mean)**	**KAPPA temp (current/mean)**	**RAM HOUR**
NB	75,30/73,37	48,86/39,97	-73,94/-90,16	4,657	84,38/73,80	69,16/39,64	-2,78/-82,56	1,113
NB-DDM	86,60/81,23	73,04/60,27	5,63/-60,27	2,622	**89,17/78,26**	**78,26/59,77**	28,74/-27/25	4,194
NB-EDDM	86,20/84,85	72,39/68,46	2,82/-6,77	1,781	84,63/84,99	69,34/67,30	-1,10/-3,23	2,200
NB-RDDM	87,20/84,23	74,48/67,05	9,86/-11,57	6,173	86,05/84,01	72,04/64,96	8,25/-10,57	8,818
NB-HDDM_A	**90,20/84,99**	**80,38/68,71**	30,99/-5,36	2,059	88,07/84,75	76,05/66,84	21,51/-4,40	2,880
NB-HDDM_w	86,50/84,12	73,08/66,86	4,93/-12,43	1,895	84,69/84,12	69,21/65,14	-0,68/-9,54	3,061
NB-SEED	**82,90/82,15**	**65,51/62,56**	**-20,42/-25,99**	**2,707**	**92,01/82,40**	**84,07/62,17**	**47,47/-21,36**	**3,664**
NB-ADWIN	78,80/80,00	57,18/57,64	-49,30/-41,48	2,162	85,07/79,74	70,06/54,68	1,79/-39,97	3,490
NB-STEPD	87,30/84,52	74,60/67,77	10,56/-9,30	2,032	85,40/84,68	70,72/66,59	3,93/-5,47	2,707
NB-CUSUM	76,20/79,20	52,04/55,43	-67,61/-48,45	1,940	71,88/78,81	43,38/52,53	-85,01/-47,54	2,327
NB-PH	76,10/78,00	51,40/52,85	-68,31/-56,94	1,790	84,38/77,83	68,51/50,53	-2,78/-54,27	2,479
HT	81,60/63,19	63,19/56,41	-29,58/-47,49	2,733	79,98/78,33	59,79/53,83	-31,66/-50,37	3,659
HAT	72,90/83,23	44,94/64,27	-90,85/-19,62	1,103	73,30/82,71	46,19/62,06	-75,66/-19,80	1,134
AdaHOT	**89,70/87,41**	79,53/73,56	28,17/10,57	1,029	88,92/87,05	77,77/72,75	27,14/9,94	1,167
OzaBag	88,10/82,50	76,10/63,06	16,20/-24,30	8,941	87,17/83,05	74,19/63,79	15,59/-17,17	9,629
OzaBagASHT	83,00/83,23	65,59/63,97	-19,72/-19,73	1,036	83,70/82,15	67,35/61,08	-7,25/-24,37	1,096
OzaBagAdwin	79,00/84,28	57,58/66,39	-47,89/-13,08	1,764	84,38/83,25	68,61/62,69	-2,78/-17,13	1,761
LeveragingBag	**92,10/89,82**	**84,17/78,50**	**44,37/27,60**	4,613	**90,98/89,46**	81,98/77,29	40,65/27,25	4,793
OzaBoost	88,50/86,78	77,07/72,32	19,01/6,05	9,001	86,60/86,20	73,09/70,43	11,87/4,15	1,051
OCBoost	**92,20/89,80**	**84,37/78,63**	45,07/27,29	1,364	89,58/89,09	79,13/76,25	31,48/24,16	1,136
AWE(10model)	78,20/71,01	55,80/40,20	-53,52/-106,84	5,212	**88,89/68,86**	77,78/36,39	26,91/-116,73	5,285
AUE(10 model)	**86,40/77,54**	72,62/52,27	4,23/-58,94	1,605	**89,33/77,32**	78,62/50,18	29,83/-57,07	1,638
DWM	75,70/79,60	50,93/56,38	-71,13/-45,30	4,187	75,28/78,75	50,23/52,66	-62,58/-47,95	4,912
LearnNSE(S=10)	71,50/71,14	42,65/41,22	-100,7/-104,06	5,958	63,07/70,87	25,97/40,38	-142,94/-101,11	1,744
PairerdLearners	**88,80/87,15**	77,50/73,26	21,13/9,23	1,048	87,17/87,44	74,33/72,98	15,60/13,63	1,141

Πίνακας 14. Αποτελέσματα ELE

7.2.2 Συζήτηση

Στο σύνολο αυτό των δεδομένων η καλύτερη μέθοδος είναι η LeveragingBag. Η δεύτερη καλύτερη είναι η OzaBoost, φαίνεται όμως να είναι καλύτερη από άποψη κόστους RAM/hours. Οι αλγόριθμοι αυτοί επιτυγχάνουν μια τέλεια συμφωνία. Μια σχεδόν τέλεια συμφωνία επιτυγχάνει και ο ανιχνευτής αλλαγών HDDM_A. Οι υπόλοιποι ανιχνευτές αλλαγών επιτυχαίνουν μια ουσιώδες συμφωνία όπως και οι μέθοδοι συνόλου bagging καθώς και η μέθοδος συνόλου AUE και η μέθοδος PairerdLearners. Οι μέθοδοι συνόλου AWE και DWM επιτυγχάνουν μια μέτρια συμφωνία και σύμφωνα με την στατιστική kappa σε ορισμένες στιγμές η απόδοση τους είναι χειρότερη από έναν κατηγοριοποιητή χωρίς αλλαγή.

Μια σημαντική παρατήρηση, στο πείραμα αυτό, είναι το πόσο σημαντικό, είναι το δείγμα δεδομένων που επιλέγουμε για να διατηρήσουμε το μοντέλο συναφής με τη φύση της ροής. Η μέθοδος ανίχνευσης αλλαγών SEED και η μέθοδος συνόλου AWE έχουν καλύτερη απόδοση σε συνδυασμό με ένα συρόμενο παράθυρο προσαρμοστικού μεγέθους. Στην παρακάτω εικόνα απεικονίζεται πώς μεταβάλλεται η ακρίβεια του ανιχνευτή αλλαγών SEED .



Εικόνα 29: NB –SEED SLIDING WINDOW AND ADWIN SLIDING WINDOWS

Κεφάλαιο 8 – Συμπεράσματα Μελλοντική Ερευνά

Στην παρούσα εργασία πραγματοποιήθηκε μια παρουσίαση της κατηγοριοποίησης ροών δεδομένων, ενώ δόθηκε έμφαση στο φαινόμενο της εννοιολογικής απόκλισης (concept drift). Έχουν προταθεί πολλές τεχνικές αναγνώρισης μιας εννοιολογικής απόκλισης, πολλές μέθοδοι ενός αλγόριθμου κατηγοριοποίησης και πολλοί συνδυαστικοί αλγόριθμοι μάθησης (ensembles). Στόχος της παρούσας εργασίας αποτελεί η σύγκριση της απόδοσης των αλγορίθμων στο περιβάλλον του MOA. Μια διαπίστωση είναι ότι αν κάποιος γνωρίζει το είδος της εννοιολογικής απόκλισης θα μπορούσε να επιλέξει και τους καταλληλότερους αλγόριθμους μάθησης. Οι ανιχνευτές αλλαγών έχουν καλές επιδόσεις σε ροές δεδομένων με απότομες και σταδιακές εννοιολογικές αποκλίσεις όπως και η μέθοδος Leveraging Bagging, αλλά όχι όταν είναι πολύ αργά σταδιακές. Οι διαδικτυακοί αλγόριθμοι (Bagging, Boosting, AUE) αποδίδουν καλά σε εφαρμογές που αντιμετωπίζουν γρήγορες και επαναλαμβανόμενες εννοιολογικές αποκλίσεις. Οι μέθοδοι συνόλου που βασίζονται σε παρτίδες (DWM, AWE, Learn.NSE) παρουσιάζουν καλύτερα αποτελέσματα σε ροές δεδομένων που περιέχουν βαθμιαία σταδιακές εννοιολογικές αποκλίσεις. Η μέθοδος PairerdLearners έχει καλή απόδοση εφόσον οριστεί το σωστό μέγεθος παραθύρου του αντιδραστικού κατηγοριοποιητή. Στην πραγματικότητα όμως, είναι δύσκολο, να γίνει πρόβλεψη του είδους μιας εννοιολογικής απόκλισης, καθώς αλλάζουν τα δεδομένα των ροών γρήγορα και με απρόβλεπτο τρόπο, όπως για παράδειγμα, οι προτιμήσεις των καταναλωτών.

Σε μελλοντική έρευνα θα γίνει η μελέτη των αλγορίθμων συσταδοποίησης ροών δεδομένων. Θα μελετηθούν οι διαχωριστικές μέθοδοι συσταδοποιήσης CluStream, STREAM k-Means και Leader. Η μέθοδος συσταδοποίησης βάσει πυκνότητας DenStream και η μέθοδος DStream (Grid-base method). Επιπλέον, θα γίνει μια πειραματική μελέτη στο περιβάλλον του MOA σε συνθετικά και πραγματικά σύνολα δεδομένων με εννοιολογικές αποκλίσεις. Στη συνέχεια, θα πραγματοποιηθεί μια σύγκριση της απόδοσης τους και θα παρουσιαστούν τα αποτέλεσμα τους.

Κεφάλαιο 9 – Βιβλιογραφία

[1] Nishida, K., & Yamauchi, K. (n.d.). Detecting Concept Drift Using Statistical Testing. In Discovery Science (pp. 264–269). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-75488-6_27

[2] Bifet, A., & Gavaldà, R. (2007). Learning from Time-Changing Data with Adaptive Windowing. In Proceedings of the 2007 SIAM International Conference on Data Mining. Proceedings of the 2007 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611972771.42.

[3] Barros, R. S. M., Cabral, D. R. L., Gonçalves, P. M., Jr., & Santos, S. G. T. C. (2017). RDDM: Reactive drift detection method. In Expert Systems with Applications (Vol. 90, pp. 344–355). Elsevier BV. https://doi.org/10.1016/j.eswa.2017.08.023.

[4] Huang, D. T. J., Koh, Y. S., Dobbie, G., & Pears, R. (2014). Detecting Volatility Shift in Data Streams. In 2014 IEEE International Conference on Data Mining. 2014 IEEE International Conference on Data Mining (ICDM). IEEE. https://doi.org/10.1109/icdm.2014.50

[5] Greg Welch and Gary Bishop (1995). An introduction to the Kalman Filter, Manuscript. Online:https://www.cs.unc.edu/~welch/media/pdf/kalman_intro.pdf

[6] PAGE, E. S. (1954). CONTINUOUS INSPECTION SCHEMES. In Biometrika (Vol. 41, Issues 1–2, pp. 100–115). Oxford University Press (OUP). https://doi.org/10.1093/biomet/41.1-2.100

[7] Baena-Garcıa, M., del Campo-Ávila, J., Fidalgo, R., Bifet, A., Gavalda, R., & Morales-Bueno, R. (2006, September). Early drift detection method. In Fourth international workshop on knowledge discovery from data streams (Vol. 6, pp. 77-86).

[8] Gama, J., Medas, P., Castillo, G., & Rodrigues, P. (2004). Learning with Drift Detection. In Advances in Artificial Intelligence – SBIA 2004 (pp. 286–295). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-28645-5_29

[9] Domingos, P., & Hulten, G. (2000). Mining high-speed data streams. In Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining. KDD00: The Second Annual International Conference on Knowledge Discovery in Data. ACM. https://doi.org/10.1145/347090.347107

[10] Bifet, A., Gavaldà, R., Holmes, G.,, Pfahringer, B. (2018). Machine Learning for Data Streams with Practical Examples in MOA. Cambridge, MA: MIT Press. ISBN: 978-0-262-03779-2

[11] Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. ACM computing surveys (CSUR), 46(4), 1-37.

[12] Bifet, A., Holmes, G., Pfahringer, B., Kranen, P., Kremer, H., Jansen, T., & Seidl, T. (2010, September). Moa: Massive online analysis, a framework for stream classification and clustering. In Proceedings of the first workshop on applications of pattern analysis (pp. 44-50). PMLR.

[13] Bifet, A., & Kirkby, R. (2011). Data Stream Mining: A Practical Approach, University of Waikato, Waikato, New Zealand.

[14] Joao Gama, May (2010). Knowledge Discovery from Data Streams, Chapman and Hall/CRC. ISBN: 9781439826126

[15] Kolter, J. Z., & Maloof, M. A. (2007). Dynamic weighted majority: An ensemble method for drifting concepts. *The Journal of Machine Learning Research*, *8*, 2755-2790

[16] Bifet, A., Holmes, G., Pfahringer, B., & Gavalda, R. (2009, November). Improving adaptive bagging methods for evolving data streams. In Asian conference on machine learning (pp. 23-37). Springer, Berlin, Heidelberg.

17] Bifet, A., Holmes, G., & 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.

[18] Bach, S. H., & Maloof, M. A. (2008, December). Paired learners for concept drift. In 2008 Eighth IEEE International Conference on Data Mining (pp. 23-32). IEEE.

[19] Gama, J., Rocha, R., & Medas, P. (2003, August). Accurate decision trees for mining high-speed data streams. In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 523-528).

[20] Hulten, G., Spencer, L., & Domingos, P. (2001, August). Mining time-changing data streams. In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 97-106).

[21] Brzeziński, D., & Stefanowski, J. (2011). Accuracy updated ensemble for data streams with concept drift. In Hybrid Artificial Intelligent Systems: 6th International Conference, HAIS 2011, Wroclaw, Poland, May 23-25, 2011, Proceedings, Part II 6 (pp. 155-163). Springer Berlin Heidelberg.

[22] Wang, H., Fan, W., Yu, P. S., & Han, J. (2003, August). Mining concept-drifting data streams using ensemble classifiers. In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 226-235).

[23] Stanley, K. O. (2003). Learning concept drift with a committee of decision trees. Informe técnico: UT-AI-TR-03-302, Department of Computer Sciences, University of Texas at Austin, USA.

[24] Bifet, A., Holmes, G., Pfahringer, B., Kirkby, R., & Gavalda, R. (2009, June). New ensemble methods for evolving data streams. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 139-148).

[25] Nyati, A., & Bhatnagar, D. (2016). Performance Evaluation of Anonymized Data Stream Classifiers. International Journal of Computer Science and Network-IJCSN, 5(2).

[26] Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., & Zhang, G. (2018). Learning under concept drift: A review. IEEE transactions on knowledge and data engineering, 31(12), 2346-2363.

[27] Žliobaitė, I., Pechenizkiy, M., & Gama, J. (2016). An overview of concept drift applications. Big data analysis: new algorithms for a new society, 91-114.

[28] Elwell, R., & Polikar, R. (2011). Incremental learning of concept drift in nonstationary environments. IEEE Transactions on Neural Networks, 22(10), 1517-1531.

[29] Bifet, A., & Gavalda, R. (2009). Adaptive learning from evolving data streams. In Advances in Intelligent Data Analysis VIII: 8th International Symposium on Intelligent Data Analysis, IDA 2009, Lyon, France, August 31-September 2, 2009. Proceedings 8 (pp. 249-260). Springer Berlin Heidelberg.

[30] Gonçalves Jr, P. M., & De Barros, R. S. M. (2013). RCD: A recurring concept drift framework. Pattern Recognition Letters, 34(9), 1018-1025.

[31] Lecture notes. Knowledge Discovery in Databases II. Winter Semester 2012/2013. Lectures: PD Dr Matthias Schubert, Dr. Eirini Ntoutsi. Online: https://www.dbs.ifi.lmu.de/Lehre/KDD_II/WS1213/skript/KDD2-1-Introduction.pdf

[32] Bifet, A., de Francisci Morales, G., Read, J., Holmes, G., & Pfahringer, B. (2015, August). Efficient online evaluation of big data stream classifiers. In Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining (pp. 59-68).

[33] Oza, N. C. (n.d.). Online Bagging and Boosting. In 2005 IEEE International Conference on Systems, Man and Cybernetics. 2005 IEEE International Conference on Systems, Man and Cybernetics. IEEE. https://doi.org/10.1109/icsmc.2005.1571498

[34] Bifet, A. (2010). Adaptive stream mining: Pattern learning and mining from evolving data streams (Vol. 207). Ios Press.

[35] Perdikis, T., & Psarakis, S. (2019). A survey on multivariate adaptive control charts: Recent developments and extensions. Quality and Reliability Engineering International, 35(5), 1342-1362.

[36] Bifet, A., Gama, J., Pechenizkiy, M., & Žliobaitė, I. Handling Concept Drift. https://www.cs.waikato.ac.nz/~abifet/PAKDD2011/PAKDD11Tutorial_Handling_Concept_Drift.pdf

[37] Gomes, H. M., Barddal, J. P., Enembreck, F., & Bifet, A. (2017). A survey on ensemble learning for data stream classification. ACM Computing Surveys (CSUR), 50(2), 1-36.

[38] Bifet, A., Kirkby, R., Kranen, P., & Reutemann, P. (2012). Massive Online Analysis Manual https://sourceforge. net/projects/moadatastream/files/documentation. Manual. pdf.

[39] Kadwe, Y., & Suryawanshi, V. (2015). A review on concept drift. Iosr J. Comput. Eng, 17(1), 20-26.

[40] Pfahringer, B., Holmes, G., & Kirkby, R. (2007). New options for hoeffding trees. In AI 2007: Advances in Artificial Intelligence: 20th Australian Joint Conference on Artificial Intelligence, Gold Coast, Australia, December 2-6, 2007. Proceedings 20 (pp. 90-99). Springer Berlin Heidelberg.

[41] Pelossof, R., Jones, M., Vovsha, I., & Rudin, C. (2008). Online Coordinate Boosting (Version 1). arXiv. https://doi.org/10.48550/ARXIV.0810.4553.

[42] Gantz, J. F., Reinsel, D., Chute, C., Schlichting, W., Minton, S., Toncheva, A., & Manfrediz, A. (2008). The expanding digital universe: An updated forecast of worldwide information growth through 2011. International Data Corporation, sponsored by EMC Corporation.

[43]https://www.statista.com/statistics/949144/worldwide-global-datasphere-real-time-data-annual-size/

[44] Barros, R. S. M. de, & Santos, S. G. T. de C. (2019). An overview and comprehensive comparison of ensembles for concept drift. In Information Fusion (Vol. 52, pp. 213–244). Elsevier BV. https://doi.org/10.1016/j.inffus.2019.03.006

[45] Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. In Journal of Computer and System Sciences (Vol. 55, Issue 1, pp. 119–139). Elsevier BV. https://doi.org/10.1006/jcss.1997.1504

[46] Ditzler, G., Roveri, M., Alippi, C., & Polikar, R. (2015). Learning in Nonstationary Environments: A Survey. In IEEE Computational Intelligence Magazine (Vol. 10, Issue 4, pp. 12–25). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/mci.2015.2471196

[47] Aggarwal, C. C. (Ed.). (2007). Data streams: models and algorithms (Vol. 31). New York: Springer.

[48] Bifet, A., Holmes, G., Pfahringer, B., & Frank, E. (2010). Fast perceptron decision tree learning from evolving data streams. In Advances in Knowledge Discovery and Data Mining: 14th Pacific-Asia Conference, PAKDD 2010, Hyderabad, India, June 21-24, 2010. Proceedings. Part II 14 (pp. 299-310). Springer Berlin Heidelberg.

[49] Bayram, F., Ahmed, B. S., & Kassler, A. (2022). From concept drift to model degradation: An overview on performance-aware drift detectors. In Knowledge-Based Systems (Vol. 245, p. 108632). Elsevier BV. https://doi.org/10.1016/j.knosys.2022.108632

[50] Stephanie Glen. "Cohen's Kappa Statistic" From StatisticsHowTo.com: Elementary Statistics for the rest of us! https://www.statisticshowto.com/cohens-kappa-statistic/

[51] Mitchell, T. M., & Mitchell, T. M. (1997). Machine learning (Vol. 1, No. 9). New York: McGraw-hill.

[52] Gaber, M. M., Zaslavsky, A., & Krishnaswamy, S. (2005). Mining data streams: a review. *ACM Sigmod Record*, *34*(2), 18-26.

[53] João Gama, Pedro Medas, Gladys Castillo and Pedro Pereira Rodrigues. Learning with Drift Detection. In Bazzan, Ana L. C. and Labidi, Sofiane, editors, Advances in Artificial Intelligence - SBIA 2004, volume 3171 of Lecture Notes in Computer Science, pages 286-295. Springer Berlin / Heidelberg, 2004. ISBN 978-3-540-23237-7. URL http://dx.doi.org/10.1007/978-3-540-28645-5_29.