ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

Π.Μ.Σ. «ΨΗΦΙΑΚΑ ΣΥΣΤΗΜΑΤΑ ΚΑΙ ΥΠΗΡΕΣΙΕΣ»

ΚΑΤΕΥΘΥΝΣΗ: ΜΕΓΑΛΑ ΔΕΔΟΜΕΝΑ ΚΑΙ ΑΝΑΛΥΤΙΚΗ

Αυξητική μέθοδος ανάλυσης συναισθήματος σε περιβάλλον microblogging

Διπλωματική Εργασία του

Πικούνη Κωνσταντίνου

Επιβλέπουσα:

Αν. Καθηγήτρια Χαλκίδη Μαρία

Περίληψη

Τα εργαλεία μηχανικής μάθησης που στοχεύουν στην ανάλυση του συναισθήματος κειμένων, είναι από τα πλέον καινοτόμα εργαλεία μηχανικής μάθησης. Για την εκπαίδευση τέτοιων εργαλείων, συνήθως απαιτούνται πολύ μεγάλα datasets τα οποία μπορεί να μην είναι διαθέσιμα στην πληρότητά τους από την αρχή, αλλά να δημιουργούνται σταδιακά. Μία νέα και καινοτόμα προσέγγιση των εργαλείων ανάλυσης συναισθήματος, η οποία μελετάται στην παρούσα διπλωματική εργασία, είναι αυτά που εκπαιδεύονται με αυξητικό τρόπο.

Η παρούσα διπλωματική εργασία αποτελείται από τέσσερα κεφάλαια. Στο Κεφάλαιο 1 αναφέρονται οι κυριότερες μέθοδοι για την αναπαράσταση κειμένου με αριθμούς (διανύσματα), καθώς και η μέθοδος που επιλέχθηκε. Στο Κεφάλαιο 2 παρουσιάζονται διάφοροι ταξινομητές που λειτουργούν με αυξητικό τρόπο. Αυτοί μελετήθηκαν σε βάθος, ορίστηκαν κριτήρια για την επιλογή του βέλτιστου και βάσει αυτών, βρέθηκε ο βέλτιστος ταξινομητής για την κατηγοριοποίηση μικρών κειμένων (περιβάλλον microblogging) με βάση το συναίσθημά τους, ο οποίος λειτουργεί με αυξητικό τρόπο. Στο Κεφάλαιο 3 παρουσιάζεται εις βάθος μελέτη του βέλτιστου ταξινομητή. Ο ταξινομητής αυτός υποβλήθηκε σε διάφορους ελέγχους και η απόδοση ταξινόμησής του ελέγχθηκε σε διαφορετικές περιπτώσεις. Τέλος, δημιουργήθηκε ένα ολοκληρωμένο εργαλείο για την ταξινόμηση μικρών κειμένων με βάση το συναίσθημά τους, το οποίο λειτουργεί με αυξητικό τρόπο και βασίστηκε στον επιλεγμένο ταξινομητή. Το Κεφάλαιο 4 αποτελεί μια σύνοψη και την κατακλείδα της παρούσας διπλωματικής εργασίας.

Abstract

Machine learning tools that perform sentiment analysis on texts is one of the most innovative branches of machine learning. To train such tools, large datasets are usually required. In some cases, these datasets are fully available in their entirety, but in other, more realistic cases, the data can be accessed incrementally, as they are created. In order to utilize such datasets as well as perform analyses on real data without having access to the whole dataset from the beginning, machine learning tools that can be trained incrementally are used. The focus of this thesis is this novel and innovative branch of sentiment analysis machine learning tools that can be trained incrementally.

This master thesis aims to find the optimal classifier, which can be trained incrementally, for classifying short texts based on their sentiment. The thesis is divided into four chapters. In Chapter 1 the most common methods to represent text into vectors are discussed as well as the one chosen for this analysis. In Chapter 2 various classifiers that can be trained incrementally are detailed. Five criteria are devised, and the best classifier is identified. In Chapter 3 a detailed analysis of the optimal classifier is conducted. That classifier is put through various tests and its behavior is studied. Finally, an integrated tool based on the chosen classifier, is created in order to classify short texts based in their sentiment. Chapter 4 is the conclusion of this thesis.

Περιεχόμενα

1 Αναπαράσταση κειμένου

1.1. Ανάλυση συναισθήματος

Ένα από τα βασικά χαρακτηριστικά της περιόδου που διανύουμε είναι ότι τα μέσα κοινωνικής δικτύωσης αποτελούν εργαλείο, μέσω του οποίου ο κάθε χρήστης εκφράζει τη γνώμη του, ενημερώνεται αλλά εκφράζεται σχετικά με προϊόντα και υπηρεσίες. Επομένως, είναι πολύ σημαντικό να γίνεται με γρήγορο και εύκολο τρόπο εξαγωγή πληροφοριών σχετικά με τα συναισθήματα που δημιουργεί ένα προϊόν, μία υπηρεσία ή ακόμα και κάποιο γεγονός. Σε αυτό στοχεύουν οι τεχνικές ανάλυσης συναισθήματος, οι οποίες είναι τεχνικές μηχανικής μάθησης που αναλύουν πληθώρα δεδομένων με στόχο την εξαγωγή αποτελεσμάτων σχετικά με τα συναισθήματα που εκφράζονται σε ένα κείμενο. Συνεπώς, η ανάλυση συναισθήματος θεωρείται ένα πολύ καινοτόμο εργαλείο, το οποίο μπορεί να συνδυάσει και να διαχειριστεί πληροφορίες που οι χρήστες αντάλλαξαν σε διαφορετικά σημεία επαφής, για παράδειγμα σε μέσα κοινωνικής δικτύωσης, στην εξυπηρέτηση πελατών εταιριών ή σε ιστοσελίδες με κριτικές.

Η ανάλυση συναισθήματος αντιμετωπίζει και αρκετές προκλήσεις που σχετίζονται με τον τρόπο έκφρασης των ανθρώπων. Χαρακτηριστικά παραδείγματα είναι η ειρωνεία και ο σαρκασμός σε κάποιο σχόλιο, το τι μπορεί να θεωρηθεί ουδέτερη απόκριση, τα υπονοούμενα που προκύπτουν από τα συμφραζόμενα καθώς και η αντικειμενικότητα κάποιας άποψης. Όλα αυτά αποτελούν σχεδόν έμφυτες εκφάνσεις της επικοινωνίας των ανθρώπων και τα περισσότερα είναι πλήρως κατανοητά ακόμα και από μικρά παιδιά. Παρ' όλα αυτά, αποτελούν προκλήσεις για την ανάλυση συναισθήματος από υπολογιστές.

Για την δημιουργία αξιόπιστων εργαλείων ανάλυσης συναισθήματος, συνήθως απαιτούνται πολύ μεγάλα datasets με παρόμοια κείμενα με αυτά τα οποία θα αξιολογηθούν. Βέβαια η δημιουργία τέτοιων datasets απαιτεί πολλούς πόρους, όπως επίσης και πολύ μεγάλο πλήθος χαρακτηρισμένων κειμένων. Τέτοιες συλλογές κειμένων μπορεί και να μην είναι διαθέσιμες ολόκληρες από την πρώτη στιγμή, αλλά να δημιουργούνται σταδιακά. Μία νέα και καινοτόμα προσέγγιση, που στοχεύει στην αντιμετώπιση αυτού του προβλήματος, είναι η αξιοποίηση εργαλείων ανάλυσης συναισθήματος που εκπαιδεύονται με αυξητικό τρόπο. Τα εργαλεία αυτά αξιοποιούν τα διαθέσιμα δεδομένα και εκπαιδεύονται με αυτά, ενώ όταν υπάρξουν επιπλέον δεδομένα μπορούν να επανεκπαιδευτούν μόνο με το νέο δείγμα. Με τον τρόπο αυτό ενισχύουν την απόδοσή τους αξιοποιώντας τις νέες πληροφορίες, χωρίς όμως να «ξεχνούν» τις παλιές.

Η παρούσα διπλωματική εργασία έχει ως στόχο την εύρεση εργαλείου μηχανικής μάθησης με το οποίο θα είναι δυνατή η ταξινόμηση μικρών κειμένων (περιβάλλον microblogging) με βάση το συναίσθημά τους, και θα εκπαιδεύεται με αυξητικό τρόπο.

1.2. Εισαγωγή στην αναπαράσταση κειμένου

Για να είναι δυνατή η επεξεργασία κειμένου από υπολογιστές, συνήθως πρέπει το κείμενο να μετασχηματιστεί σε κάποια μορφή που οι υπολογιστές μπορούν να διαχειριστούν, συνεπώς οι λέξεις «μεταφράζονται» σε αριθμούς ή σε διανύσματα. Στο σημείο αυτό όμως υπεισέρχεται το πρόβλημα ότι ίδιες λέξεις χρησιμοποιούνται για να εκφράσουν διαφορετικά νοήματα. Για τους ανθρώπους η κατανόηση της έννοιας της λέξης είναι (συνήθως) προφανής από τα συμφραζόμενα της πρότασης. Για παράδειγμα στις ακόλουθες προτάσεις η λέξη αρχή έχει διαφορετικά νοήματα:

- Σύμφωνα με την αρχή της απροσδιοριστίας του Heisenberg … – αρχή = κανόνας
- Η κατ' αρχήν ψήφιση του νομοσχεδίου … – αρχή = βασικά σημεία
- Η αρχή του καυγά ήταν η έλλειψη εμπιστοσύνης … – αρχή = πρώτη αιτία
- Στην αρχή του έργου ο σκηνοθέτης … - αρχή = ξεκίνημα
- Στην αρχή δεν τον πίστεψα, μετά όμως … – αρχή = πρώτη φορά
- Η δικαστική αρχή της Καλαμάτας … αρχή = εξουσία

Συνεπώς η αντιστοίχιση της έννοιας της λέξης σε έναν αριθμό δεν είναι κάτι μονοσήμαντο διότι για πολλές λέξεις η έννοιά τους εξαρτάται από τον τρόπο με τον οποίο χρησιμοποιούνται στην πρόταση.

Γενικά υπάρχουν διάφορες προσεγγίσεις για τον τρόπο με τον οποίο οι λέξεις ενός κειμένου μπορούν να «μεταγλωττιστούν» σε αριθμούς διατηρώντας την έννοιά τους. Η παλιότερη βασίζεται στην χρήση λεξικών και κανόνων σύνταξης. Με χρήση των κανόνων σύνταξης είναι δυνατή η αντιστοίχιση λέξεων σε μέρη του λόγου (ιδίως στα αγγλικά όπου είναι δυνατόν η ίδια λέξη, ανάλογα με τη χρήση της και τη θέση της στην πρόταση να είναι διαφορετικό μέρος του λόγου, παραδείγματος χάριν η λέξη break – ουσιαστικό: διάλειμμα ή ρήμα: σπάζω) ενώ με τη χρήση λεξικών επιτυγχάνεται η αντιστοίχισή τους σε έννοιες.

Μια άλλη προσέγγιση είναι αυτή που χρησιμοποιεί εργαλεία επιβλεπόμενης μάθησης (supervised learning). Τα εργαλεία αυτά εξάγουν από το κείμενο το οποίο τους διατίθεται, και μόνο από αυτό, το νόημα της κάθε λέξης. Οι μηχανές διανυσμάτων στήριξης και τα LSTM (Long Short Term Memory) νευρωνικά δίκτυα είναι από τα εργαλεία επιβλεπόμενης μάθησης που έχουν καλά αποτελέσματα στον τομέα αυτό.

Εργαλεία μη επιβλεπόμενης μάθησης (unsupervised learning) αποτελούν μια ακόμα προσέγγιση. Κάποια από αυτά στοχεύουν στη συσταδοποιήση των λέξεων κειμένου, η οποία βασίζεται στην λογική ότι λέξεις με παρόμοιες έννοιες χρησιμοποιούνται με παρόμοιους τρόπους, συνεπώς μπορούν να συσταδοποιηθούν. Μια άλλη κατηγορία εργαλείων μη επιβλεπόμενης μάθησης αποτελούν τα word embeddings [1] [2]. Τα εργαλεία αυτά αντιστοιχίζουν τις λέξεις κειμένου σε πυκνά διανύσματα αριθμών αναπαριστώντας τις λέξεις σε ένα συνεχή διανυσματικό χώρο, όπου η κάθε μία βρίσκεται κοντά σε άλλες που έχουν παραπλήσιο νόημα. Με την λέξη πυκνά εννοούμαι ότι στα διανύσματα δεν υπάρχουν πολλά «0» όπως στην περίπτωση της εξαγωγής πληροφορίας κειμένου με την μέθοδο της συχνότητας εμφάνισης της κάθε λέξης στο κείμενο – αντίστροφης συχνότητας κειμένου (TF-IDF text frequency – inverted document frequency). Στα διανύσματα αυτά αποτυπώνεται τόσο η έννοια των λέξεων όσο και η σύνταξη τους και γι' αυτό τα word embeddings αποτελούν μια από τις πιο καινοτόμες προσεγγίσεις της αναπαράστασης κειμένου.

1.3. Word embeddings

Για την εύρεση του κατάλληλου word embeddings εργαλείου εξετάστηκαν τα μοντέλα word2vec [2] και GloVe (Global Vectors for word representation) [3]. Το πρώτο αποτελεί υλοποίηση ενός προβλεπτικού μοντέλου, δηλαδή ενός μοντέλου που στοχεύει στην πρόβλεψη της λέξης από τις γειτονικές της, ενώ το δεύτερο αποτελεί υλοποίηση στατιστικού μοντέλου, δηλαδή μοντέλου που υπολογίζει την πιθανότητα της εμφάνισης κάποιας λέξης βασιζόμενο στις γειτονικές τις. Και οι δυο προσεγγίσεις έχουν ως αποτέλεσμα την αντιστοίχιση των λέξεων σε πυκνά διανύσματα, και η απόδοσή τους αυξάνει όταν τα κείμενα με τα οποία εκπαιδεύονται είναι μεγάλα.

Το μοντέλο word2vec που εξετάστηκε είναι υλοποιημένο στο πακέτο TensorFlow [4] (σε γλώσσα python). Στο μοντέλο αυτό χρησιμοποιείται ένα νευρωνικό δίκτυο με δύο επίπεδα (layers) το οποίο εκπαιδεύεται ώστε να αναπαριστά την έννοια των λέξεων από τις γειτονικές τους χρησιμοποιώντας το μοντέλο Skip-gram. Το μοντέλο GloVe είναι υλοποιημένο σε γλώσσα C++ και μπορεί να βρεθεί στα [5] και [6].

Για τον πειραματισμό με τα δυο προαναφερθέντα μοντέλα ως δείγμα εκπαίδευσης χρησιμοποιήθηκε το dataset text8 [7]. Αυτό το dataset αποτελεί συλλογή κειμένων από το διαδίκτυο σε αγγλική γλώσσα το όποιο αποτελείται από 10^8 χαρακτήρες, δηλαδή από 10^8 bytes ή 100 MB, και από περίπου 17.000.000 λέξεις. Το μέγεθος αυτού του dataset θεωρείται μικρό για εκπαίδευση μοντέλων word embeddings και συνήθως χρησιμοποιείται ως test και όχι ως training sample. Παρόλα αυτά εδώ χρησιμοποιήθηκε ως ένα υποτυπώδες training sample για τον πειραματισμό με τα μοντέλα. Κατά την εκπαίδευση των μοντέλων, το dataset επεξεργάζεται πολλές φορές διαδοχικά (epochs) με στόχο την βελτιστοποίηση των αποτελεσμάτων.

Για την εκτίμηση της απόδοσης της κάθε μεθόδου χρησιμοποιείται μια σειρά από σημασιολογικούς και γραμματικούς ελέγχους. Ακολουθούν δύο παραδείγματα τέτοιων ελέγχων ενός σημασιολογικού (1) και ενός γραμματικού (2):

1) boy girl king queen
2) bad worse big bigger

Κατά τον έλεγχο, από το embedding της δεύτερης λέξης αφαιρείται αυτό της πρώτης και στο αποτέλεσμα προστίθεται αυτό της τρίτης δηλαδή:

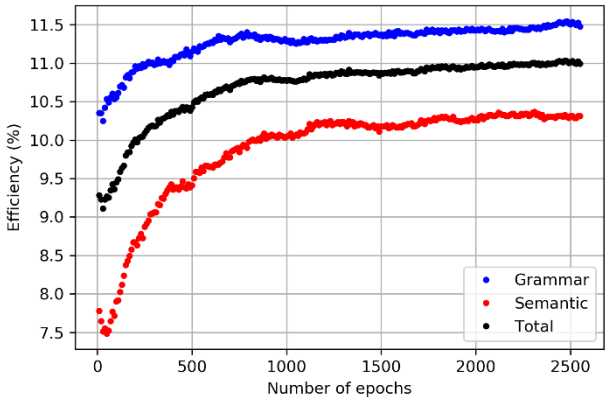
1) [girl] – [boy] + [king] [queen]
2) [worse] – [bad] + [big] [bigger]

Το διάνυσμα που βρίσκεται με αυτόν το τρόπο θα πρέπει να συμπίπτει με το embedding της τέταρτης λέξης. Συνεπώς, βρίσκεται η απόσταση του πιο πάνω διανύσματος με τα embeddings όλων των διαθέσιμων λέξεων και αν αυτό με τη μικρότερη απόσταση είναι αυτό της τέταρτης λέξης τότε το μοντέλο επιτυγχάνει, αλλιώς αποτυγχάνει. Ως απόδοση ορίζεται ο λόγος των επιτυχιών προς τον συνολικό αριθμό των ελέγχων και εκφράζεται είτε ως αριθμός μικρότερος της μονάδας είτε ως ποσοστό.

Δύο από τις σημαντικότερες παραμέτρους των μοντέλων που επηρεάζουν την απόδοση αλλά και τον χρόνο επεξεργασίας είναι ο αριθμός των στοιχείων (διάσταση) των διανυσμάτων των word embeddings και ο ρυθμός εκμάθησης (learning rate – lr). Εν γένει

όταν αυξάνει η διάσταση των διανυσμάτων, η απόδοση βελτιώνεται και γι' αυτό υπάρχουν έτοιμα εκπαιδευμένα μοντέλα word embeddings με διάσταση διανυσμάτων από 50 μέχρι και πάνω από 300. Σχετικά με τον ρυθμό εκμάθησης, οι υλοποιήσεις και των δυο μοντέλων που εξετάστηκαν έχουν ως προεπιλογή μια μέθοδο δυναμικής αναπροσαρμογής του ρυθμού εκμάθησης η οποία μεταβάλει τον ρυθμό εκμάθησης σε κάθε epoch σύμφωνα με μία αρχική τιμή που δίνεται από τον χρήστη και με τον συνολικό αριθμό των epochs της εκπαίδευσης.

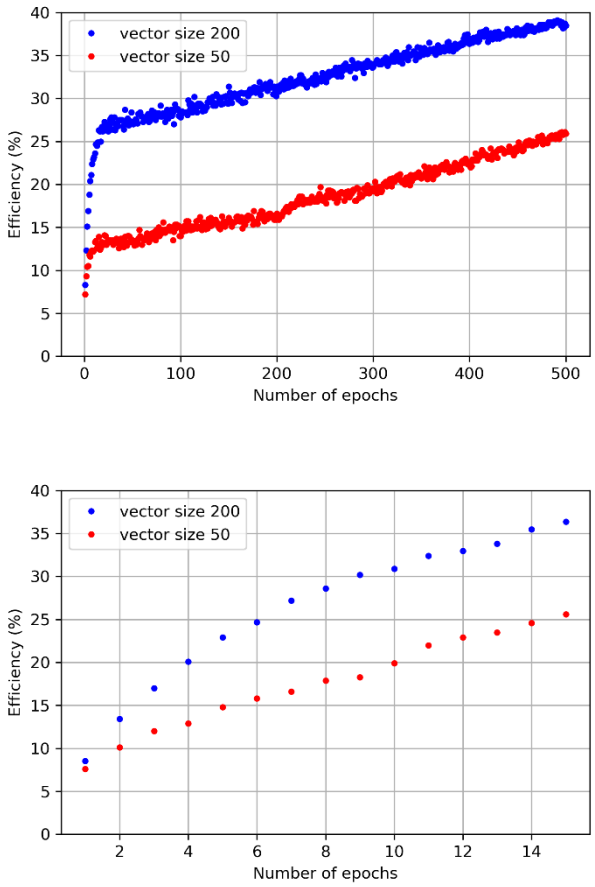
Στην **Εικόνα 1** παρουσιάζεται η απόδοση των word embeddings από το μοντέλο GloVe, ως συνάρτηση του αριθμού των epochs, όπως υπολογίζεται από τους σημασιολογικούς ελέγχους (κόκκινο), τους γραμματικούς ελέγχους (μπλε) και από το σύνολό τους (μαύρο). Κατά την εκπαίδευση του μοντέλου GloVe χρησιμοποιήθηκε διάσταση διανυσμάτων 50 και ο προεπιλεγμένος αρχικός ρυθμός μάθησης (eta = 0,05). Ο αριθμός των epochs ήταν περίπου 2550 και αντιστοιχούσε σε υπολογιστικό χρόνο περίπου 30 ωρών σε 8 πυρήνες ταυτόχρονα. Η εκπαίδευση του μοντέλου εκτελούταν σε 8 πυρήνες, η προεπεξεργασία των δεδομένων (πριν από κάθε epoch) όπως επίσης και οι έλεγχοι εκτελούταν σε 1 πυρήνα. Όπως φαίνεται από την εικόνα η απόδοση είναι καλύτερη στους γραμματικούς ελέγχους από ότι τους σημασιολογικούς. Επίσης με την πάροδο των epochs η απόδοση αρχικά αυξάνει γρήγορα ενώ στην συνέχεια συνεχίζει να έχει αυξητική πορεία αλλά με μικρότερη κλίση. Στο σημείο αυτό πρέπει να τονιστεί ότι δεν αφιερώθηκε καθόλου χρόνος στην εύρεση των βέλτιστων παραμέτρων αλλά το μοντέλο GloVe χρησιμοποιήθηκε με τις προεπιλεγμένες τιμές των μεταβλητών.



***Εικόνα 1**: Απόδοση των word embeddings από το μοντέλο GloVe συναρτήσει του αριθμού των epochs. Η διάσταση των διανυσμάτων είναι 50. Με μπλε χρώμα παρουσιάζεται η απόδοση όπως υπολογίστηκε από τους γραμματικούς ελέγχους, με κόκκινο από τους σημασιολογικούς ενώ με μαύρο από όλους.*

Στην **Εικόνα 2** παρουσιάζεται η απόδοση των word embeddings από το μοντέλο word2vec, ως συνάρτηση του αριθμού των epochs, η οποία υπολογίζεται από το σύνολο των ελέγχων (σημασιολογικών και γραμματικών). Με κόκκινο χρώμα παρουσιάζεται η απόδοση όταν η διάσταση των διανυσμάτων είναι 50 ενώ με μπλε όταν η διάστασή τους είναι 200. Στο άνω διάγραμμα παρουσιάζεται η απόδοση όταν ο αριθμός των epochs εκπαίδευσης ήταν 500 και αντιστοιχούσε σε υπολογιστικό χρόνο περίπου 30 ωρών σε 8 πυρήνες ταυτόχρονα (όπου και σε αυτή την περίπτωση 8 πυρήνες χρησιμοποιούνταν ταυτόχρονα κατά την

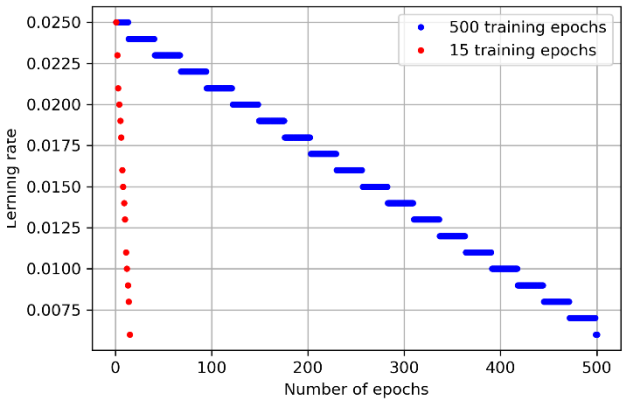
εκπαίδευση του μοντέλου, ενώ κατά την προεπεξεργασία των δεδομένων και κατά τους ελέγχους χρησιμοποιούταν 1 πυρήνας). Στο κάτω διάγραμμα παρουσιάζεται η απόδοση όταν ο αριθμός των epochs εκπαίδευσης ήταν 15.



***Εικόνα 2**: Απόδοση των word embeddings από το μοντέλο word2vec όπως υπολογίζεται από το σύνολο των ελέγχων (σημασιολογικών και γραμματικών) συναρτήσει του αριθμού των epochs. Με κόκκινο χρώμα παρουσιάζεται η απόδοση όταν η διάσταση των διανυσμάτων είναι 50 ενώ με μπλε όταν η διάσταση είναι 200. Άνω διάγραμμα: 500 epochs εκπαίδευσης, κάτω διάγραμμα 15 epochs εκπαίδευσης.*

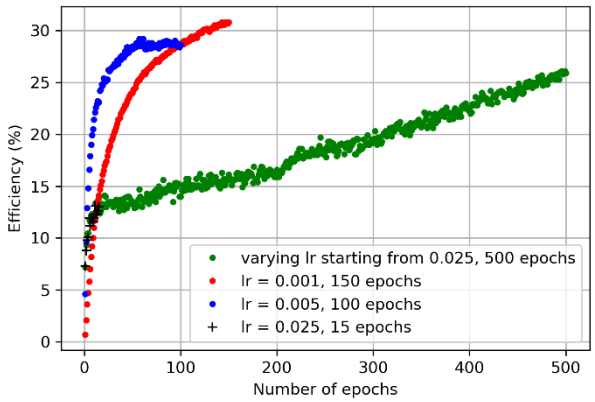
Συγκρίνοντας την **Εικόνα 1** και την **Εικόνα 2** άνω, φαίνεται ότι η απόδοση του μοντέλου word2vec είναι καλύτερη από αυτή του GloVe. Στο ίδιο υπολογιστικό χρόνο εκπαίδευσης όταν η διάσταση των διανυσμάτων είναι 50, με το word2vec επιτυγχάνεται απόδοση ~25% ενώ με το GloVe ~11%. Για το μοντέλο word2vec φαίνεται ότι με την αύξηση της διάστασης των διανυσμάτων, η απόδοση ανεβαίνει από ~25% με χρήση διανυσμάτων διάστασης 50 σε 35% με χρήση διανυσμάτων διάστασης 200. Επίσης παρατηρείται μια απότομη άνοδος της απόδοσης κατά τις πρώτες λίγες epochs (από ~7% σε ~12% στις πρώτες 6 για διάσταση διανυσμάτων 50 και από ~8% σε ~26% στις πρώτες 15 για διάσταση διανυσμάτων 200) ενώ στην συνέχεια η απόδοση συνεχίζει να αυξάνει με πιο αργό ρυθμό. Όπως έχει ήδη αναφερθεί η απόδοση του μοντέλου εξαρτάται, μεταξύ άλλων, από τη διάσταση των διανυσμάτων αλλά και από το ρυθμό εκμάθησης, ο οποίος μεταβάλλεται σε

κάθε epoch εξαρτώμενος από την αρχική τιμή που δίνεται από τον χρήστη και από τον αριθμό των epochs εκπαίδευσης. Στο κάτω διάγραμμα στην **Εικόνα 2** παρουσιάζεται η απόδοση των word embeddings από το μοντέλο word2vec, ως συνάρτηση του αριθμού των epochs, όταν ο αριθμός των epochs ορίστηκε σε 15. Συγκρίνοντας τα δυο διαγράμματα στην **Εικόνα 2** φαίνεται ότι η αρχική και η τελική απόδοση συμπίπτουν ανεξαρτήτως του αριθμού των epochs. Στην **Εικόνα 3** παρουσιάζεται ο ρυθμός εκμάθησης του μοντέλου word2vec ως συνάρτηση του αριθμού των epochs. Με μπλε χρώμα για τη περίπτωση που ο αριθμός των epochs εκπαίδευσης ορίστηκε στις 500 ενώ με κόκκινο στην περίπτωση που ορίστηκε ίσος με 15. Και στις δύο περιπτώσεις η αρχική τιμή του ρυθμού εκμάθησης ήταν 0,025.



***Εικόνα 3**: Ρυθμός εκμάθησης του μοντέλου word2vec ως συνάρτηση του αριθμού των epochs. Με μπλε παρουσιάζεται ο ρυθμός εκμάθησης όταν ο συνολικός αριθμός των epochs ορίστηκε σε 500 ενώ με κόκκινο σε 15.*

Το γεγονός ότι είτε εκτελεστούν 15 είτε 500 epochs εκπαίδευσης τα τελικά αποτελέσματα είναι συγκρίσιμα, αποτελεί αιτία προβληματισμών σχετικά με τον τρόπο χρήσης της μεθόδου αυτής και κατ' επέκταση και της μεθόδου GloVe, όπως επίσης και με τον τρόπο χρήσης του ρυθμού εκμάθησης. Για τον λόγο αυτό εκτελέστηκαν πειραματισμοί αλλάζοντας την αρχική τιμή του ρυθμού εκμάθησης αλλά και σταθεροποιώντας την τιμή του για όλες τις epochs εκπαίδευσης. Στην **Εικόνα 4** παρουσιάζονται τα αποτελέσματα αυτών των πειραματισμών. Από την εικόνα αυτή φαίνεται ότι για μικρότερο ρυθμό εκμάθησής στις πρώτες epochs η αύξηση της απόδοσης είναι μικρότερη. Βέβαια μετά την αρχική «απότομη» αύξηση τη απόδοσης όταν ο ρυθμός απόδοσης είναι 0,005 παρατηρείται σταθεροποίηση της, ενώ όταν ο ρυθμός απόδοσης είναι 0,001 η αρχική αύξηση είναι λιγότερο «θεαματική» αλλά δεν παρατηρείται σταθεροποίησή της απόδοσης ακόμα και μετά από 150 epochs. Το ασφαλέστερο συμπέρασμα που μπορεί να εξαχθεί από τη μελέτη που έχει διεξαχθεί μέχρι στιγμής είναι ότι απαιτείται περαιτέρω έρευνα και εις βάθος μελέτη όλων των παραμέτρων των μοντέλων.



***Εικόνα 4**: Απόδοση των word embeddings από το μοντέλο word2vec για διάσταση διανυσμάτων 50 όπως υπολογίζεται από το σύνολο των ελέγχων (σημασιολογικών και γραμματικών) συναρτήσει του αριθμού των epochs. Πράσινο: μεταβλητός ρυθμός εκμάθησης που ξεκινά από 0,025 και αριθμός epochs 500. Κόκκινο: σταθερός ρυθμός εκμάθησης 0,005. Μπλε: σταθερός ρυθμός εκμάθησης 0,001. Μαύρο: σταθερός ρυθμός εκμάθησης 0,025.*

1.4. Επιλογή μοντέλου αναπαράστασης κειμένου

Σε αυτή την ενότητα εξετάστηκαν και μελετήθηκαν τα μοντέλα word embeddings glove και word2vec. Από τη μελέτη που έχει διεξαχθεί μέχρι στιγμής φαίνεται ότι απαιτείται μελέτη εις βάθος των μοντέλων με στόχο την βελτιστοποίηση των παραμέτρων τους. Επιπλέον, όπως ήδη αναφέρθηκε, για την βέλτιστή απόδοσή τους απαιτείται το σύνολο των κειμένων εκπαίδευσης να είναι πολύ μεγάλο (αυτό που χρησιμοποιήθηκε ως δείγμα εκπαίδευσης στους πειραματισμούς και αποτελείται από μόλις 17 εκατομμύρια λέξεις συνήθως χρησιμοποιείται ως test sample). Βέβαια στο διαδίκτυο υπάρχουν διαθέσιμα μοντέλα ήδη εκπαιδευμένα. Πιο συγκεκριμένα στην ηλεκτρονική τοποθεσία [5] είναι διαθέσιμα διάφορα, ήδη εκπαιδευμένα, word embeddings με το μοντέλο GloVe. Ένα από αυτά είναι το glove6B. Η εκπαίδευση έγινε με συλλογή κειμένων που αποτελούνται συνολικά από 6 δισεκατομμύρια λέξεις (στις οποίες όλα τα κεφαλαία γράμματα είχαν μετατραπεί σε μικρά). Το αποτέλεσμα είναι word embeddings 400.000 λέξεων με διάσταση διανυσμάτων 50, 100, 200 και 300. Η απόδοση των μοντέλων παρουσιάζεται στον **Πίνακας 1**. Η ποσοστιαία αύξηση της απόδοσης που επιτυγχάνεται από τη χρήση διανυσμάτων διάστασης 100 σε σχέση με διανύσματα με διάσταση 50 είναι μεγαλύτερη από 35%, από τη χρήση διανυσμάτων διάστασης 200 σε σχέση με διανύσματα με διάσταση 100 είναι της τάξης του 10% ενώ από τη χρήση διανυσμάτων διάστασης 300 σε σχέση με διανύσματα με διάσταση 200 είναι της τάξης του 3%.

Διάσταση διανυσμάτων	Απόδοση		
	σημειολογική	γραμματική	ολική
50	48,46%	44,36%	46,22%
100	65,34%	61,26%	63,11%
200	74,13%	66,15%	69,77%
300	77,44%	67,00%	71,74%

***Πίνακας 1**: Απόδοση των pretrained word embeddings glove6B.*

Ο τελικός στόχος της παρούσας εργασία, είναι η εύρεση εργαλείου το οποίο θα εκτελεί ανάλυση συναισθήματος σε μικρά κείμενα και θα λειτουργεί με αυξητικό τρόπο. Τα word embedding αποτελούν το πρώτο βήμα, δηλαδή τη μετατροπή του κειμένου σε τέτοια μορφή ώστε να είναι εύκολα επεξεργάσιμο από εργαλεία μηχανικής μάθησης. Συνυπολογίζοντας λοιπόν την απόδοση των word embeddings που παρουσιάζονται στον **Πίνακας 1** με το γεγονός ότι με την αύξηση της διάστασης των διανυσμάτων που δίνονται ως είσοδος σε εργαλεία μηχανικής μάθησης η απόδοσή τους (σε αρκετά από αυτά) ελαττώνεται, αποφασίστηκε η χρήση του του pretrained globe6B μοντέλου με διάσταση διανύσματος 100.

2 Αναπαράσταση κειμένου

2.1. Γενικά

Ο στόχος της παρούσας διπλωματικής εργασίας είναι η εύρεση εργαλείου για την ταξινόμηση μικρών κειμένων (περιβάλλον microblogging) με βάση το συναίσθημα, το οποίο να λειτουργεί με αυξητικό τρόπο. Το εργαλείο αυτό μπορεί να είναι είτε κάποιος ταξινομητής είτε συλλογή ταξινομητών (ensemble of classifiers). Οι ταξινομητές και τα εργαλεία που εξετάστηκαν περιγράφονται στο παρόν κεφάλαιο.

Για την εύρεση του καταλληλότερου εργαλείου χρησιμοποιήθηκε το Dataset «Large Movie Review Dataset v1.0» [8] στο οποίο εμπεριέχονται 50.000 σύντομες κριτικές από ταινίες από το IMDB [9]. Οι κριτικές είναι χωρισμένες σε 25.000 θετικές (που έχουν λάβει βαθμό από 7 έως 10) και σε 25.000 αρνητικές (που έχουν λάβει κριτικές από 1 έως 4) ενώ δεν περιέχονται ουδέτερες κριτικές με βαθμούς 5 και 6. Το κείμενο της κάθε κριτικής βρίσκεται σε ξεχωριστό αρχείο, στον τίτλο του οποίου υπάρχει η βαθμολόγια που συνοδεύει την κριτική. Στο παρόν τμήμα της ανάλυσης (για την εύρεση του βέλτιστου ταξινομητή) δεν έχει ληφθεί υπόψιν ο βαθμός της κριτικής αλλά μόνο αν η κριτική είναι θετική ή αρνητική. Ο αριθμός των κριτικών που αφορούν στην ίδια ταινία μπορεί να είναι το πολύ μέχρι 30, διότι έχει παρατηρηθεί ότι οι κριτικές για την ίδια ταινία τείνουν να έχουν παρόμοιες βαθμολογίες. Τα training και test samples αποτελούνται από 25.000 παραδείγματα ισομοιρασμένα σε 12.500 παραδείγματα με θετική κριτική και σε 12.500 με αρνητική. Επίσης οι κριτικές που υπάρχουν στο training και στο test set αφορούν σε διαφορετικές ταινίες με σκοπό να μην είναι δυνατή η πρόβλεψη του αποτελέσματος λόγω «απομνημόνευσης» λέξεων που σχετίζονται αποκλειστικά με κάποια ταινία. Στο ίδιο dataset εμπεριέχονται και 50.000 κριτικές, στις οποίες δεν έχει καταγραφεί ο βαθμός, με σκοπό την χρήση τους σε εργαλεία που εκτελούν unsupervised learning. Αυτές οι κριτικές δεν χρησιμοποιήθηκαν στην παρούσα εργασία.

Για την μετατροπή του κειμένου σε διανύσματα χρησιμοποιήθηκαν τα pre-trained word embeddings vectors με διάσταση 100 με το μοντέλο GloVe, glove6B_100, όπως αναφέρθηκε στην ενότητα 1.4. Πριν από το στάδιο της μετατροπής των λέξεων σε διανύσματα, προηγήθηκε προεπεξεργασία του κειμένου κάθε κριτικής. Το πρώτο στάδιο της προεπεξεργασίας ήταν η ανάπτυξη των συγκοπτόμενων τύπων (για παράδειγμα η λέξη can’t μετατράπηκε στις λέξεις can not) και επιτεύχθηκε με το εργαλείο επεξεργασίας κειμένου Ekphrasis [10]. Στη συνέχεια το κείμενο κατατμήθηκε σε λεξικογραφικές μονάδες (λέξεις, σημεία στίξεις, κλπ) χρησιμοποιώντας το πακέτο Natural Language Toolkit (nlkt) [11]. Από τις λεξικογραφικές μονάδες αφαιρέθηκαν τα σημεία στίξης και επιλέχθηκαν αυτές που αποτελούνταν αποκλειστικά από χαρακτήρες (και δεν εμπεριείχαν αριθμούς ή σύμβολα), με σκοπό την επιλογή λέξεων και όχι συμβόλων ή αριθμών. Στη συνέχεια, τα κεφαλαία γράμματα των επιλεγμένων λέξεων μετατράπηκαν σε πεζά. Ακολουθεί παράδειγμα (αρνητικής) κριτικής πριν από την προεπεξεργασία (κείμενο) και μετά (συλλογή από λέξεις).

Αρχικό κείμενο (πριν την προεπεξεργασία):

> Robert DeNiro plays the most unbelievably intelligent illiterate of all time. This movie is so wasteful of talent, it is truly disgusting. The script is unbelievable. The dialog is unbelievable. Jane Fonda's character is a caricature of herself, and not a funny one. The movie moves at a snail's pace, is photographed in an ill-advised manner, and is insufferably preachy. It also plugs in every cliche in the book. Swoozie Kurtz is excellent in a supporting role, but so what?

Equally annoying is this new IMDB rule of requiring ten lines for every review. When a movie is this worthless, it doesn't require ten lines of text to let other readers know that it is a waste of time and tape. Avoid this movie.

Συλλογή από λέξεις (μετά την προεπεξεργασία):

> ['robert', 'deniro', 'plays', 'the', 'most', 'unbelievably', 'intelligent', 'illiterate', 'of', 'all', 'time', 'this', 'movie', 'is', 'so', 'wasteful', 'of', 'talent', 'it', 'is', 'truly', 'disgusting', 'the', 'script', 'is', 'unbelievable', 'the', 'dialog', 'is', 'unbelievable', 'jane', 'fonda', 's', 'character', 'is', 'a', 'caricature', 'of', 'herself', 'and', 'not', 'a', 'funny', 'one', 'the', 'movie', 'moves', 'at', 'a', 'snail', 's', 'pace', 'is', 'photographed', 'in', 'an', 'illadvised', 'manner', 'and', 'is', 'insufferably', 'preachy', 'it', 'also', 'plugs', 'in', 'every', 'cliche', 'in', 'the', 'book', 'swoozie', 'kurtz', 'is', 'excellent', 'in', 'a', 'supporting', 'role', 'but', 'so', 'what', 'equally', 'annoying', 'is', 'this', 'new', 'imdb', 'rule', 'of', 'requiring', 'ten', 'lines', 'for', 'every', 'review', 'when', 'a', 'movie', 'is', 'this', 'worthless', 'it', 'does', 'not', 'require', 'ten', 'lines', 'of', 'text', 'to', 'let', 'other', 'readers', 'know', 'that', 'it', 'is', 'a', 'waste', 'of', 'time', 'and', 'tape', 'avoid', 'this', 'movie']

Στο τελευταίο βήμα αναζητήθηκαν οι λέξεις (επιλεγείσες λεξικογραφικές μονάδες) στη συλλογή των pre-trained word embeddings glove6B_100. Για τις λέξεις που υπήρχαν στη glove6B_100 βρέθηκαν τα αντίστοιχα διανύσματα τα οποία προστέθηκαν και αποθηκεύτηκε και ο αριθμός τους, με σκοπό την εύρεση διανύσματος που οι συντεταγμένες του είναι ο μέσος όρος τον διανυσμάτων των λέξεων. Συνεπώς το αποτέλεσμα ήταν ένα διάνυσμα διάστασης 101 (διάσταση 100 των word embeddings και 1 του πλήθους των λέξεων που υπήρχαν στη glove6B_100).

2.2. Ταξινομητής Gaussian Naïve Bayesian

Ο ταξινομητής Gaussian Naïve Bayesian προϋποθέτει ότι οι μεταβλητές των παραδειγμάτων που χρησιμοποιούνται μπορούν να προσεγγιστούν με κανονική κατανομή (gaussian). Χρησιμοποιεί τα παραδείγματα του training set για να κατασκευάσει τις κανονικές κατανομές που ταιριάζουν καλύτερα στα χαρακτηριστικά των παραδειγμάτων εκπαίδευσης για κάθε κλάση. Από αυτές εξάγει τις αντίστοιχες κατανομές πυκνότητας πιθανότητας. Τέλος χωρίζει τον φασικό χώρο σε περιοχές με υπερεπίπεδα, τα οποία ονομάζονται και υπερεπίπεδα απόφασης (decision hyperplanes), έτσι ώστε σε κάθε περιοχή η πιθανότητα κάποιας κλάσης να είναι μεγαλύτερη από αυτές των άλλων. Με αυτόν το τρόπο, όλος ο φασικός χώρος χωρίζεται και κάθε περιοχή του χαρακτηρίζεται με το «όνομα» της κλάσης με την μεγαλύτερη πιθανότητα. Για κάθε άγνωστο παράδειγμα βρίσκεται η περιοχή του χώρου στον οποίο ανήκει σύμφωνα με τα χαρακτηριστικά του και κατατάσσεται στην κλάση με την οποία έχει χαρακτηριστεί η περιοχή αυτή [12].

Γενικά οι ταξινομητές Naïve Bayesian είναι πολύ απλοί στην υλοποίησή τους, είναι πολύ εύκολο να ερμηνευθούν τα αποτελέσματά τους και από υπολογιστικής πλευράς δεν χρειάζονται πολλούς πόρους. Επιπλέον μπορούν να χειριστούν δεδομένα που αποτελούνται από περισσότερες από 2 κλάσεις (ιδιότητα που δεν ισχύει για όλα τα είδη των ταξινομητών). Από την άλλη, είναι μια από τις απλούστερες υλοποιήσεις ταξινομητών και δεν λαμβάνουν υπόψιν συσχετίσεις που μπορεί να υπάρχουν στα χαρακτηριστικά των παραδειγμάτων εκπαίδευσης. Κατασκευάζεται απλώς η κατανομή πυκνότητας πιθανότητας Ν μεταβλητών, όπου Ν το πλήθος των χαρακτηριστικών, χρησιμοποιώντας το δείγμα εκπαίδευσης. Επίσης για την ικανοποιητική κατασκευή της κατανομής πυκνότητας πιθανότητας το πλήθος των παραδειγμάτων εκπαίδευσης αυξάνει εκθετικά σε σχέση με το πλήθος των χαρακτηριστικών [12].

Ο Gaussian Naïve Bayesian ταξινομητής μπορεί να εκπαιδευτεί με αυξητικό τρόπο χωρίς να έχει πρόσβαση σε προγενέστερα δεδομένα, χάρη στην υπόθεση ότι τα δεδομένα ακολουθούν την κανονική κατανομή. Υποθέτοντας ότι ο ταξινομητής εκπαιδεύτηκε με κάποιο σετ δεδομένων (s_1), τα οποία αποτελούνται από n_1 παραδείγματα με ένα χαρακτηριστικό x με μέση τιμή μ_1 και διακύμανση σ_1^2, η πυκνότητα πιθανότητας δίνεται από τη σχέση 1.

$$f(x; \mu,\sigma^2)\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (1)$$

Υποθέτοντας ότι στη συνέχεια εκπαιδεύεται με ένα άλλο σετ δεδομένων (s_2) με αριθμό παραδειγμάτων n_2 η νέα πυκνότητα πιθανότητας θα έχει μέση τιμή μ_2 και διακύμανση σ_2^2. Οι κατανομές αυτές μπορούν να συνδυαστούν (προστεθούν) σε μία κατανομή που θα έχει μέση τιμή μ και διακύμανση σ^2 που δίνονται από τη σχέση 2.

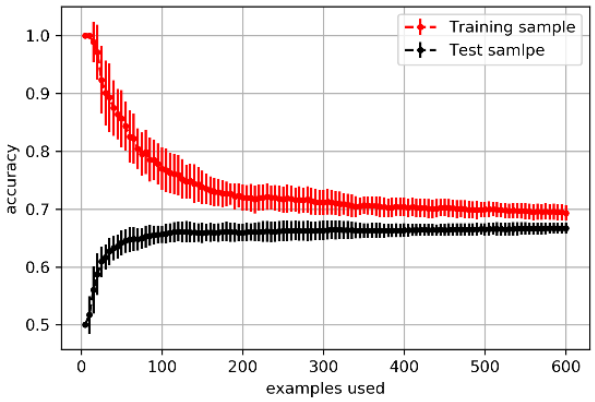
$$\mu = \frac{\mu_1 n_1 + \mu_2 n_2}{n_1 + n_2}$$
$$\sigma^2 = \frac{({\sigma_1}^2 + {\mu_1}^2)n_1 + ({\sigma_2}^2 + {\mu_2}^2)n_2}{n_1 + n_2} - \mu \quad (2)$$

Συνεπώς, το μόνο επιπλέον στοιχείο που χρειάζεται να γνωρίζουμε είναι ο αριθμός των παραδειγμάτων που έχουν χρησιμοποιηθεί μέχρι στιγμής για την εκπαίδευση του ταξινομητή [13]. Τα πιο πάνω εύκολα γενικεύονται και στην περίπτωση που τα παραδείγματα δεν αποτελούνται από ένα αλλά από m χαρακτηριστικά.

Η υλοποίηση του ταξινομητή Gaussian Naïve Bayesian που εξετάστηκε [14] είναι γραμμένη σε γλώσσα Python και εμπεριέχεται στο πακέτο scikit-learn [15]. Αξίζει να αναφερθεί ότι ο ταξινομητής αυτός μπορεί να χρησιμοποιηθεί για την ταξινόμηση παραδειγμάτων σε περισσότερες από δύο κλάσεις. Μετά από κάποιές δοκιμές με αυτόν τον ταξινομητή έγινε φανερό ότι η απόδοσή του εξαρτιόταν σε μεγάλο βαθμό από το δείγμα εκπαίδευσης. Για τον λόγο αυτό οι διαδικασίες που περιγράφονται πιο κάτω επαναλήφθηκαν 30 φορές και πριν από κάθε επανάληψη της διαδικασίας, το δείγμα εκπαίδευσης των 25.000 παραδειγμάτων ανακατευόταν με τυχαίο τρόπο. Στην πιο κάτω ανάλυση στο test sample εμπεριέχονται και τα 25.000 παραδείγματα του dataset.

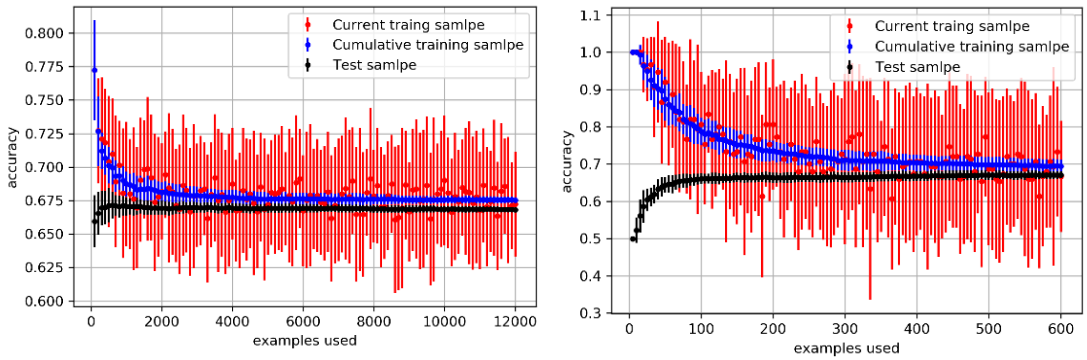
Στην **Εικόνα 5** παρουσιάζεται η μέση ακρίβεια του ταξινομητή (από 30 επαναλήψεις), ο οποίος δεν έχει εκπαιδευτεί με αυξητικό τρόπο, συναρτήσει του αριθμού των παραδειγμάτων που έχουν χρησιμοποιηθεί για την εκπαίδευσή του. Με κόκκινο χρώμα παρουσιάζεται η ακρίβεια του ταξινομητή στο δείγμα που χρησιμοποιήθηκε για την

εκπαίδευσή του (training samples) ενώ με μαύρο σε ένα άγνωστο δείγμα (test sample). Από την εικόνα αυτή είναι φανερό ότι όταν για την εκπαίδευση έχει χρησιμοποιηθεί μικρός αριθμός παραδειγμάτων ο ταξινομητής πάσχει από overtraining, ενώ όταν ο αριθμός των παραδειγμάτων αυξάνει (πάνω από 500) η ακρίβεια του training και του test sample συγκλίνουν. Ως overtraining χαρακτηρίζεται η κατάσταση κατά την οποία ο ταξινομητής έχει «μάθει» πάρα πολύ καλά το σύνολο των παραδειγμάτων με τα οποία έχει εκπαιδευτεί, αλλά δεν είναι σε θέση να γενικεύσει την «γνώση» του σε νέα παραδείγματα (πάντα υπό την προϋπόθεση ότι τα νέα παραδείγματα διαφέρουν, αλλά πολύ λίγο, από αυτά που έχουν χρησιμοποιηθεί για την εκπαίδευσή του).



Εικόνα 5: *Η ακρίβεια του ταξινομητή που δεν έχει εκπαιδευτεί με αυξητικό τρόπο σε σχέση με τον αριθμό των παραδειγμάτων που έχουν χρησιμοποιηθεί για την εκπαίδευσή του. Κόκκινο: η ακρίβεια του ταξινομητή με βάση το training sample. Μαύρο: η ακρίβεια με βάση το test sample. Τα σημεία αντιπροσωπεύουν τη μέση τιμή της ακρίβειας για τις 30 επαναλήψεις ενώ το μήκος της γραμμής σφάλματος την τυπική απόκλιση.*

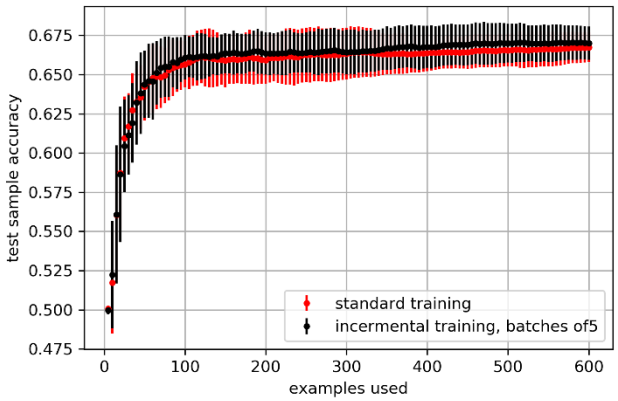
Στην **Εικόνα 6** παρουσιάζεται η μέση ακρίβεια του ταξινομητή (από 30 επαναλήψεις), ο οποίος έχει εκπαιδευτεί με αυξητικό τρόπο, συναρτήσει του αριθμού των παραδειγμάτων που έχουν χρησιμοποιηθεί συνολικά για την εκπαίδευσή του. Με κόκκινο χρώμα παρουσιάζεται η ακρίβεια του ταξινομητή από το δείγμα που έχει χρησιμοποιηθεί για την τελευταία εκπαίδευσή του, με μπλε η ακρίβεια του ταξινομητή με βάση το σύνολο του δείγματος εκπαίδευσης που έχει χρησιμοποιηθεί από την πρώτη μέχρι και την τελευταία εκπαίδευση, ενώ με μαύρο με βάση το test sample. Αριστερά παρουσιάζεται η ακρίβεια όταν χρησιμοποιήθηκαν 120 δείγματα εκπαίδευσης των 100 παραδειγμάτων ενώ δεξιά όταν χρησιμοποιήθηκαν 120 δείγματα εκπαίδευσης των 5 παραδειγμάτων. Από την εικόνα αυτή εξάγονται τα ίδια συμπεράσματα όπως και από την **Εικόνα 5**.



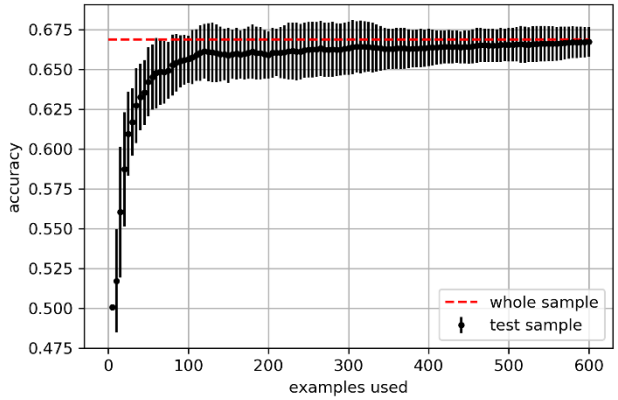
Εικόνα 6: *Η ακρίβεια του ταξινομητή που έχει εκπαιδευτεί με αυξητικό τρόπο σε σχέση με τον αριθμό των παραδειγμάτων που έχουν χρησιμοποιηθεί για την εκπαίδευσή του. Κόκκινο χρώμα: η ακρίβεια του ταξινομητή με βάση το δείγμα που έχει χρησιμοποιηθεί για την τελευταία εκπαίδευσή του. Μπλε χρώμα: η ακρίβεια με βάση το σύνολο του δείγματος εκπαίδευσης που έχει χρησιμοποιηθεί από την πρώτη μέχρι και την τελευταία εκπαίδευση. Μαύρο χρώμα: η ακρίβεια με βάση το test sample. Τα σημεία αντιπροσωπεύουν τη μέση τιμή της ακρίβειας για τις 30 επαναλήψεις ενώ το μήκος της γραμμής σφάλματος την τυπική απόκλιση. Αριστερά: για την εκπαίδευση του ταξινομητή έχουν χρησιμοποιηθεί (έως 120) δείγματα των 100 παραδειγμάτων (συνολικά 12.000 παραδείγματα). Δεξιά: για την εκπαίδευση του ταξινομητή έχουν χρησιμοποιηθεί (έως 120) δείγματα των 50 παραδειγμάτων (συνολικά 600 παραδείγματα).*

Συγκρίνοντας την **Εικόνα 5** με την **Εικόνα 6** φαίνεται ότι η κατά μέσο όρο ακρίβεια του ταξινομητή που εκπαιδεύτηκε με αυξητικό τρόπο είναι περίπου ίδια με αυτή του ταξινομητή που εκπαιδεύτηκε με τον συνήθη τρόπο. Εκπαιδεύοντας όμως τον ταξινομητή με αυξητικό τρόπο η διακύμανση της ακρίβειάς του είναι πολύ μεγαλύτερη (δηλαδή η «σταθερότητα» του ταξινομητή είναι χειρότερη). Στην **Εικόνα 7** παρατίθεται η σύγκριση της ακρίβειας του ταξινομητή που έχει εκπαιδευτεί με τον συνήθη τρόπο (κόκκινο) και με αυξητικό τρόπο (μαύρο) όπως υπολογίστηκε από το test sample.

Στην **Εικόνα 8** παρατίθεται με μαύρο χρώμα η ακρίβεια ενός Gaussian Naïve Bayesian ταξινομητή που έχει εκπαιδευτεί με αυξητικό τρόπο, ενώ με την κόκκινη διακεκομμένη γραμμή ενός ταξινομητή που έχει εκπαιδευτεί με τον συνήθη τρόπο. Από αυτό το διάγραμμα γίνεται φανερό ότι ο Gaussian Naïve Bayesian ταξινομητής που εκπαιδεύεται με αυξητικό τρόπο, με την πάροδο ικανού αριθμού δειγμάτων εκπαίδευσης, έχει εξίσου καλή ακρίβεια με ταξινομητή που έχει αξιοποιήσει το πλήρες δείγμα εκπαίδευσης χωρίς να είναι κατακερματισμένη η πληροφορία που χρησιμοποιεί κατά την εκπαίδευσή του. Αυτή είναι ίσως η πιο κρίσιμη ιδιότητα που αναζητείται σε κάθε ταξινομητή που εκπαιδεύεται με αυξητικό τρόπο. Παρόλα αυτά, η ακρίβεια του Gaussian Naïve Bayesian ταξινομητή στο test sample φτάνει περίπου στο 67,5%. Η τιμή αυτή δεν είναι αρκετά υψηλή, γεγονός το οποίο υποδεικνύει ότι αυτός ο ταξινομητής δεν είναι η βέλτιστη επιλογή για το συγκεκριμένο πρόβλημα.



Εικόνα 7: *Σύγκρισή της ακρίβειας που προκύπτει από το test sample για ταξινομητή που έχει εκπαιδευτεί με τον συνήθη – μη αυξητικό (κόκκινο) και με αυξητικό τρόπο (μαύρο). Τα σημεία αντιπροσωπεύουν τη μέση τιμή της ακρίβειας για τις 30 επαναλήψεις ενώ το μήκος του σφάλματος την τυπική απόκλιση.*



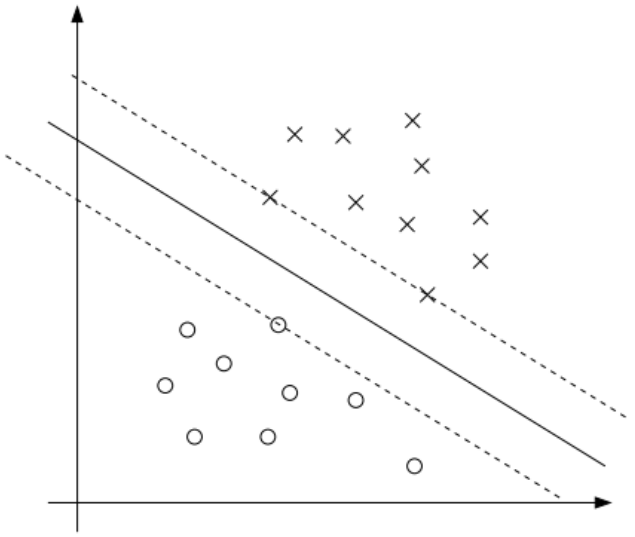
Εικόνα 8: *Ακρίβεια ταξινομητή που έχει εκπαιδευτεί με αυξητικό τρόπο. Τα σημεία αντιπροσωπεύουν τη μέση τιμή της ακρίβειας ενώ το μήκος της γραμμής σφάλματος την τυπική απόκλιση. Με την κόκκινη διακεκομμένη γραμμή παρουσιάζεται η τιμή της ακριβείας του Gaussian Naïve Bayesian ταξινομητή που έχει εκπαιδευτεί με τον συνήθη (μη αυξητικό) τρόπο χρησιμοποιώντας και τα 25.000 παραδείγματα εκπαίδευσης, όπως αυτή προκύπτει από το test sample.*

2.3.Support Vector Machine με Stochastic Gradient Descent

Οι Μηχανές Διανυσμάτων Στήριξης – Support Vector Machines είναι εργαλεία μηχανικής μάθησης που χρησιμοποιούνται τόσο για κατηγοριοποίηση όσο και για παλινδρόμηση. Στην περίπτωση της κατηγοριοποίησης ο στόχος των μηχανών διανυσμάτων στήριξης είναι η εύρεση της βέλτιστης υπερεπιφάνειας ώστε να μπορεί να επιτευχθεί ο διαχωρισμός των διαφορετικών κλάσεων. Στην περίπτωση που στον φασικό χώρο δεν υπάρχει αλληλοεπικάλυψη των κλάσεων, η εύρεση αυτής της υπερεπιφάνειας είναι

τετριμμένο πρόβλημα. Στην **Εικόνα 9** παρουσιάζεται μια σχηματική απεικόνιση αυτής της περίπτωσης όταν τα παραδείγματα που έχουν χρησιμοποιηθεί αποτελούνται από 2 χαρακτηριστικά, δηλαδή είναι διανύσματα διάστασης 2. Τα παραδείγματα κάθε κλάσης που απέχουν τη μικρότερη απόσταση από την διαχωριστική επιφάνεια ονομάζονται διανύσματα στήριξης – support vectors. Στην περίπτωση όπου οι κλάσεις δεν αλληλεπικαλύπτονται, η πρόκληση βρίσκεται στην εύρεση εκείνης της επιφάνειας που μεγιστοποιεί το εύρος του περιθωρίου μεταξύ των διαφορετικών κλάσεων, όπως παρουσιάζεται στην **Εικόνα 10** .

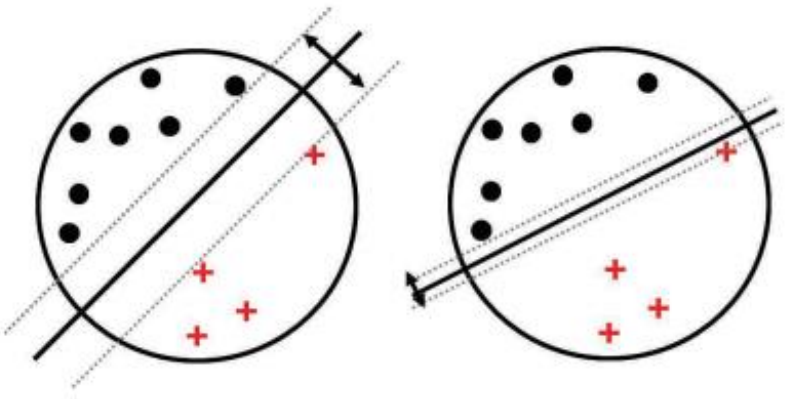
Η περίπτωση, όμως, που οι κλάσεις δεν αλληλεπικαλύπτονται δεν είναι η γενική περίπτωση. Στην γενικότερη, και πιο συνηθισμένη, περίπτωση υπάρχει αλληλοεπικάλυψη των παραδειγμάτων των διαφορετικών κλάσεων όπως φαίνεται στην **Εικόνα 11**, συνεπώς δεν είναι δυνατόν να βρεθεί τέτοια επιφάνεια ώστε να διαχωρίζονται πλήρως τα παραδείγματα των διαφορετικών κλάσεων. Στην **Εικόνα 11** παρουσιάζεται η επιφάνεια διαχωρισμού των κλάσεων και μια ζώνη γύρω από αυτή. Από τα σημεία που βρίσκονται εντός της ζώνης, αυτά που είναι εγγεγραμμένα σε τετράγωνα είναι σωστά ταξινομημένα, ενώ αυτά που είναι εγγεγραμμένα σε κύκλο είναι λάθος ταξινομημένα. Όλα τα υπόλοιπα σημεία είναι σωστά ταξινομημένα. Τα σημεία που βρίσκονται εντός της ζώνης αλλά από τη σωστή πλευρά της επιφάνειας επιβαρύνονται με κάποια (σχετικά μικρή) τιμή σφάλματος, η οποία αυξάνει όσο πιο κοντά βρίσκονται τα σημεία στην διαχωριστική επιφάνεια. Τα σημεία που βρίσκονται από την λάθος πλευρά της επιφάνειας επιβαρύνονται με μεγαλύτερη τιμή σφάλματος. Ο στόχος των μηχανών διανυσμάτων στήριξης σε αυτή την περίπτωση είναι η εύρεση της μέγιστης δυνατής ζώνης με το ελάχιστο δυνατό συνολικό σφάλμα.



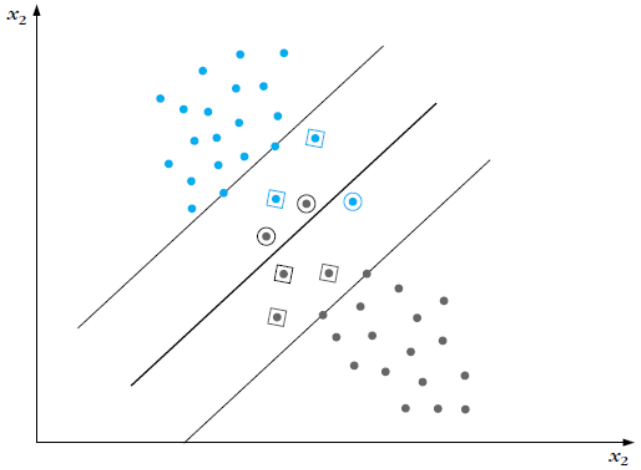
***Εικόνα 9**: Η περίπτωση όπου οι διαφορετικές κλάσεις δεν αλληλεπικαλύπτονται στον φασικό χώρο. Εικόνα από [16].*

Οι μηχανές διανυσμάτων στήριξης που αναζητούν την βέλτιστη υπερεπιφάνεια στον φασικό χώρο που ορίζεται από τα χαρακτηριστικά των παραδειγμάτων ονομάζονται γραμμικές. Είναι όμως δυνατό, και συνηθισμένο, να δύναται να επιτευχθεί καλύτερος διαχωρισμός των μεταβλητών, αν οι τιμές των χαρακτηριστικών μετασχηματιστούν μέσω κάποιας συνάρτησης. Τέτοιοι μετασχηματισμοί των τιμών των χαρακτηριστικών ονομάζονται

μετασχηματισμοί πυρήνα (kernels) και χρησιμοποιούνται ώστε να αναπαρασταθεί το κάθε παράδειγμα σε περισσότερες διαστάσεις ή με διαφορετικές μεταβλητές. Με τον τρόπο αυτό μπορούν να αξιοποιηθούν μη γραμμικές συσχετίσεις των χαρακτηριστικών των παραδειγμάτων [17].



***Εικόνα 10**: Επιλογή υπερεπιφάνειας διαχωρισμού των κλάσεων. Αριστερά: Μέγιστη απόσταση του περιθωρίου μεταξύ των διανυσμάτων στήριξης. Δεξιά: Ελάχιστη απόσταση του περιθωρίου μεταξύ των διανυσμάτων στήριξης. Εικόνα από [18].*



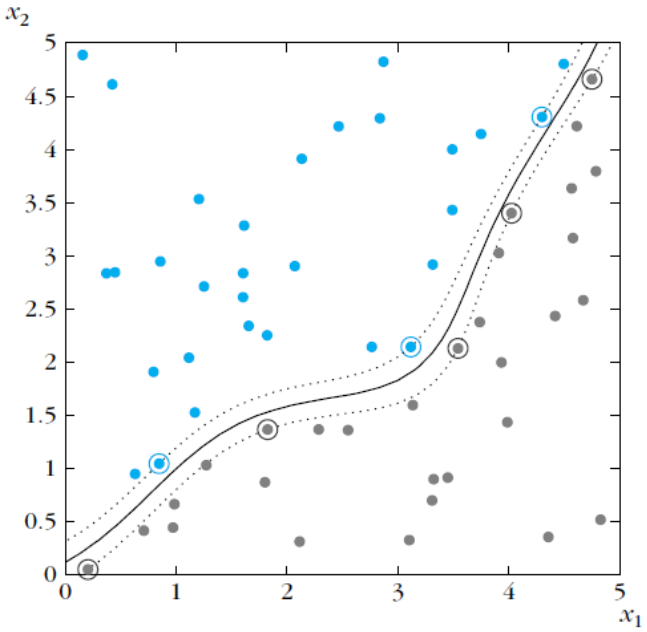
***Εικόνα 11**: Η περίπτωση όπου οι διαφορετικές κλάσεις αλληλεπικαλύπτονται στον φασικό χώρο. Εικόνα από [12].*

Μία από τις πιο συνηθισμένες συναρτήσεις μετασχηματισμού πυρήνα είναι η radial basis function (RBF) kernel (σχέση 3) ή γκαουσιανή συνάρτηση πυρήνα [19].

$$k(x_1, x_2) = exp\left(-\frac{\|x_1 - x_2\|}{2\sigma^2}\right) \quad (3)$$

Με ||…|| συμβολίζεται η ευκλείδεια απόσταση των διανυσμάτων

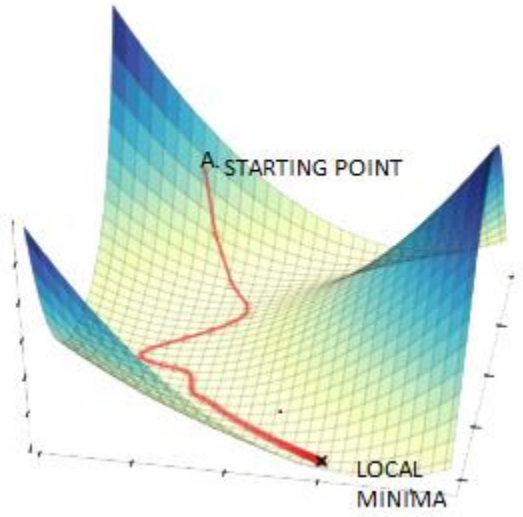
Το σύνολο τιμών αυτής της συνάρτησης είναι το [0,1]. Όταν δυο διανύσματα (παραδείγματα) απέχουν μεγάλη απόσταση, αντιστοιχίζονται σε πολύ μικρές τιμές (οριακά για άπειρη απόσταση η τιμή της συνάρτησης γίνεται 0) ενώ όταν τα διανύσματα ταυτίζονται, το αποτέλεσμα της συνάρτησης είναι 1. Με τον τρόπο αυτό δίνεται ένα μέτρο ομοιότητας των παραδειγμάτων. Χρησιμοποιώντας τη συνάρτηση μετασχηματισμού πυρήνα RBF σε μηχανές διανυσμάτων στήριξης, η οποία αντιστοιχίζει με μη γραμμικό τρόπο τα χαρακτηριστικά των παραδειγμάτων σε φασικό χώρο με περισσότερες διαστάσεις, είναι δυνατή η εύρεση και αξιοποίηση μη γραμμικών συσχετίσεων των χαρακτηριστικών για την καλύτερη κατηγοριοποίηση των παραδειγμάτων. Στην **Εικόνα 12** παρουσιάζεται μια περίπτωση όπου οι δυο κλάσεις δεν μπορούν να διαχωριστούν με γραμμικό τρόπο, αλλά με την χρήση της συνάρτησης μετασχηματισμού πυρήνα RBF ο διαχωρισμός τους παρουσιάζει άριστα αποτελέσματα.



***Εικόνα 12**: Αποτέλεσμα κατηγοριοποίησης δύο κλάσεων με μηχανή διανυσμάτων στήριξης όπου έχει χρησιμοποιηθεί η συνάρτηση μετασχηματισμού πυρήνα RBF. Τα παραδείγματα κάθε κλάσης παρουσιάζονται με διαφορετικό χρώμα και τα διανύσματα στήριξης είναι εγγεγραμμένα σε κύκλους. Εικόνα από [12].*

Μια από τις πιο γνωστές μεθόδους για την ελαχιστοποίηση της συνάρτησης σφάλματος (ή συνάρτησης κόστους) είναι η μέθοδος της απότομης καθόδου ή αλλιώς της σύγκλισης με ελάττωση της παραγώγου (Gradient Descent). Για την εύρεση του ελαχίστου βρίσκεται η τιμή τις παραγώγου για κάποιο σημείο του φασικού χώρου. Στην συνέχεια επιλέγεται επόμενο σημείο, το οποίο βρίσκεται αν στις συντεταγμένες του πρώτου προστεθεί το γινόμενο της παραγωγού επί κάποια τιμή – βήμα. Συνεπώς κάθε σημείο επιλέγεται με τέτοιο τρόπο ώστε να βρίσκεται πιο κοντά στο ελάχιστο σε σχέση με το προηγούμενο. Μια σχηματική απεικόνιση παρουσιάζεται στην **Εικόνα 13**. Βέβαια η επιτυχία της εύρεσης ελαχίστου αλλά και ο αριθμός των επαναλήψεων που απαιτείται για τη σύγκλιση εξαρτάται σε μεγάλο βαθμό τόσο από την επιλογή του βήματος όσο και από την γεωμετρία της υπό

εξέταση συνάρτησης. Αν χρησιμοποιηθεί πολύ μεγάλο βήμα είναι πιθανό να μην εντοπιστεί το ελάχιστο ή ακόμα και κάθε επανάληψη να καταλήγει σε σημεία μακρύτερα από το ελάχιστο. Αν όμως χρησιμοποιηθεί πολύ μικρό βήμα απαιτείται μεγαλύτερος αριθμός επαναλήψεων μέχρι να επιτευχθεί σύγκλιση και είναι πιθανός ο εγκλωβισμός σε τοπικά ελάχιστα.



***Εικόνα 13**: Σχηματική απεικόνιση της εύρεσης ελαχίστου με την μέθοδο Gradient Descent σε συνάρτηση δύο μεταβλητών. Εικόνα από [20].*

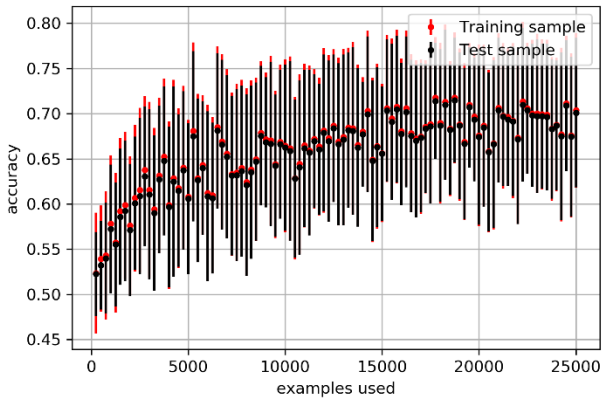
Όταν η μέθοδος Gradient Descent χρησιμοποιείται για την εύρεση ελαχίστου συνάρτησης πολλών μεταβλητών, η πολυπλοκότητα του υπολογισμού της παραγωγού σε κάθε σημείο, άρα και ο χρόνος υπολογισμού, αυξάνει καθώς αυξάνουν οι διαστάσεις. Πιο συγκεκριμένα σε προβλήματα μηχανικής μάθησης, οι διαστάσεις της συνάρτησης σφάλματος (που πρέπει να ελαχιστοποιηθεί) αυξάνουν καθώς αυξάνει το πλήθος των παραδειγμάτων εκπαίδευσης και ο αριθμός των χαρακτηριστικών τους. Σε τέτοιες περιπτώσεις χρησιμοποιείται η μέθοδος Stochastic Gradient Descent (SGD). Όπως υποδηλώνει και το όνομα στην μέθοδο αυτή εμπεριέχεται κάποιος βαθμός τυχαιότητας. Πιο συγκεκριμένα αντί να χρησιμοποιούνται όλα τα παραδείγματα για την εύρεση της παραγώγου σε κάθε επανάληψη, χρησιμοποιείται μόνο ένα παράδειγμα απλοποιώντας πολύ την διαδικασία και ελαττώνοντας κατά πολύ τον χρόνο που απαιτείται για τη σύγκλιση [21].

Στο πακέτο scikit-learn υπάρχει ο ταξινομητής Stochastic Gradient Descent [22] ο οποίος υλοποιεί, σε γλώσσα Python, μία γραμμική μηχανή διανυσμάτων στήριξης αξιοποιώντας την μέθοδο Stochastic Gradient Descent για την ελαχιστοποίηση της συνάρτησης κόστους. Ο ταξινομητής αυτός μπορεί να εκπαιδευτεί με αυξητικό τρόπο. Βέβαια, όταν εκπαιδεύεται με αυξητικό τρόπο, εκτελείται μόνο μία επανάληψη κατά την εύρεση των διανυσμάτων στήριξης και είναι πιθανό να μην βρεθεί το ολικό ελάχιστο της συνάρτησης. Αξίζει να αναφερθεί ότι ο ταξινομητής αυτός μπορεί να χρησιμοποιηθεί για την ταξινόμηση παραδειγμάτων σε περισσότερες από δύο κλάσεις.

Μετά από δοκιμές με αυτόν τον ταξινομητή φάνηκε ότι η ακρίβειά του εξαρτιόταν σε μεγάλο βαθμό από το δείγμα εκπαίδευσης. Συνεπώς, για την εύρεση των αποτελεσμάτων

που περιγράφονται πιο κάτω, εκτελέστηκαν 30 επαναλήψεις της διαδικασίας εκπαίδευσης και ελέγχου της ακρίβειας, όπου πριν από κάθε επανάληψη της διαδικασίας το δείγμα εκπαίδευσης των 25.000 παραδειγμάτων ανακατευόταν με τυχαίο τρόπο. Στην πιο κάτω ανάλυση στο test sample εμπεριέχονται και τα 25.000 παραδείγματα του dataset.

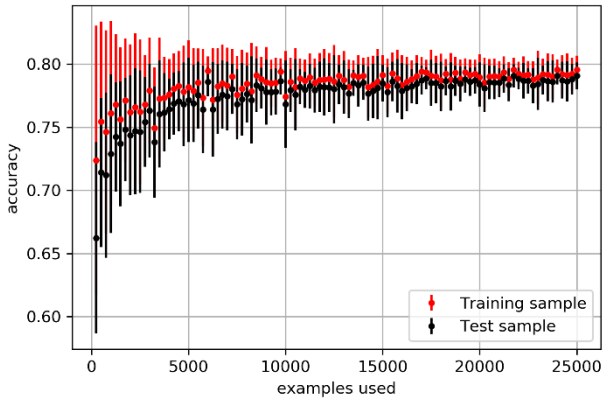
Στην **Εικόνα 14** παρουσιάζεται η μέση ακρίβεια του ταξινομητή (από 30 επαναλήψεις), ο οποίος δεν έχει εκπαιδευτεί με αυξητικό τρόπο (αλλά τα υπόλοιπα χαρακτηριστικά του ήταν ίδια με αυτά του ταξινομητή που εκπαιδεύτηκε με αυξητικό τρόπο), συναρτήσει του αριθμού των παραδειγμάτων που έχουν χρησιμοποιηθεί για την εκπαίδευσή του. Με κόκκινο χρώμα παρουσιάζεται η ακρίβεια του ταξινομητή όπως βρέθηκε από το training sample ενώ με μαύρο όπως υπολογίστηκε από το test sample.



***Εικόνα 14**: Η ακρίβεια ταξινομητή που δεν έχει εκπαιδευτεί με αυξητικό τρόπο σε σχέση με τον αριθμό των παραδειγμάτων που έχουν χρησιμοποιηθεί για την εκπαίδευσή του. Κόκκινο: η ακρίβεια του ταξινομητή με βάση το training sample. Μαύρο: η ακρίβεια με βάση το test sample. Τα σημεία αντιπροσωπεύουν τη μέση τιμή της ακρίβειας για τις 30 επαναλήψεις ενώ το μήκος της γραμμής σφάλματος την τυπική απόκλιση.*

Σε αντίθεση με τον ταξινομητή Gaussian Naïve Bayesian που πάσχει από overtraining για μικρό αριθμό δειγμάτων (**Εικόνα 5**), ο ταξινομητής SGD δεν παρουσιάζει τέτοια συμπεριφορά. Παρόλα αυτά, η απόδοση του ταξινομητή παρουσιάζει πολύ μεγάλες διακυμάνσεις όταν αλλάζει το δείγμα εκπαίδευσης, ακόμα και για δείγματα με πολύ μεγάλο αριθμό παραδειγμάτων. Βέβαια, η συμπεριφορά αυτή ήταν αναμενόμενη, δεδομένου ότι τα υπόλοιπα χαρακτηριστικά αυτού του ταξινομητή ήταν ίδια με αυτά ταξινομητή που εκπαιδεύτηκε με αυξητικό τρόπο, με πιο σημαντικό τον αριθμό των επαναλήψεων για την ελαχιστοποίηση της συνάρτησης σφάλματος με την μέθοδο SGD για την εύρεση των διανυσμάτων στήριξης (max iterations = 1), όπως αναφέρθηκε και πιο πάνω. Η ακρίβεια του ταξινομητή SGD για τον οποίο δεν τέθηκε περιορισμός στον αριθμό των επαναλήψεων ώστε να επιτευχθεί σύγκλιση, παρουσιάζεται στην **Εικόνα 15**. Σε αυτή την περίπτωση η εξάρτηση της ακρίβειας από το δείγμα εκπαίδευσης είναι πολύ μικρότερη και όσο περισσότερα παραδείγματα χρησιμοποιούνται τόσο μικρότερη είναι η διακύμανση (όπως και αναμένεται). Βέβαια, δεδομένου ότι ο στόχος της παρούσας εργασίας είναι η εύρεση ταξινομητή που λειτουργεί με αυξητικό τρόπο, η εκπαίδευση με μη αυξητικό τρόπο έχει μοναδικό σκοπό την σύγκριση της ακρίβειας του ταξινομητή με αυτόν που εκπαιδεύεται με αυξητικό τρόπο,

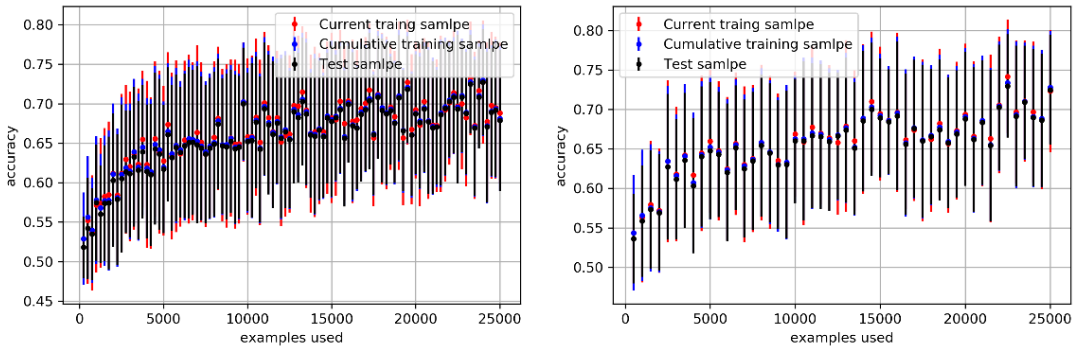
συνεπώς για να γίνει μια σύγκριση υπό ίσους όρους, ο αριθμός των επαναλήψεων κατά την εύρεση των διανυσμάτων στήριξης (max iterations) πρέπει να είναι 1.



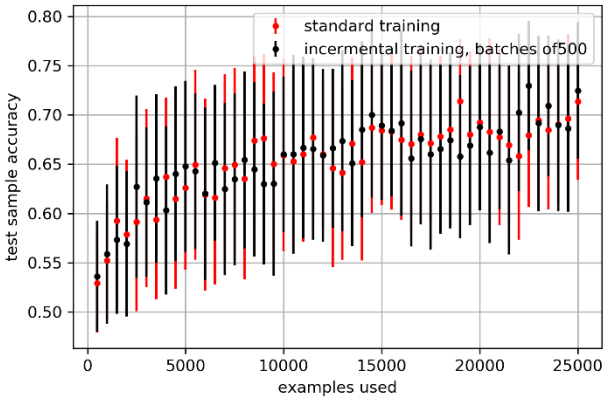
***Εικόνα 15**: Η ακρίβεια ταξινομητή SGD για τον οποίο επιτυγχάνεται σύγκλιση και ο αλγόριθμος δεν σταματά λόγω περιορισμού του μέγιστου αριθμού επαναλήψεων (max iterations), σε σχέση με τον αριθμό των παραδειγμάτων που έχουν χρησιμοποιηθεί για την εκπαίδευσή του. Με κόκκινο χρώμα παρουσιάζεται η ακρίβεια του ταξινομητή με βάση το training sample ενώ με μαύρο με βάση το test sample. Τα σημεία αντιπροσωπεύουν τη μέση τιμή της ακρίβειας για τις 30 επαναλήψεις ενώ το μήκος της γραμμής σφάλματος την τυπική απόκλιση.*

Στην **Εικόνα 16** παρουσιάζεται η μέση ακρίβεια του ταξινομητή (από 30 επαναλήψεις), ο οποίος έχει εκπαιδευτεί με αυξητικό τρόπο, συναρτήσει του αριθμού των παραδειγμάτων που έχουν χρησιμοποιηθεί συνολικά για την εκπαίδευσή του. Με κόκκινο χρώμα παρουσιάζεται η ακρίβεια του ταξινομητή όπως βρέθηκε από το δείγμα που έχει χρησιμοποιηθεί για την τελευταία εκπαίδευσή του, με μπλε όπως βρέθηκε από το σύνολο του δείγματος εκπαίδευσης που έχει χρησιμοποιηθεί από την πρώτη μέχρι και την τελευταία εκπαίδευση, ενώ με μαύρο όπως βρέθηκε από το test sample. Στα αριστερά παρουσιάζεται η ακρίβεια του ταξινομητή όταν το κάθε δείγμα εκπαίδευσης (batch) που χρησιμοποιήθηκε αποτελούταν από 250 παραδείγματα ενώ στα δεξιά όταν το κάθε δείγμα εκπαίδευσης (batch) που χρησιμοποιήθηκε αποτελούταν από 500 παραδείγματα. Από την εικόνα αυτή εξάγονται τα ίδια συμπεράσματα όπως και από την **Εικόνα 14**.

Συγκρίνοντας την **Εικόνα 14** με την **Εικόνα 16** φαίνεται ότι η κατά μέσο όρο η εκπαίδευση με αυξητικό τρόπο έχει την ίδια απόδοση με την συνήθη (μη αυξητική) εκπαίδευση του ταξινομητή όπως επίσης και συγκρίσιμη διακύμανση (δηλαδή η «σταθερότητα» του ταξινομητή είναι εξίσου κακή και στις δυο περιπτώσεις). Στην **Εικόνα 17** παρατίθεται η σύγκριση της ακρίβειας του ταξινομητή που έχει εκπαιδευτεί με τον συνήθη τρόπο (κόκκινο) και με αυξητικό τρόπο (μαύρο) όπως υπολογίστηκε από το δείγμα εκπαίδευσης, ενισχύοντας τον πιο πάνω ισχυρισμό.



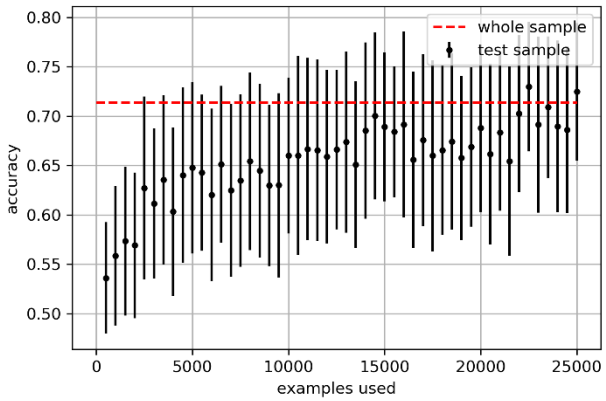
Εικόνα 16: *Η ακρίβεια του ταξινομητή που έχει εκπαιδευτεί με αυξητικό τρόπο σε σχέση με τον αριθμό των παραδειγμάτων που έχουν χρησιμοποιηθεί για την εκπαίδευσή του. Κόκκινο χρώμα: η ακρίβεια του ταξινομητή με βάση το δείγμα που έχει χρησιμοποιηθεί για την τελευταία εκπαίδευσή του. Μπλε χρώμα: η ακρίβεια του ταξινομητή με βάση το σύνολο του δείγματος εκπαίδευσης που έχει χρησιμοποιηθεί από την πρώτη μέχρι και την τελευταία εκπαίδευση. Μαύρο χρώμα: η ακρίβεια του ταξινομητή με βάση το test sample. Τα σημεία αντιπροσωπεύουν τη μέση τιμή της ακρίβειας για τις 30 επαναλήψεις ενώ το μήκος της γραμμής σφάλματος την τυπική απόκλιση. Αριστερά: Για την εκπαίδευση του ταξινομητή έχουν χρησιμοποιηθεί δείγματα των 250 παραδειγμάτων. Δεξιά: για την εκπαίδευση του ταξινομητή έχουν χρησιμοποιηθεί δείγματα των 500 παραδειγμάτων.*



Εικόνα 17: *Σύγκριση της ακρίβειας που προκύπτει από το test sample για ταξινομητή που έχει εκπαιδευτεί με τον συνήθη (κόκκινο) και με αυξητικό τρόπο (μαύρο). Τα σημεία αντιπροσωπεύουν τη μέση τιμή της ακρίβειας για τις 30 επαναλήψεις ενώ το μήκος της γραμμής σφάλματος την τυπική απόκλιση.*

Στην **Εικόνα 18** παρατίθεται με μαύρο χρώμα η ακρίβεια του ταξινομητή που έχει εκπαιδευτεί με αυξητικό τρόπο, ενώ με την κόκκινη διακεκομμένη γραμμή η μέση τιμή της ακρίβειας ταξινομητή που έχει εκπαιδευτεί με τον συνήθη (μη αυξητικό) τρόπο χρησιμοποιώντας όλο το δείγμα εκπαίδευσης. Από αυτό το διάγραμμα φαίνεται ότι ο ταξινομητής που εκπαιδεύεται με αυξητικό τρόπο, με την πάροδο ικανού αριθμού δειγμάτων εκπαίδευσης, έχει εξίσου καλή ακρίβεια με ταξινομητή που έχει αξιοποιήσει το πλήρες δείγμα εκπαίδευσης. Ενώ η μέση τιμή της ακρίβειας του SGD ταξινομητή στο test sample

υπερβαίνει το 70%, άρα κατά μέσο όρο είναι καλύτερη από αυτή του Gaussian Naïve Bayesian ταξινομητή, η διακύμανσή της είναι πολύ μεγάλη. Το γεγονός αυτό υποδεικνύει ότι ούτε αυτός ο ταξινομητής αποτελεί καλή επιλογή για την επίλυση του συγκεκριμένου προβλήματος.

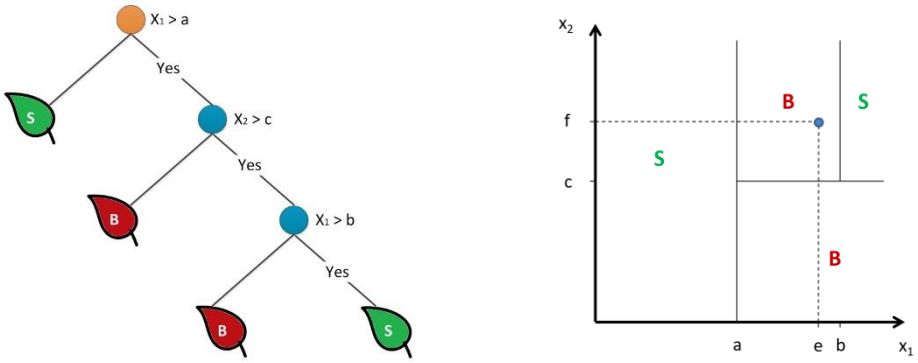


Εικόνα 18: *Ακρίβεια ταξινομητή που έχει εκπαιδευτεί με αυξητικό τρόπο. Με την κόκκινη διακεκομμένη γραμμή παρουσιάζεται η μέση τιμή της ακριβείας SGD ταξινομητή που έχει εκπαιδευτεί με τον συνήθη (μη αυξητικό) τρόπο χρησιμοποιώντας και τα 25.000 παραδείγματα εκπαίδευσης, όπως αυτή προκύπτει από το test sample.*

2.4.Online Random Forest

Ο ταξινομητής Online Random Forest αποτελεί μια συλλογή από ταξινομητές τυχαίων (randomized) δένδρων απόφασης που εκπαιδεύονται με online τρόπο. Ο online τρόπος αποτελεί ειδική περίπτωση της αυξητικής προσέγγισης καθώς ο ταξινομητής επανεκπαιδεύεται με κάθε καινούριο παράδειγμα που δέχεται. Ο ταξινομητής [23] που εξετάστηκε και αναπτύσσεται πιο κάτω είναι υλοποιημένος σε C++ και περιγράφεται στα [24] [25].

Τα δένδρα απόφασης αποτελούν μια μεγάλη κατηγορία μη γραμμικών ταξινομητών [12]. Χρησιμοποιούνται τόσο για κατηγοριοποίηση όσο και για παλινδρόμηση. Στην περίπτωση της κατηγοριοποίησης τα δυαδικά δένδρα απόφασης (binary decision trees) τμηματοποιούν τον φασικό χώρο διαδοχικά καθώς σε κάθε κόμβο απόφασης (όπως επίσης και στην ρίζα – πρώτος κόμβος) ελέγχεται μια συνθήκη που σχετίζεται με κάποιο χαρακτηριστικό. Η συνθήκη βρίσκεται με τέτοιο τρόπο ώστε σε κάθε κόμβο να μεγιστοποιείται ο διαχωρισμός των κλάσεων. Τελικά κάθε τμήμα του φασικού χώρου χαρακτηρίζεται με το «όνομα» κάποιας κλάσης, **Εικόνα 19**. Όταν κάποιο νέο παράδειγμα πρέπει να ταξινομηθεί, εξετάζεται η θέση του στον φασικό χώρο και κατατάσσεται στην κλάση με το «όνομα» της οποίας έχει «τιτλοδοτηθεί» η περιοχή του φασικού χώρου στην οποία βρίσκεται. Παράδειγμα αποτελεί το μπλε σημείο στην **Εικόνα 19** (αριστερά). Με βάση τα χαρακτηριστικά του (e, f) βρίσκεται στην περιοχή του χώρου που έχει χαρακτηριστεί ως κλάση «B», συνεπώς κατηγοριοποιείται στην κλάση «B».



***Εικόνα 19**: Αριστερά: Εικονική αναπαράσταση ενός δυαδικού δένδρου απόφασης με μέγιστο βάθος 3. Ο πορτοκαλί κύκλος αντιπροσωπεύει τη «ρίζα» ενώ κάθε μπλε κύκλος έναν κόμβο απόφασης. Η υπόθεση που ελέγχεται σε κάθε κόμβο απόφασης και στη ρίζα αναγράφεται στα δεξιά. Τα «φύλλα» έχουν χαρακτηριστεί με το όνομα των κλάσεων «S» και «B». Δεξιά: Αναπαράσταση του φασικού χώρου του δυαδικού δένδρου απόφασης που απεικονίζεται στα αριστερά. Ο φασικός χώρος έχει τμηματοποιηθεί και κάθε τμήμα φέρει το όνομα της κλάσης («S» και «B») στην οποία κατατάσσεται.*

Μια γραφική προσέγγιση των δυαδικών δένδρων απόφασης παρουσιάζεται στην **Εικόνα 19** δεξιά. Κάθε δένδρο αποτελείται από μια ρίζα (πορτοκαλί), κάποιους κόμβους απόφασης (μπλε) και φύλλα (πράσινα για την κλάση «S» και κόκκινα για την κλάση «B»). Για την ταξινόμηση ενός νέου παραδείγματος, με βάση τα χαρακτηριστικά του ακολουθείται καθοδική πορεία από την ρίζα προς τα φύλλα ακολουθώντας κάθε φορά την κατεύθυνση που καθορίζεται από τον έλεγχο της υπόθεσης κάθε κόμβου. Τελικά η κλάση στην οποία κατατάσσεται είναι αυτή με την οποία έχει χαρακτηριστεί το φύλλο στο οποίο καταλήγει. Το παράδειγμα που χρησιμοποιήθηκε πιο πάνω με χαρακτηριστικά (e, f) ξεκινώντας από τη ρίζα ακολουθεί τον δεξί δρόμο (καθώς e > a), στην συνέχεια στον πρώτο κόμβο απόφασης ακολουθεί τον δεξιό δρόμο (καθώς f > c) και στον επόμενο κόμβο ακολουθεί τον αριστερό δρόμο (καθώς e < b) καταλήγοντας στο φύλλο όπου χαρακτηρίζεται ως παράδειγμα της κλάσης «B».

Στην περίπτωση που τα χαρακτηριστικά των παραδειγμάτων είναι πολλά ή που μπορεί στην ίδια περιοχή του χώρου να συνυπάρχουν χαρακτηριστικά που ανήκουν σε περισσότερες από μια κλάσεις, τα δένδρα απόφασης που παρουσιάζουν τα καλύτερα αποτελέσματα αποτελούνται από πολλούς διαδοχικούς κόμβους απόφασης (ο αριθμός των διαδοχικών κόμβων απόφασης λέγεται και βάθος – depth – του δένδρου). Παρόλα αυτά τα δένδρα απόφασης με μεγάλο βάθος είναι επιρρεπή σε overtraining. Για την αποφυγή του overtraining, αντί για την χρήση ενός δένδρου με μεγάλο βάθος συνήθως χρησιμοποιούνται συλλογές από δένδρα με μικρό βάθος τα οποία έχουν εκπαιδευτεί με πιο έξυπνές τεχνικές. Οι πιο ευρέως διαδομένοι ταξινομητές που χρησιμοποιούν συλλογές δένδρων απόφασης είναι οι: Boosted Decision Trees [26], Random Forests [27] και Extremely Randomized Trees [27].

Τα Random Forests αποτελούν συλλογές δένδρων απόφασης. Κάθε δένδρο της συλλογής (που καλείται forest – δάσος) εκπαιδεύεται με διαφορετικό δείγμα εκπαίδευσης, το οποίο όμως έχει προκύψει από το αρχικό δείγμα επιλέγοντας τα παραδείγματα

εκπαίδευσης με τυχαίο τρόπο και κατά κανόνα με επανατοποθέτηση. Εκτός αυτού, η τυχαιότητα (όπως υποδηλώνει και το όνομα – random) έγκειται και στο ότι το χαρακτηριστικό βάσει του οποίο θα εξαχθεί η συνθήκη διαχωρισμού, επιλέγεται με τυχαίο τρόπο (και όχι με εύρεση του βέλτιστου χαρακτηριστικού που θα επιφέρει τον μέγιστο διαχωρισμό των κλάσεων). Η συνθήκη βέβαια επιλέγεται έτσι ώστε να επιτευχθεί ο βέλτιστος διαχωρισμός. Τα Extremely Randomized Trees είναι μια επέκταση των Random Forests όπου και η συνθήκη διαχωρισμού (εκτός από τα χαρακτηριστικά στα οποία θα εφαρμοστεί) βρίσκεται και αυτή με τυχαίο τρόπο.

Τα δένδρα απόφασης και οι συλλογές τους (π.χ. Extremely Randomized Trees) που έχουν πρόσβαση σε όλο το δείγμα εκπαίδευσης μπορούν να βρουν τις βέλτιστες συνθήκες για τον διαχωρισμό των κλάσεων αξιοποιώντας όλη την διαθέσιμη πληροφορία. Αντιθέτως σε μια online εφαρμογή τους αυτό δεν είναι δυνατό. Στην βιβλιογραφία έχουν προταθεί τεχνικές στις οποίες τα παραδείγματα που έχουν χρησιμοποιηθεί αποθηκεύονται στα φύλλα των δένδρων, με σκοπό να χρησιμοποιηθούν στο μέλλον. Αυτή όμως δεν είναι μια σωστή προσέγγιση ενός ταξινομητή που εκπαιδεύεται με αυξητικό τρόπο. Σε έναν τέτοιο ταξινομητή όσα παραδείγματα έχουν ήδη χρησιμοποιηθεί πρέπει να απορρίπτονται και το μοντέλο να κρατά την απαραίτητη πληροφορία υπό τη μορφή παραμέτρων (και όχι αυτούσιων των παραδειγμάτων). Εξαίρεση μπορεί να αποτελέσει η προσωρινή αποθήκευση πολύ μικρού τμήματος (χαρακτηριστικών) παραδειγμάτων αλλά όχι όλων ή της μεγάλης πλειοψηφίας τους. Η υλοποίηση Online Random Forest που ελέγχθηκε εδώ ικανοποιεί πλήρως τις απαιτήσεις αυτές.

Ο ταξινομητής Online Random Forest [23] που χρησιμοποιήθηκε αποτελεί μία online επέκταση ταξινομητή Extremely Randomized Trees. Κάθε παράδειγμα εκπαίδευσής μπορεί να χρησιμοποιηθεί σε κάθε δένδρο από καμία έως και πολλές φόρες επιλέγοντας τυχαία τον αριθμό τον φορών από μια κατανομή Poisson με μέση τιμή το 1. Για την δημιουργία του ταξινομητή αρχικά προαποφασίζεται ο αριθμός των δένδρων όπως επίσης και άλλων παραμέτρων. Κάθε δένδρο αποτελείται μόνο από τη ρίζα. Για την ρίζα ορίζεται τυχαία το χαρακτηριστικό (ή τα χαρακτηριστικά) που θα χρησιμοποιηθούν για την υπόθεση διαχωρισμού καθώς επίσης με τυχαίο τρόπο και η υπόθεση. Όσο φτάνουν παραδείγματα εκπαίδευσης αποθηκεύεται η κλάση όπως επίσης και το αν ικανοποιείται η συνθήκη ή όχι. Όταν συμπληρωθεί ο ικανός αριθμός παραδειγμάτων και το αποτέλεσμα της διάσπασης του κόμβου αποφέρει καλύτερα αποτελέσματα κατηγοριοποίησης από αυτά που προκύπτουν από την κατηγοριοποίηση πριν την διάσπαση, ο κόμβος διασπάται και τα παραδείγματα μεταφέρονται στο επόμενο επίπεδο. Επειδή συγκρατείται η κλάση των παραδειγμάτων που πέρασαν στο επόμενο επίπεδο είναι δυνατή η κατηγοριοποίηση νέων παραδειγμάτων από την πρώτη στιγμή. Η διαδικασία αυτή επαναλαμβάνεται για κάθε κόμβο απόφασης. Επίσης για την ευκολότερη προσαρμογή του ταξινομητή σε περιβάλλοντα όπου το δείγμα αλλάζει με την πάροδο του χρόνου, είναι δυνατή η απόρριψη ολόκληρου δένδρου. Για κάθε δένδρο υπολογίζεται το σφάλμα ταξινόμησης και η απόρριψη γίνεται με τυχαίο τρόπο και με πιθανότητα που αυξάνει όσο αυξάνει το σφάλμα ταξινόμησης. Δεδομένου ότι ο ταξινομητής αποτελείται από ένα σύνολο δένδρων, η απόρριψη ενός δένδρου δεν επιφέρει αρνητικά αποτελέσματα στον ταξινομητή, ενώ ταυτόχρονα η απόρριψη των δένδρων με την μικρότερη διαχωριστική δύναμη επιτρέπει την ευκολότερη προσαρμογή της συλλογής σε περιβάλλοντα που αλλάζουν. Τέλος αξίζει να σημειωθεί ότι ο ταξινομητής Online Random Forest μπορεί να χρησιμοποιηθεί για την ταξινόμηση παραδειγμάτων σε περισσότερες από δύο κλάσεις.

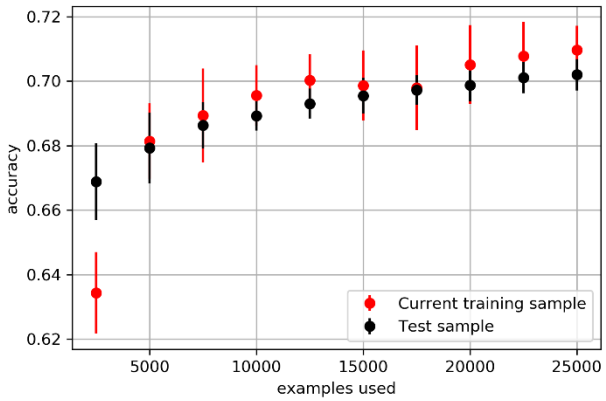
Ο ταξινομητής που εξετάστηκε δεν βελτιστοποιήθηκε με κάποιο τρόπο, αλλά χρησιμοποιήθηκε με τις προτεινόμενες (default) παραμέτρους οι οποίες είναι:

- Αριθμός δένδρων = 10
- Μέγιστο βάθος δένδρων = 20
- Ελάχιστος αριθμός παραδειγμάτων πριν τη διάσπαση κόμβου = 200

Επίσης έγιναν μικρές τροποποιήσεις στον πηγαίο κώδικα της υλοποίησης του ταξινομητή ώστε να είναι δυνατή η αξιολόγηση του test sample κατά τη διάρκεια της online εκπαίδευσης χωρίς να απαιτείται η αποθήκευση και επαναφόρτιση του εκπαιδευμένου (μέχρι εκείνη τη στιγμή) ταξινομητή. Δεδομένου ότι ο ταξινομητής είναι κατασκευασμένος ώστε να εκπαιδεύεται με online τρόπο, ως μέτρο σύγκρισης της ακρίβειας που θα είχε αντίστοιχος ταξινομητής αν δεν εκπαιδευόταν με online (αλλά χρησιμοποιώντας όλο το δείγμα εκπαίδευσής) χρησιμοποιήθηκε ταξινομητής ExtraTreesClassifier [28] από το πακέτο scikit-learn, ο οποίος αποτελεί υλοποίηση ταξινομητή Extremely Randomized Trees. Οι παράμετροι που χρησιμοποιήθηκαν για αυτόν τον ταξινομητή ήταν ίδιοι (όσο ήταν δυνατό) με τις παραμέτρους του ταξινομητή Online Random Forest. Βέβαια δεδομένου ότι οι υλοποιήσεις των δυο ταξινομητών είναι διαφορετικές και ότι οι παράμετροι δεν ήταν μια προς μια κοινές και στους δυο ταξινομητές αναμένονται αποκλίσεις την τελική απόδοση.

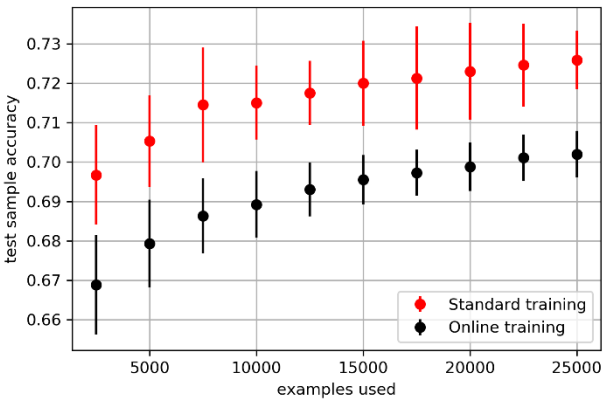
Σε αντίθεση με τους δυο προηγούμενους ταξινομητές οι αρχικοί πειραματισμοί έδειξαν ότι η χρήση διαφορετικού τμήματος του δείγματος εκπαίδευσης δεν είχε μεγάλη επίπτωση στην ακρίβεια του ταξινομητή. Για αυτό τον λόγο εκτελέστηκαν 10 επαναλήψεις από τις οποίες εξήχθη η μέση τιμή και η τυπική απόκλιση (σημείο και μήκος σφάλματος στα ακόλουθα διαγράμματα), ενώ πριν από κάθε επανάληψη της διαδικασίας το δείγμα εκπαίδευσης ανακατευόταν με τυχαίο τρόπο.

Στην **Εικόνα 20** παρουσιάζεται η μέση ακρίβεια του ταξινομητή Online Random Forest, συναρτήσει του αριθμού των παραδειγμάτων που έχουν χρησιμοποιηθεί μέχρι τη στιγμή του τεστ για την εκπαίδευσή του. Με κόκκινο χρώμα παρουσιάζεται η ακρίβεια του ταξινομητή όπως βρέθηκε από το δείγμα που έχει χρησιμοποιηθεί για την εκπαίδευσή του, ενώ με μαύρο όπως βρέθηκε από το test sample (25.000 παραδείγματα). Από την εικόνα αυτή φαίνεται ότι ο ταξινομητής δεν παρουσιάζει σημαντικές αλλαγές στην ακρίβεια όταν χρησιμοποιείται για την εκπαίδευσή του διαφορετικό τμήμα του δείγματος εκπαίδευσης, καθώς η τυπική απόκλισή είναι της τάξης του 1% και μικρότερη. Επίσης φαίνεται ότι ο ταξινομητής δεν υποφέρει από overtraining καθώς η ακρίβεια όπως βρίσκεται από το δείγμα εκπαίδευσης είναι ίδια (εντός σφάλματος) με αυτή που υπολογίζεται από το test sample.



Εικόνα 20: *Η ακρίβεια του ταξινομητή Online Random Forest σε σχέση με τον αριθμό των παραδειγμάτων που έχουν χρησιμοποιηθεί για την εκπαίδευσή του. Κόκκινο χρώμα: η ακρίβεια του ταξινομητή με βάση το δείγμα που χρησιμοποιήθηκε για την εκπαίδευσή του. Μαύρο: η ακρίβεια του ταξινομητή με βάση το test sample. Τα σημεία αντιπροσωπεύουν τη μέση τιμή της ακρίβειας για 10 επαναλήψεις ενώ το μήκος της γραμμής σφάλματος την τυπική απόκλιση.*

Στην **Εικόνα 21** με μαύρο χρώμα παρουσιάζεται η ακρίβεια του ταξινομητή Online Radom Forest ενώ με κόκκινο η ακρίβεια του ταξινομητή ExtraTreesClassifier. Η ακρίβεια και τον δύο ταξινομητών φαίνεται να αυξάνει με τον ίδιο τρόπο όταν αυξάνει ο αριθμός των παραδειγμάτων του δείγματος εκπαίδευσης. Η διαφορά έγκειται στο ότι η ακρίβεια του ταξινομητή ExtraTreesClassifier φαίνεται να είναι σταθερά 2.5% με 3% καλύτερη από αυτή του Online Radom Forest. Αυτή η διαφορά μπορεί να οφείλεται είτε στον διαφορετικό τρόπο εκπαίδευσής των δυο ταξινομητών είτε στην χρήση διαφορετικών παραμέτρων (όπως αναφέρθηκε πιο πάνω).



Εικόνα 21: *Σύγκρισή της ακρίβειας με βάση το test sample για τον ταξινομητή Online Random Forest (μαύρο) και για τον ταξινομητή ExtraTreesClassifier (κόκκινο). Τα σημεία αντιπροσωπεύουν τη μέση τιμή της ακρίβειας και το μήκος των γραμμών σφάλματος την τυπική απόκλιση (10 επαναλήψεις).*

Συνοψίζοντας, ο ταξινομητής Online Random Forest δεν φαίνεται να υποφέρει από overrating ακόμα και όταν το δείγμα εκπαίδευσης που έχει χρησιμοποιηθεί είναι μικρό. Επίσης η ακρίβειά του, όπως υπολογίζεται από το test sample, αυξάνει όταν αυξάνει ο αριθμός των παραδειγμάτων εκπαίδευσης, αγγίζοντας το 70%. Επιπλέον, φαίνεται ότι είναι αρκετά σταθερή ανεξάρτητα από το πιο τμήμα του δείγματος εκπαίδευσης έχει χρησιμοποιηθεί καθώς η τυπική της απόκλιση (σε 10 επαναλήψεις), ακόμα και για το μικρότερο δείγμα που έχει χρησιμοποιηθεί για την εκπαίδευση του, είναι της τάξης του 1%. Συνεπώς στον ταξινομητή Online Ramdom Forest εμπεριέχονται όλα τα καλά χαρακτηριστικά που συνηθώς αναζητούνται σε ταξινομητές που εκπαιδεύονται με αυξητικό τρόπο.

2.5.LaSVM

Ο ταξινομητής LaSVM [29] αποτελεί μια online υλοποίηση SVM ταξινομητή (σε γλώσσα C++) και μπορεί να βρεθεί στο [30]. Για την ελαχιστοποίηση της συνάρτησης κόστους χρησιμοποιεί την μέθοδο Sequential Minimal Optimization (SMO) [31]. Εκτός από τα διανύσματα στήριξης φυλάσσεται και αριθμός «υποψηφίων διανυσμάτων στήριξης» που καθορίζεται από τον χρήστη. Για τα υποψήφια διανύσματα στήριξης φυλάσσεται επιπλέον και η τιμή του σφάλματος λόγω της θέσης τους και της τρέχουσας τοπολογίας της διαχωριστικής υπερεπιφάνειας αλλά και η τιμή της μερικής παραγώγου της συνάρτησης κόστους στο συγκεκριμένο διάνυσμα. Τα πρώτα παραδείγματα γίνονται αυτόματα διανύσματα στήριξης και τα επόμενα, που δεν είναι ικανά να γίνουν διανύσματα στήριξης, γίνονται υποψήφια διανύσματα στήριξης. Στην συνέχεια για κάθε νέο παράδειγμα εκπαίδευσης ακολουθούνται τα πιο κάτω βήματα:

1) Το νέο παράδειγμα εισάγεται ως υποψήφιο διάνυσμα στήριξης (στην περίπτωση που δεν είναι ήδη). Αξιοποιώντας την μέθοδο SMO επανυπολογίζονται τα διανύσματα στήριξης στην περίπτωση που ικανοποιούνται οι συνθήκες της μεθόδου. Σε αντίθετη περίπτωση τα διανύσματα στήριξης μένουν αμετάβλητα.
2) Από τα πιθανά διανύσματα στήριξης αφαιρούνται όσα πληρούν κάποιες αυστηρές συνθήκες (βρίσκονται εκτός της ζώνης γύρω από την διαχωριστική επιφάνεια των κλάσεων που αναφέρθηκε στην Ενότητα 2.3)
3) Αν ο αριθμός των υποψηφίων διανυσμάτων στήριξης έχει υπερβεί τον αριθμό που έχει καθοριστεί από τον χρήστη, το βήμα 2) επαναλαμβάνεται, αφαιρώντας τα υποψήφια διανύσματα στήριξης που έχουν την μικρότερη δυνατότητα να αξιοποιηθούν στο μέλλον, έως ότου ο αριθμός τους γίνει ίσος με αυτόν που έχει οριστεί από τον χρήστη.

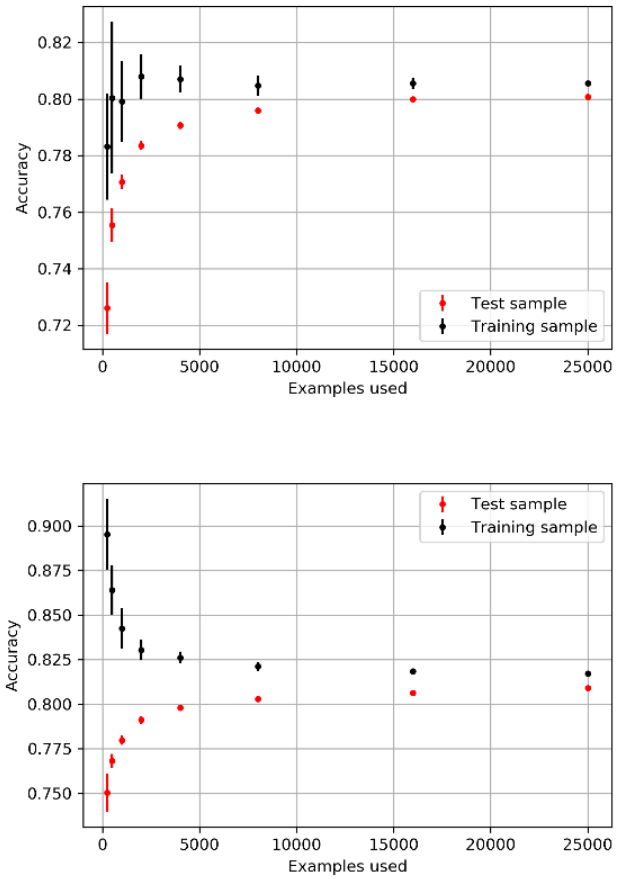
Ο ταξινομητής La SVM παρέχει την δυνατότητα να χρησιμοποιηθούν οι ακόλουθες συναρτήσεις μετασχηματισμού πυρήνα:

1) γραμμική,
2) πολυωνυμική,
3) RBF και
4) σιγμοειδής

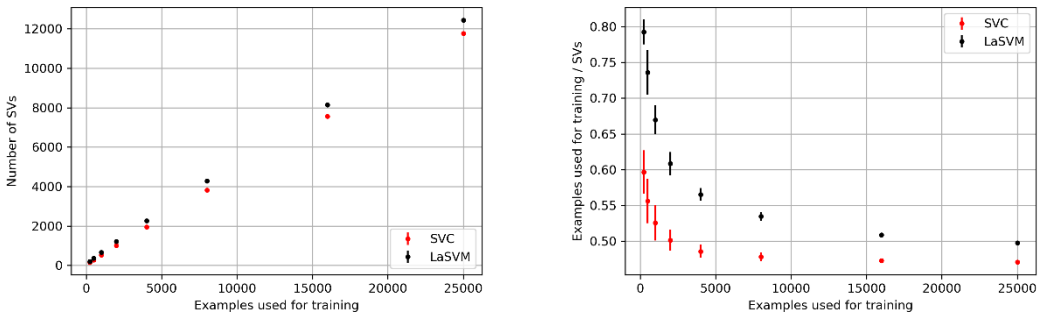
Τέλος πρέπει να σημειωθεί ότι ο ταξινομητής LaSVM δεν μπορεί να χρησιμοποιηθεί για την ταξινόμηση παραδειγμάτων σε περισσότερες από δύο κλάσεις, αλλά περιορίζεται μόνο σε δυαδική ταξινόμηση. Για την παραμετροποίηση του ταξινομητή LaSVM που εξετάστηκε χρησιμοποιήθηκε ο μετασχηματισμός πυρήνα RBF (που παρουσιάστηκε στην Ενότητα 2.3) τρίτου βαθμού, μέγιστος αριθμός πιθανών διανυσμάτων στήριξης 50 και τιμή ποινής σφάλματος 100. Για την σύγκριση των αποτελεσμάτων με ταξινομητή που δεν εκπαιδεύεται με online τρόπο αλλά αξιοποιώντας όλο το διαθέσιμο δείγμα εκπαίδευσης χρησιμοποιήθηκε ο ταξινομητής SVC [32] από το πακέτο scikit-learn. Ο ταξινομητής SVC αποτελεί υλοποίηση μηχανής διανυσμάτων στήριξης που χρησιμοποιεί την βιβλιοθήκη libsvm [33]. Για την παραμετροποίηση του ταξινομητή SVC χρησιμοποιήθηκαν οι ίδιες παράμετροι με αυτές που χρησιμοποιήθηκαν στον LaSVM.

Στην **Εικόνα 22** παρουσιάζεται η ακρίβεια των ταξινομητών LaSVM (πάνω) και SVC (κάτω) όπως υπολογίζεται τόσο από το δείγμα εκπαίδευσης (μαύρο) όσο και από το test sample (κόκκινο). Για την παραγωγή των γραφημάτων, η διαδικασία της εκπαίδευσης και ελέγχου του ταξινομητή επαναλήφθηκε 10 φορές. Από το άνω διάγραμμα φαίνεται ότι ο ταξινομητής LaSVM, που εκπαιδεύεται με αυξητικό τρόπο, παρουσιάζει overtraining κυρίως όταν ο αριθμός των παραδειγμάτων που έχουν χρησιμοποιηθεί για την εκπαίδευσή του είναι μικρός. Από την άλλη ο ταξινομητής SVC που δεν εκπαιδεύεται με αυξητικό τρόπο (κάτω διάγραμμα) παρουσιάζει πιο έντονη συμπεριφορά overtraining όταν εκπαιδεύεται με μικρό αριθμό παραδειγμάτων.

Στην **Εικόνα 23** παρουσιάζεται ο αριθμός των διανυσμάτων στήριξης που δημιουργεί το κάθε μοντέλο σε σχέση με τον αριθμό των παραδειγμάτων εκπαίδευσης που έχουν χρησιμοποιηθεί. Όπως αναμενόταν, όσο αυξάνει ο αριθμός των παραδειγμάτων εκπαίδευσης αυξάνει και ο αριθμός των διανυσμάτων στήριξης. Επίσης, ο λόγος του αριθμού των διανυσμάτων στήριξης ως προς τον αριθμό των παραδειγμάτων που έχουν χρησιμοποιηθεί για την εκπαίδευση του ταξινομητή φαίνεται να ελαττώνεται, καθώς ο αριθμός των παραδειγμάτων εκπαίδευσης αυξάνει και προσεγγίζει τιμές περίπου στο 50%. Αυτό σημαίνει βέβαια ότι από το δείγμα εκπαίδευσης περίπου τα μισά παραδείγματα γίνονται διανύσματα στήριξης. Τέλος αξίζει να σημειωθεί ότι ο αριθμός των διανυσμάτων στήριξης που δημιουργούνται με τον ταξινομητή LaSVM (ο οποίος εκπαιδεύεται με online μέθοδο) είναι μεγαλύτερος από αυτόν δημιουργούνται με τον ταξινομητή SVC.

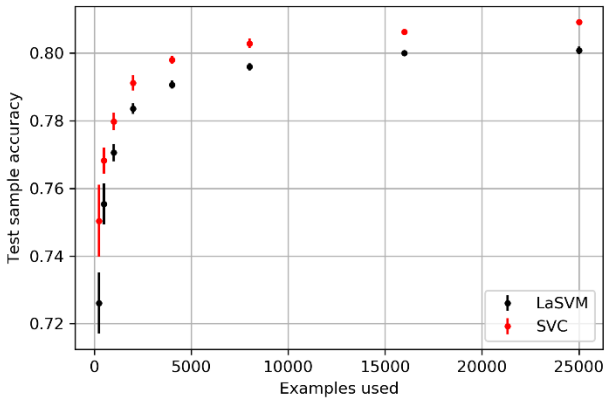


Εικόνα 22: *Ακρίβεια των ταξινομητών LaSVM (πάνω) και SVC (κάτω). Κόκκινο χρώμα: η ακρίβεια του ταξινομητή με βάση το δείγμα που χρησιμοποιήθηκε για την εκπαίδευσή του. Μαύρο: η ακρίβεια του ταξινομητή με βάση το test sample. Τα σημεία αντιπροσωπεύουν τη μέση τιμή της ακρίβειας για 10 επαναλήψεις ενώ το μήκος της γραμμής σφάλματος την τυπική απόκλιση.*



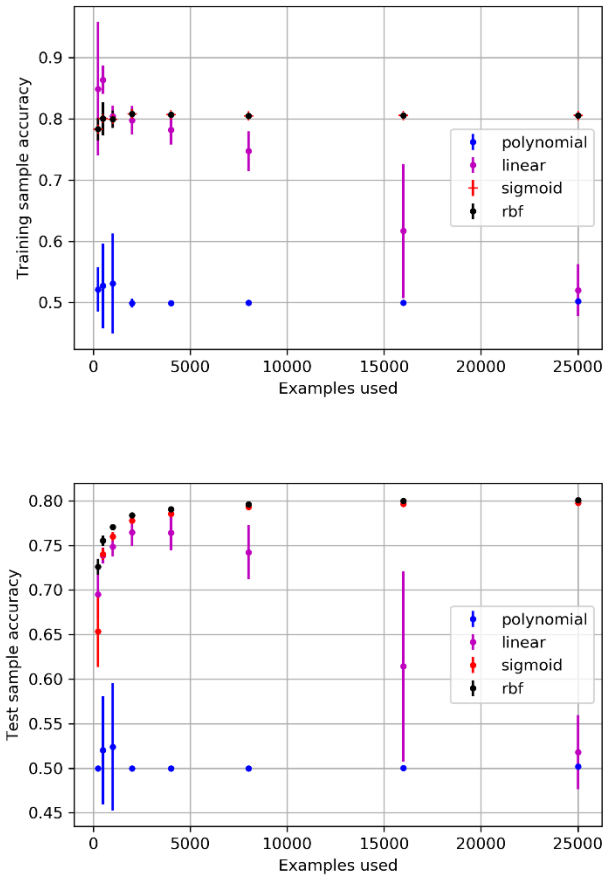
Εικόνα 23: *Αριθμός διανυσμάτων στήριξης (αριστερά) και λόγος του αριθμού διανυσμάτων στήριξης ως προς τον αριθμό των παραδειγμάτων εκπαίδευσης (δεξιά) ως προς τον αριθμό των παραδειγμάτων εκπαίδευσης. Με μαύρο χρώμα παρουσιάζεται ο ταξινομητής LaSVM ενώ με κόκκινο ο ταξινομητής SVC. Το κάθε σημείο αντιπροσωπεύει τη μέση τιμή των 10 επαναλήψεων ενώ το μήκος του σφάλματος της τυπική απόκλιση.*

Στην **Εικόνα 24** παρουσιάζεται η ακρίβεια του ταξινομητή LaSVM με μαύρο χρώμα ενώ με κόκκινο χρώμα η ακρίβεια του ταξινομητή SVC, όπως υπολογίστηκε από το test sample, ως συνάρτηση του αριθμού των παραδειγμάτων εκπαίδευσης. Για μικρό αριθμό παραδειγμάτων εκπαίδευσης η διαφορά της μέσης τιμής της ακρίβειας είναι της τάξης του 2%, αλλά τα στατιστικά σφάλματα είναι αρκετά μεγάλα (η τυπική απόκλιση είναι της τάξης του 1%). Όταν ο αριθμός των παραδειγμάτων εκπαίδευσης αυξάνει, η μέση τιμή της ακρίβειας αυξάνει με τον ίδιο τρόπο και για τους δύο ταξινομητές. Επιπλέον η διαφορά τους είναι της τάξης του 1% ενώ η τυπική απόκλιση του δείγματος των 10 επαναλήψεων είναι τόσο μικρή που τα σφάλματα δεν διακρίνονται στο διάγραμμα.



***Εικόνα 24**: Σύγκριση της ακρίβειας με βάση το test sample για τον ταξινομητή LaSVM (μαύρο) και για τον ταξινομητή SVC (κόκκινο) ως συνάρτηση του αριθμού των παραδειγμάτων εκπαίδευσης. Τα σημεία αντιπροσωπεύουν τη μέση τιμή της ακρίβειας και το μήκος των γραμμών σφάλματος την τυπική απόκλιση (10 επαναλήψεις).*

Για λόγους πληρότητας, η απόδοση του ταξινομητή LaSVM εξετάστηκε χρησιμοποιώντας όλες τις διαθέσιμες συναρτήσεις μετασχηματισμού πυρήνα (δηλαδή εκτός από την RBF χρησιμοποιήθηκαν και η γραμμική, πολυωνυμική και σιγμοειδής). Στην **Εικόνα 25** παρουσιάζεται η ακρίβεια των ταξινομητών όπως βρέθηκε τόσο από το δείγμα εκπαίδευσης (πάνω) όσο και από το test sample (κάτω) σε σχέση με τον αριθμό των παραδειγμάτων εκπαίδευσης που χρησιμοποιήθηκαν. Για την παραγωγή των γραφημάτων η διαδικασία της εκπαίδευσης και ελέγχου του ταξινομητή επαναλήφθηκε 10 φορές.



Εικόνα 25: *Ακρίβεια ταξινομητών LaSVM όπως έχει υπολογιστεί με βάση το δείγμα εκπαίδευσης (πάνω) και το test sample (κάτω). Μαύρο: συνάρτηση μετασχηματισμού πυρήνα RBF, κόκκινο: σιγμοειδής, μπλε: πολυωνυμική και μωβ: γραμμική . Τα σημεία αντιπροσωπεύουν τη μέση τιμή της ακρίβειας για 10 επαναλήψεις ενώ το μήκος της γραμμής σφάλματος την τυπική απόκλιση.*

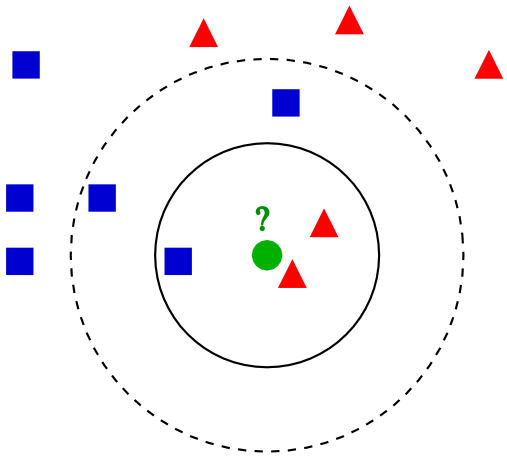
Η χρήση τόσο της σιγμοειδούς συνάρτησης όσο και της RBF έχουν σχεδόν τα ίδια αποτελέσματα (με ελάχιστα καλύτερη ακρίβεια στο test sample η RBF). Αντιθέτως, η χρήση της γραμμικής συνάρτησης και της πολυωνυμικής συνάρτησης δεν ενδείκνυνται για την λύση του προβλήματος που εξετάζεται εδώ. Με χρήση της πολυωνυμικής συνάρτησης τα αποτελέσματα της κατηγοριοποίησης μοιάζουν να είναι τυχαία, καθώς η ακρίβεια είναι (σχεδόν) σταθερή 50%, ανεξαρτήτως του αριθμού των παραδειγμάτων εκπαίδευσης που έχουν χρησιμοποιηθεί. Τέλος με χρήση της πολυωνυμικής συνάρτησης μετασχηματισμού πυρήνα τα αποτελέσματα είναι μη αναμενόμενα. Όσο ο αριθμός των παραδειγμάτων εκπαίδευσης αυξάνει, η ακρίβεια όπως υπολογίζεται από το δείγμα εκπαίδευσης που έχει χρησιμοποιηθεί μειώνεται από ~85% για 250 ή 500 παραδείγματα σε ~50% για 25.000 παραδείγματα! Επίσης η ακρίβεια όπως υπολογίζεται από το test sample αυξάνει μέχρι ~76% για 4000 παραδείγματα εκπαίδευσης και στη συνέχεια μειώνεται μέχρι ~50% για 25.000 παραδείγματα. Συνεπώς η συνάρτηση RBF αποτελεί την βέλτιστη συνάρτηση μετασχηματισμού πυρήνα για τον ταξινομητή LaSVM για την επίλυση του προβλήματος που εξετάζεται.

Όπως ο ταξινομητής Online Random Forest, έτσι και ο ταξινομητής LaSMV συγκεντρώνει πολλά από τα χαρακτηριστικά που αναζητούμε σε ταξινομητές που εκπαιδεύονται με αυξητικό τρόπο. Καταρχάς όσο αυξάνει το δείγμα εκπαίδευσης αυξάνει και η ακρίβεια του ταξινομητή φτάνοντας στο 80%, ενώ ταυτόχρονα η διαφορά της ακρίβεια από παρόμοιο ταξινομητή (SVC) που δεν εκπαιδεύεται με αυξητικό τρόπο είναι πολύ μικρή (της τάξης του 1%). Επίσης, η ακρίβειά του είναι αρκετά σταθερή και δεν φαίνεται να εξαρτάται από το τμήμα του δείγματος εκπαίδευσης με το οποίο εκπαιδεύεται, ιδίως για μεγάλα δείγματα. Τέλος, ενώ όταν εκπαιδεύεται με μικρά δείγματα παρουσιάζει σημεία overtraining, όσο το δείγμα μεγαλώνει το overtraining φαίνεται να υποχωρεί.

2.6.ILVQ

Ο ταξινομητής Incremental Learning Vector Quantization (ILVQ) [34] αποτελεί μία online υλοποίηση ταξινομητή που βασίζεται σε πρότυπα (σε γλώσσα C++) και μπορεί να βρεθεί στο [35]. Γενικά οι ταξινομητές που βασίζονται σε πρότυπα, κατηγοριοποιούν τα άγνωστα παραδείγματα με βάση την εγγύτητά τους σε πρότυπα παραδείγματα των διαφορετικών κλάσεων. Ένα από το γνωστότερα παραδείγματα τέτοιου ταξινομητή αποτελεί ο ταξινομητής «k εγγύτερων γειτόνων (k-NN)» σύμφωνα με τον οποίο για την κατηγοριοποίηση άγνωστου παραδείγματος, βρίσκονται τα k εγγύτερα παραδείγματα σε αυτό και κατατάσσεται στην πλειοψηφούσα κλάση. Στην **Εικόνα 26** παρουσιάζεται η κατηγοριοποίηση ενός άγνωστου παραδείγματος (πράσινος κύκλος) σύμφωνα με την μέθοδο των k εγγύτερων γειτόνων. Για k = 3 το άγνωστο παράδειγμα θα καταταγεί στην κλάση των κόκκινων τριγώνων ενώ για k = 5 θα καταταγεί στη κλάση των μπλε κύκλων.

Γενικά οι ταξινομητές που βασίζονται σε πρότυπα, μπορούν να χωριστούν σε δύο κατηγορίες με βάση τον τρόπο που επιλέγουν τα πρότυπα: στους ταξινομητές που «συμπυκνώνουν» όλη τη διαθέσιμη πληροφορία (όπως ο ταξινομητής των k εγγύτερων γειτόνων) αξιοποιώντας όλα τα δυνατά παραδείγματα εκπαίδευσης (condensing scheme) και στους ταξινομητές που επεξεργάζονται τα δεδομένα εκπαίδευσης (editing scheme) με σκοπό την απόρριψη θορυβωδών δεδομένων. Ο ταξινομητής ILVQ είναι ο πρώτος ταξινομητής που βασίζεται σε πρότυπα και εκπαιδεύεται με αυξητικό τρόπο. Ο τρόπος εκπαίδευσης του ταξινομητή αξιοποιεί και τις δυο μεθόδους επιλογής προτύπων που προαναφέρθηκαν. Τέλος αξίζει να σημειωθεί ότι μπορεί να χρησιμοποιηθεί για την κατηγοριοποίηση σε παραπάνω από 2 κλάσεις, όπως επίσης και ότι δεν χρειάζεται να είναι γνωστός εκ των προτέρων ο αριθμός των κλάσεων.



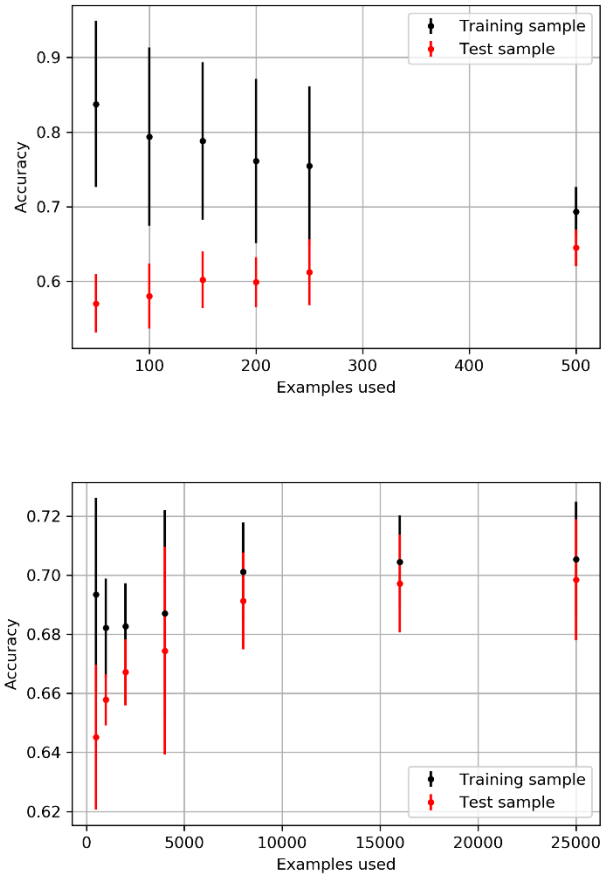
Εικόνα 26: *Σχηματική απεικόνιση κατηγοριοποίηση με τη μέθοδο των k εγγύτερων γειτόνων. Εικόνα από [36].*

Κατά την εκπαίδευση του ταξινομητή ILVQ, για κάθε νέο παράδειγμα εκπαίδευσης ακολουθούνται τα πιο κάτω βήματα:

I. Εάν το παράδειγμα είναι το πρώτο παράδειγμα της κλάσης, εισάγεται ως πρότυπο.
II. Εάν δεν είναι το πρώτο παράδειγμα της κλάσης:
    a. Βρίσκονται οι δύο εγγύτεροι γείτονες.
    b. Αν η απόσταση από κάποιον από αυτούς είναι μεγαλύτερη από μια τιμή που επιλέγει ο χρήστης, το παράδειγμα εισάγεται ως πρότυπο. Στην περίπτωση που είναι «πολύ μακριά» από τους εγγύτερους γείτονες, τότε αντιπροσωπεύει τμήμα της κατανομής των παραδειγμάτων της κλάσης στην οποία ανήκει και για το οποίο δεν υπάρχει αρκετή πληροφορία. Στην περίπτωση που βρίσκεται κοντά στον εγγύτερο γείτονα αλλά μακριά από τον επόμενο, είναι πολύ πιθανό να βρίσκεται σε κάποιο όριο της κατανομής, συνεπώς η αναγωγή του σε πρότυπο παράδειγμα είναι πολύ χρήσιμη. Το βήμα αυτό είναι ανεξάρτητο τόσο από την κλάση στην οποία ανήκει το παράδειγμα εκπαίδευσης που εξετάζεται όσο και από την κλάση στη οποία ανήκουν οι εγγύτεροι γείτονες.
    c. Αν το παράδειγμα δεν εισαχθεί στα πρότυπα, τότε αναπροσαρμόζεται η θέση των προτύπων που βρίσκονται κοντά σε αυτό. Αρχικά ελέγχεται αν η κλάση του παραδείγματος εκπαίδευσης και του εγγύτερου γείτονα είναι ίδιες (σωστή κατηγοριοποίηση), ή διαφορετικές (λάθος κατηγοριοποίηση). Στην πρώτη περίπτωση ο εγγύτερος μετατοπίζεται πιο κοντά στο νέο παράδειγμα εκπαίδευσης ενώ όλα τα πρότυπα που ανήκουν σε άλλες κλάσεις απομακρύνονται από αυτό. Στην δεύτερη περίπτωση τόσο η θέση του εγγύτερου γείτονα όσο και όλων των κοντινών παραδειγμάτων που ανήκουν στην ίδια κλάση απομακρύνονται από το παράδειγμα εκπαίδευσης.
III. Τα πρότυπα που δεν έχουν ανανεωθεί τις τελευταίες n φορές που ενεργοποιήθηκε το βήμα ii-c απομακρύνονται ως «θορυβώδη» πρότυπα.

Με την διαδικασία αυτή επιτυγχάνεται τόσο η ενημέρωση εντός των κλάσεων όσο και μεταξύ των κλάσεων.

Για τον έλεγχο του ταξινομητή ILVQ ως απόσταση των παραδειγμάτων χρησιμοποιήθηκε τόσο η ευκλείδεια όσο και το συνημίτονο της μεταξύ τους γωνίας (εσωτερικό γινόμενο). Πιο κάτω παρατίθενται τα αποτελέσματα όπου ως μέτρο απόστασης έχει χρησιμοποιηθεί η ευκλείδεια απόσταση των παραδειγμάτων εκπαίδευσης, ενώ τα αποτελέσματα με χρήση του εσωτερικού γινομένου ως μέτρο απόστασης ήταν μη αποδεκτά (ακρίβεια του ταξινομητή με βάση το δείγμα ελέγχου αλλά και το δείγμα εκπαίδευσης κάτω του 50%) και δεν παρουσιάζονται. Για την βελτιστοποίηση των παραμέτρων του ταξινομητή εκτελέστηκε έλεγχος της ακρίβειάς του στο test sample, χρησιμοποιώντας όλο το δείγμα εκπαίδευσης σε μεγάλο τμήμα του φασικού χώρου και επιλέχθηκε το set των παραμέτρων με τη μέγιστη ακρίβεια. Τέλος, πρέπει να σημειωθεί ότι δεν παρουσιάζεται σύγκριση με ταξινομητή ο οποίος να εκπαιδεύεται με μη αυξητικό παρεμφερή τρόπο, καθώς δεν βρέθηκε τέτοιος ταξινομητής.



Εικόνα 27: *Ακρίβεια του ταξινομητή ILVQ με βάση το δείγμα εκπαίδευσης (μαύρο) και το test sample (κόκκινο). Τα σημεία αντιπροσωπεύουν τη μέση τιμή της ακρίβειας για 10 επαναλήψεις ενώ το μήκος της γραμμής σφάλματος την τυπική απόκλιση. Στο άνω διάγραμμα παρουσιάζονται οι περιπτώσεις όπου ο αριθμός των παραδειγμάτων εκπαίδευσης είναι μέχρι και 500 ενώ στο κάτω από 500 μέχρι 25.000.*

Στην **Εικόνα 27** παρουσιάζεται η ακρίβεια του ταξινομητή ILVQ (όπου ως μέτρο απόσταση έχει χρησιμοποιηθεί η ευκλείδεια απόσταση των παραδειγμάτων) όπως

υπολογίζεται από το δείγμα εκπαίδευσης (κόκκινο) και από το test sample (μαύρο). Για την εύρεση της ακρίβειας έχουν εκτελεστεί 30 επαναλήψεις (όπου πριν από κάθε επανάληψη το αρχικό δείγμα εκπαίδευσης των 25.000 παραδειγμάτων, από το όποιο εξήχθησαν τα δείγματα εκπαίδευσης, ανακατευόταν με τυχαίο τρόπο), τα σημεία αναπαριστούν τη μέση τιμή της ακρίβειας ενώ το μήκος των σφαλμάτων την τυπική απόκλιση. Όπως γίνεται φανερό, για μικρό αριθμό παραδειγμάτων εκπαίδευσης ο ταξινομητής πάσχει από overtraining, το οποίο όμως εξαλείφεται καθώς ο αριθμός των παραδειγμάτων εκπαίδευσης αυξάνει. Η ακρίβειά του όμως, όπως υπολογίζεται από test sample, είναι αρκετά σταθερή και ανεξάρτητη του αριθμού των παραδειγμάτων εκπαίδευσης που χρησιμοποιήθηκαν (η τυπική απόκλιση ξεκινά από τιμές ~4% όταν στην εκπαίδευση έχουν χρησιμοποιηθεί μόλις 100 παραδείγματα ενώ φτάνει στο ~2% όταν έχουν χρησιμοποιηθεί 25.000 παραδείγματα). Τέλος, φαίνεται ότι η ακρίβεια του ταξινομητή ILVQ, όπως υπολογίζεται από το test sample, αυξάνει καθώς αυξάνει ο αριθμός των παραδειγμάτων εκπαίδευσης, ξεκινώντας από τιμές ~60% όταν στην εκπαίδευση έχουν χρησιμοποιηθεί μόλις 100 παραδείγματα και αγγίζοντας το ~70% όταν έχουν χρησιμοποιηθεί 25.000 παραδείγματα.

Συνοψίζοντας ο ταξινομητής ILVQ συγκεντρώνει πολλά από τα χαρακτηριστικά που αναζητούμε σε ταξινομητές που εκπαιδεύονται με αυξητικό τρόπο. Η ακρίβεια του ταξινομητή φτάνει στο 70%, ενώ ταυτόχρονα φαίνεται να μην εξαρτάται σε μεγάλο βαθμό από το αριθμό των παραδειγμάτων εκπαίδευσης που χρησιμοποιούνται. Τέλος, ενώ όταν εκπαιδεύεται με μικρά δείγματα φαίνεται να υποφέρει σημαντικά από overtraining, όταν ο αριθμός των παραδειγμάτων μεγαλώνει η συμπεριφορά αυτή εκλείπει.

2.7.Learn++NSE

Η Learn++NSE [37] αποτελεί συλλογή ταξινομητών η οποία εκπαιδεύεται με αυξητικό τρόπο. Το μοντέλο που κατασκευάζεται, συνδυάζει τα αποτελέσματα του κάθε ταξινομητή σε ένα αποτέλεσμα κατηγοριοποίησης με τη χρήση βαρών.

Τα σημαντικότερα προαπαιτούμενα για τη δημιουργία της συλλογής είναι το είδος των ταξινομητών και ο αριθμός των παραδειγμάτων που θα χρησιμοποιηθούν κάθε φορά που επανεκπαιδεύεται το μοντέλο. Για την εκπαίδευση του μοντέλου ακολουθείται η πιο κάτω διαδικασία για κάθε νέο δείγμα εκπαίδευσης:

I. Όλα τα παραδείγματα (N) του δείγματος εκπαίδευσης αξιολογούνται από το τρέχον μοντέλο, προβλέπεται η κλάση τους και υπολογίζεται το συνολικό (κανονικοποιημένο) σφάλμα E. Σε κάθε παράδειγμα αντιστοιχίζεται ένα βάρος. Αν το παράδειγμα έχει ταξινομηθεί σωστά το βάρος που του αντιστοιχίζεται είναι E/N ενώ αν έχει ταξινομηθεί λάθος 1/N. Τέλος όλα τα βάρη κανονικοποιούνται. Συνεπώς, αν το συνολικό σφάλμα είναι μεγάλο (δηλαδή πολλά παραδείγματα έχουν κατηγοριοποιηθεί λανθασμένα) η διαφορά του βάρους των λάθος και των σωστά κατηγοριοποιημένων παραδειγμάτων είναι μικρή. Στην αντίθετη περίπτωση, όπου το συνολικό σφάλμα είναι μικρό (λίγα παραδείγματα έχουν κατηγοριοποιηθεί λανθασμένα), η διαφορά του βάρους των λάθος και των σωστά κατηγοριοποιημένων παραδειγμάτων είναι (συγκριτικά) μεγάλη.

II. Εκπαιδεύεται ένας νέος ταξινομητής χρησιμοποιώντας όλο το νέο δείγμα εκπαίδευσης (αγνοώντας τα βάρη που έχουν αντιστοιχιστεί στα παραδείγματα εκπαίδευσης).

III. Τα παραδείγματα του δείγματος εκπαίδευσης αξιολογούνται από όλους του ταξινομητές της συλλογής (συμπεριλαμβανομένου και αυτού που μόλις δημιουργήθηκε) και βρίσκεται το σφάλμα (Σ) ταξινόμησης του κάθε ταξινομητή (με βάση το τρέχον δείγμα εκπαίδευσης). Ως σφάλμα (Σ) ορίζεται το άθροισμα των βαρών (βήμα I) των λάθος ταξινομημένων παραδειγμάτων. Συνεπώς στην περίπτωση όπου λίγα παραδείγματα έχουν κατηγοριοποιηθεί λανθασμένα από τη συλλογή και έχουν συγκριτικά μεγάλο βάρος (βήμα I), οι ταξινομητές που τα κατατάσσουν λανθασμένα αποκτούν μεγάλο σφάλμα. Στην αντίθετη περίπτωση, όπου πολλά παραδείγματα έχουν κατηγοριοποιηθεί λανθασμένα από τη συλλογή (βήμα I) και έχουν συγκριτικά μικρό βάρος, οι ταξινομητές που τα κατηγοριοποιούν λανθασμένα δεν δέχονται (συγκριτικά) τόσο μεγάλη ποινή. Αν το σφάλμα του νέου ταξινομητή είναι μεγαλύτερο από 0,5 τότε απορρίπτεται και στη θέση του δημιουργείται ένας νέος (βήμα II) και η διαδικασία αυτού του βήματος επαναλαμβάνεται. Αν το σφάλμα οποιουδήποτε άλλου ταξινομητή υπερβεί το 0,5 τότε μετατρέπεται σε 0,5.

IV. Σε κάθε ταξινομητή αντιστοιχίζεται ένα τρέχον βάρος w_i σύμφωνα με τη σχέση (4).

$$w_i = \frac{\Sigma}{1-\Sigma} \tag{4}$$

Το βάρος αυτό είναι κατασκευασμένο έτσι ώστε για μικρές τιμές του σφάλματος Σ να λαμβάνει μικρές τιμές κοντά στο 0, ενώ όταν το Σ τείνει στο 0,5 (μεγάλες τιμές σφάλματος) να λαμβάνει τιμές κοντά στο 1.

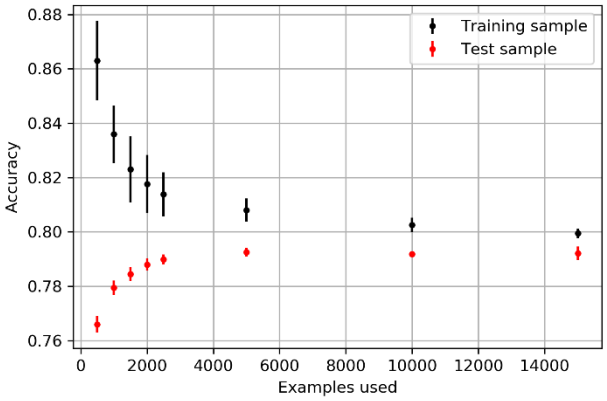
V. Όλα τα βάρη w_i που είχε λάβει ο ταξινομητής σε όσες επανεκπαιδεύσεις του μοντέλου έχει λάβει μέρος συνυπολογίζονται σε ένα βάρος w_t (ο ταξινομητής που δημιουργήθηκε με το εκάστοτε δείγμα εκπαίδευσης δεν έχει προγενέστερα βάρη w_i, ο ταξινομητής που εκπαιδεύτηκε με το προηγούμενο δείγμα εκπαίδευσης έχει το τρέχον και ένα προγενέστερο, κλπ). Για τον υπολογισμό όμως του w_t τα βάρη w_i συνυπολογίζονται με σιγμοειδή κατανομή, με αποτέλεσμα το τρέχον βάρος να παίζει το σημαντικότερο ρόλο και όλα τα υπόλοιπα να παίζουν πολύ μικρό ρολό.

VI. Τέλος το βάρος με το όποιο λαμβάνεται υπόψιν η εκτίμηση του κάθε ταξινομητή είναι ο αντίστροφος του w_t . Η αντιστροφή απαιτείται διότι τα w_i και κατ' επέκταση τα w_t παίρνουν μικρές τιμές (κοντά στο 0) όταν οι ταξινομητές κατηγοριοποιούν σωστά τα παραδείγματα εκπαίδευσης και μεγάλες (κοντά στο 1) όταν τα κατηγοριοποιούν λανθασμένα.

Η μέθοδος βάσει της οποίας υπολογίζονται τα βάρη που απονέμονται σε κάθε ταξινομητή δίνει ιδιαίτερη σημασία στο τελευταίο δείγμα με το οποίο εκπαιδεύεται το μοντέλο. Αυτό συμβαίνει διότι η μέθοδος Learn++NSE (όπου NSE είναι τα αρχικά των λέξεων NonStationary Environments) αναπτύχθηκε για την εκπαίδευση μοντέλων με αυξητικό τρόπο για την ταξινόμηση δειγμάτων που αλλάζουν με τον χρόνο. Οι ταξινομητές του μοντέλου που δεν είναι σε θέση να προβλέψουν καλά το τρέχον δείγμα εκπαίδευσης δεν καταστρέφονται, αλλά «αδρανοποιούνται» καθώς αποκτούν πολύ μικρό βάρος με αποτέλεσμα το προβλεπτικό μοντέλο να μην τους λαμβάνει υπόψιν. Αν κάποια στιγμή, σε μελλοντική επανεκπαίδευση του μοντέλου, αποδειχθεί ότι απέκτησαν ξανά καλή προβλεπτική ισχύ (λόγω μεταβολής των δεδομένων) το νέο βάρος που θα τους ανατεθεί θα «επανενεργοποιήσει» την συνεισφορά τους στο μοντέλο. Στη μέθοδο Learn++NSE εμπεριέχεται και η επιλογή του pruning, δηλαδή της απόρριψης κάποιων από τους

ταξινομητές που έχουν δημιουργηθεί. Ο χρήστης επιλέγει τον μέγιστο αριθμό των ταξινομητών και όταν ο αριθμός αυτός ξεπεραστεί, απορρίπτονται είτε οι ταξινομητές με το μεγαλύτερο σφάλμα είτε αυτοί που δημιουργήθηκαν παλαιότερα (σύμφωνα με την αντίστοιχη επιλογή του χρήστη). Ενώ η επιλογή αυτή προσφέρεται, οι συγγραφείς [37] προτρέπουν να μην χρησιμοποιείται, διότι απορρίπτοντας ταξινομητές μπορεί να χαθεί πληροφορία που εν δυνάμει στο μέλλον θα προσέφερε καλύτερη κατηγοριοποίηση. Ο μόνος λόγος που έχει νόημα ο περιορισμός των ταξινομητών είναι η γρηγορότερη απόκριση του μοντέλου τόσο κατά την αξιολόγηση νέων παραδειγμάτων όσο και κατά την επανεκπαίδευσή του (καθώς για την επανεκπαίδευση απαιτείται πρώτα η αξιολόγηση όλων των παραδειγμάτων εκπαίδευσης από τους ταξινομητές του μοντέλου – βήματα I και III).

Η υλοποίηση της συλλογής ταξινομητών Learn++NSE που εξετάστηκε, περιέχεται στο πακέτο scikit-multiflow [38] (σε γλώσσα python) το οποίο μπορεί να βρεθεί [39]. Ως ταξινομητές μπορούν να χρησιμοποιηθούν αυτοί που είναι διαθέσιμοι στο πακέτο scikit-multiflow αλλά και όσοι είναι διαθέσιμοι στο πακέτο scikit-learn. Ως προεπιλεγμένος ταξινομητής χρησιμοποιείται ένα δένδρο απόφασης (decision Tree Classifier) αλλά από την μελέτη που έχει διαταχθεί μέχρι αυτό το σημείο φαίνεται ότι τα καλύτερα αποτελέσματα επιτυγχάνονται με μηχανές διανυσμάτων στήριξης με χρήση της συνάρτησης μετασχηματισμού πυρήνα RBF. Συνεπώς ως ταξινομητής χρησιμοποιήθηκε ο ταξινομητής SVC [32] από το πακέτο scikit-learn, με χρήση της συνάρτησης μετασχηματισμού πυρήνα RBF. Ο αριθμός των παραδειγμάτων που χρησιμοποιούνται για κάθε επανεκπαίδευση του μοντέλου επιλέχθηκε στα 500 και δεν χρησιμοποιήθηκε pruning των ταξινομητών. Ως ταξινομητής που δεν εκπαιδεύεται με αυξητικό τρόπο επιλέχθηκε ο ίδιος ταξινομητής SVC.

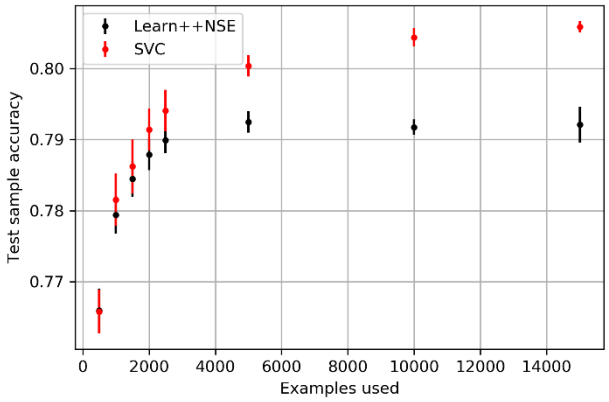


***Εικόνα 28**: Ακρίβεια του Learn++NSE. Κόκκινο: η ακρίβεια με βάση το δείγμα που χρησιμοποιήθηκε για την εκπαίδευση. Μαύρο: η ακρίβεια με βάση το test sample. Τα σημεία αντιπροσωπεύουν τη μέση τιμή της ακρίβειας για 10 επαναλήψεις ενώ το μήκος της γραμμής σφάλματος την τυπική απόκλιση.*

Στην **Εικόνα 28** παρουσιάζεται η ακρίβεια της Learn++NSE όπως υπολογίζεται τόσο από το δείγμα εκπαίδευσης (μαύρο) όσο και από το test sample (κόκκινο), όπου η διαδικασία της εκπαίδευσης και ελέγχου του ταξινομητή επαναλήφθηκε 10 φορές. Από το διάγραμμα φαίνεται ότι το μοντέλο παρουσιάζει overtraining κυρίως όταν ο αριθμός των παραδειγμάτων που έχουν χρησιμοποιηθεί για την εκπαίδευσή του είναι μικρός, ενώ για

μεγαλύτερα δείγματα εκπαίδευσης η τάση αυτή ελαττώνεται σημαντικά. Επιπλέον φαίνεται ότι η ακρίβεια του μοντέλου με βάση το test sample είναι πολύ σταθερή καθώς η τυπική απόκλιση των 10 επαναλήψεων είναι μικρότερη από 0,5%.

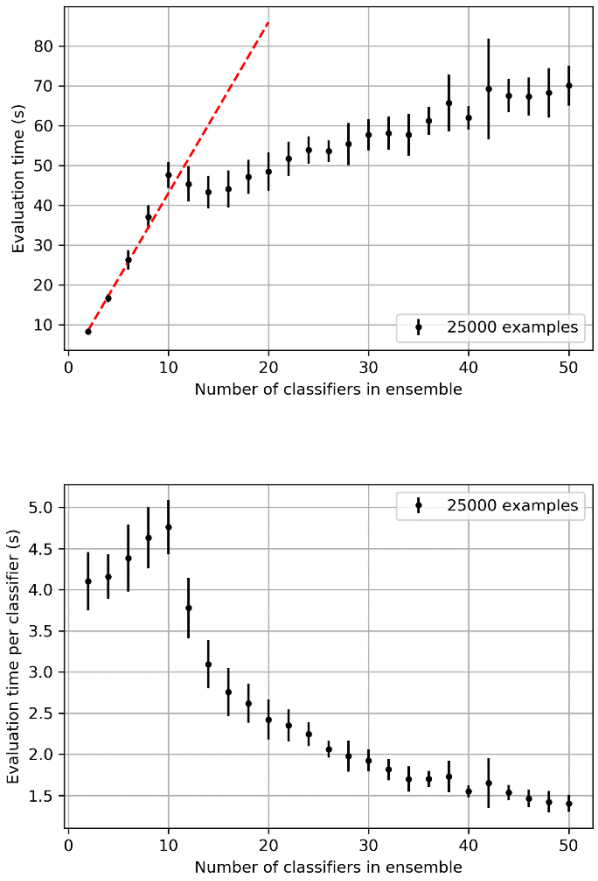
Στην **Εικόνα 29** παρουσιάζεται η σύγκριση της ακρίβειας του μοντέλου learn++NSE το όποιο για κάθε 500 παραδείγματα εκπαίδευσης επανεκπαιδεύεται (ακολουθώντας την διαδικασία που αναφέρθηκε πιο πάνω) και ενός ταξινομητή SVC στον οποίο έχουν χρησιμοποιηθεί όλα τα παραδείγματα εκπαίδευσης που αναγράφονται στον οριζόντιο άξονα του διαγράμματος, όπως υπολογίζεται από το test sample. Όπως ήταν αναμενόμενο, η ακρίβεια του μοντέλου learn++NSE και του ταξινομητή SVC είναι ίδια όταν έχει χρησιμοποιηθεί δείγμα εκπαίδευσης 500 παραδειγμάτων, καθώς η συλλογή ταξινομητών learn++NSE αποτελείται από μόνο 1 ταξινομητή SVC. Όσο το δείγμα εκπαίδευσης αυξάνει, η ακρίβεια τόσο του μοντέλου learn++NSE όσο και του ταξινομητή SVC αυξάνει, με την ακρίβεια του ταξινομητή SVC είναι μεγαλύτερη από αυτή του μοντέλου learn++NSE. Αυτό όμως δεν αποτελεί έκπληξη καθώς ο ταξινομητής SVC αξιοποιεί όλη την πληροφορία ενώ το μοντέλο learn++NSE δεν έχει διαθέσιμη όλη την πληροφορία στην ίδια χρονική στιγμή αλλά έχει πρόσβαση σε αυτή με αυξητικά. Παρόλα αυτά, για 15.000 παραδείγματα εκπαίδευσης (για τα οποία το μοντέλο learn++NSE έχει επανεκπαιδευτεί 30 φορές με δείγματα των 500 παραδειγμάτων και έχει δημιουργήσει αντιστοίχως 30 ταξινομητές) η διαφορά στην ακρίβεια είναι της τάξης του 1%.



***Εικόνα 29**: Σύγκρισή της ακρίβειας με βάση το test sample για τη συλλογή ταξινομητών Learn++NSE (μαύρο) και για τον ταξινομητή SVC (κόκκινο). Τα σημεία αντιπροσωπεύουν τη μέση τιμή της ακρίβειας και το μήκος των γραμμών σφάλματος την τυπική απόκλιση (10 επαναλήψεις).*

Όπως αναφέρθηκε πιο πάνω, στην υλοποίηση της μεθόδου Learn++NSE που εξετάστηκε δεν τέθηκε περιορισμός στον αριθμό των ταξινομητών SVC που μπορούν να δημιουργηθούν. Στην **Εικόνα 30** (άνω) παρουσιάζεται ο χρόνος που απαιτείται για την αξιολόγηση των 25.000 παραδειγμάτων του test sample που χρησιμοποιήθηκε, ως συνάρτηση του αριθμού των ταξινομητών, ενώ κάτω παρουσιάζεται ο χρόνος που απαιτείται ανά ταξινομητή. Από το πάνω διάγραμμα φαίνεται ότι για μικρό αριθμό ταξινομητών (έως 10) ο μέσος χρόνος ανά ταξινομητή είναι γραμμική συνάρτηση του αριθμού των ταξινομητών ενώ όσο ο αριθμός των ταξινομητών αυξάνεται η κλίση της ελαττώνεται. Από το κάτω

διάγραμμα στην **Εικόνα 30**, όπου απεικονίζεται ο χρόνος που απαιτείται για την αξιολόγηση των 25.000 παραδειγμάτων του test sample ανά ταξινομητή, είναι φανερό ότι όσο αυξάνει ο αριθμός των ταξινομητών ο χρόνος ανά ταξινομητή συνεχώς ελαττώνεται. Συνεπώς, λαμβάνοντας υπόψιν ότι για 50 ταξινομητές απαιτούνται κατά μέσο όρο 60 δευτερόλεπτα για την αξιολόγηση 25.000 παραδειγμάτων και συνυπολογίζοντας το γεγονός ότι όσο αυξάνει ο αριθμός των ταξινομητών (τουλάχιστον μέχρι τους 50 ταξινομητές που αποτελούν τη τελική συλλογή που μελετήθηκε), ο χρόνος που απαιτείται για την αξιολόγηση γεγονότων ανά ταξινομητή ελαττώνεται, δεν θα τεθεί περιορισμός στον αριθμό των ταξινομητών SVC που μπορούν να δημιουργηθούν.



***Εικόνα 30**: Άνω: Ο χρόνος που απαιτείται για την αξιολόγηση των 25.000 παραδειγμάτων του test sample συναρτήσει του αριθμού των ταξινομητών. Κάτω: Ο χρόνος που απαιτείται ανά ταξινομητή για την αξιολόγηση των 25.000 παραδειγμάτων του test sample συναρτήσει του αριθμού των ταξινομητών. Τα σημεία αντιπροσωπεύουν τη μέση τιμή της ακρίβειας για 10 επαναλήψεις ενώ το μήκος της γραμμής σφάλματος την τυπική απόκλιση.*

Συνοψίζοντας η Learn++NSE συλλογή ταξινομητών μηχανών διανυσμάτων στήριξης SVC με χρήση της συνάρτησης μετασχηματισμού πυρήνα RBF συγκεντρώνει πολλά από τα χαρακτηριστικά που αναζητούμε σε ταξινομητές που εκπαιδεύονται με αυξητικό τρόπο. Η ακρίβεια του ταξινομητή φτάνει σχεδόν στο 80%, ενώ ταυτόχρονα φαίνεται να μην εξαρτάται

από το αριθμό των παραδειγμάτων εκπαίδευσης που χρησιμοποιούνται. Επίσης, όταν εκπαιδεύεται με μικρά δείγματα φαίνεται να υποφέρει από overtraining ενώ όταν ο αριθμός των παραδειγμάτων μεγαλώνει η συμπεριφορά αυτή εκλείπει. Τέλος φαίνεται ότι όσο αυξάνει ο αριθμός των παραδειγμάτων εκπαίδευσης η απόδοση της συλλογής όπως υπολογίζεται από το test sample αυξάνει.

2.8. Επιλογή ταξινομητή

Από τους ταξινομητές που εξετάστηκαν σε αυτό το κεφάλαιο υπάρχουν ορισμένοι που έχουν πολύ καλά αποτελέσματα όπως ακρίβεια και σταθερότητα. Για την επιλογή του ταξινομητή με τον όποιο θα συνεχιστεί η παρούσα εργασία θα χρησιμοποιηθούν τα ακόλουθα κριτήρια:

1) Αυξητική συμπεριφορά. Με την αύξηση του δείγματος εκπαίδευσης θα πρέπει να αυξάνεται και η απόδοση όπως υπολογίζεται από το test sample – ποιοτικό κριτήριο.
2) Σταθερότητα. Θα πρέπει η διακύμανση της ακρίβειας (όπως υπολογίζεται από το test sample) να είναι μικρή (της τάξης του 1%) ιδίως όταν το δείγμα εκπαίδευσης είναι μεγάλο (μεγαλύτερο από 10000 παραδείγματα) – ποσοτικό κριτήριο
3) Overtraining. Θα πρέπει να μην παρουσιάζει overtraining όταν το δείγμα εκπαίδευσης είναι μεγάλο – ποιοτικό κριτήριο.
4) Ακρίβεια. Θα πρέπει να έχει την μέγιστη δυνατή ακρίβεια όταν έχει χρησιμοποιηθεί όλο το διαθέσιμο δείγμα εκπαίδευσης – ποιοτικό κριτήριο.
5) Τρόπος εκπαίδευσης. Οι ταξινομητές που εξετάστηκαν εκπαιδεύονται αυξητικά με δύο τρόπους. Ο πρώτος τρόπος είναι ο online, δηλαδή με κάθε νέο παράδειγμα ο ταξινομητής επανεκπαιδεύεται. Ο δεύτερος τρόπος είναι αυτός κατά τον οποίο η επανεκπαίδευση συμβαίνει με ομάδες παραδειγμάτων. Με τον Online τρόπο ο ταξινομητής είναι πάντα εκπαιδευμένος, χρησιμοποιώντας όλα τα διαθέσιμα παραδείγματα μέχρι εκείνη τη στιγμή. Βέβαια χρησιμοποιώντας ένα μόνο παράδειγμα σε κάθε επανεκπαίδευση ίσως να μην εύκολη ή ακόμα και δυνατή η σύγκλισή του και η εύρεση της βέλτιστης δυνατής λύσης. Από τους δυο τρόπους θα προτιμηθεί ο τρόπος που χρησιμοποιεί ομάδες παραδειγμάτων κατά την επανακπαίδευση – ποιοτικό κριτήριο.

Στον **Πίνακας 2** παρουσιάζονται οι έξι ταξινομητές που εξετάστηκαν σε αυτό το κεφάλαιο και η αξιολόγησή τους με βάση τα πιο πάνω κριτήρια. Ο ταξινομητής LaSVM (ενότητα 2.5) εξετάστηκε χρησιμοποιώντας συναρτήσεις μετασχηματισμού πυρήνα: RBF, σιγμοειδή, πολυωνυμική και γραμμική. Με χρήση των δυο τελευταίων, τα αποτελέσματα δεν παρουσιάζουν αυξητική συμπεριφορά και η ακρίβειά τους είναι μικρή, ενώ η ακρίβεια που επιτεύχθηκε με χρήση της RBF ήταν (λίγο) μεγαλύτερη από αυτή που επιτεύχθηκε με χρήση της σιγμοειδούς. Συνεπώς θα αξιολογηθεί ο ταξινομητής LaSVM μόνο με χρήση της συνάρτησης μετασχηματισμού πυρήνα RBF.

Από τους 6 ταξινομητές απορρίπτονται οι Gaussian Naïve Bayesian, Support Vector Machine με Stochastic Gradient Descent και ILVQ λόγω της μεγάλης τυπικής απόκλισης που παρουσιάζει η ακρίβειά τους. Από τους τρείς εναπομείναντες ο online random forest έχει την

μικρότερη ακρίβεια (70%) και έτσι απορρίπτεται. Ο ταξινομητής LaSVM με συνάρτηση μετασχηματισμού πυρήνα RBF και η συλλογή Learn++NSE χρησιμοποιώντας τον ταξινομητή SVC με συνάρτηση μετασχηματισμού πυρήνα RBF έχουν συγκρίσιμη ακρίβεια (81% και 79% αντίστοιχα). Και οι δύο παρουσιάζουν σημεία overtraining όταν εκπαιδεύονται με μικρά δείγματα ενώ όταν το δείγμα εκπαίδευσης αυξάνει το overtraining εξαλείφεται. Η μόνη αξιοσημείωτη διαφορά είναι ο τόπος εκπαίδευσής τους (online και με ομάδες παραδειγμάτων αντίστοιχα). Συνεπώς ο ταξινομητής που επιλέγεται είναι η Learn++NSE συλλογή ταξινομητών SVC με συνάρτηση μετασχηματισμού πυρήνα RBF.

Ταξινομητής	Κριτήρια				
	Αυξητική συμπεριφορά	Σταθερότητα	Overtraining	Ακρίβεια	Τρόπος εκπαίδευσης
Gaussian Naïve Bayesian	Ναι	2%	Όχι	67%	Ομάδες / Online
Support Vector Machine με Stochastic Gradient Descent	Ναι	8%	Όχι	72%	Ομάδες
Online Random Forest	Ναι	> 0,5%	Όχι	70%	Online
LaSVM με συνάρτηση μετασχηματισμού πυρήνα RBF	Ναι	> 0,5%	Όχι (για μεγάλα δείγματα)	81%	Online
ILVQ	Ναι	2%	Όχι (για μεγάλα δείγματα)	70%	Online
Learn++NSE χρησιμοποιώντας τον ταξινομητή SVC με RBF	Ναι	> 0,5%	Όχι (για μεγάλα δείγματα)	79%	Ομάδες

***Πίνακας 2**: Οι ταξινομητές που εξετάστηκαν και η αξιολόγησή τους με βάση τα 4 κριτήρια που τέθηκαν.*

3 Διερεύνηση της συλλογής ταξινομητών Learn++NSE

3.1. Γενικά

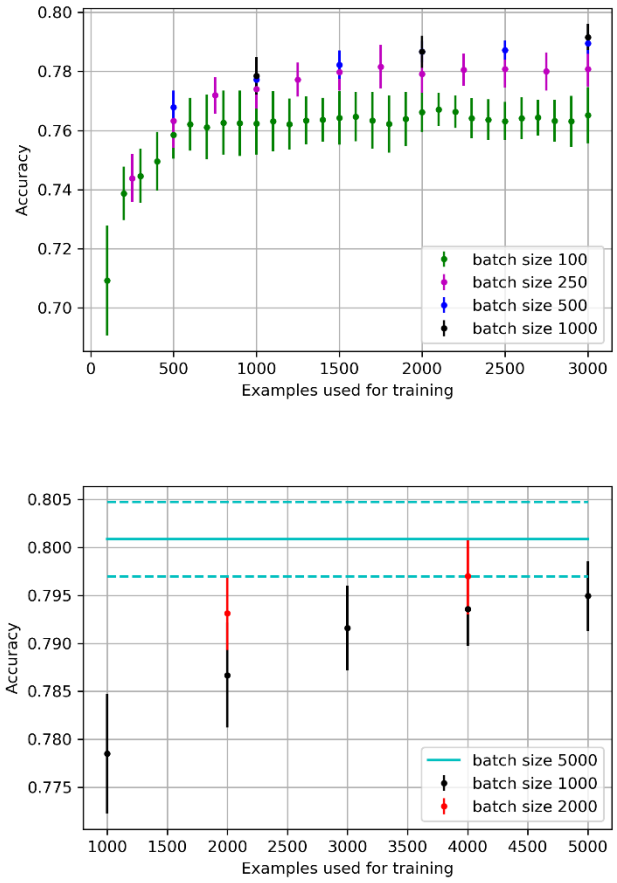
Στην προηγούμενη ενότητα εξετάστηκαν 6 ταξινομητές με στόχο την εύρεση του βέλτιστου για την κατηγοριοποίηση μικρών κειμένων, με βάση το συναίσθημά τους, με αυξητικό τρόπο και επιλέχθηκε η συλλογή ταξινομητών Learn++NSE (για χάρη συντομίας στο εξής θα αναφέρεται και ως "ταξινομητής Learn++NSE"). Σε αυτή την ενότητα θα εξεταστεί περισσότερο λεπτομερώς η συμπεριφορά του ταξινομητή Learn++NSE στις περιπτώσεις όπου η σύσταση των κλάσεων μεταβάλλονται με την πάροδο του χρόνου και όπου ο αριθμός των παραδειγμάτων των κλάσεων διαφέρει σημαντικά και μεταβάλλεται με τον χρόνο.

3.2. Βελτιστοποίηση των παραμέτρων του ταξινομητή Learn++NSE

Όπως ήδη αναφέρθηκε στην ενότητα 2.7, από τα σημαντικότερα προαπαιτούμενα για τη δημιουργία της συλλογής ταξινομητών learn++NSE είναι το είδος των ταξινομητών και ο αριθμός των παραδειγμάτων που θα χρησιμοποιηθούν κάθε φορά που επανεκπαιδεύεται το μοντέλο. Ως ταξινομητές της συλλογής έχουν ήδη επιλεγεί μηχανές διανυσμάτων στήριξης SVC με μετασχηματισμό πυρήνα RBF. Για την επιλογή του βέλτιστου αριθμού παραδειγμάτων τα οποία θα χρησιμοποιηθούν κάθε φορά που επανεκπαιδεύεται το μοντέλο (batch size) διεξήχθησαν πειραματισμοί. Πιο συγκεκριμένα, ελέγχθηκε η απόδοση του ταξινομητή ο οποίος εκπαιδεύτηκε με τμήμα του δείγματος εκπαίδευσης που αποτελούταν από 5.000 παραδείγματα, όταν χρησιμοποιήθηκαν batch sizes 100, 250, 500, 1.000 και 2.000. Για τον έλεγχο της απόδοσης χρησιμοποιήθηκε τμήμα του test sample που αποτελούταν από 15.000 παραδείγματα. Η πιο πάνω διαδικασία επαναλήφθηκε 10 φορές και πριν από κάθε επανάληψη η επιλογή των δειγμάτων εκπαίδευσης και ελέγχου έγινε με τυχαίο τρόπο από τα δείγματα εκπαίδευσης και ελέγχου του dataset «Large Movie Review Dataset v1.0».

Στην **Εικόνα 31** άνω, παρουσιάζεται η ακρίβεια του ταξινομητή Learn++NSE, όπως βρέθηκε από το δείγμα ελέγχου, ως προς τον αριθμό των παραδειγμάτων εκπαίδευσης που έχουν χρησιμοποιηθεί, όταν κατά την εκπαίδευση χρησιμοποιήθηκε batch size 100 (πράσινο), 250 (μοβ), 500 (μπλε) και 1000 (μαύρο). Στην **Εικόνα 31** κάτω, παρουσιάζονται τα αντίστοιχα αποτελέσματα, όταν κατά την εκπαίδευση χρησιμοποιήθηκε batch size 1000 (μαύρο,) 2000 (κόκκινο) ενώ με τις γαλάζιες γραμμές παρουσιάζεται η ακρίβεια ταξινόμησης όταν ο ταξινομητής εκπαιδεύτηκε με όλο το δείγμα εκπαίδευσης (αριθμός παραδειγμάτων = batch size = 5.000). Από τα διαγράμματα στην **Εικόνα 31** φαίνεται ότι ανεξαρτήτως του batch size, η ακρίβεια αυξάνει όσο αυξάνει ο αριθμός των παραδειγμάτων εκπαίδευσης και μετά σταθεροποιείται (όπως έχει ήδη διαπιστωθεί από την **Εικόνα 28** και την **Εικόνα 29**). Επίσης παρατηρείται ότι όσο αυξάνει το batch size αυξάνει και η μέγιστη ακρίβεια του ταξινομητή. Όταν το batch size είναι 100, η μέγιστη ακρίβεια είναι λίγο μεγαλύτερη από 76% ενώ όταν είναι 500, η μέγιστη ακρίβεια φτάνει στο 79%. Όταν το batch size είναι 1.000, η

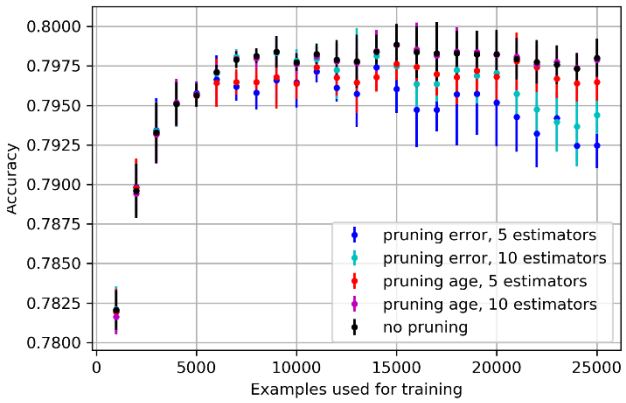
μέγιστη ακρίβεια φτάνει στο 79.5% ενώ για batch size 2.000, έχουν εκτελεστεί μόνο 2 επανεκπαιδεύσεις και η μέγιστη τιμή της ακρίβειας ξεπερνά το 79,5% χωρίς να έχει επιτευχθεί η μεγιστοποίηση της τιμή της. Τελικά για την εύρεση του βέλτιστου batch size πρέπει να συνεκτιμηθεί η πιο πάνω συμπεριφορά με την απαίτηση της αυξητικής εκπαίδευσης του ταξινομητή (ειδικά δεδομένου ότι το δείγμα εκπαίδευσης που χρησιμοποιείται αποτελείται από 25.000 παραδείγματα) η τιμή του batch size δεν πρέπει να είναι συγκριτικά μεγάλη. Τελικά αποφασίστηκε η χρήση του batch size = 1.000, ώστε να ικανοποιείται ταυτόχρονα η επιθυμία για τη μέγιστη δυνατή ακρίβεια με την απαίτηση του περιορισμένου batch size.



***Εικόνα 31**: Ακρίβεια του ταξινομητή Learn++NSE όπως βρέθηκε από το δείγμα ελέγχου ως προς τον αριθμό των παραδειγμάτων εκπαίδευσης. Άνω: batch size 100 (πράσινο), 250 (μοβ), 500 (μπλε) και 1000 (μαύρο). Κάτω batch size 1.000 (μαύρο), 2.000 (κόκκινο) και 5.000 (μπλε). Τα σημεία αντιπροσωπεύουν τη μέση τιμή της ακρίβειας για 10 επαναλήψεις, ενώ το μήκος της γραμμής σφάλματος την τυπική απόκλιση. Η γαλάζια συνεχής γραμμή αντιπροσωπεύει τη μέση τιμή της ακρίβειας (10 επαναλήψεις), ενώ οι διακεκομμένες γραμμές τη μέση τιμή στην οποία έχει προστεθεί η αφαιρεθεί η τιμή της τυπικής απόκλισης.*

Ήδη στην ενότητα 2.7, αναφέρθηκε ότι κατά τη δημιουργία της συλλογής ταξινομητών Learn++NSE δίνεται η επιλογή της απόρριψης ταξινομητών από τη συλλογή (pruning), αλλά συνιστάται από τους συγγραφείς [37] η αποφυγή της. Επίσης από την **Εικόνα**

30 όπου παρουσιάζεται ο χρόνος αξιολόγησης του δείγματος ελέγχου (για 25.000 παραδείγματα) ως προς τον αριθμό των ταξινομητών της συλλογής, κρίθηκε ότι η χρονική επιβάρυνση κατά την διαδικασία ελέγχου του test sample που προκύπτει από τη χρήση πολλών ταξινομητών δεν είναι αξιόλογη. Για λόγους πληρότητας, στην **Εικόνα 32** παρουσιάζεται σύγκριση της ακρίβειας (όπως υπολογίζεται από δείγμα ελέγχου) ως προς τον αριθμό των παραδειγμάτων εκπαίδευσης όταν έχει ενεργοποιηθεί η απόρριψη ταξινομητών από τη συλλογή με βάση το σφάλμα τους ή με βάση την παλαιότητά τους αλλά και όταν δεν απορρίπτεται κανένας ταξινομητής. Το batch size που χρησιμοποιήθηκε ήταν 1.000. Για την παραγωγή του διαγράμματος η διαδικασία εκπαίδευσης και ελέγχου επαναλήφθηκε 5 φορές και πριν από κάθε επανάληψη τα δείγματα ανακατεύονταν με τυχαίο τρόπο.



***Εικόνα 32**: Ακρίβεια του ταξινομητή learn++NSE όπως υπολογίζεται από το δείγμα ελέγχου, ως συνάρτηση του αριθμού παραδειγμάτων εκπαίδευσης. Μαύρο: δεν απορρίπτονται ταξινομητές από τη συλλογή. Μπλε: απορρίπτονται οι ταξινομητές με το μεγαλύτερο σφάλμα και ο μέγιστος αριθμός ταξινομητών στην συλλογή είναι 5. Γαλάζιο: απορρίπτονται οι ταξινομητές με το μεγαλύτερο σφάλμα και ο μέγιστος αριθμός ταξινομητών στην συλλογή είναι 10. Κόκκινο: απορρίπτονται οι παλαιότεροι ταξινομητές και ο μέγιστος αριθμός ταξινομητών στην συλλογή είναι 5. Μοβ: απορρίπτονται οι παλαιότεροι ταξινομητές και ο μέγιστος αριθμός ταξινομητών στην συλλογή είναι 10. Σε όλες τις περιπτώσεις χρησιμοποιήθηκε batch size 1.000. Τα σημεία αντιπροσωπεύουν την μέση τιμή 5 επαναλήψεων ενώ το μήκος των σφαλμάτων την τυπική απόκλιση.*

Από την **Εικόνα 32** φαίνεται ότι επιλέγοντας τη μη απόρριψη ταξινομητών από τη συλλογή, η ακρίβεια ταξινόμησης είναι υψηλότερη. Επιπλέον στις περιπτώσεις που απορρίπτονται ταξινομητές (είτε με βάση το σφάλμα τους είτε με βάση την παλαιότητά τους) προκύπτουν καλύτερα αποτελέσματα όταν ο αριθμός των ταξινομητών της συλλογής είναι μεγαλύτερος.

Σύμφωνα με την διερεύνηση που διεξήχθη σε αυτή την παράγραφο, στο εξής η συλλογή ταξινομητών learn++NSE θα χρησιμοποιείται με batch size 1.000 και χωρίς απόρριψη ταξινομητών από τη συλλογή. Σε οποιαδήποτε άλλη περίπτωση οι παράμετροι θα αναφέρονται ρητά στο κείμενο.

3.3. Αξιολόγηση του ταξινομητή σε μεταβαλλόμενο περιβάλλον – Μεταβολή της σύστασης των κλάσεων

Η απόδοση του ταξινομητή learn++NSE που εκπαιδεύεται με αυξητικό τρόπο στην περίπτωση που τα δείγματα εκπαίδευσης και ελέγχου είναι σχετικά ομογενή εξετάστηκε στις ενότητες 2.7 και 3.2. Ενδιαφέρον όμως παρουσιάζει και η μελέτη της απόδοσης του ταξινομητή όταν τα δείγματα εκπαίδευσης και ελέγχου δεν είναι ομογενή, αλλά μεταβάλλονται με τον χρόνο.

Ένα από τα κριτήρια στα οποία βασίστηκε η επιλογή του ταξινομητή learn++NSE ήταν ο τρόπος εκπαίδευσης. Για την επιλογή του βέλτιστου ταξινομητή μεγαλύτερη βαρύτητα δόθηκε στον αυξητικό τρόπο με ομάδες δεδομένων και όχι στον online τρόπο. Αυτή η προσέγγιση δίνει την δυνατότητα να επιλεγεί η στιγμή που θα εκτελεστεί η επανεκπαίδευση του ταξινομητή, αντί να επανεκπαιδεύεται με κάθε νέο παράδειγμα. Από την **Εικόνα 31** και την **Εικόνα 32** φαίνεται ότι όταν ο ταξινομητής εκπαιδευτεί με δείγμα που αποτελείται από περίπου 5.000 παραδείγματα επιτυγχάνεται η μέγιστη δυνατή ακρίβεια ενώ όταν χρησιμοποιούνται περισσότερα παραδείγματα, η τιμή της ακρίβειας παραμένει πρακτικά σταθερή. Βέβαια, στην περίπτωση που μελετήθηκε μέχρι στιγμής, τόσο το δείγμα εκπαίδευσης όσο και το δείγμα ελέγχου είναι ομογενή δείγματα. Αυτή η περίπτωση δεν είναι η γενική, στην γενικότερη όμως περίπτωση, είναι πιθανό τα παραδείγματα που αποτελούν κάθε κλάση να μεταβάλλονται με τον χρόνο. Για παράδειγμα θα ήταν δυνατόν στα πρώτα δείγματα με τα οποία εκπαιδεύτηκε ο ταξινομητής τα παραδείγματα που αποτελούν τη θετική κλάση να είναι κατά πλειοψηφία κριτικές με βαθμό 10, ενώ με την πάροδο του χρόνου η θετική κλάση να αλλάζει και να αποτελείται, κατά πλειοψηφία, από παραδείγματα με βαθμό 9, 8 ή και 7. Σε αυτή την περίπτωση η ακρίβεια αναμένεται να μεταβάλλεται καθώς μεταβάλλεται και η σύσταση των κλάσεων. Συνεπώς πρέπει να βρεθεί κάποιο κριτήριο το οποίο θα λειτουργεί ως εκκινητής της διαδικασίας επανεκπαίδευσης του ταξινομητή.

Ο εκκινητής της διαδικασίας επανεκπαίδευσης θα πρέπει να είναι τέτοιος ώστε η διαδικασία να ξεκινά όταν η σύσταση των κλάσεων του υπό εξέταση δείγματος έχει μεταβληθεί τόσο, ώστε η μεταβολή της ακρίβειας να είναι σημαντική. Για να εξεταστεί το πόσο μεταβάλλεται η ακρίβεια, λόγω των στατιστικών διακυμάνσεων τόσο του δείγματος εκπαίδευσης όσο και του δείγματος ελέγχου, εκπαιδεύτηκε μία συλλογή ταξινομητών learn++NSE με δείγμα 5.000 παραδειγμάτων (συνεπώς αποτελούταν από 5 ταξινομητές) και στην συνέχεια βρέθηκε η ακρίβεια ταξινόμησης του δείγματος ελέγχου το οποίο αποτελούταν από 2.000 παραδείγματα. Η διαδικασία επαναλήφθηκε 100 φόρες και πριν από κάθε επανάληψη τόσο το δείγμα εκπαίδευσης όσο και το δείγμα ελέγχου επιλέγονταν με τυχαίο τρόπο από τα δείγματα εκπαίδευσης και ελέγχου, αντίστοιχα, του dataset «Large Movie Review Dataset v1.0». Από την διαδικασία αυτή, βρέθηκε ότι η τυπική απόκλιση της ακρίβειας ταξινόμησης ήταν ~0,9%. Για την εκκίνηση της διαδικασίας επανεκπαίδευσης αποφασίστηκε η ακόλουθη διαδικασία:

1) Βρίσκεται η ακρίβεια ταξινόμησης με το τελευταίο δείγμα εκπαίδευσης (κρίσιμη ακρίβεια).
2) Εκτελείται έλεγχος του ταξινομητή με 2.000 παραδείγματα για τα οποία είναι γνωστή η κλάση τους.

3) Η διαδικασία επανεκπαίδευσης ξεκινά αν η απόλυτη τιμή της διαφοράς της ακρίβειας, όπως υπολογίστηκε από τα βήματα 1 και 2 είναι μεγαλύτερη από 4.5% (δηλαδή 5 φορές η τιμή της τυπικής απόκλισης).

Κατά την πρώτη εκπαίδευση του ταξινομητή χρησιμοποιείται δείγμα 5.000 παραδειγμάτων, ώστε να έχει επιτευχθεί (όσο είναι δυνατόν) η μέγιστη ακρίβεια.

Στον **Πίνακας 3** παρουσιάζεται ο αριθμός των παραδειγμάτων εκπαίδευσης και ελέγχου που έχουν λάβει κριτική με βαθμολογία 1, 2, 3, 4 (αρνητική) και 7, 8, 9 και 10 (θετική) του dataset «Large Movie Review Dataset v1.0». Για την προσομοίωση της μεταβολής της σύστασης των κλάσεων με την πάροδο του χρόνου, ενοποιήθηκαν το training και το test sample και στην συνέχεια το dataset χωρίστηκε σε δείγματα των 2.000 παραδειγμάτων με τέτοιο τρόπο ώστε η σύσταση και των δυο κλάσεων να αλλάζουν με τον χρόνο.

Κλάση	Βαθμός κριτικής	Αριθμός παραδειγμάτων δείγματος εκπαίδευσης	Αριθμός παραδειγμάτων δείγματος ελέγχου	άθροισμα
Αρνητική	1	5.100	5.022	10.122
	2	2.284	2.302	4.586
	3	2.420	2.541	4.961
	4	2.696	2.635	5.331
	σύνολο	12.500	12.500	25.000
Θετική	7	2.496	2.307	4.803
	8	3.009	2.850	5.859
	9	2.263	2.344	4.607
	10	4.372	4.999	9.371
	σύνολο	12.500	12.500	25.000

***Πίνακας 3**: Αριθμός των παραδειγμάτων των δειγμάτων εκπαίδευσης και ελέγχου που έχουν λάβει κριτική με βαθμολογία 1, 2, 3, 4 (αρνητική) και 7, 8, 9 και 10 (θετική) του dataset «Large Movie Review Dataset v1.0».*

Στο πρώτο dataset που δημιουργήθηκε (στο εξής θα αναφέρεται και ως dataset d1) η θετική (αρνητική) κλάση αποτελείται αρχικά από κριτικές με πολύ θετική (πολύ αρνητική) βαθμολογία 10 (βαθμολογία 1) και με την πάροδο του χρόνου οι βαθμολογίες των κριτικών των παραδειγμάτων γίνονταν χειρότερες καταλήγοντας στο 7 (καλύτερες καταλήγοντας στο 4). Με τον τρόπο αυτό επιτεύχθηκε η προσομοίωση κλάσεων που η σύστασή τους αλλάζει αργά με τον χρόνο. Σε κάθε δείγμα ο αριθμός των παραδειγμάτων με βαθμολογίες 1 και 10, 2 και 9, 3 και 8, 4 και 7 ήταν ίδιος. Στον **Πίνακας 4** παρουσιάζεται η σύσταση των κλάσεων σε κάθε δείγμα του dataset d1. Το d1 αποτελείται από 5.000 παραδείγματα με βαθμολογία 1, 2, 3, 8, 9 και 10 και από 3.000 παραδείγματα με βαθμολογία 4 και 7. Λόγω του περιορισμένου αριθμού των παραδειγμάτων με βαθμολογίες 2, 3, και 9, για την δημιουργία του dataset d1 επιλέχθηκαν με τυχαίο τρόπο 414, 39 και 393 παραδείγματα με βαθμολογίες 2, 3 και 9 αντίστοιχα και επανατοποθετήθηκαν σε αυτό.

Το dataset d1 χρησιμοποιήθηκε για να βρεθεί η ακρίβεια ταξινόμησης του Learn++NSE που επανεκπαιδευόταν όταν ικανοποιείτο η συνθήκη του εκκινητή που περιγράφηκε πιο πάνω. Αυτή συγκρίθηκε με την ακρίβεια ταξινομητή Learn++NSE που δεν επανεκπαιδεύτηκε και με αυτή ενός ταξινομητή svm (ταξινομητής SVC με ίδια

χαρακτηριστικά με αυτούς που απαρτίζουν την συλλογή ταξινομητών learn++NSE) που εκπαιδευόταν με όλα τα (μέχρι τότε) διαθέσιμα παραδείγματα.

Για τον ταξινομητή Learn++NSE ακολουθήθηκε η πιο κάτω διαδικασία:

a) Το δείγμα πρώτης εκπαίδευσης χωρίστηκε σε δυο τμήματα, ένα των 5.000 παραδειγμάτων και ένα των 1.000. Ο ταξινομητής εκπαιδεύτηκε με το πρώτο δείγμα, το οποίο χρησιμοποιήθηκε για να βρεθεί και η ακρίβεια εκπαίδευσης, ενώ το δεύτερο δείγμα χρησιμοποιήθηκε μόνο για την εύρεση της κρίσιμης ακρίβειας του εκκινητή.
b) Κάθε επόμενο δείγμα χρησιμοποιήθηκε για την εύρεση της ακρίβειας ταξινόμησης. Στην περίπτωση που η ακρίβεια είχε τέτοια τιμή που ικανοποιούσε την συνθήκη του εκκινητή, ο ταξινομητής επανεκπαιδευόταν χρησιμοποιώντας το εκάστοτε δείγμα. Επίσης, η ακρίβεια ταξινόμησης (δηλαδή η νέα κρίσιμη ακρίβεια του εκκινητή) βρισκόταν με το ίδιο δείγμα.

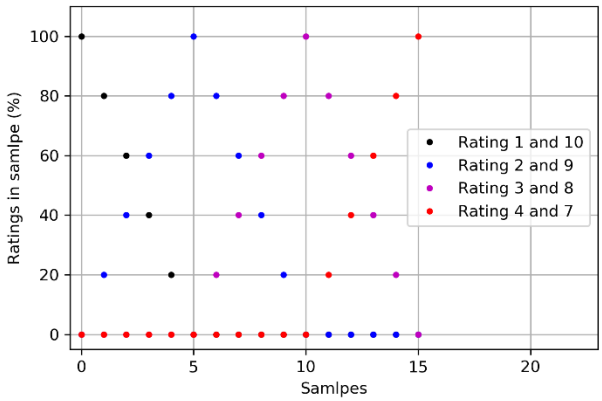
Ο ταξινομητής learn++NSE που δεν επανεκπαιδευτηκε, εκπαιδεύτηκε μόνο μια φορά με το τμήμα των 5.000 παραδειγμάτων του δείγματος πρώτης εκπαίδευσης και όλα τα υπόλοιπα δείγματα χρησιμοποιήθηκαν μόνο για την εύρεση της ακρίβειας ταξινόμησης.

Τέλος, ο ταξινομητής svm εκπαιδευόταν με όλα τα προηγούμενα δείγματα και η ακρίβεια ταξινόμησης υπολογιζόταν με βάση το τρέχον δείγμα. Για παράδειγμα κατά την επεξεργασία του Δείγματος 5, ο ταξινομητής svm εκπαιδεύτηκε χρησιμοποιώντας όλα τα παραδείγματα των δειγμάτων πρώτης εκπαίδευσης, 1, 2, 3 και 4 και η ακρίβεια ταξινόμησής του υπολογίστηκε από το δείγμα 5.

Dataset d1									
	Θετική κλάση					Αρνητική κλάση			
Βαθμολογία	10	9	8	7		4	3	2	1
Πρώτη Εκπαίδευση	3000	0	0	0		0	0	0	3000
Δείγμα 1	800	200	0	0		0	0	200	800
Δείγμα 2	600	400	0	0		0	0	400	600
Δείγμα 3	400	600	0	0		0	0	600	400
Δείγμα 4	200	800	0	0		0	0	800	200
Δείγμα 5	0	1000	0	0		0	0	1000	0
Δείγμα 6	0	800	200	0		0	200	800	0
Δείγμα 7	0	600	400	0		0	400	600	0
Δείγμα 8	0	400	600	0		0	600	400	0
Δείγμα 9	0	200	800	0		0	800	200	0
Δείγμα 10	0	0	1000	0		0	1000	0	0
Δείγμα 11	0	0	800	200		200	800	0	0
Δείγμα 12	0	0	600	400		400	600	0	0
Δείγμα 13	0	0	400	600		600	400	0	0
Δείγμα 14	0	0	200	800		800	200	0	0
Δείγμα 15	0	0	0	1000		1000	0	0	0

***Πίνακας 4**: Αριθμός παραδειγμάτων από κάθε βαθμολογία που συγκροτούν τα δείγματα του dataset d1.*

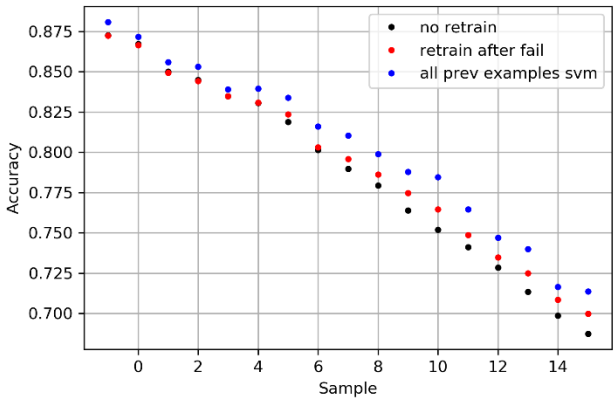
Στην **Εικόνα 33** απεικονίζεται η ποσοστιαία σύσταση της βαθμολογίας των κριτικών που αποτελούν τις δύο κλάσεις κάθε δείγματος. Το δείγμα πρώτης εκπαίδευσης παρουσιάζεται στη θέση 0. Η διαδικασία που περιεγράφηκε πιο πάνω εκτελέστηκε 5 φορές και πριν από κάθε εκτέλεση τα δείγματα του dataset d1 εξάγονταν από το dataset «Large Movie Review Dataset v1.0» με τυχαίο τρόπο.



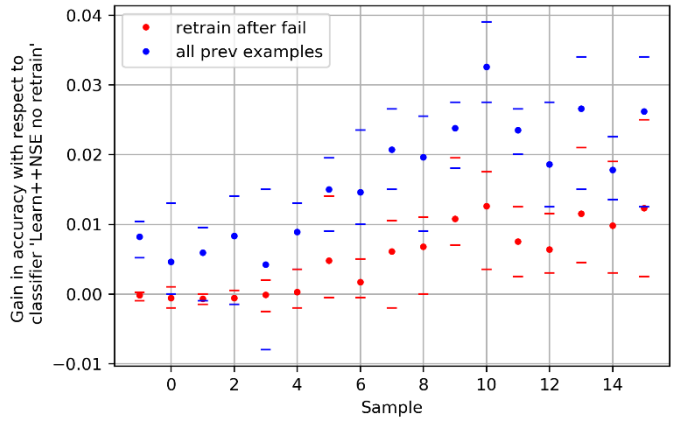
***Εικόνα 33**: Ποσοστιαία σύσταση της βαθμολογίας των κριτικών που αποτελούν τις δύο κλάσεις κάθε δείγματος του dataset d1. Ως sample 0 παρουσιάζεται το δείγμα πρώτης εκπαίδευσης.*

Στην **Εικόνα 34** παρουσιάζεται η μέση ακρίβεια των τριών ταξινομητών όπως υπολογίστηκε από την πιο πάνω διαδικασία, ως προς το δείγμα που χρησιμοποιήθηκε για τον υπολογισμό της. Με μαύρο παρουσιάζεται η ακρίβεια του ταξινομητή learn++NSE που εκπαιδεύτηκε μόνο μια φορά, με κόκκινο αυτή του ταξινομητή learn++NSE που επανεκπαιδευόταν σύμφωνα με τη συνθήκη του εκκινητή ενώ με μπλε αυτή του ταξινομητή svm που εκπαιδευόταν με όλο το διαθέσιμο (μέχρι τότε) δείγμα. Η μέση ακρίβεια που αντιστοιχεί στη τιμή sample -1 είναι αυτή που βρέθηκε από τα 5.000 παραδείγματα που χρησιμοποιήθηκαν για την πρώτη εκπαίδευση των ταξινομητών και στη τιμή sample 0 αυτή που βρέθηκε από τα υπόλοιπα 1.000 παραδείγματα του δείγματος πρώτης εκπαίδευσης.

Από την **Εικόνα 34** φαίνεται ότι η ακρίβεια ταξινόμησης, όπως υπολογίζεται από τα 1.000 παραδείγματα του δείγματος πρώτης εκπαίδευσης που δεν χρησιμοποιήθηκαν στην εκπαίδευση των ταξινομητών, είναι πρακτικά ίδια και για τους τρεις ταξινομητές. Στην συνέχεια, ενώ η σύσταση των κλάσεων αλλάζει, η ακρίβεια όλων των ταξινομητών ελαττώνεται. Αυτό είναι αναμενόμενο καθώς για την πρώτη εκπαίδευση επιλέχθηκαν τα παραδείγματα που αναμένεται να έχουν τη μεγαλύτερη διαφορά (κριτικές με βαθμούς 1 και 10) και συνεπώς η ακρίβεια ταξινόμησης αναμένεται να είναι καλύτερη. Για να είναι πιο εύκολη η σύγκριση της ακρίβειας ταξινόμησης, στην **Εικόνα 35** παρουσιάζεται η διαφορά της ακρίβειας ταξινόμησης του ταξινομητή learn++NSE που επανεκπαιδευόταν όταν ικανοποιείτο η συνθήκη του εκκινητή και του ταξινομητή svm, από αυτή του ταξινομητή learn++NSE που εκπαιδεύτηκε μόνο μια φορά, δηλαδή παρουσιάζεται η βελτίωση στην ακρίβεια κατηγοριοποίησης των δύο πρώτων ταξινομητών σε σχέση με τον τρίτο. Τα σημεία αναπαριστούν τη μέση τιμή των 5 επαναλήψεων ενώ οι παύλες πάνω και κάτω από τα σημεία τη μέγιστη και την ελάχιστη τιμή.



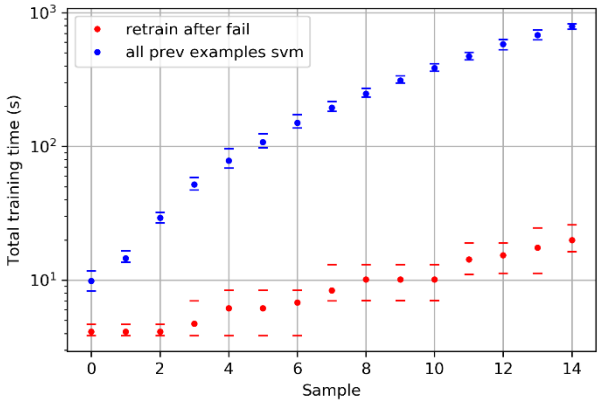
***Εικόνα 34**: Μέση ακρίβεια ταξινόμησης (5 επαναλήψεις) σε σχέση με το δείγμα που χρησιμοποιήθηκε για τον υπολογισμό της, από το dataset d1. Με μαύρο παρουσιάζεται η ακρίβεια του ταξινομητή learn++NSE που εκπαιδεύτηκε μόνο μια φορά, με κόκκινο αυτή του ταξινομητή learn++NSE που επανεκπαιδευόταν όταν ικανοποιείτο η συνθήκη του εκκινητή ενώ με μπλε αυτή του ταξινομητή svm που εκπαιδευόταν με όλο το (μέχρι τότε) διαθέσιμο δείγμα. Η μέση ακρίβεια που αντιστοιχεί στη τιμή sample -1 είναι αυτή που βρέθηκε από τα 5.000 παραδείγματα που χρησιμοποιήθηκαν για την πρώτη εκπαίδευση των ταξινομητών και αυτή που αντιστοιχεί στη τιμή sample 0 βρέθηκε από τα υπόλοιπα 1.000 παραδείγματα του δείγματος πρώτης εκπαίδευσης.*



***Εικόνα 35**: Διαφορά της ακρίβειας ταξινόμησης του ταξινομητή α) learn++NSE που επανεκπαιδευόταν όταν ικανοποιείτο η συνθήκη του εκκινητή (κόκκινο) και του ταξινομητή β) svm (μπλε) από αυτή του ταξινομητή learn++NSE που εκπαιδεύτηκε μόνο μια φορά, ως προς το δείγμα που υπολογίστηκε η ακρίβεια από το dataset d1. Τα σημεία αναπαριστούν τη μέση τιμή των 5 επαναλήψεων ενώ οι παύλες, πάνω και κάτω από αυτά, τη μέγιστη και την ελάχιστη τιμή.*

Το τελευταίο στοιχείο που πρέπει να συνεκτιμηθεί πριν την εξαγωγή συμπερασμάτων είναι ο χρόνος που απαιτείται για την εκπαίδευση ενός ταξινομητή με όλα τα παραδείγματα σε σχέση με αυτόν που απαιτείται για την εκπαίδευση και την επανεκπαίδευση όταν ικανοποιείται η συνθήκη του εκκινητή. Στην **Εικόνα 36** παρουσιάζεται ο χρόνος που χρειάστηκε να εκπαιδευτεί ο ταξινομητής svm χρησιμοποιώντας όλα τα μέχρι τότε δείγματα (μπλε) και ο χρόνος που χρειάστηκε για να εκπαιδευτεί ο ταξινομητής

learn++NSE που επανεκπαιδευόταν όταν ικανοποιείτο η συνθήκη του εκκινητή (κόκκινο). Τα σημεία αναπαριστούν τη μέση τιμή των 5 επαναλήψεων ενώ οι παύλες, πάνω και κάτω από αυτή, τη μέγιστη και την ελάχιστη τιμή. Ο χρόνος που αντιστοιχεί στο sample 0 είναι αυτός που χρειάστηκε για την εκπαίδευση με τα 5.000 παραδείγματα του δείγματος πρώτης εκπαίδευσης. Στο σημείο αυτό πρέπει να τονιστεί ότι η διαδικασία εκπαίδευσης – ελέγχου των ταξινομητών πραγματοποιήθηκε στον ίδιο υπολογιστή και υπό τις ίδιες συνθήκες (workload του υπολογιστή) με σκοπό τη διεξαγωγή ασφαλούς σύγκρισης των αποτελεσμάτων. Προφανώς αν η ίδια διαδικασία επαναληφθεί σε άλλον υπολογιστή ή ακόμα και στον ίδιο υπολογιστή αλλά με διαφορετικό workload ή και διαφορετικό ορισμό προτεραιοτήτων των προγραμμάτων που εκτελούνται, τα απόλυτα αποτελέσματα θα είναι διαφορετικά, αναμένεται όμως ότι τα συμπεράσματα που θα εξαχθούν από τη σύγκρισή τους θα είναι παρόμοια.



***Εικόνα 36**: Ο χρόνος που απαιτείται για την εκπαίδευση των ταξινομητών svm χρησιμοποιώντας όλα τα μέχρι τότε δείγματα (μπλε) και learn++NSE που επανεκπαιδευόταν όταν ικανοποιείτο η συνθήκη του εκκινητή (κόκκινο) Τα σημεία αναπαριστούν τη μέση τιμή των 5 επαναλήψεων ενώ οι παύλες, πάνω και κάτω από αυτή, τη μέγιστη και την ελάχιστη τιμή. Το διάγραμμα προέκυψε με χρήση του dataset d1.*

Όπως αναφέρθηκε και νωρίτερα στα πιο πάνω διαγράμματα παρουσιάζονται τα αποτελέσματα από 5 επαναλήψεις της διαδικασίας εκπαίδευσης και ελέγχου. Πριν από κάθε επανάληψη τα δείγματα ανακατεύονται με τυχαίο τρόπο, συνεπώς υπάρχει ένας βαθμός τυχαιότητας στα αποτελέσματα. Από τα κόκκινα σημεία στην **Εικόνα 36** τα οποία αναπαριστούν τον χρόνο που χρειάζεται για την εκπαίδευσή του ο ταξινομητής learn++NSE, φαίνεται ότι μέχρι και το δείγμα 2 δεν έχει υπάρξει μεταβολή του χρόνου και συνεπώς δεν έχει ενεργοποιηθεί η διαδικασία της επανεκπαίδευσης. Στο δείγμα 3 φαίνεται μια μικρή άνοδος της μέσης τιμής του χρόνου εκπαίδευσης ενώ η ελάχιστη τιμή παραμένει σταθερή. Αυτό συμβαίνει διότι μόνο σε μία από τις πέντε επαναλήψεις ενεργοποιήθηκε η διαδικασία επανεκπαίδευσής στο δείγμα 3. Μέχρι και το δείγμα 6 η ελάχιστη τιμή παραμένει σταθερή, γεγονός που δείχνει ότι σε μία τουλάχιστον από τις επαναλήψεις δεν έχει ενεργοποιηθεί η διαδικασία, ενώ στο δείγμα 7 (όπου και τις πέντε επαναλήψεις έχει ενεργοποιηθεί τουλάχιστον μία φορά η διαδικασία επανεκπαίδευσης) έχει αλλάξει η ελάχιστη τιμή.

Από την **Εικόνα 34** φαίνεται ότι πιο γρήγορα ελαττώνεται η ακρίβεια για τον ταξινομητή που δεν επανεκαιδεύτηκε ενώ πιο αργά για τον ταξινομητή svm που εκπαιδευόταν κάθε φορά με όλο το διαθέσιμο δείγμα. Η ακρίβεια του ταξινομητή learn++NSE που επανεκπαιδευόταν σύμφωνα με τη συνθήκη του εκκινητή, είναι μεταξύ των δυο άλλων, δηλαδή παρουσιάζει σαφώς καλύτερα αποτελέσματα από τον ταξινομητή που δεν επανεκπαιδεύτηκε αλλά όχι τόσο καλά όσο αυτός που εκπαιδευόταν με όλα τα διαθέσιμα παραδείγματα. Πιο συγκεκριμένα από την **Εικόνα 35** φαίνεται ότι η διαφορά της ακρίβειας ταξινόμησης του ταξινομητή α) learn++NSE που επανεκπαιδευόταν όταν ικανοποιείτο η συνθήκη του εκκινητή και του ταξινομητή β) svm, από αυτή του ταξινομητή learn++NSE που εκπαιδεύτηκε μόνο μια φορά, αυξάνει όσο αλλάζει η σύσταση των κλάσεων. Για τον ταξινομητή svm αυξάνει από τα πρώτα δείγματα φτάνοντας σε τιμές μεγαλύτερες του 2% ενώ για το ταξινομητή learn++NSE παραμένει σχεδόν σταθερή στα πρώτα δείγματα, καθώς δεν έχει ενεργοποιηθεί η διαδικασία επανεκπαίδευσης, ενώ μετά αυξάνει και φτάνει σε τιμές μεγαλύτερες του 1%. Όπως αναφέρθηκε και πιο πάνω, σε μία από τις πέντε επαναλήψεις η πρώτη επανεκπαίδευση του ταξινομητή συμβαίνει με το δείγμα 3. Τα αποτελέσματα αυτής της επανεκπαίδευσης όμως, δεν είναι εμφανή στη μέση ακρίβεια του δείγματος 4 (**Εικόνα 34**), καθώς η βελτίωση της ακρίβειας σε μόνο μία από τις πέντε επαναλήψεις, δεν έχει σημαντικό αντίκτυπο στη μέση τιμή. Στην **Εικόνα 35** όμως, παρατηρείται αύξηση του εύρους των τιμών της διαφοράς της ακρίβειας όπως υπολογίστηκε από το δείγμα 4. Στη συνέχεια, με την επεξεργασία του δείγματος 4 ενεργοποιείται η διαδικασία της επανεκπαίδευσης σε άλλες δύο από τις πέντε επαναλήψεις και τα αποτελέσματα είναι φανερά στη μέση ακρίβεια ταξινόμησης του δείγματος 5 στην **Εικόνα 34**.

Όσον αφορά στο χρόνο εκπαίδευσης (**Εικόνα 36**) είναι φανερό ότι για τον συγκεκριμένο ταξινομητή svm (ταξινομητής SVC όπου η εύρεση του ελαχίστου της συνάρτησης κόστους υπολογίζεται πλήρως και όχι προσεγγιστικά, όπως για παράδειγμα την περίπτωση του ταξινομητή SGD) η εκπαίδευση με μεγάλο πλήθος παραδειγμάτων είναι εξαιρετικά χρονοβόρα. Αντιθέτως χρησιμοποιώντας τη learn++NSE συλλογή ταξινομητών SVC, όπου ο κάθε ταξινομητής εκπαιδεύεται με 1.000 παραδείγματα, ο χρόνος που απαιτείται για την επανεκπαίδευσή του είναι μέχρι και σχεδόν 50 φορές μικρότερος, όταν ο αριθμός των παραδειγμάτων εκπαίδευσης είναι μεγάλος.

Τα αποτελέσματα που προέκυψαν με χρήση του dataset d1 συνοψίζονται στα:

1) Η ακρίβεια του ταξινομητή learn++NSE ο οποίος επανεκπαιδευόταν όταν ικανοποιείτο η συνθήκη του εκκινητή, είναι σαφώς καλύτερη από αυτή που επιτυγχάνεται όταν δεν επανεκπαιδεύεται, αλλά είναι υποδεέστερη αυτής που επιτυγχάνεται με ταξινομητή svm που εκπαιδευόταν με όλα τα (μέχρι τότε) διαθέσιμα παραδείγματα.
2) Ο χρόνος που απαιτείται για την εκπαίδευση και επανεκπαίδευση του ταξινομητή learn++NSE είναι μέχρι και 50 φορές μικρότερος από αυτόν που απαιτείται για την εκπαίδευση ενός ταξινομητή svm με τα ίδια χαρακτηριστικά με αυτούς που περιέχονται στη συλλογή learn++NSE.

Η συνεχής πτωτική τάση της ακρίβειας ταξινόμησης όλων των ταξινομητών καθώς αλλάζει η σύσταση των κλάσεων χρησιμοποιώντας το dataset d1 (**Εικόνα 34**) είναι αναμενόμενη, καθώς οι κλάσεις αποτελούνται αρχικά από παραδείγματα των οποίων οι διαφορές αναμένονται να είναι μεγάλες ενώ σταδιακά οι διαφορές ελαττώνονται καταλήγοντας στη μικρότερη δυνατή διαφορά στο τελευταίο δείγμα. Για να ελεγχθεί και σε

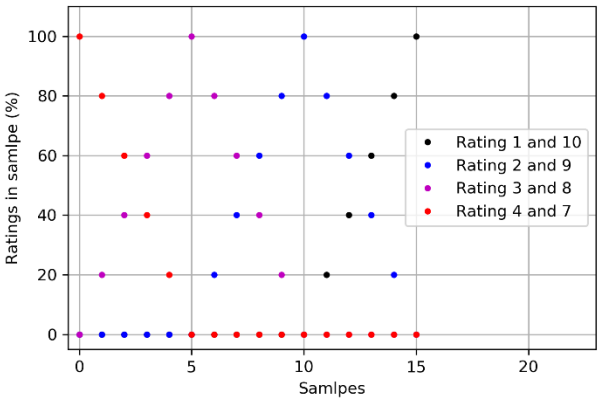
διαφορετικές συνθήκες μεταβολής της σύστασης των κλάσεων η απόδοση του ταξινομητή learn++NSE που επανεκπαιδεύεται σύμφωνα με την συνθήκη του εκκινητή, και να συγκριθεί με αυτή ενός ταξινομητή learn++NSE που δεν επανεκπαιδεύεται και με αυτή ενός ταξινομητή svm που εκπαιδεύεται με όλο το (μέχρι τότε) διαθέσιμο δείγμα εκπαίδευσης δημιουργήθηκε το dataset d2. Η σύσταση των δειγμάτων του d2 είναι η αντίστροφη αυτής του d1, δηλαδή η θετική (αρνητική) κλάση αποτελείται αρχικά από κριτικές με τη λιγότερο θετική (αρνητική) βαθμολογία 7 (βαθμολογία 4) και με την πάροδο του χρόνου οι βαθμολογίες των κριτικών των παραδειγμάτων γίνονταν καλύτερες (χειρότερες) με την θετική (αρνητική) κλάση του τελευταίου δείγματος να αποτελείται εξ ολοκλήρου από παραδείγματα με βαθμολογία 10 (βαθμολογία 1). Καθώς η σύσταση των κλάσεων των δειγμάτων μεταβάλλεται, οι διαφορές των κλάσεων εντείνονται με αποτέλεσμα να αναμένεται η ακρίβεια ταξινόμησης όλων των ταξινομητών να παρουσιάζει αυξητική τάση.

Όπως στο dataset d1 έτσι και στο d2 σε κάθε δείγμα ο αριθμός των παραδειγμάτων με βαθμολογίες 1 και 10, 2 και 9, 3 και 8, 4 και 7 ήταν ίδιος. Στον **Πίνακας 5** παρουσιάζεται η σύσταση των κλάσεων σε κάθε δείγμα του dataset d2. Το d2 αποτελείται από 5.000 παραδείγματα με βαθμολογία 2, 3, 4, 7, 8 και 9 και από 3.000 παραδείγματα με βαθμολογία 1 και 10. Λόγω του περιορισμένου αριθμού των παραδειγμάτων με βαθμολογίες 2, 3, 7 και 9 για την δημιουργία του d2 επιλέχθηκαν με τυχαίο τρόπο 414, 39, 197 και 393 παραδείγματα με βαθμολογίες 2, 3, 7 και 9 και επανατοποθετήθηκαν σε αυτό. Η εύρεση ακρίβειας ταξινόμησης με χρήση του dataset d2 έγινε με τον ίδιο τρόπο όπως και με το d1.

Dataset d2									
	Θετική κλάση					Αρνητική κλάση			
Βαθμολογία	10	9	8	7		4	3	2	1
Πρώτη Εκπαίδευση	0	0	0	3000		3000	0	0	0
Δείγμα 1	0	0	200	800		800	200	0	0
Δείγμα 2	0	0	400	600		600	400	0	0
Δείγμα 3	0	0	600	400		400	600	0	0
Δείγμα 4	0	0	800	200		200	800	0	0
Δείγμα 5	0	0	1000	0		0	1000	0	0
Δείγμα 6	0	200	800	0		0	800	200	0
Δείγμα 7	0	400	600	0		0	600	400	0
Δείγμα 8	0	600	400	0		0	400	600	0
Δείγμα 9	0	800	200	0		0	200	800	0
Δείγμα 10	0	1000	0	0		0	0	1000	0
Δείγμα 11	200	800	0	0		0	0	800	200
Δείγμα 12	400	600	0	0		0	0	600	400
Δείγμα 13	600	400	0	0		0	0	400	600
Δείγμα 14	800	200	0	0		0	0	200	800
Δείγμα 15	1000	0	0	0		0	0	0	1000

Πίνακας 5*: Αριθμός παραδειγμάτων από κάθε βαθμολογία που συγκροτούν τα δείγματα του dataset d2.*

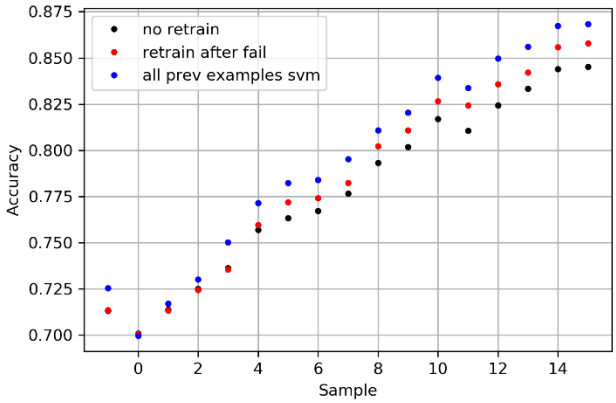
Στην **Εικόνα 37** απεικονίζεται η ποσοστιαία σύσταση της βαθμολογίας των κριτικών που αποτελούν τις δύο κλάσεις κάθε δείγματος του dataset d2. Το δείγμα πρώτης εκπαίδευσης παρουσιάζεται στη θέση 0.



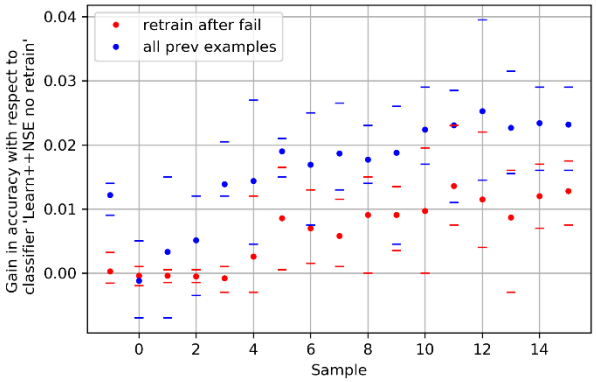
***Εικόνα 37**: Ποσοστιαία σύσταση της βαθμολογίας των κριτικών που αποτελούν τις δύο κλάσεις κάθε δείγματος του dataset d2. Ως sample 0 παρουσιάζεται το δείγμα πρώτης εκπαίδευσης.*

Στην **Εικόνα 38** παρουσιάζεται η μέση ακρίβεια των τριών ταξινομητών όπως υπολογίστηκε με χρήση του dataset d2 ως προς το δείγμα που χρησιμοποιήθηκε για τον υπολογισμό της. Με μαύρο παρουσιάζεται η ακρίβεια του ταξινομητή learn++NSE που εκπαιδεύτηκε μόνο μια φορά, με κόκκινο αυτή του ταξινομητή learn++NSE που επανεκπαιδευόταν σύμφωνα με τη συνθήκη του εκκινητή ενώ με μπλε αυτή του ταξινομητή svm που εκπαιδευόταν με όλο το διαθέσιμο (μέχρι τότε) δείγμα. Η μέση ακρίβεια που αντιστοιχεί στη τιμή sample -1 είναι αυτή που βρέθηκε από τα 5.000 παραδείγματα που χρησιμοποιήθηκαν για την πρώτη εκπαίδευση των ταξινομητών και στη τιμή sample 0 αυτή που βρέθηκε από τα υπόλοιπα 1.000 παραδείγματα του δείγματος πρώτης εκπαίδευσης. Η πτώση που παρατηρείται στην ακρίβεια όπως υπολογίζεται από το δείγμα -1 και από το δείγμα 0 είναι αναμενόμενη καθώς το δείγμα -1 χρησιμοποιήθηκε για την εκπαίδευση των ταξινομητών ενώ το δείγμα 0 (ίδιας σύστασης με το -1) είναι άγνωστο σε αυτούς. Στο dataset d1 παρατηρείται η ίδια συμπεριφορά καθώς η ακρίβεια ταξινόμησης όπως υπολογίστηκε από τα 5.000 παραδείγματα που χρησιμοποιήθηκαν για την εκπαίδευση των ταξινομητών είναι υψηλότερη από αυτήν που υπολογίστηκε με τα υπόλοιπα 1.000 παραδείγματα του ίδιου δείγματος (πρώτο και δεύτερο σημείο **Εικόνα 34**). Βέβαια αυτή δεν γίνεται αμέσως αντιληπτή (όπως με το dataset d2 – **Εικόνα 38**) καθώς «κρύβεται» στην ευρύτερη πτωτική τάση της ακρίβειας ταξινόμησης λόγω κατασκευής του dataset d1.

Στην **Εικόνα 39** παρουσιάζεται η διαφορά της ακρίβειας ταξινόμησης του ταξινομητή learn++NSE που επανεκπαιδευόταν όταν ικανοποιείτο η συνθήκη του εκκινητή και του ταξινομητή svm, από αυτή του ταξινομητή learn++NSE που εκπαιδεύτηκε μόνο μια φορά (βελτίωση της ακρίβειας κατηγοριοποίησης των δύο πρώτων ταξινομητών σε σχέση με τον τρίτο) με χρήση του dataset d2. Τα σημεία αναπαριστούν τη μέση τιμή των 5 επαναλήψεων ενώ οι παύλες πάνω και κάτω από τα σημεία τη μέγιστη και την ελάχιστη τιμή.



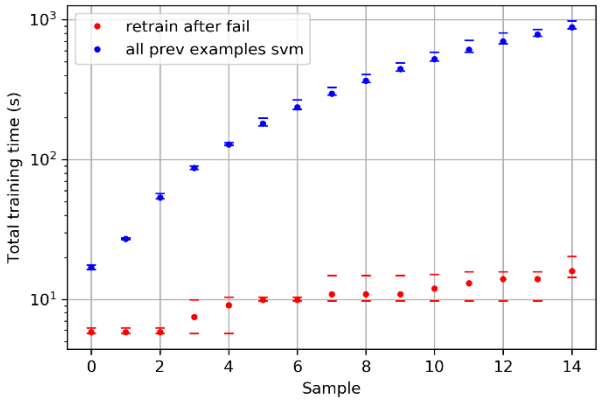
***Εικόνα 38**: Μέση ακρίβεια ταξινόμησης (5 επαναλήψεις) σε σχέση με το δείγμα που χρησιμοποιήθηκε για τον υπολογισμό της, από το dataset d2. Με μαύρο παρουσιάζεται η ακρίβεια του ταξινομητή learn++NSE που εκπαιδεύτηκε μόνο μια φορά, με κόκκινο αυτή του ταξινομητή learn++NSE που επανεκπαιδευόταν όταν ικανοποιείτο η συνθήκη του εκκινητή ενώ με μπλε αυτή του ταξινομητή svm που εκπαιδευόταν με όλο το (μέχρι τότε) διαθέσιμο δείγμα. Η μέση ακρίβεια που αντιστοιχεί στη τιμή sample -1 είναι αυτή που βρέθηκε από τα 5.000 παραδείγματα που χρησιμοποιήθηκαν για την πρώτη εκπαίδευση των ταξινομητών και αυτή που αντιστοιχεί στη τιμή sample 0 βρέθηκε από τα υπόλοιπα 1.000 παραδείγματα του δείγματος πρώτης εκπαίδευσης.*



***Εικόνα 39**: Διαφορά της ακρίβειας ταξινόμησης του ταξινομητή α) learn++NSE που επανεκπαιδευόταν όταν ικανοποιείτο η συνθήκη του εκκινητή (κόκκινο) και του ταξινομητή β) svm (μπλε) από αυτή του ταξινομητή learn++NSE που εκπαιδεύτηκε μόνο μια φορά, ως προς το δείγμα που υπολογίστηκε η ακρίβεια από το dataset d2. Τα σημεία αναπαριστούν τη μέση τιμή των 5 επαναλήψεων ενώ οι παύλες, πάνω και κάτω από αυτά, τη μέγιστη και την ελάχιστη τιμή.*

Στην **Εικόνα 40** παρουσιάζεται ο χρόνος που χρειάστηκε να εκπαιδευτεί ο ταξινομητής svm χρησιμοποιώντας όλα τα (μέχρι τότε) διαθέσιμα δείγματα και ο χρόνος που χρειάστηκε για να εκπαιδευτεί ο ταξινομητής learn++NSE που επανεκπαιδεύεται σύμφωνα με τη συνθήκη του εκκινητή με χρήση του dataset d2. Τα σημεία αναπαριστούν τη μέση τιμή των 5 επαναλήψεων ενώ οι παύλες, πάνω και κάτω από αυτή, τη μέγιστη και την ελάχιστη

τιμή. Ο χρόνος που αντιστοιχεί στο sample 0 είναι αυτός που χρειάστηκε για την εκπαίδευση με τα 5.000 παραδείγματα του δείγματος πρώτης εκπαίδευσης.



***Εικόνα 40**: Ο χρόνος που απαιτείται για την εκπαίδευση των ταξινομητών svm χρησιμοποιώντας όλα τα μέχρι τότε δείγματα (μπλε) και learn++NSE που επανεκπαιδευόταν όταν ικανοποιείτο η συνθήκη του εκκινητή (κόκκινο) Τα σημεία αναπαριστούν τη μέση τιμή των 5 επαναλήψεων ενώ οι παύλες, πάνω και κάτω από αυτή, τη μέγιστη και την ελάχιστη τιμή. Το διάγραμμα προέκυψε με χρήση του dataset d3.*

Τα συμπεράσματα που προκύπτουν χρησιμοποιώντας το dataset d2 για την αξιολόγηση των ταξινομητών, είναι ίδια με αυτά που προέκυψαν από τη χρήση του dataset d1. Συνεπώς συνοψίζοντας τα συμπεράσματα που εξήχθησαν από αυτή την ενότητα:

1) Η ακρίβεια του ταξινομητή learn++NSE ο οποίος επανεκπαιδευόταν όταν ικανοποιείτο η συνθήκη του εκκινητή, είναι σαφώς καλύτερη από αυτή που επιτυγχάνεται όταν δεν επανεκπαιδεύεται.
2) Η διαφορά της ακρίβειας του ταξινομητή που εκπαιδεύεται με αυξητικό τρόπο από αυτή του ταξινομητή που εκπαιδεύεται με όλο το διαθέσιμο δείγμα, είναι της τάξης του 1%.
3) Ο χρόνος που απαιτείται για την εκπαίδευση και επανεκπαίδευση του ταξινομητή learn++NSE είναι μέχρι και 50 φορές μικρότερος, από αυτόν που απαιτείται για την εκπαίδευση ενός ταξινομητή svm που έχει τα ίδια χαρακτηριστικά με αυτούς που περιέχονται στη συλλογή learn++NSE.

3.4. Αξιολόγηση του ταξινομητή σε μεταβαλλόμενο περιβάλλον – κλάσεις με άνισο αριθμό παραδειγμάτων

Στην προηγούμενη ενότητα μελετήθηκε η απόδοση ταξινόμησης της συλλογής ταξινομητών learn++NSE, όταν η σύσταση των κλάσεων αλλάζει με τον χρόνο. Στην παρούσα

ενότητα θα μελετηθεί η απόδοσή της στην περίπτωση όπου ο αριθμός των παραδειγμάτων των διαφορετικών κλάσεων διαφέρει σε κάθε δείγμα.

Για την μελέτη της απόδοσης της συλλογής ταξινομητών learn++NSE στην περίπτωση όπου οι κλάσεις του dataset αποτελούνται από διαφορετικό (άνισο) αριθμό παραδειγμάτων (μη ισοσκελισμένο ή imbalanced dataset), δημιουργήθηκαν νέα datastets από το dataset «Large Movie Review Dataset v1.0». Τα datasets αυτά (d3, d4 και d5) κατασκευάστηκαν έτσι ώστε να αποτελούνται από 4 κλάσεις. Για τη δημιουργία τεσσάρων κλάσεων συνενώθηκαν οι κριτικές, τόσο του δείγματος εκπαίδευσης όσο και του δείγματος ελέγχου με βαθμολογία 1 και 2 (κλάση 0 – πολύ αρνητική), 3 και 4 (κλάση 1 – αρνητική), 7 και 8 (κλάση 2 – θετική) και τέλος 9 και 10 (κλάση 3 – πολύ θετική). Ο αριθμός των παραδειγμάτων της κάθε κλάσης παρουσιάζονται στον **Πίνακας 6**.

Βαθμός κριτικής	Αριθμός παραδειγμάτων δείγματος εκπαίδευσης	Αριθμός παραδειγμάτων δείγματος ελέγχου	Άθροισμα	Κλάση	Σύνολο διαθέσιμων παραδειγμάτων
1	5.100	5.022	10.122	0	14.708
2	2.284	2.302	4.586		
3	2.420	2.541	4.961	1	10.292
4	2.696	2.635	5.331		
7	2.496	2.307	4.803	2	10.662
8	3.009	2.850	5.859		
9	2.263	2.344	4.607	3	13.978
10	4.372	4.999	9.371		

Πίνακας 6*: Αριθμός των παραδειγμάτων των κλάσεων 0, 1, 2 και 3 των datasets d3, d4 και d5 που δημιουργήθηκαν από το dataset «Large Movie Review Dataset v1.0».*

Πριν την δημιουργία και την χρήση των datasets, ελέγχθηκε πόσο καλά ξεχωρίζουν οι κλάσεις ανά δυο. Από τους πειραματισμούς που διεξήχθησαν στην προηγούμενη ενότητα, αναμένεται ότι η ακρίβεια ταξινόμησης των ακραίων κλάσεων (πολύ θετική με πολύ αρνητική , 0 με 3) θα είναι μεγάλη (πάνω από 80%), η ακρίβεια ταξινόμησης των μεσαίων κλάσεων (θετική με αρνητική , 1 με 2) θα είναι αρκετά μικρότερη (κάτω από 70%) ενώ στις υπόλοιπες περιπτώσεις η ακρίβεια θα είναι ενδιάμεση. Για την μελέτη αυτή χρησιμοποιήθηκε ταξινομητής svm (SVC με ίδια χαρακτηριστικά με αυτούς που αποτελούν τη συλλογή learn++NSE). Ως δείγμα εκπαίδευσης, κάθε φορά, χρησιμοποιήθηκαν 2000 παραδείγματα από κάθε κλάση, επιλεγμένα με τέτοιο τρόπο ώστε τα παραδείγματα με διαφορετική βαθμολογία να είναι από 1000. Με την ίδια μέθοδο επιλέχθηκε και το δείγμα ελέγχου. Η πιο πάνω διαδικασία επαναλήφθηκε 10 φορές και πριν από κάθε φορά το δείγμα εκπαίδευσης και ελέγχου επιλέχθηκαν με τυχαίο τρόπο. Στους πιο κάτω πίνακες (**Πίνακας 7** έως και **Πίνακας 12**) παρουσιάζονται οι μήτρες σύγχυσης για τις δυαδικές κατηγοριοποιήσεις των κλάσεων, στις οποίες παρατίθεται ο μέσος αριθμός γεγονότων από τις 10 επαναλήψεις. Επίσης παρατίθεται και η ακρίβεια ταξινόμησης. Όπως αναμενόταν, η μέγιστη ακρίβεια ταξινόμησης ~ 85% επιτυγχάνεται για τις ακραίες κλάσεις 0 – 3. Ακολουθεί αυτή των κλάσεων 0 – 2 ~82% και 1 – 3 ~81% και μετά αυτή των κλάσεων 1 – 2 ~74%. Οι κλάσεις 0 – 1 και 2 – 3 παρουσιάζουν τις χαμηλότερες ακρίβειες ταξινόμησης ~66% και ~65% αντίστοιχα.

		Predicted class	
True Class		0	3
	0	3.420,0	580,0
	3	625,5	3.47,7
Accuracy: 84,9%			

Πίνακας 7*: Μήτρα σύγχυσης δυαδικής ταξινόμησης κλάσεων 0 και 3, στην οποία παρουσιάζεται ο μέσος αριθμός από 10 επαναλήψεις. Επίσης παρουσιάζεται και η ακρίβεια ταξινόμησης.*

		Predicted class	
True Class		1	2
	1	3.003,5	996,5
	2	1.095,2	2.904,9
Accuracy: 73,9%			

Πίνακας 8*: Μήτρα σύγχυσης δυαδικής ταξινόμησης κλάσεων 1 και 2, στην οποία παρουσιάζεται ο μέσος αριθμός από 10 επαναλήψεις. Επίσης παρουσιάζεται και η ακρίβεια ταξινόμησης.*

		Predicted class	
True Class		0	2
	0	3.276,0	724,0
	2	707,8	3.292,2
Accuracy: 82,1%			

Πίνακας 9*: Μήτρα σύγχυσης δυαδικής ταξινόμησης κλάσεων 0 και 2, στην οποία παρουσιάζεται ο μέσος αριθμός από 10 επαναλήψεις. Επίσης παρουσιάζεται και η ακρίβεια ταξινόμησης.*

		Predicted class	
True Class		1	3
	1	3.288,2	711,8
	3	843,1	3.156,9
Accuracy: 80,6%			

Πίνακας 10*: Μήτρα σύγχυσης δυαδικής ταξινόμησης κλάσεων 1 και 3, στην οποία παρουσιάζεται ο μέσος αριθμός από 10 επαναλήψεις. Επίσης παρουσιάζεται και η ακρίβεια ταξινόμησης.*

		Predicted class	
True Class		0	1
	0	2.563,7	1.463,3
	1	1.267,5	2.732,5
Accuracy: 66,2%			

Πίνακας 11*: Μήτρα σύγχυσης δυαδικής ταξινόμησης κλάσεων 0 και 1, στην οποία παρουσιάζεται ο μέσος αριθμός από 10 επαναλήψεις. Επίσης παρουσιάζεται και η ακρίβεια ταξινόμησης.*

		Predicted class	
True Class		2	3
	2	2.800,7	1.199,3
	3	1.612,2	2.387,8
Accuracy: 64,9%			

Πίνακας 12*: Μήτρα σύγχυσης δυαδικής ταξινόμησης κλάσεων 0 και 1, στην οποία παρουσιάζεται ο μέσος αριθμός από 10 επαναλήψεις. Επίσης παρουσιάζεται και η ακρίβεια ταξινόμησης.*

Το πρώτο dataset (d3) που δημιουργήθηκε, είχε ως στόχο την διερεύνηση της απόδοσης του ταξινομητή στην περίπτωση κατά την οποία κάποιες κλάσεις αρχικά αποτελούνταν από πολύ λίγα παραδείγματα ενώ, με την πάροδο του χρόνου, ο αριθμός των παραδειγμάτων από τα οποία αποτελείτο η κάθε κλάση εξισορροπείτο. Στον **Πίνακας 13** παρουσιάζεται ο αριθμός των παραδειγμάτων από κάθε κλάση των δειγμάτων του dataset d3. Το dataset d3 αποτελείται από 9.900 παραδείγματα από τις κλάσεις 0 και 3 και από 4.100 παραδείγματα από τις κλάσεις 1 και 2. Για την δημιουργία του dataset d3 δεν υπήρξε περιορισμός των παραδειγμάτων των διαφορετικών κλάσεων, συνεπώς δεν υπήρξε η ανάγκη της επαναχρησιμοποίησης παραδειγμάτων (όπως στην περίπτωση των datasets d1 και d2). Αντιθέτως, τα διαθέσιμα παραδείγματα ήταν περισσότερα από αυτά που απαιτούνταν για την δημιουργία του. Αυτό επέτρεψε την επιλογή παραδειγμάτων από το αρχικό dataset με τέτοιο τρόπο ώστε η κάθε κλάση να αποτελείται από, όσο το δυνατόν, ίδιο αριθμό παραδειγμάτων με διαφορική βαθμολογία:

1) Σύσταση κλάσης 0: 53,6% βαθμολογία 1 και 46,4% βαθμολογία 2
2) Σύσταση κλάσης 1: 50% βαθμολογία 3 και 50% βαθμολογία 4

3) Σύσταση κλάσης 2: 50% βαθμολογία 7 και 50% βαθμολογία 8
4) Σύσταση κλάσης 3: 46,5% βαθμολογία 9 και 53,5% βαθμολογία 10

	Dataset d3			
Κλάσεις	0	1	2	3
Πρώτη Εκπαίδευση	1.900	100	100	1.900
Δείγμα 1	875	125	125	875
Δείγμα 2	875	125	125	875
Δείγμα 3	750	250	250	750
Δείγμα 4	750	250	250	750
Δείγμα 5	750	250	250	750
Δείγμα 6	750	250	250	750
Δείγμα 7	625	375	375	625
Δείγμα 8	625	375	375	625
Δείγμα 9	500	500	500	500
Δείγμα 10	500	500	500	500
Δείγμα 11	500	500	500	500
Δείγμα 12	500	500	500	500
Σύνολο	9.900	4.100	4.100	9.900

***Πίνακας 13**: Αριθμός παραδειγμάτων από κλάση που συγκροτούν τα δείγματα του dataset d3.*

Το dataset d3 χρησιμοποιήθηκε για την αξιολόγηση ταξινομητή learn++NSE ο οποίος εκπαιδευόταν με αυξητικό τρόπο. Ο ταξινομητής που χρησιμοποιήθηκε είχε τα ίδια χαρακτηριστικά με αυτόν που περιγράφηκε στη προηγούμενη ενότητα (αποτελούταν από SVC ταξινομητές και δεν χρησιμοποιήθηκε pruning) εκτός του batch size. Σε αυτή την περίπτωση, όπου οι υπό εξέταση κλάσεις είναι 4, το batch size επιλέχθηκε 2.000. Επίσης δεν χρησιμοποιήθηκε η προσέγγιση της επανεκπαίδευσης του ταξινομητή όταν ικανοποιείτο η συνθήκη κάποιου εκκινητή, αλλά ο ταξινομητής επανεκπαιδευόταν με κάθε νέο δείγμα. Η απόδοση του ταξινομητή αυτού συγκρίθηκε με την απόδοση τριών ακόμα ταξινομητών. Ενός ταξινομητή learn++NSE ο οποίος εκπαιδεύτηκε μόνο με το δείγμα πρώτης εκπαίδευσης, ενός ταξινομητή smv (SVC με ίδια χαρακτηριστικά με αυτούς που αποτελούν τη συλλογή learn++NSE) ο οποίος εκπαιδευόταν κάθε φορά με όλα τα παραδείγματα των δειγμάτων που προηγούνταν του δείγματος βάσει του οποίου ελεγχόταν και τέλος ενός ίδιου ταξινομητή svm ο οποίος εκπαιδευόταν κάθε φορά μόνο με το προηγούμενο δείγμα από αυτό βάσει του οποίου ελεγχόταν. Ο τελευταίος ταξινομητής χρησιμοποιήθηκε καθώς αναμένεται να είναι πολύ ευαίσθητος σε αλλαγές του dataset με τον χρόνο. Για την αξιολόγηση των ταξινομητών ακολουθήθηκε η πιο κάτω διαδικασία:

1) Όλοι οι ταξινομητές εκπαιδεύτηκαν με το δείγμα εκπαίδευσής και αξιολογήθηκαν με χρήση του ίδιου δείγματος.
2) Αξιολογήθηκαν με το δείγμα 1.
3) Ο ταξινομητής learn++NSE που εκπαιδευόταν με αυξητικό τρόπο επανεκπαιδεύτηκε με το δείγμα 1. Ο ταξινομητής svm που εκπαιδευόταν με όλο το διαθέσιμο δείγμα εκπαιδεύτηκε με το σύνολο του δείγματος πρώτης εκπαίδευσης και του δείγματος 1. Ο ταξινομητής svm που εκπαιδευόταν κάθε

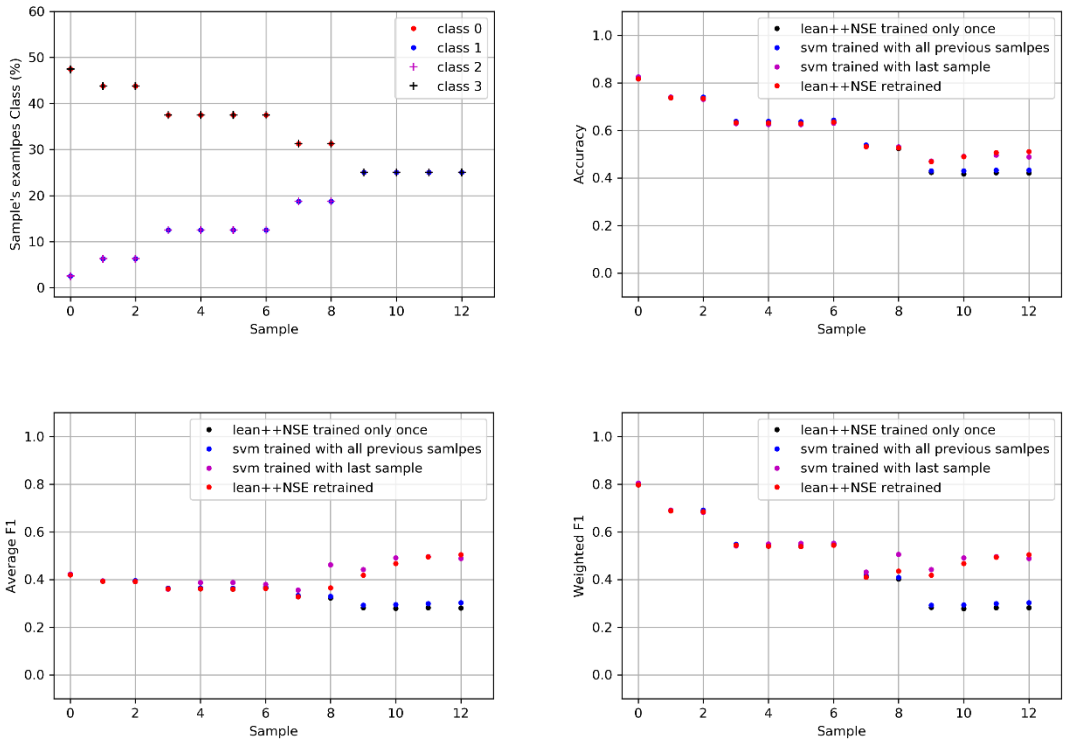
φορά μόνο με το προηγούμενο δείγμα, εκπαιδεύτηκε με το δείγμα 1. Και οι 4 ταξινομητές αξιολογήθηκαν με το δείγμα 2.

4) Το βήμα 3 επαναλήφθηκε για όλα τα επόμενα δείγματα.

Η διαδικασία αυτή επαναλήφθηκε 10 φορές και πριν από κάθε επανάληψη τα δείγματα του dataset εξήχθησαν από το dataset «Large Movie Review Dataset v1.0»με τυχαίο τρόπο.

Η απόδοση των ταξινομητών ελέγχθηκε χρησιμοποιώντας τρία κριτήρια, την ακρίβεια ταξινόμησης (accuracy), τον μέσο δείκτη F_1 (average F_1) και τον σταθμισμένο δείκτη F_1 (weighted F_1). Η ακρίβεια ταξινόμησης στην περίπτωση των τεσσάρων κλάσεων υπολογίστηκε ως το πηλίκο των σωστά ταξινομημένων παραδειγμάτων προς το σύνολο όλων των παραδειγμάτων που εξετάστηκαν. Για την εύρεση του μέσου και του σταθμισμένου δείκτη F_1 βρέθηκε πρώτα ο δείκτης F_1 για κάθε κλάση. Ο δείκτης F_1 κάθε κλάσης ορίζεται ως ο αρμονικός μέσος της ακρίβειας – precision και της ανάκλησης – recall (Precision: ο λόγος των σωστά ταξινομημένων παραδειγμάτων της κλάσης προς όλα τα παραδείγματα που έχουν ταξινομηθεί στην κλάση αυτή. Recall: ο λόγος των σωστά ταξινομημένων παραδειγμάτων της κλάσης προς όλα τα παραδείγματα που πραγματικά ανήκουν στην κλάση αυτή). Ο μέσος δείκτης F_1 είναι η μέση τιμή των F_1 δεικτών της κάθε κλάσης ενώ ο σταθμισμένος δείκτης F_1 είναι η σταθμισμένη μέση τιμή των δεικτών F_1 της κάθε κλάσης με τον αριθμό των παραδειγμάτων που πραγματικά ανήκουν στην κλάση αυτή. Τόσο η ακρίβεια ταξινόμησης, όσο και ο σταθμισμένος δείκτης F_1 επηρεάζονται σημαντικά από τις πιο πολυπληθείς κλάσεις. Αντιθέτως, ο μέσος δείκτης F_1 δίνει ίση βαρύτητα σε όλες τις κλάσεις ανεξαρτήτως του πλήθους των παραδειγμάτων από τα οποία αποτελούνται. Στο Παράρτημα Ι παρουσιάζεται παράδειγμα στο οποίο καταδεικνύεται αυτή η συμπεριφορά. Ο μέσος δείκτης F_1 ενδείκνυται στην περίπτωση που εξετάζεται πόσο καλά κατηγοριοποιούνται τα παραδείγματα τόσο των ολιγομελών κλάσεων όσο και των πολυπληθών. Στην περίπτωση όπου οι ολιγομελείς κλάσεις είναι αναλογικά μικρότερης σημασίας από τις πολυπληθείς ενδείκνυται ο σταθμισμένος δείκτης F_1.

Στην **Εικόνα 41** άνω αριστερά, παρουσιάζεται το ποσοστό των παραδειγμάτων από κάθε κλάση των δειγμάτων του dataset d3, ενώ άνω δεξιά, κάτω αριστερά και κάτω δεξιά παρουσιάζεται αντίστοιχα η ακρίβεια ταξινόμησης, ο μέσος δείκτης F_1 και ο σταθμισμένος δείκτης F_1, ως προς το δείγμα βάσει του οποίου έχουν υπολογιστεί. Στη θέση 0 παρουσιάζεται η αξιολόγηση των ταξινομητών με το δείγμα πρώτης εκπαίδευσης με το οποίο και εκπαιδεύτηκαν. Με κόκκινο παρουσιάζεται ο ταξινομητής learn++NSE ο οποίος επανεκπαιδευόταν με κάθε νέο δείγμα, με μαύρο ο ταξινομητής learn++NSE ο οποίος εκπαιδεύτηκε μόνο με το δείγμα πρώτης εκπαίδευσης, με μπλε ο ταξινομητή smv ο οποίος εκπαιδευόταν με όλα τα (μέχρι τότε) διαθέσιμα παραδείγματα και με μοβ ο ταξινομητής svm ο οποίος εκπαιδευόταν κάθε φορά μόνο με το προηγούμενο δείγμα από αυτό βάσει του οποίου ελεγχόταν.



***Εικόνα 41**: Άνω αριστερά: Ποσοστό των παραδειγμάτων από κάθε κλάση των δειγμάτων του dataset d3. Άνω δεξιά, κάτω αριστερά και κάτω δεξιά παρουσιάζεται αντίστοιχα η ακρίβεια ταξινόμησης, ο μέσος δείκτης F_1 και ο σταθμισμένος δείκτης F_1,ως προς το δείγμα βάσει του οποίου έχουν υπολογιστεί. Στη θέση 0 παρουσιάζεται η αξιολόγηση των ταξινομητών με το δείγμα πρώτης εκπαίδευσης με το οποίο και εκπαιδεύτηκαν. Με κόκκινο παρουσιάζεται ο ταξινομητής learn++NSE ο οποίος επανεκπαιδευόταν με κάθε νέο δείγμα με αυξητικό τρόπο, με μαύρο ο ταξινομητής learn++NSE ο οποίος εκπαιδεύτηκε μόνο με το δείγμα πρώτης εκπαίδευσης, με μπλε ο ταξινομητής smv ο οποίος εκπαιδευόταν με όλα τα (μέχρι τότε) διαθέσιμα παραδείγματα και με μοβ ο ταξινομητής svm ο οποίος εκπαιδευόταν κάθε φορά μόνο με το προηγούμενο δείγμα από αυτό βάσει του οποίου ελεγχόταν.*

Από την **Εικόνα 41** φαίνεται ότι όλοι οι δείκτες έχουν (σχεδόν) τις ίδιες, πτωτικές, τιμές όταν εξήχθησαν από τα δείγματα 0 έως και 7 (με τον μέσο δείκτη F_1 του ταξινομητή svm που εκπαιδεύεται με το προηγούμενο δείγμα, να λαμβάνει ελαφρώς υψηλότερες τιμές στα δείγματα 4 με 7). Αυτό σημαίνει ότι η απόδοση των ταξινομητών κατά την εκπαίδευση μέχρι και το δείγμα 6 δεν παρουσιάζει ουσιαστική διαφορά. Στην συνέχεια, παρατηρείται απόκλιση των τιμών σε όλους τους δείκτες (με τις τιμής της ακρίβειας ταξινόμησης να παρουσιάζουν σημαντικά μικρότερο εύρος απόκλισης), με αύξηση των τιμών του ταξινομητή learn++NSE ο οποίος επανεκπαιδευόταν με κάθε νέο δείγμα και του ταξινομητή svm ο οποίος εκπαιδευόταν κάθε φορά μόνο με το προηγούμενο δείγμα από αυτό βάσει του οποίου ελεγχόταν. Από τους ταξινομητές οι δείκτες των οποίων παρουσιάζουν ανοδική τάση, οι δείκτες F1 του ταξινομητή svm ο οποίος εκπαιδευόταν κάθε φορά μόνο με το προηγούμενο δείγμα από αυτό βάσει του οποίου ελεγχόταν, παρουσιάζουν ένα άλμα στο δείγμα 8 (εκπαίδευση με το δείγμα 7), ενώ αυτοί του ταξινομητή learn++NSE παρουσιάζουν μια ομαλότερη ανοδική πορεία και φτάνουν τις τιμές του svm. Φαίνεται λοιπόν ότι όταν ο αριθμός των παραδειγμάτων των διαφορετικών κλάσεων έχει μεγάλη διαφορά, οι

ταξινομητές δεν είναι ικανοί να «μάθουν» καλά τα παραδείγματα των ολιγομελών κλάσεων και τα κατατάσσουν λανθασμένα. Όταν όμως η διαφορά του αριθμού των παραδειγμάτων με βάση τα οποία εκπαιδεύονται ελαττωθεί, οι ταξινομητές μπορούν να κατατάξουν καλύτερα τα άγνωστα παραδείγματα. Στο Παράρτημα ΙΙ παρουσιάζεται οπτικοποίηση των μητρών σύγχυσης που προέκυψαν από την αξιολόγηση κάθε δείγματος και για τους 4 ταξινομητές. Από αυτές ενισχύεται το πιο πάνω συμπέρασμα.

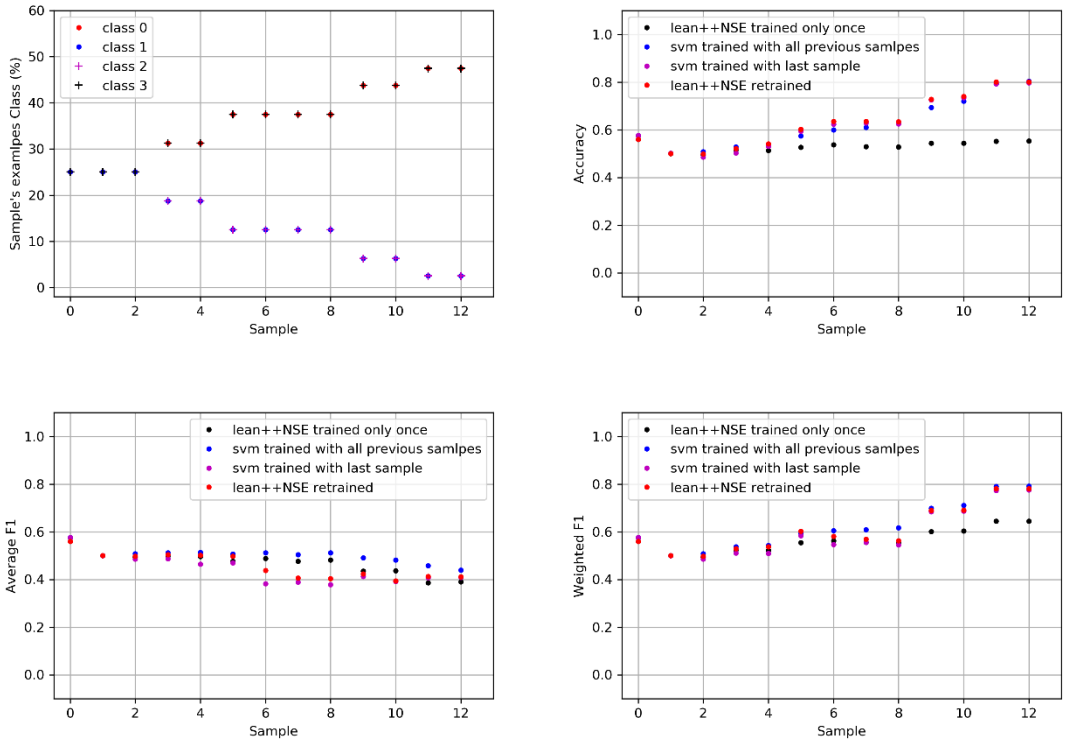
Η καλύτερη απόδοση του ταξινομητή svm ο οποίος εκπαιδευόταν κάθε φορά μόνο με το προηγούμενο δείγμα είναι αναμενόμενη, σύμφωνα με τα πιο πάνω συμπεράσματα, καθώς αυτός εκπαιδεύτηκε με το λιγότερο imbalanced δείγμα. Αντίστοιχα η απόδοση του ταξινομητή smv ο οποίος εκπαιδευόταν με όλα τα (μέχρι τότε) διαθέσιμα παραδείγματα είναι συνεχώς πτωτική καθώς πάντα εκπαιδευόταν με imbalanced δείγματα. Από όλα τα δείγματα με τα οποία εκπαιδεύτηκε, το λιγότερο imbalanced ήταν αυτό της τελευταίας εκπαίδευσης (άθροισμα όλων εκτός του δείγματος 12) το οποίο αποτελείται από παραδείγματα που ανήκουν κατά 36,2% σε κάθε μια από τις κλάσεις 0 και 3 και κατά 13,8% σε παραδείγματα που ανήκουν στις κλάσεις 1 και 2. Πιο πολύ ενδιαφέρον όμως, παρουσιάζει η απόδοση του ταξινομητή learn++NSE, η οποία είναι σχεδόν τόσο καλή όσο αυτή του ταξινομητή svm ο οποίος εκπαιδευόταν κάθε φορά μόνο με το προηγούμενο δείγμα. Αυτό συμβαίνει διότι με κάθε νέα επανεκπαίδευση, ένας νέος ταξινομητής προστίθεται στη συλλογή και τα βάρη ανακατανέμονται με τέτοιο τρόπο, ώστε να προβλέπεται καλύτερα το τελευταίο δείγμα εκπαίδευσης.

Για να ελεγχθεί η συμπεριφορά του ταξινομητή learn++NSE που εκπαιδεύεται με αυξητικό τρόπο στην αντίστροφη περίπτωση, δηλαδή όταν αρχικά ο αριθμός των παραδειγμάτων των διαφορετικών κλάσεων είναι περίπου ίδιος αλλά με τον χρόνο αλλάζει, δημιουργήθηκε το dataset d4. Σε αυτό το dataset προσομοιώνεται η περίπτωση όπου ένα balanced dataset με την πάροδο του χρόνου γίνεται imbalanced. Στον **Πίνακας 14** παρουσιάζεται ο αριθμός των παραδειγμάτων που ανήκουν σε κάθε κλάση των δειγμάτων του dataset d4. Ο αριθμός των δειγμάτων που αποτελούν την κάθε κλάση του dataset όπως και η σύσταση των κλάσεων από παραδείγματα με διαφορετικές βαθμολογίες είναι ίδια με αυτά του dataset d3. Οι ταξινομητές, η διαδικασία καθώς και οι δείκτες για την αξιολόγηση των ταξινομητών, είναι ίδιοι με αυτούς που χρησιμοποιήθηκαν πιο πάνω.

Στην **Εικόνα 42** άνω αριστερά, παρουσιάζεται το ποσοστό των παραδειγμάτων από κάθε κλάση των δειγμάτων του dataset d3, ενώ άνω δεξιά, κάτω αριστερά και κάτω δεξιά παρουσιάζεται αντίστοιχα η ακρίβεια ταξινόμησης, ο μέσος δείκτης F_1 και ο σταθμισμένος δείκτης F_1, ως προς το δείγμα βάσει του οποίου έχουν υπολογιστεί. Στη θέση 0 παρουσιάζεται η αξιολόγηση των ταξινομητών με το δείγμα πρώτης εκπαίδευσης με το οποίο και εκπαιδεύτηκαν. Με κόκκινο παρουσιάζεται ο ταξινομητής learn++NSE ο οποίος επανεκπαιδευόταν με κάθε νέο δείγμα, με μαύρο ο ταξινομητής learn++NSE ο οποίος εκπαιδεύτηκε μόνο με το δείγμα πρώτης εκπαίδευσης, με μπλε ο ταξινομητή smv ο οποίος εκπαιδευόταν με όλα τα (μέχρι τότε) διαθέσιμα παραδείγματα και με μοβ ο ταξινομητής svm ο οποίος εκπαιδευόταν κάθε φορά μόνο με το προηγούμενο δείγμα από αυτό βάσει του οποίου ελεγχόταν.

	Dataset d4			
Κλάσεις	0	1	2	3
Πρώτη Εκπαίδευση	1.000	1.000	1.000	1.000
Δείγμα 1	500	500	500	500
Δείγμα 2	500	500	500	500
Δείγμα 3	625	375	375	625
Δείγμα 4	625	375	375	625
Δείγμα 5	750	250	250	750
Δείγμα 6	750	250	250	750
Δείγμα 7	750	250	250	750
Δείγμα 8	750	250	250	750
Δείγμα 9	875	125	125	875
Δείγμα 10	875	125	125	875
Δείγμα 11	950	50	50	950
Δείγμα 12	950	50	50	950
Σύνολο	9.900	4.100	4.100	9.900

***Πίνακας 14**: Αριθμός παραδειγμάτων από κλάση που συγκροτούν τα δείγματα του dataset d4.*



***Εικόνα 42**: Άνω αριστερά: Ποσοστό των παραδειγμάτων από κάθε κλάση των δειγμάτων του dataset d4. Άνω δεξιά, κάτω αριστερά και κάτω δεξιά παρουσιάζεται αντίστοιχα η ακρίβεια ταξινόμησης, ο μέσος δείκτης F_1 και ο σταθμισμένος δείκτης F_1,ως προς το δείγμα βάσει του οποίου έχουν υπολογιστεί. Στη θέση 0 παρουσιάζεται η αξιολόγηση των ταξινομητών με το δείγμα πρώτης εκπαίδευσης με το οποίο και εκπαιδεύτηκαν.*

Τα συμπεράσματα που εξάγονται από την **Εικόνα 42** είναι διαφορετικά από αυτά που εξήχθησαν από την αξιολόγηση του dataset d3. Αρχικά, παρατηρείται ότι η ακρίβεια ταξινόμησης όπως υπολογίστηκε από τα δείγματα 4 έως και 12 παρουσιάζει εμφανή αυξητική τάση για όλους τους ταξινομητές εκτός αυτής του ταξινομητή που εκπαιδεύτηκε μόνο μια φορά. Αυτή η συμπεριφορά αρχίζει από το δείγμα 4 (εκπαίδευση με το δείγμα 3 και μετά) όπου το dataset ξεκίνησε να γίνεται imbalanced και οφείλεται στο ότι οι ταξινομητές μαθαίνουν καλύτερα τις πιο πολυπληθείς κλάσεις των οποίων ο αριθμός αυξάνει με τον χρόνο. Συνεπώς η αύξηση της ακρίβειας απλώς δείχνει ότι οι κλάσεις που μαθαίνουν καλύτερα οι ταξινομητές, αυξάνονται. Τα ίδια συμπεράσματα προκύπτουν και από το διάγραμμα του σταθμισμένου δείκτη F_1. Ο μέσος δείκτης F_1 όμως, ο οποίος δεν επηρεάζεται περισσότερο από τις πολυπληθείς κλάσεις αλλά λαμβάνει υπόψιν εξίσου όλες τις κλάσεις, δείχνει κάτι διαφορετικό. Ενώ μέχρι και το δείγμα 5 οι τιμές του δείκτη και των τεσσάρων ταξινομητών είναι περίπου ίδιες, από το δείγμα 6 και μετά (εκπαίδευση με το δείγμα 5) παρουσιάζεται μια απόκλιση όπου, αυτή τη φορά, οι τιμές του ταξινομητή learn++NSE ο οποίος επανεκπαιδευόταν με κάθε νέο δείγμα και του ταξινομητή svm ο οποίος εκπαιδευόταν κάθε φορά μόνο με το προηγούμενο δείγμα από αυτό βάσει του οποίου ελεγχόταν, παρουσιάζουν πτώση. Αυτό σημαίνει ότι η απόδοση ταξινόμησης των ολιγομελών κλάσεων φθίνει. Ο λόγος είναι ο ίδιος με αυτόν που έκανε αυτούς τους ταξινομητές να παρουσιάζουν καλύτερη απόδοση στην ταξινόμηση των ολιγομελών κλάσεων του dataset d3! Ο ταξινομητής svm εκπαιδεύεται κάθε φορά με όλο και πιο imbalanced δείγματα και ο ταξινομητής learn++NSE δίνει ιδιαίτερη βαρύτητα στο τελευταίο δείγμα εκπαίδευσης το όποιο είναι όλο και πιο imbalanced, συνεπώς σταδιακά χάνουν την ικανότητα καλής ταξινόμησης των ολιγομελών δειγμάτων. Στο Παράρτημα ΙΙΙ παρουσιάζεται οπτικοποίηση των μητρών σύγχυσης που προέκυψαν από την αξιολόγηση κάθε δείγματος και για τους 4 ταξινομητές.

Τα συμπεράσματα που προέκυψαν για τον ταξινομητή learn++NSE με χρήση του dataset d4, δείχνουν ότι όταν ο ταξινομητής εκπαιδεύεται με imbalanced δείγματα δεν είναι σε θέση να κατατάσσει σωστά τις ολιγομελείς κλάσεις αν δεν επανεκπαιδευτεί. Βέβαια η συνολική ακρίβεια ταξινόμησης αυξάνει πάραυτα. Συνεπώς ο χρήστης θα πρέπει να αποφασίσει αν για το πρόβλημα που θέλει να λύσει με τον ταξινομητή learn++NSE, είναι πιο σημαντικό να επιτυγχάνει υψηλή ακρίβεια (επανεκπαιδεύσεις, ανεξαρτήτως αν τα δείγματα επανεκπαίδευσης είναι balanced ή imbalanced) ή να επιτυγχάνει καλύτερη ταξινόμηση των ολιγομελών κλάσεων (αποφυγή επανεκπαιδεύσεων με imbalanced δείγματα). Πρέπει επίσης να επισημανθεί, ότι αν η επανεκπαίδευση του ταξινομητή γινόταν με κάποιον εκκινητή βασισμένο στην ακρίβεια ταξινόμησης των νέων δειγμάτων (όπως στην ενότητα 3.3), πιθανώς να μην ενεργοποιούταν καμία επανεκπαίδευση. Από το πάνω δεξιά διάγραμμα στην **Εικόνα 42** φαίνεται ότι η ακρίβεια ταξινόμησης του ταξινομητή που δεν επανεκπαιδεύεται παρουσιάζει πολύ μικρή αύξηση. Πιο συγκεκριμένα, αύξηση της ακρίβειας παρατηρείται κυρίως όταν αλλάζει η σύσταση του υπό εξέταση δείγματος και είναι της τάξης του 1 με 2%. Επίσης η διαφορά της ακρίβειας όπως υπολογίστηκε από το δείγμα 1 και από το δείγμα 12 παρουσιάζει αύξηση που αγγίζει μόλις το 5%.

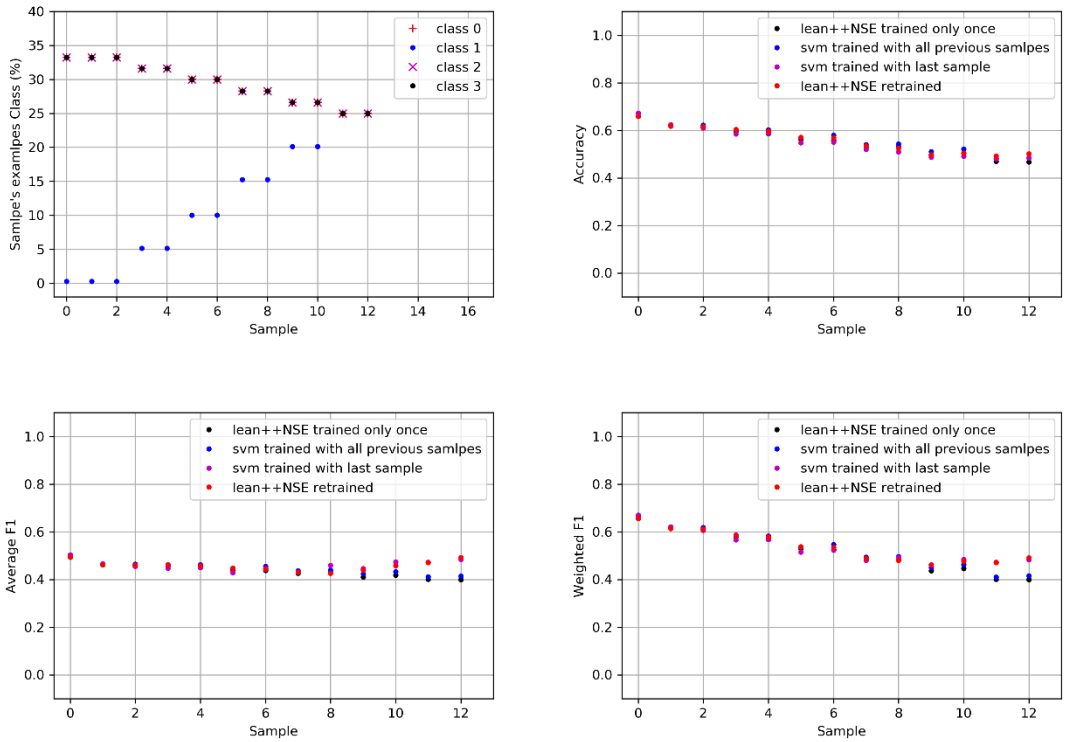
Λόγω των συμπερασμάτων που έχουν εξαχθεί μέχρι στιγμής, κρίθηκε σκόπιμο να ελεγχθεί η συμπεριφορά του ταξινομητή learn++NSE που εκπαιδεύεται με αυξητικό τρόπο στην περίπτωση κατά την οποία στο dataset όλες οι κλάσεις αποτελούνται από ισάριθμα παραδείγματα εκτός από μία. Αυτή η κλάση αρχικά αποτελείται από ελάχιστα παραδείγματα (πρακτικά δεν υπάρχουν παραδείγματα της κλάσης αυτής στο δείγμα) ενώ με την πάροδο του χρόνου ο αριθμός των παραδειγμάτων αυξάνει. Για τον λόγο αυτό δημιουργήθηκαν δύο

datasets τα d5 και d6. Στον **Πίνακας 15** παρουσιάζεται ο αριθμός των παραδειγμάτων που ανήκουν σε κάθε κλάση των δειγμάτων των datasets d5 και d6. Ο ολικός αριθμός των παραδειγμάτων αυτών των datasets, επιτρέπει την επιλογή τους με τέτοιο τρόπο, ώστε κάθε κλάση να αποτελείται κατά 50% από παραδείγματα και των δύο διαφορετικών βαθμολογιών που την αποτελούν (π.χ. η κλάση 0 αποτελείται κατά 50% από παραδείγματα με βαθμολογία 1 και κατά 50% από παραδείγματα με βαθμολογία 50%). Οι ταξινομητές, η διαδικασία καθώς και οι δείκτες για την αξιολόγηση των ταξινομητών, είναι ίδιοι με αυτούς που χρησιμοποιήθηκαν προηγουμένως.

	Dataset d5					Dataset d6			
Κλάσεις	0	1	2	3		0	1	2	3
Πρώτη Εκπαίδευση	1.330	10	1.330	1.330		10	1.330	1.330	1.330
Δείγμα 1	665	5	665	665		5	665	665	665
Δείγμα 2	665	5	665	665		5	665	665	665
Δείγμα 3	633	102	632	633		102	632	632	633
Δείγμα 4	632	103	633	632		103	633	633	632
Δείγμα 5	600	200	600	600		200	600	600	600
Δείγμα 6	600	200	600	600		200	600	600	600
Δείγμα 7	565	305	565	565		305	565	565	565
Δείγμα 8	565	305	565	565		305	565	565	565
Δείγμα 9	532	402	533	533		402	533	533	533
Δείγμα 10	533	403	532	532		403	532	532	532
Δείγμα 11	500	500	500	500		500	500	500	500
Δείγμα 12	500	500	500	500		500	500	500	500
Σύνολο	8.320	3.040	8.320	8.320		3.040	8.320	8.320	8.320

***Πίνακας 15**: Αριθμός παραδειγμάτων από κάθε κλάση που συγκροτούν τα δείγματα των datasets d4 και d5.*

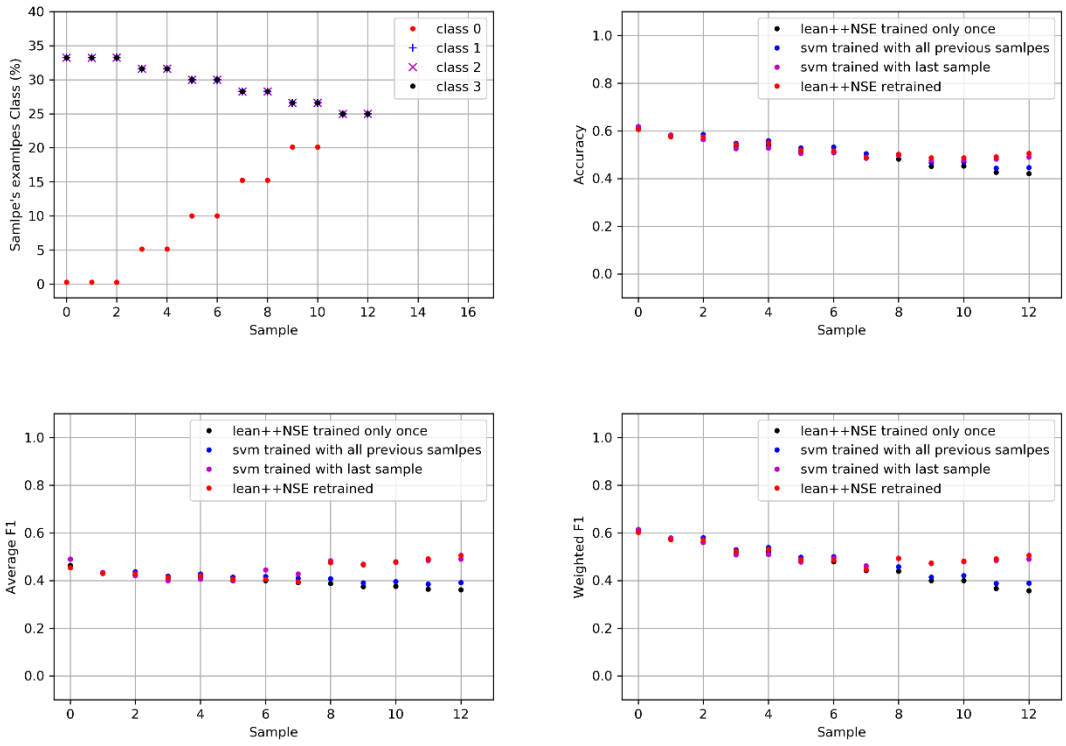
Στην **Εικόνα 43** και στην **Εικόνα 44** παρουσιάζονται τα αποτελέσματα της αξιολόγησης των ταξινομητών με τα datasets d5 και d6 αντίστοιχα. Πιο συγκεκριμένα στα άνω αριστερά διαγράμματα παρουσιάζεται το ποσοστό των παραδειγμάτων από κάθε κλάση των δειγμάτων των datasets, ενώ στα διαγράμματα στα άνω δεξιά, κάτω αριστερά και κάτω δεξιά παρουσιάζεται αντίστοιχα η ακρίβεια ταξινόμησης, ο μέσος δείκτης F_1 και ο σταθμισμένος δείκτης F_1, ως προς το δείγμα βάσει του οποίου έχουν υπολογιστεί. Στη θέση 0 παρουσιάζεται η αξιολόγηση των ταξινομητών με το δείγμα πρώτης εκπαίδευσης με το οποίο και εκπαιδεύτηκαν. Με κόκκινο παρουσιάζεται ο ταξινομητής learn++NSE ο οποίος επανεκπαιδευόταν με κάθε νέο δείγμα, με μαύρο ο ταξινομητής learn++NSE ο οποίος εκπαιδεύτηκε μόνο με το δείγμα πρώτης εκπαίδευσης, με μπλε ο ταξινομητή smv ο οποίος εκπαιδευόταν με όλα τα (μέχρι τότε) διαθέσιμα παραδείγματα και με μοβ ο ταξινομητής svm ο οποίος εκπαιδευόταν κάθε φορά μόνο με το προηγούμενο δείγμα από αυτό βάσει του οποίου ελεγχόταν.



***Εικόνα 43**: Άνω αριστερά: Ποσοστό των παραδειγμάτων από κάθε κλάση των δειγμάτων του dataset d5. Άνω δεξιά, κάτω αριστερά και κάτω δεξιά παρουσιάζεται αντίστοιχα η ακρίβεια ταξινόμησης, ο μέσος δείκτης F_1 και ο σταθμισμένος δείκτης F_1,ως προς το δείγμα βάσει του οποίου έχουν υπολογιστεί. Στη θέση 0 παρουσιάζεται η αξιολόγηση των ταξινομητών με το δείγμα πρώτης εκπαίδευσης με το οποίο και εκπαιδεύτηκαν.*

Από τα αποτελέσματα που παρουσιάζονται τόσο στην **Εικόνα 43** όσο και στην **Εικόνα 44** εξάγονται τα ίδια συμπεράσματα. Καταρχας, η ακρίβεια όλων των ταξινομητών όπως υπολογίστηκε από τα δείγματα του dataset d5 παρουσιάζει συνεχή πτωτική τάση, ενώ από το dataset d6 από τα δείγματα 10, 11 και 12 παρατηρείται μια σταθεροποίηση και ίσως μια αμυδρά ανοδική τάση για τον ταξινομητή learn++NSE που εκπαιδευόταν με αυξητικό τρόπο και για τον ταξινομητή svm που εκπαιδευόταν με το τελευταίο δείγμα. Οι τιμές του μέσου δείκτη F_1 για όλους τους ταξινομητές είναι σχεδόν ταυτόσημες στα πρώτα δείγματα. Στη συνέχεια όμως παρατηρείται απόκλιση των τιμών, με αυτές του ταξινομητή learn++NSE που εκπαιδευόταν με αυξητικό τρόπο και του ταξινομητή svm που εκπαιδευόταν με το τελευταίο δείγμα να είναι υψηλότερες από τις τιμές των άλλων δύο ταξινομητών. Στο το dataset d5, η απόκλιση αρχίζει να εμφανίζεται από τα δείγματα 9 με 10 ενώ στο το dataset d6 είναι φανερή από το δείγμα 8. Τέλος, παρόμοια είναι και η συμπεριφορά των τιμών του σταθμισμένου δείκτη F_1. Στο dataset d5, η απόκλιση ξεκινά από το δείγμα 8 για τον ταξινομητή svm και από το δείγμα 9 για τον ταξινομητή learn++NSE ενώ στο dataset d6 ξεκινά από το δείγμα 6 για τον ταξινομητή svm και από το δείγμα 8 για τον ταξινομητή learn++NSE. Φαίνεται λοιπόν ότι ο ταξινομητής learn++NSE που εκπαιδευόταν με αυξητικό τρόπο και ο ταξινομητής svm που εκπαιδευόταν με το τελευταίο δείγμα μπορούν και εντοπίζουν την ύπαρξη της «νέας» κλάσης, με τον ταξινομητή svm να είναι ο πρώτος που την εντοπίζει ενώ ο learn++NSE χρειάζεται μία με δύο επανεκπαιδεύσεις περισσότερες. Αυτά τα αποτελέσματα

επιβεβαιώνουν τα αποτελέσματα που εξήχθησαν χρησιμοποιώντας το dataset d3 και τα αντίστοιχα συμπεράσματα.



***Εικόνα 44**: Άνω αριστερά: Ποσοστό των παραδειγμάτων από κάθε κλάση των δειγμάτων του dataset d6. Άνω δεξιά, κάτω αριστερά και κάτω δεξιά παρουσιάζεται αντίστοιχα η ακρίβεια ταξινόμησης, ο μέσος δείκτης F_1 και ο σταθμισμένος δείκτης F_1,ως προς το δείγμα βάσει του οποίου έχουν υπολογιστεί. Στη θέση 0 παρουσιάζεται η αξιολόγηση των ταξινομητών με το δείγμα πρώτης εκπαίδευσης με το οποίο και εκπαιδεύτηκαν.*

Οι διαφορές που παρατηρούνται στα αποτελέσματα των δύο datasets οφείλονται στην κατασκευή των κλάσεων και στο πόσο διαφέρουν τα παραδείγματα της κάθε κλάσης από αυτά των άλλων κλάσεων. Στο dataset d5 η κλάση που ουσιαστικά λείπει από τα πρώτα δείγματα και εμφανίζεται σταδιακά στα επόμενα είναι η κλάση 1, ενώ στο d6 είναι η κλάση 0. Στην αρχή της ενότητας, παρουσιάστηκε η ικανότητα των ταξινομητών να ξεχωρίζουν την κάθε κλάση από τις άλλες, και φαίνεται ότι η ακρίβεια (δυαδικής) ταξινόμησης της κλάσης 1 από την κλάση 0 είναι ~66%, της κλάσης 1 από την κλάση 2 είναι ~74% ενώ της κλάσης 0 από την κλάση 2 είναι ~82%. Αυτό έχει ως αποτέλεσμα στην περίπτωση του dataset d5 τα παραδείγματα της κλάσης 1 να κατατάσσονται αρχικά (λανθασμένα) κατά ~55% στην κλάση 0 και κατά 35% στην κλάση 2. Όταν ο ταξινομητής αρχίζει να ανιχνεύει την ύπαρξη παραδειγμάτων της κλάσης 1 πρέπει να τα ξεχωρίσει τόσο από την κλάση 0 όσο και από την 2. Στην περίπτωση του dataset d6, τα παραδείγματα της κλάσης 0 κατατάσσονται αρχικά (λανθασμένα) κατά ~80% στην κλάση 1 και κατά 10% στην κλάση 2, συνεπώς για την αύξηση των σωστά ταξινομημένων παραδειγμάτων της κλάσης 0 ο ταξινομητής πρέπει, κατά κύριο λόγο, να τα ξεχωρίσει μόνο από την κλάση 1. Αυτός είναι ο λόγος για τον οποίο στο dataset d5 η κλάση που λείπει από τα αρχικά δείγματα ανιχνεύεται πιο εύκολα από ότι στο dataset

d6, και για τον οποίο εμφανίζονται οι διαφορές στα διαγράμματα στην **Εικόνα 43** και στην **Εικόνα 44**.

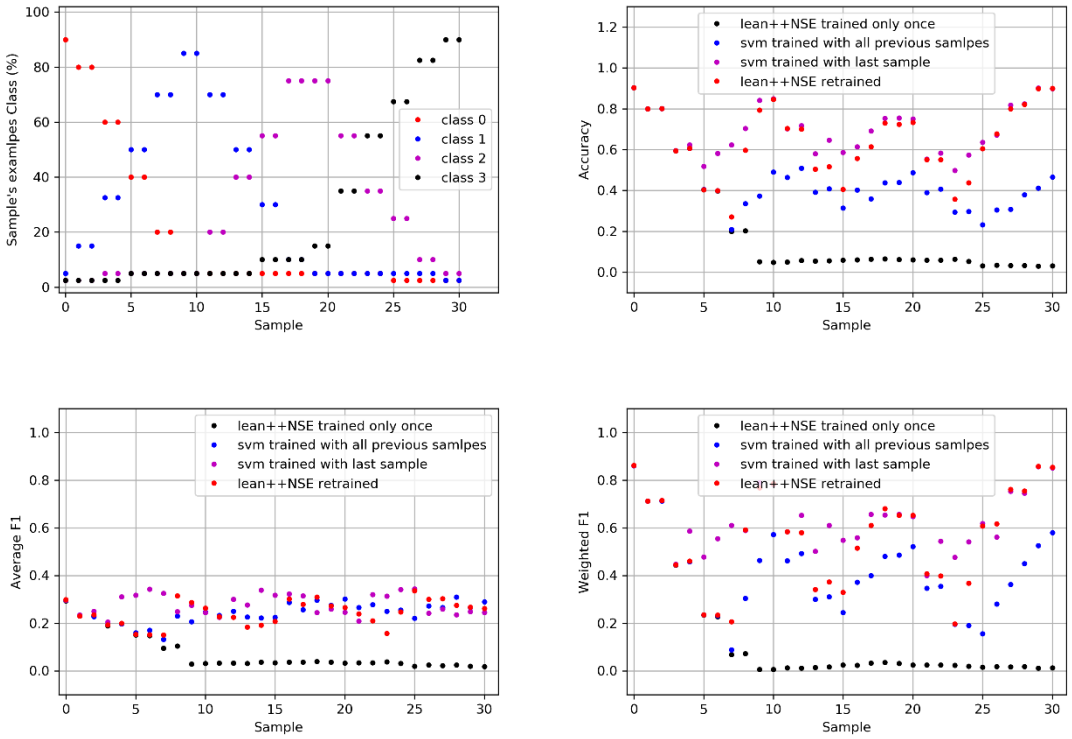
Ο τελευταίος έλεγχος του ταξινομητή learn++NSE ο οποίος εκπαιδεύεται με αυξητικό τρόπο, διεξήχθη με dataset στο οποίο με την πάροδο του χρόνου το ποσοστό των παραδειγμάτων που ανήκαν σε μία κλάση αύξανε, άγγιζε σχεδόν το 100% των παραδειγμάτων του δείγματος και στη συνέχεια μειωνόταν, ενώ ταυτόχρονα αύξανε το ποσοστό των παραδειγμάτων που ανήκαν σε άλλη κλάση. Για την προσομοίωση αυτής της περίπτωσης δημιουργήθηκε το dataset d7. Στον **Πίνακας 16** παρουσιάζεται ο αριθμός των παραδειγμάτων που ανήκουν σε κάθε κλάση των δειγμάτων του dataset d7. Όπως και στα datasets d5 και d6, ο ολικός αριθμός των παραδειγμάτων αυτού του dataset, επιτρέπει την επιλογή τους με τέτοιο τρόπο, ώστε κάθε κλάση να αποτελείται κατά 50% από παραδείγματα και των δύο διαφορετικών βαθμολογιών που την αποτελούν. Για τον ταξινομητή learn++NSE που χρησιμοποιήθηκε σε αυτό τον έλεγχο, χρησιμοποιήθηκε batch size 1000 (όπως δηλαδή και σε αυτούς του προηγούμενου κεφαλαίου). Οι υπόλοιποι ταξινομητές, η διαδικασία καθώς και οι δείκτες για την αξιολόγηση των ταξινομητών, είναι ίδιοι με αυτούς που χρησιμοποιήθηκαν προηγουμένως.

	Dataset d7			
Κλάσεις	0	1	2	3
Πρώτη Εκπαίδευση	1.800	50	25	25
Δείγματα 1, 2	800	150	25	25
Δείγματα 3, 4	600	325	50	25
Δείγματα 5, 6	400	500	50	50
Δείγματα 7, 8	200	700	50	50
Δείγματα 9, 10	50	850	50	50
Δείγματα 11, 12	50	700	200	50
Δείγματα 13, 14	50	500	400	50
Δείγματα 15, 16	50	300	550	100
Δείγματα 17, 18	50	100	750	100
Δείγματα 19, 20	50	50	750	150
Δείγματα 21, 22	50	50	550	350
Δείγματα 23, 24	50	50	350	550
Δείγματα 25, 26	25	50	250	675
Δείγματα 27, 28	25	50	100	825
Δείγματα 29, 30	25	25	50	900
Σύνολο	6.750	8.900	8.400	7.950

Πίνακας 16*: Αριθμός παραδειγμάτων από κάθε κλάση που συγκροτούν τα δείγματα του dataset d7.*

Στην **Εικόνα 45** άνω αριστερά, παρουσιάζεται το ποσοστό των παραδειγμάτων από κάθε κλάση των δειγμάτων του dataset d7, ενώ άνω δεξιά, κάτω αριστερά και κάτω δεξιά παρουσιάζεται αντίστοιχα η ακρίβεια ταξινόμησης, ο μέσος δείκτης F_1 και ο σταθμισμένος δείκτης F_1, ως προς το δείγμα βάσει του οποίου έχουν υπολογιστεί. Στη θέση 0 παρουσιάζεται η αξιολόγηση των ταξινομητών με το δείγμα πρώτης εκπαίδευσης με το οποίο και εκπαιδεύτηκαν. Με κόκκινο παρουσιάζεται ο ταξινομητής learn++NSE ο οποίος

επανεκπαιδευόταν με κάθε νέο δείγμα, με μαύρο ο ταξινομητής learn++NSE ο οποίος εκπαιδεύτηκε μόνο με το δείγμα πρώτης εκπαίδευσης, με μπλε ο ταξινομητή smv ο οποίος εκπαιδευόταν με όλα τα (μέχρι τότε) διαθέσιμα παραδείγματα και με μοβ ο ταξινομητής svm ο οποίος εκπαιδευόταν κάθε φορά μόνο με το προηγούμενο δείγμα από αυτό βάσει του οποίου ελεγχόταν.



***Εικόνα 45**: Άνω αριστερά: Ποσοστό των παραδειγμάτων από κάθε κλάση των δειγμάτων του dataset d7. Άνω δεξιά, κάτω αριστερά και κάτω δεξιά παρουσιάζεται αντίστοιχα η ακρίβεια ταξινόμησης, ο μέσος δείκτης F_1 και ο σταθμισμένος δείκτης F_1,ως προς το δείγμα βάσει του οποίου έχουν υπολογιστεί. Στη θέση 0 παρουσιάζεται η αξιολόγηση των ταξινομητών με το δείγμα πρώτης εκπαίδευσης με το οποίο και εκπαιδεύτηκαν.*

Ξεκινώντας τον σχολιασμό των πιο πάνω διαγραμμάτων με αυτό της ακρίβειας ταξινόμησης, φαίνεται ότι ο ταξινομητής learn++NSE που εκπαιδεύεται μόνο με το πρώτο δείγμα, αποτυγχάνει στην ταξινόμηση των δειγμάτων που δεν περιέχουν την κλάση με την οποία εκπαιδεύτηκε (όπως και αναμενόταν). Επίσης η ακρίβεια ταξινόμησης που επιτυγχάνεται με τον ταξινομητή svm που εκπαιδεύεται με όλο το διαθέσιμο δείγμα, είναι φανερά υποδεέστερη αυτής που επιτυγχάνεται με τον ταξινομητή learn++NSE που επανεκπαιδευόταν με κάθε δείγμα και αυτής που επιτυγχάνεται με τον ταξινομητή svm που εκπαιδευόταν κάθε φορά με το προηγούμενο δείγμα. Η κατανομή της ακρίβειας του ταξινομητή learn++NSE που επανεκπαιδευόταν με κάθε νέο δείγμα φαίνεται ότι ακολουθεί την κατανομή της κλάσης με το μεγαλύτερο ποσοστό. Βέβαια, κάθε φορά που υπάρχει αλλαγή της κλάσης με το μεγαλύτερο ποσοστό, η κατανομή που ακολουθεί η ακρίβεια αλλάζει (από την προηγούμενη κλάση με το μεγαλύτερο ποσοστό στη νέα) μετά από περίπου δύο επανεκπαιδεύσεις (π.χ. η κλάση 1 γίνεται επικρατούσα στο δείγμα 5 ενώ πριν

επικρατούσα κλάση ήταν η 0, η ακρίβεια ξεκινά να ακολουθεί την κατανομή της κλάσης 1 αντί της κατανομής της κλάσης 0 στο δείγμα 8 – εκπαίδευση με δείγμα 7). Η κατανομή της ακρίβειας του ταξινομητή svm που εκπαιδευόταν κάθε φορά με το προηγούμενο δείγμα ακολουθεί την κατανομή της κλάσης με το μεγαλύτερο ποσοστό χωρίς όμως να φαίνεται η υστέρηση που επιδεικνύει ο ταξινομητής learn++NSE. Επίσης οι τιμές τις ακρίβειας είναι υψηλότερες από αυτές του ταξινομητή learn++NSE στις περιοχές όπου υπάρχει η προαναφερθείσα υστέρηση. Από τον μέσο δείκτη F_1 εξάγονται παρόμοια συμπεράσματα. Ιδιαίτερο ενδιαφέρον παρουσιάζει η κατανομή των τιμών του σταθμισμένου δείκτη F_1. Για τον ταξινομητή learn++NSE, φαίνεται ότι οι τιμές του σταθμισμένου δείκτη F_1 ελαττώνονται σταδιακά, ενώ δύο επανεκπαιδεύσεις μετά το σημείο στο όποιο αλλάζει η κυρίαρχη κλάση παρατηρείται ένα άλμα, το οποίο ακολουθείται από σταδιακή πτώση. Αυτό συμβαίνει διότι όσο αυξάνεται το ποσοστό της κυρίαρχης κλάσης, ο ταξινομητής εκπαιδεύεται με imbalanced δείγματα, συνεπώς δεν μπορεί να αναγνωρίσει καλά τις άλλες κλάσεις (όπως έχει ήδη διαπιστωθεί από προηγούμενους πειραματισμούς). Αφού το ποσοστό της κυρίαρχης κλάσεις ελαττωθεί και το δείγμα δεν είναι πλέον imbalanced, ο ταξινομητής δεν είναι σε θέση να αναγνωρίσει τις υπόλοιπες κλάσεις πριν εκτελεστούν μία με δύο επανεκπαιδεύσεις. Τότε είναι σε θέση να αναγνωρίσει και την πρώην αλλά και τη νυν κυρίαρχη κλάση και παρατηρείται το άλμα στην τιμή του σταθμισμένου δείκτη F_1. Αλλά καθώς το δείγμα αλλάζει πάλι, ο ταξινομητής εκπαιδεύεται ξανά με imbalanced δείγματα και η απόδοσή του πέφτει ξανά. Οι τιμές του σταθμισμένου δείκτη F_1 του ταξινομητή svm που εκπαιδευόταν κάθε φορά με το προηγούμενο δείγμα παρουσιάζουν παρόμοια συμπεριφορά χωρίς όμως την υστέρηση που επιδεικνύει ο ταξινομητής learn++NSE. Η κατανομή του είναι πιο ομαλή, λαμβάνει μέγιστες τιμές λίγο μετά τα σημεία στα οποία αλλάζει η κλάση με το μεγαλύτερο ποσοστό (όπου έχει εκπαιδευτεί με δείγματα λιγότερο imbalanced) και ελάχιστες τιμές λίγο μετά τα σημεία μεγιστοποίησης του ποσοστού κάποιας κλάσης (όπου έχει εκπαιδευτεί με πιο imbalanced δείγματα).

Συμπερασματικά από αυτή την ενότητα προκύπτει ότι ο ταξινομητής learn++NSE που εκπαιδεύεται με αυξητικό τρόπο, έχει καλή απόδοση όταν χρησιμοποιείται σε δείγματα στα οποία οι κλάσεις αποτελούνται από άνισο αριθμό παραδειγμάτων ο οποίος μεταβάλλεται με τον χρόνο. Πιο συγκεκριμένα, όταν ο ταξινομητής έχει εκπαιδευτεί με ομογενές δείγμα, είναι ικανός να ξεχωρίσει παραδείγματα που ανήκουν σε διαφορετικές κλάσεις ανεξαρτήτως του μεγέθους των κλάσεων. Αν αναγκαστεί σε εκπαίδευση ή επανεκπαίδευση με ανομοιογενή δείγματα (τα οποία όμως ακολουθούν την τάση του dataset) η ακρίβεια ταξινόμησης των ολιγομελών κλάσεων ελαττώνεται σημαντικά, αλλά η συνολική ακρίβεια ταξινόμησης αυξάνεται. Επιπλέον μπορεί να εντοπίσει σε πολύ καλό βαθμό, κλάσεις που αρχικά αποτελούνται από ελάχιστα παραδείγματα ενώ με την πάροδο του χρόνου ο αριθμός των παραδειγμάτων τους αυξάνει. Τέλος επιδεικνύει καλά αποτελέσματα και σε δείγματα στα οποία με την πάροδο του χρόνου, το ποσοστό των παραδειγμάτων που ανήκουν σε μία κλάση αυξάνει, αγγίζει σχεδόν το 100% και στη συνέχεια μειώνεται ενώ ταυτόχρονα αυξάνει το ποσοστό των παραδειγμάτων που ανήκουν σε άλλη κλάση. Σε όλους του ελέγχους που διεξήχθησαν, τα αποτελέσματα ήταν σαφώς καλύτερα από αυτά ταξινομητή svm που εκπαιδευόταν με όλο το διαθέσιμο (μέχρι τότε) δείγμα και συγκρίσιμα ή ακόμα και εξίσου καλά με αυτά ταξινομητή svm που εκπαιδευόταν μόνο με το προηγούμενο δείγμα (λόγω της εκπαίδευσης μόνο με το προηγούμενο δείγμα, αυτός ο ταξινομητής αναμένεται να έχει βέλτιστα αποτελέσματα σε δείγματα που μεταβάλλονται ομαλά με την πάροδο του χρόνου).

3.5. Συμπεράσματα

Στο προηγούμενο κεφάλαιο επιλέχθηκε η συλλογή ταξινομητών Learn++NSE ως βέλτιστο εργαλείο για την κατηγοριοποίηση μικρών κειμένων με βάση το συναίσθημά τους, με τρόπο αυξητικό. Σε αυτό το κεφάλαιο εξετάστηκε λεπτομερώς η συμπεριφορά του ταξινομητή Learn++NSE στις περιπτώσεις όπου οι κλάσεις μεταβάλλονται με την πάροδο του χρόνου και όπου ο αριθμός των παραδειγμάτων των κλάσεων διαφέρει σημαντικά.

Αρχικά επιβεβαιώθηκε ότι με την εκπαίδευση και επανεκπαίδευση του ταξινομητή με πολλά παραδείγματα από το ίδιο ομογενές δείγμα, η ακρίβεια ταξινόμησής του αυξάνει, προσεγγίζοντας την ακρίβεια ταξινομητή που εκπαιδεύεται μία φόρα (όχι αυξητικά) με το σύνολο των παραδειγμάτων αυτών. Στην συνέχεια μελετήθηκε η συμπεριφορά του, όταν η σύσταση των κλάσεων αλλάζει με τον χρόνο. Από αυτούς τους πειραματισμούς διαπιστώθηκε ότι η ακρίβειά του είναι σαφώς καλύτερη από αυτή ταξινομητή που δεν επανεκπαιδεύεται. Επίσης, η ακρίβειά του είναι μόλις 1% χαμηλότερη από αυτή που επιτυγχάνεται από ταξινομητή που εκπαιδεύεται μία φόρα με το σύνολο των παραδειγμάτων, ενώ ο υπολογιστικός χρόνος που απαιτείται για την εκπαίδευσή του μπορεί να είναι έως και 50 φορές μικρότερος για τα δείγματα που μελετήθηκαν. Για πιο πολυπληθή δείγματα, αναμένεται ότι η χρονική διαφορά θα είναι ακόμα μεγαλύτερη. Τέλος η απόδοσή του σε δείγματα στα οποία ο αριθμός των παραδειγμάτων των διαφορετικών κλάσεων μεταβάλλεται με τον χρόνο, είναι συγκρίσιμη με αυτή ταξινομητή που εκπαιδεύεται μόνο με το πιο πρόσφατο δείγμα και σαφώς καλύτερη από ταξινομητή που εκπαιδεύεται μία φόρα με το σύνολο των παραδειγμάτων. Διαπιστώθηκε όμως, ότι όταν ο ταξινομητής εκπαιδεύεται με imbalanced δείγματα δεν είναι σε θέση να κατατάσσει σωστά τις ολιγομελείς κλάσεις αν δεν επανεκπαιδευτεί. Βέβαια, η συνολική ακρίβεια ταξινόμησης αυξάνει καθώς ενισχύεται η σωστή κατηγοριοποίηση των πολυπληθών κλάσεων.

3.6. Ολοκληρωμένο εργαλείο ταξινόμησης μικρών κειμένων με βάση το συναίσθημα με τη συλλογή ταξινομητών learn++NSE.

Αξιοποιώντας την έρευνα που διεξήχθη, κατασκευάστηκε ένα ολοκληρωμένο εργαλείο σε γλώσσα python, το οποίο κατηγοριοποιεί μικρά κείμενα με βάση το συναίσθημά τους, το οποίο βασίζεται σε συλλογή learn++NSE ταξινομητών SVC.

Ο χρήστης επιλέγει το batch size και το αν θα χρησιμοποιηθεί pruning και με ποιον τρόπο (με βάση το σφάλμα ή την παλαιότητα). Επίσης πρέπει να επιλέξει το αριθμό των κλάσεων. Για την επανεκπαίδευση του ταξινομητή δίνονται 2 επιλογές:

1) Συνεχής επανεκπαίδευσης, δηλαδή επανεκπαίδευση κάθε φορά που ο αριθμός των νέων παραδειγμάτων με γνωστή κλάση γίνει ίσος με το batch size.
2) Επανεκπαίδευση όταν η ακρίβεια των τελευταίων παραδειγμάτων, αριθμού ίσου με το batch size, διαφέρει κατά συγκεκριμένο ποσοστό (το οποίο επιλέγεται από τον χρήστη) από την κρίσιμη ακρίβεια. Ως κρίσιμη ορίζεται η ακρίβεια που επιτυγχάνεται από το τελευταίο δείγμα με το οποίο έχει επανεκπαιδευτεί ο

ταξινομητής (ή από το δείγμα πρώτης εκπαίδευσης, αν δεν έχει εκτελεστεί επανεκπαίδευση).

Για το δείγμα με το οποίο θα εκτελεστεί η επανεκαπαιδεύση δίνονται δύο δυνατότητες:

1) Επανεκπαίδευση με το πιο πρόσφατο διαθέσιμο δείγμα.
2) Επανεκπαίδευση όταν το ποσοστό των δειγμάτων που ανήκουν σε κάθε κλάση γίνει ίσο με κάποιο προκαθορισμένο από τον χρήστη ποσοστό. Από κάθε κλάση «αποθηκεύονται» τα τελευταία N παραδείγματα (όπου N είναι το προκαθορισμένο ποσοστό του batch size). Κάθε φορά που αξιολογείται ένα νέο παράδειγμα ελέγχου, αν δεν έχει συμπληρωθεί ο αριθμός των N παραδειγμάτων της κλάσης στην οποία ανήκει, το παράδειγμα αυτό απλώς συγκρατείται, στην αντίθετη περίπτωση συγκρατείται το νέο παράδειγμα ενώ ταυτόχρονα απορρίπτεται το παλαιότερο παράδειγμα της κλάσης αυτής (και όχι το παλαιότερο γενικά). Αυτός ο τρόπος επανεκπαίδευσης κρύβει τον κίνδυνο να μην εκτελείται επανεκπαίδευση (αν δεν έχουν συμπληρωθεί τα προαπαιτούμενα παραδείγματα από κάθε κλάση) ακόμα και αν η ακρίβεια ταξινόμησης έχει ελαττωθεί πολύ. Αυτή η δυνατότητα επανεκπαίδευσης παρέχεται ώστε να είναι δυνατή η επανεκπαίδευση με balanced δείγμα ακόμα και στην περίπτωση που το dataset είναι imbalanced .

Το δείγμα εκπαίδευσης πρέπει να δοθεί από τον χρήστη υπό τη μορφή ενός .csv αρχείου χωρίς επικεφαλίδες, στο οποίο στην πρώτη στήλη θα αναγράφεται η κλάση ενώ μετά θα ακολουθεί το κείμενο. Η αρίθμηση των κλάσεων θα πρέπει να ξεκινά από το 0 και να είναι διαδοχικοί ακέραιοι αριθμοί. Η ύπαρξη κομμάτων “,” στο κείμενο, δεν επηρεάζει την λειτουργία του εργαλείου. Πιο κάτω παρουσιάζεται παράδειγμα ενός τέτοιου αρχείου:

1, This was the best movie ever! The scenery was amazing, the director was brilliant and …

0, Do not see this movie. I was constantly looking at the time, hoping for the movie to end …

Με τον ίδιο τρόπο πρέπει να δοθούν και τα δεδομένα ελέγχου, καθώς και τα άγνωστα προς ταξινόμηση δεδομένα. Σε αυτά, ως κλάση (πρώτο στοιχείο στο .svc αρχείο) θα πρέπει να έχει δοθεί ένας αρνητικός ακέραιος αριθμός. Και στις τρεις περιπτώσεις, τα διαφορετικά παραδείγματα πρέπει να βρίσκονται σε διαφορετική γραμμή. Συνεπώς όταν αλλάζει γραμμή, το εργαλείο θα αντιμετωπίζει τη νέα γραμμή σαν νέο παράδειγμα. Για αυτό τον λόγο, ο χρήστης θα πρέπει να απαλείψει τους χαρακτήρες νέας γραμμής από το κείμενο κάθε παραδείγματος. Τα δεδομένα του δείγματος εκπαίδευσης ανακατεύονται με τυχαίο τρόπο προτού χρησιμοποιηθούν για την εκπαίδευση του ταξινομητή, ενώ τα υπόλοιπα παραδείγματα (ελέγχου και προς ταξινόμηση) επεξεργάζονται με την σειρά που βρίσκονται στο αρχείο.

Τα στάδια επεξεργασίας κάθε γραμμής των .csv αρχείων είναι τα ακόλουθα:

1) Από τη συμβολοσειρά επιλέγονται όλοι οι χαρακτήρες μέχρι το πρώτο κόμμα ”,” οι οποίοι αποτελούν την κλάση στην οποία ανήκει το παράδειγμα. Όλοι οι υπόλοιποι χαρακτήρες αποτελούν το κείμενο.
2) Το κείμενο προεπεξεργάζεται με τον τρόπο που αναλύθηκε στην ενότητα 2.1 και χρησιμοποιείται το μοντέλο glove6B_100 για την μετατροπή του σε διάνυσμα διάστασης 100.

Το εργαλείο αποτελείται από πολλές συναρτήσεις, μια προτεινόμενη ροή είναι η ακόλουθη:

1) Αρχικοποίηση της συλλογής ταξινομητών learn++NSE σύμφωνα με τις παραμέτρους που δόθηκαν από τον χρήστη.
2) Εκπαίδευση του ταξινομητή με το δείγμα εκπαίδευσης.
3) Αξιολόγηση και επανεκπαίδευση του ταξινομητή με το δείγμα ελέγχου, ή εύρεση της κλάσης άγνωστων παραδειγμάτων.

4 Σύνοψη

Η ανάλυση συναισθήματος σε περιβάλλον microblogging αποτελεί ένα πολύ χρήσιμο εργαλείο στη σύγχρονη εποχή, δεδομένου ότι πολύ μεγάλο τμήμα του παγκόσμιου πληθυσμού εκφράζεται στα μέσα κοινωνικής δικτύωσης ή σε άλλες ιστοσελίδες. Μία καινοτόμα προσέγγιση της ανάλυσης συναισθήματος είναι η αξιοποίηση ταξινομητών που εκπαιδεύονται με αυξητικό τρόπο. Ο στόχος της παρούσας διπλωματικής εργασίας είναι η μελέτη ενός τέτοιου ταξινομητή.

Αρχικά εξετάστηκε ο τρόπος με τον οποίο το κείμενο μπορεί να μετατραπεί σε αριθμούς (διανύσματα), ώστε να μπορούν να αξιοποιηθούν μέθοδοι μηχανικής μάθησης για την ταξινόμησή του με βάση το συναίσθημα. Από τις διαθέσιμες μεθόδους επιλέχθηκε αυτή των word embedings και πιο συγκεκριμένα χρησιμοποιήθηκε το ήδη εκπαιδευμένο μοντέλο Glove6B_100.

Κατόπιν, μελετήθηκε η απόδοση έξι διαφορετικών ταξινομητών που εκπαιδεύονταν με αυξητικό τρόπο. Οι ταξινομητές που μελετήθηκαν ήταν i) ο Gaussian Naïve Bayesian, ii) ταξινομητής Support Vector Machine με Stochastic Gradient Descent, iii) ο Online Random Forest, iv) ο LaSVM, v) ο ILVQ και τέλος vi) η συλλογή ταξινομητών Learn++NSE. Για την αξιολόγηση των ταξινομητών χρησιμοποιήθηκε το Dataset «Large Movie Review Dataset v1.0» στο οποίο εμπεριέχονται 50.000 κριτικές από ταινίες από το IMDB. Από τους έξι ταξινομητές ως βέλτιστος κρίθηκε η συλλογή ταξινομητών Learn++NSE καθώς παρουσίαζε υψηλή ακρίβεια ταξινόμησης, η οποία αύξανε με την επανεκπαίδευσή του με νέα παραδείγματα, σταθερότητα, δεν παρουσίαζε overtraining όταν εκπαιδευόταν με ικανό αριθμό δειγμάτων και η επανεκπαίδευσή του εκτελούταν με ομάδες παραδειγμάτων.

Στην συνέχεια μελετήθηκε η απόδοση του ταξινομητή Learn++NSE σε περιβάλλοντα τα οποία άλλαζαν με τον χρόνο. Στην περίπτωση όπου η σύσταση των κλάσεων άλλαζε με την πάροδο του χρόνου, διαπιστώθηκε ότι η ακρίβεια του ταξινομητή που εκπαιδευόταν με αυξητικό τρόπο ήταν καλύτερη από αυτή ταξινομητή που δεν επανεκπαιδευόταν, ενώ ήταν μόλις 1% χαμηλότερη από αυτή που επετεύχθη από ταξινομητή που εκπαιδευόταν με το σύνολο των διαθέσιμων παραδειγμάτων. Επίσης ο υπολογιστικός χρόνος που απαιτείτο για την εκπαίδευσή του ήταν έως και 50 φορές μικρότερος από αυτόν που απαιτείτο για την εκπαίδευση του ταξινομητή που εκπαιδευόταν με το σύνολο των διαθέσιμων παραδειγμάτων. Στην περίπτωση όπου ο αριθμός των παραδειγμάτων των διαφορετικών κλάσεων μεταβαλλόταν με τον χρόνο, η απόδοσή του ήταν συγκρίσιμη με αυτή ταξινομητή που εκπαιδευόταν μόνο με το πιο πρόσφατο δείγμα (ο οποίος αναμένεται να έχει πολύ μεγάλη ευαισθησία στη συγκεκριμένη περίπτωση). Τέλος κατασκευάστηκε ολοκληρωμένο εργαλείο ταξινόμησης μικρών κειμένων με βάση το συναίσθημά τους με αυξητικό τρόπο, το οποίο βασίζεται στον ταξινομητή Learn++NSE.

Βιβλιογραφία

[1] R. Collobert και et al., «A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning,» *Proceedings of the 25th International Conference on Machine Learning,* 2008.

[2] T. Mikolov και et al., «Distributed Representations of Words and Phrases and their Compositionality,» *Advances in Neural Information Processing Systems 26,* pp. 3111--3119.

[3] J. Pennington και et al., «GloVe: Global Vectors for Word Representation,» *Empirical Methods in Natural Language Processing (EMNLP),* pp. 1532--1543, 2014.

[4] M. Abadi και e. al., Επιμ.«{TensorFlow}: Large-Scale Machine Learning on Heterogeneous Systems».

[5] «Glove,» [Ηλεκτρονικό]. Available: https://nlp.stanford.edu/projects/glove/. [Πρόσβαση 29 12 2019].

[6] «GloVe GitHub,» [Ηλεκτρονικό]. Available: https://github.com/stanfordnlp/GloVe. [Πρόσβαση 29 12 2019].

[7] «About the Test Data,» [Ηλεκτρονικό]. Available: http://mattmahoney.net/dc/textdata.html. [Πρόσβαση 21 12 2019].

[8] A. Maas και et al., «Learning Word Vectors for Sentiment Analysis,» *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies,* pp. 142-150, June 2011.

[9] «IMBD,» [Ηλεκτρονικό]. Available: https://www.imdb.com/. [Πρόσβαση 20 12 2019].

[10] C. Baziotis και et al., «DataStories at SemEval-2017 Task 4: Deep LSTM with Attention for Message-level and Topic-based Sentiment Analysis,» *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017),* pp. 747-754, 8 2017.

[11] S. Bird και et al., Natural Language Processing with Python, O'Reilly Media Inc., 2009.

[12] K. Koutroumbas και S. Theodoridis, Pattern Recognition, Elsevier, 2008.

[13] S. Ren και et al., «Incremental Naïve Bayesian Learning Algorithm based on Classification Contribution Degree,» *Journal of Computers,* τόμ. 9, 08 2014.

[14] [Ηλεκτρονικό]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html. [Πρόσβαση 1 12 2019].

[15] F. Pedregosa και et al., «Scikit-learn: Machine Learning in {P}ython,» *Journal of Machine Learning Research,* τόμ. 12, pp. 2825--2830, 2011.

[16] A. Ng, *Σημειώσεις από το μάθημα CS-229.*

[17] M. C. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

[18] . K. P. Murphy, *Machine Learning A Probabilistic Perspective,* The MIT Press, 2012.

[19] C.-W. Hsu, «A Practical Guide to Support Vector Classification,» 2003.

[20] [Ηλεκτρονικό]. Available: https://easyai.tech/en/ai-definition/gradient-descent/. [Πρόσβαση 6 12 2019].

[21] L. Bottou, «Large-Scale Machine Learning with Stochastic Gradient Descent,» *Proceedings of COMPSTAT'2010,* 2010.

[22] [Ηλεκτρονικό]. Available: https://scikit-learn.org/stable/modules/sgd.html. [Πρόσβαση 7 12 2019].

[23] A. Saffari. [Ηλεκτρονικό]. Available: https://github.com/amirsaffari/online-random-forests. [Πρόσβαση 11 12 2019].

[24] A. Saffari και et al, «Online Random Forests,» *3rd IEEE ICCV Workshop on On-line Computer Vision,* 2009.

[25] A. Saffari και e. al., «Online Multi-Class LPBoost,» *IEEE Conference on Computer Vision and Patter Recognition,* 2010.

[26] A. Hocker και et al., «TMVA - Toolkit for Multivariate Data Analysis,» *PoS, vol. ACAT,,* p. 040, 2007.

[27] [Ηλεκτρονικό]. Available: https://scikit-learn.org/stable/modules/ensemble.html#forest. [Πρόσβαση 11 12 2019].

[28] [Ηλεκτρονικό]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesClassifier.html. [Πρόσβαση 15 12 2019].

[29] A. Bordes και et al., «Fast Kernel Classifiers with Online and Active Learning,» *Journal of Machine Learning Research,* τόμ. 6, pp. 1579-1619, September 2006.

[30] L. Bottou, «https://leon.bottou.org/projects/lasvm,» 15 12 2019. [Ηλεκτρονικό]. Available: https://leon.bottou.org/projects/lasvm.

[31] J. C. Platt, «Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines,» *ADVANCES IN KERNEL METHODS - SUPPORT VECTOR LEARNING,* 1998.

[32] «https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html,» [Ηλεκτρονικό]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html. [Πρόσβαση 15 12 2019].

[33] C.-C. Chang και et al., «LIBSVM: A Library for Support Vector Machines,» *ACM Trans. Intell. Syst. Technol.,* τόμ. 2, pp. 27:1--27:27, 05 2011.

[34] X. Ye και et al., «An incremental learning vector quantization algorithm for pattern classification,» *Neural Computing and Applications,* τόμ. 21, pp. 1205 -- 1215, September 2012.

[35] A. v. Rossum, «https://github.com/mrquincle/ilvq,» 21 12 2019. [Ηλεκτρονικό]. Available: https://github.com/mrquincle/ilvq.

[36] C. B.-S. 3. h. By Antti Ajanki AnAj - Own work, «Wikipedia,» [Ηλεκτρονικό]. Available: https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm. [Πρόσβαση 21 12 2019].

[37] R. Elwell και et al., «Incremental Learning of Concept Drift in Nonstationary Environments,» *IEEE Transactions on Neural Networks,* τόμ. 22, pp. 1517-1531, 2011.

[38] J. Montiel και et al., «Scikit-Multiflow: A Multi-output Streaming Framework,» *Journal of Machine Learning Research,* τόμ. 19, pp. 1-5, 2018.

[39] «https://scikit-multiflow.github.io/,» [Ηλεκτρονικό]. Available: https://scikit-multiflow.github.io/. [Πρόσβαση 22 12 2019].

Παράρτημα Ι

Σε αυτό το παράρτημα παρατίθεται παράδειγμα όπου υπολογίζονται η ακρίβεια, ο μέσος και ο σταθμισμένος δείκτης F_1 σε imbalanced dataset. Υποθέτουμε ότι το δείγμα αποτελείται από τρεις κλάσεις (0, 1 και 2). Οι κλάσεις 0 και 2 αποτελούνται από 450 παραδείγματα η κάθε μία, ενώ η κλάση 1 από 100 παραδείγματα. Ακολουθούν δύο μήτρες σύγχυσης για δύο διαφορετικές υποθετικές ταξινομήσεις του δείγματος. Στην πρώτη ταξινόμηση (αριστερά) τα παραδείγματα των κλάσεων 0 και 2 έχουν αναγνωριστεί πολύ καλά (recall 88.8%) ενώ αυτά της κλάσης 1 έχουν αναγνωριστεί πολύ λίγο (recall 20%). Στην δεύτερη ταξινόμηση (δεξιά) τα παραδείγματα των κλάσεων 0 και 2 έχουν αναγνωριστεί ελαφρώς χειρότερα συγκριτικά με την πρώτη ταξινόμηση (recall 84.4%) ενώ αυτά της κλάσης 1 έχουν αναγνωριστεί πολύ καλύτερα (recall 60%).

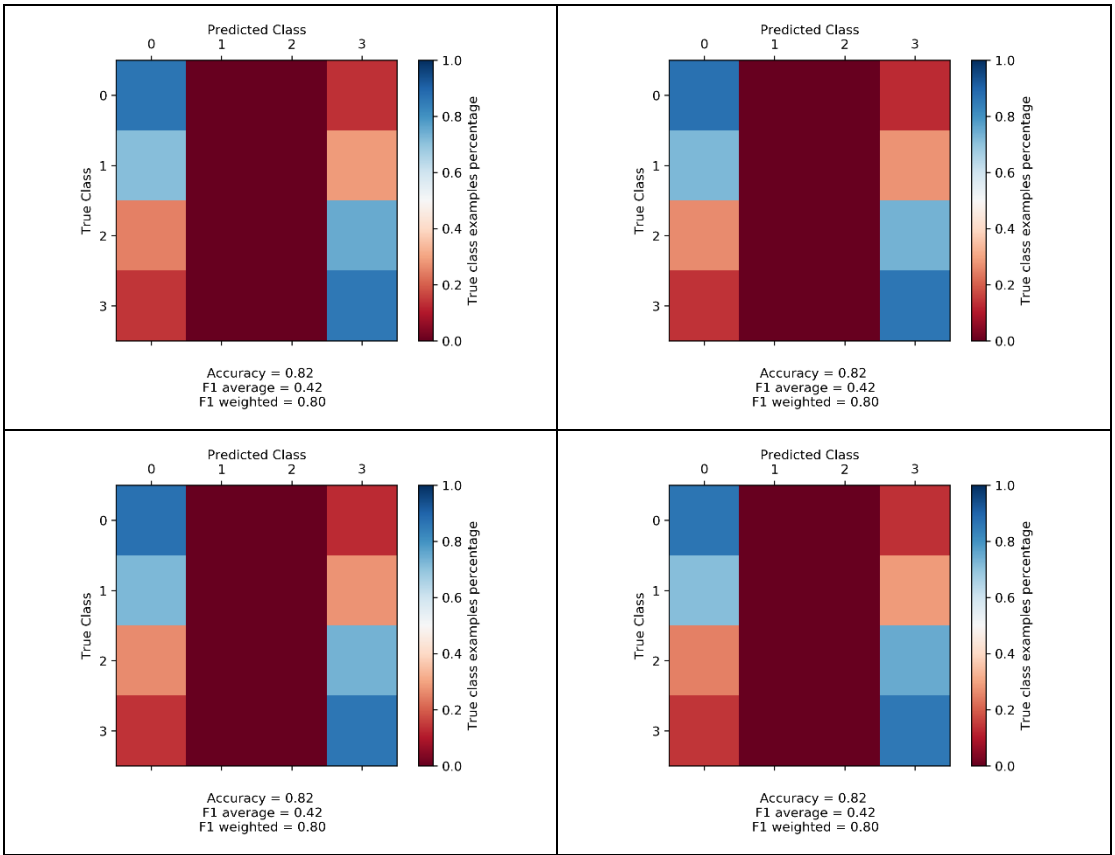
		Predicted class		
True Class		0	1	2
	0	400	30	20
	1	40	20	40
	2	20	30	400
Accuracy: 82% Weighted F_1: 80.5% Average F_1: 66%				

		Predicted class		
True Class		0	1	2
	0	380	40	90
	1	20	60	20
	2	30	40	380
Accuracy: 82% Weighted F_1: 81.8% Average F_1: 82%				

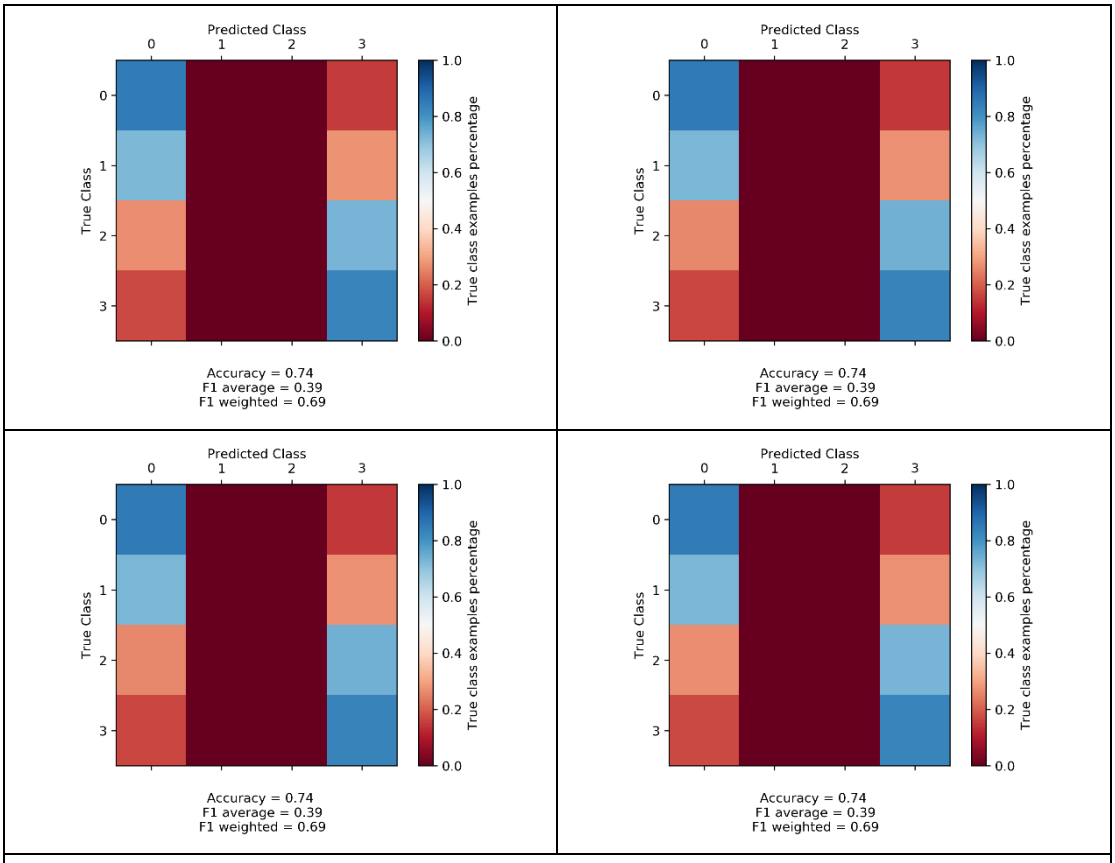
Όπως παρατηρείται από τους δείκτες κάτω από τις μήτρες σύγχυσης, και στις δυο περιπτώσεις η ακρίβεια ταξινόμησης του δείγματος είναι 82%. Συνεπώς η μεγάλη βελτίωση στην ταξινόμηση των παραδειγμάτων της κλάσης 1 εξισορροπείται από την ελαφρά επιδείνωση της ταξινόμησης των κλάσεων 0 και 2. Ο δείκτης Weighted F_1 παρουσιάζει πολύ μικρή βελτίωση από την πρώτη στη δεύτερη περίπτωση καθώς επηρεάζεται περισσότερο από τις πιο πολυπληθείς κλάσεις. Τέλος ο δείκτης Average F_1 παρουσιάζει μεγάλη βελτίωση από την πρώτη στη δεύτερη περίπτωση καθώς επηρεάζεται εξίσου από όλες τις κλάσεις.

Παράρτημα ΙΙ

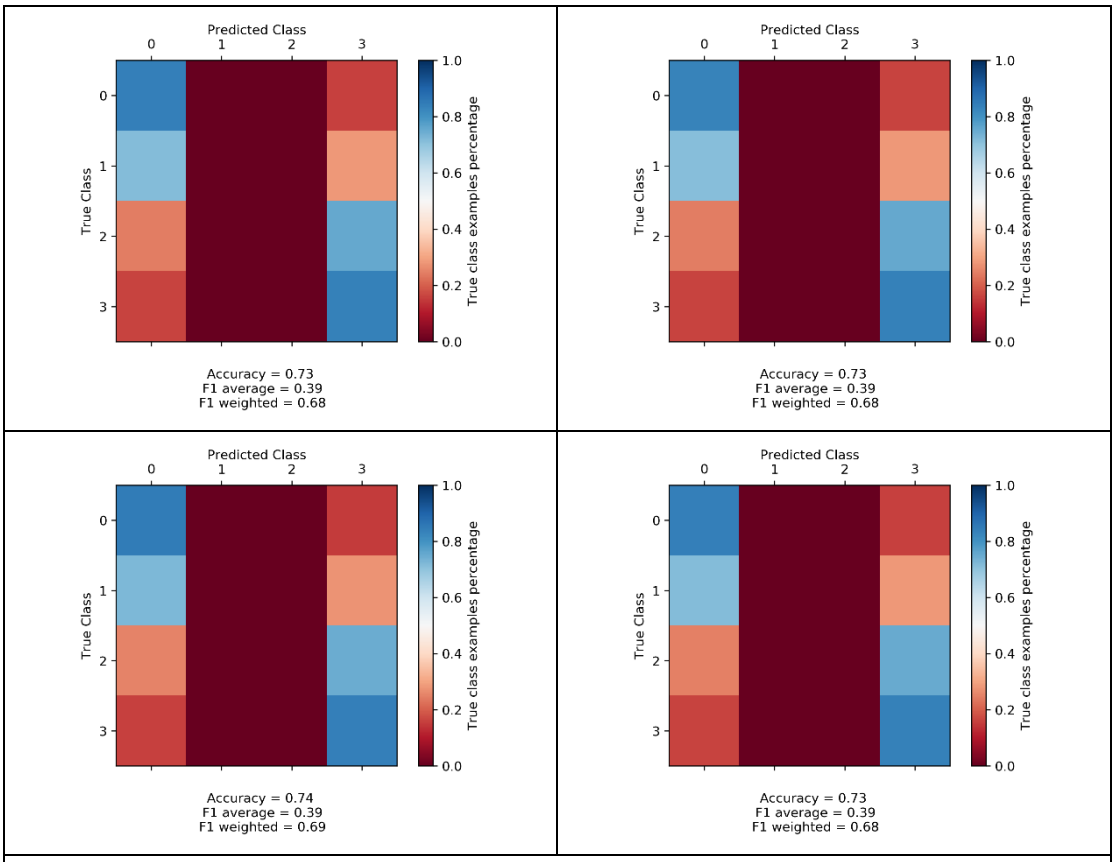
Σε αυτό το παράρτημα παρουσιάζονται γραφικές απεικονίσεις των μητρών σύγχυσης κατά την κατηγοριοποίηση του dataset d3 από τους 4 ταξινομητές όπως παρουσιάστηκε στην ενότητα 3.4. Οι μήτρες σύγχυσης που παρουσιάζονται είναι ο μέσος όρος που προέκυψε από τις 10 επαναλήψεις. Επίσης, επειδή το dataset είναι imbalanced και ο αριθμός των παραδειγμάτων που αποτελούν την κάθε κλάση παρουσιάζει πολύ μεγάλες διαφορές σε κάθε δείγμα, τα αποτελέσματα παρουσιάζονται κανονικοποιημένα ως προς τον αριθμό των παραδειγμάτων που πραγματικά ανήκουν στην κάθε κλάση. Σε κάθε τετράδα εικόνων στην θέση επάνω αριστερά παρουσιάζεται η μήτρα σύγχυσης ταξινόμησης με τον learn++NSE που εκπαιδεύεται με αυξητικό τρόπο, στη θέση επάνω δεξιά ο ταξινομητής svm ο οποίος εκπαιδευόταν κάθε φορά μόνο με το προηγούμενο δείγμα από αυτό βάση του οποίου ελεγχόταν, στην κάτω αριστερά ο ταξινομητή smv ο οποίος εκπαιδευόταν με όλα τα (μέχρι τότε) διαθέσιμα παραδείγματα και τέλος στην θέση κάτω δεξιά ο ταξινομητής svm ο οποίος εκπαιδεύτηκε μόνο με το δείγμα πρώτης εκπαίδευσης.



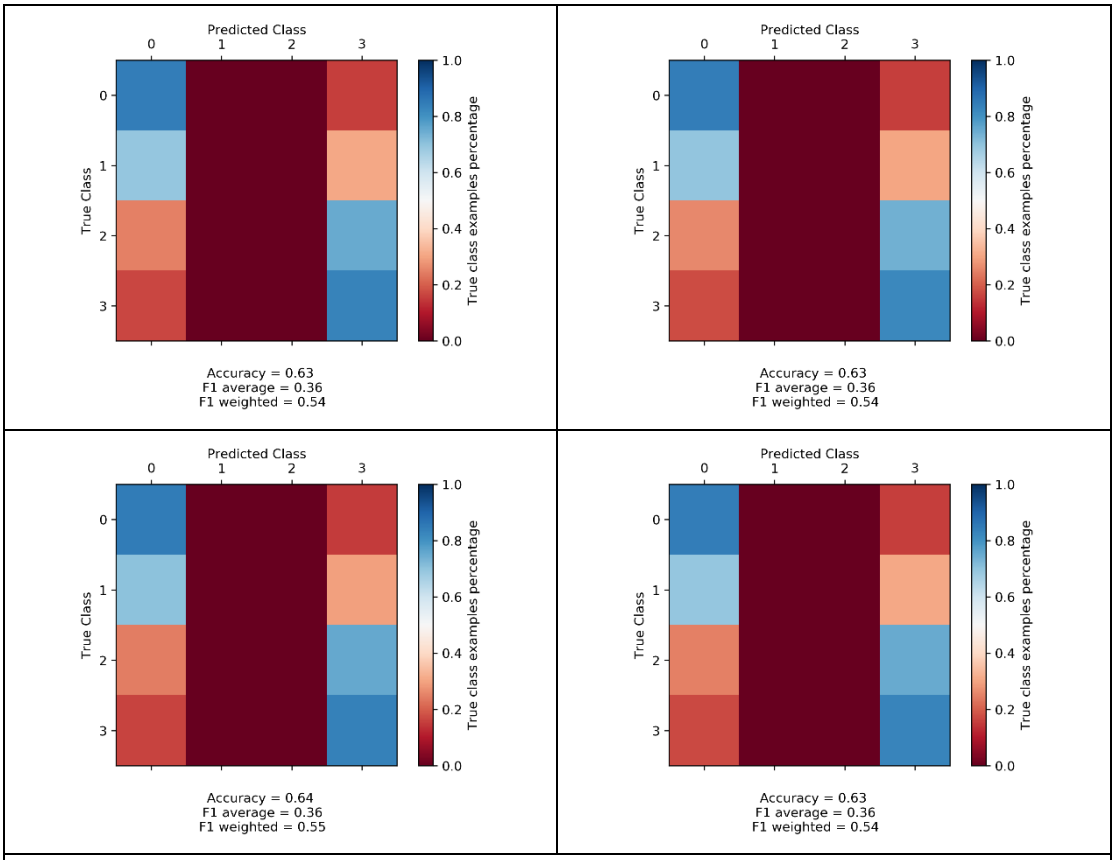
Αξιολόγηση με δείγμα πρώτης εκπαίδευσης



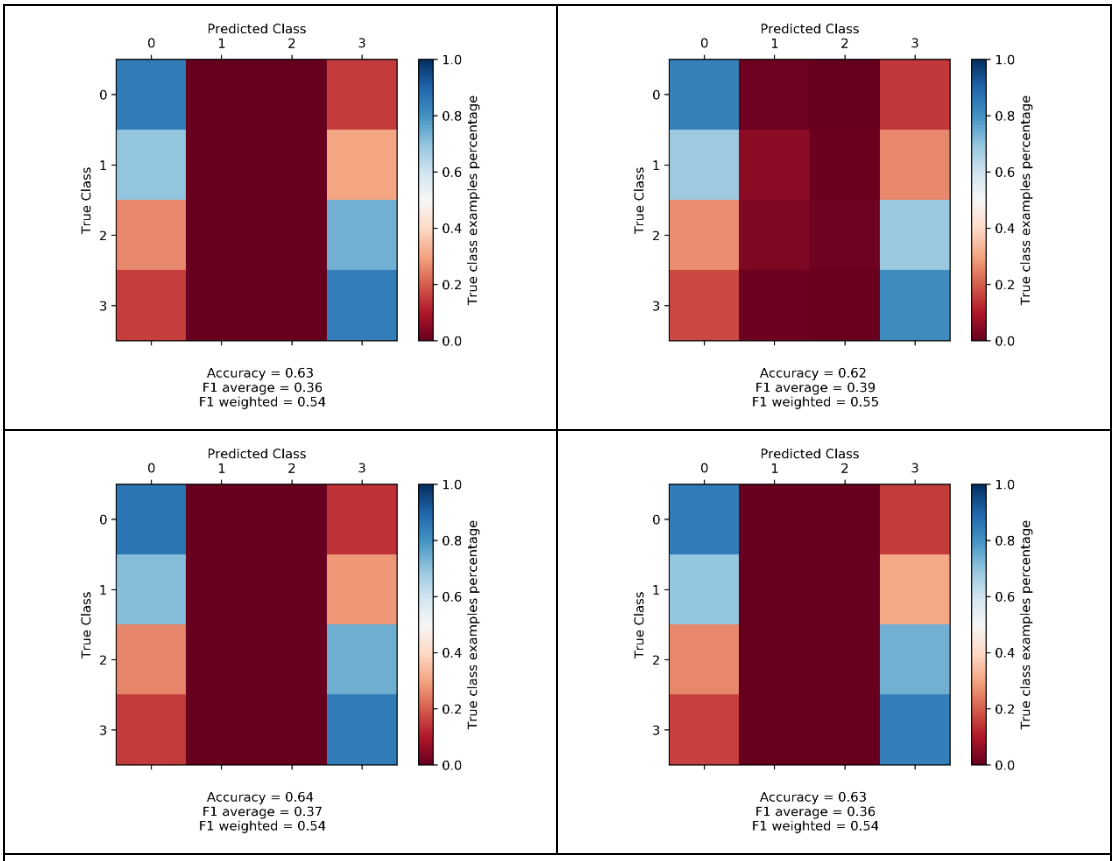
Αξιολόγηση με δείγμα 1



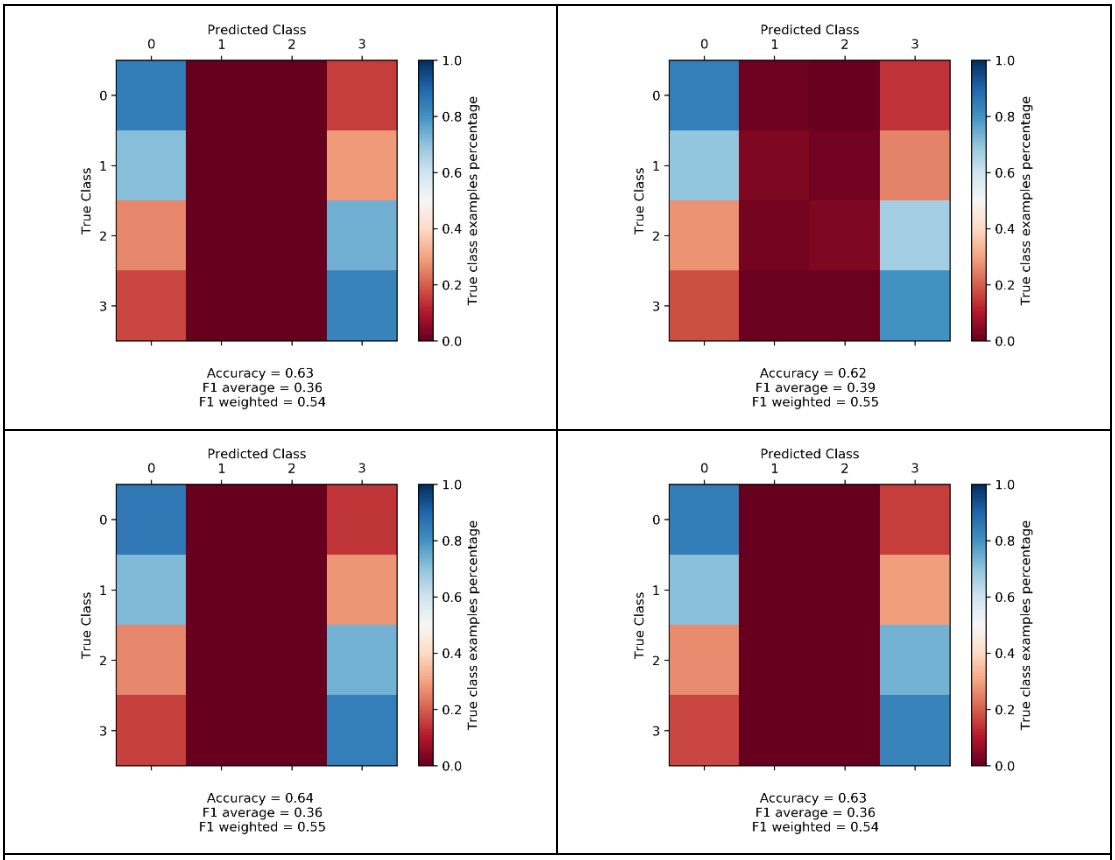
Αξιολόγηση με δείγμα 2



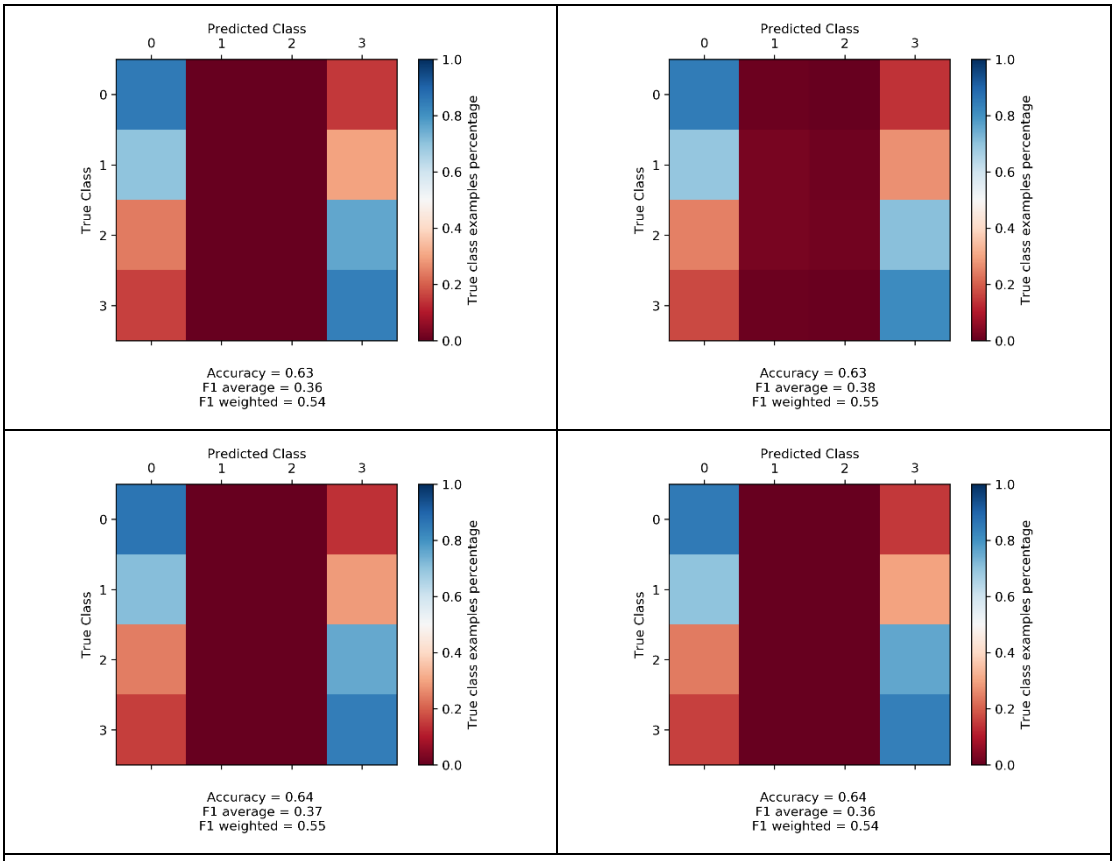
Αξιολόγηση με δείγμα 3



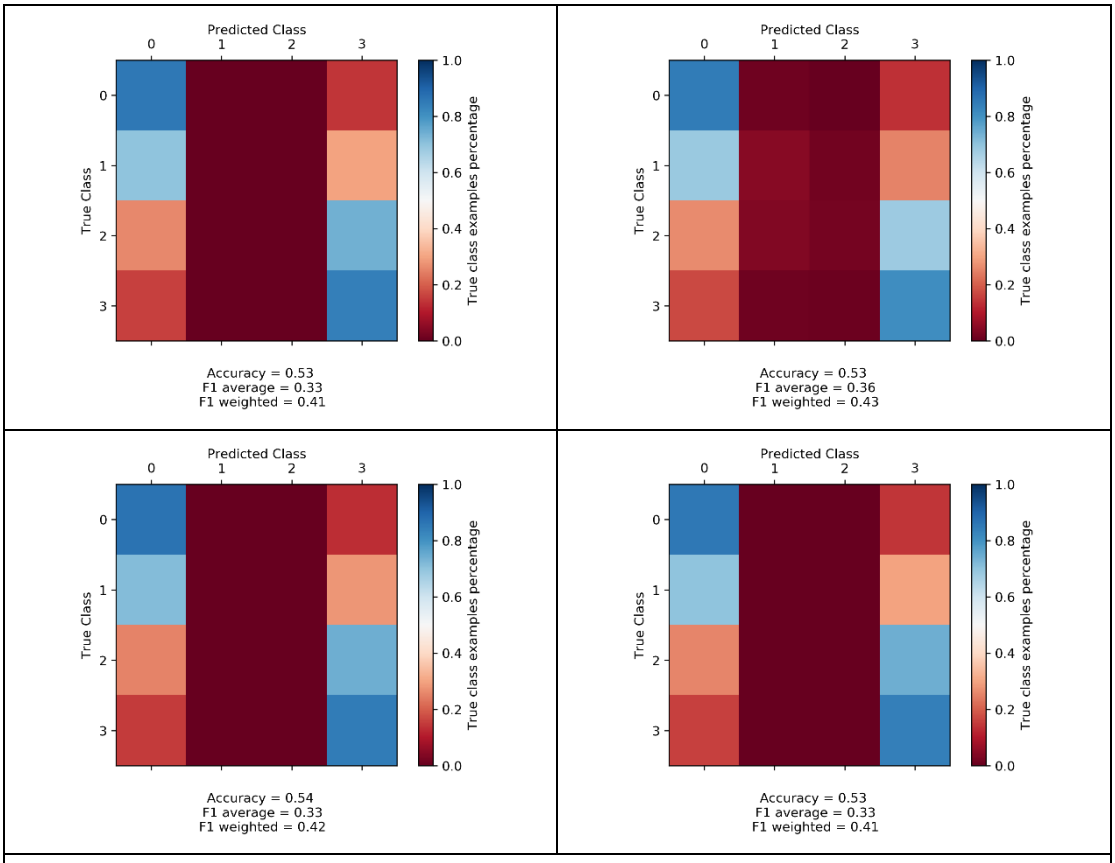
Αξιολόγηση με δείγμα 4



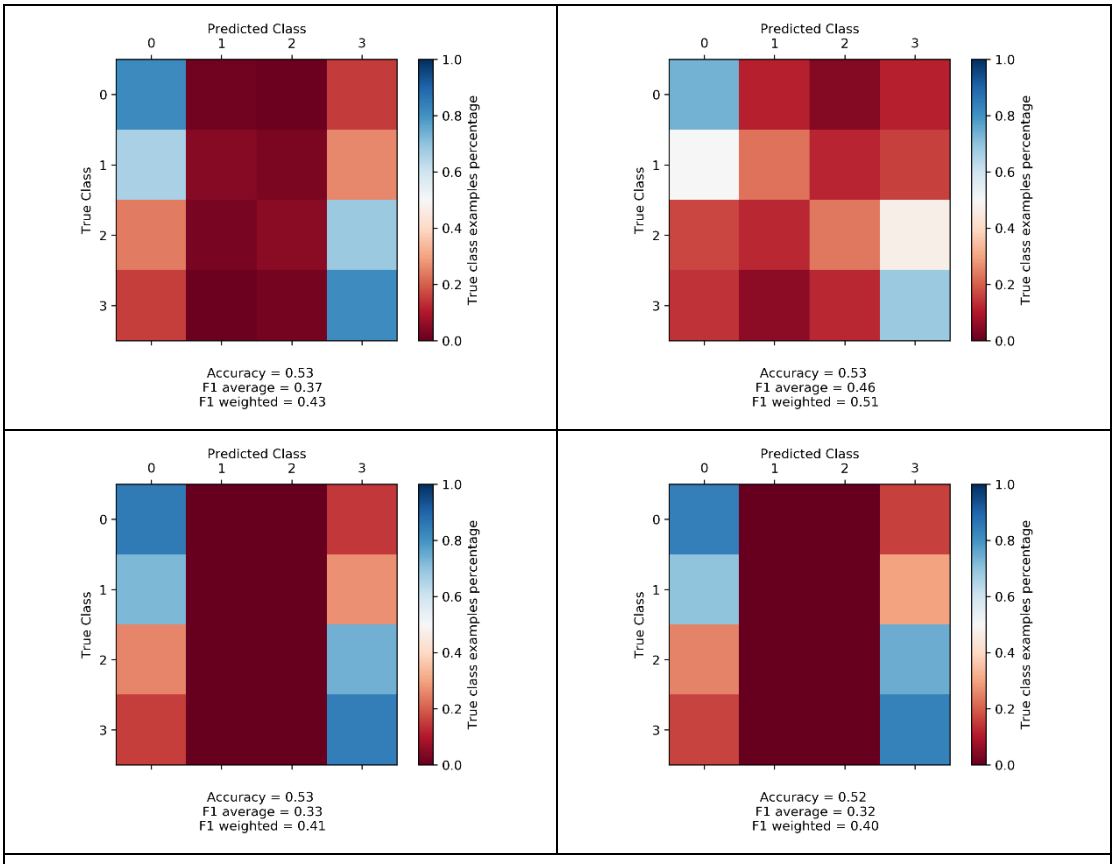
Αξιολόγηση με δείγμα 5



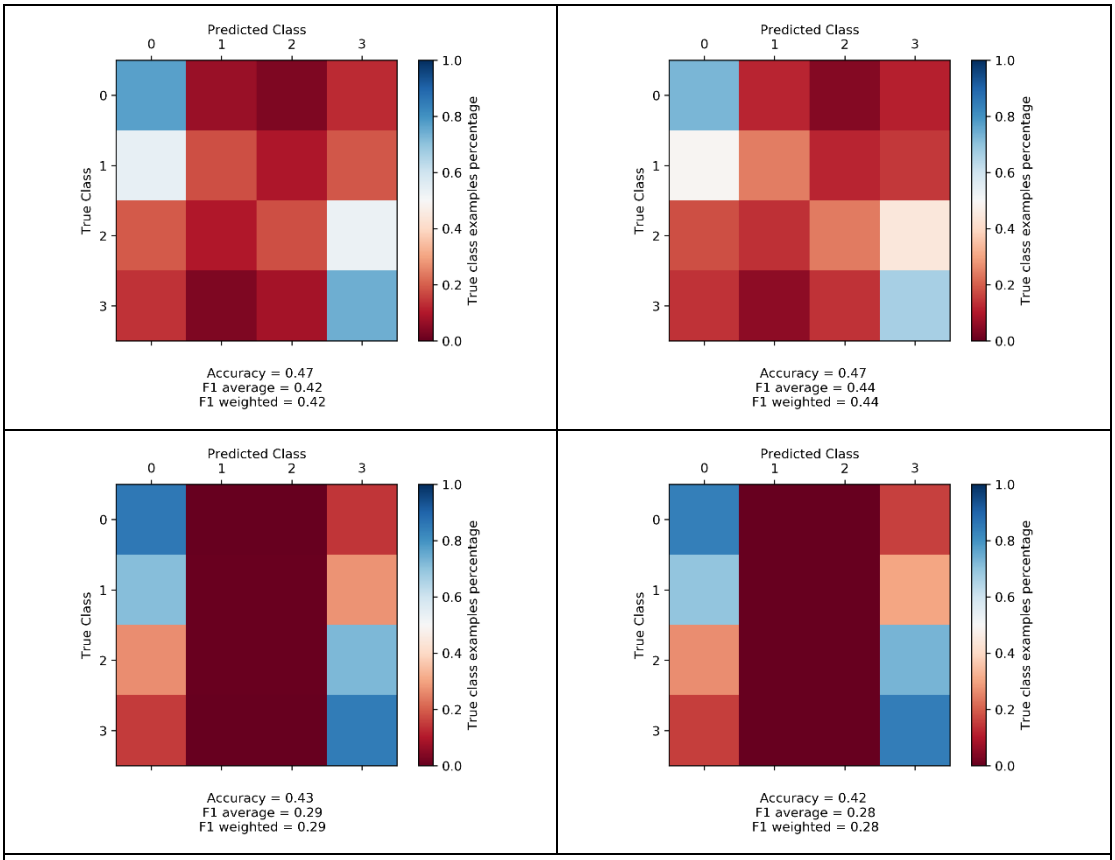
Αξιολόγηση με δείγμα 6



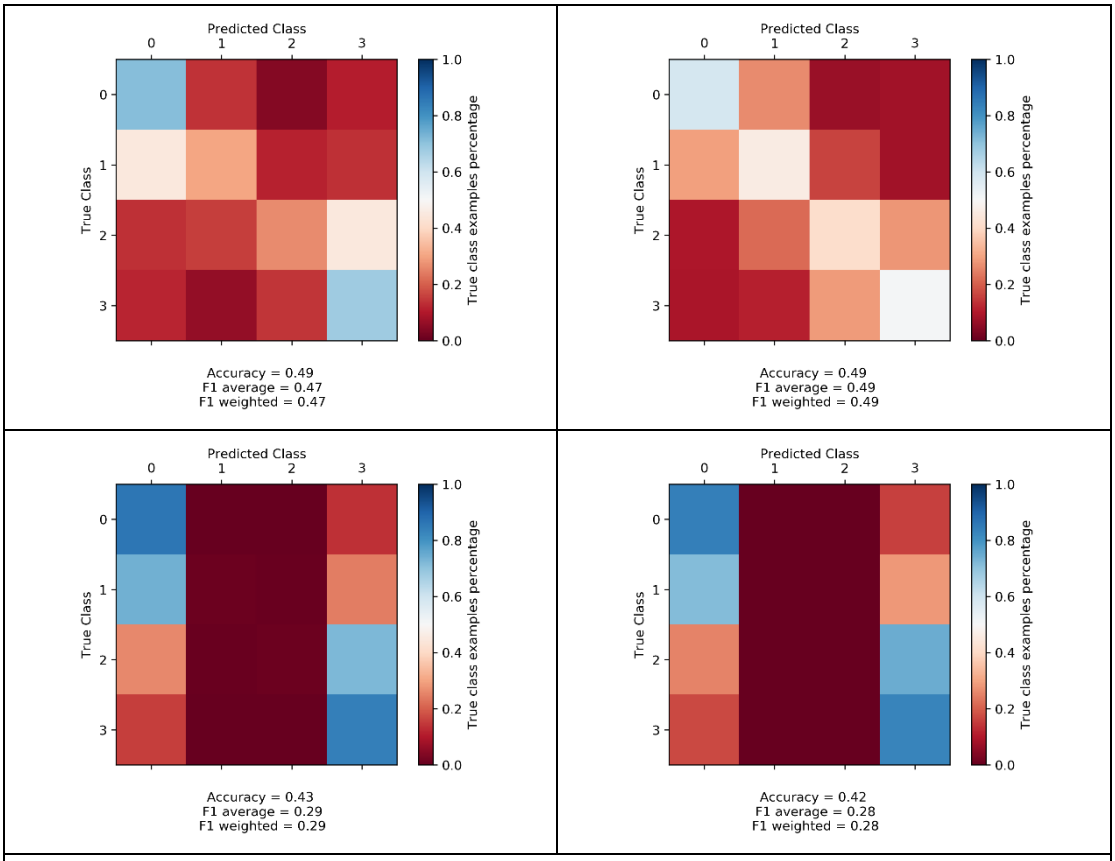
Αξιολόγηση με δείγμα 7



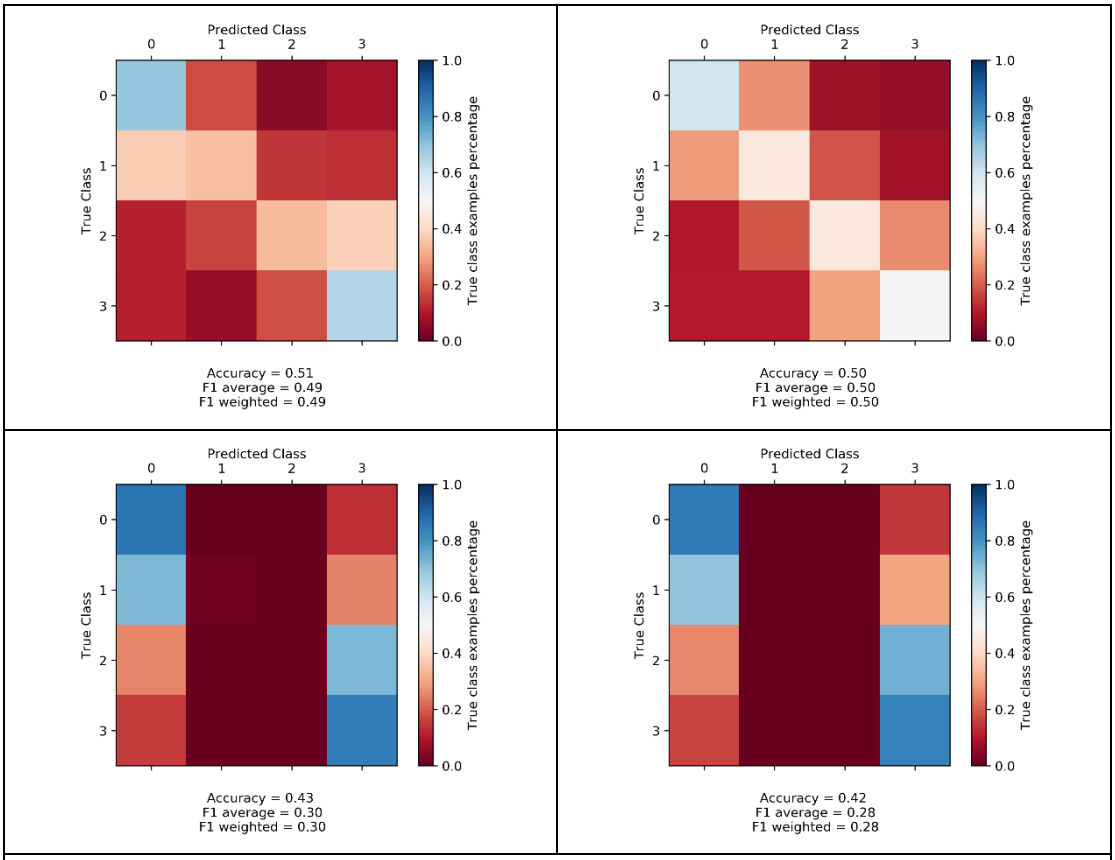
Αξιολόγηση με δείγμα 8



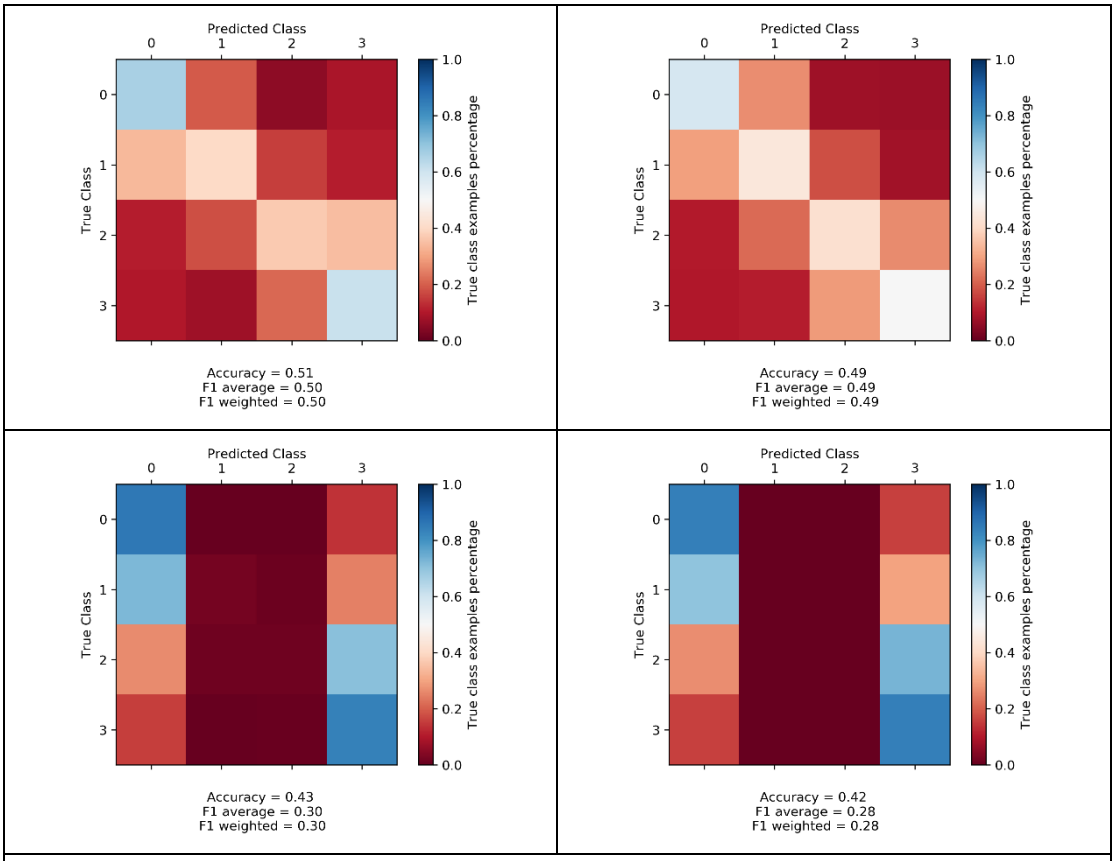
Αξιολόγηση με δείγμα 9



Αξιολόγηση με δείγμα 10



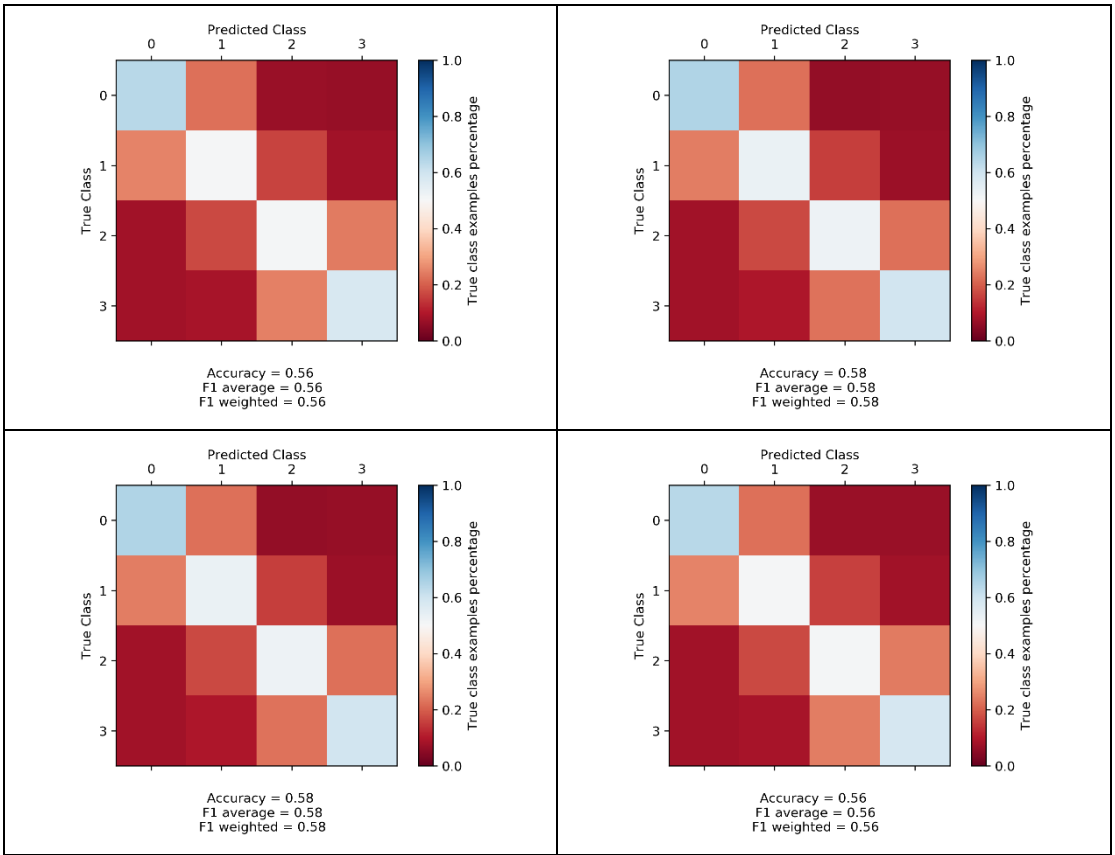
Αξιολόγηση με δείγμα 11



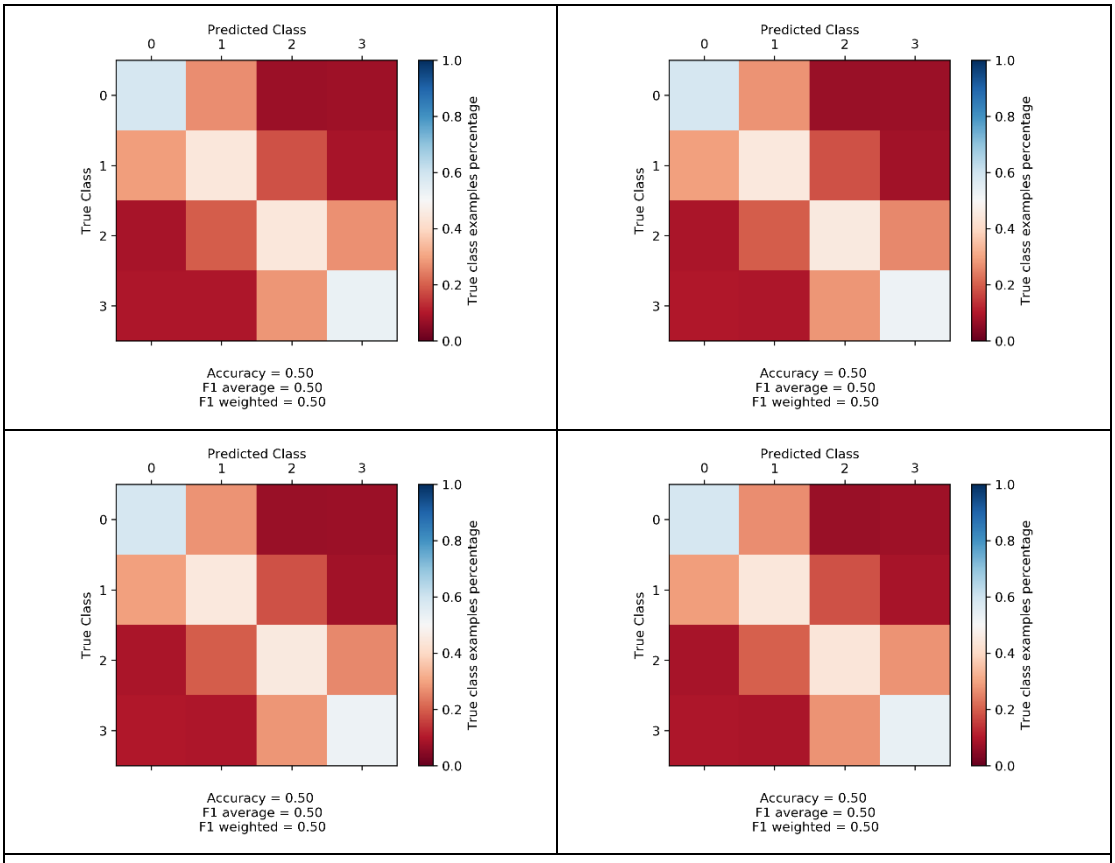
Αξιολόγηση με δείγμα 12

Παράρτημα III

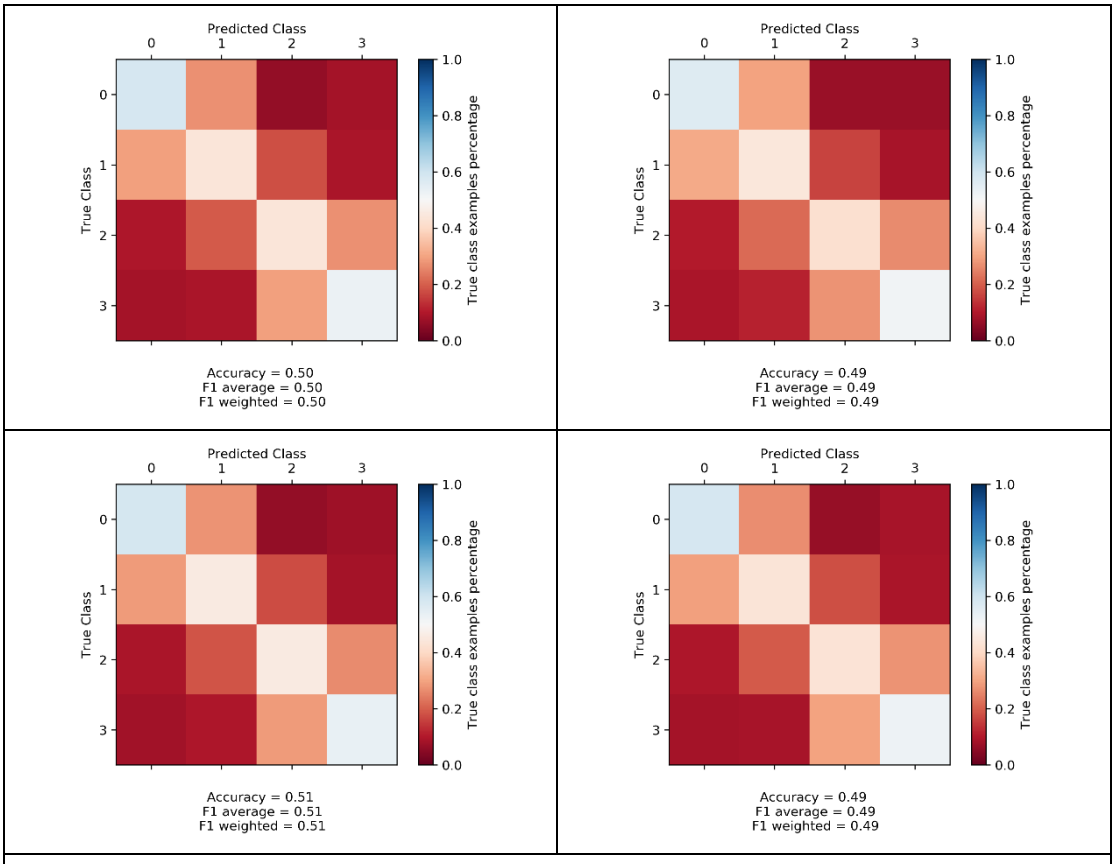
Σε αυτό το παράρτημα παρουσιάζονται γραφικές απεικονίσεις των μητρών σύγχυσης κατά την κατηγοριοποίηση του dataset d4 από τους 4 ταξινομητές όπως παρουσιάστηκε στην ενότητα 3.4. Οι μήτρες σύγχυσης που παρουσιάζονται είναι ο μέσος όρος που προέκυψε από τις 10 επαναλήψεις. Επίσης, επειδή κεφάτο dataset είναι imbalanced και ο αριθμός των παραδειγμάτων που αποτελούν την κάθε κλάση παρουσιάζει πολύ μεγάλες διαφορές σε κάθε δείγμα, τα αποτελέσματα παρουσιάζονται κανονικοποιημένα ως προς τον αριθμό των παραδειγμάτων που πραγματικά ανήκουν στην κάθε κλάση. Σε κάθε τετράδα εικόνων στην θέση επάνω αριστερά παρουσιάζεται η μήτρα σύγχυσης ταξινόμησης με τον learn++NSE που εκπαιδεύεται με αυξητικό τρόπο, στη θέση επάνω δεξιά ο ταξινομητής svm ο οποίος εκπαιδευόταν κάθε φορά μόνο με το προηγούμενο δείγμα από αυτό βάση του οποίου ελεγχόταν, στην κάτω αριστερά ο ταξινομητή smv ο οποίος εκπαιδευόταν με όλα τα (μέχρι τότε) διαθέσιμα παραδείγματα και τέλος στην θέση κάτω δεξιά ο ταξινομητής svm ο οποίος εκπαιδεύτηκε μόνο με το δείγμα πρώτης εκπαίδευσης.



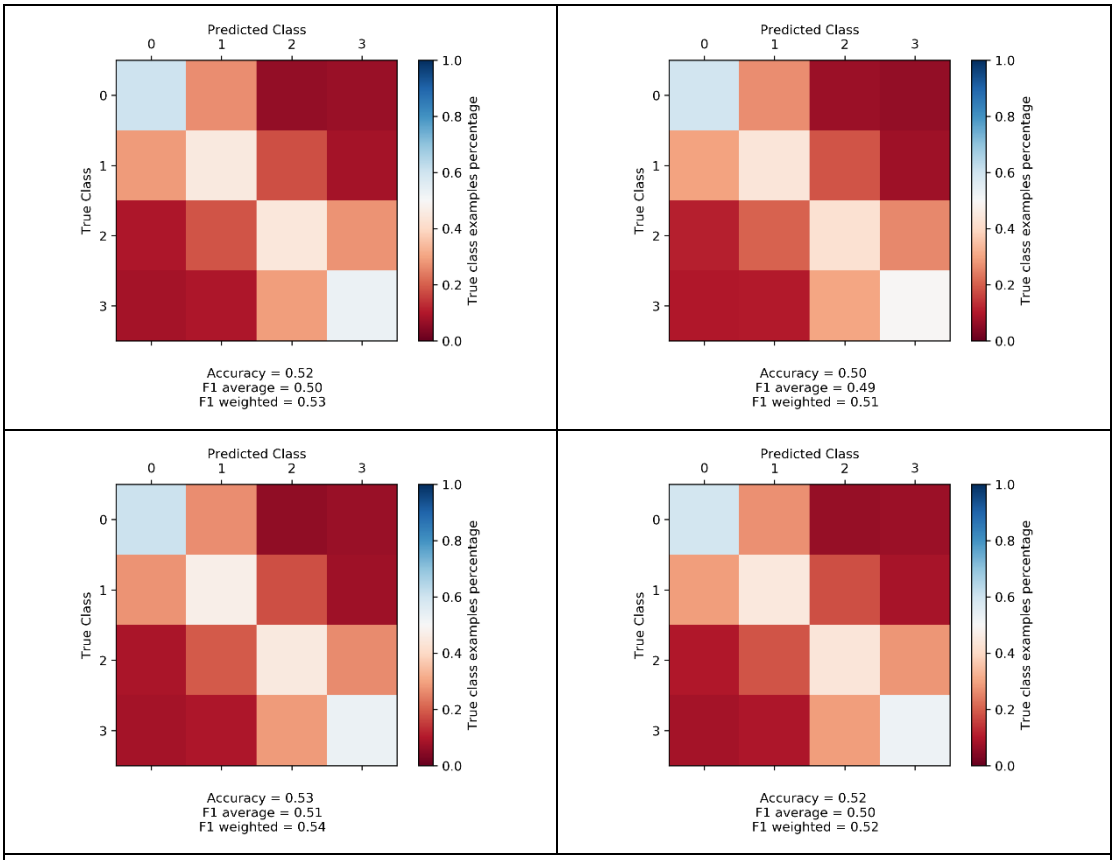
Αξιολόγηση με δείγμα πρώτης εκπαίδευσης



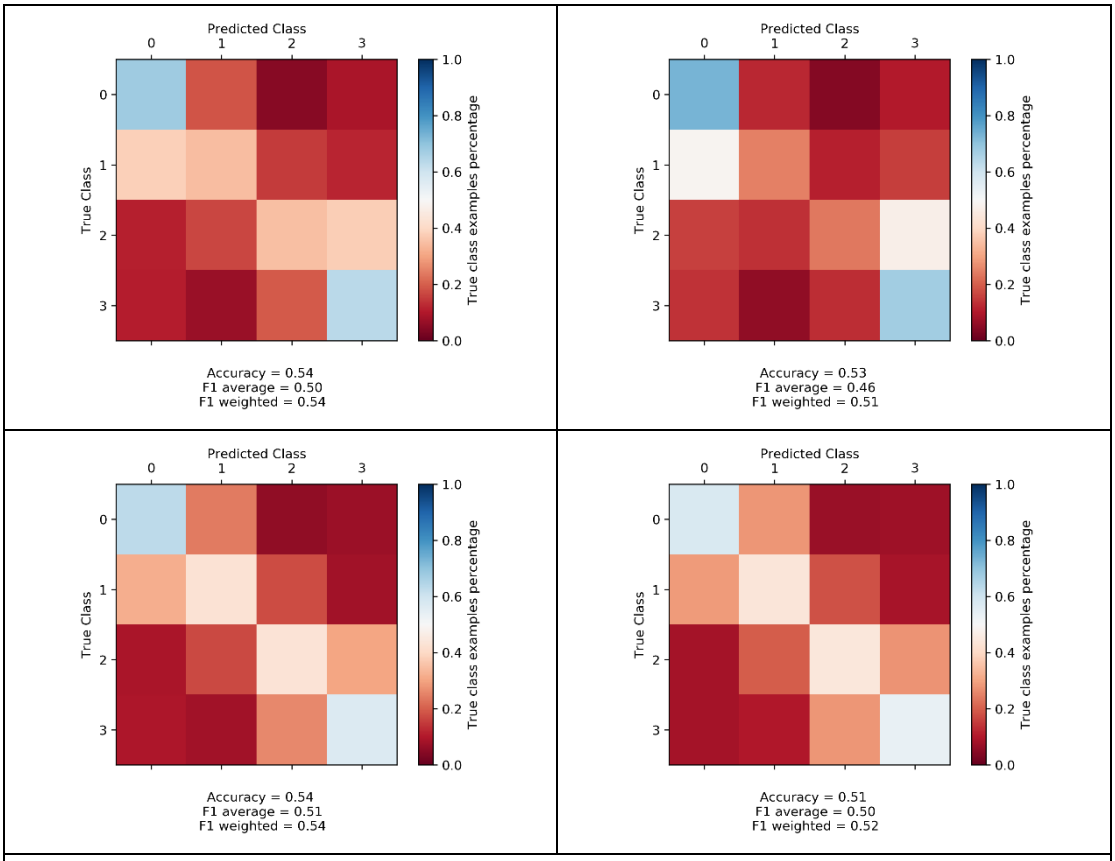
Αξιολόγηση με δείγμα 1



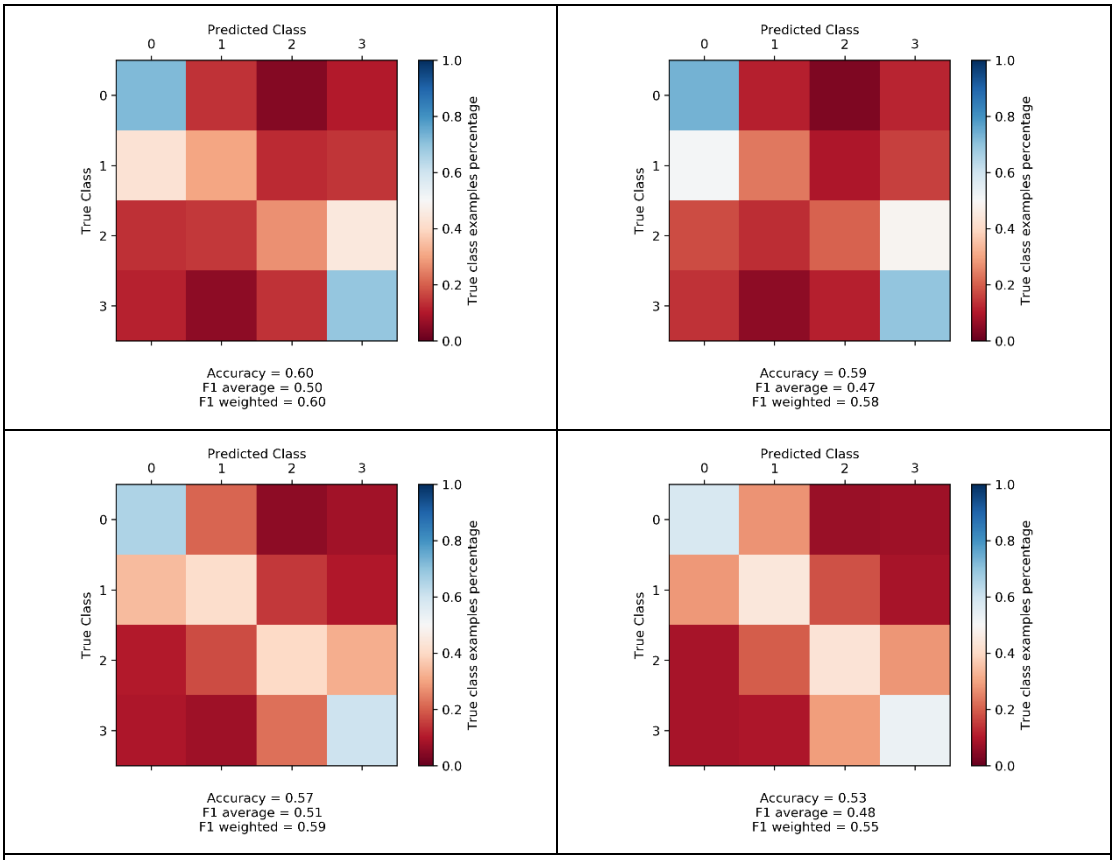
Αξιολόγηση με δείγμα 2



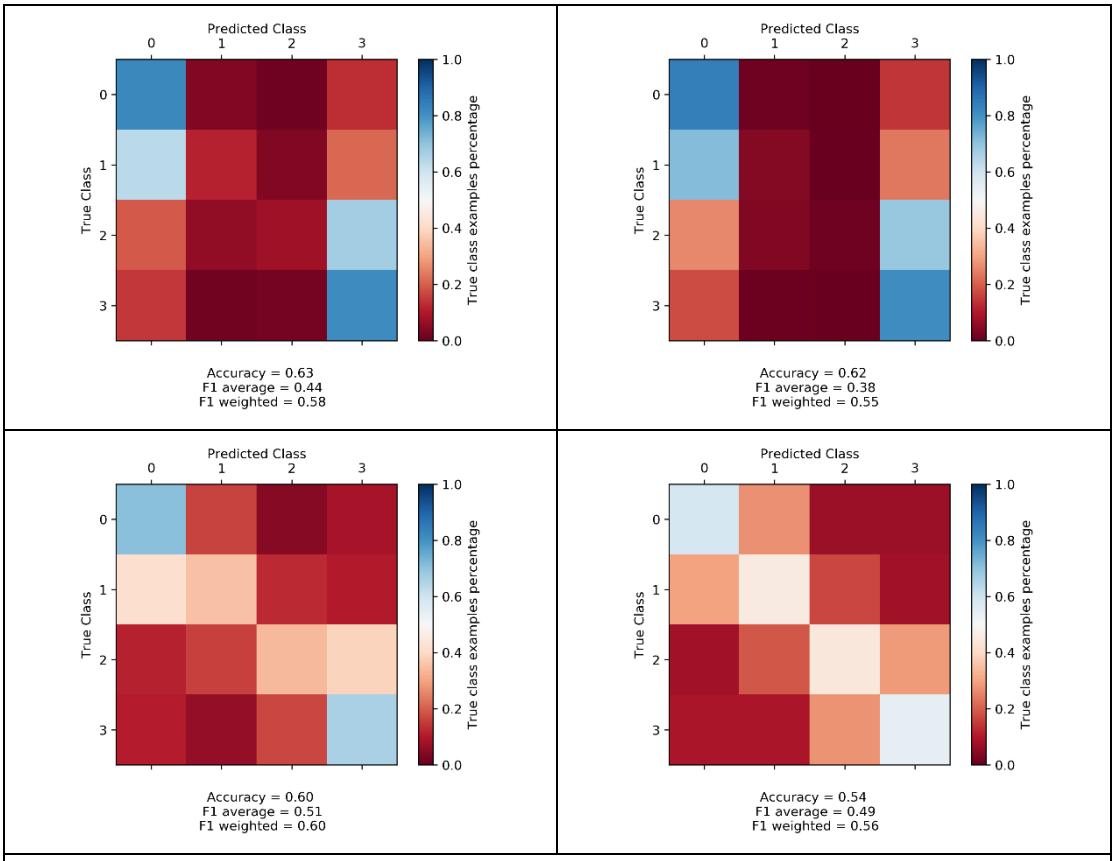
Αξιολόγηση με δείγμα 3



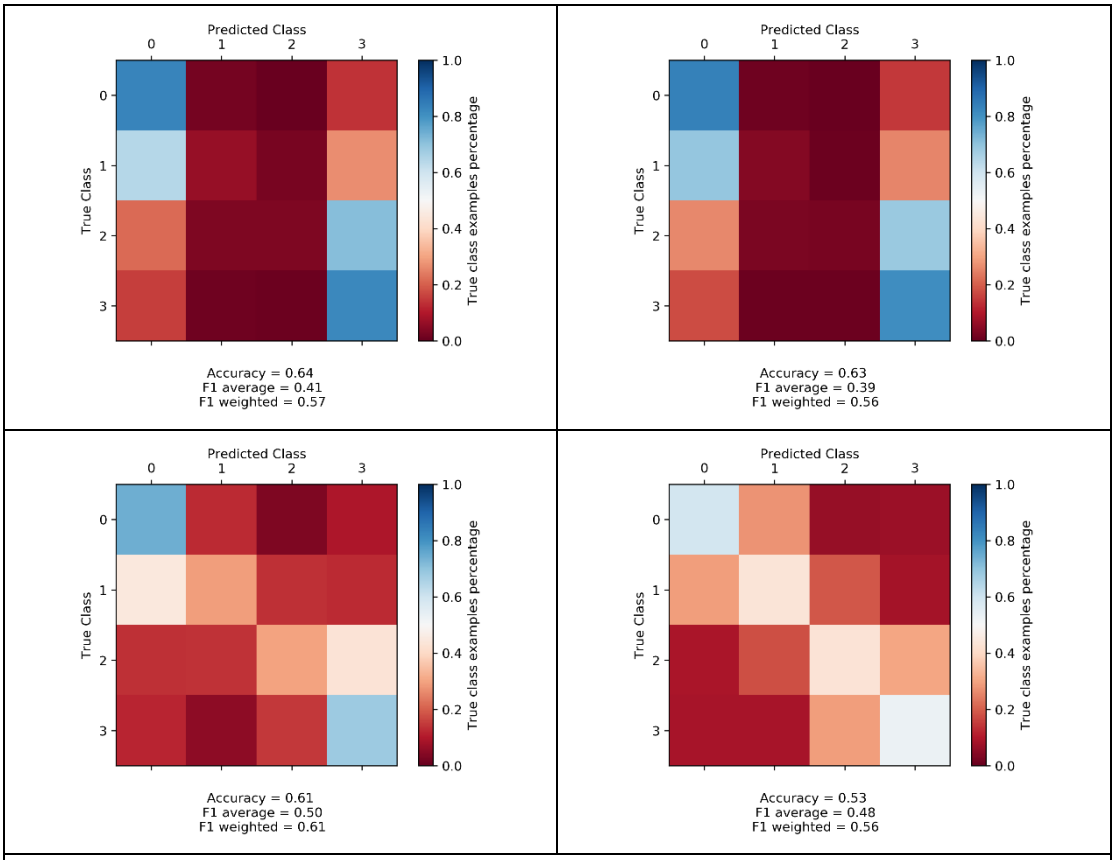
Αξιολόγηση με δείγμα 4



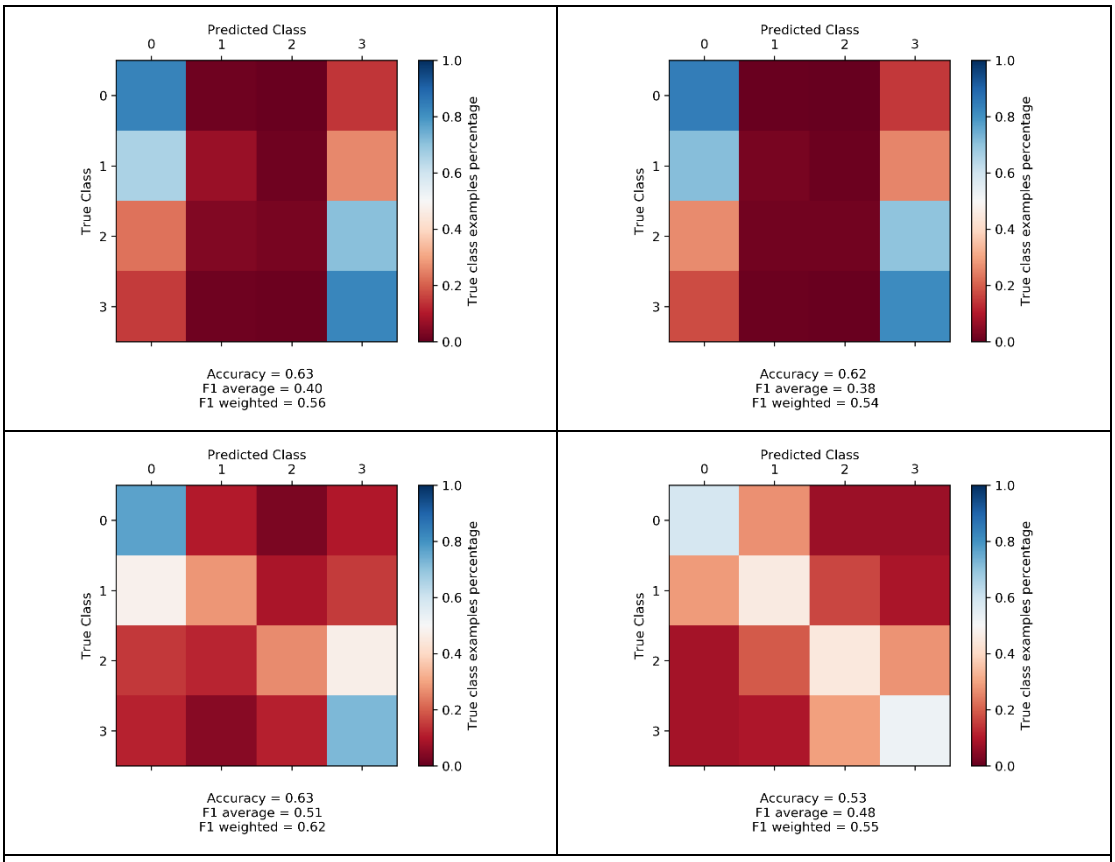
Αξιολόγηση με δείγμα 5



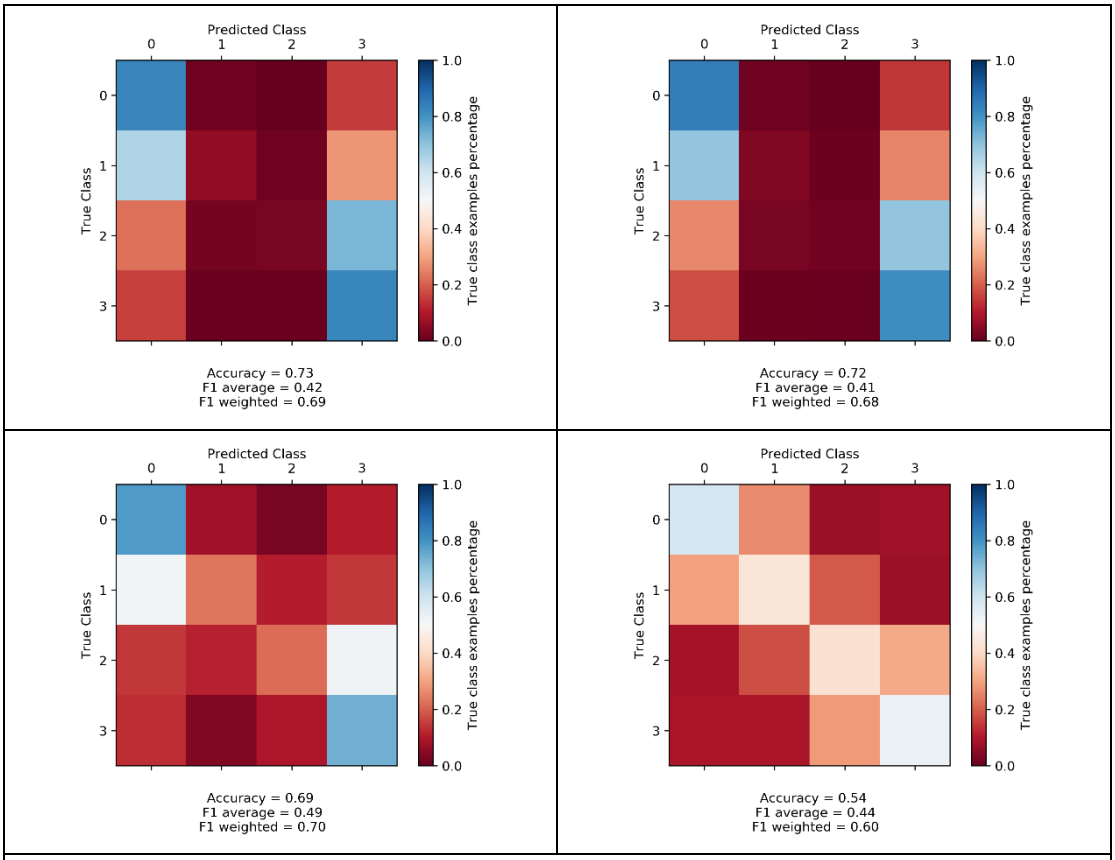
Αξιολόγηση με δείγμα 6



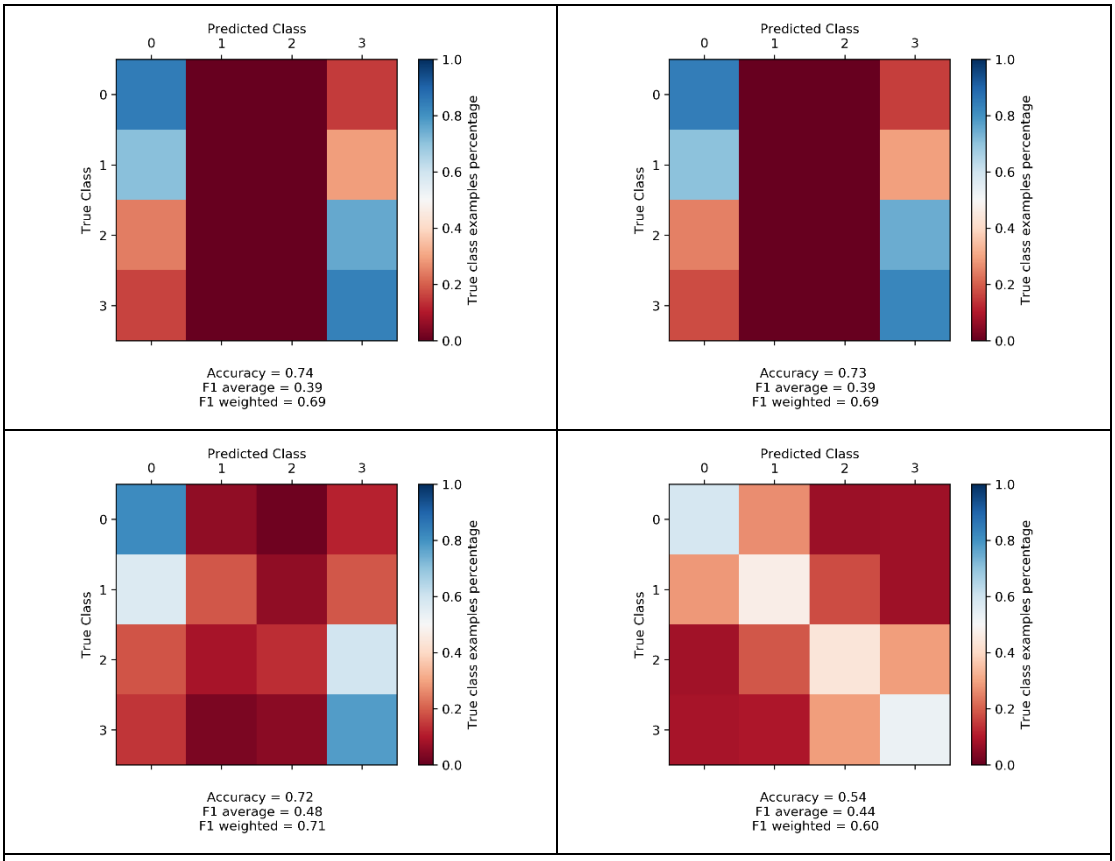
Αξιολόγηση με δείγμα 7



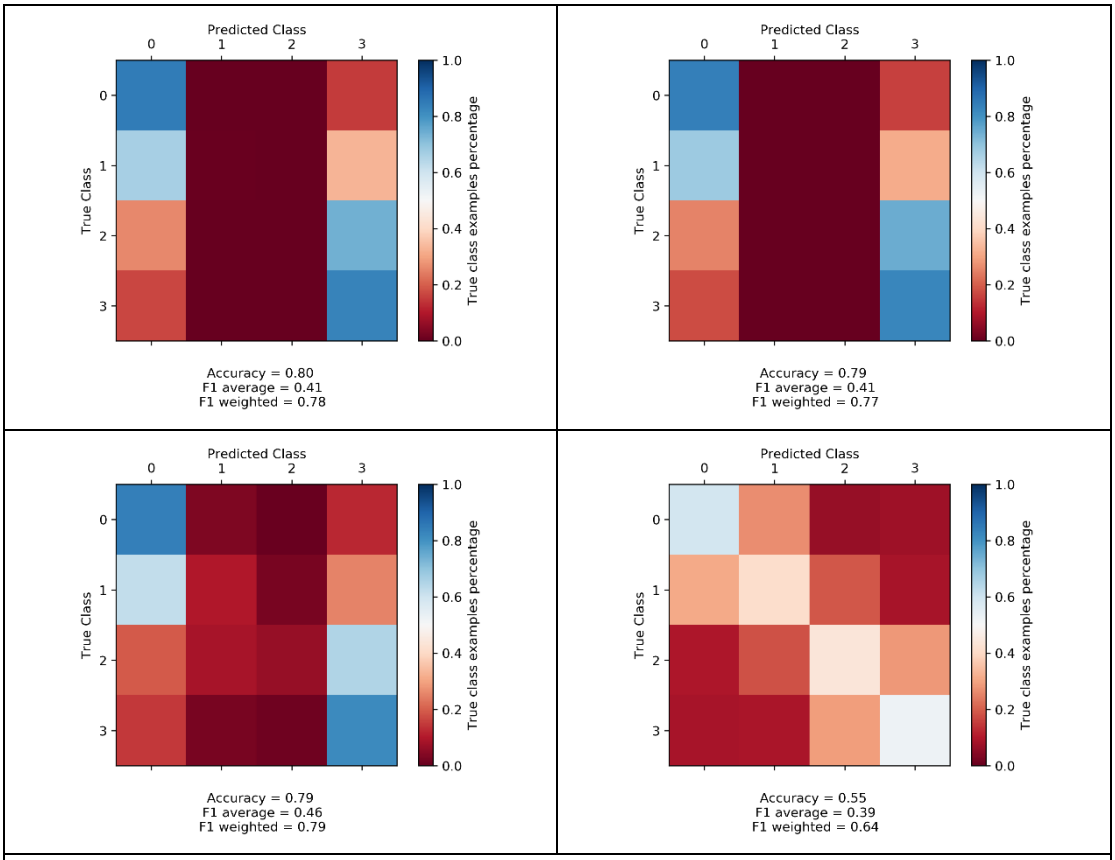
Αξιολόγηση με δείγμα 8



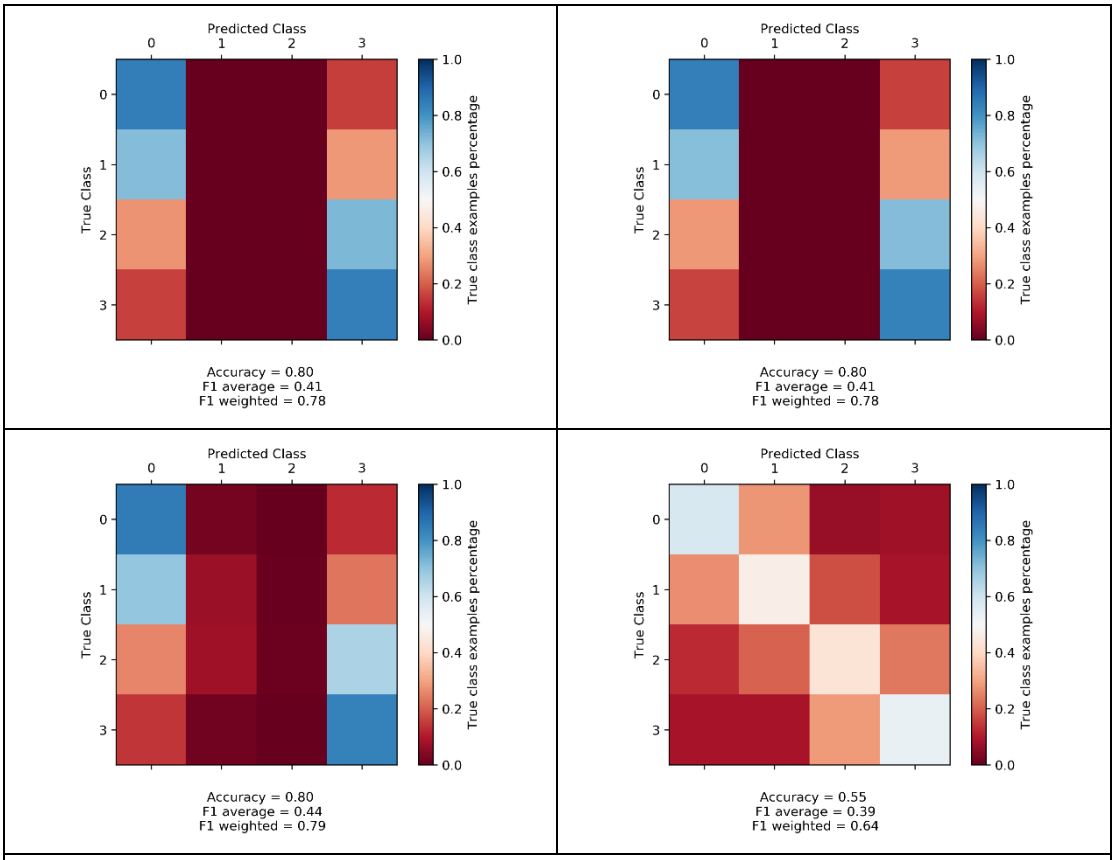
Αξιολόγηση με δείγμα 9



Αξιολόγηση με δείγμα 10



Αξιολόγηση με δείγμα 11



Αξιολόγηση με δείγμα 12