Ηχητικός Εντοπισμός Μουσικής/Ομιλίας και Κατηγοριοποίηση

Μελεζιάδης Ιωάννης, Πηλιανίδης Αριστείδης
Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Η/Υ
Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
Θεσσαλονίκη 3/1/2018
meleziadisgiannis@gmail.com, apiliani@auth.gr

ΠΕΡΙΛΗΨΗ

Το πρόβλημα της κατηγοριοποίησης ενός αρχείου ήχου σε ομιλία ή μουσική έχει απασχολήσει τη βιβλιογραφία τα τελευταία χρόνια. Αυτή η εργασία ασχολείται με το αντικείμενο της κατάτμησης των αρχείων σε παράθυρα, της ταξινόμησης των εν λόγω παραθύρων και της τελικής τους αναγνώρισης. Επίσης γίνεται απόπειρα ταξινόμησης ολόκληρων αρχείων. Αναλύονται θεωρητικά κάποιες βασικές έννοιες της **μηχανικής μάθησης**(*Machine Learning*)όπως η κανονικοποίηση δεδομένων και η **ανάλυση κυρίων συνιστωσών**(*Principal Component Analysis*) με χρήση παραδειγμάτων από το πραγματικό πρόβλημα που έγινε προσπάθεια να επιλυθεί. Το κύριο αναγνωστικό κοινό που θα επωφεληθεί από την εργασία είναι προπτυχιακοί φοιτητές που θα δουν στην πράξη μια απόπειρα ταξινόμησης πραγματικών δεδομένων.

Λέξεις κλειδιά—*speech/music classification; mirex; machine learning; principal component analysis; data normalization;*

I. ΕΙΣΑΓΩΓΗ

Θέμα της εργασίας είναι ο **Ηχητικός Εντοπισμός Μουσικής/ Ομιλίας και Κατηγοριοποίηση** *(Speech/Music Detection and Classification)* στα πλαίσια του διαγωνισμού MIREX 2018 [1]. Το αρχικό **σύνολο δεδομένων** (*dataset*), που θα χρησιμοποιηθεί για την επίλυση του προβλήματος, είναι από την συλλογή Music-Speech *GTZAN* [2] που περιέχει 128 κομμάτια μουσικής/ ομιλίας όπου το καθένα διαρκεί 30 δευτερόλεπτα και περιέχει 60 παραδείγματα από κάθε κλάση. Η συχνότητα όλων των κομματιών είναι 22,050Hz(Mono) των 16-bit σε μορφή .wav . Τα δεδομένα μετά από εξαγωγή χαρακτηριστικών βρίσκονται στο αρχείο myDataFinal.txt . Επιπλέον χρησιμοποιήθηκαν τα 3 πρώτα λεπτά από το αρχείο theconcert16 [3] για πρόσθετο έλεγχο. Στα 3 αυτά λεπτά, από την αρχή μέχρι το 1:39 υπάρχει ομιλία και από εκεί μέχρι το τέλος μουσική. Το αρχείο ονομάζεται mySong.wav και μετά από εξαγωγή χαρακτηριστικών τα δεδομένα είναι στο test3.txt. Τέλος για την αντιμετώπιση του προβλήματος χρησιμοποιείται το περιβάλλον του Matlab2015b καθώς και το Rstudio.

II. ΕΞΑΓΩΓΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ

Η επιλογή χαρακτηριστικών είναι από τα πιο σημαντικά βήματα για έναν πετυχημένο ταξινομητή. Τα χαρακτηριστικά πρέπει να περιγράφουν όσο το δυνατόν καλύτερα τις οντότητες προς ταξινόμηση αλλά και να είναι όσο το δυνατόν λιγότερα σε αριθμό. Το τελευταίο ζητείται εξαιτίας της **'κατάρας της διαστατικότητας'** (*Curse of Dimensionality*) η οποία ορίζει ότι αντί για ένα μεγάλο διάνυσμα χαρακτηριστικών προτιμάται ένα μικρό που έχει τα βασικότερα χαρακτηριστικά. Αυτό συμβαίνει διότι τα σύνολα δεδομένων που έχουν πολλές διαστάσεις συνήθως είναι αραιά με αποτέλεσμα νέες εγγραφές να μην μπορούν να κατηγοριοποιηθούν σωστά. Με τη μείωση των διαστάσεων έχει παρατηρηθεί καλύτερη ταξινόμηση καθώς και μικρότερη υπολογιστική πολυπλοκότητα [4]. Με βάση τα παραπάνω επιλέγονται τα εξής χαρακτηριστικά που κρίνονται και ως βασικότερα σύμφωνα με αλγορίθμους βαθμολόγησης της σημαντικότητας των χαρακτηριστικών [5] : RMS Energy, Entropy, Roll-off Frequency, Zerocross, Spectral Brightness, Spectral Centroid, Spectral Spread, Spectral Skewness, Spectral Kurtosis, Spectral Flatness, MFCC1–MFCC13. Μοναδική αλλαγή είναι η αφαίρεση του Spectral Irregularity και η πρόσθεση του Zerocross στη θέση του.

Για την εξαγωγή των χαρακτηριστικών χρησιμοποιήθηκε το περιβάλλον της Matlab2015b σε συνδυασμό με τη βιβλιοθήκη MIRtoolbox που παρέχει ρουτίνες για την δημιουργία παραθύρων καθώς και για την εξαγωγή χαρακτηριστικών ήχου από αρχεία. Συγκεκριμένα, κάθε αρχείο διασπάστηκε σε παράθυρα ενός δευτερολέπτου με 50% επικάλυψη. Επομένως κάθε αρχείο ήχου 30 δευτερολέπτων καταλήγει να έχει 60 παράθυρα. Στη συνέχεια δημιουργήθηκε ένας πίνακας μεγέθους 7680x24 (7680 / 60 = 128 αρχεία) που περιέχει ανά 60άδες αρχεία ομιλίας και μουσικής εναλλάξ. Ως **τάξη ή κλάση**(*Class*) ορίζεται η τελευταία στήλη του πίνακα με όνομα Class και τιμές μηδέν ή ένα όπου το μηδέν αντιστοιχεί σε ομιλία και το ένα σε μουσική. Τέλος δημιουργήθηκε ένα αρχείο myDataFinal.txt το οποίο θα διαβαστεί από το Rstudio στο οποίο και υλοποιείται η ταξινόμηση. Το αρχείο περιέχει εναλλάξ 60άδες μουσικής-ομιλίας ξεκινώντας από μουσική.

III. ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ

Εφόσον έχει γίνει επιλογή των χαρακτηριστικών απομένει η προεπεξεργασία τους. Συγκεκριμένα, αφου προστεθούν όλα σε ένα διάνυσμα χαρακτηριστικών θα ακολουθήσει η αναζήτηση ελλιπών τιμών. Παρατηρήθηκαν λίγες τέτοιες τιμές(όλες ήταν στη στήλη Centroid) όποτε και αντικαταστάθηκαν από τη μέση τιμή της στήλης ώστε να μην επηρεάσουν τους αλγορίθμους ταξινόμησης. Στη συνέχεια ακολούθησε min-max κανονικοποίηση η οποία αναλύεται παρακάτω.

A. Κανονικοποίηση

Η κανονικοποίηση των δεδομένων είναι ένα πολύ σημαντικό βήμα στην επιστήμη της μηχανικής μάθησης καθώς σε πολλούς αλγορίθμους διευκολύνει την ταξινόμηση. Αυτό επιτυγχάνεται καθώς μετασχηματίζει τα δεδομένα σε κοινή κλίμακα οπότε όλα έχουν την ίδια βαρύτητα. Ένα σημαντικό ζήτημα που εμφανίζεται συχνά στη πράξη και οδηγεί σε εσφαλμένα αισιόδοξα αποτελέσματα ταξινόμησης είναι ο τρόπος με τον οποίο θα γίνει η κανονικοποίηση. Συγκεκριμένα είναι λανθασμένη λογική να γίνεται κανονικοποίηση στο πλήρες σύνολο των δεδομένων, προτού γίνει διαχωρισμός σε

σύνολο εκπαίδευσης και ελέγχου, καθώς οδηγεί σε διαρροή πληροφορίας. Κάτι τέτοιο δικαιολογείται εύκολα καθώς στην περίπτωση της Min-max κανονικοποίησης :

$$Normalized_{Min-max}(x) = \frac{x - min(x)}{max(x) - min(x)}$$

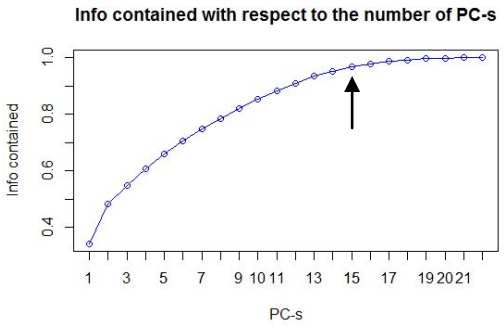
οι συναρτήσεις min και max περιέχουν πληροφορία για όλο το σύνολο δεδομένων άρα και για το σύνολο ελέγχου που θα προκύψει αργότερα από αυτό. Είναι αναγκαίο συνεπώς, πρώτα να γίνεται διαχωρισμός σε σύνολα εκπαίδευσης και ελέγχου και μετά να γίνεται ξεχωριστά στο κάθε ένα η κανονικοποίηση. Ενώ στην πράξη η ανάλυση σταματάει πολλές φορές εδώ, υπάρχει ακόμα ένα ζήτημα διαρροής δεδομένων. Η κανονικοποίηση του συνόλου ελέγχου δεν μπορεί να γίνει με βάση όλα τα δεδομένα του αφού αν συμβεί κάτι τέτοιο διαρρέει πληροφορία από τις μελλοντικές περιπτώσεις που θέλουμε να ταξινομήσουμε. Επομένως σαν min και max χρησιμοποιούνται οι τιμές από το σύνολο εκπαίδευσης και όχι αυτές από το σύνολο δεδομένων για να αποφευχθεί η **μεροληψία**(*bias*) που εισάγεται από την γνώση ολόκληρου του συνόλου ελέγχου.

Για τη κανονικοποίηση όπως ορίστηκε παραπάνω χρησιμοποιήθηκε, όπου γινόταν, μια ειδική συνάρτηση στην R, με όνομα myNormalization( ) και ορίσματα τα δεδομένα εκπαίδευσης και ελέγχου, η οποία καλείται αφού γίνει διάσπαση του συνόλου των δεδομένων. Επιπλέον χρησιμοποιήθηκαν δύο είδη κανονικοποίησης ανάλογα με τον αλγόριθμο, min-max και standardization.

B. Ανάλυση Κυρίων Συνιστωσών

Η ανάλυση κυρίων συνιστωσών είναι από τις πιο διάσημες τεχνικές για τη μείωση των διαστάσεων σε ένα σύνολο δεδομένων που όπως προαναφέρθηκε είναι μείζων σημασίας. Λειτουργεί κατασκευάζοντας νέα χαρακτηριστικά, ασυσχέτιστα μεταξύ τους, τα οποία είναι γραμμικοί συνδυασμοί των αρχικών χαρακτηριστικών. Αυτά τα νέα χαρακτηριστικά, ή αλλιώς κύριες συνιστώσες προσπαθούν, να διατηρήσουν όσο το δυνατόν περισσότερη διακύμανση από τα αρχικά χαρακτηριστικά. Οι **κύριες συνιστώσες**(*PC-s*) είναι ίδιες στον αριθμό με τα αρχικά χαρακτηριστικά με τη διαφορά όμως ότι οι πρώτες κύριες συνιστώσες φέρουν μεγάλο ποσοστό της συνολικής πληροφορίας του συνόλου δεδομένου, ενώ οι τελευταίες μπορούν να μη χρησιμοποιηθούν χωρίς να χάνεται σημαντική πληροφορία. Έτσι για την μείωση των διαστάσεων αρκεί να επιλεχθούν οι πρώτες n κύριες συνιστώσες ,που φτάνουν σε διατήρηση του 90-95% της αρχικής διακύμανσης, και να ακολουθήσει προβολή του αρχικού συνόλου δεδομένων σε αυτές. Οι κύριοι λόγοι για την εφαρμογή της ανάλυσης κυρίων συνιστωσών είναι η επιτάχυνση των αλγορίθμων ταξινόμησης(λόγω λιγότερων δεδομένων), η πιθανή αφαίρεση θορύβου από το σύνολο των δεδομένων(αντιστοιχεί στην πληροφορία που χάθηκε) και η συμπίεση [6]. Στην εργασία μας, έγινε ανάλυση για του πρώτους δυο λόγους.

Μια βασική προϋπόθεση για να λειτουργήσει σωστά ο αλγόριθμος της ανάλυσης σε κύριες συνιστώσες είναι τα δεδομένα να έχουν ως κέντρο τους την αρχή των αξόνων [6] κάτι το οποίο προσέχεται στον κώδικά μας. Παρακάτω παρουσιάζεται το διάγραμμα διακύμανσης/διαστάσεων :



1. Διάγραμμα Πληροφορίας/Διαστάσεων

Το βέλος δείχνει τον αριθμό των PC-s που διατηρούν συνολική πληροφορία μεγαλύτερη του 95% της αρχικής. Με βάση το διάγραμμα, κρίθηκε μη σκόπιμη η χρήση του μετασχηματισμού PCA καθώς δεν μειώνονται αρκετά οι διαστάσεις(από 23 σε 15) κάτι που καταδεικνύει τη σωστή αρχική επιλογή χαρακτηριστικών. Κάτι τέτοιο είναι λογικό καθώς η επιλογή τους έγινε με βάση τα αποτελέσματα αλγορίθμων σημαντικότητας χαρακτηριστικών, συνεπώς δεν έχουν σημαντική περιττή πληροφορία.

IV. ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ

Όσον αφορά το πρόβλημα της ταξινόμησης των δεδομένων, η επιλογή του αλγορίθμου ταξινόμησης εξαρτάται από το πρόβλημα προς επίλυση. Στο πρόβλημα του Ηχητικού Εντοπισμού Μουσικής/Ομιλίας, σύμφωνα με τη βιβλιογραφία, δεν εμφανίζεται κάποιος αλγόριθμος που είναι ανώτερος από όλους τους άλλους. Για αυτό το λόγο αποφασίστηκε να γίνει έλεγχος αρκετών μεθόδων και να επιλεχθεί η καλύτερη. Συγκεκριμένα δοκιμάστηκαν οι παρακάτω αλγόριθμοι : **Κ-κοντινότεροι γείτονες**(*KNN*), **Απλός ταξινομητής Bayes**(*Naive Bayes*) και **Μηχανές Διανυσμάτων Υποστήριξης** (*SVM*), **Συλλογική Μάθηση**(*Ensemble Learning*).

A. Κ-Κοντινότεροι Γείτονες

Ο KNN είναι ένας από τους πιο συχνά χρησιμοποιούμενους αλγορίθμους επειδή είναι διαισθητικός, απλός στην εφαρμογή και προσφέρει αξιοσημείωτα καλή απόδοση ακόμη και για απαιτητικές εφαρμογές. Μειονέκτημά του είναι ότι πρέπει να έχει αποθηκευμένο συνεχώς ολόκληρο το **σύνολο των δεδομένων**(*dataset*).

B. Απλός ταξινομητής Bayes

Η μέθοδος Bayes δίνει τον βέλτιστο ταξινομητή αν είναι γνωστή η κατανομή της (υπό-όρους) πιθανότητας των δεδομένων να ανήκουν σε μια κλάση(=P(x|ωj), όπου x = το διάνυσμα γνωρισμάτων ενός δείγματος για ταξινόμηση , ωj = η κλάση j). Το παραπάνω μειονέκτημα αντιμετωπίζεται από τη μέθοδο Naive Bayes που προϋποθέτει ανεξαρτησία των γνωρισμάτων ενός δείγματος δεδομένου ότι ανήκει σε μια κλάση ω. Αυτή η μέθοδος, έχει αποδειχθεί στην πράξη, ότι παρέχει αξιόλογα αποτελέσματα που είναι συγκρινόμενα με αυτά των νευρωνικών δικτύων.

C. Μηχανές Διανυσμάτων Υποστήριξης

Τα SVMs έχουν την ικανότητα να λύνουν προβλήματα ταξινόμησης μεγάλων διαστάσεων χωρίς να εκτελούν πράξεις στις μεγαλύτερες διαστάσεις χάρη στο **κόλπο πυρήνα**(*kernel trick*).

D. *Συλλογική Μάθηση*

Στη συλλογική μάθηση χρησιμοποιούνται περισσότεροι του ενός αλγόριθμοι ταξινόμησης ώστε να παραχθούν καλύτερα αποτελέσματα.

V. ΑΞΙΟΛΟΓΗΣΗ

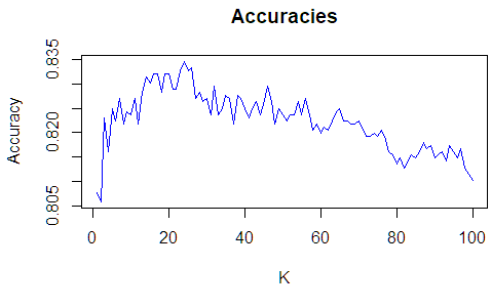
Για την αξιολόγηση της απόδοσης των παραπάνω αλγορίθμων τα κριτήρια που θα χρησιμοποιηθούν είναι η ***Ακρίβεια*** *(Accuracy)*, και η **βαθμολογία-F1***(F1-score)* σε επίπεδο παραθύρων όπως ορίζονται από το διαγωνισμό. Επίσης μέσω της μεθόδου *k-fold cross validation* επιτυγχάνεται καλύτερη αξιολόγησης της απόδοσης γενίκευσης των αλγορίθμων(δηλαδή της σωστής ταξινόμησης δεδομένων που δεν έχουν δει), εφόσον λαμβάνεται ο μέσος όρος και η διακύμανση των παραπάνω κριτηρίων. Η παρατήρηση της διακύμανσης είναι πολύ σημαντική καθώς η κρίση με χρήση μόνο του μέσου όρου μπορεί να δώσει εσφαλμένα αισιόδοξα αποτελέσματα.

VI. ΥΛΟΠΟΙΗΣΗ

Ακολουθεί αναλυτικά η διαδικασία που ακολουθήσαμε. Ο αναγνώστης καλείται να έχει ανοιχτό τα αρχεία του κώδικα διότι είναι επαρκώς σχολιασμένα και διευκολύνουν την κατανόηση.

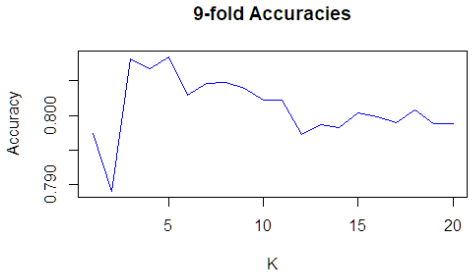
A. *K-Κοντινότεροι Γείτονες*

Αρχικά έγινε διαχωρισμός σε σύνολα εκπαίδευσης και ελέγχου με το 80% του αρχικού συνόλου να χρησιμοποιείται για εκπαίδευση και το 20% για έλεγχο(*80-20 split*). Τα συγκεκριμένα ποσοστά επιλέχθηκαν καθώς διατηρούν τις 60-άδες, που αντιστοιχούν σε ένα αρχείο, σε ένα μόνο σύνολο. Αν ένα αρχείο είχε τα παράθυρα του και στα δυο σύνολα τότε θα οδηγούσε και πάλι σε εσφαλμένα αισιόδοξα αποτελέσματα. Ακολούθησε κανονικοποίηση με τη συνάρτηση myNormalization() και έτρεξε ο αλγόριθμος KNN για K από 1 έως 100 παράγοντας το παρακάτω διάγραμμα :

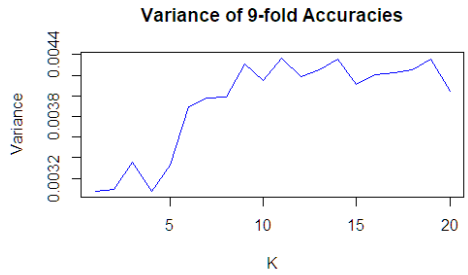


2. Ακρίβεια για διαφορετικά K

Ακολουθεί η εφαρμογή της μεθοδου *9fold cross validation*(για K από 1 εώς 20) με 85-15 split που επιλέχθηκε ώστε τα folds να έχουν πλήρεις 60άδες για τον λόγο που προαναφέρθηκε. Για κάθε επανάληψη αποθηκεύουμε τη μέση τιμή και τη διακύμανση, για όλες τις εναλλαγές των folds, καθώς η μέση τιμή δεν μας δίνει όλη την πληροφορία για την απόδοση γενίκευσης του αλγορίθμου. Παρακάτω παρουσιάζονται τα αποτελέσματα :

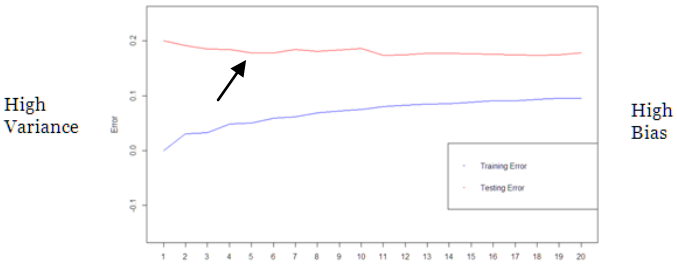


3. 9fold Ακρίβεια για διαφορετικά K



4. 9fold Διακύμανση για διαφορετικά K

Παρατηρείται μέγιστη Ακρίβεια για K = 5 και αμελητέα διακύμανση. Συμπεραίνουμε ότι η τιμή K = 5 είναι καλή επιλογή. Ακολουθεί ένας ακόμα έλεγχος όπου δημιουργείται το διάγραμμα σφάλματος εκπαίδευσης και ελέγχου για 80-20 split:



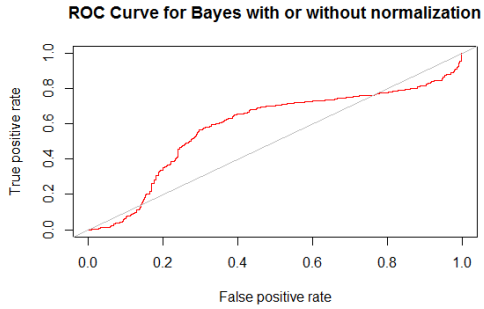
5. Διάγραμμα σφάλματος εκπαίδευσης και ελέγχου

Όπως φαίνεται για K = 5 το σφάλμα ελέγχου αρχίζει να αυξάνεται ενώ το σφάλμα εκπαίδευσης συνεχίζει να μειώνεται. Όσο το K μικραίνει τόσο αυξάνει η προσαρμοστικότητα του αλγορίθμου, άρα το διάγραμμα φαίνεται ανεστραμμένο [7]. Έτσι επιβεβαιώνεται η ορθότητα της επιλογής. Τελικά, στο αρχικό σύνολο δεδομένων ,η Ακρίβεια(για μουσική και ομιλία) ισούται με 0,84. Όσον αφορά το αρχείο mySong.wav που περιέχει τα 3 πρώτα λεπτά του theconcert16 προκύπτουν τα εξής αποτελέσματα : Ακρίβεια(για ομιλία) = 0,54 , Ακρίβεια(για μουσική) = 0,98 , Ακρίβεια(για μουσική και ομιλία) = 0,74 και Βαθμολογία-f1(για μουσική και ομιλία) = 0,69. Όλες οι παραπάνω μετρικές αναφέρονται σε ξεχωριστά παράθυρα άρα επιτυγχάνουν ανίχνευση ορίων. Ομοίως και στους υπόλοιπους αλγορίθμους.

B. *Απλός ταξινομητής Bayes*

Όσον αφορά την μέθοδο Naive - Bayes, ακολουθήθηκε η ίδια λογική. Επιπλέον εγινε απόπειρα απόδειξης της μη αναγκαίας κανονικοποίησης πριν την εφαρμογή του αλγορίθμου. Αυτό συμβαίνει διότι η ταξινόμηση γίνεται με βάση τις πιθανότητες των ανεξάρτητων χαρακτηριστικών και όχι με βάση την ευκλείδεια

απόσταση τους από ένα σημείο αναφοράς. Παρακάτω εμφανίζεται το διάγραμμα με τις ROC καμπύλες(επικαλύπτει η μια την άλλη) με κανονικοποίηση και χωρίς, που επαληθεύει τη θεωρία :



6. Καμπύλες ROC με και χωρίς κανονικοποίηση

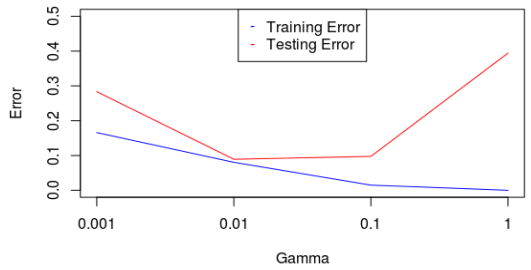
Επίσης , είναι χρήσιμο σε αυτό το σημείο να αναφερθεί ότι χρησιμοποιήθηκε Laplace smoothing ,έτσι ώστε να αποφύγουμε τον λανθασμένο μηδενισμό της πιθανότητας μιας ολόκληρης εγγραφής του σετ ελέγχου να ανήκει σε μια κλάση, επειδή κάποιο χαρακτηριστικό αυτής της εγγραφής είχε μηδενική πιθανότητα, αλλά αντ' αυτού να πάρουμε μια μικρή πιθανότητα, που αντιστοιχεί περισσότερο στην πραγματικότητα.

Τελικά, εκπαιδεύοντας το μοντέλο μας και κάνοντας προβλέψεις στα δεδομένα ελέγχου του αρχικού συνόλου δεδομένων, καταλήγουμε σε Ακρίβεια(για μουσική και ομιλία) ίση με 0,383 και Βαθμολογία-f1(για μουσική και ομιλία) ίση με 0,427 είτε κανονικοποιώντας τα δεδομένα είτε όχι. Όσον αφορά το αρχείο mySong.wav τα αποτελέσματα είναι τα εξής : Ακρίβεια(για ομιλία) = 0,53 , Ακρίβεια(για μουσική) = 1 , Ακρίβεια(για μουσική και ομιλία) = 0,74 και Βαθμολογία-f1(για μουσική και ομιλία) = 0,69. Όλες οι παραπάνω μετρικές αναφέρονται σε ξεχωριστά παράθυρα άρα επιτυγχάνουν ανίχνευση ορίων.

C. Μηχανές Διανυσμάτων Υποστήριξης

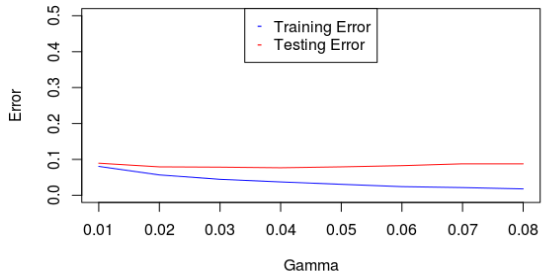
Όσον αφορά τα SVMs, για την παραγωγή των καλύτερων δυνατών αποτελεσμάτων ταξινόμησης, πρέπει πρώτα να αποφασίστει το είδος του **πυρήνα**(*kernel*) που θα χρησιμοποιηθεί, ανάλογα αν το πρόβλημα είναι γραμμικό ή όχι , καθώς και να γίνει ρύθμιση των τυχόν παραμέτρων. Επειδή το πρόβλημα είναι πολυδιάστατο και δεν υπήρχε δυνατότητα για μια εποπτική εκτίμηση των δεδομένων, ακολουθήθηκε η τακτική που αναλύεται στην επόμενη παράγραφο.

Αρχικά, υπολογίστηκε το σφάλμα εκπαίδευσης και το σφάλμα ελέγχου για το **γραμμικό**(*linear*), **ακτινικό**(*radial*) και **σιγμοειδή** (*sigmoid*) πυρήνα για μια πληθώρα από παραμέτρους . Το σφάλμα ελέγχου γραμμικού πυρήνα ήταν 0,236. Παρακάτω εμφανίζεται το διάγραμμα σφαλμάτων για τον ακτινικό πυρήνα :



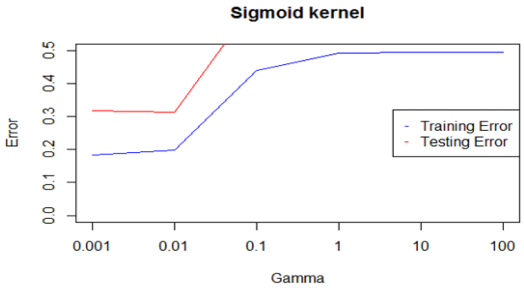
7. Σφάλμα εκπαίδευσης(μπλε) και σφάλμα ελέγχου (κόκκινο) για ακτινικό πυρήνα

Ακόμη, όπως φαίνεται από το παρακάτω διάγραμμα, δεν υπάρχει κάποια βέλτιστη τιμή στην περιοχή 0,01 έως 0,1 για την παράμετρο gamma :



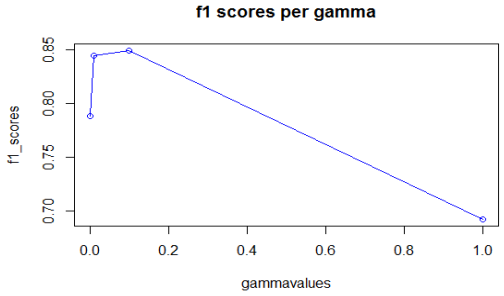
8. Σφάλμα εκπαίδευσης(μπλε) και σφάλμα ελέγχου (κόκκινο) για ακτινικό πυρήνα στην περιοχή 0,01 έως 0,1.

Παρακάτω παρουσιάζονται τα αποτελέσματα για τον σιγμοειδή πυρήνα :



9. Σφάλμα εκπαίδευσης(μπλε) και σφάλμα ελέγχου (κόκκινο) για σιγμοειδή πυρήνα

Έπειτα, εφόσον ο ακτινικός πυρήνας είχε το χαμηλότερο σφάλμα ελέγχου χρησιμοποιήθηκε για περαιτέρω ανάλυση. Πιο συγκεκριμένα, για να αποφασίσουμε την τελική τιμή για την παράμετρο gamma, εφαρμόσαμε την μέθοδο k-fold Cross Validation και συγκεκριμένα για k = 9 ,για μια πληθώρα από παραμέτρους ,δίνοντας ιδιαίτερη προσοχή στα δεδομένα εκπαίδευσης και επικύρωσης, ώστε αυτά να περιέχουν και τα 60 πακέτα από κάθε αρχείο ήχου που περιλαμβάνουν. Παρακάτω εμφανίζεται το διάγραμμα μετά την εκτέλεση του αλγορίθμου :



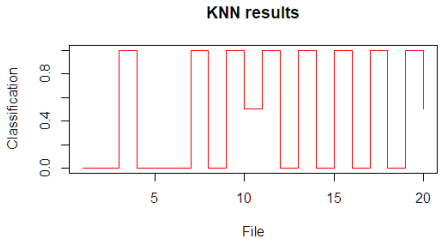
10. Μετρική F1_Score για ακτινικό πυρήνα (radial kernel)

Παρατηρήθηκε για ποια τιμή εμφανίζεται η μεγαλύτερη Βαθμολογία-f1, που όπως φαίνεται από το διάγραμμα είναι η gamma = 0,1. Έτσι, εκπαιδεύοντας το μοντέλο με αυτήν την τιμή και κάνοντας προβλέψεις στα δεδομένα ελέγχου του αρχικού συνόλου, καταλήγουμε σε Ακρίβεια(για μουσική και ομιλία) ίση με 0,902 και Βαθμολογία-f1 ίση με 0,897. Δεν χρησιμοποιήθηκε η συνάρτηση myNormalization() καθώς με το όρισμα scale = TRUE στη συνάρτηση svm() γίνεται σωστά η κανονικοποίηση. Τα αποτελέσματα στο αρχείο mySong.wav είναι τα παρακάτω : Ακρίβεια(για ομιλία) = 0,63 , Ακρίβεια(για μουσική) = 1 , Ακρίβεια(για μουσική και ομιλία) = 0,79 και Βαθμολογία-f1(για μουσική και ομιλία) = 0,77.

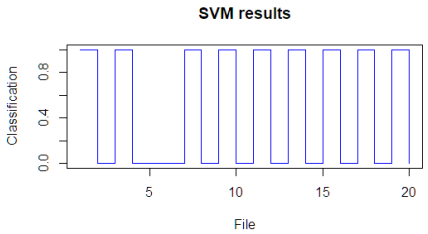
D. Συλλογική Μάθηση

Στη συλλογική μάθηση συμμετείχαν οι αλγόριθμοι KNN και SVMs διότι εμφάνισαν τα καλύτερα αποτελέσματα. Καθώς δεν υπήρχε τρίτος αλγόριθμος για να χρησιμοποιηθεί **ψήφος πλειοψηφίας**(*majority voting*), έγινε αντ'αυτού προσπάθεια ταξινόμησης ολόκληρων των αρχείων. Αρχικά γίνεται καταμέτρηση του αριθμού των παραθύρων κάθε κλάσης σε 60άδες(δηλαδή ολόκληρα αρχεία) όπως προκύπτουν από τα αποτελέσματα των δυο αλγορίθμων. Αν εμφανίζεται μεγαλύτερο από 75% ποσοστό παραθύρων μιας κλάσης τότε υποθέτουμε ότι το αρχείο είναι αυτή η κλάση(1 αν είναι μουσική και 0 αν είναι ομιλία), διαφορετικά δεν είναι σίγουρο το αποτέλεσμα και παίρνει την τιμή 0,5. Το ίδιο γίνεται και για τον άλλον αλγόριθμο. Η τελική απόφαση λαμβάνεται σύμφωνα με τα παρακάτω : αν και οι δυο αλγόριθμοι βρήκαν το ίδιο τότε αυτή είναι και η σωστή τιμή, αν ο KNN δεν έχει σίγουρο αποτέλεσμα(δηλαδή 0,5) τότε σαν σωστή λαμβάνεται η τιμή του SVM, αν έχουν διαφορετικά αποτελέσματα(δηλαδή το ένα 1 και το άλλο 0) τότε η τελική τιμή είναι το 0,5 και έτσι επιτυγχάνεται πιθανή καλύτερη ανίχνευση μουσικής και ομιλίας συγχρόνως.

Παρακάτω εμφανίζονται τα αποτελέσματα για το αρχικό σύνολο ελέγχου. Πρόκειται για 20 αρχεία που είναι εναλλάξ μουσική-ομιλία ξεκινώντας από μουσική. Σκοπίμως αλλάχτηκε η ταξινόμηση του πρώτου αρχείου από τον KNN ώστε να παρατηρηθεί αργότερα το τελικό αποτέλεσμα 0,5 :



11. Ταξινόμηση αρχείων (1 = μουσική, 0 = ομιλία, 0,5 = μη σίγουρο αποτέλεσμα)



12. Αποτελέσματα SVM

13. Τελικά αποτελέσματα

Παρατηρείται η διόρθωση των αποτελεσμάτων του KNN αλλά και η επιλογή του 0,5(στην αρχή) όταν δεν είναι σίγουρο το μοντέλο.

VII. Συμπερασμα

Όσον αφορά τον έλεγχο που έγινε με βάση τη διάσπαση του αρχικού συνόλου δεδομένων, τα SVMs μοντελοποιούν καλύτερα το πρόβλημα με δεύτερο να έρχεται ο KNN. Ο απλός ταξινομητής Bayes απορρίπτεται λόγω πολύ κακών αποτελεσμάτων στο αρχικό σύνολο δεδομένων. Σχετικά με τον έλεγχο στο αρχείο mySong.wav , όπως φαίνεται και από τα αποτελέσματα, τα SVMs παρέμειναν η βέλτιστη επιλογή. Ιδιαίτερης σημασίας είναι η παρατήρηση καλύτερης ανίχνευσης παραθύρων μουσικής από ότι ομιλίας. Συνεπώς το μοντέλο μας φαίνεται να είναι πιο ευαίσθητο στη μουσική. Επιπροσθέτως, με χρήση συλλογικής μάθησης, γίνεται προσπάθεια ταξινόμησης ολόκληρων αρχείων αλλά και ανίχνευση περιπτώσεων που πιθανώς είναι μουσική και ομιλία συγχρόνως.

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Music-ir.org. (2019). *2018:Music and/or Speech Detection - MIREX Wiki*. [online] Available at: https://www.music-ir.org/mirex/wiki/2018:Music_and/or_Speech_Detection [Accessed 18 Jan. 2019].
2. *Data Sets GTZAN Genre Collection*(Approximately 297MB) Marsyasweb.appspot.com. (2018). *Marsyas*. [online] Available at:

http://marsyasweb.appspot.com/download/data_sets/ [Accessed 12 Nov. 2018].

3. Mirg.city.ac.uk. (2019). *Index of /datasets/muspeak*. [online] Available at : http://mirg.city.ac.uk/datasets/muspeak [Accessed 18 Jan. 2019].

4. Kotsakis, R., Kalliris, G. and Dimoulas, C. Investigation of salient audio-features for patternbased semantic content analysis of radio productions

5. Kotsakis, R., Kalliris, G. and Dimoulas, C. (2012). Investigation of broadcast-audio semantic analysis scenarios employing radio-programme-adaptive pattern classification. *Speech Communication*, 54(6), pp.743-762.

6. Géron, A. (2018). *Hands-on machine learning with Scikit-Learn and TensorFlow*. Beijing: O'Reilly.

7. James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). *An Introduction to Statistical Learning*. New York, NY: Springer