Πτυχιακή εργασία του φοιτητή Παναγιωτόπουλου Ανδρέα

ΑΛΕΞΑΝΔΡΕΙΟ Τ.Ε.Ι. ΘΕΣΣΑΛΟΝΙΚΗΣ
ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΚΩΝ ΕΦΑΡΜΟΓΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Μείωση του πληθυσμού δεδομένων εκπαίδευσης πολλαπλών ετικετών

(Data Reduction for multi-label classifications datasets)

Του φοιτητή **Ανδρέα Παναγιωτόπουλου** (Αρ. Μητρώου: 123892)

Επιβλέπων: Στέφανος Ουγιάρογλου

Θεσσαλονίκη 2019

ΠΡΟΛΟΓΟΣ

Στα πλαίσια της εκπόνησης της πτυχιακής μου εργασίας επέλεξα μαζί με τον επιβλέπων καθηγητή μου το συγκεκριμένο θέμα. Το θεώρησα σημαντική ευκαιρία για να πάρω μια πρώτη εικόνα γύρω από το ευρύτερο κομμάτι της επιστήμης των δεδομένων. Μέσα από αυτή την εργασία επιβεβαιώθηκε τον ενδιαφέρον μου για το συγκεκριμένο κομμάτι συνειδητοποιώντας τις δυνατότητες που υπάρχουν και που προσφέρει η συγκεκριμένη περιοχή.

Στη συγκεκριμένη εργασία μελετάμε κάποιους αλγόριθμους μείωσης των δεδομένων πάνω σε σύνολα δεδομένων πολλαπλών ετικετών και πως μπορούν να επηρεάσουν τα αποτελέσματα στις μετρήσεις απόδοσης της κατηγοριοποίησης.

ΠΕΡΙΛΗΨΗ

Τα τελευταία χρόνια πολλές τεχνικές μείωσης του πληθυσμού των δεδομένων εκπαίδευσης έχουν προταθεί και είναι διαθέσιμες στη βιβλιογραφία. Οι τεχνικές αυτές σκοπεύουν στη γρήγορη κατηγοριοποίηση νέων στιγμιοτύπων με υψηλή ακρίβεια. Ωστόσο, οι τεχνικές δεν έχουν εφαρμοστεί σε προβλήματα κατηγοριοποίησης πολλαπλών ετικετών (multilabel classification). Σε αυτού του είδους τα προβλήματα, ένα αντικείμενο μπορεί να ανήκει ταυτόχρονα σε περισσότερες από μια κλάσεις. Στα πλαίσια της πτυχιακής εργασίας, εφαρμόζονται τεχνικές μείωσης του πληθυσμού των δεδομένων (data reduction) σε τέτοιου είδους σύνολα πολλαπλών ετικετών. Συγκεκριμένα, γνωστοί αλγόριθμοι παραγωγής και επιλογής πρωτοτύπων (prototype abstraction and selection algorithm) εφαρμόζονται σε τέτοιου είδους σύνολα δεδομένων με στόχο την επίτευξη τόσο υψηλής ακρίβειας στην κατηγοριοποίηση όσο θα είχε επιτευχθεί χωρίς τη μείωση του πληθυσμού αλλά με πολύ μικρότερο υπολογιστικό κόστος. Η εργασία παρουσιάζει τα αποτελέσματα πειραμάτων σε τρία σύνολα δεδομένων πολλαπλών ετικετών τα οποία προέκυψαν εφαρμόζοντας τη τεχνική επικύρωσης five-fold-cross-validation.

ABSTRACT

In recent years, many techniques for reducing the population of training data have been proposed and are available in the bibliography. These techniques aim at quickly classify new instances with high precision. However, the techniques have not been applied to multi-label problems. In this type of problem, an instance may belong to more than one class. In this thesis, data population reduction techniques (data reduction) are applied to such types of multi-label data-sets. In Particular, known algorithms for data condnensing and abstraction are applied to such type of data in order to achieve as high a classification as would be achieved without a population reduction but with much lower cost. The thesis presents the test results in three data-sets of multi-label data obtained by applying the 5-fold cross validation.

Πίνακας Περιεχομένων

Ευρετήριο πινάκων

Ευρετήριο εικόνων

ΚΕΦΑΛΑΙΟ 1: ΕΙΣΑΓΩΓΗ

1.1 Τι είναι η κατηγοριοποίηση

Η αποδοτικότητα και η αποτελεσματικότητα των αλγορίθμων εξόρυξης δεδομένων είναι ένα σημαντικό πρόβλημα έρευνας που έχει τραβήξει την προσοχή τόσο του ακαδημαϊκού κλάδου όσο και του βιομηχανικού. Η κατηγοριοποίηση αποτελεί σημαντικό έργο για την εξόρυξη δεδομένων. Το πρόβλημα της κατηγοριοποίησης, κατά τις προηγούμενες δεκαετίες κέντρισε το ενδιαφέρον πολλών ερευνητών από διαφορετικά πεδία έρευνας της επιστήμης υπολογιστών και αυτό έχει ως συνέπεια να είναι διαθέσιμοι στη βιβλιογραφία διάφοροι αλγόριθμοι κατηγοριοποίησης (αλγόριθμος κατηγοριοποίησης).

Οι αλγόριθμοι κατηγοριοποίησης έχουν ως σκοπό να αναθέσουν νέα, μη κατηγοριοποιημένα δεδομένα σε ένα σύνολο προκαθορισμένων κλάσεων, με βάση τα διαθέσιμα δεδομένα εκπαίδευσης, δηλαδή ένα σύνολο ήδη ταξινομημένων στιγμιοτύπων. Τυπικό παράδειγμα κατηγοριοποίησης είναι ο καθορισμός ενός email είτε ως “spam” είτε ως “not spam”.

Οι κατηγοριοποιητές μπορούν να χωριστούν σε δύο βασικές κατηγορίες: (α) eager αλγόριθμος κατηγοριοποίησης και (β) lazy αλγόριθμος κατηγοριοποίησης.

Κοινός σκοπός και των δύο είναι η πρόβλεψη κλάσεων με ακρίβεια. Παρόλα αυτά διαφέρουν στο τρόπο που δουλεύουν. Σημαντικό ρόλο για την αποτελεσματικότητα των αλγορίθμων παίζει το σύνολο εκπαίδευσης(training set). Ένας κατηγοριοποιητής eager αλγόριθμος κατηγοριοποίησης προεπεξεργάζεται τα δεδομένα εκπαίδευσης και δημιουργεί ένα μοντέλο κατηγοριοποίησης που χρησιμοποιείται μετά για την κατηγοριοποίηση νέων, μη κατηγοριοποιημένων αντικειμένων. Αντίθετα, ένας κατηγοριοποιητής lazy αλγόριθμος κατηγοριοποίησης δεν δημιουργεί κανένα μοντέλο. Θεωρεί το σύνολο εκπαίδευσης ως το μοντέλο κατηγοριοποίησης. Ένας lazy αλγόριθμος κατηγοριοποιεί ένα νέο αντικείμενο ψάχνοντας στο σύνολο εκπαίδευσης τι στιγμή που το λαμβάνει.

Από τη στιγμή που ένας eager αλγόριθμος κατηγοριοποίησης δημιουργεί ένα μοντέλο κατηγοριοποίησης πριν την έλευση οποιουδήποτε νέου αντικειμένου, η διαδικασία κατηγοριοποίησης είναι πολύ γρήγορη. Παρόλο που οι lazy αλγόριθμος κατηγοριοποίησης δεν ξοδεύουν χρόνο στη δημιουργία μοντέλου η διαδικασία κατηγοριοποίησης είναι περισσότερο χρονοβόρα από αυτή ενός eager αλγόριθμος κατηγοριοποίησης. Ένα μειονέκτημα ενός eager αλγόριθμος κατηγοριοποίησης είναι ότι πρέπει να παράγει μια υπόθεση που θα καλύψει ολόκληρο το σύνολο εκπαίδευσης. Αυτό δεν είναι πάντα εφικτό καθώς μπορεί να επηρεάσει την ακρίβεια της κατηγοριοποίησης και μπορεί να καθορίσει την κατασκευή του μοντέλου ως μια πολύ χρονοβόρα και περίπλοκη εργασία προεπεξεργασίας. Οι lazy αλγόριθμος κατηγοριοποίησης χρησιμοποιούν ολόκληρο το σύνολο εκπαίδευσης με αποτέλεσμα να μπορούν να υιοθετήσουν πιο περίπλοκες

υποθέσεις για τα δεδομένα. Κατά συνέπεια, μπορούν να βελτιώσουν την ακρίβεια της κατηγοριοποίησης. Ένα μειονέκτημα των lazy αλγόριθμος κατηγοριοποίησης είναι ότι απαιτούν όλα τα δεδομένα εκπαίδευσης να είναι πάντα διαθέσιμα το οποίο απαιτεί μεγάλες απαιτήσεις σχετικά με την αποθήκευση ενώ οι eager αλγόριθμος κατηγοριοποίησης, μετά την κατασκευή του μοντέλου κατηγοριοποίησης, τα δεδομένα εκπαίδευσης μπορούν να αφαιρεθούν για να ελευθερωθεί χώρος.

Η κατηγορία των lazy αλγόριθμος κατηγοριοποίησης περιλαμβάνει τον γνωστό κατηγοριοποιητή των k εγγύτερων γειτόνων (k-Nearest Neighbours) ο οποίος θεωρείται ένας από τους πιο απλούς αλγορίθμους όπου τα δεδομένα που ανήκουν στο σύνολο εκπαίδευσης επεξεργάζονται όταν εμφανιστεί ένα νέο αντικείμενο. Κάθε φορά που ένα αντικείμενο πρόκειται να κατηγοριοποιηθεί, υπολογίζεται η ομοιότητά του με κάθε ένα από τα δεδομένα του συνόλου εκπαίδευσης. Σαν μέτρο ομοιότητας του κατηγοριοποιητή k εγγύτερων γειτόνων(k-NN) χρησιμοποιείται μια συνάρτηση που υπολογίζει την απόσταση ανάμεσα σε κάθε αντικείμενο του συνόλου εκπαίδευσης και του αντικειμένου που πρόκειται να κατηγοριοποιηθεί. Πιο συγκεκριμένα, ο κατηγοριοποιητής ψάχνει ανάμεσα στα k πιο κοντινά αντικείμενα του συνόλου εκπαίδευσης και τοποθετεί σε αυτά μια ετικέτα βάση της κλάσης στην οποία ανήκουν τα αντικείμενα.

1.2 Αποτίμηση της απόδοσης των κατηγοριοποιητών

Αυτό που εξετάζουμε για να δούμε την απόδοση ενός κατηγοριοποιητή είναι η ακρίβεια(accuracy) κατηγοριοποίησης, την ικανότητα του μοντέλου δηλαδή στο να κάνει σωστές προβλέψεις σε νέα δεδομένα. Ο ρόλος της ακρίβειας είναι πολύ σημαντικός καθώς μας φανερώνει την ικανότητα ενός κατηγοριοποιητή στο να προβλέπει σωστά νέα δεδομένα για τα οποία δεν έχει εκπαιδευτεί. Ακόμα, η εκτίμηση της ακρίβειας μας βοηθάει και μας επιτρέπει να συγκρίνουμε διαφορετικούς αλγόριθμους κατηγοριοποίησης.

Ωστόσο, πέρα από την ακρίβεια(accuracy) υπάρχουν και άλλοι τρόποι στο να συγκρίνουμε κατηγοριοποιητές μεταξύ τους. Αυτοί είναι οι:

a) **Ταχύτητα:** Το κόστος υπολογισμού, συνυπολογίζοντας και την παραγωγή και την χρήση του μοντέλου.

b) **Rebustness:** Η σωστή πρόβλεψη ενώ τα δεδομένα είναι είτε ελλειπή είτε περιέχουν θόρυβο.

c) **Scalability:** Όταν η ποσότητα των δεδομένων είναι μεγάλη, η κατασκευή του μοντέλου να είναι αποδοτική.

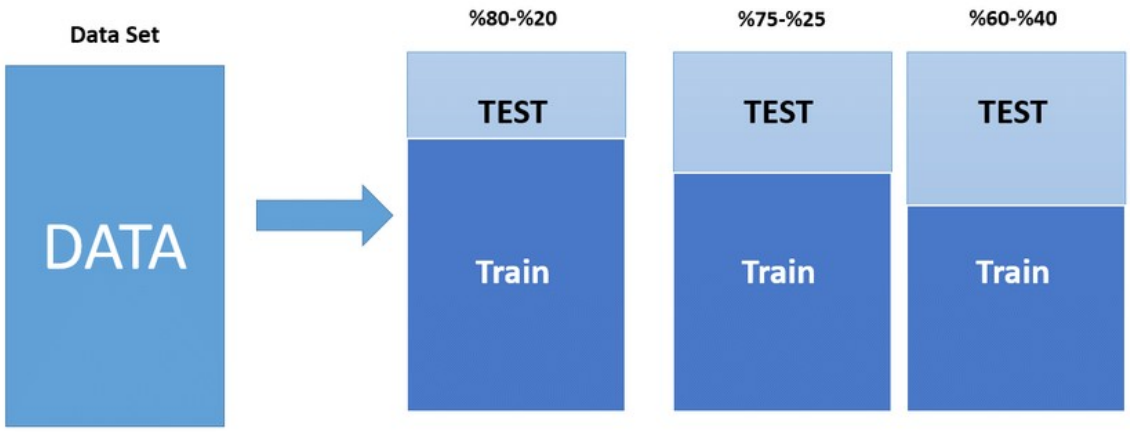
d) **Interpretability**: Το πόσο κατανοητό είναι το μοντέλο και το επίπεδο γνώσης που παρέχεται από αυτό. (Η εκτίμηση μπορεί να γίνει π.χ από τον αριθμό των κόμβων των δέντρων απόφασης)

Παρόλο που η ακρίβεια αποτελεί το σημαντικότερο μέτρο μέτρησης της απόδοσης ενός κατηγοριοποιητή δεν θα πρέπει να υπολογίζεται ξεχωριστά από τα υπόλοιπα μέτρα καθώς δεν έχει νόημα να έχουμε έναν αλγόριθμο που μας επιστρέφει αποτελέσματα με υψηλή ακρίβεια αλλά να είναι χρονοβόρος. Πολλές φορές προτιμάται ο αλγόριθμος κατηγοριοποίησης να επιστρέφει αποτελέσματα με χαμηλότερη ακρίβεια αλλά να είναι πιο γρήγορος. Το αποτέλεσμα της κατηγοριοποίησης θεωρείται ακριβές σύμφωνα με τον καθορισμό ενός ποσοστού αντικειμένων τα οποία κατηγοριοποιήθηκαν στη σωστή κλάση.

Οι τρόποι για να εκτιμήσουμε το πόσο ακριβής είναι ένας αλγόριθμος κατηγοριοποίησης είναι οι εξής:

Αρχικά, μπορούμε να χρησιμοποιήσουμε ένα σύνολο δεδομένων για να εκπαιδεύσουμε τον αλγόριθμό μας και στη συνέχεια να χρησιμοποιήσουμε το ίδιο σύνολο δεδομένων για να εκτιμήσουμε την ακρίβεια του αλγορίθμου. Αυτό το σενάριο είναι αρκετά αισιόδοξο καθώς ο αλγόριθμος εκπαιδεύεται και στη συνέχεια δοκιμάζεται στο ίδιο σύνολο δεδομένων.

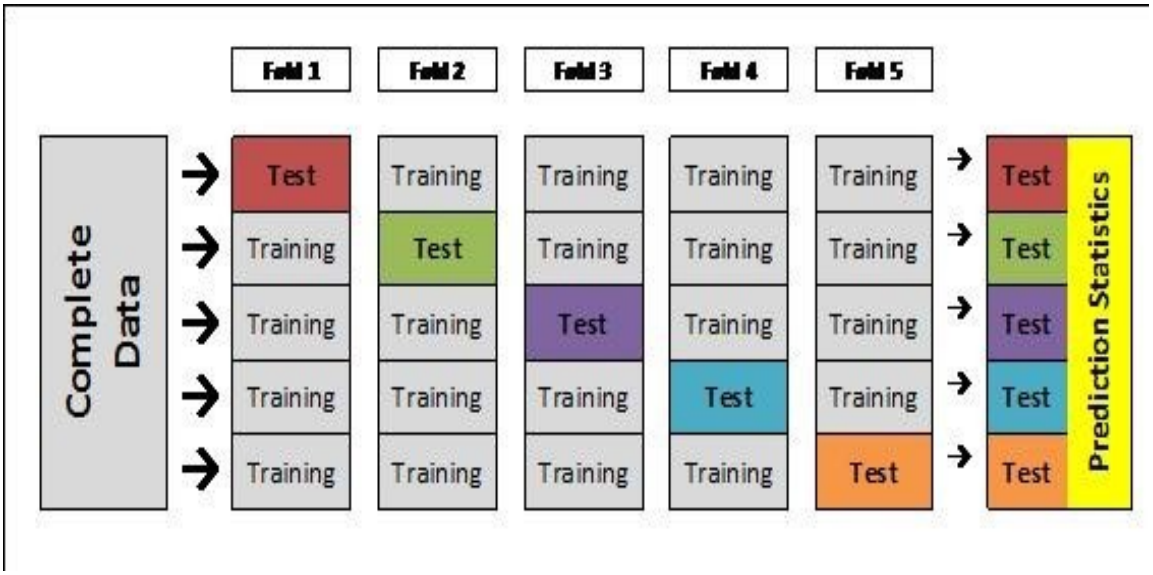
Ένας άλλος τρόπος εκτίμησης της ακρίβειας ενός αλγορίθμου είναι η μέθοδος κατακράτησης(hold-out method). Με αυτή τη μέθοδο, το σύνολο δεδομένων χωρίζεται τυχαία σε δύο σύνολα δεδομένων. Το πρώτο σύνολο ονομάζεται σύνολο δεδομένων εκπαίδευσης(train test) και χρησιμοποιείται για την εκπαίδευση του αλγορίθμου ενώ το δεύτερο σύνολο ονομάζεται σύνολο δεδομένων δοκιμής(test set) και χρησιμοποιείται για τη δοκιμή του αλγορίθμου. Συνήθως ο τρόπος που χωρίζεται το αρχικό σύνολο δεδομένων είναι τα 2/3 του αρχικού σαν σύνολο εκπαίδευσης και το υπόλοιπο 1/3 σαν σύνολο δοκιμής.



Εικόνα 1: "Σύνολο δεδομένων εκπαίδευσης και δοκιμής με τη μέθοδο hold out"

Σημαντικό μέτρο για την εκτίμηση της απόδοσης, είναι ο cross validation. Στη συγκεκριμένη περίπτωση, το αρχικό σύνολο δεδομένων χωρίζεται σε σύνολο δεδομένων εκπαίδευσης και δοκιμής αρκετές φορές χρησιμοποιώντας κάθε φορά διαφορετικά υποσύνολα και στο τέλος βγαίνει ο μέσος όρος από όλες τις διαφορετικές περιπτώσεις. Στη μέθοδο cross validation υπάρχουν δύο παραλλαγές:

K-fold cross validation: Στη συγκεκριμένη μέθοδο, το αρχικό σύνολο δεδομένων χωρίζεται σε k υποσύνολα. Από τα k υποσύνολα, ένα χρησιμοποιείται για δοκιμή ενώ τα υπόλοιπα k-1 υποσύνολα χρησιμοποιούνται για εκπαίδευση. Η διαδικασία επαναλαμβάνεται k φορές, όπου κάθε ένα από τα k υποσύνολα χρησιμοποιείται μια φορά σαν σύνολο δοκιμής. Ένα από τα πλεονεκτήματα αυτής της μεθόδου είναι ότι όλα τα αντικείμενα του συνόλου δεδομένων χρησιμοποιούνται τόσο για εκπαίδευση όσο και για δοκιμή. Όπως είναι λογικό, αυτή η μέθοδος χρειάζεται k φορές περισσότερο χρόνο σε σχέση με την hold-out μέθοδο. Συνήθως, στις περισσότερες εργασίες σαν k χρησιμοποιείται ή το 5 ή το 10.



Εικόνα 2: "Μέθοδος cross validation με k=5"

Leave-one-out cross validation: Όπως μπορούμε να καταλάβουμε και από το όνομα, σε αυτή τη μέθοδο χρησιμοποιείται ένα μόνο αντικείμενο από το αρχικό σύνολο δεδομένων για δοκιμή ενώ όλα τα υπόλοιπα αντικείμενα χρησιμοποιούνται για εκπαίδευση. Η διαδικασία επαναλαμβάνεται μέχρι να χρησιμοποιηθούν όλα τα αντικείμενα για δοκιμή τουλάχιστον μία φορά. Δεν υπάρχουν πολλές διαφορές με το k-fold cross validation μόνο που εδώ το k ισούται με τον αριθμό των αντικειμένων. Αυτή η μέθοδος δίνει τα καλύτερα αποτελέσματα αλλά είναι φανερό ότι το κόστος της είναι μεγάλο καθώς χρειάζεται μεγάλη υπολογιστική ισχύς λόγω των επαναλήψεων που απαιτούνται.

	ΠΡΟΒΛΕΨΗ	
	Χαλαζόπτωση	Όχι χαλαζόπτωση
Χαλαζόπτωση	TP	FN
Όχι χαλαζόπτωση	FP	TN

Εικόνα 3: "Πίνακας σύγχυσης (Confussion Matrix)"

1.3 Τι είναι τα σύνολα δεδομένων πολλαπλών ετικετών

Τα προβλήματα κατηγοριοποίησης μπορούν να διακριθούν σε τρεις κατηγορίες. Η πρώτη κατηγορία αφορά τα δυαδικά προβλήματα κατηγοριοποίησης (binary classification). Σε αυτή την περίπτωση, στο σύνολο δεδομένων υπάρχουν δύο κλάσεις. Έτσι, ένα στιγμιότυπο του συνόλου δεδομένων ανήκει σε μια από αυτές τις δύο κλάσεις. Η δεύτερη κατηγορία προβλημάτων είναι αυτή των πολλαπλών κλάσεων (multiclass classification) όπου στο σύνολο δεδομένων υπάρχουν περισσότερες από δύο κλάσεις και ένα στιγμιότυπο ανήκει πάντα σε μόνο μια από αυτές. Τέλος, η τρίτη κατηγορία αφορά τα προβλήματα πολλαπλών ετικετών όπου στο σύνολο δεδομένων υπάρχουν περισσότερες από δύο κλάσεις και ένα στιγμιότυπο μπορεί να ανήκει σε περισσότερες από μια κλάσεις.

Οι λόγοι που οδήγησαν στην χρήση της multi-label κατηγοριοποίησης ήταν κυρίως για κατηγοριοποίηση κειμένου και ιατρικές διαγνώσεις. Επίσης, τα έγγραφα κειμένου ανήκουν συνήθως σε πάνω από μια κλάσεις ενώ ένα άλλο παράδειγμα είναι η κατηγοριοποίηση ταινιών όπου μια ταινία μπορεί να είναι ταυτόχρονα δράσης και μυστηρίου.

Εφαρμογές που εξαρτώνται περισσότερο από αλγορίθμους multi-label κατηγοριοποίησης στις μέρες μας, είναι η κατηγοριοποίηση της λειτουργίας των πρωτεϊνών, τα μουσικά κομμάτια, η σημασιολογική κατηγοριοποίηση εικόνων ενώ το πιο κλασσικό παράδειγμα στις μέρες μας είναι η κατηγοριοποίηση ταινιών οι οποίες ανήκουν σε πάνω από μια κλάσεις, όπως για παράδειγμα ότι μια ταινία μπορεί να ανήκει ταυτόχρονα και στη κλάση περιπέτεια και στη κλάση θρίλερ.

Εικόνα 4: "Η εικόνα απεικονίζει ταυτόχρονα ένα σπίτι,σύννεφα και δέντρα"

Για την εικόνα 4, στην ερώτηση εάν περιλαμβάνει ένα σπίτι, η απάντηση θα είναι ναι ή όχι. Παρόλα αυτά, περιλαμβάνονται και άλλα στοιχεία στην εικόνα τα οποία μπορούν να την περιγράψουν όπως δέντρο,σύννεφα ή βουνό. Αυτού του είδους τα προβλήματα, όπου υπάρχει ένα σύνολο ετικετών, είναι γνωστά ως προβλήματα multi-label κατηγοριοποίησης και δεν πρέπει να συνδέονται με προβλήματα που χαρακτηρίζονται ως multi-class. Πιο συγκεκριμένα, ένα παράδειγμα για να φανεί η διαφορά ανάμεσα στα δύο. Μια ταινία μπορεί να πάρει κάποια ετικέτα κλάσης που να σχετίζεται με το ποιος μπορεί να την δει, όπως να είναι ακατάλληλη για παιδιά ηλικίας κάτω των 12 ή να είναι κατάλληλη μόνο για ενήλικες, αλλά το σίγουρο είναι ότι κάθε ταινία μπορεί να κατηγοριοποιηθεί με έναν μόνο τύπο τέτοιου πιστοποιητικού. Από την άλλη, η ίδια ταινία μπορεί να χαρακτηριστεί ως ρομαντική και ως κωμωδία, ανήκοντας σε δύο κατηγορίες ταυτόχρονα.

1.3.1 Πόσο multi-label είναι ένα σύνολο δεδομένων

Τα σύνολα δεδομένων δεν είναι όλα το ίδιο multi-label. Ο αριθμός των ετικετών σε κάθε πρόβλημα μπορεί να είναι πολύ μικρός ή πολύ μεγάλος σε σχέση με το μέγεθος του συνόλου των ετικετών *L*. Αυτή η παράμετρος είναι αρκετά σημαντική για την αποτελεσματικότητα των διάφορων multi-label κατηγοριοποιητών διότι κάποιοι αλγόριθμοι έχουν διαφορετική απόδοση ανάλογα με τα χαρακτηριστικά του συνόλου δεδομένων στα οποία εφαρμόζονται.

Επίσης, έχουν οριστεί οι έννοιες πολλαπλότητα ετικετών(label cardinality) και πυκνότητα ετικετών(label density). Αν D είναι ένα multi-label σύνολο αποτελούμενο από |D| multi-label παραδείγματα, τότε ως πολλαπλότητα ετικετών του D ορίζεται ο μέσος αριθμός ετικετών των παραδειγμάτων στο D ενώ, ως πυκνότητα ετικετών του D ορίζεται ο μέσος αριθμός ετικετών των παραδειγμάτων στο D διαιρούμενος με το |*L*|.

Σε ένα πρόβλημα κατηγοριοποίησης, η πολλαπλότητα ετικετών είναι ανεξάρτητη του αριθμού των ετικετών |*L*| και η χρήση της απευθύνεται στην ποσοτικοποίηση του αριθμού των διαφορετικών ετικετών που χαρακτηρίζει ένα αντικείμενο σε ένα multi-label σύνολο εκπαίδευσης.

Αντίθετα, η πυκνότητα ετικετών, είναι ανεξάρτητη του αριθμού των ετικετών |*L*| στο πρόβλημα κατηγοριοποίησης και έχει να κάνει με την ποσοτικοποίηση του αριθμού των διαφορετικών ετικετών που αφορούν τα παραδείγματα ενός multi-label συνόλου εκπαίδευσης.

Από την πλευρά τους οι αλγόριθμοι κατηγοριοποίησης παρουσιάζουν διαφορετικές επιδόσεις όταν εφαρμόζονται σε σύνολα δεδομένων που έχουν την ίδια πολλαπλότητα ετικετών αλλά διαφέρουν στο μέγεθος του συνόλου των ετικετών.

1.3.2 Κατηγορίες μεθόδων multi-label κατηγοριοποίησης

Οι μέθοδοι multi-label κατηγοριοποίησης μπορούν να χωριστούν σε δύο κατηγορίες: α) μέθοδοι μετασχηματισμού προβλήματος και β) μέθοδοι προσαρμογής αλγορίθμων.

Η πρώτη κατηγορία μεθόδων είναι ανεξάρτητες από τον αλγόριθμο κατηγοριοποίησης που χρησιμοποιείται. Αυτό που κάνουν είναι ότι μετατρέπουν ένα πρόβλημα multi-label κατηγοριοποίησης σε πρόβλημα single-label κατηγοριοποίησης

Στην δεύτερη κλάση, προκειμένου οι μέθοδοι να μπορέσουν να διαχειριστούν multi-label δεδομένα απευθείας, επεκτείνουν συγκεκριμένους αλγόριθμους κατηγοριοποίησης. Επεκτάσεις υπάρχουν για δέντρα αποφάσεων(decision trees), μηχανές διανυσμάτων υποστήριξης(support vector machines), νευρωνικά δίκτυα και άλλα.

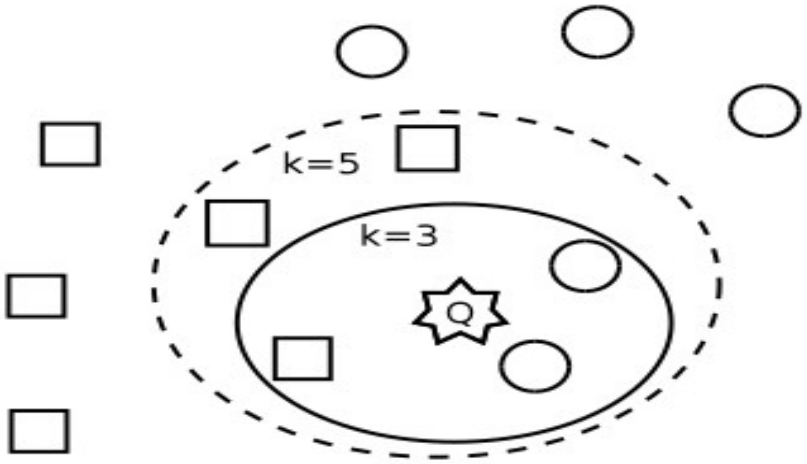
Ειδικότερα, στη συγκεκριμένη πτυχιακή εργασία, επικεντρωθήκαμε στη πρώτη μέθοδο κατηγοριοποίησης, σε αυτή του μετασχηματισμού προβλήματος και τα πειράματα τα οποία κάναμε έγιναν με βάση αυτή τη μέθοδο.

1.4 Ο Κατηγοριοποιητής k-Nearest Neighbor

Ο λόγος για τον οποίο δίνεται μεγαλύτερη σημασία στους αλγόριθμους κατηγοριοποίησης(classification) και λιγότερη σημασία στους αλγόριθμους παλινδρόμησης(regression) είναι ότι πολλά από τα προβλήματα που αντιμετωπίζουμε στην καθημερινή μας ρουτίνα ανήκουν σε εργασίες κατηγοριοποίησης. Για παράδειγμα, στην ιατρική, θα θέλαμε να μάθουμε εάν ένας όγκος είναι καλοήθης ή κακοήθης, στα θέματα μιας επιχείρησης θέλουμε να γνωρίζουμε εάν το προϊόν που πούλησε μια επιχείρηση είχε θετική ή αρνητική απήχηση στους καταναλωτές και πολλά άλλα παραδείγματα. Ο αλγόριθμος των Κ Εγγύτερων Γειτόνων (k – Nearest Neighbours) είναι ένας τέτοιος αλγόριθμος κατηγοριοποίησης και ανήκει στην κλάση των lazy αλγόριθμος κατηγοριοποίησης το οποίο σημαίνει ότι δεν κατασκευάζει κάποιο μοντέλο κατηγοριοποίησης.

Η διαδικασία κατηγοριοποίησης με τον συγκεκριμένο αλγόριθμο το μόνο που προϋποθέτει είναι η αποθήκευση των δεδομένων εκπαίδευσης. Τα δεδομένα εκπαίδευσης επεξεργάζονται όταν κάνει την εμφάνισή του ένα νέο αντικείμενο το οποίο, όταν πρόκειται να κατηγοριοποιηθεί, υπολογίζεται η ομοιότητά του με κάθε ένα από τα αντικείμενα που ανήκουν στο σύνολο δεδομένων εκπαίδευσης.

Πιο συγκεκριμένα, ο αλγόριθμος ψάχνει στο σύνολο δεδομένων εκπαίδευσης και ανακτά τα *k* πιο κοντινά αντικείμενα(γείτονες) του νέου αντικειμένου σύμφωνα με μια μετρική απόστασης. Τότε το νέο αντικείμενο κατατάσσεται στην κλάση που επικρατεί ανάμεσα στις κλάσεις των γειτόνων του. Συνήθως, αυτή η κλάση χαρακτηρίζεται ως η κύρια κλάση και καθορίζεται μέσω μιας διαδικασίας γνωστής ως ψηφοφορία των εγγύτερων γειτόνων. Αξίζει να σημειωθεί πως όταν το $k = 1$, ο αλγόριθμος είναι γνωστός ως κατηγοριοποιητής κοντινότερου γείτονα(1-NN).



Εικόνα 5: “Κατηγοριοποίηση αντικειμένου με k-NN για *k*=3 και *k*=5”

Στην εικόνα 5, υπάρχει ένα σύνολο δεδομένων με δύο κλάσεις, τα τετράγωνα και τους κύκλους και ένα νέο αντικείμενο, το Q, το οποίο χρειάζεται να κατηγοριοποιηθεί σε μία από τις δύο κλάσεις. Εάν το *k* πάρει την τιμή 3, τότε το Q κατηγοριοποιείτε στην κλάση των κύκλων επειδή δύο από τους τρεις κοντινότερους γείτονές του είναι κύκλοι. Από την άλλη πλευρά, εάν το k πάρει την τιμή 5, η ετικέτα κλάσης του Q θα είναι τα τετράγωνα καθώς τρεις από τους πέντε γείτονές του είναι τετράγωνα.

Η απόδοση της κατηγοριοποίησης εξαρτάται σε μεγάλο βαθμό από την επιλογή της τιμής του *k*. Η τιμή του *k* που πετυχαίνει μεγαλύτερη ακρίβεια κατηγοριοποίησης εξαρτάται από το σύνολο δεδομένων που θα χρησιμοποιηθεί και ο καθορισμός του προσδιορίζεται μέσω δαπανηρών εργασιών προεπεξεργασίας. Παρόλα αυτά, ο καθορισμός του *k* δεν ακολουθάει κάποιον γενικό κανόνα και το “καλύτερο” *k* μπορεί να είναι εντελώς διαφορετικό για διαφορετικά σύνολα δεδομένων. Μεγαλύτερες τιμές του k είναι κατάλληλες για σύνολα δεδομένων με θόρυβο από τη στιγμή που γειτονιές που εξετάζονται είναι μεγαλύτερες. Ωστόσο δεν καθορίζουν με ακρίβεια τα όρια ανάμεσα σε διακριτές κλάσεις. Αντίθετα οι πιο μικρές τιμές της παραμέτρου κάνει τον κατηγοριοποιητή πιο ευαίσθητο στο θόρυβο. Άρα, σε περιπτώσεις που το σύνολο δεδομένων εκπαίδευσης που περιλαμβάνει θόρυβο, η κατηγοριοποίηση θα είναι λιγότερο ακριβής.

Σε περιπτώσεις που το πρόβλημα είναι δυαδικής κατηγοριοποίησης, δηλαδή σύνολα δεδομένων με δύο κλάσεις, το *k* πρέπει να έχει περιττή τιμή με σκοπό να αποφευχθούν οι ισοπαλίες. Εάν το πρόβλημα δεν είναι δυαδικό, το k μπορεί να έχει οποιαδήποτε τιμή και αν υπάρξουν περιπτώσεις ισοπαλίας το πρόβλημα μπορεί να επιλυθεί είτε επιλέγοντας μια τυχαία κλάση ή την κλάση του κοντινότερου γείτονα.

Ένα άλλο σημαντικό θέμα το οποίο πρέπει να αναφερθεί, είναι η επιλογή της μετρικής που θα χρησιμοποιηθεί για τον υπολογισμό των αποστάσεων μεταξύ των αντικειμένων. Με αυτή την απόφαση πρέπει να ληφθεί στα υπόψιν ο τύπος των δεδομένων των χαρακτηριστικών του συνόλου δεδομένων. Εάν τα χαρακτηριστικά είναι πραγματικοί ή/και ακέραιοι αριθμοί, η Ευκλείδεια απόσταση είναι αυτή που χρησιμοποιείται πιο συχνά ως μετρική απόστασης.

$$d(p,q)=d(q,p)=\sqrt{(q1-p1)^2+(q2-p2)^2+...+(qn-pn)^2}=\sqrt{\sum_{i=1}^{n}(qi-pi)^2} \quad (1.4.1)$$

Μια εναλλακτική της Ευκλείδειας απόστασης, είναι η απόσταση Manhattan όπου η διαφορά μεταξύ των τιμών των χαρακτηριστικών δεν υψώνεται στο τετράγωνο αλλά αθροίζεται.

d1(p,q)= $\|p-q\|1=\sum_{i=1}^{n}|pi-qi|$ (1.4.2)

Τα διαφορετικά χαρακτηριστικά μετρώνται σε διαφορετικές κλίμακες τιμών. Η επίδραση κάποιων χαρακτηριστικών στην απόσταση δύο περιπτώσεων μπορεί να επηρεαστεί από την ύπαρξη χαρακτηριστικών με μεγάλες κλίμακες τιμών. Επομένως, αποτελεί καλή τεχνική η κανονικοποίηση των τιμών όλων των χαρακτηριστικών στο διάστημα από 0 έως 1. Για κάθε χαρακτηριστικό ενός παραδείγματος, αφαιρούμε από την τιμή του την ελάχιστη τιμή για αυτό το χαρακτηριστικό και μετά διαιρούμε με το εύρος τιμών για το συγκεκριμένο χαρακτηριστικό. Αυτές οι αποστάσεις είναι κατάλληλες μόνο για αριθμητικά δεδομένα και δεν μπορούν να χρησιμοποιηθούν για κατηγορικά δεδομένα. Ωστόσο, υπάρχουν μέτρα ομοιότητας και για κατηγορικά δεδομένα.

Σε περιπτώσεις που τα χαρακτηριστικά δεν είναι αριθμητικά αλλά διακριτά τα οποία λαμβάνουν συμβολικές και όχι αριθμητικές τιμές, η μέθοδος που συνηθίζεται είναι να ορίζεται ως 1 η απόσταση μεταξύ δύο τιμών που δεν ταυτίζονται και με 0 η απόσταση όταν οι τιμές ταυτίζονται. Εδώ δεν απαιτείται κανονικοποίηση καθώς οι μόνες τιμές που χρησιμοποιούνται είναι 0 και 1.

Εάν στο πρόβλημα που καλούμαστε να αντιμετωπίσουμε οι τιμές είναι ελλιπείς και πρόκειται για διακριτά χαρακτηριστικά, υποθέτουμε ότι η τιμή που λείπει έχει τη μέγιστη διαφορά με κάθε άλλη τιμή ενός χαρακτηριστικού. Έτσι είτε απουσιάζει η μια είτε και οι δύο τιμές ή αν οι τιμές είναι διαφορετικές, η απόσταση μεταξύ τους είναι ίση με 1. Η διαφορά ισούται με 0 στην περίπτωση που υπάρχουν και οι δύο τιμές και είναι όμοιες. Εάν τα χαρακτηριστικά είναι αριθμητικά η διαφορά μεταξύ δύο τιμών που λείπουν είναι επίσης 1.

1.4.1 Προβλήματα και Αδυναμίες του Κατηγοριοποιητή k-NN

Ο k-NN υπολογίζει την απόσταση μεταξύ δύο περιπτώσεων με βάση όλα τα χαρακτηριστικά, πράγμα το οποίο μπορεί να επηρεάσει την ακρίβεια του αλγορίθμου όταν ο αριθμός των χαρακτηριστικών που δεν επηρεάζουν την εξαρτημένη μεταβλητή είναι μεγάλος. Ας υποθέσουμε για παράδειγμα ένα πρόβλημα με 30 χαρακτηριστικά από τα οποία σημαντικά για τη πρόβλεψη μιας νέας περίπτωσης είναι μόνο τα 5. Εάν μία περίπτωση και ένα αντικείμενο έχουν ίδιες τιμές σε αυτά τα 5 χαρακτηριστικά τότε η Ευκλείδεια απόσταση αυτών μπορεί να είναι πολύ μεγάλη. Το πρόβλημα αυτό της εκθετικής αύξησης της δυσκολίας μάθησης σχετικά με την αύξηση του αριθμού των χαρακτηριστικών ονομάζεται κατάρα των διαστάσεων(curse of dimensionality) και για την αντιμετώπισή του έχουν γίνει προτάσεις που αφορούν είτε την επιλογή ενός υποσυνόλου των χαρακτηριστικών είτε τη στάθμιση των χαρακτηριστικών. Και στις δύο περιπτώσεις, για την πρόβλεψη της εξαρτημένης μεταβλητής απαιτείται εύρεση των σημαντικότερων χαρακτηριστικών. Στην περίπτωση όπου κάποια χαρακτηριστικά είναι πιο σημαντικά για κάποια συγκεκριμένη κλάση και κάποια άλλα λιγότερο σημαντικά, αποτελεί καλή επιλογή, ο διαχωρισμός των χαρακτηριστικών στα πιο σημαντικά που αφορούν κάθε διακριτή τιμή της κλάσης. Η συγκεκριμένη τεχνική στην περιοχή της Μηχανικής Μάθησης ονομάζεται Επιλογή Χαρακτηριστικών(Feature Selection).

Ενώ ο k-NN θεωρείται μια αρκετά αποτελεσματική μέθοδος έχει κάποιες αδυναμίες που μπορούν να επηρεάσουν τη χρήση του και μια από αυτές είναι το υψηλό υπολογιστικό κόστος του. Ο κατηγοριοποιητής πρέπει να υπολογίσει όλες τις αποστάσεις μεταξύ των αντικειμένων που δεν έχουν κατηγοριοποιηθεί και των αντικειμένων που βρίσκονται στο σύνολο δεδομένων εκπαίδευσης. Σε περιπτώσεις που το σύνολο δεδομένων είναι μεγάλο, το μειονέκτημα αυτό κάνει την διαδικασία χρονοβόρα και σχεδόν απαγορευτική. Για παράδειγμα, εάν έχουμε αποθηκεύσει 100.000 αντικείμενα στο σύνολο εκπαίδευσης και θέλουμε να κατηγοριοποιήσουμε περίπου 50.000 αντικείμενα με τον k-NN, αυτό σημαίνει ότι πρέπει να υπολογιστούν πέντε δισεκατομμύρια διαστάσεις. Παρόλο που τα σημερινά συστήματα διαθέτουν πανίσχυρους επεξεργαστές αυτοί οι υπολογισμοί συνεχίζουν να είναι αρκετά χρονοβόροι και σε περιπτώσεις που το περιβάλλον είναι περιοριστικό σε σχέση με το χρόνο αυτές οι περιπτώσεις είναι μη αποδεκτές. Ακόμα, εκτός από το μέγεθος του συνόλου δεδομένων εκπαίδευσης, το υπολογιστικό κόστος εξαρτάται και από τη διάσταση των δεδομένων, καθώς όσο υψηλότερη είναι τόσο περισσότεροι υπολογισμοί πρέπει να γίνουν.

Οι μεγάλες απαιτήσεις όσο αφορά το χώρο που καταλαμβάνει το σύνολο δεδομένων εκπαίδευσης είναι μια ακόμη αδυναμία του κατηγοριοποιητή k-NN. Αντίθετα με τους eager αλγόριθμος κατηγοριοποίησης οι οποίοι μπορούν να αφαιρέσουν τα δεδομένα εκπαίδευσης μετά την κατασκευή του μοντέλου κατηγοριοποίησης, ο k-NN χρειάζεται τα δεδομένα εκπαίδευσης να είναι πάντα

διαθέσιμα. Άρα, υπολογιστικά συστήματα εξοπλισμένα με αρκετή κύρια μνήμη για την αποθήκευση των δεδομένων εκπαίδευσης πρέπει να χρησιμοποιηθούν για την εκτέλεση του k-NN.

Τέλος, ένα ακόμη μειονέκτημα του κατηγοριοποιητή k-NN, όπως και άλλων μεθόδων κατηγοριοποίησης, είναι η ευαισθησία στο θόρυβο. Πιο συγκεκριμένα, η ακρίβεια της κατηγοριοποίησης εξαρτάται σε μεγάλο βαθμό από την ποιότητα των δεδομένων εκπαίδευσης. Εάν τα δεδομένα περιλαμβάνουν θόρυβο ή δεν έχουν ετικέτα κλάσης το αποτέλεσμα της κατηγοριοποίησης είναι λιγότερο ακριβές. Όπως αναφέρθηκε και παραπάνω, η χρήση υψηλών τιμών του *k* οδηγεί στην εξέταση μεγαλύτερης γειτονιάς και μπορεί εν μέρη να καλύψει το παραπάνω μειονέκτημα. Όμως, αυτό σημαίνει μεγαλύτερο αριθμό εκτελέσεων για το καθορισμό της κατάλληλης τιμής του k και ο θόρυβος κατανέμεται ομοιόμορφα στο σύνολο εκπαίδευσης.

Συνοψίζοντας, η επιλογή του καταλληλότερου k συνδέεται άμεσα με τον αριθμό των αντικειμένων που βρίσκονται στο σύνολο εκπαίδευσης. Όσο μεγαλύτερο είναι το σύνολο τόσο περισσότεροι γείτονες πρέπει να επιλέγονται ενώ αντίθετα όταν το σύνολο εκπαίδευσης είναι μικρό, κάτι τέτοιο είναι επικίνδυνο. Μεγαλύτερη τιμή του k σημαίνει συμβιβασμός με την κλάση που είναι πλειοψηφία στο σύνολο δεδομένων και μικρή τιμή του k σημαίνει μικρότερο σύνολο για εξέταση άρα ο αλγόριθμος είναι και πιο επιρρεπής στο θόρυβο.

Η κατηγοριοποίηση με βάση τους εγγύτερους γείτονες είναι μια απλή μέθοδος με καλή απόδοση. Από τα δυνατά της σημεία είναι και η κατηγοριοποίηση πολύπλοκων συναρτήσεων λόγω του ότι επικεντρώνεται σε πιο απλές περιοχές του χώρου των περιπτώσεων.

1.5 Μείωση Δεδομένων (Data Reduction)

Όταν δουλεύουμε με δεδομένα κάποιες φορές μπορεί να συναντάμε εμπόδια τα οποία μας δυσκολεύουν και άλλες φορές η αναζήτηση αυτή μπορεί να μας οδηγήσει σε αποτελέσματα αρκετά ενθαρρυντικά. Το μέγεθος και η πολυπλοκότητα ενός συνόλου δεδομένων μπορεί να κάνει την εργασία μας πιο δύσκολη αλλά όσο μεγαλύτερο είναι το σύνολο δεδομένων τόσο πιο πλούσιο θα είναι σε πληροφορίες.

Ένας τρόπος ο οποίος είναι αρκετά αποτελεσματικός για την απόδοση των αλγορίθμων κατηγοριοποίησης είναι η μείωση των δεδομένων εκπαίδευσης. Οι τεχνικές μείωσης του όγκου των δεδομένων στοχεύουν στην μείωση του αρχικού συνόλου δεδομένων και στην εξαγωγή ενός νέου συνόλου το οποίο θα αντιπροσωπεύει όσο το δυνατόν καλύτερα το αρχικό. Το νέο σύνολο που προκύπτει ονομάζεται συμπυκνωμένο σύνολο (condensing set). Στην ουσία, αυτό που προσπαθούμε να καταφέρουμε με τη μείωση των δεδομένων είναι να μειώσουμε το υπολογιστικό

κόστος αναζήτησης διατηρώντας την ακρίβεια της κατηγοριοποίησης σε όσο το δυνατόν υψηλότερα επίπεδα. Ένα άλλο πλεονέκτημα της μείωσης δεδομένων είναι ότι περιορίζονται σημαντικά οι απαιτήσεις μνήμης για την αποθήκευση των δεδομένων που έχει σαν αποτέλεσμα και σαν όφελος την εκτέλεση εργασιών κατηγοριοποίησης από συστήματα με μικρότερη μνήμη χωρίς να απαιτείται η χρήση μεγάλων και ισχυρών υπολογιστών.

Οι κατηγορίες στις οποίες μπορούν να χωριστούν οι τεχνικές μείωσης δεδομένων είναι δύο: α) Αλγόριθμοι επιλογής δεδομένων (data selection algorithms) β) Αλγόριθμοι σύνοψης δεδομένων (data abstraction algorithms). Και στις δύο κατηγορίες ο σκοπός είναι η παραγωγή ενός νέου συμπυκνωμένου συνόλου δεδομένων που θα αντιπροσωπεύει το αρχικό, μεγαλύτερο σύνολο. Οι αλγόριθμοι που ανήκουν στη κλάση επιλογής δεδομένων, επιλέγουν ως αντιπροσώπους κάποια στιγμιότυπα του αρχικού συνόλου δεδομένων τα οποία αποθηκεύονται στο νέο συμπυκνωμένο σύνολο. Από την άλλη πλευρά, οι αλγόριθμοι που ανήκουν στην κλάση σύνοψης δεδομένων, δημιουργούν νέους αντιπροσώπους και τους τοποθετούν στο νέο σύνολο συνοψίζοντας όμοια στιγμιότυπα του αρχικού συνόλου. Αυτό το πετυχαίνουν είτε βρίσκοντας το μέσο όρο των χαρακτηριστικών είτε με την εφαρμογή αλγορίθμων συσταδοποίησης.

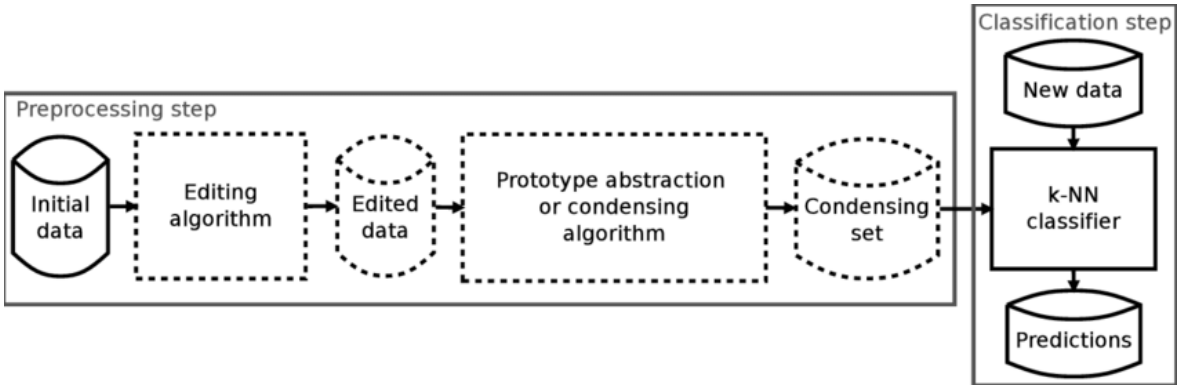
Οι μέθοδοι μείωσης δεδομένων μπορούν να κατηγοριοποιηθούν επίσης ως αυξητικοί και μη αυξητικοί αλγόριθμοι. Οι πρώτοι, αρχικά έχουν ένα μικρό όγκο δεδομένων ο οποίος αυξάνεται σταδιακά μέχρι να ικανοποιηθούν τα κριτήρια που έχει θέσει ο ίδιος ο αλγόριθμος. Στους μη αυξητικούς αλγόριθμους, όλα τα δεδομένα εκπαίδευσης λαμβάνονται υπόψη στην αρχή και μειώνονται σταδιακά μέχρι να ικανοποιηθούν τα κριτήρια.

Αυτό το οποίο κάνουν οι περισσότερες τεχνικές μείωσης των δεδομένων είναι να βρουν τα σύνορα μεταξύ των διαφορετικών κλάσεων στα δεδομένα. Τα περισσότερα αντικείμενα που τοποθετούνται στο νέο σύνολο είναι αυτά που βρίσκονται κοντά στα σύνορα των κλάσεων. Η κίνηση αυτή βασίζεται στην ιδέα ότι οι προβλέψεις της κατηγοριοποίησης δεν επηρεάζονται από τα δεδομένα που δεν ορίζουν τα όρια απόφασης για αυτό στο νέο σύνολο τοποθετούνται μόνο αυτά τα δεδομένα που ορίζουν τα όρια.

Επίσης, μια υποκλάση των αλγόριθμων επιλογής δεν έχουν ως στόχο την μείωση του κόστους αλλά την αύξηση της ακρίβειας κατηγοριοποίησης και ονομάζονται αλγόριθμοι επεξεργασίας δεδομένων (editing algorithms). Ο τρόπος με τον οποίο οι αλγόριθμοι αυτοί προσπαθούν να αυξήσουν την ακρίβεια είναι με την απομάκρυνση των δεδομένων με θόρυβο και με την εξομάλυνση των ορίων απόφασης των κλάσεων.

Σημαντικό ρόλο στο πόσο θα μειωθεί το σύνολο δεδομένων εκπαίδευσης παίζει το πλήθος των κλάσεων που υπάρχει σε αυτό και το πόσο θόρυβος υπάρχει στα δεδομένα. Στο νέο σύνολο,

αποθηκεύονται περισσότερα δεδομένα όσο περισσότερο θόρυβος υπάρχει ενώ τα όρια απόφασης είναι περισσότερα όσο περισσότερες είναι και οι κλάσεις. Άρα είναι σημαντική προϋπόθεση να έχει γίνει σωστός “καθαρισμός” των δεδομένων πριν εφαρμοστεί μια τεχνική μείωσης του όγκου σε αυτά.



Εικόνα 6: “Διαδικασία προεπεξεργασίας δεδομένων και εφαρμογής αλγορίθμου κατηγοριοποίησης k-NN”

Όπως φαίνεται και στην εικόνα 7 η διαδικασία κατηγοριοποίησης περιλαμβάνει μια ακολουθία ενεργειών. Σε πρώτο στάδιο, εφαρμόζεται στα δεδομένα ένας αλγόριθμος επεξεργασίας που έχει ως σκοπό την απομάκρυνση του θορύβου και το σύνολο που προκύπτει καλείται επεξεργασμένο σύνολο (Edited data). Στη συνέχεια εφαρμόζεται μια τεχνική μείωσης δεδομένων και προκύπτει το συμπυκνωμένο σύνολο (Condensing set). Ο κατηγοριοποιητής k-NN για κάθε αντικείμενο εφαρμόζεται σε αυτό το συμπυκνωμένο σύνολο και κάνει τη πρόβλεψή του. Τέλος, για τα πειράματα της συγκεκριμένης πτυχιακής εργασίας επικεντρωθήκαμε περισσότερο στους αλγόριθμους condensing και abstraction και όχι τόσο στους editing.

1.6 Κίνητρο και Συνεισφορά

Ο αριθμός των εφαρμογών των οποίων τα σύνολα δεδομένων είναι μεγάλα και τα αντικείμενα τους ανήκουν σε πάνω από μια κατηγορίες αυξάνεται συνεχώς. Τέτοια παραδείγματα αποτελούν πολλά κείμενα τα οποία ανήκουν σε πάνω από ένα είδος συγγραφής, είτε είναι λογοτεχνικά είτε επιστημονικά. Επίσης, πολλές φωτογραφίες τις περισσότερες φορές απεικονίζουν κάτι το οποίο μπορεί να περιγραφεί με πολλούς τρόπους και να κατηγοριοποιηθεί σε διαφορετικές κλάσεις καθώς ακόμα και πολλές ταινίες οι οποίες μπορούν να κατηγοριοποιηθούν σε πάνω από μια κλάσεις,

όπως το ότι μια ταινία μπορεί να χαρακτηριστεί ως επιστημονικής φαντασίας και δράσης ταυτόρονα.

Η κατηγοριοποίηση πολλαπλών ετικετών απασχολεί την κοινότητα της εξόρυξης δεδομένων και της μηχανικής μάθησης. Παρόλο που ο βασικός στόχος και σε αυτού του είδους τα δεδομένα είναι η επίτευξη υψηλών ποσοστών ακρίβειας κατηγοριοποίησης και ή όσο το δυνατό ταχύτερη εκτέλεση της κατηγοριοποίησης, οι διάφοροι αλγόριθμοι μείωσης τους πληθυσμού των δεδομένων δεν έχουν εφαρμοστεί σε τέτοιου είδους σύνολα δεδομένων. Η παρατήρηση αυτή, αποτελεί το κίνητρο για την εκπόνησης της παρούσας εργασίας.

Στα πλαίσια εκπόνησης της εργασίας, εφαρμόζονται γνωστοί αλγόριθμοι μείωσης των δεδομένων εκπαίδευσης σε τέτοια σύνολα δεδομένων πολλαπλών ετικετών. Το αποτέλεσμα τους, είναι ένα μικρό, συμπυκνωμένο σύνολο πολλαπλών ετικετών που αντιπροσωπεύει το αρχικό. Σε αυτά τα συμπυκνωμένα σύνολα που προκύπτουν, εφαρμόζεται ο αλγόριθμος κατηγοριοποίησης κ εγγύτερων γειτόνων με τέσσερις διαφορετικές τιμές στην παράμετρο κ. Οι μετρήσεις που εκτιμώνται μέσω των πειραμάτων είναι η ακρίβεια κατηγοριοποίησης, ο λόγος συμπύκνωσης (reduction rate) καθώς και το κόστος προεπεξεργασίας (pre-processing cost)

1.7 Οργάνωση της Πτυχιακής

Στη συνέχεια της συγκεκριμένης πτυχιακής εργασίας και πιο συγκεκριμένα στο κεφάλαιο 2, θα γίνει παρουσίαση των αλγόριθμων μείωσης του πληθυσμού των δεδομένων που χρησιμοποιήθηκαν. Στο κεφάλαιο 3, θα αναφέρουμε το τρόπο με τον οποίο έγινε η διαχείριση των δεδομένων πολλαπλών ετικετών και έπειτα στο κεφάλαιο 4 θα γίνει παρουσίαση της πειραματικής μελέτης. Τα σύνολα δεδομένων που χρησιμοποιήσαμε για την εφαρμογή των παραπάνω αλγορίθμων θα παρουσιαστούν όπως και οι παράμετροι και τα αποτελέσματα των πειραμάτων, καταλήγοντας στο σχολιασμό των αποτελεσμάτων.

Επίλογος

Σε αυτή την εισαγωγική ενότητα περιγράψαμε βασικές έννοιες σχετικά με την κατηγοριοποίηση, την αποτίμηση της απόδοσης των κατηγοριοποιητών καθώς και τα σύνολα δεδομένων πολλαπλών ετικετών. Αναφερθήκαμε στον κατηγοριοποιητή k-NN και σε κάποια γενικά πράγματα για τις τεχνικές μείωσης του πληθυσμού των δεδομένων. Στο επόμενο κεφάλαιο θα δούμε πιο αναλυτικά κάποια πράγματα για τους αλγόριθμους μείωσης του πληθυσμού που χρησιμοποιήθηκαν στην συγκεκριμένη εργασία.

ΚΕΦΑΛΑΙΟ 2: ΑΛΓΟΡΙΘΜΟΙ ΜΕΙΩΣΗΣ ΤΟΥ ΠΛΗΘΥΣΜΟΥ ΤΩΝ ΔΕΔΟΜΕΝΩΝ

ΕΙΣΑΓΩΓΗ

Οι κατηγορίες των τεχνικών μείωσης του πληθυσμού των δεδομένων είναι δύο: (α) επιλογή πρωτοτύπων (prototype selection) και (β) αφαίρεση πρωτοτύπων (prototype abstraction). Οι αλγόριθμοι επιλογής πρωτοτύπων χωρίζονται σε αλγόριθμους συμπύκνωσης(condensing algorithms) και αλγόριθμους επεξεργασίας (editing algorithms). Οι αδυναμίες του κατηγοριοποιητή k-NN μπορούν να αντιμετωπιστούν με τους αλγόριθμους συμπύκνωσης και σύνοψης πρωτοτύπων λόγω του υψηλού υπολογιστικού κόστους της κατηγοριοποίησης και των απαιτήσεων αποθήκευσης. Αυτό το πετυχαίνουν με την δημιουργία ενός νέου μικρότερου αντιπροσωπευτικού συνόλου δεδομένων το οποίο καλείται συμπυκνωμένο σύνολο(condensing set) και περιέχει μόνο τα βασικά αντικείμενα. Εφαρμόζοντας τον κατηγοριοποιητή k-NN χρησιμοποιώντας το συμπυκνωμένο σύνολο, τα πλεονεκτήματα που προκύπτουν είναι ότι μειώνεται το υπολογιστικό κόστος και οι απαιτήσεις μνήμης ενώ το ποσοστό της ακρίβειας παραμένει υψηλό.

Οι τεχνικές μείωσης του πληθυσμού δεδομένων μπορούν να αξιολογηθούν με τη χρήση τριών κριτηρίων. Το πρώτο είναι ο βαθμός μείωσης που πετυχαίνουν και υποδεικνύει το πόσο μικρότερο θα είναι το μέγεθος του συμπυκνωμένου συνόλου σε σχέση με το αρχικό σύνολο. Όσο μεγαλύτερος είναι ο βαθμός μείωσης τόσο πιο γρήγορα θα εκτελεστεί ο k-NN. Ένα άλλο σημαντικό κριτήριο είναι η ακρίβεια της κατηγοριοποίησης που πετυχαίνει ο k-NN όταν εκτελείται στο συμπυκνωμένο σύνολο. Το τρίτο κριτήριο είναι το υπολογιστικό κόστος της προεπεξεργαασίας το οποίο είναι και από τα πιο βασικά για τη δημιουργία του συμπυκνωμένου συνόλου.

Παρόλο που η επεξεργασία έχει διαφορετικό στόχο από τις τεχνικές μείωσης του πληθυσμού, μπορεί να χρησιμοποιηθεί για την βελτίωση της απόδοσης είτε αυξάνοντας το βαθμό μείωσης(reduction rate) είτε τα επίπεδα της ακρίβειας. Ο βαθμός μείωσης πολλών αλγορίθμων, είτε σύνοψης πρωτοτύπων είτε συμπύκνωσης, εξαρτάται από τα επίπεδα θορύβου που βρίσκονται στο σύνολο δεδομένων εκπαίδευσης. Υψηλά επίπεδα θορύβου στο σύνολο εκπαίδευσης εμποδίζουν τους αλγόριθμους να πετύχουν υψηλά ποσοστά βαθμού μείωσης.

Σκοπός της διαδικασίας προεπεξεργασίας της μείωσης του πληθυσμού των δεδομένων είναι να φτιάξει ένα συμπυκνωμένο σύνολο χωρίς θόρυβο κρατώντας παράλληλα ή δημιουργώντας για κάθε κλάση έναν αριθμό πρωτοτύπων που είναι απαραίτητα για την κατηγοριοποίηση k-NN.

2.1 Ο Αλγόριθμος Condensing Nearest Neighbour – CNN

Ο συγκεκριμένος αλγόριθμος[13] χρονολογείται από το 1968 και αποτελεί τον πρώτο αλγόριθμο συμπύκνωσης. Αντίθετα από τους μεταγενέστερους αλγόριθμους, ο CNN δεν έχει ως στόχο την αφαίρεση σημείων που βρίσκονται μακρυά από τα όρια απόφασης. Αντί αυτού, στοχεύει στην αφαίρεση των σημείων τα οποία δεν απαιτούνται για την σωστή κατηγοριοποίηση του συνόλου δεδομένων εκπαίδευσης.

Βασική ιδέα του αλγορίθμου είναι ότι τα στιγμιότυπα που βρίσκονται πιο κοντά στη περιοχή μιας κλάσης και είναι πιο απομακρυσμένα από τα σύνορά της, είναι αχρείαστα κατά την διάρκεια της κατηγοριοποίησης και μπορούν να αφαιρεθούν χωρίς να υπάρχει κάποια επίδραση στην ακρίβεια της κατηγοριοποίησης. Αυτό έχει ως αποτέλεσμα ο αλγόριθμος να τοποθετεί στο συμπυκνωμένο σύνολο μόνο αυτά τα αντικείμενα που βρίσκονται πιο μακρυά από τα σύνορα μιας κλάσης.

Αρχικά, ο αλγόριθμος τοποθετεί ένα αντικείμενο από το σύνολο εκπαίδευσης στο συμπυκνωμένο σύνολο και εφαρμόζει στη συνέχεια τον αλγόριθμο του εγγύτερου γείτονα με τιμή του k=1(1-NN) όπου και κατηγοριοποιεί τα αντικείμενα του συνόλου εκπαίδευσης ελέγχοντας τα αντικείμενα του συμπυκνωμένου συνόλου. Εάν ένα αντικείμενο έχει κατηγοριοποιηθεί λάθος, μεταφέρεται στο συμπυκνωμένο σύνολο από το σύνολο εκπαίδευσης. Η διαδικασία αυτή συνεχίζεται μέχρι να μην υπάρχουν άλλες κινήσεις οι οποίες να μπορούν να γίνουν από το σύνολο εκπαίδευσης στο συμπυκνωμένο σύνολο. Αυτή η διαδικασία εξασφαλίζει ότι το περιεχόμενο του συνόλου εκπαίδευσης είναι σωστά κατηγοριοποιημένο ενώ τα αντικείμενα που περισσεύουν στο σύνολο εκπαίδευσης, απορρίπτονται.

Input: TS **Output**: CS

1: $CS \leftarrow \varnothing$
2: pick an item of TS and move it to CS
3: **repeat**
4: $stop \leftarrow TRUE$
5: **for each** $x \in TS$ **do**
6: $NN \leftarrow$ Nearest Neighbour of x in CS
7: **if** $NN_{class} \neq x_{class}$ **then**
8: $CS \leftarrow CS \cup \{x\}$
9: $TS \leftarrow TS - \{x\}$
10: $stop \leftarrow FALSE$
11: **end if**
12: **end for**
13: **until** $stop == TRUE$ {no move during a pass of TS}
14: discard TS
15: **return** CS

Εικόνα 7: "Τμήμα ψευδοκώδικα για τον αλγόριθμο CNN"

Ένα πλεονέκτημα του αλγόριθμου CNN είναι ότι δεν χρειάζεται κάποια παράμετρο. Ο αριθμός των πρωτοτύπων αποφασίζεται αυτόματα χωρίς να απαιτείται κάποια παράμετρος από το χρήστη. Επίσης, πολύ σημαντικό πλεονέκτημα του αλγόριθμου, είναι ότι μέσω των πολλαπλών περασμάτων που κάνει στα δεδομένα εγγυάται ότι τα αντικείμενα εκπαίδευσης μπορούν να κατηγοριοποιηθούν σωστά με την εκτέλεση του αλγόριθμου εγγύτερου γείτονα με παράμετρο k=1 (1-NN).

Δυστυχώς ο αλγόριθμος είναι ευαίσθητος στη σειρά παρουσίασης των αντικειμένων εκπαίδευσης και η απόδοση του κατηγοριοποιητή μπορεί να διαφέρει εάν δώσουμε το ίδιο σύνολο εκπαίδευσης αλλά με διαφορετική σειρά.

Επίσης, μια αδυναμία του αλγόριθμου είναι η ευαισθησία του στο θόρυβο με αποτέλεσμα να επιλέγει λανθασμένα τα θορυβώδη στιγμιότυπα μιας γειτονιάς. Επομένως, ο θόρυβος επηρεάζει το ποσοστό μείωσης.

2.2 Ο Αλγόριθμος IB2

Ο αλγόριθμος IB2[13] ανήκει στην κλάση των Instance-Based Learning (IBL) αλγορίθμων. Είναι ένας αυξητικός αλγόριθμος συμπύκνωσης ενός περάσματος και βασίζεται στον CNN. Αντίθετα από τον CNN και άλλων αλγορίθμων συμπύκνωσης και σύνοψης πρωτοτύπων, ο IB2 μπορεί να χτίσει δυναμικά το συμπυκνωμένο σύνολο. Με άλλα λόγια, είναι ένας αυξητικός αλγόριθμος. Μπορεί να λάβει υπόψη νέα αντικείμενα εκπαίδευσης μετά την κατασκευή του συμπυκνωμένου συνόλου χωρίς να χρειάζεται να αφαιρέσει τα παλιά αντικείμενα εκπαίδευσης. Έτσι, το υπάρχων συμπυκνωμένο σύνολο μπορεί να βελτιωθεί καθώς, κάθε νέο αντικείμενο εκπαίδευσης μπορεί αφού πρώτα εξεταστεί, να τοποθετηθεί ή όχι στο συμπυκνωμένο σύνολο.

Με τρόπο παρόμοιο με αυτό του αλγόριθμου CNN, ο IB2 δεν απαιτεί κάποια παράμετρο και το συμπυκνωμένο σύνολο που προκύπτει εξαρτάται σε μεγάλο βαθμό από τη σειρά των αντικειμένων στο σύνολο εκπαίδευσης. Αντίθετα όμως από τον CNN, ο IB2 δεν εγγυάται ότι όλα τα αντικείμενα που απορρίπτονται μπορούν να κατηγοριοποιηθούν σωστά στο τελικό συμπυκνωμένο σύνολο. Παρόλα αυτά από τη στιγμή που πρόκειται για έναν αλγόριθμο ενός περάσματος, είναι αρκετά πιο γρήγορος κατά την εκτέλεσή του και το κόστος της προεπεξεργασίας είναι σημαντικά μικρότερο.

Ο αλγόριθμος IB2, αρχικά δημιουργεί το συμπυκνωμένο σύνολο σταδιακά. Κάθε αντικείμενο που ανήκει στο σύνολο εκπαίδευσης, κατηγοριοποιείται με τη χρήση του κατηγοριοποιητή εγγύτερου γείτονα με παράμετρο k=1(1-NN) στο συπυκνωμένο σύνολο. Εάν το αντικείμενο έχει κατηγοριοποιηθεί σωστά αφαιρείται και σε αντίθετη περίπτωση μεταφέρεται στο συμπυκνωμένο σύνολο. Ακόμα, ένα πλεονέκτημα του αλγόριθμου IB2 είναι ότι μπορεί να διαχειριστεί νέες κλάσεις αντικειμένων, πράγμα που τον κάνει κατάλληλο για εφαρμογές όπου νέα αντικείμενα εκπαίδευσης μπορεί να προστεθούν σταδιακά.

2.3 Ο Αλγόριθμος AIB2

Μια έκδοση σύνοψης πρωτοτύπων που βασίζεται στον IB2, αποτελεί ο αλγόριθμος AIB2[13]. Με επιρροές από τους IB2 και CNN, ο AIB2 προτείνει, ότι τα αντικείμενα που βρίσκονται στην εσωτερική περιοχή δεδομένων μιας κλάσης μπορούν να αφαιρεθούν χωρίς να χαθεί η ακρίβεια. Επομένως, ο AIB2 είναι ένας μη παραμετρικός, γρήγορος και μονής κατεύθυνσης αλγόριθμος. Είναι επίσης κατάλληλος για δυναμικά περιβάλλοντα και μπορεί να χειριστεί νέες ετικέτες κλάσεων. Όπως ο IB2, μπορεί να εφαρμοστεί σε πολύ μεγάλα σύνολα δεδομένων τα οποία δεν μπορούν να αποθηκευτούν στην κύρια μνήμη ή σε συσκευές με περιορισμένη κύρια μνήμη χωρίς να μεταφέρουν δεδομένα σε έναν εξυπηρέτη στο διαδίκτυο για επεξεργασία, που είναι μια διαδικασία χρονοβόρα και με υψηλό κόστος.

Η κύρια διαφορά ανάμεσα στον AIB2 και στον IB2 είναι πως τα αντικείμενα τα οποία έχουν ταξινομηθεί σωστά από τον 1-NN δεν παραμελούνται. Στην πραγματικότητα, συνεισφέρουν στην

κατασκευή του συμπυκνωμένου συνόλου με την επανατοποθέτηση των κοντινότερων πρωτοτύπων στο συμπυκνωμένο σύνολο. Η κύρια ιδέα πίσω από τον αλγόριθμο AIB2, είναι ότι τα πρωτότυπα πρέπει να βρίσκονται στο κέντρο της περιοχής δεδομένων που αντιπροσωπεύουν. Για να το πετύχει αυτό ο AIB2, υιοθετεί το σενάριο του βάρους των πρωτοτύπων. Κάθε πρωτότυπο έχει μια τιμή βάρους ως ένα επιπλέον χαρακτηριστικό το οποίο δείχνει τον αριθμό των αντικειμένων εκπαίδευσης που αντιπροσωπεύει. Αυτή η τιμή βάρους χρησιμοποιείται για την ενημέρωση των χαρακτηριστικών των πρωτοτύπων στον πολυδιάστατο χώρο.

Το παρακάτω τμήμα ψευδοκώδικα*(Εικόνα 11)*, παρουσιάζει την λειτουργία του AIB2. Αρχικά, το συμπυκνωμένο σύνολο(Condensing Set , CS) αποτελείται από τυχαία αντικείμενα του συνόλου εκπαίδευσης (Training Set , TS) του οποίου το βάρος αρχικοποιείται σε 1. Για κάθε αντικείμενο *x* του *TS*, ο αλγόριθμος ψάχνει στο συμπυκνωμένο σύνολο(CS) και ανακτά το κοντινότερο πρωτότυπο *NN*. Εάν το x είναι λάθος ταξινομημένο, τοποθετείται στο συμπυκνωμένο σύνολο και το βάρος του αρχικοποιείται σε 1. Εάν το x είναι σωστά ταξινομημένο τα χαρακτηριστικά των εγγύτερων γειτόνων ενημερώνονται λαμβάνοντας υπόψη το τρέχων βάρος και τα χαρακτηριστικά του x. Οι εγγύτεροι γείτονες κινούνται προς το x στο πολυδιάστατο χώρο. Τέλος, το βάρος των εγγύτερων γειτόνων αυξάνεται κατά ένα και το x αφαιρείται. Το συμπυκνωμένο σύνολο αποτελεί και το τελικό Condensing Set(CS). Κάθε φορά που ένα αντικείμενο εκπαίδευσης είναι διαθέσιμο είτε εισάγεται στο συμπυκνωμένο σύνολο είτε επανατοποθετεί το κοντινότερο πρωτότυπο.

Όπως οι IB2 και CNN, ο αλγόριθμος AIB2 θεωρεί ότι τα λάθος ταξινομημένα αντικείμενα είναι πιθανόν κοντά στα όρια απόφασης και πρέπει να τοποθετηθούν στο συμπυκνωμένο σύνολο. Παρόλο που τα σωστά ταξινομημένα αντικείμενα δεν τοποθετούνται στο συμπυκνωμένο σύνολο, δεν αγνοούνται. Συνεισφέρουν ενημερώνοντας το κοντινότερα πρωτότυπα χωρίς να μειώνεται ο βαθμός μείωσης. Η ιδέα είναι, ότι ένα συμπυκνωμένο σύνολο που έχει κατασκευαστεί από τον AIB2, θα περιέχει καλύτερα πρωτότυπα σε σχέση με ένα συμπυκνωμένο σύνολο που έχει κατασκευαστεί από τον IB2 και μπορεί να σημειώσει μεγαλύτερα ποσοστά ακρίβειας. Ακόμη, με την ενημέρωση των πρωτοτύπων θα μειωθεί ο αριθμός των αντικειμένων που εισέρχονται στο συμπυκνωμένο σύνολο και έτσι ο AIB2 θα σημειώσει υψηλότερο βαθμό μείωσης και χαμηλότερο κόστος προεπεξεργασίας σε σχέση με τον IB2.

Algorithm 13 AIB2

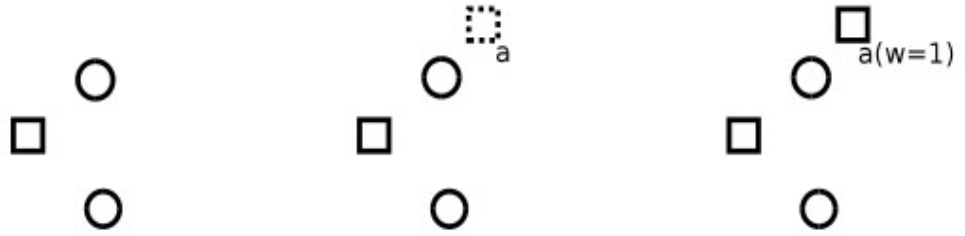
Input: TS

Output: CS

1: $CS \leftarrow \varnothing$
2: **pick an item** y **of** TS **and move it to** CS
3: $y_{weight} \leftarrow 1$
4: **for each** $x \in TS$ **do**
5: $NN \leftarrow$ **Nearest Neighbour of** x **in** CS
6: **if** $NN_{class} \neq x_{class}$ **then**
7: $x_{weight} \leftarrow 1$
8: $CS \leftarrow CS \cup \{x\}$
9: **else**
10: **for each attribute** $attr(i)$ **do**
11: $NN_{attr(i)} \leftarrow \frac{NN_{attr(i)} \times NN_{weight} + x_{attr(i)}}{NN_{weight}+1}$
12: **end for**
13: $NN_{weight} \leftarrow NN_{weight} + 1$
14: **end if**
15: $TS \leftarrow TS - \{x\}$
16: **end for**
17: **return** CS

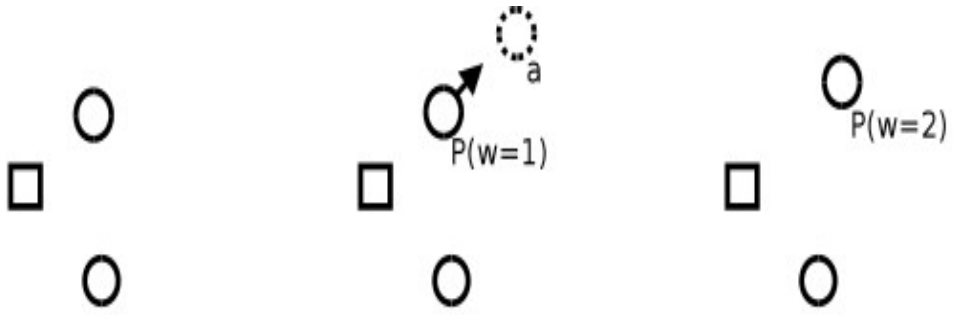
Εικόνα 8: "Τμήμα ψευδοκώδικα για τον αλγόριθμο AIB2"

Στις παρακάτω εικόνες (*Εικόνα 12, Εικόνα 13*) παρουσιάζονται κάποια παραδείγματα εκτέλεσης του AIB2. Υποθέτουμε ότι το τρέχων συμπυκνωμένο σύνολο περιλαμβάνει τρία πρωτότυπα, δύο που ανήκουν στην κλάση κύκλος και ένα που ανήκει στη κλάση τετράγωνο. Όταν ένα νέο τετράγωνο αντικείμενο φθάνει, ο αλγόριθμος AIB2 πρέπει να αποφασίσει εάν το αντικείμενο θα εισαχθεί στο συμπυκνωμένο σύνολο ή θα χρησιμοποιηθεί για την ενημέρωση ενός υπάρχων πρωτοτύπου. Από τη στιγμή που το νέο αντικείμενο είναι πιο κοντά σε διαφορετική κλάση, εισάγεται στο συμπυκνωμένο σύνολο και η τιμή βάρους του αρχικοποιείται σε ένα. Εάν αντίθετα, το νέο αντικείμενο ανήκει στην κλάση κύκλος και βρίσκεται πιο κοντά σε ένα πρωτότυπο της ίδιας κλάσης, το νέο αντικείμενο ενημερώνει το πρωτότυπο. Επομένως το πρωτότυπο μετακινείται πιο κοντά στο νέο αντικείμενο και η τιμή βάρους του αυξάνεται κατά ένα.



(a) Current condensing set (3 prototypes) (b) New item arrival (c) The new item enters the condnensing set

Εικόνα 9: "Παράδειγμα αλγορίθμου AIB2: Ένα νέο αντικείμενο εισέρχεται στο συμπυκνωμένο σύνολο"



(a) Current condensing set (3 prototypes) (b) New item arrival (c) Updating of nearest prototype

Εικόνα 10: "Παράδειγμα αλγορίθμου AIB2: Επανατοποθέτηση υπάρχων πρωτοτύπου"

Με την ενημέρωση των πρωτοτύπων ο AIB2 εξασφαλίζει ότι κάθε πρωτότυπο βρίσκεται κοντά στο κέντρο της περιοχής δεδομένων που αντιπροσωπεύει. Το βάρος ενός πρωτοτύπου δηλώνει το πληθυσμό των αρχικών αντικειμένων που αναπαριστούν. Παρόλο που κάθε φορά που ένα νέο αντικείμενο εκπαίδευσης φθάνει και καταχωρείται σε ένα υπάρχων πρωτότυπο, το οποίο μετακινείται προς το νέο αντικείμενο, όσο μεγαλύτερο είναι το βάρος του πρωτοτύπου τόσο πιο μικρή είναι η κίνησή του προς το νέο αντικείμενο εκπαίδευσης.

Παρόμοια με τον IB2, στον AIB2 τα πρωτότυπα εξαρτώνται από τη σειρά των αντικειμένων εκπαίδευσης. Είναι πολύ πιθανό λόγω της ανομοιόμορφης άφιξης των αντικειμένων εκπαίδευσης,

ένα πρωτότυπο του AIB2 μπορεί να απομακρύνεται από κάποιο αρχικό αντικείμενο εκπαίδευσης που αντιπροσωπεύει. Παρόλα αυτά, τα πρωτότυπα του AIB2 με μεγάλα βάρη θα μετακινούνται πολύ αργά προς τα νέα αντικείμενα εκπαίδευσης. Αυτό αποτελεί ένα πρόβλημα για τον IB2, πως αντίθετα με τον CNN δεν δίνει κάποια εγγύηση σωστής κατηγοριοποίησης όλων των δεδομένων εκπαίδευσης που εξετάζονται.

2.4 Ο αλγόριθμος RSP3 – Reduction by Space Partitioning

Πρόγονος των Rduction by Space Partitioning(RSP)[13] αλγορίθμων είναι ο αλγόριθμος PG που προτάθηκε από τους Chen και Jozwik (Chen's algorithm). Ο αλγόριθμος αυτός ανακτά τα στιγμιότυπα που καθορίζουν τη διάμετρο του συνόλου εκπαίδευσης, δηλαδή τα δύο πιο απομακρυσμένα σημεία. Στη συνέχεια, χωρίζει τα δεδομένα εκπαίδευσης σε δύο υποσύνολα τοποθετώντας τα στιγμιότυπα σε συστάδες ανάλογα σε ποιο άκρο είναι πιο κοντά. Ο αλγόριθμος του Chen επιλέγει να χωρίσει τα μη ομοιογενή υποσύνολα με τη μεγαλύτερη διάμετρο. Μη ομοιογενή είναι τα υποσύνολα με στιγμιότυπα από πάνω από μια κλάση. Εάν υπάρχουν μη ομοιογενή υποσύνολα ο αλγόριθμος αφαιρεί τα ομοιογενή υποσύνολα με τη μεγαλύτερη διάμετρο και η διαδικασία αυτή τερματίζεται όταν ο αριθμός των υποσυνόλων φτάσει τον αριθμό που έχει δώσει ο χρήστης. Τελικό βήμα αποτελεί η παραγωγή των πρωτοτύπων, με κάθε υποσύνολο *C* να αντικαθιστάτε από το αντικείμενο μέσο. Η ετικέτα κλάσης αυτού του αντικείμενου είναι η μεγαλύτερη κλάση στο *C*. Το σύνολο των στιγμιότυπων μέσων αποτελεί το συμπυκνωμένο σύνολο.

Ο αλγόριθμος του Chen παράγει το ίδιο συμπυκνωμένο σύνολο ανεξάρτητα από τη σειρά των στιγμιότυπων. Ένα μειονέκτημα είναι ότι ο χρήστης πρέπει να ορίσει το μέγεθος του συμπυκνωμένου συνόλου. Με αυτό το τρόπο πιστεύετε ότι επιτρέπει στο χρήστη να καθορίσει το αντάλλαγμα ανάμεσα στο βαθμό μείωσης και στην ακρίβεια αλλά στην πράξη αυτό αποτελεί μια δαπανηρή διαδικασία. Ακόμα ένα αδύνατο σημείο είναι πως τα στιγμιότυπα που δεν ανήκουν στη κύρια κλάση ενός τελικού υποσυνόλου δεν εκπροσωπούνται στο συμπυκνωμένο σύνολο.

Παρόμοιος με τον αλγόριθμο του Chen αλλά χωρίς να αγνοεί τα στιγμιότυπα, είναι ο αλγόριθμος RSP1. Ο αλγόριθμος υπολογίζει τόσα μέσα όσες και οι διαφορετικές κλάσεις στα μη ομοιογενή υποσύνολα με αποτέλεσμα να κατασκευάζει μεγαλύτερα συμπυκνωμένα σύνολα από ότι ο αλγόριθμος του Chen. Παρόλα αυτά, η ποιότητα του συμπυκνωμένου συνόλου βελτιώνεται λαμβάνοντας υπόψη όλα τα στιγμιότυπα εκπαίδευσης.

Από τη πλευρά του, ο αλγόριθμος RSP2 επιλέγει τα υποσύνολα τα οποία θα χωριστούν πρώτα εξετάζοντας το βαθμό επικάλυψης τους. Ο βαθμός επικάλυψης ενός υποσυνόλου είναι η αναλογία της μέσης απόστασης μεταξύ στιγμιοτύπων που ανήκουν σε διαφορετικές κλάσεις και της μέσης απόστασης μεταξύ στιγμιοτύπων που ανήκουν στην ίδια κλάση. Το κριτήριο για τον διαχωρισμό

προϋποθέτει τα στιγμιότυπα που ανήκουν σε μια δοθείσα κλάση είναι κοντά το ένα με το άλλο ενώ τα στιγμιότυπα που ανήκουν σε διαφορετικές κλάσεις βρίσκονται μακρυά.

Ο μόνος RSP αλγόριθμος που κατασκευάζει το συμπυκνωμένο σύνολο χωρίς καμία παράμετρο ορισμένη από το χρήστη, είναι ο RSP3. Ο αλγόριθμος εξαλείφει τις αδυναμίες του αλγόριθμου του Chen. Χωρίζει τα μη ομοιογενή υποσύνολα και τερματίζει όταν όλα τα υποσύνολα γίνονται ομογενή. Επίσης ο RSP3 μπορεί να χρησιμοποιήσει είτε τη διάμετρο είτε το βαθμό επικάλυψης σαν κριτήριο χωρισμού. Η επιλογή του κριτηρίου είναι θέμα δευτερεύουσας σημασίας διότι όλα τα μη ομοιογενή υποσύνολα τελικά χωρίζονται όπως επίσης, η σειρά των στιγμιοτύπων εκπαίδευσης δεν παίζει κάποιο ρόλο. Ο αλγόριθμος RSP3 παράγει πολλά πρωτότυπα για τις περιοχές που βρίσκονται κοντά στα όρια και λίγα πρωτότυπα για τις πιο εσωτερικές περιοχές. Το μέγεθος του συμπυκνωμένου συνόλου εξαρτάται από το επίπεδο του θορύβου στα δεδομένα. Όσο πιο ψηλό είναι το επίπεδο του θορύβου, τόσο πιο μικρό είναι το υποσύνολο που κατασκευάζεται και η μείωση που επιτυγχάνεται είναι μικρότερη. Σημαντική παρατήρηση αποτελεί ότι, η εύρεση των πιο απομακρυσμένων στιγμιοτύπων είναι μια χρονοβόρα διαδικασία από τη στιγμή που όλες οι αποστάσεις μεταξύ των στιγμιοτύπων του υποσυνόλου πρέπει να υπολογιστούν. Επομένως, η χρήση του αλγόριθμου RSP3 είναι απαγορευτική για μεγάλα σύνολα δεδομένων. Τέλος, στα πειράματά μας, χρησιμοποιούμε τον RSP3 και όχι τους RSP1,RSP2,Chen καθώς εκμεταλλευόμαστε το πλεονέκτημα του RSP3 όπου δεν δέχεται κάποια παράμετρο.

2.5 Ο Αλγόριθμος RHC – Reduction through Homogenous Clusters

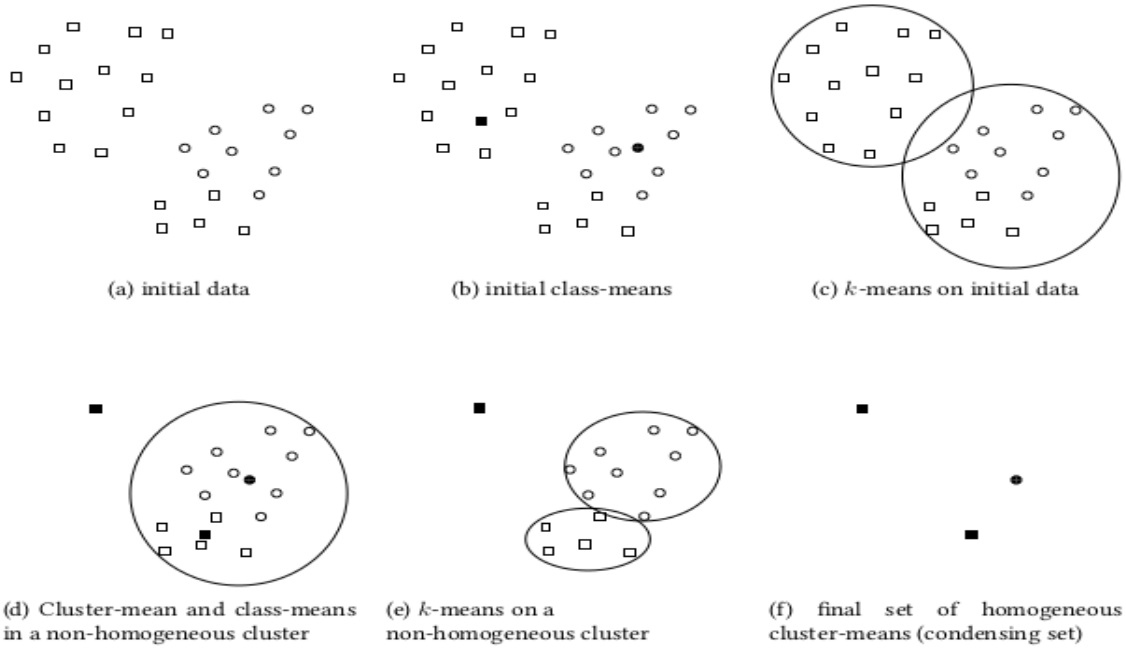
Ο RHC[13] είναι ένας αλγόριθμος σύνοψης πρωτοτύπων. Βασίζεται στην ιδέα των αναδρομικών εφαρμογών της συσταδοποίησης k-means. Πιο συγκεκριμένα, ο RHC κατασκευάζει συστάδες μέχρι όλες τους να είναι ομοιογενής, δηλαδή να περιέχουν αντικείμενα τα οποία ανήκουν σε συγκεκριμένη κλάση. Ο RHC είναι μη παραμετρικός. Δηλαδή το condensing set παράγεται αυτόματα χωρίς ο χρήσης να πρέπει να προσδιορίσει τιμή σε κάποια παράμετρο.

Σε πρώτη φάση, ο RHC θεωρεί ολόκληρο το σύνολο εκπαίδευσης ως μια μη ομοιογενή συστάδα. Υπολογίζει το μέσο κάθε κλάσης κατά τον μέσο όρο των τιμών των χαρακτηριστικών των αντικειμένων του συνόλου εκπαίδευσης. Στη συνέχεια, για ένα σύνολο δεδομένων με *n* κλάσεις ο αλγόριθμος υπολογίζει *n* μέσα και εκτελεί την συσταδοποίηση k-means χρησιμοποιώντας τα μέσα των κλάσεων που αναφέρθηκαν παραπάνω ως αρχικά μέσα και φτιάχνει *n* συστάδες. Για κάθε ομοιογενή συστάδα (δηλαδή συστάδα που περιέχει μόνο στιγμιότυπα μιας κλάσης) , το μέσο της τοποθετείται στο συμπυκνωμένο σύνολο ως πρωτότυπο. Για κάθε μη ομοιογενή συστάδα, η διαδικασία που αναφέρθηκε εφαρμόζεται αναδρομικά. Ο αλγόριθμος σταματάει όταν όλες οι συστάδες είναι ομοιογενής. Στο τέλος, το συμπυκνωμένο σύνολο περιλαμβάνει όλα τα μέσα των

ομοιογενών κλάσεων. Αναφέρεται πως η χρήση των μέσων των κλάσεων ως αρχικά μέσα για την συσταδοποίηση k-means, ο αριθμός των συστάδων καθορίζεται αυτόματα.

Το μέσο αντικείμενο *m* κάθε συστάδας ή κλάσης *C*, υπολογίζεται από το μέσο όρο των n τιμών των χαρακτηριστικών για τα αντικείμενα xi, i = 1,2,... |C| που ανήκουν στην κλάση C. Ο υπολογισμός γίνεται με τον παρακάτω τύπο:

$$m.dj = \frac{1}{|C|} \sum_{xi \in C} xi, dj, j = 1,2,..,n \quad (2.1.1)$$



Εικόνα 11: "Σύνοψη Δεδομένων με χρήση του RHC"

Η εικόνα 9 περιγράφει ένα παράδειγμα εκτέλεσης του RHC. Στο συγκεκριμένο παράδειγμα υποθέτουμε ότι το σύνολο δεδομένων περιέχει 26 αντικείμενα, κύκλους και τετράγωνα. Αρχικά, ο RHC υπολογίζει ένα μέσο κλάσης για τα τετράγωνα και ένα για τους κύκλους. Έπειτα, η μέθοδος συσταδοποίησης k-means χρησιμοποιεί τα δύο μέσα των κλάσεων ως αρχικά μέσα και κατασκευάζει δύο συστάδες. Η μία περιλαμβάνει μόνο τετράγωνα και η άλλη περιλαμβάνει αντικείμενα και από τις δύο κλάσεις. Για την ομοιογενή συστάδα, ο αλγόριθμος αποθηκεύει το μέσο της στο συμπυκνωμένο σύνολο ως πρωτότυπο της κλάσης των τετραγώνων. Για τα αντικείμενα της μη ομοιογενής κλάσης, ο RHC αναδρομικά κατασκευάζει δύο ομοιογενή συστάδες. Άρα, άλλα δύο πρωτότυπα αποθηκεύονται στο συμπυκνωμένο σύνολο. Το τελικό συμπυκνωμένο σύνολο περιέχει τρία πρωτότυπα αντί για τα 26 αντικείμενα του αρχικού συνόλου εκπαίδευσης.

Σύμφωνα και με το παραπάνω παράδειγμα, βλέπουμε ότι ο αλγόριθμος RHC παράγει πολλά πρωτότυπα για περιοχές δεδομένων των οποίων τα όρια είναι κοντά και λιγότερα πρωτότυπα για τις πιο κεντρικές περιοχές δεδομένων. Επομένως, όσο περισσότερες κλάσεις και θόρυβος υπάρχουν στα δεδομένα, τόσο περισσότερα σύνορα υπάρχουν με αποτέλεσμα να επιτυγχάνεται μικρότερο ποσοστό μείωσης. Όταν ο αλγόριθμος εκτελείται σε ένα σύνολο δεδομένων χωρίς θόρυβο, σχηματίζει μεγαλύτερες συστάδες ενώ αν το σύνολο έχει πολύ θόρυβο τότε σχηματίζονται πολλές μικρές συστάδες. Ο RHC αυξάνει τη πιθανότητα της εύρεσης μεγάλων ομοιογενών συστάδων και πετυχαίνει μεγάλο ποσοστό μείωσης.

2.6 Ο Αλγόριθμος ERHC – Editing and Reduction through Homogenous Clusters

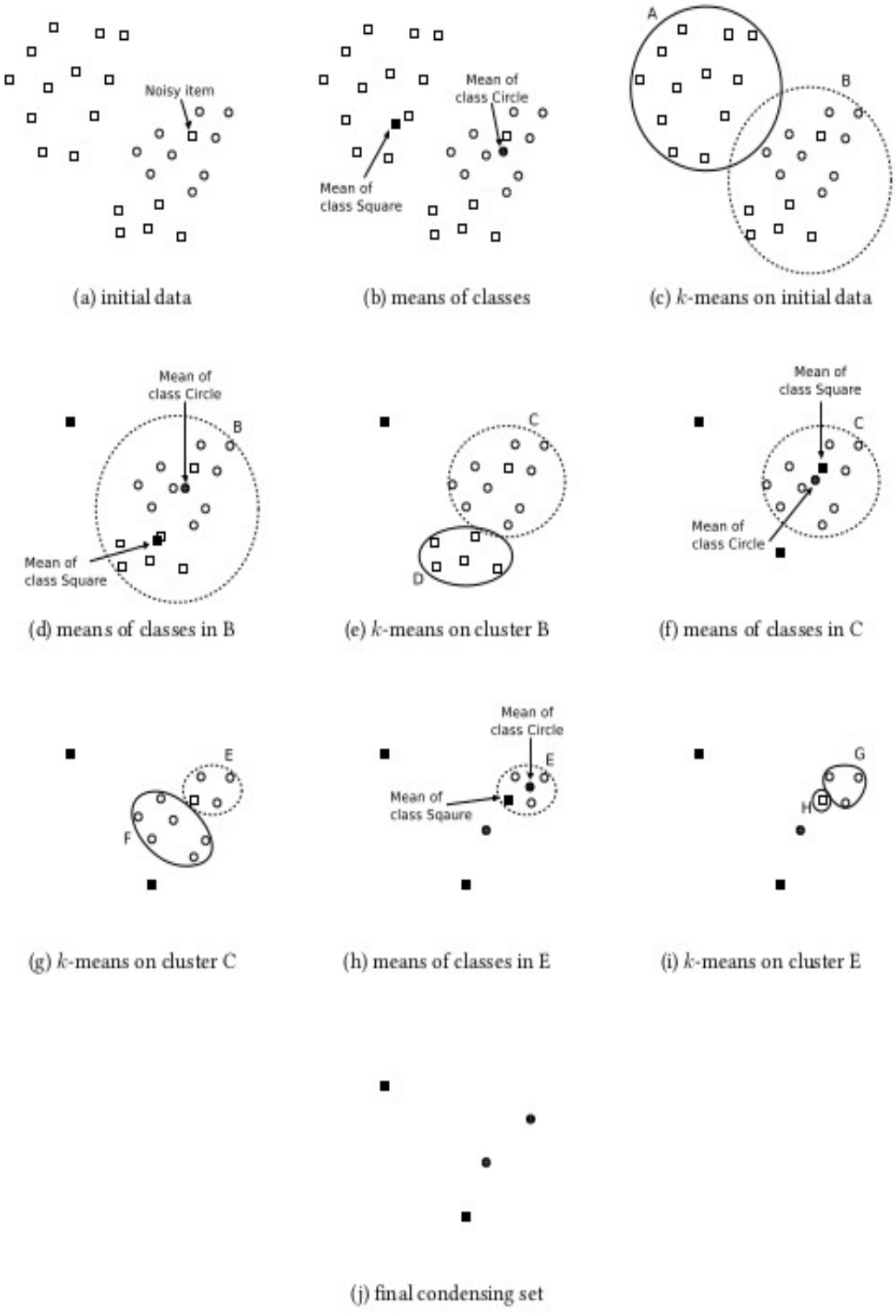
Μια απλή παραλλαγή του RHC αποτελεί ο αλγόριθμος ERHC[13] ο οποίος μπορεί να αντιμετωπίσει περιπτώσεις όπου τα δεδομένα περιέχουν θόρυβο. Στον ERHC, οι ομοιογενής συστάδες με ένα μόνο αντικείμενο απορρίπτονται για το λόγο ότι πιθανόν να αναπαριστούν θόρυβο. Το τελικό σύνολο συμπύκνωσης περιλαμβάνει τα μέσα των ομοιογενών συστάδων που περιέχουν πάνω από ένα αντικείμενο. Ταυτόχρονα, ο ERHC αφαιρεί το θόρυβο και μειώνει το μέγεθος του συνόλου εκπαίδευσης. Έτσι, μπορεί να χαρακτηριστεί ως ένας υβριδικός αλγόριθμος παραγωγής πρωτοτύπων.

Ο αλγόριθμος ERHC δουλεύει με έναν τρόπο παρόμοιο με τους RHC και EHC. Πιο συγκεκριμένα, αρχικά ολόκληρο το σύνολο εκπαίδευσης θεωρείται ως μια μη ομοιογενή συστάδα. Ο αλγόριθμος ξεκινάει υπολογίζοντας τα μέσα των κλάσεων για κάθε κλάση υπολογίζοντας το μέσο όρο των αντίστοιχων αντικειμένων κάθε συστάδας. Επομένως για ένα σύνολο δεδομένων με *n* κλάσεις ο ERHC υπολογίζει *n* μέσα κλάσεων. Ο αλγόριθμος συνεχίζει με την εκτέλεση της k-means συσταδοποίησης υιοθετώντας τα *n* μέσα των κλάσεων ως αρχικά μέσα και το αποτέλεσμα είναι η κατασκευή n συστάδων. Εάν μια ομοιογενής συστάδα περιέχει μόνο ένα αντικείμενο, ο ERHC το αφαιρεί. Διαφορετικά, το μέσο της αποτελεί ένα πρωτότυπο και τοποθετείται στο συμπυκνωμένο σύνολο. Η διαδικασία αυτή συσταδοποίησης εφαρμόζεται αναδρομικά στα αντικείμενα κάθε μη ομοιογενούς συστάδας. Ο αλγόριθμος τερματίζει όταν όλες οι συστάδες γίνουν ομοιογενής. Το τελικό συμπυκνωμένο σύνολο που κατασκευάστηκε από τον ERHC περιλαμβάνει τα μέσα των κλάσεων των ομοιογενών συστάδων που περιέχουν πάνω από ένα αντικείμενα.

Η διαδικασία που περιγράφηκε παραπάνω απεικονίζεται στην παρακάτω εικόνα *(Εικόνα 14)* . Ας υποθέσουμε, ότι το αρχικό σύνολο εκπαίδευσης περιέχει δύο κλάσεις, οι οποίες είναι τα τετράγωνα και οι κύκλοι. Ο ERHC υπολογίζει τα δύο μέσα των κλάσεων. Ο *k-means* εφαρμόζεται στο σύνολο δεδομένων εκπαίδευσης και κατασκευάζει δύο συστάδες, τις A και B. Η συστάδα A είναι ομοιογενής και περιλαμβάνει πάνω από ένα αντικείμενα. Επομένως, το μέσο της συστάδας τοποθετείται στο συμπυκνωμένο σύνολο. Η B συστάδα αντίθετα, είναι μη ομοιογενής από τη στιγμή που

περιλαμβάνει αντικείμενα και από τις δύο κλάσεις. Έτσι, ο ERHC υπολογίζει δύο μέσα για τις κλάσεις και στη συνέχεια ο *k-means* εφαρμόζεται στη συστάδα B και κατασκευάζει τις συστάδες C και D. Εφόσον η συστάδα D είναι ομοιογενής και περιλαμβάνει πάνω από ένα αντικείμενα, το μέσο της κλάσης τοποθετείται στο συμπυκνωμένο σύνολο. Η συστάδα C που είναι μη ομοιογενής, το μέσο της υπολογίζεται και εφαρμόζεται ο *k-means* που με τη σειρά του κατασκευάζει άλλες δύο συστάδες, τις E,F. Το μέσο της συστάδας F τοποθετείται στο συμπυκνωμένο σύνολο ενώ, τα μέσα των κλάσεων υπολογίζονται για την μη ομοιογενή συστάδα E. Εφαρμόζεται ο k-means στην E και τα αποτελέσματα είναι οι συστάδες G και H. Και οι δύο συστάδες είναι ομοιογενής. Όμως, η H περιέχει ένα μόνο αντικείμενο. Το μέσο της συστάδας G τοποθετείται στο συμπυκνωμένο σύνολο. Το τελικό συμπυκνωμένο σύνολο περιλαμβάνει μόνο 4 αντικείμενα. Η διαφορά μεταξύ του RHC και του ERHC, είναι ότι ο RHC θα τοποθετήσει στο συμπυκνωμένο σύνολο ένα πρωτότυπο για την συστάδα H το οποίο επηρεάζει την ποιότητα του τελικού συμπυκνωμένου συνόλου.

Οι ομοιότητες του αλγόριθμου ERHC είναι αρκετές με τον RHC. Παρόλα αυτά, ο απλός μηχανισμός επεξεργασίας στον ERHC μπορεί να βελτιώσει αποτελεσματικά την απόδοση της κατηγοριοποίησης, ειδικά όταν τα δεδομένα περιέχουν θόρυβο. Ο ERHC κληρονομεί όλα τα πλεονεκτήματα του αλγορίθμου για την εύρεση ομοιογενών συστάδων. Αυτό τον κάνει αρκετά γρήγορο ειδικά από τη στιγμή που βασίζεται στον k-means που επιταχύνεται μέσω της αρχικοποίησης των μέσων των κλάσεων. Επιπλέον, ο αλγόριθμος ERHC δεν εξαρτάται από τη σειρά των δεδομένων στο σύνολο εκπαίδευσης. Η εκτέλεσή του δεν είναι ισοδύναμη με αυτή του αλγορίθμου RHC. Διαφορετικές συστάδες κατασκευάζονται από τους δύο αλγόριθμους με αποτέλεσμα τα επίπεδα μείωσης που επιτυγχάνονται να είναι διαφορετικά. Ακόμα, ο RHC έχει υψηλότερο κόστος προεπεξεργασίας συγκριτικά με τον ERHC καθώς η διαδικασία εύρεσης ομοιογενών συστάδων γίνεται δύο φορές, μια για τα δεδομένα που δεν έχουν υποστεί επεξεργασία και μια για τα επεξεργασμένα δεδομένα. Τέλος, ο ERHC μπορεί να απομακρύνει το θόρυβο από τα δεδομένα και να μειώνει το μέγεθος του συνόλου δεδομένων εκπαίδευσης ταυτόχρονα.



Εικόνα 12: "ERHC: Διαδικασία σύνοψης και επεξεργασίας δεδομένων"

Algorithm 12 ERHC

Input: TS

Output: CS

```
1: Queue ← ∅
2: Enqueue(Queue, TS)
3: CS ← ∅
4: repeat
5:   C ← Dequeue(Queue)
6:   if C is homogeneous then
7:     if |C| > 1 then
8:       r ← mean of C
9:       CS ← CS ∪ {r}
10:    end if
11:  else
12:    M ← ∅ {M is the set of class-means}
13:    for each class L in C do
14:      m_L ← mean of L
15:      M ← M ∪ {m_L}
16:    end for
17:    NewClusters ← K-MEANS(C, M)
18:    for each cluster C ∈ NewClusters do
19:      Enqueue(Queue, C)
20:    end for
21:  end if
22: until IsEmpty(Queue)
23: return CS
```

Εικόνα 13: "Τμήμα ψευδοκώδικα του αλγορίθμου ERHC"

ΕΠΙΛΟΓΟΣ

Κλείνοντας το συγκεκριμένο κεφάλαιο, έχουμε καλύψει βασικές και χρήσιμες πληροφορίες για τους αλγόριθμους μείωσης δεδομένων οι οποίοι χρησιμοποιήθηκαν στην εργασία ενώ αναφερθήκαμε και στα μειονεκτήματα και τα πλεονεκτήματα αυτών. Στη συνέχεια, θα σχολιάσουμε το πρόβλημα της κατηγοριοποίησης των δεδομένων πολλαπλών ετικετών και τρόπους με τους οποίους μπορούμε να τα διαχειριστούμε.

ΚΕΦΑΛΑΙΟ 3: ΔΙΑΧΕΙΡΙΣΗ ΔΕΔΟΜΕΝΩΝ ΠΟΛΛΑΠΛΩΝ ΕΤΙΚΕΤΩΝ

ΕΙΣΑΓΩΓΗ

Η κατηγοριοποίηση στιγμιοτύπων, παραδοσιακά, αναφέρεται σε στιγμιότυπα τα οποία έχουν μια ετικέτα μέσα από ένα σύνολο ανεξάρτητων ετικετών L, όπου L > 1. Όταν L = 2, τότε το πρόβλημα της κατηγοριοποίησης καλείται δυαδική κατηγοριοποίηση. Στις περιπτώσεις όμως όπου το L είναι μεγαλύτερο του 2, τότε έχουμε να κάνουμε με multi-class κατηγοριοποίηση. Στη multi-label κατηγοριοποίηση, το σύνολο των ετικετών των στιγμιοτύπων είναι υποσύνολο του συνόλου L. Οι περιπτώσεις στις οποίες συναντάμε σύνολα των οποίων τα στιγμιότυπα διαθέτουν πάνω από μια ετικέτες, αποτελούν τα δεδομένα κειμένου, όπως τα έγγραφα, καθώς και ιστοσελίδες. Η κατηγοριοποίηση δεδομένων κειμένου είναι από τις πιο γνωστές εφαρμογές κατηγοριοποίησης πολλαπλών ετικετών. Άλλες εφαρμογές που κίνησαν και το ενδιαφέρον πολλών ερευνητών του χώρου είναι ο σημασιολογικός σχολιασμός εικόνων και βίντεο καθώς και η κατηγοριοποίηση της μουσικής σε συναισθήματα.

3.1 Κατηγορίες Μεθόδων κατηγοριοποίησης Δεδομένων Πολλαπλών Ετικετών

Οι μέθοδοι κατηγοριοποίησης multi-label δεδομένων μπορούν να χωριστούν σε δύο ομάδες. Η πρώτη αφορά τις μεθόδους μετασχηματισμού του προβλήματος ενώ η δεύτερη αφορά τις μεθόδους προσαρμογής αλγορίθμων. Στη πρώτη ομάδα, οι μέθοδοι είναι ανεξάρτητες από τον αλγόριθμο κατηγοριοποίησης καθώς μετατρέπουν το πρόβλημα της multi-label κατηγοριοποίησης σε ένα ή περισσότερα προβλήματα single-label κατηγοριοποίησης, παλινδρόμησης ή κατάταξης. Αντίθετα, η δεύτερη ομάδα μεθόδων, προσπαθούν να επεκτείνουν αλγόριθμους κατηγοριοποίησης με σκοπό να μπορέσουν να χειριστούν multi-label δεδομένα κατευθείαν. Αυτό επιτυγχάνεται με τις υπάρχουσες επεκτάσεις σχετικά με δέντρα αποφάσεων, μηχανών διανυσμάτων υποστήριξης(support vector machines), νευρωνικών δικτύων και άλλων αλγορίθμων κατηγοριοποίησης. Στη συνέχεια του κεφαλαίου θα γίνει περιγραφή δύο μεθόδων που ανήκουν στην ομάδα μετασχηματισμού προβλήματος οι οποίες χρησιμοποιήθηκαν στην συγκεκριμένη εργασία.

3.2 Μέθοδος Δυαδικής Σχετικότητας – Binary Relevance

Η μέθοδος μετασχηματισμού που χρησιμοποιείται στις περισσότερες περιπτώσεις, θεωρεί την πρόβλεψη κάθε ετικέτας και την αντιμετωπίζει σαν ανεξάρτητο πρόβλημα. Σε αυτή τη περίπτωση, ένα σύνολο από single-label δυαδικών κατηγοριοποιητών εκπαιδεύεται, ένας για κάθε κλάση. Κάθε κατηγοριοποιητής προβλέπει εάν ένα αντικείμενο ανήκει σε μια κλάση ή όχι. Η ένωση όλων των κλάσεων που προβλέφθηκαν λαμβάνεται ως η multi-label έξοδος. Αυτή η προσέγγιση είναι πιο

γνωστή επειδή είναι πιο εύκολο να υλοποιηθεί αλλά, αγνοεί τις πιθανές συσχετίσεις μεταξύ των ετικετών των κλάσεων. Με άλλα λόγια, το multi-label σύνολο δεδομένων *D* πρώτα αποσυντίθεται σε *q* δυαδικά σύνολα δεδομένων τα οποία χρησιμοποιούνται για την κατασκευή *q* ανεξάρτητων δυαδικών κατηγοριοποιητών. Σε κάθε δυαδικό πρόβλημα κατηγοριοποίησης, παραδείγματα τα οποία συσχετίζονται με την αντίστοιχη ετικέτα θεωρούνται θετικά και τα υπόλοιπα παραδείγματα θεωρούνται αρνητικά.

	Περιπέτεια	Όχι Περιπέτεια		Κωμωδία	Όχι Κωμωδία		Δράσης	Όχι Δράσης		Φαντασίας	Όχι Φαντασίας
1	X		1	X		1		X	1		X
2	X		2		X	2		X	2	X	
3	X		3		X	3	X		3		X
4		X	4		X	4	X		4	X	

Εικόνα 14: "Νέα Μετασχηματισμένα σύνολα δεδομένων που προέκυψαν από τη μέθοδο Binary Relevance"

Στην περίπτωση κατηγοριοποίησης ένα νέο mluti-label αντικείμενου, η μέθοδος Binary Relevance εξάγει το σύνολο των ετικετών που προβλέφθηκαν σωστά από τους *q* ανεξάρτητους δυαδικούς κατηγοριοποιητές. Καθώς η μέθοδος κλιμακώνεται γραμμικά σχετικά με το μέγεθος q του συνόλου των ετικετών, είναι καταλληλότερη για όχι πολύ μεγάλα q. Παρόλα αυτά, η μέθοδος υστερεί από την έλλειψη ότι η συσχέτιση μεταξύ των ετικετών δεν λαμβάνεται υπόψη.

X	Y_1	Y_2	Y_3	Y_4
$x^{(1)}$	0	1	1	0
$x^{(2)}$	1	0	0	0
$x^{(3)}$	0	1	0	0
$x^{(4)}$	1	0	0	1
$x^{(5)}$	0	0	0	1

X	Y_1
$x^{(1)}$	0
$x^{(2)}$	1
$x^{(3)}$	0
$x^{(4)}$	1
$x^{(5)}$	0

X	Y_2
$x^{(1)}$	1
$x^{(2)}$	0
$x^{(3)}$	1
$x^{(4)}$	0
$x^{(5)}$	0

X	Y_3
$x^{(1)}$	1
$x^{(2)}$	0
$x^{(3)}$	0
$x^{(4)}$	0
$x^{(5)}$	0

X	Y_4
$x^{(1)}$	0
$x^{(2)}$	0
$x^{(3)}$	0
$x^{(4)}$	1
$x^{(5)}$	1

Εικόνα 15: "Παράδειγμα μετασχηματισμού της μεθόδου Binary Relevance. Όπου X τα στιγμιότυπα και Y οι διαφορετικές ετικέτες."

3.3 Μέθοδος Δυναμοσύνολο Ετικέτας – Label Powerset

Μία άλλη μέθοδος μετασχηματισμού, λιγότερο διαδεδομένη, είναι η μέθοδος Label Powerset. Η συγκεκριμένη μέθοδος θεωρεί κάθε υποσύνολο του συνόλου ετικετών *L* που βρίσκεται στο σύνολο εκπαίδευσης ως μία διαφορετική ετικέτα. Είναι αρκετά ενδιαφέρουσα για μελέτη η συγκεκριμένη μέθοδος καθώς έχει το πλεονέκτημα να λαμβάνει υπόψη τις συσχετίσεις μεταξύ των ετικετών. Με αυτό το τρόπο, σε μερικές περιπτώσεις, μπορεί να πετύχει καλύτερη επίδοση συγκριτικά με πιο απλές υπολογιστικά προσεγγίσεις όπως η μέθοδος Binary Relevance η οποία μαθαίνει ένα δυαδικό μοντέλο για κάθε ετικέτα ανεξάρτητα από τις υπόλοιπες.

Παρόλα αυτά, η μέθοδος Label Powerset αμφισβητείται εφαρμογές με μεγάλο αριθμό ετικετών και παραδείγματα εκπαίδευσης λόγω του μεγάλου αριθμού δυναμοσυνόλων που εμφανίζονται στο σύνολο εκπαίδευσης. Αυτός ο μεγάλος αριθμός αυξάνει το υπολογιστικό κόστος της μεθόδου Label Powerset από τη μια κάνει τη διαδικασία μάθησης αρκετά δύσκολη και από την άλλη πολλά από αυτά τα δυναμοσύνολα συνήθως σχετίζονται με πολύ λίγα στιγμιότυπα. Επιπλέον, η μέθοδος Label Powerset μπορεί μόνο να προβλέψει δυναμοσύνολα που παρατηρούνται στο σύνολο δεδομένων εκπαίδευσης. Αυτό αποτελεί ένα σημαντικό περιορισμό καθώς νέα δυναμοσύνολα τυπικά

εμφανίζονται στα σύνολα δεδομένων δοκιμής και αναπαριστούν νέα δεδομένα τα οποία δεν έχουν εμφανιστεί ξανά.

Έχουν γίνει κάποιες προσπάθειες για την αντιμετώπιση των παραπάνω προβλημάτων της μεθόδου Label Powerset. Σε μία από αυτές, το αρχικό σύνολο ετικετών χωρίζεται τυχαία σε μικρότερα δυναμοσύνολα και στη συνέχεια καλείται η μέθοδος Label Powerset για την εκπαίδευση των αντίστοιχων multi-label κατηγοριοποιητών. Με αυτό το τρόπο τα αποτελέσματα των single-label κατηγοριοποιητών, υπολογίζονται πιο εύκολα και η κατανομή των τιμών των κλάσεων είναι μικρότερη. Η μέθοδος που περιγράφηκε παραπάνω καλείται Ra*k*EL (RAndom *k* labELsets) όπου k είναι μια παράμετρος η οποία καθορίζει το μέγεθος του δυναμοσυνόλου.

Παρακάτω παρουσιάζονται κάποιες εικόνες για την καλύτερη εξήγηση της μεθόδου. Συγκεκριμένα πρόκειται για παραδείγματα ταινιών που ανήκουν σε μία ή περισσότερες από τις κλάσεις Περιπέτεια, Κωμωδία, Δράση, Φαντασίας

	Περιπέτεια	Κωμωδία	Δράσης	Φαντασίας
1	X	X		
2	X			X
3	X		X	
4			X	X

Εικόνα 16: "Πίνακας στιγμιοτύπων που ανήκουν σε πάνω από μια κλάσεις"

Η Εικόνα 17 δείχνει τα αποτελέσματα του μετασχηματισμού του συνόλου δεδομένων με τη χρήση της μεθόδου Label Powerset.

	Περιπέτεια & Κωμωδία	Περιπέτεια & Φαντασίας	Περιπέτεια & Δράσης	Δράσης & Φαντασίας
1	X			
2		X		
3			X	
4				X

Εικόνα 17: "Νέο μετασχηματισμένο σύνολο δεδομένων που προέκυψε από τη μέθοδο Label Powerset"

3.4 Μέτρα για την αξιολόγηση Multi-label κατηγοριοποιητών

Στα προβλήματα που συναντάμε κατά την διαδικασία της κατηγοριοποίησης, όπως για παράδειγμα σε multi-class προβλήματα ή σε προβλήματα με δεδομένα πολλαπλών ετικετών, το κριτήριο που είναι πιο συχνό για να εκτιμήσουμε την απόδοση ενός κατηγοριοποιητή, είναι η ακρίβεια(accuracy).

Κατά την multi-label κατηγοριοποίηση, οι προβλέψεις για ένα στιγμιότυπο είναι ένα σύνολο ετικετών με αποτέλεσμα η πρόβλεψη που θα γίνει να είναι είτε πλήρως σωστή, είτε μερικώς σωστή είτε τελείως λάθος. Ανάλογα από το ποιο πρόβλημα αντιμετωπίζεται κάθε φορά, οι τεχνικές αξιολόγησης για multi-label δεδομένα μπορούν να χωριστούν στις κατηγορίες: evaluating partitions, evaluating ranking, using label hierarchy. Η πρώτη κατηγορία αξιολογεί την ποιότητα της κατηγοριοποίησης σε κλάσεις, η δεύτερη αφορά την αξιολόγηση εάν οι κλάσεις βρίσκονται διατεταγμένες στη σειρά και η τρίτη κατηγορία, αξιολογεί πόσο αποτελεσματικά το σύστημα μάθησης είναι ικανό να λάβει υπόψιν μια υπάρχουσα δομή των ετικετών.

Η αξιολόγηση για κάποιους αλγόριθμους μάθησης, είναι η μέτρηση του πόσο μακρυά από τις πραγματικές τιμές των ετικετών των κλάσεων είναι οι προβλέψεις που κάνει ο αλγόριθμος κατηγοριοποίησης όταν δοκιμάζεται σε δεδομένα τα οποία δεν έχει ξαναδεί. Πιο συγκεκριμένα, μπορούμε να αξιολογήσουμε τη μέση διαφορά ανάμεσα στις ετικέτες τις οποίες πρόβλεψε ο αλγόριθμος και στις πραγματικές ετικέτες για κάθε παράδειγμα δοκιμής και μετά για όλα τα παραδείγματα. Αυτή η μέθοδος καλείται example based evaluations.

Μια τέτοια example based μετρική χρησιμοποιήσαμε και εμείς στα πειράματά μας και η μετρική αυτή είναι το Exact Match Ratio(EMR). Με άλλα λόγια, εκτιμούμε το πόσο ακριβείς είναι ένας αλγόριθμος στο να προβλέψει όλες τις κλάσεις στις οποίες ανήκει ένα στιγμιότυπο. Επιπρόσθετα, η αξιολόγηση ενός multi-label αλγόριθμου κατηγοριοποίησης είναι δύσκολη κυρίως γιατί οι multi-label προβλέψεις έχουν την ιδιαιτερότητα του να είναι μερικώς σωστές. Έτσι, στα πειράματά μας χρησιμοποιήσαμε μια ακόμα μετρική, την ακρίβεια(accuracy) η οποία θεωρεί ότι μια μερικώς σωστή πρόβλεψη δεν είναι τελείως λάθος (όπως στη περίπτωση του Exact Match Ratio). Συνεπώς, η ακρίβεια για κάθε στιγμιότυπο ορίζεται ως το σύνολο των ετικετών που έχουν προβλεφθεί σωστά προς το συνολικό αριθμό των ετικετών αυτού του στιγμιότυπου. Άρα, το accuracy είναι ο μέσος όρος αυτών των προβλέψεων.

ΕΠΙΛΟΓΟΣ

Σε αυτό το κεφάλαιο μιλήσαμε για τις κατηγορίες των μεθόδων κατηγοριοποίησης δεδομένων πολλαπλών ετικετών και πιο συγκεκριμένα είδαμε τις μεθόδους Binary Relevance και Label Powerset. Ακόμα αναφερθήκαμε στα μέτρα με τα οποία γίνεται η αξιολόγηση των multi-label αλγορίθμων κατηγοριοποίησης και καλύψαμε κάποια βασικά πράγματα για κάποια από αυτά. Το κεφάλαιο που ακολουθεί, αφορά την πειραματική μελέτη που έγινε για την συγκεκριμένη πτυχιακή εργασία. Πιο συγκεκριμένα, θα μιλήσουμε για τα σύνολα δεδομένων που χρησιμοποιήσαμε για τα πειράματα και τις παραμέτρους αυτών ενώ τέλος θα παρουσιάσουμε τα αποτελέσματα που πήραμε και θα γίνει σχολιασμός αυτών.

ΚΕΦΑΛΑΙΟ 4: ΠΕΙΡΑΜΑΤΙΚΗ ΜΕΛΕΤΗ

4.1 ΕΙΣΑΓΩΓΗ

Στο κεφάλαιο αυτό θα γίνει παρουσίαση των συνόλων δεδομένων που χρησιμοποιήθηκαν στη συγκεκριμένη πτυχιακή εργασία καθώς και των παραμέτρων που χρησιμοποιήθηκαν στα πειράματά μας. Ακόμη, θα δούμε τα αποτελέσματα που πήραμε από τα πειράματα και τέλος θα σχολιάσουμε τα αποτελέσματα αυτά για να δούμε τελικά πόσο κοντά βρεθήκαμε στο στόχο μας.

4.2 Περιβάλλον Πειραματικής Μελέτης

Το περιβάλλον στο οποίο πραγματοποιήθηκαν τα πειράματα είναι ο προσωπικός μου υπολογιστής ο οποίος διαθέτει επεξεργαστή *Intel(R) Core(TM) i3-4005U* με ταχύτητα 1.70GHz και μνήμη RAM 4GB. Σκοπός των πειραμάτων που πραγματοποιήθηκαν στην εργασία είναι, με την εφαρμογή αλγορίθμων μείωσης δεδομένων και στη συνέχεια με την εφαρμογή της τεχνικής 5-fold cross validation, να ελέγξουμε την απόδοση και την αποτελεσματικότητα του κατηγοριοποιητή εγγύτερων γειτόνων σε δεδομένα πολλαπλών ετικετών.

Οι αλγόριθμοι που χρησιμοποιήθηκαν για την μείωση των δεδομένων και περιγράφηκαν σε προηγούμενο κεφάλαιο της εργασίας είναι οι: RHC, CNN, IB2, AIB2, ERHC και RSP3. Όλοι οι αλγόριθμοι υλοποιήθηκαν σε γλώσσα C από μέλη της ομάδας διαχείρισης πληροφορίας του Πανεπιστημίου Μακεδονίας, της οποίας μέλος αποτελεί και ο επιβλέπων της πτυχιακής εργασίας.

Η κατηγοριοποίηση των παραδειγμάτων πραγματοποιήθηκε αρχικά, πριν την εφαρμογή των παραπάνω αλγορίθμων αλλά και μετά την εφαρμογή αυτών για κάθε συμπυκνωμένο σύνολο που παρήγαγε ο καθένας τους. Όλοι οι αλγόριθμοι που χρησιμοποιήθηκαν είναι μη παραμετρικοί το οποίο σημαίνει πως για την παραγωγή του συμπυκνωμένου συνόλου δεν ορίσαμε κάποια παράμετρο.

4.2.1 Αποτίμηση της Απόδοσης

Για κάθε σύνολο δεδομένων και για τους αλγόριθμους που εφαρμόστηκαν σε αυτά, έγινε υπολογισμός δύο μετρήσεων όσο αφορά το στάδιο της προεπεξεργασίας στο οποίο έγινε και η κατασκευή των συμπυκνωμένων συνόλων. Αυτές οι μετρήσεις είναι ο βαθμός μείωσης των δεδομένων και το πλήθος των αποστάσεων που υπολογίζει κάθε αλγόριθμος μείωσης των δεδομένων. Με αυτό το τρόπο λαμβάνεται υπόψη και το υπολογιστικό κόστος που είχαμε. Σχετικά με την κατηγοριοποίηση, η αποτελεσματικότητα υπολογίστηκε με τη χρήση του κατηγοριοποιητή k-NN, όπως αναφέραμε και παραπάνω, χρησιμοποιώντας κάθε φορά το συμπυκνωμένο σύνολο που είχαμε σαν έξοδο από κάθε αλγόριθμο. Υπολογίστηκαν επίσης δύο μετρήσεις. Αυτές είναι η

ακρίβεια και το Exact Match Ratio. Συνόψίζοντας, οι μετρήσεις που έγιναν για την αποτίμηση της απόδοσης, συνοπτικά είναι:

- Βαθμός Μείωσης – Reduction Rate
- Ακρίβεια Κατηγοριοποίησης – Classification Accuracy
- Exact Match Ratio (EMR)
- Κόστος Προεπεξεργασίας – Preprocessing Cost (Το κόστος της προεπεξεργασίας το εκτημήσαμε μέσω των αποστάσεων που υπολογίζει κάθε αλγόριθμος μείωσης δεδομένων)

Φυσικά, το Exact Match Ratio βρίσκει εφαρμογή και στην περίπτωση της προσέγγισης Label Powerset αλλά και σε αυτή της Binary Relevance. Αντίθετα, η ακρίβεια δεν μπορεί να εφαρμοστεί στην προσέγγιση Label Powerset και έτσι, σε αυτή την περίπτωση, υπολογίζουμε μόνο το Exact Match Ratio.

Για την κατηγοριοποίηση χρησιμοποιήσαμε τον αλγόριθμο των k εγγύτερων γειτόνων (k-NN) ενώ οι τιμές του k που χρησιμοποιήσαμε είναι οι 1,5,9 για να εκτιμήσουμε την συμπεριφορά του αλγορίθμου, τα ποσοστά ακρίβειας (accuracy) και την αναλογία ακριβούς αντιστοίχησης (Exact Match Ratio – EMR) που μπορεί να πετύχει με διαφορετικές τιμές του *k*. Το EMR αποτελεί την πιο αυστηρή μετρική καθώς υποδεικνύει το ποσοστό των παραδειγμάτων τα οποία έχουν σωστά ταξινομημένες όλες τις ετικέτες τους. Αντίθετα το accuracy εκτιμάει το πόσες σωστές προβλέψεις ετικετών πραγματοποιήθηκαν προς το πλήθος όλων των προβλέψεων ετικετών.

4.3 Παρουσίαση Συνόλων Δεδομένων

Τα σύνολα δεδομένων που χρησιμοποιήθηκαν στην εργασία είναι 3. Πρόκειται για το σύνολο δεδομένων *yeast* το οποίο αφορά την κατηγοριοποίηση γονιδίων σε λειτουργικές κλάσεις. Το σύνολο δεδομένων *emotions* το οποίο είναι ένα μουσικό σύνολο και αφορά την κατηγοριοποίηση μουσικών κομματιών ανάλογα με τα συναισθήματα που προκαλούν. Το τελευταίο σύνολο που χρησιμοποιήθηκε είναι το *scene* και αφορά τη σημασιολογική κατηγοριοποίηση ακίνητων εικόνων. Το σύνολο δεδομένων emotions, προέκυψε από τη συνεργασία των τμημάτων Δημοσιογραφίας & Μ.Μ.Ε και Μουσικών Σπουδών του Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης ενώ τα υπόλοιπα 2 ανακτήθηκαν από την ιστοσελίδα της βιβλιοθήκης μηχανών διανυσμάτων υποστήριξης LIBSVM (Support Vector Classification Library).

4.3.1 Σύνολο Δεδομένων Yeast

Το σύνολο δεδομένων yeast έχει να κάνει με την πρόβλεψη των λειτουργικών κλάσεων στις οποίες ανήκουν τα γονίδια του μύκητα Saccharomyces cerevisiae. Κάθε γονίδιο περιγράφεται από την συνένωση δεδομένων έκφρασης μικρο-συστοιχιών και δεδομένων φυλογενετικού προφίλ και συσχετίζεται με ένα σύνολο λειτουργικών κλάσεων των οποίων το μέγιστο μέγεθος μπορεί να υπερβεί τις 190. Το σύνολο των λειτουργικών κλάσεων μπορεί να δομήσει μια ιεραρχία βάθους τεσσάρων επιπέδων αλλά όπως συμβαίνει συχνά έτσι και εδώ χρησιμοποιούμε μόνο τις κλάσεις του ανώτερου επιπέδου. Το τελικό σύνολο δεδομένων που προκύπτει περιλαμβάνει 2417 γονίδια από τα οποία το καθένα αντιπροσωπεύεται από ένα διάνυσμα χαρακτηριστικών με 103 διαστάσεις. Οι πιθανές κλάσεις που υπάρχουν είναι 14.

4.3.2 Σύνολο Δεδομένων Scene

Το συγκεκριμένο σύνολο δεδομένων αφορά τη σημασιολογική κατηγοριοποίηση εικόνων και εφαρμόζεται σε περιοχές όπως η ανάλυση εικόνων με βάση το περιεχόμενο καθώς και η ενίσχυση εικόνων. Το Scene αποτελείται από 2400 στιγμιότυπα τα οποία έχουν προέλευση από τη βιβλιοθήκη φωτογραφιών της Corel καθώς και από προσωπικές συλλογές εικόνων. Τα στιγμιότυπα μπορεί να ανήκουν σε μία από τις παρακάτω πιθανές κλάσεις : beach, sunset, foliage, field. Mountain, urban. Αρχικά οι εικόνες επιλέχθηκαν με σκοπό κάθε κύρια κλάση να περιλαμβάνει 400 εικόνες έπειτα όμως έγινε νέα κατανομή στις εικόνες με χρήση πολλαπλών ετικετών. Στο τέλος της διαδικασίας, το ποσοστό των εικόνων που ανήκαν σε πολλαπλές κλάσεις ήταν κατά προσέγγιση 7.4%. Οι εικόνες αντιπροσωπεύονται από ένα διάνυσμα χαρακτηριστικών που περιέχει πληροφορίες για το χώρο και τα χρώματα της εικόνας οι οποίες είναι αρκετά αποτελεσματικές για τη διάκριση συγκεκριμένων τύπων εξωτερικών σκηνών. Κάθε εικόνα μετασχηματίζεται στο χώρο CIEL*U*V και μετά διαιρείται σε 49 περιοχές με τη χρήση ενός πλέγματος 7X7. Γίνεται υπολογισμός του μέσου όρου για κάθε περιοχή και η διακύμανση κάθε μπάντας που αντιστοιχούν σε εικόνες χαμηλής ανάλυσης και σε χαμηλού υπολογιστικού κόστους χαρακτηριστικά υφής. Το τελικό αποτέλεσμα είναι ένα 49 x 2 x 3 = 294-διάστατο διάνυσμα χαρακτηριστικών ανά εικόνα.

4.3.3 Σύνολο Δεδομένων Emotions

Το τελευταίο σύνολο δεδομένων, το emotions, αφορά την πρόβλεψη των συναισθημάτων που προκαλούνται από το άκουσμα μουσικών κομματιών. Αποτελείται από 593 κομμάτια τα οποία με τη σειρά τους μπορεί να ανήκουν σε 6 πιθανές κλάσεις συναισθημάτων οι οποίες είναι: amazed-

surprised, happy-pleased, relaxing-calm, quiet-still, sad-lonely και angry-fearful. Το σύνολο δεδομένων όταν αρχικά δημιουργήθηκε αποτελούνταν από 100 κομμάτια από 7 διαφορετικά είδη: classical, reggae, rock, pop, hip-hop, techno, jazz. Επιλέχθηκαν 3 κομμάτια από 233 μουσικά άλμπουμ και από το κάθε κομμάτι κρατήθηκε μια διάρκεια 30 δευτερολέπτων ενώ τα ηχητικά τμήματα που προέκυψαν μετατράπηκαν σε αρχεία wave. Από κάθε μουσικό αρχείο έγινε εξαγωγή 2 ειδών χαρακτηριστικών από τα οποία 8 αφορούν το ρυθμό και 64 έχουν σχετίζονται με τη χροιά. Το τελικό στάδιο της διαδικασίας αφορά την απόδοση συναισθημάτων σε κάθε κομμάτι. Έτσι, τα παραπάνω οδήγησαν σε ένα σύνολο δεδομένων που αποτελείται από 593 στιγμιότυπα τα οποία ανήκουν σε 6 πιθανές κλάσεις.

4.3.4 Χαρακτηριστικά των Συνόλων Δεδομένων

Παρακάτω παρουσιάζεται ο πίνακας με τα χαρακτηριστικά των συνόλων δεδομένων. Περιλαμβάνει χαρακτηριστικά όπως ο αριθμός των παραδειγμάτων του κάθε συνόλου, ο αριθμός των αριθμητικών και των διακριτών χαρακτηριστικών και ο αριθμός των ετικετών. Ακόμη περιλαμβάνονται κάποια multi-label στοιχεία όπως ο αριθμός των διαφορετικών υποσυνόλων ετικετών καθώς και η πολλαπλότητα και η πυκνότητα των ετικετών.

	Σύνολα Δεδομένων		
Attributes	Emotions	Scene	Yeast
Domain	Music	Image	Biology
Instances	593	2407	2417
Nominal	0	0	0
Numeric	72	294	103
Labels	6	6	14
Cardinality	1.869	1.074	4.237
Density	0.311	0.179	0.303
Distinct	27	15	198

Πίνακας 1: "Σύνολα Δεδομένων"

Πιο συγκεκριμένα, το σύνολο δεδομένων Emotions αποτελείται από 593 στιγμιότυπα και 72 αριθμητικά χαρακτηριστικά ενώ υπάρχουν 6 ετικέτες κλάσεων. Η πολλαπλότητα των ετικετών είναι 1.869 και η πυκνότητα των ετικετών είναι 0.311 ενώ υπάρχουν και 27 υποσύνολα.

Αντίστοιχα, για το σύνολο δεδομένων Scene, αποτελείται από 2.407 στιγμιότυπα με 294 αριθμητικά χαρακτηριστικά ενώ και εδώ συναντάμε 6 ετικέτες κλάσεων. Σε αυτό το σύνολο δεδομένων η πολλαπλότητα των ετικετών είναι 1074 και η πυκνότητα των ετικετών είναι 0.179 και 15 υποσύνολα.

Τέλος, για το σύνολο δεδομένων Yeast, το οποίο αποτελείται από 2417 στιγμιότυπα με 103 αριθμητικά χαρακτηριστικά και 14 ετικέτες, ο αριθμός των υποσυνόλων είναι 198 και η πολλαπλότητα των ετικετών είναι 4.237 ενώ η πυκνότητα των ετικετών είναι 0.303.

Κάποια συμπεράσματα που μπορούμε να βγάλουμε από το παραπάνω πίνακα χαρακτηριστικών είναι για παράδειγμα, ότι το σύνολο δεδομένων *scene*, με πυκνότητα ετικετών 0.179, είναι ένα αραιό σύνολο δεδομένων με μέσο όρο ετικετών κοντά στο 1. Τα σύνολα δεδομένων *emotions*, με πυκνότητα ετικετών 0.311 και *yeast*, με πυκνότητα ετικετών 0.303, είναι πιο πυκνά σύνολα με πάνω

από 1.5 ετικέτες ανά αντικείμενο με αποτέλεσμα τα δύο αυτά σύνολα να αποτελούν κατάλληλα παραδείγματα για την αξιολόγηση multi-label αλγορίθμων κατηγοριοποίησης.

4.4 Παρουσίαση Αποτελεσμάτων

Παρακάτω παρουσιάζονται οι πίνακες με τους αλγόριθμους μείωσης των συνόλων δεδομένων που χρησιμοποιήσαμε καθώς και με τις τιμές που κατέγραψαν όσον αφορά τη διαδικασία της προ-επεξεργασίας, τα ποσοστά μείωσης που πέτυχαν και τις αποστάσεις που υπολόγισαν, καθώς και τις τιμές σχετικά με την κατηγοριοποίηση μέσω του αλγόριθμου k-NN. Οι πίνακες αυτοί αφορούν την πρώτη από τις δύο μεθόδους διαχείρισης των δεδομένων πολλαπλών ετικετών, την Label Powerset. Ο αλγόριθμος conv k-NN πο παρουσιάζεται και στις δύο μεθόδους, είναι ο αλγόριθμος του εγγύτερου γείτονα που εφαρμόζεται στα σύνολα δεδομένων πριν επεξεργαστούν από οποιοδήποτε άλλο αλγόριθμο μείωσης των δεδομένων.

RHC					
	Pre-processing		k-NN Classification		
	RR	Distances	EMR(k=1)	EMR(k=5)	EMR(k=9)
emotions	50.905	163853.4	26.324	29.342	30.191
scene	81.776	1145103	66.142	67.472	66.557
yeast	39.928	4868407.4	23.127	24.533	24.699

Πίνακας 2: "Πίνακας αποτελεσμάτων του αλγόριθμου RHC(Label Powerset)"

ERHC					
	Pre-processing		k-NN Classification		
	RR	Distances	EMR(k=1)	EMR(k=5)	EMR(k=9)
emotions	79.453	163853.4	27.673	segmentation fault	segmentation fault
scene	88.463	1145103	66.931	67.638	67.305
yeast	82.523	4868407.4	23.003	Aborted	

Πίνακας 3: "Πίνακας αποτελεσμάτων του αλγόριθμου ERHC(Label Powerset)"

RSP3					
	Pre-processing		k-NN Classification		
	RR	Distances	EMR(k=1)	EMR(k=5)	EMR(k=9)
emotions	10.147	300049	24.622	30.694	32.211
scene	31.526	6950213.5	65.062	69.631	69.049
yeast	6.267	3859596.75	21.348	23.955	Aborted

Πίνακας 4: "Πίνακας αποτελεσμάτων του αλγόριθμου RSP3(Label Powerset)"

CNN					
	Pre-processing		k-NN Classification		
	RR	Distances	EMR(k=1)	EMR(k=5)	EMR(k=9)
emotions	15.453	175162	23.774	29.182	29.856
scene	43.219	3770134.5	57.915	65.227	66.639
yeast	12.037	2569968.75	18.369	21.721	Aborted

Πίνακας 5: "Πίνακας αποτελεσμάτων του αλγόριθμου CNN(Label Powersett)"

IB2					
	Pre-processing		k-NN Classification		
	RR	Distances	EMR(k=1)	EMR(k=5)	EMR(k=9)
emotions	21.937	88030.6	23.099	27.996	30.026
scene	53.749	893284.8	56.211	63.398	65.932
yeast	15.346	1602090.8	17.459	21.307	Aborted

Πίνακας 6: "Πίνακας αποτελεσμάτων του αλγόριθμου IB2(Label Powerset)"

AIB2					
	Pre-processing		k-NN Classification		
	RR	Distances	EMR(k=1)	EMR(k=5)	EMR(k=9)
emotions	26.863	83550.398	27.493	28.664	30.37
scene	62.285	731208.812	65.434	68.593	65.767
yeast	19.648	1529340.625	22.548	24.575	Aborted

Πίνακας 7: "Πίνακας αποτελεσμάτων του αλγόριθμου AIB2(Label Powerset)"

conv k-NN			
	k-NN Classification		
	EMR(k=1)	EMR(k=5)	EMR(k=9)
emotions	25.123	30.836	31.037
scene	62.36	68.176	68.342
yeast	19.983	22.879	Aborted

Πίνακας 8: "Πίνακας αποτελεσμάτων του αλγόριθμου conv k-NN(Label Powerset)"

4.4.1 Συμπεράσματα Πειραμάτων για τη Μέθοδο Label Powerset

Από τους παραπάνω πίνακες μπορούμε να βγάλουμε χρήσιμα συμπεράσματα για τη λειτουργία των αλγορίθμων μείωσης των δεδομένων. Πιο συγκεκριμένα, το μεγαλύτερο ποσοστό μείωσης για το σύνολο δεδομένων *Emotions* το πετυχαίνει ο αλγόριθμος ERHC όπως επίσης πετυχαίνει το υψηλότερο ποσοστό και στο σύνολο δεδομένων Scene και Yeast.

Όσο αφορά το κόστος προεπεξεργασίας που υπολογίζει ο κάθε αλγόριθμος για, για το σύνολο δεδομένων Emotions, οι περισσότερες αποστάσεις υπολογίζονται από τον RSP3. Για το σύνολο δεδομένων Scene, οι περισσότερες αποστάσεις υπολογίζονται από τον IB2 όπως και για το σύνολο

δεδομένων Yeast. Σχετικά με το βαθμό μείωσης που πετυχαίνει κάθε αλγόριθμος στα σύνολα δεδομένων, στο σύνολο Emotions οι ERHC(79.453), RHC(50.905) και AIB2(26.863) πετυχαίνουν τα καλύτερα ποσοστά. Στο σύνολο Scene, τα καλύτερα ποσοστά είναι από τους ERHC(88.463), RHC(81.776) και AIB2(62.285). Τέλος, στο Yeast οι ERHC(82.523), RHC(39.928) και AIB2(19.648) σημελιωσαν τα καλύτερα ποσοστά.

Σχετικά με την κατηγοριοποίηση, χρησιμοποιήσαμε τον αλγόριθμο των k εγγύτερων γειτόνων (k-NN) για όλα τα σύνολα δεδομένων με διαφορετικές τιμές στο k. Για k=1, στο σύνολο δεδομένων Emotions καλύτερα ποσοστά EMR σημειώνουν οι αλγόριθμοι ERHC(27.673), AIB2(27.493) και RHC(26.324). Στο σύνολο δεδομένων Scene, καλύτερα ποσοστά EMR πετυχαίνουν οι ERHC(66.931), RHC(66.142) και AIB2(65.434). Τέλος, για το σύνολο δεδομένων Yeast και τιμή του k=1 τα πιο υψηλά ποσοστά τα σημειώνουν οι RHC(23.127), ERHC(23.003) και AIB2(22.548).

Τα υψηλότερα ποσοστά EMR που λάβαμε από τον k-NN για k=5 για το σύνολο δεδομένων *Emotions,* είναι από τους RSP3(30.694), RHC(29.342) και CNN(29.182). Σχετικά με το σύνολο δεδομένων *Scene* τα υψηλότερα ποσοστά σημειώθηκαν από τους RSP3(69.631), AIB2(68.593) και ERHC(67.638). Στο σύνολο δεδομένων *Yeast*, τα πιο υψηλά ποσοστά EMR σημειώθηκαν από τους AIB2(24.575), RHC(24.533) και RSP3(23.955).

Η τελευταία τιμή του k που δοκιμάσαμε με τον k-NN είναι 9. Με τη συγκεκριμένη τιμή για το σύνολο δεδομένων Emotions οι αλγόριθμοι με τα υψηλότερα ποσοστά μείωσης είναι οι RSP3(32.211), AIB2(30.37) και IB2(30.026). Στο σύνολο δεδομένων Scene, οι αλγόριθμοι είναι οι RSP3(69.049), ERHC(67.305) και CNN(66.639). Στο σύνολο δεδομένων Yeast, τα υψηλότερα ποσοστά τα συναντάμε στον αλγόριθμο RHC(24.699).

Από τα παραπάνω, μπορούμε εύκολα να βγάλουμε το συμπέρασμα ότι οι αλγόριθμοι που πρωταγωνιστούν καθώς πέτυχαν υψηλά ποσοστά και και στα 3 σύνολα δεδομένων και με τις 3 διαφορετικές τιμές του *k*, είναι οι RHC,ERHC και RSP3. Επίσης αξίζει να σημειωθεί ότι σε όλα τα σύνολα δεδομένων για τις διαφορετικές τιμές του *k*, οι παραπάνω αλγόριθμοι σημείωσαν καλύτερα ποσοστά όσον αφορά το EMR από τα αντίστοιχα ποσοστά που πέτυχε ο k-NN πριν εφαρμοστούν οι αλγόριθμοι μείωσης των συνόλων δεδομένων.

Οι πίνακες που παρουσιάζονται παρακάτω αφορούν την δεύτερη μέθοδο που χρησιμοποιήσαμε, την Binary Relevance. Για τη συγκεκριμένη μέθοδο, οι τιμές που εμφανίζονται στους παρακάτω πίνακες προήλθαν ύστερα από ξεχωριστούς υπολογισμούς. Πιο συγκεκριμένα, τόσο στο στάδιο της προεπεξεργασίας όσο και στο στάδιο της κατηγοριοποίησης στη συνέχεια μέσω του k-NN, υπολογίστηκαν ξεχωριστά για κάθε κλάση σε κάθε σύνολο δεδομένων ο βαθμός μείωσης(Reduction Rate – RR) που πετυχαίνει κάθε αλγόριθμος όπως και οι αποστάσεις που υπολογίζονται. Η ίδια διαδικασία έγινε και για την κατηγοριοποίηση των στιγμιοτύπων μέσω του k-

ΝΝ πάλι με τρεις διαφορετικές τιμές στο *k*, τις 1,5 και 9. Για κάθε κλάση ξεχωριστά υπολογίστηκαν οι μέσοι όροι οι οποίοι και τοποθετήθηκαν στους παρακάτω πίνακες στα αντίστοιχα πεδία.

RHC								
	Pre-processing		k-NN Classification					
	RR	Distances	Acc(k=1)	EMR(k=1)	Acc(k=5)	EMR(k=5)	Acc(k=9)	EMR(k=9)
Emotions	86.603667	33932	75.8055	22.93	75.352167	18.55	74.9845	17.54
Scene	94.524	278589.2	88.82416	48.23	89.170333	49.81	88.66533	48.73
Yeast	90.801429	297404.114	74.48428	9.72	77.341857	9.39	78.05421	8.48

Πίνακας 9: "Πίνακας αποτελεσμάτων του αλγόριθμου RHC(Binary Relevance)"

ERHC								
	Pre-processing		k-NN Classification					
	RR	Distances	Acc(k=1)	EMR(k=1)	Acc(k=5)	EMR(k=5)	Acc(k=9)	EMR(k=9)
Emotions	89.7545	33932	76.1975	23.44	76.307	20.40	76.165	17.54
Scene	95.5935	278589.2	89.15633	49.52	89.558167	51.10	89.1983	50.64
Yeast	92.5825	297404.114	74.58764	9.81	77.501642	9.60	78.22571	8.65

Πίνακας 10: "Πίνακας αποτελεσμάτων του αλγόριθμου ERHC(Binary Relevance)"

RSP3								
	Pre-processing		k-NN Classification					
	RR	Distances	Acc(k=1)	EMR(k=1)	Acc(k=5)	EMR(k=5)	Acc(k=9)	EMR(k=9)
Emotions	52.2385	295943.833	76.33633	24.28	79.1755	29.34	80.41567	31.20
Scene	69.068833	6843961.25	89.86983	52.76	90.756	55.92	90.86	55.88
Yeast	49.593714	3841028.964	79.09428	18.78	80.646071	18.37	80.64607	17.63

Πίνακας 11: "Πίνακας αποτελεσμάτων του αλγόριθμου RSP3(Binary Relevance)"

CNN								
	Pre-processing		k-NN Classification					
	RR	Distances	Acc(k=1)	EMR(k=1)	Acc(k=5)	EMR(k=5)	Acc(k=9)	EMR(k=9)
Emotions	57.607167	212198.6353	71.07933	15.35	77.4315	23.78	77.57616	22.93
Scene	75.117833	2600823.438	84.933	38.01	89.025	49.65	89.82816	52.01
Yeast	56.428	3870576.214	70.80457	10.67	76.04443	11.42	77.67014	12.62

Πίνακας 12: "Πίνακας αποτελεσμάτων του αλγόριθμου CNN(Binary Relevance)"

IB2								
	Pre-processing		k-NN Classification					
	RR	Distances	Acc(k=1)	EMR(k=1)	Acc(k=5)	EMR(k=5)	Acc(k=9)	EMR(k=9)
Emotions	69.4245	35170.7	69.286	12.65	75.26833	17.88	75.80333	16.86
Scene	82.142667	343305.6667	82.676	31.99	87.467167	43.50	88.43667	47.24
Yeast	67.637857	611235.9857	68.72107	7.28	74.135714	8.81	75.90592	9.35

Πίνακας 13: "Πίνακας αποτελεσμάτων του αλγόριθμου IB2(Binary Relevance)"

AIB2								
	Pre-processing		k-NN Classification					
	RR	Distances	Acc(k=1)	EMR(k=1)	Acc(k=5)	EMR(k=5)	Acc(k=9)	EMR(k=9)
Emotions	75.284167	29033.3335	75.44216	23.44	71.455467	15.35	70.10867	14.17
Scene	87.603	242850.1692	88.5335	47.86	87.89633	46.49	86.22716	41.05
Yeast	75.539357	468008.003	76.82714	14.27	76.659785	6.58	75.5965	5.71

Πίνακας 14: "Πίνακας αποτελεσμάτων του αλγόριθμου AIB2(Binary Relevance)"

conv k-NN						
	k-NN Classification					
	Acc(k=1)	EMR(k=1)	Acc(k=5)	EMR(k=5)	Acc(k=9)	EMR(k=9)
Emotions	75.60116	25.13	79.3433	30.19	80.27283	30.86
Scene	88.59566	47.99	90.5625	55.26	90.73533	55.30
Yeast	75.60392	19.98	79.357428	21.43	80.03714	20.07

Πίνακας 15: "Πίνακας αποτελεσμάτων του αλγόριθμου conv k-NN(Binary Relevance)"

4.4.2 Συμπεράσματα Πειραμάτων για τη Μέθοδο Binary Relevance

Πιο αναλυτικά, στο σύνολο δεδομένων *Emotions*, τα υψηλότερα ποσοστά μείωσης του συνόλου δεδομένων τα πέτυχαν οι αλγόριθμοι ERHC(89.7545), RHC(86.60367) και AIB2(75.2841667). Στο *Scene* αντίστοιχα, οι αλγόριθμοι ERHC(95.5935), RHC(94.524) και AIB2(87.603). Τέλος, στο σύνολο *Yeast* καλύτερη επίδοση είχαν οι ERHC(92.5825), RHC(90.80143) και AIB2(75.5393571).

Στο δεύτερο κομμάτι της προεπεξεργασίας των συνόλων και τον υπολογισμό των αποστάσεων αυτών, οι 3 πρώτοι αλγόριθμοι με τα καλύτερα ποσοστά όσον αφορά το σύνολο δεδομένων *Emotions* είναι οι IB2(35170.7) και οι RHC,ERHC(33932). Στο σύνολο δεδομένων *Scene*, είναι οι RSP3(6843961.25), IB2(34.3305667) και ERHC,RHC(278589.2). Στο *Yeast*, οι περισσότερες αποστάσεις υπολογίστηκαν από τους IB2(61.123598), AIB2(46.8008) και CNN(38.70576).

Για τη κατηγοριοποίηση των παραδειγμάτων και για *k*=1 στο σύνολο δεδομένων *Emotions* η μεγαλύτερη ακρίβεια επιτεύχθηκε από τους RSP3(76.33633), ERHC(76.1975) και AIB2(75.442166). Στο σύνολο *Scene* από τους RSP3(89.8698), ERHC(89.1563) και RHC(88.82416) ενώ στο σύνολο δεδομένων *Yeast* τα υψηλότερα ποσοστά πέτυχαν οι RSP3(79.0942), AIB2(76.8271) και ERHC(74.5876).

Στην περίπτωση που το k είναι 5, τα υψηλότερα ποσοστά ακρίβειας που σημειώθηκαν από τους αλγορίθμους στο σύνολο δεδομένων *Emotions* είναι από τους RSP3(79.1755), CNN(77.4315) και ERHC(76.307). Στο σύνολο δεδομένων *Scene* αντίστοιχα οι αλγόριθμοι με τα καλύτερα ποσοστά

είναι οι RSP3(90.756), ERHC(89.5581) και RHC(89.1703). Τέλος, στο *Yeast* είναι οι αλγόριθμοι RSP3(80.6460), ERHC(77.5016) και RHC(77.3418).

Η τελευταία περίπτωση που εξετάσαμε για τον κατηγοριοποιητή k-NN είναι για k=9 όπου και λάβαμε τα εξής αποτελέσματα. Στο σύνολο δεδομένων Emotions, υψηλότερα ποσοστά καταγράφηκαν από τους RSP3(80.4156), CNN(77.5761) και ERHC(76.165). Για το σύνολο *Scene*, έχουμε τους RSP3(90.86), CNN(89.8281) και ERHC(89.1983) και τέλος για το σύνολο δεδομένων *Yeast*, οι αλγόριθμοι που πέτυχαν τα υψηλότερα ποσοστά είναι οι RSP3(80.6460), ERHC(78.2257) και RHC(78.0542).

Τα συμπεράσματα που μπορούμε να βγάλουμε από τα παραπάνω αποτελέσματα είναι ότι και στα 3 σύνολα δεδομένων για τις διαφορετικές τιμές του k, οι αλγόριθμοι που συμπεριφέρονται με μεγαλύτερη αποτελεσματικότητα και σημείωσαν υψηλότερα ποσοστά, είναι οι RSP3, ERHC και RHC. Τα ποσοστά αυτά είναι σε αρκετές περιπτώσεις βελτιωμένα σε σχέση με την εφαρμογή του k-NN όταν δεν έχει προηγηθεί κάποιος αλγόριθμος μείωσης του συνόλου δεδομένων. Ένα άλλο κοινό χαρακτηριστικό σε όλα τα σύνολα, είναι ότι η ακρίβεια της κατηγοριοποίησης βελτιώνεται σημαντικά όσο μεγαλύτερος είναι ο αριθμός των εγγύτερων γειτόνων που εξετάζουμε.

ΕΠΙΛΟΓΟΣ

Σε αυτό το κεφάλαιο έγινε περιγραφή του περιβάλλοντος στο οποίο πραγματοποιήθηκαν τα πειράματα της πτυχιακής εργασίας. Είδαμε αναλυτικά τα σύνολα δεδομένων που χρησιμοποιήθηκαν καθώς και τα χαρακτηριστικά τους. Στη συνέχεια, παρουσιάστηκαν με λεπτομέρεια οι πίνακες με τα αποτελέσματα της πειραματικής μελέτης και για τις δύο μεθόδους που χρησιμοποιήσαμε. Τέλος, έγινε σχολιασμός και εξαγωγή συμπερασμάτων σχετικά με το ποιοι αλγόριθμοι μείωσης δεδομένων πέτυχαν καλύτερα ποσοστά στη φάση της προεπεξεργασίας των συνόλων, ποιοι αλγόριθμοι πέτυχαν μεγαλύτερα ποσοστά σχετικά με το Exact Match και αν είχαμε καλύτερα ποσοστά ακρίβειας κατά την κατηγοριοποίηση μέσω του k-NN.

Κεφάλαιο 5: Επίλογος

Σκοπός της συγκεκριμένης πτυχιακής εργασίας ήταν η μελέτη συγκεκριμένων αλγορίθμων μείωσης του συνόλου δεδομένων στην περιοχή των multi-label συνόλων.

Στα πλαίσια της εργασίας, είδαμε τις μεθόδους με τις οποίες μπορούμε να μετρήσουμε την απόδοση των κατηγοριοποιητών και πιο συγκεκριμένα μιλήσαμε για τον κατηγοριοποιητή των k Εγγύτερων Γειτόνων(k-NN), τις επιδράσεις που έχει η παράμετρος k στην κατηγοριοποίηση αλλά και τα προβλήματα που μπορούμε να συναντήσουμε.

Επίσης, εγινε λεπτομερής σχολιασμός των αλγορίθμων μείωσης των δεδομένων που χρησιμοποιήσαμε. Αναφερθήκαμε στα πλεονεκτήματα και στα μειονεκτήματα αυτών όπως επίσης και στο κόστος που έχουν οι αλγόριθμοι στην προεπεξεργασία των δεδομένων, είτε υπολογιστικό κόστος είτε χρονικό.

Ακόμα, σχολιάσαμε τις δύο μεθόδους μετασχηματισμού του προβλήματος που εφαρμόσαμε στην πτυχιακή εργασία και είδαμε το τρόπο που λειτουργούν αυτές οι μέθοδοι όπως και τα πλεονεκτήματα και τα μειονεκτήματά τους.

Τέλος, όλα τα παραπάνω μπορέσαμε και τα είδαμε σε εφαρμογή μέσω των πειραμάτων που πραγματοποιήσαμε με σκοπό να δούμε την αποτελεσματικότητα της μείωσης δεδομένων σε σύνολα με πάνω από μία ετικέτες. Καταλήξαμε σε πολύ ικανοποιητικά αποτελέσματα καθώς στις περισσότερες περιπτώσεις των αλγορίθμων για όλα τα σύνολα δεδομένων τα ποσοστά ήταν καλύτερα από αυτά στα οποία δεν είχε γίνει μείωση του συνόλου. Αυτό το γεγονός μπορεί να αποτελέσει πολύ μεγάλη βοήθεια/βελτίωση σε τέτοιου είδους προβλήματα του χώρου μειώνοντας παράλληλα σε σημαντικό βαθμό το κόστος επεξεργασίας τέτοιων προβλημάτων.

Έτσι, μπορούμε με ευκολία να συμπεράνουμε, ότι οι αλγόριθμοι επιλογής και σύνοψης αντιπροσώπων μπορούν να χρησιμοποιηθούν αποτελεσματικά και σε περιπτώσεις συνόλων δεδομένων πολλαπλών ετικετών.

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Τατόγλου, Χ.(2013). Εφαρμογή Αλγορίθμων Μείωσης Όγκου Δεδομένων στη Κατηγοριοποίηση Χρονοσειρών. Πτυχιακή Εργασία, Τμήματος Πληροφορικής Α.Τ.Ε.Ι Θεσσαλονίκη.
2. Bronshtein, A. (2017). A Quick Introduction to K-Nearest Neighbors Algorithm. [online] Medium. Available at: https://medium.com/@adi.bronshtein/a-quick-introduction-to-k-nearest-neighbors-algorithm-62214cea29c7.
3. Cerri, Ricardo & de Carvalho, Andre & Freitas, Alex. (2011). Adapting non-hierarchical multilabel classification methods for hierarchical multilabel classification. Intelligent Data Analysis. 15. 861-887. 10.3233/IDA-2011-0500.
4. Cherman, E., Monard, M. and Metz, J. (2011). Multi-label Problem Transformation Methods: a Case Study. [online] Scielo.edu.uy. Available at: http://www.scielo.edu.uy/scielo.php?pid=S0717-50002011000100005&script=sci_arttext.
5. Gandhi, R. (2018). K Nearest Neighbours — Introduction to Machine Learning Algorithms. [online] Towards Data Science. Available at: https://towardsdatascience.com/k-nearest-neighbours-introduction-to-machine-learning-algorithms-18e7ce3d802a.
6. Jain, S. (2017). Solving Multi-Label Classification problems (Case studies included). [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2017/08/introduction-to-multi-label-classification/.
7. Khade, S. and Balwan, S. (2017). Study and Analysis of Multi-Label Classification Methods in Data Mining. [online] Pdfs.semanticscholar.org. Available at: https://pdfs.semanticscholar.org/a928/fb1f6279a5c05e9a98bea20fe712372118c6.pdf.
8. Nooney, K. (2018). Deep dive into multi-label classification..! (With detailed Case Study). [online] Towards Data Science. Available at: https://towardsdatascience.com/journey-to-the-center-of-multi-label-classification-384c40229bff.
9. Ougiaroglou, Stefanos & Evangelidis, Georgios. (2013). AIB2: An abstraction data reduction technique based on IB2. ACM International Conference Proceeding Series. 13-16. 10.1145/2490257.2490260.

10. Ougiaroglou, S., Diamantaras, K., Evangelidis, G. (2017), Exploring the effect of data reduction on Neural Network and Support Vector Machine Classification, Thessaloniki, GR.

11. Srivastava, T. (2018). Introduction to KNN, K-Nearest Neighbors : Simplified. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/.

12. Tsoumakas, G., Katakis, I. and Vlahavas, I. (2019). Mining Multi-label Data. [online] Lkm.fri.uni-lj.si. Available at: http://lkm.fri.uni-lj.si/xaigor/slo/pedagosko/dr-ui/tsoumakas09-dmkdh.pdf.

13. Ougiaroglou, S. (2014). Algorithms and Techniques for Efficient and Effective Nearest Neighbours Classification, Διδακτορική Διατριβή, Τμήμα Εφαρμοσμένης Πληροφορικής Πανεπιστήμιο Μακεδονίας.