ΔΙΕΘΝΕΣ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΤΗΣ ΕΛΛΑΔΟΣ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Τεχνικές Μείωσης των Δεδομένων Εκπαίδευσης με Διατήρηση των Σπάνιων Κλάσεων

Του φοιτητή:
Κωνσταντίνου Ξουβερούδη
Αρ. Μητρώου: 144175

Επιβλέπων:
Στέφανος Ουγιάρογλου

26 Ιουνίου 2020

Τίτλος Π.Ε: Τεχνικές Μείωσης των Δεδομένων Εκπαίδευσης
με Διατήρηση των Σπάνιων Κλάσεων
Κωδικός Π.Ε: 19065
Ονοματεπώνυμο φοιτητή: Κωνσταντίνος Ξουβερούδης
Ονοματεπώνυμο εισηγητή: Στέφανος Ουγιάρογλου
Ημερομηνία ανάληψης Π.Ε: 6 Δεκεμβρίου 2019
Ημερομηνία περάτωσης Π.Ε: 26 Ιουνίου 2020

Βεβαιώνω ότι είμαι ο συγγραφέας αυτής της εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, έχω καταγράψει τις όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών, εικόνων και κειμένου, είτε αυτές αναφέρονται ακριβώς είτε παραφρασμένες. Επιπλέον, βεβαιώνω ότι αυτή η εργασία προετοιμάστηκε από εμένα προσωπικά, ειδικά ως πτυχιακή εργασία, στο Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του ΔΙ.ΠΑ.Ε.

Η έγκριση της πτυχιακής εργασίας από το Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του Διεθνούς Πανεπιστημίου της Ελλάδος, δεν υποδηλώνει απαραιτήτως και αποδοχή των απόψεων του συγγραφέα, εκ μέρους του Τμήματος.

Πρόλογος

Η παρούσα πτυχιακή εργασία με τίτλο "Τεχνικές Μείωσης των Δεδομένων Εκπαίδευσης με Διατήρηση των Σπάνιων Κλάσεων" συντάχθηκε από τον φοιτητή Κωνσταντίνο Ξουβερούδη στα πλαίσια της ολοκλήρωσης των προϋποθέσεων για την λήψη του πτυχίου του φοιτητή για το τετραετές πρόγραμμα σπουδών από το Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του ΔΙ.ΠΑ.Ε. με έδρα στην Θεσσαλονίκη, το οποίο αντιστοιχεί στο πρόγραμμα σπουδών του πρώην Τμήματος Μηχανικών Πληροφορικής του Α.Τ.Ε.Ι. Θεσσαλονίκης πριν από την συγχώνευση του με το ΔΙ.ΠΑ.Ε. στις αρχές του ακαδημαϊκού έτους 2019-2020.

Η εκπόνηση της εργασίας ξεκίνησε τον Νοέμβριο του 2019 και η ανάληψη της ορίστηκε επίσημα από το ΔΙ.ΠΑ.Ε. τον Δεκέμβριο του 2019, ενώ ολοκληρώθηκε με επιτυχία τον Ιούνιο του 2020. Ως υπεύθυνος και επιβλέπων καθηγητής ορίστηκε ο κ. Στέφανος Ουγιάρογλου, μέλος Ε.ΔΙ.Π. του Τμήματος Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του ΔΙ.ΠΑ.Ε.

Θα ήθελα να ευχαριστήσω τον κ. Ουγιάρογλου, που η αρωγή του ήτανε πολύτιμη κατά την σύνταξη της πτυχιακής μου εργασίας, καθώς και τον κ. Γεώργιο Ευαγγελίδη του Τμήματος Εφαρμοσμένης Πληροφορικής του ΠΑ.ΜΑΚ. και τον κ. Δημήτριο Δέρβο του Τμήματος Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του ΔΙ.ΠΑ.Ε, οι οποίοι μέσω της επιστημονικής τους γνώσης πάνω στο θέμα της εξόρυξης δεδομένων βοήθησαν στην περαιτέρω προώθηση των πειραμάτων της εργασίας.

Επιπλέον, θέλω να ευχαριστήσω την οικογένεια και τους φίλους μου οι οποίοι κατά την διάρκεια αυτής της κρίσιμης περιόδου των σπουδών μου, μου έδειξαν θερμή συμπαράσταση.

Περίληψη

Πολλοί αλγόριθμοι μείωσης του πληθυσμού των δεδομένων εκπαίδευσης (*data reduction*) για προβλήματα κατηγοριοποίησης έχουν προταθεί και είναι διαθέσιμοι στην βιβλιογραφία. Ωστόσο, η εφαρμογή τέτοιων αλγορίθμων δεν ενδείκνυται για σύνολα δεδομένων που παρουσιάζουν μεγάλη ανισοκατανομή κλάσεων, ή "τάξεων" όπως επίσης αποκαλούνται. Τα αντικείμενα που ανήκουν σε σπάνιες κλάσεις (αυτές που έχουν λίγα αντικείμενα), είναι συνήθως σημαντικά αφού μπορεί να ορίζουν για παράδειγμα ακραία καιρικά φαινόμενα ή φυσικές καταστροφές. Για παράδειγμα, σε ένα σύστημα κατηγοριοποίησης για την προφύλαξη καλλιεργειών από το χαλάζι, το σύνολο δεδομένων εκπαίδευσης σίγουρα θα περιλαμβάνει πολλά αντικείμενα που ανήκουν στην κλάση "Μη Χαλαζόπτωση" και ελάχιστα αντικείμενα που ανήκουν στην κλάση "Χαλαζόπτωση". Αυτό είναι ένα σύνολο δεδομένων με μεγάλη ανισοκατανομή κλάσεων. Σε αυτές τις περιπτώσεις, ένας κατηγοριοποιητής που προβλέπει πάντα μη χαλαζόπτωση θα επιτυγχάνει υψηλή ακρίβεια αλλά στην πραγματικότητα είναι ακατάλληλος. Σε αυτό το πρόβλημα, ο αλγόριθμος θα πρέπει να προβλέπει σωστά την χαλαζόπτωση. Εξάλλου είναι σημαντικότερο για τους γεωργούς το να προβλεφθεί χαλαζόπτωση και τελικά να μην πραγματοποιηθεί παρά το αντίθετο. Στην μία περίπτωση θα προφυλάξουν, όπως θα αποδειχθεί χωρίς λόγο, τις καλλιέργειες τους, ενώ στην άλλη, θα καταστραφούν οι καλλιέργειες.

Η απόδοση των κατηγοριοποιητών σε τέτοιου είδους προβλήματα εκτιμάται υπολογίζοντας την ορθότητα (*precision*) και την ευαισθησία (*recall*). Αν επιχειρηθεί εφαρμογή αλγορίθμων μείωσης του πληθυσμού των δεδομένων σε τέτοιου είδους σύνολα δεδομένων, οι σπάνιες κλάσεις θα εξασθενήσουν ακόμη περισσότερο ενώ είναι πιθανή η ολοκληρωτική εξάλειψη των αντικειμένων που ανήκουν σε αυτές.

Στα πλαίσια της παρούσας πτυχιακής εργασίας, ο φοιτητής παρουσιάζει την σχετική βιβλιογραφία και εκτελεί πειράματα για την εκτίμηση της απόδοσης των αλγορίθμων μείωσης του πληθυσμού των δεδομένων σε σύνολα δεδομένων εκπαίδευσης με ανισοκατανομή κλάσεων. Προτείνονται τρεις διαφορετικές προσεγγίσεις, οι οποίες ονομάστηκαν Μέθοδοι Διατήρησης Σπάνιων Κλάσεων (*Rare Class Preservation Methods*), μέσω των οποίων επιδιώκεται η βελτιστοποίηση των αποτελεσμάτων αυτών των πειραμάτων. Η πρώτη μέθοδος είναι η RCPM1, η οποία εκτελεί μείωση δεδομένων μόνο στις μη-σπάνιες κλάσεις του συνόλου εκπαίδευσης. Επιπλέον έχουμε και την RCPM2, η οποία εκτελεί μείωση δεδομένων σε όλο το σύνολο εκπαίδευσης, αλλά έπειτα αντικαθιστά τα αντικείμενα των σπάνιων κλάσεων που έμειναν στο συμπυκνωμένο σύνολο με όλα τα αντίστοιχα αντικείμενα τους στο σύνολο εκπαίδευσης. Τέλος, υπάρχει και η RCPM-SMOTE, η οποία πριν την κατηγοριοποίηση χρησιμοποιεί την μέθοδο SMOTE για να εξισορροπήσει την κατανομή των κλάσεων.

Αυτές οι μέθοδοι δοκιμάστηκαν χρησιμοποιώντας δώδεκα σύνολα δεδομένων και τα πειραματικά αποτελέσματα έδειξαν πως τα αντικείμενα των σπάνιων κλάσεων αποφεύγουν τον κίνδυνο της πιθανής εξάλειψης, ενώ παράλληλα στις περισσότερες περιπτώσεις αυξάνεται το ποσοστό της ευαισθησίας.

Abstract

Many data reduction algorithms for training data regarding classification problems have been suggested and are available in the bibliography. However, the usage of these algorithms is not deemed appropriate for datasets that showcase significant class imbalance. Any items that belong to rare classes (which is what we call the classes with too few items), are usually important since they can, for example, define extreme weather events or natural disasters. For example, in a classification system meant to help protect crops from hail, the dataset used will surely include many items that belong in the "Not Hail" class and very few items that belong in the "Hail" class. This is a dataset with high class imbalance. In such cases, a classifier that always predicts "Not Hail" will always achieve high accuracy, yet will prove to be unreliable. For this problem, the algorithm needs to predict the cases of hail correctly. That's because to the farmers it's preferable to predict that there will be hail and in the end to not actually have any. In one case they will protect, admittedly for no reason, their crops, while in the other their crops will be destroyed.

In problems such as these, the classifiers' performance is estimated by calculating the precision, as well as recall. If attempts are made to implement data reduction algorithms on datasets of this kind, the amount of items in rare classes will be reduced even further. In addition, it is also possible that the rare classes might be extinct altogether.

In this thesis paper, the student showcases the relevant bibliography and conducts experiments in order to estimate the performance of data reduction algorithms on the training sets of datasets that contain imbalanced classes. In order to improve the results of said experiments, three different methods, which were named Rare Class Preservation Methods, are proposed. The first is RCPM1, which uses data reduction only on the non-rare classes of a training set. In addition there is also RCPM2, which uses data reduction on the entire training set as normal, but then replaces any remaining items of the rare classes in the condensing set with all their respective items in the training set. Finally there is RCPM-SMOTE, which uses the SMOTE method to balance the class distribution before the classification.

These methods were tested on twelve different datasets and the experimental results showed that the rare class items manage to successfully avoid the danger of being possibly deleted. In addition, in most cases recall is increased.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κατάλογος Αλγορίθμων

1 Εισαγωγή

1.1 Κατηγοριοποίηση Δεδομένων

Στο κόσμο της εξόρυξης δεδομένων διατίθενται σύνολα δεδομένων από τα οποία μπορούν να εξαχθούν χρήσιμες πληροφορίες. Αυτές οι πληροφορίες μπορούν να έχουν ένα μεταβαλλόμενο βαθμό σημαντικότητας, για παράδειγμα ένα σύνολο δεδομένων μπορεί να χρησιμοποιείται για να βοηθήσει στην πρόβλεψη του καιρού ενώ ένα άλλο σύνολο μπορεί να επιχειρεί να προβλέψει ποιος συνδυασμός χαρτιών είναι πιο πιθανό να προκύψει σε ένα παιχνίδι πόκερ. Τα αποτελέσματα αυτών, καθώς και άλλων παρόμοιων θεμάτων τα αποκτάμε μέσω της κατηγοριοποίησης των διαθέσιμων δεδομένων.

Με τον όρο Κατηγοριοποίηση [1] αναφερόμαστε στην διαδικασία της ανάθεσης κλάσεων σε μη-κατηγοριοποιημένα δεδομένα. Η επιλογή των κλάσεων, πραγματοποιείται μέσω διαθέσιμων δεδομένων εκπαίδευσης (*training data*).

Η Κατηγοριοποίηση είναι ένα βασικό μέρος πολλαπλών εφαρμογών του σημερινού κόσμου. Για παράδειγμα, οι εφαρμογές που χρησιμοποιούνται από τραπεζικές και ασφαλιστικές υπηρεσίες διαθέτουν μία υλοποιημένη διαδικασία ανίχνευσης απατών (*fraud detection*), έτσι ώστε να προστατεύονται από πιθανές κλοπές. Ένα επιπλέον, πιο απλό παράδειγμα είναι πως μπορεί να έχουμε μία σειρά φωτογραφιών με σκυλιά και να θέλουμε να καθορίσουμε τι ράτσα είναι το καθένα τους.

Αν και όλοι οι αλγόριθμοι κατηγοριοποίησης μοιράζονται τον ίδιο σκοπό, χωρίζονται σε δύο διαφορετικές κατηγορίες. Η κατηγορία στην οποία ανήκει ο κάθε αλγόριθμος εξαρτάται από τον τρόπο που λειτουργεί. Αυτές οι κατηγορίες είναι: (*i*) πρώτον οι πρόθυμοι κατηγοριοποιητές (*eager classifiers*), οι οποίοι επεξεργάζονται τα δεδομένα εκπαίδευσης πριν την κατηγοριοποίηση και μέσω αυτών δημιουργούν ένα μοντέλο κατηγοριοποίησης το οποίο χρησιμοποιείται για την εκπαίδευση των δεδομένων που δεν έχουν κατηγοριοποιηθεί ακόμα. (*ii*) Επιπλέον, υπάρχουν οι οκνηροί κατηγοριοποιητές (*lazy classifiers*), οι οποίοι δεν χτίζουν κάποιο μοντέλο, αλλά αναλύουν εκ νέου τα δεδομένα εκπαίδευσης κάθε φορά που λαμβάνουν ένα καινούριο μη-κατηγοριοποιημένο αντικείμενο έτσι ώστε να καθορίσουν σε ποια κλάση ανήκει. Οι οκνηροί κατηγοριοποιητές είναι επίσης γνωστοί και ως "κατηγοριοποιητές βασισμένοι σε στιγμιότυπα" (*instance based classifiers*). Προφανώς, τα δεδομένα εκπαίδευσης είναι ένα βασικός και σημαντικός παράγοντας που χρειάζεται κάθε αλγόριθμος κατηγοριοποίησης. Τα "στιγμιότυπα" είναι ένας εναλλακτικός όρος για τα αντικείμενα.

Και οι δύο κατηγορίες έχουν τα πλεονεκτήματα και τα μειονεκτήματα τους. Λόγω του μοντέλου που παράγουν, οι πρόθυμοι αλγόριθμοι εκτελούν την κατηγοριοποίηση πολύ πιο γρήγορα. Αξίζει να σημειωθεί πως αφού χτιστεί το μοντέλο, τα δεδομένα εκπαίδευσης μπορούν να αφαιρεθούν από την μνήμη καθώς δεν χρειάζονται πια. Όμως, ως αντάλλαγμα

οι κατηγοριοποιητές πρέπει να χρησιμοποιούν αυτό το μοντέλο για όλα τα δεδομένα. Λόγω αυτού, υπάρχει η περίπτωση να χρειαστεί να χτίσουν ένα υπερβολικά πολύπλοκο μοντέλο για συγκεκριμένα σύνολα δεδομένων, κάτι το οποίο είναι εξαιρετικά χρονοβόρο.

Αντίθετα, χρησιμοποιώντας όλα τα δεδομένα εκπαίδευσης για κάθε καινούριο αντικείμενο που πρέπει να κατηγοριοποιηθεί, οι οκνηροί κατηγοριοποιητές μπορούν να εφαρμόσουν πιο πολύπλοκους υπολογισμούς. Έτσι, αν και οι κατηγοριοποιήσεις τους απαιτούν περισσότερο υπολογιστικό κόστος και είναι γενικά πιο χρονοβόρες αφού δεν χρησιμοποιείται κάποιο μοντέλο κατηγοριοποίησης, μπορούν να πετύχουν υψηλή ακρίβεια. Το μειονέκτημα των οκνηρών κατηγοριοποιητών είναι η ανάγκη να υπάρχει αρκετή ελεύθερη μνήμη στον υπολογιστή, καθώς πρέπει να έχει όλα τα δεδομένα εκπαίδευσης διαθέσιμα κατά την διάρκεια της κατηγοριοποίησης και φυσικά, το υψηλό υπολογιστικό κόστος κατά την διάρκεια της κατηγοριοποίησης. Στα πλαίσια αυτής της εργασίας κλείνουμε προς την χρήση οκνηρών κατηγοριοποιητών, πιο συγκεκριμένα τον κατηγοριοποιητή των κ εγγύτερων γειτόνων (k-Nearest Neighbours classifier).

1.2 Instance-Based Classification (k-NN)

Ο k-Nearest Neighbours (***k-NN***) [2, 3] είναι ένας ευρέως γνωστό κατηγοριοποιητής. Λόγω της απλότητας του, καθώς και του γεγονότος ότι παράγει καλύτερα αποτελέσματα από πολλούς άλλους κατηγοριοποιητές, χρησιμοποιείται σε πολλές εφαρμογές.

Ο k-NN λειτουργεί ως εξής: κάθε φορά που πρέπει να κατηγοριοποιήσει ένα νέο αντικείμενο x, ψάχνει να βρει τα k πιο κοντινά αντικείμενα (οι λεγόμενοι γείτονες) που περιέχονται στα δεδομένα εκπαίδευσης, σύμφωνα με μια μετρική απόστασης. Μετά αναλύει σε ποια κλάση ανήκουν οι περισσότεροι από αυτούς τους k γείτονες και το x εκχωρείται σε αυτήν. Επαναλαμβάνει αυτήν την διαδικασία για κάθε νέο αντικείμενο.

Ο τρόπος με τον οποίο υπολογίζεται η απόσταση εξαρτάται από τον τύπο των δεδομένων. Στα πλαίσια αυτής της εργασίας, τα σύνολα δεδομένων αποτελούνται όλα από ακέραιους και πραγματικούς αριθμούς. Επομένως, η πιο κατάλληλη συνάρτηση για τον υπολογισμό της απόστασης είναι η Ευκλείδεια μετρική (*Euclidean distance*):

$$d(p,q) = d(q,p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + ... + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \qquad (1.1)$$

όπου το p και το q είναι δύο σημεία μέσα στο σύνολο δεδομένων και n είναι ο αριθμός των αντικειμένων

Όμως, ακόμα και αν όλα τα δεδομένα βρίσκονται σε μορφή ακέραιων ή πραγματικών αριθμών, υπάρχει και το επιπλέον ζήτημα του εύρους. Για παράδειγμα, μπορεί να έχουμε

δύο διαφορετικά αντικείμενα τα οποία έχουν το ίδιο βάρος, έστω ο μισθός και ο αριθμός παιδιών ενός ατόμου. Σε αυτήν την περίπτωση το εύρος τιμών του μισθού θα είναι πολύ μεγαλύτερο και λόγω αυτού θα συνεισφέρει περισσότερο στον υπολογισμό της απόστασης.

Για αυτόν τον λόγο, συνιστάται η εκτέλεση μία διαδικασίας κανονικοποίησης (*normalization*), η οποία θα επεξεργαστεί ανάλογα τις τιμές των δεδομένων για να έχουν όλες το ίδιο εύρος:

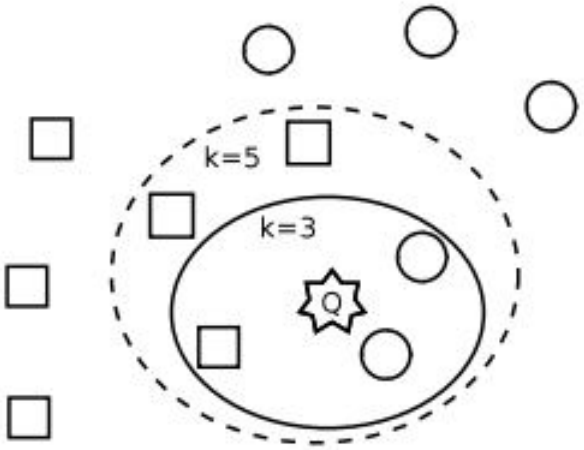
$$normalized(e_i) = \frac{e_i - E_{min}}{E_{max} - E_{min}} \tag{1.2}$$

όπου e_i είναι ένα χαρακτηριστικό (*attribute*) του i-οστού αντικειμένου του συνόλου δεδομένων, $i = 1, 2, ..., n$, και τα E_{max}, E_{min} είναι η μεγαλύτερη και η μικρότερη τιμή που έχει πάρει το χαρακτηριστικό e αντίστοιχα.

Στο Σχήμα 1.1 μπορούμε να δούμε με γραφικό τρόπο παράδειγμα της εκτέλεσης του k-NN. Έστω ότι έχουμε ένα σύνολο δεδομένων με αντικείμενα που ανήκουν σε δύο διαφορετικές κλάσεις: την κλάση Τετράγωνο και την κλάση Κύκλος. Έχουμε ένα αντικείμενο Q που πρέπει να κατηγοριοποιηθεί σε μία από τις δύο κλάσεις. Στην μία περίπτωση θέτουμε $k = 3$ και από τους τρεις κοντινότερους γείτονες, οι δύο είναι κύκλοι. Επομένως το Q θα κατηγοριοποιηθεί ως κύκλος. Εναλλακτικά, αν θέσουμε το $k = 5$, τότε οι περισσότεροι από τους πέντε κοντινότερους γείτονες είναι τετράγωνα, άρα το Q κατηγοριοποιείται ως τετράγωνο.

Βέβαια, το ερώτημα που τίθεται είναι ποια είναι η βέλτιστη τιμή για το k. Στην πραγματικότητα, η καλύτερη τιμή που μπορούμε να θέσουμε μπορεί να διαφέρει ανάλογα με το σύνολο δεδομένων που πρέπει να αναλυθεί. Για παράδειγμα, αν έχουμε ένα σύνολο δεδομένων όπου παρατηρείται ένας μεγάλος αριθμός δεδομένων που αποτελούν θόρυβο [4], τότε προτιμάται μία μεγάλη τιμή για το k, αφού θα μας επιτρέψει να αναλύσουμε περισσότερους γείτονες και πιθανόν να ξεχωρίσουμε τα δεδομένα που είναι "σωστά" από τα αυτά που αποτελούν θόρυβο. Ωστόσο, επιλέγοντας μία υψηλή τιμή για το k, δεν μπορούμε να ξεχωρίσουμε το ίδιο καλά τα σύνορα μεταξύ των κλάσεων. Αυτή είναι μόνο μία από τις πολλές πιθανές περιπτώσεις. Η βέλτιστη τιμή για το k συνήθως βρίσκεται δοκιμάζοντας πολλές τιμές και διαλέγοντας αυτήν που παράγει το καλύτερο αποτέλεσμα. Αξίζει να σημειωθεί ότι ένας παρόμοιος αλγόριθμος, ο 1-NN, είναι στην πραγματικότητα ο k-NN με $k = 1$.

Στην περίπτωση που κατά την κατηγοριοποίηση ενός αντικειμένου, βρούμε ίσο αριθμό γειτόνων για δύο ή περισσότερες διαφορετικές κλάσεις, ο k-NN διαλέγει σε ποια κλάση θα κατηγοριοποιηθεί το νέο αντικείμενο είτε τυχαία, είτε διαλέγοντας την ίδια κλάση που έχει ο πιο κοντινός γείτονας από όλους. Η μόνη περίσταση όπου δεν ισχύει αυτό είναι όταν έχουμε μόνο δύο κλάσεις στο σύνολο δεδομένων. Σε αυτήν την περίπτωση, θα πρέπει να θέσουμε έναν περιττό αριθμό ως τιμή για το k έτσι ώστε να αποφύγουμε κάθε περίπτωση "ισοπαλίας".



Σχήμα 1.1: Παράδειγμα της διαδικασίας του k-Nearest Neighbours με $k = 3$ και $k = 5$.

Παρά την χρησιμότητα του όμως, ο k-NN έχει και διάφορα μειονεκτήματα. Ένα από αυτά είναι πως μπορεί να χρησιμοποιήσει μόνο μία τιμή για το k. Υπάρχει περίπτωση όμως το σύνολο δεδομένων να είναι ασταθής, και να χρειάζεται άλλες τιμές για το k σε διαφορετικές περιοχές του συνόλου δεδομένων που έχουν σημαντικά διαφορετική δομή για να παράγει αυτό που θεωρείται πως είναι το βέλτιστο αποτέλεσμα.

Εκτός αυτού υπάρχει και το ζήτημα του υπολογιστικού κόστους, το οποίο είναι πιθανόν και το πιο σημαντικό μειονέκτημα από όλα. Το k-NN πρέπει αναγκαστικά να υπολογίσει όλες τις αποστάσεις μεταξύ κάθε μη-κατηγοριοποιημένου αντικειμένου και όλων των αντικειμένων μέσα στο σύνολο δεδομένων εκπαίδευσης. Σαφώς, αυτό σημαίνει πως όταν έχουμε μεγάλα σύνολα δεδομένων χάνεται πολύς χρόνος στους υπολογισμούς αυτούς. Για παράδειγμα, έστω ότι έχουμε ένα σύνολο δεδομένων με 200.000 αντικείμενα και σκοπεύουμε, μέσω αυτών, να κατηγοριοποιήσουμε 50.000 αντικείμενα χρησιμοποιώντας τον k-NN. Επομένως, θα χρειαστεί να γίνουν συνολικά δέκα δισεκατομμύρια υπολογισμοί, κάτι το οποίο είναι προφανώς υπερβολικά χρονοβόρο ακόμα και για υπολογιστές με δυνατούς επεξεργαστές. Πέρα αυτού, το υπολογιστικό κόστους μπορεί να αυξηθεί και σε περιπτώσεις όπου έχουμε δεδομένα με μεγάλο αριθμό διαστάσεων (*data dimensionality*).

Η πηγή ενός άλλου μειονεκτήματος είναι ο θόρυβος που αναφέραμε πριν. Ο κατηγοριοποιητής k-NN είναι πολύ ευαίσθητος σε δεδομένα που αποτελούν θόρυβο, ειδικά όταν το k έχει χαμηλή τιμή. Αυτό μπορεί να οδηγήσει σε κατηγοριοποιήσεις με μικρότερη ακρίβεια. Για να το αποφύγουμε, μπορούμε να δοκιμάσουμε να βάλουμε ένα μεγαλύτερο k, επιδιώκοντας υψηλότερη ακρίβεια.

Τέλος, ο k-NN πρέπει να έχει όλα τα δεδομένα εκπαίδευσης διαθέσιμα κατά την εκτέλεση του, σε αντίθεση με άλλους αλγόριθμους που ανήκουν στην κατηγορία των πρόθυμων κατηγοριοποιητών οι οποίοι μπορούν να έχουν την πολυτέλεια να αφαιρούν δεδομένα αφού τα

αναλύσουν. Επομένως, για να χρησιμοποιήσουμε τον k-NN χρειαζόμαστε και υπολογιστές με πολύ μεγάλη μνήμη, έτσι ώστε να έχουμε όλα τα αντικείμενα του συνόλου δεδομένων που χρησιμοποιούμε συνεχώς διαθέσιμα.

Ωστόσο, μπορούμε να αντιμετωπίσουμε τα προβλήματα του k-NN με μία πολύ συγκεκριμένη τεχνική, την Μείωση Δεδομένων (*Data Reduction*) [5, 6, 7, 8, 9, 10]. Πρόκειται για μία τεχνική η οποία όπως προτείνει και το όνομα της, συρρικνώνει ένα σύνολο δεδομένων, επιχειρώντας να κρατήσει μόνο τα πιο σημαντικές πληροφορίες τους και να ξεφορτωθεί τυχόν περιττά δεδομένα. Χρησιμοποιείται σε πολυπληθής σύνολα δεδομένων έτσι ώστε να μειωθεί το υπολογιστικό κόστος της κατηγοριοποίησης. Υπάρχει ένας αρκετά μεγάλος αριθμός αλγορίθμων που πραγματοποιούν μείωση δεδομένων, μερικοί από τους οποίος χρησιμοποιούνται και σε αυτήν την εργασία. Τέτοιοι αλγόριθμοι ονομάζονται Τεχνικές Μείωσης Δεδομένων, ή αλλιώς ΤΜΔ (*Data Reduction Techniques, DRTs*). Οι ΤΜΔ θα αναλυθούν περισσότερο παρακάτω.

1.3 Μη-Ισορροπημένα Δεδομένα (Imbalanced Data)

Το θέμα της ανισοκατανομής κλάσεων [11] αναφέρεται σε περιπτώσεις όπου σε ένα σύνολο δεδομένων, ο αριθμός των διαθέσιμων αντικειμένων για μία ή περισσότερες κλάσεις είναι σημαντικά μικρότερος σε σύγκριση με άλλες. Πρόκειται για ένα πρόβλημα που παρουσιάζεται συχνά σε εφαρμογές διάφορων ιδιοτήτων, όπως είναι για παράδειγμα οι προγνώσεις καιρού ή η ανίχνευση απάτης (*fraud detection*).

Ας σκεφτούμε για παράδειγμα τι μπορεί να γίνει με ένα σύστημα πρόγνωσης καιρού εάν χρησιμοποιεί ένα μη-ισορροπημένο σύνολο δεδομένων. Όπως ξέρουμε, το χαλάζι είναι ένα εξαιρετικά σπάνιο καιρικό φαινόμενο. Επομένως, αν το σύνολο δεδομένων έχει μία κλάση για κάθε πιθανό καιρό (λιακάδα, βροχή, χιόνι, κ.τ.λ.), η κλάση "Χαλάζι" θα περιέχει πολύ λιγότερα αντικείμενα σε σχέση με τις άλλες. Κάτι τέτοιο μπορεί να βλάψει σημαντικά την κατηγοριοποίηση, καθώς πιθανότατα δεν θα προβλέπονται σωστά οι χαλαζοπτώσεις και ορισμένες ομάδες ανθρώπων, όπως για παράδειγμα οι αγρότες, θα πάθουν μεγάλη ζημιά, καθώς θα χάσουν τις καλλιέργειες τους από τον καιρό επειδή δεν προειδοποιήθηκαν πως πρέπει να τις προστατεύσουν.

1.4 Κίνητρο και Συνεισφορά

Το κεντρικό πρόβλημα που παρουσιάζουν τα μη-ισορροπημένα σύνολα δεδομένων, είναι πως εφόσον εφαρμοστεί ένας αλγόριθμος μείωσης δεδομένων πάνω τους, δεν θα καταφέρουμε να εκτιμήσουμε σωστά πόσο σημαντικές είναι οι κλάσεις με μικρότερο αριθμό αντικειμένων, ή σπάνιες κλάσεις (*rare classes*) για συντομία, καθώς τα αντικείμενα τους είτε θα μειωθούν

ακόμα περισσότερο, είτε θα εξαλειφθούν τελείως. Το γεγονός αυτό αποτελεί το κίνητρο της παρούσας πτυχιακής εργασίας.

Συνεπώς, η παρούσα πτυχιακή εργασία έχει εκπονηθεί με σκοπό την αντιμετώπιση αυτού του θέματος. Για αυτόν τον λόγο, η εργασία προτείνει νέες μεθόδους που επιχειρούν να εξισορροπήσουν τα σύνολα δεδομένων. Συγκεκριμένα προτείνονται δύο διαφορετικές μέθοδοι που διαχωρίζουν τις σπάνιες κλάσεις από τις υπόλοιπες. Έτσι, επιχειρούν να χρησιμοποιήσουν μία ΤΜΔ μόνο πάνω στα αντικείμενα που ανήκουν σε μεγαλύτερες κλάσεις. Με αυτόν τον τρόπο τα αντικείμενα των σπάνιων κλάσεων δεν μειώνονται περαιτέρω. Συνεπώς, με αυτόν τον τρόπο επιτυγχάνεται εξισορροπούνται οι κλάσεις του συνόλου δεδομένων. Για να κατασκευαστεί το τελικό σύνολο, στο οποίο θα πραγματοποιηθεί η διαδικασία κατηγοριοποίησης k-NN, οι μέθοδοι συνενώνουν το υποσύνολο με τον μειωμένο αριθμό αντικειμένων των μεγαλύτερων κλάσεων, με αυτό των σπάνιων κλάσεων το οποίο δεν έχει τροποποιηθεί. Επιπρόσθετα, η παρούσα εργασία προτείνει μια τρίτη μέθοδος που αξιοποιεί την μέθοδο SMOTE [12] ώστε να ενισχυθούν οι σπάνιες κλάσεις και στην συνέχεια εφαρμόζονται οι τεχνικές μείωσης των δεδομένων.

Στην πειραματική μελέτη που εκτελέστηκε στα πλαίσια αυτής της εργασίας, εκτός από την ακρίβεια (*accuracy*) της κατηγοριοποίησης, υπολογίζονται επίσης οι τιμές της ορθότητας (*precision*), της ευαισθησίας (*recall*) καθώς και οι τιμές του F-Measure. Μέσω αυτών θα μπορέσουμε να εκτιμήσουμε καλύτερα την απόδοση που επιτυγχάνει ο κατηγοριοποιητής στις σπάνιες κλάσεις και τελικά ποια από τις προτεινόμενες μεθόδους είναι καλύτερη. Και τα τέσσερα κριτήρια που μόλις αναφέρθηκαν αναλύονται παρακάτω.

1.5 Οργάνωση της Πτυχιακής

Στο Κεφάλαιο 2 αναλύονται όλες οι έννοιες εξόρυξης δεδομένων που είναι σχετικές με το θέμα της πτυχιακής εργασίας. Το κεφάλαιο χωρίζεται σε δύο μέρη. Στο πρώτο μέρος αναλύονται οι πολλαπλοί τρόποι μέτρησης της αποτελεσματικότητας των μεθόδων κατηγοριοποίησης, τα είδη μεθόδων κατηγοριοποίησης που υπάρχουν καθώς και μία εξήγηση του λόγου που, για τα πειράματα που εκτελέστηκαν για τους σκοπούς αυτής της εργασίας, επιλέχθηκαν σύνολα δεδομένων που ανήκουν σε ένα συγκεκριμένο από αυτά τα είδη. Τέλος, παρουσιάζονται μερικοί από τους πιο γνωστούς τρόπους με τους οποίους αντιμετωπίζεται το πρόβλημα της ανισοκατανομής δεδομένων. Στο δεύτερο μέρος γίνεται μία λεπτομερή ανάλυση των ΤΜΔ που χρησιμοποιήθηκαν στα πειράματα μας. Παράλληλα δίνεται ο ψευδοκώδικας του κάθε αλγορίθμου.

Στο Κεφάλαιο 3 παρουσιάζεται μία ανάλυση της συλλογής των προτεινόμενων μεθόδων για την εξισορρόπηση της ανισοκατανομής κλάσεων σε ένα σύνολο δεδομένων εκπαίδευσης και την εφαρμογή τεχνικών μείωσης δεδομένων. Οι δύο πρώτες χρησιμοποιούν μία μέθοδο

που ξεχωρίζει τις σπάνιες κλάσεις από τις υπόλοιπες, η οποία αναπτύχθηκε στα πλαίσια της παρούσας εργασίας. Η τρίτη μέθοδος χρησιμοποιεί την SMOTE για να εξισορροπήσει τα δεδομένα, η οποία είναι διαθέσιμη στην βιβλιογραφία.

Στο Κεφάλαιο 4 δίνεται μία σύντομη περιγραφή των συνόλων δεδομένων που επιλέχτηκαν για τα πειράματα μας. Στην συνέχεια παρουσιάζονται τα πειραματικά αποτελέσματα κάθε μεθόδου με κάθε ξεχωριστό ΤΜΔ μέσω μίας σειράς αναλυτικών πινάκων που περιέχουν το ποσοστό μείωσης, την ακρίβεια, την ορθότητα, την ευαισθησία και το F-Measure. Έπειτα, τα αποτελέσματα συγκρίνονται και αξιολογούνται.

Τέλος, στο Κεφάλαιο 5 δίνεται το τελικό συμπέρασμα που σχηματίστηκε μέσω των πειραμάτων που εκτελέστηκαν σε αυτήν την εργασία ενώ παράλληλα παρουσιάζονται κατευθύνσεις για μελλοντική έρευνα.

2 Γνωστικό Υπόβαθρο

2.1 Εισαγωγή

Σε αυτό το κεφάλαιο θα επικεντρωθούμε στις έννοιες και τις μεθόδους που έχουν ήδη παρουσιαστεί λεπτομερώς από διάφορους ερευνητές και χρησιμοποιούνται για στην εκπόνηση της παρούσας εργασίας. Το κεφάλαιο έχει χωριστεί σε δύο μέρη.

Το πρώτο μέρος είναι η Ενότητα 2.2, όπου αναλύεται το πρόβλημα που παρουσιάζουν τα μη-ισορροπημένα σύνολα δεδομένων. Παρουσιάζονται τα διάφορα είδη μεθόδων κατηγοριοποίησης, καθώς και τα κριτήρια μέσω των οποίων μπορεί να μετρηθεί η αποτελεσματικότητα τους. Επιπλέον, εξηγείται ποια από αυτά τα κριτήρια είναι πιο χρήσιμα σε περιπτώσεις που παρατηρείται το προαναφερόμενο πρόβλημα. Τέλος, γίνεται μία σύντομη περιγραφή της υπερδειγματοληψίας, της υποδειγματοληψίας και της μεθόδου SMOTE, οι οποίες αποτελούν μεθόδους αντιμετώπισης αυτής της ανισοκατανομής των δεδομένων.

Το δεύτερο μέρος είναι η Ενότητα 2.3, η οποία αφορά στις Τεχνικές Μείωσης Δεδομένων (ΤΜΔ). Ξεκινάει με μία ανάλυση των διάφορων διαθέσιμων κατηγοριών ΤΜΔ που υπάρχουν και έπειτα παρουσιάζονται οι συγκεκριμένες τεχνικές που χρησιμοποιήθηκαν στα πειράματα της παρούσας εργασίας.

2.2 Ανάλυση του Προβλήματος

2.2.1 Μετρηση της Αποτελεσματικότητας

Στην μηχανική μάθηση και στην εξόρυξη γνώσης από δεδομένα, ο στόχος της κατηγοριοποίησης είναι να χτίσει ένα σύστημα το οποίο μπορεί να "προβλέψει" σε ποια κλάση ανήκει κάθε αντικείμενο σε ένα σύνολο δεδομένων. Ο αριθμός των προβλέψεων είναι τεράστιος, οπότε θεωρητικά είναι αδύνατον να είναι όλες οι προβλέψεις του σωστές. Επομένως, ένας κατηγοριοποιητής θεωρείται αποτελεσματικός εφόσον καταφέρει να πετύχει ένα μεγάλο ποσοστό ακρίβειας (*accuracy)*, δηλαδή να προβλέψει σωστά την κλάση αρκετών αντικειμένων. Για να μετρηθεί η ακρίβεια, πρέπει να χωριστεί το σύνολο δεδομένων σε δύο υποσύνολα: το σύνολο εκπαίδευσης (*training set*) και το σύνολο δοκιμών (*testing set*). Το μοντέλο που χρησιμοποιείται, στην περίπτωση μας, ο k-NN, για κάθε αντικείμενο που πρέπει να κατηγοριοποιηθεί, θα αναλύσει πρώτα το σύνολο εκπαίδευσης. Στην ουσία, με αυτόν τον τρόπο, "μαθαίνει" πως να κατηγοριοποιεί σωστά. Το σύνολο δοκιμών περιέχει τα αντικείμενα για τα οποία προσπαθούμε να προβλέψουμε τις κλάσεις.

Αν συγκρίνουμε τα αποτελέσματα των προβλέψεων της κατηγοριοποίησης με τις πραγμα-

τικές κλάσεις στις οποίες ανήκουν τα δεδομένα του συνόλου δοκιμών, μπορούμε για κάθε αντικείμενο να βρούμε ένα από τα παρακάτω τέσσερα αποτελέσματα:

- True Positive (TP): Ο κατηγοριοποιητής πρόβλεψε πως το αντικείμενο ανήκει σε μία κλάση x και στην πραγματικότητα όντως ανήκει σε αυτήν.
- True Negative (TN): Ο κατηγοριοποιητής πρόβλεψε πως το αντικείμενο δεν ανήκει σε μία κλάση x και στην πραγματικότητα όντως δεν ανήκει σε αυτήν.
- False Positive (FP): Ο κατηγοριοποιητής πρόβλεψε πως το αντικείμενο ανήκει σε μία κλάση x, αλλά στην πραγματικότητα δεν ανήκει σε αυτήν.
- False Negative (FN): Ο κατηγοριοποιητής πρόβλεψε πως το αντικείμενο δεν ανήκει σε μία κλάση x, αλλά στην πραγματικότητα ανήκει σε αυτήν.

Επιπλέον, τα παραπάνω κριτήρια μπορούν να χρησιμοποιηθούν για να υπολογίσουμε την ακρίβεια που αναφέραμε πριν. Η ακρίβεια αποτελεί τον αριθμό των σωστά κατηγοριοποιημένων αντικειμένων προς το συνολικό αριθμό των προβλέψεων που έκανε ο κατηγοριοποιητής:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.1}$$

2.2.2 Είδη Κατηγοριοποίησης

Κάθε σύνολο δεδομένων μπορεί να κατηγοριοποιεί τα δεδομένα με έναν από τους τρεις παρακάτω τρόπους: Binary Classification, Multi-class Classification [13] και Multi-label Classification [14, 15].

Οι πρώτες δύο περιπτώσεις είναι άμεσα συσχετιζόμενες μεταξύ τους. Το Binary Classification [13], ή αλλιώς η Δυαδική Κατηγοριοποίηση αναφέρεται σε περιπτώσεις όπου έχουμε μόνο δύο κλάσεις με τις οποίες κατηγοριοποιούμε τα δεδομένα. Για παράδειγμα αν έχουμε ένα σύνολο δεδομένων με σκύλους στο οποίο γίνεται κατηγοριοποίηση με βάση το φύλο τους, τότε έχουμε δύο κλάσεις "Αρσενικό" και "Θηλυκό". Αυτό το σύνολο θεωρείται Δυαδικό.

Αντίθετα, το Multi-class Classfication [13], ή αλλιώς η Κατηγοριοποίηση Πολλαπλών Κλάσεων αφορά περιστάσεις όπου έχουμε πάνω από δύο διαθέσιμες κλάσεις. Για παράδειγμα, πάνω στο ίδιο σύνολο δεδομένων, θα έχουμε Κατηγοριοποίηση Πολλαπλών Κλάσεων εφόσον ξεχωρίζουμε τους σκύλους από τις ράτσες τους και ένας σκύλος ανήκει αποκλειστικά σε μια ράτσα. Επομένως θα υπάρχει μία κλάση για κάθε διαθέσιμη ράτσα. Οι ράτσες είναι ασφαλώς περισσότερες από δύο.

Τέλος, υπάρχει το Multi-label Classfication [14, 15], ή αλλιώς Κατηγοριοποίηση Πολλαπλών Ετικετών. Η διαφορά αυτού του είδους κατηγοριοποίησης είναι πως μέχρι τώρα θεωρούσαμε ότι κάθε αντικείμενο μπορεί να ανήκει σε μόνο μία κλάση. Όμως, στο συγκεκριμένο είδος κατηγοριοποίησης ένα αντικείμενο μπορεί να κατηγοριοποιηθεί σε δύο, τρεις, ίσως και όλες τις διαθέσιμες κλάσεις του συνόλου δεδομένων. Σε αυτήν την περίπτωση, ένας σκύλος μπορεί να ανήκει σε περισσότερες από μια ράτσες, δηλαδή κλάσεις. Ένα κλασικό παράδειγμα προβλήματος πολλαπλών ετικετών είναι το ότι μια κινηματογραφική ταινία μπορεί ταυτόχρονα να ανήκει και στην κλάση "Περιπέτεια" και στην κλάση "Θρίλερ". Ένα άλλο παράδειγμα μπορεί να είναι μια φωτογραφία στην οποία φαίνονται πολλά διαφορετικά πράγματα. Αν η φωτογραφία περιέχει, για παράδειγμα ένα σπίτι, ένα δέντρο και μία κούνια, τότε ανήκει ταυτόχρονα και στις τρεις κλάσεις που αντιστοιχούν σε αυτά τα πράγματα, δηλαδή τις κλάσεις "Σπίτι", "Δέντρο" και "Κούνια".

Με άλλα λόγια, όταν ένα αντικείμενο ενδέχεται να ανήκει σε περισσότερες από μια κλάσεις, τότε το πρόβλημα ονομάζεται πρόβλημα πολλαπλών ετικετών. Αντίθετα, αν ένα αντικείμενο ανήκει σε μια από τις πολλές κλάσεις που είναι διαθέσιμες στο σύνολο δεδομένων, τότε το πρόβλημα ονομάζεται πρόβλημα πολλαπλών κλάσεων. Τέλος, όταν οι κλάσεις είναι μόνο δύο, και τα αντικείμενα ανήκουν σε μια από τις δύο κλάσεις, το πρόβλημα ονομάζεται δυαδικό πρόβλημα κατηγοριοποίησης.

Στα πλαίσια αυτής της εργασίας θα ασχοληθούμε με τα δυαδικά προβλήματα κατηγοριοποίησης και με τα προβλήματα κατηγοριοποίησης πολλαπλών κλάσεων. θα εκτελέσουμε ΤΜΔ, για να παράγουμε σημαντικά μικρότερα υποδείγματα των συνόλων εκπαίδευσης κάθε συνόλου δεδομένων. Αυτά τα υποδείγματα είναι γνωστά ως συμπυκνωμένα σύνολα (*condensing sets*). Λόγω του μικρότερου μεγέθους των συμπυκνωμένων συνόλων, ο k-NN απαιτεί λιγότερο αποθηκευτικό χώρο και χαμηλότερο υπολογιστικό κόστος. Παράλληλα, η ακρίβεια κατηγοριοποίησης παραμένει υψηλή παρά την σημαντική μείωση των δεδομένων.

2.2.3 Μη-Ισορροπημένα Δεδομένα (Imbalanced Data)

Ο κίνδυνος των μη-ισορροπημένων συνόλων δεδομένων αποτελεί ένα από τα πιο σημαντικά προβλήματα στον κόσμο της εξόρυξης δεδομένων εδώ και πολλά χρόνια [11]. Πολλές φορές μπορεί να υπάρξει μεγάλη ανισοκατανομή σε ένα σύνολο δεδομένων, δηλαδή να υπάρχουν κλάσεις που έχουν έναν πολύ μικρό αριθμό αντικειμένων σε σχέση με άλλες κλάσεις που έχουν πολλά. Οι πρώτες ονομάζονται κλάσεις μειοψηφίας (*minority classes*), ενώ παράλληλα οι δεύτερες ονομάζονται κλάσεις πλειοψηφίας (*majority classes*) [16]. Επίσης, σε ορισμένες έρευνες αναφέρονται ως θετικές και αρνητικές κλάσεις αντίστοιχα ή σπάνιες (rare) και μη-σπάνιες κλάσεις. Στα πλαίσια των πειραμάτων αυτής της εργασίας όμως, αναφέρουμε τις μη-σπάνιες κλάσεις ως κοινές κλάσεις (*common classes*).

Αν ο αριθμός των αντικειμένων μίας σπάνιας κλάσης είναι εξαιρετικά μικρός, υπάρχει η περίπτωση να μειωθούν ακόμη περισσότερο κατά την διαδικασία της μείωσης δεδομένων ή και να εξαλειφθούν εντελώς. Επομένως, η συγκεκριμένη κλάση δεν θα συνεισφέρει καθόλου στην κατηγοριοποίηση. Ως συνέπεια, η ακρίβεια της κατηγοριοποίησης αντικειμένων αυτής της κλάσης μειώνεται. Επίσης, πιθανότατα αυτές οι κλάσεις είναι πιο σημαντικές [17] γιατί "κρύβουν" κάποια σημαντική πληροφορία. Για αυτό και οι σπάνιες κλάσεις θεωρούνται πως είναι οι "θετικές" κλάσεις, επειδή ο χαμηλότερος αριθμός αντικειμένων πιθανότατα σημαίνει κάτι σημαντικό. Για παράδειγμα, τα αντικείμενα που ανήκουν στην κλάση "Χαλαζόπτωση" είναι σημαντικό να προβλεφθούν σωστά σε σχέση με τα αντικείμενα που ανήκουν στην κλάση "Μη Χαλαζόπτωση". Αυτό ισχύει επειδή αν δεν προβλεφθεί μια χαλαζόπτωση πιθανότατα οι καλλιέργειες θα καταστραφούν. Αντίθετα αν προβλεφθεί λανθασμένα μια χαλαζόπτωση, απλά οι αγρότες θα προστατεύσουν τις καλλιέργειες χωρίς, όπως θα αποδειχθεί, λόγο. Με άλλα λόγια, δεν υπάρχει μεγάλο κόστος αν ο κατηγοριοποιητής προβλέψει λανθασμένα χαλαζόπτωση και τελικά δεν πραγματοποιηθεί, παρά να προβλέψει μη χαλαζόπτωση και τελικά να πραγματοποιηθεί χαλαζόπτωση.

Το πρόβλημα της ανισοκατανομής προκαλείται από διάφορους παράγοντες. Ένας από τους πιο συχνούς λόγους είναι πως οι αλγόριθμοι κατηγοριοποίησης που χρησιμοποιούνται από τις εφαρμογές δεν θεωρούν σε καμία περίπτωση πως μπορεί να υπάρχει ανισοκατανομή. Με άλλα λόγια θεωρούν πως οι διαφορές μεταξύ των μεγεθών των κλάσεων είναι μικρές. Παράλληλα θεωρούν πως τα σφάλματα που προκαλούνται σε κάθε κλάση έχουν όλα το ίδιο κόστος.

Ο κεντρικός λόγος όμως και η πηγή του προβλήματος είναι πως βασιζόμαστε μόνο στον υπολογισμό της ακρίβειας (*accuracy*). Δηλαδή, από τον συνολικό αριθμό δεδομένων υπολογίζονται μόνο πόσα αντικείμενα έχουν κατηγοριοποιηθεί σωστά. Όσο μεγαλύτερη είναι η ακρίβεια, τόσο λιγότερα σφάλματα υπάρχουν και ο σκοπός των αλγορίθμων είναι να μειώσουν την συνολική ύπαρξη σφαλμάτων/λαθών όσο περισσότερο γίνεται. Το θέμα όμως είναι πως τα αντικείμενα που προσφέρουν οι σπάνιες κλάσεις στους υπολογισμούς συμβάλλουν πολύ λίγο σε αυτό το σύνολο. Επομένως αντιμετωπίζονται σαν να είναι θόρυβος [4] και αφαιρούνται κατά την ανάλυση από λάθος.

Για παράδειγμα, ένας αλγόριθμος κατηγοριοποίησης που προβλέπει πάντα "Μη Χαλαζόπτωση" θα επιτυγχάνει υψηλή ακρίβεια. Σίγουρα όμως δεν είναι αποδεκτός. Ένα άλλο πιθανό πρόβλημα είναι η ανίχνευση μία φυσικής καταστροφής, όπως είναι οι σεισμοί ή τα τσουνάμι. Καθώς η κλάση "Σεισμός" είναι σπάνια, ο κατηγοριοποιητής θα προβλέψει "Όχι Σεισμός" με μεγάλη ακρίβεια, ωστόσο όμως μπορεί να γίνει όντως σεισμός και να μην προειδοποιηθεί ο πληθυσμός.

Με τον όρο θόρυβο, εννοούμε ύποπτα δεδομένα τα οποία εντοπίζονται μέσα στο σύνολο. Συνήθως αποτελούνται από έναν μικρό αριθμό αντικειμένων τα οποία ανήκουν σε διαφορε-

τικές κλάσεις από τα υπόλοιπα αντικείμενα τριγύρω τους, παρεμποδίζοντας έτσι την ομοιογένεια. Εκτός αυτού, δεδομένα τα οποία είναι αλλοιωμένα ή απλά δεν είναι χρήσιμα μπορούν επίσης να ταυτιστούν ως θόρυβος. Φυσικά, αν τα αντικείμενα αυτά ανήκουν σε σπάνιες κλάσεις, τότε σε καμία περίπτωση δεν αποτελούν θόρυβο. Από την άλλη, αν τα αντικείμενα αυτά ανήκουν σε κοινές κλάσεις και απλά βρίσκονται σε περιοχές όπου τα υπόλοιπα αντικείμενα τριγύρω τους ανήκουν σε διαφορετικές κλάσεις, τότε πιθανότατα αποτελούν θόρυβο.

Το πρόβλημα της ανισοκατανομής μπορεί να προκύψει σε περιπτώσεις κατηγοριοποίησης και δυαδικών και πολλαπλών κλάσεων. Σαφώς, όσες περισσότερες σπάνιες κλάσεις υπάρχουν, τόσο πιο πολύ αυξάνεται η βαρύτητα του προβλήματος [18, 19, 20].

Η αλήθεια είναι πως πολλές εφαρμογές του σημερινού κόσμου, οι οποίες χρησιμοποιούνται κιόλας από υπηρεσίες και επιχειρήσεις, π.χ. εφαρμογές σε νοσοκομεία για ιατρικές διαγνώσεις, αντιμετωπίζουν το πρόβλημα των μη-ισορροπημένων συνόλων δεδομένων. Επομένως αποτελεί πρόβλημα που απαιτεί άμεση αντιμετώπιση. Για αυτόν τον λόγο έχουν προταθεί διάφορες μέθοδοι που επιδιώκουν να περιορίσουν αυτό το φαινόμενο [18]. Αυτές είναι και οι μέθοδοι που αναλύουμε παρακάτω, οι οποίες ανάλογα με την μεθοδολογία τους και τους παράγοντες που επηρεάζουν μπορούν να υπάγονται σε μία από τρεις διαφορετικές κατηγορίες:

- Το εξωτερικό επίπεδο, ή εναλλακτικά επίπεδο δεδομένων (*data and/or external level*) [21]. Σε αυτήν την κατηγορία ανήκουν μέθοδοι που επεξεργάζονται το σύνολο εκπαίδευσης πριν την εκτέλεση των αλγορίθμων μάθησης με σκοπό να εξισορροπήσουν τα μεγέθη των κλάσεων. Οι πιο γνωστές μέθοδοι εξωτερικού επιπέδου είναι αυτές της δειγματοληψίας (*sampling*).

- Το αλγοριθμικό ή αλλιώς εσωτερικό επίπεδο (*algorithmic and/or internal level*) [22], στο οποίο ανήκουν νέοι αλγόριθμοι ή τροποποιημένες εκδόσεις των παλιών οι οποίες δεν αγνοούν τις σπάνιες κλάσεις.

- Η μάθηση με ευαισθησία στο κόστος (*cost-sensitive learning*) [23, 24], μέσω της οποίας μπορούμε να ενσωματώσουμε προσεγγίσεις και από το δύο επίπεδα (αλγοριθμικό και δεδομένων) είτε ταυτόχρονα είτε ξεχωριστά εφόσον το θελήσουμε. Όπως υποδηλώνει και το όνομα, μία μέθοδος που ανήκει σε αυτήν την κατηγορία μπορεί να αναθέσει και διαφορετικό κόστος σε κάθε κλάση για τις εσφαλμένες κατηγοριοποιήσεις της. Επομένως, το πρόβλημα του ίσου κόστους για κάθε κλάση άσχετα από το μέγεθος τους εξαφανίζεται.

Στα πλαίσια αυτής της εργασίας, θα ασχοληθούμε με την πρώτη περίπτωση, δηλαδή το εξωτερικό επίπεδο λύσεων. Θα αναλύσουμε τα πιο διαδεδομένα είδη δειγματοληψίας, η οποία πρόκειται για μία διαδικασία τροποποίησης δεδομένων όπου ο σκοπός της είναι η επίτευξη μεγαλύτερης ισορροπίας μεταξύ των κλάσεων του συνόλου δεδομένων. Ανάλογα

με τον τύπο της δειγματοληψίας, μπορούμε είτε να αυξήσουμε τον αριθμό των αντικειμένων για τις κλάσεις μειοψηφίας (υπερδειγματοληψία) ή αντίστοιχα να μειώσουμε τον αριθμό αντικειμένων των κλάσεων πλειοψηφίας (υποδειγματοληψία).

Βέβαια, υπάρχει πάντα η περίπτωση να μην είναι ξεκάθαρο πια μέθοδος δειγματοληψίας πρέπει να χρησιμοποιηθεί. Ο Andrew Estabrooks έχει παρουσιάσει μία μέθοδο ειδικά σχεδιασμένη για να χειρίζεται τέτοιες περιπτώσεις [25], η οποία διαλέγει αν η υπερδειγματοληψία ή η υποδειγματοληψία θα είναι πιο αποτελεσματική πάνω σε ένα σύνολο δεδομένων, καθώς και ποιος είναι ο βέλτιστος ρυθμός εκτέλεσης τους.

2.2.4 Ορθότητα, Ευαισθησία και F-Measure

Εφόσον θέλουμε να αξιολογήσουμε την αξία των σπάνιων κλάσεων, το ποσοστό της ακρίβειας δεν επαρκεί, καθώς όπως αναφέραμε και παραπάνω ένας κατηγοριοποιητής που προβλέπει πάντα "Μη Χαλαζόπτωση" έχει 99% ακρίβεια, αλλά δεν είναι αποδεκτός. Άμα θέλουμε να ανιχνεύσουμε ποιες μέρες θα έχει χαλαζόπτωση, δεν μπορούμε απλά να θέσουμε σχεδόν όλα τα αντικείμενα ως "Μη Χαλαζόπτωση" μόνο και μόνο επειδή πρόκειται για ένα σπάνιο καιρικό φαινόμενο. Ο κατηγοριοποιητής απλά δεν θα ανιχνεύει ποτέ "Χαλαζόπτωση" και έτσι όταν όντως συμβεί, οι αγρότες δεν θα έχουν ιδέα και δεν θα προστατεύσουν τις σοδειές τους. Για αυτόν τον λόγο, δίνεται προσοχή σε δύο άλλα κριτήρια: την ορθότητα (*precision*) και την ευαισθησία (*recall*), των οποίων τα αποτελέσματα θεωρούνται πιο πολύτιμα σε τέτοιες περιστάσεις.

Η ορθότητα πρόκειται για ένα ποσοστό που αντιπροσωπεύει πόσα από τα θετικά αντικείμενα που λάβαμε είναι όντως θετικά. Μπορεί να υπολογιστεί ως εξής:

$$Precision = \frac{TP}{TP + FP} \tag{2.2}$$

όπου TP είναι τα True Positives, δηλαδή τα αντικείμενα που είναι θετικά και κατηγοριοποιήθηκαν όντως ως θετικά και FP είναι τα False Positives, δηλαδή τα αντικείμενα που κατηγοριοποιήθηκαν ως θετικά αλλά στην πραγματικότητα είναι αρνητικά. Στο πρόβλημα της πρόβλεψης για χαλαζοπτώσεις, η ορθότητα υπολογίζει πόσες προβλέψεις για χαλαζόπτωση είναι σωστές.

Η ευαισθησία, επιπλέον, αντιπροσωπεύει πόσα από τα θετικά αντικείμενα που είναι διαθέσιμα κατηγοριοποιήθηκαν όντως ως θετικά και υπολογίζεται με τον παρακάτω τύπο:

$$Recall = \frac{TP}{TP + FN} \tag{2.3}$$

όπου FN είναι τα False Negatives, δηλαδή τα αντικείμενα που κατηγοριοποιήθηκαν ως αρνητικά αλλά στην πραγματικότητα είναι θετικά και TP είναι τα True Positives. Στο πρόβλημα

της πρόβλεψης για χαλαζοπτώσεις, η ευαισθησία υπολογίζει πόσες από τις χαλαζοπτώσεις που πραγματοποιήθηκαν, προβλέφθηκαν από τον κατηγοριοποιητή.

Επιπλέον, υπάρχει και ένα τρίτο κριτήριο το οποίο συνδυάζει την ορθότητα και την ευαισθησία. Αυτό το κριτήριο ονομάζεται F-Measure, ή εναλλακτικά F1 Score [26] και είναι χρήσιμο σε περιπτώσεις όπου δεν μπορεί να αποφασιστεί πιο από τα άλλα δύο κριτήρια είναι πιο σημαντικό. Το F-Measure υπολογίζεται ως εξής:

$$F = \frac{Precision * Recall}{Precision + Recall} \tag{2.4}$$

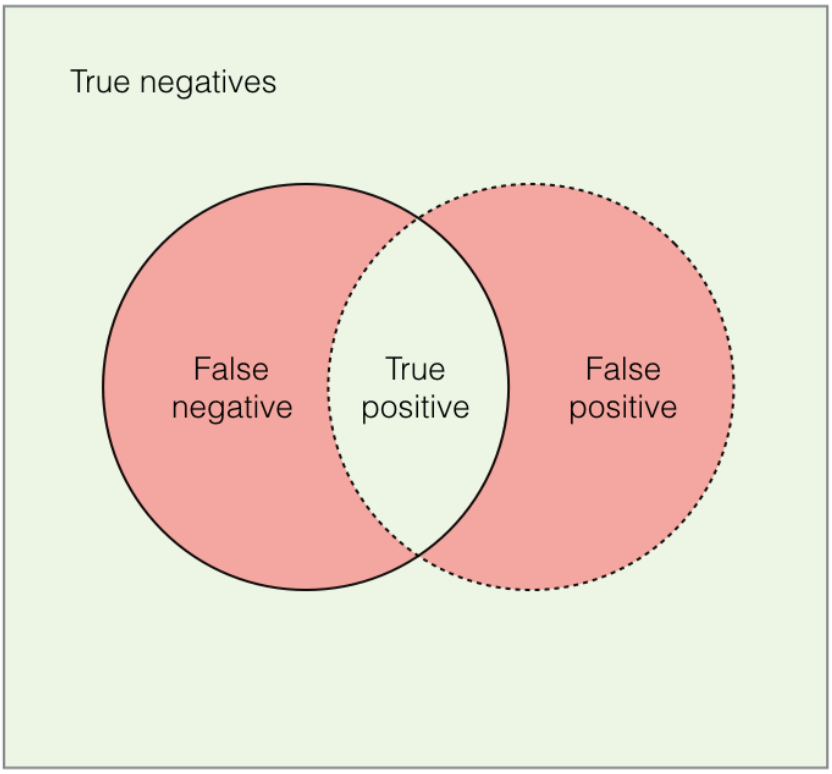
Για να εξηγήσουμε καλύτερα την ορθότητα και την ευαισθησία, ας χρησιμοποιήσουμε ένα παράδειγμα. Έστω ότι έχουμε ένα σύνολο δεδομένων που περιέχει σκυλιά με διάφορες ράτσες. Ας θεωρήσουμε πως θέλουμε να βρούμε τα σκυλιά μίας συγκεκριμένης ράτσας (π.χ. λαμπραντόρ). Επομένως χρειαζόμαστε έναν κατηγοριοποιητή για να ξεχωρίσουμε την συγκεκριμένη ράτσα από τις υπόλοιπες. Χρησιμοποιώντας μία ομάδα σχημάτων που παρασχέθηκαν από το [27], θα επιχειρήσουμε να εξηγήσουμε περαιτέρω την ορθότητα και την ευαισθησία μέσω του προ-αναφερόμενου συνόλου δεδομένων.

Στο Σχήμα 2.1 βλέπουμε την χειρότερη δυνατή περίπτωση όπου και η ορθότητα και η ευαισθησία είναι χαμηλές. Μέσα στο τετράγωνο περιέχονται όλα τα δεδομένα του συνόλου δεδομένων. Μέσα στην διακεκομμένη γραμμή είναι τα σκυλιά που ο κατηγοριοποιητής αναγνωρίζει ως λαμπραντόρ, ενώ μέσα στην συνεχόμενη γραμμή είναι τα σκυλιά που είναι λαμπραντόρ στην πραγματικότητα. Επομένως, εδώ έχουμε πολλά λαμπραντόρ λάθος κατηγοριοποιημένα σε άλλες ράτσες (False Negatives, FN) και ταυτόχρονα έχουμε πολλά σκυλιά από άλλες ράτσες κατηγοριοποιημένα σαν λαμπραντόρ (False Positives, FP). Το τετράγωνο και οι δύο γραμμές συμβολίζουν τα ίδια πράγματα και στα επόμενα τρία Σχήματα (2.2, 2.3, 2.4).

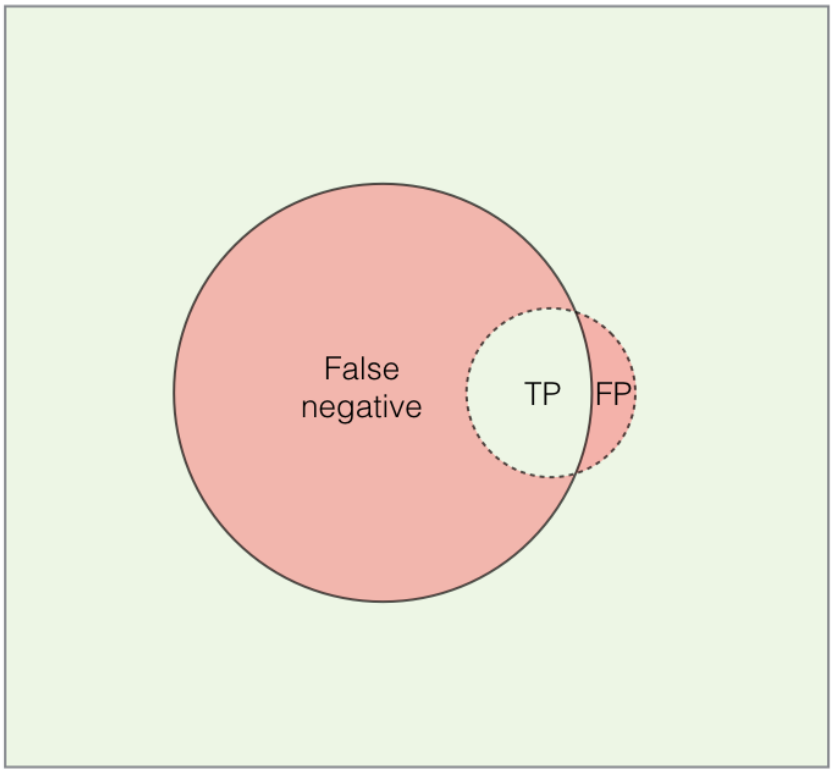
Στο Σχήμα 2.2 βλέπουμε το αποτέλεσμα ενός αυστηρού κατηγοριοποιητή, ο οποίος αναγνωρίζει ένα πολύ μικρό ποσοστό των σκυλιών ως λαμπραντόρ. Από τα αντικείμενα που εξάγει, τα περισσότερα είναι σωστά κατηγοριοποιημένα, αλλά παράλληλα υπάρχουν πολλά περισσότερα λαμπραντόρ στο σύνολο δεδομένων που δεν βρήκε.

Στο Σχήμα 2.3 ο κατηγοριοποιητής έχει χαμηλή ορθότητα, αλλά μεγάλη ευαισθησία. Επομένως, καταφέρνει να κατηγοριοποιήσει πολλά από τα λαμπραντόρ στην σωστή κλάση. Παράλληλα όμως, θεωρεί πως και πολλά άλλα σκυλιά που είναι διαφορετικές ράτσες είναι λαμπραντόρ.

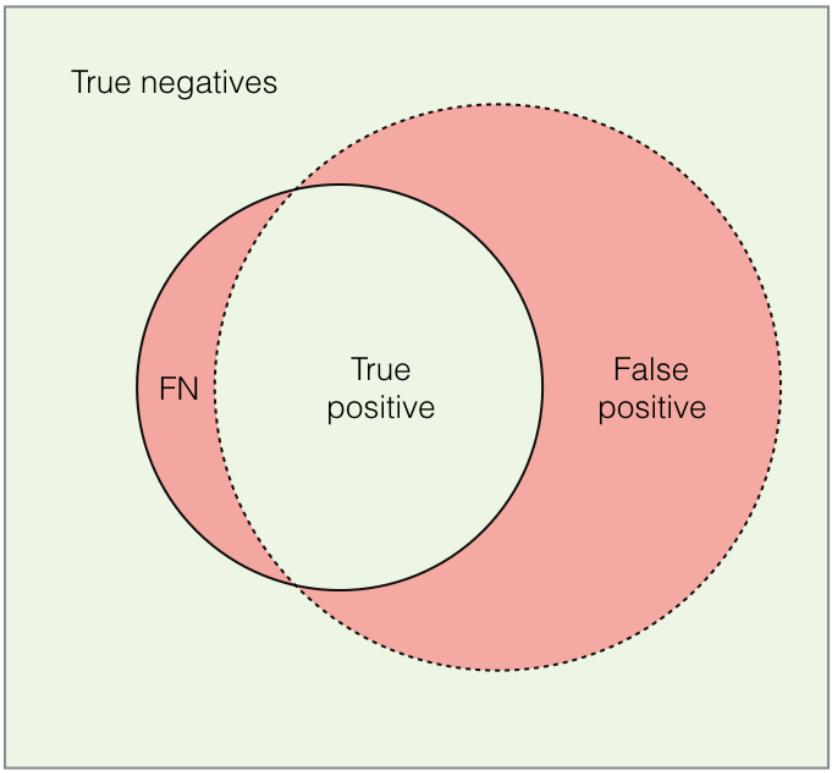
Τέλος, στο Σχήμα 2.4 παρουσιάζεται η πιο ιδανική περίπτωση, όπου έχουμε και υψηλή ορθότητα και υψηλή ευαισθησία. Ο κατηγοριοποιητής καταφέρνει να κατηγοριοποιήσει τα περισσότερα λαμπραντόρ στην σωστή κλάση, με ελάχιστες απώλειες και λάθη.



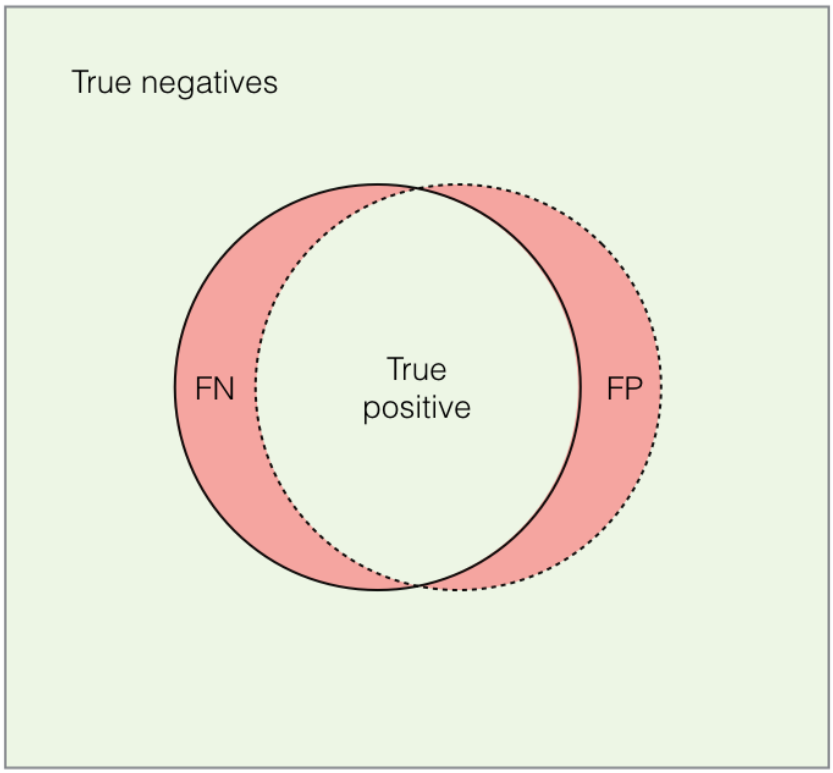
Σχήμα 2.1: Χαμηλή ορθότητα, χαμηλή ευαισθησία.



Σχήμα 2.2: Υψηλή ορθότητα, χαμηλή ευαισθησία.



Σχήμα 2.3: Χαμηλή ορθότητα, υψηλή ευαισθησία.



Σχήμα 2.4: Υψηλή ορθότητα, υψηλή ευαισθησία.

Να σημειώσουμε πως η ορθότητα και η ευαισθησία δεν είναι πάντα το ίδιο σημαντικές. Ανάλογα με το σύνολο δεδομένων που χρησιμοποιείται το ένα κριτήριο μπορεί να αποδειχθεί πιο ωφέλιμο από το άλλο. Για παράδειγμα, ας υποθέσουμε πως έχουμε ένα σύνολο για την πρόβλεψη του καιρού και η σπάνια κλάση του αφορά την χαλαζόπτωση. Αν υπολογιστεί χαμηλή ευαισθησία, αυτό σημαίνει πως έχουμε πολλά False Negatives, δηλαδή πολλές προβλέψεις για μη χαλαζόπτωση ενώ στην πραγματικότητα πραγματοποιείται χαλαζόπτωση. Κάτι τέτοιο μπορεί να είναι καταστροφικό για ορισμένες ομάδες ανθρώπων, όπως είναι για παράδειγμα οι αγρότες, καθώς δεν θα προστατεύσουν τις καλλιέργειες τους από το χαλάζι γιατί δεν προειδοποιήθηκαν. Εναλλακτικά, αν έχουμε χαμηλή ορθότητα, τότε αυτό σημαίνει ότι έχουμε πολλά False Positives, άρα πολλές λανθασμένες προβλέψεις ότι θα υπάρξει χαλαζόπτωση ενώ στην πραγματικότητα αυτό δεν ισχύει και οι αγρότες θα προφυλάξουν τις καλλιέργειες τους χωρίς λόγο. Σαφώς, οι πιθανές συνέπειες της πρώτης περίπτωσης είναι πολύ πιο επικίνδυνες από αυτές της δεύτερης, επομένως σε αυτό το σενάριο εκτιμάμε την ευαισθησία πιο πολύ.

Βέβαια, αν ένας κατηγοριοποιητής προβλέπει πάντα "Χαλαζόπτωση", θα έχουμε πολλά FP και άρα χαμηλή ορθότητα. Από την άλλη, δεν θα υπάρχει ούτε ένα FN. Άρα θα έχουμε ($recall = 1$). Με βάση αυτό το παράδειγμα, η ευαισθησία είναι πιο σημαντική αλλά σε καμία περίπτωση δεν πρέπει να έχουμε εξαιρετικά χαμηλή ορθότητα. Αν ισχύει κάτι τέτοιο, οι αγρότες θα προστατεύουν καθημερινά τις καλλιέργειες τους χωρίς λόγο και κάποια στιγμή, θα σταματήσουν να εμπιστεύονται τον κατηγοριοποιητή.

Αντίθετα, αν είχαμε μια μηχανή αναζήτησης που ανακτά μαζί με τα σχετικά προς τα κριτήρια αναζήτησης έγγραφα και πολλά άσχετα, θα έχουμε χαμηλή ορθότητα. Σε αυτήν την περίπτωση μας ενδιαφέρει περισσότερο η ορθότητα παρά η ευαισθησία. Για παράδειγμα, αν αναζητήσουμε πόσα γκολ πέτυχε ο Μέσι το 2020 και η μηχανή αναζήτησης μας επιστρέψει πολλά άσχετα αποτελέσματα μαζί με όλα τα έγγραφα που περιέχουν την απάντηση στην ερώτηση μας, υπάρχει ο κίνδυνος να "χαθούμε" μέσα στα FP και να μην βρούμε τελικά αυτό που ψάχνουμε. Δεν είναι ανάγκη όλα τα σχετικά άρθρα που αφορούν τα πόσα γκολ έβαλε ο Μέσι να ανακτηθούν. Μπορούμε να πληροφορηθούμε για το πόσα γκολ έβαλε ο Μέσι από ένα άρθρο που ανακτήθηκε από την μηχανή αναζήτησης. Άρα, σε αυτήν την περίπτωση μπορούμε να δεχτούμε χαμηλή ευαισθησία αλλά όχι χαμηλή ορθότητα.

2.2.5 Υπερδειγματοληψία

Υπάρχουν διάφορες κατηγορίες μεθόδων δειγματοληψίας. Μία από τις πρώτες, καθώς και τις πιο βασικές είναι η αποκαλούμενη υπερδειγματοληψία (*oversampling*) [18, 16]. Οι μέθοδοι που περιλαμβάνονται σε αυτήν την κατηγορία χτίζουν ένα υπερσύνολο του αρχικού σύνολου δεδομένων δημιουργώντας αντίγραφα ή παραλλαγές των αντικειμένων που ανήκουν στις σπάνιες κλάσεις. Όμως είναι πολύ πιθανό να προκαλέσουν υπερμοντελοποιηση (*over-*

fitting), καθώς όπως είπαμε δεν δημιουργούν καινούρια αντικείμενα, αλλά αντίγραφα άλλων αντικειμένων που υπάρχουν ήδη. Επίσης, κατά συνέπεια η υπερδειγματοληψία αυξάνει την διάρκεια μάθησης μεγαλώνοντας τον αριθμό των αντικειμένων που ανήκουν στο σύνολο δεδομένων εκπαίδευσης.

Μία από τις μεθόδους αυτής της κατηγορίας είναι η τυχαία υπερδειγματοληψία (*random oversampling*). Πρόκειται για μία μέθοδο που ισορροπεί την κατανομή των κλάσεων μέσω της τυχαίας αντιγραφής θετικών αντικειμένων στις σπάνιες κλάσεις. Επιδιώκει να κάνει πιο αισθητή την ύπαρξη αυτών των κλάσεων.

Μερικές άλλες μέθοδοι οι οποίες αναφέρονται στις βιβλιογραφικές πηγές [16, 28], είναι η κατευθυνόμενη υπερδειγματοληψία (*directed oversampling*) όπου μπορούμε να διαλέξουμε τα αντικείμενα που θα αντικατασταθούν, καθώς και η υπερδειγματοληψία με μία ενημερωμένη γενιά νέων δειγμάτων.

Επιπλέον μπορούμε να χρησιμοποιήσουμε έναν συνδυασμό των αναφερόμενων μεθόδων. Στο [25] παρουσιάζεται μία τροποποιημένη μέθοδος υπερδειγματοληψίας η οποία δίνει περισσότερη προσοχή στις μικρότερες συστάδες δεδομένων (*data clusters*) που μπορεί να δημιουργηθούν, ενώ παράλληλα χειρίζεται καλύτερα την ανισορροπία των κλάσεων (*class imbalance*) είτε είναι εσωτερική είτε μεταξύ διαφορετικών κλάσεων. Τέλος, μία ακόμα μέθοδος που χρησιμοποιεί την υπερδειγματοληψία προς όφελος της είναι η SMOTE, η οποία θα αναλυθεί με λεπτομέρεια παρακάτω.

2.2.6 Υποδειγματοληψία

Η υποδειγματοληψία (*undersampling*) [18, 16] αποτελεί την αντίθετη διαδικασία της υπερδειγματοληψίας, δηλαδή πρόκειται για μία μέθοδο που εξάγει ένα υποσύνολο από τις μη-σπάνιες κλάσεις και το χρησιμοποιεί για να εκπαιδεύσει τον κατηγοριοποιητή. Οι μέθοδοι υποδειγματοληψίας είναι καλύτερες στον χειρισμό του θορύβου. Έτσι τείνουν να καταλήγουν σε καλύτερα αποτελέσματα από την υπερδειγματοληψία, ειδικά όταν επικρατούν υψηλά επίπεδα τέτοιων περιττών αντικειμένων. Ωστόσο, υπάρχει ο κίνδυνος της πιθανής διαγραφής σημαντικών δεδομένων.

Μία από αυτές τις μεθόδους, όπως αναφέρεται και στο [18], θεωρείται πως είναι ο πρώτος και πιο βασικός αλγόριθμος επεξεργασίας δεδομένων, ο Edited Nearest Neighbor Rule (***ENN-Rule***) [29], ο οποίος όταν βρίσκει αντικείμενα τα οποία δεν ταιριάζουν με την πλειοψηφία των κοντινότερων γειτόνων τους, τα αφαιρεί. Χρησιμοποιείται συχνά για την αφαίρεση δεδομένων που αποτελούν θόρυβο.

Επιπλέον, μία ακόμη από τις πρώτες και πιο γνωστές μεθόδους είναι η τυχαία υποδειγματοληψία (*random undersampling,* ***RUS***) [16]. Εκτελεί την διαδικασία της υποδει-

γματοληψίας και αφαιρεί τυχαία αντικείμενα από τις μη-σπάνιες κλάσεις. Είναι μία από τις πιο αποτελεσματικές μεθόδους και μπορεί να παράγει καλά αποτελέσματα άσχετα από το ποσοστό θορύβου. Εκτός αυτού, σύμφωνα με τις πηγές [16, 28], υπάρχει και η κατευθυνόμενη υποδειγματοληψία (*directed undersampling*) όπου μπορούμε να διαλέξουμε τα αντικείμενα που θα αντικατασταθούν.

Αφού ο κατηγοριοποιητής εκπαιδεύεται μόνο με ένα υποσύνολο των μη-σπάνιων κλάσεων, πολλά από τα αντικείμενα που ανήκουν σε αυτές αγνοούνται ή αφαιρούνται. Η εκπαίδευση επομένως επιταχύνεται σημαντικά, καθώς το σύνολο εκπαίδευσης σταθεροποιείται. Κατά συνέπεια όμως, η τυχαία αφαίρεση αντικειμένων από τις μη-σπάνιες κλάσεις σημαίνει ότι είναι πιθανόν να χαθούν σημαντικές πληροφορίες χωρίς να επεξεργαστούν ποτέ. Οπότε η ίδια η φύση της διαδικασίας της υποδειγματοληψίας δημιουργεί και ένα πολύ σημαντικό μειονέκτημα. Ευτυχώς όμως, έχουν υλοποιηθεί διάφοροι μέθοδοι που επιδιώκουν να περιορίσουν αυτό το πρόβλημα.

Η βελτιστοποίηση της υποδειγματοληψίας, καθώς και η υψηλή αντίσταση της στον θόρυβο και γενικά περιττές πληροφορίες οφείλεται στην υιοθέτηση διάφορων τεχνικών. Μία από αυτές τις τεχνικές είναι τα Tomek Links [30] για παράδειγμα, τα οποία ανιχνεύουν αντικείμενα κοντά στα σύνορα των κλάσεων που ο πιο κοντινός του γείτονας είναι στην "άλλη πλευρά", επομένως τουλάχιστον το ένα έχει κατηγοριοποιηθεί λανθασμένα. Τέτοια αντικείμενα αφαιρούνται.

Η Μονόπλευρη Επιλογή (*One-Sided Selection,* ***OSS***) [31] έχει υλοποιηθεί από τους Miroslav Kubat και Stan Matwin. Αφαιρεί από τις σπάνιες κλάσεις μόνο τα αντικείμενα που κατά τους υπολογισμούς της είναι πλεονάζοντα. Χρησιμοποιεί σε συνδυασμό τον αλγόριθμο 1-NN Rule και τα Tomek Links, όπου το πρώτο βοηθάει με την κατηγοριοποίηση των δεδομένων και έπειτα μέσω του δεύτερου ανιχνεύονται τα δεδομένα που μπορούν να αφαιρεθούν. Σύμφωνα με τους Kubat και Matwin, ως πλεονάζοντα ορίζονται τα ομοιογενή δεδομένα τα οποία δεν βρίσκονται κοντά στα σύνορα κλάσεων. Αν και δεν συνεισφέρουν στην λάθος κατηγοριοποίηση δεδομένων, μπορούν να αυξήσουν σημαντικά το κόστος της διαδικασίας και γι΄ αυτό συνιστάται η αφαίρεση τους. Επίσης, το ίδιο συνιστάται και για αντικείμενα που βρίσκονται πολύ κοντά, ή ακόμα και πάνω στα σύνορα κλάσεων, καθώς είναι πιθανόν να κατηγοριοποιηθούν λανθασμένα αν υπάρχει έστω και ένα παραμικρό επίπεδο θορύβου.

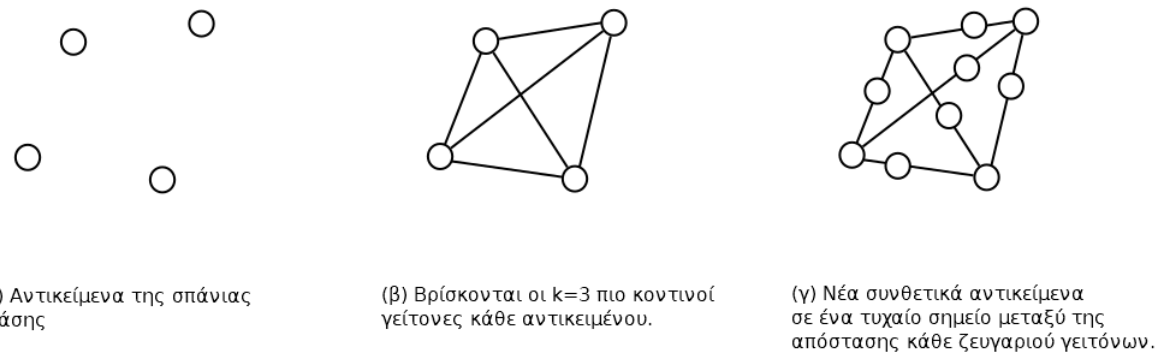
Επιπλέον, μία παραλλαγή του OSS είναι ο Κανόνας Καθαρισμού Γειτονιάς (*Neighborhood Cleaning Rule,* ***NCL***) [32]. Η μεγαλύτερη διαφορά μεταξύ αυτών των δύο είναι ότι ενώ ο κύριος στόχος του OSS είναι η μείωση δεδομένων, το NCL επικεντρώνεται περισσότερη στον καθαρισμό τους.

2.2.7 SMOTE

Η Συνθετική Τεχνική για την Υπερδειγματοληψία των Μειονοτήτων (*Synthetic Minority Oversampling Technique,* ***SMOTE***) [12] πρόκειται για μία μέθοδο που δημιουργεί νέα, συνθετικά αντικείμενα στις σπάνιες κλάσεις, προκαλώντας έτσι υπερδειγματοληψία σε αυτές. Επειδή τα νέα αντικείμενα που προσθέτει δεν είναι αντίγραφα, αποφεύγεται ο κίνδυνος της υπερμοντελοποίησης που έχουν άλλες μέθοδοι υπερδειγματοληψίας. Η SMOTE υιοθετεί επομένως μία εντελώς διαφορετική προσέγγιση από τις υπόλοιπες τεχνικές υπερδειγματοληψίας, οι οποίες προσθέτουν διπλότυπα στο σύνολο δεδομένων. Η SMOTE παίρνει κάθε αντικείμενο που ανήκει σε μία σπάνια κλάση και εισάγει συνθετικά αντικείμενα στα γραμμικά τμήματα που ενώνουν τους k πιο κοντινούς γείτονες που ανήκουν στην ίδια κλάση.

Μπορούμε να δούμε ένα παράδειγμα της SMOTE στο Σχήμα 2.5. Αρχικά έχουμε τέσσερα αντικείμενα, τα οποία ανήκουν όλα τους στην ίδια σπάνια κλάση, όπως φαίνεται στο 2.5(α). Έχοντας θέσει την μεταβλητή που αντιπροσωπεύει τον αριθμό των πιο κοντινών γειτόνων που θέλουμε να βρούμε σε 3 (k = 3), το κάθε αντικείμενο "βρίσκει" τα άλλα 3 και η SMOTE σχηματίζει μία διαδρομή μεταξύ τους. Η κάθε διαδρομή συμβολίζεται από τα γραμμικά τμήματα που φαίνονται στο 2.5(β). Τέλος, βλέπουμε στο 2.5(γ) πως για κάθε μία από αυτές της διαδρομές δημιουργείται ένα καινούριο σύνθετο αντικείμενο σε κάποιο τυχαίο σημείο πάνω τους. Η τιμή του k, καθώς και ο αριθμός νέων συνθετικών αντικειμένων που δημιουργούνται με κάθε πέρασμα της SMOTE μπορούν να μεταβληθούν ανάλογα με τις ανάγκες μας.

Αν και είναι αποτελεσματική, η μέθοδος SMOTE δεν μπορεί να αντεπεξέλθει καλά πάνω σε σύνολα δεδομένων που περιέχουν συνεχή (*continuous*) και ονομαστικά (*nominal*) χαρακτηριστικά. Για αυτόν τον λόγο αναπτύχθηκε μία επέκταση της, η SMOTE-NC (Synthetic Minority Oversampling Technique Nominal Continuous), η οποία μπορεί να χειριστεί μικτά σύνολα δεδομένων που περιέχουν και τις δύο από τις προαναφερόμενες κατηγορίες. Μία



Σχήμα 2.5: Χρήση της SMOTE με k = 3.

επιπλέον γνωστή επέκταση είναι η SMOTE-N (Synthetic Minority Oversampling Technique Nominal), η οποία μπορεί να χειριστεί σύνολα που αποτελούνται αποκλειστικά από ονομαστικά δεδομένα, κάτι που δεν μπορεί να κάνει η SMOTE-NC.

Ένα μειονέκτημα της SMOTE και πολλών άλλων μεθόδων υπερδειγματοληψίας είναι ότι αυξάνεται η επικάλυψη (*overlapping*) [33] μεταξύ κλάσεων. Το SMOTE συγκεκριμένα, για κάθε αντικείμενο των σπάνιων κλάσεων δημιουργεί τον ίδιο αριθμό συνθετικών αντικειμένων, χωρίς να νοιάζεται αν υπάρχουν γειτονικά αντικείμενα και που ακριβώς βρίσκονται [34]. Για τον περιορισμό αυτού του θέματος έχουν αναπτυχθεί ορισμένες μέθοδοι, όπως το Borderline-SMOTE1 και το Borderline-SMOTE2 [35] που προκαλούν υπερδειγματοληψία μόνο σε αντικείμενα των σπάνιων κλάσεων που βρίσκονται στα σύνορα. Άλλοι παρόμοιοι μέθοδοι είναι οι το Safe-Level SMOTE [36], το Adaptive Synthetic Sampling (ADASYN) [37] και το SPIDER12 [38].

2.3 Τεχνικές Μείωσης Δεδομένων (Data Reduction Techniques)

2.3.1 Κατηγορίες των Τεχνικών

Ως Τεχνικές Μείωσης Δεδομένων (ΤΜΔ), ορίζονται αλγόριθμοι οι οποίοι χρησιμοποιούνται για να μειώσουμε τον αριθμό των αντικειμένων σε ένα υπερβολικά μεγάλο σύνολο δεδομένων. Ουσιαστικά, η διαδικασία μείωσης δεδομένων (*data reduction*) επιχειρεί να αφαιρέσει διπλότυπα αντικείμενα και πληροφορίες που δεν θεωρούνται χρήσιμες. Ο κύριος λόγος που χρησιμοποιούνται οι ΤΜΔ είναι για να μειωθεί το υπολογιστικό κόστος του κατηγοριοποιητή k-NN, χωρίς να μειωθεί το ποσοστό της ακρίβειας τους.

Οι ΤΜΔ χωρίζονται σε δύο κατηγορίες: (*i*) τους Αλγορίθμους Σύνοψης Προτύπων (*Prototype Abstraction Algorithms*) [8] και (*ii*) τους Αλγορίθμους Επιλογής Προτύπων (*Prototype Selection Algorithms*) [7], οι οποίοι με την σειρά τους μπορούν να χωριστούν σε μία από τις παρακάτω υποκατηγορίες, (*ii.a*) τους Αλγορίθμους Συμπύκνωσης (*Condensing Algorithms*) και (*ii.b*) τους Αλγορίθμους Επεξεργασίας (*Editing Algorithms*).

Οι Αλγόριθμοι Επεξεργασίας επιχειρούν να πετύχουν μεγαλύτερο ποσοστό ακρίβειας παρά να επιτύχουν μεγάλη μείωση των δεδομένων. Αυτό το κάνουν με δύο τρόπους, πρώτον μέσω της αφαίρεσης ορισμένων δεδομένων εκπαίδευσης τα οποία θεωρούνται πως είναι θόρυβος και δεύτερον μέσω της εξομάλυνσης των ορίων απόφασης μεταξύ των κλάσεων.

Οι Αλγόριθμοι Συμπύκνωσης και οι Αλγόριθμοι Σύνοψης Προτύπων δημιουργούν ένα μικρό αντιπροσωπευτικό υποσύνολο του αρχικού συνόλου δεδομένων, το οποίο ονομάζεται συμπυκνωμένο σύνολο (*condensing set).* Αυτό το συμπυκνωμένο σύνολο, έχει το πλεονέκτημα του χαμηλότερου υπολογιστικού κόστους και των μειωμένων απαιτήσεων για μνήμη

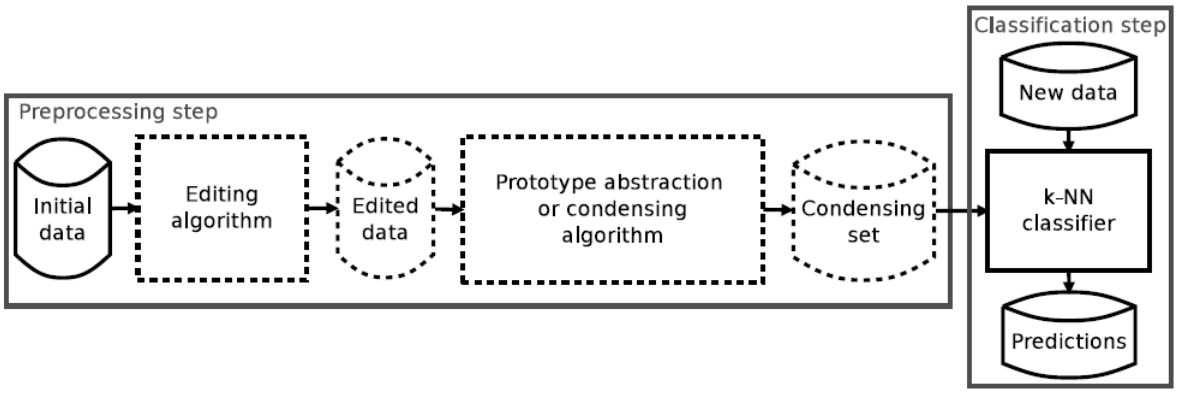
ενώ παράλληλα, η ακρίβεια του κατηγοριοποιητή k-NN παραμένει σε υψηλά επίπεδα, όμοια με αυτά που θα επιτύγχανε ο κατηγοριοποιητής αν χρησιμοποιούσε το αρχικό σύνολο εκπαίδευσης, αλλά σαφώς, ξοδεύοντας πολύ λιγότερο υπολογιστικό κόστος κατά την κατηγοριοποίηση. Παράλληλα, λόγω του μικρότερου μεγέθους του αρχείου δεν χρειαζόμαστε και τόσο πολύ αποθηκευτικό χώρο. Οι ΤΜΔ μπορούν να αξιολογηθούν από τρία διαφορετικά κριτήρια:

- Το ποσοστό μείωσης (*reduction rate*), το οποίο πρόκειται για την αναλογία του αριθμού των αφαιρούμενων αντικείμενων προς τον συνολικό αριθμό αντικειμένων. Συμβολίζει πόσο μικρότερο είναι το μέγεθος του συμπυκνωμένου συνόλου σε σχέση με το σύνολο εκπαίδευσης. Όσο μεγαλύτερο είναι το ποσοστό μείωσης, τόσο λιγότερο κόστος κατά την κατηγοριοποίηση απαιτείται.
- Το υπολογιστικό κόστος προεπεξεργασίας (*preprocessing computational cost*), το οποίο αναφέρεται στο κόστος που απαιτείται δεδομένου να χτιστεί το συμπυκνωμένο σύνολο.
- Και την ακρίβεια της κατηγοριοποίησης (*classification accuracy*) που καταγράφει το ποσοστό των αντικειμένων που έχουν κατηγοριοποιηθεί σωστά σε σχέση με το συνολικό αριθμό προβλέψεων που πραγματοποιήθηκαν.

Αν και γενικά θεωρείται πως και τα τρία κριτήρια είναι το ίδιο σημαντικά, υπό ορισμένες περιπτώσεις το ένα μπορεί να αποδειχθεί πιο χρήσιμο από τα άλλα. Η αποτελεσματικότητα των Αλγορίθμων Συμπύκνωσης, καθώς και των Αλγορίθμων Σύνοψης Προτύπων εξαρτάται σε μεγάλο βαθμό από το επίπεδο ύπαρξης θορύβου μέσα στο σύνολο δεδομένων. Όσο περισσότερα τέτοια δεδομένα υπάρχουν, τόσο πιο πολύ φθείρεται το τελικό αποτέλεσμα των αλγορίθμων.

Ωστόσο, μέσω των Αλγορίθμων Επεξεργασίας μπορούμε να μειώσουμε, ή ακόμα και να εξαφανίσουμε τον θόρυβο από ένα σύνολο δεδομένων. Η πιο αποτελεσματική προσέγγιση επομένως, είναι να χρησιμοποιήσουμε σαν πρώτο βήμα έναν Αλγόριθμο Επεξεργασίας για να αφαιρέσουμε τον θόρυβο χωρίς να μειώσουμε το ποσοστό ακρίβειας της κατηγοριοποίησης, ενώ παράλληλα αυξάνουμε το ποσοστό μείωσης. Έπειτα, με το υποσύνολο που παράγεται εκτελούμε έναν Αλγόριθμο Συμπύκνωσης ή αλγόριθμο Σύνοψης Προτύπων. Μέσω αυτής της διαδικασίας παράγουμε ένα ποιοτικό συμπυκνωμένο σύνολο.

Στο Σχήμα 2.6 παρουσιάζεται διαγραμματικά η διαδικασία της κατηγοριοποίησης k-NN μέσω μείωσης δεδομένων. Χωρίζεται σε δύο φάσεις: την προεπεξεργασία και την κατηγοριοποίηση. Η προεπεξεργασία είναι προαιρετική και έχει τέσσερις διαφορετικούς τύπους: (i) χωρίς προεπεξεργασία, (ii) μόνο επεξεργασία, (iii) μόνο συμπύκνωση και (iv) επεξεργασία και συμπύκνωση. Προφανώς ο κάθε ένας από αυτούς τους τύπους αποδεικνύεται χρήσιμος σε διαφορετικά είδη συνόλων δεδομένων. Αν το σύνολο εκπαίδευσης έχει μικρό μέγεθος και



Σχήμα 2.6: Κατηγοριοποίηση k-NN μέσω μείωσης δεδομένων.

δεν περιέχει καθόλου θόρυβο, τότε δεν χρειάζεται να γίνει προεπεξεργασία. Όταν το σύνολο εκπαίδευσης είναι μικρό, αλλά περιέχει θόρυβο τότε πρέπει να κάνουμε μόνο επεξεργασία και αντιθέτως, όταν έχουμε μεγάλα σύνολα εκπαίδευσης χωρίς θόρυβο είναι πιο ιδανικό να κάνουμε μόνο συμπύκνωση. Τέλος, όταν ισχύουν και οι δύο περιπτώσεις ταυτόχρονα, έχουμε δηλαδή ένα μεγάλο σύνολο εκπαίδευσης με θόρυβο, πρέπει να χρησιμοποιήσουμε πρώτα έναν αλγόριθμο επεξεργασίας και έπειτα έναν συμπύκνωσης. Μέσω της παραπάνω διαδικασίας σκοπεύουμε να χτίσουμε ένα συμπυκνωμένο σύνολο χωρίς θόρυβο, παράγοντας έναν επαρκή αριθμό αντικειμένων για κάθε κλάση τα οποία θα αποδειχτούν ουσιώδη κατά την εκτέλεση της κατηγοριοποίησης k-NN. Σημειώνουμε πως πολλές ΤΜΔ έχουν ενσωματωθεί στο λογισμικό KEEL [39].

Οι ΤΜΔ που χρησιμοποιούμε στα πλαίσια αυτής της εργασίας είναι Αλγόριθμοι Συμπύκνωσης και Αλγόριθμοι Σύνοψης Προτύπων, καθώς ο σκοπός μας δεν αφορά το θέμα του θορύβου. Αξίζει να σημειωθεί πως δεν χτίζουν όλες οι ΤΜΔ τα συμπυκνωμένα σύνολα τους με τον ίδιο τρόπο. Ανάλογα με την μεθοδολογία που ακολουθεί η κάθε τεχνική μπορεί είτε: (i) να δημιουργεί ένα καινούριο, άδειο συμπυκνωμένο σύνολο και να προσθέτει αντικείμενα σε αυτό κατά την ανάλυση, ή (ii) να χρησιμοποιεί το αρχικό σύνολο εκπαίδευσης και να αφαιρεί από αυτό τα αντικείμενα που υπολογίζει πως είναι περιττά.

2.3.2 CNN-Rule

Ο Κανόνας Συμπύκνωσης Κοντινότερου Γείτονα (*Condensing Nearest Neighbor Rule,* ***CNN-Rule***) [40], πρόκειται για τον πρώτο αλγόριθμο συμπύκνωσης και είναι μία χρήσιμη ΤΜΔ. Όπως και διάφορες άλλες ΤΜΔ, ακολουθεί την λογική ότι τα αντικείμενα που δεν βρίσκονται κοντά στα "σύνορα" των κλάσεων δεν είναι χρήσιμα κατά την διάρκεια της κατηγοριοποίησης των δεδομένων. Συνεπώς τα διαγράφει, καθώς δεν μπορούν να επηρεάσουν την ακρίβεια (*accuracy*). Δημιουργεί ένα συμπυκνωμένο σύνολο που αποτελείται από τα δεδομένα που βρίσκονται κοντά στα σύνορα, τα μόνα δεδομένα δηλαδή που θεωρεί πως

Αλγόριθμος 1 CNN-Rule

Input: Training Set (TS), **Output:** Condensing Set (CS)

```
1: CS ← ∅
2: pick an item of TS and move it to CS
3: repeat
4:    stop ← TRUE
5:    for each x ∈ TS do
6:       NN ← Nearest Neighbour of x in CS
7:       if NN_class ≠ x_class then
8:          CS ← CS ∪ {x}
9:          TS ← TS − {x}
10:         stop ← FALSE
11:      end if
12:   end for
13: until stop == TRUE {no move during a pass of TS}
14: discard TS
15: return CS
```

μπορούν να συνεισφέρουν στον υπολογισμό της ακρίβειας. Ο CNN-Rule αποτελεί την βάση, καθώς και ένα μέσο σύγκρισης για τους υπόλοιπους αλγορίθμους του είδους του.

Ο ψευδοκώδικας του CNN-Rule φαίνεται στον Αλγόριθμο 1. Αρχικά, ένα αντικείμενο από το σύνολο εκπαίδευσης (TS) μεταφέρεται στο μέχρι τώρα κενό συμπυκνωμένο σύνολο (CS). Έπειτα, ο CNN-Rule εκτελεί τον 1-NN rule και κατηγοριοποιεί τα αντικείμενα στο TS σκανάροντας το CS. Εφόσον ένα αντικείμενο δεν είναι σωστά κατηγοριοποιημένο, μετακινείται από το TS στο CS. Σημειώνουμε πως το αντικείμενο φεύγει τελείως από το TS, δεν αντιγράφεται. Ο αλγόριθμος συνεχίζει να εκτελείται μέχρι να μην υπάρχουν άλλα αντικείμενα στο TS ή μέχρι να μην μετακινηθεί κανένα αντικείμενο από το TS στο CS για ένα ολόκληρο πέρασμα του TS. Όσα αντικείμενα μένουν στο TS απορρίπτονται.

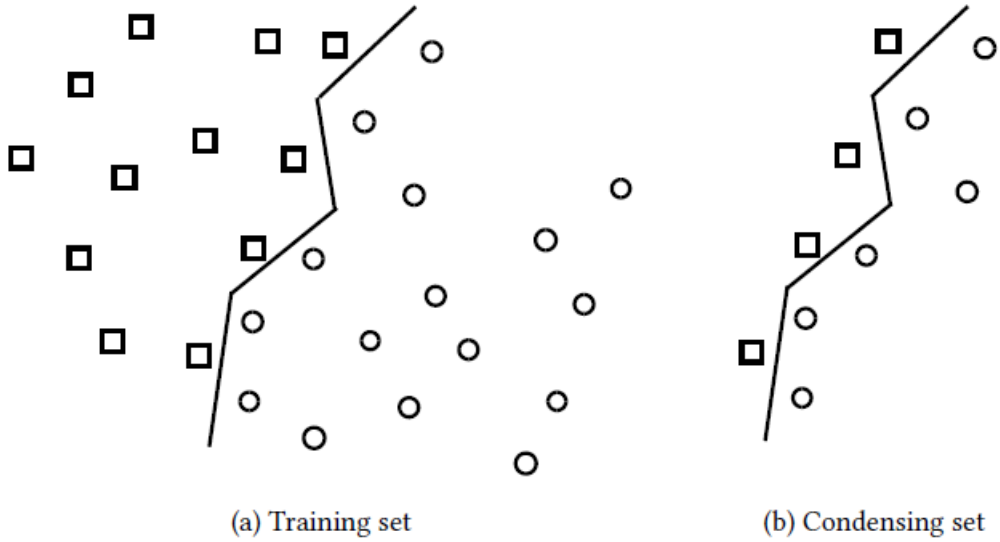
Αξίζει να αναφερθεί πως επειδή ο CNN-Rule εκτελείται πολλές φορές, εγγυάται πως τα αφαιρούμενα αντικείμενα εκπαίδευσης (*training items*) μπορούν να κατηγοριοποιηθούν σωστά στα πλαίσια του τελικού συμπυκνωμένου συνόλου που δημιουργήθηκε. Δηλαδή, αν προσπαθήσουμε να εκτελέσουμε τον αλγόριθμο 1-NN στα αντικείμενα εκπαίδευσης που δεν μπήκαν το συμπυκνωμένο σύνολο, ψάχνοντας για τον έναν κοντινότερο γείτονα σε αυτό το συμπυκνωμένο σύνολο θα προκύπτει ακρίβεια 100%. Επιπλέον ο CNN-Rule είναι μη-παραμετρικός αλγόριθμος, δηλαδή δεν δέχεται κάποια παράμετρο από τον χρήστη. Συνεπώς, ο CNN-Rule ορίζει τον αριθμό των αντικειμένων που παράγει αυτόματα, χωρίς να χρειάζεται να εισάγει κάποια πληροφορία ο χρήστης.

Παρόλα αυτά, όπως αναφέραμε και πριν ο CNN-Rule είναι ο πρώτος αλγόριθμος συμπύκνωσης, και αν και θεωρείται πετυχημένος, έχει και διάφορα μειονεκτήματα. Ένα από αυτά είναι ότι η ποιότητα του τελικού συμπυκνωμένου συνόλου εξαρτάται από την σειρά των αντικειμένων εκπαίδευσης. Είναι πιθανόν επομένως ο CNN-Rule να δημιουργήσει διαφορετικά

συμπυκνωμένα σύνολα αναλύοντας τα ίδια δεδομένα αλλά με διαφορετική σειρά. Επιπλέον, ο CNN-Rule δεν μπορεί να χειριστεί καινούρια δεδομένα εκπαίδευσης, οπότε δεν μπορεί να ενημερώσει ένα συμπυκνωμένο σύνολο με δεδομένα που έλαβε μετά την δημιουργία του. Επομένως, δεν είναι κατάλληλος για δυναμικά περιβάλλοντα όπου νέα δεδομένα εκπαίδευσης γίνονται σταδιακά διαθέσιμα, καθώς ο μόνος τρόπος για να επεξεργαστεί καινούρια δεδομένα είναι να ξανα-εκτελέσουμε τον αλγόριθμο από την αρχή, περιλαμβάνοντας αυτήν την φορά και αυτά.

Όπως αναφέραμε, ο CNN-Rule επιλέγει τα δεδομένα που βρίσκονται κοντά στα σύνορα κλάσεων, τα οποία προσπαθεί να τα εντοπίσει ψάχνοντας για αντικείμενα που δεν είναι σωστά κατηγοριοποιημένα χρησιμοποιώντας τον 1-NN. Θεωρεί επομένως πως όσα αντικείμενα βρίσκει να είναι τοποθετημένα κοντά σε αυτούς τους διαχωρισμούς πρέπει να αποθηκευτούν μέσα στο συμπυκνωμένο σύνολο. Αυτό όμως αποτελεί και ένα σημαντικό μειονέκτημα, καθώς μπορεί έτσι να δεχτεί αντικείμενα που είναι θόρυβος. Τα δεδομένα που αποτελούν θόρυβο ανήκουν σε μια διαφορετική κλάση από τα αντικείμενα τριγύρω τους. Λανθασμένα λοιπόν, αυτά τα αντικείμενα θα αποθηκευτούν στο συμπυκνωμένο σύνολο μαζί με τους πιο κοντινούς γείτονες τους αφού ο CNN-Rule θα τα θεωρήσει αντικείμενα που βρίσκονται κοντά στα όρια των κλάσεων. Όσο περισσότερος θόρυβος υπάρχει ανάμεσα στο σύνορα δεδομένων, τόσο πιο πολύ επηρεάζεται το αποτέλεσμα και τόσο πιο πολύ μικραίνει το ποσοστό μείωσης (*reduction rate*) των δεδομένων.

Τέλος, ένας επιπλέον παράγοντας ο οποίος μπορεί να επηρεάσει το ποσοστό μείωσης δεδομένων είναι ο αριθμός των διαθέσιμων κλάσεων.Όσο περισσότερες κλάσεις υπάρχουν, τόσο περισσότερο αυξάνεται το πλήθος των ορίων απόφασης και επομένως τόσο αυξάνεται



Σχήμα 2.7: Διαφορά μεγέθους μεταξύ του συνόλου εκπαίδευσης και του συμπυκνωμένου συνόλου.

και ο αριθμός των αντικειμένων που βρίσκονται κοντά στα όρια και αποθηκεύονται στο συμπυκνωμένο σύνολο.

Για διευκόλυνση, στο Σχήμα 2.7 μπορούμε να δούμε ένα σύνολο δεδομένων πριν και μετά την εκτέλεση του CNN-Rule. Παρατηρούμε πως επειδή αφαιρέθηκαν όσα αντικείμενα δεν ήταν κοντά στα σύνορα, το μέγεθος του συνόλου δεδομένων μίκρυνε πολύ, χωρίς να χαλάσει η δομή του. Ο κατηγοριοποιητής k-NN θα επιτύχει την ίδια ακρίβεια είτε χρησιμοποιήσει το σύνολο εκπαίδευσης, είτε χρησιμοποιήσει το συμπυκνωμένο σύνολο. Ωστόσο, η εκτέλεση του k-NN στο συμπυκνωμένο σύνολο θα είναι πολύ πιο γρήγορη, καθώς δεν χρειάζεται να υπολογιστούν τόσες αποστάσεις όσες με το αρχικό σύνολο εκπαίδευσης.

2.3.3 IB2

Ο δεύτερος αλγόριθμος συμπύκνωσης που μελετάμε ονομάζεται Αλγόριθμος Μάθησης Μέσω Παραδειγμάτων (*Instance-Based Learning,* ***IBL***) [41, 42]. Υπάρχουν τρεις διαθέσιμοι IBL αλγόριθμοι, ωστόσο εμείς συγκεντρωνόμαστε στον δεύτερο. Ο αλγόριθμος IB2 είναι αρκετά παρόμοιος με τον CNN-Rule, τόσο πολύ που θα μπορούσαν να θεωρηθεί παραλλαγή του. Όπως και ο CNN-Rule συγκρίνει δεδομένα με τον πιο κοντινό τους γείτονα, με την κύρια διαφορά ότι τα δεδομένα αναλύονται μόνο μία φορά, όχι επαναληπτικά. Ο IB2 λοιπόν, όπως και ο CNN-Rule, προσπαθεί να συλλέξει στο συμπυκνωμένο σύνολο μόνο τα αντικείμενα που βρίσκονται κοντά στα σύνορα των κλάσεων.

Η εκτέλεση του IB2 έχει ως εξής: ξεκινάμε με ένα κενό συμπυκνωμένο σύνολο (CS). Βάζουμε ένα τυχαίο αντικείμενο του συνόλου εκπαίδευσης (TS) σε αυτό. Κάθε αντικείμενο x_i του TS κατηγοριοποιείται χρησιμοποιώντας τον 1-NN Rule με το CS που έχουμε μέχρι τώρα. Αν το x_i κατηγοριοποιηθεί σωστά, τότε το απορρίπτουμε, αλλιώς μεταφέρεται στο CS. Ο ψευδοκώδικας του IB2 φαίνεται στον Αλγόριθμο 2. Ο IB2 έχει έναν σημαντικό αριθμό χαρακτηριστικών που πολλές άλλες ΤΜΔ στερούνται:

Αλγόριθμος 2 IB2

Input: Training Set (TS), **Output:** Condensing Set (CS)

$CS \leftarrow \varnothing$
pick an item of TS and move it to CS
for each $x \in TS$ **do**
  $NN \leftarrow$ Nearest Neighbour of x in CS
  if $NN_{class} \neq x_{class}$ **then**
    $CS \leftarrow CS \cup \{x\}$
  end if
  $TS \leftarrow TS - \{x\}$
end for
return CS

- Το κύριο πλεονέκτημα είναι η ταχύτητα του. Αφού πραγματοποιεί μόνο ένα πέρασμα, ο χρόνος εκτέλεσης του είναι πολύ χαμηλότερος σε σύγκριση με αλγόριθμους όπως ο CNN-Rule.

- Δεν απαιτεί όλα τα δεδομένα εκπαίδευσης να βρίσκονται στην κεντρική μνήμη. Επομένως, μπορεί να εκτελεστεί και σε συσκευές που δεν έχουν αρκετή μνήμη για να διατηρούν όλα τα δεδομένα εκπαίδευσης που χρειαζόμαστε ταυτόχρονα.

- Εκτός αυτού, ο IB2 μπορεί να χτίζει σταδιακά το συμπυκνωμένο σύνολο του, οπότε μπορεί να υπολογίζει και αντικείμενα εκπαίδευσης που έγιναν διαθέσιμα μετά από την κατασκευή αυτού του συμπυκνωμένου συνόλου.

- Μπορεί να χειριστεί νέες ετικέτες κλάσεων (*class labels*), κάτι που σε συνδυασμό με την μεγαλύτερη ελευθερία όσο αφορά το συμπυκνωμένο σύνολο, καθιστά τον αλγόριθμο κατάλληλο για δυναμικά περιβάλλοντα, όπως είναι το streaming.

Όμως, αφού εκτελείτε μόνο ένα σκανάρισμα των δεδομένων, ο IB2 δεν μπορεί να εγγυηθεί πως όλα τα δεδομένα έχουν κατηγοριοποιηθεί σωστά. Αυτή είναι η "ανταλλαγή" που κάνει για την ταχύτητα του. Επιπλέον, ο IB2 μοιράζεται μερικά από τα μειονεκτήματα του CNN-Rule. Πρώτον η δομή του συμπυκνωμένου συνόλου βασίζεται πολύ στην σειρά των δεδομένων στο σύνολο εκπαίδευσης. Δεύτερον, όπως αναφέραμε και πριν το συμπυκνωμένο σύνολο κρατάει μόνο τα αντικείμενα που βρίσκονται κοντά στα σύνορα. Ωστόσο όμως μπορούν να είναι και θόρυβος. Αυτό σημαίνει ότι ο IB2 είναι πολύ ευαίσθητος στην ύπαρξη θορύβου. Και καθώς ο θόρυβος σχετίζεται άμεσα με το ποσοστό μείωσης δεδομένων, όσο περισσότερο υπερισχύει το πρώτο, τόσο πιο πολύ επιδεινώνεται το δεύτερο και ως συνέπεια επηρεάζεται αρνητικά το τελικό αποτέλεσμα του αλγορίθμου. Παρόλα αυτά τα πλεονεκτήματα που διαθέτει τον καθιστούν έναν χρήσιμο αλγόριθμο που προσφέρει αξιόλογα αποτελέσματα.

2.3.4 RSP3

Πρόκειται για έναν από τους τρεις αλγορίθμους Μείωσης με Διαχωρισμού Χώρου (*Reduction by Space Partitioning,* ***RSP***) [43], οι οποίοι είναι βασισμένοι στον αλγόριθμο CJA (*Chen and Jozwik Algorithm*) [44]. Όλοι αυτοί οι αλγόριθμοι ανήκουν στην κατηγορία αλγορίθμων σύνοψης δεδομένων και λειτουργούν με την ίδια βασική λογική: χωρίζουν το σύνολο εκπαίδευσης σε πολλά μικρότερα υποσύνολα μέχρι να ικανοποιούν κάποιο συγκεκριμένο κριτήριο. Στην περίπτωση του RSP3, το κριτήριο αυτό είναι η ομοιογένεια. Ένα υποσύνολο θεωρείται ομοιογενές όταν όλα τα αντικείμενα που περιέχει ανήκουν στην ίδια κλάση.

Ο αλγόριθμος CJA λειτουργεί χωρίζοντας το σύνολο εκπαίδευσης σε μικρότερα υποσύνολα. Αυτό το κάνει ψάχνοντας τα δύο πιο απομακρυσμένα αντικείμενα μέσα στο σύνολο

εκπαίδευσης, έστω πως αυτά είναι δύο αντικείμενα A και B. Όταν το σύνολο εκπαίδευσης χωρίζεται σε δύο υποσύνολα, έστω Σ_A και Σ_B, τα αντικείμενα που είναι τοποθετημένα πιο κοντά στο A τοποθετούνται στο πρώτο υποσύνολο, ενώ αυτά που είναι πιο κοντά στο B εισάγονται στο δεύτερο υποσύνολο. Αυτή η διαδικασία συνεχίζεται επαναληπτικά, επομένως Σ_A και Σ_B χωρίζονται και αυτά. Όσο εκτελείται ο CJA συνεχίζονται να δημιουργούνται όλο και μικρότερα υποσύνολα. Για να αποφασίσει ποιο από τα υποσύνολα θα χωριστεί πρώτο, ο CJA ψάχνει για το μη-ομοιογενές υποσύνολο με την μεγαλύτερη διάμετρο, η οποία ορίζεται από τα δύο πιο απομακρυσμένα αντικείμενα μέσα σε αυτό, και το διαλέγει. Αν όλα τα υποσύνολα που έχουμε είναι ομοιογενή, τότε αυτό που χωρίζεται είναι το υποσύνολο με την μεγαλύτερη διάμετρο. Ο CJA συνεχίζει να εκτελεί αυτήν την διαδικασία μέχρι να παράγει έναν αριθμό υποσυνόλων ίσο με μία παράμετρο που έχει ορίσει ο χρήστης. Ο λόγος που χρησιμοποιείται η διάμετρος ως το κριτήριο επιλογής για το επόμενο υποσύνολο, είναι επειδή θεωρείται πως το υποσύνολο με την μεγαλύτερη διάμετρο θα περιέχει περισσότερα αντικείμενα από τα υπόλοιπα και λόγω αυτού θα πετύχει μεγαλύτερο ποσοστό μείωσης δεδομένων.

Αφού ολοκληρωθεί αυτή η διαδικασία, ο CJA βρίσκει την πλειοψηφούσα κλάση σε κάθε υποσύνολο. Έπειτα, για κάθε υποσύνολο δημιουργεί μια σύνοψη των αντικειμένων του, βρίσκοντας το μέσο όρο των χαρακτηριστικών (*attributes*) τους. Μέσω αυτής της διαδικασίας δημιουργείται ένα αντιπροσωπευτικό αντικείμενο για κάθε υποσύνολο. Στην κατασκευή αυτού του νέου αντικειμένου, λαμβάνονται υπόψη μόνο τα αντικείμενα της πλειοψηφούσας κλάσης του κάθε υποσυνόλου. Τα υπόλοιπα αγνοούνται. Αυτά τα αντιπροσωπευτικά αντικείμενα εισάγονται στο συμπυκνωμένο σύνολο.

Συνεχίζοντας θα εξηγήσουμε τον αλγόριθμο RSP1, ο οποίος είναι αρκετά παρόμοιος με τον CJA. Η διαφορά μεταξύ τους είναι πως σε αντίθεση με τον CJA, ο RSP1 για κάθε υποσύνολο δημιουργεί ένα αντιπροσωπευτικό αντικείμενο για κάθε ξεχωριστή κλάση μέσα σε αυτό. Σε κάθε υποσύνολο βρίσκει επομένως τον μέσο όρο των χαρακτηριστικών των αντικειμένων κάθε κλάσης του. Ως αποτέλεσμα δημιουργεί ένα μεγαλύτερο συμπυκνωμένο σύνολο από τον CJA. Καθώς ο RSP1 δεν αγνοεί αντικείμενα, τα συμπυκνωμένα σύνολα που παρέχει στον κατηγοριοποιητή k-NN τον βοηθάνε να πετύχει μεγαλύτερα ποσοστά ακρίβειας.

Ο αλγόριθμος RSP2 διαφέρει με τους δύο προηγούμενους αλγορίθμους όσο αφορά τον τρόπο επιλογής του επόμενου υποσυνόλου που θα χωριστεί. Με τον CJA και τον RSP1 όπως αναφέραμε είδη, χωρίζουμε το υποσύνολο με την μεγαλύτερη διάμετρο καθώς θεωρητικά αυτό θα έχει τα περισσότερα αντικείμενα και έτσι θα επιτύχουμε μεγαλύτερο ποσοστό μείωσης δεδομένων. Αυτό όμως δεν ισχύει σε κάθε περίπτωση, καθώς μπορεί να έχουμε ένα σχετικά μικρότερο υποσύνολο που περιέχει περισσότερα αντικείμενα από άλλα υποσύνολα με μεγαλύτερη διάμετρο. Για αυτόν τον λόγο, ο RSP2 χρησιμοποιεί ένα διαφορετικό κριτήριο από την διάμετρο. Αυτό το κριτήριο είναι ο βαθμός επικάλυψης, ο οποίος υπολογίζεται μέσω ενός κλάσματος, για το οποίο ο αριθμητής είναι ο μέσος όρος των αποστάσεων των

αντικειμένων που βρίσκονται σε διαφορετικές κλάσεις και ο παρονομαστής είναι ο μέσος όρος των αποστάσεων από αντικείμενα που ανήκουν στην ίδια κλάση. Η μόνη διαφορά του RSP2 με τον RSP1 είναι πως διαλέγει να χωρίσει το υποσύνολο με την μεγαλύτερη επικάλυψη, αντί για αυτό με την μεγαλύτερη διάμετρο. Ο στόχος του RSP2 είναι να πετύχει μεγαλύτερα ποσοστά μείωσης δεδομένων σε σύγκριση με τους προηγούμενους δύο αλγορίθμους.

Τέλος έχουμε τον αλγόριθμο RSP3, ο οποίος δεν υιοθετεί κάποιο συγκεκριμένο κριτήριο για τη διαίρεση των υποσυνόλων ή πιο συγκεκριμένα, δεν είναι σημαντικό το ποιο υποσύνολο θα διασπαστεί πρώτο. Όπως είπαμε και πριν, ο RSP3 υιοθετεί την έννοια της ομοιογένειας των υποσυνόλων. Επομένως, η διαδικασία του χωρισμού του συνόλου εκπαίδευσης σε μικρότερα υποσύνολα συνεχίζεται μέχρι όλα τα διαθέσιμα υποσύνολα να είναι ομοιογενή, δηλαδή να περιλαμβάνουν αντικείμενα μόνο μίας κλάσης.

Ο ψευδοκώδικας για τον RSP3 φαίνεται παρακάτω στον Αλγόριθμο 3. Στην αρχή, δημιουργεί μία δομή δεδομένων S στην οποία τοποθετούνται τα υποσύνολα που δεν έχουν επεξεργαστεί ακόμα. Επομένως, το TS σαν σύνολο μπαίνει μέσα στο S. Σε κάθε επανάληψη, ο RSP3 διαλέγει το υποσύνολο C με το μεγαλύτερο κριτήριο διαχωρισμού και ελέγχει αν το C είναι ομοιογενές. Αν είναι, υπολογίζεται το μέσο αντικείμενο βρίσκοντας τον μέσο όρο όλων των αντικειμένων στο C και μετακινείται στο συμπυκνωμένο σύνολο (CS), ως ένα πρότυπο (*prototype)*. Αλλιώς, το C χωρίζεται σε δύο υποσύνολα $D1$ και $D2$. Αυτά τα δύο υποσύνολα εισάγονται στο S, ενώ το C αφαιρείται από το S. Η διαδικασία επαναλαμβάνεται όταν το S έχει αδειάσει. Τότε καταλαβαίνουμε πως όλα τα διαθέσιμα υποσύνολα είναι πλέον ομοιογενή.

Αλγόριθμος 3 RSP3

Input: Training Set (TS), **Output:** Condensing Set (CS)

$S \leftarrow \varnothing$
add(S, TS)
$CS \leftarrow \varnothing$
repeat
  $C \leftarrow$ select the subset ϵ S with the highest splitting criterion value
  if C is homogeneous **then**
    $r \leftarrow$ calculate the mean item by averaging the items in C
    $r.label \leftarrow$ class of items in C
    $CS \leftarrow CS \cup \{r\}$
  else
    $(D_1, D_2) \leftarrow$ divide C into two subsets
    add(S, D_1)
    add(S, D_2)
    remove(S, C)
  end if
until IsEmpty(S)
return CS

Ο RSP3 διαθέτει μερικά πλεονεκτήματα που υστερούν οι υπόλοιποι αλγόριθμοι διαχωρισμού χώρου. Ένα από αυτά τα πλεονεκτήματα είναι πως δεν χρειάζεται την εισαγωγή κάποιας παραμέτρου για να ορίσει το μέγεθος του συμπυκνωμένου συνόλου. Είναι μη-παραμετρικός αλγόριθμος και το κάνει αυτόματα, σε αντίθεση με τους προηγούμενους. Έτσι αποφεύγονται οι επαναλαμβανόμενες δοκιμές για τον καθορισμού παραμέτρων.

Ο RSP3 πραγματοποιεί πιο πετυχημένη κατηγοριοποίηση, αλλά συνήθως χαμηλότερα ποσοστά μείωσης σε σχέση με άλλους αλγορίθμους. Αυτό συμβαίνει επειδή το ποσοστό μείωσης των δεδομένων του RSP3 μπορεί να επιβαρυνθούν σημαντικά από υψηλά επίπεδα θορύβου. Ο θόρυβος παρεμποδίζει την ομοιογένεια, επομένως όσο περισσότερος θόρυβος υπάρχει στο σύνολο δεδομένων, τόσο πιο πολύ αργεί η ολοκλήρωση της εκτέλεσης του RSP3 καθώς συνεχίζει να δημιουργεί όλο και πιο μικρά υποσύνολα. Το ίδιο ισχύει και αν τα δεδομένα κάποιας κλάσης είναι διαχωρισμένα σε πολλαπλά σημεία του υποσυνόλου.

Μελετώντας τον RSP3 γίνεται αντιληπτό πως ο αλγόριθμος δημιουργεί πολλά αντιπροσωπευτικά αντικείμενα για τις περιοχές δεδομένων που βρίσκονται κοντά στα σύνορα μεταξύ κλάσεων και ελάχιστα αντιπροσωπευτικά αντικείμενα για τις εσωτερικές περιοχές των κλάσεων, καθώς είναι ομοιογενείς στις περισσότερες περιπτώσεις.

Η ταχύτητα εκτέλεσης του RSP3 μπορεί να επηρεαστεί από το μέγεθος του συνόλου δεδομένων, καθώς ο αλγόριθμος θα πρέπει να υπολογίσει όλες τις αποστάσεις μεταξύ κάθε ζεύγους αντικειμένων μέσα σε κάθε υποσύνολο. Επομένως, όσο μεγαλύτερο είναι το σύνολο δεδομένων τόσο αυξάνεται το κόστος προεπεξεργασίας. Υπό ορισμένες περιπτώσεις μπορεί να μην συνιστάται καν η εκτέλεση του RSP3.

Τέλος, αξίζει να αναφερθεί πως το συμπυκνωμένο σύνολο που παράγεται από τους τρεις αλγορίθμους RSP, καθώς και τον CJA, δεν επηρεάζεται από την σειρά των δεδομένων στο σύνολο εκπαίδευσης.

2.3.5 AIB2

Πρόκειται για μία παραλλαγή του IB2 που ανήκει στην κατηγορία αλγορίθμων σύνοψης δεδομένων. Ο AIB2 [45] προσπαθεί να πετύχει μεγαλύτερη ακρίβεια προσθέτοντας "βάρος" στα αντικείμενα. Ουσιαστικά, σε κάθε αντικείμενο που μπαίνει στο συμπυκνωμένο σύνολο δίνεται μία τιμή που αντιπροσωπεύει το βάρος του. Σε αντίθεση με τον IB2, τα αντικείμενα που κατηγοριοποιούνται σωστά από τον 1-NN δεν αγνοούνται. Συμβάλλουν στον υπολογισμό του συμπυκνωμένου συνόλου, μεταβάλλοντας τη θέση του πιο κοντινού πρότυπου μέσα σε αυτό. Το αναφερόμενο πρότυπο μετακινείται προς την θέση του αντικείμενου που κατηγοριοποιήθηκε σωστά και το βάρος του αυξάνεται.

Μπορούμε να δούμε πως δουλεύει ο AIB2 με την βοήθεια του Αλγορίθμου 4. Στην

Αλγόριθμος 4 AIB2

Input: Training Set (TS), **Output:** Condensing Set (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: $NN_{weight} = NN_{weight} + 1$
14: **end if**
15: $TS \leftarrow TS - \{x\}$
16: **end for**
17: **return** CS

αρχή διαλέγει ένα τυχαίο αντικείμενο από το σύνολο εκπαίδευσης (TS) και το βάζει στο συμπυκνωμένο σύνολο (CS), με βάρος ίσο με 1. Έπειτα, για κάθε αντικείμενο, ο AIB2 ψάχνει μέσα στο CS και επιστρέφει το πιο κοντινό γείτονα του (NN, Nearest Neighbour). Αν το x δεν είναι σωστά κατηγοριοποιημένο, το τοποθετεί στο CS και του θέτει και αυτού βάρος ίσο με 1. Αντιθέτως, αν το x είναι σωστά κατηγοριοποιημένο, τότε δεν αγνοείται όπως στην περίπτωση του IB2. Αντιθέτως, τα χαρακτηριστικά του NN ενημερώνονται ανάλογα, παίρνοντας υπόψη τα χαρακτηριστικά του x. Συγκεκριμένα, το NN μετακινείται προς το x. Στο τέλος, το βάρος του NN αυξάνεται κατά 1 και το x φεύγει. Έτσι, το x αντιπροσωπεύεται από το NN, αφού το NN μετακινήθηκε προς το x.

Αυτή η διαδικασία δεν χαμηλώνει το ποσοστό μείωσης δεδομένων. Για την ακρίβεια, επειδή το τελικό συμπυκνωμένο σύνολο περιέχει λιγότερα αντικείμενα από ότι αυτά που παράγει ο IB2, το ποσοστό μείωσης βελτιώνεται. Εκτός αυτού, παράλληλα επιτυγχάνει και χαμηλότερο κόστος προεπεξεργασίας.

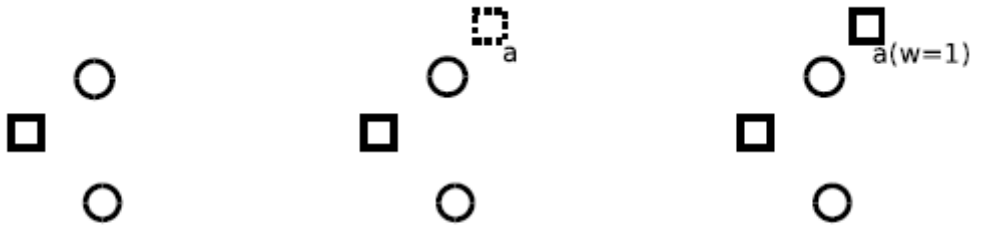
Αξίζει επίσης να σημειωθεί πως όσο μεγαλύτερο είναι το βάρος του NN, τόσο λιγότερο μετακινείται προς το x. Για παράδειγμα, αν το NN έχει βάρος 1 θα διασχίσει την μισή απόσταση προς το x. Ενώ αν έχει βάρος 2, θα διασχίσει μόνο το 1/3 και ούτω καθεξής. Αυτό φυσικά σημαίνει ότι τα αντικείμενα με πολύ μεγάλο βάρος κινούνται πολύ αργά. Το βάρος μπορεί επίσης να χρησιμοποιηθεί και ως μετρητής για τον αριθμό των αντικειμένων που αντιπροσωπεύονται από ένα πρότυπο που αποθηκεύεται στο συμπυκνωμένο σύνολο. Η φιλοσοφία του AIB2 είναι ότι το κάθε αντικείμενο που εισάγεται μέσα στο συμπυκνωμένο σύνολο θα πρέπει να βρίσκεται στο κέντρο της περιοχής την οποία αντιπροσωπεύει. Με την

συνεχή ενημέρωση των αντικειμένων φτάνουμε όλο και πιο κοντά σε αυτήν την κατάσταση.

Η λειτουργία του AIB2 παρουσιάζεται διαγραμματικά στα Σχήματα 2.8 και 2.9. Και στα δύο ξεκινάμε έχοντας διαθέσιμο ένα συμπυκνωμένο σύνολο με τρία αντικείμενα, δύο από τα οποία ανήκουν σε μία κλάση Κύκλος και ένα που ανήκει σε μία κλάση Τετράγωνο. Στο Σχήμα 2.8(b), εισάγετε ένα νέο τετράγωνο. Αφού ο πιο κοντινός τους γείτονας είναι κύκλος, το νέο τετράγωνο εισάγετε στο συμπυκνωμένο σύνολο, όπως φαίνεται στο Σχήμα 2.8(c).

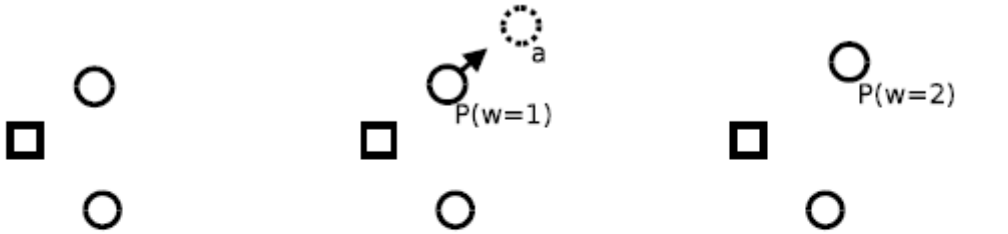
Εναλλακτικά, στο Σχήμα 2.9(b), το νέο αντικείμενο είναι κύκλος αντί για τετράγωνο. Σε αυτήν την περίπτωση, αφού το πιο κοντινό αντικείμενο μέσα στο συμπυκνωμένο σύνολο είναι το P, το οποίο ανήκει στην ίδια κλάση, το νέο αντικείμενο δεν μπαίνει στο συμπυκνωμένο σύνολο. Αντίθετα, ενημερώνει την θέση του P, όπως φαίνεται στο Σχήμα 2.9(c). Παράλληλα το βάρος του P αυξάνεται κατά ένα.

Πέρα από αυτά, ο AIB2 είναι παρόμοιος με τους αλγόριθμους στους οποίους είναι βασισμένος, οπότε συνεπώς μοιράζεται και πολλά χαρακτηριστικά μαζί τους. Για παράδειγμα



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

Σχήμα 2.8: AIB2 - Εισαγωγή ενός νέου αντικειμένου μέσα στο συμπυκνωμένο σύνολο.



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

Σχήμα 2.9: AIB2 - Ενημέρωση της τοποθεσίας και του βάρους ενός αντικειμένου μέσα στο συμπυκνωμένο σύνολο.

όπως ο CNN-Rule και ο IB2, θεωρεί πως τα μόνα αντικείμενα που δεν είναι περιττά είναι αυτά που βρίσκονται κοντά στα σύνορα μεταξύ κλάσεων και προσπαθεί να συλλέξει μόνο αυτά. Πολύ συχνά τα αντικείμενα που βρίσκονται κοντά σε σύνορα είναι λάθος κατηγοριοποιημένα. Οπότε ο AIB2 χρησιμοποιεί αυτό το κριτήριο για να τα εντοπίσει και να τα εισάγει στο συμπυκνωμένο σύνολο. Όμως είναι πιθανόν να υπάρχουν αντικείμενα που αποτελούν θόρυβο. Η παρουσία θορύβου χαμηλώνει τα ποσοστά μειώσεις αφού τα αντικείμενα που αποτελούν θόρυβο μαζί με τα γειτονικά τους μεταφέρονται λανθασμένα στο συμπυκνωμένο σύνολο.

Όπως και ο IB2, ο AIB2 είναι μη-παραμετρικός αλγόριθμος. Επίσης είναι πολύ γρήγορος επειδή εκτελεί μόνο ένα πέρασμα των διαθέσιμων δεδομένων. Όμως, εξαιτίας του δεύτερου δεν μπορεί να εγγυηθεί την σωστή κατηγοριοποίηση όλων των δεδομένων που δεν μεταφέρθηκαν στο συμπυκνωμένο σύνολο, σε αντίθεση με τον CNN-Rule. Αυτό αποτελεί και το κεντρικό μειονέκτημα του αλγορίθμου. Όμως, χάρις στην ταχύτητα του και το γεγονός ότι μπορεί να χειριστεί νέες ετικέτες κλάσεων (*class labels*), ο αλγόριθμος θεωρείται κατάλληλος για τον βέλτιστο χειρισμό δυναμικών περιβαλλόντων (π.χ. streaming) όπου τα δεδομένα εκπαίδευσης γίνονται σταδιακά διαθέσιμα. Όπως και ο IB2, μπορούμε να τον χρησιμοποιήσουμε σε συσκευές που δεν απαιτούν όλα τα δεδομένα τους να είναι στην κεντρική μνήμη, οπότε ο AIB2 είναι ένας χρήσιμος και για πολύ μεγάλα σύνολα δεδομένων.

Αν κατά την εκτέλεση του αλγορίθμου γίνουν διαθέσιμα νέα αντικείμενα εκπαίδευσης, ο AIB2 μπορεί να τα χρησιμοποιήσει. Παρόλα αυτά όμως, τα πρότυπα που θα χτίσουμε με τον AIB2 εξαρτώνται από την σειρά των αντικειμένων εκπαίδευσης. Λόγου αυτού, υπάρχει η περίπτωση ένα από τα πρότυπα του να μετακινηθεί σταδιακά όλο και πιο μακριά από μερικά από τα αντικείμενα που αντιπροσωπεύει.

2.3.6 RHC

Επιστρέφοντας στο θέμα της ομοιογένειας, ο επόμενος αλγόριθμος είναι η Μείωση μέσω Ομοιογενών Συστάδων (*Reduction through Homogenous Clusters,* ***RHC***) [45, 46]. Και αυτός ο αλγόριθμος ανήκει στην κατηγόρια αλγορίθμων σύνοψης δεδομένων. Ουσιαστικά, ο RHC πρόκειται για έναν αλγόριθμο που εκτελεί επαναληπτικά την γνωστή συσταδοποίηση k-Means (*k-Means clustering*) [47] μέχρι να έχουμε μόνο ομοιογενείς συστάδες.

Στην αρχή αναλύεται όλο το σύνολο εκπαίδευσης και βρίσκεται ο αριθμός των κλάσεων. Για κάθε κλάση, ο RHC μαζεύει όλα τα διαθέσιμα αντικείμενα και υπολογίζει τον μέσο όρο των τιμών των χαρακτηριστικών τους (*attribute values*). Έπειτα, ξεκινάει η εκτέλεση του k-Means Clustering χρησιμοποιώντας αυτούς τους μέσους όρους ως αρχικά κέντρα. Από τις συστάδες που βρίσκουμε, για όσες είναι ομοιογενείς σταματάει η διαδικασία, ενώ για την αντίθετη περίπτωση συνεχίζουμε να εκτελούμε τον k-Means για κάθε μη-ομοιογενή συστάδα

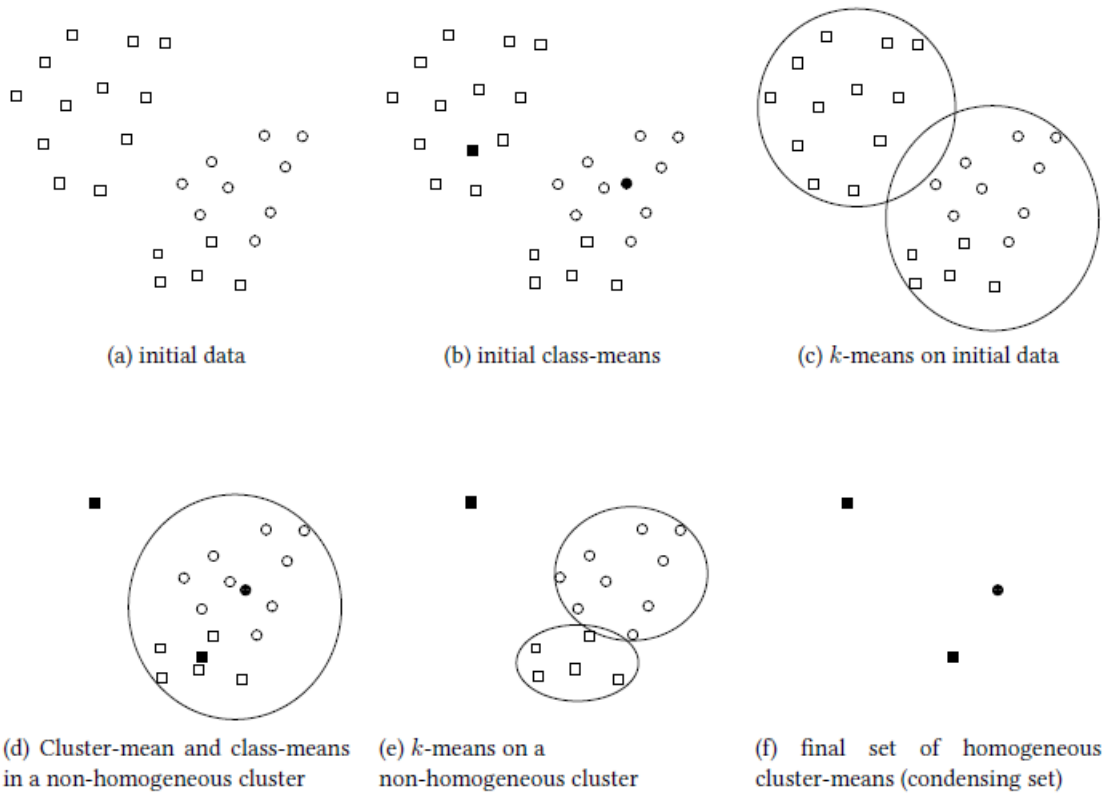
και αυτό γίνεται μέχρι να εξαφανιστεί κάθε ένδειξη μη-ομοιογένειας. Όταν βρίσκουμε μία ομοιογενής συστάδα, χτίζουμε το αντιπροσωπευτικό αντικείμενο της και το βάζουμε στο συμπυκνωμένο σύνολο.

Ένα χαρακτηριστικό του RHC είναι η ταχύτητα του. Υπό άλλες συνθήκες, πριν αρχίσει τους υπολογισμούς της η συσταδοποίηση k-Means χάνει χρόνο διαλέγοντας δεδομένα για να χρησιμοποιήσει σαν κέντρα. Εδώ όμως αντικαταστάθούμε αυτό το βήμα με τον υπολογισμό των μέσων τιμών των χαρακτηριστικών για τα αντικείμενα κάθε κλάσης, κάτι που επιταχύνει την εκτέλεση του αλγορίθμου δραστικά. Αυτό σημαίνει επίσης πως στην αρχή της συσταδοποίησης k-Means πολλές φορές μπορούν να χτιστούν μεγάλες ομοιογενείς συστάδες σχεδόν κατευθείαν. Ο αλγόριθμος είναι επίσης μη-παραμετρικός. Δεν απαιτεί από τον χρήστη να εισάγει κάποια παράμετρο.

Αρχικά μπορεί να μας φαίνεται πως στο τέλος έχουμε μία ομοιογενής συστάδα για κάθε κλάση και το συμπυκνωμένο σύνολο αποτελείται από τα αντιπροσωπευτικά αντικείμενα της κάθε συστάδας. Αυτό δεν ισχύει όμως. Τα δεδομένα των κλάσεων συνήθως είναι διασκορπισμένα με τέτοιο τρόπο που δεν μπορούμε να χτίσουμε μόνο ομοιογενείς συστάδες από το πρώτο πέρασμα. Έτσι, ο k-means εκτελείται στα δεδομένα κάθε μη ομοιογενής συστάδας που προκύπτει. Συνεπώς, δημιουργούνται πολλές ομοιογενείς συστάδες για κάθε κλάση. Βέβαια, υπάρχει πάντα η περίπτωση να υπάρχει θόρυβος ανάμεσα στα δεδομένα μας. Σε αυτήν την περίπτωση, θα δημιουργηθεί μια συστάδα για κάθε αντικείμενο που είναι θόρυβος.

Ας υποθέσουμε για παράδειγμα πως έχουμε δύο κλάσεις A και B όπου τα αντικείμενα τους είναι αρκετά καλά χωρισμένα, εξαιρώντας έναν μικρό αριθμό αντικειμένων της A που βρίσκονται ανάμεσα σε πολλά αντικείμενα της B. Σε αυτήν την περίπτωση θα χτιστούν τρεις συστάδες, μία για την κλάση B και δύο για την A. Επομένως, όσο μεγαλύτερη ποικιλία υπάρχει, καθώς και όσο πιο "ανακατεμένα" είναι τα αντικείμενα μέσα στο σύνολο εκπαίδευσης, τόσο πιο πολλά σύνορα υπάρχουν και κατά συνέπεια τόσες πιο πολλές συστάδες θα δημιουργηθούν. Ο συνολικός αριθμός συνόρων μεταξύ κλάσεων μπορεί να επηρεάσει και το ποσοστό μείωσης δεδομένων. Σαφώς, όσο μεγαλύτερος είναι ο τελικός αριθμός συστάδων, τόσο μικρότερος είναι ο ρυθμός μείωσης.

Στο Σχήμα 2.10 φαίνεται άλλο ένα τέτοιο παράδειγμα και παρουσιάζεται διαγραμματικά η εκτέλεση του RHC. Έστω ότι έχουμε ένα σύνολο δεδομένων με αντικείμενα που μπορούν να ανήκουν σε δύο διαφορετικές κλάσεις, το Τετράγωνο ή τον Κύκλο (Σχήμα 2.10a). Ο RHC θα υπολογίσει ένα αντιπροσωπευτικό αντικείμενο για την κάθε κλάση (Σχήμα 2.10b). Βλέπουμε ότι με το πρώτο πέρασμα, από τις δύο συστάδες που φτιάχνει η μία περιέχει μόνο αντικείμενα της κλάσης Τετράγωνο, ενώ η άλλη περιέχει αντικείμενα και από τις δύο κλάσεις (Σχήμα 2.10c). Ο RHC υπολογίζει το αντιπροσωπευτικό αντικείμενο της ομοιογενής συστάδας και το βάζει στο συμπυκνωμένο σύνολο (Σχήμα 2.10d). Έπειτα, η συστάδα που μένει χωρίζεται σε δύο μικρότερες, ομοιογενείς υποσυστάδες (Σχήμα 2.10d, e).



Σχήμα 2.10: Αφαίρεση δεδομένων μέσω του RHC.

Επομένως, τα αντιπροσωπευτικά αντικείμενα των δύο υποσυστάδων αποθηκεύονται και αυτά στο συμπυκνωμένο σύνολο. Το τελικό συμπυκνωμένο σύνολο φαίνεται στο (Σχήμα 2.10f) και αποτελείται από μόνο τρία αντικείμενα.

Μπορούμε να βρούμε το αντιπροσωπευτικό αντικείμενο m κάθε συστάδας ή κλάσης C υπολογίζοντας τον μέσο όρο των τιμών των n χαρακτηριστικών κάθε διαθέσιμου αντικειμένου $x_i, i = 1, 2, ..., |C|$ που ανήκει στο C. Συγκεκριμένα, τα n χαρακτηριστικά $m.d_j$ του m υπολογίζονται ως εξής:

$$m.d_j = \frac{1}{|C|} \sum_{x_i \epsilon C} x_i.d_j, j = 1, 2, ..., n \tag{2.5}$$

Λόγω της αναζήτησης του RHC για ομοιογένεια, τα περισσότερα αντικείμενα που περιέχει το συμπυκνωμένο σύνολο προέρχονται από τις περιοχές κοντά στα σύνορα κλάσεων, όπως και στην περίπτωση του RSP3. Κάτι το οποίο είναι λογικό, καθώς σε εκείνα τα σημεία υπάρχουν και οι περισσότερες πιθανότητες να βρεθούν αντικείμενα από τις κοντινές γειτονικές κλάσεις.

Μπορούμε να δούμε τον ψευδοκώδικα για τον RHC στον Αλγόριθμο 5. Κρατάει όλα τα μη-επεξεργασμένα αντικείμενα μέσα σε μία Ουρά ($Queue$). Στην αρχή της εκτέλεσης

Αλγόριθμος 5 RHC

Input: Training Set (TS), **Output:** Condensing Set (CS)

```
{Stage 1: Queue Initialization}
Queue ← ∅
Enqueue(Queue, TS)
{Stage 2: Construction of condensing set}
CS ← ∅
repeat
   C ← Dequeue(Queue)
   if C is homogeneous then
      r ← mean of C
      CS ← CS ∪ {r}
   else
      M ← ∅ {M is the set of class-means}
      for each class L in C do
         m_L ← mean of L
         M ← M ∪ {m_L}
      end for
      NewClusters ← K-MEANS(C, M)
      for each cluster C ∈ NewClusters do
         Enqueue(Queue, C)
      end for
   end if
until IsEmpty(Queue)
return CS
```

του RHC, το $Queue$ περιέχει όλο το σύνολο εκπαίδευσης (TS), καθώς το δεύτερο το αντιμετωπίζουμε σαν μία μεγάλη μη-επεξεργασμένη συστάδα. Σε κάθε επανάληψη, ο RHC αφαιρεί μία συστάδα C που βρίσκεται στην αρχή του $Queue$ και ελέγχει αν είναι ομοιογενής ή όχι. Αν είναι, τότε εισάγετε στο συμπυκνωμένο σύνολο (CS) και τα αντικείμενα του αφαιρούνται. Αλλιώς αν δεν είναι, τότε ο RHC υπολογίζει μία λίστα από μέσα αντικείμενα (M), ένα για κάθε κλάση που υπάρχει μέσα στο C. Μετά, ο RHC καλεί τον k-Means χρησιμοποιώντας την μη-ομοιογενή συστάδα C ως παράμετρο και την λίστα μέσων αντικειμένων M για να χρησιμοποιηθούν ως αρχικά κέντρα στον k-Means. Αυτή η διαδικασία παράγει ένα καινούριο σύνολο από μη-επεξεργασμένες συστάδες ($NewClusters$) οι οποίες εισάγονται στο $Queue$. Η διαδικασία εκτελείται επαναληπτικά μέχρι να αδειάσει το $Queue$, κάτι το οποίο σημαίνει πως δεν υπάρχουν άλλες μη-ομοιογενείς συστάδες.

3 Μείωση Δεδομένων σε Μη-Ισορροπημένα Σύνολα

3.1 Εισαγωγή

Ως μη-ισορροπημένα σύνολα δεδομένων (*imbalanced datasets*) ορίζονται σύνολα για τα οποία αν συγκρίνουμε τον αριθμό αντικειμένων κάθε κλάσης τους, θα παρατηρήσουμε ότι τουλάχιστον μία κλάση έχει πολύ λιγότερα ή πολύ περισσότερα αντικείμενα από τις υπόλοιπες. Σε αυτό το κεφάλαιο, προτείνουμε μία συλλογή μεθόδων για την διαχείριση τέτοιου είδους δεδομένων σε συνδυασμό με τεχνικές μείωσης των δεδομένων. Το πρόβλημα των μη-ισορροπημένων δεδομένων έχει παρατηρηθεί σε σύνολα δεδομένων με δύο κλάσεις (*binary*), καθώς και σε σύνολα με πολλαπλό αριθμό κλάσεων (*multi-class*). Στα πλαίσια αυτής της εργασίας έχουμε πειραματιστεί με σύνολα δεδομένων που ανήκουν και στις δύο κατηγορίες.

Η χρήση τεχνικών μείωσης δεδομένων (ΤΜΔ) πάνω σε μη-ισορροπημένα σύνολα δεδομένων επιφέρει ως συνέπεια την περαιτέρω μείωση, καθώς και την πιθανή εξάλειψη όλων των αντικειμένων των σπάνιων κλάσεων. Για αυτόν τον λόγο ακολουθήθηκε μία συγκεκριμένη προσέγγιση ώστε να μην παρατηρηθεί αυτό το φαινόμενο.

Στην Ενότητα 3.2 περιγράφεται η εν λόγω προσέγγιση, η οποία πρόκειται για τον διαχωρισμό κάθε συνόλου εκπαίδευσης σε δύο υποσύνολα, ένα για τις σπάνιες (*rare*) κλάσεις και ένα για τις μη-σπάνιες, ή αλλιώς κοινές (*common*) κλάσεις όπως της αποκαλούμε στα πλαίσια της εργασίας. Χρησιμοποιώντας αυτήν την τεχνική ως βάση, αναπτύχθηκαν δύο μέθοδοι δημιουργίας ενός νέου τροποποιημένου συμπυκνωμένου συνόλου το οποίο διατηρεί τα αντικείμενα των σπάνιων κλάσεων.

Στις Ενότητες 3.3 και 3.4 αναλύονται αυτές οι δύο μέθοδοι, καθώς και οι διαδικασίες που ακολουθούν για να πετύχουν τον σκοπό τους. Παράλληλα, δίνονται και παραδείγματα με σχήματα, τα οποία βοηθάνε στην περαιτέρω εξήγηση των μεθόδων.

Τέλος, εκτός από τις μεθόδους που αναφέρθηκαν ήδη, πραγματοποιήθηκαν επίσης πειράματα μέσω της χρήσης της μεθόδου SMOTE [12]. Τα αποτελέσματα των πειραμάτων που δόθηκαν από τα ειδικά τροποποιημένα σύνολα εκπαίδευσης που παράγει η SMOTE περιλαμβάνονται επίσης με τα υπόλοιπα.

Στην Ενότητα 3.5 περιγράφεται η μέθοδος η οποία χρησιμοποιεί την SMOTE για να εξισορροπήσει την κατανομή κλάσεων των συνόλων δεδομένων. Εξηγείται η διαδικασία που ακολουθεί, μαζί με ένα διάγραμμα το οποίο παρουσιάζει ένα απλό παράδειγμα της.

Η οικογένεια αυτών των τριών μεθόδων, δηλαδή οι δύο που χρησιμοποιούν τον διαχωρισμό των δεδομένων και η μία που χρησιμοποιεί την SMOTE, ονομάστηκε Μέθοδοι Διατήρησης Σπάνιων Κλάσεων (*Rare Class Preservation Methods*).

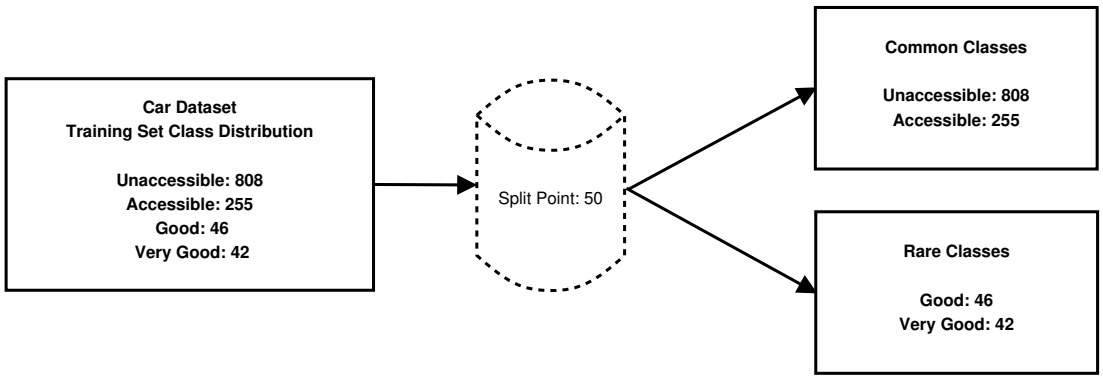
3.2 Διαχωρισμός των Συνόλων Δεδομένων

Όπως αναφέραμε και παραπάνω, σε περιπτώσεις μεγάλης ανισοκατανομής δεδομένων σε ένα σύνολο υπάρχει ο πιθανός κίνδυνος να μειωθούν ακόμη περισσότερο τα αντικείμενα μίας σπάνιας κλάσης κατά την διαδικασία μείωσης δεδομένων (*data reduction*). Ωστόσο, όπως έχει ήδη αναφερθεί, η σωστή πρόβλεψη των αντικειμένων που ανήκουν σε αυτές τις κλάσεις, είναι πιθανότατα σημαντικότερη (π.χ. σωστή πρόβλεψη χαλαζόπτωσης) από ότι η σωστή πρόβλεψη των αντικειμένων που ανήκουν σε κοινές κλάσεις. Ως "σπάνιες" ορίζουμε τις κλάσεις που έχουν εξαιρετικά χαμηλό αριθμό αντικειμένων. Ένα εναλλακτικό όνομα τους είναι "κλάσεις μειοψηφίας".

Στην προσπάθεια αντιμετώπισης αυτού του προβλήματος ακολουθήθηκε μία προσέγγιση. Η προτεινόμενη αυτή προσέγγιση πρόκειται για τον διαχωρισμό μίας κλάσης σε δύο υποσύνολα, ένα από τα οποία περιέχει μόνο τα αντικείμενα των σπάνιων κλάσεων και ένα για τα υπόλοιπα. Για αυτόν τον λόγο αναπτύχθηκε μία μέθοδος η οποία αναλύει την κατανομή κάθε κλάσης στο σύνολο εκπαίδευσης ενός συνόλου δεδομένων και έχει την δυνατότητα να διαφοροποιήσει τις σπάνιες κλάσεις από τις υπόλοιπες, δηλαδή τις κοινές κλάσεις.

Για να λειτουργήσει σωστά ο αλγόριθμος, θα πρέπει να ξέρουμε την κατανομή των κλάσεων του συνόλου εκπαίδευσης (*training set*) που θέλουμε να χωρίσουμε. Έπειτα, θα πρέπει να διαλέξουμε έναν αριθμό ο οποίος θα καθορίζει σε ποια από τις δύο κατηγορίες ανήκει η κάθε κλάση.

Στο Σχήμα 3.1 φαίνεται ένα παράδειγμα της διαδικασίας. Έστω ότι χρησιμοποιούμε το σύνολο δεδομένων Car, το οποίο έχει τέσσερις κλάσεις, δύο από τις οποίες θεωρούνται σπάνιες. Στο σύνολο εκπαίδευσης και οι δύο σπάνιες κλάσεις έχουν λιγότερα από πενήντα αντικείμενα. Επομένως ορίζουμε των αριθμό πενήντα ως το "σημείο διαχωρισμού" (*split point*). Στην συνέχεια θα αναλυθεί το σύνολο εκπαίδευσης σειριακά. Αν ένα αντικείμενο ανήκει σε μία κλάση με συνολικό αριθμό δεδομένων μεγαλύτερο από το σημείο διαχωρισμού, τότε αντιγράφεται σε ένα καινούριο υποσύνολο για όλες τις κοινές κλάσεις. Στην αντίθετη



Σχήμα 3.1: Χωρισμός των κοινών και των σπάνιων κλάσεων.

περίπτωση αντιγράφεται σε ένα διαφορετικό αρχείο το οποίο περιέχει μόνο τα αντικείμενα των σπάνιων κλάσεων. Χρησιμοποιώντας αυτά υποσύνολα, αναπτύχθηκαν οι δύο μέθοδοι που περιγράφονται στις επόμενες ενότητες.

3.3 Rare Class Preservation Method 1

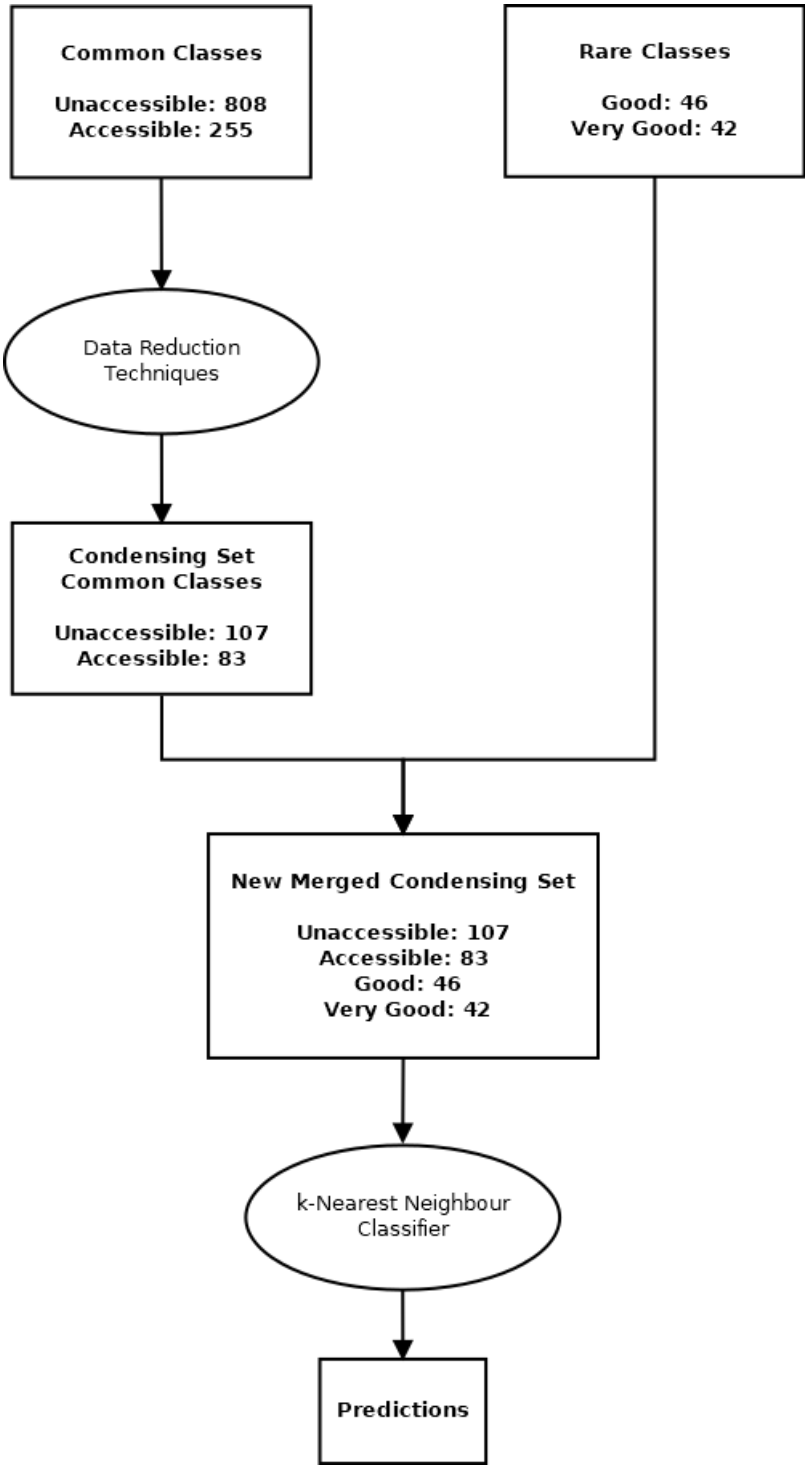
Η πρώτη στρατηγική αντιμετώπισης του προβλήματος που παρουσιάζουν τα μη-ισορροπημένα σύνολα δεδομένων είναι απλή. Πρόκειται για μία μέθοδο διατήρησης, την οποίο ονομάσαμε Rare Class Preservation Method 1 (RCPM1). Όπως αναφέρθηκε και πριν, οι ανησυχίες μας προέρχονται από την πιθανότητα να διαγραφούν δεδομένα των σπάνιων κλάσεων όταν μειωθεί ο αριθμός δεδομένων.

Επομένως με την μέθοδο RCPM1, αφού ξεχωρίσουμε αυτές τις σπάνιες κλάσεις από τις κοινές, εκτελούμε μία τεχνική μείωσης δεδομένων (ΤΜΔ) μόνο στο δεύτερο από αυτά τα δύο υποσύνολα. Δηλαδή, στο υποσύνολο που περιέχει τα αντικείμενα που ανήκουν στις κοινές κλάσεις. Έπειτα, ενώνουμε το νέο συμπυκνωμένο σύνολο, το οποίο περιέχει μόνο αντικείμενα των κοινών κλάσεων, με το αρχείο που περιέχει όλα τα αντικείμενα των σπάνιων κλάσεων. Έτσι δημιουργούμε ένα νέο επεξεργασμένο συμπυκνωμένο σύνολο που περιέχει μία πιο ισορροπημένη κατανομή μεταξύ όλων των διαθέσιμων κλάσεων. Τέλος, πάνω σε αυτό συμπυκνωμένο σύνολο εκτελούμε μία υλοποίηση του αλγορίθμου κατηγοριοποίησης k-Nearest Neighbors (k-NN) μέσω της οποίας υπολογίζουμε την ακρίβεια για όλες τις κλάσεις μαζί, καθώς και την ορθότητα, την ευαισθησία και το F-Measure για τις σπάνιες κλάσεις.

Η διαδικασία που ακολουθεί η μέθοδος RCPM1 παρουσιάζεται διαγραμματικά στο Σχήμα 3.2. Στο συγκεκριμένο παράδειγμα χρησιμοποιούμε το σύνολο δεδομένων Car, το οποίο αναλύουμε παρακάτω στο Κεφάλαιο 4. Έχουμε ήδη χωρίσει το σύνολο εκπαίδευσης του Car σε δύο υποσύνολα για τις κοινές και τις σπάνιες κλάσεις του. Όπως βλέπουμε, οι σπάνιες κλάσεις περιέχουν έναν σημαντικά μεγαλύτερο αριθμό αντικειμένων από τις σπάνιες. Η κλάση Accessible είναι πέντε φορές μεγαλύτερη από τις σπάνιες κλάσεις, ενώ η κλάση Unaccessible είναι εικοσαπλάσια. Έτσι, για να εξισορροπήσουμε την κατανομή των κλάσεων παίρνουμε το αρχείο των κοινών κλάσεων και εκτελούμε μείωση δεδομένων πάνω του. Από αυτό παράγεται ένα συμπυκνωμένο σύνολο. Στην συνέχεια, ενώνουμε αυτό το αποτέλεσμα με το αρχείο των σπάνιων κλάσεων, το οποίο είναι άθικτο. Έχουμε επομένως δημιουργήσει ένα καινούριο συμπυκνωμένο σύνολο, το οποίο περιέχει έναν ικανοποιητικό αριθμό αντικειμένων για όλες τις κλάσεις. Ολοκληρώνουμε την διαδικασία τρέχοντας τον κατηγοριοποιητή k-NN πάνω του.

Προσοχή, πρέπει να σημειωθεί πως αν το σύνολο δεδομένων που χρησιμοποιούμε είναι δυαδικό, δεν μπορούμε να χρησιμοποιήσουμε την μέθοδο RCPM1 πάνω του. Αφού θα είχαμε μόνο δύο κλάσεις, θα είμαστε υποχρεωμένοι να χωρίσουμε το σύνολο εκπαίδευσης σε δύο αρ-

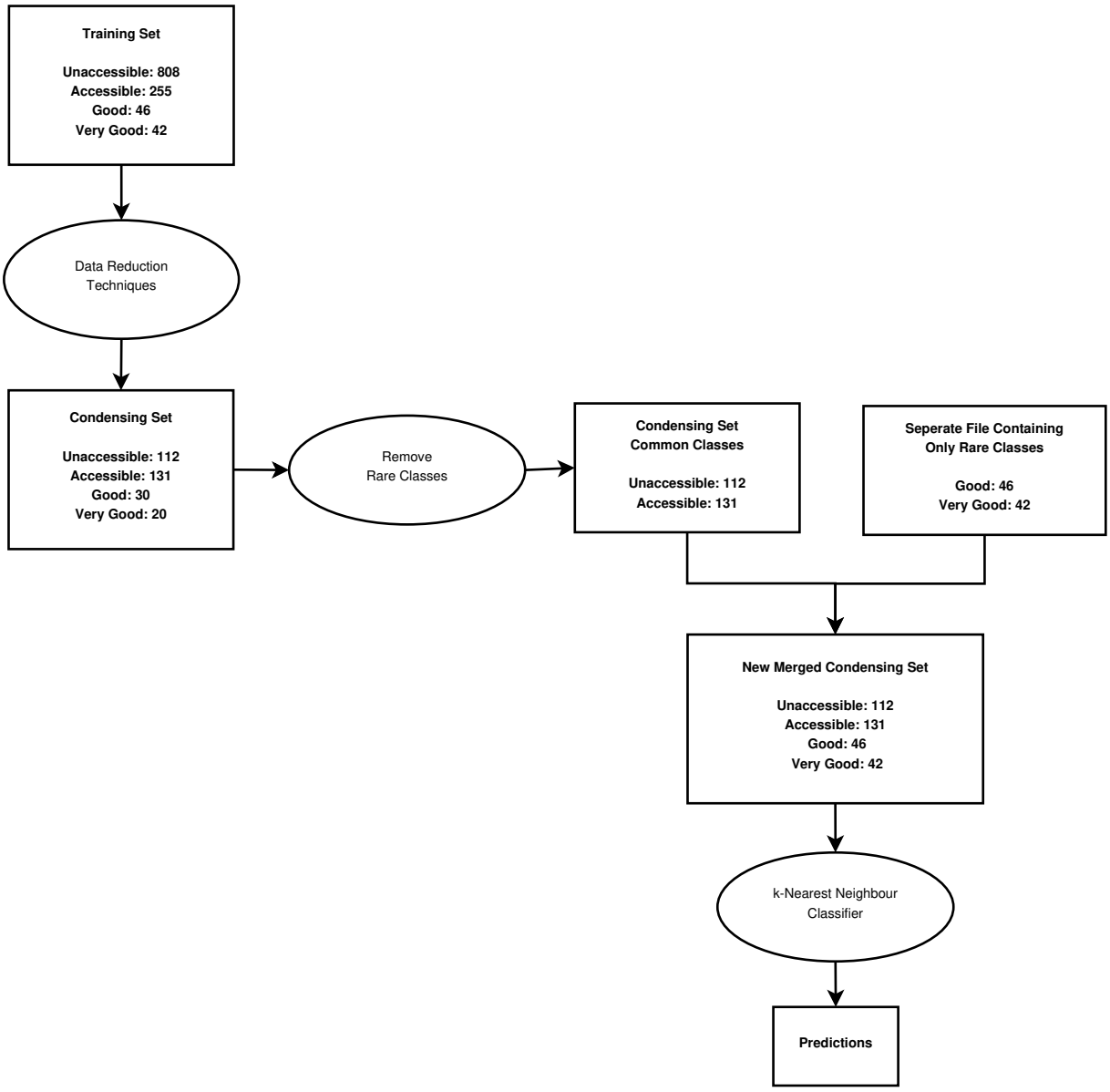
χεία με μία κλάση το καθένα. Αναγκαστικά, θα πραγματοποιήσουμε μείωση δεδομένων πάνω σε ένα υποσύνολο με μόνο μία κλάση. Κάτι τέτοιο θα παράγει ένα αλλοιωμένο αποτέλεσμα που περιέχει μόνο ένα αντικείμενο που θα ανήκει στην κοινή κλάση. Κάτι τέτοιο, δεν θα μπορούσε να χρησιμοποιηθεί στην κατηγοριοποίηση και να παράγουμε σωστά αποτελέσματα.



Σχήμα 3.2: Μέθοδος RCPM1 - Δημιουργία του νέου συμπυκνωμένου συνόλου και εκτέλεση κατηγοριοποίησης πάνω του.

3.4 Rare Class Preservation Method 2

Η δεύτερη στρατηγική είναι αρκετά παρόμοια με την πρώτη. Η κύρια διαφορά είναι ότι αφού χωρίσουμε το σύνολο εκπαίδευσης, δεν χρησιμοποιούμε την ΤΜΔ στο υποσύνολο των κοινών κλάσεων. Πιο συγκεκριμένα, η ΤΜΔ εκτελείται στο αρχικό σύνολο εκπαίδευσης που περιλαμβάνει όλες τις κλάσεις. Στην συνέχεια, το συμπυκνωμένο σύνολο που παράγεται ενώνεται με το υποσύνολο που περιλαμβάνει τα αντικείμενα που ανήκουν σε σπάνιες κλάσεις. Η παρούσα στρατηγική ονομάστηκε Rare Class Preservation Method 2 (RCPM2), και όπως και η προηγούμενη πρόκειται για μία μέθοδο διατήρησης των αντικειμένων που ανήκουν σε σπάνιες κλάσεις.



Σχήμα 3.3: Μέθοδος RCPM2 - Δημιουργία του νέου συμπυκνωμένου συνόλου και εκτέλεση κατηγοριοποίησης πάνω του.

Μπορούμε να δούμε ένα παράδειγμα της χρήσης του RCPM2 στο Σχήμα 3.3, στο οποίο χρησιμοποιούμε το σύνολο δεδομένων Car. Η διαδικασία έχει ως εξής: έχουμε δημιουργήσει ήδη το αρχείο των σπάνιων κλάσεων, αλλά το αφήνουμε στην άκρη προς το παρόν. Στην συνέχεια, παίρνουμε το αρχικό σύνολο εκπαίδευσης, το οποίο περιέχει τα αντικείμενα όλων των κλάσεων και χρησιμοποιούμε μία ΤΜΔ πάνω του όπως συνήθως. Έπειτα, από το συμπυκνωμένο σύνολο που δημιουργείται αφαιρούμε τα αντικείμενα των σπάνιων κλάσεων που έχουν μείνει. Θα αντικατασταθούν από όλα τα αντικείμενα των σπάνιων κλάσεων από το σύνολο εκπαίδευσης, όπως ήταν δηλαδή και στην αρχή. Έτσι δημιουργούμε ένα νέο συμπυκνωμένο σύνολο, το οποίο χρησιμοποιούμε όταν εκτελούμε τον αλγόριθμο k-Nearest Neighbors (k-NN) μέσω του οποίου υπολογίζουμε την ακρίβεια για όλες τις κλάσεις, καθώς και την ορθότητα, την ευαισθησία και το F-Measure για τις σπάνιες κλάσεις.

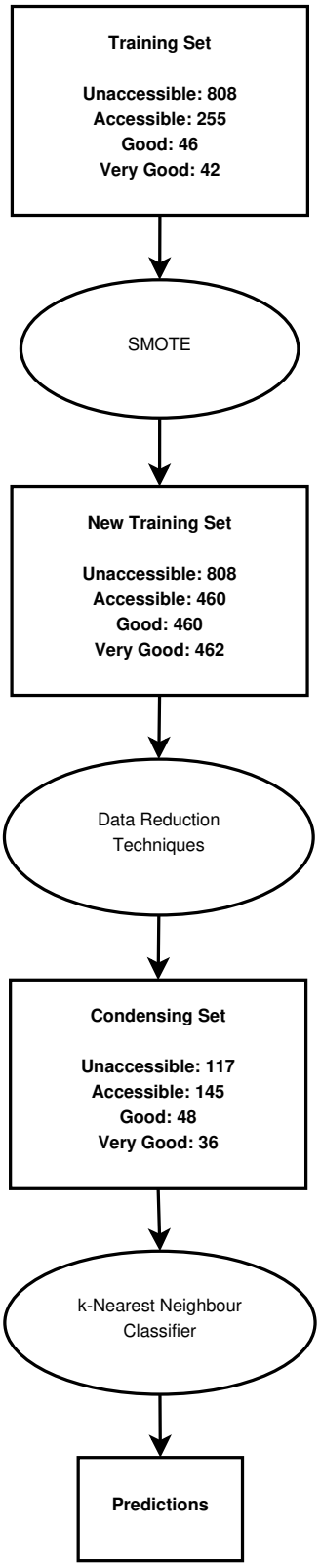
Ο λόγος που μειώνουμε τα δεδομένα του αρχικού συνόλου εκπαίδευσης, είναι επειδή κάνοντας μείωση δεδομένων σε ένα υποσύνολο που περιέχει αντικείμενα αποκλειστικά των κοινών κλάσεων, μπορεί στην τελική να χάσουμε περισσότερα δεδομένα από αυτές τις κλάσεις. Έχει παρατηρηθεί σε πολλές περιστάσεις ότι με αυτόν τον τρόπο καταλήγουμε με περισσότερα False Positive δεδομένα και έτσι υπολογίζεται χαμηλότερο ποσοστό ορθότητας. Θεωρήθηκε πως ο λόγος που συμβαίνει αυτό, είναι επειδή δεν υπάρχουν αρκετά σύνορα στο σύνολο εκπαίδευσης των κοινών κλάσεων, καθώς αφαιρέθηκαν τα μετρημένα αντικείμενα των σπάνιων κλάσεων ανάμεσα τους. Επομένως, χρησιμοποιώντας την μέθοδο RCPM2 ελπίζουμε να αυξήσουμε τα ποσοστά ορθότητας σε σχέση με την RCPM1.

Τέλος σε αντίθεση με την RCPM1, η μέθοδος RCPM2 μπορεί να χρησιμοποιηθεί και σε δυαδικά σύνολα, καθώς και σε σύνολα πολλαπλών κλάσεων. Αφού εκτελούμε μείωση δεδομένων πάνω στο αρχικό σύνολο εκπαίδευσης, δεν υπάρχει πια ο κίνδυνος να παράγουμε ένα αλλοιωμένο σύνολο μέσω αυτής της δράσης.

3.5 Rare Class Preservation Method - SMOTE

Στα πειράματα μας, εκτός από τις προαναφερόμενες δύο μεθόδους που εκμεταλλεύονται τον διαχωρισμού των συνόλων δεδομένων και διατηρούν τα αντικείμενα που ανήκουν στις σπάνιες κλάσεις, δοκιμάστηκε και μία άλλη μέθοδος η οποία χρησιμοποιεί την διαθέσιμη και πολύ διαδομένη τεχνική SMOTE [12]. Η συγκεκριμένη μέθοδος, πριν κάνει οτιδήποτε άλλο χρησιμοποιεί την SMOTE πάνω στο σύνολο εκπαίδευσης. Έτσι, αυξάνεται ο αριθμός των αντικειμένων των κλάσεων που βλέπουμε πως έχουν μικρό αριθμό αντικειμένων μέχρι να πετύχουν όλες τους ένα επίπεδο ισορροπίας μεταξύ τους. Έπειτα, μπορούμε να προχωρήσουμε στην χρήση μίας ΤΜΔ χωρίς να ανησυχούμε πως μπορεί να δημιουργήσουμε ένα συμπυκνωμένο σύνολο που στερείται σημαντικών δεδομένων. Τέλος εκτελούμε κατηγοριοποίηση πάνω σε αυτό το συμπυκνωμένο σύνολο. Η μέθοδος που μόλις περιγράψαμε έχει ονομαστεί Rare Class Preservation Method - SMOTE (RCPM-SMOTE).

Ένα διάγραμμα της εκτέλεσης της μεθόδου φαίνεται στο Σχήμα 3.4, χρησιμοποιώντας το σύνολο δεδομένων Car. Όπως βλέπουμε το σύνολο εκπαίδευσης έχει δύο κλάσεις με



Σχήμα 3.4: Μέθοδος RCPM-SMOTE - Επεξεργασία του συνόλου εκπαίδευσης μέσου της SMOTE πριν την χρήση των τεχνικών μείωσης δεδομένων και της κατηγοριοποίησης.

808 και 255 αντικείμενα, καθώς και δύο ακόμα με μόνο 46 και 42 αντικείμενα. Οι δύο τελευταίες είναι και οι σπάνιες κλάσεις σε αυτήν την περίπτωση. Χρησιμοποιώντας την SMOTE πολλαπλασιάσουμε τον αριθμό των αντικειμένων, παράγοντας ένα νέο σύνολο εκπαίδευσης όπου οι τέσσερις κλάσεις έχουνε 808, 460, 460 και 462 αντικείμενα αντίστοιχα. Έπειτα, η διαδικασία συνεχίζεται όπως συνήθως, με την μόνη διαφορά να είναι το γεγονός πως χρησιμοποιούμε αυτό το νέο σύνολο εκπαίδευσης που μόλις δημιουργήσαμε. Χρησιμοποιούμε μία ΤΜΔ, η οποία παράγει το συμπυκνωμένο σύνολο μας, πάνω στο οποίο εκτελούμε και την κατηγοριοποίηση k-NN.

Από το αποτελέσματα που δόθηκαν από τον αλγόριθμο k-NN υπολογίστηκε η ακρίβεια για όλες τις κλάσεις, καθώς και η ορθότητα, η ευαισθησία και το F-Measure για τις σπάνιες κλάσεις. Είναι σημαντικό να αναφερθεί πως η SMOTE χρησιμοποιείται μόνο στο σύνολο εκπαίδευσης και όχι σε όλο το σύνολο δεδομένων, καθώς δεν υπάρχει σε καμία περίπτωση ο κίνδυνος να εξαλειφθούν τα δεδομένα των σπάνιων κλάσεων από το σύνολο δοκιμών. Επίσης, το ποσοστό αύξησης των δεδομένων ή αλλιώς το ποσοστό υπερδειγματοληψίας που χρησιμοποιήθηκε κάθε φορά διαφέρει ανάλογα με το κάθε σύνολο δεδομένων.

Όπως και η μέθοδος RCPM2, η RCPM-SMOTE είναι κατάλληλη και για δυαδικά προβλήματα, καθώς και για προβλήματα πολλαπλών κλάσεων. Αφού εκτελούμε μείωση δεδομένων πάνω στο αρχικό σύνολο εκπαίδευσης, δεν υπάρχει πια ο κίνδυνος που αναφέραμε για την RCPM1, όπου υπάρχει η πιθανότητα να παράγουμε ένα αλλοιωμένο σύνολο κάνοντας μείωση πάνω σε ένα υποσύνολο που περιέχει μόνο μία κλάση. Να σημειώσουμε πως σε αυτό το παράδειγμα χρησιμοποιήσαμε την SMOTE και σε μία κλάση που δεν ήταν αρκετά μικρή για να θεωρείται σπάνια, αυτό όμως δεν χρειάζεται να γίνεται πάντα. Σε αυτήν την περίσταση θεωρήθηκε πως αυτό θα ήταν το βέλτιστο μέγεθος για το σύνολο εκπαίδευσης, αλλά σε άλλα θα μπορούσαμε να επεξεργαστούμε μόνο τις σπάνιες κλάσεις.

3.6 Επίλογος

Σε αυτό το κεφάλαιο προτείναμε τον διαχωρισμό των αντικειμένων που ανήκουν σε κοινές κλάσεις από αυτά που ανήκουν σε σπάνιες ως μία πιθανή λύση στο πρόβλημα που παρουσιάζουν τα μη-ισορροπημένα σύνολα δεδομένων σε συνδυασμό με τις ΤΜΔ, καθώς και μία συλλογή μεθόδων, τρεις συνολικά, οι οποίες επιχειρούν να διατηρήσουν τα αντικείμενα των σπάνιων κλάσεων μετά την ολοκλήρωση της διαδικασίας της μείωσης δεδομένων. Δύο από αυτές τις μεθόδους, οι RCPM1 και RCPM2, χρησιμοποιούν τον διαχωρισμό συνόλων έτσι ώστε να μην χρειαστεί να συρρικνωθούν παραπάνω οι σπάνιες κλάσεις. Επιπλέον η τρίτη μέθοδος, η RCPM-SMOTE, χρησιμοποιεί την SMOTE για να πολλαπλασιάσει τον αριθμό των αντικειμένων στις σπάνιες κλάσεις έτσι ώστε να μην κινδυνεύουν πλέον με περαιτέρω μείωση πριν προχωρήσει στις διαδικασίες της μείωσης δεδομένων και της κατηγοριοποίησης.

4 Πειραματική Ανάλυση

4.1 Εισαγωγή

Σε αυτό το κεφάλαιο παρουσιάζουμε όλες τις πληροφορίες οι οποίες είναι σχετικές με τα πειράματα που πραγματοποιήθηκαν στα πλαίσια της παρούσας πτυχιακής εργασίας. Ο σκοπός αυτών των πειραμάτων είναι να συγκρίνουμε την απόδοση των τριών μεθόδων RCPM σε σχέση με την κλασική προσέγγιση της κατηγοριοποίησης μέσω μείωσης δεδομένων χωρίς την χρήση κάποιας μεθόδου για την διατήρηση των σπάνιων κλάσεων.

Στην Ενότητα 4.2 δίνεται μία αναλυτική περιγραφή των συνόλων δεδομένων που χρησιμοποιήσαμε στα πειράματα μας. Εκτός αυτού, αναφέρονται επίσης τα κριτήρια που έπρεπε να εκπληρώνουν αυτά τα σύνολα για να μπορέσουμε να τα χρησιμοποιήσουμε.

Στην Ενότητα 4.3 παρουσιάζονται τα αποτελέσματα των πειραμάτων μας σε μία σειρά αναλυτικών πινάκων. Αναφέρουμε τις κατηγορίες στις οποίες χωρίστηκαν τα πειράματα, καθώς και τα κριτήρια τα οποία υπολογίζουμε. Επιπλέον, επισημαίνουμε παρατηρήσεις για τα πιο αξιοσημείωτες πειραματικές μετρήσεις των αποτελεσμάτων.

Στην Ενότητα 4.4 συγκρίνουμε τα αποτελέσματα των τεσσάρων μεθόδων, δηλαδή τις μεθόδους RCPM και την απλή κατηγοριοποίηση μέσω μείωσης δεδομένων (Original) μεταξύ τους, σημειώνοντας παράλληλα τα υπέρ και τα κατά που συμπεράναμε.

4.2 Σύνολα Δεδομένων

Τα σύνολα δεδομένων που χρειάζονταν για τα πειράματα έπρεπε να πληρούν ορισμένα κριτήρια. Το πρώτο και πιο σημαντικό από αυτά, είναι σαφώς το ότι θα πρέπει να έχουν άνιση κατανομή των δεδομένων στις κλάσεις. Δηλαδή να παρατηρείται το πρόβλημα των μη-ισορροπημένων συνόλων δεδομένων πάνω τους. Εκτός αυτού, θα πρέπει να είναι εξειδικευμένα σύνολα δεδομένων τα οποία προορίζονται για θέματα κατηγοριοποίησης.

Τα σύνολα δεδομένων μπορούν να είναι είτε δυαδικά (*binary*), να έχουν δηλαδή μόνο δύο κλάσεις όπου η μία είναι σημαντικά μεγαλύτερα από την άλλη, ή εναλλακτικά να έχουν περισσότερες κλάσεις (*multi-class*) και μία τουλάχιστον από αυτές να διαθέτει έναν υπερβολικά μικρό αριθμό αντικειμένων σε σύγκριση με τις υπόλοιπες.

Για να εκτελέσουμε τις ΤΜΔ καθώς και τον κατηγοριοποιητή εγγύτερων γειτόνων χρησιμοποιώντας την Ευκλείδεια απόσταση, θα πρέπει όλα τα χαρακτηριστικά να είναι αριθμητικά. Αυτό, βοηθάει και στο να αξιολογήσουμε τα αποτελέσματα μεταξύ των διαφορετικών συνόλων δεδομένων.

Προσοχή, είναι απαραίτητο να έχουν όλα τα χαρακτηριστικά το ίδιο εύρος τιμών, καθώς θεωρούμε πως είναι όλα τους το ίδιο σημαντικά. Ας υποθέσουμε, για παράδειγμα, πως έχουμε δύο χαρακτηριστικά για ένα αντικείμενο που αντιπροσωπεύει έναν εργαζόμενο. Τα δύο χαρακτηριστικά καταγράφουν τον αριθμό των παιδιών του και τον μισθό του. Είναι επομένως αδύνατον να έχουν το ίδιο εύρος, καθώς η κοινή λογική θέτει πως ο εργαζόμενος μπορεί να έχει έναν μονοψήφιο/διψήφιο αριθμό παιδιών, ενώ ο αριθμός του μισθού του είναι τριψήφιος ή τετραψήφιος.

Για να αντιμετωπιστεί αυτό το θέμα, πραγματοποιήθηκε μία διαδικασία κανονικοποίησης (*normalization*) έτσι ώστε να προσαρμοστούν οι τιμές ανάλογα. Μέσω της κανονικοποίησης, οι τιμές μεταβλήθηκαν έτσι ώστε όλα τα χαρακτηριστικά να αποτελούν πραγματικούς αριθμούς που κυμαίνονται από το μηδέν μέχρι το ένα $(0, 1)$. Έτσι δίνουμε σε όλα τα χαρακτηριστικά το ίδιο βάρος χωρίς να αλλοιωθεί η σημασία των τιμών τους.

Όσα σύνολα δεδομένων έχουν ονομαστικά χαρακτηριστικά δυστυχώς δεν μπορούν να χρησιμοποιηθούν σε ζητήματα που αφορούν κάποια ΤΜΔ. Ο λόγος είναι επειδή οι προαναφερόμενες τεχνικές χρειάζονται κάποιο κριτήριο για να υπολογίζουν πόσο απέχουν τα αντικείμενα μεταξύ τους και στην περίπτωση μας αυτό το κριτήριο είναι η Ευκλείδεια απόσταση (*Euclidean distance*). Βέβαια, θα μπορούσαμε να πραγματοποιήσουμε κάποια μετατροπή πάνω σε ονομαστικά δεδομένα για να τους δώσουμε αριθμητική μορφή, όμως αυτή θα ήταν μία αναξιόπιστη λύση. Αν για παράδειγμα αλλάζαμε μερικά ονομαστικά χαρακτηριστικά, έστω (A, B, C) σε αριθμούς έστω $(0, 1, 2)$, τότε ναι θα μπορούσαν να χρησιμοποιηθούν στους υπολογισμούς των ΤΜΔ. Όμως, δεν μπορούμε να είμαστε βέβαιοι ότι το C είναι πιο μακριά από το A από ότι το B ή ότι η απόσταση των (A, C) είναι όντως διπλάσια από αυτή των (A, B). Για τον λόγο αυτόν, αποφεύγουμε την χρήση ονομαστικών δεδομένων. Φυσικά, μία εναλλακτική λύση θα ήταν να χρησιμοποιήσουμε κάποιο μέτρο απόστασης που είναι κατάλληλο για ονομαστική δεδομένα. Ωστόσο, αυτό δεν αποτελεί στόχο της παρούσας εργασίας και έτσι επικεντρωθήκαμε σε σύνολα δεδομένων που περιέχουν αποκλειστικά αριθμητικές τιμές.

Τέλος, θεωρήθηκε ότι για να βρεθούν αξιόλογα αποτελέσματα θα ήταν προτιμότερο να πειραματιστούμε πάνω σε σύνολα δεδομένων που περιέχουν τουλάχιστον χίλια αντικείμενα. Τα σύνολα δεδομένων που τελικά χρησιμοποιήθηκαν φαίνονται στον Πίνακα 4.1. Με την εξαίρεση του μεγέθους σε δύο περιπτώσεις, πληρούν και τα δώδεκα τους όλα τα κριτήρια που αναφέραμε. Η στήλη "Σημείο Διαχωρισμού", ή αλλιώς Split Point, αφορά μόνο τις μεθόδους RCPM1 και RCPM2. Οι αριθμοί της συμβολίζουν το μέγεθος για κάθε σύνολο δεδομένων που καθορίζει αν μία κλάση είναι σπάνια ή όχι. Για παράδειγμα στο σύνολο δεδομένων Car, θεωρούμε πως όσες κλάσεις του συνόλου εκπαίδευσης έχουν λιγότερα από εξήντα αντικείμενα είναι σπάνιες. Το σημείο διαχωρισμού αποφασίστηκε μελετώντας την κατανομή των κλάσεων του συνόλου εκπαίδευσης κάθε συνόλου δεδομένων.

Πίνακας 4.1: Περιγραφή των Συνόλων Δεδομένων που χρησιμοποιήθηκαν.

Σύνολα Δεδομένων (Datasets)	**Αριθμός Αντικειμένων**	**Χαρακτηριστικά (Attributes)**	**Κλάσεις**	**Σπάνιες Κλάσεις**	**Σημείο Διαχωρισμού**
Avila	20867	10	12	3	200
Balance	625	4	3	1	100
Car	1.728	6	4	2	60
KDD-BigData	141.481	36	23	12	100
Page-Blocks	5.473	10	5	3	100
Shuttle	57.999	9	7	4	150
Yeast	1.484	8	10	6	100
Page-Blocks 0	5472	10	2	1	500
Segment 0	2308	18	2	1	300
Shuttle: c0 vs c4	1829	9	2	1	100
Vowel 0	988	13	2	1	100
Wine Quality Red 4	1599	11	2	1	100

4.2.1 Τα Χρησιμοποιημένα Σύνολα Δεδομένων Πολλαπλών Κλάσεων

Ακολουθεί μία σύντομη περιγραφή των συνόλων δεδομένων που χρησιμοποιήθηκαν στα πειράματα. Συνολικά χρησιμοποιήθηκαν δώδεκα σύνολα δεδομένων, από τα οποία τα επτά ανήκουν στην κατηγορία των πολλαπλών κλάσεων. Όλα τα σύνολα δεδομένων είναι διαθέσιμα στο KEEL [39] και το UCI Machine Learning Repository [48]. Αυτά με πολλαπλό αριθμό κλάσεων είναι:

- Το σύνολο δεδομένων **Balance**, είναι το μικρότερο σύνολο που χρησιμοποιήθηκε στην παρούσα εργασία. Επιχειρεί να μιμηθεί ένα ψυχολογικό πείραμα σχετικά με την ισορροπία. Για αυτόν το λόγο μοντελοποιεί όλα τα πιθανά αποτελέσματα που μπορεί να παράγει το αναφερόμενο πείραμα. Έχει τέσσερα χαρακτηριστικά (*attributes*) και τρεις κλάσεις.

  Τα τέσσερα χαρακτηριστικά του περιγράφουν το βάρος και την απόσταση της κάθε μεριάς: Left-weight, Left-distance, Right-weight, Right-distance. Οι τρεις κλάσεις του αντιπροσωπεύουν την τελική τοποθεσία του αντικειμένου (Left, Right, Balanced). Αν ο υπολογισμός του $(Left-weight * Left-distance)$ είναι μεγαλύτερος από τον υπολογισμό του $(Right-weight * Right-distance)$, τότε το αντικείμενο ανήκει στην κλάση Left, ενώ στην αντίθετη περίπτωση ανήκει στην κλάση Right. Αν οι δύο υπολογισμοί είναι ίσοι, τότε το αντικείμενο είναι στο κέντρο και ανήκει στην κλάση Balanced. Η κλάση Balanced είναι και η σπάνια κλάση αυτού το συνόλου, καθώς περιέχει μόνο το ένα δέκατο των αντικειμένων σε σύγκριση με τις άλλες δύο.

- Το σύνολο δεδομένων **Car**, γνωστό και ως Car Evaluation Database είναι ένα ευρέως

γνωστό και χρησιμοποιημένο σύνολο δεδομένων το οποίο περιγράφει την κατάσταση και την ποιότητα αυτοκινήτων.

Έχει έξι χαρακτηριστικά: buying, maint, doors, persons, lug boot, safety. Χωρίζει τα αυτοκίνητα σε τέσσερις διαφορετικές κλάσεις οι οποίες αντιπροσωπεύουν την προσιτότητα τους: unaccessible (unacc), accessible (acc), good, very good (vgood). Από αυτές τις τέσσερις κλάσεις, οι δύο (good, very good) έχουν έναν σημαντικά μικρότερο αριθμό αντικειμένων, επομένως θεωρούνται σπάνιες.

- Το σύνολο δεδομένων **Avila** περιγράφει ένα λατινικό αντίγραφο ολόκληρης της Βίβλου από τον δωδέκατο αιώνα. Οι πληροφορίες στο σύνολο δεδομένων εξήχθησαν από ψηφιακές εικόνες του βιβλίου και το σύνολο έχει επίσης χρησιμοποιηθεί για τους σκοπούς της έρευνας στην πηγή [49]. Έχει δέκα χαρακτηριστικά, καθώς και δώδεκα διαφορετικές κλάσεις.

  Υποθετικά, υπήρξαν δώδεκα διαφορετικοί αντιγραφείς οι οποίοι είχαν εργαστεί πάνω στην συγκεκριμένη έκδοση της Βίβλου. Οι δώδεκα κλάσεις επομένως, αντιπροσωπεύουν αυτά τα άτομα. Τα δέκα χαρακτηριστικά κάθε αντικειμένου καθορίζουν την μορφή του γραψίματος (writing pattern) και χρησιμοποιούνται για να προσδιορίσουμε ποιος από τους αντιγραφείς έχει συνεισφέρει στο συγκεκριμένο σημείο.

  Σκοπός του συνόλου δεδομένων είναι να εντοπίσει πόσο έχει συνεισφέρει ο κάθε αντιγραφέας στην σύνταξη αυτής της έκδοσης της Βίβλου. Σαφώς δεν έχουν συνεισφέρει όλοι τους στον ίδιο βαθμό, για την ακρίβεια μερικοί έχουν γράψει πολύ λιγότερο από άλλους. Οι αντιγραφείς που έχουν γράψει λιγότερο από όλους αντιπροσωπεύουν και τις σπάνιες κλάσεις για την συγκεκριμένη περίπτωση.

- Το **KDD** πρόκειται για ένα από τα πιο διαδεδομένα σύνολα στο κόσμο της εξόρυξης δεδομένων. Έχει πολλές εκδόσεις, αλλά η συγκεκριμένη που χρησιμοποιούμε σε αυτήν την εργασία είναι η KDD Cup 1999 [50], η οποία δημιουργήθηκε για τον Τρίτο Διεθνή Διαγωνισμό Ανίχνευσης Γνώσης και Εργαλείων Εξόρυξης Δεδομένων (*Third International Knowledge Discovery and Data Mining Tools Competition*). Σκοπός του συνόλου είναι να δημιουργήσει έναν ανιχνευτή εισβολής δικτύου, ο οποίος μπορεί να ξεχωρίσει τις καλές συνδέσεις σε ένα δίκτυο από τις κακές, δηλαδή τις πιθανές επιθέσεις από κακόβουλες ομάδες.

  Αξίζει να σημειωθεί ότι το δείγμα του KDD που χρησιμοποιήθηκε σε αυτήν την εργασία είναι ένα απλό υποσύνολο, από το οποίο έχουν αφαιρεθεί τα διπλότυπα, τα οποία αποτελούν την πλειονότητα των αρχικών δεδομένων. Το πλήρες σύνολο δεδομένων περιέχει έναν τεράστιο αριθμό αντικειμένων, ο οποίος φτάνει στα εκατομμύρια. Το συγκεκριμένο υποσύνολο έχει τριάντα-έξι χαρακτηριστικά και είκοσι-τρεις κλάσεις,

από τις οποίες οι δώδεκα είναι σπάνιες.

- Το **Page-Blocks** περιέχει δεδομένα για την διάταξη των σελίδων ενός έγγραφου. Κάθε σελίδα είναι χωρισμένη σε πολλαπλά μπλοκ και κάθε αντικείμενο μέσα στο σύνολο δεδομένων αντιπροσωπεύει ένα από αυτά τα μπλοκ. Έχει δέκα χαρακτηριστικά και πέντε διαφορετικές κλάσεις.

  Έχουμε επομένως μία κλάση για κάθε πιθανό είδος μπλοκ: Κείμενο (Text), Σχήμα (Graphic), Εικόνα (Picture), Οριζόντια Γραμμή (Horizontal Line) και Κάθετη Γραμμή (Vertical Line). Τα δέκα χαρακτηριστικά χρησιμοποιούνται για να προσδιοριστεί η κλάση κάθε αντικειμένου. Αντιπροσωπεύουν διάφορες πληροφορίες όπως το μάκρος ενός μπλοκ, ή σε ποιο μέρος της σελίδας βρίσκεται.

- Το **Shuttle** πρόκειται για ένα σύνολο δεδομένων χτισμένο για τον συγκεκριμένο σκοπό να βοηθάει τα διαστημικά σκάφη κατά την διάρκεια της διαδικασίας προσγείωσης. Πιο συγκεκριμένα, εξάγει κανόνες μέσω των οποίων μπορούμε να προσδιορίσουμε αν οι συνθήκες είναι κατάλληλες για να επιτρέψουμε να γίνει αυτόματη προσγείωση ή αν θα ήταν προτιμότερο να γίνει χειροκίνητη προσγείωση.

  Έχει εννέα χαρακτηριστικά τα οποία έχουν όλα αριθμητική μορφή, καθώς και επτά κλάσεις (Rad Flow, Fpv Close, Fpv Open, High, Bypass, Bpv Close, Bpv Open).

- Τέλος, το **Yeast** είναι ένα σύνολο δεδομένων που αφορά τους ζυμομύκητες (yeast cells). Ο σκοπός του είναι να προσδιορίσει τον υποκυτταρικό εντοπισμό κάθε κυττάρου. Έχει οκτώ χαρακτηριστικά και δέκα κλάσεις.

  Τα οκτώ χαρακτηριστικά του κάθε αντικείμενου είναι τα αποτελέσματα οκτώ διαφορετικών μεθόδων που χρησιμοποιούνται στην πρόβλεψη του υποκυτταρικού εντοπισμού (Mcg, Gvh, Alm, Mit, Erl, Pox, Vac, Nuc). Οι κλάσεις αντιπροσωπεύουν τα δέκα πιθανά αποτελέσματα (MIT, NUC, CYT, ME1, ME2, ME3, EXC, VAC, POX, ERL).

4.2.2 Τα Χρησιμοποιημένα Δυαδικά Σύνολα Δεδομένων

Στα πλαίσια αυτής της εργασίας, χρησιμοποιήθηκαν πέντε σύνολα δεδομένων που περιέχουν μόνο δύο κλάσεις, είναι δηλαδή δυαδικά. Είναι όλα τους διαθέσιμα στο KEEL. Ακολουθεί μία σύντομη περιγραφή των συγκεκριμένων συνόλων:

- Το **Page-Blocks 0** είναι μία δυαδική παραλλαγή του κανονικού Page-Blocks, όπου τα αντικείμενα που ανήκουν στην κλάση "Text" τίθενται ως τα αρνητικά και αυτά των υπόλοιπων κλάσεων τίθενται ως τα θετικά.

- Το **Segment 0** είναι μία τροποποιημένη εκδοχή του απλού Segment. Το συγκεκριμένο σύνολο δεδομένων περιέχει δεδομένα για επτά εικόνες, οι οποίες αποτελούν τις κλάσεις του. Οι εικόνες είναι χωρισμένες με το χέρι, έτσι ώστε να μπορέσει να κατηγοριοποιηθεί κάθε pixel ξεχωριστά και κάθε αντικείμενο αποτελείται από μία περιοχή $3x3$. Ο σκοπός του συνόλου δεδομένων, είναι να καταφέρει να ξεχωρίσει το είδος της επιφάνειας που φαίνεται σε κάθε περιοχή μίας εικόνας.

  Στο Segment 0 όμως έχουμε μόνο δύο κλάσεις, αντί για μία για κάθε εικόνα. Τα αρνητικά αντικείμενα ανήκουν στην Κλάση 0 και τα θετικά ανήκουν στην Κλάση 1.

- Το **Shuttle (c0 vs c4)** είναι μία δυαδική παραλλαγή του κανονικού Shuttle. Τα αρνητικά αντικείμενα ανήκουν στην Κλάση 4 και τα θετικά ανήκουν στην Κλάση 0.

- Το Vowel περιέχει δεδομένα ανεξάρτητης αναγνώρισης ομιλητή για τα φωνήεντα της Αγγλικής γλώσσας. Το αρχικό σύνολο δεδομένων διαθέτει έντεκα κλάσεις για τα φωνήεντα σταθερής κατάστασης των Βρετανικών Αγγλικών. Ωστόσο, εμείς χρησιμοποιούμε μία δυαδική εκδοχή του, το **Vowel 0**, όπου τα θετικά αντικείμενα ανήκουν στην Κλάση 0 και τα αρνητικά στην Κλάση 1. Επιπλέον, το Vowel 0 έχει αφαιρέσει μερικά διπλότυπα.

- Το Wine Quality Red είναι ένα σύνολο δεδομένων το οποίο συγκρατεί πληροφορίες για την αξιολόγηση της ποιότητας ενός κρασιού, το κόκκινο Portuguese Vinho Verde. Έχει έντεκα κλάσεις οι οποίες αντιπροσωπεύουν την βαθμολογία κάθε αντικειμένου μέσα στο σύνολο, η οποία κυμαίνεται από το μηδέν μέχρι το δέκα.

  Το **Wine Quality Red 4**, είναι μία δυαδική παραλλαγή του κανονικού συνόλου δεδομένων. Τα δεδομένα που ανήκαν στην Κλάση 4 εισάγονται στην κλάση των θετικών αντικειμένων, ενώ όλα τα υπόλοιπα στην κλάση των αρνητικών.

4.3 Πειραματική Μελέτη

Σε αυτήν την ενότητα, θα παρουσιάσουμε και θα αναλύσουμε τα αποτελέσματα των πειραμάτων μας. Συνολικά, έχουμε να συγκρίνουμε πέντε διαφορετικά κριτήρια: (i) το Ποσοστό Μείωσης, (ii) την Ακρίβεια, (iii) την Ορθότητα, (iv) την Ευαισθησία και (v) το F-Measure. Παρακάτω φαίνεται μία σειρά πινάκων, οι οποίοι καλύπτουν όλες αυτές τις κατηγορίες. Σε όλους αυτούς του πίνακες συγκρίνουμε τα αποτελέσματα που βρήκαμε για το κάθε ένα από τα δώδεκα σύνολα δεδομένων που αναφέραμε παραπάνω. Σε αντίθεση με τα (i) και (ii), τα κριτήρια (iii), (iv) και (v) μπορούν να μετρηθούν για κάθε κλάση ξεχωριστά. Σε αυτήν την εργασία, τα μετράμε μόνο για τις σπάνιες κλάσεις, καθώς αυτές μας ενδιαφέρουν. Ωστόσο, στην περίπτωση του σύνολο KDD, δεν συμπεριλαμβάνουμε μία

από τις σπάνιες κλάσεις σε αυτούς τους πίνακες, επειδή σε όλα τα πειράματα υπολόγιζε είτε μηδέν (0), είτε NaN (not a number). Αυτό συνέβη επειδή μόνο το σύνολο δοκιμής περιέχει αντικείμενα της συγκεκριμένης σπάνιας κλάσης.

Οι ΤΜΔ και οι προτεινόμενες μέθοδοι διατήρησης σπάνιων κλάσεων είναι υλοποιημένες σε C++. Τα πειράματα εκτελέστηκαν σε έναν εξυπηρετητή (server) του Τμήματος Μηχανικής Πληροφορικής και Ηλεκτρονικών Συστημάτων του Διεθνούς Πανεπιστημίου της Ελλάδος. Οι προδιαγραφές του συγκεκριμένου εξυπηρετητή έχουν ως εξής: 12 πυρήνες, 16 GB RAM, λειτουργικό σύστημα Debian GNU/Linux 10 (buster) με 2.40 GHz CPU.

Για να ελέγξουμε την απόδοση των προτεινόμενων μεθόδων μέσω των πειραμάτων μας, έπρεπε να χωρίσουμε τα σύνολα δεδομένων σε σύνολα εκπαίδευσης (*training sets*) και σύνολα δοκιμών (*testing sets*). Το ποσοστό του χωρισμού που επιλέχθηκε είναι 67% για το πρώτο και 33% για το δεύτερο, αυτό ισχύει για όλα τα σύνολα δεδομένων που χρησιμοποιήσαμε.

Τα πειράματα χωρίζονται σε τέσσερις κατηγορίες. Η πρώτη κατηγορία (Original) μας δείχνει τι αποτελέσματα λάβαμε από τον k-Nearest Neighbours (k-NN) αλγόριθμο χρησιμοποιώντας τα συμπυκνωμένα σύνολα (*condensing sets*) κάθε συνόλου δεδομένων. Τα αναφερόμενα συμπυκνωμένα σύνολα δημιουργήθηκαν μέσω των τεχνικών μείωσης δεδομένων (CNN, IB2, RSP3, AIB2, RHC). Οι υπόλοιπες τρεις κατηγορίες μας δείχνουν τα αντίστοιχα αποτέλεσμα για τα δικά μας επεξεργασμένα σύνολα που δημιουργήθηκαν μέσω των μεθόδων RCPM1, RCPM2 και RCPM-SMOTE.

Επίσης, σε όλα τα κριτήρια εκτός από το ποσοστό μείωσης παρουσιάζεται και το μη-επεξεργασμένο αποτέλεσμα του k-NN, δηλαδή αυτό που υπολογίζεται χρησιμοποιώντας το αρχικό σύνολο εκπαίδευσης στο οποίο δεν έχει εκτελεστεί καμία τεχνική μείωσης δεδομένων. Στους πίνακες της ακρίβειας, της ορθότητας, της ευαισθησίας και του F-Measure, όσο πιο κοντά είναι ένας αριθμός στο ένα (1,000), τόσο πιο πετυχημένη θεωρείται η κατηγοριοποίηση.

4.3.1 Πειράματα Ποσοστών Μείωσης Δεδομένων

Αναλύοντας τους Πίνακες σε αυτό το κεφάλαιο (4.2 - 4.6) παρατηρείται πως στις μεθόδους που χρησιμοποιούμε στα πειράματα μας (RCPM1, RCPM2, RCPM-SMOTE), για τα περισσότερα σύνολα, το ποσοστό μείωσης των δεδομένων είναι μικρότερο σε σχέση με αυτό της αρχικής διαδικασίας, κάτι που σημαίνει πως ο συνολικός αριθμός των δεδομένων στο συμπυκνωμένο σύνολο είναι μεγαλύτερος από ότι στην περίπτωση των αντίστοιχων προβλέψεων k-NN όπου εκτελέστηκε μείωση δεδομένων σε όλες τις κλάσεις. Αυτό είναι και το λογικό αποτέλεσμα λόγου του ότι τα δεδομένα που ανήκουν σε σπάνιες κλάσεις δεν μειώνονται.

Πίνακας 4.2: Ποσοστό % Μείωσης Δεδομένων - CNN

Datasets	Original	RCPM1	RCPM2	RCPM-SMOTE
Avila	61.671	61.246	60.873	58.723
Balance	67.308	74.038	67.308	43.990
Car	74.544	75.847	71.242	69.939
KDD	99.060	99.025	98.996	98.457
Page-Blocks	89.200	89.940	87.664	84.501
Shuttle	99.566	99.472	99.255	99.366
Yeast	31.749	31.446	28.514	-13.448
Page-Blocks 0	90.625	-	83.936	83.032
Segment 0	96.684	-	81.925	96.684
Shuttle: c0 vs c4	99.672	-	94.258	99.590
Vowel 0	94.529	-	86.778	93.769
Wine Quality Red 4	85.336	-	85.178	65.291

Πίνακας 4.3: Ποσοστό % Μείωσης Δεδομένων - IB2

Datasets	Original	RCPM1	RCPM2	RCPM-SMOTE
Avila	67.386	66.868	66.537	64.755
Balance	71.875	78.125	71.635	55.769
Car	80.626	80.278	76.455	76.021
KDD	99.198	99.154	99.133	98.834
Page-Blocks	91.530	91.091	89.419	74.589
Shuttle	99.604	99.478	99.289	99.501
Yeast	43.883	42.568	39.838	10.920
Page-Blocks 0	92.873	-	85.307	88.734
Segment 0	97.594	-	82.575	97.529
Shuttle: c0 vs c4	99.672	-	94.258	99.590
Vowel 0	95.137	-	87.234	94.883
Wine Quality Red 4	88.743	-	88.462	81.426

Πίνακας 4.4: Ποσοστό % Μείωσης Δεδομένων - RSP3

Datasets	Original	RCPM1	RCPM2	RCPM-SMOTE
Avila	47.581	47.150	47.085	38.358
Balance	61.058	67.308	60.817	25.240
Car	67.420	69.244	65.595	56.733
KDD	99.118	98.432	98.411	97.211
Page-Blocks	86.157	88.158	84.814	49.260
Shuttle	99.312	99.403	99.059	98.849
Yeast	26.997	26.694	24.469	-45.602
Page-Blocks 0	87.747	-	81.880	76.974
Segment 0	94.083	-	80.429	93.888
Shuttle: c0 vs c4	99.590	-	94.258	99.508
Vowel 0	90.578	-	83.739	90.729
Wine Quality Red 4	84.897	-	84.803	46.154

Πίνακας 4.5: Ποσοστό % Μείωσης Δεδομένων - AIB2

Datasets	Original	RCPM1	RCPM2	RCPM-SMOTE
Avila	69.104	68.349	68.227	66.846
Balance	73.077	81.971	72.115	63.462
Car	83.406	82.103	78.454	81.060
KDD	99.104	99.067	99.035	98.686
Page-Blocks	91.913	91.283	89.446	49.260
Shuttle	99.563	99.457	99.250	99.421
Yeast	45.501	44.388	40.950	13.852
Page-Blocks 0	92.708	-	85.280	88.816
Segment 0	97.854	-	82.705	97.724
Shuttle: c0 vs c4	99.754	-	94.340	99.754
Vowel 0	95.745	-	87.234	95.259
Wine Quality Red 4	89.587	-	89.024	84.240

Πίνακας 4.6: Ποσοστό % Μείωσης Δεδομένων - RHC

Datasets	Original	RCPM1	RCPM2	RCPM-SMOTE
Avila	70.175	69.758	69.334	66.933
Balance	79.087	81.490	76.202	53.606
Car	85.491	83.753	80.799	80.712
KDD	99.118	99.063	99.054	98.598
Page-Blocks	90.680	90.707	88.651	77.961
Shuttle	99.573	99.496	99.294	99.426
Yeast	48.534	46.714	44.186	11.426
Page-Blocks 0	91.859	-	84.923	85.307
Segment 0	98.049	-	83.030	97.984
Shuttle: c0 vs c4	99.754	-	94.340	99.754
Vowel 0	97.112	-	88.298	96.960
Wine Quality Red 4	92.026	-	91.651	76.642

Όπως φαίνεται, στα μεγαλύτερα σύνολα δεδομένων όπως το KDD και το Shuttle το ποσοστό μείωσης έχει επηρεαστεί ελάχιστα, ενώ στα μικρότερα σύνολα όπως το Balance και το Wine Quality Red 4 μπορούμε να δούμε αρκετά σημαντικές αλλαγές. Σαφώς, όταν έχουμε έναν μικρό αριθμό αντικειμένων στην διάθεση μας ακόμα και οι μικρότερες μετατροπές του μεγέθους ενός συνόλου μπορούν να μεταβάλουν δραστικά το ποσοστό μείωσης του.

Επιπλέον, σε ορισμένες περιπτώσεις θα δούμε πως για την μέθοδο RCPM-SMOTE υπολογίστηκε ένας αρνητικός αριθμός για το ποσοστό μείωσης. Ο λόγος που συμβαίνει αυτό, είναι επειδή υπολογίζουμε το ποσοστό μείωσης σε σχέση με το αρχικό εκπαιδευτικό σύνολο και όχι αυτό που δημιουργήσαμε μέσω της SMOTE. Επομένως, σε μερικές περιστάσεις το συμπυκνωμένο σύνολο έχει περισσότερα αντικείμενα σε σύγκριση με το αρχικό εκπαιδευτικό σύνολο, βλάπτοντας έτσι το ποσοστό μείωσης.

Όσο αφορά την σύγκριση του ποσοστού μείωσης μεταξύ των πέντε ΤΜΔ, αυτή με τις πιο βέλτιστες αποδώσεις είναι ο αλγόριθμος RHC, ωστόσο σε μερικές περιπτώσεις τον ξεπερνάνε ο IB2 και ο AIB2. Ο αλγόριθμος RSP3 έχει γενικά την χαμηλότερη απόδοση.

Τέλος, από τις τέσσερις διαφορετικές προσεγγίσεις στην κατηγοριοποίηση, η απλή μείωση δεδομένων (Original), καθώς και η RCPM1 πετυχαίνουν τα υψηλότερα ποσοστά μείωσης στις περισσότερες περιπτώσεις. Αυτό είναι και λογικό για το θέμα του ποσοστού μείωσης, καθώς οι μέθοδοι RCPM2 και RCPM-SMOTE έχουν περισσότερα δεδομένα να επεξεργαστούν.

4.3.2 Πειράματα Ακρίβειας

Στους Πίνακες 4.8 - 4.12 φαίνονται τα ποσοστά ακρίβειας των συνόλων δεδομένων με κάθε μέθοδο. Ο κάθε πίνακας αντιπροσωπεύει και μία ξεχωριστή ΤΜΔ. Επιπλέον, ο Πίνακας 4.7 μας δείχνει τι ακρίβεια πετυχαίνει κάθε σύνολο όταν χρησιμοποιείται μόνο η κατηγοριοποίηση k-NN, χωρίς να μειώσουμε καθόλου τα δεδομένα.

Παρατηρώντας αυτούς τους πίνακες, βλέπουμε πως σε πολλές περιπτώσεις η ακρίβεια μειώνεται, είτε σε μικρό είτε σε μεγάλο βαθμό. Για παράδειγμα στο σύνολο Avila, παρατηρείται μία αξιοσημείωτη μείωση της ακρίβειας σε όλες τις περιπτώσεις μας. Όμως, στο σύνολο Car οι μειώσεις είναι ελάχιστες, ενώ παράλληλα έχουμε και μερικές περιπτώσεις που πετυχαίνουν και λίγο μεγαλύτερη ακρίβεια από την σκέτη κατηγοριοποίηση. Μια ιδιαίτερα εν-

Πίνακας 4.7: Ακρίβεια - Μόνο Κατηγοριοποίηση k-NN

Dataset	k-NN only
Avila	0,813
Balance	0,813
Car	0,859
KDD	0,997
Page-Blocks	0,954
Shuttle	0,999
Yeast	0,475
Page-Blocks 0	0,951
Segment 0	0,995
Shuttle: c0 vs c4	1,000
Vowel 0	1,000
Wine Quality Red 4	0,934

Πίνακας 4.8: Ακρίβεια - CNN

Datasets	Original	RCPM1	RCPM2	RCPM-SMOTE
Avila	0,776	0,772	0,775	0,779
Balance	0,718	0,746	0,751	0,708
Car	0,865	0,844	0,861	0,852
KDD	0,997	0,985	0,997	0,996
Page-Blocks	0,945	0,839	0,946	0,912
Shuttle	0,999	0,464	0,990	0,999
Yeast	0,448	0,430	0,448	0,426
Page-Blocks 0	0,947	-	0,936	0,939
Segment 0	0,992	-	0,974	0,991
Shuttle: c0 vs c4	1,000	-	1,000	1,000
Vowel 0	0,991	-	0,979	0,991
Wine Quality Red 4	0,906	-	0,906	0,857

Πίνακας 4.9: Ακρίβεια - IB2

Datasets	Original	RCPM1	RCPM2	RCPM-SMOTE
Avila	0,756	0,753	0,754	0,756
Balance	0,703	0,737	0,703	0,670
Car	0,861	0,812	0,851	0,847
KDD	0,995	0,984	0,995	0,987
Page-Blocks	0,925	0,820	0,926	0,739
Shuttle	0,997	0,464	0,988	0,948
Yeast	0,432	0,412	0,434	0,394
Page-Blocks 0	0,925	-	0,917	0,858
Segment 0	0,988	-	0,951	0,986
Shuttle: c0 vs c4	1,000	-	1,000	1,000
Vowel 0	0,982	-	0,964	0,982
Wine Quality Red 4	0,859	-	0,856	0,659

Πίνακας 4.10: Ακρίβεια - RSP3

Datasets	Original	RCPM1	RCPM2	RCPM-SMOTE
Avila	0,792	0,794	0,792	0,791
Balance	0,737	0,746	0,764	0,703
Car	0,856	0,849	0,849	0,870
KDD	0,996	0,994	0,996	0,996
Page-Blocks	0,941	0,895	0,942	0,911
Shuttle	0,996	0,587	0,995	0,995
Yeast	0,481	0,463	0,477	0,432
Page-Blocks 0	0,940	-	0,932	0,933
Segment 0	0,988	-	0,975	0,988
Shuttle: c0 vs c4	1,000	-	1,000	0,998
Vowel 0	0,985	-	0,982	0,988
Wine Quality Red 4	0,889	-	0,887	0,854

Πίνακας 4.11: Ακρίβεια - AIB2

Datasets	Original	RCPM1	RCPM2	RCPM-SMOTE
Avila	0,767	0,764	0,767	0,765
Balance	0,785	0,718	0,766	0,770
Car	0,873	0,816	0,851	0,863
KDD	0,994	0,992	0,994	0,992
Page-Blocks	0,922	0,885	0,921	0,808
Shuttle	0,990	0,679	0,986	0,982
Yeast	0,469	0,451	0,463	0,412
Page-Blocks 0	0,921	-	0,922	0,889
Segment 0	0,982	-	0,948	0,981
Shuttle: c0 vs c4	1,000	-	0,997	1,000
Vowel 0	0,997	-	0,997	0,997
Wine Quality Red 4	0,865	-	0,850	0,777

Πίνακας 4.12: Ακρίβεια - RHC

Datasets	Original	RCPM1	RCPM2	RCPM-SMOTE
Avila	0,759	0,764	0,760	0,759
Balance	0,732	0,713	0,742	0,722
Car	0,833	0,674	0,806	0,818
KDD	0,995	0,993	0,994	0,994
Page-Blocks	0,931	0,868	0,928	0,913
Shuttle	0,997	0,460	0,988	0,997
Yeast	0,420	0,438	0,416	0,416
Page-Blocks 0	0,930	-	0,922	0,920
Segment 0	0,987	-	0,962	0,984
Shuttle: c0 vs c4	1,000	-	0,997	1,000
Vowel 0	1,000	-	0,985	1,000
Wine Quality Red 4	0,856	-	0,856	0,824

διαφέρουσα περίπτωση, είναι αυτή του συνόλου Shuttle, όπου βλέπουμε πως η μέθοδος RCPM1 βλάπτει σημαντικά την ακρίβεια του, αλλά οι άλλες δύο (RCPM2, RCPM-SMOTE) καταφέρνουν να την διατηρήσουν ως επί το πλείστον.

Όσο αφορά την σύγκριση της ακρίβειας μεταξύ της κλασικής κατηγοριοποίησης με μείωση κλάσεων (Original) και τις τρεις μεθόδους RCPM φαίνεται πως στις περισσότερες περιπτώσεις το ποσοστό ακρίβειας μένει αρκετά σταθερό. Ωστόσο υπάρχουν και μερικές εξαιρέσεις όπου χάνεται ένα μεγάλο ποσοστό ακρίβειας, κυρίως στο RCPM-SMOTE.

Τέλος, από τις πέντε ΤΜΔ, ο RSP3 φαίνεται να πετυχαίνει την μεγαλύτερη ακρίβεια σε αρκετές περιπτώσεις, ωστόσο υπάρχουν και κάποιες εξαιρέσεις όπου οι CNN και AIB2 αποδίδουν καλύτερα.

4.3.3 Πειράματα Ορθότητας, Ευαισθησίας και F-Measure

Στην συνέχεια θα ασχοληθούμε με τα αποτελέσματα της ορθότητας, της ευαισθησίας και του F-Measure, τα οποία είναι και τα πιο σημαντικά κριτήρια όσο αφορά το θέμα των μη-ισορροπημένων συνόλων δεδομένων. Στον Πίνακα 4.13 βλέπουμε τα αποτελέσματα τους για τις σπάνιες κλάσεις αν εκτελέσουμε τον κατηγοριοποιητή k-NN από μόνο του, ενώ στους Πίνακες 4.14 - 4.18 φαίνονται τα ίδια αποτελέσματα, αλλά με τις τέσσερις διαφορετικές προσεγγίσεις προς την μείωση δεδομένων (Original, RCPM1, RCPM2, RCPM-SMOTE), όπου με Original εννοούμε την απλή μείωση δεδομένων πάνω σε όλο το εκπαιδευτικό σύνολο. Κάθε Πίνακας αντιπροσωπεύει και μία ξεχωριστή ΤΜΔ. Να θυμίσουμε πως η ορθότητα, η ευαισθησία και το F-Measure δεν έχουν όλα τους πάντα την ίδια σημασία. Ανάλογα με το σύνολο που χρησιμοποιείται, τα αποτελέσματα του ενός κριτηρίου μπορεί να αποδείξουν πιο ωφέλιμα από τα άλλα.

Ξεκινάμε με την ορθότητα, η οποία είναι πιο χαμηλή για τις τρεις μεθόδους RCPM σε σύγκριση με την απλή μείωση δεδομένων. Η μέθοδος με την μεγαλύτερη πτώση είναι με διαφορά η RCPM1. Η μέθοδος RCPM2, αν και στις περισσότερες περιπτώσεις υπολογίζει κι αυτή μικρότερο ποσοστό ορθότητας, παράγει τιμές που είναι πολύ πιο κοντά σε αυτές

της Original. Τέλος, το RCPM-SMOTE έχει τα πιο μικτά αποτελέσματα από όλες τις προσεγγίσεις. Οι τιμές που υπολογίζει μπορούν να πέφτουν όσο χαμηλά όσο αυτές της RCPM1, ή να ξεπερνάνε την Original. Ωστόσο, στις περισσότερες περιπτώσεις φαίνεται να έχει παρόμοιες τιμές με την RCPM2.

Όσο αφορά την απόδοση των ΤΜΔ, πάνω στο θέμα της ορθότητας φαίνεται να είναι σχετικά ισορροπημένες μεταξύ τους. Βέβαια, σε μερικές περιπτώσεις κάποιες είναι αρκετά καλύτερες. Ένα τέτοιο παράδειγμα είναι η ορθότητα του συνόλου Shuttle με την μέθοδο RCPM2, όπου η τεχνική μείωσης RSP3 πετυχαίνει πολύ μεγαλύτερο αριθμό από τις υπόλοιπες τέσσερις. Ωστόσο, με τα συγκεκριμένα αποτελέσματα δεν φαίνεται να υπάρχει κάποιο οριστικό συμπέρασμα όσο αφορά ποια τεχνική είναι η καλύτερη για την ορθότητα, δεδομένου πως καμία τους δεν έχει έναν σημαντικό αριθμό καλύτερων υπολογισμών από τις άλλες.

Προχωρώντας, θα εξετάσουμε τα ποσοστά ευαισθησίας των πειραμάτων μας. Πρώτον, θα χωρίσουμε τα αποτελέσματα σε τέσσερις διαφορετικές κατηγορίες και θα τα συγκρίνουμε μεταξύ τους: την απλή διαδικασία μείωσης δεδομένων (Original) και τις τρεις μεθόδους διατήρησης (RCPM1, RCPM2, RCPM-SMOTE). Σε αυτήν την περίπτωση, τα αποτελέσματα φαίνονται να είναι ακριβώς αντίθετα από αυτά της ορθότητας. Δηλαδή, βλέπουμε πως σε αρκετά σύνολα δεδομένων η ευαισθησία των σπάνιων κλάσεων αυξάνεται και από τις τρεις μεθόδους σε σχέση με την Original. Η μέθοδος με τις μεγαλύτερες βελτιώσεις είναι η RCPM1. Η RCPM2 παρουσιάζει και αυτή βελτιώσεις, αλλά όχι στο ίδιο επίπεδο με την RCPM1, ενώ η RCPM-SMOTE είναι η πιο ελαστική από όλες. Αυτό ισχύει για την πλειονότητα των συνόλων δεδομένων.

Στην συνέχεια θα συγκρίνουμε τις ΤΜΔ μεταξύ τους για να δούμε ποια από από τις πέντε μας δίνει καλύτερα ποσοστά ευαισθησίας. Αν και όπως με την ορθότητα, η μέτρηση της ευαισθησίας είναι σχετικά ισορροπημένη μεταξύ των ΤΜΔ στις περισσότερες περιπτώσεις, έχουμε και μερικά αποτελέσματα όπου η ευαισθησία μίας ΤΜΔ προεξέχει αρκετά από τις αντίστοιχες του, κυρίως για τις μεθόδους AIB2 και RHC.

Τέλος, θα κοιτάξουμε τα αποτελέσματα που μας βγάζει το F-Measure, το οποίο προσφέρει το αρμονικό μέσο της ορθότητας και της ευαισθησίας. Ξεκινάμε με τις διαφορές μεταξύ της απλής μείωσης δεδομένων (Original), τις των τριών μεθόδων RCPM. Αναλύοντας τους πίνακες, θα δούμε πως σε πολλές από τις σπάνιες κλάσεις, το Original F-Measure είναι το μεγαλύτερο. Αυτό μας δείχνει πως δυστυχώς οι μειώσεις της ορθότητας είναι μεγαλύτερες από τις αυξήσεις της ευαισθησίας που παρατηρούνται στις άλλες μεθόδους. Ωστόσο παρατηρούνται και μερικές περιπτώσεις με αρκετά ικανοποιητικές βελτιώσεις, για παράδειγμα και οι τρεις μέθοδοι δίνουν πολύ καλά αποτελέσματα για το σύνολο δεδομένων Car. Οι μέθοδοι RCPM1 και RCPM-SMOTE δίνουν καλά αποτελέσματα και σε κάποιες άλλες περιπτώσεις.

Στην σύγκριση των ΤΜΔ, βλέπουμε μερικές περιπτώσεις όπου ο CNN-Rule, καθώς και

ο RSP3 πετυχαίνουν μεγαλύτερα αποτελέσματα από τις υπόλοιπες τεχνικές. Ωστόσο, η τεχνική που υπολογίζει τα μεγαλύτερα αποτελέσματα είναι ο αλγόριθμος RHC.

Πίνακας 4.13: Ορθότητα/Ευαισθησία/F-measure - Μόνο Κατηγοριοποίηση k-NN

Dataset	Rare Classes	k-NN Only		
		Precision	Recall	F-Measure
Avila	Class [1]	1,000	1,000	1,000
	Class [2]	0,828	0,679	0,746
	Class [9]	0,839	0,929	0,882
Balance	Class [1]	0,100	0,050	0,067
Car	Class [2]	0,579	0,478	0,524
	Class [3]	0,640	0,696	0,667
KDD	Class [1]	0,625	0,769	0,690
	Class [2]	1,000	0,500	0,667
	Class [3]	1,000	0,905	0,950
	Class [4]	0,833	0,833	0,833
	Class [6]	1,000	0,714	0,833
	Class [7]	0,000	0,000	0,000
	Class [8]	0,250	0,200	0,222
	Class [12]	0,500	0,500	0,500
	Class [13]	1,000	1,000	1,000
	Class [16]	0,000	0,000	0,000
	Class [22]	1,000	0,833	0,909
Page-Blocks	Class [2]	0,875	0,538	0,666
	Class [3]	0,609	0,737	0,667
	Class [4]	0,524	0,550	0,537
Shuttle	Class [1]	1,000	1,000	1,000
	Class [2]	0,964	0,841	0,898
	Class [5]	nan	0,000	nan
	Class [6]	0,500	0,500	0,500
Yeast	Class [3]	0,692	0,750	0,720
	Class [4]	0,350	0,389	0,368
	Class [6]	0,455	0,385	0,417
	Class [7]	0,000	0,000	0,000
	Class [8]	0,364	0,800	0,500
	Class [9]	1,000	1,000	1,000
Page-Blocks 0	positive	0,739	0,764	0,751
Segment 0	positive	0,988	0,966	0,977
Shuttle: c0 vs c4	positive	1,000	1,000	1,000
Vowel 0	positive	1,000	1,000	1,000
Wine Quality Red 4	positive	0,130	0,167	0,146

Πίνακας 4.14: Ορθότητα/Ευαισθησία/F-measure - CNN

Datasets	Rare Classes	Precision				Recall				F-Measure			
		Original	RCPM1	RCPM2	RCPM-SMOTE	Original	RCPM1	RCPM2	RCPM-SMOTE	Original	RCPM1	RCPM2	RCPM-SMOTE
Avila	Class [1]	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
	Class [2]	0,764	0,589	0,724	0,727	0,705	0,718	0,705	0,821	0,733	0,647	0,714	0,771
	Class [9]	0,735	0,578	0,643	0,722	0,893	0,929	0,964	0,929	0,806	0,713	0,771	0,813
Balance	Class [1]	0,200	0,200	0,300	0,148	0,200	0,150	0,150	0,200	0,200	0,171	0,200	0,170
Car	Class [2]	0,571	0,472	0,556	0,552	0,522	0,739	0,652	0,696	0,545	0,576	0,600	0,616
	Class [3]	0,640	0,500	0,594	0,613	0,696	0,957	0,826	0,826	0,667	0,657	0,691	0,704
KDD	Class [1]	0,625	0,458	0,588	0,611	0,769	0,846	0,769	0,846	0,690	0,594	0,666	0,710
	Class [2]	1,000	1,000	1,000	1,000	0,500	0,500	0,500	0,500	0,667	0,667	0,667	0,667
	Class [3]	1,000	0,040	1,000	1,000	0,905	0,952	0,952	0,905	0,950	0,077	0,975	0,950
	Class [4]	0,833	0,455	0,714	0,714	0,833	0,833	0,833	0,833	0,833	0,589	0,769	0,769
	Class [6]	1,000	1,000	1,000	1,000	0,714	0,714	0,714	0,857	0,833	0,833	0,833	0,923
	Class [7]	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	Class [8]	0,250	0,250	0,250	0,250	0,200	0,200	0,200	0,200	0,222	0,222	0,222	0,222
	Class [12]	0,500	0,500	5,000	1,000	0,500	0,500	0,500	0,500	0,500	0,500	0,909	0,667
	Class [13]	1,000	0,333	1,000	1,000	1,000	1,000	0,100	1,000	1,000	0,500	0,182	1,000
	Class [16]	0,000	0,021	0,000	0,083	0,000	0,167	0,000	0,167	0,000	0,037	0,000	0,111
	Class [22]	1,000	1,000	1,000	1,000	0,833	0,833	0,833	0,833	0,909	0,909	0,909	0,909
Page-Blocks	Class [2]	0,800	0,562	0,800	0,667	0,615	0,692	0,615	0,769	0,695	0,620	0,695	0,714
	Class [3]	0,538	0,389	0,583	0,342	0,737	0,737	0,737	0,684	0,622	0,509	0,651	0,456
	Class [4]	0,438	0,125	0,451	0,262	0,525	0,775	0,575	0,675	0,478	0,215	0,506	0,377
Shuttle	Class [1]	1,000	0,007	0,153	1,000	1,000	1,000	1,000	1,000	1,000	0,014	0,265	1,000
	Class [2]	0,964	0,007	0,368	0,982	0,857	1,000	0,889	0,873	0,907	0,014	0,521	0,924
	Class [5]	nan	0,000	nan	1,000	0,000	0,000	0,000	0,250	nan	0,000	0,000	0,400
	Class [6]	0,500	0,667	0,500	0,067	0,500	1,000	0,500	1,000	0,500	0,800	0,500	0,125
Yeast	Class [3]	0,800	0,692	0,692	0,889	0,667	0,750	0,750	0,667	0,727	0,720	0,720	0,762
	Class [4]	0,300	0,318	0,350	0,225	0,333	0,389	0,389	0,500	0,316	0,350	0,368	0,310
	Class [6]	0,455	0,455	0,455	0,267	0,385	0,385	0,385	0,308	0,417	0,417	0,417	0,286
	Class [7]	0,000	0,059	0,000	0,021	0,000	0,200	0,000	0,200	0,000	0,091	0,000	0,038
	Class [8]	0,364	0,333	0,364	0,176	0,800	0,800	0,800	0,600	0,500	0,470	0,500	0,272
	Class [9]	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
Page-Blocks 0	positive	0,707	-	0,642	0,649	0,787	-	0,775	0,820	0,745	-	0,702	0,725
Segment 0	positive	0,966	-	0,833	0,955	0,966	-	0,966	0,966	0,966	-	0,895	0,960
Shuttle: c0 vs c4	positive	1,000	-	1,000	1,000	1,000	-	1,000	1,000	1,000	-	1,000	1,000
Vowel 0	positive	0,923	-	0,781	0,923	0,960	-	1,000	0,960	0,941	-	0,877	0,941
Wine Quality Red 4	positive	0,079	-	0,079	0,097	0,167	-	0,167	0,389	0,107	-	0,107	0,155

Πίνακας 4.15: Ορθότητα/Ευαισθησία/F-measure - IB2

Datasets	Rare Classes	Precision				Recall				F-Measure			
		Original	RCPM1	RCPM2	RCPM-SMOTE	Original	RCPM1	RCPM2	RCPM-SMOTE	Original	RCPM1	RCPM2	RCPM-SMOTE
Avila	Class [1]	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
	Class [2]	0,707	0,564	0,629	0,650	0,679	0,731	0,718	0,833	0,693	0,637	0,671	0,730
	Class [9]	0,735	0,565	0,643	0,703	0,893	0,929	0,964	0,929	0,806	0,703	0,771	0,800
Balance	Class [1]	0,190	0,185	0,176	0,125	0,200	0,250	0,150	0,200	0,195	0,213	0,162	0,154
Car	Class [2]	0,522	0,400	0,455	0,516	0,522	0,783	0,652	0,696	0,522	0,783	0,652	0,696
	Class [3]	0,679	0,468	0,553	0,625	0,826	0,957	0,913	0,870	0,826	0,957	0,913	0,870
KDD	Class [1]	0,471	0,458	0,476	0,478	0,615	0,846	0,769	0,846	0,533	0,594	0,588	0,611
	Class [2]	1,000	1,000	1,000	0,500	0,500	0,500	0,500	0,500	0,667	0,667	0,667	0,500
	Class [3]	1,000	0,040	1,000	1,000	0,905	0,952	0,952	0,905	0,950	0,077	0,975	0,950
	Class [4]	0,500	0,455	0,625	0,500	0,667	0,833	0,833	0,667	0,572	0,589	0,714	0,572
	Class [6]	1,000	1,000	1,000	1,000	0,714	0,714	0,714	0,857	0,833	0,833	0,833	0,923
	Class [7]	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	Class [8]	0,250	0,250	0,250	0,250	0,200	0,200	0,200	0,200	0,222	0,222	0,222	0,222
	Class [12]	0,500	0,500	0,500	1,000	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,667
	Class [13]	1,000	0,333	1,000	1,000	1,000	1,000	0,100	1,000	1,000	0,500	0,182	1,000
	Class [16]	0,000	0,021	0,000	0,010	0,000	0,167	0,000	0,167	0,000	0,037	0,000	0,019
	Class [22]	1,000	1,000	1,000	0,833	0,833	0,833	0,833	0,833	0,909	0,909	0,909	0,833
Page-Blocks	Class [2]	0,600	0,529	0,615	0,600	0,462	0,692	0,615	0,692	0,522	0,600	0,615	0,643
	Class [3]	0,520	0,341	0,600	0,259	0,684	0,737	0,789	0,789	0,591	0,466	0,682	0,390
	Class [4]	0,344	0,123	0,342	0,078	0,525	0,775	0,625	0,700	0,416	0,212	0,442	0,140
Shuttle	Class [1]	0,241	0,007	0,099	0,027	1,000	1,000	1,000	1,000	0,388	0,014	0,180	0,053
	Class [2]	0,966	0,007	0,377	0,098	0,889	1,000	0,921	0,905	0,926	0,014	0,535	0,177
	Class [5]	0,000	0,000	0,000	1,000	0,000	0,000	0,000	0,250	0,000	0,000	0,000	0,400
	Class [6]	0,333	0,667	0,500	0,667	0,250	1,000	0,500	1,000	0,286	0,800	0,500	0,800
Yeast	Class [3]	0,750	0,692	0,692	0,875	0,500	0,750	0,750	0,583	0,600	0,720	0,720	0,700
	Class [4]	0,316	0,280	0,318	0,237	0,333	0,389	0,389	0,500	0,324	0,326	0,350	0,322
	Class [6]	0,417	0,455	0,455	0,250	0,385	0,385	0,385	0,380	0,400	0,417	0,417	0,302
	Class [7]	0,000	0,048	0,000	0,017	0,000	0,200	0,000	0,200	0,000	0,077	0,000	0,031
	Class [8]	0,500	0,333	0,364	0,150	0,600	0,800	0,800	0,600	0,545	0,470	0,500	0,240
	Class [9]	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
Page-Blocks 0	positive	0,598	-	0,549	0,396	0,702	-	0,815	0,865	0,646	-	0,656	0,543
Segment 0	positive	0,934	-	0,708	0,914	0,966	-	0,966	0,966	0,950	-	0,817	0,939
Shuttle: c0 vs c4	positive	1,000	-	1,000	1,000	1,000	-	1,000	1,000	1,000	-	1,000	1,000
Vowel 0	positive	0,828	-	0,676	0,828	0,960	-	1,000	0,960	0,889	-	0,807	0,889
Wine Quality Red 4	positive	0,062	-	0,060	0,044	0,222	-	0,222	0,444	0,097	-	0,094	0,080

Πίνακας 4.16: Ορθότητα/Ευαισθησία/F-measure - RSP3

Datasets	Rare Classes	Precision				Recall				F-Measure			
		Original	RCPM1	RCPM2	RCPM-SMOTE	Original	RCPM1	RCPM2	RCPM-SMOTE	Original	RCPM1	RCPM2	RCPM-SMOTE
Avila	Class [1]	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
	Class [2]	0,788	0,743	0,783	0,713	0,667	0,705	0,692	0,795	0,722	0,724	0,735	0,752
	Class [9]	0,818	0,711	0,794	0,844	0,964	0,964	0,964	0,964	0,885	0,818	0,871	0,900
Balance	Class [1]	0,059	0,077	0,077	0,154	0,050	0,050	0,050	0,200	0,054	0,061	0,061	0,174
Car	Class [2]	0,667	0,515	0,591	0,652	0,522	0,739	0,565	0,652	0,586	0,607	0,578	0,652
	Class [3]	0,609	0,514	0,533	0,680	0,609	0,783	0,696	0,739	0,609	0,621	0,604	0,708
KDD	Class [1]	0,524	0,370	0,524	0,647	0,846	0,769	0,846	0,846	0,647	0,500	0,647	0,733
	Class [2]	0,667	1,000	0,667	0,667	0,500	0,500	0,500	0,500	0,572	0,667	0,572	0,572
	Class [3]	1,000	0,370	1,000	0,950	0,905	0,952	0,905	0,905	0,950	0,533	0,950	0,927
	Class [4]	0,714	0,455	0,714	0,714	0,833	0,833	0,833	0,833	0,769	0,589	0,769	0,769
	Class [6]	1,000	1,000	1,000	1,000	0,857	1,000	1,000	0,857	0,923	1,000	1,000	0,923
	Class [7]	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	Class [8]	0,250	0,250	0,250	0,250	0,200	0,200	0,200	0,200	0,222	0,222	0,222	0,222
	Class [12]	0,500	0,500	0,500	1,000	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,667
	Class [13]	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
	Class [16]	0,000	0,031	0,000	0,100	0,000	0,167	0,000	0,167	0,000	0,052	0,000	0,125
	Class [22]	1,000	1,000	1,000	1,000	0,833	0,833	0,833	0,833	0,909	0,909	0,909	0,909
Page-Blocks	Class [2]	0,800	0,769	0,818	0,733	0,615	0,769	0,692	0,846	0,695	0,769	0,750	0,785
	Class [3]	0,524	0,412	0,636	0,389	0,579	0,737	0,737	0,737	0,550	0,529	0,683	0,509
	Class [4]	0,458	0,200	0,434	0,261	0,550	0,725	0,575	0,725	0,500	0,314	0,495	0,384
Shuttle	Class [1]	0,684	0,007	0,565	0,867	1,000	1,000	1,000	1,000	0,812	0,014	0,722	0,929
	Class [2]	0,859	0,010	0,814	0,465	0,873	1,000	0,905	0,937	0,866	0,020	0,857	0,622
	Class [5]	nan	0,000	nan	nan	0,000	0,000	0,000	0,000	nan	0,000	0,000	0,000
	Class [6]	0,500	0,667	0,500	0,667	0,500	1,000	0,500	1,000	0,500	0,800	0,500	0,800
Yeast	Class [3]	0,750	0,750	0,692	0,750	0,750	0,750	0,750	0,750	0,750	0,750	0,720	0,750
	Class [4]	0,368	0,318	0,350	0,235	0,389	0,389	0,389	0,444	0,378	0,350	0,368	0,307
	Class [6]	0,500	0,417	0,455	0,235	0,462	0,385	0,385	0,308	0,480	0,400	0,417	0,267
	Class [7]	0,167	0,000	0,167	0,042	0,200	0,000	0,200	0,400	0,182	0,000	0,000	0,076
	Class [8]	0,364	0,308	0,364	0,143	0,800	0,800	0,800	0,600	0,500	0,445	0,500	0,231
	Class [9]	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
Page-Blocks 0	positive	0,657	-	0,619	0,618	0,798	-	0,787	0,809	0,721	-	0,693	0,701
Segment 0	positive	0,934	-	0,842	0,934	0,966	-	0,966	0,966	0,950	-	0,900	0,950
Shuttle: c0 vs c4	positive	1,000	-	1,000	0,982	1,000	-	1,000	1,000	1,000	-	1,000	0,991
Vowel 0	positive	0,833	-	0,806	0,862	1,000	-	1,000	1,000	0,909	-	0,893	0,926
Wine Quality Red 4	positive	0,098	-	0,096	0,083	0,278	-	0,278	0,333	0,145	-	0,143	0,133

Πίνακας 4.17: Ορθότητα/Ευαισθησία/F-measure - AIB2

Datasets	Rare Classes	Precision				Recall				F-Measure			
		Original	RCPM1	RCPM2	RCPM-SMOTE	Original	RCPM1	RCPM2	RCPM-SMOTE	Original	RCPM1	RCPM2	RCPM-SMOTE
Avila	Class [1]	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
	Class [2]	0,754	0,667	0,718	0,645	0,667	0,718	0,718	0,769	0,708	0,692	0,718	0,702
	Class [9]	0,774	0,683	0,692	0,812	0,857	1,000	0,964	0,929	0,813	0,812	0,806	0,867
Balance	Class [1]	0,000	0,194	0,143	0,185	0,000	0,300	0,150	0,250	0,000	0,236	0,146	0,213
Car	Class [2]	0,667	0,353	0,472	0,533	0,609	0,783	0,739	0,696	0,637	0,487	0,576	0,604
	Class [3]	0,680	0,488	0,556	0,586	0,739	0,913	0,870	0,739	0,708	0,636	0,678	0,654
KDD	Class [1]	0,474	0,414	0,407	0,550	0,692	0,923	0,846	0,846	0,563	0,572	0,550	0,667
	Class [2]	0,667	0,667	0,667	0,667	0,500	0,500	0,500	0,500	0,572	0,572	0,572	0,572
	Class [3]	0,950	0,299	0,870	0,952	0,905	0,952	0,952	0,952	0,927	0,455	0,909	0,952
	Class [4]	0,714	0,455	0,714	0,714	0,833	0,833	0,833	0,833	0,769	0,589	0,769	0,769
	Class [6]	1,000	1,000	1,000	1,000	0,857	1,000	1,000	1,000	0,923	1,000	1,000	1,000
	Class [7]	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	Class [8]	0,250	0,250	0,250	0,250	0,200	0,200	0,200	0,200	0,222	0,222	0,222	0,222
	Class [12]	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500
	Class [13]	1,000	0,333	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,500	1,000	1,000
	Class [16]	0,000	0,017	0,000	0,000	0,000	0,167	0,000	0,000	0,000	0,031	0,000	0,000
	Class [22]	1,000	1,000	1,000	0,714	0,833	0,833	0,833	0,833	0,909	0,909	0,909	0,769
Page-Blocks	Class [2]	0,875	0,692	0,800	0,769	0,538	0,692	0,615	0,769	0,666	0,692	0,695	0,769
	Class [3]	0,609	0,593	0,615	0,309	0,737	0,842	0,842	0,895	0,667	0,696	0,711	0,459
	Class [4]	0,388	0,370	0,382	0,129	0,650	0,750	0,725	0,800	0,486	0,496	0,500	0,222
Shuttle	Class [1]	0,078	0,011	0,075	0,075	1,000	1,000	1,000	1,000	0,145	0,022	0,140	0,140
	Class [2]	0,812	0,012	0,401	0,261	0,889	1,000	0,937	0,921	0,849	0,024	0,562	0,407
	Class [5]	0,000	0,000	0,000	nan	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	Class [6]	0,500	0,667	0,500	0,667	0,500	1,000	0,500	1,000	0,500	0,800	0,500	0,800
Yeast	Class [3]	0,700	0,692	0,692	0,875	0,583	0,750	0,750	0,583	0,636	0,720	0,720	0,700
	Class [4]	0,375	0,286	0,308	0,229	0,500	0,444	0,444	0,444	0,429	0,348	0,364	0,302
	Class [6]	0,500	0,417	0,417	0,263	0,462	0,385	0,385	0,385	0,480	0,400	0,400	0,313
	Class [7]	0,000	0,050	0,000	0,019	0,000	0,200	0,000	0,200	0,000	0,080	0,000	0,035
	Class [8]	0,429	0,308	0,286	0,150	0,600	0,800	0,800	0,600	0,500	0,445	0,421	0,240
	Class [9]	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
Page-Blocks 0	positive	0,576	-	0,566	0,463	0,702	-	0,843	0,843	0,633	-	0,677	0,598
Segment 0	positive	0,885	-	0,688	0,876	0,966	-	1,000	0,966	0,924	-	0,815	0,919
Shuttle: c0 vs c4	positive	1,000	-	0,965	1,000	1,000	-	1,000	1,000	1,000	-	0,982	1,000
Vowel 0	positive	0,962		0,962	0,962	1,000		1,000	1,000	0,981	-	0,981	0,981
Wine Quality Red 4	positive	0,078	-	0,081	0,068	0,278	-	0,333	0,444	0,122	-	0,130	0,118

Πίνακας 4.18: Ορθότητα/Ευαισθησία/F-measure - RHC

Datasets	Rare Classes	Precision				Recall				F-Measure			
		Original	RCPM1	RCPM2	RCPM-SMOTE	Original	RCPM1	RCPM2	RCPM-SMOTE	Original	RCPM1	RCPM2	RCPM-SMOTE
Avila	Class [1]	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
	Class [2]	0,738	0,585	0,688	0,686	0,615	0,705	0,705	0,756	0,671	0,639	0,696	0,719
	Class [9]	0,800	0,684	0,737	0,833	0,857	0,929	1,000	0,893	0,828	0,788	0,849	0,862
Balance	Class [1]	0,091	0,156	0,130	0,100	0,100	0,250	0,150	0,100	0,095	0,192	0,139	0,100
Car	Class [2]	0,556	0,260	0,425	0,607	0,652	0,826	0,736	0,739	0,600	0,396	0,539	0,667
	Class [3]	0,654	0,389	0,500	0,667	0,739	0,913	0,913	0,696	0,694	0,546	0,646	0,681
KDD	Class [1]	0,500	0,324	0,588	0,611	0,538	0,846	0,769	0,864	0,518	0,469	0,666	0,716
	Class [2]	0,667	0,667	0,667	1,000	0,500	0,500	0,500	0,500	0,572	0,572	0,572	0,667
	Class [3]	0,952	0,357	0,800	0,950	0,952	0,952	0,952	0,905	0,952	0,519	0,869	0,927
	Class [4]	0,714	0,714	0,714	0,625	0,833	0,833	0,833	0,833	0,769	0,769	0,769	0,714
	Class [6]	1,000	1,000	1,000	1,000	0,714	1,000	0,714	0,714	0,833	1,000	0,833	0,833
	Class [7]	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	Class [8]	0,250	0,250	0,250	0,250	0,200	0,200	0,200	0,200	0,222	0,222	0,222	0,222
	Class [12]	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500
	Class [13]	1,000	0,333	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,500	1,000	1,000
	Class [16]	0,000	0,020	0,000	0,000	0,000	0,167	0,000	0,000	0,000	0,036	0,000	0,000
	Class [22]	1,000	0,714	1,000	0,714	0,833	0,833	0,833	0,833	0,909	0,769	0,909	0,769
Page-Blocks	Class [2]	0,615	0,321	0,615	0,867	0,615	0,692	0,615	1,000	0,615	0,439	0,615	0,929
	Class [3]	0,577	0,410	0,577	0,457	0,789	0,842	0,789	0,842	0,667	0,551	0,667	0,592
	Class [4]	0,391	0,183	0,357	0,257	0,625	0,775	0,625	0,725	0,481	0,296	0,454	0,379
Shuttle	Class [1]	0,929	0,006	0,542	0,382	1,000	1,000	1,000	1,000	0,963	0,012	0,703	0,553
	Class [2]	0,609	0,008	0,214	0,887	0,889	1,000	0,889	0,837	0,723	0,016	0,345	0,861
	Class [5]	nan	0,000	nan	nan	0,000	0,000	0,000	0,000	nan	0,000	0,000	0,000
	Class [6]	0,500	0,667	0,500	0,667	0,500	1,000	0,500	1,000	0,500	0,800	0,500	0,800
Yeast	Class [3]	0,800	0,750	0,692	0,889	0,667	0,750	0,750	0,667	0,727	0,750	0,720	0,762
	Class [4]	0,348	0,250	0,350	0,229	0,444	0,389	0,389	0,444	0,390	0,304	0,368	0,302
	Class [6]	0,500	0,333	0,417	0,333	0,462	0,385	0,385	0,385	0,480	0,357	0,400	0,357
	Class [7]	0,000	0,056	0,000	0,000	0,000	0,200	0,000	0,000	0,000	0,088	0,000	0,000
	Class [8]	0,364	0,267	0,364	0,176	0,800	0,800	0,800	0,600	0,500	0,400	0,500	0,272
	Class [9]	1,000	1,000	1,000	1,000	1,000	0,000	1,000	1,000	1,000	0,000	1,000	1,000
Page-Blocks 0	positive	0,615	-	0,571	0,562	0,764	-	0,815	0,809	0,681	-	0,672	0,663
Segment 0	positive	0,924	-	0,761	0,904	0,966	-	0,977	0,966	0,945	-	0,856	0,934
Shuttle: c0 vs c4	positive	1,000	-	0,965	1,000	1,000	-	1,000	1,000	1,000	-	0,982	1,000
Vowel 0	positive	1,000	-	0,833	1,000	1,000	-	1,000	1,000	1,000	-	0,909	1,000
Wine Quality Red 4	positive	0,117	-	0,117	0,068	0,500	-	0,500	0,333	0,190	-	0,190	0,113

4.4 Σύγκριση των Αποτελεσμάτων

Σε αυτήν την ενότητα θα συγκεντρώσουμε και θα συγκρίνουμε τα αποτελέσματα των πειραμάτων όλων των μεθόδων μεταξύ τους. Ακολουθεί μία σειρά πινάκων σύγκρισης για το ποσοστό μείωσης, την ακρίβεια, την ορθότητα, την ευαισθησία και το F-Measure.

Η λογική με την οποία συγκρίθηκαν τα αποτελέσματα έχει ως εξής: για κάθε σύνολο δεδομένων, υπολογίστηκαν οι μέσοι όροι της κάθε μεθόδου. Θυμίζουμε ότι οι μέθοδοι αυτοί είναι: η αρχική μείωση δεδομένων χωρίς να ξεχωρίζουμε τις σπάνιες από τις κοινές κλάσεις (Original), η RCPM1, η RCPM2, και τέλος η RCPM-SMOTE. Για κάθε σύγκριση, η μέθοδος με τον μεγαλύτερο μέσο όρο θεωρήθηκε και η πιο αποτελεσματική. Οι αριθμοί επομένως συμβολίζουν σε πόσα σύνολα δεδομένων η κάθε μέθοδος είχε μεγαλύτερο μέσο όρο. Θυμίζουμε πως το RCPM1 δεν μπορεί να χρησιμοποιηθεί με δυαδικά σύνολα δεδομένων, επομένως σε κάθε σύγκριση της συγκεκριμένης μεθόδου με τις άλλες το άθροισμα των συνόλων δεδομένων είναι επτά, δηλαδή όσο και τα σύνολα πολλαπλών κλάσεων αντί για δώδεκα, δηλαδή τον συνολικό αριθμό των συνόλων που χρησιμοποιήθηκαν. Επίσης, σε ορισμένες περιπτώσεις έχουμε το ίδιο αποτέλεσμα για δύο μεθόδους. Αυτές εισάγονται στην στήλη 'Ισοπαλίες'. Οι συγκρίσεις αυτού το είδους πραγματοποιήθηκαν και για τα πέντε προαναφερόμενα κριτήρια (ποσοστό μείωσης, ακρίβεια, ορθότητα ευαισθησία, F-Measure) και τα αποτελέσματα φαίνονται στους Πίνακες 4.19 - 4.23.

Αναλύοντας αυτούς τους πίνακες μπορούμε να εξάγουμε τις παρακάτω πληροφορίες: με μια πρώτη ματιά, όσο αφορά το ποσοστό μείωσης, την ακρίβεια καθώς και την ορθότητα των σπάνιων κλάσεων, η Original φαίνεται να είναι η πιο αποτελεσματική μέθοδος. Ωστόσο, οι τρεις RCPM μέθοδοι καταφέρνουν όλες τους να βελτιώσουν το ποσοστό ευαισθησίας σε

Πίνακας 4.19: Σύγκριση Αποτελεσμάτων - Ποσοστό Μείωσης

Μέθοδοι	Σκορ (Αριστερά)	Σκορ (Δεξιά)	Ισοπαλίες
Original vs RCPM1	5	2	0
Original vs RCPM2	12	0	0
Original vs RCPM-SMOTE	12	0	0
RCPM1 vs RCPM2	7	0	0
RCPM1 vs RCPM-SMOTE	7	0	0
RCPM2 vs RCPM-SMOTE	7	5	0

Πίνακας 4.20: Σύγκριση Αποτελεσμάτων - Ακρίβεια

Μέθοδοι	Σκορ (Αριστερά)	Σκορ (Δεξιά)	Ισοπαλίες
Original vs RCPM1	7	0	0
Original vs RCPM2	8	1	3
Original vs RCPM-SMOTE	9	1	2
RCPM1 vs RCPM2	0	7	0
RCPM1 vs RCPM-SMOTE	3	4	0
RCPM2 vs RCPM-SMOTE	7	4	1

Πίνακας 4.21: Σύγκριση Αποτελεσμάτων - Ορθότητα

Μέθοδοι	Σκορ (Αριστερά)	Σκορ (Δεξιά)	Ισοπαλίες
Original vs RCPM1	6	1	0
Original vs RCPM2	8	3	1
Original vs RCPM-SMOTE	8	4	0
RCPM1 vs RCPM2	0	7	0
RCPM1 vs RCPM-SMOTE	2	5	0
RCPM2 vs RCPM-SMOTE	6	6	0

Πίνακας 4.22: Σύγκριση Αποτελεσμάτων - Ευαισθησία

Μέθοδοι	Σκορ (Αριστερά)	Σκορ (Δεξιά)	Ισοπαλίες
Original vs RCPM1	0	7	0
Original vs RCPM2	2	9	1
Original vs RCPM-SMOTE	0	9	3
RCPM1 vs RCPM2	5	2	0
RCPM1 vs RCPM-SMOTE	5	2	0
RCPM2 vs RCPM-SMOTE	4	7	1

Πίνακας 4.23: Σύγκριση Αποτελεσμάτων - F-Measure

Μέθοδοι	Σκορ (Αριστερά)	Σκορ (Δεξιά)	Ισοπαλίες
Original vs RCPM1	6	1	0
Original vs RCPM2	9	3	0
Original vs RCPM-SMOTE	7	5	0
RCPM1 vs RCPM2	1	6	0
RCPM1 vs RCPM-SMOTE	2	5	0
RCPM2 vs RCPM-SMOTE	4	8	0

σχέση με την αρχική μέθοδο. Στην τελική όμως, παρατηρώντας τον πίνακα του F-Measure καταλήγουμε στο συμπέρασμα πως στις περισσότερες περιπτώσεις η μείωση της ορθότητας υπερβαίνει την αύξηση της ευαισθησίας.

Στην συνέχεια, συγκρίνοντας το ποσοστό μείωσης μόνο για τις μεθόδους RCPM μεταξύ τους, βλέπουμε πως η RCPM1 πετυχαίνει το μεγαλύτερο ποσοστό μείωσης. Επιπλέον, στην σύγκριση μεταξύ της RCPM2 και της RCPM-SMOTE, αξίζει να σημειωθεί πως για τα περισσότερα σύνολα δεδομένων πολλαπλών κλάσεων κέρδισε το πρώτο, ενώ για τα περισσότερα δυαδικά σύνολα κέρδιζε το δεύτερο.

Αν εξαιρέσουμε την Original πάνω στο θέμα της ακρίβειας και συγκρίνουμε μόνο τις άλλες τρεις, θα δούμε πως η RCPM2 δίνει τα καλύτερα αποτελέσματα μεταξύ τους. Όπως έγινε και με το ποσοστό μείωσης, οι περισσότερες νίκες της RCPM-SMOTE κατά της RCPM2 στην περίπτωση της ακρίβειας ήταν πάνω σε δυαδικά σύνολα.

Όσο αφορά την ευαισθησία, όπως αναφέραμε είδη και στις τρεις μεθόδους παρατηρείται μία βελτίωση, η μέθοδος όμως με τις περισσότερες 'νίκες' είναι η RCPM1. Εκ των υστέρων, φαίνεται και οι τρεις μέθοδοι μειώνουν την ορθότητα και αυξάνουν την ευαισθησία, μόνο που το RCPM1 το κάνει σε μεγαλύτερο βαθμό. Αυτό είναι κάτι το θετικό, καθώς η αύξηση της ευαισθησίας σημαίνει πως υπό τις κατάλληλες συνθήκες μειώνεται ο κίνδυνος να μην προβλέψουμε κάποιον κίνδυνο. Για παράδειγμα σε ένα σύστημα σεισμών όπου η κλάση "Σεισμός", δηλαδή η πρόβλεψη πως θα γίνει όντως σεισμός, είναι σπάνια τότε αν η ευαισθησία που υπολογίζεται είναι χαμηλή, αυτό σημαίνει πως πολλά αντικείμενα της κλάσης "Σεισμός" κατηγοριοποιήθηκαν στην κλάση "Όχι Σεισμός".

Μία τελευταία παρατήρηση, είναι πως αφού μέσω των μεθόδων RCPM διατηρούμε όλα τα αντικείμενα των σπάνιων κλάσεων, αποφεύγεται ο κίνδυνος της πιθανής εξάλειψης των ελάχιστων αυτών δεδομένων.

4.5 Επίλογος

Σε αυτό το κεφάλαιο αρχικά παρουσιάστηκαν τα σύνολα δεδομένων που χρησιμοποιήθηκαν στην πειραματική μελέτη η οποία εκπονήθηκε στα πλαίσια της παρούσας εργασίας. Στην συνέχεια παρουσιάστηκαν τα αποτελέσματα των πειραμάτων. Μέσω των μεθόδων μας, καταφέραμε με επιτυχία να διατηρήσουμε τα αντικείμενα των σπάνιων κλάσεων. Στην συνέχεια σχολιάστηκαν και συγκρίθηκαν τα αποτελέσματα μεταξύ τους. Παρατηρήσαμε πως αν και για τις τρεις μεθόδους RCPM το ποσοστό μείωσης, η ακρίβεια και η ορθότητα μειώνονται σε σχέση με την κλασική προσέγγιση στην μείωση δεδομένων, η ευαισθησία αυξάνεται. Κάτι το οποίο είναι θετικό, καθώς η ευαισθησία αποτελεί και το πιο κύριο κριτήριο που πρέπει να προσέχουμε σε θέματα εξαιρετικής σημαντικότητας όπως για παράδειγμα η πρόβλεψη ακραίων καιρικών φαινομένων.

5 Συμπεράσματα και Μελλοντική Έρευνα

Στόχος αυτής της πτυχιακής εργασίας ήταν η ανάπτυξη τεχνικών που μπορούν να περιορίσουν το πρόβλημα της περαιτέρω μείωσης των αντικειμένων που ανήκουν σε σπάνιες κλάσεις σε σύνολα δεδομένων με μεγάλη ανισοκατανομή κλάσεων από τεχνικές μείωσης των δεδομένων, καθώς και η σύγκριση αυτών των τεχνικών με την κλασική κατηγοριοποίηση συνόλων δεδομένων μέσω αλγορίθμων μείωσης δεδομένων. Εκτός αυτών, η μέθοδος υπερδειγματοληψίας SMOTE προστέθηκε επίσης στις μεθόδους που συγκρίναμε στα πειράματα μας.

Μέσω αυτών των πειραμάτων συγκρίναμε τα ποσοστά της ορθότητας και της ευαισθησίας, τα οποία μας ενδιαφέρουν πιο πολύ όσο αφορά τις σπάνιες κλάσεις, που υπολογίστηκαν από τις προτεινόμενες μεθόδους διατήρησης σπάνιων κλάσεων: τις RCPM1, RCPM2 και RCPM-SMOTE με τα αντίστοιχα αποτελέσματα που λαμβάνουμε από την κλασική κατηγοριοποίηση μέσω μείωσης δεδομένων. Τα αποτελέσματα μας έδειξαν πως οι μέθοδοι RCPM ουσιαστικά καταφέρνουν να μειώσουν το πρώτο και να αυξήσουν το δεύτερο από αυτά τα δύο κριτήρια. Τα επιπρόσθετα κριτήρια του ποσοστού μείωσης και της ακρίβειας επίσης μειώνονται ελαφρά όταν χρησιμοποιούμε μία από τις μεθόδους RCPM.

Επιπλέον, οι τρεις αυτές μέθοδοι δεν βλάπτουν τις σπάνιες κλάσεις. Επομένως, σαν τελικό συμπέρασμα μπορούμε να πούμε πως ο κίνδυνος της εξαφάνισης σημαντικών πληροφοριών που μπορούν να συγκρατούν οι σπάνιες κλάσεις έχει αντιμετωπιστεί με επιτυχία, αλλά ως συνέπεια όλα τα κριτήρια μέτρησης της αποτελεσματικότητας της κατηγοριοποίησης k-NN εκτός από την ευαισθησία έχουν ελαττωθεί.

Ωστόσο η αύξηση της ευαισθησίας, έστω με αντάλλαγμα μία μικρή πτώση στην ορθότητα, μπορεί να αποδείξει μία πολύ αξιοσημείωτη εξέλιξη, καθώς η ευαισθησία είναι ένα εξαιρετικά χρήσιμο κριτήριο σε περιπτώσεις μεγάλης σημαντικότητας, όπως είναι για παράδειγμα η χαλαζόπτωση, ή οι φυσικές καταστροφές σαν τον σεισμό και το τσουνάμι. Σε τέτοιες περιπτώσεις, αν έχουμε υπολογίσει χαμηλό ποσοστό για την ευαισθησία μπορεί να υποστούμε σοβαρές συνέπειες. Στο παράδειγμα με την χαλαζόπτωση ας πούμε, κάτι τέτοιο σημαίνει πως υπάρχουν πολλές προβλέψεις για "Μη Χαλαζόπτωση" που στην πραγματικότητα είναι "Χαλαζόπτωση" και άτομα που θα πρέπει να πάρουν μέτρα για να προστατευτούνε από το χαλάζι, όπως οι αγρότες, δεν θα το κάνουν γιατί δεν θα έχουν προειδοποιηθεί. Σε αυτές τις περιπτώσεις μπορούμε να δεχθούμε ελαφρώς χαμηλότερες τιμές ορθότητας αλλά σε καμία περίπτωση χαμηλές τιμές ευαισθησίας και αυτό επιτυγχάνεται από τις προτεινόμενες μεθόδους RCPM.

Εκτός αυτού, οι προτεινόμενες μέθοδοι RCPM είναι ικανές να χρησιμοποιηθούν και σε άλλα σενάρια τα οποία αφορούν σύνολα δεδομένων με πιο ισορροπημένη κατανομή κλάσεων, εφόσον ο σκοπός μας είναι να αυξήσουμε το ποσοστό της ευαισθησίας.

Ένα ακόμη συμπέρασμα της πειραματικής μελέτης της παρούσας εργασίας είναι το γεγονός ότι οι ΤΜΔ τελικά δεν μειώνουν τόσο πολύ τις τιμές ευαισθησίας όσο θα περιμέναμε. Συνεπώς, μπορούν να χρησιμοποιηθούν χωρίς την χρήση κάποιας μεθόδου διατήρησης αντικειμένων που ανήκουν σε σπάνιες κλάσεις όταν ελαφρώς χαμηλότερες τιμές ευαισθησίας μπορεί να επιτρέπονται σε συγκεκριμένα προβλήματα κατηγοριοποίησης.

Όπως είναι τώρα, οι τρεις μέθοδοι RCPM αποτελούν ένα επιπρόσθετο βήμα στην διαδικασία της μείωσης του μεγέθους και την κατηγοριοποίηση των αντικειμένων των συνόλων δεδομένων. Οι τεχνικές μείωσης δεδομένων από μόνες τους δεν βλάπτουν σοβαρά την απόδοση του καγηγοριοποιητή k-NN. Ωστόσο, μπορεί να βλάψουν τις σπάνιες κλάσεις. Επομένως, μία πιθανή κατεύθυνση για μελλοντικές έρευνες πάνω στο θέμα των μη-ισορροπημένων συνόλων δεδομένων θα μπορούσε να είναι η ανάπτυξη νέων τεχνικών μείωσης δεδομένων, οι οποίες μπορούν να προστατεύσουν τα αντικείμενα που ανήκουν σε σπάνιες κλάσεις χωρίς την βοήθεια κάποιας εξωτερικής μεθόδου διατήρησης δεδομένων.

Βιβλιογραφία

[1] M. James, *Classification algorithms*. New York, NY, USA: Wiley-Interscience, 1985.

[2] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, pp. 21–27, January 1967.

[3] B. V. Dasarathy, *Nearest neighbor (NN) norms : NN pattern classification techniques*. IEEE Computer Society Press, 1991.

[4] L. P. F. Garćia, A. C. P. L. F. de Carvalho, and A. C. Lorena, "Noisy data set identification," in *Hybrid Artificial Intelligent Systems* (J.-S. Pan, M. M. Polycarpou, M. Woźniak, A. C. P. L. F. de Carvalho, H. Quintián, and E. Corchado, eds.), (Berlin, Heidelberg), pp. 629–638, Springer Berlin Heidelberg, 2013.

[5] H. Brighton and C. Mellish, "Advances in instance selection for instance-based learning algorithms," *Data Min. Knowl. Discov.*, vol. 6, pp. 153–172, 04 2002.

[6] M. Grochowski and N. Jankowski, "Comparison of instance selection algorithms ii. results and comments," in *Artificial Intelligence and Soft Computing - ICAISC 2004* (L. Rutkowski, J. H. Siekmann, R. Tadeusiewicz, and L. A. Zadeh, eds.), (Berlin, Heidelberg), pp. 580–585, Springer Berlin Heidelberg, 2004.

[7] S. García, J. Derrac, J. Cano, and F. Herrera, "Prototype selection for nearest neighbor classification: Taxonomy and empirical study.," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, pp. 417–435, 01 2012.

[8] I. Triguero, J. Derrac, S. Garcia, and F. Herrera, "A taxonomy and experimental study on prototype generation for nearest neighbor classification," *Trans. Sys. Man Cyber Part C*, vol. 42, p. 86–100, Jan. 2012.

[9] M. Lozano, "Data reduction techniques in classification processes.." PhD Thesis, Universitat Jaume I, 2007.

[10] D. Wilson and T. Martinez, "Reduction techniques for instance-based learning algorithms," *Machine Learning*, vol. 38, pp. 257–286, 01 2000.

[11] Q. YANG and X. WU, "10 challenging problems in data mining research," *International Journal of Information Technology & Decision Making*, vol. 05, no. 04, pp. 597–604, 2006.

[12] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *J. Artif. Int. Res.*, vol. 16, p. 321–357, June 2002.

[13] A. Lorena, A. Carvalho, and J. Gama, "A review on the combination of binary classifiers in multiclass problems," *Artificial Intelligence Review*, vol. 30, pp. 19–37, 12 2008.

[14] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *Int J Data Warehousing and Mining*, vol. 2007, pp. 1–13, 2007.

[15] A. de Carvalho and A. A. Freitas, "A tutorial on multi-label classification techniques," 2009.

[16] V. Ganganwar, "An overview of classification algorithms for imbalanced datasets," *International Journal of Emerging Technology and Advanced Engineering*, vol. 2, pp. 42–47, 01 2012.

[17] M. Galar, A. Fernández, E. B. Tartas, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, pp. 463–484, 2012.

[18] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics," *Inf. Sci.*, vol. 250, pp. 113–141, 2013.

[19] M. Lin, K. Tang, and X. Yao, "Dynamic sampling approach to training neural networks for multiclass imbalance classification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, pp. 647–660, 2013.

[20] A. Fernández, V. López, M. Galar, M. J. del Jesus, and F. Herrera, "Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches," *Knowledge-Based Systems*, vol. 42, pp. 97 – 110, 2013.

[21] B. Zadrozny and C. Elkan, "Learning and making decisions when costs and probabilities are both unknown," *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 07 2001.

[22] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *SIGKDD Explor. Newsl.*, vol. 6, p. 20–29, June 2004.

[23] P. Domingos, "Metacost: A general method for making classifiers cost-sensitive," in *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '99, (New York, NY, USA), p. 155–164, Association for Computing Machinery, 1999.

[24] Yanminsun, A. Wong, and M. S. Kamel, "Classification of imbalanced data: a review," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 23, 11 2011.

[25] A. Estabrooks, T. Jo, and N. Japkowicz, "A multiple resampling method for learning from imbalanced data sets," *Computational Intelligence*, vol. 20, no. 1, pp. 18–36, 2004.

[26] Will Koehrsen, "Beyond accuracy: Precision and recall." `https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c`. [Published March 3, 2017].

[27] Andreas Klintberg, "Explaining precision and recall." `https://medium.com/@klintcho/explaining-precision-and-recall-c770eb9c69e9`. [Published May 22, 2017].

[28] N. Chawla, N. Japkowicz, and A. Kołcz, "Editorial: Special issue on learning from imbalanced data sets," *SIGKDD Explorations*, vol. 6, pp. 1–6, 06 2004.

[29] D. L, "Asymptotic properties of nearest neighbor rules using edited data," in *Cybern ., vol . SMC2, pp*, pp. 408–421, 1972.

[30] I. Tomek, "Two modifications of cnn," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-6, pp. 769–772, Nov 1976.

[31] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: One-sided selection," in *ICML*, 1997.

[32] J. Laurikkala, "Improving identification of difficult small classes by balancing class distribution," in *Artificial Intelligence in Medicine* (S. Quaglini, P. Barahona, and S. Andreassen, eds.), (Berlin, Heidelberg), pp. 63–66, Springer Berlin Heidelberg, 2001.

[33] H. Xiong, J. Wu, and L. Liu, "Classification with classoverlapping: A systematic study," 12 2010.

[34] B. Wang and N. Japkowicz, "Imbalanced data set learning with synthetic samples," *Proceedings of the IRIS Machine Learning Workshop*, 2004.

[35] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-smote: A new over-sampling method in imbalanced data sets learning," vol. 3644, pp. 878–887, 09 2005.

[36] C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap, "Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem," in *Advances in Knowledge Discovery and Data Mining* (T. Theeramunkong, B. Kijsirikul, N. Cercone, and T.-B. Ho, eds.), (Berlin, Heidelberg), pp. 475–482, Springer Berlin Heidelberg, 2009.

[37] H. He, Y. Bai, E. Garcia, and S. Li, "Adasyn: Adaptive synthetic sampling approach for imbalanced learning," pp. 1322 – 1328, 07 2008.

[38] J. Stefanowski and S. Wilk, "Selective pre-processing of imbalanced data for improving classification performance," in *Data Warehousing and Knowledge Discovery* (I.-Y. Song, J. Eder, and T. M. Nguyen, eds.), (Berlin, Heidelberg), pp. 283–292, Springer Berlin Heidelberg, 2008.

[39] J. Alcala-Fdez, L. Sanchez, S. García, M. J. Del Jesus, S. Ventura, J.-M. Garrell, J. Otero, C. Romero, J. Bacardit, V. Rivas Santos, J. C. Fernández, and F. Herrera, "Keel: A software tool to assess evolutionary algorithms for data mining problems," *Soft Comput.*, vol. 13, pp. 307–318, 02 2009.

[40] P. Hart, "The condensed nearest neighbor rule (corresp.)," *IEEE Transactions on Information Theory*, vol. 14, pp. 515–516, May 1968.

[41] D. W. Aha, "Tolerating noisy, irrelevant and novel attributes in instance-based learning algorithms," *International Journal of Man-Machine Studies*, vol. 36, no. 2, pp. 267 – 287, 1992. Symbolic problem solving in noisy and novel task environments.

[42] D. W. Aha, D. Kibler, and M. K. Albert, "Instance-based learning algorithms," *Mach. Learn.*, vol. 6, p. 37–66, Jan. 1991.

[43] J. Sánchez, "High training set size reduction by space partitioning and prototype abstraction," *Pattern Recognition*, vol. 37, p. 1561–1564, 07 2004.

[44] C. H. Chen and A. Józwik, "A sample set condensation algorithm for the class sensitive artificial neural network," *Pattern Recognit. Lett.*, vol. 17, pp. 819–823, 1996.

[45] S. Ougiaroglou, *Algorithms and Techniques for Efficient and Effective Nearest Neighbour Classification*. PhD thesis, 06 2014.

[46] S. Ougiaroglou and G. Evangelidis, "RHC: a non-parametric cluster-based data reduction for efficient k-nn classification," *Pattern Anal. Appl.*, vol. 19, no. 1, pp. 93–109, 2016.

[47] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, (Berkeley, Calif.), pp. 281–297, University of California Press, 1967.

[48] D. Dua and C. Graff, "UCI machine learning repository," 2017.

[49] C. De Stefano, F. Fontanella, M. Maniaci, and A. Scotto di Freca, "A method for scribe distinction in medieval manuscripts using page layout features," in *Image Analysis and Processing – ICIAP 2011* (G. Maino and G. L. Foresti, eds.), (Berlin, Heidelberg), pp. 393–402, Springer Berlin Heidelberg, 2011.

[50] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the kdd cup 99 data set," in *Proceedings of the Second IEEE International Conference on Computational Intelligence for Security and Defense Applications*, CISDA'09, p. 53–58, IEEE Press, 2009.