ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΕΥΦΥΕΙΣ ΤΕΧΝΟΛΟΓΙΕΣ ΔΙΑΔΙΚΤΥΟΥ – WEB INTELLIGENCE

Τεχνικές μείωσης του πληθυσμού των δεδομένων με ανεκτικότητα στις απούσες τιμές

(Data reduction techniques with missing values tolerance)

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Κουκάρα Πολυχρόνη
Α.Μ. : 14/2018

Επιβλέπων : Δρ. Στέφανος Ουγιάρογλου
Μέλος Ε.ΔΙ.Π, ΔΙ.ΠΑ.Ε.

ΔΙΕΘΝΕΣ ΠΑΝΕΠΙΣΤΗΜΙΟ ΤΗΣ ΕΛΛΑΔΟΣ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ ΕΥΦΥΕΙΣ ΤΕΧΝΟΛΟΓΙΕΣ ΔΙΑΔΙΚΤΥΟΥ – WEB INTELLIGENCE

Τεχνικές μείωσης του πληθυσμού των δεδομένων με ανεκτικότητα στις απούσες τιμές

(Data reduction techniques with missing values tolerance)

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Κουκάρα Πολυχρόνη

Επιβλέπων : Δρ. Στέφανος Ουγιάρογλου
Μέλος Ε.ΔΙ.Π, ΔΙ.ΠΑ.Ε.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή στις 4 Ιουλίου 2020.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Στέφανος Ουγιάρογλου Μέλος Ε.ΔΙ.Π, ΔΙ.ΠΑ.Ε.	Κωνσταντίνος Διαμαντάρας Καθηγητής ΔΙ.ΠΑ.Ε.	Δημήτριος Δέρβος Καθηγητής ΔΙ.ΠΑ.Ε.

Θεσσαλονίκη, Ιούλιος 2020

.....

(Υπογραφή)

*Η παρούσα εργασία αποτελεί πνευματική ιδιοκτησία του μεταπτυχιακού φοιτητή Πολυχρόνη Κουκάρα που την εκπόνησε. Στο πλαίσιο της πολιτικής ανοικτής πρόσβασης, ο συγγραφέας/δημιουργός εκχωρεί στο Διεθνές Πανεπιστήμιο της Ελλάδος άδεια χρήσης του δικαιώματος αναπαραγωγής, δανεισμού, παρουσίασης στο κοινό και ψηφιακής διάχυσης της εργασίας διεθνώς, σε ηλεκτρονική μορφή και σε οποιοδήποτε μέσο, για διδακτικούς και ερευνητικούς σκοπούς, άνευ ανταλλάγματος. Η ανοικτή πρόσβαση στο πλήρες κείμενο της εργασίας, δεν σημαίνει καθ' οποιονδήποτε τρόπο παραχώρηση δικαιωμάτων διανοητικής ιδιοκτησίας του συγγραφέα/δημιουργού, ούτε επιτρέπει την αναπαραγωγή, αναδημοσίευση, αντιγραφή, πώληση, εμπορική χρήση, διανομή, έκδοση, μεταφόρτωση (*downloading*), ανάρτηση (*uploading*), μετάφραση, τροποποίηση με οποιονδήποτε τρόπο, τμηματικά ή περιληπτικά της εργασίας, χωρίς τη ρητή προηγούμενη έγγραφη συναίνεση του συγγραφέα/δημιουργού. Η έγκριση της μεταπτυχιακής 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.*

Πολυχρόνης Κουκάρας

Μηχανικός Πληροφορικής Α.Τ.Ε.Ι.Θ.

Ευχαριστίες

Κατά τη διάρκεια των μεταπτυχιακών μου σπουδών και ιδιαίτερα κατά την υλοποίηση αυτής της μεταπτυχιακής εργασίας μού δόθηκε η ευκαιρία να αποκτήσω νέες γνώσεις και ιδέες σε μια περιοχή αρκετά εξελισσόμενη, αυτή της Εξόρυξης Δεδομένων. Για αυτό το λόγο θα ήθελα να ευχαριστήσω τους ανθρώπους που με βοήθησαν σε αυτή την προσπάθειά μου.

Αρχικά, θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου κύριο Στέφανο Ουγιάρογλου για την εμπιστοσύνη που μου έδειξε αλλά και για την ουσιαστική βοήθειά του με τις πολύτιμες συμβουλές και κατευθύνσεις που μου έδινε καθ' όλη τη διάρκεια της μεταπτυχιακής μου εργασίας.

Επίσης, θα ήθελα να ευχαριστήσω τους καθηγητές κυρίους Κωνσταντίνο Διαμαντάρα και Δημήτριο Δέρβο για την τιμή που μου έκαναν να διαβάσουν την μεταπτυχιακή μου εργασία και να συμμετάσχουν στην επιτροπή αξιολόγησης αυτής.

Επίσης, τους ευχαριστώ για την καλή συνεργασία που είχαμε τόσο κατά την διάρκεια των σπουδών μου όσο και κατά την διάρκεια εκπόνησης της εργασίας μου καθώς και για τη βοήθεια που μου προσέφεραν, όποτε τους τη ζήτησα, και για το ότι μου έδωσαν την ευκαιρία να ασχοληθώ με αυτόν τον ενδιαφέροντα τομέα της Εξόρυξης Δεδομένων.

Τέλος, οφείλω να ευχαριστήσω όλους εκείνους τους ανθρώπους, οι οποίοι, αν και δεν χρειάστηκε να κάνουν κάτι για αυτή την εργασία, παρ' όλα αυτά στάθηκαν δίπλα μου και με στήριξαν με περίσσεια υπομονή και αγάπη σε όλη μου την πορεία μέχρι σήμερα. Επίσης, ένα μεγάλο ευχαριστώ οφείλω στους γονείς μου, που με στηρίζουν όλα αυτά τα χρόνια μέχρι σήμερα, στα παιδιά μου Γιώργο και Ράνια για την υπομονή τους όλες αυτές τις ώρες που στερήθηκαν την παρουσία μου, καθώς και στην σύντροφό μου Αλεξάνδρα για την υπομονή, την ηθική στήριξη και την εμπιστοσύνη που έδειξε προς εμένα.

Πολυχρόνης Κουκάρας

Θεσσαλονίκη, Ιούλιος 2020

Περίληψη

Τα τελευταία χρόνια, μεγάλες ποσότητες δεδομένων εκπαίδευσης γίνονται καθημερινά διαθέσιμες από διάφορες πηγές. Αυτές οι ποσότητες, συνήθως δεν είναι δυνατό να χρησιμοποιηθούν από τους αλγόριθμους κατηγοριοποίησης εξαιτίας του υψηλού υπολογιστικού κόστους καθώς και των υψηλών απαιτήσεων αποθήκευσης στη μνήμη. Συνεπώς, συχνά τα δεδομένα αυτά προ-επεξεργάζονται από τεχνικές μείωσης του πληθυσμού των δεδομένων εκπαίδευσης (Data Reduction Techniques) με στόχο τη μείωση του υπολογιστικού κόστους αλλά και των απαιτήσεων σε μνήμη. Πολλές τεχνικές μείωσης του πληθυσμού των δεδομένων έχουν προταθεί και είναι διαθέσιμες στη βιβλιογραφία. Οι τεχνικές αυτές αφορούν κυρίως τον κατηγοριοποιητή των κ εγγύτερων γειτόνων (k Nearest Neighbor classifier). Ωστόσο, οι τεχνικές αυτές δεν μπορούν να διαχειριστούν τις απούσες τιμές (Missing Values) που σχεδόν πάντα εμφανίζονται στα πραγματικά σύνολα δεδομένων εκπαίδευσης. Έτσι, πριν από την προ-επεξεργασία από μια τεχνική μείωσης του πληθυσμού των δεδομένων, είναι απαραίτητη η εφαρμογή ενός ακόμη βήματος προ-επεξεργασίας για την συμπλήρωση των απουσών τιμών (Missing Values Imputation). Στη βιβλιογραφία, συναντάμε διάφορες τέτοιες μεθόδους και η παρούσα εργασία παρουσιάζει τις σημαντικότερες. Ωστόσο, η εφαρμογή ενός ακόμη βήματος προ-επεξεργασίας είναι ένα σημαντικό μειονέκτημα που προσθέτει υπολογιστικό κόστος. Αυτό αποτελεί το κίνητρο εκπόνησης της παρούσας διπλωματικής εργασίας. Η παρούσα εργασία προτείνει μια νέα παραλλαγή μιας τεχνικής μείωσης του πληθυσμού των δεδομένων που μπορεί να διαχειριστεί τις απούσες τιμές χωρίς να απαιτείται το επιπρόσθετο βήμα προ-επεξεργασίας για την συμπλήρωση τους. Η τεχνική αυτή είναι ένας αλγόριθμος παραγωγής προτύπων (Prototype Generation) και ονομάζεται αλγόριθμος Επεξεργασίας και Μείωσης μέσω Ομοιογενών συστάδων (Editing and Reduction through Homogeneous Clusters – ERHC). Η νέα παραλλαγή του ERHC διαχειρίζεται τις απούσες τιμές αξιοποιώντας την τεχνική της μερικής ευκλείδειας απόστασης (partial distance) και εφαρμόζοντας συσταδοποίηση κ-μέσων (k-means) που δεν λαμβάνει υπ' όψιν τις απούσες τιμές. Επιπρόσθετα, η απόδοση του ERHC ελέγχθηκε αφού πρώτα οι απούσες τιμές συμπληρώθηκαν από τη μέθοδο της συμπλήρωσης του μέσου όρου του χαρακτηριστικού ανά κλάση. Οι δύο προαναφερθέντες ERHC αλγόριθμοι συγκρίνονται μεταξύ τους αλλά και με τον αλγόριθμο των κ εγγύτερων γειτόνων χωρίς μείωση του πληθυσμού των δεδομένων εκτελώντας πειράματα σε 13 σύνολα δεδομένων και εκτιμώντας την ακρίβεια κατηγοριοποίησης και το λόγο μείωσης (Reduction Rate) που επιτυγχάνουν οι δύο ERHC αλγόριθμοι. Τα πειραματικά αποτελέσματα δείχνουν αξιοσημείωτη απόδοση και για τις δύο παραλλαγές του ERHC αλγορίθμου.

Λέξεις Κλειδιά: τεχνικές μείωσης του πληθυσμού των δεδομένων εκπαίδευσης, Data Reduction Techniques, κατηγοριοποιητής κ εγγύτερων γειτόνων, συμπλήρωση απουσών τιμών, Missing Values Imputation, καταλογισμός, αλγόριθμος Επεξεργασίας και Μείωσης μέσω Ομοιογενών συστάδων, ERHC, συσταδοποίηση κ-μέσων, k-means, μερική ευκλείδεια απόσταση, partial distance.

Abstract

In recent years, large amounts of training data, from various sources, become available on a daily basis. These quantities are usually not possible to be used by classification algorithms due to the high cost of computing as well as the high memory storage requirements. Therefore, this data is often pre-processed by Data Reduction Techniques in order to reduce computing costs and memory requirements. Many data reduction techniques have been proposed and are available in the literature. These techniques mainly concern the 'k Nearest Neighbor classifier'. However, these techniques cannot manage the Missing Values that always appear in real training data sets. Thus, before pre-processing by a data reduction technique, it is necessary to apply another pre-processing step to complete the Missing Values Imputation. In the literature, we come across to several such methods and this paper presents the most important ones. However, by applying an extra pre-processing step is a major drawback that adds computational cost. This is the motivation for this thesis. This thesis proposes a new variant of a data reduction technique that can manage missing values without requiring the additional pre-processing step for data imputation. This technique is a Prototype Generation algorithm and is called the Editing and Reduction through Homogeneous Clusters (ERHC) algorithm. The new ERHC variant manages the missing values using the partial distance technique and applying k-means clustering that does not take into account the missing values. In addition, the performance of ERHC has been tested after the imputation of missing values by the average per class imputation method. The two aforementioned ERHC variants are compared to each other and to the algorithm of the nearest neighbors without reducing the population of data by performing experiments on 13 data sets and estimating the accuracy of classification and reduction ratio (Reduction Rate) achieved by the two ERHC algorithms. The experimental results show remarkable performance for both variants of the ERHC algorithm.

Keywords: Data Reduction Techniques, Categorization of Neighboring Neighbors, Incomplete Pricing, Missing Values Imputation, Calculation, Processing and Reduction Algorithm through Homogeneous Clusters, ERHC, k-means Clustering, partial distance.

Πίνακας περιεχομένων

Πίνακας εικόνων

Πίνακας πινάκων

1

Εισαγωγικές έννοιες

Στο κεφάλαιο αυτό γίνεται μια εισαγωγή στην διαδικασία της κατηγοριοποίησης (classification), η οποία εμφανίζεται σε πολλά ερευνητικά πεδία της Επιστήμης των Υπολογιστών. Αρχικά, παρουσιάζονται κάποια εισαγωγικά θέματα γύρω από την έννοια και τις τεχνικές της κατηγοριοποίησης, ενώ στην συνέχεια παρουσιάζεται ο κατηγοριοποιητής των κ εγγύτερων γειτόνων (k-NN) ο οποίος χρησιμοποιείται ευρέως από τις περισσότερες τεχνικές μείωσης του πληθυσμού των δεδομένων εκπαίδευσης. Ακολουθεί μία σύντομη αναφορά στις τεχνικές μείωσης του πληθυσμού των δεδομένων εκπαίδευσης και στο τέλος παρουσιάζονται το κίνητρο που οδήγησε στην εκπόνηση της παρούσας διπλωματικής εργασίας, η συνεισφορά της, καθώς και η δομή που ακολουθείται στα επόμενα κεφάλαιά της.

1.1 Κατηγοριοποίηση

Η κατηγοριοποίηση (classification) [1] είναι μια από τις βασικότερες εργασίες της Εξόρυξης δεδομένων, με μεγάλο αριθμό εφαρμογών τόσο στον ακαδημαϊκό χώρο, όσο και στον χώρο της οικονομίας και της βιομηχανίας [2][3]. Ο όρος κατηγοριοποίηση συναντάται στη βιβλιογραφία και ως ταξινόμηση και είναι μία τεχνική, κατά την οποία ένα στοιχείο ανατίθεται σε ένα προκαθορισμένο σύνολο κατηγοριών. Γενικότερα, ο στόχος της

διαδικασίας αυτής είναι η ανάπτυξη ενός μοντέλου, το οποίο αργότερα θα μπορεί να χρησιμοποιηθεί για την κατηγοριοποίηση μελλοντικών δεδομένων.

Ο σκοπός των τεχνικών κατηγοριοποίησης είναι η ομαδοποίηση δεδομένων στην απαιτούμενη δομή με βάση κοινά χαρακτηριστικά. Η κατηγοριοποίηση καθιστά δυνατό τον προσδιορισμό εκείνων των δεδομένων των οποίων η συνδεδεμένη κλάση είναι άγνωστη. Οι μέθοδοι κατηγοριοποίησης ορίζονται ως μοντέλα που παράγουν διαφορετικά αποτελέσματα. Όλες αυτές οι μέθοδοι περιλαμβάνουν ανάλυση και κατηγοριοποίηση βάσει των στιγμιότυπων στο σύνολο δεδομένων εκπαίδευσης [4]. Η τεχνική κατηγοριοποίησης (ή αλλιώς αλγόριθμοι ταξινόμησης ή ταξινομητές), χρησιμοποιείται σε πολλές περιοχές ενδιαφερόντων. Τέτοια παραδείγματα συναντάμε στην ιατρική με την πρόβλεψη καρκινικών κυττάρων χαρακτηρίζοντάς τα ως καλοήθη ή κακοήθη, στην οικονομία με την κατηγοριοποίηση των πελατών μιας τράπεζας ανάλογα με την πιστωτική τους ικανότητα, στην βιομηχανία με τη βελτιστοποίηση και την ποιοτική ανάλυση των διαδικασιών παραγωγής σε διάφορες ενεργειακές εφαρμογές, τον προσδιορισμό της αντοχής των δομικών υλικών ακόμη και στον διαχωρισμό των μηνυμάτων ηλεκτρονικού ταχυδρομείου με βάση την επικεφαλίδα τους ή το περιεχόμενό τους είτε σε κατηγορίες τύπου "spam" ή κατηγορία "μη-spam". Σε πολλές περιπτώσεις λοιπόν, μας ενδιαφέρει μια απάντηση του τύπου ναι ή όχι οπότε μιλάμε και για δυαδικό πρόβλημα κατηγοριοποίησης. Αλλά οι κατηγορίες δεν είναι πάντοτε δύο, οπότε μιλάμε πλέον για προβλήματα πολλαπλών κατηγοριών (multi-class classification).

Η κατηγοριοποίηση μπορεί να περιγραφεί ως μία διαδικασία δύο κύριων βημάτων:

1. **Εκμάθηση** (Learning): Στο πρώτο βήμα της διαδικασίας δημιουργείται/προσδιορίζεται το μοντέλο κατηγοριοποίησης [5] με βάση ένα σύνολο προ-κατηγοριοποιημένων στιγμιότυπων, που ονομάζονται δεδομένα εκπαίδευσης (training data). Τα δεδομένα εκπαίδευσης αναλύονται από ένα αλγόριθμο κατηγοριοποίησης, προκειμένου να σχηματιστεί το μοντέλο. Λόγω του ότι τα δεδομένα εκπαίδευσης ανήκουν σε μία προκαθορισμένη κατηγορία, η οποία είναι γνωστή, η κατηγοριοποίηση αποτελεί μέθοδος εποπτευομένης μάθησης (supervised learning). Το μοντέλο, που λέγεται και αλλιώς κατηγοριοποιητής (classifier), αναπαρίσταται με τη μορφή κανόνων κατηγοριοποίησης (classification rules), δέντρων απόφασης (decision trees) ή μαθηματικών τύπων.
2. **Κατηγοριοποίηση** (Classification) [1]: Μετά την δημιουργία του μοντέλου, το επόμενο βήμα είναι η αξιολόγησή του. Για να επιτευχθεί αυτό, χρησιμοποιούμε τα δοκιμαστικά δεδομένα (test data) για να υπολογίσουν την ακρίβεια του μοντέλου. Το μοντέλο κατηγοριοποιεί τα δοκιμαστικά δεδομένα. Έπειτα, η κατηγορία που σχηματίστηκε με βάση τα δοκιμαστικά δεδομένα συγκρίνεται με την πρόβλεψη που έγινε για τα δεδομένα εκπαίδευσης, τα οποία είναι ανεξάρτητα από αυτά της δοκιμής. Η εκτίμηση της

απόδοσης ενός μοντέλου-αλγορίθμου κατηγοριοποίησης βασίζεται στο πλήθος των ορθών και των εσφαλμένων προβλέψεών του, τα οποία μπορούν να εκφραστούν με τον πίνακα συσχέτισης ή το ρυθμό σφάλματος.

Στην περίπτωση που το μοντέλο κριθεί αποδεκτό, τότε μπορεί να χρησιμοποιηθεί για την κατηγοριοποίηση μελλοντικών δεδομένων, των οποίων η κατηγορία είναι άγνωστη.

Χωρίς να αναλυθεί λεπτομερώς ο τρόπος λειτουργίας τους, αναφέρονται παρακάτω κάποια παραδείγματα κατηγοριοποιητών:

- Ένας πολύ απλός αλλά αποδοτικός αλγόριθμος είναι αυτός των **k εγγύτερων γειτόνων** (k-NN) [11],[12], κατά τον οποίο ορίζουμε ένα μέτρο ομοιότητας ή απόστασης μεταξύ των στιγμιότυπων και θεωρούμε ότι αυτά τα στιγμιότυπα που βρίσκονται αρκετά κοντά μεταξύ τους θα μοιράζονται την ίδια κλάση, επομένως υπολογίζουμε τις αποστάσεις μεταξύ τους, βρίσκουμε τους k κοντινότερους γείτονες και συμβουλευόμαστε τη δική τους γνωστή κλάση έτσι ώστε να προβλέψουμε την κλάση του νέου δείγματος. Ο συγκεκριμένος αλγόριθμος δεν κατασκευάζει κάποιο μοντέλο κατηγοριοποίησης. Αντίθετα, το σύνολο δεδομένων εκπαίδευσης παίζει τον ρόλο του μοντέλου κατηγοριοποίησης.
- Τα **δένδρα απόφασης** [6],[13] είναι μια άλλη γενική κατηγορία τέτοιων αλγορίθμων που χτίζουν ένα δένδρο με τη βοήθεια του συνόλου εκπαίδευσης έτσι ώστε κάθε φύλλο να αντιστοιχεί σε μια από τις προκαθορισμένες κλάσεις. Μετά αρκεί να το διασχίσουν με βάση τα χαρακτηριστικά των νέων στιγμιότυπων, τα οποία και κατηγοριοποιούνται ανάλογα με το φύλλο του δένδρου στο οποίο θα καταλήξουν. Το μοντέλο δέντρου απόφασης ορίζεται επίσης ως μάθηση βασισμένη σε κανόνες [16].
- Η τεχνική κατηγοριοποίησης του **Bayes** [8] είναι μια μέθοδος υπολογισμού της πιθανότητας να συμπεριληφθούν νέα δεδομένα σε μία από τις τρέχουσες κατηγορίες με βάση τα υπάρχοντα, διαθέσιμα και ήδη κατηγοριοποιημένα δεδομένα [10]. Η ανεξαρτησία των δεδομένων θεωρείται προϋπόθεση για την επιτυχή κατηγοριοποίηση [17],[18]. Το θεώρημα του Bayes εισάγεται από τον βρετανό μαθηματικό Thomas Bayes και αναπτύσσεται περαιτέρω μετά το θάνατό του. Βασίζεται στην υπό όρους πιθανότητα. Δηλαδή, εάν η εμφάνιση ενός συμβάντος Β εξαρτάται από ένα γεγονός Α, τότε αυτή η κατάσταση μπορεί να εξηγηθεί με πιθανότητα υπό όρους. Εφαρμόζεται σε διάφορους τομείς, από την ιατρική έως την οικονομία, τις στατιστικές, την αρχαιολογία, το δίκαιο, τις επιστήμες της ατμόσφαιρας, στις δοκιμές και στην αξιολόγηση, τη φυσική και τη γενετική.
- Οι **μηχανές διανυσμάτων υποστήριξης** (SVM) κατηγοριοποιούν με τη βοήθεια γραμμικής ή μη γραμμικής λειτουργίας. Η μέθοδος της μηχανής διανυσμάτων υποστήριξης βασίζεται στην εκτίμηση της καταλληλότερης λειτουργίας για τον διαχωρισμό των δεδομένων [5]. Η μέθοδος SVM στοχεύει στην εύρεση μιας ειδικής

διαχωριστικής γραμμής που χωρίζει τις κλάσεις μεταξύ τους. Υπάρχει η δυνατότητα να δημιουργηθεί και να μετακινηθεί ανάλογα αυτή η γραμμή περισσότερες από μία φορές κατά τη διάρκεια της κατηγοριοποίησης. Η μηχανή διανυσμάτων υποστήριξης προσδιορίζει την πιο μακρινή γραμμή και στις δύο κλάσεις και έτσι καθορίζεται η μέγιστη ανοχή σφάλματος. Κατά τον προσδιορισμό των δεδομένων εκπαίδευσης και της συνοριακής γραμμής, τα δεδομένα δοκιμών κατηγοριοποιούνται με βάση τις θέσεις τους σε σχέση με τα όρια της γραμμής [7].

- Τα **τεχνητά νευρωνικά δίκτυα** [19][21] μπορούν επίσης να χρησιμοποιηθούν ως κατηγοριοποιητές. Αν και συνήθως έχουν πολύπλοκη αρχιτεκτονική και απαιτούν πολύπλοκη υλοποίηση έχουν δείξει πολύ καλά αποτελέσματα σε μια πληθώρα προβλημάτων και εφαρμογών.
- **Παραγωγή κανόνων κατηγοριοποίησης**. Η γνώση που αποκτούμε κατά την διαδικασία της κατηγοριοποίησης μπορεί να αναπαρασταθεί και με τη χρήση κανόνων κατηγοριοποίησης. Οι κανόνες κατηγοριοποίησης [16], σε σχέση με τα δέντρα απόφασης, γίνονται ευκολότερα κατανοητοί όταν το δέντρο που παράχθηκε είναι μεγάλο. Έτσι μπορούμε να μετατρέψουμε ένα δέντρο απόφασης σε ένα σύνολο κανόνων κατηγοριοποίησης. Αυτό μπορεί να επιτευχθεί εάν θεωρήσουμε ότι κάθε κανόνας αντιστοιχεί σε ένα μονοπάτι του δέντρου από τη ρίζα μέχρι ένα κόμβο-φύλλο. Άρα κάθε φύλλο παράγει ένα κανόνα. Οι συνθήκες που θα μας οδηγήσουν στο φύλλο (υπόθεση) αποτελούν το αριστερό μέρος του κανόνα, ενώ το φύλλο (αποτέλεσμα) αντιστοιχεί στο δεξιό μέρος του κανόνα.

Τα περισσότερα μοντέλα κατηγοριοποίησης που έχουμε αναφέρει παραπάνω (εξαιρουμένου του **k εγγύτερων γειτόνων**-k-NN) είναι όλα μοντέλα πρόθυμης όπως λέγεται μάθησης και αφορούν τους πρόθυμους κατηγοριοποιητές (eager classifiers). Αυτό σημαίνει ότι όταν δίνεται ένα σύνολο δεδομένων για εκπαίδευση θα κατασκευαστεί ένα γενικευμένο μοντέλο πριν ακόμη γίνουν δεκτά νέα στιγμιότυπα προς κατηγοριοποίηση. Μπορούμε να θεωρήσουμε ότι το μοντέλο είναι έτοιμο και πρόθυμο να κατηγοριοποιήσει τα άγνωστα μέχρι πριν στιγμιότυπα. Αντιθέτως στην οκνηρή προσέγγιση (lazy approach) περιμένουμε μέχρι την τελευταία στιγμή, πριν γίνει κατασκευή οποιουδήποτε μοντέλου, να κατηγοριοποιηθεί το σύνολο των δοκιμαστικών στιγμιότυπων. Ένας οκνηρός λοιπόν κατηγοριοποιητής (ή τεμπέλης όπως αλλιώς λέγεται - lazy classifier) απλά αποθηκεύει ένα στιγμιότυπο με γνωστή κατηγορία, ή κάνει μια πολύ μικρή επεξεργασία του και περιμένει μέχρι να λάβει ένα δοκιμαστικό στιγμιότυπο. Μόνο όταν δει το δοκιμαστικό στιγμιότυπο εφαρμόζει γενίκευση για να κατηγοριοποιηθεί με βάση την ομοιότητά του με τα αποθηκευμένα εκπαιδευμένα στιγμιότυπα.

Σε αντίθεση με τους πρόθυμους κατηγοριοποιητές, οι οκνηροί κατηγοριοποιητές κάνουν λιγότερη δουλειά όταν εμφανίζεται ένα στιγμιότυπο προς εκπαίδευση και περισσότερη όταν είναι να γίνει κατηγοριοποίηση ή αριθμητική πρόβλεψη. Επειδή λοιπόν οι οκνηροί κατηγοριοποιητές αποθηκεύουν τα στιγμιότυπα αναφέρονται και ως κατηγοριοποιητές βασισμένοι σε στιγμιότυπα, διότι σε αυτά βασίζεται όλη η διαδικασία της μάθησης. Για να γίνει μια κατηγοριοποίηση ή μια αριθμητική πρόβλεψη η χρήση οκνηρών κατηγοριοποιητών μπορεί να είναι υπολογιστικά ακριβή. Απαιτούν αποδοτικές τεχνικές αποθήκευσης και θα πρέπει να προσαρμόζονται σε εφαρμογές με υλικό που λειτουργεί με παραλληλοποίηση, ενώ δεν δίνουν πολλές λεπτομέρειες για την εσωτερική δομή των δεδομένων. Παρόλα αυτά όμως υποστηρίζουν από τη φύση τους τη στοιχειώδη εκπαίδευση. Μπορούν να μοντελοποιήσουν πολύπλοκους χώρους αποφάσεων που έχουν υπερπολυγωνικά σχήματα, τα οποία δεν είναι εύκολα να περιγραφούν από άλλους αλγόριθμους.

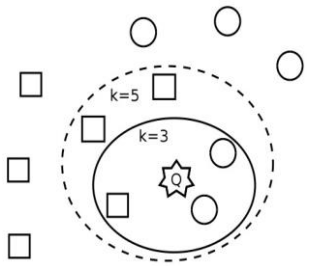
Έτσι λοιπόν, τα δένδρα αποφάσεων κατηγοριοποίησης [13] αποτελούν μια πολύ γνωστή υποκατηγορία πρόθυμων κατηγοριοποιητών. Άλλοι πρόθυμοι κατηγοριοποιητές βασίζονται σε τεχνητά νευρωνικά δίκτυα. Ένα χαρακτηριστικό παράδειγμα ενός πιθανοτικού κατηγοριοποιητή είναι ο αφελής κατηγοριοποιητής Bayes. Από την άλλη πλευρά, η κατηγορία των οκνηρών – "lazy" κατηγοριοποιητών περιλαμβάνει τον γνωστό κατηγοριοποιητή k Nearest Neighbors [11][12].

1.2 Κατηγοριοποίηση με βάση τους k εγγύτερους γείτονες (k-Nearest Neighbors)

Η πιο κοινή και απλή, βασισμένη σε στιγμιότυπα, μέθοδος κατηγοριοποίησης χρησιμοποιεί τον αλγόριθμο των k-εγγύτερων (ή κοντινότερων) γειτόνων [11][12] καθώς είναι και μη παραμετρική. Είναι ένας αποτελεσματικός και ευρέως χρησιμοποιούμενος οκνηρός αλγόριθμος κατηγοριοποίησης. Είναι ένας απλός και εύχρηστος κατηγοριοποιητής, καθώς μπορεί να χρησιμοποιηθεί σε πολλούς τομείς εφαρμογών και να ενσωματωθεί εύκολα σε πολλά συστήματα.

Ο αλγόριθμος αυτός υποθέτει ότι όλα τα στιγμιότυπα απεικονίζονται σε σημεία του n-διάστατου χώρου. Δεδομένου ότι ο κατηγοριοποιητής k-κοντινότερων γειτόνων (k-NN) είναι ένας οκνηρός κατηγοριοποιητής, συνεπώς, δεν δημιουργεί κανένα μοντέλο κατηγοριοποίησης αλλά παράγει τη βάση της γνώσης του αποθηκεύοντας όλα τα δεδομένα εκπαίδευσης.

Για την κατηγοριοποίηση ενός νέου στιγμιότυπου χρησιμοποιούνται τα αποθηκευμένα δεδομένα για να βρεθεί ένα συγκεκριμένο πλήθος (k) των πιο όμοιων στιγμιότυπων εκπαίδευσης (κοντινότεροι γείτονες), σύμφωνα με μια μετρική απόστασης. Οι εγγύτεροι γείτονες ενός στιγμιότυπου ορίζονται συνήθως με την έννοια της Ευκλείδειας απόστασης. Το νέο στιγμιότυπο ανατίθεται στην κατηγορία που είναι επικρατέστερη μεταξύ των κοντινότερων γειτόνων του. (Εικόνα 1). Ο αριθμός k είναι συνήθως ένας μικρός περιττός αριθμός (π.χ. 3 ή 5 ή 7). Στην περίπτωση που το k = 1, ο αλγόριθμος είναι επίσης γνωστός ως κατηγοριοποιητής πλησιέστερου γείτονα (ή κανόνας 1-NN).



Εικόνα 1: Καθορισμός κλάσης νέου στιγμιότυπου για k=3 και k=5

k-NN παράδειγμα:

Για k = 3, το σημείο αναζήτησης προσδιορίζεται στην κλάση "κύκλος"

Για k = 5, το σημείο αναζήτησης προσδιορίζεται στην κλάση "τετράγωνο"

Βασικό χαρακτηριστικό ενός αλγορίθμου κατηγοριοποίησης είναι η απόδοση του. Και αυτή φυσικά εξαρτάται από την τιμή του k που επιτυγχάνει την υψηλότερη ακρίβεια κατηγοριοποίησης και επηρεάζεται από το σύνολο δεδομένων που χρησιμοποιεί. Πειραματικές μετρήσεις έχουν δείξει ότι οι μεγαλύτερες τιμές k είναι κατάλληλες για μεγάλα σύνολα δεδομένων που περιέχουν ενδεχομένως και θόρυβο και δίνουν ικανοποιητική ακρίβεια, ενώ αντίθετα οι μικρές τιμές του k καθιστούν τον κατηγοριοποιητή περισσότερο ευαίσθητο στο θόρυβο και κατά συνέπεια μικρότερη ακρίβεια. Βέβαια ακόμα και η καλύτερη τιμή k μπορεί να μην είναι και η βέλτιστη για όλες τις περιοχές του χώρου δεδομένων.

Άλλος παράγοντας που επηρεάζει την απόδοση του κατηγοριοποιητή είναι η επιλογή της μέτρησης που θα χρησιμοποιηθεί για τους υπολογισμούς της απόστασης μεταξύ των στιγμιότυπων. Αναφέρθηκε προηγουμένως ότι θα χρησιμοποιηθεί η Ευκλείδεια απόσταση λόγω του τύπου δεδομένων των χαρακτηριστικών του συνόλου δεδομένων. Φυσικά, για να μπορεί να υπολογιστεί η Ευκλείδεια απόσταση, βασική προϋπόθεση είναι το σύνολο δεδομένων να μην περιέχει απούσες τιμές.

Όπως έχουμε πει ο κατηγοριοποιητής k-NN έχει το πλεονέκτημα ότι μπορεί πολύ εύκολα να εφαρμοστεί, είναι ανθεκτικός στα θορυβώδη δεδομένα εκπαίδευσης και μπορεί να

επιτυγχάνει υψηλότερη ακρίβεια εάν τα δεδομένα εκπαίδευσης είναι πολύ μεγάλα. Ωστόσο παρουσιάζει ορισμένα μειονεκτήματα όπως είναι ότι πρέπει πάντα να καθορίζεται από την αρχή η τιμή του k που μπορεί να είναι σύνθετη για ορισμένα σύνολα δεδομένων. Επίσης το υπολογιστικό κόστος είναι πολύ υψηλό έως απαγορευτικό ορισμένες φορές, λόγο του ότι χρειάζεται να υπολογίσει όλες οι αποστάσεις μεταξύ ενός μη κατηγοριοποιημένου στιγμιότυπου και των δεδομένων εκπαίδευσης, πράγμα το οποίο πολλές φορές γίνεται ασύμφορο για την ολοκλήρωση της διαδικασίας.

Ο τρόπος λειτουργίας του κατηγοριοποιητή k-NN θα μπορούσε να χαρακτηριστεί και αδυναμία γιατί υποδηλώνει υψηλές απαιτήσεις αποθήκευσης για τα δεδομένα εκπαίδευσης. Η βάση δεδομένων εκπαίδευσης πρέπει να είναι πάντα διαθέσιμη με αποτέλεσμα να δεσμεύεται μόνιμα μεγάλος αποθηκευτικός χώρος. Κατά συνέπεια ο κατηγοριοποιητής πρέπει να εκτελείται σε υπολογιστικά συστήματα εξοπλισμένα με μεγάλη κύρια μνήμη για να μπορούν να αποθηκεύονται τα δεδομένα εκπαίδευσης.

Μία τελευταία αδυναμία του κατηγοριοποιητή k-NN είναι το γεγονός ότι είναι ένας αλγόριθμος ευαίσθητος στον θόρυβο. Ο θόρυβος και τα λανθασμένα δεδομένα, καθώς και οι αποκλίσεις και οι αλληλοεπικαλύψεις μεταξύ περιοχών δεδομένων διαφορετικών κατηγοριών επηρεάζουν την ακρίβεια κατηγοριοποίησης με αποτέλεσμα μια λιγότερο ακριβή κατηγοριοποίηση.

1.3 Τεχνικές μείωσης του πληθυσμού των δεδομένων (Data Reduction Techniques)

Όπως είδαμε κάθε κατηγοριοποιητής έχει τα προτερήματα και τα μειονεκτήματα του. Έτσι λοιπόν και ο κατηγοριοποιητής k-NN έχει και αυτός κάποια μειονεκτήματα. Ένα βασικό μειονέκτημα θα μπορούσαμε να πούμε ότι είναι το υπολογιστικό του κόστος που απαιτεί και πώς αυτό μπορεί να μειωθεί. Έχουν γίνει πολλές έρευνες με τις οποίες προσπαθούν να αντιμετωπίσουν αυτά τα αδύνατα σημεία του.

Ένα κύριο χαρακτηριστικό είναι η ανάπτυξη **τεχνικών μείωσης του πληθυσμού των δεδομένων** (Data Reduction Techniques - DRT). Οι τεχνικές αυτές οδηγούν σε πιο ευέλικτη παρουσίαση των δεδομένων, καθώς μειώνεται κατά πολύ ο όγκος χωρίς να χάνεται η ακεραιότητα των αρχικών δεδομένων.

Οι τεχνικές αυτές μπορούν να χωριστούν σε δύο κατηγορίες : α) **αλγόριθμοι επιλογής προτύπων** (prototype selection algorithms) [14] και β) **αλγόριθμοι παραγωγής**

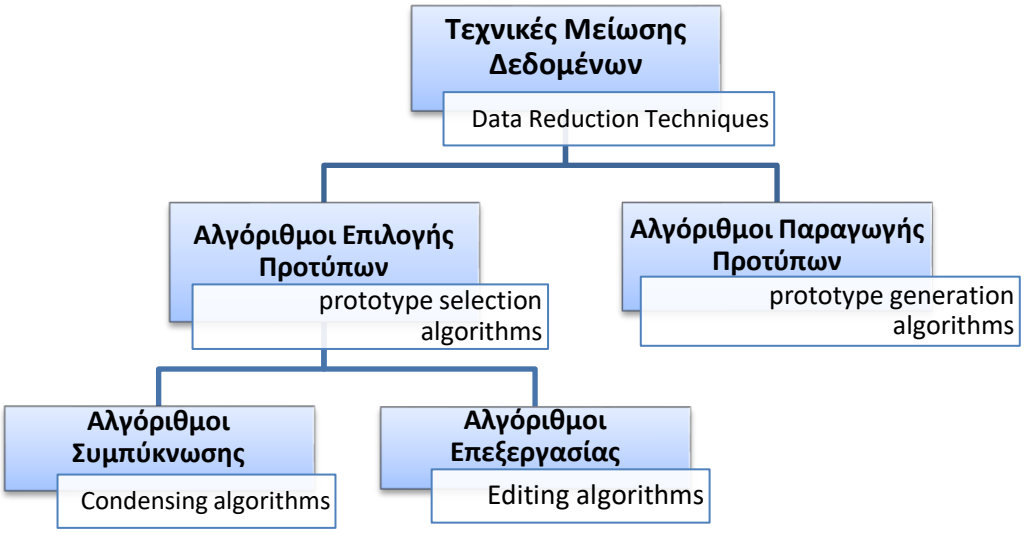
προτύπων (prototype generation algorithms) [20]. Στην πρώτη κατηγορία, οι αλγόριθμοι επιλογής προτύπων επιλέγουν κάποια αντιπροσωπευτικά στιγμιότυπα, ή αλλιώς πρότυπα, από το σύνολο δεδομένων ενώ στην δεύτερη κατηγορία οι αλγόριθμοι παραγωγής προτύπων δημιουργούν καινούρια στιγμιότυπα συνενώνοντας παρόμοια στιγμιότυπα εκπαίδευσης και έτσι δημιουργούν καινούργια πρότυπα. Στην πραγματικότητα κάθε πρότυπο αντιπροσωπεύει μία συγκεκριμένη περιοχή δεδομένων του πολυδιάστατου χώρου.

Οι αλγόριθμοι επιλογής προτύπων διασπώνται σε δύο υποκατηγορίες αλγορίθμων. Η πρώτη είναι οι **αλγόριθμοι συμπύκνωσης** (Condensing algorithms) και η δεύτερη είναι οι **αλγόριθμοι επεξεργασίας** (Editing algorithms)[46].

Οι αλγόριθμοι συμπύκνωσης έχουν στόχο να δημιουργήσουν ένα μικρό αντιπροσωπευτικό σύνολο δεδομένων από το αρχικό σύνολο εκπαίδευσης το οποίο ονομάζεται και σύνολο συμπύκνωσης (Condensing Set). Με αυτό τον τρόπο επιτυγχάνεται χαμηλότερο υπολογιστικό κόστος σε πρώτη φάση, ενώ παράλληλα δεν υπάρχουν μεγάλες απαιτήσεις αποθήκευσης, χωρίς παράλληλα να επηρεάζεται αρνητικά η ακρίβεια της κατηγοριοποίησης.

Οι αλγόριθμοι επεξεργασίας, σε αντίθεση με τους αλγόριθμους συμπύκνωσης, έχουν σαν στόχο την βελτίωση της ακρίβειας παρά την επίτευξη υψηλών ποσοστών μείωσης. Προσπαθούν δηλαδή να βελτιώσουν την ποιότητα των δεδομένων εκπαίδευσης με διάφορες τεχνικές, όπως την αφαίρεση του θορύβου, την αφαίρεση των ακραίων τιμών και λανθασμένων στοιχείων και την εξομάλυνση των ορίων απόφασης μεταξύ των τάξεων. Σαν αποτέλεσμα, ένας αλγόριθμος απομάκρυνσης θορύβου δημιουργεί ένα επεξεργασμένο σύνολο εκπαίδευσης χωρίς επικαλύψεις μεταξύ των τάξεων.

Ορισμένοι αλγόριθμοι συμπύκνωσης ενσωματώνουν και την τεχνική της επεξεργασίας. Οι αλγόριθμοι αυτοί ονομάζονται υβριδικοί αλγόριθμοι. Στην Εικόνα 2 που ακολουθεί παρουσιάζονται οι κατηγορίες που αναφέρθηκαν παραπάνω σε ιεραρχική σχεδίαση.



Εικόνα 2: Κατηγορίες τεχνικών μείωσης του πληθυσμού των δεδομένων

1.4 Κίνητρο και συνεισφορά

Ένας αποτελεσματικός και ευρέως χρησιμοποιούμενος οκνηρός αλγόριθμος κατηγοριοποίησης είναι ο k-NN. Έχει αναγνωριστεί ως απλός και εύχρηστος κατηγοριοποιητής, καθώς μπορεί να χρησιμοποιηθεί σε πολλούς τομείς εφαρμογών και να ενσωματωθεί εύκολα σε πολλά συστήματα.

Το αρνητικό του σημείο όμως είναι πως προϋποθέτει ότι τα σύνολα δεδομένων, τα οποία χειρίζεται, είναι πλήρη και δεν περιέχουν απούσες τιμές. Όταν όμως περιέχουν απούσες τιμές, που σχεδόν πάντα εμφανίζονται στα πραγματικά σύνολα δεδομένων εκπαίδευσης, τότε δεν μπορεί να υπολογίσει τις αποστάσεις μεταξύ των στιγμιότυπων, διότι δεν γνωρίζει πώς να τις χειριστεί, χαρακτηρίζοντάς τον απαγορευτικό για την εκτέλεσή του. Ακόμη και η εφαρμογή των τεχνικών μείωσης του πληθυσμού των δεδομένων προϋποθέτει τη μη ύπαρξη απουσών τιμών στο σύνολο δεδομένων. Έτσι, πριν από την προ-επεξεργασία από μια τεχνική μείωσης του πληθυσμού των δεδομένων, είναι απαραίτητη η εφαρμογή ενός ακόμη βήματος προ-επεξεργασίας για την συμπλήρωση των απουσών τιμών (Missing Values Imputation). Ωστόσο, η εφαρμογή ενός ακόμη βήματος προ-επεξεργασίας είναι άλλο ένα αρνητικό σημείο καθώς προσθέτει επιπλέον υπολογιστικό κόστος. Αυτό το σημείο αποτελεί και το κίνητρο για την εκπόνηση της παρούσας διπλωματικής εργασίας, δηλαδή πώς μπορούμε να αντιμετωπίσουμε τις απούσες τιμές, με ποιους τρόπους μπορούμε να τις συμπληρώσουμε ή να τις αγνοήσουμε ώστε να πετύχουμε παράλληλα καλύτερες επιδόσεις του κατηγοριοποιητή και πώς θα μπορούσε μία τεχνική μείωσης του πληθυσμού των

δεδομένων να μειώσει το απαιτούμενο επεξεργαστικό κόστος που προσθέτουν οι τεχνικές συμπλήρωσης των απουσών τιμών.

Η παρούσα διπλωματική εργασία προτείνει μια νέα παραλλαγή μιας τεχνικής μείωσης του πληθυσμού των δεδομένων που μπορεί να διαχειριστεί τις απούσες τιμές χωρίς να απαιτείται το επιπρόσθετο βήμα προ-επεξεργασίας για την συμπλήρωση τους και να είναι ανθεκτική στα δεδομένα με θόρυβο. Η τεχνική που προτείνεται είναι ένας αλγόριθμος παραγωγής προτύπων (Prototype Generation) και ονομάζεται αλγόριθμος Επεξεργασίας και Μείωσης μέσω Ομοιογενών συστάδων (Editing and Reduction through Homogeneous Clusters – ERHC). Συγκεκριμένα, η νέα παραλλαγή του ERHC, την οποία ονομάζουμε ERHC-PD, δεν απαιτεί την εκ των προτέρων συμπλήρωση των απουσών τιμών με κάποια τεχνική συμπλήρωσης απουσών τιμών (missing value imputation) [28] και αυτό επιτυγχάνεται αξιοποιώντας την τεχνική της μερικής απόστασης (partial distance) και εφαρμόζοντας συσταδοποίηση κ-μέσων (k-means) που δεν λαμβάνει υπ' όψιν τις απούσες τιμές. Έτσι, αποφεύγεται η επιπλέον επιβάρυνση υπολογιστικού κόστους που θα απαιτούσε η συμπλήρωση των απουσών τιμών και συνεπώς επεξεργάζεται απευθείας τα σύνολα δεδομένων που περιέχουν απούσες τιμές.

Η συνεισφορά μας, λοιπόν, με την εργασία αυτή έχει να κάνει και με την ανάπτυξη μιας εφαρμογής, η οποία δέχεται ένα πλήρες σύνολο δεδομένων και δημιουργεί σε αυτό, με τυχαίο τρόπο, απούσες τιμές σε ποσοστά που επιθυμούμε. Για τα δικά μας πειράματα δημιουργήθηκαν σύνολα δεδομένων με απούσες τιμές σε ποσοστό 10% και 20%. Στη συνέχεια δημιουργήσαμε μία νέα εφαρμογή, η οποία συμπληρώνει τις απούσες τιμές κάνοντας χρήση της τεχνικής του καταλογισμού μέσης τιμής κάθε κλάσης, κατά την οποία συμπληρώνουμε τις απούσες τιμές με τον μέσο όρο του κάθε χαρακτηριστικού της αντίστοιχης κλάσης.

Με πειραματικό έλεγχο δοκιμάζεται η λειτουργία και η απόδοση του αλγόριθμου ERHC-PD στα πλήρη σύνολα με απούσες τιμές κάνοντας χρήση του υπολογισμού της μερικής απόστασης καθώς επίσης και στα σύνολα που δημιουργήθηκαν από τους δύο παραπάνω αλγορίθμους χρησιμοποιώντας την προτεινόμενη παραλλαγή, την οποία ονομάζουμε ERHC-IMP. Οι δύο προαναφερθέντες παραλλαγές του ERHC αλγόριθμου συγκρίνονται μεταξύ τους αλλά και με τον αλγόριθμο των k εγγύτερων γειτόνων, χωρίς μείωση του πληθυσμού των δεδομένων, εκτελώντας πειράματα σε 13 σύνολα δεδομένων και εκτιμώντας την ακρίβεια κατηγοριοποίησης αλλά και το λόγο μείωσης (Reduction Rate) που επιτυγχάνουν. Τα πειραματικά αποτελέσματα δείχνουν αξιοσημείωτη απόδοση και για τις δύο παραλλαγές του ERHC αλγόριθμου.

1.5 Οργάνωση της διπλωματικής εργασίας

Αυτή η μεταπτυχιακή εργασία επικεντρώνεται σε δύο ερευνητικές διαστάσεις της εξόρυξης δεδομένων που αναφέρονται στην κατηγοριοποίηση. Η μία διάσταση είναι οι μέθοδοι συμπλήρωσης απουσών τιμών που συναντάμε σε σύνολα δεδομένων, και η άλλη διάσταση μελετά τις τεχνικές μείωσης του πληθυσμού των δεδομένων. Η εργασία συνεισφέρει μια νέα παραλλαγή μείωσης του πληθυσμού των δεδομένων, προτείνοντας βελτιώσεις σε μία υπάρχουσα τεχνική μείωσης του πληθυσμού των δεδομένων. Έτσι, στο πρώτο κεφάλαιο της εργασίας συνοψίζουμε τα κύρια σημεία που χρειάζονται για να εξερευνήσουμε τον χώρο της κατηγοριοποίησης τα οποία μας είναι απαραίτητα, ειδικά ο κατηγοριοποιητής k-NN και μας έδωσαν το κίνητρο για να δημιουργήσουμε και να συνεισφέρουμε με τον τρόπο μας στην υλοποίηση της προτεινόμενης βελτίωσης τεχνικής μείωσης προτύπων καθώς και στην πειραματική μελέτη της. Ταυτόχρονα παρουσιάζουμε την οργάνωση της εργασίας μας.

Στο κεφάλαιο 2 γίνεται μία μελέτη ανασκόπησης για τις τεχνικές συμπλήρωσης απουσών τιμών καθώς και στους μηχανισμούς και τρόπους αντιμετώπισης των απουσών τιμών με την μέθοδο του καταλογισμού. Στην συνέχεια γίνεται μία μελέτη για τις τεχνικές μείωσης του πληθυσμού των δεδομένων, τις κατηγορίες τους, τον τρόπο αξιολόγησής τους, και παράλληλα παρουσιάζονται οι πιο αντιπροσωπευτικές τεχνικές από κάθε κατηγορία. Κλείνοντας το κεφάλαιο, γίνεται αναφορά στον αλγόριθμο συσταδοποίησης k-Means, ο οποίος είναι βασικό δομικό στοιχείο για την τεχνική μείωσης του πληθυσμού των δεδομένων που εστιάζει η παρούσα εργασία και κατά συνέπεια για τη νέα προτεινόμενη παραλλαγή τους αλγορίθμου.

Στο κεφάλαιο 3 γίνεται αναφορά σε δύο τεχνικές μείωσης του πληθυσμού των δεδομένων που βασίζονται στην έννοια του σχηματισμού ομοιογενών συστάδων. Τα δεδομένα εκπαίδευσης δηλαδή, δημιουργούν ομάδες, ή αλλιώς συστάδες, που περιέχουν μόνο στιγμιότυπα της ίδιας κλάσης. Συγκεκριμένα, το κεφάλαιο 3 παρουσιάζει δύο αλγορίθμους. Αναφερόμαστε στον αλγόριθμο Μείωσης μέσω Ομοιογενών Συστάδων - RHC (Reduction through Homogeneous Clusters [55],[56] και τον αλγόριθμο Επεξεργασίας και Μείωσης μέσω Ομογενών Συστάδων - ERHC (Editing and Reduction through Homogeneous Clusters - ERHC) [57]. Ο RHC είναι ένας αποτελεσματικός αλγόριθμος παραγωγής προτύπων ο οποίος έχει χαμηλό κόστος προ-επεξεργασίας και παράλληλα επιτυγχάνει υψηλά ποσοστά μείωσης χωρίς να μειώνεται ιδιαίτερα η ακρίβεια κατηγοριοποίησης σε μεγάλα σύνολα δεδομένων. Αντίστοιχα ο ERHC είναι μια παραλλαγή του RHC που μπορεί να διαχειριστεί αποτελεσματικά σύνολα δεδομένων με θόρυβο. Λόγω του ότι έχει ως βάση τον RHC θεωρείται εξίσου γρήγορος, και πειραματικές μελέτες δείχνουν ότι επιτυγχάνει υψηλότερα

ποσοστά μείωσης στοιχείων και υψηλότερο ποσοστό ακρίβειας από τον RHC, ειδικά όταν το σύνολο δεδομένων περιέχει θόρυβο.

Στο κεφάλαιο 4 παρουσιάζονται οι δύο αλγόριθμοι μείωσης του πληθυσμού των δεδομένων που έχουν αναπτυχθεί για τις ανάγκες της διπλωματικής μας. Με τον πρώτο αλγόριθμο που προτείνουμε, τον ERHC-PD, για να διαχειριστούμε τις απούσες τιμές, κάνουμε χρήση της μερικής ευκλείδειας απόστασης κατά τον υπολογισμό των αποστάσεων από τα πλησιέστερα κέντρα κατά την διαδικασία της συσταδοποίησης του αλγορίθμου k-means. Ο δεύτερος αλγόριθμος που προτείνουμε, ο ERHC-IMP, εφαρμόζεται πάνω σε σύνολα δεδομένων, στα οποία έχουμε συμπληρώσει τις απούσες τιμές με τον μέσο όρο του κάθε χαρακτηριστικού της κάθε κλάσης, ώστε να μπορεί να ελέγχεται ως προς την απόδοσή του πάνω σε σύνολα δεδομένων που περιέχουν θόρυβο.

Στο κεφάλαιο 5 παρουσιάζονται όλες οι ρυθμίσεις και οι παράμετροι που χρησιμοποιήθηκαν και πως αξιοποιήθηκαν για την πειραματική μελέτη που έγινε πάνω σε 13 γνωστά σύνολα δεδομένων. Στην συνέχεια παρουσιάζονται και αναλύονται τα πειραματικά αποτελέσματα που προέκυψαν από τις μετρήσεις των προτεινόμενων αλγορίθμων όσον αφορά την ακρίβεια κατηγοριοποίησης αλλά και τον λόγο μείωσης του πληθυσμού των δεδομένων που επιτυγχάνουν.

Τέλος, στο κεφάλαιο 6 παρουσιάζονται τα συμπεράσματα και γίνονται προτάσεις για μελλοντική εργασία.

2

Θεωρητικό υπόβαθρο

Το πρόβλημα των απουσών τιμών είναι ένα συχνό πρόβλημα, το οποίο οι ερευνητές συχνά καλούνται να επιλύσουν. Στο κεφάλαιο αυτό γίνεται αναφορά στο πρόβλημα των απουσών δεδομένων, στις κατηγορίες που χωρίζονται ανάλογα με τον μηχανισμό που τις προκαλεί καθώς και στους τρόπους αντιμετώπισής τους. Παρουσιάζεται μία μελέτη ανασκόπησης για τις τεχνικές συμπλήρωσης απουσών τιμών με τη μέθοδο του καταλογισμού. Στην συνέχεια γίνεται μία μελέτη για τις τεχνικές μείωσης του πληθυσμού των δεδομένων, τις κατηγορίες τους, τον τρόπο αξιολόγησής τους και παράλληλα παρουσιάζονται οι πιο αντιπροσωπευτικές τεχνικές από κάθε κατηγορία. Τέλος, παρουσιάζεται ο αλγόριθμος συσταδοποίησης k-means, ο οποίος είναι βασικό δομικό στοιχείο για τις τεχνικές μείωσης του πληθυσμού των δεδομένων που χρησιμοποιήθηκαν στην παρούσα εργασία και παρουσιάζονται στο κεφάλαιο 3.

2.1 Συμπλήρωση Απουσών Τιμών

Στην σημερινή εποχή ο όγκος πληροφοριών που διακινούνται καθημερινά αυξάνεται με γοργούς ρυθμούς και καλύπτει διάφορους τομείς της ανθρώπινης δραστηριότητας. Πολλές εταιρείες, κυβερνήσεις, οργανισμοί καθώς και πολλά εκατομμύρια ανθρώπων, έχουν ψηφιοποιήσει τις δραστηριότητες τους με αποτέλεσμα να συνδέονται καθημερινά στο διαδίκτυο, μέσω των έξυπνων συσκευών που χρησιμοποιούν και να παράγουν τεράστιες

ποσότητες πληροφοριών. Πολλοί τομείς όπως η βιομηχανία, η υγεία, η οικονομία κ.α. συλλέγουν δεδομένα και τα αναλύουν για να παίρνουν αποφάσεις σχετικά με τις ενέργειες που πρέπει να ακολουθήσουν με στόχο την ανάπτυξή τους. Έτσι η εύκολη συλλογή και η πρόσβαση σε τέτοιου είδους δεδομένα ανοίγει το δρόμο για μεγάλες ευκαιρίες και επιτυχίες. Όμως πολύ συχνά εμφανίζεται το πρόβλημα των απουσών δεδομένων, ένα πρόβλημα το οποίο αποτελεί αντικείμενο εντατικής έρευνας από πολλούς ερευνητές και αναλυτές δεδομένων σε όλο τον κόσμο, το οποίο πρόβλημα αποτελεί ένα σοβαρό εμπόδιο στο στάδιο της προ-επεξεργασίας των δεδομένων.

Τα απόντα δεδομένα ή αλλιώς απούσες τιμές σε ένα σύνολο δεδομένων εμφανίζονται όταν για κάποιο λόγο δεν αποθηκεύεται κάποια τιμή για την μεταβλητή ενός χαρακτηριστικού ή αφήνοντας την πραγματική τιμή της παρατήρησης άγνωστη.

Πολλοί είναι οι λόγοι που μπορούν να οδηγήσουν στην εμφάνιση απουσών τιμών σε μία έρευνα. Συνήθως οι απούσες τιμές εμφανίζονται είτε λόγω της φύσης της έρευνας που πραγματοποιείται μέσω κάποιου ερωτηματολογίου που πιθανώς να μην είναι αναγκαίο (ανά περίπτωση) να απαντηθούν όλες οι ερωτήσεις ή σε περίπτωση που κάποιοι δεν θέλουν να απαντήσουν σε συγκεκριμένες ερωτήσεις για διάφορους προσωπικούς λόγους, είτε λόγω ανεπαρκούς γνώσης, είτε λόγω έλλειψης χρόνου, είτε λόγω παραβίαση ιδιωτικού απορρήτου κ.α. Άλλος λόγος που θα μπορούσαμε να έχουμε απόντα δεδομένα έχει να κάνει και με πιθανή αποτυχία εξοπλισμού που χρησιμοποιείται σε πειράματα της μελέτης που πραγματοποιείται. Ακόμη και ο ανθρώπινος παράγοντας είτε μέσω λανθασμένης χρήσης του εξοπλισμού είτε μέσω λανθασμένων εκτιμήσεων μπορεί να οδηγήσει σε εσφαλμένες μετρήσεις ή απώλεια τιμών. Πολλές φορές και οι τιμές από τέτοια λάθη μπορεί να θεωρούνται ως ακατάλληλες και θα πρέπει να αφαιρεθούν και να αντιμετωπιστούν σαν να λείπουν για να μην οδηγήσουν σε παραπλανητικά αποτελέσματα [23].

Λόγω του ότι το κάθε σύνολο δεδομένων προέρχεται από διαφορετική πηγή, γι αυτό και η ασυνέπεια των δεδομένων (απούσες τιμές) είναι διαφορετικών τύπων, οι οποίοι μπορούν να κατηγοριοποιηθούν σε τρεις μεγάλες κατηγορίες, ανάλογα με τον μηχανισμό που τους προκαλεί [28], όπως αναφέρονται παρακάτω:

- **Λείπουν εντελώς τυχαία** (Missing completely at random-MCAR) [28]. Οι απούσες τιμές εμφανίζονται εντελώς τυχαία και κατανέμονται ομοιόμορφα στα χαρακτηριστικά των στιγμιότυπων. Με άλλα λόγια, όλα τα χαρακτηριστικά έχουν τις ίδιες πιθανότητες να θεωρηθούν απούσες τιμές. Ο λόγος έλλειψης δεν σχετίζεται με τα παρατηρούμενα ή μη χαρακτηριστικά. Σε αυτήν την περίπτωση, οι απούσες τιμές είναι ένα τυχαίο υποσύνολο του συνόλου δεδομένων και δεν σχετίζονται με άλλα χαρακτηριστικά (που λείπουν ή παρατηρούνται).

- **Λείπουν τυχαία** (Missing at random-MAR) [28]. Οι τιμές που λείπουν δεν σχετίζονται με τα χαρακτηριστικά, αλλά σχετίζονται με ορισμένα από τα παρατηρούμενα χαρακτηριστικά. Αυτό σημαίνει ότι οι απούσες τιμές σχετίζονται με ένα ή περισσότερα χαρακτηριστικά του συνόλου δεδομένων. Για να γίνει πιο συγκεκριμένο, μια τιμή είναι MAR, όταν η πιθανότητα να λείπει εξαρτάται μόνο από τις διαθέσιμες πληροφορίες. Οι τιμές MAR είναι πιο συχνές από τις MCAR.
- **Να μην λείπουν τυχαία** (Not missing at random - NMAR) [28]. Η τιμή των χαρακτηριστικών που λείπουν σχετίζεται με το λόγο της έλλειψης. Το φαινόμενο ότι λείπουν όλες οι τιμές ενός χαρακτηριστικού λόγω των τιμών τους αναφέρεται ως λογοκρισία [29], αλλά σε πραγματικές συνθήκες είναι εξαιρετικά δύσκολο να πραγματοποιηθεί. [24][25][26].

Το γεγονός ότι ένα μέρος του πραγματικού συνόλου δεδομένων λείπει σημαίνει ότι μειώνεται η ποσότητα των πληροφοριών που μπορούν να αντληθούν από τα δεδομένα με αποτέλεσμα να επηρεάζεται η ικανότητα κατανόησης και η ικανότητα ανάλυσης των δεδομένων και να δημιουργούνται έτσι ερωτηματικά ως προς την αξιοπιστία των αποτελεσμάτων της έρευνας [22].

Αν ανατρέξουμε στη βιβλιογραφία θα παρατηρήσουμε ότι έχουν πραγματοποιηθεί αρκετές ερευνητικές προσπάθειες για την αντιμετώπιση του προβλήματος των απουσών τιμών. Παρόλα αυτά αποδεικνύεται ότι δεν υπάρχει ένας συγκεκριμένος τρόπος που να μπορεί να χειριστεί το θέμα των απουσών τιμών σε οποιοδήποτε σύνολο δεδομένων γιατί κάθε περίπτωση θεωρείται ότι είναι ξεχωριστή. Οι απούσες τιμές ανήκουν σε σύνολα δεδομένων που προκύπτουν από διαφορετικούς τομείς και είναι λογικό η φύση των δεδομένων να έχει διαφορετικά χαρακτηριστικά οπότε είναι δύσκολο να υπάρχει μία σωστή και ενιαία προσέγγιση στο θέμα των απουσών τιμών. Έχουν προταθεί πολλές μέθοδοι για την αντιμετώπιση των απουσών τιμών και σε γενικές γραμμές μπορούν να ομαδοποιηθούν σε δύο κυρίως τύπους. Ο πρώτος τύπος και σαν απλούστερη μέθοδος προτείνει την αγνόηση ή την απόρριψη των δεδομένων που λείπουν [30] ενώ ο δεύτερος τύπος, σαν δεύτερη μέθοδος, προτείνει την αντικατάσταση των τιμών που λείπουν με κάποια άλλη τιμή, μεθοδολογία γνωστή και ως "καταλογισμός".

Με βάση την προσέγγιση της πρώτης μεθόδου, η αγνόηση των απουσών τιμών σε ένα σύνολο δεδομένων θεωρείται και η πιο απλή διότι χρησιμοποιεί τους πιο απλούς και παραδοσιακούς μηχανισμούς για την αντιμετώπιση των απουσών τιμών σύμφωνα με τους οποίους διαγράφονται όλες οι τιμές των χαρακτηριστικών της κλάσης στην οποία υπάρχουν απούσες τιμές. Κατά συνέπεια, διαγράφεται ολόκληρο το στιγμιότυπο από το σύνολο δεδομένων. Με τον τρόπο αυτό όμως χάνεται η πραγματική αξία των υπολοίπων χαρακτηριστικών με αποτέλεσμα να μικραίνει το σύνολο δεδομένων που επεξεργαζόμαστε.

Σε πολλές περιπτώσεις όμως η μέθοδος αυτή θα πρέπει να αποφεύγεται διότι αν το αρχικό σύνολο δεδομένων είναι μικρό ή περιέχει μεγάλο ποσοστό απουσών τιμών, για παράδειγμα μεγαλύτερο από το 20%, τότε θα έχουμε μεγάλη και σημαντική απώλεια πληροφοριών, καθώς υπάρχει ο κίνδυνος απόρριψης τιμών που ενδεχομένως να παίζουν σημαντικό ρόλο στην τελική απόφαση, με αποτέλεσμα να οδηγηθούμε σε ανακριβή συμπεράσματα [27].

Με βάση την προσέγγιση της δεύτερης μεθόδου, κάνοντας δηλαδή χρήση του καταλογισμού, μιας μεθοδολογίας που θεωρείται ευκολότερη και πιο ελκυστική από την προηγούμενη, με βάση την οποία γίνεται αντικατάσταση των απουσών τιμών με μία συγκεκριμένη τιμή ή σταθερά. Βέβαια υπάρχει το ενδεχόμενο αυτή η συγκεκριμένη τιμή να θεωρηθεί ως παραπλανητική διότι μπορεί να μην αντιπροσωπεύει τις πραγματικές ή επιθυμητές τιμές του χαρακτηριστικού της κλάσης στο σύνολο δεδομένων. Έχουν προταθεί πιο έξυπνες και πιο εύκολες στην εφαρμογή μεθοδολογίες με βάση τις οποίες αντί να χρησιμοποιείται μία τυχαία τιμή ή σταθερά, χρησιμοποιούνται προβλεπόμενες τιμές για το σύνολο δεδομένων. Υπολογίζεται έτσι μία αντιπροσωπευτική τιμή, αν πρόκειται για αριθμητικά χαρακτηριστικά, δηλαδή μία τιμή που αντιπροσωπεύει το μέσο όρο των τιμών του συγκεκριμένου χαρακτηριστικού στο σύνολο δεδομένων ή χρησιμοποιείται μία τιμή η οποία εμφανίζεται τις περισσότερες φορές στο χαρακτηριστικό του συνόλου δεδομένων. Αν θεωρήσουμε ότι γίνεται και μία βελτίωση της μεθόδου τότε θα ήταν πιο σωστό να χρησιμοποιηθεί για παράδειγμα ο μέσος όρος ο οποίος αντιπροσωπεύει τον μέσο όρο του χαρακτηριστικού της συγκεκριμένης κλάσης στο σύνολο των δεδομένων ή να επιλεχθεί με τον ίδιο τρόπο η τιμή που εμφανίζεται πιο πολλές φορές σε κάθε κλάση αντίστοιχα [30].

Στην βιβλιογραφία αναφέρεται πως μια απούσα τιμή αντικαθίσταται με μια εκτίμηση της τιμής, είτε εναλλακτικά αντικαθίσταται από αντίστοιχες προβλέψεις μοντέλων που πιθανώς συνδυάζονται για την εκτίμηση νέας τιμής. Υπάρχουν διαθέσιμες πολλές μεθοδολογίες καταλογισμών για απούσες τιμές στα στιγμιότυπα εκπαίδευσης. Ωστόσο, ορισμένες μεθοδολογίες όπως ο πολλαπλός καταλογισμός [37] (ή επαναλαμβανόμενος καταλογισμός) είναι μια προσέγγιση Monte Carlo που παράγει πολλαπλές προσομοιωμένες εκδόσεις ενός συνόλου δεδομένων που κάθε μία αναλύεται και τα αποτελέσματα συνδυάζονται για να δημιουργήσουν συμπεράσματα. Αντίθετα με την παραπάνω μεθοδολογία, στην διπλωματική αυτή μελετήσαμε τεχνικές απλού καταλογισμού που μπορούν να παράγουν απ’ ευθείας σύνολα δεδομένων και συγκεκριμένα θα χρησιμοποιήσουμε από την μέθοδο απλής εισαγωγής τον καταλογισμό μέσης τιμής κάθε κλάσης. Ας αναφέρουμε όμως εδώ πως οι πιο ευρέως χρησιμοποιούμενες μέθοδοι καταλογισμού εμπίπτουν σε τρεις κύριες κατηγορίες και για κάθε μία κατηγορία θα αναφέρουμε τις πιο αντιπροσωπευτικές στατιστικές τεχνικές [83]:

1. **Μέθοδοι διαγραφής** - Deletion methods (διαγραφή κατά παραδοχή, δηλαδή ανάλυση πλήρους περίπτωσης, διαγραφή κατά ζεύγη, δηλαδή ανάλυση διαθέσιμης περίπτωσης)

   Καταλογισμός Διαγραφής δεδομένων [30]. Σε αυτήν την τεχνική απλώς διαγράφονται όλα τα στιγμιότυπα από το σύνολο δεδομένων που περιέχουν απούσες τιμές. Στην περίπτωση ενός πολύ μεγάλου συνόλου δεδομένων με πολύ λίγες απούσες τιμές, η προσέγγιση αυτή θα μπορούσε ενδεχομένως να λειτουργήσει πραγματικά καλά. Ωστόσο, εάν οι απούσες τιμές είναι σε διαφορετικές σειρές, αυτή η μέθοδος μπορεί να αλλοιώσει αρκετά σοβαρά τα αποτελέσματα που θα εξαχθούν από το σύνολο δεδομένων. Ένα άλλο σημαντικό πρόβλημα με αυτήν την προσέγγιση είναι ότι δεν θα είναι σε θέση να επεξεργαστεί τυχόν μελλοντικά δεδομένα που περιέχουν απούσες τιμές. Αν η μέθοδος αυτή χρησιμοποιηθεί σε ένα μοντέλο πρόβλεψης που έχει σχεδιαστεί για παραγωγική διαδικασία, τότε αυτό θα μπορούσε να δημιουργήσει σοβαρά προβλήματα στο στάδιο της ανάπτυξης.

2. **Μέθοδοι απλής εισαγωγής** - Single Imputation Methods (αντικατάσταση μέσου όρου/συχνής τιμής, γραμμική παρεμβολή, Hot deck, καταλογισμός χρησιμοποιώντας τον k-NN)

   Καταλογισμός Μέσης τιμής (Mean imputation) [30]: Αυτή η τεχνική λειτουργεί υπολογίζοντας τον μέσο όρο των τιμών που δεν λείπουν σε μια στήλη ξεχωριστά και στη συνέχεια αντικαθιστά τις απούσες τιμές σε κάθε στήλη από τον αντίστοιχο μέσο όρο της στήλης που ανήκει.

   Καταλογισμός μέσης τιμής κάθε κλάσης (Class mean imputation ή Group Mean Imputation ή class-conditional mean imputation) [30]: Για προβλήματα κατηγοριοποίησης, μια από τις πιο γνωστές και απλές μεθόδους είναι ο καταλογισμός μέσης τιμής κάθε κλάσης. Αποτελεί επέκταση της τεχνικής καταλογισμού μέσης τιμής. Στόχος της είναι να υπολογισθούν οι μέσοι όροι για κάθε χαρακτηριστικό της κάθε κλάσης και να γίνει αντικατάσταση του αντίστοιχου μέσου όρου στην θέση της απούσας τιμής του χαρακτηριστικού της κάθε κλάσης ξεχωριστά [63][64][65]. Η τεχνική αυτή είναι σχετικά απλή, εύκολη, γρήγορη και μπορεί να χρησιμοποιηθεί μόνο με αριθμητικά δεδομένα. Δεν επηρεάζει τις συσχετίσεις μεταξύ των χαρακτηριστικών. Λειτουργεί μόνο στο επίπεδο της στήλης των χαρακτηριστικών. Αυτό το μοντέλο είναι χρήσιμο μόνο για MAR, αλλά δεν είναι χρήσιμο για MCAR [24][32][33]. Για τις ανάγκες της τρέχουσας διπλωματικής ακολουθήθηκε η συγκεκριμένη τεχνική.

   Καταλογισμός πιο συχνής τιμής: (Most Frequent imputation) [30]: Αυτή η προσέγγιση καταλογισμού λειτουργεί με κατηγορηματικά χαρακτηριστικά (συμβολοσειρές ή αριθμητικές αναπαραστάσεις) και χρησιμοποιεί την τιμή του χαρακτηριστικού που σε όλα τα παρατηρούμενα δεδομένα εμφανίζεται πιο συχνά. Έτσι τοποθετεί στην θέση της

απούσας τιμής την πιο συχνά εμφανιζόμενη τιμή για το κάθε χαρακτηριστικό. Δεν επηρεάζει τις συσχετίσεις μεταξύ χαρακτηριστικών και μπορεί να δημιουργεί μία προκατάληψη για τις τιμές των χαρακτηριστικών.

Καταλογισμός Μηδενικής ή σταθερής τιμής (Zero or Constant imputation) [30]: όπως υποδηλώνει το όνομα - αντικαθιστά τις τιμές που λείπουν είτε με μηδέν είτε με οποιαδήποτε σταθερή τιμή που προκαθορίζεται.

Καταλογισμός θερμού καταστρώματος (Hot deck imputation) [31]: αυτή η τεχνική χρησιμοποιείται για κατηγορηματικά δεδομένα και είναι αποτελεσματική για μεγάλα σύνολα δεδομένων και όχι για μικρά σύνολα δεδομένων. Σε αυτήν τη τεχνική, η απούσα τιμή αντικαθίσταται από παρόμοιες τιμές αυτού του χαρακτηριστικού. Δηλαδή για κάθε στιγμιότυπο που περιέχει απούσες τιμές, βρίσκει το πιο κοντινό και παρόμοιο στιγμιότυπο και οι απούσες τιμές συμπληρώνονται από αυτό το στιγμιότυπο. Εάν το πιο παρόμοιο στιγμιότυπο περιέχει επίσης απούσες τιμές για τα ίδια χαρακτηριστικά με αυτά του αρχικού στιγμιότυπου, τότε αυτό απορρίπτεται και βρίσκει ένα δεύτερο πιο κοντινό στιγμιότυπο. Η διαδικασία επαναλαμβάνεται έως ότου όλες οι απούσες τιμές καταλογιστούν επιτυχώς ή ολοκληρωθεί η αναζήτηση σε ολόκληρο το σύνολο δεδομένων. Σε περίπτωση που δεν υπάρχει παρόμοιο στιγμιότυπο με τις απαιτούμενες τιμές, το πλησιέστερο στιγμιότυπο με τον ελάχιστο αριθμό απουσών τιμών επιλέγεται για να αντικαταστήσει τα χαρακτηριστικά που περιέχουν απούσες τιμές. Υπάρχουν διάφοροι τρόποι εύρεσης του πιο παρόμοιου στιγμιότυπου με το αρχικό στιγμιότυπο που περιέχει απούσες τιμές [62]. Αυτή η μέθοδος γίνεται προβληματική όταν δεν υπάρχουν άλλα παρόμοια στιγμιότυπα [24][34][35].

Καταλογισμός χρησιμοποιώντας τον k-NN (k-nearest Neighbor imputation) [81]: Αυτή η τεχνική χρησιμοποιεί ευκλείδειες αποστάσεις για να προσδιορίσει την ομοιότητα μεταξύ δύο στιγμιότυπων και να αντικαταστήσει την απούσα τιμή με τον σταθμισμένο μέσο όρο των τιμών που λαμβάνονται από τους k πλησιέστερους γείτονες. Η τεχνική αυτή χρησιμοποιείται για σύνολα δεδομένων που έχουν ποσοτικές τιμές χαρακτηριστικών. Δεν υπάρχει ανάγκη δημιουργίας κάποιου προγνωστικού μοντέλου για κάθε χαρακτηριστικό των απουσών τιμών και βοηθάει ιδιαίτερα στα σύνολα δεδομένων με πολλαπλές απούσες τιμές. Ακόμη, μπορεί να είναι πολύ ακριβέστερη από τη τεχνική της μέσης τιμής ή τις πιο συχνής τιμής (εξαρτάται από το σύνολο δεδομένων). Το κύριο μειονέκτημα της προσέγγισης k-NN είναι ότι, κάθε φορά που ο k-NN αναζητά τις πιο παρόμοιες περιπτώσεις, ο αλγόριθμος κάνει αναζήτηση σε όλο το σύνολο δεδομένων [33]. Αυτό σημαίνει ότι είναι υπολογιστικά δαπανηρή. Λειτουργεί αποθηκεύοντας ολόκληρο το σύνολο δεδομένων κατάρτισης στη μνήμη. Το k-NN είναι αρκετά ευαίσθητο και στις αποκλίσεις-θόρυβο που συναντάμε στα σύνολα δεδομένων.

3. **Μέθοδοι βασιζόμενοι σε μοντέλο** (πολλαπλός καταλογισμός, καταλογισμός παλινδρόμησης).

   Πολλαπλός καταλογισμός (Multiple imputation) [31],[36]: Ο πολλαπλός καταλογισμός προτάθηκε για πρώτη φορά από τον Rubin [37] και τώρα είναι ένας ολοένα και πιο δημοφιλής τρόπος αντιμετώπισης απουσών τιμών. Παράγει m πλήρη σύνολα δεδομένων και στη συνέχεια κάθε σύνολο δεδομένων αναλύεται με κάποια μέθοδο αξιόπιστων εργαλείων ανάλυσης δεδομένων, όπως t-test ή ANOVA. Το impute σημαίνει να "συμπληρώσετε". Με μεθόδους μοναδικού καταλογισμού, χρησιμοποιείται ο μέσος όρος, ο διάμεσος ή κάποιο άλλο στατιστικό στοιχείο για να καταλογιστούν οι απούσες τιμές. Εντούτοις, χρησιμοποιώντας μεμονωμένες τιμές δημιουργείται ένα επίπεδο αβεβαιότητας σχετικά με τις τιμές που πρέπει να καταλογιστούν. Ο πολλαπλός καταλογισμός περιορίζει την αβεβαιότητα για τις απούσες τιμές υπολογίζοντας πολλές και διαφορετικές τιμές καταλογισμού. Δημιουργούνται έτσι διάφορες εκδόσεις του ίδιου συνόλου δεδομένων. Στο τέλος τα αποτελέσματα που προέρχονται από αυτά τα m σύνολα δεδομένων, συνδυάζονται για να προκύψουν οι "καλύτερες" τιμές. Έτσι λοιπόν ο πολλαπλός καταλογισμός έχει την ιδιότητα να εξισορροπεί το επίπεδο αβεβαιότητας σχετικά με τις τιμές που πρέπει να καταλογιστούν.

   Καταλογισμός παλινδρόμησης (Regression imputation) [82]: Η τεχνική αυτή εφαρμόζεται με τη χρήση γνωστών τιμών για την κατασκευή του μοντέλου και υπολογίζει την παλινδρόμηση μεταξύ των μεταβλητών και στη συνέχεια εφαρμόζει αυτό το μοντέλο για να υπολογίσει τις απούσες τιμές. Έρευνες έχουν δείξει πως αυτή η τεχνική δίνει ακριβέστερα αποτελέσματα από ότι ο καταλογισμός μέσου όρου [32].

2.2 Τεχνικές μείωσης του πληθυσμού των δεδομένων (DRTs)

Όπως έχουμε δει στο κεφάλαιο 1.3, οι τεχνικές μείωσης του πληθυσμού των δεδομένων (DRT - Data Reduction Techniques) οδηγούν σε μία πιο ευέλικτη παρουσίαση των δεδομένων καθώς στοχεύουν στη μείωση του υπολογιστικού κόστους αναζήτησης και διατήρηση ταυτόχρονα της ακρίβειας της κατηγοριοποίησης σε υψηλά επίπεδα. Αυτό επιτυγχάνεται με την δημιουργία ενός μικρού συνόλου δεδομένων το οποίο αντιπροσωπεύει ένα μεγάλο σε όγκο σύνολο δεδομένων που τους δίνεται. Το σύνολο δεδομένων που προκύπτει ονομάζεται συμπυκνωμένο σύνολο (CS - Condensing Set) και ανάλογα με την μέθοδο πού χρησιμοποιείται προσπαθούν να μειώσουν κατά πολύ τον όγκο χωρίς όμως να χάνεται η ακεραιότητα των αρχικών δεδομένων. Αυτό συνεπάγεται ότι εκτός από τη γρήγορη

κατηγοριοποίηση νέων στιγμιότυπων θα μειώσει αισθητά το υπολογιστικό κόστος δηλαδή τις μεγάλες απαιτήσεις σε μνήμη, αποθηκευτικό χώρο, και χρόνο εκτέλεσης των αλγορίθμων κατηγοριοποίησης.

Οι τεχνικές μείωσης του πληθυσμού των δεδομένων μπορούν να ομαδοποιηθούν σε δύο μεγάλες κατηγορίες αλγορίθμων:

(i) **αλγόριθμοι επιλογής προτύπων** (prototype selection algorithms- PS) [14] [46]και

(ii) **αλγόριθμοι παραγωγής προτύπων** (prototype abstraction (or generation) algorithms - PA) [20][44][45]. Και οι δύο κατηγορίες έχουν στόχο τη δημιουργία ενός μικρού αντιπροσωπευτικού συνόλου δεδομένων, αλλά διαφέρουν ως προς την μεθοδολογία που χρησιμοποιούν για να το κατασκευάσουν.

Οι τεχνικές επιλογής προτύπων [46] επιλέγουν ορισμένα στιγμιότυπα από το αρχικό σύνολο εκπαίδευσης και τα χρησιμοποιούν ως αντιπροσώπους και τα αποθηκεύουν στο συμπυκνωμένο σύνολο, ενώ οι αλγόριθμοι παραγωγής προτύπων [47] δημιουργούν νέους αντιπροσώπους, συνοψίζοντας όμοια στιγμιότυπα του αρχικού συνόλου και τα τοποθετούν στο συμπυκνωμένο σύνολο. Οι αναφορές [46][47] παρουσιάζουν και συγκρίνουν εκτός των άλλων, τις παραπάνω κατηγοριοποιήσεις.

Να υπενθυμίσουμε εδώ πως οι αλγόριθμοι επιλογής προτύπων[46], διασπώνται σε δύο υποκατηγορίες αλγορίθμων (βλ. Εικόνα 2). Η πρώτη είναι οι **αλγόριθμοι συμπύκνωσης** (Condensing algorithms) και η δεύτερη είναι οι **αλγόριθμοι επεξεργασίας** (Editing algorithms). Ο σκοπός και ο στόχος και των δύο περιγράφονται στην 3^{η} παράγραφο του 1^{ου} κεφαλαίου.

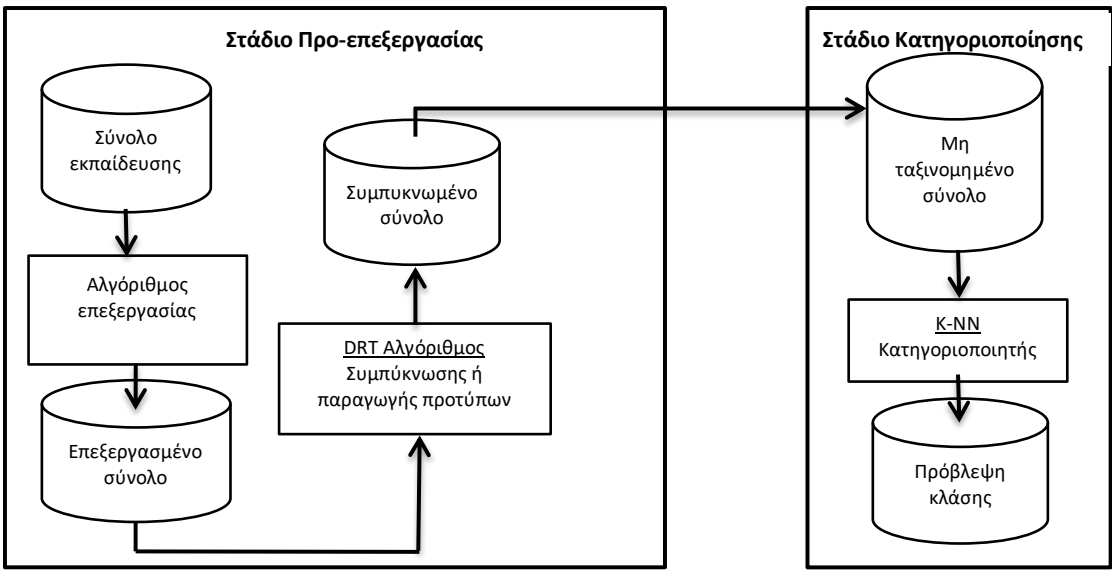
Όσον αφορά το θέμα της αξιολόγησης των τεχνικών μείωσης του πληθυσμού των δεδομένων (DRT) χρησιμοποιούνται 3 κριτήρια αξιολόγησης :

α) το πρώτο είναι ο **λόγος μείωσης (reduction rate)** ο οποίος εστιάζει στο πόσο μικρότερο γίνεται το μέγεθος του συνόλου συμπύκνωσης σε σχέση με το αρχικό σύνολο εκπαίδευσης

β) το δεύτερο κριτήριο είναι η **ακρίβεια κατηγοριοποίησης** που επιτυγχάνεται, κάνοντας χρήση του κατηγοριοποιητή k-NN όταν εκτελείται πάνω στο σύνολο συμπύκνωσης, και

γ) τρίτο κριτήριο είναι το **υπολογιστικό κόστος** της προ-επεξεργασίας που απαιτείται για την κατασκευή του συνόλου συμπύκνωσης. Ανάλογα τον τομέα για τον οποίο γίνεται η κατηγοριοποίηση, κάθε ένα κριτήριο έχει και άλλη βαρύτητα ανάλογα βέβαια και με τις απαιτήσεις που πρέπει να πληρούνται.

Στην Εικόνα 3 που ακολουθεί παρουσιάζεται η διαδικασία που απαιτείται για το στάδιο της προ-επεξεργασίας κάνοντας χρήση κάποιας τεχνικής μείωσης του πληθυσμού των δεδομένων και η διαδικασία που απαιτείται για το στάδιο της κατηγοριοποίησης κάνοντας χρήση του κατηγοριοποιητή k-NN.

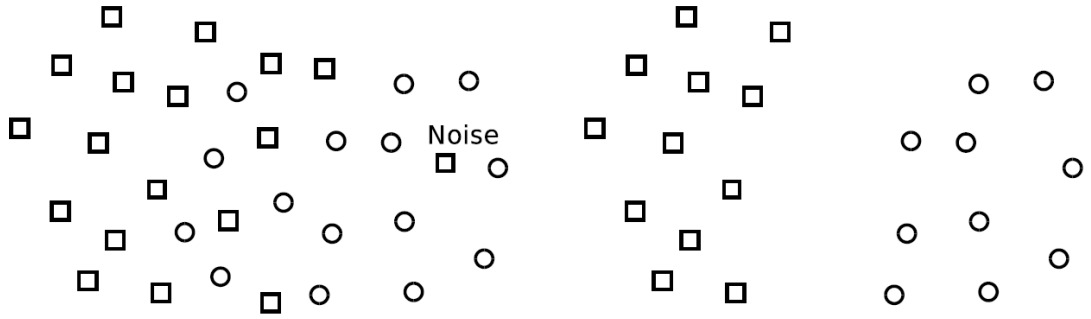


Εικόνα 3: Διαδικασία κατηγοριοποίησης μέσω της μείωσης του πληθυσμού των δεδομένων

Στόχος λοιπόν των τεχνικών μείωσης του πληθυσμού των δεδομένων (DRT) είναι να δημιουργήσουν ένα σύνολο συμπύκνωσης χωρίς θόρυβο και να κρατήσουν ή να παράγουν για κάθε κλάση επαρκή πρότυπα για να κάνουν τη δουλειά του k-NN κατηγοριοποιητή πιο εύκολη και πιο γρήγορη. Παρουσιάζονται στην συνέχεια ανά κατηγορία κάποιες βασικές τεχνικές μείωσης του πληθυσμού των δεδομένων και ο σκοπός που επιτελούν.

2.2.1 Αλγόριθμοι απομάκρυνσης θορύβου

Οι αλγόριθμοι αυτής της κατηγορίας προσπαθούν να βελτιώσουν την ποιότητα του συνόλου εκπαίδευσης απομακρύνοντας τον θόρυβο, τα λανθασμένα στιγμιότυπα καθώς και τα ακραία στιγμιότυπα που βρίσκονται στα όρια μιας κλάσης. Στόχος τους είναι να δημιουργήσουν ένα σύνολο εκπαίδευσης το οποίο θα χωρίζει τις κλάσεις έτσι ώστε να μην υπάρχουν επικαλύψεις στα σημεία των ορίων των κλάσεων (Εικόνα 4). Έχουν αναπτυχθεί διάφοροι αλγόριθμοι που προσπαθούν που κάνουν αυτή τη δουλειά. Οι κυριότεροι αναφέρονται παρακάτω:



Σύνολο εκπαίδευσης Σύνολο επεξεργασίας.

Εικόνα 4: Διαχωρισμός κλάσεων και αφαίρεση θορύβου

Αλγόριθμος επεξεργασίας πλησιέστερου γείτονα (ENN)

Ο κανόνας επεξεργασίας πλησιέστερου γείτονα έχει αποτελέσει τη βάση πάνω στην οποία αναπτύχθηκαν και οι υπόλοιποι αλγόριθμοι επεξεργασίας. Έχει δημιουργηθεί από τον Wilson [38] και είναι ο πιο απλός αλγόριθμος. Η λειτουργία του έχει να κάνει με τη δημιουργία ενός συνόλου επεξεργασίας το οποίο αρχικά ορίζεται να είναι το ίδιο με το σύνολο εκπαίδευσης. Ομοίως αρχικά ορίζεται και το k δηλαδή μία τιμή που χαρακτηρίζει το πλήθος των πλησιέστερων γειτόνων που θα εξετάζονται κάθε φορά. Έτσι για κάθε στιγμιότυπο x του συνόλου εκπαίδευσης ο αλγόριθμος ελέγχει όλα τα k πλησιέστερα στιγμιότυπα γύρω από το στιγμιότυπο x. Σε περίπτωση που το στιγμιότυπο x έχει κατηγοριοποιηθεί εσφαλμένα ή βρίσκεται σε ακραίο σημείο ανάμεσα στα κ στιγμιότυπα που ελέγχονται τότε χαρακτηρίζεται ως θόρυβος και απομακρύνεται από το σύνολο επεξεργασίας. Η διαδικασία επαναλαμβάνεται για κάθε στιγμιότυπο x του αρχικού συνόλου εκπαίδευσης με σκοπό να προκύψει το τελικό σύνολο επεξεργασίας.

Η αποτελεσματικότητα αυτού του αλγορίθμου χαρακτηρίζεται από δύο σημεία. Πρώτον όταν έχουμε μεγάλο σύνολο εκπαίδευσης τότε το κόστος επεξεργασίας είναι πολύ υψηλό διότι ο κανόνας ENN θεωρείται χρονοβόρος και με μεγάλο υπολογιστικό κόστος λόγω του ότι για κάθε στιγμιότυπο x υπολογίζονται οι αποστάσεις του για όλους τους k γείτονες που έχουν οριστεί.

Το δεύτερο χαρακτηριστικό έχει να κάνει με την τιμή του k που ορίζεται από την αρχή και χαρακτηρίζει το σύνολο των γειτόνων που θα ελέγχονται για κάθε στιγμιότυπο x [39]. Έρευνες έχουν δείξει ότι η τιμή 3 (k=3) είναι μία τυπική τιμή που επιτυγχάνει την καλύτερη απόδοση αλλά οι δοκιμές με διαφορετικές τιμές συντελούν στην καλύτερη απόδοση μιας και κάθε σύνολο δεδομένων είναι διαφορετικό [40]. Αν ληφθεί υπόψη και η κατανομή των στιγμιότυπων στον πολυδιάστατο χώρο, υπάρχει πιθανότητα η χρήση της σταθερής τιμής του k να μην είναι η βέλτιστη καθώς ο αλγόριθμος να θεωρεί ψευδώς ως

θόρυβο στιγμιότυπα που χρειάζονται και να τα διαγράψει ή το ανάποδο, να κρατήσει στιγμιότυπα που είναι στην πραγματικότητα θόρυβος.

Αλγόριθμος All k-NN

Μία βελτιωμένη παραλλαγή του κανόνα ENN παρουσιάζει ο κανόνας All k-NN [41]. Η διαφορά τους έγκειται στο ότι χρησιμοποιεί όχι μία αλλά πολλές τιμές για το k τις οποίες ορίζει ως kmax και επαναλαμβάνει τον κανόνα ENN για αυτές τις kmax διαφορετικές τιμές. Με παρόμοιο τρόπο ορίζει το σύνολο επεξεργασίας ίδιο με το σύνολο εκπαίδευσης και για κάθε στιγμιότυπο του x από το σύνολο εκπαίδευσης εφαρμόζει τον k-NN κατηγοριοποιητή για όλα τα στιγμιότυπα του συνόλου εκπαίδευσης. Ξεκινάει για k=1 και προσπαθεί να αφαιρέσει το στιγμιότυπο x από το σύνολο επεξεργασίας όπως γινόταν και στον ENN. Στην περίπτωση που το στιγμιότυπο x έχει κατηγοριοποιηθεί εσφαλμένα τότε το διαγράφει και συνεχίζεται η διαδικασία με το επόμενο στιγμιότυπο x. Διαφορετικά αυξάνει το k κατά ένα και ο αλγόριθμος προσπαθεί να αφαιρέσει το στιγμιότυπο x. Αν γίνουν όλες οι kmax επαναλήψεις και δεν καταφέρει να διαγράψει το στιγμιότυπο x τότε αυτό παραμένει στο σύνολο επεξεργασίας και ο αλγόριθμος συνεχίζει με το επόμενο στιγμιότυπο x ορίζοντας το k ίσον με ένα.

Όσον αφορά τον τρόπο λειτουργίας αυτού του αλγορίθμου, όπως και πριν θα πρέπει να οριστεί το πλήθος του kmax για να έχουμε την καλύτερη απόδοση, το οποίο βέβαια γίνεται κατόπιν δοκίμων και τελικών σφαλμάτων για την καλύτερη τιμή του. Δεδομένου ότι χρησιμοποιεί περισσότερες τιμές για το k, αφαιρεί και περισσότερα στιγμιότυπα από ότι ο κανόνας ENN. Ο υπολογισμός των αποστάσεων είναι εξίσου ίδιος όπως και στον κανόνα ENN και θεωρείται παραμετρικός.

Αλγόριθμος Multiedit

Ο αλγόριθμος Multiedit [42] είναι μία άλλη παραλλαγή που βασίζεται στον κανόνα ENN αλλά χρησιμοποιεί τον κανόνα 1-NN. Με παρόμοιο τρόπο ορίζει το σύνολο επεξεργασίας ίδιο με το σύνολο εκπαίδευσης και διαιρεί το σύνολο εκπαίδευσης σε n τυχαία υποσύνολα ίδιου μεγέθους και εφαρμόζει τον ENN κατηγοριοποιητή για όλα τα στιγμιότυπα x του κάθε υποσυνόλου αλλά αναζητώντας μόνο τον έναν πλησιέστερο γείτονα (κανόνας 1-NN) στο επόμενο υποσύνολο. Στην περίπτωση που το στιγμιότυπο x έχει κατηγοριοποιηθεί εσφαλμένα τότε το διαγράφει από το σύνολο επεξεργασίας. Αν αφαιρεθεί τουλάχιστον ένα στιγμιότυπο τότε ορίζεται το σύνολο επεξεργασίας ως σύνολο εκπαίδευσης και επαναλαμβάνεται η όλη διαδικασία. Διαφορετικά αν δεν γίνει καμία επεξεργασία στο σύνολο, τότε ο αλγόριθμος Multiedit τερματίζεται.

Όσον αφορά τον τρόπο λειτουργίας αυτού του παραμετρικού αλγορίθμου, όπως και πριν θα πρέπει να οριστεί από την αρχή το πλήθος των τυχαίων υποσυνόλων n και των επαναλήψεων R, για να έχουμε την καλύτερη απόδοση, το οποίο βέβαια γίνεται κατόπιν δοκιμών. Για αρκετά μεγάλα σύνολα εκπαίδευσης, το κύριο πλεονέκτημα της επαναληπτικής διαδικασίας είναι ότι η συμπεριφορά της είναι σημαντικά καλύτερη λόγω του γεγονότος ότι δεν έχει εξάρτηση από την παράμετρο k, αντίθετα από τον προηγούμενο αλγόριθμο. Η συμπεριφορά των προσεγγίσεων επεξεργασίας με βάση τα τυχαία υποσύνολα χειροτερεύει καθώς μειώνεται το μέγεθος του συνόλου εκπαίδευσης. Αυτή η υποβάθμιση της αποτελεσματικότητας γίνεται πιο σημαντική όταν αυξάνεται ο αριθμός των υποσυνόλων[43]. Στην πραγματικότητα, για σχετικά μικρά σύνολα, ο αλγόριθμός ENN του Wilson λειτουργεί πολύ καλύτερα από ο αλγόριθμος Multiedit. Να σημειωθεί και το γεγονός ότι οι επαναλαμβανόμενες εφαρμογές του αλγορίθμου ενδέχεται να δημιουργήσουν διαφορετικό σύνολο επεξεργασίας .

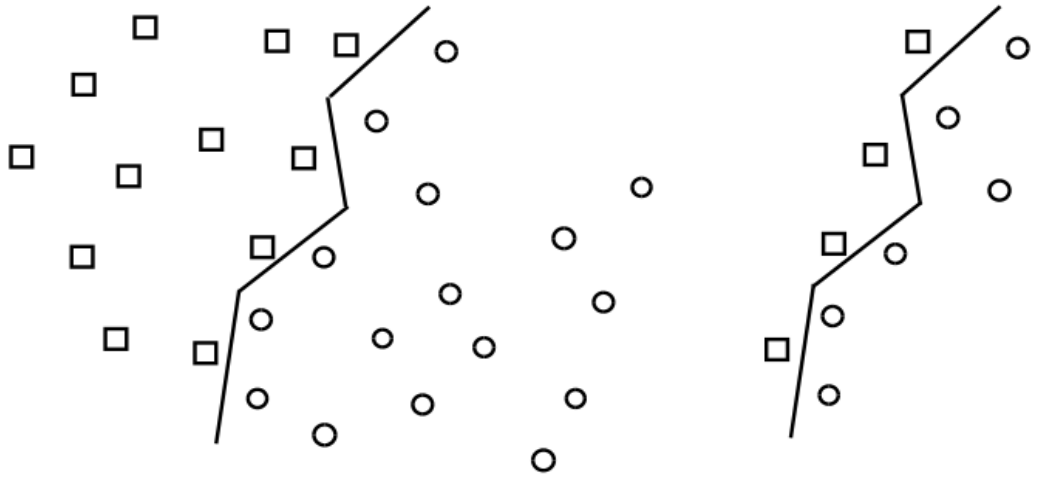
2.2.2 Αλγόριθμοι επιλογής προτύπων για συμπύκνωση δεδομένων

Με την χρήση των αλγορίθμων συμπύκνωσης επιτυγχάνεται χαμηλότερο υπολογιστικό κόστος σε πρώτη φάση, χωρίς να υπάρχουν μεγάλες απαιτήσεις αποθήκευσης και χωρίς παράλληλα να επηρεάζεται αρνητικά η ακρίβεια της κατηγοριοποίησης.

Έχουν αναπτυχθεί διάφοροι αλγόριθμοι που προσπαθούν που κάνουν αυτή τη δουλειά. Οι κυριότεροι αναφέρονται παρακάτω:

Αλγόριθμος συμπύκνωσης εγγύτερου γείτονα (Condensing Nearest Neighbor rule - CNN-rule)

Ένας από τους πιο γνωστούς, μη-παραμετρικούς και με μεγάλη συχνότητα χρήσης, αλγορίθμους μείωσης του πληθυσμού των δεδομένων εκπαίδευσης εισήχθη από τον Hart [48], ο οποίος πρώτος παρουσίασε την ιδέα ότι τα στιγμιότυπα του συνόλου εκπαίδευσης που δε βρίσκονται κοντά στα σύνορα απόφασης των κλάσεων (όρια) μπορούν να απομακρυνθούν με ασφάλεια, με αποτέλεσμα το κόστος της σειριακής αναζήτησης των γειτόνων να μειωθεί σε μεγάλο βαθμό. Έχει αποδειχθεί ότι αποτελεί έναν από τους πιο αποδοτικούς αλγορίθμους και πολλοί μεταγενέστεροι αλγόριθμοι έχουν βασιστεί στην ιδέα του Hart.



(α) Σύνολο εκπαίδευσης TS (β) Σύνολο συμπύκνωσης CS

Εικόνα 5: Στιγμιότυπα του συνόλου εκπαίδευσης και στιγμιότυπα στα όρια των κλάσεων

Ο αλγόριθμός CNN-rule δουλεύει ως εξής: Αρχικά, ένα στιγμιότυπο του συνόλου εκπαίδευσης TS τοποθετείται στο σύνολο συμπύκνωσης CS. Στη συνέχεια, ο κανόνας CNN προσπαθεί να ταξινομήσει το περιεχόμενο του TS χρησιμοποιώντας τον κανόνα 1-NN σαρώνοντας τα στιγμιότυπα του συνόλου συμπύκνωσης CS. Αν κάποιο στιγμιότυπο δεν κατηγοριοποιηθεί σωστά, θεωρείται ότι βρίσκεται σε μια περιοχή δεδομένων κοντά στα όρια απόφασης και έτσι μεταφέρεται από το σύνολο εκπαίδευσης TS στο σύνολο συμπύκνωσης CS. Η διαδικασία επαναλαμβάνεται μέχρις ότου να μην πραγματοποιούνται μετακινήσεις από το σύνολο εκπαίδευσης στο σύνολο συμπύκνωσης. Τα υπόλοιπα στιγμιότυπα του συνόλου εκπαίδευσης απορρίπτονται και προκύπτει το σύνολο συμπύκνωσης.

Αλγόριθμος συμπύκνωσης IB2

Μία γρήγορη παραλλαγή του αλγορίθμου CNN-rule αποτελεί ο IB2 [49] . Είναι κι αυτός μη παραμετρικός αλγόριθμός και το σύνολο συμπύκνωσης εξαρτάται σε μεγάλο βαθμό από τη σειρά των στιγμιότυπων στο σύνολο εκπαίδευσης.

Ο αλγόριθμός CNN-rule δουλεύει ως εξής: Κάθε στιγμιότυπο x από το σύνολο εκπαίδευσης ταξινομείται χρησιμοποιώντας τον κατηγοριοποιητή 1-NN στο τρέχον σύνολο συμπύκνωσης (CS). Εάν το στιγμιότυπο x έχει κατηγοριοποιηθεί σωστά, τότε απορρίπτεται. Διαφορετικά, το στιγμιότυπο x μεταφέρεται στο σύνολο συμπύκνωσης (CS). Σε αντίθεση με τον κανόνα CNN ο IB2 δεν διασφαλίζει ότι όλα τα απορριπτόμενα στιγμιότυπα μπορούν να κατηγοριοποιηθούν σωστά από το σύνολο συμπύκνωσης CS. Λόγω του ότι είναι ένας αλγόριθμος ενός περάσματος, είναι πολύ γρήγορος, κατά συνέπεια με χαμηλό υπολογιστικό κόστος προ-επεξεργασίας και δεν απαιτεί όλα τα δεδομένα εκπαίδευσης να βρίσκονται στην

κύρια μνήμη. Ακόμη, θεωρείται κατάλληλος για δυναμικά περιβάλλοντα ροών δεδομένων όπου νέα δεδομένα εκπαίδευσης γίνονται σταδιακά διαθέσιμα , λόγω του ότι κατασκευάζει το σύνολο συμπύκνωσης διαδοχικά, δηλαδή, νέα στιγμιότυπα εκπαίδευσης μπορούν να ληφθούν υπόψη μετά τη δημιουργία του συνόλου συμπύκνωσης. Συνεπώς, τα νέα στοιχεία εκπαίδευσης μπορούν να ενημερώσουν ένα υπάρχον σύνολο συμπύκνωσης με απλό τρόπο και χωρίς να λάβουν υπόψη τα προηγούμενα στιγμιότυπα που είχαν χρησιμοποιηθεί για την κατασκευή του αρχικού συνόλου συμπύκνωσης.

2.2.3 Αλγόριθμοι Παραγωγής Προτύπων

Κινούμενοι πάνω στην ίδια λογική με τους αλγόριθμους συμπύκνωσης έτσι και οι αλγόριθμοι παραγωγής προτύπων χτίζουν το σύνολο συμπύκνωσης με έναν διαφορετικό τρόπο. Έτσι, δημιουργούν πρότυπα στιγμιότυπα από παρόμοια στιγμιότυπα του συνόλου εκπαίδευσης, αντί να επιλέξουν ως πρότυπο κάποιο πραγματικό στιγμιότυπο από το σύνολο εκπαίδευσης. Στην πραγματικότητα ένας k-NN κατηγοριοποιητής που υιοθετεί την ιδέα της παραγωγής προτύπων εκτελείται πάνω από ένα τεχνητό σύνολο εκπαίδευσης. Στην συνέχεια αναφέρουμε τους πιο χαρακτηριστικούς αλγόριθμους παραγωγής προτύπων.

Αλγόριθμος Αφαίρεσης IB2 (AIB2)

Ο αλγόριθμος AIB2 αποτελεί μια παραλλαγή του IB2. Επομένως, κληρονομεί όλες τις προαναφερθείσες ιδιότητες του IB2. Η λειτουργικότητά του έχει ως εξής: τα πρότυπα πρέπει να βρίσκονται στο επίκεντρο της περιοχής δεδομένων που αντιπροσωπεύουν. Επομένως, τα σωστά κατηγοριοποιημένα στοιχεία δεν αγνοούνται. Χρησιμοποιούνται για να μετακινήσουν το πλησιέστερο πρότυπο. Στην πραγματικότητα, συμβάλλουν στην τελική ρύθμιση συμπύκνωσης αντικαθιστώντας το πλησιέστερο πρότυπο αποδίδοντας σε αυτό μία τιμή βάρους, η οποία υποδηλώνει τον αριθμό των αντικειμένων που αντιπροσωπεύει. Όταν το υπό-εξέταση στιγμιότυπο x βρίσκεται σε ένα πρότυπο ίδιας κλάσης, τότε το πρότυπο "μετακινείται" προς την κατεύθυνση του x. Αυτό επιτυγχάνεται αξιοποιώντας την τιμή βάρος. Με αυτό τον τρόπο Ο AIB2 στοχεύει στη βελτίωση της αποτελεσματικότητας του IB2 κατασκευάζοντας ένα σύνολο συμπύκνωσης στο οποίο κάθε πρότυπο να βρίσκεται κοντά στο κέντρο της περιοχής δεδομένων που αντιπροσωπεύει. Επομένως, ο αλγόριθμος AIB2 είναι σε θέση να επιτύχει υψηλότερη ακρίβεια κατηγοριοποίησης και ο παράλληλα υψηλότερα ποσοστά μείωσης και ακόμη χαμηλότερο επεξεργαστικό κόστος από το IB2.

Αλγόριθμος των Chen και Jozwik (CJA)

Ένας άλλος αποτελεσματικός αλγόριθμος παραγωγής προτύπων είναι ο αλγόριθμος Chen και Jozwik (CJA)[52]. Η λειτουργικότητά του έχει ως εξής: Αρχικά ανακτά τα πιο απομακρυσμένα στιγμιότυπα τα x και y στο σύνολο εκπαίδευσης (η απόσταση αυτή ορίζεται ως διάμετρος του συνόλου δεδομένων) και διαιρεί το σύνολο εκπαίδευσης σε δύο υποσύνολα: τα στιγμιότυπα που βρίσκονται πιο κοντά στο x τοποθετούνται στο υποσύνολο Sx ενώ τα στιγμιότυπα που βρίσκονται πιο κοντά στο y τοποθετούνται στο υποσύνολο Sy. Ο αλγόριθμος CJA προχωρά επιλέγοντας να διαιρέσει υποσύνολα που περιέχουν στιγμιότυπα περισσότερων από μία κλάσεων (μη ομοιογενή υποσύνολα). Διαιρείται πρώτα το μη ομοιογενές υποσύνολο με τη μεγαλύτερη διάμετρο. Εάν όλα τα υποσύνολα είναι ομοιογενή, τότε ο αλγόριθμος CJA συνεχίζει διαιρώντας τα ομοιογενή υποσύνολα. Αυτή η διαδικασία επαναλαμβάνεται μέχρι ο αριθμός των υποσυνόλων να είναι ίσος με μια καθορισμένη από το χρήστη τιμή. Για κάθε δημιουργημένο υποσύνολο S, ο αλγόριθμος CJA καταμετρά τα στιγμιότυπα του υποσυνόλου S και δημιουργεί ένα νέο στιγμιότυπο μέσου όρου στο οποίο αντιστοιχίζεται η κλάση που αντιπροσωπεύει το υποσύνολο S. Αυτά τα στιγμιότυπα μέσου όρου που δημιουργούνται από κάθε υποσύνολο, αποτελούν το τελικό σύνολο συμπύκνωσης. Η σειρά διαίρεσης των υποσυνόλων εξαρτάται από την διάμετρό τους, δηλαδή η ιδέα είναι ότι ένα υποσύνολο με μεγάλη διάμετρο πιθανότατα περιέχει περισσότερα στιγμιότυπα εκπαίδευσης. Επομένως, αν πρώτα υποδιαιρείται αυτό το υποσύνολο, θα επιτευχθεί υψηλότερος λόγος μείωσης. Να αναφέρουμε εδώ πως ο αλγόριθμος CJA είναι παραμετρικός αλγόριθμος, κατασκευάζει το σύνολο συμπύκνωσης ανεξάρτητα από την σειρά των στιγμιότυπων στο σύνολο εκπαίδευσης και πως τα στιγμιότυπα που δεν ανήκουν στην πιο κοινή κλάση του υποσυνόλου δεν αντιπροσωπεύονται στο σύνολο συμπύκνωσης.

Αλγόριθμοι μείωση με χωρισμό διαστήματος (Αλγόριθμοι RSP1- RSP2- RSP3)

Ο αλγόριθμος CJA ήταν η βάση πάνω στην οποία δημιουργήθηκε μία άλλη οικογένεια αλγορίθμων γνωστή ως οικογένεια αλγορίθμων μείωσης με χωρισμό διαστήματος (Reduction by Space Partitioning algorithms- RSP)[51] και θεωρείται έτσι ως πρόγονός τους. Στην συνέχεια αναφέρονται κάποια βασικά χαρακτηριστικά και συγκρίσεις μεταξύ τους.

Ο αλγόριθμος RSP1 (α) υπολογίζει τόσα στιγμιότυπα μέσου όρου όσα είναι ο αριθμός των διαφορετικών κλάσεων σε κάθε υποσύνολο (δεν αγνοεί στιγμιότυπα), (β) δημιουργεί μεγαλύτερο σύνολο συμπύκνωσης CS από ότι ο αλγόριθμος CJA (γ) προσπαθεί να βελτιώσει την ακρίβεια αφού λαμβάνει υπόψη όλα τα στιγμιότυπα εκπαίδευσης (δ) παρόμοια με τον αλγόριθμο CJA, χρησιμοποιεί τη διάμετρο του υποσυνόλου ως κριτήριο διάσπασης, με βάση την ιδέα ότι το υποσύνολο με τη μεγαλύτερη διάμετρο μπορεί να

περιέχει περισσότερα στιγμιότυπα εκπαίδευσης και έτσι μπορεί να επιτευχθεί υψηλότερος λόγος μείωσης.

Οι αλγόριθμοι RSP1 και RSP2 διαφέρουν ως προς τον τρόπο με τον οποίο επιλέγουν το επόμενο υποσύνολο που πρέπει να διαιρεθεί. Έτσι ο αλγόριθμος RSP2 χρησιμοποιεί ως κριτήριο διαίρεσης τον υψηλότερο βαθμό επικάλυψης ενός υποσυνόλου [51]. Ως βαθμός επικάλυψης ενός υποσυνόλου ορίζεται ο λόγος της μέσης απόστασης μεταξύ των στιγμιότυπων που ανήκουν σε διαφορετικές κλάσεις και της μέσης απόστασης μεταξύ στιγμιότυπων που ανήκουν στην ίδια κλάση.

Ο αλγόριθμος RSP3 υιοθετεί την έννοια της ομοιογένειας. Διαφέρει στο ότι συνεχίζει να διαιρεί τα μη ομοιογενή υποσύνολα και να τερματίζει όταν όλα αυτά γίνονται ομοιογενή. Ως κρίσιμο κριτήριο μπορεί να χρησιμοποιήσει είτε τη μεγαλύτερη διάμετρο είτε τον υψηλότερο βαθμό αλληλοεπικάλυψης.

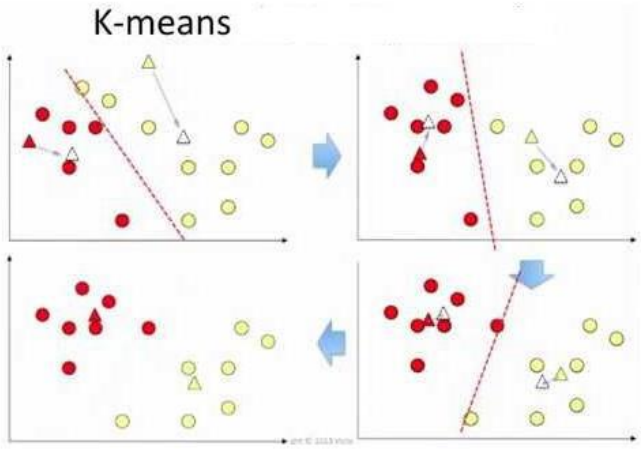
Ο αλγόριθμος RSP3 (όπως και ο αλγόριθμος CJA) είναι ο μόνος αλγόριθμος της οικογένειας RSP που καθορίζει αυτόματα το μέγεθος του συνόλου συμπύκνωσης CS και όπως ισχύει και για όλους τους προηγούμενους αλγορίθμους (CJA, RSP1 και RSP2), το σύνολο συμπύκνωσης CS που κατασκευάζει, είναι ανεξάρτητο από την σειρά των στιγμιότυπων στο σύνολο εκπαίδευσης TS.

Κλείνοντας θα αναφέρουμε τις βασικές ιδιότητες του αλγορίθμου RSP3 όπου (α) δημιουργεί λίγα πρότυπα για την αντιπροσώπευση των κοντινότερων στην κλάση στιγμιότυπων και πολλά πρότυπα για την αντιπροσώπευση των απομακρυσμένων και οριακών στιγμιότυπων της κλάσης του υποσυνόλου (β) ο λόγος μείωσης που επιτυγχάνεται από τον αλγόριθμο RSP3 εξαρτάται σε μεγάλο βαθμό από το επίπεδο θορύβου στα δεδομένα του συνόλου και (γ) η εύρεση των πιο απομακρυσμένων στιγμιότυπων σε κάθε υποσύνολο συνεπάγεται τον υπολογισμό όλων των αποστάσεων μεταξύ των στιγμιότυπων του υποσυνόλου.

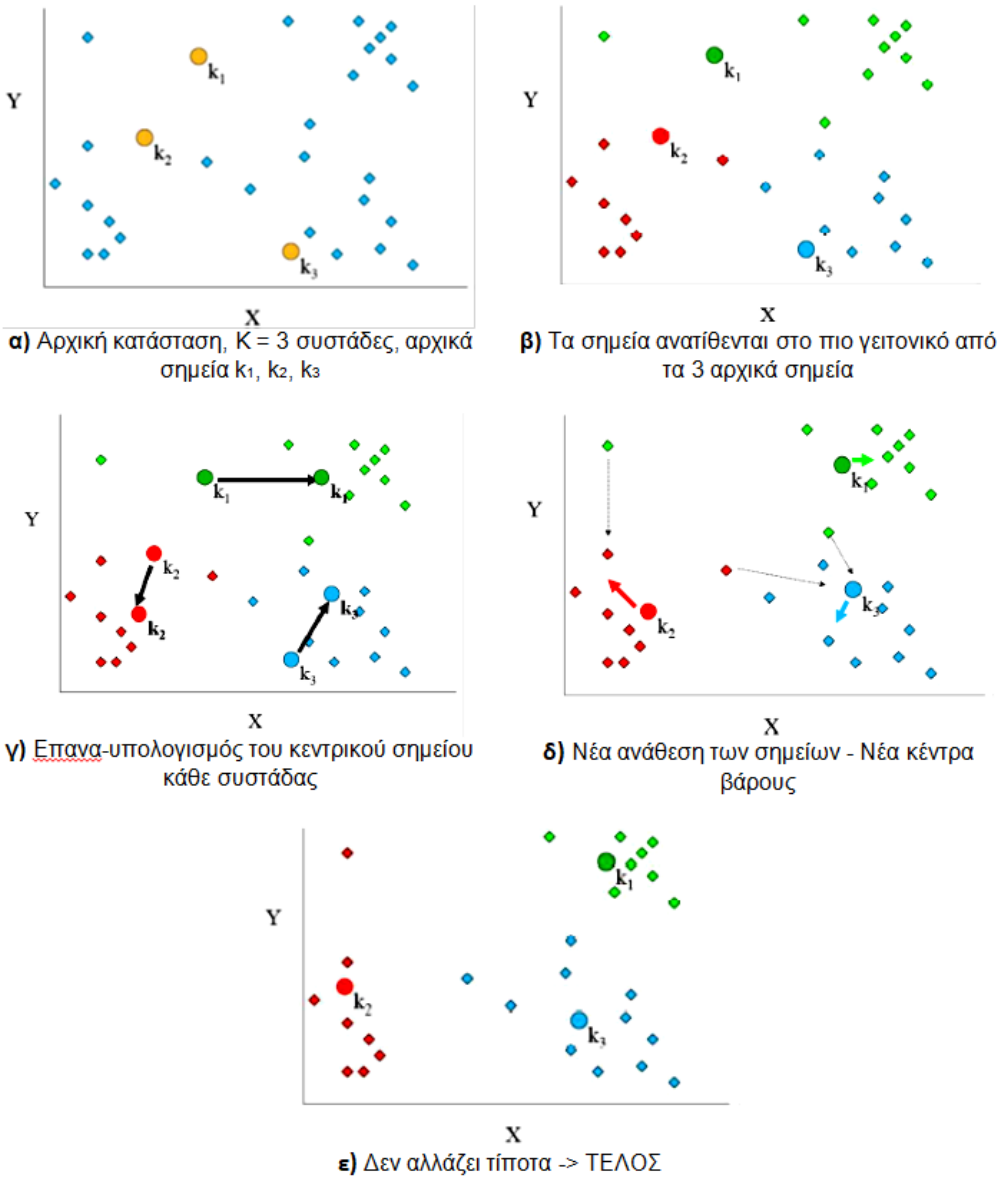
2.3 Συσταδοποίηση k-Means

Ο αλγόριθμος συσταδοποίησης k-means [53][54] είναι ένας από τους πιο γνωστούς και συχνότερα χρησιμοποιούμενους αλγορίθμους συσταδοποίησης. Είναι ένας απλός διαχωριστικός αλγόριθμος και είναι δημοφιλής εξαιτίας της απλότητας της υλοποίησής του. Στόχος του είναι να κατατάσσει τα στιγμιότυπα ενός συνόλου δεδομένων σε προκαθορισμένο k αριθμό ομάδων (clusters), έτσι ώστε τα στιγμιότυπα που ανήκουν στην ίδια ομάδα να είναι όσο το δυνατόν πιο ομοιόμορφα (να έχουν δηλαδή υψηλή ενδοκλασική ομοιότητα), ενώ τα στιγμιότυπα σε διαφορετικές ομάδες να είναι όσο το δυνατόν πιο διαφορετικά. Για να γίνει

αυτό θα πρέπει να προσδιοριστούν αρχικά k κεντρικά σημεία ή κέντρα (centroids), ένα για κάθε ομάδα (cluster). Η επιλογή των αρχικών κέντρων θέλει ιδιαίτερη προσοχή και επιδεξιότητα, καθώς η αρχική θέση των κέντρων επηρεάζει το αποτέλεσμα που θα παρουσιάσει ο αλγόριθμος. Γι αυτό και θεωρείται καλύτερη η επιλογή εκείνων των κέντρων που θα απέχουν μεταξύ τους όσο περισσότερο γίνεται. Στην συνέχεια γίνεται επιλογή κάθε στιγμιότυπου από το σύνολο δεδομένων και συσχετίζεται με το κοντινότερο σε αυτό κέντρο. Συνήθως για την συσχέτιση αυτή χρησιμοποιείται ως μέτρο απόστασης, η Ευκλείδεια απόσταση. Όταν ολοκληρωθεί η ίδια συσχέτιση για όλα τα στιγμιότυπα του συνόλου δεδομένων, τότε προκύπτει μία πρώτη και προσωρινή ομαδοποίηση. Ο αλγόριθμος συσταδοποίησης k-means επαναϋπολογίζει ξανά k νέα κέντρα, τα οποία αντιπροσωπεύονται από το κέντρο της ομάδας (δηλαδή το κέντρο βάρους της) που αντιστοιχεί στο μέσο όρο των συντεταγμένων των σημείων που αντιστοιχούν στην ομάδα, το οποίο μπορεί να μην είναι ένα από τα στιγμιότυπα του συνόλου δεδομένων της ομάδας. Αφού λοιπόν οριστούν τα k νέα κέντρα, ακολουθεί και πάλι η ίδια διαδικασία ανάθεσης καθενός από τα στιγμιότυπα του συνόλου δεδομένων στο νέο πλέον, κοντινότερο με αυτό, κέντρο. Έτσι, η διαδικασία επαναλαμβάνεται με αποτέλεσμα σε κάθε βήμα τα κέντρα να αλλάζουν θέση (να ορίζονται νέα) και τα στιγμιότυπα να ανατίθενται στην κατάλληλη ομάδα κάθε φορά με βάση το κοντινότερο κέντρο. Όταν σε κάποια επανάληψη δεν σημειωθούν συσχετίσεις στιγμιότυπων, δηλαδή οι ομάδες μένουν αμετάβλητες, τότε τερματίζεται η εκτέλεση του αλγορίθμου. Πολλές φορές ορίζεται ο αλγόριθμος να σταματά μετά από έναν ορισμένο αριθμό επαναλήψεων, ανάλογα με την κρίση του ερευνητή. Το αποτέλεσμα που προκύπτει είναι και η τελική ομαδοποίηση του συνόλου δεδομένων σε k ομάδες. Στην Εικόνα 6 και στην Εικόνα 7 παρουσιάζεται σχηματικά ο τρόπος λειτουργίας του αλγόριθμου συσταδοποίησης k-means με δύο και τρεις κλάσεις αντίστοιχα.



Εικόνα 6: Σχηματική αναπαράσταση μεθόδου λειτουργίας του k-means με δύο κλάσεις



Εικόνα 7: Σχηματική αναπαράσταση μεθόδου λειτουργίας του k-means με τρεις κλάσεις

Ο αλγόριθμος k-Means διαθέτει τα παρακάτω πλεονεκτήματα:

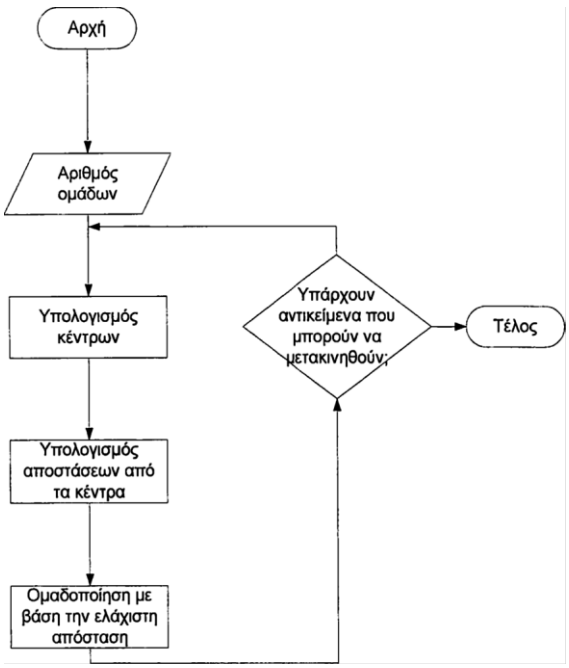
- Είναι απλός και κατανοητός.
- Τα στιγμιότυπα μοιράζονται σε συστάδες με αυτόματο τρόπο.
- Είναι αρκετά γρήγορος, τουλάχιστον σε σχέση με τις ιεραρχικές μεθόδους. Ο χρόνος εκτέλεσης του αλγορίθμου εξαρτάται γραμμικά από τα στοιχεία του προβλήματος, όπως το πλήθος των συστάδων k, το πλήθος των στιγμιότυπων n, το πλήθος των επαναλήψεων l καθώς και ο αριθμός των γνωρισμάτων του (διάσταση). Η υπολογιστική πολυπλοκότητα του αλγορίθμου είναι O(n*k*l*d). Για τον λόγο αυτό,

είναι πιο κατάλληλος από άλλες μεθόδους για την ομαδοποίηση μεγάλων συνόλων στιγμιότυπων.

Τα βασικά μειονεκτήματα του k-Means είναι τα ακόλουθα:

- Ο αριθμός των συστάδων πρέπει να προκαθοριστεί από τον χρήστη.
- Το τελικό αποτέλεσμα εξαρτάται σε σημαντικό βαθμό από την επιλογή των αρχικών κέντρων. Επιλογή διαφορετικών κέντρων μπορεί να οδηγήσει σε σημαντικά διαφορετικές συστάδες.
- Είναι πολύ ευαίσθητος στην ύπαρξη στιγμιότυπων με ακραίες τιμές (outliers). Λίγα στιγμιότυπα με πολύ μεγάλες τιμές μπορούν να επηρεάσουν σημαντικά τον υπολογισμό των νέων κέντρων και κατά συνέπεια τη διαμόρφωση των τελικών συστάδων.
- Έχει την τάση να δημιουργεί σφαιρικές και ίσου μεγέθους συστάδες. Για τον λόγο αυτό, δεν είναι κατάλληλος για συστάδες με περίπλοκα σχήματα ή με πολύ διαφορετικά μεγέθη.

Η διαγραμματική αναπαράσταση του αλγορίθμου k-means όσο και τα βήματα εκτέλεσής του μπορούν να συνοψιστούν ως εξής:



Βήμα 1: Καθορίζουμε τον αριθμό των ομάδων (k) που θα δημιουργηθούν (καθορισμός από τον ερευνητή)

Βήμα 2: Επιλέγουμε τυχαία k στιγμιότυπα ως αρχικά κέντρα των ομάδων.

Βήμα 3: Αντιστοιχίζουμε κάθε στιγμιότυπο στο πλησιέστερο κέντρο βάρους (με βάση συνήθως την ευκλείδεια απόσταση μεταξύ του στιγμιότυπου και του κέντρου βάρους)

Βήμα 4: Για κάθε ομάδα, επαναϋπολογίζουμε το κέντρο της χρησιμοποιώντας τις νέες μέσες τιμές όλων των στιγμιότυπων στην ομάδα.

Βήμα 5: Προσπαθούμε να ελαχιστοποιήσουμε το συνολικό άθροισμα τετραγώνων των σφαλμάτων μέσα στις ομάδες. Δηλαδή, επαναλαμβάνουμε τα βήματα 3 και 4 μέχρι να σταματήσουν οι ανακατανομές των στιγμιότυπων στις ομάδες που δημιουργούνται ή να επιτευχθεί ένα μέγιστος (προεπιλεγμένος) αριθμός επαναλήψεων.

Θα παρουσιάσουμε τώρα ένα παράδειγμα στον δισδιάστατο χώρο.

Έστω ότι θέλουμε να εφαρμόσουμε τον αλγόριθμο k-means με k=2 για τη συσταδοποίηση των παρακάτω 12 σημείων. Τα βήματα του αλγόριθμου παρουσιάζονται αναλυτικά με τις απαραίτητες επεξηγήσεις.

Σημεία	**X**	**Y**
1	7	4
2	6	4
3	5	6
4	4	2
5	6	3
6	5	2
7	3	3
8	4	5
9	6	5
10	3	6
11	4	4
12	8	2

1η ΕΠΑΝΑΛΗΨΗ

Επιλέγονται αυθαίρετα τα σημεία 1 και 8 ως τα αρχικά κέντρα των συστάδων Α και Β.

Υπολογίζονται οι αποστάσεις (Ευκλείδεια απόσταση) κάθε σημείου από τα αντίστοιχα κέντρα των συστάδων Α και Β.

Κάθε σημείο ανατίθεται στην κοντινότερη συστάδα.

Σημεία	**Απόσταση από Συστάδα A**	**Απόσταση από Συστάδα B**	**Ανάθεση στη Συστάδα:**
1	0.000	3.162	**A**
2	1.000	2.236	**B**
3	2.828	1.414	**B**
4	3.606	3.000	**B**
5	1.414	2.828	**A**
6	2.828	3.162	**A**
7	4.123	2.236	**B**
8	3.162	0.000	**B**
9	1.414	2.000	**A**
10	4.472	1.414	**B**
11	3.000	1.000	**B**
12	2.236	5.000	**A**

Υπολογίζονται οι τιμές για τα νέα κέντρα των δύο συστάδων ως ο μέσος όρος των αντίστοιχων τιμών των σημείων που ανήκουν στην κάθε συστάδα.

Κέντρα των συστάδων	x	y
A	5.333	2.667
B	4.143	4.286

Υπολογίζονται οι αποστάσεις (Ευκλείδεια απόσταση) κάθε σημείου από τα αντίστοιχα κέντρα των συστάδων Α και Β.

Κάθε σημείο ανατίθεται στην κοντινότερη συστάδα.

Σημεία	**Απόσταση από Συστάδα Α**	**Απόσταση από Συστάδα Β**	**Ανάθεση στη Συστάδα**
1	2.134	2.871	**Α**
2	1.491	1.879	**Α**
3	3.350	1.917	**Β**
4	1.491	2.290	**Α**
5	0.745	2.259	**Α**
6	0.745	2.441	**Α**
7	2.357	1.720	**Β**
8	2.687	0.728	**Β**
9	2.427	1.990	**Β**
10	4.069	2.060	**Β**
11	1.886	0.319	**Β**
12	2.749	4.484	**Α**

Υπολογίζονται οι τιμές για τα νέα κέντρα των δύο συστάδων ως ο μέσος όρος των αντίστοιχων τιμών των σημείων που ανήκουν στην κάθε συστάδα.

Κέντρα των συστάδων	x	y
A	6.000	2.833
B	4.167	4.833

Υπολογίζονται οι αποστάσεις (Ευκλείδεια απόσταση) κάθε σημείου από τα αντίστοιχα κέντρα των συστάδων Α και Β.

Κάθε σημείο ανατίθεται στην κοντινότερη συστάδα.

Σημεία	**Απόσταση από Συστάδα Α**	**Απόσταση από Συστάδα Β**	**Ανάθεση στη Συστάδα:**
1	1.537	2.953	**Α**
2	1.167	2.014	**Α**
3	3.321	1.434	**Β**
4	2.167	2.838	**Α**
5	0.167	2.593	**Α**
6	1.302	2.953	**Α**
7	3.005	2.173	**Β**

8	2.949	0.236	**B**
9	2.167	1.841	**B**
10	4.362	1.650	**B**
11	2.315	0.850	**B**
12	2.167	4.767	**A**

Δεν υπάρχει μετακίνηση σημείου μεταξύ των δύο συστάδων. **Ο αλγόριθμός τερματίζεται.**

3

Τεχνικές μείωσης του πληθυσμού των δεδομένων μέσω Ομοιογενών Συστάδων

Όπως έχουμε ήδη αναφέρει στην παράγραφο 1.3, οι τεχνικές μείωσης του πληθυσμού των δεδομένων (DRTs) έχουν στόχο τη μείωση των στιγμιότυπων ενός συνόλου δεδομένων. Το κεφάλαιο αυτό παρουσιάζει δύο τεχνικές μείωσης του πληθυσμού των δεδομένων που βασίζονται στην έννοια του σχηματισμού ομοιογενών συστάδων. Τα δεδομένα εκπαίδευσης δηλαδή, δημιουργούν ομάδες, ή αλλιώς συστάδες, clusters όπως αναφέρονται στη βιβλιογραφία, που περιέχουν μόνο στιγμιότυπα της ίδιας κλάσης. Στόχος αυτών των δύο τεχνικών είναι να πετύχουν μία γρήγορη προ-επεξεργασία των δεδομένων εκπαίδευσης χωρίς να απαιτείται από τον χρήστη να εισάγει κάποια παράμετρο και διατηρώντας παράλληλα την υψηλή απόδοση της κατηγοριοποίησης.

Ο πρώτος αλγόριθμος παραγωγής προτύπων που παρουσιάζεται στην παράγραφο 3.2, είναι ο αλγόριθμος Μείωσης μέσω Ομοιογενών Συστάδων - RHC (Reduction through Homogeneous Clusters [55],[56]). Είναι ένας αποτελεσματικός αλγόριθμος παραγωγής προτύπων ο οποίος έχει χαμηλό κόστος προ-επεξεργασίας και παράλληλα επιτυγχάνει υψηλά ποσοστά μείωσης χωρίς να μειώνεται ιδιαίτερα η ακρίβεια κατηγοριοποίησης σε μεγάλα σύνολα δεδομένων. Η λειτουργία του βασίζεται σε μία γρήγορη αναδρομική διαδικασία συσταδοποίησης που δημιουργεί ομοιογενείς συστάδες, τα μέσα των οποίων αποτελούν το τελικό σύνολο συμπύκνωσης. Λόγω του ότι βασίζεται στον γνωστό αλγόριθμο

συσταδοποίησης k-means, μπορεί εύκολα να ενσωματωθεί σε πολλά υπάρχοντα περιβάλλοντα.

Ο δεύτερος αλγόριθμος παραγωγής προτύπων που παρουσιάζεται στην παράγραφο 3.3, είναι ο αλγόριθμος Επεξεργασίας και Μείωσης μέσω Ομογενών Συστάδων - ERHC (Editing and Reduction through Homogeneous Clusters - ERHC) [57][58]. Είναι μια παραλλαγή του RHC που μπορεί να διαχειριστεί αποτελεσματικά σύνολα δεδομένων με θόρυβο. Ο αλγόριθμος ERHC λειτουργεί παρόμοια με τον RHC και θεωρείται απόγονός του. Χρησιμοποιεί την διαδικασία συσταδοποίησης του RHC, με την διαφορά ότι τις συστάδες που περιέχουν μόνο ένα στιγμιότυπο τις θεωρεί ως θόρυβο και τις αφαιρεί. Συνεπώς, θεωρεί πρότυπα τα μέσα των ομογενών συστάδων που περιέχουν περισσότερα από ένα στιγμιότυπα, τα οποία τα τοποθετεί στο σύνολο συμπύκνωσης.

3.1 Κίνητρο για την ανάπτυξη του RHC και του ERHC

Στόχος του RHC και του ERHC είναι να γίνουν πιο αποδοτικοί από τους υπολοίπους αλγορίθμους μείωσης του πληθυσμού των δεδομένων που αναφέρονται στην βιβλιογραφία. Για να το πετύχουν αυτό προσπαθούν να καλύψουν τις αδυναμίες των άλλων αλγορίθμων, οι οποίες αναφέρονται συνοπτικά παρακάτω:

- Η προ-επεξεργασία που απαιτούν πολλοί αλγόριθμοι είναι ένα χαρακτηριστικό που θεωρείται χρονοβόρο με μεγάλο υπολογιστικό κόστος και το οποίο πολλές φορές μπορεί να είναι απαγορευτικό για μεγάλα σύνολα δεδομένων.
- Πολλοί αλγόριθμοι συμπύκνωσης και παραγωγής προτύπων περιλαμβάνουν παραμέτρους που πρέπει να οριστούν από τον χρήστη. Αυτό σημαίνει ότι απαιτούν εκ των προτέρων από το χρήστη να δίνει προκαθορισμένες τιμές σε κάποιες παραμέτρους. Αυτό αποτελεί μία χρονοβόρα επαναληπτική εκτέλεση μιας διαδικασίας δοκιμής-σφάλματος ώστε να επιτευχθεί ο καλύτερος συνδυασμός τιμών για τις παραμέτρους έτσι ώστε να προκύψουν μεγαλύτερα ποσοστά μείωσης και ακρίβειας κατηγοριοποίησης.
- Κάθε αλγόριθμος συμπύκνωσης ή παραγωγής προτύπων δημιουργεί το σύνολο συμπύκνωσης ανάλογα με τη σειρά των στιγμιότυπων στο σύνολο εκπαίδευσης. Κατά συνέπεια, εάν κάθε αλγόριθμος διαβάζει τα στιγμιότυπα του ίδιου συνόλου δεδομένων με διαφορετική σειρά, τότε θα παρουσιάζει κάθε ένας και διαφορετικό σύνολο συμπύκνωσης.
- Δεν υπάρχει ικανοποιητική αναλογία μεταξύ του ποσοστού μείωσης του πληθυσμού των δεδομένων και του ποσοστού της ακρίβειας κατηγοριοποίησης. Υπάρχουν

αλγόριθμοι που πετυχαίνουν υψηλό λόγο μείωσης δεδομένων αλλά παρουσιάζουν μικρή ακρίβεια κατηγοριοποίησης. Αντίθετα υπάρχουν αλγόριθμοι που πετυχαίνουν μεγάλα ποσοστά ακρίβειας κατηγοριοποίησης αλλά το ποσοστό μείωσης του συνόλου εκπαίδευσης είναι χαμηλότερο από το αναμενόμενο.

Λαμβάνοντας υπόψη και τις τέσσερις αδυναμίες των άλλων αλγορίθμων καθώς και την απαίτηση για την πιο γρήγορη εκτέλεση του κατηγοριοποιητή k-NN σε μεγάλα σύνολα δεδομένων, αναπτύχθηκαν αλγόριθμοι RHC και ERHC που περιγράφονται στην συνέχεια.

3.2 Ο αλγόριθμος Μείωσης μέσω Ομοιογενών Συστάδων - RHC

Ο αλγόριθμος μείωσης μέσω ομοιογενών συστάδων είναι ένας αποτελεσματικός και μη παραμετρικός αλγόριθμος παραγωγής προτύπων. Η βασική ιδέα του είναι να εφαρμόζει αναδρομικά το γνωστό αλγόριθμο συσταδοποίησης k-means. Επεκτείνει την χρήση του αλγορίθμου συσταδοποίησης k-means, κατασκευάζοντας συνεχώς συστάδες μέχρις ότου να πετύχει την ομοιογένεια σε όλες τις συστάδες.

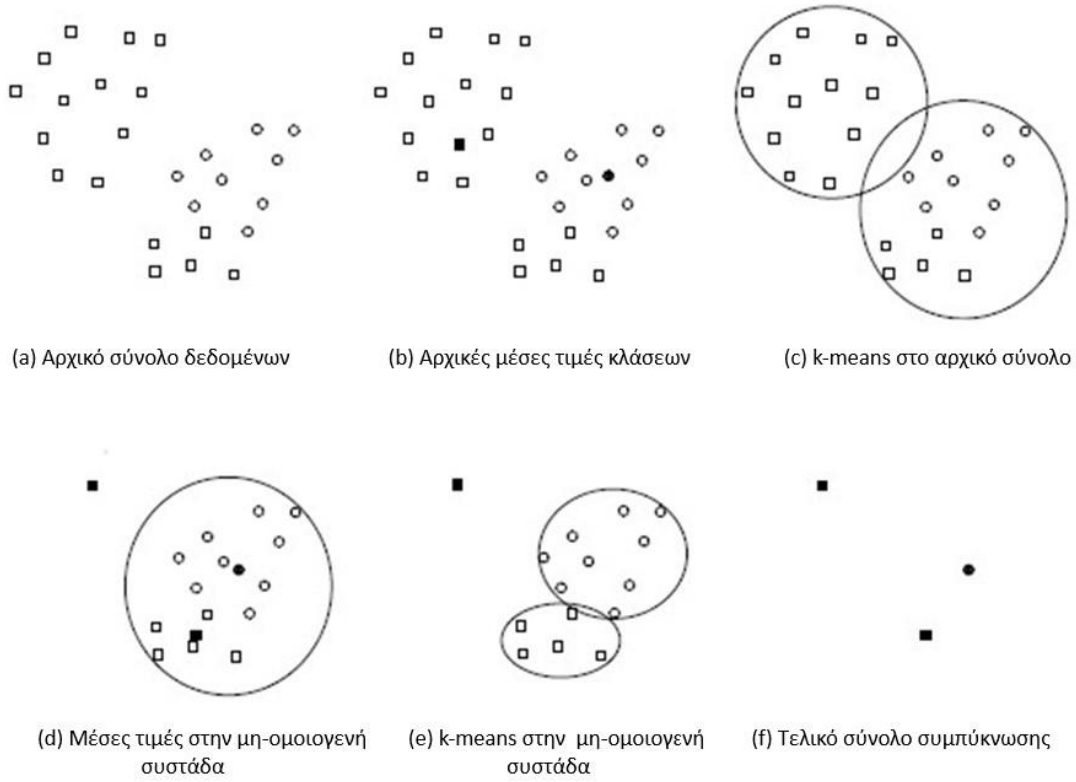
Η λειτουργικότητα του έχει ως εξής: Για αρχή κάνει την παραδοχή ότι όλο το σύνολο εκπαίδευσης είναι μία μη ομοιογενής συστάδα. Ξεκινάει υπολογίζοντας την μέση τιμή για κάθε κλάση, βρίσκοντας τον μέσο όρο των χαρακτηριστικών των αντίστοιχων στιγμιότυπων στο σύνολο εκπαίδευσης. Επομένως, αν ένα σύνολο εκπαίδευσης έχει n κλάσεις τότε θα δημιουργηθούν n μέσες τιμές που θα αντιπροσωπεύουν τις n αρχικές ομάδες ή συστάδες. Στη συνέχεια ο RHC, χρησιμοποιώντας τις μέσες τιμές που έχει υπολογίσει ως αρχικά κέντρα, κάνει συσταδοποίηση των στιγμιότυπων του συνόλου εκπαίδευσης καλώντας τον αλγόριθμο συσταδοποίησης k-means με αποτέλεσμα από τις n αρχικά μέσες τιμές, να προκύψουν n αρχικά συστάδες. Εάν κάποια συστάδα είναι ομοιογενής τότε παίρνει το κέντρο της και το τοποθετεί στο σύνολο συμπύκνωσης ως πρότυπο. Για κάθε άλλη μη ομοιογενής συστάδα επαναλαμβάνει αναδρομικά την παραπάνω διαδικασία. Ο αλγόριθμος τερματίζεται όταν θα επιτευχθεί η ομοιογένεια όλων των συστάδων. Κατά συνέπεια το σύνολο συμπύκνωσης αποτελείται από όλα τα μέσα των ομοιογενών συστάδων. Καταλαβαίνουμε πως η χρήση της μέσης τιμής για κάθε κλάση σαν αρχικά n μέσα κάθε κλάσης έχει ως αποτέλεσμα την αυτόματη δημιουργία των n συστάδων.

Στη συνέχεια παρουσιάζεται ο τρόπος υπολογισμού με τον οποίο εκτιμάται η μέση τιμή m για κάθε κλάση C, υπολογίζοντας το μέσο όρο των n τιμών των χαρακτηριστικών των

στιγμιότυπων x_i,i=1,2...|C| που ανήκουν στην C. Δηλαδή τα n χαρακτηριστικά αποδίδουν το $m.d_j$ της μέσης τιμής m, ως εξής:

$$m.d_j = \frac{1}{|C|} \sum_{x_i \in C} x_i.d_j, j = 1, 2, \ldots, n$$

Για να γίνει κατανοητός ο τρόπος εκτέλεσης του RHC παραθέτουμε ένα παράδειγμα εκτέλεσής του στον δισδιάστατο χώρο. Θεωρούμε ότι έχουμε ένα σύνολο δεδομένων το οποίο αποτελείται από είκοσι έξι στιγμιότυπα τα οποία υπάγονται σε δύο κλάσεις: τετράγωνα και κύκλοι (Εικόνα 8(a)). Ο αλγόριθμος RHC υπολογίζει μία μέση τιμή για την κλάση που αναφέρεται στα τετράγωνα και μία μέση τιμή για την κλάση που αναφέρεται στους κύκλους, όπως φαίνονται στην Εικόνα 8(b) με μαύρο τετράγωνο και μαύρο κύκλο. Στην συνέχεια καλείται ο αλγόριθμος συσταδοποίησης k-means, ο οποίος χρησιμοποιεί τα δύο κέντρα των κλάσεων ως αρχικά κέντρα και κατασκευάζει έτσι δύο συστάδες. Η μία συστάδα περιλαμβάνει τους κύκλους και η άλλη συστάδα περιλαμβάνει τα τετράγωνα. Παρατηρούμε ότι η συστάδα που περιλαμβάνει τα τετράγωνα έχει ομοιογένεια (Εικόνα 8(c)) με αποτέλεσμα να τοποθετεί την μέση τιμή που έχει υπολογιστεί στο σύνολο συμπύκνωσης ως πρότυπο της συστάδας των τετραγώνων. Για την δεύτερη συστάδα που περιλαμβάνει τους κύκλους, παρατηρούμε ότι δεν υπάρχει ομοιογένεια γιατί περιλαμβάνει και τετράγωνα (Εικόνα 8(d)). Αυτό σημαίνει ότι πρέπει να εφαρμοστεί η παραπάνω διαδικασία αναδρομικά και έτσι δημιουργούνται δύο καινούργιες συστάδες, οι οποίες είναι ομοιογενείς (Εικόνα 8(e)). Κριτήριο τερματισμού του αλγορίθμου RHC είναι η ανεύρεση ομοιογενών συστάδων. Κατά συνέπεια, αποθηκεύονται οι δύο νέες μέσες τιμές ως πρότυπα στο σύνολο συμπύκνωσης. Το τελικό σύνολο συμπύκνωσης περιέχει πλέον τρία πρότυπα αντί των είκοσι έξι (26) στιγμιότυπων που υπήρχαν στο αρχικό σύνολο εκπαίδευσης (Εικόνα 8(f)).



Εικόνα 8: Λειτουργία Μείωσης μέσω Ομοιογενών Συστάδων - RHC

Όπως καταλαβαίνουμε από το παραπάνω παράδειγμα ο αλγόριθμος RHC δημιουργεί περισσότερα πρότυπα σε σχέση με τον αριθμό των κλάσεων όταν τα στιγμιότυπα του συνόλου εκπαίδευσης βρίσκονται σε κοντινές θέσεις στα όρια των κλάσεων και λιγότερα πρότυπα όταν τα στιγμιότυπα του συνόλου εκπαίδευσης βρίσκονται συγκεντρωμένα πιο κοντά στο κέντρο των κλάσεων. Κατά συνέπεια όσο πιο πολλές κλάσεις υπάρχουν και αν υπάρχει πολύς θόρυβος στα δεδομένα, τόσο πιο πολλά συνοριακά στιγμιότυπα θα βρίσκονται στο σύνολο εκπαίδευσης επομένως επιτυγχάνεται χαμηλότερος λόγος μείωσης. Είναι εύκολα κατανοητό πως ο ορισμός της μέσης τιμής μιας κλάσης σαν αρχικό μέσο για την λειτουργία του k-means αυξάνει κατά πολύ την πιθανότητα εύκολης και γρήγορης εύρεσης μεγάλων ομοιογενών συστάδων, πετυχαίνοντας έτσι υψηλό λόγο μείωσης του συνόλου εκπαίδευσης. Ειδικά όταν το σύνολο εκπαίδευσης έχει χαμηλό ποσοστό θορύβου, τότε δημιουργούνται μεγάλες συστάδες και γίνεται πολύ πιο γρήγορος και εύκολος ο εντοπισμός προτύπων για το σύνολο συμπύκνωσης. Αντίθετα, αν το σύνολο εκπαίδευσης έχει υψηλό ποσοστό θορύβου, τότε δημιουργούνται μικρές συστάδες και γίνεται χρονοβόρος και πιο δύσκολος ο εντοπισμός προτύπων για το σύνολο συμπύκνωσης με αποτέλεσμα να επιτυγχάνεται έτσι χαμηλός λόγος μείωσης του συνόλου εκπαίδευσης.

Στη συνέχεια παρουσιάζεται ο αλγόριθμος RHC σε μία μη αναδρομική έκδοσή του. Η λειτουργία του βασίζεται στην χρήση της δομής δεδομένων Ουράς την οποία χρησιμοποιεί

για να διατηρεί ανεξάρτητες τις συστάδες που δημιουργούνται. Για αρχή θεωρεί ότι το σύνολο εκπαίδευσης αποτελείται από μία μη επεξεργασμένη συστάδα C και την οποία τοποθετεί στην ουρά (Queue-Γραμμή 3). Στη συνέχεια ξεκινά μία επαναληπτική διαδικασία κατά την οποία παίρνει από την αρχή της ουράς την πρώτη συστάδα και ελέγχει αν είναι ομοιογενή ή όχι. Αν είναι ομοιογενής (Γραμμή 8) τότε υπολογίζει την μέση τιμή των στιγμιότυπων της και τοποθετεί το κέντρο της στο σύνολο συμπύκνωσης ως πρότυπο (Γραμμή 10) και παράλληλα αφαιρούνται τα στιγμιότυπά της από το σύνολο εκπαίδευσης. Διαφορετικά, αν δεν είναι ομοιογενής, ο τότε ο αλγόριθμος υπολογίζει μία νέα λίστα M που περιέχει τις μέσες τιμές των κλάσεων που ανήκουν στη συγκεκριμένη συστάδα (Γραμμές 13-16). Στη συνέχεια καλεί τον αλγόριθμο συσταδοποίησης k-means δίνοντάς του παραμετρικά την τρέχουσα μη ομοιογενή συστάδα και την λίστα των μέσων των κλάσεων M που δημιουργήθηκαν για να τα χρησιμοποιήσει ως αρχικά μέσα. Ο αλγόριθμος k-means παράγει νέες ανεξάρτητες συστάδες (NewClusters - Γραμμή 17) και τοποθετούνται όλες μαζί στην ουρά Queue (Γραμμές 18-20). Η διαδικασία επαναλαμβάνεται μέχρις ότου να μην τοποθετούνται πλέον μη ομοιογενείς συστάδες στην ουρά δηλαδή να έχει επιτευχθεί ομοιογένεια σε όλες τις συστάδες.

Αλγόριθμος RHC - Reduction through Homogeneous Clusters

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

Ο αλγόριθμος RHC έχει σχεδιαστεί έτσι ώστε να βελτιώνει τις αδυναμίες των υπολοίπων αλγορίθμων και συγκεντρώνει κάποια θετικά στοιχεία σε σχέση με τους υπόλοιπους αλγόριθμους που έχουμε δει. Ένα βασικό πλεονέκτημά του είναι ότι σε αντίθεση με άλλους αλγόριθμους, όπως για παράδειγμα τον CNN και IB2, στηρίζει την αποτελεσματικότητά του στο γεγονός ότι η απόδοση του δεν εξαρτάται από τη σειρά των στιγμιότυπων του συνόλου εκπαίδευσης. Επίσης συνδυάζοντας τη βασική ιδέα του RSP3 της οικογένειας αλγορίθμων RSP [59] (αλγόριθμοι μείωσης με χωρισμό διαστήματος, όπως έχουμε αναφέρει στην παράγραφο 2.2.3) εκμεταλλεύεται τα πλεονεκτήματα τους και αποφεύγει τις αδυναμίες τους. Για παράδειγμα παρόλο που ο RSP3 [51] είναι ένας μη παραμετρικός αλγόριθμος έχει υψηλό κόστος προ-επεξεργασίας εξαιτίας της συνεχόμενης εύρεσης των πιο απομακρυσμένων στιγμιότυπων σε κάθε συστάδα. Αυτό τον κάνει αποτρεπτικό για μεγάλα σύνολα δεδομένων. Αντίθετα ο RHC σαν μη παραμετρικός αλγόριθμος κι αυτός, είναι πιο γρήγορος στηριζόμενος στην παραδοχή του υπολογισμού των μέσων τιμών κάθε κλάσης για αρχικοποίηση και των μέσων τιμών κάθε συστάδας κατά την εκτέλεση, κάνοντας χρήση του αλγορίθμου συσταδοποίησης k-means.

3.3 Ο αλγόριθμος Επεξεργασίας και Μείωσης μέσω Ομοιογενών Συστάδων - ERHC

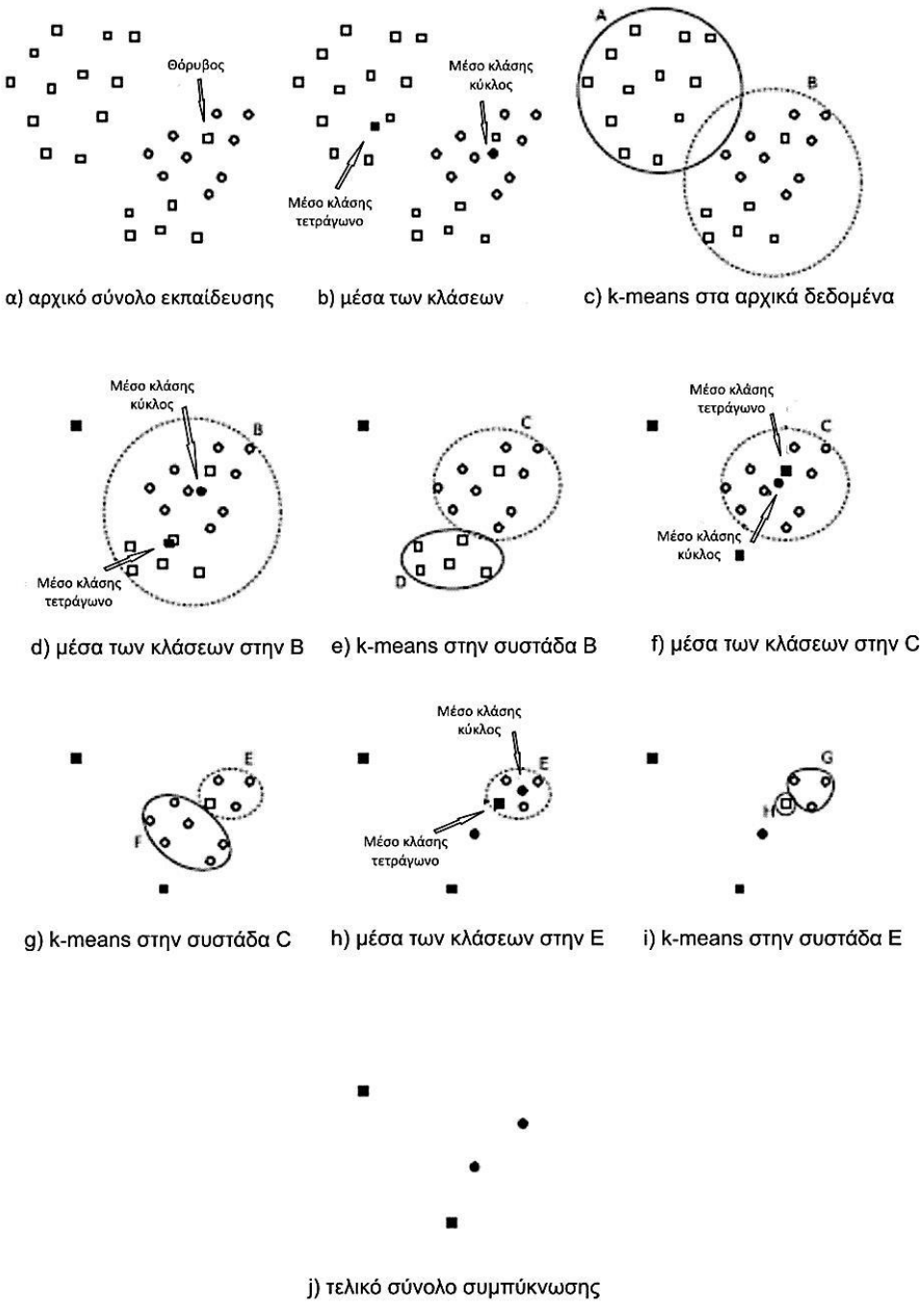
Όπως έχει αναφερθεί στην παράγραφο 2.2, οι αλγόριθμοι επεξεργασίας προσπαθούν να βελτιώσουν την ποιότητα του συνόλου εκπαίδευσης απομακρύνοντας τον θόρυβο, τα λανθασμένα στιγμιότυπα καθώς και τα ακραία στιγμιότυπα που βρίσκονται στα όρια μιας κλάσης. Στόχος τους είναι να δημιουργήσουν ένα σύνολο εκπαίδευσης το οποίο θα χωρίζει τις κλάσεις έτσι ώστε να μην υπάρχουν επικαλύψεις στα όρια των κλάσεων. Η εφαρμογή τους βέβαια αποτελεί ένα αρνητικό σημείο μιας και αυξάνουν το κόστος προ-επεξεργασίας των δεδομένων.

Παρόλα αυτά, αναπτύχθηκε ο αλγόριθμος Επεξεργασίας και Μείωσης μέσω Ομοιογενών Συστάδων – ERHC, ο οποίος είναι ένας αποτελεσματικός, μη παραμετρικός αλγόριθμος παραγωγής προτύπων και εύκολος στην χρήση του. Η βασική ιδέα του είναι να εφαρμόζει αναδρομικά το γνωστό αλγόριθμο συσταδοποίησης k-means. Επεκτείνει την χρήση του αλγορίθμου συσταδοποίησης k-means, κατασκευάζοντας συνεχώς συστάδες μέχρις ότου να πετύχει την ομοιογένεια σε όλες τις συστάδες.

Η λειτουργικότητα του έχει ως εξής: Για αρχή κάνει την παραδοχή ότι όλο το σύνολο εκπαίδευσης είναι μία μη ομοιογενής συστάδα. Ξεκινάει υπολογίζοντας την μέση τιμή για κάθε κλάση, βρίσκοντας τον μέσο όρο των χαρακτηριστικών των αντίστοιχων στιγμιότυπων στο σύνολο εκπαίδευσης. Επομένως, αν ένα σύνολο εκπαίδευσης έχει n κλάσεις τότε θα δημιουργηθούν n μέσες τιμές που θα αντιπροσωπεύουν τις n αρχικές ομάδες ή συστάδες. Στη συνέχεια ο ERHC, χρησιμοποιώντας τις μέσες τιμές που έχει υπολογίσει ως αρχικά κέντρα, κάνει συσταδοποίηση των στιγμιότυπων του συνόλου εκπαίδευσης καλώντας τον αλγόριθμο συσταδοποίησης k-means με αποτέλεσμα από τις n αρχικά μέσες τιμές, να προκύψουν n αρχικά συστάδες. Εάν κάποια συστάδα είναι ομοιογενής τότε παίρνει το κέντρο της και το τοποθετεί στο σύνολο συμπύκνωσης ως πρότυπο. Εάν μία συστάδα αποτελείται από ένα μόνο στιγμιότυπο τότε αυτό θεωρείται θόρυβος και αφαιρείται από το σύνολο εκπαίδευσης. Για κάθε άλλη μη ομοιογενής συστάδα, με περισσότερα από ένα στιγμιότυπα, επαναλαμβάνει αναδρομικά την παραπάνω διαδικασία. Ο αλγόριθμος ERHC τερματίζεται όταν θα επιτευχθεί η ομοιογένεια όλων των συστάδων. Κατά συνέπεια το σύνολο συμπύκνωσης αποτελείται από όλα τα μέσα των ομοιογενών συστάδων. Καταλαβαίνουμε πως η χρήση της μέσης τιμής για κάθε κλάση στα αρχικά n μέσα κάθε κλάσης έχει ως αποτέλεσμα την αυτόματη δημιουργία των n συστάδων. Η ομοιότητα λοιπόν των αλγορίθμων RHC και ERHC είναι προφανής, διαφέρουν μόνο ως προς τον τρόπο αντιμετώπισης των ομοιογενών συστάδων που περιλαμβάνουν μόνο ένα στιγμιότυπο.

Για να γίνει κατανοητός ο τρόπος εκτέλεσης του ERHC παραθέτουμε ένα παράδειγμα εκτέλεσής του στον δισδιάστατο χώρο. Θεωρούμε ότι έχουμε ένα σύνολο δεδομένων το οποίο αποτελείται από είκοσι έξι στιγμιότυπα τα οποία υπάγονται σε δύο κλάσεις: τετράγωνα και κύκλοι (δεκαεπτά τετράγωνα και εννέα κύκλοι – Εικόνα 9(α)). Ο αλγόριθμος ERHC υπολογίζει μία μέση τιμή (κέντρο) για την κλάση που αναφέρεται στα τετράγωνα και μία δεύτερη μέση τιμή (κέντρο) για την κλάση που αναφέρεται στους κύκλους, όπως φαίνονται στην Εικόνα 9(b) με μαύρο τετράγωνο και μαύρο κύκλο. Στην συνέχεια καλείται ο αλγόριθμος συσταδοποίησης k-means, ο οποίος χρησιμοποιεί τα δύο κέντρα των κλάσεων ως αρχικά κέντρα και κατασκευάζει έτσι δύο συστάδες, την συστάδα Α και την συστάδα Β. Παρατηρούμε ότι η συστάδα Α περιλαμβάνει μόνο τετράγωνα οπότε έχει ομοιογένεια (Εικόνα 9(c)) με αποτέλεσμα να τοποθετεί το κέντρο που έχει υπολογιστεί στο σύνολο συμπύκνωσης ως πρότυπο της συστάδας Α. Για την δεύτερη συστάδα Β, παρατηρούμε ότι δεν υπάρχει ομοιογένεια γιατί περιλαμβάνει και κύκλους και τετράγωνα (Εικόνα 9(c)). Αυτό σημαίνει ότι ο αλγόριθμος θα εφαρμόσει την παραπάνω διαδικασία αναδρομικά και έτσι θα δημιουργηθούν εκ νέου δύο νέα κέντρα (Εικόνα 9(d)) και με την χρήση του αλγόριθμου συσταδοποίησης k-means δημιουργούνται δύο καινούργιες συστάδες C και D. Η συστάδα D περιλαμβάνει περισσότερα από ένα στιγμιότυπα και είναι όλα

τετράγωνα, δηλαδή έχει ομοιογένεια, κατά συνέπεια τοποθετείται το κέντρο της στο σύνολο συμπύκνωσης ως πρότυπο της συστάδας D. Στην συστάδα C όμως δεν υπάρχει ομοιογένεια με αποτέλεσμα να επαναληφθεί η διαδικασία από την αρχή και να δημιουργηθούν νέα κέντρα (Εικόνα 9(f)), να γίνει εκ νέου συσταδοποίηση από τον αλγόριθμο k-means με αποτέλεσμα να προκύψουν οι νέες συστάδες E και F (Εικόνα 9(g)). Ομοίως το κέντρο της συστάδας F, λόγο ομοιογένειας, τοποθετείται στο σύνολο συμπύκνωσης ως πρότυπο της συστάδας F, ενώ για την συστάδα E, λόγο ανομοιογένειας, υπολογίζονται εκ νέου νέα κέντρα (Εικόνα 9(h)), γίνεται πάλι συσταδοποίηση με αποτέλεσμα να προκύψουν οι συστάδες G και H (Εικόνα 9(i)). Παρατηρούμε ότι και οι δύο συστάδες είναι ομοιογενής οπότε τοποθετείται το κέντρο της συστάδας G στο σύνολο συμπύκνωσης ως πρότυπο της συστάδας G αλλά η συστάδα H θα αφαιρεθεί από το σύνολο, επειδή περιέχει μόνο ένα στιγμιότυπο. Κατά συνέπεια δεν θα μεταφερθεί το κέντρο της στο σύνολο συμπύκνωσης. Ο αλγόριθμος ERHC τερματίζεται όταν δεν θα δημιουργούνται πλέον συστάδες με λιγότερα από δύο στιγμιότυπα. Έτσι προκύπτει το τελικό σύνολο συμπύκνωσης, το οποίο αποτελείται από τέσσερα μόνο στιγμιότυπα (Εικόνα 9(j)). Αυτή είναι η μόνη και ουσιαστική διαφορά του ERHC από τον RHC. Δηλαδή ο RHC θα τοποθετούσε το κέντρο της συστάδας H στο σύνολο συμπύκνωσης. Αυτός είναι και ο λόγος που ο ERHC πετυχαίνει καλύτερη ποιότητα στο σύνολο συμπύκνωσης και σαφώς μεγαλύτερα ποσοστά μείωσης.



Εικόνα 9: Λειτουργία Επεξεργασίας και Μείωσης μέσω Ομοιογενών Συστάδων – ERHC

Στη συνέχεια παρουσιάζεται ο αλγόριθμος ERHC. Η λειτουργία του βασίζεται στην χρήση της δομής δεδομένων Ουράς την οποία χρησιμοποιεί για να διατηρεί ανεξάρτητες τις συστάδες που δημιουργούνται. Για αρχή θεωρεί ότι το σύνολο εκπαίδευσης αποτελείται από μία μη ομοιογενή συστάδα C και την οποία τοποθετεί στην ουρά (Queue – Γραμμές 1-2). Στη συνέχεια ξεκινά μία επαναληπτική διαδικασία κατά την οποία παίρνει από την αρχή της ουράς την πρώτη συστάδα που συναντάει και ελέγχει αν είναι ομοιογενής και δεν αποτελείται από ένα μόνο στιγμιότυπο (Γραμμές 6-7). Αν είναι τότε υπολογίζει την μέση τιμή των

στιγμιότυπών της και τοποθετεί το κέντρο της στο σύνολο συμπύκνωσης ως πρότυπο (Γραμμή 8-9) ενώ παράλληλα αφαιρούνται τα στιγμιότυπά της από το σύνολο εκπαίδευσης. Διαφορετικά, αν δεν είναι ομοιογενής (Γραμμή 11), ο τότε αλγόριθμος υπολογίζει μία νέα λίστα M που περιέχει τις μέσες τιμές των κλάσεων που ανήκουν στη συγκεκριμένη συστάδα (Γραμμές 12-16). Στη συνέχεια καλεί τον αλγόριθμο συσταδοποίησης k-means δίνοντάς του παραμετρικά την τρέχουσα μη ομοιογενή συστάδα και την λίστα των μέσων των κλάσεων M που δημιουργήθηκαν για να τα χρησιμοποιήσει ως αρχικά μέσα (Γραμμές 17-20). Ο αλγόριθμος συσταδοποίησης k-means παράγει νέες ανεξάρτητες συστάδες (NewClusters - Γραμμή 17) και τοποθετούνται όλες μαζί στην ουρά Queue (Γραμμές 18-20). Η διαδικασία επαναλαμβάνεται (Γραμμές 4-22) μέχρις ότου να μην τοποθετούνται πλέον μη ομοιογενείς συστάδες στην ουρά δηλαδή να έχει επιτευχθεί ομοιογένεια σε όλες τις συστάδες και συνεπώς η ουρά να είναι άδεια, οπότε ο αλγόριθμος τερματίζεται.

Αλγόριθμος ERHC - Editing and Reduction through Homogeneous Clusters

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

Ο αλγόριθμος ERHC - Editing and Reduction through Homogeneous Clusters

Προφανώς, ο ERHC έχει αναπτυχθεί σε μεγάλο βαθμό πάνω στον RHC και ενσωματώνει έναν μηχανισμό απομάκρυνσης θορύβου και μειώνει το μέγεθος του συνόλου

εκπαίδευσης. Επομένως, αναμένεται ότι μπορεί να βελτιώσει αποτελεσματικά την απόδοση της κατηγοριοποίησης και ταυτόχρονα να είναι πιο γρήγορος, δεδομένου ότι βασίζεται στην κατηγοριοποίηση k-means καθώς και στην τεχνική του υπολογισμού των αρχικών κέντρων. Επίσης ο ERHC δεν επηρεάζεται από την σειρά των δεδομένων του συνόλου εκπαίδευσης και με βάση τα παραπάνω μπορεί να χαρακτηριστεί και ως υβριδικός αλγόριθμος. Οι πειραματικές μελέτες δείχνουν ότι επιτυγχάνει υψηλότερα ποσοστά μείωσης και υψηλότερο ποσοστό ακρίβειας από τον RHC, ειδικά όταν το σύνολο δεδομένων περιέχει θόρυβο.

Ο ERHC, όπως και ο RHC, είναι καλός αλγόριθμος, βγάζει καλά αποτελέσματα, όπως αναφέρεται σε πολλές αναφορές, έχει πολύ καλή απόδοση και είναι πολύ πιο γρήγορος από τον RSP3 και από τον CNN, έχει πολύ καλό λόγο μείωσης και πετυχαίνει ακρίβεια τόσο υψηλή όσο και ο CNN.

4

Προτεινόμενοι αλγόριθμοι

Μία από τις βασικές αδυναμίες του ERHC είναι ότι δεν μπορεί να ανταπεξέλθει όταν τα σύνολα δεδομένων περιέχουν απούσες τιμές γιατί χρησιμοποιεί τον αλγόριθμο συσταδοποίησης k-means ο οποίος χρησιμοποιεί αριθμητικά δεδομένα και δεν μπορεί να διαχειριστεί τις απούσες τιμές. Για τον λόγο αυτό ψάξαμε στη βιβλιογραφία να βρούμε τρόπους ώστε χωρίς να χρειάζεται συμπλήρωση των απουσών τιμών εκ των προτέρων, να είναι σε θέση ο ERHC να τις διαχειρίζεται από μόνος του. Στην παράγραφο 4.1 προτείνουμε μία παραλλαγή του ERHC ανεκτική στις απούσες τιμές. Η προτεινόμενη μέθοδος ονομάζεται ERHC-PD. Επιπρόσθετα, στην παράγραφο 4.2, προτείνεται μία δεύτερη παραλλαγή του ERHC, η οποία συμπληρώνει τις απούσες τιμές με τη μέθοδο του καταλογισμού του μέσου όρου του κάθε χαρακτηριστικού ανά κλάση. Αυτή η προτεινόμενη μέθοδος ονομάζεται ERHC-IMP. Η απόδοση των δύο παραλλαγών ελέγχεται πειραματικά στο κεφάλαιο 5.

4.1 Μια παραλλαγή του ERHC ανεκτική στις απούσες τιμές – ERHC-PD

Μετά από εκτενή έρευνα στη βιβλιογραφία ανακαλύψαμε ότι ένας ενδεδειγμένος τρόπος είναι να χρησιμοποιήσουμε την μερική απόσταση (Partial Distance) ή μερική

ευκλείδεια απόσταση, όπως αλλιώς αποκαλείται. Στην πραγματικότητα αποτελεί μία παραλλαγή του τρόπου υπολογισμού της ευκλείδειας απόστασης μεταξύ δύο στιγμιότυπων, αλλά διαχειρίζεται τις απούσες τιμές. Με βάση την λειτουργικότητά της, εκτιμάται η μέση τιμή m για κάθε κλάση C, υπολογίζοντας το μέσο όρο των n τιμών των χαρακτηριστικών των στιγμιότυπων x_i, όπου i=1,2...|C| που ανήκουν στην συστάδα C. Δηλαδή ο αλγόριθμος υπολογίζει την απόσταση μεταξύ δύο στιγμιότυπων λαμβάνοντας υπόψη στον υπολογισμό του μέσου όρου των τιμών των χαρακτηριστικών, μόνο εκείνες τις τιμές που δεν είναι απούσες. Για παράδειγμα, όπως αναφέρεται και στο scikit-learn.org [60], υποθέστε ότι έχουμε δύο στιγμιότυπα A και B και ότι NB είναι ο αριθμός των χαρακτηριστικών που είναι απόντα (στο ένα στιγμιότυπο ή στο άλλο ή και στα δύο), οπότε η απόσταση μεταξύ των στιγμιότυπων υπολογίζεται ως εξής:

$$\mathrm{DAB}_j = \begin{cases} 0, & \text{if } X_{a,j} \text{ or } X_{b,j} \text{ is blank} \\ (X_{a,j} - X_{b,j}), & \text{otherwise} \end{cases}$$

$$\mathrm{DISTANCE}_{ab} = \frac{N}{N - NB} \sum_{j=1}^{N} (\mathrm{DAB}_j)^2.$$

Το αποτέλεσμα είναι το τετράγωνο της ευκλείδειας απόστασης αν δεν υπάρχουν απόντα χαρακτηριστικά. Πιο απλά, υπολογίζει την ευκλείδεια απόσταση κάθε ζεύγους ομοειδών στιγμιότυπων X και Y. Κατά τον υπολογισμό της απόστασης μεταξύ ενός ζεύγους στιγμιότυπων, η τυποποίηση αυτή αγνοεί τα χαρακτηριστικά που έχουν απούσα τιμή σε κάθε στιγμιότυπο και αυξάνει το βάρος των υπόλοιπων χαρακτηριστικών ως εξής:

$$Απόσταση(x, y) = \sqrt{βάρος * (\sum_{j=1}^{N} (x_{a,j} - x_{b,j})^2)}$$

όπου,

$$βάρος = \frac{\text{Πλήθος χαρακτηριστικών (N)}}{\text{Πλήθος υπολογίσιμων χαρακτηριστικών (N − NB)}}$$

Για παράδειγμα, η απόσταση μεταξύ του στιγμιότυπου A=[3, na, na, 6] και του στιγμιότυπου B=[1, na, 4, 5] είναι:

$$Απόσταση(A, B) = \sqrt{\frac{4}{(4-2)}((3-1)^2 + (6-5)^2)}$$

Αν για παράδειγμα, η απόσταση μεταξύ του στιγμιότυπου A=[4, na, na, na] και του στιγμιότυπου B=[na, 3, 4, 5] είναι:

$$Απόσταση(A, B) = \sqrt{\frac{4}{(4-4)}(na)^2 + (na)^2)} \quad \rightarrow \text{NaN}$$

Εάν λείπουν όλα τα χαρακτηριστικά ή εάν δεν υπάρχουν κοινά υπάρχουσα χαρακτηριστικά τότε το NaN (Not a Number) επιστρέφεται για το συγκεκριμένο ζεύγος.

Αν για παράδειγμα, η απόσταση μεταξύ του στιγμιότυπου A=[4, 6, 6, 3] και του στιγμιότυπου B=[2, 3, 4, 5] είναι:

$$Απόσταση(A,B) = \sqrt{\frac{4}{(4-0)}((4-2)^2 + (6-3)^2 + (6-4)^2 + (3-5)^2)}$$

Και όπως καταλαβαίνουμε η περίπτωση αυτή συμπίπτει με την ευκλείδεια απόσταση μιας και η παράμετρος βάρος μας δίνει την μονάδα, οπότε έχουμε τον ορισμό της ευκλείδειας απόστασης.

Στον προτεινόμενο αλγόριθμο ERHC-PD τροποποιούμε τον αλγόριθμο ERHC και τον προσαρμόσουμε έτσι ώστε να μπορεί να χρησιμοποιήσει την μερική απόσταση. Οπότε, για τον υπολογισμό του αρχικού κέντρου της κάθε κλάσης, γίνεται υπολογισμός του μέσου στιγμιότυπου για κάθε κλάση, έτσι ώστε όταν υπάρχει ένα στιγμιότυπο που έχει απούσες τιμές τότε δεν λαμβάνεται υπόψη η διάσταση, το χαρακτηριστικό δηλαδή που δεν έχει τιμή. Στην συνέχεια γίνεται χρήση του κατηγοριοποιητή k-means ο οποίος υπολογίζει τις αποστάσεις του κάθε στιγμιότυπου από το πλησιέστερο κέντρο. Για να υπολογισθεί η απόσταση αυτή, γίνεται χρήση της μερικής ευκλείδειας απόστασης. Για παράδειγμα στην Εικόνα 9(b) για τον αρχικό προσδιορισμό των μέσων των κλάσεων (τετράγωνο και κύκλος) υπολογίζονται για κάθε ένα χαρακτηριστικό ο μέσος όρος όλων των στιγμιότυπων τα οποία όμως δεν περιέχουν απούσες τιμές. Αν υποθέσουμε δηλαδή ότι έχουμε ένα σύνολο με χίλια στιγμιότυπα και στο δεύτερο χαρακτηριστικό έχουμε εκατό στιγμιότυπα με απούσες τιμές, τότε ο μέσος όρος για τον υπολογισμό του αρχικού κέντρου θα υπολογιστεί με βάση τα εννιακόσια στιγμιότυπα. Ομοίως αν θεωρήσουμε ότι το τρίτο χαρακτηριστικό έχει πενήντα στιγμιότυπα τα οποία έχουν απούσα τιμή στο συγκεκριμένο χαρακτηριστικό τότε ο μέσος όρος θα υπολογιστεί με βάση τα εννιακόσια πενήντα στιγμιότυπα. Κατά συνέπεια το αντίστοιχο άθροισμα και πλήθος για κάθε ένα χαρακτηριστικό δεν είναι σταθερό και σίγουρα δεν είναι το σύνολο των στιγμιότυπων του συνόλου δεδομένων για τα χαρακτηριστικά με απούσες τιμές. Στη συνέχεια εκτελούμε τον k-means, ο οποίος κάνοντας χρήση της μερικής ευκλείδειας απόστασης, υπολογίζει την απόσταση όλων των στιγμιότυπων της συστάδας από το κέντρο. Στη συνέχεια επαναλαμβάνει την ίδια διαδικασία αλλά μόνο για τις συστάδες που δεν είναι ομοιογενής. Βέβαια όσο μικραίνει το πλήθος των στιγμιότυπων σε κάθε συστάδα είναι πολύ πιθανόν απούσες τιμές να έχει όχι μόνο κάποιο στιγμιότυπο της συστάδας αλλά ακόμη και το κέντρο της συστάδας. Ας υποθέσουμε δηλαδή ότι στην Εικόνα 9(h) και στην συστάδα E τυχαίνει να έχουν απούσα τιμή και τα τρία στιγμιότυπα στο ίδιο χαρακτηριστικό. Κατά συνέπεια και το κέντρο που θα προκύψει υποχρεωτικά στο τρίτο χαρακτηριστικό θα έχει και αυτό απούσα τιμή. Συνεπώς το χαρακτηριστικό αυτό δεν θα ληφθεί καθόλου υπόψη.

Αν υποθέσουμε ακόμα ότι από τα τρία αυτά στιγμιότυπα το ένα έχει απούσα τιμή και στο πέμπτο χαρακτηριστικό τότε από το συγκεκριμένο στιγμιότυπο δεν θα ληφθεί υπόψη ούτε το τρίτο χαρακτηριστικό ούτε και το πέμπτο χαρακτηριστικό για τον υπολογισμό του κέντρου, διότι για το τρίτο χαρακτηριστικό δεν υπάρχει τιμή ούτε στο στιγμιότυπο ούτε στο κέντρο ενώ για το πέμπτο χαρακτηριστικό δεν υπάρχει τιμή στο στιγμιότυπο.

4.2 Μία παραλλαγή του ERHC με καταλογισμό δεδομένων - ERHC-IMP

Ο προτεινόμενος αλγόριθμος ERHC-IMP αποτελεί παραλλαγή του ERHC και ελέγχεται ως προς την απόδοσή του πάνω σε σύνολα δεδομένων που περιέχουν θόρυβο. Δεν χρησιμοποιήσαμε έτοιμα σύνολα τα οποία περιέχουν απούσες τιμές, όπως έχουν παραχθεί και προσομοιώνουν δεδομένα του πραγματικού κόσμου, αλλά μπήκαμε στην διαδικασία να τα προκαλέσουμε εμείς τις απούσες τιμές για να δούμε την αντιμετώπισή τους από τον ERHC-IMP. Στην συνέχεια οι απούσες τιμές των συνόλων δεδομένων συμπληρώθηκαν με τη μέθοδο του καταλογισμού του μέσου όρου των χαρακτηριστικών ανά κλάση. Για να επιτευχθεί αυτό, πραγματοποιήσαμε μία διαδικασία, βάση της οποίας δημιουργήσαμε και εκτελέσαμε για αρχή έναν αλγόριθμο παραγωγής απουσών τιμών, ο οποίος δέχεται ένα σύνολο δεδομένων και ένα ποσοστό επί τις εκατό, με βάση το οποίο θα δημιουργήσει απούσες τιμές στο σύνολο δεδομένων και δημιουργεί ένα νέο σύνολο δεδομένων στο οποίο έχουν αντικατασταθεί με τυχαίο τρόπο οι τιμές κάποιων χαρακτηριστικών με την τιμή -1 (δηλαδή παρουσία απούσας τιμής).

Στη συνέχεια εφαρμόζουμε έναν δεύτερο αλγόριθμο που έχουμε δημιουργήσει, ο οποίος παίρνει το νέο σύνολο δεδομένων που δημιουργήθηκε από τον προηγούμενο αλγόριθμο, το οποίο περιλαμβάνει απούσες τιμές στο ποσοστό που ζητήσαμε και στη συνέχεια διαβάζει όλα τα χαρακτηριστικά από τα στιγμιότυπα του συνόλου δεδομένων και με βάση τις τιμές που δεν περιέχουν απούσες τιμές δημιουργεί τους μέσους όρους για κάθε χαρακτηριστικό της κάθε κλάσης. Στη συνέχεια διαβάζει το σύνολο δεδομένων από την αρχή, εντοπίζει τα χαρακτηριστικά που έχουν απούσες τιμές (δηλαδή όπου υπάρχει το -1), αναγνωρίζει την κλάση πού ανήκει το συγκεκριμένο χαρακτηριστικό και αντικαθιστά την απούσα τιμή με τον αντίστοιχο μέσο όρο του συγκεκριμένου χαρακτηριστικού της αντίστοιχης κλάσης που έχει υπολογίσει νωρίτερα. Έτσι λοιπόν με τυχαίο τρόπο δημιουργούνται απούσες τιμές και γίνεται αντικατάσταση τους με τον αντίστοιχο μέσο όρο του κάθε χαρακτηριστικού της αντίστοιχης κλάσης.

Με βάση λοιπόν την παραπάνω διαδικασία, έχουν δημιουργηθεί για κάθε σύνολο δεδομένων δύο νέα σύνολα. Στο πρώτο σύνολο δεδομένων έχουμε εφαρμόσει τον αλγόριθμο δημιουργίας απουσών τιμών όπου ορίσαμε το ποσοστό του 10% ενώ παράλληλα στο δεύτερο σύνολο δεδομένων έχουμε εφαρμόσει τον ίδιο αλγόριθμο δημιουργίας απουσών τιμών ορισμένο σε ποσοστό 20%. Με την μέθοδο που χρησιμοποιούμε στον δεύτερο αλγόριθμο, κάνουμε αντικατάσταση των απουσών τιμών με τους αντίστοιχους μέσους όρους των τιμών των αντίστοιχων χαρακτηριστικών για την συγκεκριμένη κλάση. Η αναφορά σε κάθε ένα σύνολο δεδομένων γίνεται με βάση το όνομα του συνόλου δεδομένων βάζοντας τον αριθμό 10 ή τον αριθμό 20 στο τέλος, ο οποίος χαρακτηρίζει το ποσοστό των απουσών τιμών που έχουμε εφαρμόσει και καταλογίσει στο συγκεκριμένο σύνολο δεδομένων. Για παράδειγμα για το πρώτο σύνολο δεδομένων που χρησιμοποιούμε με όνομα Balance ή BL έχουμε δημιουργήσει τα σύνολα δεδομένων BL 10 και BL 20.

5

Πειραματική Μελέτη

Σε αυτό το κεφάλαιο παρουσιάζεται η πειραματική μελέτη που εκπονήθηκε στα πλαίσια της παρούσας διπλωματικής εργασίας. Η πειραματική μελέτη βασίστηκε σε 13 γνωστά σύνολα δεδομένων. Στην παράγραφο 5.1 παρουσιάζονται όλες οι ρυθμίσεις και οι παράμετροι που χρησιμοποιήθηκαν για την εκτέλεση των πειραμάτων. Στην συνέχεια, στην παράγραφο 5.2, παρουσιάζονται και αναλύονται τα πειραματικά αποτελέσματα που προέκυψαν και επιχειρείται μία σύγκριση μεταξύ των προτεινόμενων μεθόδων με τις διαφορετικές προσεγγίσεις κατηγοριοποίησης που βασίζονται στην αναζήτηση εγγύτερων γειτόνων. Η σύγκριση πραγματοποιήθηκε ως προς την ακρίβεια κατηγοριοποίησης και τον λόγο μείωσης του πληθυσμού των δεδομένων που επιτυγχάνουν.

5.1 Πειραματικές ρυθμίσεις

Οι αλγόριθμοι που δημιουργήθηκαν για τις ανάγκες της διπλωματικής γράφτηκαν στην γλώσσα C++ και συμμετείχαν στα πειράματά μας με τις παρακάτω συντομογραφίες :

KNN – IMP (ACC): ο κλασικός κατηγοριοποιητής k-NN με εφαρμογή σε σύνολο δεδομένων στο οποίο έχουν συμπληρωθεί οι απούσες τιμές με την μέθοδο του καταλογισμού μέσης τιμής κάθε κλάσης υπολογίζοντας την ακρίβεια-Acc.

KNN – PD (ACC): ο κλασικός k-NN με εφαρμογή σε σύνολο δεδομένων που περιέχει απούσες τιμές κάνοντας χρήση της μερικής απόστασης -partial distance- υπολογίζοντας την ακρίβεια-Acc.

ERHC-IMP (ACC): η προτεινόμενη παραλλαγή του αλγόριθμου ERHC με εφαρμογή σε σύνολο δεδομένων στο οποίο έχουν συμπληρωθεί οι απούσες τιμές με την μέθοδο του καταλογισμού μέσης τιμής κάθε κλάσης υπολογίζοντας την ακρίβεια-Acc.

ERHC-PD (ACC): η προτεινόμενη παραλλαγή του αλγόριθμου ERHC με εφαρμογή σε σύνολο δεδομένων που περιέχει απούσες τιμές κάνοντας χρήση της μερικής απόστασης -partial distance- υπολογίζοντας την ακρίβεια-Acc.

ERHC-IMP (RR): η προτεινόμενη παραλλαγή του αλγόριθμου ERHC με εφαρμογή σε σύνολο δεδομένων στο οποίο έχουν συμπληρωθεί οι απούσες τιμές με την μέθοδο του καταλογισμού μέσης τιμής κάθε κλάσης υπολογίζοντας το ποσοστό μείωσης-RR.

ERHC-PD (RR): η προτεινόμενη παραλλαγή του αλγόριθμου ERHC με εφαρμογή σε σύνολο δεδομένων που περιέχει απούσες τιμές κάνοντας χρήση της μερικής απόστασης -partial distance- υπολογίζοντας το ποσοστό μείωσης-RR.

Ο υπολογιστής στον οποίο εκτελέστηκαν οι αλγόριθμοι και όλα τα πειράματα ήταν ένας επιτραπέζιος υπολογιστής ο οποίος χρησιμοποιούσε έναν i5-Intel επεξεργαστή, μνήμη RAM 4GB και λειτουργικό σύστημα Windows 10. Δεν ήταν και ότι καλύτερο αλλά βοήθησε η υπομονή και η επιμονή για την επιτυχία των πειραμάτων.

Για τον έλεγχο της απόδοσης των αλγορίθμων που μελετήσαμε καθώς και για τον υπολογισμό της ακρίβειάς τους χρησιμοποιήθηκαν δεκατρία σύνολα δεδομένων που διανεμήθηκαν τα περισσότερα από το αποθετήριο KEEL [66] καθώς και από το αποθετήριο UC Irvine Machine Learning Repository (UCI) [67] και παρουσιάζονται συνοπτικά στον πίνακα 1 που βρίσκεται παρακάτω. Οι τιμές που μετρήθηκαν είναι η ακρίβεια (Acc- Accuracy) και το ποσοστό μείωσης (RR- Reduction Rate). Οι τιμές Acc και RR μετρήθηκαν ως ποσοστά (%). Για λόγους σύγκρισης, χρησιμοποιούμε την συμβατική κατηγοριοποίηση k-NN με ορισμό του k στο ένα (k = 1).

Σε όλα τα σύνολα δεδομένων που χρησιμοποιήσαμε έχει προηγηθεί μία επεξεργασία κανονικοποίησης (normalization) κατά την οποία όλες οι τιμές του συνόλου δεδομένων έχουν μετατραπεί στην κλίμακα [0 - 1]. Έχει αποδειχθεί ότι η τιμή της απόστασης μπορεί εύκολα να επηρεαστεί από τα χαρακτηριστικά του συνόλου δεδομένων. Είτε τα χαρακτηριστικά είναι παρόμοια είτε έχουν την ίδια σημασία, η ευρεία κλίμακα που μπορεί να υποστηρίζουν επηρεάζει άμεσα την τιμή της απόστασης. Κατά συνέπεια χαρακτηριστικά που έχουν ευρεία κλίμακα έχουν μεγαλύτερη επίδραση στην τιμή της απόστασης από εκείνα τα

χαρακτηριστικά με μικρότερο εύρος τιμών. Έτσι αν υποθέσουμε ότι χρησιμοποιούμε το χαρακτηριστικό «ταχύτητα κίνησης» είναι κατανοητό πως άλλο είναι το εύρος της ταχύτητας με την οποία κινείται ένα αυτοκίνητο και άλλο εύρος της ταχύτητας κίνησης ενός ποδηλάτη. Αντίστοιχα άλλο το εύρος του χαρακτηριστικού «μισθός» ενός υπαλλήλου και άλλο το εύρος του χαρακτηριστικού «αριθμός παιδιών» ενός υπαλλήλου. Όλα τα χαρακτηριστικά μπορεί να έχουν την ίδια σημασία αλλά έχουν διαφορετικό αντίκτυπο στον υπολογισμό της απόστασης κατά συνέπεια το εύρος των χαρακτηριστικών πρέπει να προσαρμοστεί σε μία ενιαία κλίμακα, να κανονικοποιηθεί δηλαδή σε ένα συγκεκριμένο εύρος διαστήματος (π.χ. [0,1]). Ας υποθέσουμε ότι ένα σύνολο δεδομένων περιέχει n στιγμιότυπα και ένα χαρακτηριστικό e θα πρέπει να κανονικοποιηθεί σε [0,1]. Η τιμή του κάθε χαρακτηριστικού i, όπου i = 1, ..., n κανονικοποιείται ως εξής:

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

όπου οι E_{min} και E_{max} είναι οι ελάχιστες και μέγιστες τιμές για το χαρακτηριστικό e, αντίστοιχα.

Η κανονικοποίηση δεδομένων είναι μια κοινή διαδικασία προ-επεξεργασίας σε πολλές εργασίες εξόρυξης δεδομένων. Ορισμένες εφαρμογές εξόρυξης δεδομένων υιοθετούν την κανονικοποίηση ως προεπιλεγμένη διεργασία.

Οι μετρικές απόστασης που χρησιμοποιήθηκαν είναι η ευκλείδεια απόσταση και η μερική ευκλείδεια απόσταση. Ως γνωστών, δοθέντος ενός αγνώστου σημείου η μέθοδος των k-κοντινότερων γειτόνων βασίζει τις προβλέψεις της στα k πιο κοντινά στιγμιότυπα. Για αυτό το λόγο για να κάνουμε πρόβλεψη θα πρέπει να καθορίσουμε ένα μέτρο που να υπολογίζει την απόσταση ανάμεσα στο άγνωστο σημείο και στα k κοντινότερα στιγμιότυπα του συνόλου δεδομένων. Το πιο σύνηθες μέτρο που χρησιμοποιούμε για να μετρήσουμε αυτή την απόσταση είναι η ευκλείδεια απόσταση. Υπάρχουν όμως και άλλα μέτρα για να μετρήσουμε αυτή την απόσταση, αλλά εμείς θα ασχοληθούμε με την ευκλείδεια απόσταση και με την μερική ευκλείδεια απόσταση.

Ο τρόπος υπολογισμού της ευκλείδειας απόστασης μεταξύ δύο σημείων, χωρίς την παρουσία απουσών τιμών, εκτιμάται υπολογίζοντας το μέσο όρο των n τιμών των χαρακτηριστικών των στιγμιότυπων x_i,i=1,2...|C| που ανήκουν σε μία συστάδα C. Χρησιμοποιούμε την Ευκλείδεια απόσταση ως μέτρο ομοιότητας δύο στιγμιότυπων x_a και x_b εάν όλα τα γνωρίσματα των στιγμιότυπων έχουν αριθμητικές τιμές. Θεωρούμε ότι τα στιγμιότυπων έχουν n γνωρίσματα. Η απόσταση μεταξύ των στιγμιότυπων x_a και x_b συμβολίζεται ως $d(x_a, x_b)$.

Η Ευκλείδεια απόσταση των στιγμιότυπων x_a και x_b δίνεται από τη εξίσωση:

$$d(x_a, x_b) = \sqrt{\sum_{j=1}^{n} (x_{aj} - x_{bj})^2}$$

όπου x_{aj} είναι η τιμή της μεταβλητής j του στιγμιότυπου x_a.

Η Ευκλείδεια απόσταση έχει το πλεονέκτημα ότι η απόσταση μεταξύ δύο οποιονδήποτε στιγμιότυπων δεν επηρεάζεται από τον ύπαρξη στοιχείων με μεγάλες αποστάσεις (ακραίες τιμές).

Γενίκευση της Ευκλείδειας απόστασης είναι η απόσταση Manhattan και η απόσταση Minkowski, όμως δεν αναλύονται μιας και δεν αποτελούν μέτρο υπολογισμού στην παρούσα μεταπτυχιακή εργασία.

Μία παραλλαγή του τρόπου υπολογισμού της ευκλείδειας απόστασης μεταξύ δύο στιγμιότυπων, παρουσία απουσών τιμών, αποτελεί η μερική ευκλείδεια απόσταση, με βάση την οποία εκτιμάται η μέση τιμή για κάθε κλάση, υπολογίζοντας το μέσο όρο των τιμών των χαρακτηριστικών των στιγμιότυπων που ανήκουν στην ίδια συστάδα. Δηλαδή υπολογίζει την απόσταση μεταξύ δύο στιγμιότυπων λαμβάνοντας υπόψη στον υπολογισμό του μέσου όρου των τιμών των χαρακτηριστικών, μόνο εκείνες τις τιμές που δεν είναι απούσες. Εάν δεν υπάρχουν απούσες τιμές στα χαρακτηριστικά τότε το αποτέλεσμα είναι το τετράγωνο της ευκλείδειας απόστασης. Πιο απλά, υπολογίζει την ευκλείδεια απόσταση κάθε ζεύγους ομοειδών στιγμιότυπων X και Y. Κατά τον υπολογισμό της απόστασης μεταξύ ενός ζεύγους στιγμιότυπων, η τυποποίηση αυτή αγνοεί τα χαρακτηριστικά που έχουν απούσα τιμή και αυξάνει το βάρος των υπολοίπων συντεταγμένων.

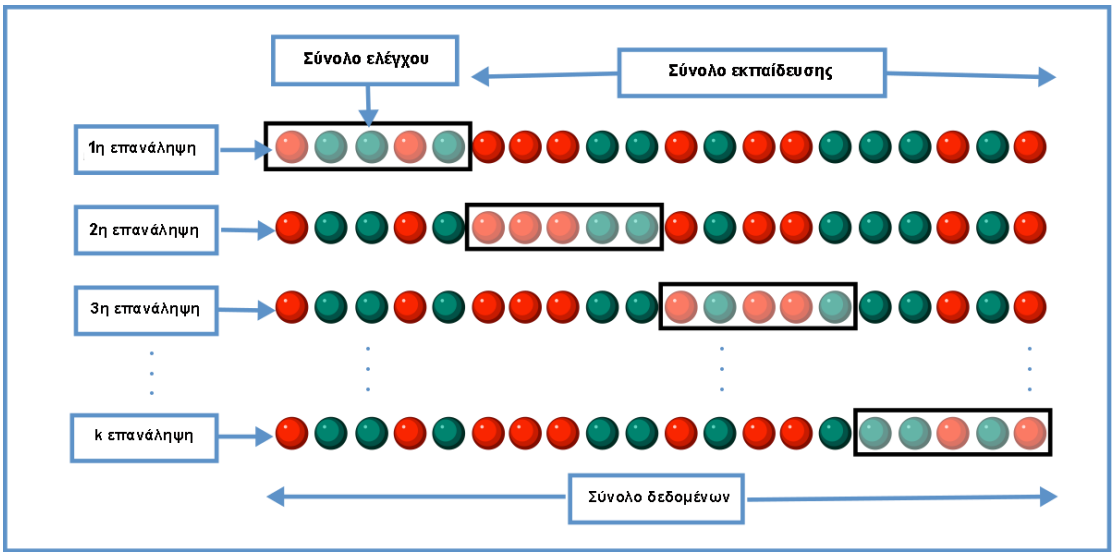
Για την πρόβλεψη και την εκτίμηση του πόσο ακριβή θα είναι ένα μοντέλο πρόβλεψης στην πράξη χρησιμοποιήσαμε την μέθοδο διασταυρούμενης επικύρωσης (cross validation) [61]. Πρόκειται για μία τεχνική επικύρωσης ενός μοντέλου για την αξιολόγηση των αποτελεσμάτων μιας στατιστικής ανάλυσης κατά πως θα γενικευτούν σε ένα ανεξάρτητο σύνολο δεδομένων.

Σε ένα πρόβλημα πρόβλεψης, ένα μοντέλο λαμβάνει συνήθως δύο σύνολα δεδομένων. Το ένα είναι το σύνολο δεδομένων στο οποίο εκτελείται η εκπαίδευση (σύνολο γνωστών δεδομένων εκπαίδευσης) και το άλλο ένα σύνολο δεδομένων ελέγχου (ή άγνωστων δεδομένων που τα συναντάμε πρώτη φορά) έναντι των οποίων δοκιμάζεται το μοντέλο (ονομάζεται σύνολο δεδομένων επικύρωσης ή δοκιμής). Ο στόχος της διασταυρούμενης επικύρωσης είναι να δοκιμάσει την ικανότητα του μοντέλου να προβλέψει νέα δεδομένα που

δεν χρησιμοποιήθηκαν για την εκτίμησή του, προκειμένου να επισημανθούν τυχόν προβλήματα και να δοθεί μια εικόνα για το πώς το μοντέλο θα γενικευθεί σε ένα ανεξάρτητο σύνολο δεδομένων (δηλαδή σε ένα άγνωστο σύνολο δεδομένων που πιθανώς να αποτελεί ένα πραγματικό πρόβλημα).

Σύμφωνα με τη διαδικασία αυτή το σύνολο των δεδομένων χωρίζεται σε k υποσύνολα ίσου μεγέθους. Κάθε φορά, ένα υποσύνολο από αυτά χρησιμοποιείται ως σύνολο ελέγχου, ενώ τα υπόλοιπα k-1 υποσύνολα αποτελούν το σύνολο εκπαίδευσης που χρησιμοποιείται για την ανάπτυξη του μοντέλου. Συνοπτικά, η διασταυρούμενη επικύρωση συνδυάζει μέτρα καταλληλότητας στην πρόβλεψη (μέσους όρους) για να αντλήσει μια πιο ακριβή εκτίμηση της απόδοσης πρόβλεψης του μοντέλου.

Όταν η διαδικασία διασταυρούμενης επικύρωσης επαναλαμβάνεται k φορές, γνωστή και ως k-fold cross validation, τότε καθένα από τα k υποσύνολα χρησιμοποιείται ακριβώς μία φορά ως σύνολο επικύρωσης. Τα k αποτελέσματα μπορούν στη συνέχεια και υπολογίζονται κατά μέσο όρο για να παράγουν μία μόνο εκτίμηση. Το πλεονέκτημα αυτής της μεθόδου είναι ότι όλα τα δεδομένα χρησιμοποιούνται τόσο για εκπαίδευση όσο και για επικύρωση και κάθε δεδομένο για επικύρωση ακριβώς μία φορά. Συνήθως χρησιμοποιείται 10πλή διασταυρούμενη επικύρωση, αλλά γενικά το k παραμένει μια μη καθορισμένη παράμετρος. Για τις ανάγκες αυτής της διπλωματικής χρησιμοποιήθηκε για k η τιμή πέντε (k=5). Στην παρακάτω Εικόνα 10 παρουσιάζεται σχηματικά η διαδικασία της διασταυρούμενης επικύρωσης k-fold (k-fold cross validation).



Εικόνα 10: Σχηματική αναπαράσταση της k-fold cross validation

Στην διαδικασία k-fold cross validation τα δεδομένα αρχικά διαμερίζονται σε k τυχαία υποσύνολα (folds) f1, f2, .., fk ίδιου μεγέθους. Από αυτά τα υποσύνολα επιλέγεται ένα τυχαία και οι τιμές που περιέχει θα αποτελέσουν το σύνολο επικύρωσης ενώ τα υπόλοιπα υποσύνολα θα χρησιμοποιηθούν αποκλειστικά και μόνο για εκπαίδευση. Αξιοποιώντας τον διαχωρισμό των υποσυνόλων, υπάρχει η δυνατότητα εκπαίδευσης και επικύρωσης του κατηγοριοποιητή με διαφορετικά δεδομένα κάθε φορά από το ίδιο όμως αρχικό σετ δεδομένων. Με αυτόν τον τρόπο μελετάται πιο συστηματικά η συμπεριφορά του μοντέλου κατηγοριοποίησης.

Περιγραφή Datasets

Στο σημείο αυτό θα γίνει μία σύντομη αναφορά στο τι αντιπροσωπεύει κάθε ένα σύνολο δεδομένων που έχουμε χρησιμοποιήσει στα πειράματά μας. Αλφαβητικά λοιπόν είναι τα εξής:

Balance Scale Data Set [68]: είναι ένα σύνολο δεδομένων που δημιουργήθηκε για να μοντελοποιήσει ψυχολογικά πειραματικά αποτελέσματα. Κάθε παράδειγμα ταξινομείται ότι έχει το άκρο της κλίμακας ισορροπίας προς τα δεξιά, το άκρο προς τα αριστερά ή είναι ισορροπημένο. Τα χαρακτηριστικά είναι το αριστερό βάρος, η αριστερή απόσταση, το σωστό βάρος και η σωστή απόσταση. Ο σωστός τρόπος για να βρείτε την κλάση είναι ο μεγαλύτερος (αριστερή απόσταση * αριστερό βάρος) και (δεξιά απόσταση * δεξιά βάρος). Εάν είναι ίσοι, είναι ισορροπημένο.

Ecoli Data Set [69]: Ο στόχος από την χρήση αυτού του συνόλου είναι η πρόβλεψη της θέσης εντοπισμού των πρωτεϊνών χρησιμοποιώντας ορισμένα μέτρα σχετικά με το κύτταρο (κυτταρόπλασμα, εσωτερική μεμβράνη, υπερπλασία, εξωτερική μεμβράνη, λιποπρωτεΐνη εξωτερικής μεμβράνης, εσωτερική μεμβράνη λιποπρωτεΐνης εσωτερικής μεμβράνης, διασπώμενη αλληλουχία σήματος).

Kdd Cup Data Set [70]: Πρόκειται για ένα υποσύνολο 10% του συνόλου δεδομένων που χρησιμοποιήθηκε για τον διαγωνισμό Third Third Knowledge Discovery και Data Mining Tools, ο οποίος πραγματοποιήθηκε σε συνδυασμό με το KDD-99 Το πέμπτο διεθνές συνέδριο για τη γνώση και την εξόρυξη δεδομένων. Η αποστολή ανταγωνισμού ήταν να δημιουργήσει έναν ανιχνευτή εισβολής δικτύου, ένα μοντέλο πρόβλεψης ικανό να διακρίνει μεταξύ κακών συνδέσεων, που ονομάζονται εισβολές ή επιθέσεις και καλών κανονικών συνδέσεων, χρησιμοποιώντας τόσο ονομαστικά όσο και συνεχή δεδομένα.

Statlog (Landsat Satellite) Data Set [71]: Η βάση δεδομένων αποτελείται από τις πολυφασματικές τιμές των pixel σε γειτονιές 3x3 σε μια δορυφορική εικόνα και την σχετική

κατηγοριοποίησή τους με το κεντρικό pixel σε κάθε γειτονιά. Ο στόχος είναι να προβλεφθεί αυτή η κατηγοριοποίηση, δεδομένων των πολυφασματικών τιμών. Στη βάση δεδομένων του δείγματος, η κλάση ενός pixel κωδικοποιείται ως αριθμός. Τα δορυφορικά δεδομένα Landsat είναι μία από τις πολλές πηγές πληροφοριών που διατίθενται για μια σκηνή. Η ερμηνεία μιας σκηνής ενσωματώνοντας χωρικά δεδομένα διαφορετικών τύπων και αναλύσεων, συμπεριλαμβανομένων πολυφασματικών και δεδομένων από ραντάρ, χάρτες που υποδεικνύουν τοπογραφία, χρήση γης κ.λπ. αναμένεται να αποκτήσει σημαντική σημασία με την έναρξη μιας εποχής που χαρακτηρίζεται από ολοκληρωμένες προσεγγίσεις στην τηλεπισκόπηση (για παράδειγμα , Το Σύστημα Παρατήρησης της Γης της NASA ξεκινά αυτή τη δεκαετία). Οι υπάρχουσες στατιστικές μέθοδοι δεν είναι καλά εξοπλισμένες για το χειρισμό τέτοιων διαφορετικών τύπων δεδομένων. Σημειώστε ότι αυτό δεν ισχύει για δεδομένα Landsat MSS που εξετάζονται μεμονωμένα (όπως σε αυτό το δείγμα βάσης δεδομένων). Αυτά τα δεδομένα ικανοποιούν τις σημαντικές απαιτήσεις της αριθμητικής και σε μία μόνο ανάλυση, και η τυπική κατηγοριοποίηση μέγιστης πιθανότητας αποδίδει πολύ καλά. Κατά συνέπεια, για αυτά τα δεδομένα, θα πρέπει να είναι ενδιαφέρον να συγκρίνουμε την απόδοση άλλων μεθόδων με τη στατιστική προσέγγιση.

Letter Image Recognition Data Set [72]: Ο στόχος είναι να προσδιοριστεί καθένα από έναν μεγάλο αριθμό ασπρόμαυρων ορθογώνιων εικονοστοιχείων ως ένα από τα 26 κεφαλαία γράμματα στο αγγλικό αλφάβητο. Οι εικόνες των χαρακτήρων βασίστηκαν σε 20 διαφορετικές γραμματοσειρές και κάθε γράμμα σε αυτές τις 20 γραμματοσειρές παραμορφώθηκε τυχαία για να παράγει ένα αρχείο 20.000 μοναδικών καταχωρήσεων. Κάθε καταχώρηση μετατράπηκε σε 16 αριθμητικά χαρακτηριστικά (στατιστικές στιγμές και μετρήσεις ακραίων τιμών) τα οποία στη συνέχεια κλιμακώθηκαν ώστε να χωρέσουν σε μια σειρά ακέραιων τιμών από 0 έως 15.

MAGIC Gamma Telescope Data Set [73]: Αυτό το σύνολο δεδομένων περιέχει παραγόμενα δεδομένα που δημιουργούνται από την προσομοίωση της καταγραφής σωματιδίων γ υψηλής ενέργειας από ένα επίγειο τηλεσκόπιο Cherenkov, χρησιμοποιώντας τη τεχνική της απεικόνισης. Το τηλεσκόπιο Cherenkov παρατηρεί ακτίνες γ υψηλής ενέργειας, εκμεταλλεύοντας την ακτινοβολία που εκπέμπεται από φορτισμένα σωματίδια που παράγονται μέσα στα ηλεκτρομαγνητικά φάσματα που ξεκινούν οι ακτίνες γ και αναπτύσσονται στην ατμόσφαιρα. Αυτή η ακτινοβολία Cherenkov (ορατή σε μήκη κύματος UV) διαρρέει από την ατμόσφαιρα και καταγράφεται στον ανιχνευτή, επιτρέποντας την ανακατασκευή των παραμέτρων του φάσματος. Οι διαθέσιμες πληροφορίες αποτελούνται από παλμούς που αφήνονται από τα εισερχόμενα φωτόνια Cherenkov στους σωλήνες φωτοπολλαπλασιαστή, τοποθετημένα σε ένα επίπεδο, την κάμερα. Ανάλογα με την ενέργεια της πρωτογενούς ακτίνας γ, συλλέγονται συνολικά μερικές εκατοντάδες έως περίπου 10000

φωτόνια Cherenkov, σε μοτίβα (που ονομάζεται εικόνα φάσματος), επιτρέποντας να διακρίνονται στατιστικά εκείνα που προκαλούνται από πρωτογενή ακτίνα γ (σήμα) από τις εικόνες αδρονικών φασμάτων που ξεκίνησαν από τις αχανής ακτίνες στην ανώτερη ατμόσφαιρα (φόντο, ετικέτα κλάσης h). Το σύνολο δεδομένων δημιουργήθηκε από ένα πρόγραμμα Monte Carlo, Corsika, που περιγράφεται στο: D. Heck et al., CORSIKA, A Monte Carlo code to simulate extensive air showers, Forschungszentrum Karlsruhe FZKA 6019 (1998). Ο στόχος είναι η διάκριση στατιστικών εικόνων που δημιουργούνται από πρωτογενή ακτίνες γ (σήμα, ετικέτα κλάσης g) από τις εικόνες αδρονικών ντους που ξεκινούν από κοσμικές ακτίνες στην ανώτερη ατμόσφαιρα (φόντο, ετικέτα κλάσης h).

Pen-Based Recognition of Handwritten Digits Data Set [74]: Μια ψηφιακή βάση δεδομένων που δημιουργήθηκε με τη συλλογή 250 δειγμάτων από 44 συγγραφείς, χρησιμοποιώντας μόνο (x, y) πληροφορίες συντεταγμένων που αντιπροσωπεύονται ως διανύσματα χαρακτηριστικών σταθερού μήκους, τα οποία επαναπροσδιορίστηκαν σε 8 σημεία ανά ψηφίο (επομένως το σύνολο δεδομένων περιέχει 8 σημεία x 2 συντεταγμένες = 16 γνωρίσματα). Η κλάση αντιπροσωπεύει τον κωδικό του ψηφίου που γράφτηκε.

Phoneme data set [75]:Ο στόχος αυτού του συνόλου δεδομένων είναι να γίνει διάκριση μεταξύ ρινικών (κλάσης 0) και προφορικών ήχων (κλάση 1). Η κατανομή των κλάσεων είναι 3.818 στιγμιότυπα στην κλάση 0 και 1.586 στιγμιότυπα στην κλάση 1.

Pima (PM) Indians Diabetes data set [76]: Ο στόχος αυτού του συνόλου δεδομένων είναι να γίνει πρόβλεψη για την πιθανή δημιουργία διαβήτη βάσει διαγνωστικών μετρήσεων. Προέρχεται από το Εθνικό Ινστιτούτο Διαβήτη και Πεπτικού και Νεφροπάθειες. Αρκετοί περιορισμοί τέθηκαν κατά την επιλογή αυτών των περιπτώσεων από μια πολύ μεγάλη βάση δεδομένων. Συγκεκριμένα, όλοι οι ασθενείς εδώ είναι γυναίκες ηλικίας τουλάχιστον 21 ετών από την ινδική κληρονομιά της Πίμα.

Statlog (Shuttle) Data Set [77]: Το σύνολο δεδομένων shuttle περιέχει 9 χαρακτηριστικά όλα τα οποία είναι αριθμητικά. Περίπου το 80% των δεδομένων ανήκει στην κλάση 1. Επομένως, η προεπιλεγμένη ακρίβεια είναι περίπου 80%. Ο στόχος εδώ είναι να επιτευχθεί ακρίβεια 99 - 99,9%. Τα στιγμιότυπα στο αρχικό σύνολο δεδομένων ήταν σε χρονολογική σειρά, και αυτή η χρονολογική σειρά θα μπορούσε πιθανόν να είναι σχετική με την κατηγοριοποίηση. Ωστόσο, αυτό δεν θεωρήθηκε σωστό για τους σκοπούς του StatLog, επομένως η σειρά των στιγμιότυπων στο αρχικό σύνολο δεδομένων άλλαξε με τυχαίο ρυθμό και ένα τμήμα του αρχικού συνόλου δεδομένων καταργήθηκε για λόγους επικύρωσης.

Texture (TXT) data set [78]: Ο στόχος αυτού του συνόλου δεδομένων είναι να γίνει διάκριση μεταξύ 11 διαφορετικών υφών (Grass lawn, Pressed calf leather, Handmade paper, Raffia looped to a high pile, Cotton canvas, ...), κάθε σχέδιο (pixel) περιγράφεται από 40

χαρακτηριστικά με την εκτίμηση των τεσσάρων διαφορετικών κλάσεων ανά προσανατολισμό της τάξης των 0, 45, 90 και 135 μοιρών αντίστοιχα.

Twonorm (TN) data set [79]: Πρόκειται για μια εφαρμογή του Twonorm παραδείγματος του Leo Breiman (Breiman L. Bias, variance and arcing classifiers. Tec. Report 460, Statistics department. University of california. April 1996). Είναι ένα 20-διάστατο παράδειγμα κατηγοριοποίησης δύο κατηγοριών. Κάθε κλάση προέρχεται από μια κανονική διανομή πολλαπλών παραλλαγών με διακύμανση μονάδας. Η κλάση 1 έχει μέση τιμή (a, a, .. a) ενώ η κλάση 2 έχει μέση τιμή (-a, -a, ..- a), όπου a = 2/sqrt(20). Ο Breiman αναφέρει το θεωρητικό αναμενόμενο ποσοστό εσφαλμένης κατηγοριοποίησης ως 2,3%. Χρησιμοποίησε 300 στιγμιότυπα εκπαίδευσης με το CART και βρήκε σφάλμα 22,1%.

Yeast Data Set [80]: Ο στόχος αυτού του συνόλου δεδομένων είναι να γίνει πρόβλεψη των κυτταρικών περιοχών εντοπισμού πρωτεϊνών. Το σύνολο Yeast αποτελείται από 1484 στιγμιότυπα με 8 χαρακτηριστικά γνωρίσματα μετρήσεων και αναλύσεων σημάτων για περιοχές εντοπισμού πρωτεϊνών.

Συνοπτική παρουσίαση των συνόλων δεδομένων με τις βασικές πληροφορίες τους φαίνεται στον Πίνακα 1 που ακολουθεί.

Σύνολο Δεδομένων		**Μέγεθος**	**Χαρακτηριστικά**	**Κλάσεις**
1.	Balance (BL)	625	4	3
2.	Ecoli (ECL)	336	7	8
3.	KddCup (KDD)	141481	36	23
4.	Landsat Satellite (LS)	6435	36	6
5.	Letter Image Recognition (LIR)	20000	16	26
6.	Magic Gamma Telescope (MGT)	19020	10	2
7.	Pen-Digits (PD)	10992	16	10
8.	Phoneme (PH)	5404	5	2
9.	Pima (PM)	768	8	2
10.	Statlog (Shuttle) (SH)	58000	9	7
11.	Texture (TXR)	5500	40	11
12.	Twonorm (TN)	7400	20	2
13.	Yeast (YS)	1484	8	10

Πίνακας 1: Συνοπτική παρουσίαση των συνόλων δεδομένων

5.2 Πειραματικές μετρήσεις

Η απόδοση του κατηγοριοποιητή k-NN συγκρίθηκε με τις βελτιστοποιημένες εκδόσεις του ERHC που προτείναμε, τον ERHC-PD και τον ERHC-IMP, χρησιμοποιώντας τα 13 σύνολα δεδομένων που παρουσιάστηκαν νωρίτερα και σε δύο εκδόσεις του κάθε συνόλου, ανάλογα με το ποσοστό των απουσών τιμών που επιλέγαμε. Μετρήθηκε η ακρίβεια του κατηγοριοποιητή σε σύνολα που έχει γίνει συμπλήρωση των απουσών τιμών με τους αντίστοιχους μέσους όρους, καθώς και σε σύνολα δεδομένων όπου έχουν απούσες τιμές και γίνεται χρήση της μερικής απόστασης. Τα αποτελέσματα που προέκυψαν παρουσιάζονται στον Πίνακα 2. Παράλληλα μετρήθηκε και το ποσοστό μείωσης που επιτυγχάνεται μετά από κάθε πειραματική μέτρηση.

	KNN - IMP (Acc)	**KNN - PD (Acc)**	**ERHC-IMP (Acc)**	**ERHC-PD (Acc)**	**ERHC-IMP (RR)**	**ERHC-PD (RR)**
BL 10	77.561	63.308	79.004	75.961	87.600	88.800
BL 20	75.320	60.890	73.885	74.357	88.600	88.440
ECL 10	78.507	70.746	78.806	73.731	85.279	84.104
ECL 20	78.806	63.881	78.507	73.433	87.955	83.657
KDD 10	99.647	99.379	99.144	96.314	99.485	99.259
LIR 10	93.670	94.075	91.235	92.460	90.945	90.802
LIR 20	90.455	91.185	88.414	91.020	90.096	89.528
LS 10	89.198	89.400	89.369	88.576	94.740	93.217
LS 20	88.064	89.757	88.172	88.529	95.874	93.430
MGT 10	79.868	75.351	78.353	76.970	87.053	85.793
MGT 20	78.521	73.085	77.586	74.147	87.593	85.071
PD 10	99.072	99.045	98.253	98.508	97.098	96.725
PD 20	98.108	98.271	97.671	98.453	97.128	95.803
PH 10	87.248	83.860	84.860	83.084	87.609	88.272
PH 20	83.398	73.034	81.677	79.383	88.659	87.347
PM 10	71.447	67.411	70.535	68.972	83.122	81.661
PM 20	72.109	65.713	71.192	68.188	84.715	81.336
SH 10	99.897	99.697	99.374	99.484	99.445	99.475
SH 20	99.853	98.145	99.614	98.084	99.406	99.134
TN 10	95.432	93.256	92.066	91.445	97.895	96.895
TN 20	95.878	91.580	98.544	90.553	98.544	96.230
TXR 10	97.363	98.854	96.763	97.036	96.845	95.895
TXR 20	93.926	98.891	95.908	96.781	97.459	96.018
YS 10	51.312	41.398	52.997	51.515	79.040	79.680

YS 20	51.245	30.075	51.651	44.438	79.714	79.579
MO-10	86.171	82.752	8 5.443	84.158	91.243	90.814
MO-20	83.807	77.876	83.568	81.447	91.312	89.631
MO	85.036	80.411	84.543	82.857	91.276	90.246

Πίνακας 2: Αποτελέσματα πειραματικών μετρήσεων

Στη συνέχεια θα διασπάσουμε τον πίνακα 2 σε τέσσερις πίνακες ώστε να μπορέσουμε να πραγματοποιήσουμε τις κατάλληλες συγκρίσεις και να εστιάσουμε με έντονους χαρακτήρες τις μετρήσεις τις οποίες έχουμε πετύχει την καλύτερη απόδοση. Η ακρίβεια που προκύπτει από την εφαρμογή του κατηγοριοποιητή k-NN (όπου k=1) στα μη επεξεργασμένα δεδομένα εκπαίδευσης (χωρίς μείωση του πληθυσμού των των δεδομένων) αναγράφεται σε ποσοστό επί τις εκατό. Αναφέρουμε μόνο ότι η διαδικασία υπολογισμού των αποστάσεων ήταν μία χρονοβόρα διαδικασία. Ενδεικτικά θα αναφέρουμε ότι για το για σύνολο δεδομένων KDD η μία μόνο διασταυρούμενη επικύρωση (fold1) πετύχαινε κατά μέσο όρο πολλάκις εκατομμύρια υπολογισμούς. (π.χ. Fold1. Accuracy: 99.6572 Computations: 18.446.744.072.617.238.784)

	KNN - IMP (Acc)	**KNN - PD (Acc)**	**ΔΙΑΦΟΡΑ ΑΚΡΙΒΕΙΑΣ**
BL 10	**77.561**	63.308	14,25%
BL 20	**75.320**	60.890	14,43%
ECL 10	**78.507**	70.746	7,76%
ECL 20	**78.806**	63.881	14,93%
KDD 10	**99.647**	99.379	0,27%
LIR 10	93.670	**94.075**	0,41%
LIR 20	90.455	**91.185**	0,73%
LS 10	89.198	**89.400**	0,20%
LS 20	88.064	**89.757**	1,69%
MGT 10	**79.868**	75.351	4,52%
MGT 20	**78.521**	73.085	5,44%
PD 10	**99.072**	99.045	0,03%
PD 20	98.108	**98.271**	0,16%
PH 10	**87.248**	83.860	3,39%
PH 20	**83.398**	73.034	10,36%
PM 10	**71.447**	67.411	4,04%
PM 20	**72.109**	65.713	6,40%
SH 10	**99.897**	99.697	0,20%
SH 20	**99.853**	98.145	1,71%
TN 10	**95.432**	93.256	2,18%
TN 20	**95.878**	91.580	4,30%
TXR 10	97.363	**98.854**	1,49%

TXR 20	93.926	**98.891**	4,97%
YS 10	**51.312**	41.398	9,91%
YS 20	**51.245**	30.075	21,17%
MO-10	**86.171**	82.752	3,42%
MO-20	**83.807**	77.876	5,93%
MO	**85.036**	80.411	4,62%
Wins	18	7	
Losses	7	18	

Πίνακας 3: KNN - IMP vs KNN - PD

Στον παραπάνω πίνακα 3, συγκρίνουμε τον συμβατικό κατηγοριοποιητή k-NN και υπολογίζουμε την ακρίβεια του όταν εφαρμόζεται σε σύνολο δεδομένων στο οποίο έχουμε συμπληρωμένες τις απούσες τιμές καθώς και όταν εφαρμόζεται σε σύνολο δεδομένων το οποίο περιλαμβάνει απούσες τιμές αλλά κάνει υπολογισμό με τη χρήση της μερικής απόστασης. Όπως είναι προφανές είτε χρησιμοποιούμε το σύνολο δεδομένων με συμπλήρωση 10% είτε με συμπλήρωση 20%, η περίπτωση της συμπλήρωσης των απουσών τιμών, υπερέχει κατά πολύ σε σχέση με τη χρήση της μερικής απόστασης. Μόνο σε τρία σύνολα από τα δεκατρία σύνολα (LIR, LS, TXR και PD στο 20%) υπερέχει η δεύτερη περίπτωση με μέσο όρο βελτίωσης ακρίβειας κατά 0,015%. Αν καταμετρήσουμε το σύνολο υπεροχής στις μετρήσεις όλων των συνόλων δεδομένων που χρησιμοποιήθηκαν, τότε έχουμε υπεροχή δεκαοκτώ έναντι επτά. Αν υπολογίσουμε και τους μέσους όρους ανά περίπτωση 10% και 20% και τον γενικό μέσο όρο, παρατηρούμε πως υπερέχουν τουλάχιστον κατά τέσσερις με έξι μονάδες ακρίβειας. Συγκεκριμένα έχουμε 3,42%, 5,93% και 4,62% για τον μέσο όρο.

	ERHC-PD (Acc)	**KNN - PD (Acc)**	**ΔΙΑΦΟΡΑ ΑΚΡΙΒΕΙΑΣ**
BL 10	**75.961**	63.308	12,65%
BL 20	**74.357**	60.890	13,47%
ECL 10	**73.731**	70.746	2,99%
ECL 20	**73.433**	63.881	9,55%
KDD 10	96.314	**99.379**	3,07%
LIR 10	92.460	**94.075**	1,62%
LIR 20	91.020	**91.185**	0,17%
LS 10	88.576	**89.400**	0,82%
LS 20	88.529	**89.757**	1,23%
MGT 10	**76.970**	75.351	1,62%
MGT 20	**74.147**	73.085	1,06%
PD 10	98.508	**99.045**	0,54%
PD 20	**98.453**	98.271	0,18%
PH 10	83.084	**83.860**	0,78%

PH 20	**79.383**	73.034	6,35%
PM 10	**68.972**	67.411	1,56%
PM 20	**68.188**	65.713	2,48%
SH 10	99.484	**99.697**	0,21%
SH 20	98.084	**98.145**	0,06%
TN 10	91.445	**93.256**	1,81%
TN 20	90.553	**91.580**	1,03%
TXR 10	97.036	**98.854**	1,82%
TXR 20	96.781	**98.891**	2,11%
YS 10	**51.515**	41.398	10,12%
YS 20	**44.438**	30.075	14,36%
MO-10	**84.158**	82.752	1,41%
MO-20	**81.447**	77.876	3,57%
MO	**82.857**	80.411	2,45%
Wins	12	13	
Losses	13	12	

Πίνακας 4: ERHC-PD vs KNN - PD

Στον παραπάνω πίνακα 4, συγκρίνουμε τον προτεινόμενο αλγόριθμο ERHC-PD με χρήση της μερικής απόστασης με τον συμβατικό κατηγοριοποιητή k-NN κάνοντας χρήση της μερικής απόστασης. Παρατηρούμε μία ισοκατανομή της υπεροχής του κατηγοριοποιητή k-NN με χρήση της μερικής απόστασης πετυχαίνοντας 13 υπεροχές έναντι 12 του ERHC-PD. Από αυτήν την οπτική, φαίνεται καλύτερος ο k-NN κάνοντας χρήση της μερικής απόστασης, αλλά αν παρατηρήσουμε το ποσοστό της ακρίβειας που πετυχαίνουν και οι δυο τους, θα δούμε ότι στις περιπτώσεις που υπερέχει ο k-NN με χρήση της μερικής απόστασης, παρουσιάζει πολύ μικρή αύξηση της ακρίβειάς του έναντι του προτεινόμενου αλγόριθμου ERHC-PD, σε αντίθεση με την ακρίβεια που πετυχαίνει ο προτεινόμενος αλγόριθμος όπου υπερέχει, παρουσιάζει μεγάλη διαφορά ακριβείας σε σχέση με του k-NN με χρήση της μερικής απόστασης. Αν υπολογίσουμε και τους μέσους όρους ακριβείας ανά σύνολο δεδομένων με ποσοστό συμπλήρωσης 10% και 20% αλλά και ανά γενικό μέσο όρο, τότε παρατηρούμε ότι ο προτεινόμενος αλγόριθμος υπερέχει ξεκάθαρα στο συγκεντρωτικό ποσοστό ακριβείας που αποδίδει δηλαδή 1,41%, 3,57% και 2,45% για τον μέσο όρο αντίστοιχα. Μπορούμε να θεωρήσουμε λοιπόν ότι ο προτεινόμενος αλγόριθμός μας, υπερέχει έναντι του k-NN με χρήση της μερικής απόστασης, δίνοντας καλύτερη ακρίβεια. Παρατηρούμε ακόμα ότι όπως και πριν στα σύνολα δεδομένων LIR, LS, TXR συνεχίζουμε να έχουμε πλήρης υπέροχη του k-NN και ενσωματώνει και τα KDD, SH και τα PD και PH στο 10%.

	ERHC-IMP (Acc)	KNN - IMP (Acc)	ΔΙΑΦΟΡΑ ΑΚΡΙΒΕΙΑΣ
BL 10	**79.004**	77.561	1,44%
BL 20	73.885	**75.320**	1,44%
ECL 10	**78.806**	78.507	0,30%
ECL 20	78.507	**78.806**	0,30%
KDD 10	99.144	**99.647**	0,50%
LIR 10	91.235	**93.670**	2,44%
LIR 20	88.414	**90.455**	2,04%
LS 10	**89.369**	89.198	0,17%
LS 20	**88.172**	88.064	0,11%
MGT 10	78.353	**79.868**	1,52%
MGT 20	77.586	**78.521**	0,94%
PD 10	98.253	**99.072**	0,82%
PD 20	97.671	**98.108**	0,44%
PH 10	84.860	**87.248**	2,39%
PH 20	81.677	**83.398**	1,72%
PM 10	70.535	**71.447**	0,91%
PM 20	71.192	**72.109**	0,92%
SH 10	99.374	**99.897**	0,52%
SH 20	99.614	**99.853**	0,24%
TN 10	92.066	**95.432**	3,37%
TN 20	**98.544**	95.878	2,67%
TXR 10	96.763	**97.363**	0,60%
TXR 20	**95.908**	93.926	1,98%
YS 10	**52.997**	51.312	1,69%
YS 20	**51.651**	51.245	0,41%
MO-10	85.443	**86.171**	0,73%
MO-20	83.568	**83.807**	0,24%
MO	84.543	**85.036**	0,49%
Wins	8	17	
Losses	17	8	

Πίνακας 5: ERHC-IMP vs KNN-IMP

Στον παραπάνω πίνακα 5, συγκρίνουμε τον προτεινόμενο αλγόριθμο ERHC-IMP όταν εφαρμόζεται σε σύνολο δεδομένων στο οποίο έχουμε συμπληρωμένες τις απούσες τιμές με τους αντίστοιχους μέσους όρους, με τον συμβατικό κατηγοριοποιητή k-NN όταν εφαρμόζεται σε σύνολο δεδομένων στο οποίο έχουμε συμπληρωμένες τις απούσες τιμές με τους αντίστοιχους μέσους όρους. Παρατηρούμε πως οι περιπτώσεις στις οποίες υπερτερεί ο

k-NN είναι περισσότερες και σε πλήθος είναι 17 έναντι 8 του αλγόριθμου ERHC-IMP. Η βελτίωση στο ποσοστό ακρίβειας που πετυχαίνει εξαρτάται από το σύνολο δεδομένων. Παρουσιάζει διακυμάνσεις αλλά είναι εμφανής η υπεροχή τόσο στα δύο επιμέρους σύνολα του 10% και του 20% όσο και στο γενικό μέσο όρο που παρουσιάζει. Εμφανής υπεροχή παρουσιάζει ο αλγόριθμος ERHC-IMP στα σύνολα δεδομένων LS και YS καθώς και στα σύνολα BL και ECL στο ποσοστό 10% και στα σύνολα TN και TXR στο ποσοστό 20%.

	ERHC-IMP (Acc)	**ERHC-PD (Acc)**	**ΔΙΑΦΟΡΑ ΑΚΡΙΒΕΙΑΣ**
BL 10	**79.004**	75.961	3,04%
BL 20	73.885	**74.357**	0,47%
ECL 10	**78.806**	73.731	5,08%
ECL 20	**78.507**	73.433	5,07%
KDD 10	**99.144**	96.314	2,83%
LIR 10	91.235	**92.460**	1,23%
LIR 20	88.414	**91.020**	2,61%
LS 10	**89.369**	88.576	0,79%
LS 20	88.172	**88.529**	0,36%
MGT 10	**78.353**	76.970	1,38%
MGT 20	**77.586**	74.147	3,44%
PD 10	98.253	**98.508**	0,26%
PD 20	97.671	**98.453**	0,78%
PH 10	**84.860**	83.084	1,78%
PH 20	**81.677**	79.383	2,29%
PM 10	**70.535**	68.972	1,56%
PM 20	**71.192**	68.188	3,00%
SH 10	99.374	**99.484**	0,11%
SH 20	**99.614**	98.084	1,53%
TN 10	**92.066**	91.445	0,62%
TN 20	**98.544**	90.553	7,99%
TXR 10	96.763	**97.036**	0,27%
TXR 20	95.908	**96.781**	0,87%
YS 10	**52.997**	51.515	1,48%
YS 20	**51.651**	44.438	7,21%
MO-10	**85.443**	84.158	1,28%
MO-20	**83.568**	81.447	2,12%
MO	**84.543**	82.857	1,69%
Wins	16	9	
Losses	9	16	

Πίνακας 6: ERHC-IMP vs ERHC-PD

Στον παραπάνω Πίνακα 6, συγκρίνουμε τον προτεινόμενο αλγόριθμο ERHC-IMP όταν εφαρμόζεται σε σύνολο δεδομένων στο οποίο έχουμε συμπληρωμένες τις απούσες τιμές με τους αντίστοιχους μέσους όρους, με τον προτεινόμενο αλγόριθμο ERHC-PD όταν εφαρμόζεται σε σύνολο δεδομένων το οποίο περιλαμβάνει απούσες τιμές αλλά κάνει υπολογισμό με τη χρήση της μερικής απόστασης. Παρατηρούμε ότι ο αλγόριθμος ERHC-IMP παρουσιάζει εμφανή υπεροχή έναντι του ίδιου αλγορίθμου κάνοντας χρήση της μερικής απόστασης. Η αναλογία στην οποία υπερτερεί ο αλγόριθμος ERHC-IMP είναι 16 προς 9. Τα σύνολα δεδομένων στα οποία ο αλγόριθμος ERHC-PD υπερτερεί του αλγορίθμου ERHC-IMP είναι το LIR, το PD και το TXR καθώς και στα BL και LS σε ποσοστό 20% και στο SH στο ποσοστό 10%. Αν δούμε την αντιστοιχία ως προς τους μέσους όρους των δύο συνόλων καθώς και το γενικό μέσο όρο τότε παρατηρούμε ότι έχουμε αύξηση της ακρίβειας του ERHC-IMP κατά 1,69% σε σχέση με τον αλγόριθμο ERHC-PD. Παρατηρώντας το ποσοστό μεγαλύτερης ακρίβειας που παρουσιάζει ο αλγόριθμος ERHC-IMP σε σχέση με τον αλγόριθμο ERHC-PD στα επιμέρους σύνολα παρατηρούμε μία αισθητή διαφορά. Κοιτώντας προσεκτικά ένα-ένα τα σύνολα δεδομένων παρατηρούμε ότι ο αλγόριθμος ERHC-IMP στα 10 από τα 13 σύνολα δεδομένων, επιτυγχάνει μεγαλύτερη αύξηση του ποσοστού ακρίβειας όσο πιο πολλά απόντα δεδομένα έχουν συμπληρωθεί. Δηλαδή η χρήση της μερικής απόστασης επιτυγχάνει μεγαλύτερη ακρίβεια όσο οι απούσες τιμές είναι πιο πολλές. Συμπεραίνουμε ότι ο αλγόριθμος ERHC-PD με χρήση της μερικής απόστασης ίσως βοηθάει σε σύνολα με μικρό ποσοστό απουσών τιμών και κρίνεται απαγορευτική για σύνολα δεδομένων με μεγάλα ποσοστά απουσών τιμών.

	ERHC-IMP (RR)	**ERHC-PD (RR)**	**ΔΙΑΦΟΡΑ ΑΚΡΙΒΕΙΑΣ**
BL 10	87.600	**88.800**	1,20%
BL 20	**88.600**	88.440	0,16%
ECL 10	**85.279**	84.104	1,18%
ECL 20	**87.955**	83.657	4,30%
KDD 10	**99.485**	99.259	0,23%
LIR 10	**90.945**	90.802	0,14%
LIR 20	**90.096**	89.528	0,57%
LS 10	**94.740**	93.217	1,52%
LS 20	**95.874**	93.430	2,44%
MGT 10	**87.053**	85.793	1,26%
MGT 20	**87.593**	85.071	2,52%
PD 10	**97.098**	96.725	0,37%
PD 20	**97.128**	95.803	1,33%
PH 10	87.609	**88.272**	0,66%
PH 20	**88.659**	87.347	1,31%
PM 10	**83.122**	81.661	1,46%
PM 20	**84.715**	81.336	3,38%

SH 10	99.445	**99.475**	0,03%
SH 20	**99.406**	99.134	0,27%
TN 10	**97.895**	96.895	1,00%
TN 20	**98.544**	96.230	2,31%
TXR 10	**96.845**	95.895	0,95%
TXR 20	**97.459**	96.018	1,44%
YS 10	79.040	**79.680**	0,64%
YS 20	**79.714**	79.579	0,14%
MO-10	**91.243**	90.814	0,43%
MO-20	**91.312**	89.631	1,68%
MO	**91.276**	90.246	1,03%

Wins	21	4
Losses	4	21

Πίνακας 7: ERHC-IMP vs ERHC-PD (RR)

Καλό θα ήταν να βλέπαμε και την αντίδραση του αλγόριθμου ERHC-IMP κάνοντας χρήση της συμπλήρωσης των απουσών τιμών και του αλγόριθμου ERHC-PD κάνοντας χρήση της μερικής απόστασης (Πίνακας 7). Αυτή τη φορά όμως ως προς το ποσοστό μείωσης που επιτυγχάνουν. Παρατηρώντας τον παραπάνω πίνακα βλέπουμε ότι ο αλγόριθμος ERHC-IMP επιτυγχάνει σαφώς μεγαλύτερα ποσοστά μείωσης στο σύνολο σχεδόν των περιπτώσεων. Παρατηρούμε ότι υπερτερεί του αλγόριθμου ERHC-PD σε αναλογία 21 προς 4. Αν υπολογίσουμε για άλλη μία φορά τους μέσους όρους τον δύο συνόλων που έχουμε ορίσει καθώς και το γενικό μέσο όρο παρατηρούμε ότι πετυχαίνει καλύτερο ποσοστό μείωσης κατά ένα τουλάχιστον τοις εκατό σε σχέση με τη χρήση της μερικής απόστασης. Για άλλη μία φορά σε σύνολα δεδομένων με μεγάλο ποσοστό απουσών τιμών η χρήση της μερικής απόστασης επιτυγχάνει μικρότερα ποσοστά μείωσης. Το γεγονός αυτό οφείλεται στο ότι θεωρούνται πολλά στιγμιότυπα ως θόρυβος με αποτέλεσμα να χάνεται η ακρίβεια των μετρήσεων. Συμπεραίνουμε λοιπόν ότι ο αλγόριθμος που κάνει χρήση της μερικής απόστασης θεωρείται απαγορευτικός για σύνολα δεδομένων με μεγάλα ποσοστά απουσών τιμών.

6

Συμπεράσματα και μελλοντικές εργασίες

Η παρούσα διπλωματική εργασία επικεντρώνεται στο πρόβλημα των απουσών τιμών (missing values) που σχεδόν πάντα εμφανίζονται στα πραγματικά σύνολα δεδομένων εκπαίδευσης. Οι τεχνικές μείωσης του πληθυσμού των δεδομένων εκπαίδευσης (Data Reduction Techniques) που στοχεύουν στη μείωση του υπολογιστικού κόστους αλλά και των απαιτήσεων σε μνήμη, δεν μπορούνε να τις διαχειριστούνε και η αντιμετώπισή τους δεν είναι εφικτή. Αναζητώντας στη σχετική διεθνή βιβλιογραφία βρήκαμε αφενός μεθόδους συμπλήρωσης των απουσών τιμών (καταλογισμός) και αφετέρου τρόπους αντιμετώπισης τους από τις τεχνικές μείωσης του πληθυσμού των δεδομένων χωρίς να γίνει συμπλήρωσή τους. Δική μας συνεισφορά συνιστούν οι αλγόριθμοι ERHC-IMP και ERHC-PD. Με τον αλγόριθμο ERHC-IMP πειραματιστήκαμε χρησιμοποιώντας την μέθοδο της συμπλήρωσης του μέσου όρου κάθε χαρακτηριστικού ανά κλάση, ενέργεια όμως που απαιτεί ένα βήμα προ-επεξεργασίας και προσθέτει επιπλέον υπολογιστικό κόστος. Με τον αλγόριθμο ERHC-PD, μελετήσαμε τις τεχνικές RHC και ERHC, με βάση τις οποίες προτείναμε μια νέα παραλλαγή μιας τεχνικής μείωσης του πληθυσμού των δεδομένων που μπορεί να διαχειριστεί τις απούσες τιμές χωρίς να απαιτείται κάποιο επιπρόσθετο βήμα προ-επεξεργασίας για την συμπλήρωση τους.

Ολοκληρώνοντας την πειραματική διαδικασία, συμπεραίνουμε ότι συμβατικός κατηγοριοποιητής k-NN - IMP επιτυγχάνει μεγαλύτερη ακρίβεια στην περίπτωση που χειρίζεται σύνολα δεδομένων στα οποία έχει γίνει καταλογισμός στις απούσες τιμές με την

μέθοδο του μέσου όρου ανά κλάση απ' ότι όταν υπολογίζονται οι αποστάσεις κάνοντας χρήση της μερικής απόστασης.

Σε σχέση με αυτόν, ο αλγόριθμος ERHC-IMP παρουσιάζει μικρότερη αλλά σχετικά αμελητέα διαφορά στην ακρίβεια που επιτυγχάνει όταν και οι δύο χρησιμοποιούν σύνολα δεδομένων στα οποία έχουν συμπληρωθεί οι απούσες τιμές.

Από την άλλη μεριά, ο αλγόριθμος ERHC-PD όταν χρησιμοποιεί την μερική απόσταση παρουσιάζει αύξηση κατά 2,45 μονάδες επί τις εκατό, δηλαδή παρουσιάζει μεγαλύτερη ακρίβεια από τον κατηγοριοποιητή k-NN - PD όταν γίνεται χρήση της μερικής απόστασης.

Συγκεντρωτικά ο αλγόριθμος ERHC-IMP παρουσίασε καλύτερα ποσοστά στην ακρίβεια κάνοντας χρήση της συμπλήρωσης των απουσών τιμών από τον αλγόριθμο ERHC-PD, όταν δηλαδή γίνεται χρήση της μερικής απόστασης. Επίσης ο λόγος μείωσης που επιτυγχάνει ο αλγόριθμος ERHC-IMP είναι εξίσου μεγαλύτερος σε σχέση με το λόγο μείωσης που παρουσιάζει ο αλγόριθμος ERHC-PD κάνοντας χρήση της μερικής απόστασης.

Καταλήγουμε λοιπόν στο συμπέρασμα ότι ο αλγόριθμος ERHC-IMP είναι οριακά «ισάξιος» του συμβατικού k-NN – IMP αλλά ο τρόπος υπολογισμού των αποστάσεων με τη χρήση της μερικής απόστασης δεν μας έδωσε τα αποτελέσματα που θα επιθυμούσαμε. Κατά συνέπεια συνίσταται η χρήση του αλγόριθμου ERHC-IMP χρησιμοποιώντας την τεχνική της συμπλήρωσης των απουσών τιμών. Αυτό βέβαια απαιτεί και το κατάλληλο υπολογιστικό κόστος, από την άποψη ότι πρέπει να γίνει μία σχετική προ-επεξεργασία, να συμπληρωθούν με κάποια τεχνική οι απούσες τιμές και στη συνέχεια να γίνει η κατηγοριοποίηση. Αποτελεί βέβαια μία διαδικασία χρονοβόρα, την οποία προσπαθούμε να την αποφύγουμε κάνοντας χρήση του αλγόριθμου ERHC-PD με τον υπολογισμό της μερικής απόστασης. Παρόλα αυτά, χάνοντας λίγο σε ακρίβεια μπορούμε να έχουμε κέρδος, όσον αφορά τον χρόνο προ-επεξεργασίας και ολοκλήρωσης της κατηγοριοποίησης.

Ίσως, με περαιτέρω έρευνα, θα μπορούσαμε να κάνουμε συμπλήρωση των απουσών τιμών, όχι με το μέσο όρο ανά κλάση που χρησιμοποιήσαμε, αλλά κάνοντας χρήση κάποιας άλλης τεχνικής συμπλήρωσης, όπως για παράδειγμα την Hot Deck, αλλά αυτό αφήνεται για περαιτέρω έρευνα σε κάποια άλλη έρευνα μας.

7

Βιβλιογραφία

[1] Μιχάλης Βαζιργιάννης, Μαρία Χαλκίδη, Εξόρυξη Γνώσης από Βάσεις Δεδομένων και τον Παγκόσμιο Ιστό, Εκδ. Gutenberg

[2] J. Han, M. Kamber, and J. Pei. Data Mining: Concepts and Techniques. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science, 2011

[3] R. Roiger and M. W. Geatz. Data Mining: A Tutorial Based Primer. Addison Wesley, 2003

[4] I. Witten, E. H Frank, Practical Machine Learning Tools and Techniques Second Edition, USA:Morgan Kaufmann Publications, 2005

[5] Y. Ozkan, Data Mining Methods, istanbul, Turkey:Papatya Publications, 2008

[6] M.F. Amasyah, New Machine Learning Methods and Drug Design Applications, 2008

[7] M.F. Amasyah, Introduction to Machine Learning, 2010, [online] Available: http://www.ce.yildiz.edu.tr/mygetfile.php?id=868

[8] G. Silahtaroglu, Basic Concepts and Algorithms of Data Mining, 2008

[9] N. Murat, The Use Of Bayesian Approaches To Model Selection, 2007

[10] H.I Bülbül, ö Ünsal, Determination of Vocational Fields With Machine Learning Algorithm, pp. 710-713, 2010

[11] T. Cover and P. Hart. Nearest neighbor pattern classification. IEEE Trans. Inf. Theor., 13(1):21–27, September 2006

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

[13] L. Rokach. Data Mining with Decision Trees: Theory and Applications. Series in

machine perception and artificial intelligence. World Scientific Publishing Company, Incorporated, 2007

[14] Salvador Garcia, Joaquin Derrac, Jose Cano, and Francisco Herrera. Prototype selection for nearest neighbor classification: Taxonomy and empirical study. IEEE Trans. Pattern Anal. Mach. Intell., 34(3):417–435, March 2012

[15] Janet Kolodner. Case-based Reasoning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993

[16] Fadi Thabtah. A review of associative classification mining. Knowl. Eng. Rev., 22(1):3765, March 2007

[17] Pedro Domingos and Michael Pazzani. On the optimality of the simple bayesian classifier under zero-one loss. Mach. Learn., 29(2-3):103–130, November 1997

[18] Harry Zhang. The optimality of naive bayes. In Valerie Barr and Zdravko Markov, editors, Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, Miami Beach, Florida, USA, pages 562–567. AAAI Press, 2004

[19] Simon Haykin. Neural Networks: A Comprehensive Foundation. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2nd edition, 1998

[20] Isaac Triguero, Joaquín Derrac, Salvador Garcia, and Francisco Herrera. A taxonomy and experimental study on prototype generation for nearest neighbor classification. Trans. Sys. Man Cyber Part C, 42(1):86–100, January 2012

[21] G. P. Zhang. Neural networks for classification: A survey. Trans. Sys. Man Cyber Part C, 30(4):451–462, November 2000

[22] Acuña, E. and Rodriguez, C. 2004. The Treatment of Missing Values and its Effect on Classifier Accuracy. Classification, Clustering, and Data Mining Applications. 1995 (2004), 639–647. DOI:https://doi.org/10.1007/978-3-642-17103-1_60

[23] Kwak, S.K. and Kim, J.H. 2017. Statistical data preparation: Management of missing values and outliers. Korean Journal of Anesthesiology. 70, 4 (2017), 407–411. DOI:https://doi.org/10.4097/kjae.2017.70.4.407

[24] Irfan Pratama, A. E. (2016). A Review of Missing Values Handling Methods on Time-Series Data. International Conference on Information Technology Systems and Innovation (ICITSI) (p. 6). Bandung-Bali : IEEE.

[25] Julián Luengo, S. G. (2012). On the choice of the best imputation methods for missingvalues considering three groups of classification methods. , Knowledge Information System , 77–108.

[26] Shichao Zhang, Z. J. (2011). Missing data imputation by utilizing information within incomplete instances. The Journal of Systems and Software , 452–459.

[27] Langkamp, D.L. et al. 2010. Techniques for handling missing data in secondary analyses of large surveys. Academic Pediatrics. 10, 3 (2010), 205–210. DOI:https://doi.org/10.1016/j.acap.2010.01.005

[28] van Buuren, S. (2018). Flexible Imputation of Missing Data, Second Edition. New York: Chapman and Hall/CRC, https://doi.org/10.1201/9780429492259

[29] Haukoos, J.S. and Newgard, C.D. 2007. Advanced Statistics: Missing Data in Clinical Research-Part 1: An Introduction and Conceptual Framework. Academic Emergency Medicine. 14, 7 (2007), 662–668. DOI:https://doi.org/10.1197/j.aem.2006.11.037

[30] Grzymala-Busse, J.W. et al. 2005. Handling missing attribute values in preterm birth data sets. Lecture Notes in Computer Science. 3642 LNAI, (2005), 342–351.

DOI:https://doi.org/10.1007/11548706_36

[31] Joenssen, D. W., & Bankhofer, U. (2012). Hot Deck Methods for Imputing Missing Data. Machine Learning and Data Mining in Pattern Recognition, 63–75. https://doi.org/10.1007/978-3-642-31537-4_6

[32] Jason Van Hulse, T. M. (2008). A comprehensive empirical evaluation of missing value imputation in noisy software measurement data. Journal of System and Software , 691-708.

[33] Sasi, T. A. (2016). Intelligent Imputation Technique for Missing Values . International Conference on Advances in Computing, Communications and Informatics (ICACCI), (pp. 2441-2445). Jaipur, India.

[34] Alireza Farhangfara, L. K. (2008). Impact of imputation of missing values on classification error for discrete data. Pattern Recognition , 3692-3705.

[35] Esther-Lydia Silva-Ramírez, R. P.-M.-C.-D.-d.-l.-V. (2011). Missing value imputation on missing completely at random data using multilayer perceptrons. Neural Networks , 121–129.

[36] Harel O, Zhou XH (2007) Multiple imputation: Review of theory, implementation and software. Stat Med 26(16):3057–3077

[37] Rubin DB (1976) Inference and missing data. Biometrika 63(3):581–59

[38] Dennis L. Wilson. Asymptotic properties of nearest neighbor rules using edited data. IEEE trans. on systems, man, and cybernetics, 2(3):408–421, July 1972

[39] D. Randall Wilson and Tony R. Martinez. Reduction techniques for instance based learning algorithms. Mach. Learn., 38(3):257–286, March 2000

[40] J. S. Sánchez, R. Barandela, A. I. Marqués, R. Alejo, and J. Badenas. Analysis of new techniques to obtain quality training sets. Pattern Recogn. Lett., 24(7):1015–1022, April 2003

[41] I. Tomek. An experiment with the edited nearest-neighbor rule. IEEE Transactions on Systems, Man, and Cybernetics, 6:448–452, 1976

[42] Devijver, P. A. (1986). On the editing rate of the Multiedit algorithm. Pattern Recognition Letters, 4(1), 9–12. doi:10.1016/0167-8655(86)90066-8

[43] Vázquez, F., Sánchez, J. S., & Pla, F. (2005). A Stochastic Approach to Wilson's Editing Algorithm. Lecture Notes in Computer Science, 35–42. doi:10.1007/11492542_5

[44] M. Lozano. Data Reduction Techniques in Classification processes (Phd Thesis). Universitat Jaume I, 2007

[45] E. Namey, G. Guest, L. Thairu, L. Johnson, Data reduction techniques for large qualitative data sets. Handbook for team-based qualitative research, 137–163 (2007)

[46] Garcia, S., Derrac, J., Cano, J., Herrera, F.: Prototype selection for nearest neighbor classification: Taxonomy and empirical study. IEEE Trans. Pattern Anal. Mach. Intell. 34(3), 417–435 (2012), http://dx.doi.org/10.1109/TPAMI.2011.142, doi:10.1109/TPAMI.2011.14

[47] Triguero, I., Derrac, J., Garcia, S., Herrera, F.: A taxonomy and experimental study on prototype generation for nearest neighbor classification. Trans. Sys. Man Cyber Part C 42(1), 86–100 (2012), http://dx.doi.org/10.1109/TSMCC.2010.2103939, doi:10.1109/TSMCC.2010.210393

[48] Hart, P. E. (1968), 'The condensed nearest neighbor rule', IEEE Transactions on Information Theory 14(3), 515-516

[49] David W. Aha, Dennis Kibler, and Marc K. Albert. Instance-based learning algorithms. Mach. Learn., 6(1):37–66, January 1991

[50] P E Hart. The condensed nearest neighbor rule. IEEE Transactions on Information Theory,14(3):515–516, 1968

[51] José Salvador Sánchez. High training set size reduction by space partitioning and prototype abstraction. Pattern Recognition, 37(7):1561–1564, 2004

[52] C. H. Chen and Adam Jóźwik. A sample set condensation algorithm for the class sensitive artificial neural network. Pattern Recogn. Lett., 17(8):819–823, July 1996

[53] Pham, D. & Dimov, Stefan & Nguyen, Cuong. (2005). Selection of K in K -means clustering. Proceedings of The Institution of Mechanical Engineers Part C-journal of Mechanical Engineering Science - PROC INST MECH ENG C-J MECH E. 219. 103-119. 10.1243/095440605X8298

[54] Likas, A., Vlassis, N., & J. Verbeek, J. (2003). The global k-means clustering algorithm. Pattern Recognition, 36(2), 451–461. doi:10.1016/s0031-3203(02)00060-2

[55] Stefanos Ougiaroglou and Georgios Evangelidis. RHC: Non-parametric cluster-based data reduction for efficient k-nn classification. Pattern Analysis and Applications, pages (accepted, to appear)

[56] Stefanos Ougiaroglou and Georgios Evangelidis. Efficient dataset size reduction by finding homogeneous clusters. In Proceedings of the Fifth Balkan Conference in Informatics, BCI '12, pages 168–173, New York, NY, USA, 2012. ACM

[57] Ougiaroglou, S., & Evangelidis, G. (2015). Efficient editing and data abstraction by finding homogeneous clusters. Annals of Mathematics and Artificial Intelligence, 76(3-4), 327–349. doi:10.1007/s10472-015-9472-8

[58] Ougiaroglou, S., Evangelidis, G.: EHC: Non-parametric editing by finding homogeneous clusters. In: Beierle, C., Meghini, C. (eds.) Foundations of Information and Knowledge Systems, Lecture Notes in Computer Science, vol. 8367, pp. 290–304. Springer International Publishing (2014). doi:10.1007/978-3-319-04939-7 1

[59] Olvera-Lopez JA, Carrasco-Ochoa JA, Trinidad JFM (2010) A new fast prototype selection method based on clustering. Pattern Anal Appl 13(2):131–14

[60] sklearn.metrics.pairwise.nan_euclidean_distances — scikit-learn 0.23.1 documentation. (n.d.). Retrieved June 6, 2020, from scikit-learn.org website: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.nan_euclidean_distances.html

[61] Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. Proceedings of the 14th International Joint Conference on Artificial Intelligence – Volume 2, 1137- 1143. San Francisco, CA: Morgan Kaufmann

[62] G. Sande, Hot Deck Imputation Procedures, Incomplete Data in Sample Surveys, vol. 3, Academic Press, New York, 198

[63] Sim, J., Lee, J. S., & Kwon, O. (2015). Missing Values and Optimal Selection of an Imputation Method and Classification Algorithm to Improve the Accuracy of Ubiquitous Computing Applications. Mathematical Problems in Engineering, 2015, 1–14. doi:10.1155/2015/538613

[64] Engels, J. (2003). Imputation of missing longitudinal data: a comparison of methods. Journal of Clinical Epidemiology, 56(10), 968–976. doi:10.1016/s0895-4356(03)00170-7

[65] Schafer JL (1997) Analysis of incomplete multivariate data. Chapman & Hall, Florida

[66] KEEL: A software tool to assess evolutionary algorithms for Data Mining problems (regression, classification, clustering, pattern mining and so on). (n.d.). Retrieved June 6, 2020, from sci2s.ugr.es website: http://sci2s.ugr.es/keel/datasets.php

[67] UCI Machine Learning Repository: Data Sets. (2009). Retrieved June 6, 2020, from Uci.edu website: https://archive.ics.uci.edu/ml/datasets.php

[68] KEEL: A software tool to assess evolutionary algorithms for Data Mining problems (regression, classification, clustering, pattern mining and so on). (n.d.). Retrieved June 6, 2020, from sci2s.ugr.es website: https://sci2s.ugr.es/keel/dataset.php?cod=54

[69] KEEL: A software tool to assess evolutionary algorithms for Data Mining problems (regression, classification, clustering, pattern mining and so on). (n.d.). Retrieved June 6, 2020, from sci2s.ugr.es website: https://sci2s.ugr.es/keel/dataset.php?cod=61

[70] KEEL: A software tool to assess evolutionary algorithms for Data Mining problems (regression, classification, clustering, pattern mining and so on). (n.d.). Retrieved June 6, 2020, from sci2s.ugr.es website: https://sci2s.ugr.es/keel/dataset.php?cod=196

[71] Dua, Dheeru and Graff, Casey. (2017). UCI Machine Learning Repository: Statlog (Landsat Satellite) Data Set. Retrieved June 6, 2020, from archive.ics.uci.edu website: https://archive.ics.uci.edu/ml/datasets/Statlog+(Landsat+Satellite)

[72] KEEL: A software tool to assess evolutionary algorithms for Data Mining problems (regression, classification, clustering, pattern mining and so on). (n.d.). Retrieved June 6, 2020, from sci2s.ugr.es website: https://sci2s.ugr.es/keel/dataset.php?cod=198

[73] Dua, Dheeru and Graff, Casey. (2017). UCI Machine Learning Repository: MAGIC Gamma Telescope Data Set. Retrieved June 6, 2020, from archive.ics.uci.edu website: http://archive.ics.uci.edu/ml/datasets/magic+gamma+telescope

[74] Dua, Dheeru and Graff, Casey. (2017). UCI Machine Learning Repository: Pen-Based Recognition of Handwritten Digits Data Set. Retrieved June 6, 2020, from archive.ics.uci.edu website: https://archive.ics.uci.edu/ml/datasets/Pen-Based+Recognition+of+Handwritten+Digits

[75] KEEL: A software tool to assess evolutionary algorithms for Data Mining problems (regression, classification, clustering, pattern mining and so on). (n.d.). Retrieved June 6, 2020, from sci2s.ugr.es website: https://sci2s.ugr.es/keel/dataset.php?cod=105

[76] Data Society. (2016). Pima Indians Diabetes Database - dataset by data-society. Retrieved June 6, 2020, from data.world website: https://data.world/data-society/pima-indians-diabetes-database

[77] Dua, Dheeru and Graff, Casey. (2017). UCI Machine Learning Repository: Statlog (Shuttle) Data Set. Retrieved June 6, 2020, from archive.ics.uci.edu website: https://archive.ics.uci.edu/ml/datasets/Statlog+%28Shuttle%29

[78] KEEL: A software tool to assess evolutionary algorithms for Data Mining problems (regression, classification, clustering, pattern mining and so on). (n.d.). Retrieved June 6, 2020, from sci2s.ugr.es website: https://sci2s.ugr.es/keel/dataset.php?cod=72

[79] KEEL: A software tool to assess evolutionary algorithms for Data Mining problems (regression, classification, clustering, pattern mining and so on). (n.d.). Retrieved June 6, 2020, from sci2s.ugr.es website: https://sci2s.ugr.es/keel/dataset.php?cod=110

[80] KEEL: A software tool to assess evolutionary algorithms for Data Mining problems (regression, classification, clustering, pattern mining and so on). (n.d.). Retrieved June 6, 2020, from sci2s.ugr.es website: https://sci2s.ugr.es/keel/dataset.php?cod=112

[81] Batista G, Monard M (2003) An analysis of four missing data treatment methods for supervised learning. Appl Artif Intell 17(5):519–533

[82] Grace-Martin, K. (2009, February 4). Seven Ways to Make up Data: Common Methods to Imputing Missing Data. The Analysis Factor. (n.d.). Retrieved July 4, 2020, from theanalysisfactor.com website: https://www.theanalysisfactor.com/seven-ways-to-make-up-data-common-methods-to-imputing-missing-data/

[83] Spark, C. (2019, September 3). Tutorial: Introduction to Missing Data Imputation. Retrieved July 4, 2020, from Medium website: https://medium.com/@Cambridge_Spark/tutorial-introduction-to-missing-data-imputation-4912b51c34eb

[84] Garcia, S., Derrac, J., Cano, J. R., & Herrera, F. (2012). Prototype selection for nearest neighbor classification: Taxonomy and empirical study. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(3), 417–435. https://doi.org/10.1109/TPAMI.2011.142