ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

«Τεχνικές δειγματοληψίας και μείωσης δεδομένων για κατηγοριοποίηση βασισμένη σε στιγμιότυπα»

Του φοιτητή
Κεσόπουλου Βασιλείου
Αρ. Μητρώου: 175081

Επιβλέπων
Ουγιάρογλου Στέφανος
Επ. Καθηγητής.

23 Φεβρουαρίου 2023

Τίτλος Δ.Ε. Τεχνικές δειγματοληψίας και μείωσης δεδομένων για κατηγοριοποίηση βασισμένη σε στιγμιότυπα
Κωδικός Δ.Ε. 22279
Ονοματεπώνυμο φοιτητή : Κεσόπουλος Βασίλειος
Ονοματεπώνυμο εισηγητή : Ουγιάρογλου Στέφανος
Ημερομηνία ανάληψης Δ.Ε. : 21-10-2022
Ημερομηνία περάτωσης Δ.Ε. ....

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

Η παρούσα εργασία αποτελεί πνευματική ιδιοκτησία του φοιτητή Βασίλειου Κεσόπουλου που την εκπόνησε/αν. 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' οιονδήποτε 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 διπλωματικής 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.

«Στους γονείς μου»

Πρόλογος

Η πρόκληση του ενδιαφέροντος μου για την επιλογή και την εκπόνηση της συγκεκριμένης διπλωματικής εργασίας οφείλεται στα ερεθίσματα που πήρα από την παρακολούθηση μαθημάτων γύρω από το πεδίο της στατιστικής, της μηχανικής μάθησης και μαθημάτων ανάλυσης δεδομένων στα πλαίσια της παρακολούθησης του προπτυχιακού προγράμματος σπουδών του τμήματος Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων. Αφορμή υπήρξε επίσης η συνάφεια του αντικειμένου με τις προσωπικές μου επαγγελματικές βλέψεις αλλά και η διεύρυνση της γνώσης μου σε θέματα χρήσιμα κατά τη γνώμη μου.

Περίληψη

Πολλές τεχνικές μείωσης δεδομένων εκπαίδευσης έχουν προταθεί στη βιβλιογραφία για κατηγοριοποιητές που βασίζονται σε στιγμιότυπα (instance-based classifiers) όπως ο αλγόριθμος κατηγοριοποιήσης k εγγύτερων γειτόνων. Οι τεχνικές αυτές προσπαθούν δημιουργήσουν ένα μικρό σε μέγεθος αντιπροσωπευτικό σύνολο δεδομένων ώστε οι αλγόριθμοι κατηγοριοποίησης να μπορούν να εφαρμοστούν με μικρό υπολογιστικό κόστος. Ωστόσο, οι τεχνικές αυτές δεν έχουν συγκριθεί με τις τεχνικές δειγματοληψίας. Στόχος της εργασίας είναι να παρουσιάσει τις τεχνικές δειγματοληψίας για προβλήματα κατηγοριοποίησης καθώς και η συγκριτική μελέτη των τεχνικών δειγματοληψίας και τεχνικών μείωσης δεδομένων με στόχο την κατηγοριοποίηση. Η εργασία περιγράφει τα χαρακτηριστικά των συνόλων δεδομένων όπου οι τεχνικές δειγματοληψίας μπορούν να εφαρμοστούν αντί των τεχνικών μείωσης δεδομένων. Στο πειραματικό σκέλος εκτελέστηκαν διαφορετικές τεχνικές μείωσης προτύπων, τεχνικές δειγματοληψίας και εξετάστηκε η αποτελεσματικότητα τους με τον κατηγοριοποιητή των εγγύτερων γειτόνων πάνω σε διάφορα σύνολα δεδομένων. Στη μελέτη αυτή παρατηρείτε πώς η ύπαρξη θορύβου επηρεάζει την ποιότητα των τεχνικών αυτών και πως οι μέθοδοι πιθανοτικής δειγματοληψίας παρουσιάζουν καλύτερη επίδοση σε συνθήκες θορύβου 10%, 30% και 50% σε σχέση με τους αλγορίθμους condensed nearest neighbor(CNN) και reduction by space partitioning 3(RSP3).

«Sampling and Data Reduction Techniques for instance-based Classification»

«Vasileios Kesopoulos»

Abstract

Many training data reduction techniques have been proposed in the literature for instance-based classifiers such as the k-nearest neighbor classification algorithm. These techniques attempt to create a small representative data set so that classification algorithms can be applied with little computational cost. However, these techniques have not been compared to sampling techniques. The aim of the paper is to present the sampling techniques for categorization problems as well as the comparative study of sampling techniques and data reduction techniques aimed for classification. The work describes the characteristics of data sets where sampling techniques can be applied instead of data reduction techniques. Several data reduction techniques and sampling techniques were implemented in the experimental analysis of this thesis and their effectiveness were tested with the nearest neighbor classifier on various datasets. In this study it is observed how the existence of noise affects the quality of these techniques and how probability sampling methods perform better in 10%, 30% and 50% noise level with respect to condensed nearest neighbor(CNN) and reduction by space partitioning 3(RSP3) algorithms.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Συντομογραφίες

Δ.Ε. Διπλωματική Εργασία
ΔΙΠΑΕ Διεθνές Πανεπιστήμιο Ελλάδος

Κεφάλαιο 1ο: Εισαγωγή

1.1 Κατηγοριοποίηση

Κατηγοριοποίηση είναι το πρόβλημα προσδιορισμού της κατηγορίας(class) στην οποία ανήκει μια παρατήρηση, από ένα σύνολο κατηγοριών. Οι παρατηρήσεις, γνωστές και ως στιγμιότυπα(instances) ή πρότυπα(prototypes), αποτελούνται απο επιμέρους ιδιότητες οι οποίες είναι μετρήσιμες και συχνά ονομάζονται χαρακτηριστικα(features) [1]. Τα χαρακτηριστικα αυτά, όσον αφορά την φύση τους μπορουν να ειναι μεταβλητες διαφόρων τύπων και παρουσιάζονται παρακάτω :

- κατηγορική μεταβλητή(categorical variable) πχ. “Red”, “Green”, “Yello”
- τακτική μεταβλητή(ordinal variable) πχ. “low income”, “middle income”, “high income”
- ακέραια μεταβλητή(integer variable) πχ. ηλικία ανθρώπου : 20, 25, 30 ετών
- πραγματική μεταβλητή(real-valued variable) πχ. μέτρηση θερμοκρασίας : 2.031 Celsius

Οι αλγόριθμοι που λύνουν ή προσπαθούν να λύσουν το πρόβλημα κατηγοριοποίησης ονομάζονται κατηγοριοποιητές(classifiers) και γενικώς ο τρόπος που προσεγγίζουν το προβλήμα είναι με την εφαρμογή της μεθόδου μάθησης με επίβλεψη(supervised learning). Στην μάθηση με επίβλεψη, τα δεδομένα αποτελούνται από τα στιγμιότυπα, το κάθε στιγμιότυπο συνήθως έχει μια ετικέτα(label) η οποία δείχνει την κατηγορία ή κλάση στην οποία ανήκει και τελικά στόχος είναι η δυναμική δημιουργία μιας συνάρτησης που δέχεται τα στιγμιότυπα ως είσοδο και τα αντιστοιχίζει με ετικέτες κλάσης, μια διαδικασία που για να πραγματοποιηθεί με επιτυχία προυποθέτει την προγενέστερη τροφοδότηση ζευγών στιγμιοτύπουετικέτας κλάσης, κοινώς γνωστά και ως στιγμιότυπα εκπαίδευσης(training instances) και ανάλυση τους. Το πρόβλημα της κατηγοριοποίησης ονομάζεται πολλών κλάσεων(multiclass classification) όταν υπάρχουν 3 ή περισσότερες τιμές από τις οποίες μπορεί να αντιστοιχηθεί ένα στιγμιότυπο ενώ υπάρχει επίσης η δυαδική κατηγοριοποίηση κατά την οποία υπάρχουν μόνο 2 δυνατές τιμές.

Τα διάφορα μοντέλα κατηγοριοποίησης έρχονται συχνά αντιμέτωπα με δύο επιμέρους προβλήματα. Το πρόβλημα της υπερμοντελοποίησης(overfitting), όπου παρουσιάζεται όταν υπάρχει ένα περίπλοκο μοντέλο που ταιριάζει υπερβολικά με ένα σύνολο δεδομένων και κατά συνέπεια αποτυγχάνει να εξάγει ακριβείς προβλέψεις πάνω σε καινούργια δεδομένα [2]. Επίσης, υπάρχει το πρόβλημα της υπομοντελοποίησης(underfitting), το οποίο συμβαίνει όταν η ερμήνευση της δομής των δεδομένων είναι αδύνατη γιατί το μοντέλο είναι υπερβολικά απλό και έχει παρομοίως ως αποτέλεσμα αδύναμη προγνωστική απόδοση.

Το πεδίο εφαρμογών της στατιστικής κατηγοριοποίσης είναι πολύ ευρύ [3]. Η αναγνώριση προτύπων(pattern recognition) είναι ένα από τα πεδία εφαρμογών, η οποία αντιστοιχίζει ετικέτες κλάσεις για δεδομένες τιμές εισόδου και συνεπώς καταφέρνει για παράδειγμα την ανίχνευση κακόβουλων μηνυμάτων ηλεκτρονικού ταχυδρομείου και κατηγοριοποίησης τους ως ανεπιθύμητα. Η αναγνώριση γλώσσας(speech recognition) είναι ένα άλλο πεδίο εφαρμογής το οποίο υποστηρίζει τη υπολογιστική δυνατότητα αναγνώρισης προφορικού λόγου και μετάφρασης του. Συναντάται συχνά σαν τεχνολογική εφαρμογή στα συστήματα αυτοκινήτων καθώς υποστηρίζουν υπηρεσίες αυτόματης αναπαραγωγής ραδιαφω-

νικών σταθμών, μουσικής, τηλεφωνικών κλήσεων κλπ, μέσω της αναγνώρισης εντολών προφορικού λόγου.

1.2 Τύποι κατηγοριοποιητών

Υπάρχουν δύο τύποι μάθησης που χρησιμοποιούνται στα προβλήματα κατηγοριοποίησης, η οκνηρή μάθηση(lazy learning) και η πρόθυμη μάθηση(eager learning). Η διακριτή τους διαφορά έγκειται στο χρονικό σημείο που υλοποιείται η διαδικασία της γενίκευσης(generalization) των δεδομένων εκπαίδευσης. Παρακάτω θα γίνει μια περιγραφή αυτών των τύπων μάθησης καθώς και μια αναφορά σε μερικά μοντέλα κατηγοριοποίησης.

Στην οκνηρή(lazy) μέθοδο μάθησης, αλλιώς γνωστή και ως μάθηση βασισμένη σε στιγμιότυπα(instance based learning) ή μάθηση κατ'απαίτηση(learning on demand), η διαδικασία μάθησης της βασίζεται στην αποθήκευση των δεδομένων εκπαίδευσης και τη ανάκτηση αυτών από τη μνήμη με σκοπό την κατηγοριοποίηση ενός νέου στιγμιοτύπου [4]. Η υπολογιστική διαδικασία καθυστερεί να γίνει πράξη μέχρι την εμφάνιση νέου στιγμιοτύπου, γι'αυτό και δόθηκε το όνομα αυτό σ'αυτούς τους αλγορίθμους. Με αυτόν τον τύπο μάθησης μπορούν να δημιουργηθούν διαφορετικοί προσεγγιτές συναρτήσεων στόχου για διακριτές περιπτώσεις νέων προτύπων. Για παράδειγμα, καθιστούν δυνατό έναν τοπικής κατασκευής προσεγγιτή συνάρτησης στόχου που εφαρμόζεται στην γειτονιά του προς κατηγοριοποίηση στιγμιότυπου και οχι με πρόθεση να αποδίδει καλά σε ολόκληρο τον χώρο στιγμιοτύπων. Πάνω στην 'οκνηρή' συλλογιστική έχουν βρει έδαφος πληθώρα εφαρμογών όπως επιχειρηματολογίες για νομικά θέματα βασιζόμενες από προηγούμενες νομικές υποθέσεις, βοηθητικές ιατρικές διαγνώσεις μέσω σύγκρισης συμπτωμάτων με μια βάση δεδομένων αποτελούμενη από γνωστές αρρώστιες ή και επιλύσεις προβλήματων χρονοπρογραμματισμού με την επαναχρησιμοποίηση τμημάτων από παλαιότερες σχετικές λυμένες περιπτώσεις. Μερικοί δημοφιλείς οκνηροί αλγόριθμοι είναι ο κατηγοριοποιητής των k εγγύτερων γειτόνων, οι μηχανές kernel και τα δίκτυα RBF.

Στην πρόθυμη(eager) μέθοδο μάθησης, η γενίκευση των δεδομένων εκπαίδευσης επιτελειτε κατά την φάση της εκπαίδευσης του μοντέλου και συνεπώς το μοντέλο είναι γενικευμένο προτού δεχτεί ερωτήματα(queries) το σύστημα [5]. Τα τεχνητά νευρωνικά δίκτυα για παράδειγμα είναι μοντέλα πρόθυμης μάθησης και ένα σημαντικό πλεονέκτημα έναντι των οκνηρών αλγορίθμων είναι η μικρότερη χωρητικότητα δεδομένων που καταλαμβάνουν καθώς και η καλύτερη αντιμετώπιση του θορύβου όταν τα δεδομένα εκπαίδευσης έχουν θόρυβο. Τα μοντέλα αυτά είναι αποτελεσματικά και γρήγορα γιατί αποθηκεύουν μονάχα μια εσωτερική αναπαράσταση των δεδομένων στις παραμέτρους του μοντέλου και δεν έχουν ανάγκες επεξεργασίας ολόκληρου του συνόλου δεδομένων για να κάνουν εκτιμήσεις. Ακόμη, αυτή η αποθήκευση της αναπαράστασης καθιστά τα μοντέλα αυτά εύκολα στην ερμηνεία κατά κύριο λόγο και διευκολύνει τους αναλυτές δεδομένων στο κομμάτι της αποσφαλμάτωσης και στην ανίχνευση τυχών σφαλμάτων μεροληψίας. Μερικά επιπλέον μοντέλα πρόθυμης μάθησης είναι τα δέντρα απόφασης(Decision Tree) και ο κατηγοριοποιητής Naive Bayes.

Τα τεχνητά νευρωνικά δίκτυα είναι μοντέλα μάθησης με επίβλεψη τα οποία έχουν εμπνευστεί από τη λειτουργία των βιολογικών νευρωνικών δικτύων του εγκεφάλου. Ένα τέτοιο δίκτυο απαρτίζεται από ένα συνδεδεμένους κόμβους που ονομάζονται τεχνητοί νευρώνες και έχουν λειτουργία παρόμοια με αυτή των βιολογικών νευρώνων. Ουσιαστικά, νευρώνες συνδέονται με άλλους νευρώνες και στέλνουν σή-

ματα. Τα σήματα αυτά είναι πραγματικοί αριθμοί και αφού δέχτουν επεξεργασία στέλνουν έξοδο που συνήθως υπολογίζεται από κάποια μη γραμμική συνάρτηση του αθροίσματος των επιμέρους εισόδων. Η διαδικασία μάθησης των δικτύων αυτού του τύπου επιτελείται από τα βάρη τα οποία προσαρμόζονται και ανάλογα των τιμών τους καθορίζεται και η ισχύ σήματος εξόδου που θα παράγουν. Επίσης υπάρχει τιμή κατώφλι η οποία μπορεί να εμποδίσει πλήρως την έξοδο κάποιου σήματος από κάποιον νευρώνα άμα η έξοδος δεν υπερβαίνει τη τιμή αυτή.

Ένας άλλος γνωστός κατηγοριοποιητής πρόθυμου τύπου μάθησης ονομάζεται Naive Bayes. Ο κατηγοριοποιτής αυτός βασίζεται στο θεώρημα του Bayes για να κάνει προβλέψεις και θεωρεί ότι δεν υπάρχουν συσχετίσεις μεταξύ των χαρακτηριστικών κατά τη διαδικασία εκτέλεσης του. Αυτός είναι και ο λόγος που του δόθηκε το όνομα του 'αφελή' καθότι είναι σπάνιο να μην υπάρχουν συσχετίσεις μεταξύ των χαρακτηριστικών στη πραγματική ζωή και αυτός τους παραβλέπει. Ένα σημαντικό του πλεονέκτημα πάντως είναι ότι απαιτεί λίγα μόνο δεδομένα εκπαίδευσης για την εκτίμηση των σημαντικών παραμέτρων για τη κατηγοριοποίηση. Ακολουθεί παρακάτω ο τύπος του θεωρήματος Bayes :

$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x} \mid C_k)}{p(\mathbf{x})} \tag{1}$$

Τα δέντρα απόφασης(Decision trees) είναι μοντέλα μάθησης με επίβλεψη τα οποία χρησιμοποιούνται για να κάνουν προβλέψεις και να εξάγουν συμπεράσματα από ένα σύνολο παρατηρήσεων. Αρχικά αναλύεται ένα σύνολο δεδομένων σε μικρότερα υποσύνολα και τελικά αναπτύσσεται το δέντρο απόφασης που έχει κόμβους, κλάδους και φύλλα. Κάθε κόμβος έχει τουλάχιστον δύο κλάδους και τα φύλλα του δέντρου αντιπροσωπεύουν ουσιαστικά μία απόφαση ή κατηγοριοποίηση. Ένα σημαντικό πλεονέκτημα του αλγορίθμου είναι η εύκολη ερμηνευσιμότητα και κατανόηση του καθώς οι κανόνες του μοντέλου αναπαρίστανται οπτικά όπως τα διαγράμματα ροής και αυτό είναι πολύ βοηθητικό γιατί παρουσιάζει πολλές πληροφορίες με όμορφο τρόπο.

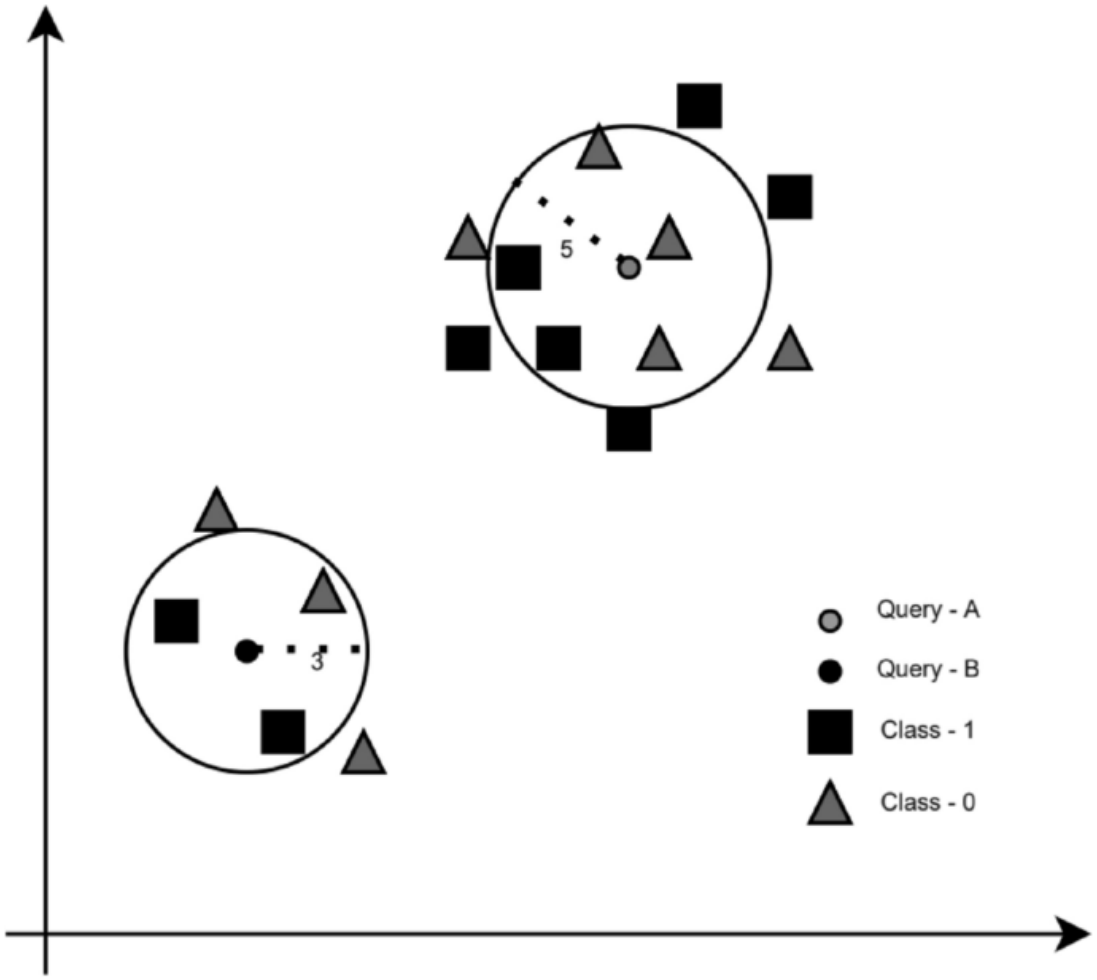
1.3 Ο Κατηγοριοποιητής των κ εγγύτερων γειτόνων

Ο κατηγοριοποιητής των k εγγύτερων γειτόνων(k-NN) είναι μια μη παραμετρική μέθοδος μάθησης με επίβλεψη(supervised learning) και χρησιμοποιείται για προβλήματα κατηγοριοποίησης αλλα και παλινδρόμησης(regression) [6]. Στην ενότητα αυτή γίνεται η περιγραφή της k-NN μεθόδου ως κατηγοριοποιητής καθότι αυτό είναι το είδος μοντελοποίησης που λαμβάνει μέρος σε επόμενα κεφάλαια.

Τα στιγμιότυπα εκπαίδευσης είναι συνήθως πολυδιάστατα διανύσματατα και ειναι εκχωρημένη μια κλάση σε κάθε στιγμιότυπο. Στο πλάισιο εκπαίδευσης του μοντέλου, είναι αποθηκεύονται όλα τα στιγμιότυπα ενώ στο πλαίσιο επικύρωσης ορίζεται η τιμή της παραμέτρου k για να κατηγοριοποιησει ενα νέο στιγμιότυπο που δεν έχει ετικέτα κλάσης.

Αυτό το μοντέλο κατηγοριοποίησης δέχεται για είσοδο τα k κοντινότερα στιγμιότυπα εκπαίδευσης από ένα συνολο δεδομένων(dataset), ενώ η έξοδος που παράγει ειναι μια ετικέτα κλάσης η οποία υπολογίζεται από την τιμή της παραμέτρου k. Συνεπώς, το στιγμιότυπο θα κατηγοριοποιηθεί στην κλάση πλειοψηφιας των k στιγμιοτύπων που βρίσκονται στο χώρο πιο κοντά του. Στο σχήμα 1.1 παρουσιάζεται ένα παράδειγμα εφαρμογής του αλγορίθμου στο δυσδιάστατο χώρο. Πιο αναλυτικά, παρουσιάζει ενα πρόβλημα

δυαδικής κατηγοριοποίσης με την γραφική απεικόνιση της κλάσης 1 ως ένα σχήμα τετραγώνου, την κλάση 0 και δύο μη κατηγοριοποιημένα στιγμιότυπα, το A και το B σε σχήμα κύκλου. Στη περίπτωση του A, έγινε η ρύθμιση της παραμέτρου k να είναι ίση με 5 και αποφασίζεται η κατηγοριοποίηση του στιγμιοτύπου A στη κλάση 1 γιατί βρίσκονται πιο κοντά του κατά απόσταση περισσότερα τετράγωνα παρά τρίγωνα(τρια τρίγωνα έναντι 2 τετραγώνων). Τέλος, το B κατατάζεται στη κλάση 0 χρησιμοποιώντας την ίδια αλγοριθμική συλλογιστική με αυτή στη προηγούμενη περίπτωση.



Σχήμα 1.1: διαδικασία k εγγύτερων γειτόνων για k=3 και k=5

Ο αλγόριθμος αυτός βασίζεται στην απόσταση για να κάνει την κατηγοριοποίηση και η πιο γνωστή μετρική απόστασης που χρησιμοποιείται είναι η ευκλείδια η οποία υπολογίζεται ως εξής :

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

όπου p και q είναι τα διανύσματα και d(p,q) είναι η μεταξύ τους απόσταση.

Μια άλλη μετρική απόστασης που χρησιμοποιείται είναι η Μανχάταν(Manhattan distance) η οποία υπολογίζει την απόσταση δύο διανυσμάτων σύμφωνα με το άθροισμα διαφορών των καρτεσιανών συντεταγμένων τους κατά απόλυτη τιμή. Η μετρική αυτή είθιστε να προτιμάται για σύνολα δεδομένων με υψηλό

αριθμό διαστάσεων. Ο υπολογισμός της απόστασης γίνεται ως εξής :

$$d(x,y) = \sum_{i=1}^{n} |x_i - y_i| \tag{3}$$

Η απόσταση Minkowski είναι ακόμα μια γνωστή μετρική η οποία θεωρείται ότι είναι μια γενικευμένη περίπτωση της Ευκλείδιας και της Μανχάταν. Παρακάτω παρουσιάζεται ο τύπος :

$$\left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}} \tag{4}$$

Παρατηρούμε ότι για p=1 προκύπτει ισοδύναμα ο τύπος υπολογισμού της απόστασης Μανχάταν ενώ για p=2 προκύπτει ο τύπος της Ευκλείδιας απόστασης. Ωστόσο για p<1 παραβιάζεται η τριγωνική ανισότητα και επομένως δεν μπορεί να θεωρηθεί ως μετρική. Για παράδειγμα, στο δυσδιάστατο χώρο, η απόσταση μεταξύ των σημείων (0,0) και (1,1) είναι $2^{\frac{1}{p}} > 2$ ενώ το σημείο (0,1) έχει απόσταση 1 και από τα δύο άλλα σημεία. Επίσης, μια άλλη περίπτωση Minkowski για p=∞ ονομάζεται η μετρική Chebyshev ή αλλιώς, μέγιστη μετρική(maximum metric). Η απόσταση Chebyshev ορίζεται ως η μεγαλύτερη απο τις διαφορές μεταξύ δύο διανυσμάτων κατά μήκος οποιαδήποτε διάστασης συντεταγμένων ή διαφορετικά, είναι η μέγιστη απόσταση κατά μήκος ενός άξονα. Πολλές φορές ονομάζεται και ως απόσταση σκακιέρας(Chessboard distance) καθώς το πλήθος των ελάχιστων κινήσεων που μπορεί να κάνει ένας βασιλιάς για να πάει από το ένα τετράγωνο στο άλλο ισούται με την απόσταση Chebyshev.

Επίσης, είναι συχνά επιθυμητή η κανονικοποίηση(normalization) των δεδομένων εκπαίδευσης γιατί μπορεί η αναπαράσταση των διανυσμάτων των χαρακτηριστικών να έρχεται σε πολύ διαφορετικές κλίμακες και αυτό είναι ένα εν δυνάμει πρόβλημα για την αποδοτικότητα του μοντέλου [7]. Ας υποθέσουμε ότι ένα σύνολο δεδομένων αποτελείται από n στιγμιότυπα και ότι πρέπει να κανονικοποιηθεί το χαρακτηριστικό e στο εύρος διαστήματος [0,1]. Τότε, η τιμή του χαρακτηριστικού του m-οστού στιγμιοτύπου, όπου m = 0,1,2,3....,n υπολογίζεται από τον παρακάτω τύπο :

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{5}$$

1.4 Μειονεκτήματα του κατηγοριοποιητή εγγύτερων γειτόνων

Παρά τα οφέλη και τα πλεονεκτήματα του αλγορίθμου πρέπει να επισημανθούν και τα σημαντικά αδύνατα σημεία του. Ενα απο αυτά είναι ο ορισμός της τιμής της παραμέτρου k [8]. Η επιλογή της τιμής του πρέπει να γίνετε με προσοχή διότι ο αλγόριθμος είναι ευαίσθητος σε δεδομένα με θόρυβο όταν η τιμή του k είναι μικρή ενώ αντίθετα, μια μεγάλη τιμή του k μπορεί να οδηγήσει σε αναξιόπιστες προβλέψεις. Έτσι, πρέπει να επιδιώκεται ο καθορισμός μιας ισορροπημένης τιμής μεταξύ των δύο ακραίων περιπτώσεων.

Ένα άλλο μειονέκτημα είναι το υπολογιστικό κόστος χρήσης του αλγορίθμου μπορεί να είναι υψηλό, ειδικά σε μεγάλα σύνολα δεδομένων(datasets), καθώς πρέπει να υπολογίζεται η απόσταση μεταξύ του

προς εξέταση προτύπου και όλων των υπολοίπων προτύπων που υπάρχουν στο σετ εκπαίδευσης(training set) με σκοπό τη πρόβλεψη της ετικέτας κλάσης του νέου προτύπου.

Ακόμη, η κατάρα της διάστασης(curse of dimensionality) είναι ένα μειονέκτημα το οποίο παρουσιάζεται σαν πρόβλημα όταν είναι υψηλής διάστασης τα δεδομένα [9]. Σαν αποτέλεσμα, η ευκλείδια απόσταση σταματάει να είναι χρήσιμη επειδή οι πλέον απστάσεις μεταξύ του προς εξέταση προτύπου και των προτύπων του σετ εκπαίδευσης είναι όλες περίπου ίσες.

Η ύπαρξη άσχετων χαρακτηριστικών(features) στα δεδομένα μπορεί να δημιουργήσει σημαντικό πρόβλημα στη διαδικασία κατηγοριοποίησης ή πρόβλεψης με τον k-NN καθώς μπορεί να θεωρήσει ότι είναι εξίσου σημαντικά και ενώ δεν θα έπρεπε να λαμβάνονται υπόψιν [10]. Για παράδειγμα, έστω ότι υπάρχει ένα σύνολο δεδομένων με πληροφορίες για διαφορετικά σπίτια και αποτελείτε από χαρακτηριστικά όπως χρώμα του σπιτιού ή μάρκες των συσκευών που περιλαμβάνονται στο ακίνητο. Η πρόβλεψη της τιμής των ακινήτων συνυπολογίζοντας ισοδύναμα την χρησιμότητα όλων των χαρακτηριστικών είναι μια ακατάλληλη προσέγγιση και οδηγεί σε κακά προγνωστικά αποτελέσματα.

Ο αλγόριθμος των k εγγύτερων γειτόνων χρησιμοποιείται συνήθως σε αριθμητικά δεδομένα για να υπολογίζει τις αποστάσεις μεταξύ σημείων και γι'αυτό είναι αδύναμος στη χρήση του πάνω σε κατηγορικά δεδομένα όπως "κόκκινο" , "πράσινο", "μπλε". Ένας τρόπος αντιμετώπισης είναι η χρήση one-hot κωδικοποίηση(one-hot encoding). Ωστόσο μπορεί να προκαλέσει αύξηση των διαστάσεων των δεδομένων και τελικά να αυξηθεί το υπολογιστικό κόστος.

1.5 Κίνητρο και Συνεισφορά

Το κεντρικό αντικείμενο της διπλωματικής αυτής είναι η μελέτη τεχνικών μείωσης δεδομένων και πιο συγκεκριμένα, η εφαρμογή των τεχνικών αυτών σε προβλήματα κατηγοριοποιητών βασιζόμενων σε στιγμιότυπα εκπαίδευσης για την αντιμετώπιση των αδυναμιών του. Οι κατηγοριοποιητές αυτού του τύπου χρειάζονται συχνά την εφαρμογή ενός βήματος προεπεξεργασίας για να μειωθεί το υψηλό υπολογιστικό κόστος χωρητικότητας ή και ειδικά για περιπτώσεις όπου τα σύνολα δεδομένων είναι πολύ μεγάλα. Επίσης, υπάρχει μια πληθώρα τεχνικών που έχουν σκοπό την επιλογή ενός υποσυνόλου δεδομένων που να είναι ικανό να αντιπροσωπευει τον συνολικό πλυθησμό των δεδομένων. Αυτές οι τεχνικές δεν είναι άλλες από τις τεχνικές δειγματοληψίας οι οποίες δεν εως σήμερα συγκριθεί ως προς την αποτελεσματικότητα τους με τις μεθόδους μείωσης δεδομένων(DRT) και κίνητρο λοιπόν της διπλωματικής είναι η συγκριτική μελέτη αυτών των δύο διαφορετικών κατηγοριών με τελικό στόχο την ανάδειξη της καλύτερης ή και υπό ποιες προϋποθέσεις είναι αληθές.

1.6 Οργάνωση της εργασίας

Στο τρέχων κεφάλαιο κάναμε μια εισαγωγή στο πρόβλημα της κατηγοριοποίησης και στις προκλήσεις που προκύπτουν κατά την επίλυση του προβλήματος. Επίσης, περιγράψαμε τα χαρακτηριστικά των διαφορετικών τύπων μάθησης ενώ εστιάσαμε περισσότερο στον οκνηρό τύπο μάθησης(lazy learning) και πιο συγκεκριμένα στο μοντέλο των k εγγύτερων γειτόνων και σε μερικά από τα αδύνατα σημεία του.

Στο 2ο κεφάλαιο θα γίνει η περιγραφή της στατιστικής δειγματοληψίας, ο σκοπός της και οι δυνατότητες

της ως άξιο εργαλείο άντλησης των χρήσιμων πληροφοριών από ένα ευρύτερο πλυθησμό ατόμων ή δεδομένων γενικότερα. Ακόμη, θα περιγραφούν διάφορες δειγματοληπτικές τεχνικές όσον αφορά τον τρόπο λειτουργίας τους, τα πλεονεκτήματα αλλά και τους περιορισμούς που υπάρχουν στην αποτελεσματική εφαρμογή των τεχνικών αυτών.

Στο 3ο κεφάλαιο θα εισαγάγουμε την έννοια της μείωσης δεδομένων(data reduction) που είναι συχνά μια πολύ σημαντική διαδικασία που εφαρμόζεται στα δεδομένα πριν από τη διαδικασία της κατηγοριοποίησης. Θα εξηγήσουμε την συνεισφορά της, τις διαφορετικές κατηγορίες της ενώ θα δωθεί μεγαλύτερη προσοχή στην ιδέα της αριθμητικής μείωσης δεδομένων και τους επιμέρους τρόπους επίτευξης του σκοπού αυτού. Τέλος, θα περιγράψουμε τη διαδοχική εφαρμογή μείωσης δεδομένων αποτελούμενη από αλγορίθμους διαφορετικών φιλοσοφιών και το κίνητρο για αυτή τη προσέγγιση.

Στο 4ο κεφάλαιο παραθέτουμε τον κώδικα που συγγράφηκε σε μορφή python για την υλοποίηση πειραμάτων κατηγοριοποίησης με το μοντέλο του εγγύτερου γείτονα(1-NN). Παρατίθεται κώδικας για την εφαρμογή τεχνικών μείωσης δεδομένων(DRT) με μεθόδους δειγματοληψίας και άλλων μεθόδων μείωσης προτύπων. Επίσης, παρουσιάζεται κώδικας για την προσθήκη θορύβου στα σύνολα δεδομένων και ο κώδικας για την αξιολόγηση της ακρίβειας του κατηγοριοποιητή με το μοντέλο της μετρικής k-fold cross-validation. Όλα οι κώδικες αναλύονται ως προς τον τρόπο λειτουργίας τους.

Τα πειραματικά αποτελέσματα της συγκριτικής μελέτης των τεχνικών μείωσης δεδομένων(DRT) με των τεχνικών δειγματοληψίας και η ανάλυση παρουσιάζεται στο 5ο κεφάλαιο. Η σύνοψη των αποτελεσμάτων παρουσιάζεται σε συγκεντρωτικούς πίνακες των ποσοστών μείωσης δεδομένων και της αποτελεσματικότητας του κατηγοριοποιητή 1-NN όταν τα δεδομένα υπάγονται σε διαφορετικά επίπεδα θορύβου. Τέλος, γίνετε μια συζήτηση σχετικά με τη συνεισφορά των αποτελεσμάτων.

Κεφάλαιο 2ο: Τεχνικές δειγματοληψίας

2.1 Δειγματοληψία πληθυσμού

Η δειγματοληψία πληθυσμού είναι κομμάτι του κλάδου της στατιστικής και παίζει κεντρικό ρόλο στο πεδίο της μεθοδολογίας έρευνας και στην διασφάλιση ποιότητας για θέματα βιομηχανιών παραγωγής αλλα και παροχής υπηρεσιών. Δειγματοληψία πληθυσμού ονομάζεται η διαδικασία επιλογής ενός υποπληθυσμού ή αλλιώς υποσυνόλου μέσα από έναν στατιστικό πληθυσμό [11]. Σκοπός της δειγματοληψίας είναι εξαγωγή των ιδιοτήτων του πληθυσμού με την επιλογή ενός δείγματος που θα αντιπροσωπεύει το προς μελέτη πλαίσιο δειγματοληψίας όσο το δυνατόν καλύτερα.

Μια σημαντική παράμετρος της δειγματοληψίας είναι η σωστή πηγή προέλευσης από όπου θα δημιουργηθεί το δείγμα μας [12]. Η πλήρης αξιοποίηση ή συλλογή πληροφοριών από τον συνολικό πληθυσμό είναι μια σχεδόν αδύνατη προσέγγιση του προβλήματος καθώς απαιτεί μεγάλο οικονομικό κόστος και μπορεί να είναι εξαιρετικά χρονοβόρο. Επίσης, υπάρχουν περιπτώσεις όπου ο πληθυσμός προέλευσης του δείγματος και ο πληθυσμός για τον οποίον μας ενδιαφέρει η συγκέντρωση πληροφοριών, να είναι δύο διαφορετικά μεταξύ τους σύνολα. Για παράδειγμα, ο πειραματισμός και η μελέτη πάνω σε αρουραίους μπορεί να εξάγει χρήσιμες πληροφορίες για ανθρώπινη υγεία. Μια άλλη σχετική παράμετρος του προβλήματος είναι η σωστή επιλογή του δειγματοληπτικού πλαισίου(sampling frame). Το δειγματοληπτικό πλαίσιο είναι μια λίστα και απαρτίζεται από τον πλυθησμό ενδιαφέροντος που μπορεί να υποστεί δειγματοληψία.

Υπάρχει πληθώρα εφαρμογών αξιοποίησης της δειγματοληψίας. Για παράδειγμα, στην διασφάλιση ή έλεγχου ποιότητας, όπου ένας παραγωγός προμηθεύει στον καταναλωτή μεγάλες ποσότητες προιόντων. Η απόφαση αποδοχής ή απόρριψης της συναλλάγης θα καθοριστεί απο την εκ των προτέρων λήψη ενός δείγματος και σύγκριση της επιμέρους ποιότητας του δείγματος σύμφωνα με ενα καθορισμένο όριο αποδοχής ελαττωματικών ειδών.

Η προσπάθεια εκτίμησης των στατιστικών χαρακτηριστικών κάποιου πληθυσμού και δημιουργίας ενός αντιπροσωπευτικού δείγματος είναι μια διαδικασία της οποίας τα αποτελέσματα συνήθως υπόκεινται σε διαφόρων φύσεως σφάλματα και μεροληψίες [13]. Το σφάλμα της μεροληψίας επιλογής(selection bias), που προκύπτει λόγω της μεθόδου συλλογής δειγμάτων και έχει σαν αποτέλεσμα τη διαφορά της πραγματικής πιθανότητας στην επιλογή ατόμων ή δεδομένων με την πιθανότητα που εμείς τελικά υποθέτουμε στους υπολογισμούς. Επίσης, το σφάλμα τυχαίας δειγματοληψίας(sampling error), που υφίσταται εξαιτίας της αποτυχίας ενός υποπληθυσμού να αντιπροσωπεύει πλήρως τον συνολικό πλυθησμό. Για παράδειγμα, η εκτίμηση του μέσου όρου του δείκτη μάζας σώματος των ανδρών με λήψη δείγματος πληθυσμού μιας χώρας με πολύ υψηλό δείκτη πείνας θα οδηγούσε σε υποεκτίμηση. Αυτό το σφάλμα υπολογίζεται από την διαφορά μεταξύ της εκτίμησης μια παραμέτρου και της πραγματικής αλλά άγνωστης τιμής της παραμέτρου. Η σωστή επιλογή των ατόμων γίνετε αποτελεσματικά όταν οι επιμέρους πιθανότητες επιλογής των ατόμων είναι ισοδύναμες αλλά στην πραγματικότητα η απόκτηση ενός αμερόληπτου δείγματος είναι δύσκολη καθώς δεν θα πρέπει κάποια από τις παραμέτρους να εμπλακούν στη διαδικασία της επιλογής για το δείγμα. Τέλος, είναι σημαντικό να αναφερθεί και η παρουσία των μη δειγματοληπτικών σφαλμάτων(Non-sampling error) τα οποία εμφανίζονται από τη προβληματική συλλογή, επεξεργασία ή σχεδιασμό του δείγματος, για παράδειγμα :

- συμπερίληψη δεδομένων έξω από τον πληθυσμό(overcoverage), όταν δηλαδή συμπεριλαμβάνεται το ίδιο άτομο πάνω απο 1 φορά
- μη συμπερίληψη δεδομένων του πληθυσμού από το δειγματοληπτικό πλαίσιο(undercoverage), από το γεγονός ότι το δείγμα δεν αποτελείτε από το συνολικό πλυθησμό
- παρερμήνευση των ερωτήσεων ή δυσκολίες στην απάντηση
- κακή κωδικοποίηση των δεδομένων
- μεροληψία συμμετοχής, όταν το δείγμα είναι διαφορετικό από το πληθυσμό στόχου λόγω δυσανάλογων κοινωνικών, πολιτικών ή αλλων χαρακτηριστικών που υπάρχουν μεταξύ των ατόμων

2.2 Απλή τυχαία δειγματοληψία

Η απλή τυχαία δειγματοληψία (Simple Random Sampling) είναι η πιο βασική τεχνική δειγματοληψίας. Είναι ένα υποσύνολο στοιχείων που συγκεντρώθηκε απο ενα μεγαλύτερο συνολο πληθυσμού και η επιλογή των στοιχείων πραγματοποιήθηκε με τυχαίο τρόπο [14]. Βασικό του χαρακτηριστικό είναι ότι οποιοδήποτε υποσύνολο από k στοιχεία έχει την ίδια πιθανότητα να επιλεγεί για το δείγμα με οποιοδήποτε άλλο k υποσύνολο στοιχείων. Για παράδειγμα, έστω ότι γίνετε μια κλήρωση X εισιτηρίων για τη παρακολούθηση μια κινηματογραφικής προβολής και έστω ότι η εκδήλωση ενδιαφέροντος συμμετοχής αποτελείτε από ένα πλήθος N ατόμων, όπου N > X. Έπειτα, δίνετε ένας αριθμός σε κάθε άτομο στο εύρος περιοχής από 0 έως N-1. Οι αριθμοί που θα αγνοηθούν είναι εκείνοι που τυχόν επιλέχτηκαν προηγουμένως αλλά και οι αριθμοί εκτός του εύρου 0 έως N-1 και νικητές της κλήρωσης αναγνωρίζονται όσοι έχουν καποιον απο τους πρώτους X αριθμούς. Γενικά, συνηθίζετε να αποφεύγεται η δειγματοληψία "με αντικατάσταση" σε περιπτώσεις όπου έχουμε μικρούς αλλά συχνά ακόμα και σε μεγάλους πληθυσμούς, που σημαίνει ότι δεν επιτρέπουμε συμπεριλαμβάνουμε στο δείγμα μας την επιλογή του ίδιου στοιχείου του πληθυσμού περισσότερες απο μία φορές. Ωστόσο, για ένα μικρό δείγμα απο ένα πολύ μεγάλο πλυθησμό, η πιθανότητα να επιλεγεί ένα στοιχείο μπορει να είναι η ίδια είτε "με αντικατάσταση" είτε "χωρίς αντικατάσταση", καθώς γίνετε σπάνια πλέον η επιλογή του ίδιου στοιχείου δύο φορές.

Η δειγματοληπτική μέθοδος αυτή έχει το προσόν να μην χρειάζεται εκ των προτέρων γνώση για τον πλυθησμό εκτός από αυτόν του δειγματοληπτικού πλαισίου. Επίσης, εξαιτίας της απλότητας της, διευκολύνει σημαντικά την ανάλυση των δεδομένων που συλλέγει. Για τους παραπάνω λόγους, είναι καλή πρακτική η επιλογή αυτής της μεθόδου όταν οι διαθέσιμες πληροφορίες που έχουμε για τον πλυθησμό είναι ελλιπείς ή όταν μας ενδιαφέρει περισσότερο το χαμηλό υπολογιστικό κόστος παρά η αποτελεσματικότητα.

Υπάρχουν διάφοροι αποτελεσματικοί αλγόριθμοι δημιουργίας δειγμάτων με τυχαίο τρόπο. Ένας από αυτούς ονομάζεται αφελής και λειτουργεί ως εξής. Εξετάζει σειριακά τα στοιχεία του πληθυσμού και σε κάθε βήμα αφαιρεί το στοιχείο με ίση πιθανότητα και το τοποθετεί στο δείγμα. Η διαδικασία ολοκληρώνει όταν έχει τοποθετηθεί στο δείγμα ένας πλήθος k στοιχείων που ορίστηκε από τον σχεδιαστή της μελέτης. Ο αλγόριθμος του Sunter [15] είναι ένας άλλος αλγόριθμος κατά τον οποίο εκχωρείται ένας τυχαίος αριθμός k για κάθε στοιχείο, που προέρχεται απο μια ομοιόμορφη κατανομή (0,1) και χρησιμοποιείται ως δείκτης. Έπειτα γίνετε ταξινόμηση των στοιχείων βάσει των k και στο τέλος επιλέγονται για το δείγμα τα x μικρότερα στοιχεία.

Σχήμα 2.1: Απλή τυχαία δειγματοληψία

2.3 Στρωματοποιημένη δειγματοληψία

Η στρωματοποιημένη δειγματοληψία(stratified sampling) είναι μια άλλη δημοφιλής τεχνική η οποία διαιρεί τον πληθυσμό πριν από την δειγματοληψία, σε ομοιογενείς υποπληθυσμούς ή αλλιώς γνωστούς ως στρώματα(strata), δηλαδή υποπληθυσμούς με μεταξύ τους διακριτά χαρακτηριστικά [16]. Μία προϋποθέση για την δημιουργία των στρωμάτων είναι η ύπαρξη αμοιβαία αποκλειστικότητας(mutually exclusive), δηλαδή η μη ταυτόχρονη παρουσία του ίδιου στοιχείου του πληθυσμού σε περισσότερα από ένα στρώμα. Επίσης πρέπει να είναι συλλογικά εξαντλητικά(collectively exhaustive), για παράδειγμα έστω ότι ο συνολικός πληθυσμός έχει χωρίστεί σε 5 στρώματα και το κάθε στρώμα είναι ένα σύνολο(subset). Τότε, από την θεωρία πιθανοτήτων, το αποτέλεσμα της ένωσης των επιμέρους συνόλων θα πρέπει να είναι ίσο με το συνολικό πληθυσμό.

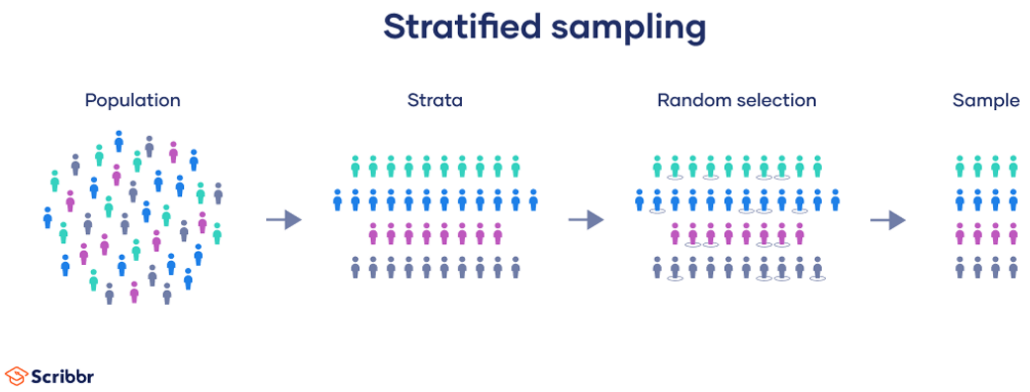
Γενικά υπάρχουν δύο στρατηγικές στρωματοποιημένης δειγματοληψίας. Η στρατηγική αναλογικής κατανομής(proportionate allocation), όπου για την δημιουργία των στρωμάτων λαμβάνεται υπόψιν ενα δειγματοληπτικό κλάσμα, που είναι το πηλίκο του πληθυσμού δειγματοληψίας προς το συνολικό πλυθησμό του στρώματος. Με αυτό τον τρόπο καταφέρνεται η δημιουργία μικρότερου σφάλματος στις εκτιμήσεις. Ακόμη, υπάρχει η στρατηγική της δυσανάλογης κατανομής(disproportionate allocation), στην οποία το κλάσμα δειγματοληψίας των στρωμάτων υπολογίζεται με τον τρόπο της προηγούμενη στρατηγικής αλλά και με την αξιοποίηση της τυπικής απόκλισης της κατανομής. Ένα παράδειγμα αναλογικής κατανομής είναι το παρακάτω. Έστω ότι, μια εταιρία αποτελείται απο το εξής προσωπικό :

- άνδρας, πλήρους απασχόλησης: 90
- άνδρας, μερικής απασχόλησης: 18
- γυναίκα, πλήρους απασχόλησης: 9
- γυναίκα, μερικής απασχόλησης: 63

Δηλαδή, αποτελείται συνολικά από 180 άτομα και μας ζητήτε η λήψη δείγματος προσωπικού 40 ατόμων. Το αρχικό βήμα είναι η δημιουργία των στρωμάτων και ο υπολογισμός που καταλαμβάνει το κάθε στρώμα.Τα ποσοστά είναι τα εξής :

- άνδρες, πλήρους απασχόλησης = 90 ÷ 180 = 50%
- άνδρες, μερικής απασχόλησης = 18 ÷ 180 = 10%
- γυναίκες, πλήρους απασχόλησης = 9 ÷ 180 = 5%
- γυναίκες, μερικής απασχόλησης = 63 ÷ 180 = 35%

Τελικά, για δεδομένο μέγεθος δειγματοληψίας ίσο με 40, συμπεραίνουμε ότι το 50% του δείγματος θα αποτελείτε από 20 άτομα του 1ου στρώματος(20*0,5=10), 4 άτομα του 2ου στρώματος(20*0.1=2), 2 άτομα του 3ου στρώματος και τέλος, 14 άτομα από το 4ο στρώμα.



Σχήμα 2.2: στρωματοποιημένη δειγματοληψία

Οι λόγοι για την επιλογή εφαρμογής της στρωματοποιημένης δειγματοληψίας και τα πλεονεκτήματα της είναι πολυάριθμα. Καταρχάς, δημιουργεί αντιπροσωπευτικά δείγματα του πληθυσμού καθώς διασφαλίζεται η συμμετοχή ατόμων απο κάθε στρώμα. Επίσης, καθιστά πιο διαχειρίσμες και πιο φθηνές τις μετρήσεις του πληθυσμού σε στρώματα αν και αυτό εξαρτάται από την εφαρμογή. Ακόμη, το σφάλμα εκτίμησης είναι μικρό όταν οι τυπικές αποκλίσεις εσωτερικά των στρωμάτων είναι μικρότερες σε σχέση με την τυπική απόκλιση του συνολικού πληθυσμού.

2.4 Συστηματική δειγματοληψία

Συστηματική δειγματοληψία(systematic sampling) είναι μια στατιστική μέθοδος η οποία δημιουργεί το δείγμα μέσω της επιλογής στοιχείων από ένα διατεταγμένο πλαίσιο(ordinal frame). Συνήθως εφαρμόζεται ώστε να υποστηρίζει την ιδιότητα της ισοπιθανότητας(Equiprobability), όπου σημαίνει ότι η πιθανότητα επιλογής του κάθε στοιχείου είναι γνωστή και ίση με οποιοδήποτε άλλο στοιχείο, γνωστό και ως equal probability of selection(epsem). Αυτή η ιδιότητα το καθιστά ως μια παρόμοια τεχνική με αυτή της απλής τυχαίας δειγματοληψίας, ωστόσο, διαφέρουν λόγω του γεγονότος ότι δεν είναι πιθανή η επιλογή δύο γειτονικών στοιχείων του συνόλου γιατί η συστηματική δειγματοληψία χρησιμοποιεί το διάστημα

Σχήμα 2.3: Συστηματική δειγματοληψία

δειγματοληψίας που δεν επιτρέπει αυτο το συμβάν. Διάστημα δειγματοληψίας είναι μία παράμετρος της μεθόδου και ορίζεται από το πηλίκο του μεγέθους του πληθυσμού K προς το μέγεθος του δείγματος n. Έστω οτι η παράμετρος διαστήματος είναι k, τότε η εξίσωση θα είναι : k = N / n Το σημείο αφετηρίας για την δημιουργία του δείγματος σε ένα διατεταγμένο σύνολο, θα είναι ένα τυχαίο στοιχείο μεταξύ του 1 και του k, και στη συνέχεια συνεχίζεται η διαδικασία με την επιλογή κάθε kο στοιχείου από το πλυθησμό [17]. Αμα θεωρήσουμε το διατεταγμένο σύνολο οτί είναι μια δομή δεδομένων λίστας, τότε το σημείο τερματισμού της διαδικασίας θα είναι η επιστροφή του δείκτη στην κορυφή της λίστας.

Πιο αναλυτικά, δίνετε παρακάτω ένα παράδειγμα υλοποίησης της διαδικασίας για την καλύτερη περιγραφή της. Έστω ότι, θέλετε να διεξάγετε μια έρευνα σχετικά με τη συχνότητα αλλιλεπίδρασης των μαθητών με το διάβασμα στον ελεύθερο τους χρόνο και έστω ότι θέλετε να μελετήσετε ένα δείγμα των 100 μαθητών από εναν ευρύτερο πλυθησμό των 1000. Στο πρώτο βήμα, δημιουργείται μια λίστα αποτελούμενη απο 1000 μαθητές και ορίζεται ένας αριθμός για τον καθένα από το εύρος 1 έως N(όπου N στο παράδειγμα μας ο συνολικός πληθυσμός των 1000). Μετά, επιλέγεται το σημείο αφετηρίας της διαδικασίας, το οποίο είναι ένας τυχαίος αριθμός x μεταξύ του 1 και της τιμής του διαστήματος δειγματοληψίας k που υπολογίζεται απο τον τύπο που παρουσιάστηκε πιο πριν. Τελικά, ο μαθητής αφετηρίας θα αποτελέσει το 1ο στοιχείο της λίστας του δείγματος ενώ επόμενα στοιχεία θα είναι ο κάθε k μαθητής. Για παράδειγμα, αν x=4 και k=10, τότε θα γίνει η επιλογή του : 4ου, 14ου, 24ου, 34ου κλπ. Η διαδικασία θα διαρκέσει μέχρις ότου γεμίσει η λίστα στο επιθυμητό μέγεθος δείγματος των 100. Στο σχήμα 2.3, παρουσιάζεται οπτικά η συμπεριφορά της μεθόδου πάνω σε ένα διαφορετικό παράδειγμα.

Ένας από τους λόγους που ερευνητές προτιμούν αυτό τον τρόπο δημιουργίας δειγμάτων έναντι αλλών είναι λόγω της εύκολης υλοποίησης της καθότι εξαρτάται μόνον από την λειτουργία της περιοδικής λήψης δειγμάτων. Ακόμη, είναι λιγότερο κοστοβόρα καθώς απαιτεί λιγότερους πόρους για την εφαρμογή της συνήθως, σε αντίθεση με άλλες πιθανοτικές τεχνικές. Επίσης, δεν πρέπει να παραλείψουμε οτι είναι δημιουργεί μια άξια αντιπροσώπευση του πληθυσμού με την στοχευμένη επιλογή του διαστήματος δειγματοληψίας k αλλά και ότι έχει μικρότερο σφάλμα δειγματοληψίας από την απλή τυχαία δειγματοληψία.

2.5 Άλλες τεχνικές δειγματοληψίας

Η δειγματοληψία σε συστάδες(cluster sampling) ή αλλιώς γνωστή ως δειγματοληψία ενός σταδίου(one-stage cluster), είναι μια τεχνική δειγματοληψίας που συχνά χρησιμοποιείται στην έρευνα μάρκετινγ. Συστάδες ονομάζονται οι ομάδες που κατασκευάζει για να διαιρέσει τον συνολικό πλυθησμό από τις οποίες στη συνέχεια επιλέγει ένα πλήθος συστάδων με τυχαίο τρόπο όπου και θα αποτελέσουν το δείγμα [18]. Μια άλλη προσέγγιση της ονομάζεται δειγματοληψία δύο σταδίων(two-stage), όπου μετά την δημιουργία των συστάδων και της φάσης της τυχαίας επιλογής των συστάδων, πραγματοποιείται στο τέλος και μια απλή τυχαία δειγματοληψία σε κάθε επιλεχθέν συστάδα.

Ένα παράδειγμα που μπορεί να εφαρμοστεί η δειγματοληψία σε συστάδες δύο σταδίων είναι το ακόλουθο. Ας υποθέσουμε ότι ερευνητές ενδιαφέρονται να μελετήσουν νέους ανθρώπους που αντιμετωπίζουν θέματα ψυχικής υγείας σε μια συγκεκριμένη πόλη και αποφασίζουν να πάρουν συνέντευξη από ένα δείγμα 50 ανθρώπων. Ο συνολικός πληθυσμός αποτελείται από τη λίστα καταλόγου όλων των κλινικών ψυχικής υγείας της πόλης και κάθε κλινική είναι μία συστάδα. Οι ερευνητές διαλέγουν από τον κατάλογο 5 τυχαίες συστάδες και εφαρμόζουν απλή τυχαία δειγματοληψία σε κάθε μία από αυτές ώστε να συλλέξουν 10 νέους από κάθε συστάδα και τελικά να διεξάγουν τις συνεντεύξεις.

Σχήμα 2.4: Δειγματοληψία σε συστάδες

Στο σχήμα 2.4 βλέπουμε την εφαρμογή της δειγματοληψίας δύο-σταδίων και την συμπερίληψη 4 ατόμων

απο τις τυχαία επιλεγμένες συστάδες. Βέβαια, ο πληθυσμός μπορεί να είναι διαφορετικός από συστάδα σε συστάδα πράγμα που μπορεί να οδηγήσει σε σφάλματα μεροληψίας. Όμως ο τύπος δύο-σταδίων αντιμετωπίζει αρκετά αποτελεσματικά αυτό το πρόβλημα. Ένας άλλος τρόπος αντιμετώπισης του προβλήματος είναι η δειγματοληψία με πιθανότητα ανάλογη του μεγέθους(Probability-proportional-to-size sampling), στην οποία γίνετε ανάθεση πιθανότητας ανάλογης του μεγέθους της κάθε συστάδας. Οι μεγαλύτερες σε μέγεθος συστάδες θα έχουν μεγαλύτερη πιθανότητα να επιλεγούν και το πλεονέκτημα της έγκειται στο ότι η επιλογή μεγάλων συστάδων θα οδηγήσει στην συμπερίληψη του τελικού δείγματος απο κατά προσέγγιση ισομεγέθη συστάδες.

Τα πλεονεκτήματα της δειγματοληψίας σε συστάδες είναι πολυάριθμα [19]. Το σημαντικότερο του ίσως χαρακτηριστικό είναι οι μικρές απαιτήσεις σε πόρους εφόσον επιλέγονται μόνο ορισμένες συστάδες για την ολοκλήρωση κάποιου ερευνητικού έργου και γενικά τα πιθανά έξοδα ταξιδιού είναι μειώνονται. Επίσης, η τεχνική αυτή μειώνει τις διακυμάνσεις στα αποτελέσματα δεδομένου ότι η κάθε συστάδα αποτελεί μια αντιπροσώπευση του συνολικού πληθυσμού και εφόσον είναι ομοιογενείς επιτυγχάνεται καλύτερη ακρίβεια σε σχέση με άλλες μεθόδους. Γενικά, η τεχνική δειγματοληψίας αυτή περιλαμβάνει όλα τα πλεονεκτήματα της στρωματοποιημένης και της τυχαίας προσέγγισης χωρίς να έχει τόσα πολλά μειονεκτήματα. Αυτό βοηθάει στη μείωση της πιθανότητας μεροληψίας στα συλλεγόμενα δεδομέν. Επίσης, επειδή υπάρχουν λιγότεροι κίνδυνοι δυσμενών επιδράσεων που δημιουργούν τυχαίες διακυμάνσεις, τα αποτελέσματα της εργασίας μπορούν να δημιουργήσουν αποκλειστικά συμπεράσματα όταν εφαρμόζονται στο συνολικό πληθυσμό.

Η δειγματοληψία σε συστάδες μπορεί ωστόσο να είναι αναποτελεσματική καθώς η επιλογή των συστάδων γίνεται τυχαία και ορισμένες συστάδες μπορεί να είναι μικρότερες σε μέγεθος σε σχέση με άλλες. Συνεπώς, η δυσανάλογη κατανομή των πόρων από τον ερευνητή μπορεί να οδηγήσει σε σπατάλες πόρων. Ένα άλλο μειονέκτημα είναι το ενδεχόμενο υποτίμησης του σφάλματος δειγματοληψίας(sampling error) καθ'ότι είναι αληθές ότι οι παρατηρήσεις εντός μιας συστάδας τείνουν να είναι πιο όμοιες μεταξύ τους και λιγότερο όμοιες μεταξύ τους οι παρατηρήσεις από συστάδα σε συστάδα.

Μία σύνθετη μορφή της δειγματοληψίας σε συστάδες είναι η πολυσταδιακή δειγματοληψία(multi-stage sampling). Ονομάζεται πολυσταδιακή γιατί η διαδικασία εφαρμογής της περιλαμβάνει πολλά βήματα [20]. Γενικά, χωρίζει τον πλυθησμό σε συστάδες, έπειτα επιλέγει με τυχαίο τρόπο μία ή ένα πλήθος συστάδων και τέλος εξάγει ένα δείγμα απο την ή τις επιμέρους συστάδες. Στο πρώτο βήμα-στάδιο πρέπει να γίνει η δημιουργία των συστάδων ενώ στο δεύτερο βήμα πρέπει να αποφασιστεί ποια στοιχεία θα επιλεγούν από τις συστάδες. Η τεχνική αυτή προτιμάται συνήθως όταν δεν υπάρχει νόημα αναγνώρισης όλων των μελών του πληθυσμού ή όταν δεν έχουμε γνώση αυτής την παραμέτρου.

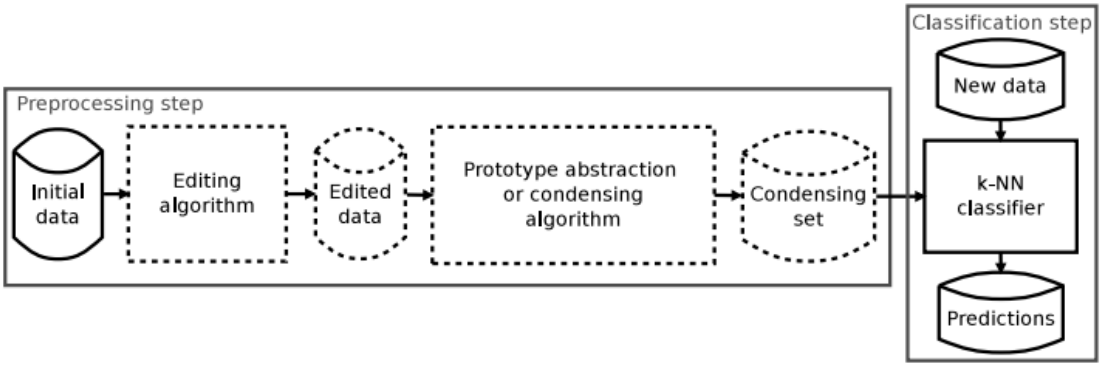
Στο ακόλουθο παράδειγμα περιγράφεται η αλληλουχία βημάτων εκτέλεσης πολυσταδιακης δειγματοληψίας. Στο παράδειγμα αυτό, εφαρμόστηκε η πολυσταδιακή δειγματοληψία με κίνητρο την επιλογή ενός δείγματος αποτελούμενο από μαθητές γυμνασίου και λυκείου των Ηνωμένων Πολιτειών. Στο πρώτο στάδιο δημιουργήθηκαν στρώματα πληθυσμού όπως στη μέθοδο στρωματοποίησης και προσδιορίστηκαν με βάση γεωγραφικών περιοχών(βορειοανατολικά, νότια, μεσοδυτικά και δυτικά), μετά επιλέχτηκε τυχαίο δείγμα 10% των σχολείων από τη κάθε γεωγραφική διαίρεση(στρώμα) και στη συνέχεια επιλέχτηκε ξανά τυχαίο δείγμα 10% αλλά αυτή τη φορά, τάξεων από το κάθε σχολείο. Στο τέλος, οι μαθητές που διαλέχτηκαν για την μελέτη προσκαλέθηκαν να συμπληρώσουν το ερωτηματολόγιο έρευνας.

Κεφάλαιο 3ο: Τεχνικές Μείωσης Δεδομένων

3.1 Εισαγωγή

Μείωση δεδομένων ονομάζεται οποιαδήποτε διαδικασία μετατροπής αριθμητικών πληροφοριών ή χαρακτήρων σε μια διορθωμένη και απλοποιημένη αναπαράσταση χωρίς την απώλεια σημαντικών δεδομένων. Οι τεχνικές που είναι υπεύθυνες για την υλοποίηση της μείωσης δεδομένων εντοπίζουν και αφαιρούν τις άσχετες ή περιττές πληροφορίες ενώ διατηρούν την ακεραιότητα και την ικανότητα των δεδομένων να τεθούν υπο ανάλυση [21].

Τα πλεονεκτήματα της εργασίας της μείωσης δεδομένων είναι πολυάριθμα και την καθιστούν μια πολύ σημαντική τεχνική στο πεδίο της αναλυτικής των δεδομένων. Το κυριότερο της πλεονέκτημα είναι ότι βελτιώνει σημαντικά την απόδοση και τον ρυθμό εκτέλεσης της επεξεργασίας και ανάλυσης των δεδομένων καθώς μειώνει το μέγεθος του συνόλου δεδομένων. Ένα άλλο πλεονέκτημα είναι ότι μπορεί να μειώσει τις απαιτήσεις αποθήκευσης ενός συνόλου δεδομένων και αυτό είναι χρήσιμο όταν δουλεύουμε με πολύ μεγάλα σύνολα δεδομένων που μπορεί να απαιτούν ακριβές λύσεις αποθήκευσης. Έτσι, μπορεί να μειωθεί το κόστος αποθήκευσης με τεχνικές συμπύκνωσης(condensation) και παραγωγής(generation) οι οποίες μειώνουν το μέγεθος των δεδομένων. Ακόμη, μπορεί να βελτιωθεί η ακρίβεια(accuracy) και η ποιότητα της ανάλυσης με τη κατάλληλη αφαίρεση θορύβου ή μη σχετικών δεδομένων και τη διατήρηση μόνο των ουσιαστικών πληροφοριών που χρειάζονται.



Σχήμα 3.1: Μείωση δεδομένων για τον k-NN κατηγοριοποιητή

Σε αυτό το σημείο είναι επιτακτική και η περιγραφή μερικών προκλήσεων και δυσκολιών που είναι πιθανό να παρουσιαστουν από την εφαρμογή των τεχνικών αυτών. Η μείωση των δεδομένων μπορεί πολλές φορές να απορρίψει εκτός απο περιττά αλλά και χρήσιμα δεδομένα για την μελέτη με αποτέλεσμα δυσκολία ερμηνείας τους και την επακόλουθη έκπτωση της ακρίβειας της μελέτης. Ακόμη, υπάρχει το ενδεχόμενο εισαγωγής μεροληψιών στα δεδομένα όταν η συμπίεση τους πραγματοποιηθεί από κάποια δειγματοληπτική μέθοδο που δεν έχει δημιουργήσει αντιπροσωπευτικά δείγματα. Αξίζει να σημειωθεί ότι γενικά έχει σημασία ο ορισμός του πλαισίου αλλά και η φύση των δεδομένων πάνω στα οποία αποφασίζεται να εφαρμοστεί κάποια τεχνική και συνεπώς δεν μπορούν τα μειονεκτήματα αυτά να χαρακτηρίσουν όλες τις υπάρχουσες περιπτώσεις.

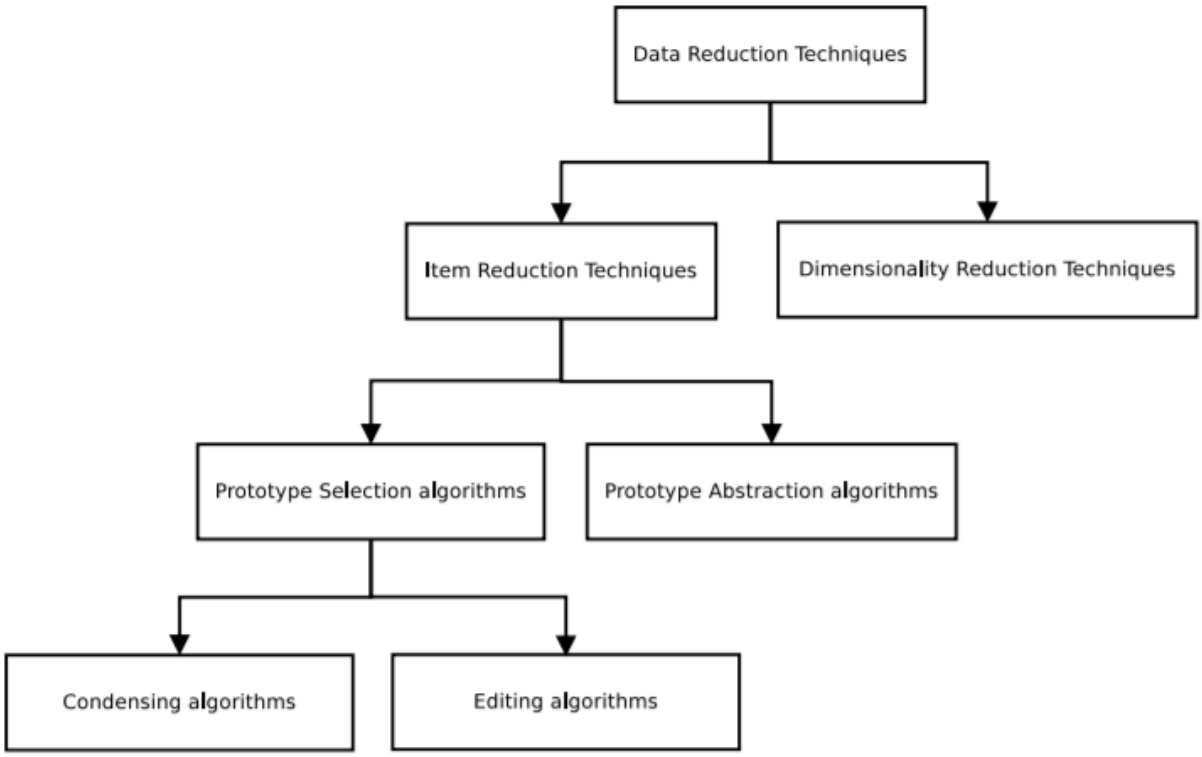
3.2 Κατηγορίες Τεχνικών Μείωσης Δεδομένων

Οι τεχνικές μείωσης δεδομένων(Data Reduction Techniques) μπορούν να διαιρεθούν σε δύο γενικές κατηγορίες με βάση το κριτήριο του τύπου πληροφορίας που πρόκειται να υποστεί συρρίκνωση. Η μία από αυτές είναι η αριθμητική μείωση των δεδομένων(Numerosity Reduction) η οποία, με όρους της αναλυτικής των δεδομένων, αναφέρεται στη μείωση των συνολικών στοιχείων ή των γραμμών από κάποιο σύνολο δεδομένων [22]. Ο παράγοντας του πλήθους των στοιχείων παίζει σημαντικό ρόλο στην ανάλυση των δεδομένων επειδή καθορίζει το επίπεδο δυσκολίας της επεξεργασίας, ανάλυσης και εξαγωγής συμπερασμάτων. Η μείωση αυτού του όγκου μπορεί να δημιουργήσει τη προοπτική μιας πιο διαχειρίσιμης δομής με απουσία θορύβου και καλύτερο εντοπισμό των μοτίβων που κρύβονται στα δεδομένα. Οι τεχνικές αυτές βρίσκουν ευρεία εφαρμογή σε προβλήματα κατηγοριοποίησης και μπορούν να χωριστούν σε δύο επιμέρους τύπους, τους αλγορίθμους επιλογής προτύπων(Prototype Selection algorithms) και τους αλγορίθμους παραγωγής προτύπων(Prototype Abstraction algorithms), όπου και θα γίνει η περιγραφή τους στα υποκεφάλαια που θα ακολουθήσουν. Ακόμη, η μείωση των διαστάσεων(Dimensionality reduction) είναι η δεύτερη κατηγορία μείωσης δεδομένων και χαρακτηρίζεται ως η διαδικασία μείωσης του πλήθους των διαστάσεων ενός σετ δεδομένων, γνωστών και ως χαρακτηριστικά(features), με ταυτόχρονη διατήρηση των χρήσιμων πληροφοριών όσο αυτό είναι εφικτό. Χρησιμοποιείται συχνά για την αναπαράσταση δεδομένων υψηλών διαστάσεων και έχει κίνητρο τη βελτίωση της αποτελεσματικότητας διαφόρων αλγορίθμων μηχανικής μάθησης. Τέλος, ο διαχωρισμός του σε κατηγορίες μπορεί να γίνει ως εξής :

- Επιλογή χαρακτηριστικών(feature selection), η διαδικασία επιλογής ενός υποσυνόλου μεταβλητών για τη χρήση του σε κάποιο μοντέλο
- εξαγωγή χαρακτηριστικών(feature extraction), η διαδικασία δημιουργίας νέων χαραρακτηριστικών μέσω του μετασχηματισμού raw δεδομένων σε αριθμητικά χαρακτηριστικά
- (Manifold learning algorithms), αναφέρεται σε τεχνικές ή αλγορίθμους που έχουν στόχο τη προβολή δεδομένων υψηλών διαστάσεων σε χαμηλότερες διαστάσεις

3.3 Επιλογή προτύπων (Prototype Selection)

Η επιλογή προτύπων, που ανήκει στην ευρύτερη κατηγορία αριθμητικής μείωσης των δεδομένων, είναι η εργασία εύρεσης των προτύπων εκπαίδευσης και δημιουργίας ενός σετ που να μπορεί να είναι αντιπροσωπευτικό του αρχικού συνόλου δεδομένων εκπαίδευσης [23]. Γενικά, η επιλογή προτύπων υποστηρίζεται από μια οικογένεια αλγορίθμων που μπορεί να διαχωριστεί σε δύο κατηγορίες, τους αλγορίθμους συμπύκνωσης και τους αλγορίθμους επεξεργασίας και πολλές φορές, το σετ που κατασκευάζουν ονομάζεται σετ συμπύκνωσης(condensing set). Έχουνε κοινό κίνητρο την εξαγωγή ενός λιγότερο ογκώδη σετ δεδομένων, όμως, ο τρόπος λειτουργίας και τα οφέλη τους διαφέρουν ουσιαστικά, καθώς, οι αλγόριθμοι συμπύκνωσης έχουν χαμηλές ανάγκες χωρητικότητας και μικρό υπολογιστικό κόστος με παράλληλη διατήρηση της απόδοσης στη φάση αξιολόγησης της κατηγοριοποίησης κάποιου μοντέλου. Στην άλλη περίπτωση, η επεξεργασία προτύπων δεν έχει σκοπό την επίτευξη υψηλών ποσοστών μείωσης δεδομένων αλλά την αύξηση της αποτελεσματικότητας(accuracy) και το καταφέρνουν αυτό διαχειρίζοντας τον



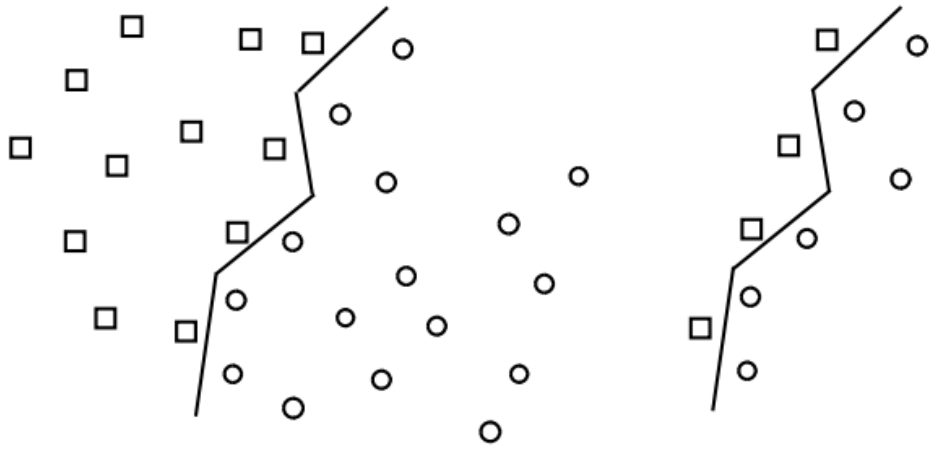
Σχήμα 3.2: Κατηγορίες τεχνικών μείωσης δεδομένων

θόρυβο στα δεδομένα και τα λανθασμένως κατηγοριοποιημένα δεδομένα με την εξομάλυνση των ορίων στο χώρο μεταξύ των κλάσεων με τελικό αποτέλεσμα την απότρεψη αλληλοεπικαλίψεων μεταξύ των κλάσεων.

3.3.1 Συμπύκνωση δεδομένων (CNN)

Ο παλαιότερος αλγόριθμος συμπύκνωσης δεδομένων που χρησιμοποείται και ως αναφορά για άλλους πιο εξιδεικευμένους αλγορίθμους είναι ο κανόνας συμπύκνωσης του πλησιέστερου γείτονα(Condensing Nearest Neighbor) και στηρίζεται στο ακόλουθο σκεπτικό [24]. Τα στιγμιότυπα που βρίσκονται στη πιο "εσωτερική" περιοχή της κάθε κλάσης αφαιρούνται με την οπτική ότι δεν συνεισφέρουν ουσιαστικά κατά την φάση κατηγοριοποίσης και συνεπώς η ακρίβεια μένει ανεπηρέαστη. Στο τέλος, έχει δημιουργήσει ένα συμπυκνωμένο σετ δεδομένων αποτελούμενο μόνο από στιγμιότυπα που είναι εγγύτερα στο όριο μεταξύ των κλάσεων και έχει καταφέρει να μειώσει ικανοποιητικά΄το υπολογιστικό κόστος και τις απαιτήσεις αποθήκευσης. Η αλγοριθμική διαδικασία σε μορφή ψευδοκώδικα παρουσιάζεται παρακάτω.

Αρχικά, θεωρείται άδειο το σύνολο συμπύκνωσης CS και προστίθεται εντός του ένα στιγμιότυπο. Στην συνέχεια, εφαρμόζεται ο κανόνας του εγγύτερου γείτονα(1-NN) για κάθε στιγμιότυπο που βρίσκετε στο σετ εκπαίδευσης TS με την σάρωση των στοιχείων που υπάρχουν στο CS. Εάν ο πλησιέστερος γείτονας στιγμιότυπο δεν είναι της ίδιας κλάσης με το στιγμιότυπο του TS, τότε θα μεταφερθεί το τελευταίο στο σύνολο συμπύκνωσης ενώ θα αφαιρεθεί από το TS. Η διαδικασία αυτή θα συνεχίζεται μέχρις ότου δεν υπάρχουν άλλες μετακινήσεις προτύπων από το TS στο CS σε ένα ολόκληρο πέρασμα δεδομένων του TS και στο τέλος το TS απορρίπτεται. Ο αλγόριθμος αυτός επιλέγει για το σύνολο συμπύκνωσης τα στοιχεία που είναι λανθασμένα κατηγοριοποιημένα γιατί θεωρεί ότι βρίσκονται κοντά στο όριο απόφασης και το



(a) Training set (b) Condensing set

Σχήμα 3.3: Αρχικά δεδομένα εκπαίδευσης και δεδομένα κοντά στο όριο κλάσεων

κάνει και για τα γειτονικά στοιχεία. Είναι ευαίσθητος σε υψηλά επίπεδα θορύβου και δεδομένου αυτού επηρεάζεται αρνητικά το ποσοστό μείωσης. Αυτό πρακτικά σημαίνει επίσης ότι όσες περισσότερες είναι οι κλάσεις, τόσες περισσότερα θα είναι και τα όρια απόφασης. Επομένως, το CS θα αυξηθεί με επακόλουθο την ελάττωση του ποσοστού μείωσης(reduction rate) κατα συνέπεια.

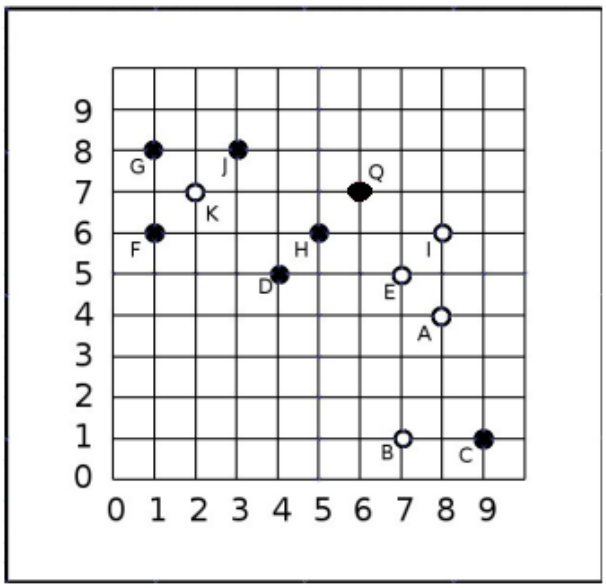
Algorithm 4 CNN-rule

Input: TS **Output:** CS

$CS \leftarrow \varnothing$
pick an item of TS and move it to CS
repeat
  $stop \leftarrow TRUE$
  for each $x \in TS$ **do**
    $NN \leftarrow$ Nearest Neighbour of x in CS
    if $NN_{class} \neq x_{class}$ **then**
      $CS \leftarrow CS \cup \{x\}$
      $TS \leftarrow TS - \{x\}$
      $stop \leftarrow FALSE$
    end if
  end for
until $stop == TRUE$ {no move during a pass of TS}
discard TS
return CS

Παρακάτω δίνετε ένα παράδειγμα εφαρμογής του CNN αλγορίθμου. Στο σχήμα τάδε έχουμε έναν σύνολο 12 στιγμιοτύπων στον Ευκλείδιο επίπεδο που ανήκουν στη κλάση 'λευκός κύκλος' είτε στη κλάση 'μαύρος κύκλος' και ζητείται η δημιουργία ενός συνόλου συμπύκνωσης. Η μετρική απόστασης που χρησιμοποιείται είναι η Ευκλείδια ενώ οι τιμές των χαρακτηριστικών των στιγμιοτύπων είναι ακέραιες ώστε να είναι οφθαλμοφανείς οι εγγύτεροι γείτονες για λόγους ευκολίας κατανόησης του τρόπου λειτουργίας του αλγορίθμου. Έστω ότι εξετάζονται τα στιγμιότυπα εκπάιδευσης με αλφαβητική σειρά. Συνεπώς,

εξετάζεται αρχικά το A ενώ το σύνολο συμπύκνωσης είναι κενό. Στη συνέχεια τοποθετείται το A στο CS γιατί είναι το πρώτο στιγμιότυπο. Μετά, εξετάζεται το επόμενο σε αλφαβητική σειρά σημείο B και ελέγχεται με ποιο από τα στιγμιότυπα του CS έχει την μικρότερη απόσταση. Προστοπαρών μόνο το A βρίσκεται στο CS και επειδή έχουν την ίδια κλάση, τότε απορρίπτεται το B. Επόμενο στιγμιότυπο είναι το C και τοποθετείτε στο CS επειδή είναι διαφορετικής κλάσης από το A. Η διαδικασία συνεχίζεται και για τα υπόλοιπα στιγμιότυπα με τον ίδιο τρόπο. Έπειτα ξεκινάει το επόμενο πέρασμα για τα υπόλοιπα στιγμιότυπα που απορρίφθηκαν προηγουμένως. Το πρώτο στιγμιότυπο που απορρίφθηκε στο προηγούμενο πέρασμα ήταν το B, άρα θα ελεγθεί ποιος είναι ο εγγύτερος γείτονας του σύμφωνα με τα στιγμιότυπα του CS που έχει κατασκευαστεί(A, C, D, K). Ο εγγύτερος γείτονας του B είναι το C, συνεπώς προστίθεται στο CS και εξετάζεται το επόμενο και ούτω καθεξής. Τελικά η διαδικασία τερματίζει καθ'ότι δεν συμβαίνει καμία μετακίνηση κατά το 3ο πέρασμα. Το τελικό συμπυκνωμένο σύνολο αποτελείται από τα : A, C, D, K, B, F, G, J.



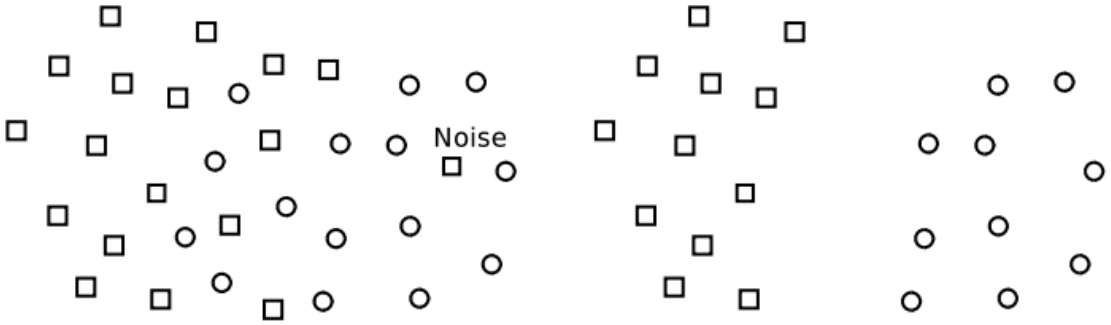
Σχήμα 3.4: Στιγμιότυπα δυαδικής κλάσης στο Ευκλείδιο επίπεδο

Ένα από τα πλεονεκτήματα του CNN αλγορίθμου είναι ο αυτόματος καθορισμός του πλήθους των προτύπων για το σετ συμπύκνωσης που σημαίνει ότι δεν προαπαιτεί ορισμό κάποιας παραμέτρου. Επιπλέον, είναι ικανός για τη σωστή κατηγοριοποίηση των απορριπτόμενων προτύπων εκπαίδευσης με την εφαρμογή του μοντέλου 1-NN στο CS. Παρ'όλα τα θετικά του, πρέπει να αναφερθούμε και στα αδύνατα σημεία του. Η διαφορετική σειρά σάρωσης των δεδομένων εκπαίδευσης παράγει ένα διαφορετικό σετ συμπύκνωσης και συνεπώς, η διαφορετική ταξινόμηση των στοιχείων έχει σημασία. Ακόμη, είναι ακατάλληλος για χρήση σε περιβάλλοντα πραγματικού χρόνου διότι τα υποσύνολα συμπύκνωσης που δημιουργεί είναι αρρηκτά συνδεδεμένα με τα αρχικά TS δεδομένα. Συνεπώς, η τροφοδότηση νέων δεδομένων εκπαίδευσης θα γεννήσουν διαφορετικά CS και για να αντιμετωπιστεί αυτό θα πρέπει να γίνει επανεκτέλεση του αλγορίθμου με είσοδο το TS που θα περιλαμβάνει τα παλιά και τα νέα δεδομένα. Τέλος, μια προυπόθεση για την λειτουργία του είναι η μόνιμη διατήρηση των προτύπων εκπαίδευσης στη μνήμη.

3.3.2 Επεξεργασία δεδομένων για απομάκρυνση θορύβου (ENN)

Οι αλγόριθμοι επεξεργασίας δεδομένων επεξεργάζονται τα δεδομένα εκπαίδευσης δημιουργώντας βελτιωμένα σετ δεδομένων μέσω της αφαίρεσης απομακρυσμένων προτύπων που βρίσκονται στο χώρο, αφαίρεσης λανθασμένων κατηγοριοποιήσεων και του θορύβου. Παρακάτω γίνεται η περιγραφή του πιο ίσως δημοφιλούς αλγορίθμου επεξεργασίας, τον λεγόμενο επεξεργασμένο εγγύτερο γείτονα(Edited Nearest Neighbor).

Η βάση και η συλλογιστική όλων τον αλγορίθμων επεξεργασίας επικεντρώνεται στον ENN κανόνα, έναν πολύ απλό στην υλοποίηση του αλγόριθμο [25]. Παρακάτω επισυνάπτεται ο κώδικας του σε μορφή ψευδογλώσσας. Στο πρώτο βήμα, ορίζεται σαν είσοδος η παράμετρος k ως ο παράγοντας αφαίρεσης ή διατήρησης του προς εξέταση προτύπου και επίσης, φυσικά η είσοδος του σετ εκπαίδευσης TS. Το σετ αυτό αρχικά είναι ταυτισμένο με το επεξεργασμένο σύνολο ES και η διαδικασία προχωράει ως εξής. Για κάθε κάθε στιγμιότυπο, ανιχνεύονται οι k εγγύτεροι γείτονες του και αν αποκαλυφθεί ότι οι γείτονες αυτοί είναι διαφορετικά κατηγοριοποιημένοι στην πλειοψηφία τους, σε σύγκριση με το στιγμιότυπο εξέτασης, τοτε αποφασίζεται η αφαίρεση του στιγμιοτύπου από το ES, ενώ διαφορετικά, συνεχίζεται η διαδικασία για τα υπόλοιπα. Είναι σημαντικό να σημειωθεί ότι οι υπολογισμοί απόστασης των στιγμιοτύπων στον επαναληπτικό βρόχο για την εύρεση των γειτόνων γίνετε με το αρχικό σύνολο εκπαίδευσης και όχι με το προς κατασκευή σύνολο.



Σχήμα 3.5: Ομαλοποίηση των ορίων απόφασης και επεξεργασία θορύβου

Είναι γεγονός ότι το κόστος επεξεργασίας αυτού του αλγορίθμου είναι στενά συνδεδεμένο με το μέγεθος του σετ εκπαίδευσης διότι όσο είναι μεγαλύτερο τόσο περισσότεροι είναι και οι υπολογισμοί απόστασης μεταξύ των στιγμιοτύπων εκπαίδευσης και συγκεκριμένα, το πλήθος υπολογισμών μπορεί να οριστεί από τον τύπο:

$$\frac{N \times (N-1)}{2} \tag{6}$$

όπου N, ο αριθμός των αντικειμένων στο σετ εκπαίδευσης. Επίσης, ένα άλλο σημείο άξιο προσοχής είναι η απόφαση της τιμής της παραμέτρου k, καθώς μη κατάλληλες τιμές του k μπορούν να προκαλέσουν την αφαίρεση στιγμιοτύπων που δεν είναι θόρυβος. Πολλές μελέτες και πειράματα προτείνουν ότι το k=3 είναι μια καλή προσέγγιση ή αλλιώς, ορισμένες φορές προτείνεται η κατάλληλη εύρεση του k μέσω της μεθόδου δοκιμής και σφάλματος(trial and error). Όμως γενικά, μπορεί να είναι αποτελεσματικότερο να έχουμε διαφορετικές τιμές k για διαφορετικές περιοχές στο χώρο.

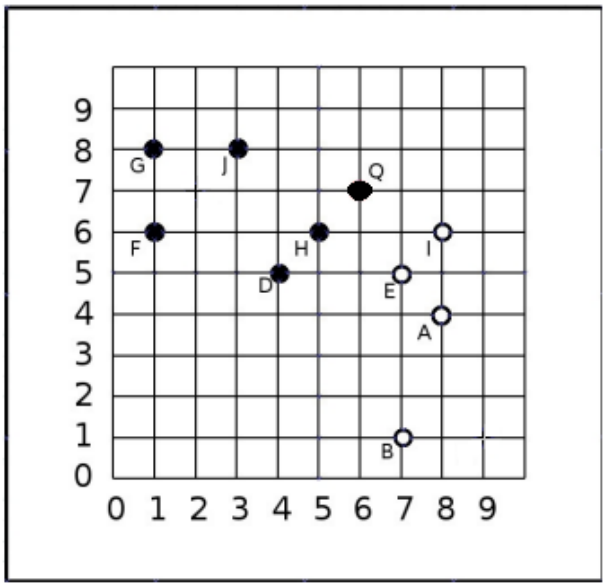
Algorithm 1 ENN-rule

Input: TS, k
Output: ES

1: $ES \leftarrow TS$
2: **for each** $x \in TS$ **do**
3: $NNs \leftarrow$ find the k nearest to x neighbors in $TS - \{x\}$
4: $majorClass \leftarrow$ find the most common class of NNs
5: **if** $x_{class} \neq majorClass$ **then**
6: $ES \leftarrow ES - \{x\}$
7: **end if**
8: **end for**
9: **return** ES

Σχήμα 3.6: Ψευδοκώδικας του ENN κανόνα

Ένα απλό παράδειγμα εφαρμογής της διαδικασίας του αλγορίθμου ENN για παράμετρο n=3 είναι το εξής. Έστω ότι έχουμε 12 στιγμιότυπα στον Ευκλείδιο χώρο με συντεταγμένες ακέραιων αριθμών και δύο κλάσεις, τον λευκό και μαύρο κύκλο, όπως φαίνεται στο 3.4. Αρχικά εξετάζονται αλφαβητικά τα στιγμιότυπα και το επεξεργασμένο σύνολο ES περιέχει όλα τα στιγμιότυπα. Το σημείο A που είναι το πρώτο αλφαβητικά σημείο έχει δύο εγγύτερους γείτονες κλάσης λευκό και έναν μαύρο. Η πλειοψηφία των γειτόνων είναι ίδιας κλάσης με το A, συνεπώς δεν γίνεται κάποια αλλαγή και προχωράει η διαδικασία με το B. Το στιγμιότυπο B έχει εγγύτερους γείτονες τους E, A και C. Η πλειοψηφία των γειτόνων έχει την ίδια κλάση με το B άρα ομοίως δεν θα γίνει κάποια αλλαγή. Ωστόσο, το στιγμιότυπο C(κλάσης μαύρο) με εγγύτερους γείτονες τους A, B, E θα αφαιρεθεί από το CS γιατί τα A, B είναι διαφορετικής κλάσης από το C. Η διαδικασία συνεχίζεται με την ίδια συλλογιστική και ολοκληρώνει όταν γίνει έλεγχος για όλα τα στιγμιότυπα. Επίσης, ο έλεγχος για κάθε στιγμιότυπο γίνετε με το συνολικό σύνολο εκπαίδευσης και όχι με το σύνολο προς κατασκευή. Στο τέλος θα έχουν αφαιρεθεί τα C και K από το CS.



Σχήμα 3.7: Σύνολο συμπύκνωσης με τον ENN αλγόριθμο

3.4 Παραγωγή προτύπων (Prototype Generation)

Το κίνητρο της εφαρμογής αλγορίθμων παραγωγής προτύπων είναι η ελάττωση του υπολογιστικού κόστους και η αντιμετώπιση των υψηλών απαιτήσεων χώρου αποθήκευσης στις διαδικασίες της κατηγοριοποίσης. Οι αλγόριθμοι συμπύκνωσης έχουν εξίσου το ίδιο σκοπό όμως διαφέρουν ως προς τον τρόπο προσέγγισης της κατασκευής του μειωμένου υποσυνόλου καθώς όπως περιγράφτηκε και σε προηγούμενο κεφάλαιο, οι αλγόριθμοι συμπύκνωσης επιλέγουν ένα δείγμα από ευρύτερο σύνολο εκπαίδευσης. Οι αλγόριθμοι παραγωγής προτύπων από την άλλη, δημιουργούν νέα στιγμιότυπα παρόμοιας κατασκευής, με μια διαδικασία σύνοψης των προϋπαρχόντων προτύπων εκπαίδευσης [26]. Συνεπώς, μπορούμε να ειπωθεί ότι το κατασκευασμένο σύνολο που θα εκτελεστεί από ένα k-NN μοντέλο αποτελεί ένα τεχνητό σύνολο δεδομένων. Στο κεφάλαιο αυτό, θα γίνει η περιγραφή του RSP3 αλγορίθμου καθώς και των γεννήτορων του.

3.4.1 Chen and Jozwik αλγόριθμος

Ένας αποτελεσματικός αλγόριθμος παραγωγής προτύπων είναι αυτός του Chen και Jozwik(CJA) [27]. Αρχικά, ο χρήστης ορίζει μια παράμετρο n για να προσδιορίζει το επιθυμητό πλήθος προτύπων που χρειάζεται να παραχθούν και φυσικά τροφοδοτεί σαν είσοδο τα στιγμιότυπα εκπαίδευσης. Στο πρώτο βήμα, εντοπίζονται τα δύο πιο απομακρυσμένα στιγμιότυπα(στιγμιότυπο x και y) του συνόλου εκπαίδευσης με βάση μιας μετρικής απόστασης, συνήθως την ευκλείδια, και αφού βρεθούν, διαχωρίζονται τα συνολικά δεδομένα σε δύο επιμέρους υποσύνολα. Στο σύνολο Sx κατατάζονται εκείνα τα στιγμιότυπα που βρίσκονται πιο κοντά στο x ενώ εκείνα που βρίσκονται πιο κοντά στο y καταλήγουν στο σύνολο Sy. Στη συνέχεια, ο CJA επιλέγει να διαιρέσει σε επιμέρους υποσύνολα με βάση κάποιου κριτηρίου ομοιότητας, δηλαδή της ύπαρξης περισσότερων από μιας κλάση προτύπων της ίδιας περιοχής και η διαδικασία διαίρεσης θα ξεκινήσει πρώτα με τα μεγαλύτερης απόστασης πιο απομακρυσμένα στιγμιότυπα του χώρου αυτού. Ύστερα, συνεχίζει με τη δημιουργία υποσυνόλων εντός των ομοιογενών συνόλων εφόσον δεν υπάρχουν άλλα ομοιογενή και θα σταματήσει μέχρις ότου το πλήθος των υποσυνόλων συνολικά είναι ίσο με τι τιμή n που έχει ορίστει από τον χρήστη. Τέλος, δημιουργεί ένα στιγμιότυπο εκπαίδευσης για κάθε επιμέρους υποσύνολο S με τον υπολογισμό της μέσης τιμής των στοιχείων του, τα οποία στοιχεία μέσης τιμής θα αποτελέσουν το τελικό σετ συμπύκνωσης και η ετικέτα κλάσης του εκάστοτε στοιχείου προσδιορίζεται από τη πλειοψηφική ετικέτα κλάσης του αντίστοιχου συνόλου προέλευσης S. Πιο συγκεκριμένα, τα συνοψισμένα αντικείμενα παράγονται ως εξής :

$$m.dj = \frac{1}{|S|} \sum_{x_i \in S} x_i.d_j, j = 1, 2, ..., t \tag{7}$$

όπου m είναι το μέσο στιγμιότυπο του κάθε υποσυνόλου S και η τιμή για το κάθε χαρακτηριστικό(feature) του προκύπτει από τον μέσο όρο χαρακτηριστικού σύμφωνα με το σύνολο των στιγμιοτύπων στο σετ δεδομένων.

Παρακάτω παρατίθεται η αλγοριθμική διαδικασία σε μορφή ψευδοκώδικα.

Όπως αναφέραμε προηγουμένως, η μεθοδολογία επιλογής των υποσυνόλων που θα υποδιαιρεθούν βασίζεται στον υπολογισμό της διαμέτρου αλλά δεν εξηγήσαμε πως αυτή η λογική στηρίζεται στην ιδέα

Algorithm 6 CJA

Input: TS, n
Output: CS

$S \leftarrow \varnothing$
add(S, TS)
for $i = 2$ to n **do**
  $C \leftarrow$ select the non-homogeneous subset $\in S$ with the largest diameter
  if $C == \varnothing$ {All subsets are homogeneous} **then**
    $C \leftarrow$ select the homogeneous subset $\in S$ with the largest diameter
  end if
  $(S_x, S_y) \leftarrow$ divide C into two subsets
  add(S, S_x)
  add(S, S_y)
  remove(S, C)
end for
$CS \leftarrow \varnothing$
for each subset $T \in S$ **do**
  $r \leftarrow$ compute the mean item by averaging the items in T
  $r.label \leftarrow$ find the most common class label in T
  $CS \leftarrow CS \cup \{r\}$
end for
return CS

Σχήμα 3.8: Ψευδοκώδικας του αλγορίθμου Chen and Jozwik

ότι τα υποσύνολα με τη μεγαλύτερη διάμετρο θα απαρτίζονται και από μεγαλύτερο πλήθος στοιχείων εκπαίδευσης. Συνεπώς, είναι πιο επιθυμητά γιατί επιτυγχάνουν μεγαλύτερα ποσοστά μείωσης με την διαίρεση τους σε σειρά προτεραιότητας. Ένα από τα πλεονεκτήματα του CJA αλγορίθμου είναι η ικανότητα του να δημιουργεί το ίδιο σύνολο συμπύκνωσης ανεξαρτήτως ταξινόμησης των δεδομένων. Ένα ομως αδύνατο σημείο του είναι ανάγκη εισόδου του αριθμού των παραγόμενων προτύπων που ναι μεν μπορεί ο προσδιορισμός του μεγέθους να είναι ενα οφέλιμο σε μερικές περιπτώσεις χαρακτηριστικό, ωστόοσο δεν έχει την ικανότητα να προσδιορίζει τα μεγέθη βάσει συμπερίληψης της ιδιότητας της φύσης του εκάστοτε συνόλου δεδομένων.

3.4.2 Reduction by Space Partitioning αλγόριθμοι (RSP)

Οι reduction by space partitioning αλγόριθμοι, γνωστοί και ως RSP, είναι μία ομάδα τριών αλγορίθμων παραγωγής προτύπων(RSP1, RSP2, RSP3) και είναι προιόντα δημιουργίας του CJA, του οποίου ο τρόπος λειτουργίας περιγράφτηκε στο προηγούμενο υποκεφάλαιο [28]. Σε αντίθεση με τον CJA, ο RSP1, για κάθε υποσύνολο δημιουργεί τόσα στιγμιότυπα εκπαίδευσης όσες είναι και οι διαφορετικές κλάσεις που υπάρχουν εντός του, με αποτέλεσμα να πετυχαίνει μεγαλύτερα ποσοστά ακρίβειας γιατί δεν αγνοεί τις μειονοτικές περιπτώσεις εκπαίδευσης. Αυτό έρχεται όμως με κόστος την αύξηση του τελικού συνόλου συμπύκνωσης εφόσον εφόσον αυξάνεται ο αριθμός των μέσων στοιχείων παραγωγής. Η κύρια διαφορά μεταξύ του RSP1 και του RSP2 έγκειται στο κριτήριο επιλογής του επόμενου προς διαίρεση υποσυνόλου. Ο RSP1, όπως ακριβώς και ο CJA, στηρίζεται στο κριτήριο της μεγαλύτερης διαμέτρου(diameter

criterion) υποθέτοντας ότι το υποσύνολο που ικανοποιεί αυτή τη προυπόθεση θα συνεπάγεται μεγαλύτερο μέγεθος συνόλου. Με μια διαφορετική προσέγγιση, ο RSP2 εφαρμόζει το κριτήριο της μεγαλύτερης αλληλοεπικάλυψης(overlapping degree criterion), το οποίο υποθέτει ότι τα στοιχεία της ίδιας κλάσης βρίσκονται κοντά μεταξύ τους ενώ τα στοιχεία διαφορετικών κλάσεων βρίσκονται μακρία. Ο βαθμός αλληλοεπικάλυψης υπολογίζεται από το πηλίκο της μέσης απόστασης των αντικειμένων διαφορετικής κλάσης με την μέση απόσταση των αντικειμένων της ίδιας κλάσης [29].

Όσον αφορά τον RSP3 αλγόριθμο, η στρατηγική διαίρεσης των ανομοιογενών συνόλων σε επιμέρους υποσύνολα ακολουθεί παρομοίως την έννοια της ομοιογένειας και μπορεί να γίνει είτε με το κριτήριο της μεγαλύτερης διαμέτρου είτε με τον μεγαλύτερο βαθμό αλληλοεπικάλυψης, με συνθήκη τερματισμού στην κατάσταση εκείνη όπου όλα τα υποσύνολα είναι ομοιογενή [30]. Ο αλγόριθμος RSP3 είναι καλύτερος των προηγουμένων καθως είναι μη παρεμετρικός. Αυτό σημαίνει δηλαδή ότι το πλήθος των παραγώμενων προτύπων διαμορφώνεται αυτόματα και όχι με τη χρήση ορίσματος από τον χρήστη. Αναφέρεται ακόμη, ότι το τελικό σετ συμπύκνωσης δεν αλλάζει ανεξαρτήτως της σειράς που έχουν τα δεδομένα εκπαίδευσης και το ίδιο χαρακτηριστικό ισχύει και με την εφαρμογή των αλγορίθμων CJA, RSP1 και RSP2.

Algorithm 7 RSP3

Input: TS
Output: CS

1: $S \leftarrow \varnothing$
2: add(S, TS)
3: $CS \leftarrow \varnothing$
4: **repeat**
5: $C \leftarrow$ select the subset $\in S$ with the highest splitting criterion value
6: **if** C is homogeneous **then**
7: $r \leftarrow$ calculate the mean item by averaging the items in C
8: $r.label \leftarrow$ class of items in C
9: $CS \leftarrow CS \cup \{r\}$
10: **else**
11: $(D_1, D_2) \leftarrow$ divide C into two subsets
12: add(S, D_1)
13: add(S, D_2)
14: remove(S, C)
15: **end if**
16: **until** IsEmpty(S)
17: **return** CS

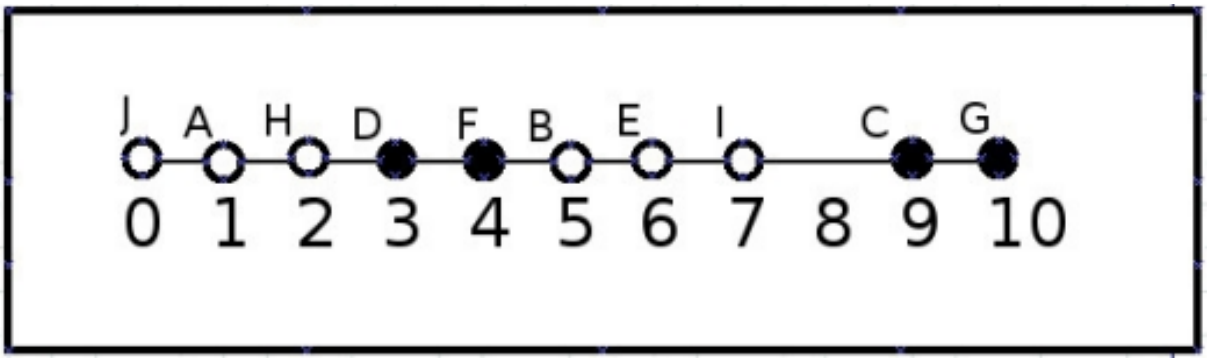
Σχήμα 3.9: Ψευδοκώδικας του RSP3 αλγορίθμου

Παραπάνω παρατίθεται ο κώδικας ψευδογλώσσας του RSP3. Στην αρχή, χρησιμοποιείται μια δομή δεδομένων με όνομα μεταβλητής S στην οποία αποθηκεύονται όλα τα στιγμιότυπα εκπαίδευσης και μια άδεια δομή δεδομένων με όνομα CS όπου πρόκειται να καταχωρηθούν τα παραγόμενα στιγμιότυπα. Μετά σαρώνεται το S για την εύρεση του υποσυνόλου C αυτού με το κριτήριο της μεγαλύτερης τιμής διαχωρισιμότητας(highest criterion value) και εξετάζεται αν το C είναι ομοιογενές ή όχι. Στην περίπτωση που είναι ομοιογενές, δημιουργείται το μέσο στιγμιότυπο από τον υπολογισμό του μέσου όρου των προτύπων του C, εκχωρείται η ετικέτα κλάσης του ίση με αυτή του C και τέλος προστίθεται το στιγμιό-

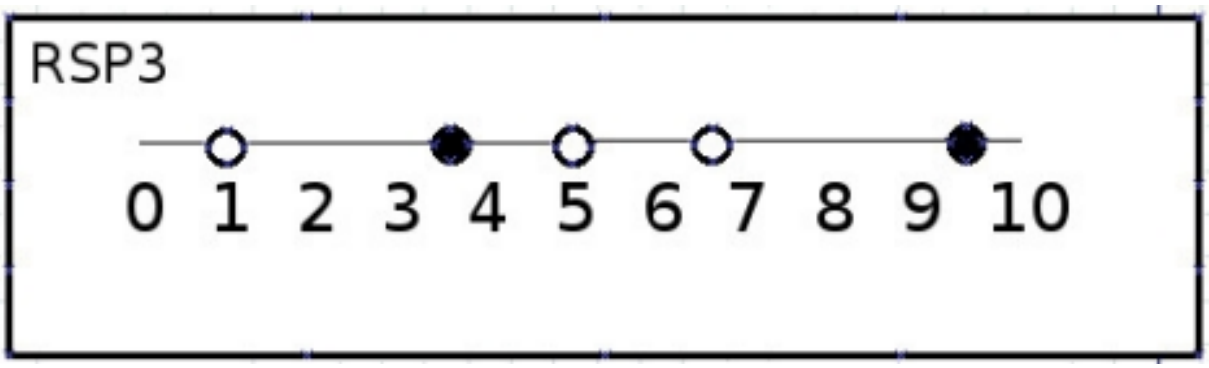
τυπο παραγωγής στο σύνολο συμπύκνωσης CS. Διαφορετικά, αν το υποσύνολο C δεν είναι ομοιογενές, τότε διαιρείται σε δύο υποσύνολα(D1, D2) με τον ίδιο τρόπο όπως γίνετε στον CJA και προστίθενται στη δομή S ενώ το C αφαιρείται. Η επαναληπτική διαδικασία τερματίζεται όταν η δομή S γίνει άδεια, δηλαδή όταν όλα τα υποσύνολα γίνουν ομοιογενή.

Ο θόρυβος στα δεδομένα επηρεάζει άμεσα το ποσοστό μείωσης των δεδομένων από τον RSP3 αλγόριθμο. Ο αλγόριθμος αυτός τείνει να καταφέρνει μικρότερα ποσοστά μείωσης δεδομένων όταν μεγαλώνει το ποσοστό θορύβου γιατί τα υποσύνολα που δημιουργούνται τείνουν να είναι μικρότερα σε αυτή τη περίπτωση. Ακόμη, μια αδυναμία του RSP3 είναι το υψηλό υπολογιστικό κόστος διότι η διαδικασία εύρεσης των πιο απομακρυσμένων στοιχείων στο υποσύνολο πραγματοποιείται με τον υπολογισμό όλων των αποστάσεων μεταξύ των στοιχείων που περιλαμβάνει. Το γεγονός αυτό μας οδηγεί στο συμπέρασμα ότι η προτίμηση τέτοιων αλγορίθμων μπορεί να είναι απαγορευτική για σύνολα δεδομένων μεγάλου μεγέθους.

Στο σχήμα 3.10 βλέπουμε ένα σύνολο δεδομένων κατηγοριοποίησης δυαδικής κλάσης στο μονοδιάστατο χώρο όπου θα εφαρμοστεί ο αλγόριθμος RSP3. Αρχικά διαιρείται όλο το σύνολο δεδομένων σε δύο υποσύνολα και τοποθετούνται στο καθένα εκείνα τα στιγμιότυπα που βρίσκονται πιο κοντά στο δύο πιο απομακρυσμένα στιγμιότυπα του συνόλου. Για παράδειγμα τα στιγμιότυπα A, J, H, D, F, B βρίσκονται πιο κοντά στο J οπότε θα δημιουργήσουν ένα υποσύνολο ενώ τα E, I, C βρίσκονται πιο κοντά στο G και αντίστοιχα θα δημιουργήσουν ένα υποσύνολο. Στο επόμενο βήμα επιλέγεται εκείνο το υποσύνολο που έχει τη μεγαλύτερη διάμετρο, δηλαδή το {J,A,H,D,F,B} καθότι η απόσταση μεταξύ των στιγμιοτύπων J και B έχει μήκος 5-0=5 ενώ το υποσύνολο E,I,C,G έχει μήκος μόλις 10-6=4. Το υποσύνολο που διαλέχθηκε δεν είναι ομοιογενές αφού περιέχει στιγμιότυπα διαφορετικών κλάσεων εντός του, άρα ο αλγόριθμος συνεχίζει διαιρώντας το υποσύνολο σε δύο υποσύνολα με τον ίδιο τρόπο που έγινε και προηγουμένως. Ο αλγόριθμος πριν τερματίσει θα έχει δημιουργήσει ένα πλήθος υποσυνόλων τα οποία θα είναι όλα ομοιογενή και σε εκείνο το χρονικό σημείο θα κατασκευάσει ένα μέσο στιγμιότυπο για κάθε διαφορετικό υποσύνολο. Τα υποσύνολα αυτά θα είναι τα εξής : {J,A,H} , {E,I} , {C,G} , {D,F} και {B}. Στο σχήμα 3.11 φαίνεται το τελικό σύνολο συμπύκνωσης.



Σχήμα 3.10: Δεδομένα στο μονοδιάστατο χώρο



Σχήμα 3.11: Σύνολο συμπύκνωσης με τον RSP3 αλγόριθμο

3.5 Συνδυασμός εκτέλεσης τεχνικών μείωσης δεδομένων

Η μοντελοποίηση ενός προβλήματος κατηγοριοποίησης είναι μια διαδικασία που μπορεί να χωριστεί σε δύο φάσεις εκτέλεσης. Την φάση της προεπεξεργασίας(data pre-processing), που μπορεί να αναλάβει αρμοδιότητες για την αφαίρεση ανεπιθύμητων δεδομένων, θορύβου και γενικά δραστηριότητες μείωσης του όγκου των δεδομένων. Η δεύτερη φάση εκτέλεσης περιλαμβάνει την εκπαίδευση του μοντέλου με τα καινούργια δεδομένα εκπαίδευσης και τελικά η δημιουργία προβλέψεων.Η προεπεξεργασία είναι ένα προαιρετικό βήμα και γενικά υπάρχουν τέσσερις ξεχωριστές περιπτώσεις υλοποίησης στην κατηγοριοποίηση :

1. Καμία προεπεξεργασία, αν το σετ εκπαίδευσης δεν περιέχει θόρυβο ούτε ανεπιθύμητα δεδομένα και αν το σύνολο δεδομένων είναι μικρό
2. μόνο επεξεργασία, αν το σύνολο δεδομένων είναι μικρό αλλά περιέχει θόρυβο
3. μόνο συμπύκνωση, αν το σύνολο δεδομένων είναι μεγάλο αλλά είναι καθαρό απο θόρυβο
4. επεξεργασία και συμπύκνωση, αν το σύνολο δεδομένων είναι μεγάλο και υπάρχει θόρυβος

Όπως αναφέραμε σε προηγούμενη ενότητα, οι αλγόριθμοι συμπύκνωσης και επεξεργασίας είναι διαφορετικές ως προς το κίνητρο τους καθώς οι μεν αλγόριθμοι συμπύκνωσης στοχεύουν στην επίτευξη υψηλών ποσοστών μείωσης των δεδομένων για την μείωση του υψηλού υπολογιστικού κόστους και απαιτήσεων αποθήκευσης ενώ οι αλγόριθμοι επεξεργασίας στοχεύουν στη βελτίωση της αποτελεσματικότητας μέσω της αφαίρεσης θορύβου, ακραίων τιμών και λανθασμένων κατηγοριοποιήσεων. Για παράδειγμα, σε ένα περιβάλλον με θόρυβο δεν θα ανταποκρίνεται καλά ο αλγόριθμος CNN γιατί κανονικά κατασκευάζει το σύνολο συμπύκνωσης από στιγμιότυπα που βρίσκονται κοντά στο όριο απόφασης. Η ύπαρξη θορύβου θα εμποδίσει την ομαλή διαδικασία με αποτέλεσμα τη μείωση του ποσοστού μείωσης δεδομένων. Όμως, ένας αλγόριθμος σαν τον ENN, μπορεί να εντοπίσει τον θόρυβο και να τον αφαιρέσει αποτρέποντας έτσι το πρόβλημα. Επίσης, η εφαρμογή μόνο ενός αλγορίθμου επεξεργασίας στο βήμα προεπεξεργασίας μπορεί να μην είναι μια ιδανική λύση για περιπτώσεις μεγάλων συνόλων δεδομένων, γι'αυτό συμβουλεύεται να ακολουθήσει μια διαδικασία συμπύκνωσης. Συνεπώς, είναι αντιληπτό ότι πολλές φορές επιδιώκεται κατά την φάση προεπεξεργασίας μια συνδυαστική υλοποίηση ενός αλγορίθμου αφαίρεσης θορύβου και μετά ενός αλγορίθμου συμπύκνωσης για την καταπολέμηση των αδυναμιών και ιδαιτεροτήτων του κάθε αλγορίθμου.

Κεφάλαιο 4ο: Υλοποιήσεις σε Python

Για την υλοποίηση της πειραματικής μελέτης σε python έγινε η αξιοποίηση διαφόρων βιβλιοθηκών. Η βιβλιοθήκη numpy φάνηκε χρήσιμη για την δημιουργία δομών πινάκων και την υποστήριξη της με ενσωματωμένες συναρτήσεις, όπως για παράδειγμα συνάρτηση υπολογισμού μέσων τιμών. Επίσης, το λογισμικό βιβλιοθήκης pandas βοήθησε στην αναπαράσταση των συνόλων δεδομένων σε μορφή δομής δεδομένων dataframe και στον εύκολο χειρισμό των δεδομένων με πολλές built-in συναρτήσεις. Τέλος, το μοντέλο του 1-NN κατηγοριοποιητή καθώς και η εκπαίδευση του μοντέλου υποστηρίχτηκαν από την βιβλιοθήκη μηχανικής μάθησης scikit-learn [31].

4.1 Τεχνικές Δειγματοληψίας

Σε αυτό το κεφάλαιο περιγραφεται αναλυτικά η προγραμματιστική υλοποίηση των τεχνικών δειγματοληψίας που χρησιμοποιήθηκαν στα πειράματα της συγκριτικής μελέτης.

Τυχαία απλή δειγματοληψία

```
def random_sampling(df, fraction):
    random_sample = df.sample(frac=fraction, replace=False)
    return random_sample
```

Η συνάρτηση random_sampling() δέχεται σαν όρισμα ένα pandas dataframe στο οποίο είναι αποθηκευμένο το σύνολο δεδομένων και το κλάσμα δειγματοληψίας fraction, το οποίο παίρνει τιμή ανάλογη της τιμής που υπολογίστηκε για το αναφερόμενο σετ δεδομένων. Έπειτα, καλείται η ενσωματωμένη συνάρτηση sample() η οποία επιστρέφει από το df ένα τυχαίο δείγμα χωρίς αντικατάσταση(replace=False) με μέγεθος που αποφασίζεται από το όρισμα fraction και τέλος επιστρέφει το δείγμα.

Στρωματοποιημένη δειγματοληψία

```
def stratified_sampling(df, fraction):
    stratified_sample = df.groupby(df.iloc[:, -1], group_keys=False)
    .apply(lambda x: x.sample(frac=fraction))
    return stratified_sample
```

Παρομοίως, δέχεται σαν όρισμα το σύνολο δεδομένων σε dataframe της βιβλιοθήκης pandas και δεύτερο όρισμα τη τιμή του κλάσματος δειγματοληψίας. Η δειγματοληψία σε στρώματα επιτυγχάνεται με τη

χρήση της ενσωματωμένης συνάρτησης groupby() όπου στη συγκεκριμένη εφαρμογή δέχεται δύο ορίσματα. Το πρώτο όρισμα, που είναι και το πιο σημαντικό, προσδιορίζει τη στήλη με βάση της οποίας θα δημιουργηθούν ομάδες στι οποίες θα χωριστεί όλο το σετ δεδομένων. Στην δική μας περίπτωση, η στήλη που επιλέγεται είναι αυτή που περιλαμβάνει τις ετικέτες κλάσεων όλων των δεδομένων εκπαίδευσης και συνεπώς η συνάρτηση groupby() θα χωρίσει τα δεδομένα εκπαίδευσης σε τόσες ομάδες όσες είναι οι ετικέτες κλάσεις και επίσης θα συμπεριλάβει τα στιγμιότυπα εκπαίδευσης στις αντίστοιχα στρώματα. Η συνάρτηση επιστρέφει ένα groupby αντικείμενο που περιέχει τις πληροφορίες για τα στρώματα και στο αντικείμενο αυτό έπειτα, εφαρμόζεται η συνάρτηση apply(). Η συνάρτηση αυτή επιλέγει ένα τυχαίο δείγμα στοιχείων ξεχωριστά από κάθε στρώμα και σε ποσοστό που ορίζει το κλάσμα δειγματοληψία. Τέλος, επιστρέφει ένα δείγμα αποτελούμενο από όλα τα τυχαία δείγματα που λήφθηκαν προηγουμένως.

Συστηματική δειγματοληψία

```
def systematic_sampling(df,starting_point, step):
    indexes = np.arange(starting_point, len(df), step=step)
    systematic_sample = df.iloc[indexes]
    return systematic_sample
```

Η συνάρτηση για την συστηματική δειγματοληψία έχει σαν ορίσματα το σετ δεδομένων df, το starting_point σαν δείκτης εκίνησης της διαδικασίας και τέλος το βήμα που προσδιορίζει το διάστημα δειγματοληψίας. Η διαδικασία υλοποιείται εύκολα με τη χρήση της συνάρτησης arrange() της βιβλιοθήκης numpy, στην οποία ορίζεται ένας δείκτης εκίνησης, το μέγεθος του σετ εκπαίδευσης και το διάστημα δειγματοληψίας. Με αυτές τις παραμέτρους δημιουργεί έναν πίνακα δεικτών(indexes) ίσων μεταξύ τους διαστημάτων με διάστημα τον ακέραιο αριθμό step. Τέλος, αντιστοιχίζει τον πίνακα δεικτών με το σετ δεδομένων για να αποθηκεύσει στη μεταβλητή systematic_sample τα αντίστοιχα στιγμιότυπα εκπαίδευσης.

4.2 Αξιολόγηση ακρίβειας(Accuracy evaluation)

```
knn_model = KNeighborsClassifier(n_neighbors=1)
```

```
def knnAccuracy(model, train_data, test_data, algor):
    lst_accuracy = []
    for x in range(0, 5):
        df = train_data[x]
        x_train_fold = train_data[x].iloc[:, :len(df.columns) - 1]
        y_train_fold = train_data[x].iloc[:, -1]
        x_test_fold = test_data[x].iloc[:, :len(df.columns) - 1]
```

```
        y_test_fold = test_data[x].iloc[:, -1]
        model.fit(x_train_fold, y_train_fold)
        lst_accuracy.append(model.score(x_test_fold, y_test_fold))
    print('Overall Accuracy for  :', algor, mean(lst_accuracy) * 100, '%')
```

Η μοντελοποίηση του κατηγοριοποιητή των k εγγύτερων γειτόνων έγινε από τη βιβλιοθήκη sklearn. Η παράμετρος n_neighbors=1 ορίζει το k=1, δηλαδή την εφαρμογή του αλγορίθμου του εγγύτερου γείτονα(1-NN) και για μετρική απόστασης χρησιμοποιεί την ευκλείδια απόσταση(default τιμή). Η τεχνική αξιολόγησης της ακρίβειας που εφαρμόζεται είναι η k-fold-cross-validation για k=5.Αρχικά, τα δεδομένα εκπαίδευσης και δοκιμής είναι χωρισμένα σε 5 αρχεία και κάθε αρχείο αντιστοιχεί σε ένα fold. Για κάθε fold γίνετε η ανάθεση των αντίστοιχων δεδομένων εκπαίδευσης και δοκιμής στις μεταβλητές x_train_fold, y_train_fold, x_test_fold, y_test_fold και χρησιμοποιείται η ενσωματωμένη συνάρτηση fit() που δέχεται σαν όρισματα τα χαρακτηριστικά εκπαίδευσης καθώς και τις πραγματικές ετικέτες κλάσης για την εκπαίδευση του μοντέλου. Ύστερα, χρησιμοποιείται η συνάρτηση score() για να αξιολογήσει την ακρίβεια που πετυχαίνει το μοντέλο πάνω σε νέα δεδομένα, δηλαδή στα δεδομένα δοκιμής και επιστρέφει την ακρίβεια ως αριθμό μεταξύ του 0 και του 1. Πιο συγκεκριμένα, η συνάρτηση score υπολογίζει της αποστάσεις μεταξύ των δεδομένων εκπαίδευσης και δοκιμής για να εντοπίσει τον κοντινότερο γείτονα. Αν έχουν έχουν ίδια τιμή ετικέτας κλάσης τότε το μοντέλο έχει κανει σωστή πρόβλεψη, διαφορετικά είναι λαθνασμένη. Τελος, ο υπολογισμός της ακρίβειας γίνετε από το πηλίκο των συνολικά σωστών προβλέψεων δια τον συνολικό αριθμό των προβλέψεων.

4.3 Προσθήκη θορύβου

```
cwd = os.getcwd()
data_root = "./"
noiselev = [0.3, 0.1, 0.5]
fold = 5

for DATASET in datasets:

    filebase = 'data'
    if DATASET == 'BALANCE':
        path = 'bl'
    elif DATASET == 'WINE':
        path = 'wine'
    elif DATASET == 'WF':
        path = 'WF'
    elif DATASET == 'EEG':
        path = 'eeg'
    elif DATASET == 'RING':
        path = 'ring'
```

```
    elif DATASET == 'BANANA':
        path = 'bn'
    elif DATASET == 'ECL':
        path = 'ecl'
    elif DATASET == 'KDD':
        path = 'kdd'
    elif DATASET == 'LIR':
        path = 'lir'
    elif DATASET == 'LS':
        path = 'ls'
    elif DATASET == 'MGT':
        path = 'mgt'
    elif DATASET == 'PENDIGITS':
        path = 'pd'
    elif DATASET == 'PH':
        path = 'ph'
    elif DATASET == 'SH':
        path = 'sh'
    elif DATASET == 'TWONORM':
        path = 'tn'
    elif DATASET == 'TXR':
        path = 'txr'


    for index in range(fold):
        filename = '{}/{}'.format(data_root + path, filebase + '-tr' + str(index + 1))

        data = read_csv(filename, header=None, sep='\s+').values
        NUM_CLASSES = int(np.max(data, axis=0)[-1]) + 1
        for level in noiselev:
            if NUM_CLASSES < 3 and level == 0.5:
                continue
            for i in range(len(data)):
                if random() < level :
                    x = data[i][-1]
                    data[i][-1] = randint(0, NUM_CLASSES - 1)
                    while data[i][-1] == x:
                        data[i][-1] = randint(0, NUM_CLASSES - 1)
                with open('{}/{}'.format(data_root + path, filebase + '-tr' +
                str(index + 1) + '-' + str(level)),
                          'w') as f:
                    for x in data:
                        for i in range(len(x) - 1):
```

```
                f.write('{:.6f}\t'.format(float(x[i])))
            f.write('{}\n'.format(int(x[-1])))
```

Η προσθήκη θορύβου στα δεδομένα έγινε σε τρία διαφορετικά ποσοστιαία μεγέθη : 10% , 30% και 50% θόρυβο. Στα σύνολα δεδομένων πολλαπλών κλάσεων(multi-class) δημιουργήθηκαν θορυβώδη δεδομένα(noisy data) και των τρίων μεγέθων ενώ στα σύνολα δεδομένων δυαδικών κλάσεων δεν προστέθηκε 50% θόρυβος καθώς η τυχαία αντιστροφή των ετικετών κλάσεων θα προκαλούσε προφανώς πολύ μεγάλη πτώση στην αποτελεσματικότητα των αλγορίθμων. Τα σύνολα δεδομένων στη πραγματικότητα δεν είναι τέλεια και πολλές φορές περιέχουν θόρυβο. Έτσι, η αυτόβουλη προσθήκη θορύβου έχει σκοπό να μετατρέψει σύνολα δεδομένων ώστε να είναι πιο αντιπροσωπευτικά στον πραγματικό κόσμο. Επίσης, η προσθήκη θορύβου είναι ένας τρόπος να ερευνηθεί η ανθεκτικότητα των διαφόρων μοντέλων να επιτελούν το έργο τους σε δύσκολες συνθήκες.

Αρχικά τα δεδομένα είναι ήδη χωρισμένα σε 5 αρχεία(folds). Υπάρχει ένας for βρόγχος που διαλέγει το σετ δεδομένων και αναλόγως του ονόματος του θέτει στη μεταβλητή διαδρομής path την ανάλογη τιμή. Έπειτα, δημιουργεί στο filename τη διαδρομή που αναφέρεται στο 1ο αρχείο δεδομένων εκπαίδευσης του dataset και φορτώνει το αρχείο στη μεταβλητή data με την συνάρτηση read_csv() της βιβλιοθήκης pandas η οποία δημιουργεί σωστά τη δομή του αρχείου του οποίου τα χαρακτηριστικά(features) είναι χωρισμένα με κενό διάστημα(space). Στη συνέχεια, αποθηκεύει σε μεταβλητή το πλήθος των διαφορετικων κλάσεων το οποίο αμέσως μετά χρησιμοποιεί για να ελέγξει πότε πρέπει να προστεθεί ο θόρυβος. Τα επίπεδα θορύβου είναι αποθηκευμένα σε μία λίστα και για κάθε διαφορετικό επίπεδο εκτελεί την εξης διαδικασία : Αν το σετ δεδομένων είναι δυαδικής κλάσης και το επίπεδο προσθήκης θορύβου είναι 50%, δεν κάνει τίποτα. Διαφορετικά, δημιουργεί ένα τυχαίο αριθμό μεταξύ του 0 και του 1 και τον συγκρίνει με τη τιμή του επιπέδου θορύβου. Αν ο αριθμός είναι μικρότερος από το level, τότε δημιουργεί ένα τυχαίο αριθμό από το εύρος τιμών των κλάσεων, το συγκρίνει με την ετικέτα κλάσης που είχε, μέχρις ότου οι αριθμοί αυτοί είναι διαφορετικοί μεταξύ τους και τέλος αποθηκεύει τη νέα ετικέτα κλάσης στο στιγμιότυπο. Αν όμως η random() είναι μεγαλύτερη από το level, τότε δεν θα προστεθεί θόρυβος. Η διαδικασία επαναλαμβάνεται για όλα τα στιγμιότυπα εκπαίδευσης και δημιουργείται το νέο αρχείο με θόρυβο πριν προχωρήσει στο επόμενο αρχείο.

Κεφάλαιο 5ο: Πειραματική μελέτη

Το αντικείμενο της πειραματικής μελέτης που διεξήχθη ήταν η εφαρμογή ενός κατηγοριοποιητή βασιζόμενου σε στιγμιότυπα(instance-based classifier), του κ εγγύτερου γείτονα, πάνω σε 16 διαφορετικά σύνολα δεδομένων. Επίσης, αντικείμενο της μελέτης ήταν η χρήση τεχνικών μείωσης δεδομένων(DRT) στη φάση της προεπεξεργασίας για την αριθμητική μείωση των προτύπων, η εξέταση των ποσοστών μείωσης που επιτεύχθηκαν, η αξιολόγηση της ακρίβειας του κατηγοριοποιητή καθώς και η ανάλυση της συμπεριφοράς αυτών των παραμέτρων σε συνθήκες όπου έχει γίνει προσθήκη θορύβου σε ποσοστά 10%, 30% και 50% αντίστοιχα. Ακόμη, χρησιμοποιήθηκαν διάφορες τεχνικές δειγματοληψίας με ίδιο σκοπό την μείωση των δεδομένων και έγινε η σύκριση τους ως προς την αποτελεσματικότητα. Τέλος, τα πειραματικά αποτελέσματα επικυρώθηκαν στατιστικά με τη χρήση του Wilcoxon signed-rank test και του Friedman test ενώ κατά την εφαρμογή αυτών των test υπολογίστηκε ο μέσος όρος της δειγματοληψίας για λόγους ευκολίας καθότι όλες οι αυτές τεχνικές έχουνε παρόμοια επίδοση.

5.1 Περιβάλλον εκτέλεσης πειραμάτων

Η εκτέλεση των πειραμάτων πραγματοποιήθηκε σε ένα υπολογιστικό σύστημα με επεξεργαστή CPU Intel Xeon E5620 αρχιτεκτονικής x86_64. Η CPU έχει μέγεθος φυσικής διεύθυνσης 40 bit και μέγεθος εικονικής διεύθυνσης 48 bit. Διαθέτει 12 πυρήνες, οι οποίοι κατανέμονται σε 6 υποδοχές(sockets). Κάθε πυρήνας έχει 1 νήμα(thread) και κάθε υποδοχή έχει 2 πυρήνες. Επίσης, η CPU λειτουργεί με βασική ταχύτητα ρολογιού 2,40 GHz και έχει 4800,32 BogoMIPS. Διαθέτει κρυφή μνήμη δεδομένων 32K L1, κρυφή μνήμη εντολών 32K L1, κρυφή μνήμη L2 256K και κρυφή μνήμη L3 12288K.

5.2 Σύνολα δεδομένων

Η πειραματική μελέτη πραγματοποιήθηκε πάνω σε 16 διαφορετικά σύνολα δεδομένων(datasets) κατηγοριοποίησης πολλαπλής(10 multiclass) και δυαδικής κλάσης(6 binary) τα οποία υπάρχουν διαθέσιμα στο λογισμικό του KEEL αποθετήριου [32] και στο UCI Machine Learning Repository [33].

Πίνακας 5.1: Πληροφορίες για τα Σύνολα Δεδομένων

Σύνολα δεδομένων (Datasets)	**Στιγμιότυπα (instances)**	**Χαρακτηριστικά (features)**	**Κλάσεις (Classes)**
kdd	141482	41	23
sh	58000	9	7
lir	20000	16	26
pd	10992	16	10
ls	6436	36	6
wf	7164	21	3
txr	5500	40	11
sg	2310	19	7
ecl	336	7	8
wine	178	13	3
mgt	19020	10	2
egg	14982	14	2
ph	5404	5	2
bn	5300	2	2
tn	7400	20	2
ring	7400	20	2

Ecoli Data Set (ecl)

Το dataset περιλαμβάνει πληροφορίες μετρήσεων σχετικά με κάποια χαρακτηριστικά των κυττάρων(κυτταρόπλασμα, εσωτερική μεμβράνη, περιπλάσμα, εξωτερική μεμβράνη, λιποπρωτεΐνη εξωτερικής μεμβράνης, εσωτερική μεμβράνη λιποπρωτεΐνης εσωτερικής μεμβράνης, αλληλουχία σήματος που μπορεί να διασπαστεί) και σκοπός είναι η πρόβλεψη της θέσης εντοπισμού των κυττάρων αυτών. Αποτελείται από 7 χαρακτηριστικά και 8 κλάσεις

Letter Recognition Data Set (lir)

Πρόκειται για ένα σύνολο δεδομένων που αναπαριστά τα 26 κεφαλαία γράμματα του αγγλικού αλφαβήτου σε ασπρόμαυρα ορθογώνια εικονοστοιχεία(pixels). Έχει μέγεθος 20.000 παραδειγμάτων που δημιουργήθηκαν από την τυχαία τροποποίηση 20 διαφορετικών γραμματοσειρών φωτογραφιών του κάθε γράμματος και ύστερα μετατράπηκαν σε 16 χαρακτηριστικά τα οποία κλιμακώθηκαν το κάθε ένα στο εύρος ακεραίων αριθμών από το 0 έως το 15 με τελικό στόχο την αναγνώριση στων γραμμάτων.

Pen-Based Recognition of Handwritten Digits Data Set (pd)

Το dataset αυτό δημιουργήθηκε από τη συλλογή χειρόγραφων αριθμών του δεκαδικού συστήματος από 44 συγγραφείς με την αξιοποίηση πληροφοριών σχετικών μόνο με τις συντεταγμένες(x,y) οι οποίες αναπαραστάθηκαν ως διανύσματα χαρακτηριστικών σταθερού μήκους και έπειτα επαναδειγματοληπτήθηκαν σε 8 σημεία για κάθε ψηφίο.

Phoneme data set (ph)

Αποτελείται από στιγμιότυπα φωνημάτων εκ των οποίων τα 3818 είναι στιγμιότυπα ρινικών ήχων(κλάση 0) ενώ 1586 είναι στοματικοί ήχοι(κλάση 1) και στόχος είναι η ικανότητα διάκρισης των δύο κλάσεων. Έχει 5 χαρακτηριστικά τα οποία προσδιορίζονται ως εξής : sh όπως η λέξη she, dcl όπως η λέξη dark, iy όπως το φωνήεν στο she και ao, όπως το πρώτο φωνήεν στη λέξη water.

Texture data set (txr)

Τα στιγμιότυπα του dataset αναφέρονται σε υφές διαφόρων ειδών και σκοπός είναι η διάκριση των υφών με τη κατηγοριοποίηση. Αποτελείται από 11 κλάσεις, δηλαδή 11 επιφάνειες() και κάθε στιγμιότυπο προσδιορίζεται από 40 χαρακτηριστικά τα οποία δημιουργήθηκαν από την εκτίμηση τροποποιημένων ροπών τέταρτης τάξης σε τέσσερις προσανατολισμούς: 0, 45, 90 και 135 μοίρες.

Statlog (Shuttle) data set (sh)

Το σετ δεδομένων ’σαΐτα’ κατασκευάστηκε αρχικά ώστε να εξαχθούν κανόνες που θα μπορούν να προσδιορίσουν τις συνθήκες κάτω από τις οποίες είναι επιθυμητή η αυτόματη προσγείωση ενός διαστημικού σκάφος έναντι του χειροκίνητου τρόπου προσγείωσης. Έχει 7 αριθμητικά χαρακτηριστικά και οι διαφορετικές τιμές ετικέτας κλάσης που υπάρχουν είναι οι εξής :

1. Rad Flow
2. Fpv Close
3. Fpv Open
4. High
5. Bypass
6. Bpv Close
7. Bpv Open

Image Segmentation data set (sg)

Αυτή η βάση δεδομένων περιέχει περιπτώσεις που αντλούνται τυχαία από μια βάση δεδομένων 7 εικόνων εξωτερικού χώρου (κλάσεις). Οι εικόνες τμηματοποιήθηκαν με το χέρι για να δημιουργηθεί μια ταξινόμηση για κάθε εικονοστοιχείο. Κάθε περίπτωση κωδικοποιεί μια περιοχή 3x3 και στόχος είναι να προσδιοριστεί ο τύπος της επιφάνειας κάθε περιοχής.

MAGIC Gamma Telescope data set (mgt)

Αυτό το σύνολο δεδομένων περιέχει δεδομένα που παράγονται για την προσομοίωση της καταγραφής σωματιδίων γάμμα υψηλής ενέργειας σε ένα επίγειο ατμοσφαιρικό τηλεσκόπιο γάμμα Cherenkov με χρήση της τεχνικής απεικόνισης.

Το σύνολο δεδομένων δημιουργήθηκε από ένα πρόγραμμα Monte Carlo, το 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 στατιστική διάκριση των εικόνων που δημιουργούνται από πρωτογενή γάμμα (σήμα, ετικέτα κατηγορίας g) από τις εικόνες των αδρονικών βροχών που ξεκινούν από κοσμικές ακτίνες στην ανώτερη ατμόσφαιρα (υπόβαθρο, ετικέτα κατηγορίας h).

Ringnorm data set (ring)

Πρόκειται για ένα πρόβλημα ταξινόμησης 20 διαστάσεων, 2 κλάσεων. Κάθε κλάση αντλείται από μια πολυμεταβλητή κανονική κατανομή. Η κλάση 1 έχει μέση τιμή μηδέν και συνδιακύμανση 4 φορές την ταυτότητα. Η κλάση 2 έχει μέση τιμή (a,a,..a) και μοναδιαία συνδιακύμανση. a = 2/sqrt(20).

Wafer data set (wf)

Τα δεδομένα που δημιουργήθηκαν αφορούν πλακίδια για την κατασκευή ημιαγωγών μικροηλεκτρονικής. Περιέχει μετρήσεις που έχουν καταγραφεί από αισθητήρες κατά τη διάρκεια της επεξεργασίας μιας πλακέτας από ένα εργαλείο. Αποτελείται από 2 κλάσεις, τη κανονική και την μη κανονική. Επίσης, υπάρχει μεγάλη ανισορροπία κλάσεων μεταξύ κανονικών και μη κανονικών (10,7% του συρμού είναι μη κανονικά, 12,1% της δοκιμής). Είναι ένα πολύ μεγάλο dataset με 41 χαρακτηριστικά και 23 κλάσεις.

Twonorm data set (tn)

Πρόκειται για ένα πρόβλημα ταξινόμησης 20 διαστάσεων, 2 κλάσεων. Κάθε κλάση αντλείται από μια πολυμεταβλητή κανονική κατανομή. Η κλάση 1 έχει μέση τιμή (a,a,..a) ενώ η κλάση 2 έχει μέση τιμή (-a,-a,..-a).. a = 2/sqrt(20).

KDD Cup 1999 data set (kdd)

Πρόκειται για ένα υποσύνολο ενός dataset που χρησιμοποιήθηκε σε ένα διεθνή διαγωνισμό εργαλίων ανακάλυψης γνώσης και εξόρυξης δεδομένων. Ο στόχος του διαγωνισμού ήταν η δημιουργία ενός μοντέλου που θα μπορεί να ανιχνεύει εισβολές σε ένα δίκτυο και να διακρίνει τις συνδέσεις αυτές σε κα-

κές(εισβολές ή επιθέσεις) και σε καλές(κανονικές συνδέσεις) μέσω αξιοποίησης ονομαστικών(nominal) αλλά και συνεχών δεδομένων(continuous).

Eye State Data Set (eeg)

Τα δεδομένα αναφέρονται στις καταστάσεις των ματιών και η προέλευση τους έγινε από μια συνεχή μέτρηση EEG με τη χρήση της συσκευής Emotiv EEG Neuroheadset. Η διάρκεια μέτρησης ήταν 117 δευτερόλεπτα και η ανίχνευση της κατάστασης των ματιών πραγματοποιήθηκε μέσω μιας κάμερας κατά τη διάρκεια του πειράματος. Μετά την ανάλυση των καρέ του βίντεο(video frames) προστέθηκαν χειροκίνητα οι καταστάσεις του ματιού στο αρχείο. Έχει δύο κλάσεις, η κατάσταση στην οποία το μάτι είναι ανοιχτό παίρνει την τιμή 0, ενώ η κατάσταση όπου το μάτι είναι κλειστό έχει την τιμή 1. Επίσης, αποτελείται από 15 χαρακτηριστικά.

Banana Data Set (bn)

Είναι ένα τεχνητό σύνολο δεδομένων αποτελούμενο από δύο χαρακτηριστικά, το at1 αντιστοιχεί στον x άξονα ενώ το at2 χαρακτηριστικό αντιστοιχεί στον y άξονα. Αποτελεί 2 κλάσεις και τα στιγμιότυπα του σετ δεδομένων ανήκουν σε συστάδες σχήματος μπανάνας.

Landsat Data Set (ls)

Μια βάση δεδομένων που αποτελείται από τις πολυφασματικές τιμές των εικονοστοιχείων σε 3x3 γειτονιές μιας δορυφορικής εικόνας και την ταξινόμηση που σχετίζεται με το κεντρικό εικονοστοιχείο σε κάθε γειτονιά. Στόχος είναι η πρόβλεψη αυτής της ταξινόμησης, δεδομένων των πολυφασματικών τιμών. Στη βάση δεδομένων του δείγματος, η κλάση ενός εικονοστοιχείου κωδικοποιείται ως αριθμός. Αποτελείτε από 6435 στιγμιότυπα και 36 χαρακτηριστικά.

Wine Data set (wine)

Τα δεδομένα wine δημιουργήθηκαν από την χημική ανάλυση τριών ποικιλιών οινών που καλιεργήθηκαν σε μια περιοχή της Ιταλίας. Στην ανάλυση αυτή προσδιορίστηκαν ποσότητες 13 συστατικών(Αλκοόλ, Μηλικό οξύ, Τέφρα, Αλκαλικότητα της τέφρας, Μαγνήσιο, Ολικές φαινόλες, Φλαβονοειδή, Μη φλαβανοειδείς φαινόλες, Προανθοκυανίνες, Ένταση χρώματος, Απόχρωση, OD280/OD315 των αραιωμένων οίνων, Προλίνη) και οι ετικέτες κλάσεις του συνόλου δεδομένων ειναι 3.

5.3 Αποτελέσματα πειραμάτων

Πίνακας 5.2: ποσοστά μείωσης με τον CNN αλγόριθμο

	noise			
dataset	free	10%	30%	50%
pd	95.351	59.845	24.64	8.543
ls	79.918	52.455	23.947	12.055
txr	91.959	58.982	24.368	8.414
ecl	58.507	44.627	20.075	9.552
lir	83.116	56.396	23.097	6.483
sh	99.65	64.747	26.812	10.608
wf	59.34	44.855	26.21	20.24
wine	44.562	35.443	17.964	7.788
kdd	99.123	62.432	24.802	6.912
sg	88.896	59.545	25.206	9.794
tn	82.115	55.973	35.868	-
bn	77.363	55	35.217	-
ph	75.526	55.452	35.073	-
eeg	66.105	50.719	34.781	-
mgt	64.085	48.93	34.765	-
ring	73.26	49.983	35.209	-

Πίνακας 5.3: ποσοστά μείωσης με τον RSP3 αλγόριθμο

	noise			
dataset	free	10%	30%	50%
pd	89.642	57.914	21.624	6.84
ls	72.89	48.737	20.206	8.411
txr	82.323	54.441	20.705	6.145
ecl	52.537	39.851	15.597	6.194
lir	61.884	43.067	16.719	4.301
sh	99.412	61.012	23.265	8.3
wf	57.025	41.385	21.21	14.65
wine	38.331	30.003	14.254	5.578
sg	82.695	54.351	20.693	7.002
tn	84.307	55.652	28.449	-
bn	75.09	51.811	28.368	-
ph	69.313	49.253	27.763	-
eeg	53.758	41.13	26.754	-
mgt	58.749	43.916	27.258	-
ring	56.997	42.919	26.601	-

Πίνακας 5.4: ποσοστά μείωσης με τον ENN αλγόριθμο

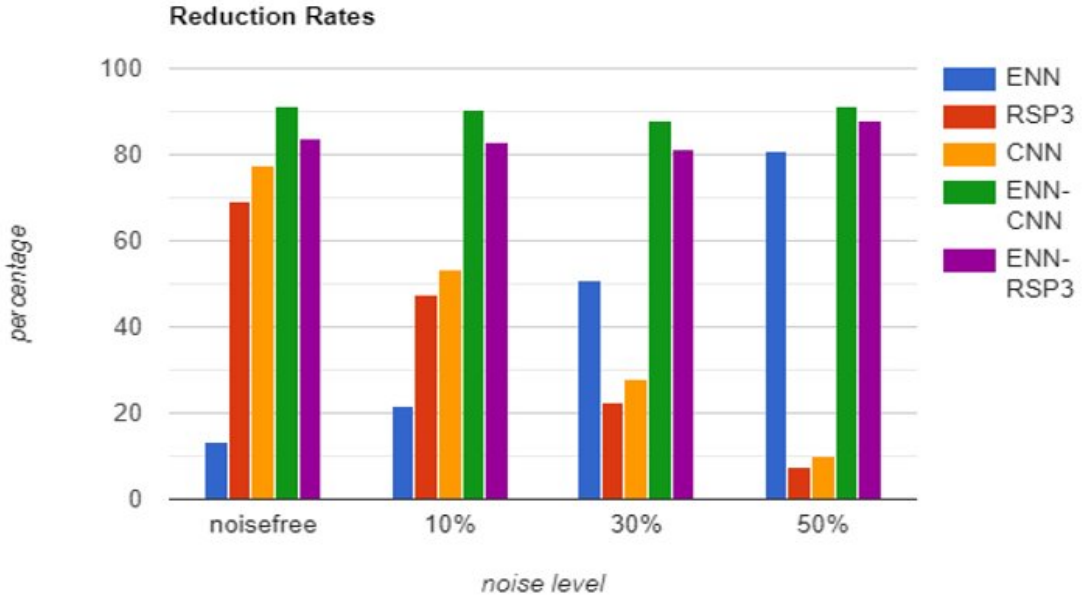
	noise			
dataset	free	10%	30%	50%
pd	0.685	12.371	49.947	82.406
ls	9.106	20.225	53.069	77.778
txr	1.323	12.877	51.314	82.905
ecl	22.313	29.403	61.418	82.239
lir	4.358	16.096	53.486	86.326
sh	0.099	11.802	49.837	79.52
wf	19.495	28.78	53.47	65.885
wine	47.165	46.9	67.966	85.721
kdd	0.291	11.998	50.9	85.559
sg	18.128	15.206	51.396	80.584
tn	3.588	15.986	43.878	-
bn	11.67	21.755	43.741	-
ph	11.719	21.476	44.502	-
eeg	14.601	24.845	45.169	-
mgt	17.029	26.149	45.43	-
ring	28.659	33.882	46.875	-

Πίνακας 5.5: ποσοστά μείωσης με συνδυασμό ENN-CNN αλγορίθμων

	noise			
dataset	free	10%	30%	50%
pd	96.496	96.703	93.878	92.156
ls	91.478	91.719	89.795	88.249
txr	93.371	93.803	93.248	92.816
ecl	88.059	86.641	90.522	91.417
lir	87.68	88.778	91.754	94.895
sh	99.776	99.569	95.492	91.893
wf	81.955	83.28	82.385	81.08
wine	87.835	83.578	88.167	92.363
kdd	99.626	99.53	97.628	95.856
sg	93.752	93.817	92.701	91.694
tn	89.227	87.886	78.55	-
bn	95.343	94.295	86.279	-
ph	90.409	89.849	85.005	-
eeg	83.249	83.699	81.353	-
mgt	90.509	88.962	81.782	-
ring	92.675	87.402	77.168	-

Πίνακας 5.6: ποσοστά μείωσης με τον ENN-RSP3 αλγόριθμο

	noise			
dataset	free	10%	30%	50%
pd	90.548	91.257	90.83	90.093
ls	85.661	86.547	86.889	84.70954
txr	84.337	85.469	88.67	90.807
ecl	80.298	79.701	87.462	89.402
lir	66.261	69.565	80.928	92.059
sh	99.583	99.353	94.7	90.350015
wf	76.035	77.51	75.47	73.315
wine	82.266	76.947	84.41	90.893
sg	87.677	87.395	88.9	89.669
tn	89.386	87.997	69.839	-
bn	93.871	92.696	82.344	-
ph	84.783	84.149	78.468	-
eeg	69.447	71.745	71.246	-
mgt	85.077	83.779	73.39	-
ring	80.003	78.385	66.685	-



Σχήμα 5.1: Ποσοστά μείωσης δεδομένων των αλγορίθμων κατά μέσο όρο

Πίνακας 5.7: Αποτελέσματα ακρίβειας για τον CNN και sampling

	noise															
	free				10%				30%				50%			
dataset	**cnn**	**random**	**stratified**	**systematic**	**cnn**	**random**	**stratified**	**systematic**	**cnn**	**random**	**stratified**	**systematic**	**cnn**	**random**	**stratified**	**systematic**
pd	98.435	95.560	95.447	95.207	82.167	89.217	89.204	89.073	56	63.570	63.674	63.365	30.743	33.183	33.178	33.000
ls	87.753	87.277	87.274	87.206	73.795	80.221	80.187	80.302	52.052	58.953	58.381	58.290	31.536	34.025	34.007	33.842
txr	97.127	93.719	93.402	93.490	81.997	87.816	88.129	87.623	54.81	61.909	61.516	61.160	30.042	32.206	32.115	32.180
ecl	72.836	75.522	73.731	71.701	65.97	70.925	68.239	69.254	42.687	47.224	46.030	46.925	26.866	28.896	28.299	26.806
lir	92.66	87.287	87.091	87.254	79.949	83.613	83.760	83.629	55.023	60.327	59.850	60.548	29.501	31.605	31.256	31.750
sh	99.919	97.985	98.172	98.191	87.465	89.993	89.972	89.804	60.028	63.497	64.199	63.360	33.615	34.821	35.620	34.823
wf	74	76.276	76.952	77.040	64.2	71.076	70.552	70.476	49.66	54.016	54.044	54.456	37.92	38.448	38.900	38.312
wine	42.835	44.219	44.349	43.578	39.302	40.267	40.773	40.286	30.156	30.867	30.446	31.153	19.6	19.801	19.838	19.801
kdd	99.664	98.141	98.151	98.051	85.451	89.687	89.681	89.634	58.289	63.009	62.892	63.085	30.81	32.519	32.421	32.428
sg	95.236	90.679	90.687	90.090	83.586	85.300	86.253	85.881	57.991	61.741	61.395	61.266	29.536	31.259	30.948	31.043
tn	89.783	93.840	93.902	93.848	75.470	84.736	84.698	84.636	59.022	63.200	63.038	63.501	-	-	-	-
bn	86.375	86.556	86.522	86.620	76.675	79.064	79.351	79.317	59.010	61.638	61.755	62.143	-	-	-	-
ph	87.694	83.653	83.354	83.150	78.035	79.094	79.090	78.420	60.77	62.457	62.206	62.272	-	-	-	-
eeg	47.236	45.510	45.358	45.502	47.437	46.467	46.578	46.409	50.327	49.017	49.143	49.449	-	-	-	-
mgt	76.781	79.655	79.485	79.266	69.178	74.116	74.017	74.299	57.285	59.871	59.709	60.092	-	-	-	-
ring	82.649	69.619	69.822	69.749	69.068	67.562	67.932	67.378	56.054	56.586	56.543	56.586	-	-	-	-

Πίνακας 5.8: Αποτελέσματα ακρίβειας για τον RSP3 και sampling

	noise															
	free				10%				30%				50%			
dataset	**rsp3**	**random**	**stratified**	**systematic**	**rsp3**	**random**	**stratified**	**systematic**	**rsp3**	**random**	**stratified**	**systematic**	**rsp3**	**random**	**stratified**	**systematic**
pd	99.163	97.629	97.611	97.545	82.049	88.887	89.209	89.211	57.729	63.594	63.279	63.978	31.68	33.105	33.169	33.158
ls	89.758	88.101	88.004	88.051	81.909	80.361	80.379	80.289	61.424	58.337	58.561	58.505	35.468	34.078	34.001	34.072
txr	98.618	96.032	96.257	96.126	84.161	87.696	87.954	88.176	58.483	61.844	61.709	61.724	31.187	32.086	32.089	31.958
ecl	74.925	74.746	75.881	75.284	69.851	67.701	69.672	70.269	44.179	46.328	45.552	44.716	28.657	28.776	28.896	28.955
lir	95.56	92.301	92.268	92.321	83.704	84.835	84.673	84.521	58.393	60.401	60.112	60.505	30.876	31.725	31.352	31.635
sh	99.483	98.438	98.657	98.552	85.886	90.126	90.018	89.854	59.524	63.570	64.298	63.311	33.61	34.879	35.646	34.859
wf	77.54	76.708	76.696	76.844	70.9	70.232	70.564	70.296	55.58	54.008	54.516	53.764	39.16	38.504	38.520	38.768
wine	44.264	44.901	44.529	45.501	40.854	40.895	40.495	40.768	30.585	31.414	30.642	31.222	19.948	20.054	19.662	19.980
sg	95.279	91.918	92.117	92.308	82.980	85.994	85.803	85.412	59.204	61.092	61.793	61.188	29.796	31.077	31.017	31.191
tn	93.107	94.097	93.832	93.945	81.673	84.793	85.011	84.965	62.143	63.000	63.306	63.217	-	-	-	-
bn	84.601	86.680	86.828	86.292	74.335	79.381	79.313	79.774	58.822	61.422	61.973	61.483	-	-	-	-
ph	86.621	84.664	84.282	85.026	76.758	79.075	79.364	79.061	60.807	62.046	62.261	62.094	-	-	-	-
eeg	47.31	45.409	45.235	45.594	48.458	47.164	46.256	46.164	49.866	48.874	49.235	49.083	-	-	-	-
mgt	77.412	79.518	79.466	79.702	69.793	74.089	74.227	74.060	57.69	59.748	59.810	59.798	-	-	-	-
ring	81.432	71.651	71.630	71.492	73.027	68.211	68.076	68.327	58.514	56.216	56.476	56.489	-	-	-	-

Πίνακας 5.9: Αποτελέσματα ακρίβειας για τον ENN και sampling

	noise															
	free				10%				30%				50%			
dataset	**enn**	**random**	**stratified**	**systematic**	**enn**	**random**	**stratified**	**systematic**	**enn**	**random**	**stratified**	**systematic**	**enn**	**random**	**stratified**	**systematic**
pd	99.235	99.327	99.321	99.321	99.09	89.457	89.437	89.344	96.406	63.226	63.696	63.041	68.556	33.275	32.827	33.116
ls	89.975	89.810	89.820	89.938	89.462	80.560	80.665	80.749	84.784	58.104	57.660	58.238	54.414	33.080	33.528	33.298
txr	98.508	98.894	98.887	98.909	98.272	88.630	88.383	88.823	94.981	61.236	61.553	61.444	71.759	32.268	32.751	32.912
ecl	80	77.672	76.896	77.851	78.805	70.149	68.776	69.672	71.94	45.552	48.000	44.299	50.149	25.672	28.478	26.687
lir	94.809	95.637	95.650	95.653	94.194	85.811	85.721	85.975	89.814	58.902	58.946	58.817	68.833	28.435	28.424	27.895
sh	99.889	99.934	99.934	99.934	99.815	90.117	90.417	90.106	95.307	63.557	63.842	63.442	63.044	34.862	34.645	34.831
wf	78.74	76.996	76.972	77.064	77.76	70.212	70.66	70.948	68.04	54.068	54.272	54.492	43.54	38.324	37.984	38.796
wine	48.96	44.080	44.010	44.488	46.06	40.613	40.474	40.662	42.814	29.739	30.389	30.091	30.074	18.404	18.473	18.380
kdd	99.674	99.714	99.713	99.713	99.654	89.755	89.743	89.757	98.377	63.092	62.890	63.045	83.482	32.446	32.307	32.487
sg	81.159	96.154	96.310	96.483	94.715	87.293	87.189	87.189	90.125	60.355	60.416	60.372	61.063	32.558	32.030	31.232
tn	95.445	94.689	94.699	94.713	94.647	85.157	85.184	85.103	77.659	63.209	63.181	63.457	-	-	-	-
bn	89.526	86.937	87.069	86.971	87.922	79.521	79.083	79.687	72.032	61.940	61.524	61.774	-	-	-	-
ph	87.657	89.267	88.983	88.912	86.362	80.800	80.911	81.033	70.411	61.532	61.658	62.199	-	-	-	-
eeg	44.786	45.630	45.561	45.566	44.886	46.557	46.461	46.573	48.017	49.242	49.077	49.089	-	-	-	-
mgt	83.074	80.320	80.321	80.311	81.781	74.364	74.304	74.256	68.353	59.563	59.598	59.720	-	-	-	-
ring	62.135	73.486	73.454	73.097	61.716	68.765	68.635	68.743	58.189	56.408	55.930	56.357	-	-	-	-

Πίνακας 5.10: Αποτελέσματα ακρίβειας για τον ENN-RSP3 και sampling

	noise															
	free				10%				30%				50%			
dataset	enn-rsp3	random	stratified	systematic	enn-rsp3	random	stratified	systematic	enn-rsp3	random	stratified	systematic	enn-rsp3	random	stratified	systematic
pd	99.127	97.278	97.445	97.267	98.808	87.879	87.919	87.204	92.603	61.838	62.366	62.082	63.57	32.650	32.665	32.885
ls	89.54	86.599	86.624	87.010	89.322	78.001	77.942	79.058	84.209	56.736	56.711	56.833	55.316	32.512	32.608	33.336
txr	98.236	95.548	95.734	95.756	98	85.964	86.041	86.223	93.381	60.738	60.502	59.953	68.395	32.628	32.948	31.915
ecl	77.015	70.149	71.881	70.567	77.313	63.522	65.134	62.388	67.463	43.821	44.358	44.597	47.761	28.537	31.582	26.866
lir	94.6	91.769	91.707	91.584	93.895	81.939	82.284	81.995	89.284	55.849	55.973	55.946	67.288	27.029	26.891	26.065
sh	99.128	98.366	98.263	98.365	99.436	88.647	89.146	88.965	92.65	62.908	63.242	62.541	61.035	34.539	34.515	34.851
wf	79.86	76.600	76.608	76.604	79.4	70.800	70.352	69.428	71.66	54.132	54.276	53.712	45.16	38.036	38.088	38.644
wine	47.001	42.018	41.711	41.568	44.386	39.841	38.997	39.547	41.874	29.478	28.519	29.041	29.482	17.754	19.037	17.616
kdd	99.623	98.112	98.121	98.155	99.3	89.065	87.969	87.990	94.973	62.561	62.524	62.442	76.805	32.331	32.275	32.069
sg	80.814	90.264	91.242	90.784	94.195	82.096	81.793	82.693	87.786	57.948	57.973	58.005	57.729	32.732	31.165	32.368
tn	93.729	93.461	93.575	93.597	92.905	84.503	84.733	84.876	76.456	63.427	63.144	62.325	-	-	-	-
bn	88.318	85.827	86.277	85.892	86.394	78.917	78.732	78.826	67.654	61.245	61.479	61.668	-	-	-	-
ph	85.77	81.995	81.740	82.236	84.031	75.097	76.088	75.785	67.579	61.114	60.784	60.285	-	-	-	-
eeg	46.175	45.499	45.642	44.778	46.121	46.371	46.613	46.770	49.159	49.143	48.858	49.355	-	-	-	-
mgt	82.139	78.362	78.448	78.628	79.883	72.836	73.010	73.242	64.762	59.171	59.542	58.973	-	-	-	-
ring	72.311	68.503	68.781	68.889	70.554	64.968	65.270	65.214	62.311	56.408	55.900	56.027	-	-	-	-

Πίνακας 5.11: Αποτελέσματα ακρίβειας για τον ENN-CNN και sampling

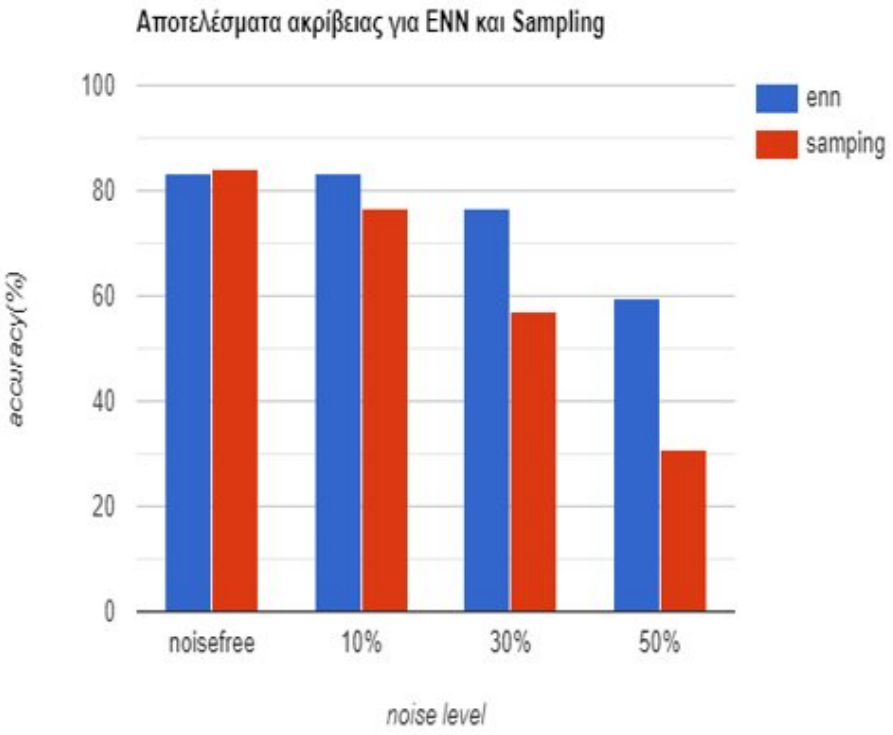
	noise															
	free				10%				30%				50%			
dataset	**enn-cnn**	**random**	**stratified**	**systematic**	**enn-cnn**	**random**	**stratified**	**systematic**	**enn-cnn**	**random**	**stratified**	**systematic**	**enn-cnn**	**random**	**stratified**	**systematic**
pd	98.644	94.572	94.610	94.310	98.399	85.701	85.637	85.022	92.594	62.024	62.697	61.245	59.849	33.366	33.267	33.946
ls	88.421	85.508	85.819	85.284	87.955	76.705	76.571	76.105	80.292	56.605	56.205	56.077	49.006	32.876	33.018	33.084
txr	96.581	92.533	92.831	92.591	96.563	82.655	82.491	83.062	91.671	58.800	59.102	58.985	63.266	32.271	32.937	31.057
ecl	77.91	68.776	69.552	66.149	75.522	63.343	63.881	61.672	67.463	44.537	41.313	42.627	48.358	27.284	28.955	29.731
lir	91.71	84.310	84.502	84.705	90.925	75.059	75.733	75.361	86.064	50.443	51.277	50.485	62.768	24.666	24.700	25.496
sh	99.893	97.564	97.868	97.464	99.8	87.654	88.651	87.981	94.189	63.019	63.267	63.547	60.723	34.733	34.659	35.257
wf	76.28	76.304	76.160	76.460	75.22	70.196	70.380	69.804	64.2	53.284	52.336	54.316	42.36	38.568	38.408	38.120
wine	45.693	41.875	41.850	41.139	44.529	38.886	38.731	38.379	40.935	29.466	28.616	29.425	28.665	18.301	17.975	18.126
kdd	99.659	97.465	97.808	97.570	99.609	87.709	88.389	88.180	97.348	62.611	62.426	62.132	77.897	32.327	32.207	32.205
sg	80.251	87.232	87.898	87.933	94.282	77.635	78.978	78.587	87.094	56.907	59.081	56.353	57.902	29.571	31.831	32.411
tn	91.688	93.694	93.596	93.629	89.445	84.179	84.319	84.679	69.171	63.454	62.665	62.298	-	-	-	-
bn	88.885	85.284	85.850	85.990	86.998	78.943	79.517	79.362	69.428	61.657	60.921	61.310	-	-	-	-
ph	86.288	80.255	80.296	80.396	84.567	75.059	74.016	74.149	68.171	60.995	61.399	60.355	-	-	-	-
eeg	45.641	45.531	45.546	45.368	45.901	47.155	46.000	45.356	47.69	49.920	49.403	48.677	-	-	-	-
mgt	81.986	77.609	78.102	78.061	80.073	72.712	72.765	72.588	64.483	59.411	59.201	59.181	-	-	-	-
ring	78.595	64.981	65.303	65.173	70.622	63.111	63.338	63.781	57.757	56.127	55.759	55.778	-	-	-	-

Πίνακας 5.12: Αποτελέσματα ακρίβειας για τον ENN-CNN και sampling στα καθαρά δεδομένα

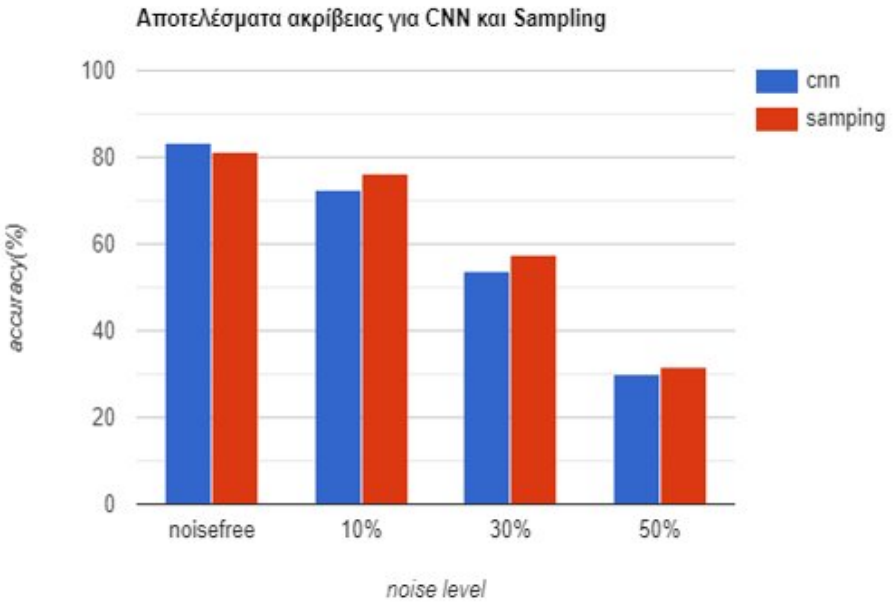
	noise															
	free				10%				30%				50%			
dataset	enn-cnn	random	stratified	systematic	enn-cnn	random	stratified	systematic	enn-cnn	random	stratified	systematic	enn-cnn	random	stratified	systematic
pd	98.644	94.659	94.818	94.656	98.399	93.810	93.853	94.150	92.594	92.214	92.265	92.199	59.849	66.014	66.258	66.292
ls	88.421	86.484	86.534	86.375	87.955	85.633	85.825	85.717	80.292	81.763	82.275	81.937	49.006	56.273	57.626	56.932
txr	96.581	92.522	92.697	92.973	96.563	91.711	91.919	91.755	91.671	87.060	87.692	85.863	63.266	61.353	64.423	63.539
ecl	77.91	69.493	69.254	69.433	75.522	70.328	71.940	66.806	67.463	59.701	65.134	57.194	48.358	41.672	44.716	42.985
lir	91.71	84.073	84.142	84.366	90.925	82.085	82.170	81.966	86.064	70.032	69.856	69.979	62.768	36.936	38.211	37.428
sh	99.893	97.589	97.883	97.154	99.8	98.331	98.294	98.271	94.189	96.135	96.522	96.420	60.723	67.029	68.036	67.224
wf	76.28	79.420	78.644	79.500	75.22	77.896	78.048	78.156	64.2	69.680	69.936	70.020	42.36	46.156	44.628	44.340
wine	45.693	48.111	46.935	47.429	44.529	44.937	45.063	44.586	40.935	41.205	42.021	41.944	28.665	28.571	30.470	29.967
kdd	99.659	97.530	97.727	97.437	99.609	97.515	97.734	97.549	97.348	97.669	97.700	97.630	77.897	84.790	84.649	85.327
sg	80.251	73.607	73.832	73.443	94.282	86.374	86.505	86.765	87.094	82.840	83.421	81.273	57.902	54.880	55.127	51.267
tn	91.688	94.402	94.597	94.526	89.445	94.148	93.718	93.967	69.171	77.302	77.632	77.921	-	-	-	-
bn	88.885	88.043	88.028	87.801	86.998	87.413	87.616	87.424	69.428	74.222	74.011	73.874	-	-	-	-
ph	86.288	81.751	81.921	81.258	84.567	81.037	81.092	80.962	68.171	70.192	70.211	69.414	-	-	-	-
eeg	45.641	44.454	44.323	44.140	45.901	44.252	44.558	44.395	47.69	47.760	47.845	47.052	-	-	-	-
mgt	81.986	81.034	81.076	81.027	80.073	80.176	80.332	80.175	64.483	68.956	68.717	68.906	-	-	-	-
ring	78.595	56.176	56.370	56.284	70.622	57.159	57.338	57.573	57.757	56.668	56.257	56.303	-	-	-	-

Πίνακας 5.13: Αποτελέσματα ακρίβειας για τον ENN-RSP3 και sampling στα καθαρά δεδομένα

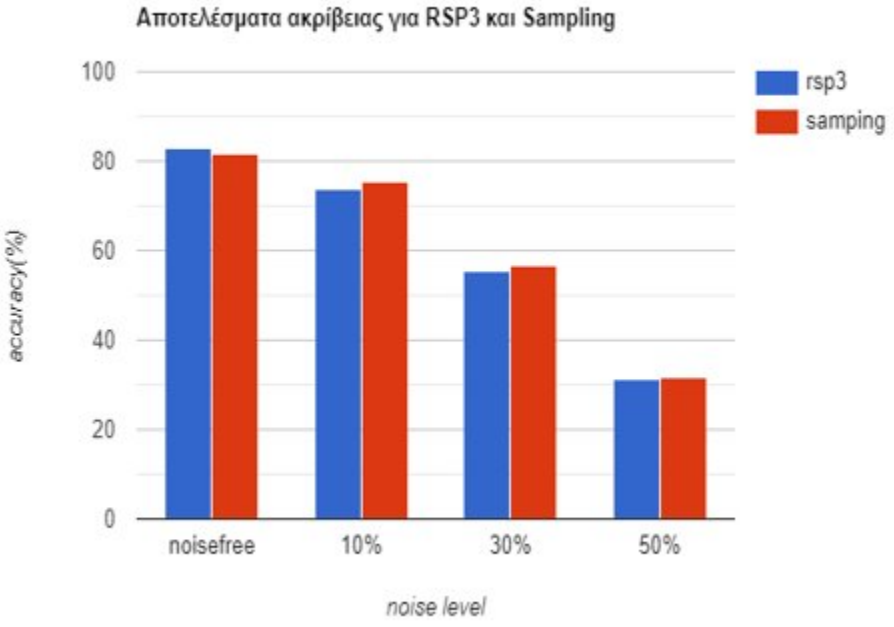
	noise															
	free				10%				30%				50%			
dataset	enn-rsp3	random	stratified	systematic	enn-rsp3	random	stratified	systematic	enn-rsp3	random	stratified	systematic	enn-rsp3	random	stratified	systematic
pd	99.127	97.411	97.511	97.345	98.808	96.945	97.167	96.872	92.603	93.415	93.742	93.407	63.57	68.379	68.168	66.365
ls	89.54	87.569	87.482	87.442	89.322	86.888	86.410	86.774	84.209	82.742	82.465	82.017	55.316	54.430	54.931	55.207
txr	98.236	95.850	95.661	95.817	98	95.126	95.017	94.854	93.381	89.918	89.838	89.795	68.395	64.627	67.780	64.379
ecl	77.015	71.761	73.075	73.433	76.716	71.403	74.507	70.866	67.463	61.194	63.881	63.761	47.761	41.970	51.522	44.478
lir	94.6	91.139	91.178	91.171	93.895	89.766	89.688	89.819	89.284	80.103	80.285	80.086	67.288	43.521	44.559	43.800
sh	99.128	98.348	98.231	98.259	99.436	98.462	98.509	98.575	92.65	96.471	96.442	96.490	61.035	67.502	67.068	67.518
wf	79.86	79.524	79.360	79.556	79.4	78.696	78.456	78.660	71.66	70.900	70.928	70.884	45.16	46.244	44.684	44.580
wine	47.001	47.588	47.915	47.711	44.386	44.619	45.117	44.978	41.874	42.193	42.328	41.237	29.482	31.735	29.861	29.711
kdd	99.623	98.196	98.254	98.169	99.3	98.099	98.173	98.149	94.973	97.699	97.691	97.646	76.805	85.535	84.871	85.095
sg	80.814	76.691	76.691	77.305	94.195	90.748	90.601	90.835	87.786	84.503	84.494	84.252	57.729	57.236	55.550	49.937
tn	93.729	94.564	94.607	94.264	92.905	93.745	93.959	94.051	76.456	77.716	77.578	77.424	-	-	-	-
bn	88.318	88.507	88.372	88.692	86.394	87.975	87.431	88.036	67.654	72.685	73.765	73.104	-	-	-	-
ph	85.77	82.757	82.591	82.746	84.031	81.984	82.002	82.062	67.579	70.273	69.833	70.573	-	-	-	-
eeg	46.175	44.471	44.398	44.645	46.121	44.252	44.660	44.314	49.159	47.734	48.096	47.296	-	-	-	-
mgt	82.139	81.520	81.433	81.575	79.883	80.531	80.567	80.649	64.762	68.878	68.544	68.875	-	-	-	-
ring	72.311	58.208	58.386	58.327	70.554	58.405	58.251	58.351	62.311	56.435	56.932	56.886	-	-	-	-



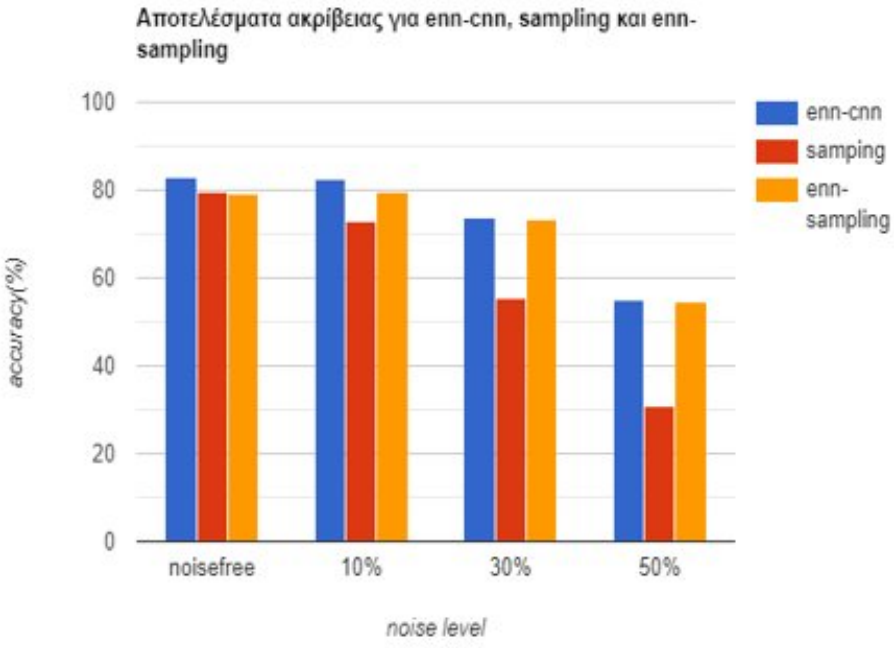
Σχήμα 5.2: Αποτελέσματα ακρίβειας του ENN και sampling κατά μέσο όρο



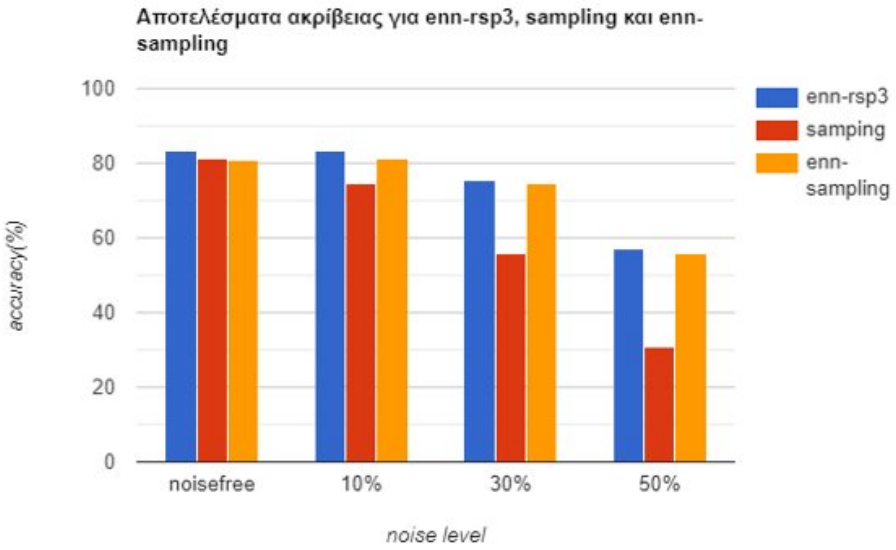
Σχήμα 5.3: Αποτελέσματα ακρίβειας του CNN και sampling κατά μέσο όρο



Σχήμα 5.4: Αποτελέσματα ακρίβειας του RSP3 και sampling κατά μέσο όρο



Σχήμα 5.5: Αποτελέσματα ακρίβειας του ENN-CNN και sampling κατά μέσο όρο



Σχήμα 5.6: Αποτελέσματα ακρίβειας του ENN-RSP3 και sampling κατά μέσο όρο

Πίνακας 5.14: Αποτελέσματα του Wilcoxon signed-rank test για noise free

Methods	**Accuracy**		**Reduction Rates**	
	w/l/t	Wilc	w/l/t	Wilc
ENN vs CNN	11/5	0.215	1/15	**0.001**
ENN vs RSP3	10/5	0.394	1/14	**0.001**
ENN vs SAMPLING	7/9	0.918	-	-
CNN vs RSP3	4/11	**0.053**	14/1	**0.001**
CNN vs Sampling	10/6	0.109	-	-
RSP3 VS SAMPLING	11/4	**0.023**	-	-
ENN-CNN vs ENN-RSP3	6/10	0.234	14/1	**0.001**
ENN-CNN vs SAMPLING	13/3	**0.010**	-	-
ENN-CNN vs ENN-SAMPLING	13/3	**0.015**	-	-
ENN-RSP3 vs SAMPLING	15/1	**0.007**	-	-
ENN-RSP3 vs ENN-SAMPLING	14/2	**0.001**	-	-

Πίνακας 5.15: Αποτελέσματα του Wilcoxon signed-rank test σε 10% θόρυβο

Methods	**Accuracy**		**Reduction Rates**	
	w/l/t	Wilc	w/l/t	Wilc
ENN vs CNN	14/2	**0.001**	1/15	**0.001**
ENN vs RSP3	13/2	**0.004**	1/14	**0.002**
ENN vs SAMPLING	14/2	**0.001**	-	-
CNN vs RSP3	5/10	**0.041**	15/0	**0.001**
CNN vs SAMPLING	2/14	**0.001**	-	-
RSP3 VS SAMPLING	7/8	0.691	-	-
ENN-CNN vs ENN-RSP3	8/8	0.179	15/1	**0.001**
ENN-CNN vs SAMPLING	15/1	**0.001**	-	-
ENN-CNN vs ENN-SAMPLING	11/5	**0.026**	-	-
ENN-RSP3 vs SAMPLING	15/1	**0.001**	-	-
ENN-RSP3 vs ENN-SAMPLING	12/4	**0.006**	-	-

Πίνακας 5.16: Αποτελέσματα του Wilcoxon signed-rank test σε 30% θόρυβο

Methods	**Accuracy**		**Reduction Rates**	
	w/l/t	Wilc	w/l/t	Wilc
ENN vs CNN	15/1	**0.001**	16/0	**0.000**
ENN vs RSP3	13/2	**0.001**	15/0	**0.001**
ENN vs SAMPLING	15/1	**0.001**	-	-
CNN vs RSP3	3/12	**0.008**	15/0	**0.001**
CNN vs SAMPLING	1/15	**0.001**	-	-
RSP3 VS SAMPLING	5/10	0.112	-	-
ENN-CNN vs ENN-RSP3	4/11	0.078	16/0	**0.000**
ENN-CNN vs SAMPLING	15/1	**0.001**	-	-
ENN-CNN vs ENN-SAMPLING	7/9	0.717	-	-
ENN-RSP3 vs SAMPLING	16/0	**0.000**	-	-
ENN-RSP3 vs ENN-SAMPLING	8/8	0.679	-	-

Πίνακας 5.17: Αποτελέσματα του Wilcoxon signed-rank test σε 50% θόρυβο

Methods	**Accuracy**		**Reduction Rates**	
	w/l/t	Wilc	w/l/t	Wilc
ENN vs CNN	10/0	**0.005**	6/4	0.139
ENN vs RSP3	9/0	**0.008**	7/2	0.139
ENN vs SAMPLING	10/0	**0.005**	-	-
CNN vs RSP3	1/8	**0.011**	7/2	0.441
CNN vs SAMPLING	0/10	**0.005**	-	-
RSP3 VS SAMPLING	3/6	0.441	-	-
ENN-CNN vs ENN-RSP3	3/7	0.059	10/0	**0.005**
ENN-CNN vs SAMPLING	10/0	**0.005**	-	-
ENN-CNN vs ENN-SAMPLING	3/7	0.059	-	-
ENN-RSP3 vs SAMPLING	10/0	**0.005**	-	-
ENN-RSP3 vs ENN-SAMPLING	5/5	0.959	-	-

Πίνακας 5.18: Αποτελέσματα του FRIEDMAN test για noise free δεδομενα

Algorithms	Mean Rank	
	Accuracy	Reduction Rate
ENN	4.33	1.14
SAMPLING	10.27	
CNN	6.87	3.21
SAMPLING	7.87	
RSP3	8.33	2.00
SAMPLING	6.60	
ENN-CNN	7.87	4.86
SAMPLING	2.53	
ENN-SAMPLING	4.47	
ENN-RSP3	9.13	3.79
SAMPLING	3.53	
ENN-SAMPLING	6.20	

Πίνακας 5.19: Αποτελέσματα του FRIEDMAN test για 10% θόρυβο

Algorithms	Mean Rank	
	Accuracy	Reduction Rate
ENN	10.6	1.14
SAMPLING	6.87	
CNN	3.07	2.93
SAMPLING	5.20	
RSP3	5	1.93
SAMPLING	5.80	
ENN-CNN	9.40	4.86
SAMPLING	2.53	
ENN-SAMPLING	7.40	
ENN-RSP3	9.67	4.14
SAMPLING	3.80	
ENN-SAMPLING	8.67	

Πίνακας 5.20: Αποτελέσματα του FRIEDMAN test για 30% θόρυβο

Algorithms	Mean Rank	
	Accuracy	Reduction Rate
ENN	10.73	3
SAMPLING	5.20	
CNN	2.40	2
SAMPLING	6.53	
RSP3	4.33	1
SAMPLING	5.80	
ENN-CNN	8.50	4.92
SAMPLING	3.00	
ENN-SAMPLING	8.73	
ENN-RSP3	9.90	4.08
SAMPLING	3.20	
ENN-SAMPLING	9.67	

Πίνακας 5.21: Αποτελέσματα του FRIEDMAN test για 50% θόρυβο

Algorithms	Mean Rank	
	Accuracy	Reduction Rate
ENN	11.00	3
SAMPLING	4.22	
CNN	1.89	2
SAMPLING	5.00	
RSP3	4.22	1
SAMPLING	5.11	
ENN-CNN	9.00	5
SAMPLING	3.78	
ENN-SAMPLING	9.56	
ENN-RSP3	10.11	4
SAMPLING	3.78	
ENN-SAMPLING	10.33	

5.3.1 Reduction Rates

Στον πίνακα 5.2, παρατηρούμε πώς επηρεάζεται η ικανότητα του CNN αλγορίθμου να παράγει μειωμένα σύνολα δεδομένων σε συνάρτηση με τα διαφορετικά επίπεδα θορύβου που μπορεί να υπάρχουν σε αυτά. Βλέπουμε ότι ο CNN είναι ευαίσθητος στο θόρυβο καθώς μειώνεται σημαντικά ο λόγος μείωσης και ειδικά, προσέχουμε ότι όταν υπάρχει 50% θόρυβος, το ποσοστό μείωσης πέφτει κάτω από 10%. Ωστόσο, με την συνδυαστική εκτέλεση πρώτα ενός αλγορίθμου επεξεργασίας-αφαίρεσης θορύβου(ENN) και ύστερα του αλγορίθμου συμπύκνωσης CNN, τα αποτελέσματα είναι πάρα πολύ διαφορετικά καθώς όπως βλέπουμε στον πίνακα 5.5, το ποσοστό μείωσης δεδομένων είναι γύρω στο 90% ακόμα και με ύπαρξη 50% θορύβου. Γενικά βλέπουμε και από τους αντίστοιχους πίνακες των αποτελεσμάτων του friedman test, για παράδειγμα σε δεδομένα χωρίς θόρυβο ότι ο συνδυασμός αλγορίθμων enn-cnn έχει τη καλύτερη απόδοση με μέση κατάταξη 4.86 ενώ ο αλγόριθμος ENN έχει τη χειρότερη μέση βαθμολογία. Το αποτέλεσμα αυτό είναι αναμενόμενο διότι ο αλγόριθμος ENN λειτουργεί με σκοπό την αφαίρεση θορύβου. Επομένως, όταν δεν υπάρχει θόρυβος δεν θα καταφέρει μεγάλους λόγους μείωσης δεδομένων.

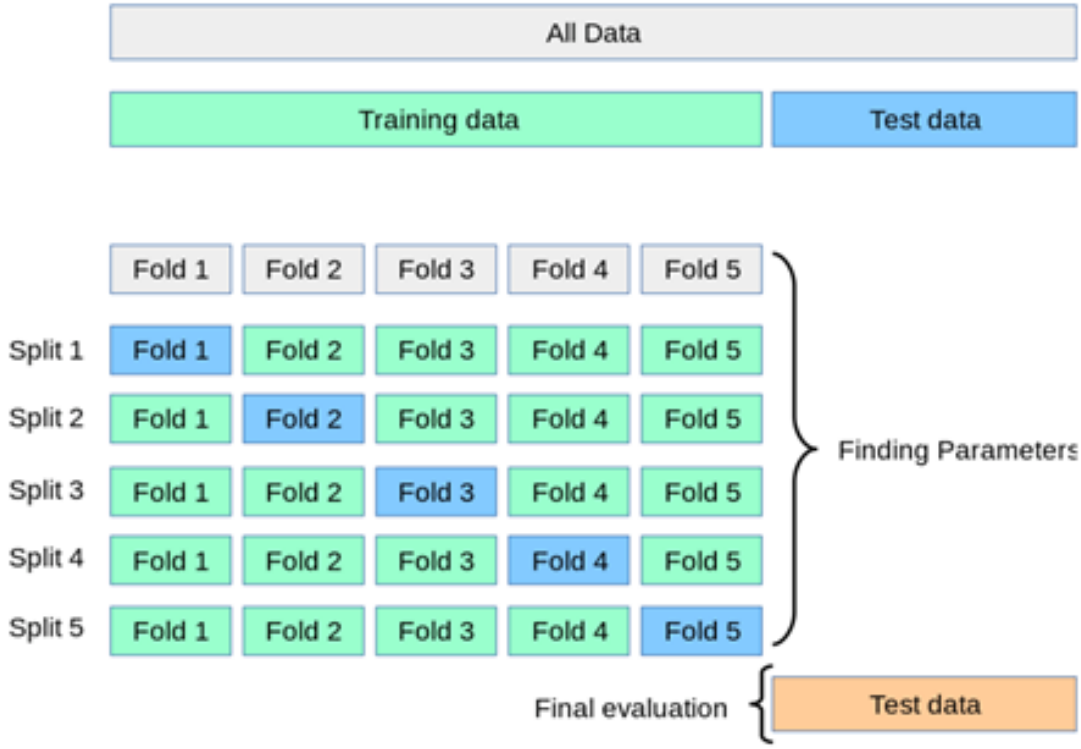
Στον πίνακα 5.4 βλέπουμε την επίδοση του αλγορίθμου ENN να αφαιρεί τα θορυβώδη δεδομένα. Παρατηρούμε ότι αυξάνει το ποσοστό μείωσης δεδομένων με την αύξηση ποσοστού του θορύβου. Η αφαίρεση του θορύβου καταλήγει να είναι μεγαλύτερη από 80% για επίπεδο θορύβου 50%

Όσον αφορά τον αλγόριθμο παραγωγής προτύπων RSP3 βλέπουμε την επίδοση του στον πίνακα 5.3. Όταν δεν υπάρχει θόρυβος στα δεδομένα βλέπουμε ότι κάνει μείωση δεδομένων μεγαλύτερη από 50%. Ωστόσο, είναι εύκολα αντιληπτή η αδυναμία του αλγορίθμου να δημιουργεί μικρότερα σύνολα συμπύκνωσης όταν υπάρχει θόρυβος της τάξης του 30% και 50% όπου τα ποσοστά μείωσης τις περισσότερες φορές μικρότερα από 25%. Ωστόσο, τα πράγματα είναι εντελώς διαφορετικά όταν προηγηθεί στην εκτέλεση ένας αλγόριθμος αφαίρεσης θορύβου όπως ο ENN. Στον πίνακα 5.3, βλέπουμε την χρησιμότητα του ENN καθ'ότι αντιμετωπίζει την αδυναμία του RSP3 με μεγάλη επιτυχία.

Εξετάζοντας συνολικά τα αποτελέσματα που παράγουν οι αλγόριθμοι, παρατηρούμε ότι ο συνδυασμός αλγορίθμων ENN-CNN επιτυγχάνει τα μεγαλύτερα ποσοστά μείωσης δεδομένων ενώ, μεταξύ των ENN,CNN και RSP3, τα καταφέρνει καλύτερα ο αλγόριθμος αφαίρεσης θορύβου ENN.

5.3.2 Accuracy

Για την αξιολόγηση της ακρίβειας εφαρμόστηκε η τεχνική 5-fold cross-validation, δηλαδή έγινε ο διαχωρισμός των δεδομένων σε 5 ισομεγέθη υποδείγματα και σε κάθε ένα υποδείγμα χρησιμοποιήθηκε ένα υποσύνολο του για την εκπαίδευση του μοντέλου ενώ το υπόλοιπο του χρησιμοποιήθηκε για την επικύρωση του. Επομένως, η διαδικασία εκπαίδευσης του μοντέλου εκτελέστηκε 5 φορές και τελικά υπολογίστηκε ο μέσος όρος των αποτελεσμάτων της ακρίβειας. Επίσης, να σημειωθεί ότι τα κλάσματα δειγματοληψίας υπολογίστηκαν αναλογικά με τα ποσοστά μείωσης δεδομένων που καθόρισαν οι τεχνικές μείωσης δεδομένων στην κάθε περίπτωση.



Σχήμα 5.7: k-fold cross-validation

Στον πίνακα 5.7 βλέπουμε τα αποτελέσματα της ακρίβειας του κατηγοριοποιητή που επιτυγχάνεται από την εκπαίδευση του συμπυκνωμένου συνόλου δεδομένων που δημιούργησε ο αλγόριθμος CNN και αντίστοιχα βλέπουμε τα αποτελέσματα που προέκυψαν όταν προστέθηκε θόρυβος 10% 30% και 50%. Παρομοίως, παρουσιάζονται τα αποτελέσματα του κατηγοριοποιητή όταν εκπαιδεύτηκε από στιγμιότυπα που επιλέχθηκαν σύμφωνα με τις δειγματοληπτικές τεχνικές: τυχαία απλή δειγματοληψία, στρωματοποιημένη και συστηματική δειγματοληψία. Όπως παρατηρούμε, ο αλγόριθμος CNN τα πηγαίνει γενικά καλύτερα απ'ότι η δειγματοληψία. Όταν όμως συγκριθούν τα αποτελέσματα για τις περιπτώσεις όπου έχουμε θόρυβο, τότε οι δειγματοληπτικές τεχνικές έχουν γενικά καλύτερες επιδόσεις, αν και δεν παρουσιάζεται μεγάλη διαφορά για 50% επίπεδο θορύβου. Επίσης, πραγματοποιήθηκαν πειράματα για τον έλεγχο της επίδοσης του CNN αλγορίθμου όταν προηγηθεί στα δεδομένα η εφαρμογή της τεχνικής αφαίρεσης θορύβου ENN. Σε αυτή τη περίπτωση(βλ. πίνακα 5.11) βελτιώνεται αισθητά η επίδοση του CNN ενώ οι δειγματοληπτικές μέθοδοι έχουνε συγκριτικά πολύ χαμηλότερες επιδόσεις. Όμως, η εφαρμογή της δειγματοληψίας πάνω στα καθαρά δεδομένα εκπαίδευσης(clean data) βελτιώνει σε πολύ μεγάλο βαθμό την αποτελεσματικότητα της καθώς φαίνεται και από τα αποτελέσματα του Wilcoxon signed-rank

test ότι κερδίζει στις περιπτώσεις 30% και 50% θορύβου.

Στον πίνακα 5.8 γίνετε η σύγκριση του αλγορίθμου παραγωγής προτύπων RSP3 με τις τεχνικές δειγματοληψίας. Προσέχουμε ότι ο RSP3 είναι γενικά καλύτερος. Ωστόσο σε πολλές περιπτώσεις(txr, sh, pd, ecl, mgt, ph, bn, tn), η δειγματοληψία είναι προτιμότερη όταν έχουμε θόρυβο 10% ή ακόμα και 30%. Για θόρυβο 50% δεν υπάρχει ουσιαστική συγκριτική διαφορά αλλά από τα αποτελέσματα του wilcoxon signed-rank test η δειγματοληψία κερδίζει έχοντας 6 νίκες έναντι 3 που έχει ο RSP3. Γενικά παρατηρούμε πόσο προβληματική είναι η ύπαρξη θορύβου καθώς ρίχνει την απόδοση του RSP3 κάτω από 60% για θόρυβο ακόμα και 30%. Όμως, όταν χρησιμοποιηθεί πάνω στα δεδομένα ο αλγόριθμος αφαίρεσης θορύβου ENN, τότε βλέπουμε ότι ο RSP3 έχει προβάδισμα έναντι της δειγματοληψίας καθώς οι επιδόσεις του πλέον είναι καλύτερες σε κάθε περίπτωση που εξετάστηκε. Τα αποτελέσματα φαίνονται αναλυτικά στον πίνακα 5.10. Όσον αφορά την ακρίβεια που επιτυγχάνει το μοντέλο με χρήση της δειγματοληψίας για την μείωση δεδομένων πάνω στα καθαρά δεδομένα εκπαίδευσης αυτή τη φορά, παρατηρούμε ότι υπάρχει σημαντική βελτίωση της ακρίβειας για κάθε επίπεδο θορύβου. Πιο συγκεκριμένα, ο enn-rsp3 είναι χειρότερος για 10% θόρυβο στα : tn, bn, mgt. Για 30% θόρυβο, η δειγματοληψία είναι καλύτερη στα : sh, tn, bn, mgt, pd ενώ για 50% θόρυβο ο enn-rsp3 είναι γενικά καλύτερος εκτός από λίγες περιπτώσεις.

Στον πίνακα 5.9 βλέπουμε την επίδοση του κατηγοριοποιητή με τη χρήση του αλγορίθμου αφαίρεσης θορύβου ENN και τη χρήση μεθόδων δειγματοληψίας. Στα καθαρά δεδομένα η ακρίβεια είναι πανομοιότυπη με αυτή των δειγματοληπτικών τεχνικών εκτός απο ορισμένα σύνολα δεδομένων. Όταν όμως συγκρίνεται η επίδοση τους σε συνθήκες θορύβου, τότε βλέπουμε μια τεράστια διαφορά υπέρ του ENN.

Στους πίνακες 5.14 - 5.21 βλέπουμε την στατιστική επικύρωση των αποτελεσμάτων με τη χρήση του Friedman και του wilcoxon signed rank test. Πιο συγκεκριμένα, το wilcoxon signed rank test συγκρίνει ανά ζεύγη τους αλγορίθμους σύμφωνα με τις επιδόσεις τους στα διάφορα σύνολα δεδομένων. Οι έλεγχοι αυτοί έγιναν συγκριτικά για τα ποσοστά επίτευξης μείωσης δεδομένων και ακρίβειας στο λογισμικό PSPP. Όπως βλέπουμε, η στήλη w/l/r προσδιορίζει το πλήθος νικών, ηττών και ισοπαλιών για κάθε ζεύγος αλγορίθμων ενώ η στήλη Wilc προσδιορίζει μια τιμή η οποία εκφράζει ποσοτικά τη σημαντικότητα της διαφοράς της επίδοσης των αλγορίθμων. Αν η τιμή είναι μικρότερη του 0.05, τότε είναι ασφαλές να ισχυριστεί κανείς ότι η διαφορά ειναι στατιστικά σημαντική. Σχετικά με την αποτελεσματικότητα μόνο των αλγορίθμων επιλογής και παραγωγής προτύπων βλέπουμε από τους αντίστοιχους πίνακες του wilcoxon ότι είναι καλύτερος ο ENN για όλες τις περιπτώσεις θορύβου αν και η δειγματοληψία είναι καλύτερη χωρις θόρυβο καθώς έχει 9 νίκες και 7 ήττες όπως βλέπουμε στον πίνακα 5.14. Δεύτερος καλύτερος έρχεται ο RSP3, με τον CNN να έρχεται τελευταίος. Για 10% θόρυβο όμως, βλέπουμε ότι τους συνδυασμούς ENN-CNN και ENN-RSP3 να έχουν ίσες νίκες που σημαίνει ότι έχουνε παρόμοια απόδοση.

Το Friedman test χρησιμοποιήθηκε για να δημιουργήσει μια κατάταξη των αλγορίθμων βάσει της ακρίβειας και των ποσοστών μείωσης δεδομένων. Επίσης, ο αλγόριθμος με τις καλύτερες επιδόσεις είναι αυτός με τη μεγαλύτερη τιμή βαθμολογίας. Για παράδειγμα, από τους πίνακες των αποτελεσμάτων του friedman test βλέπουμε ότι η δειγματοληψία με λόγο μείωσης που όρισε ο αλγόριθμος ENN έχει τη καλύτερη βαθμολογία από όλες τις άλλες τεχνικές όταν απουσίαζει θόρυβος. Μία άλλη παρατήρηση που μπορούμε να κάνουμε είναι ότι ο ENN κερδίζει σε όλα τα επίπεδα θορύβου όλους τους αλγορίθμους, ακόμα και τις συνδυαστικές περιπτώσεις εφαρμογής.

5.4 Συζήτηση

Η ανάλυση των πειραμάτων σχετικά με την αποτελεσματικότητα των διαφόρων τεχνικών μείωσης δεδομένων για το πρόβλημα της κατηγοριοποίησης με το μοντέλο του εγγύτερου γείτονα(1-NN) υποστηρίζει ότι οι τεχνικές δειγματοληψίας, σε σύγκριση με τον condensed nearest neighbor αλγόριθμο(CNN), έχουν γενικά καλύτερες επιδόσεις σε συνθήκες ύπαρξης θορύβου 10%, 30% και 50%. Στα σύνολα δεδομένων Ecoli, Wafer, Two Norm, MAGIC Gamma Telescope, έχει επίσης καλύτερα αποτελέσματα ακόμα και αν δεν υπάρχει θόρυβος. Όσο για την σύγκριση του RSP3 αλγορίθμου, παρατηρήσαμε ότι οι δειγματοληπτικές μέθοδοι είναι αποτελεσματικότεροι για ύπαρξη θορύβου 10%, 30% και 50%. Η δυσκολία του αλγορίθμου CNN να ανταπεξέλθει έχει τεκμηριωθεί στη σχετική βιβλιογραφία καθώς όσο αυξάνεται ο θόρυβος τόσο πιο ασαφή γίνονται τα όρια απόφασης και τόσο χειροτερεύει και η διαδικασία κατηγοριοποίησης. Ακόμη, η χρήση δειγματοληψίας πάνω σε καθαρά δεδομένα μέσω του ENN αλγορίθμου, αποδείχθηκε καλύτερη σε πάνω από το 50% των συνολικών dataset σε θόρυβο(10%, 30%), συγκριτικά με τον συνδυασμό της διαδοχικής εφαρμογής των ENN-CNN αλγορίθμων. Τέλος, η δειγματοληψία σε σύγκριση με τον RSP3 αλγόριθμο, σε καθαρά δεδομένα, διαπιστώθηκε ότι είναι καλύτερη στις εξής παρακάτω περιπτώσεις :

- tn, bn, mgt, wine για 10% θόρυβο
- sh, tn, ph, mgt, pd, wine, bn, kdd 30% θόρυβο
- pd, sh, wine, kdd για 50% θόρυβο

Όπως συζητήσαμε σε προηγούμενο κεφάλαιο, η υπολογιστικότητα του RSP3 αλγορίθμου είναι υψηλή, ειδικά για μεγάλα σύνολα δεδομένων. Από την άλλη πλευρά, οι δειγματοληπτικοί αλγόριθμοι είναι γενικά πιο γρήγοροι στην εκτέλεση τους και τα αποτελέσματα υποστηρίζουν ότι αποδίδουν συγκριτικά καλύτερα σε συνθήκες θορύβου. Το ίδιο συμπεραίνουμε και από την σύγκριση του CNN με τις δειγματοληπτικές μεθόδους.

Κεφάλαιο 6ο: Συμπεράσματα

Η συγκριτική πειραματική μελέτη αυτής της εργασίας είχε σκοπό την ανάλυση διάφορων τεχνικών αριθμητικής μείωσης δεδομένων με μεθόδους παραγωγής προτύπων(RSP3), επιλογής προτύπων(CNN, ENN), τεχνικών δειγματοληψίας και την εφαρμογή τους για κατηγοριοποίηση με τον κατηγοριοποιητή των k εγγύτερων γειτόνων. Τελικός στόχος ήταν η δημιουργία μικρότερων συνόλων δεδομένων με στόχο τη μείωση του υπολογιστικού κόστους για την εφαρμογή του κατηγοριοποιητή. Τα κεντρικά ερωτήματα και κίνητρο της εργασίας ήταν να διερευνηθεί ποια τεχνική είναι πιο αποτελεσματική, υπό ποιες προυποθέσεις συμβαίνει αυτό και εάν μπορεί η δειγματοληψία να τα καταφέρνει καλύτερα από τις άλλες τεχνικές όταν υπάρχουν διαφορετικά επίπεδα θορύβου(10% , 30% ,50%) στα δεδομένα. Συμπερασματικά η έρευνα αυτή έδειξε ότι οι πιθανοτικές τεχνικές δειγματοληψίας, σε όλα τα επίπεδα θορύβου, δημιουργούν μειωμένα σύνολα δεδομένων με πιο αποτελεσματικό τρόπο κατα πλειοψηφία, σε σχέση με τον αλγόριθμο CNN ενώ συγκριτικά με τον RSP3 το ίδιο συμπέρασμα είναι περισσότερο αληθές σε ύπαρξη θορύβου από 30% και πάνω.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer, 2009.

[2] T. M. Mitchell, *Machine learning*. McGraw-Hill, 1997.

[3] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, Inc., 2019.

[4] C. Sammut and G. I. Webb, *Encyclopedia of machine learning*. Springer Science & Business Media, 2011.

[5] I. Hendrickx and A. van den Bosch, "Hybrid algorithms with instance-based classification," Jan 1970.

[6] N. Altman, "A review of the k-nearest neighbor algorithm," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 22, no. 4, pp. 426–435, 1992.

[7] S. Patro and K. K. Sahu, "Normalization: A preprocessing stage," *arXiv preprint arXiv:1503.06462*, 2015.

[8] K. Taunk, S. De, S. Verma, and A. Swetapadma, "A brief review of nearest neighbor algorithm for learning and classification," *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, 2019.

[9] N. Kouiroukidis and G. Evangelidis, "The effects of dimensionality curse in high dimensional knn search," in *2011 15th Panhellenic Conference on Informatics*, pp. 41–45, IEEE, 2011.

[10] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artificial intelligence*, vol. 97, no. 1-2, pp. 245–271, 1997.

[11] S. N. Singh, "Sampling techniques & determination of sample size in applied statistics research : an overview," 2014.

[12] A. S. Acharya, A. Prakash, P. Saxena, and A. Nigam, "Sampling: Why and how of it," *Indian Journal of Medical Specialties*, vol. 4, no. 2, pp. 330–333, 2013.

[13] D. B. Chelton, "Effects of sampling errors in statistical estimation," *Deep Sea Research Part A. Oceanographic Research Papers*, vol. 30, no. 10, pp. 1083–1103, 1983.

[14] L. Martino, D. Luengo, and J. Míguez, *Independent random sampling methods*. Springer, 2018.

[15] X. Meng, "Scalable simple random sampling and stratified sampling," in *International Conference on Machine Learning*, pp. 531–539, PMLR, 2013.

[16] T. Nguyen, M.-H. Shih, D. Srivastava, S. Tirthapura, and B. Xu, "Stratified random sampling from streaming and stored data," *Distributed and Parallel Databases*, vol. 39, pp. 1–46, 09 2021.

[17] S. A. Mostafa and I. A. Ahmad, "Recent developments in systematic sampling: A review," *Journal of Statistical Theory and Practice*, vol. 12, no. 2, pp. 290–310, 2018.

[18] P. Sedgwick, "Cluster sampling," *Bmj*, vol. 348, 2014.

[19] G. Sharma, "Pros and cons of different sampling techniques," *International journal of applied research*, vol. 3, no. 7, pp. 749–752, 2017.

[20] P. Sedgwick, "Multistage sampling," *Bmj*, vol. 351, 2015.

[21] E. Namey, G. Guest, L. Thairu, and L. Johnson, "Data reduction techniques for large qualitative data sets," *Handbook for team-based qualitative research*, vol. 2, no. 1, pp. 137–161, 2008.

[22] B. Ghojogh, "Data reduction algorithms in machine learning and data science," 2021.

[23] S. Ougiaroglou and G. Evangelidis, "Fast and accurate k-nearest neighbor classification using prototype selection by clustering," in *2012 16th Panhellenic Conference on Informatics*, pp. 168–173, IEEE, 2012.

[24] O. Sutton, "Introduction to k nearest neighbour classification and condensed nearest neighbour data reduction," *University lectures, University of Leicester*, vol. 1, 2012.

[25] R. Alejo, J. M. Sotoca, R. M. Valdovinos, and P. Toribio, "Edited nearest neighbor rule for improving neural networks classifications," in *International Symposium on Neural Networks*, pp. 303–310, Springer, 2010.

[26] Y. B. Fernandez Hernandez, R. Bello, Y. Filiberto, M. Frías, L. Coello Blanco, and Y. Caballero, "An approach for prototype generation based on similarity relations for problems of classification," *Computación y Sistemas*, vol. 19, no. 1, pp. 109–118, 2015.

[27] S. Ougiaroglou, K. I. Diamantaras, and G. Evangelidis, "Exploring the effect of data reduction on neural network and support vector machine classification," *Neurocomputing*, vol. 280, pp. 101–110, 2018.

[28] S. Ougiaroglou and G. Evangelidis, "Rhc: a non-parametric cluster-based data reduction for efficient

k

k-nn classification," *Pattern Analysis and Applications*, vol. 19, no. 1, pp. 93–109, 2016.

[29] J. Sánchez, "High training set size reduction by space partitioning and prototype abstraction," *Pattern Recognition*, vol. 37, no. 7, pp. 1561–1564, 2004.

[30] S. Ougiaroglou, T. Mastromanolis, G. Evangelidis, and D. Margaris, "Fast training set size reduction using simple space partitioning algorithms," *Information*, vol. 13, no. 12, p. 572, 2022.

[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[32] J. Alcalá-Fdez, L. Sanchez, S. Garcia, M. J. del Jesus, S. Ventura, J. M. Garrell, J. Otero, C. Romero, J. Bacardit, V. M. Rivas, *et al.*, "Keel: a software tool to assess evolutionary algorithms for data mining problems," *Soft Computing*, vol. 13, no. 3, pp. 307–318, 2009.

[33] A. Asuncion and D. Newman, “Uci machine learning repository,” 2007.

[34] N. J. Salkind, *Encyclopedia of research design*. Thousand Oaks, CA: Sage Publications, 2010.

[35] J. R. Fraenkel and N. E. Wallen, *How to design and evaluate research in education*. New York, NY: McGraw-Hill, 2009.

[36] P. S. Maxim, “Sampling techniques,” *Journal of Statistics Education*, vol. 5, no. 3, 1997.