ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

«ΤΕΧΝΙΚΕΣ ΑΠΟΔΟΣΗΣ ΒΕΛΤΙΣΤΗΣ ΤΙΜΗΣ ΣΤΗΝ ΠΑΡΑΜΕΤΡΟ Κ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ Κ ΕΓΓΥΤΕΡΩΝ ΓΕΙΤΟΝΩΝ»

Της φοιτήτριας
Καραγκούνη Ελένης
Αρ. Μητρώου: 134030

Επιβλέπων
Κωνσταντίνος Γουλιάνας
Αναπληρωτής Καθηγητής

Ημερομηνία 6/2/2022

Τεχνικές απόδοσης βέλτιστης τιμής
στην παράμετρο κ του αλγορίθμου
κατηγοριοποίησης κ εγγύτερων γειτόνων
20200
Καραγκούνη Ελένη
Κωνσταντίνος Γουλιάνας
12-10-2020
6-2-2022

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

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

Στον Χρήστο

Πρόλογος

Η παρούσα Πτυχιακή Εργασία εκπονήθηκε στο πλαίσιο λήψης του τίτλου σπουδών του Μηχανικού Πληροφορικής από το τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του Διεθνούς Πανεπιστημίου Ελλάδος (Δι.Πα.Ε) κατά τη χρονική περίοδο 2020-2022.

Η εργασία πραγματοποιήθηκε υπό την επίβλεψη αρχικά του κ. Στέφανου Ουγιάρογλου, πρώην μέλος Εργαστηριακού Διδακτικού Προσωπικού (Ε.ΔΙ.Π) στο τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του Διεθνούς Πανεπιστημίου της Ελλάδος (ΔΙ.ΠΑ.Ε.) και στη συνέχεια υπό την επίβλεψη του κ. Κωνσταντίνου Γουλιάνα, Αναπληρωτή Καθηγητή του τμήματος Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του Διεθνούς Πανεπιστημίου της Ελλάδος (ΔΙ.ΠΑ.Ε.).

Αντικείμενο της εργασίας αποτελεί ο δημοφιλής κατηγοριοποιητής κ εγγύτερων γειτόνων (k Nearest Neighbors classification, k-NN) και θα παρουσιαστούν αναλυτικά οι μέθοδοι απόδοσης βέλτιστης τιμής στην παράμετρο κ.

Στόχος της εργασίας αποτελεί η εκτενής βιβλιογραφική έρευνα που αφορά τις μεθόδους απόδοσης της βέλτιστης τιμής στην παράμετρο κ για τον αλγόριθμο κατηγοριοποίησης κ εγγύτερων γειτόνων.

Περίληψη

Ένας από τους ευρέως χρησιμοποιούμενους αλγόριθμους κατηγοριοποίησης είναι ο k-Nearest Neighbors (k-NN). Η δημοτικότητά του οφείλεται κυρίως στην απλότητα, την αποτελεσματικότητα, την ευκολία εφαρμογής και την ικανότητά του στο να μπορούν να προστεθούν νέα δεδομένα στο σύνολο δεδομένων εκπαίδευσης ανά πάσα στιγμή. Ωστόσο, ένα από τα κύρια μειονεκτήματά του είναι το γεγονός ότι η απόδοσή του εξαρτάται σε μεγάλο βαθμό από τη σωστή επιλογή της παραμέτρου k, δηλαδή τον αριθμό των πλησιέστερων γειτόνων που εξετάζει ο αλγόριθμος. Η πιο συχνά χρησιμοποιούμενη τεχνική για τον προσδιορισμό του «καλύτερου» k είναι η διασταυρούμενη επικύρωση (cross validation), καθώς δεν υπάρχει γενικός κανόνας για την επιλογή της τιμής k λόγω της εξάρτησής της από το σύνολο δεδομένων εκπαίδευσης. Ωστόσο, η επιλογή μιας σταθερής τιμής k σε όλο το σύνολο δεδομένων δεν λαμβάνει υπόψη τα ιδιαίτερα χαρακτηριστικά του, όπως η κατανομή των δεδομένων, τον διαχωρισμό των κλάσεων, τις ανισορροπίες στις κατανομές των στιγμιότυπων στις κλάσεις, τις αραιές και τις πυκνές γειτονιές και υποχώρους με αυξημένα επίπεδα θορύβου. Πολλές έρευνες έχουν γίνει μέχρι σήμερα στο συγκεκριμένο πεδίο, οδηγώντας σε πολλές παραλλαγές του κατηγοριοποιητή που αφορούν στην απόδοση της βέλτιστης τιμής στην παράμετρο k. Στην παρούσα εργασία, πραγματοποιείται μια διεξοδική βιβλιογραφική ανασκόπηση προκειμένου να συνοψιστούν όλα τα επιτεύγματα που έχουν επιτευχθεί μέχρι σήμερα στον τομέα αυτό.

Best value techniques of k in kNN classifier

Karagkouni Eleni

Abstract

K-Nearest Neighbors is a popular classification algorithm (k-NN). Its simplicity, effectiveness, ease of use, and flexibility to add fresh data to the training set whenever necessary all contribute to its appeal. However, one of its key downsides is that the choice of parameter k, or the number of nearest neighbors that the algorithm looks at, has a significant impact on how well it performs. As there is no universal strategy for selecting the k value due to its dependence on the training dataset, cross validation is the method that is most usually employed to determine the "best" k value. The dataset's unique characteristics, such as data distribution, class separation, imbalanced classes, sparse and dense neighborhoods, and noisy subspaces, are not considered when a fixed k value is chosen throughout. Numerous k-NN versions have been developed as a result of extensive research in the particular area to date. To compile all the advancements made so far in this topic, a complete literature review is undertaken in the current study. In particular, a collection of twenty-eight (28) methodologies and procedures for dynamic "best" k selection is described, together with their experimental outcomes.

Ευχαριστίες

Αρχικά θα ήθελα να ευχαριστήσω τους καθηγητές μου κ. Στέφανο Ουγιάρογλου και κ. Κωνσταντίνο Γουλιάνα για την επίβλεψη της πτυχιακής εργασίας, την καθοδήγησή τους, την πρακτική και ουσιαστική βοήθεια και στήριξη που είχα και από τους δύο όχι μόνο για την ολοκλήρωση αυτής της εργασίας αλλά και καθ΄ όλη τη διάρκεια των σπουδών.

Επίσης, θα ήθελα να ευχαριστήσω και τα μέλη της επιτροπής.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένεια μου που ήταν πάντα δίπλα μου και με υποστήριζε κυρίως στις δύσκολες στιγμές. Το μεγαλύτερο όμως ευχαριστώ θέλω να πω στον γιο μου, Χρήστο, για την υπομονή που έδειξε.

Περιεχόμενα

Κεφάλαιο 1ο: Εισαγωγή

1.1 Κατηγοριοποίηση

Η αποδοτικότητα και η αποτελεσματικότητα των αλγορίθμων εξόρυξης δεδομένων είναι ένα σημαντικό ερευνητικό πρόβλημα που έχει προσελκύσει την προσοχή τόσο του ακαδημαϊκού χώρου όσο και του κλάδου της βιομηχανίας [1]. Η κατηγοριοποίηση (ή η εποπτευόμενη μάθηση σύμφωνα με την ορολογία μηχανικής μάθησης) είναι μια κρίσιμη εργασία εξόρυξης δεδομένων.

Οι αλγόριθμοι κατηγοριοποίησης (ή κατηγοριοποιητές)[2] προσπαθούν να αντιστοιχίσουν νέα, μη κατηγοριοποιημένα στοιχεία δεδομένων σε ένα σύνολο προκαθορισμένων κλάσεων, με βάση τα διαθέσιμα δεδομένα εκπαίδευσης, δηλαδή ένα σύνολο ήδη κατηγοριοποιημένων περιπτώσεων (ή στοιχείων). Η κατηγοριοποίηση είναι από τις πιο γνωστές και δημοφιλέστερες τεχνικές εξόρυξης γνώσης και Μηχανικής Μάθησης. Συστήματα κατηγοριοποίησης χρησιμοποιούνται από πολλές εταιρείες είτε του δημόσιου είτε του ιδιωτικού τομέα καθημερινά. Συστήματα όπως αναγνώρισης προτύπων, συστήματα ανίχνευσης λαθών σε βιομηχανικές εφαρμογές, συστήματα κατηγοριοποίησης των τάσεων στην οικονομία, συστήματα έγκρισης δανείων και πιστωτικών καρτών, συστήματα ιατρικών διαγνώσεων και πολλά άλλα είναι μερικά παραδείγματα τέτοιου είδους συστημάτων. Η κατηγοριοποίηση ενός εισερχόμενου email είτε στην κατηγορία “spam” είτε στην κατηγορία “non-spam” είναι ένα κλασικό παράδειγμα κατηγοριοποίησης.

Η γνώση των δεδομένων είναι βασική προϋπόθεση σε όλες τις προσεγγίσεις στην εκτέλεση της κατηγοριοποίησης. Ένα σύνολο εκπαίδευσης χρησιμοποιείται συνήθως, για τον καθορισμό συγκεκριμένων παραμέτρων που απαιτούνται από την τεχνική αλλά και για την εκπαίδευση του κατηγοριοποιητή. Με άλλα λόγια, αν το αντικείμενο της κατηγοριοποίησης είναι η κατηγοριοποίηση ενός email στη κλάση spam ή στη κλάση όχι spam, ο κατηγοριοποιητής, πρέπει να εκπαιδευτεί στο να ξεχωρίζει τα emails που είναι spam και αυτά που δεν είναι spam «δείχνοντάς» του αντιπροσωπευτικά emails και από τις δύο κατηγορίες. Αυτό το δείγμα δεδομένων εισόδου όπως και η κατηγοριοποίηση που έχει δοθεί σε αυτά τα δεδομένα αποτελούν στοιχεία των δεδομένων εκπαίδευσης.

Το πρόβλημα της κατηγοριοποίησης μπορεί να οριστεί ως εξής: Η κατηγοριοποίηση (classification), είναι η διαδικασία με την οποία ένα σύνολο δεδομένων απεικονίζεται σε προκαθορισμένες ομάδες. Οι ομάδες αυτές συνήθως αναφέρονται ως κατηγορίες ή κλάσεις. Σύμφωνα με τον παραπάνω ορισμό η κατηγοριοποίηση θεωρείται ως μια απεικόνιση από τη Βάση Δεδομένων στο σύνολο των κατηγοριών. Οι κατηγορίες δεν επικαλύπτονται, διαμερίζουν ολόκληρη τη Βάση Δεδομένων και είναι προκαθορισμένες. Επίσης, κάθε στοιχείο της Βάσης Δεδομένων τοποθετείται σε μία μόνο κατηγορία.

Δύο είναι τα βασικά στάδια τα οποία περιλαμβάνονται στην επίλυση των προβλημάτων κατηγοριοποίησης. Αρχικά, από την αξιολόγηση και την ανάλυση των δεδομένων δημιουργείται ένα μοντέλο. Ως είσοδο, σε αυτό το βήμα, χρησιμοποιούνται τα δεδομένα εκπαίδευσης και ως έξοδο ένας ορισμός του μοντέλου που αναπτύχθηκε. Τα δεδομένα εκπαίδευσης μπορούν πλέον να κατηγοριοποιηθούν με τη μεγαλύτερη δυνατή ακρίβεια από το μοντέλο που δημιουργήθηκε σε αυτό το στάδιο. Όταν το σύνολο των δεδομένων εκπαίδευσης περιλαμβάνει ένα χαρακτηριστικό το οποίο δείχνει σε ποια κλάση έχει κατηγοριοποιηθεί το κάθε αντικείμενο, δηλαδή όταν οι κατηγορίες του συνόλου δεδομένων εκπαίδευσης είναι ήδη γνωστές, τότε το βήμα αυτό ονομάζεται εποπτευόμενη μάθηση (supervised learning). Αντίθετα, όταν δηλαδή οι κατηγορίες του συνόλου δεδομένων δεν είναι γνωστές, τότε το βήμα αυτό ονομάζεται μη εποπτευόμενη μάθηση (unsupervised learning-clustering). Σε αυτή την εργασία δεν ασχολούμαστε με τη μη εποπτευόμενη μάθηση. Αντιθέτως, το αντικείμενο της παρούσας εργασίας είναι η εποπτευόμενη μάθηση, δηλαδή η κατηγοριοποίηση. Στη συνέχεια με την εφαρμογή του μοντέλου που αναπτύχθηκε στο βήμα της εκπαίδευσης γίνεται η κατηγοριοποίηση των αντικειμένων Βάσης Δεδομένων που εξετάζεται (μελλοντικές περιπτώσεις).

Οι κατηγοριοποιητές μπορούν να χωριστούν σε δύο κύριες κατηγορίες αλγορίθμων[2]:

i. κατηγοριοποιητές πρόθυμοι (eager), και
ii. κατηγοριοποιητές οκνηροί (lazy) (ή βασισμένοι σε παράδειγμα)

Και οι δύο μοιράζονται το ίδιο κίνητρο, δηλαδή την ακριβή πρόβλεψη της κλάσης. Ωστόσο, διαφέρουν ως προς τον τρόπο λειτουργίας τους. Ουσιαστικό ρόλο για την αποτελεσματικότητα των αλγορίθμων και των δύο κατηγοριών παίζει το διαθέσιμο σετ εκπαίδευσης. Ένας πρόθυμος κατηγοριοποιητής προ επεξεργάζεται τα διαθέσιμα δεδομένα εκπαίδευσης και δημιουργεί ένα μοντέλο κατηγοριοποίησης που στη συνέχεια χρησιμοποιείται για την κατηγοριοποίηση νέων, μη κατηγοριοποιημένων στοιχείων. Από την άλλη πλευρά, οι οκνηροί κατηγοριοποιητές δεν δημιουργούν κανένα μοντέλο κατηγοριοποίησης. Στην πραγματικότητα, θεωρούν το σύνολο δεδομένων εκπαίδευσης ως μοντέλο κατηγοριοποίησης. Ένας οκνηρός αλγόριθμος κατηγοριοποιεί ένα νέο αντικείμενο σαρώνοντας το σετ εκπαίδευσης τη στιγμή που φτάνει εξετάζοντας εκείνη τη στιγμή τα δεδομένα εκπαίδευσης. Με άλλα λόγια, στην περίπτωση των πρόθυμων κατηγοριοποιητών, ο κατηγοριοποιητής μαθαίνει από τα δεδομένα εκπαίδευσης στο να κάνει σωστές προβλέψεις και από τη στιγμή που ολοκληρωθεί η διαδικασία της μάθησης, τα δεδομένα εκπαίδευσης μπορούν να διαγραφούν. Από την άλλη, οι οκνηροί κατηγοριοποιητές, στην πραγματικότητα δεν εκπαιδεύονται. Όταν κάποιο στοιχείο πρέπει να κατηγοριοποιηθεί, εκείνη τη στιγμή εξετάζουν τα δεδομένα εκπαίδευσης ώστε να αποφασίσουν σε ποια κατηγορία πρέπει να κατηγοριοποιηθεί.

Δεδομένου ότι οι πρόθυμοι κατηγοριοποιητές δημιουργούν ένα μοντέλο κατηγοριοποίησης πριν από την άφιξη οποιουδήποτε νέου στιγμιότυπου, η διαδικασία κατηγοριοποίησης είναι πολύ γρήγορη. Αν και οι οκνηροί κατηγοριοποιητές δεν ξοδεύουν καθόλου χρόνο για να δημιουργήσουν μοντέλα

κατηγοριοποίησης, η διαδικασία κατηγοριοποίησής τους είναι πιο χρονοβόρα από αυτή των πρόθυμων κατηγοριοποιητών. Ένα μειονέκτημα των πρόθυμων κατηγοριοποιητών είναι ότι πρέπει να δημιουργήσουν μια ενιαία υπόθεση που καλύπτει ολόκληρο το εκπαιδευτικό σύνολο. Αυτό δεν είναι πάντα εφικτό, μπορεί να επηρεάσει την ακρίβεια κατηγοριοποίησης και μπορεί να καταστήσει την κατασκευή του μοντέλου κατηγοριοποίησης μια εξαιρετικά χρονοβόρα και περίπλοκη εργασία προ επεξεργασίας.

Από την άλλη πλευρά, οι οκνηροί κατηγοριοποιητές χρησιμοποιούν ολόκληρο το σύνολο εκπαίδευσης και, έτσι, μπορούν να υιοθετήσουν πιο σύνθετες υποθέσεις σχετικά με τα δεδομένα. Κατά συνέπεια, μπορεί να βελτιώσουν την ακρίβεια κατηγοριοποίησης. Ένα μειονέκτημα των οκνηρών κατηγοριοποιητών είναι ότι απαιτούν όλα τα δεδομένα εκπαίδευσης να είναι πάντα διαθέσιμα, γεγονός που οδηγεί σε υψηλές απαιτήσεις αποθήκευσης. Αντίθετα, στην πρόθυμη κατηγοριοποίηση, μετά την κατασκευή του μοντέλου κατηγοριοποίησης τα δεδομένα εκπαίδευσης μπορούν να αφαιρεθούν για εξοικονόμηση χώρου.

Τις τελευταίες δεκαετίες, το πρόβλημα της κατηγοριοποίησης έχει προσελκύσει το ενδιαφέρον πολλών ερευνητών από διαφορετικά ερευνητικά πεδία της επιστήμης των υπολογιστών. Ως εκ τούτου, διάφοροι πρόθυμοι και οκνηροί κατηγοριοποιητές [3] αποτελούν μια πολύ γνωστή υποκατηγορία πρόθυμων κατηγοριοποιητών. Με βάση τα διαθέσιμα δεδομένα εκπαίδευσης, αυτοί οι κατηγοριοποιητές δημιουργούν μια δομή δέντρου που χρησιμοποιείται για την κατηγοριοποίηση νέων στοιχείων.

Άλλοι πρόθυμοι (eager) κατηγοριοποιητές βασίζονται σε τεχνητά νευρωνικά δίκτυα[4]. Ένα νευρωνικό δίκτυο αρχικά εκπαιδεύεται χρησιμοποιώντας το σύνολο εκπαίδευσης και στη συνέχεια εκτελεί κατηγοριοποιήσεις. Οι πιθανοτικοί κατηγοριοποιητές ανήκουν επίσης στην κατηγορία των πρόθυμων κατηγοριοποιητών. Κατασκευάζουν ένα μοντέλο κατηγοριοποίησης που βασίζεται σε πιθανότητες. Ένα χαρακτηριστικό παράδειγμα ενός πιθανοτικού κατηγοριοποιητή είναι ο αφελής (naive) κατηγοριοποιητής Bayes [5]. Αρχικά, αυτού του είδους τους κατηγοριοποιητές δεν τους θεωρούσαν ως εργαλεία κατηγοριοποίησης. Αυτό στη συνέχεια άλλαξε καθώς ανακάλυψαν πως έχουν αυξημένες δυνατότητες κατηγοριοποίησης παρόμοιες με αυτές των νευρωνικών δικτύων και των δέντρων αποφάσεων.

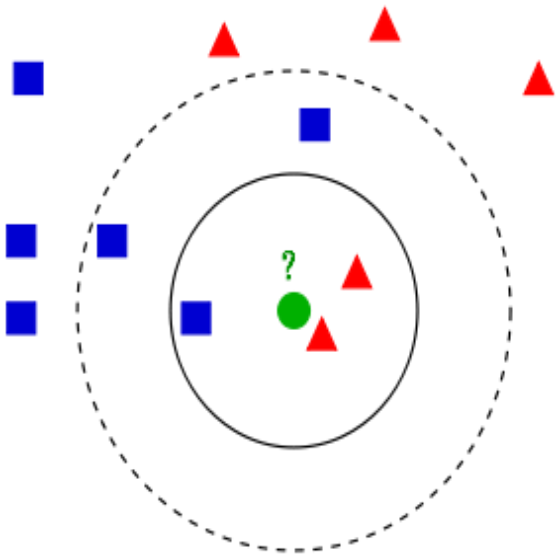
Μια άλλη υποκατηγορία πρόθυμων αλγορίθμων περιλαμβάνει τους κατηγοριοποιητές που βασίζονται στην εξόρυξη κανόνων συσχέτισης[6]. Ανακαλύπτουν κανόνες συσχέτισης μέσα στα διαθέσιμα δεδομένα εκπαίδευσης. Αυτοί οι κανόνες χρησιμοποιούνται για σκοπούς κατηγοριοποίησης. Από την άλλη πλευρά, η κατηγορία των οκνηρών κατηγοριοποιητών περιλαμβάνει τον γνωστό κατηγοριοποιητή k Nearest Neighbors[7] και τις μεθόδους κατηγοριοποίησης συλλογισμών που βασίζονται σε περιπτώσεις (case based learning) [8]. Ο κατηγοριοποιητής k Nearest Neighbors είναι το αντικείμενο μελέτης της παρούσας πτυχιακής εργασίας.

1.2 Κατηγοριοποίηση k-NN

Ο κατηγοριοποιητής k Nearest Neighbors (k-NN) [7] είναι ένας αποτελεσματικός και ευρέως χρησιμοποιούμενος οκνηρός αλγόριθμος κατηγοριοποίησης. Είναι ένας απλός και εύκολος στην εφαρμογή κατηγοριοποιητής, μπορεί να χρησιμοποιηθεί σε πολλούς τομείς εφαρμογών και μπορεί εύκολα να ενσωματωθεί σε πολλά συστήματα.

Επιπλέον, ο κατηγοριοποιητής kNN είναι αναλυτικά ισχυρός και για k=1 και απεριόριστα είδη το ποσοστό σφάλματος ασυμπτωματικά δεν είναι ποτέ χειρότερο από το διπλάσιο του ελάχιστου δυνατού, που είναι το ποσοστό του Bayes[9]. Δεδομένου ότι ο κατηγοριοποιητής k-NN είναι ένας οκνηρός κατηγοριοποιητής, δεν δημιουργεί κανένα μοντέλο κατηγοριοποίησης. Ο αλγόριθμος χρησιμοποιεί τα δεδομένα εκπαίδευσης όποτε χρειάζεται να κατηγοριοποιηθεί ένα νέο αντικείμενο. Συγκεκριμένα, κατηγοριοποιεί ένα στοιχείο x αναζητώντας στα διαθέσιμα δεδομένα εκπαίδευσης και ανακτώντας τα k πλησιέστερα στοιχεία (γείτονες) στο x σύμφωνα με μια μέτρηση απόστασης. Στη συνέχεια, το x εκχωρείται στην πιο κοινή κλάση μεταξύ των κλάσεων των ανακτημένων k πλησιέστερων γειτόνων. Αυτή η τάξη ονομάζεται συχνά η κύρια τάξη και καθορίζεται μέσω μιας διαδικασίας γνωστής ως ψηφοφορίας των πλησιέστερων γειτόνων. Σημειώστε ότι όταν k = 1, ο αλγόριθμος είναι επίσης γνωστός ως κατηγοριοποιητής k Nearest Neighbors (ή κανόνας 1-NN).

Στο παρακάτω σχήμα παρουσιάζεται ένα παράδειγμα κατηγοριοποίησης με βάση τον εν λόγω κατηγοριοποιητή. Στο σχήμα αυτό, φαίνονται οι δύο κατηγορίες στις οποίες έχουν χωριστεί τα δεδομένα εκπαίδευσης. Οι κατηγορίες είναι τα τρίγωνα και τα τετράγωνα. Το στοιχείο-ερώτημα q είναι αυτό που προορίζεται για κατηγοριοποίηση σε μία από τις δύο κλάσεις και αναπαρίσταται με τον πράσινο κύκλο. Αν για k=3 τότε ο αλγόριθμος προβλέπει ότι η κατηγορία είναι το τρίγωνο, όπως φαίνεται αντίστοιχα και στον μικρότερης διαμέτρου κύκλο. Ενώ για k=5 ο αλγόριθμος προβλέπει ότι το q θα κατηγοριοποιηθεί στην κατηγορία τετράγωνο, όπως φαίνεται και στον αντίστοιχο κύκλο με τη μεγαλύτερη διάμετρο.



Εικόνα 1: Η γειτονιά του νέου στιγμιότυπου που θα εξεταστεί με κ=3 και κ=5

1.3 Μέτρα ομοιότητας στον κατηγοριοποιητή k-NN

Ένα άλλο σημαντικό ζήτημα που πρέπει να αντιμετωπιστεί στη χρήση του κατηγοριοποιητή των κ εγγύτερων γειτόνων είναι η επιλογή της μετρικής που χρησιμοποιείται για τον υπολογισμό της απόστασης μεταξύ των στοιχείων. Ασφαλώς, αυτή η απόφαση θα πρέπει να λάβει υπόψη τους τύπους δεδομένων των χαρακτηριστικών των δεδομένων (μεταβλητές). Σε περιπτώσεις πραγματικών ή/και ακέραιων χαρακτηριστικών, η Ευκλείδεια απόσταση είναι η μετρική απόστασης που χρησιμοποιείται συνήθως. Ωστόσο, μπορούν να υιοθετηθούν και άλλες μετρικές απόστασης (π.χ. Mahalanobis, Manhatan, Minkowski, Chebyshev)[15].

Παρότι υπάρχουν αρκετές δυνατές επιλογές, οι περισσότεροι κατηγοριοποιητές που βασίζονται σε μέτρα ομοιότητας χρησιμοποιούν την Ευκλείδεια απόσταση. Η απόσταση ανάμεσα σε ένα νέο αντικείμενο x': $< a_1(x'),..., a_m(x') >$ και τα αποθηκευμένα αντικείμενα εκπαίδευσης x_i: $< a_1(x_i),..., a_m(x_i) >$ (όπου m είναι ο αριθμός των χαρακτηριστικών) ορίζεται ως:

$$d(x_i, x') = \sqrt{\sum_{j=1}^{m} \left(a_j(x_i) - a_j(x')\right)^2}$$

Όταν συγκρίνουμε αποστάσεις δεν είναι απαραίτητο να υπολογίζουμε την τετραγωνική ρίζα, μπορούμε να συγκρίνουμε απευθείας τα αθροίσματα των τετραγώνων. Αυτό θα είχε ως αποτέλεσμα την επιτάχυνση της διαδικασίας υπολογισμού της απόστασης. Μια εναλλακτική της Ευκλείδειας απόστασης

είναι η απόσταση Manhattan ή city-block, όπου η διαφορά μεταξύ των τιμών των χαρακτηριστικών δεν υψώνεται στο τετράγωνο αλλά αθροίζεται (αφού έχουμε πάρει την απόλυτη τιμή της):

$$d(x_i, x') = \sum_{j=1}^{m} \left| a_j(x_i) - a_j(x') \right|$$

Άλλα μέτρα απόστασης λαμβάνονται υψώνοντας σε μεγαλύτερες δυνάμεις από το τετράγωνο. Ένα τέτοιο μέτρο είναι η απόσταση Minkowski.

$$d(x_i, x') = \sum_{j=1}^{m} \left| a_j(x_i) - a_j(x') \right|$$

Το λ είναι ένας ακέραιος. Αν λ=1, τότε έχουμε την απόσταση Manhattan. Αν λ=2, τότε έχουμε την απόσταση Ευκλείδεια απόσταση. Ο ρόλος του λ, όταν αυξάνεται, είναι να μεγεθύνει την απόσταση ανάμεσα στο πιο ανόμοια αντικείμενα σε σχέση με τα πιο όμοια. Γενικά, η Ευκλείδεια απόσταση αντιπροσωπεύει έναν καλό συμβιβασμό. Άλλα μέτρα απόστασης μπορεί να είναι πιο κατάλληλα σε ειδικές περιστάσεις.

Τέλος θα πρέπει να αναφέρουμε ότι υπάρχουν και άλλες, λιγότερο γνωστές αποστάσεις που έχουν εμφανιστεί στην βιβλιογραφία. Θα ήταν λάθος να θεωρήσουμε ότι ως μέτρα ομοιότητας χρησιμοποιούνται μόνο οι διάφοροι μαθηματικοί τύποι αποστάσεων. Υπάρχουν και άλλα μέτρα ομοιότητας που έχουν εφαρμοστεί σε συστήματα ανάκτησης πληροφορίας και στις μηχανές αναζήτησης στο διαδίκτυο. Μερικά από τα μέτρα αυτά είναι το μέτρο Dice, Jaccard, το συνημίτονο και η επικάλυψη.

Διαφορετικά εύρη στα χαρακτηριστικά μπορούν να επηρεάσουν την τιμή της απόστασης. Ακόμη και σε περιπτώσεις όπου όλα τα χαρακτηριστικά έχουν την ίδια σημασία, τα χαρακτηριστικά με μεγάλα εύρη έχουν μεγαλύτερο αντίκτυπο στην τιμή της απόστασης από τα χαρακτηριστικά με μικρότερα εύρη. Αν χρησιμοποιούσαμε απευθείας τον τύπο της Ευκλείδειας απόστασης, η επίδραση κάποιων από τα χαρακτηριστικά στην απόσταση δύο αντικειμένων, μπορεί να υποσκιαζόταν πλήρως από την επίδραση χαρακτηριστικών με μεγαλύτερες κλίμακες τιμών. Συνεπώς είναι συνήθης πρακτική η κανονικοποίηση των τιμών όλων των χαρακτηριστικών στο διάστημα μεταξύ 0 και 1. Ας υποθέσουμε ότι το χαρακτηριστικό «μισθός» κυμαίνεται από 800 έως 5000 και το χαρακτηριστικό «αριθμός παιδιών» παίρνει τιμές από 0 έως 6. Με και τα δύο χαρακτηριστικά να έχουν την ίδια αξία, ο «μισθός» έχει μεγαλύτερη επίδραση στον υπολογισμό της απόστασης από τον «αριθμός παιδιών». Επομένως, το εύρος των χαρακτηριστικών θα πρέπει να κανονικοποιηθεί σε ένα συγκεκριμένο εύρος διαστήματος[16].

Μέχρι τώρα έχουμε υποθέσει την ύπαρξη αριθμητικών χαρακτηριστικών. Στα αριθμητικά χαρακτηριστικά, η διαφορά μεταξύ δύο τιμών είναι απλά η αριθμητική τους διαφορά. Για διακριτά χαρακτηριστικά τα οποία λαμβάνουν συμβολικές και όχι αριθμητικές τιμές, συνηθίζεται να παίρνουμε ως 1 την απόσταση μεταξύ δύο τιμών που δεν ταυτίζονται και ως 0 την απόσταση όταν οι τιμές ταυτίζονται. Σε αυτήν την περίπτωση δεν απαιτείται κανονικοποίηση αφού χρησιμοποιούνται μόνο οι τιμές 0 και 1. Αξίζει να σημειωθεί ότι πολλά άλλα μέτρα ομοιότητας έχουν προταθεί για τον χειρισμό ονομαστικών χαρακτηριστικών. Ωστόσο, δεν παρουσιάζονται στην εργασία.

1.4 Πλεονεκτήματα και μειονεκτήματα του κατηγοριοποιητή kNN

Ο κατηγοριοποιητής k-NN είναι ένας από τους δημοφιλέστερους στο είδος του. Τα πλεονεκτήματα που έχει και τον καθιστούν τόσο δημοφιλή είναι τα παρακάτω:

- Είναι πολύ απλός στην κατανόηση και εξίσου εύκολος στην εφαρμογή του. Απαιτείται μόνο η απόδοση μιας ακέραιας τιμής στην παράμετρο k, ένα σύνολο εκπαίδευσης και ένα μέτρο απόστασης.
- Έχει πολλές εφαρμογές όπως αναγνώριση προτύπων, κατηγοριοποίηση χρονοσειρών κτλ.
- Είναι αποτελεσματικός σε εφαρμογές όπου κάποιο αντικείμενο μπορεί να ανήκει σε πάνω από μια κλάση.
- Μπορεί να χρησιμοποιηθεί τόσο για προβλήματα ταξινόμησης όσο και για προβλήματα παλινδρόμησης.
- Υπάρχει ευελιξία στην επιλογή της μέτρησης της απόστασης. Η Ευκλείδεια Απόσταση, η Απόσταση Μανχάταν, η Απόσταση Hamming και η Απόσταση Minkowski είναι αυτές που χρησιμοποιούνται πιο συχνά.
- Δεν είναι απαραίτητο να σχεδιάσει ή να μάθει κάποιο μοντέλο και έτσι μπορεί να προσαρμόζεται εύκολα όταν παρουσιάζονται αλλαγές στα δεδομένα εκπαίδευσης.

Το βασικό πλεονέκτημα οφείλεται στο γεγονός ότι είναι ένας lazy αλγόριθμος, δηλαδή δεν είναι απαραίτητη καμία εκπαίδευση στο σύνολο μέχρι να φτάσει κάποιο αντικείμενο για κατηγοριοποίηση. Το παραπάνω γεγονός αποτελεί ένα από τα πλεονεκτήματα του κατηγοριοποιητή ταυτόχρονα όμως γίνεται και η αιτία για ένα από τα μειονεκτήματά του καθώς αυξάνει το υπολογιστικό κόστος του κατηγοριοποιητή. Η σειριακή αναζήτηση είναι ο πιο απλός αλγόριθμος για την υλοποίηση του κατηγοριοποιητή. Αυτό σημαίνει ότι ο κατηγοριοποιητής k-NN υπολογίζει όλες τι αποστάσεις μεταξύ των αντικειμένων x που πρόκειται να κατηγοριοποιηθούν και όλων των δεδομένων εκπαίδευσης. Με άλλα λόγια, πρέπει να υπολογιστεί η απόσταση κάθε αντικειμένου που πρόκειται να κατηγοριοποιηθεί με κάθε ένα αντικείμενο του συνόλου εκπαίδευσης. Αυτό σημαίνει ότι εάν ένα σύνολο δεδομένων έχει 15.000 αντικείμενα, από τα οποία τα 10.000 είναι τα δεδομένα εκπαίδευσης και τα 5.000 είναι τα νέα

αντικείμενα προς κατηγοριοποίηση, τότε ο αριθμός των αποστάσεων που πρέπει να υπολογιστούν είναι 10.000 * 5.000 = 50.000.000. Το κόστος των υπολογισμών αυτών αυξάνεται εκθετικά όσο μεγαλύτερη γίνεται η βάση δεδομένων. Επιπλέον, το κόστος για κάθε υπολογισμό απόστασης εξαρτάται από το πλήθος των χαρακτηριστικών (attributes) των αντικειμένων. Άρα, καθώς το σύνολο δεδομένων αυξάνεται η αποτελεσματικότητα και η ταχύτητα του κατηγοριοποιητή μειώνονται πολύ γρήγορα.

Συνεπώς, το κόστος υπολογισμού όλων των αποστάσεων είναι μεγάλο. Αυτό είναι και το πιο σημαντικό μειονέκτημα του κατηγοριοποιητή k-NN καθώς συνεχίζονται οι έρευνες περί αυτού από ερευνητές διαφόρων περιοχών της πληροφορικής, όπως η Μηχανική Μάθηση και η Τεχνητή Νοημοσύνη, η Στατιστική, η Εξόρυξη Δεδομένων και οι Βάσεις Δεδομένων. Η δημοσίευση πολλών εργασιών στις οποίες προτείνονται μέθοδοι για να αντιμετωπιστεί αυτό το σημαντικό μειονέκτημα του κατηγοριοποιητή k-NN είναι και το αποτέλεσμα της προσπάθειας των ερευνητών.

Σε δυο βασικές κατηγορίες επιτάχυνσης του κατηγοριοποιητή k-NN μπορούν να κατηγοριοποιηθούν οι προτεινόμενες μέθοδοι (i) Μέθοδοι δεικτοδότησης πολυδιάστατων δεδομένων (Multiattribute Indexing) και (ii) Μέθοδοι μείωσης του όγκου δεδομένων (Data Reduction Techniques).

Τα αποτελέσματα της έρευνας των επιστημόνων, οι οποίοι προέρχονται από την περιοχή των Βάσεων Δεδομένων είναι οι μέθοδοι δεικτοδότησης πολυδιάστατων δεδομένων[29,30]. Η προ επεξεργασία των διαθέσιμων δεδομένων εκπαίδευσης είναι απαιτούμενη γι' αυτές τις μέθοδοι ώστε να κατασκευαστεί η δομή του δείκτη, η μορφή της οποίας συνήθως είναι δενδροειδής. Λόγω της ικανότητάς τους να αποφεύγουν πολλούς υπολογισμούς αποστάσεων, οι μέθοδοι αναζήτησης κοντινότερων γειτόνων συνήθως είναι πολύ αποτελεσματικές σε τέτοιου είδους δομές.

Δεκάδες είναι οι μέθοδοι δεικτοδότησης πολυδιάστατων δεδομένων που έχουν προταθεί. Το k-d-tree, το R-tree και οι διάφορες παραλλαγές του [31], k-DB-tree [32] και το Vantage Point (VP) tree είναι μερικά χαρακτηριστικά παραδείγματα. Το πιο βασικό μειονέκτημα των περισσότερων από αυτών των μεθόδων είναι η μεγάλου βαθμού εξάρτηση της απόδοσης αναζήτησης εγγύτερων γειτόνων σε δεικτοδοτημένα δεδομένα από το πλήθος των διαστάσεων. Όταν τα δεδομένα που χρησιμοποιούνται είναι λίγων διαστάσεων, δηλαδή λιγότερα από 10, τότε, η απόδοση (η ταχύτητα εκτέλεσης) διατηρείται σε υψηλά επίπεδα. Σε μεγαλύτερες διαστάσεις υπάρχει σταδιακή μείωση της απόδοσης και μπορεί να φτάσει ή και να ξεπεράσει τα χαμηλά επίπεδα της σειριακής αναζήτησης. Το φαινόμενο αυτό είναι γνωστό ως φαινόμενο της κατάρας των διαστάσεων (dimensionality curse). Ως εκ τούτου, όταν τα διαθέσιμα δεδομένα είναι λίγων διαστάσεων η αναζήτηση εγγύτερων γειτόνων επιταχύνεται από τις μεθόδους δεικτοδότησης.

Όπως έχει ήδη αναφερθεί υπάρχουν τεχνικές μείωσης των διαστάσεων [33]. Οι τεχνικές αυτές μπορούν να αντιμετωπίσουν το πρόβλημα δεικτοδότησης δεδομένων με μεγάλο αριθμό διαστάσεων. Ωστόσο, η εφαρμογή της αποτελεί ένα επιπλέον βήμα προ επεξεργασίας και αυτό φέρνει ως αποτέλεσμα το επιπλέον κόστος. Δεν είναι όμως πάντα επιτυχής και μπορεί να οδηγήσει σε χάσιμο πληροφορίας.

Τέλος, για να κατηγοριοποιηθεί κάθε νέο στιγμιότυπο είναι απαραίτητος ο μετασχηματισμός του ώστε να είναι ίδιας διάστασης με τα μετασχηματισμένα δεδομένα εκπαίδευσης.

Αξιοσημείωτο είναι το γεγονός, ότι αν και οι μέθοδοι δεικτοδότησης πολυδιάστατων δεδομένων μπορούν να επιταχύνουν την αναζήτηση εγγύτερων γειτόνων, αντίθετα με τις τεχνικές μείωσης του όγκου δεδομένων, ωστόσο, δεν μειώνουν τις απαιτήσεις χώρου για την αποθήκευση των δεδομένων. Επομένως, αυτές οι μέθοδοι και κατ' επέκταση ο κατηγοριοποιητής k-NN, δεν βρίσκουν εφαρμογή σε συσκευές με περιορισμό στη μνήμη.

Οι τεχνικές μείωσης του όγκου των δεδομένων διακρίνονται σε δυο μεγάλες κατηγορίες: (i) τεχνικές μείωσης των δεδομένων εκπαίδευσης και (ii) τεχνικές μείωσης των διαστάσεων.

Στόχος των τεχνικών μείωσης των δεδομένων εκπαίδευσης είναι η γρήγορη κατηγοριοποίηση βασισμένη στη μέθοδο εγγύτερων γειτόνων. Ακόμη, στόχος τους είναι η κατασκευή ενός μικρού συνόλου δεδομένων εκπαίδευσης, το οποίο αντιπροσωπεύει το αρχικό μεγάλου όγκου σύνολο όσο το δυνατό περισσότερο. Το μικρό σύνολο που προκύπτει αποτελείται από λίγα και αντιπροσωπευτικά αντικείμενα εκπαίδευσης. Η εφαρμογή της σειριακής αναζήτησης εγγύτερων γειτόνων σε αυτό το μικρό, αντιπροσωπευτικό σύνολο, και όχι στο αρχικό σύνολο, μπορεί να επιτευχθεί χωρίς τη σπατάλη υψηλού υπολογιστικού κόστους.

Αντίθετα, οι τεχνικές μείωσης διαστάσεων μπορούν να χρησιμοποιηθούν ώστε να μειωθεί του κόστος υπολογισμού αποστάσεων. Όπως έχει αναφερθεί, το κόστος υπολογισμού απόστασης εξαρτάται από το πλήθος των χαρακτηριστικών των αντικειμένων. Η μείωση τους ισοδυναμεί με μείωση του κόστους, δηλαδή με την επιτάχυνση την διαδικασία κατηγοριοποίησης.

Άλλα ζητήματα σχετικά με τη χρήση του κατηγοριοποιητή εγγύτερων γειτόνων συνοψίζονται παρακάτω:

- Βέλτιστος αριθμός γειτόνων. Ένα από τα μεγαλύτερα ζητήματα του κατηγοριοποιητή k-NN είναι η επιλογή του βέλτιστου αριθμού γειτόνων που θα ληφθούν υπόψη κατά την κατηγοριοποίηση του νέου αντικειμένου.
- Μεγάλη ευαισθησία του κατηγοριοποιητή στο θόρυβο και στις άνισες κατανομές των δεδομένων εκπαίδευσης στις κλάσεις. Εάν λάβουμε υπόψη δυο κλάσεις την Α και την Β, και η πλειονότητα των δεδομένων εκπαίδευσης επισημαίνεται ως Α, τότε ο k-NN είναι πιθανό να κατηγοριοποιήσει λανθασμένα το νέο αντικείμενο στην κλάση Α.
- Μεγάλη απαίτηση αποθηκευτικού χώρου. Είναι βασική προϋπόθεση για τον κατηγοριοποιητή k-NN ότι τα δεδομένα εκπαίδευσης είναι πάντα στη διάθεσή του γι' αυτό και σε καμία περίπτωση δεν μπορούν να διαγραφούν από τη μνήμη. Αντίθετα, οι πρόθυμοι (eager)

κατηγοριοποιητές φτιάχνουν στην αρχή ένα μοντέλο κατηγοριοποίησης και στη συνέχεια τα δεδομένα εκπαίδευσης μπορούν να διαγραφούν.

1.5 Επιλογή χαρακτηριστικών

Ο αλγόριθμος k-NN υπολογίζει την απόσταση μεταξύ δύο αντικειμένων με βάση όλα τα χαρακτηριστικά. Αυτό μπορεί να μειώσει σημαντικά την ακρίβεια του αλγορίθμου όταν υπάρχουν πολλά χαρακτηριστικά που δεν επηρεάζουν την κατηγορία. Έστω ένα πρόβλημα με 30 χαρακτηριστικά, εκ των οποίων μόνο τα 5 είναι σημαντικά για την πρόβλεψη νέων περιπτώσεων. Τότε στην περίπτωση που ένα αντικείμενο έχει τις ίδιες τιμές στα 5 αυτά χαρακτηριστικά, αλλά διαφορετικές σε όλα τα άλλα, μπορεί να έχει πολύ μεγάλη Ευκλείδεια απόσταση και κατά συνέπεια να κατηγοριοποιηθεί λανθασμένα.

Για να αντιμετωπιστεί αυτό το πρόβλημα έχουν προταθεί μέθοδοι τόσο για τη στάθμιση των χαρακτηριστικών, όσο και για την επιλογή ενός υποσυνόλου χαρακτηριστικών. Και στις δύο περιπτώσεις απαιτείται η εύρεση των σημαντικότερων χαρακτηριστικών για την πρόβλεψη της κατηγορίας. Συχνά κάποια χαρακτηριστικά είναι περισσότερο σημαντικά όσον αφορά μία από τις τιμές μίας διακριτής κατηγορίας και λιγότερο σημαντικά σε σχέση με κάποια άλλη τιμή. Σε αυτήν την περίπτωση απαιτείται ένας διαχωρισμός μεταξύ των σημαντικών χαρακτηριστικών για κάθε διακριτή τιμή της κλάσης. Υπάρχει ένας τομέας της Μηχανικής Μάθησης που ασχολείται με τα παραπάνω ζητήματα ονομάζεται Επιλογή Χαρακτηριστικών (Feature Selection).

1.6 Παραλλαγές του αλγορίθμου kNN

Καθώς η απλή καταμέτρηση των γειτόνων φαίνεται να είναι ανεπαρκής για τον προσδιορισμό της κλάσης ενός στοιχείου δοκιμής[10], έχει γίνει προσπάθεια αλλαγής του κατηγοριοποιητή kNN με τρόπο που θα λαμβανόταν υπόψη ένας άλλος παράγοντας, δηλαδή αυτός της πυκνότητας. Η Δομική Πυκνότητα (ΔΠ) ορίζεται ως ο αριθμός των σημείων στη γειτονία ενός στοιχείου πάνω από τον όγκο αυτής της γειτονιάς. Η παράμετρος που εμπλέκεται είναι αυτή της ακτίνας (r) που ορίζει τη γειτονιά, η οποία προσδιορίζεται ως εξής. Αρχικά, υπολογίζεται η πυκνότητα όλων των στοιχείων ως συνάρτηση του r και τη μέση πυκνότητα ολόκληρου του συνόλου (ως συνολικός αριθμός στοιχείων επί του συνολικού όγκου του συνόλου δεδομένων, κάτι άσχετο με το r). Στη συνέχεια αναζητείται μια τιμή r έτσι ώστε ο μέσος όρος των επιμέρους πυκνοτήτων να είναι ίσος με τη μέση πυκνότητα που υπολογίστηκε νωρίτερα. Ο κατηγοριοποιητής kNN με βάση την πυκνότητα (DB-kNN) έχει δημιουργηθεί λαμβάνοντας υπόψη την έννοια της δοκιμής πυκνότητας για την αξιολόγηση της σημασίας κάθε γείτονα, μαζί με τις αποστάσεις. Αρχικά, οι πυκνότητες όλων των στοιχείων υπολογίζονται για κάθε μία από τις κλάσεις του συνόλου δεδομένων.

Στη συνέχεια, κανονικοποιούνται στο [11] αφού οι σχετικές πυκνότητες φαίνεται να έχουν μεγαλύτερη σημασία από τις απόλυτες. Με βάση αυτές τις πυκνότητες, κάθε γειτονικό στοιχείο αξιολογείται ως προς τον ρόλο του ως στοιχείου «πυρήνα» της κατηγορίας του μετρώντας τη σχετική δομική του πυκνότητα σε σχέση με την κατηγορία. Διαιρώντας το με την Ευκλείδεια απόστασή του προκύπτει μια βαθμολογία για κάθε γείτονα. Λαμβάνοντας έναν προκατειλημμένο μέσο όρο (ο οποίος επηρεάζεται περισσότερο από τους μεγαλύτερους αριθμούς) αυτών των βαθμολογιών για κάθε μια από τις τάξεις, καταλήγουμε σε q βαθμολογίες ψηφοφορίας, όπου q είναι ο αριθμός των τάξεων. Από αυτές τις τελικές βαθμολογίες προκύπτει η κατηγοριοποίηση κάθε σημείου εξέτασης, καθώς και ο Βαθμός Βεβαιότητας. Στην όχι και τόσο συνηθισμένη περίπτωση που ο ΔΠ της κατηγοριοποίησης ενός συγκεκριμένου στοιχείου δοκιμής είναι κάτω από 0,667 (δηλαδή η κατηγοριοποίηση θεωρείται αναξιόπιστη), για αυτήν την περίπτωση εφαρμόζεται η κλασική μέθοδος kNN. Το DB-kNN παρέχει μια νέα ματιά στη χρήση των γειτόνων στο kNN, επειδή διερευνά τις δυνατότητες αξιολόγησής τους αντί απλώς να τους καταμετρήσει, κάτι που, κατά την άποψή μας, είναι ένα σημαντικό βήμα προς τα εμπρός. Επίσης, χρησιμοποιείται η απόσταση, κάνοντας την αξιολόγηση των γειτόνων πιο εκλεπτυσμένη.

Δεδομένου ότι η τιμή της παραμέτρου k συχνά επηρεάζει τα αποτελέσματα της κατηγοριοποίησης, μερικές φορές σημαντικά, επινοήσαμε έναν άλλο αλγόριθμο κατηγοριοποίησης που ξεπερνά αυτό το ζήτημα. Χρησιμοποιώντας την έννοια του DC, υπολογίζεται το βέλτιστο k για κάθε κατηγοριοποίηση. Ο κατηγοριοποιητής Variable kNN (V-kNN) λειτουργεί ως εξής. Πρώτα για κάθε ένα από τα στοιχεία του συνόλου εκπαίδευσης πραγματοποιείται μια κατηγοριοποίησή του με βάση διάφορες γειτονιές. Βρίσκεται η τιμή της k που μεγιστοποιεί το DC κάθε κατηγοριοποιητή. Επομένως, για κάθε σετ εκπαίδευσης αντιστοιχεί μια συγκεκριμένη τιμή K που θεωρείται η καλύτερη διαθέσιμη. Στη συνέχεια, για κάθε άγνωστο στοιχείο, βρίσκεται ο πλησιέστερος γείτονας και υπολογίζεται η τιμή k του (βάσει του «βέλτιστου» k πίνακα). Στη συνέχεια, ο κατηγοριοποιητής kNN εφαρμόζεται σε αυτό το στοιχείο δοκιμής, χρησιμοποιώντας αυτήν την τιμή k. Ως έννοια, αυτή είναι κάτι παρόμοιο με μία από τις ιδέες που παρουσιάζονται στο [12]. Αξίζει να σημειωθεί ότι εκτός από την απόδοσή της ως κατηγοριοποιητή, η προσέγγισή V-kNN αποδίδει ορισμένες χρήσιμες πληροφορίες σχετικά με το σύνολο δεδομένων: τη μέση βέλτιστη τιμή k, την οποία είναι πολύ χρήσιμο να γνωρίζουμε για κατηγοριοποιητές τύπου kNN καθώς βελτιώνει την απόδοσή τους, ιδιαίτερα για τον κλασικό κατηγοριοποιητή kNN. Ωστόσο, για πολύ αραιά σύνολα δεδομένων το βέλτιστο k που βρέθηκε μπορεί να μην είναι έγκυρο και τα αποτελέσματα μπορεί να μην είναι καλύτερα από αυτά του kNN.

Ομοίως με τον κατηγοριοποιητή kNN με βάση την πυκνότητα, ο σταθμισμένος kNN (W-kNN) εκτελεί μια αξιολόγηση αλλά αυτή τη φορά στα χαρακτηριστικά αντί για τα μοτίβα. Κάθε χαρακτηριστικό αξιολογείται και αποδίδεται ένα βάρος με βάση το πόσο χρήσιμο είναι αυτό το χαρακτηριστικό για τη διάκριση των κατηγοριών του συνόλου δεδομένων. Για να γίνει αυτό, εισάγεται εδώ μια νέα έννοια, δηλαδή αυτή του Δείκτη Διακρίσεως. Ο Δείκτης Διακρίσεως (ΔΔ) είναι ένα μέτρο που αναπτύχθηκε για την αξιολόγηση του πόσο εύκολα διακρίνονται οι κατηγορίες ενός συνόλου δεδομένων. Αρχικά

διακρίναμε κλάσεις χρησιμοποιώντας κουτιά που τις περιείχαν, αλλά αυτό φαινόταν να μην είναι ευαίσθητο στις δομές της τάξης και να μην είναι υπολογιστικά αποτελεσματικό. Σε αυτήν την έκδοση του Δείκτη Διακρίσεως, χρησιμοποιούμε (υπερ)σφαίρες. Αυτή η προσπάθεια υποθέτει μια σταθερή ακτίνα γύρω από κάθε στοιχείο του συνόλου δεδομένων, η οποία αντιστοιχεί στη μέση απόσταση μεταξύ αυτού και των υπολοίπων στοιχείων αυτής της κλάσης. Πρέπει να σημειωθεί ότι η ακτίνα εξαρτάται από τη δομή της κλάσης, επομένως τα στοιχεία που ανήκουν σε διαφορετικές κλάσεις μπορεί να έχουν διαφορετικές ακτίνες. Μόλις καθοριστεί η ακτίνα ενός στοιχείου, αναγνωρίζονται και καταμετρώνται στοιχεία της ίδιας κατηγορίας με το εξεταζόμενο στοιχείο που ανήκουν στην (υπερ)σφαίρα του. Η ευκρίνεια αυτού του στοιχείου υπολογίζεται διαιρώντας τον αριθμό αυτών των στοιχείων με τον αριθμό των συνολικών στοιχείων στην (υπερ)σφαίρα. Ο Δείκτης Διακρίσεως ολόκληρου του συνόλου δεδομένων υπολογίζεται ως ο αριθμός των στοιχείων με διακριτικότητα μεγαλύτερη από 0,5 διαιρεμένος με τον συνολικό αριθμό στοιχείων. Ο δείκτης διακριτικότητας χρησιμοποιείται επίσης για την αξιολόγηση μεμονωμένων χαρακτηριστικών με την εφαρμογή του σε μεμονωμένες διαστάσεις του συνόλου δεδομένων.

Ο πιο σημαντικός είναι ο σταθμισμένος κανόνας k-NN[17] που χρησιμοποιεί μια συνάρτηση απόστασης-βάρους για να ζυγίζει περισσότερο τους πιο κοντινούς γείτονες από τους παραπέρα. Ο πλησιέστερος γείτονας σταθμίζεται με ένα ενώ ο πιο απομακρυσμένος γείτονας σταθμίζεται με μηδέν. Τα βάρη όλων των άλλων γειτόνων κλιμακώνονται σε αυτό το διάστημα. Ένα νέο στοιχείο κατηγοριοποιείται με σταθμισμένη πλειοψηφία: κατατάσσεται στην κατηγορία με το μεγαλύτερο άθροισμα βαρών.

Ο κατηγοριοποιητής W-kNN λειτουργεί ως εξής. Πρώτα, κάθε ένα από τα χαρακτηριστικά του σετ εκπαίδευσης αξιολογείται χρησιμοποιώντας το αναγνωριστικό. Στη συνέχεια, τα βάρη λαμβάνονται με κανονικοποίηση των αναγνωριστικών. Τέλος, τα βάρη εφαρμόζονται τόσο στο σύνολο εκπαίδευσης όσο και στο σύνολο δοκιμών και ο κατηγοριοποιητής kNN λειτουργεί στο τώρα μετασχηματισμένο σύνολο δεδομένων.

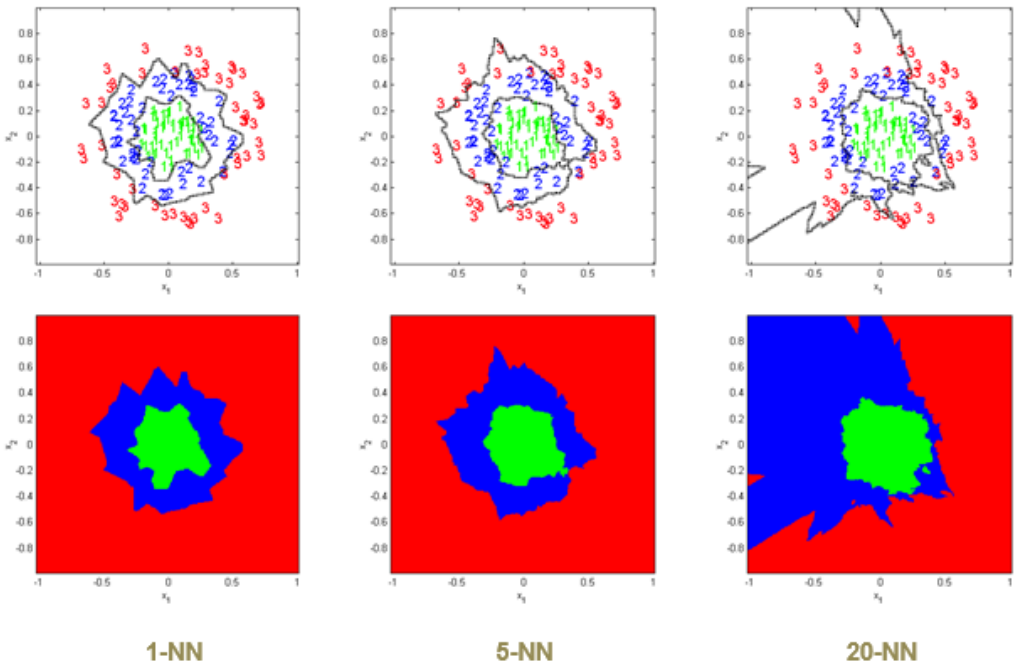
Το Class Based kNN (CB-kNN) είναι κάπως διαφορετικό ως επέκταση kNN. Αναπτύχθηκε επειδή συχνά τα σύνολα δεδομένων δεν είναι ισορροπημένα ως προς τη δομή της κλάσης τους, επομένως μπορεί να συμβαίνει ότι μια τάξη έχει πολύ λίγα στοιχεία για να «κερδίσει» την ψήφο μιας κατηγοριοποίησης του κατηγοριοποιητή kNN. Ο αλγόριθμος CB-kNN ασχολείται με αυτά τα σύνολα δεδομένων λειτουργώντας με τον ακόλουθο τρόπο. Για κάθε στοιχείο δοκιμής λαμβάνονται τα k πλησιέστερα στοιχεία κάθε κλάσης. Η τιμή του k επιλέγεται αυτόματα από τον κατηγοριοποιητή, έτσι ώστε να μεγιστοποιηθεί ο ΔΠ της κατηγοριοποίησης. Στη συνέχεια υπολογίζεται ο αρμονικός μέσος όρος των αποστάσεων αυτών των γειτόνων (ώστε να μην επηρεάζεται τόσο από τα πιο απομακρυσμένα στοιχεία). Τέλος, αυτά τα μέσα συγκρίνονται και επιλέγεται για την κατηγοριοποίηση η κλάση που δίνει τη χαμηλότερη τιμή.

1.7 Η παράμετρος k

Η επιλογή της τιμής για την παράμετρο k αποτελεί πρόκληση για τον κατηγοροποιητή k-NN, καθώς διαφορετικές τιμές μπορούν να οδηγήσουν σε διαφορετικά αποτελέσματα κατηγοριοποίησης. Το k παίζει αρκετά σημαντικό ρόλο στην αποδοτικότητα του κατηγοριοποιητή και είναι δύσκολο να προσδιοριστεί. Αν η τιμή του είναι μικρή, το αποτέλεσμα μπορεί να είναι ευαίσθητο σε θορυβώδη δεδομένα. Ως θορυβώδη ορίζονται τα δεδομένα όπου τα σύνορα μεταξύ των κλάσεων δεν είναι διακριτά. Αντίθετα, αν η τιμή του k είναι πολύ μεγάλη, το αποτέλεσμα των κοντινότερων γειτόνων μπορεί να περιέχει πολλά αντικείμενα από άλλες κατηγορίες. Το k μπορεί να οριστεί χρησιμοποιώντας διάφορες τεχνικές.

Μια ειδική περίπτωση του k-NN κατηγοριοποιητή, που χρησιμοποιείται σε πολλά ερευνητικά πεδία είναι με σταθερό k, και συγκεκριμένα με k=1. Τα μειονεκτήματα που έχει έναντι των περιπτώσεων με μεγάλο k, είναι ότι μεγάλο k σημαίνει πιο ομαλές περιοχές αποφάσεων και δίνει πιο σωστές πιθανοτικά πληροφορίες. Ωστόσο, το πολύ μεγάλο k μπορεί να χαλάσει την τοπικότητα της απόφασης και αυξάνει το υπολογιστικό κόστος. Στο παρακάτω σχήμα είναι ένα χαρακτηριστικό παράδειγμα όπου το k=1 δίνει τα καλύτερα δυνατά αποτελέσματα. Όσο αυξάνεται το k τόσο χαλάει η τοπικότητα της απόφασης.

Στα παραδείγματα έχουμε θεωρήσει ως απόσταση μεταξύ της νέας περίπτωσης και των δεδομένων εκπαίδευσης την Ευκλείδεια απόσταση. Η Ευκλείδεια απόσταση και άλλα συχνά χρησιμοποιούμενα μέτρα απόστασης σχολιάζονται στην επόμενη υποενότητα.



Εικόνα 2: Η τοπικότητα της απόφασης για κ=1, κ=5 και κ=20

Συνεπώς, η απόδοση κατηγοριοποίησης εξαρτάται σίγουρα από την επιλογή της τιμής της παραμέτρου k. Η τιμή του k που επιτυγχάνει την υψηλότερη ακρίβεια κατηγοριοποίησης εξαρτάται από το σύνολο δεδομένων που χρησιμοποιείται και ο προσδιορισμός του συνήθως συνεπάγεται με διαδικασίες tuning μέσω δαπανηρών εργασιών προ επεξεργασίας δοκιμής και σφάλματος. Ο προσδιορισμός του k δεν μπορεί να ακολουθήσει κανένα γενικό κανόνα και το «καλύτερο» k μπορεί να είναι εντελώς διαφορετικό για διαφορετικά σύνολα δεδομένων. Όπως αναφέρθηκε και προηγουμένως, μεγαλύτερες τιμές k είναι κατάλληλες για σύνολα δεδομένων με θόρυβο, καθώς εξετάζουν μεγαλύτερες γειτονιές. Ωστόσο, δεν ορίζουν ξεκάθαρα τα όρια μεταξύ διαφορετικών τάξεων.

Αντίθετα, μικρές τιμές παραμέτρων καθιστούν τον κατηγοριοποιητή πιο ευαίσθητο στον θόρυβο. Επομένως, σε περιπτώσεις δεδομένων εκπαίδευσης που περιέχουν θόρυβο, η κατηγοριοποίηση είναι πιθανώς λιγότερο ακριβής. Αξίζει να αναφέρουμε ότι ακόμη και η καλύτερη τιμή k μπορεί να μην είναι βέλτιστη. Αυτό συμβαίνει επειδή ο κατηγοριοποιητής k-NN χρησιμοποιεί μια μοναδική τιμή k. Διαφορετικές τιμές k μπορεί να είναι βέλτιστες για διαφορετικές περιοχές του χώρου δεδομένων.

Συνεπώς, μπορούν να υιοθετηθούν ευρετικές μέθοδοι για δυναμικό προσδιορισμό του k[13] που μπορούν να επιτύχουν μεγαλύτερη ακρίβεια από τον κατηγοριοποιητή k-NN με τον «καλύτερο» προσδιορισμό της τιμής k. Σε περιπτώσεις προβλημάτων δυαδικής κατηγοριοποίησης (σύνολα δεδομένων με δύο κλάσεις), το k θα πρέπει να έχει μια περιττή τιμή για να αποφευχθούν οι ισοψηφίες (και οι δύο κατηγορίες είναι οι πιο συνηθισμένες) κατά την ψηφοφορία των πλησιέστερων γειτόνων. Σε περιπτώσεις μη δυαδικών προβλημάτων, το k μπορεί να έχει οποιαδήποτε τιμή. Εδώ, οι πιθανές

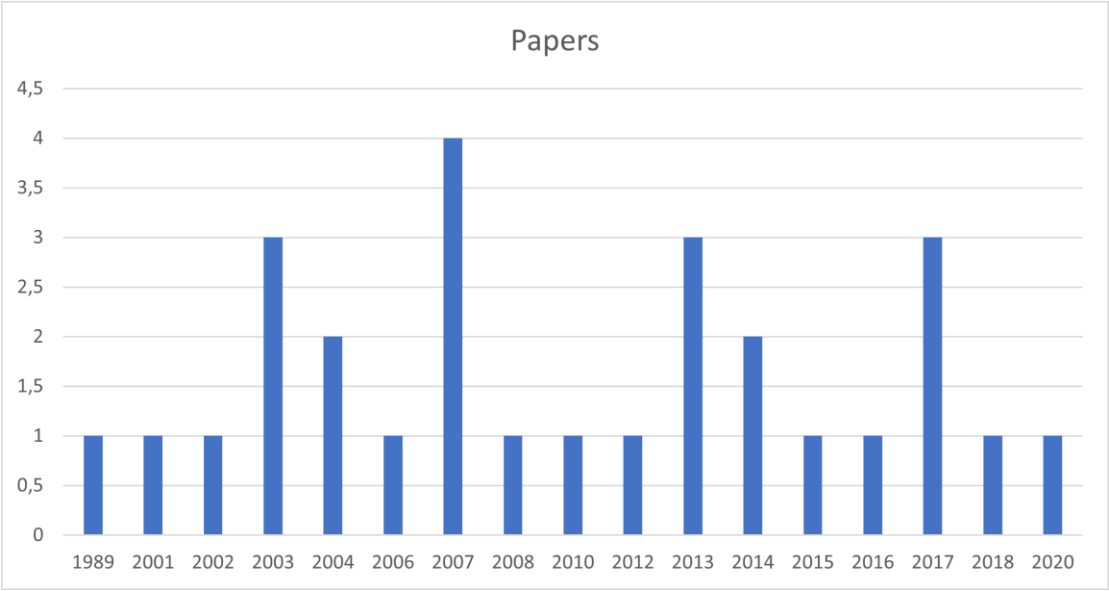
ισοψηφίες κατά τη διάρκεια της ψηφοφορίας επιλύονται επιλέγοντας είτε μια τυχαία κατηγορία «πιο συνηθισμένη» ή την τάξη του πλησιέστερου γείτονα.

Το δημοφιλές λογισμικό Weka[14] και πολλά άλλα εργαλεία λογισμικού εξόρυξης δεδομένων / μηχανικής μάθησης επιλύουν τυχαία τις ισοψηφίες.

1.8 Κίνητρο και Συνεισφορά

Αν και υπάρχουν πολλά άρθρα σχετικά με τον αλγόριθμο κ εγγύτερων γειτόνων, ωστόσο δεν υπάρχει αντίστοιχη έρευνα η οποία να περιλαμβάνει όλες τις πληροφορίες συγκεντρωμένες. Αυτό αποτελεί το κίνητρο για την εκπόνηση της παρούσας εργασίας.

Συνεισφορά της εργασίας αποτελεί μια εκτεταμένη μελέτης και ανασκόπησης για καταγραφή όλων όσων έχουν γραφτεί κατά τη χρονική περίοδο 1986 έως 2020. Παρατηρούμε ότι οι περισσότερες μελέτες έχουν καταγραφεί τις χρονικές περιόδους 2003-2007 και 2013-2017 και οι λιγότερες την τελευταία τριετία οι οποίες είναι μόνο 2.



Εικόνα 3: Μελέτες που έχουν γίνει από το 1989 έως και το 2020

1.9 Οργάνωση της πτυχιακής

Στο κεφάλαιο 2 παρουσιάζονται οι μέθοδοι ορισμού του “στατικό” k στον κατηγοριοποιητή των k εγγύτερων γειτόνων. Με άλλα λόγια, οι τεχνικές αυτές αναζητούν και βρίσουν μια καλή τιμή για τη μεταβλητή k η οποία στη συνέχεια χρησιμοποιείται για την κατηγοριοποίηση όλων των δεδομένων που πρέπει να κατηγοριοποιηθούν.

Στο Κεφάλαιο 3 παρουσιάζονται οι αλγόριθμοι που αφορούν το δυναμικό k. Αυτές οι τεχνικές χρησιμοποιούν διαφορετικό k για κάθε στοιχείο που πρέπει να κατηγοριοποιηθεί. Με άλλα λόγια, κάθε ένα στοιχείο κατηγοριοποιείται με το “δικό” του k.

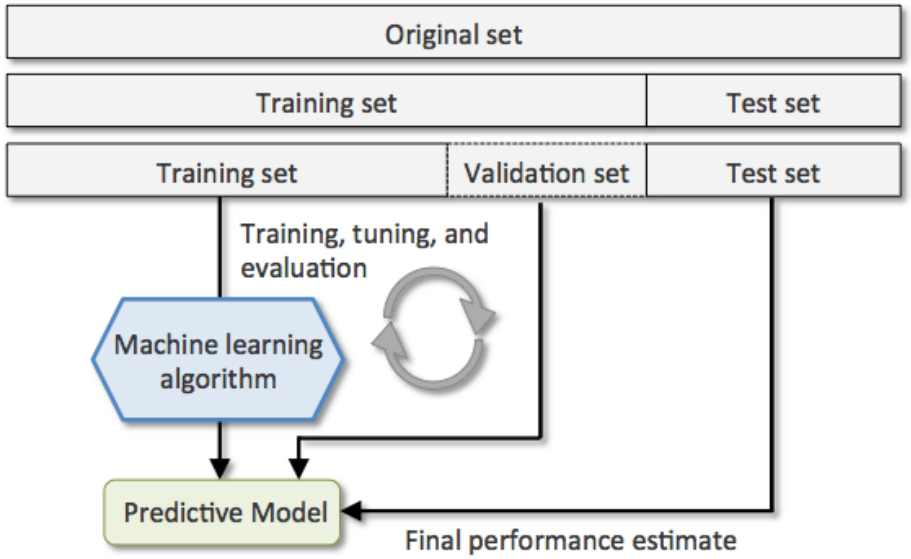
Στο Κεφάλαιο 4 καταγράφονται κάποια συμπεράσματα της βιβλιογραφικής έρευνας και δίνονται κάποιες κατευθύνσεις για μελλοντική έρευνα.

Κεφάλαιο 2ο: Μέθοδοι ορισμού του σταθερού k

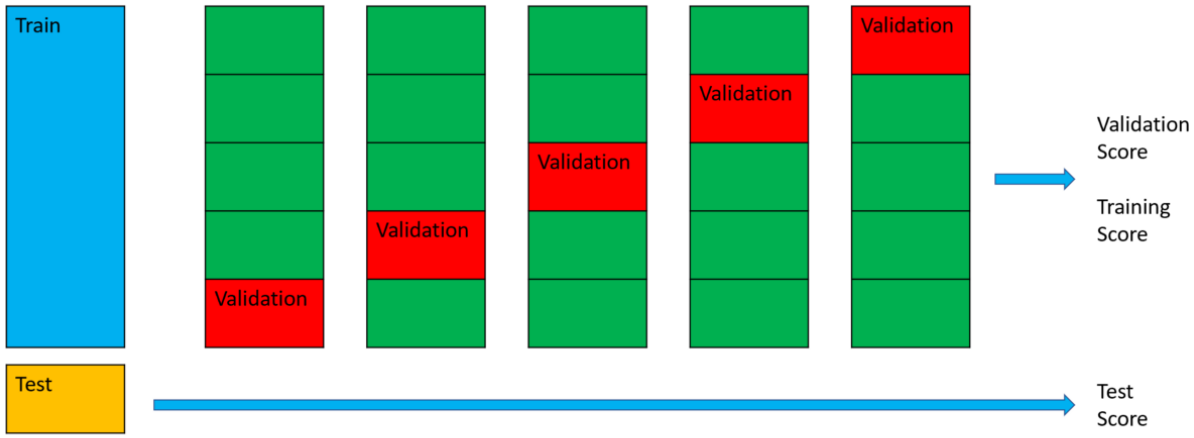
Ο κατηγοριοποιητής k-NN προβλέπει την ετικέτα κλάσης ενός άγνωστου στιγμιότυπου πραγματοποιώντας μια τοπική αναζήτηση μεταξύ των k πλησιέστερων γειτόνων του και στη συνέχεια, εφαρμόζοντας πλειοψηφία, το άγνωστο στιγμιότυπο επισημαίνεται με την κλάση της πλειοψηφίας των k πλησιέστερων γειτόνων. Ένα μειονέκτημα του k-NN είναι ότι η απόδοση εξαρτάται από την τιμή k, η οποία καθορίζει την έκταση της γειτονιάς στην οποία πραγματοποιείται η αναζήτηση. Τα τελευταία χρόνια έχουν γίνει πολλές έρευνες για την αντιμετώπιση των προαναφερθέντων μειονεκτημάτων. Στην πράξη, το k αντιμετωπίζεται ως υπερ-παράμετρος και, κατά συνέπεια, η πιο συχνά χρησιμοποιούμενη τεχνική για τον προσδιορισμό του «καλύτερου» k είναι η διασταυρούμενη επικύρωση (cross validation), καθώς δεν υπάρχει γενικός κανόνας για την επιλογή της τιμής k λόγω της εξάρτησής της από το σύνολο δεδομένων εκπαίδευσης. Παρακάτω παρουσιάζονται οι τεχνικές εύρεσης του καλύτερου στατικού κ για τον κατηγοριοποιητή.

2.1 Διασταυρούμενη επικύρωση (cross validation)

Η τεχνική διασταυρούμενης επικύρωσης για την απόδοση της βέλτιστης τιμής στην παράμετρο κ, υλοποιείται ως εξής: Αρχικά το σύνολο δεδομένων χωρίζεται δύο σύνολα: (α) σε σύνολο δεδομένων εκπαίδευσης και (β) σε σύνολο δεδομένων δοκιμής. Στη συνέχεια, το σύνολο δεδομένων εκπαίδευσης, χωρίζεται σε έναν συγκεκριμένο αριθμό τμημάτων, έστω χ (συνήθως 5 ή 10). Κάθε φορά, ο κατηγοριοποιητής χρησιμοποιεί διαφορετική τιμή για το k και η διαδικασία της κατηγοριοποίησης εκτελείται χ φορές με την ίδια τιμή k όπου κάθε φορά η αναζήτηση των εγγύτερων γειτόνων γίνεται στην ένωση των χ-1 τμημάτων. Το ένα τμήμα που δεν συμπεριλαμβάνεται, το οποίο ονομάζεται validation set, περιέχει τα δεδομένα που κατηγοριοποιούνται. Σε κάθε επανάληψη, χρησιμοποιείται άλλο τμήμα ως validation set και η ένωση των υπολοίπων τμημάτων ως σύνολο δεδομένων εκπαίδευσης. Αυτή η διαδικασία επαναλαμβάνεται πολλές φορές και τελικά, η τιμή k ορίζεται στο να είναι η τιμή που πέτυχε την υψηλότερη ακρίβεια σε μια από αυτές τις επαναλαμβανόμενες επαναλήψεις. Στο τέλος, η εκτίμηση της απόδοσης του κατηγοριοποιητή, πραγματοποιείται βάσει του συνόλου δεδομένου δοκιμής που δεν συμμετείχε στην διασταυρωμένη επικύρωση. Τα σχήματα που παρουσιάζονται παρακάτω, περιγράφουν με γραφικό τρόπο της διαδικασία απόδοσης της βέλτιστης τιμής στην παράμετρο κ.



Εικόνα 4: Διαδικασία απόδοσης βέλτιστης τιμής στην παράμετρο



Εικόνα 5: Τα τμήματα που χωρίζεται το σύνολο δεδομένων κατά τη διαδικασία cross validation

Η τιμή k που τελικά προσδιορίζεται με τη χρήση διασταυρούμενης επικύρωσης είναι μια μοναδική και σταθερή τιμή για ολόκληρο το σύνολο δεδομένων χωρίς να λαμβάνονται υπόψη τα συγκεκριμένα και μοναδικά χαρακτηριστικά που μπορεί να έχει κάθε σύνολο δεδομένων καθώς και η κατανομή των δεδομένων στους υποχώρους.

Για παράδειγμα, σε περιπτώσεις είτε κλάσεων που δεν είναι καλά διαχωρισμένες είτε θορυβωδών περιπτώσεων, μια μεγάλη τιμή k μπορεί να είναι πιο κατάλληλη για να εξεταστεί ένας εκτεταμένος υποχώρος (γειτονιά). Μια μεγάλη τιμή k έχει ως αποτέλεσμα έναν κατηγοριοποιητή με ανοχή θορύβου, καθώς η περιοχή αναζήτησής του είναι μεγάλη. Αντίθετα, σε περίπτωση διακριτών κλάσεων, μια μεγάλη τιμή k μπορεί να οδηγήσει τόσο σε υψηλότερο υπολογιστικό κόστος όσο και σε υποβάθμιση

της ακρίβειας. Σε τέτοιες περιπτώσεις, μια μικρή τιμή k μπορεί να είναι πιο κατάλληλη. Μια μικρή τιμή k έχει ως αποτέλεσμα έναν ευαίσθητο στο θόρυβο κατηγοριοποιητή, καθώς η περιοχή αναζήτησης είναι περιορισμένη. Ωστόσο, το πρόβλημα γίνεται όλο και πιο περίπλοκο σε περιπτώσεις συνόλων δεδομένων πραγματικής ζωής που μπορεί να περιέχουν ταυτόχρονα καλά και όχι καλά διαχωρισμένες τάξεις, ανισόρροπες τάξεις, αραιές και πυκνές γειτονιές και θορυβώδεις υποχώρους. Με βάση τα παραπάνω, θα πρέπει να θεωρηθεί ότι μια καθολικά καθορισμένη σταθερή τιμή k δεν είναι κατάλληλη για ένα σύνολο δεδομένων. Αντίθετα, θα πρέπει να λάβει κανείς υπόψη τα ειδικά χαρακτηριστικά του συνόλου δεδομένων και τον υποχώρο στον οποίο βρίσκεται κάθε στιγμιότυπο και να προσπαθήσει να προσδιορίσει δυναμικά μια τοπική τιμή k για κάθε στιγμιότυπο που θα κατηγοριοποιηθεί.

2.2 Ευρετική επιλογή της παραμέτρου k

Μια ευρετική παραλλαγή k-NN παρουσιάζεται στο [20], όπου η τιμή k επιλέγεται αυτόματα, χωρίς καμία παρέμβαση του χρήστη. Η ευρετική βασίζεται στην ιδέα ότι ο αλγόριθμος θα αναζητήσει αυτήν την τιμή k που κατηγοριοποιεί σωστά την πλειονότητα των περιπτώσεων εκπαίδευσης. Η προτεινόμενη προσέγγιση συγκρίθηκε, όσον αφορά την ακρίβεια, με το συμβατικό k-NN, με $k \in [1, 51]$ (μόνο περιττοί αριθμοί) σε 25 σύνολα δεδομένων. Σε δεκατρείς από τις είκοσι περιπτώσεις, ο προτεινόμενος αλγόριθμος ξεπέρασε τις επιδόσεις του ευρέως χρησιμοποιούμενου 1-NN. Επιπλέον, σε πέντε από τις δεκατρείς παραπάνω περιπτώσεις, η διαφορά ήταν στατιστικά σημαντική.

2.3 Συμμετρικός k-NN

Παρά το γεγονός ότι η προσέγγιση που παρουσιάζεται στο [21] δεν προτείνει κάποια διαφορετική διαδικασία για την επιλογή k, η περιγραφόμενη ευρετική καθιστά την απόδοση του συμβατικού k-NN λιγότερο εξαρτώμενη από την επιλεγμένη τιμή k. Εν συντομία, η ιδέα πίσω από αυτή την προσέγγιση είναι ότι (για ένα δεδομένο k) αν το x ψηφίζει για το y, τότε το y ψηφίζει επίσης για το x, ακόμα κι αν το x δεν ανήκει στους y s k πλησιέστερους γείτονες. Ο προτεινόμενος αλγόριθμος συγκρίθηκε, ως προς την ακρίβεια, με άλλους αλγορίθμους που βασίζονται στην εξέταση εγγύτερων γειτόνων, σε 29 σύνολα δεδομένων. Τα αποτελέσματα έδειξαν ότι ο προτεινόμενος αλγόριθμος ξεπέρασε τους υπόλοιπους χωρίς στατιστική διαφορά.

2.4 Πιθανολογική μέθοδος k-NN

Ανεξάρτητα από το πώς επιλέγεται το k, οι προβλέψεις που γίνονται από τον αλγόριθμο των k εγγύτερων γειτόνων δεν έχουν πιθανολογική ερμηνεία. Οι συγγραφείς στο [22] παρουσίασαν μια αναλυτική πιθανολογική παραλλαγή k-NN που ασχολείται με την αβεβαιότητα στο k χρησιμοποιώντας

μια προηγούμενη κατανομή σε αυτό. Η προτεινόμενη προσέγγιση συγκρίθηκε, όσον αφορά το ποσοστό σφάλματος κατηγοριοποίησης σε μια ποικιλία συνόλων δεδομένων, με τον συμβατικό αλγόριθμο k-NN, επιδεικνύοντας ανταγωνιστική απόδοση. Η προσέγγιση που παρουσιάζεται δεν προτείνει μια ποικίλη διαδικασία για την επιλογή k αλλά μια ταχύτερη τεχνική διασταυρούμενης επικύρωσης προκειμένου να εξεταστεί μεγαλύτερος αριθμός τιμών k εντός του ίδιου χρόνου εκτέλεσης, μειώνοντας τη χρονική πολυπλοκότητα κατά O(K∗), όπου K∗ είναι η μέγιστη τιμή k. Η προτεινόμενη τεχνική δοκιμάστηκε σε 3 σύνολα δεδομένων αποδεικνύοντας τη συμβολή της στη μείωση του χρόνου εκτέλεσης.

2.5 Η παράμετρος λείανσης και η αποτελεσματικότητα του k-NN

Μια αρκετά παλαιότερη εμπειρική προσέγγιση παρουσιάζεται στο [23]. Οι Fix και Hodges [34] εισήγαγαν μια νέα προσέγγιση στη μη παραμετρική ταξινόμηση βασιζόμενοι στην «απόσταση» μεταξύ σημείων ή κατανομών. Οι συγγραφείς υποστηρίζουν ότι η βέλτιστη τιμή k εξαρτάται από τις διαστάσεις, το μέγεθος και τη δομή του μεγέθους του δείγματος. Επιπλέον, προτείνουν μερικές εξισώσεις για τον υπολογισμό k σε συνάρτηση με τη διαφορά μεταξύ των αναλογιών του δείγματος και τη διαφορά μεταξύ των πινάκων συνδιακύμανσης.

Κεφάλαιο 3ο: Αλγόριθμοι που αποδίδουν δυναμικό k

3.1 Adaptive k-Nearest-Neighbor Classification Using a Dynamic Number of Nearest Neighbors

Το κίνητρο σε αυτή την εργασία [13] είναι το να μειωθεί το υπολογιστικό κόστος ανάκτησης των κ εγγύτερων γειτόνων όταν η τιμή του κ είναι μεγάλη. Έτσι προτείνονται τεχνικές όπου βοηθούν στη μείωση του υπολογιστικού κόστους χωρίς όμως να επηρεάζεται η ακρίβεια. Αυτό πραγματοποιείται επιλέγοντας τοπικά την καλύτερη τιμή για το k και όχι μία σταθερή τιμή για όλο το set.

Ο επαυξητικός (incremental) αλγόριθμος ανάκτησης των γειτόνων ενός στιγμιότυπου χρησιμοποιεί την δομή δεδομένων R-Tree και υπολογίζει τις αποστάσεις μόνο μεταξύ του νέου στιγμιότυπου και των εγγύτερων γειτόνων σε αυτό και όχι ολόκληρου του data set. Συγκεκριμένα, ο εν λόγω αλγόριθμος ανακτά τους εγγύτερους γείτονες με την σειρά διασχίζοντας την δομή r-tree. Πρώτα, ανακτάται ο πρώτος εγγύτερος γείτονας, στην συνέχεια ο δεύτερος κ.ο.κ. Φυσικά, όσοι περισσότεροι γείτονες ανακτηθούν τόσο αυξάνεται το υπολογιστικό κόστος.

Η προτεινόμενη μέθοδος χρησιμοποιεί τον επαυξητικό αλγόριθμο ανάκτησης εγγύτερων γειτόνων και τη δομή R-Tree. Επιπρόσθετα, εισάγει κάποιες μεθόδους τύπου heuristic που διακόπτουν (early break) τη διαδικασία αναζήτησης και ανάκτησης εγγύτερων γειτόνων τη στιγμή που οι γείτονες που έχουν ανακτηθεί, ικανοποιούν τα κριτήρια που ορίζουν οι heuristics. Για παράδειγμα, έστω ότι η καλύτερη τιμή που ορίστηκε για το k σύμφωνα με τη μέθοδο cross-validation είναι k=25, τότε, χρησιμοποιείται το r-tree και ο επαυξητικός αλγόριθμος και υπολογίζονται αυτές οι 25 αποστάσεις. Μπορεί όμως το στιγμιότυπο προς κατηγοριοποίηση να κατηγοριοποιηθεί εξετάζοντας λιγότερους από 25 γείτονες. Σε αυτή την περίπτωση θα μειωθεί το υπολογιστικό κόστος και γίνεται η χρήση μιας από τις προτεινόμενες heuristics. Με άλλα λόγια, αν η καλύτερη δυνατή τιμή για την παράμετρο k είναι ένας μεγάλος αριθμός, η διαδικασία αναζήτησης των εγγύτερων γειτόνων διακόπτεται πιο πριν την ανάκτηση και των k εγγύτερων γειτόνων αν τα κριτήρια που ορίζει η heuristic που χρησιμοποιείται ικανοποιούνται από τους γείτονες που έχουν ανακτηθεί μέχρι αυτή τη στιγμή. Με αυτό το τρόπο δεν ξοδεύεται το υπολογιστικό κόστος που απαιτείται για την ανάκτηση των υπόλοιπων εγγύτερων γειτόνων.

Στόχος της προτεινομένης μεθοδολογίας είναι το να εξεταστούν λιγότεροι γείτονες από αυτούς που εξετάζονται από τον κατηγοριοποιητή k-NN με την απόδοση της "καλύτερης" τιμής στη μεταβλητή k, μειώνοντας το υπολογιστικό κόστος και διατηρώντας την ακρίβεια σε υψηλά επίπεδα. Έτσι, κάθε στιγμιότυπο που πρέπει να κατηγοριοποιηθεί, κατηγοριοποιείται με ένα διαφορετικό αριθμό εγγύτερων γειτόνων. Ο αριθμός αυτός εξαρτάται από την κατανομή των στιγμιοτύπων σε κλάσεις στην υπο-περιοχή που βρίσκεται το υπό κατηγοριοποίηση στιγμιότυπο.

Τρία είναι τα heuristics που προτείνονται από την εργασία για τη διακοπή της διαδικασίας αναζήτησης και ανάκτησης των k εγγύτερων γειτόνων. Η απόδοση της κατηγοριοποίησης εξαρτάται από τις διάφορες παραμέτρους των heuristics. Η παράμετρος MinNN είναι κοινή σε όλα τα heuristics και καθορίζει τον μικρότερο αριθμό των εγγύτερων γειτόνων οι οποίοι πρέπει να χρησιμοποιηθούν στην κατηγοριοποίηση. Μετά από αυτόν το αριθμό ενεργοποιείται ο έλεγχος για διακοπή της διαδικασίας ανάκτησης γειτόνων (early break).

Στο πρώτο heuristic (Simple Heuristic SH) υπάρχει μία παράμετρος η οποία είναι ένα ποσοστό. Έστω ότι μια δεδομένη χρονική στιγμή έχει ανακτηθεί ένας αριθμός γειτόνων που είναι μικρότερος από k. Αν το ποσοστό των γειτόνων που "ψηφίζουν" την πλειοψηφούσα κλάση είναι μεγαλύτερο από αυτό το ποσοστό, η διαδικασία αναζήτησης και ανάκτησης των επιπλέον εγγύτερων γειτόνων δεν συνεχίζει (early break). Πιο συγκεκριμένα, έστω ότι η καλύτερη ακρίβεια επιτυγχάνεται εξετάζοντας 100 εγγύτερους γείτονες. Δίνονται τις τιμές στις παραμέτρου του heuristic PMaj=0.9 και MinNN=7, τότε ο incremental αλγόριθμος ξεκινάει να ελέγχει για early break μετά την ανάκτηση του 7ου εγγύτερου γείτονα. Αν το 90% των εγγύτερων γειτόνων που έχουν ανακτηθεί μέχρι μια δεδομένη στιγμή "ψηφίζουν" στο να κατηγοριοποιηθεί το υπο κατηγοριοποίηση στιγμιότυπο σε μια συγκεκριμένη κλάση, τότε η διαδικασία αναζήτησης εγγύτερων γειτόνων ολοκληρώνεται. Αν αυτό το ποσοστό επιτευχθεί κατά την εξέταση του 10ου γείτονα τότε δεν είναι απαραίτητος ο υπολογισμός των υπόλοιπων 90 γειτόνων.

Στο δεύτερο heuristic (Independent Class Heuristic ICH) εάν υποθέσουμε ότι το σύνολο δεδομένων περιέχει αντικείμενα 5 κλάσεων, η παράμετρος IndFactor ορίζεται ίση με 1 και ο αλγόριθμος έχει καθορίσει 100 εγγύτερους γείτονες, τότε ο επαυξητικός αλγόριθμος θα σταματήσει εάν 51 από τους εγγύτερους γείτονες κατηγοριοποιηθούν σε συγκεκριμένη κλάση και οι υπόλοιποι 49 σε άλλες. Αν η τιμή της παραμέτρου IndFactor οριστεί ίση με 2, τότε θα εκτελεστεί early-break όταν η πλειοψηφούσα κλάση έχει περισσότερα από 66 στιγμιότυπα. Η τιμή της παραμέτρου IndFactor θα πρέπει να ορίζεται λαμβάνοντας υπόψη τον αριθμό των κλάσεων και την κατανομή των κλάσεων. Σε μία κανονική κατανομή η τιμή της παραμέτρου θα πρέπει να είναι μεγάλη όταν ο αριθμός των κλάσεων είναι μικρός ενώ αντίθετα, η τιμή της παραμέτρου θα πρέπει να είναι μικρή όταν οι κλάσεις είναι πολλές.

Το τρίτο heuristic (M-times Major Classes Heuristic MMCH) σταματάει τον επαυξητικό αλγόριθμο όταν βρεθούν M συνεχόμενοι εγγύτεροι γείτονες οι οποίοι ανήκουν στην πλειοψηφούσα κλάση. Πιο συγκεκριμένα, όταν το ποσοστό των εγγύτερων γειτόνων που ανήκουν στην πλειοψηφούσα κλάση είναι μεγαλύτερο από την τιμή της παραμέτρου PMaj και υπάρχει μια ακολουθία από M εγγύτερους γείτονες που ανήκουν στην πλειοψηφούσα κλάση τότε ο αλγόριθμος σταματάει.

Η προτεινόμενη μέθοδος και τα τρία heuristics εφαρμόστηκαν σε 2 data sets. Τα αποτελέσματα των πειραμάτων έδειξαν πως η συγκεκριμένη μέθοδος μπορεί να επιτύχει σημαντική βελτίωση της απόδοσης χωρίς να μειωθεί η ακρίβεια. Σε μερικές περιπτώσεις παρατηρήθηκε ακόμη καλύτερη

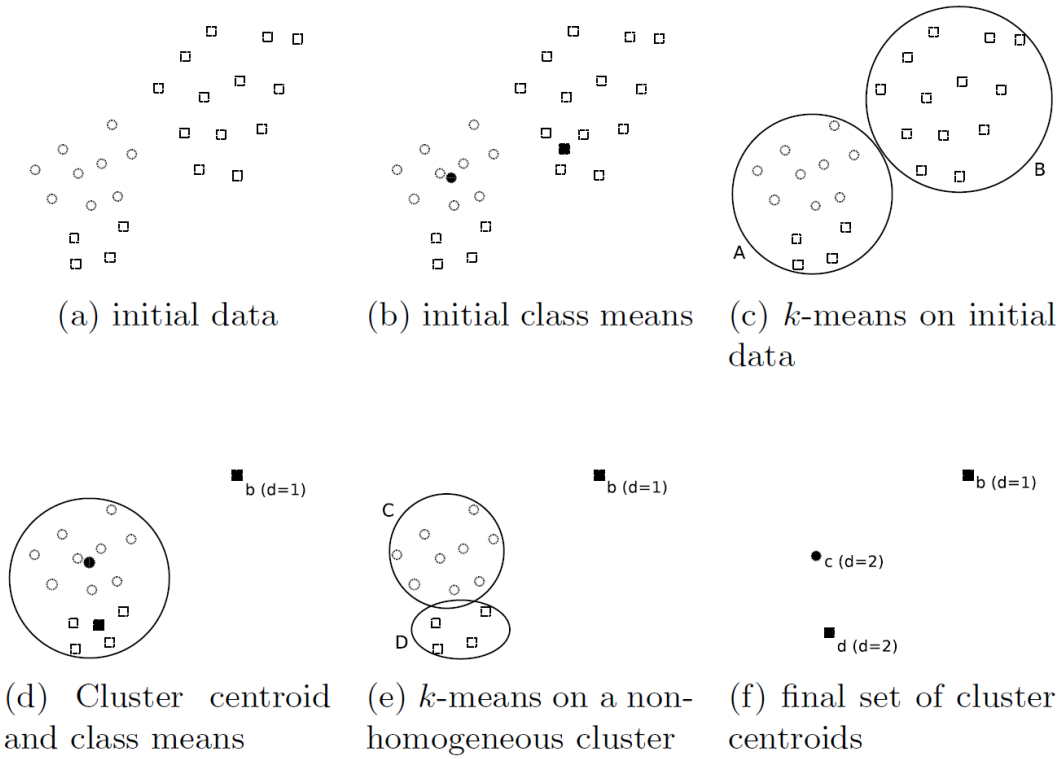
ακρίβεια σε σύγκριση με αυτή του k-NN κατηγοριοποιητή όπου το k έχει οριστεί να είναι αυτό που επιτυγχάνει την υψηλότερη δυνατή ακρίβεια.

3.2 Dynamic k-NN Classification based on Subspace Homogeneity

Γενικά, για σύνολα δεδομένων εκπαίδευσης τα οποία δεν είναι καλά διαχωρισμένα και περιέχουν θόρυβο, στην τιμή k αποδίδονται μεγάλες τιμές ώστε να εξεταστούν και μεγαλύτερες γειτονιές. Αντίθετα, σε σύνολα δεδομένων τα οποία είναι καλά διαχωρισμένα και δεν περιέχουν θόρυβο, οι τιμές που δίνονται στο k είναι μικρότερες. Στην πραγματικότητα όμως, σε ένα σύνολο δεδομένων, κάποιες κλάσεις μπορεί να είναι καλά διαχωρισμένες ενώ άλλες όχι. Επίσης, μπορεί να υπάρχει θόρυβος μόνο σε συγκεκριμένες περιοχές. Το γεγονός αυτό επηρεάζει την αποτελεσματικότητα του κατηγοριοποιητή k-NN διότι η παράμετρος k είναι σταθερή για όλες τις περιοχές. Ακόμη και αν ο κατηγοριοποιητής χρησιμοποιεί την καλύτερη σταθερή τιμή κ η οποία επιλέγεται με τη μέθοδο cross-validation, αυτή δεν είναι η βέλτιστη για συγκεκριμένες περιοχές του συνόλου δεδομένων. Άρα, σκοπός στην εργασία είναι η τιμή του k να μεταβάλλεται ανάλογα με την υποπεριοχή που βρίσκεται το υπο κατηγοριοποίηση στιγμιότυπο.

Πιο συγκεκριμένα, στόχος της εργασίας [25] είναι η ανάπτυξη μιας τεχνικής που θα αποδίδει υψηλή τιμή στην παράμετρο k όταν το υπο-κατηγοριοποίηση στιγμιότυπο βρίσκεται σε περιοχές με θόρυβο και σε περιοχές όπου οι κλάσεις δεν είναι καλά διαχωρισμένες ενώ να αποδίδει μικρότερες τιμές, αν το υπο-κατηγοριοποίηση στιγμιότυπο βρίσκεται σε περιοχές που δεν υπάρχει θόρυβος και οι κλάσεις είναι καλά διαχωρισμένες. Συνεπώς, ένας νέος αλγόριθμος με πέντε heuristics είναι η πρόταση αυτής της εργασίας ώστε το k να ορίζεται δυναμικά και όχι εκ των προτέρων από το χρήστη.

Ο προτεινόμενος αλγόριθμος ονομάζεται shd-kNN classifier (Subspace Homogenene based Dynamic k-NN) και συνδυάζεται με τις πέντε heuristics. Κοινό τους χαρακτηριστικό των heuristics είναι ότι βασίζονται στην ίδια δομή δεδομένων η οποία δημιουργεί ομοιογενή συστάδες. Η δομή κατασκευάζεται από απλή διαδικασία k-means η οποία διατηρεί τα κεντροειδή των συστάδων που δημιουργούνται. Όταν ένα νέο στιγμιότυπο έρχεται για κατηγοριοποίηση τότε γίνεται η ανάκτηση του κοντινότερου κεντροειδούς c. Με βάση αυτό το κεντροειδές ορίζεται τιμή για το k και το νέο αντικείμενο κατηγοριοποιείται εξετάζοντας τους k κοντινότερους γείτονες. Η δομή αυτή είναι γνωστή ως Structure of Homogenous Clusters (SHC) και δημιουργείται ακολουθώντας τα βήματα του παρακάτω αλγόριθμου:

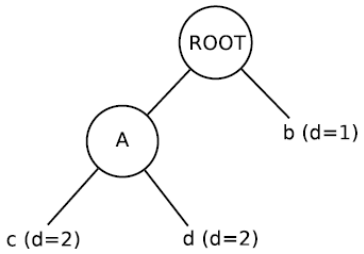


(a) initial data (b) initial class means (c) k-means on initial data

(d) Cluster centroid and class means (e) k-means on a non-homogeneous cluster (f) final set of cluster centroids

Εικόνα 6: Παραγωγή δεδομένων μέσω αναδρομικής ομαδοποίησης k-means

Αρχικά, υπολογίζεται η μέση τιμή για κάθε κλάση του training set το οποίο θεωρείται ότι είναι ανομοιογενές. Έπειτα, εφαρμόζεται ο k-means χρησιμοποιώντας αυτές τις μέσες τιμές ως αρχικά κέντρα. Έτσι δημιουργούνται τόσες συστάδες όσες και οι κλάσεις του συνόλου. Η διαδικασία επαναλαμβάνεται αναδρομικά σε κάθε ομοιογενή συστάδα που δημιουργείται. Η διαδικασία τερματίζει όταν όλες οι συστάδες που έχουν δημιουργηθεί να είναι ομοιογενείς. Ο αριθμός των αναδρομών που έγιναν για να δημιουργηθεί τελικά μια ομοιογενής συστάδα λέγεται βάθος αναδρομής d και μαζί με το κεντροειδές κάθε συστάδας αποθηκεύονται στην SCH δομή. Παρακάτω απεικονίζεται η δενδροειδής μορφή της δομής SHC και το βάθος d που οι κλάσεις έγιναν ομοιογενείς.



Εικόνα 7: Δενδροειδής μορφή της δομής SHC

Βάσει αυτής της διαδικασίας θα δημιουργηθούν μικρές ομοιογενείς συστάδες σε μεγάλο βάθος για τις περιοχές που περιέχουν θόρυβο και για τις περιοχές όπου οι κλάσεις δεν είναι καλά διαχωρισμένες. Αντίθετα, για τις "καθαρές" περιοχές που δεν βρίσκονται κοντά στα όρια των κλάσεων θα δημιουργηθούν μεγάλες συστάδες σε μικρό βάθος.

Το βάθος αναδρομής d, λοιπόν, είναι μικρό και οι συστάδες είναι μεγάλες και ομοιογενείς για τις περιοχές που περιέχουν ετικέτες μιας κλάσης. Αντίθετα στις περιοχές με θόρυβο και σε περιοχές που τα όρια των κλάσεων δεν είναι ξεκάθαρα, μικρές ομοιογενείς συστάδες συναντώνται σε μεγαλύτερο βάθος αναδρομής. Επομένως, οι πληροφορίες σχετικά με το βάθος της αναδρομής υπάρχουν στις κεντροειδείς συστάδες.

Ο shd-kNN classifier χρησιμοποιεί τη δομή δεδομένων SHC ο οποία θα καθορίσει την τιμή k. Όταν ένα νέο στιγμιότυπο x χρειάζεται να κατηγοριοποιηθεί, ο shd-kNN θα βρει το 1-κοντινότερο κεντροειδές c και το βάθος αναδρομής d αυτού. Σε αυτή τη φάση ένα από τα 5 heuristics καλείται να προσδιορίσει το k με βάση το d και τελικά ο shd-kNN βρίσκει τους k πλησιέστερους γείτονες και το x κατηγοριοποιείται.

Τα heuristics είναι:

Το k να οριστεί ίσο με το d. Για παράδειγμα, το k θα πάρει τις τιμές 1,2,3,4,5,6,7….

Το k να οριστεί ίσο με το 2d . Όταν όμως το βάθος της αναδρομής είναι μεγαλύτερο από 9 τότε ορίζουμε το k = 29 διότι το heuristic θα εξετάσει έναν πολύ μεγάλο αριθμό γειτόνων. Για παράδειγμα, το k θα πάρει τις τιμές 2,4,8,16,32,64,128…..

Το k να ισούται με το τετράγωνο του βάθους αναδρομής. Για παράδειγμα, το k θα πάρει τις τιμές 1,4,9,16,25,36,49….

Το k να ισούται με το $(d \times (d + 1))/2$ ή $\kappa = \sum_{i=1}^{d} d$. Για παράδειγμα, το k θα πάρει τις τιμές 1,3,6,10,15,21,28,….

Το k να είναι ίσο με $e\sqrt{d}$. Για παράδειγμα, , το k θα πάρει τις τιμές 2,4,5,7,9,11,14,….

Ο κατηγοριοποιητής shd-kNN δοκιμάστηκε σε 18 datasets και αποδείχθηκε ότι ο shd-kNN classifier είναι πιο ακριβής σε σχέση με τον k-NN που χρησιμοποιεί το καλύτερο βάσει cross-validation k όταν το dataset περιέχει καλά-διαχωρισμένες και όχι καλά-διαχωρισμένες περιοχές καθώς και θόρυβο σε κάποιες περιοχές. Επίσης, τα αποτελέσματα έδειξαν πως το 4ο heuristic είναι αυτό που έχει την καλύτερη απόδοση και είναι ιδανικό να εφαρμόζεται σε δεδομένα που περιέχουν θόρυβο.

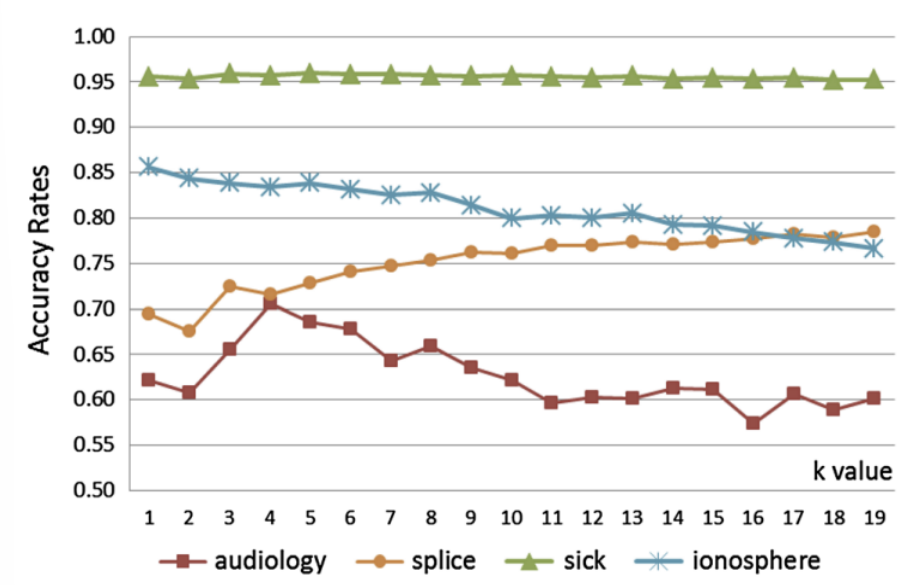
3.3 Locally adaptive k parameter selection for nearest neighbor classifier: one nearest cluster (1NC)

Σκοπός και αυτής της εργασίας είναι να προτείνει τη δυναμική επιλογή του k για κάθε στιγμιότυπο που εξετάζεται και πρέπει να κατηγοριοποιηθεί. Η νέα προτεινόμενη τεχνική [18], εκτελεί συσταδοποίηση στους γείτονες τους στιγμιότυπου που πρέπει να κατηγοριοποιηθεί. Συνεπώς, για κάθε στιγμιότυπο προς κατηγοριοποίηση, η προτεινόμενη τεχνική εκτελεί τον αλγόριθμο k-means και βρίσκει συστάδες στους κοντινότερους γείτονες.

Οι συγγραφείς της εργασίες αναφέρουν ότι η σταθερή τιμή του k μπορεί να έχει ως αποτέλεσμα χαμηλή ακρίβεια και αυτό το αποδεικνύουν παρουσιάζοντας τα 4 στοιχεία παρακάτω.

1ο Στοιχείο

Στην εικόνα παρατηρεί κανείς πως η τιμή του k επηρεάζει την απόδοση του κάθε dataset. Στην περίπτωση του συνόλου δεδομένων sick δεν παρατηρείται κάποια σημαντική μεταβολή. Αντίθετα, τα υπόλοιπα παρουσιάζουν μεγάλη ευαισθησία στην τιμή του k με το splice να πετυχαίνει μεγαλύτερη ακρίβεια όσο η τιμή αυξάνεται και τα ionosphere και audiology να σημειώνουν χαμηλότερη ακρίβεια όσο η τιμή αυξάνεται.



Εικόνα 8: Η επίδραση της παραμέτρου k στην απόδοση του k-NN

2ο Στοιχείο

Η ακρίβεια της ταξινόμησης που παρέχεται από τις διάφορες τιμές k για τα σημεία δοκιμής του συνόλου δεδομένων ακουολογίας φαίνεται στον επόμενο πίνακα. Οι σωστές και οι ψευδείς προβλέψεις συμβολίζονται με T και F, αντίστοιχα. Το T σημαίνει ότι ο κατηγοριοποιητής προβλέπει με ακρίβεια την κλάση του νέου στιγμιότυπου και το αντίστροφο. Η πιθανότητα της πρόβλεψης υποδεικνύεται από τους πλάγιους πραγματικούς αριθμούς κάτω από τα γράμματα T και F. Η πλειοψηφία των k πιο

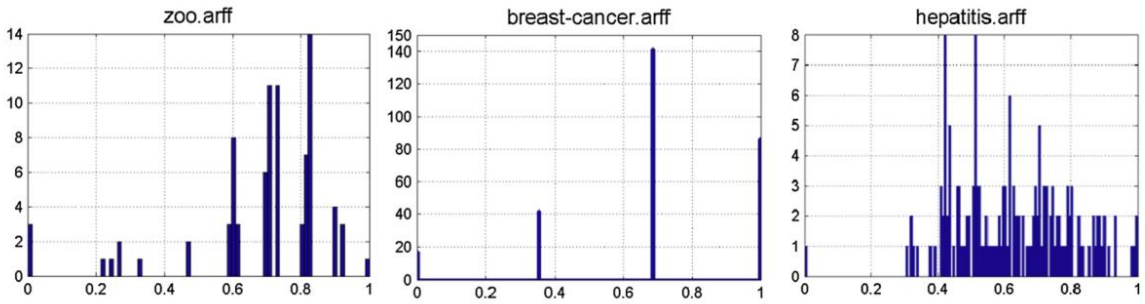
κοντινών γειτόνων ενός αντικειμένου αποφασίζει για την κατηγοριοποίησή του. Για παράδειγμα, οι πιθανότητα του 33ου στιγμιότυπου είναι περίπου 1. Εάν το k είναι μεγαλύτερο από 2, το 61ο στιγμιότυπο κατηγοριοποιείται επιτυχώς. Λόγω των μεταβαλλόμενων τιμών k, η εκτίμηση του 70ου στιγμιότυπου είναι ασταθής. Η εγγύτητα του στιγμιότυπου σε ένα όριο απόφασης μπορεί να είναι η αιτία. Σύμφωνα με αυτόν τον πίνακα, η απόδοση της κατηγοριοποίησης είναι εξαιρετικά ευαίσθητη στην ακριβή τιμή του k. Ο πίνακας δείχνει ότι μόνο μία κλάση των στιγμιότυπων προβλέπεται με ακρίβεια όταν η παράμετρος k είναι σταθερή στο 1. Όταν η τιμή k οριστεί σε 4, συνολικά 4 στιγμιότυπα ταξινομούνται σωστά. Επιπλέον, η χρήση διαφόρων τιμών k κατάλληλων για κάθε στιγμιότυπο αντί για σταθερές τιμές k βελτιώνει τη συνολική ακρίβεια. Για μεγαλύτερη ακρίβεια, η τιμή του k για κάθε στιγμιότυπο θα πρέπει να είναι μικρή, μεγάλη ή μέσα σε ένα συγκεκριμένο διάστημα.

	k parameter for *k*-NN									
	1	2	3	4	5	6	7	8	9	10
33rd test point	T	T	T	T	T	T	T	T	T	T
	1.0	*1.0*	*1.0*	*1.0*	*1.0*	*1.0*	*1.0*	*1.0*	*0.88*	*0.9*
54th test point	F	F	F	F	F	F	F	F	F	F
	1.0	*0.5*	*0.66*	*0.75*	*0.80*	*0.83*	*0.85*	*0.75*	*0.66*	*0.60*
61st test point	F	F	T	T	T	T	T	T	T	T
	1.0	*0.5*	*1.0*	*0.75*	*0.6*	*0.83*	*0.71*	*0.75*	*0.88*	*0.9*
70th test point	F	F	F	T	T	T	F	F	F	T
	1.0	*1.0*	*1.0*	*0.75*	*0.6*	*0.66*	*0.71*	*0.62*	*0.67*	*0.6*
87th test point	F	F	F	T	F	F	F	F	F	F
	1.0	*1.0*	*0.66*	*0.75*	*0.6*	*0.66*	*0.71*	*0.75*	*0.78*	*0.7*
# of true prediction	**1**	**1**	**2**	**4**	**3**	**3**	**2**	**2**	**2**	**3**

Εικόνα 9: Ανάλυση των τιμών k που επηρεάζουν την απόδοση

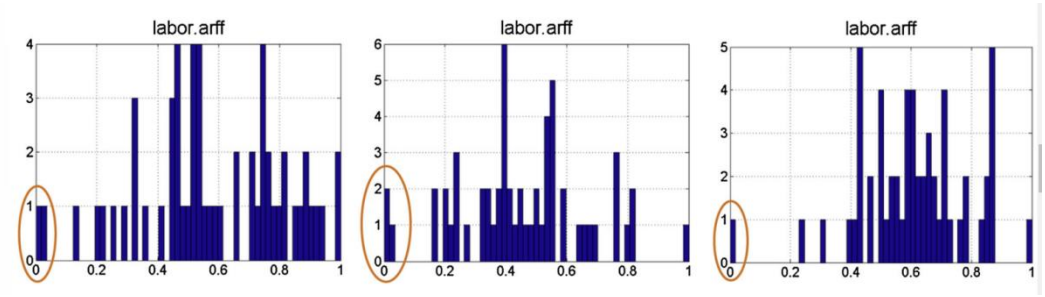
3ο Στοιχείο

Κάθε ιστόγραμμα αναπαριστά τις συχνότητες των αποστάσεων της αρχής των αξόνων και όλων των στιγμιότυπων στο χώρο καθώς και την κατανομή. Εάν έστω και ένα στιγμιότυπο αλλάξει θέση στο χώρο τα ιστογράμματα θα αλλάξουν απολύτως. Στο σύνολο δεδομένων Zoo για ένα νέο στιγμιότυπο η καλύτερη τιμή για το k που θα μπορούσε να δοθεί είναι k=3 αφού υπάρχουν 3 πλησιέστερα σημεία τα οποία βρίσκονται στην ίδια τροχιά. Παρόλο που τα τρία στιγμιότυπα έχουν μεγάλη απόσταση μεταξύ τους στο χώρο, θεωρούνται μαζί λόγω της ίδιας απόστασης και λέγεται ότι είναι στην ίδια τροχιά. Αντίστοιχα, για το σύνολο δεδομένων Breast-cancer για οποιοδήποτε νέο στιγμιότυπο η τιμή που θα έχει το k είναι περίπου 15 και για το Hepatitis είναι k = 1. Ιδανικό θα ήταν να ορίζεται τιμή για το k για κάθε ένα νέο στιγμιότυπο ξεχωριστά λόγω της κατανομής των πλησιέστερων γειτόνων του σε έναν άξονα.



Εικόνα 10: Ιστογράμματα από 3 σύνολα δεδομένων

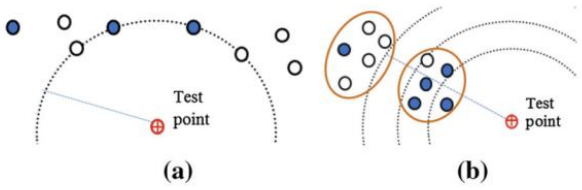
Τα τρία επόμενα ιστογράμματα αφορούν το ίδιο σύνολο δεδομένων Labor δείχνοντας το πλήθος στιγμιοτύπων ανάμεσα στην αρχή των αξόνων, το κέντρο και το επάνω δεξιότερο στιγμιότυπο του χώρου, αντίστοιχα. Αν και είναι ο χώρος είναι ο ίδιο, το ιστόγραμμα για κάθε περίπτωση δοκιμής είναι διαφορετικό. Για να είναι αξιόπιστα και ακριβή τα αποτελέσματα, οι τιμές k για κάθε στιγμιότυπο δοκιμής στον κατηγοριοποιητή k-NN θα πρέπει να οριστούν σε 2, 3 και 1, αντίστοιχα. Οι αριθμοί του πλησιέστερου στιγμιοτύπου είναι κυκλωμένοι στα διαγράμματα ως παράδειγμα.



Εικόνα 11: Πλήθος στιγμιοτύπων μεταξύ της αρχής των αξόνων, του κέντρου και του επάνω δεξιότερου στιγμιότυπου του χώρου

4ο στοιχείο

Έστω δυο διαφορετικά σενάρια στον δισδιάστατο χώρο. Το σενάριο a δείχνει ότι ίσως να υπάρχουν άλλα στιγμιότυπα των οποίων οι αποστάσεις είναι παρόμοιες με του στιγμιότυπου που εξετάζεται. Σε αυτή την περίπτωση δίνοντας τιμή 1 στο k παίρνουμε μη αξιόπιστη κατηγοριοποίηση λόγω της τυχαίας επιλογής. Η πιθανότητα βέβαια να υπάρχουν στιγμιότυπα στην ίδια απόσταση από το στιγμιότυπο που εξετάζεται είναι χαμηλή.



Εικόνα 12: Παράδειγμα κατηγοριοποίησης

Classification with one nearest clusters (1NC)

Ο αλγόριθμος 1NC (one nearest cluster) λειτουργεί ως εξής:

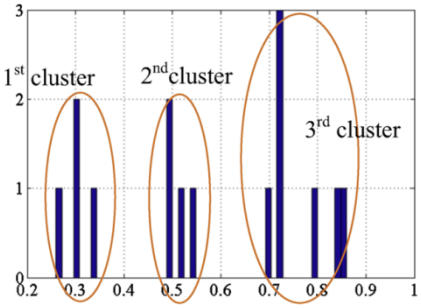
1. Αρχικά επιλέγονται τα M κοντινότερα στιγμιότυπο του συνόλου δεδομένων εκπαίδευσης γύρω από το στιγμιότυπο που πρέπει να κατηγοριοποιηθεί, οι ℓ συστάδες και οι I επαναλήψεις.

2. Στη συνέχεια υπολογίζονται οι αποστάσεις των M στιγμιοτύπων από το στιγμιότυπο, γίνεται κανονικοποίηση στο διάστημα [0,1] και τοποθετούνται τα M στιγμιότυπα στο μονοδιάστατο άξονα σύμφωνα με την απόστασή τους από το στιγμιότυπο.

3. τα τοποθετημένα δείγματα χωρίζονται σε ℓ ομάδες με τη χρήση της μεθόδου συσταδοποίησης k-means.

4. Τέλος, όλα τα στιγμιότυπα του κοντινότερου cluster εξετάζονται από τον k-NN κατηγοριοποιητή ως τα k κοντινότερα στιγμιότυπα και ο κατηγοριοποιητής διενεργεί την πρόβλεψη βάσει της ψήφου της πλειοψηφίας των στιγμιοτύπων αυτού του cluster

Τα παραπάνω βήματα επαναλαμβάνονται για κάθε στιγμιότυπο που πρέπει να κατηγοριοποιηθεί. Όπως καταλαβαίνει κανείς, πρόκειται για μια μεθοδολογία με σχετικά μεγάλο υπολογιστικό κόστος.



Εικόνα 13: Ιστόγραμμα των πλησιέστερων στιγμιότυπων στο νέο στιγμιότυπο

Ο συνδυασμός M/ℓ έχει κάποια επακόλουθα: i) το ακέραιο πηλίκο του M/ℓ θα πρέπει να είναι μικρό, αφού οποιεσδήποτε τιμές των M και ℓ δίνουν παραπλήσιο αποτέλεσμα, ii) οι μικρές τιμές των M και ℓ μειώνουν τον συνολικό υπολογιστικό χρόνο και το υπολογιστικό κόστος συνεχίζει να είναι ασήμαντο, iii) ο συνδυασμός M/ℓ ίσως θεωρηθεί ως μία μοναδική υπερπαράμετρος της μεθόδου. Αν η τιμή της ℓ είναι σταθερή τότε η τιμή της M μπορεί να είναι 4 ή 5 φορές μεγαλύτερή της. Συνεπώς, καθιερώθηκε μία συγκεκριμένη απόσταση μεταξύ του M/ℓ και της παραμέτρου k.

Οι συγγραφείς της εργασίας πραγματοποίησαν πειράματα σε 36 πραγματικά σύνολα δεδομένων και τα αποτελέσματα δείχνουν ότι η νέα μέθοδος με μια παράμετρο, την M/ℓ, αποδίδει πιο επιτυχή αποτελέσματα πρόβλεψης παρόλο που έχει μεγαλύτερη χρονική πολυπλοκότητα. Αξίζει να σημειωθεί ότι αυτό το συμπέρασμα προκύπτει χωρίς οι συγγραφείς να παρουσιάσουν πειράματα με την τιμή της σταθερής παραμέτρου κ του συμβατικού κατηγοριοποιητή k-NN να έχει οριστεί στην καλύτερη δυνατή για κάθε σύνολο δεδομένων. Οι συγγραφείς εκτέλεσαν πειράματα για τον συμβατικό αλγόριθμο κ-NN μόνο εξετάζοντας τον έναν κοντινότερο γείτονα και τους πέντε κοντινότερους γείτονες για όλα τα σύνολα δεδομένων. Το υπολογιστικό κόστος του κατηγοριοποιητή 1NC αποδείχθηκε ότι είναι τρεις φορές υψηλότερο από το κόστος του συμβατικού κατηγοριοποιητή με σταθερό κ. Αυτό είναι απολύτως λογικό, αφού για κάθε στιγμιότυπο που πρέπει να κατηγοριοποιηθεί, εκτελείται ο αλγόριθμος k-means.

3.4 An Adaptive k-Nearest Neighbor Algorithm

Ο προτεινόμενος αλγόριθμος, adaptive k-nearest neighbor algorithm (AdaNN)[26], βρίσκει το καλύτερο k, δηλαδή τον μικρότερο αριθμό κοντινότερων γειτόνων που χρειάζεται ώστε το νέο στιγμιότυπο να κατηγοριοποιηθεί στη σωστή κλάση. Η διαφορά με τον k-NN είναι ότι ο AdaNN ορίζει διαφορετικό k για κάθε ένα στιγμιότυπο που πρέπει να κατηγοριοποιηθεί αντί για ένα σταθερό.

Η μέθοδος περιλαμβάνει μια διαδικασία προ-επεξεργασίας όπου κάθε στοιχείο του συνόλου δεδομένων εκπαίδευσης κατηγοριοποιείται με τόσους κοντινότερους γείτονες όσους απαιτούνται για κατηγοριοποιηθεί στη σωστή κλάση. Όταν ένα νέο στιγμιότυπο έρχεται για κατηγοριοποίηση, αρχικά εντοπίζεται ο κοντινότερος γείτονάς του και στη συνέχεια δανείζεται το k και με αυτόν τον τρόπο κατηγοριοποιείται. Για παράδειγμα, έστω ότι το νέο στιγμιότυπο n έχει ως κοντινότερο γείτονα το x και το x έχει κατηγοριοποιηθεί στη σωστή κλάση με k=3, τότε και το n θα υιοθετήσει αυτό το k και θα κατηγοριοποιηθεί εξετάζοντας τους 3 κοντινότερους γείτονες.

Υπάρχουν κάποιες περιπτώσεις που στο νέο στιγμιότυπο δεν μπορεί να οριστεί το k του κοντινότερου γείτονα και αυτό επειδή ο κοντινότερος γείτονας δεν είναι δυνατόν να κατηγοριοποιηθεί σωστά με οποιαδήποτε τιμή k από το ένα έως το εννιά. Σε αυτές τις περιπτώσεις, οι συγγραφείς της εργασίας ορίζουν αυθαίρετα το k να είναι ίσο με 9. Είναι αξιοσημείωτο ότι ο αλγόριθμος περιορίζει την εύρεση

της τιμής k που μπορεί να κατηγοριοποιήσει σωστά το υπο-εξέταση στιγμιότυπο των δεδομένων εκπαίδευσης στους 9 κοντινότερους γείτονες με τον κίνδυνο το νέο στιγμιότυπο να κατηγοριοποιηθεί τελικά σε λάθος κλάση. Αντιθέτως, αν χρησιμοποιούσε ένα μεγαλύτερο εύρος τιμών k τότε είναι πιθανόν να εντοπίζονταν μεγαλύτερη τιμή k (k>9) που να εκτελεί σωστή κατηγοριοποίηση για το στιγμιότυπο εκπαίδευσης και έτσι, το νέα, υπο-κατηγοριοποίηση στιγμιότυπα τελικά να κατηγοριοποιούνται με υψηλότερη ακρίβεια.

Οι αλγόριθμοι 1NN-9NN και AdaNN εφαρμόστηκαν σε 15 σύνολα δεδομένων τα οποία βλέπουμε στον παρακάτω πίνακα. Το 90% του κάθε συνόλου δεδομένων έχει οριστεί ως training test και το 10% ως testing set.

dataset	classes	attributes	training	test	total
Iris	3	4	140	10	150
Protein	8	7	297	39	336
Haberman	2	3	270	36	306
Blood	2	4	666	82	748
Zoo	7	16	90	11	101
glass	6	9	189	25	214
Pima	2	8	684	84	768
Heart	2	13	243	27	270
Teaching	3	5	135	16	151
Wine	3	13	153	25	178
Balance	3	4	558	67	625
Parkinsons	2	22	171	24	195
Ionosphere	2	34	315	36	351
Contraceptive	3	9	1323	150	1473
Wisconsin	2	30	504	65	569

Εικόνα 14: Σύνολα δεδομένων που εξετάστηκαν

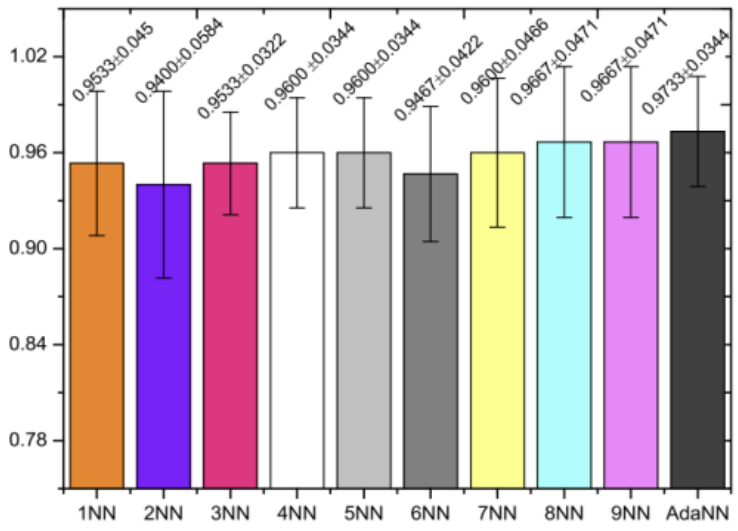
Ο AdaNN αλγόριθμος πετυχαίνει τη μεγαλύτερη ακρίβεια στα σύνολα δεδομένων Iris, Protein, Haberman και Blood. Αντίθετα η χειρότερη ακρίβεια παρατηρείται στο σύνολο δεδομένων Wisconsin

Στον πίνακα παρουσιάζονται στην πρώτη στήλη οι 10 διαφορετικοί αλγόριθμοι κατηγοριοποίησης και 6 σύνολα δεδομένων με τις αντίστοιχες μετρήσεις ακρίβειας. Σύμφωνα με την τελευταία σειρά του πίνακα ο προτεινόμενος αλγόριθμος στα σύνολα δεδομένων Protein, Haberman και Blood έχει την καλύτερη ακρίβεια ενώ στα υπόλοιπα έρχεται στη δεύτερη θέση.

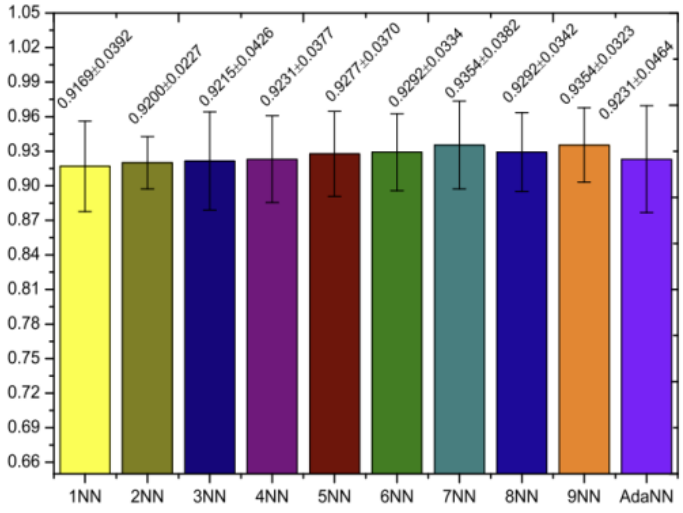
datasets	Protein	Haberman	Blood	Zoo	Glass	Pima
1NN	81.54	68.06	65.49	97.27	72.40	69.05
2NN	83.08	74.17	62.44	93.64	68.00	63.45
3NN	86.15	71.39	73.90	95.45	67.20	69.64
4NN	85.38	73.33	74.15	91.82	67.60	67.38
5NN	85.90	73.06	74.88	89.09	66.80	72.14
6NN	85.38	72.50	72.68	89.09	64.80	70.12
7NN	86.92	72.22	74.63	84.55	65.20	72.86
8NN	86.15	73.89	75.00	82.73	64.00	72.14
9NN	86.92	73.33	75.85	80.00	62.40	73.57
AdaNN	86.92	75.56	76.46	95.45	68.00	72.86
Rank	1	1	1	2	2	2

Εικόνα 15: Ποσοστά ακρίβειας 10 αλγορίθμων σε 6 σύνολα δεδομένων

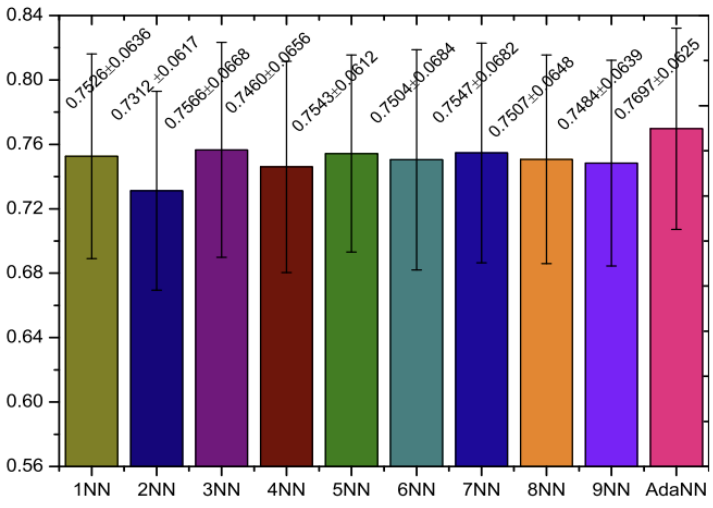
Στις επόμενες εικόνες φαίνεται ο μέσος όρος της ακρίβειας των 10 αλγορίθμων κατηγοριοποίησης που εξετάστηκαν. Στην εικόνα 16 την πρώτη θέση έχει ο AdaNN και την τελευταία ο 2NN. Στην Εικόνα 17 βλέπουμε τον 7NN και 9NN να βρίσκονται στην πρώτη θέση ενώ AdaNN να βρίσκεται στην τελευταία. Επίσης, στην Εικόνα 18 βλέπουμε τη συνολική απόδοση των αλγορίθμων που εξετάστηκαν σε όλα τα σύνολα δεδομένων.



Εικόνα 16: Ποσοστά ακρίβειας για το σύνολο δεδομένων Iris



Εικόνα 17: Ποσοστά ακρίβειας για το σύνολο δεδομένων Wisconsin



Εικόνα 18: Ποσοστά ακρίβειας για το σύνολο δεδομένων Total

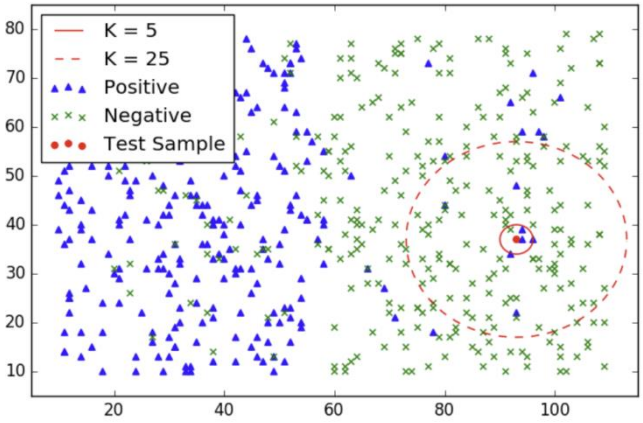
Συμπερασματικά, αν και ο AdaNN ξεπερνάει τους 1NN-9NN στα περισσότερα σύνολα δεδομένων, ειδικά στα σύνολα δεδομένων μικρής κλίμακας, ωστόσο και σε αυτή την εργασία δεν έγινε η σύγκρισή του με τον best kNN αλγόριθμο όπου η καλύτερη σταθερή τιμή k ορίζεται με τη μέθοδο cross-validation.

3.5 An Improved k-NN Classification with Dynamic k

Ένας ακόμη αλγόριθμος που προτείνεται ο οποίος χρησιμοποιεί δυναμικό k είναι ο Dk-NN[28]. Η μέθοδος περιλαμβάνει ένα διάστημα τη παραμέτρου k το οποίο συμβολίζεται ως [kmin,kmax], και έναν

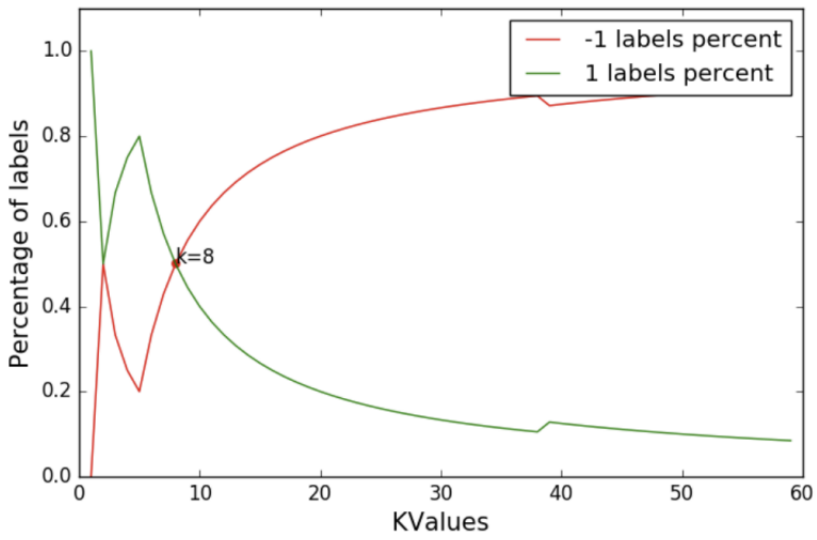
πίνακα T = [T1, T2,…, Tlen(L)]. Για κάθε στιγμιότυπο που πρέπει να κατηγοριοποιηθεί, ο πίνακας T δίνει τις καμπύλες τάσης μεταβολής για κάθε κλάση εντός του διαστήματος [kmin,kmax]. Στην πραγματικότητα, υπάρχουν τρεις διαφορετικές καταστάσεις.

Πρώτη κατάσταση: Στο επόμενο σχήμα απεικονίζεται η δυαδική κατηγοριοποίηση ενός στιγμιοτύπου και η επιλογή σταθερής τιμής k με k = 5 και k =25.



Εικόνα 19: Παράδειγμα δυαδικής κατηγοριοποίησης επιλέγοντας σταθερή τιμή της k με k = 5 και k = 25

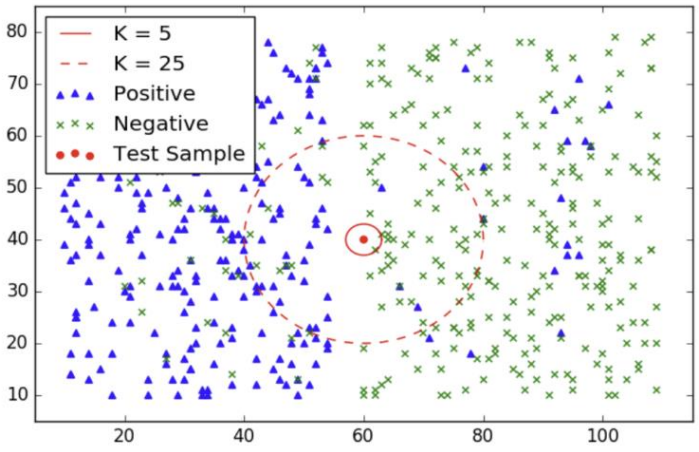
Παρακάτω βλέπουμε τη μεταβολή που παρουσιάζουν οι καμπύλες τάσης του στιγμιοτύπου αναλόγως με την τιμή που θα πάρει το k. Τα ποσοστά αρνητικής ετικέτας είναι χαμηλότερα από εκείνα της θετικής ετικέτας όταν k ανήκει στο διάστημα [1,8]. Εάν υπάρχουν μερικές τομές στο τέλος του διαστήματος [kmin,kmax], αλλά στα άλλα μέρη δεν υπάρχουν τομές και η ποσοστιαία διαφορά είναι προφανής, η ετικέτα του δείγματος πρέπει να οριστεί στην κατηγορία που έχει το υψηλότερο ποσοστό στο ενδιάμεσο τμήμα του διαστήματος [kmin,kmax].



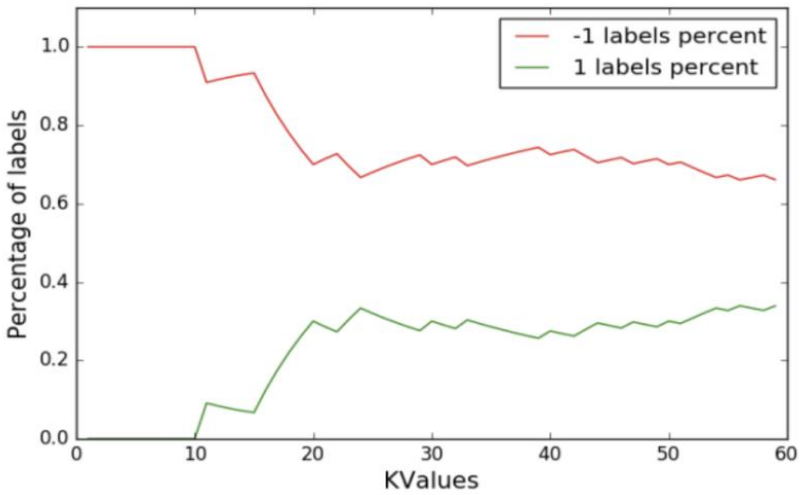
Εικόνα 20: Καμπύλες τάσης μεταβολής του στιγμιότυπου

Ο λόγος αυτού του φαινομένου είναι ότι υπάρχουν ακραίες τιμές πολύ κοντά στο δείγμα δοκιμής. Αντιθέτως, στις περιπτώσεις που k>8, το ποσοστό των δύο ετικετών αντιστρέφεται και όσο συνεχίζει η τιμή του k να αυξάνεται, το ποσοστό της αρνητικής ετικέτας τείνει να είναι κοντά στο 1, ενώ το ποσοστό της θετικής ετικέτας τείνει να είναι κοντά στο 0. Κατά συνέπεια, αναλύοντας τη μεταβολή των καμπυλών τάσης με τον πίνακα T μπορούν να αποφύγουν την επίδραση των ακραίων τιμών και να δώσουν τη σωστή απάντηση, δηλαδή να εκτελέσουν σωστή κατηγοριοποίηση.

Δεύτερη κατάσταση: Αν δεν υπάρχει καμία τιμή στο διάστημα [kmin,kmax], η ετικέτα του δείγματος δοκιμής μπορεί να οριστεί ως η κατηγορία που έχει το υψηλότερο ποσοστό στο διάστημα [kmin,kmax].

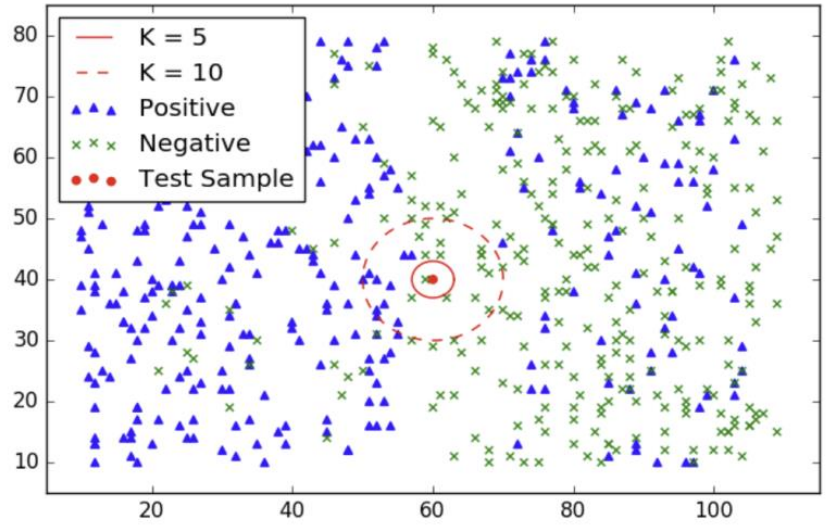


Εικόνα 21: Πραγματική κατάσταση 600 στιγμιότυπων και ενός νέου που πρόκειται να κατηγοριοποιηθεί

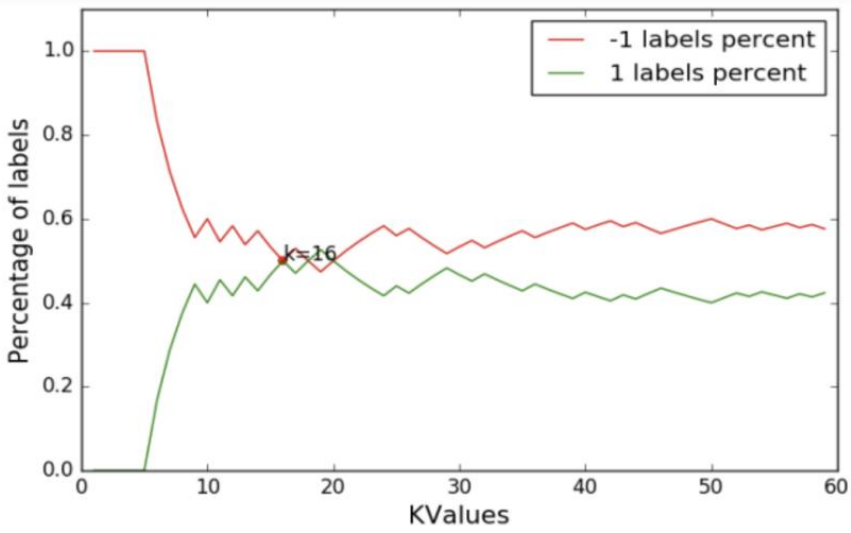


Εικόνα 22: Καμπύλες τάσης μεταβολής του στιγμιότυπου

Τρίτη κατάσταση: Το σχήμα που ακολουθεί δείχνει μια άλλη κατάσταση, στην οποία υπάρχουν μερικά ακραία σημεία γύρω από το δείγμα δοκιμής. Το σχήμα β δίνει τις αντίστοιχες καμπύλες τάσης. Στην αρχή, οι δύο καμπύλες έχουν μεγάλη απόσταση. Αλλά με την αύξηση της τιμής k, η απόσταση μεταξύ των δύο καμπυλών μειώνεται σταδιακά ώσπου τέμνονται σε 3 σημεία, για k=16, k=18 και k=20. Σε αυτήν την περίπτωση, οι ακραίες τιμές δεν έχουν μεγάλη επιρροή στα αποτελέσματα της ταξινόμησης. Εάν, λοιπόν, η ποσοστιαία διαφορά στο αριστερό διάστημα [kmin,kmax] είναι προφανής, αλλά σε άλλα μέρη, αυτή η διαφορά είναι πολύ μικρή, η ετικέτα του δείγματος δοκιμής θα πρέπει να οριστεί στην κατηγορία που έχει το υψηλότερο ποσοστό εντός του διαστήματος [kmin,kmax] αλλά για k>20.



Εικόνα 23: Πραγματική κατάσταση 600 στιγμιότυπων και ενός νέου που πρόκειται να κατηγοριοποιηθεί



Εικόνα 24: Καμπύλες τάσης μεταβολής του στιγμιότυπου

Όπως φαίνεται από τα παραπάνω, το διάστημα [kmin,kmax] θα πρέπει να υπολογιστεί πριν από την εφαρμογή αυτών των κριτηρίων. Για τον καθορισμό του διαστήματος [kmin,kmax], δίνεται ένας επαναληπτικός αλγόριθμος στον Αλγόριθμο 1. Σε κάθε επαναληπτικό, το σύνολο δεδομένων εκπαίδευσης και το σύνολο δεδομένων δοκιμής επιλέγονται τυχαία. Στη συνέχεια, υπολογίζονται οι τομές κάθε δείγματος δοκιμής. Μετά την επανάληψη, η συχνότητα των τομών μετράται προς τα πάνω και το διάστημα [kmin,kmax] καθορίζεται σύμφωνα με τη συχνότητα των τομών με κατώφλι συχνότητας.

Algorithm 1: interval selection

Input sample dataset: $D=\{(x,l):x\in R^M,l\in L\}$, Iterative times N , threshold h, percentage of random selection p, kList=NULL, fList=NULL

Output: interval: $[k_{\min},k_{\max}]$

1:For each iterative do

2: Select p of samples randomly as training dataset D_{train} , testing dataset D_{test} contains The remaining samples.

3: Calculate matrix **T** of each sample in D_{test} with samples in D_{train} .

4: for each **T** do

5: Calculate intersections, get k value list k_{x_i} and frequency list f_{x_i} of these intersection.

6: Combine k_{x_i} to kList and f_{x_i} to fList.

7:According to h, get list $\{k_i,k_{i+1},\cdots,k_j\}$ by removing the k value whose frequency is beyond h.

8: set $k_{\min}=\min\{k_i,k_2,\cdots,k_j\}$

$k_{\max}=\min\{k_i,k_2,\cdots,k_j\}$

9: Output $[k_{\min},k_{\max}]$

Εικόνα 25: Επαναληπτικός Αλγόριθμος για τον καθορισμό του διαστήματος [kmin, kmax]

Έπειτα από πειράματα που έγιναν σε πραγματικό σύνολο δεδομένων, έγινε η σύγκριση του προτεινόμενου βελτιωμένου αλγορίθμου με τιμές του k στο διάστημα [2, 232] με τον κλασικό k-NN αλγόριθμο για k=1, k=3, k=5, και k=130. Αν και η ακρίβεια του Dk-NN και του 130-NN αλγορίθμου είναι πολύ κοντά, ωστόσο την καλύτερη ακρίβεια την πετυχαίνει ο k-NN με τιμή k=130.

Table 2. Experiments result of Dk-NN and k-NN.

	k value type	k value	precision	recall	F
k-NN	Fixed	1	0.7467	0.7196	0.7329
k-NN	Fixed	3	0.8328	0.7601	0.7948
k-NN	Fixed	5	0.8560	0.7670	0.8091
k-NN	Fixed	130	**0.9081**	0.7046	0.7935
Dk-NN	dynamic	[2,232]	0.8911	**0.7800**	**0.8319**

Εικόνα 26: Αποτέλεσμα των πειραμάτων του Dk-NN και του k-NN

Μελλοντικός στόχος των συγγραφέων αποτελεί η μελέτη των προβλημάτων ταξινόμησης πολλαπλών ετικετών.

3.6 Extending Nearest Neighbor Classification with Spheres of Confidence

Μια άλλη προσέγγιση είναι ο αλγόριθμος να μην βασίζεται σε k γείτονες αλλά να μελετάει τη γειτονιά στην οποία θα βρεθεί το στιγμιότυπο και ανάλογα να ρυθμίζει το k.

Συγκεκριμένα, ο προτεινόμενος αλγόριθμος, ο οποίος ονομάζεται BuLL (Bubble Lazy Learner)[27], αρχικά αναπτύσσει γύρω από κάθε στιγμιότυπο μία σφαίρα εμπιστοσύνης. Μόλις έρθει ένα νέο στιγμιότυπο για κατηγοριοποίηση τότε καλύπτεται από κάποιες σφαίρες που δημιουργήθηκαν νωρίτερα. Από αυτές τις σφαίρες καθορίζεται η κλάση που θα κατηγοριοποιηθεί το νέο στιγμιότυπο.

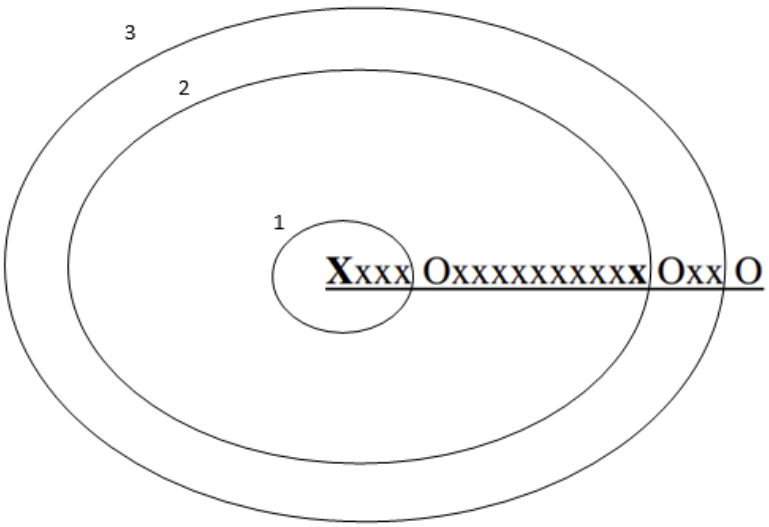
Για να δημιουργηθούν οι σφαίρες εμπιστοσύνης γύρω από κάθε στιγμιότυπο υπάρχουν διαφορετικοί τρόποι. Υπάρχουν δύο εκδοχές του βασικού αλγορίθμου. Η μία εκδοχή υπολογίζει το πλήθος των σφαιρών (sphere aggregation) και η δεύτερη υπολογίζει το πλήθος των στιγμιοτύπων της κάθε κλάσης (instance aggregation). Οι σφαίρες δημιουργούνται με τον εξής τρόπο: το στιγμιότυπο γύρω από το οποίο δημιουργείται η σφαίρα είναι το κέντρο. Ξεκινώντας λοιπόν από το κέντρο η ακτίνα μεγαλώνει για όσο τα γειτονικά στιγμιότυπα ανήκουν στην ίδια κλάση με αυτή του κέντρου και σταματάει μόλις συναντήσει στιγμιότυπο διαφορετικής κλάσης. Επομένως, τα στιγμιότυπα της σφαίρας εμπιστοσύνης έχουν όλα της ίδιας κλάσης. Συμπεραίνουμε πως σε καθαρές περιοχές οι σφαίρες θα είναι μεγάλες και θα περιέχουν πολλά στιγμιότυπα σε αντίθεση με περιοχές που περιέχουν στιγμιότυπα πολλών διαφορετικών κλάσεων τότε οι σφαίρες θα είναι πολύ μικρότερες με λιγότερα στιγμιότυπα. Όταν ένα νέο στιγμιότυπο έρχεται για κατηγοριοποίηση, οι σφαίρες που το καλύπτουν θεωρούνται η γειτονιά του. Χωρίς να υπολογίζεται ο αριθμός των στιγμιοτύπων που βρίσκονται μέσα, μετράει τις κατηγορίες των κλάσεων και αυτή με το μεγαλύτερο πλήθος είναι αυτή στην οποία θα κατηγοριοποιηθεί το νέο αντικείμενο.

Εκτός από τον βασικό αλγόριθμο υπάρχουν και 2 παραλλαγές του. Σε αυτές, οι σφαίρες δεν περιέχουν στιγμιότυπα ίδιας κλάσης, συνεπώς κατανομή 100%, αλλά δυο διαφορετικών κλάσεων και το ποσοστό της κατανομής υπολογίζεται με τη βοήθεια της εκτίμησης Laplace σύμφωνα με τον τύπο:

$$p_{class\ A} = \frac{k+1}{N+C}$$

όπου k: τα στιγμιότυπα που ανήκουν στην κυρίαρχη κλάση, N: όλα τα στιγμιότυπα που ανήκουν στη σφαίρα και C: το πλήθος των σφαιρών. Με τον τρόπο αυτό, προκύπτει ένα ποσοστό, διαφορετικό του 100%, που δείχνει την κατανομή της κάθε σφαίρας και μπορούν 2 σφαίρες να συγκριθούν ώστε να γίνει η κατηγοριοποίηση. Τελικά, το στιγμιότυπο κατηγοριοποιείται στη σφαίρα με τη μεγαλύτερη εκτίμηση Laplace.

Σύμφωνα με την 1η παραλλαγή, ο τρόπος που δημιουργούνται οι σφαίρες εμπιστοσύνης γύρω από ένα στιγμιότυπο διαφέρει από τη βασική εκδοχή καθώς δεν περιέχονται μόνο στιγμιότυπα ίδιας κλάσης αλλά στόχος είναι η ακτίνα να αυξηθεί και πέρα από το πρώτο διαφορετικής κλάσης στιγμιότυπο. Η σφαίρα σταματάει να μεγαλώνει πριν από το δεύτερο διαφορετικής κλάσης στιγμιότυπο που θα συναντήσει.



Εικόνα 27: Μια δομή eager

Όπως βλέπουμε και στο παραπάνω παράδειγμα, με κέντρο το X δημιουργούνται τρεις σφαίρες. Η πρώτη περιέχει στιγμιότυπα ίδιας κλάσης, στη συνέχεια η ακτίνα μεγαλώνει σταδιακά και κάθε φορά συμπεριλαμβάνει και ένα στιγμιότυπο διαφορετικής κλάσης. Εφαρμόζοντας τον τύπο της εκτίμησης Laplace προκύπτει το εξής συμπέρασμα: P1<P2>P3. Δηλαδή, η εκτίμηση Laplace της 2ης σφαίρας είναι

μεγαλύτερη από την πρώτη γι αυτό και δημιουργήθηκε η Τρίτη σφαίρα. Η εκτίμηση Laplace της 3ης σφαίρας είναι μικρότερη της 2ης και ο αλγόριθμος δεν δημιουργεί άλλη σφαίρα.

Η άλλη παραλλαγή δεν θα σταματήσει μόλις η εκτίμηση Laplace μειωθεί αλλά θα υπολογίσει εξ αρχής όλες τις πιθανές σφαίρες εμπιστοσύνης και στη συνέχεια θα υπολογιστεί η εκτίμηση Laplace της κάθε μίας. Το νέο στιγμιότυπο θα κατηγοριοποιηθεί σύμφωνα με τη σφαίρα που έχει τη μεγαλύτερη εκτίμηση Laplace.

Υπάρχουν περιπτώσεις που το νέο στιγμιότυπο δεν καλύπτεται από κάποια σφαίρα τότε ενεργοποιείται ο αλγόριθμος k-NN. Δεν χρησιμοποιείται όμως η καλύτερη τιμή του k έπειτα από τη μέθοδο cross validation αλλά δίνεται στο k η τιμή 5 και έτσι γίνεται η κατηγοριοποίηση ελέγχοντας τους 5 κοντινότερους γείτονες.

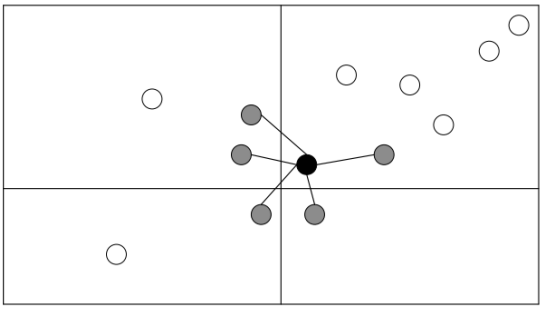
Έγιναν 2 πειράματα στα πλαίσια της συγκεκριμένης εργασίας. Στο πρώτο έγινε η σύγκριση μεταξύ της βασικής εκδοχής του αλγορίθμου και του k-NN με τιμές για το k 5, 11 και 17 οι οποίοι εφαρμόστηκαν σε 18 σύνολα δεδομένων και νικητής βγήκε ο BuLL με καλύτερη απόδοση σε 13 από τα 18 σύνολα δεδομένων. Στο δεύτερο, έγινε η σύγκριση του k-NN με όλες τις εκδοχές και παραλλαγές του αλγόριθμου BuLL και το συμπέρασμα είναι πως ο k-NN έρχεται πάλι δεύτερος. Αν συγκριθούν οι εκδοχές και παραλλαγές του αλγορίθμου BuLL μεταξύ τους προκύπτει πως αυτή που υπερτερεί είναι η βασική εκδοχή του αλγορίθμου με τις σφαίρες οι οποίες περιέχουν στιγμιότυπα ίδιας κλάσης. Αξίζει να σημειωθεί ότι οι συγγραφείς του paper δεν συμπεριέλαβαν στα πειράματα τους τον k-NN με το καλύτερο k το οποίο προέκυψε από τη μέθοδο εύρεσης του βέλτιστου σταθερού k cross validation.

3.7 Evolving a Locally Optimized Instance Based Learner

Η τελευταία τεχνική που παρουσιάζεται στα πλαίσια της παρούσας εργασίας είναι η τεχνική G-kNN[24], η οποία, όπως και σε προηγούμενες τεχνικές δεν απαιτεί να οριστεί ο αριθμός των γειτόνων ως παράμετρος. Με τη χρήση δέντρων αποφάσεων καθορίζεται ο αριθμός των γειτόνων που θα χρησιμοποιηθούν για την κατηγοριοποίηση ενός νέου στιγμιότυπου. Το δέντρο απόφασης χωρίζει των πολυδιάστατο χώρο σε υπο-χώρους. Τα φύλλα του δέντρου απόφασης δεν κατηγοριοποιούν αμέσως ένα στιγμιότυπο, αλλά χρησιμοποιούν κάποια έκδοση του kNN για όλα τα στιγμιότυπα που φθάνουν σε αυτόν τον κόμβο. Συγκεκριμένα, τα φύλλα του αφορούν μια υπο-περιοχή του συνόλου των δεδομένων εκπαίδευσης. Σε κάθε υπο-περιοχή χρησιμοποιείται άλλη τιμή k για την κατηγοριοποίηση του κάθε στιγμιότυπου. Παρακάτω παρουσιάζονται 3 διαφορετικές τεχνικές της προτεινόμενης μεθόδου.

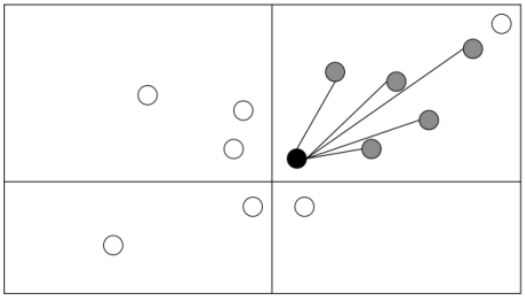
Στην πρώτη τεχνική, τη global G-kNN, για να κατηγοριοποιηθεί ένα νέο στιγμιότυπο χρησιμοποιούνται οι k κοντινότεροι γείτονες ακόμη και αν αυτοί ανήκουν σε άλλα φύλλα. Το κάθε φύλλο έχει το "δικό του καλύτερο" k. Αυτό σημαίνει ότι, π.χ. για το επάνω δεξιά τετράγωνο η καλύτερη τιμή του k είναι k=5, για να κατηγοριοποιηθεί οποιοδήποτε νέο αντικείμενο θα υπολογιστούν οι 5 κοντινότεροι γείτονές του. Επομένως, για την κατηγοριοποίηση του μαύρου στιγμιοτύπου, του παρακάτω παραδείγματος,

επιλέγονται οι 5 κοντινότεροι γείτονες από όλο το σύνολο δεδομένων παρόλο που όπως φαίνεται είναι σε διαφορετικές περιοχές οι 4 από αυτούς.



Εικόνα 28: Επιλέγονται 5 γείτονες εφαρμόζοντας global G-kNN

Στη δεύτερη τεχνική του αλγορίθμου, τη local G-kNN, οι κοντινότεροι γείτονες που λαμβάνονται υπόψη είναι μόνο από την ίδια περιοχή που βρίσκεται και το στιγμιότυπο που πρόκειται να κατηγοριοποιηθεί. Όπως φαίνεται και στο σχήμα που ακολουθεί, για να κατηγοριοποιηθεί αυτή τη φορά το μαύρο στιγμιότυπο οι γείτονες που θεωρούνται ως κοντινότεροι είναι μόνο από την περιοχή που ανήκει και το στιγμιότυπο.



Εικόνα 29: Επιλέγονται 5 γείτονες εφαρμόζοντας local G=kNN

Η global τεχνική μοιάζει περισσότερο με τον κλασικό kNN αλγόριθμο. Η βασική διαφορά είναι πως τοπικά χρησιμοποιούνται πολλές βελτιστοποιημένες τιμές για το k, δηλαδή κάθε φύλλο έχει τη δική του τιμή k που θεωρείται η καλύτερη και όχι μια τιμή για όλο το σύνολο δεδομένων. Η τεχνική local G-kNN είναι παρόμοια με τα δέντρα αποφάσεων με τη διαφορά ότι τα φύλλα δεν κατηγοριοποιούν απευθείας ένα νέο στιγμιότυπο με βάση την πλειοψηφούσα κλάση των στιγμιοτύπων αλλά, χρησιμοποιούν ένα τοπικό kNN το οποίο περιορίζεται στα στιγμιότυπα του συγκεκριμένου φύλλου.

Η Τρίτη τεχνική αποτελεί συνδυασμό των δύο προηγούμενων. Στις περιπτώσεις όπου οι κοντινότεροι γείτονες ανήκουν σε διαφορετικά φύλλα τότε ενεργοποιείται ο global G-kNN. Αντίθετα, αν οι κοντινότεροι γείτονες είναι στο ίδιο φύλλο τότε χρησιμοποιείται το βελτιστοποιημένο k.

Στον παρακάτω πίνακα γίνεται η σύγκριση των πειραμάτων μεταξύ των διαφόρων τεχνικών του G-kNN αλγορίθμου και του κλασικού k-NN αλγορίθμου. Χρησιμοποιήθηκαν μόνο περιττοί αριθμοί για την τιμή της k και συγκεκριμένα k=5, k=11 και k=17 για τον k-NN αλγόριθμο και για τον αλγόριθμο G-kNN οι πιθανές τιμές που μπορεί να πάρει η μεταβλητή k είναι από 1 έως 25.

Dataset	**kNN**			**G-kNN**		
	k=5	**k=11**	**k=17**	**Global**	**Local**	**Mix**
Bcancer	0.6927	0.7485	**0.7518**	0.7332	0.7188	0.7251
Bld	0.6116	0.6234	0.6384	**0.6505**	0.6399	0.6427
Cleve	**0.8282**	0.8247	0.8278	0.8250	0.8253	0.8206
CMC	0.4650	0.4759	0.4732	0.4666	**0.5184**	0.5168
Cred-A	**0.8710**	0.8667	0.8638	0.8652	0.8644	0.8609
Cred-G	0.7320	**0.7390**	0.7330	0.7326	0.7322	0.7296
CB	0.7315	0.7130	0.6981	**0.8006**	0.7937	0.7526
Ecoli	0.8692	0.8662	0.8335	0.8706	0.8637	**0.8709**
Glass	0.6784	0.6357	0.6177	0.6944	0.6867	**0.7022**
Haber	0.6926	**0.7452**	0.7223	0.7248	0.7133	0.7185
Heart	0.8037	0.8148	**0.8222**	0.8155	0.8181	0.8104
Hepati	**0.8458**	**0.8458**	0.8200	0.8387	0.8346	0.8341
Horse	0.8098	0.8182	0.8264	**0.8307**	0.8258	0.8288
Iono	0.8547	0.8463	0.8434	0.8697	0.8729	**0.8789**
Iris	0.9533	0.9467	**0.9667**	0.9547	0.9540	0.9567
Labor	**0.8867**	0.8367	0.8000	0.8717	0.8573	0.8530
Lymph	0.8390	0.8262	0.8400	**0.8502**	0.8417	0.8444
PID	0.7291	0.7331	**0.7382**	0.7340	0.7321	0.7318
Sick	0.9634	0.9597	0.9557	0.9644	0.9640	**0.9660**
Sonar	0.8360	0.7260	0.7060	**0.8534**	0.8113	0.8240
TAE	0.5508	0.5104	0.4979	**0.5800**	0.5295	0.5455
TTT	0.8841	0.9530	0.9813	0.9904	0.9776	**0.9917**
Vehicle	**0.6975**	0.6786	0.6845	0.6889	0.6947	0.6911
Votes	0.9335	0.9289	0.9266	0.9344	**0.9584**	0.9544
Wine	0.9660	**0.9775**	**0.9775**	0.9741	0.9666	0.9694
WBC	0.9643	**0.9657**	**0.9657**	0.9571	0.9559	0.9594
Zoo	0.9500	0.8818	0.8827	**0.9651**	0.9362	0.9523
MEAN	**0.8015**	**0.7958**	**0.7924**	**0.8162**	**0.8106**	**0.8123**

Εικόνα 30: Ακρίβειες των k-NN και G-kNN

Το συμπέρασμα των πειραμάτων είναι πως ο global G-kNN καταφέρνει σημαντικά μεγαλύτερη ακρίβεια από ότι οι τρεις εκδόσεις του kNN καθώς και η γενική απόδοση του G-kNN αλγορίθου είναι καλύτερη από του κλασικού kNN.

Όπως και σε προηγούμενες τεχνικές που είδαμε στο πλαίσιο εκπόνησης της παρούσας εργασίας, βασικό μειονέκτημα της εργασίας αποτελεί το γεγονός ότι οι συγκρίσεις μεταξύ του αλγορίθμου G-kNN και kNN γίνονται για επιλεγμένες τιμές του k και δεν γίνεται καμία σύγκριση με το καλύτερο σταθερό k το οποίο προκύπτει από τη μέθοδο cross-validation.

Κεφάλαιο 4ο: Συμπεράσματα

4.1 Συμπεράσματα

Η απόδοση τιμής στην παράμετρο k του κατηγοριοποιητή των k εγγύτερων γειτόνων έχει απασχολήσει την ερευνητική κοινότητα της τελευταίες δεκαετίες. Το πρόβλημα είναι ακόμη ανοιχτό. Συνεπώς, ακόμη και σήμερα, δημοσιεύονται άρθρα σε επιστημονικά περιοδικά και συνέδρια που να αφορούν την απόδοση τιμής σε αυτή την παράμετρο.

Η παρούσα εργασία, αρχικά παρουσίασε τον κατηγοριοποιητή των k εγγύτερων γειτόνων, τα πλεονεκτήματα και τα μειονέκτημα του καθώς και τις διάφορες παραλλαγές του. Στη συνέχεια, έγινε μια προσπάθεια παρουσίασης των ερευνητικών προσπαθειών που αφορούν την απόδοση τιμής στην παράμετρο k. Αρχικά παρουσιάστηκαν οι τεχνικές απόδοσης μιας μοναδικής και στατικής τιμής. Οι προσεγγίσεις αυτές επιτυγχάνουν την επιλογή μιας σταθερής τιμής k ανεξάρτητα από την υπο περιοχή που βρίσκεται το προς κατηγοριοποίηση αντικείμενο. Με άλλα λόγια, δεν λαμβάνουν υπόψη τα ιδιαίτερα χαρακτηριστικά των υποπεριοχών του συνολικού πολυδιάστατου χώρου του συνόλου δεδομένων εκπαίδευσης. Σε ένα σύνολο δεδομένων μπορεί να υπάρχουν υπο-περιοχές με θόρυβο και άλλες περιοχές χωρίς θόρυβο. Αυτό αγνοείται πλήρως από τις τεχνικές απόδοσης στατικής τιμής.

Στη συνέχεια παρουσιάστηκαν οι τεχνικές που αφορούν την απόδοση δυναμικής τιμής ανάλογα με τον υποχώρο που βρίσκεται το υποκατηγοριοποίηση αντικείμενο. Οι προσεγγίσεις απόδοσης δυναμικής τιμής αποδίδουν μια διαφορετική τιμή σε κάθε υποκατηγοριοποίηση αντικείμενο. Αυτό που προσπαθούν να επιτύχουν είναι την απόδοση μιας μικρής τιμής k για τις “καθαρές” περιοχές των κλάσεων που βρίσκονται μακριά από τα όρια των κλάσεων και δεν περιέχουν θόρυβο και μεγαλύτερες τιμές k, για τις περιοχές που περιέχουν θόρυβο και τις περιοχές που βρίσκονται κοντά στα όρια των κλάσεων.

Από την παρουσίαση της βιβλιογραφίας φάνηκε ότι οι τεχνικές που βασίζονται στη δυναμική απόδοση τιμής στην παράμετρο k ανάλογα με την υποπεριοχή που βρίσκεται το προς κατηγοριοποίηση αντικείμενο, φαίνεται ότι κυριαρχούν στη βιβλιογραφία με τους συγγραφείς να προσπαθούν να αποδείξουν ότι αποτελούν αποτελεσματικότερες προσεγγίσεις αφού κατορθώνουν να επιτύχουν ακρίβεια κατηγοριοποίησης που είναι υψηλότερη ακόμη και από την προσέγγιση που αποδίδει την στατική τιμή. Ωστόσο δεν είναι ξεκάθαρο αν στην πραγματικότητα είναι πιο αποτελεσματικές αφού δεν έγινε σε κανένα πείραμα η σύγκριση με την βέλτιστη τιμή που μπορεί να πάρει το k βάσει της μεθόδου cross-validation.

4.2 Μελλοντική Έρευνα

Από τη μελέτη που έγινε στη βιβλιογραφία ο κάθε ένας συγγραφέας προσπαθεί να αποδείξει ότι ο αλγόριθμος που προτείνει είναι καλύτερος από τον αλγόριθμο k-NN με στατικό k χωρίς όμως να λαμβάνεται υπόψη η βέλτιστη τιμή του k που αποδίδεται με την τεχνική cross-validation.

Επιπρόσθετα, παρά την πληθώρα των άρθρων που έχουν δημοσιευτεί, οι συγγραφείς που προτείνουν έναν νέο αλγόριθμος ή τεχνική απόδοσης τιμής στην παράμετρο k του κατηγοριοποιητή των k εγγύτερων γειτόνων, δεν λαμβάνουν υπόψη προγενέστερους αλγορίθμους και εξακολουθούν να συγκρίνονται με τον συμβατικό αλγόριθμο των k εγγύτερων γειτόνων αποδίδοντας σε αυτόν τυχαίες τιμές για την παράμετρο k.

Αρχικά θα ήταν χρήσιμο οι αλγόριθμοι που αποδίδουν δυναμική τιμή στην παράμετρο k να συγκριθούν με τον αλγόριθμο k-NN με σταθερό k το οποίο έχει οριστεί βάσει της μεθόδους cross-validation στα ίδια σύνολα δεδομένων. Οι αλγόριθμοι που επιτυγχάνουν υψηλότερη ακρίβεια είναι άξιοι αναφοράς και πιθανότατα αξίζει να ενσωματωθούν σε συστήματα κατηγοριοποίησης. Στη συνέχεια, οι αλγόριθμοι που επιτυγχάνουν υψηλότερη απόδοση, πρέπει να συγκριθούν μεταξύ τους στα ίδια σύνολα δεδομένων ώστε να βγει το συμπέρασμα ποια μέθοδος τελικά είναι αυτή που πετυχαίνει την καλύτερη ακρίβεια.

ΒΙΒΛΙΟΓΡΑΦΙΑ

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

[2]Mike James. Classification algorithms. Wiley-Interscience, New York, NY, USA, 1985.

[3]L. Rokach. Data Mining with Decision Trees: Theory and Applications. Series in machine perception and artificial intelligence. World Scientific Publishing Company, Incorporated, 2007

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

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

[6] Fadi Thabtah. A review of associative classification mining. Knowl. Eng. Rev., 22(1):37–65, March 2007

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

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

[9]Cover, T., Hart, P.: Nearest neighbor pattern classification. IEEE Transactions on Information Theory 13(1), 21–27, September 2006

[10] Wang, H. and Bell, D. Extended k-Nearest Neighbours Based on Evidence Theory. The Computer Journal, Vol. 47 (6) Nov. 2004, pp. 662-672

[11]Moreno-Seco, F., Mico, L. and Oncina, J. A Modification of the LAESA Algorithm for Approximated k-NN Classification. Pattern Recognition Letters 24 (2003), pp. 47–53

[12]Khan, M., Ding, Q. and Perrizo, W. K-Nearest Neighbors Classification of Spatial Data Streams using P-trees. Proceedings of the PAKDD, 2002, pp. 517-528

[13]Ougiaroglou, S., Nanopoulos, A., Papadopoulos, A.N., Manolopoulos, Y., WelzerDruzovec, T.: Adaptive k-Nearest-Neighbor Classification Using a Dynamic Number of Nearest Neighbors. In: Advances in Databases and Information Systems, pp. 66–82. Springer Berlin Heidelberg, Berlin, Heidelberg (2007).

[14]Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: An update. SIGKDD Explor. Newsl.,11(1):10–18, November 2009.

[15]Elena Deza and Michel M. Deza. Encyclopedia of Distances. Springer, Berlin, Heidelberg, 2009.

[16] C.C. Aggarwal. Data Streams: Models and Algorithms. Advances in Database Systems Series. Springer Science+Business Media, LLC, 2007

[17]Sahibsingh A. Dudani. The distance-weighted k-nearest-neighbor rule. Systems, Man and Cybernetics, IEEE Transactions on, SMC-6(4):325–327, 1976.

[18]Bulut, F., Amasyali, M.F.: Locally adaptive k parameter selection for nearest neighbor classifier: one nearest cluster. Pattern Analysis and Applications 20(2), 415–425

[19]Kardan, A.A., Kavian, A., Esmaeili, A.: Simultaneous feature selection and feature weighting with K selection for KNN classification using BBO algorithm. In: The 5th Conference on Information and Knowledge Technology. pp. 349–354.

[20]Ferrer-Troyano, F.J., Aguilar-Ruiz, J.S., Riquelme, J.C.: Non-parametric Nearest Neighbor with Local Adaptation. In: Brazdil, P., Jorge, A. (eds.) Progress in Artificial Intelligence. EPIA 2001. LNCS, vol. 2258, pp. 22–29. Springer, Berlin, Heidelberg (2001).

[21] Nock, R., Sebban, M., Bernard, D.: A Simple Locally Adaptive Nearest Neighbor Rule With Application To Pollution Forecasting. International Journal of Pattern Recognition and Artificial Intelligence 17(08), 1369–1382

[22]Holmes, C.C., Adams, N.M.: Journal of the Royal Statistical Society: Series B (Statistical Methodology)

[23]Enas, G.G., Choi, S.C.: Choice of the Smoothing Parameter and Efficiency of k-Nearest Neighbor Classification. In: Statistical Methods of Discrimination and Classification, vol. 12, pp. 235–244. Elsevier (1986)

[24]Johansson, U., K¨onig, R., Niklasson, L.: Evolving a locally optimized instance based learner. In: Proceedings of the 2008 International Conference on Data Mining, DMIN 2008. pp. 124–129

[25] Ougiaroglou, S., Evangelidis, G., Diamantaras, K.I.: Dynamic k-NN Classification Based on Subspace Homogeneity. In: Darmont, J., Novikov, B., Wrembel, R. (eds.) Communications in Computer and Information Science, vol. 1259 CCIS, pp. 27–37. Springer, Cham (2020)

[26] Sun, S., Huang, R.: An adaptive k-nearest neighbor algorithm. In: 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery. vol. 1, pp. 91–94. IEEE (aug 2010).

[27]Johansson, U., Bostr¨om, H., K¨onig, R.: Extending nearest neighbor classification with spheres of confidence. Proceedings of the 21th International Florida Artificial Intelligence Research Society Conference, FLAIRS-21 (Feb 2014), 282–287 (2008)

[28] Zhong, X.F., Guo, S.Z., Gao, L., Shan, H., Zheng, J.H.: An Improved k-NN Classification with Dynamic k. In: Proceedings of the 9th International Conference on Machine Learning and Computing. pp. 211–216. ACM, New York, NY, USA (feb 2017).

[29] Zhang, S., Zong, M., Sun, K., Liu, Y., Cheng, D.: Efficient kNN Algorithm Based on Graph Sparse Reconstruction. In: Luo, X., Yu, J., Li, Z. (eds.) Advanced Data Mining and Applications. ADMA 2014. LNCS, vol. 8933, pp. 356–369. Springer, Cham (2014)

[30] Zezula, P., Amato, G., Dohnal, V., Batko, M. (2006), Similarity Search - The Metric Space Approach, vol. 32. Springer, Heidelberg

[31] Samet, H. (2006), Foundations of multidimensional and metric data structures. The Morgan Kaufmann series in computer graphics. Elsevier,Morgan Kaufmann

[32] Manolopoulos Y., Nanopoulos, A. Papadopoulos, A. N., Theodoridis, Y (2006), "R-Tress: Theory and Applications", Springer

[33] Robinson J. T. (1981). The k-d-b-tree: a search structure for large multidimensional dynamic indexes. In Proceedings of the 1981 ACM SIGMOD international conference on Management of data, SIGMOD

[34] Fix, Evelyn, and Joseph Lawson Hodges. "Nonparametric discrimination: consistency properties." Randolph Field, Texas, Project (1951): 21-49.