ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ ΣΤΗΝ
ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

Ανίχνευση Ασυνήθιστης Συμπεριφοράς στην Εξόρυξη Δεδομένων

Παναγιώτης Χασάνης

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην *Εφαρμοσμένη Στατιστική*

Πειραιάς

Ιανουάριος 2019

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ ΣΤΗΝ
ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

Ανίχνευση Ασυνήθιστης Συμπεριφοράς στην Εξόρυξη Δεδομένων

Παναγιώτης Χασάνης

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην *Εφαρμοσμένη Στατιστική*

Πειραιάς

Ιανουάριος 2019

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη ΓΣΕΣ του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. …… συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

Τα μέλη της Επιτροπής ήταν:

- Καθηγητής Κούτρας Μάρκος (Επιβλέπων)
- Καθηγητής Θεοδωρίδης Ιωάννης
- Αναπληρωτής Καθηγητής Κοφίδης Ελευθέριος

Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

UNIVERSITY OF PIRAEUS

School of Finance and Statistics

Department of Statistics and Insurance Science

POSTGRADUATE PROGRAM IN

APPLIED STATISTICS

Anomaly Detection in Data Mining

Panagiotis Chasanis

Msc Dissertation

Submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus, Greece

January 2019

Στους γονείς μου

Ιωάννη και Ευδοκία

Ευχαριστίες

Αυτή η διπλωματική εργασία εκπονήθηκε στο πλαίσιο του μεταπτυχιακού προγράμματος «Εφαρμοσμένη Στατιστική» του τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς. Ευχαριστώ από καρδιάς όλους εκείνους που χωρίς την πολύτιμη βοήθειά τους θα ήταν αδύνατον να υλοποιηθούν οι στόχοι μου.

Ευχαριστώ τους καθηγητή Μάρκο Κούτρα, καθηγητή Ιωάννη Θεοδωρίδη και αναπληρωτή καθηγητή Ελευθέριο Κοφίδη, οι οποίοι δέχθηκαν να είναι μέλη της επιτροπής αξιολόγησης. Ευχαριστώ τους καθηγητή Ιωάννη Θεοδωρίδη και αναπληρωτή καθηγητή Ελευθέριο Κοφίδη που με τις σημαντικές υποδείξεις τους συνέβαλαν ουσιαστικά στην ολοκλήρωση της έρευνάς μου.

Ιδιαιτέρως ευχαριστώ τον καθηγητή Στατιστικής και Ασφαλιστικής Επιστήμης στο Πανεπιστήμιο Πειραιώς Μάρκο Κούτρα, ο οποίος μου πρότεινε το ερευνητικό θέμα, υποστήριξε ένθερμα της προσπάθειά μου, με ενέπνευσε και με εμπιστεύθηκε, επιβλέποντας ακούραστα την εργασία μου καθ’ όλη τη διάρκεια της πραγματοποίησής της.

Τέλος αισθάνομαι την ανάγκη να εκφράσω την ευγνωμοσύνη μου στον πατέρα μου Ιωάννη Χασάνη και στη μητέρα μου Ευδοκία Πατσιλινάκου για την υποστήριξή τους σε κάθε βήμα στη ζωή μου.

Παναγιώτης Χασάνης

Περίληψη

Ο όρος ανίχνευση ασυνήθιστης συμπεριφοράς ή ανωμαλιών χρησιμοποιείται για να περιγράψει τεχνικές με τις οποίες εντοπίζονται αντικείμενα, γεγονότα ή παρατηρήσεις οι οποίες δεν ακολουθούν τους γενικούς κανόνες των δεδομένων που παρατηρούμε και χαρακτηρίζονται ως ανωμαλίες. Οι ανωμαλίες προκύπτουν εξαιτίας μηχανικών βλαβών, αλλαγές στη συμπεριφορά του συστήματος, δόλια συμπεριφορά, ανθρώπινο σφάλμα, σφάλμα οργάνου ή απλά μέσω φυσικών αποκλίσεων στους πληθυσμούς. Η έγκαιρη ανίχνευση των ανωμαλιών μπορεί να εντοπίσει τα σφάλματα του συστήματος και την απάτη πριν κλιμακωθούν με δυνητικά καταστροφικές συνέπειες.

Πολλές τεχνικές ανίχνευσης ασυνήθιστης συμπεριφοράς έχουν αναπτυχθεί ειδικά για ορισμένους τομείς εφαρμογής, ενώ άλλες είναι πιο γενικές. Στη παρούσα διπλωματική εργασία προσπαθούμε να παρέχουμε μια δομημένη και ολοκληρωμένη επισκόπηση της έρευνας σχετικά με την ανίχνευση ασυνήθιστης συμπεριφοράς. Για τις διαφορετικές κατηγορίες τεχνικών ανίχνευσης ανωμαλιών παρουσιάζεται η βασική τεχνική και οι παραδοχές, οι οποίες χρησιμοποιούνται για τη διαφοροποίηση μεταξύ κανονικής και ανώμαλης συμπεριφοράς. Επίσης, για κάθε κατηγορία παρουσιάζονται διάφορες μέθοδοι ανίχνευσης ανωμαλιών, οι οποίες στηρίζονται στη αντίστοιχη βασική τεχνική, και προσδιορίζονται τα πλεονεκτήματα και τα μειονεκτήματά τους. Τέλος, γίνεται εφαρμογή και σύγκριση της απόδοσης διάφορων μεθόδων ανίχνευσης ασυνήθιστης συμπεριφοράς.

Abstract

The term outlier or anomaly detection is used to describe techniques that detect objects, events or observations that do not follow the general rules of the data we observe and are characterized as anomalies. Anomalies arise due to mechanical faults, changes in system behavior, fraudulent behavior, human error, instrument error or simply through natural deviations in populations. Their detection can identify system faults and fraud before they escalate with potentially catastrophic consequences.

Many anomaly detection techniques have been specifically developed for certain application domains, while others are more generic. This thesis tries to provide a structured and comprehensive overview of the research on anomaly detection. For each category of anomaly detection techniques, the basic technique and the assumptions used to differentiate between normal and anomaly behavior are presented. In addition, for each category various methods of anomaly detection, which are based on the corresponding basic technique are presented, and their advantages and disadvantages are determined. Finally, various methods are applied, and their performance in detecting anomalies is compared.

Περιεχόμενα

ΠΕΡΙΕΧΟΜΕΝΑ XIII
ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ XV
ΚΕΦΑΛΑΙΟ 1: ΕΙΣΑΓΩΓΗ 1
1.1 Ορισμός και χρήσεις 1
1.2 Κατηγοριοποίηση Τεχνικών 2
1.2.1 Κατηγοριοποίηση με βάση τις ετικέτες των δεδομένων 2
1.2.2 Κατηγοριοποίηση με βάση το αποτέλεσμα 4
1.3 Είδη Ανωμαλιών 4
1.4 Δομή παρούσας εργασίας 9
ΚΕΦΑΛΑΙΟ 2: ΣΤΑΤΙΣΤΙΚΕΣ ΜΕΘΟΔΟΙ ΑΝΙΧΝΕΥΣΗΣ ΑΝΩΜΑΛΙΩΝ 10
2.1 Περιγραφή 10
2.2 Παραμετρικές Μέθοδοι 10
2.2.1 Μέθοδοι βασισμένες σε μοντέλα Κανονικών Κατανομών 11
2.2.2 Μέθοδοι βασισμένες σε μοντέλα παλινδρόμησης 13
2.2.3 Μέθοδοι βασισμένες σε μίξεις παραμετρικών κατανομών 15
2.3 Μη-Παραμετρικές Μέθοδοι 17
2.3.1 Μέθοδοι βασισμένες σε ιστογράμματα 17
2.3.2 Μέθοδοι βασισμένες σε συνάρτηση πυρήνα (kernel function) 19
2.5 Πλεονεκτήματα και Μειονεκτήματα Στατιστικών Μεθόδων 22
ΚΕΦΑΛΑΙΟ 3: ΜΕΘΟΔΟΙ AD ΒΑΣΙΣΜΕΝΕΣ ΣΤΗΝ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ 23
3.1 Περιγραφή 23
3.2 Μέθοδοι βασισμένες σε Νευρωνικά Δίκτυα 24
3.2 Μέθοδοι βασισμένες σε Μπεϋζιανά Δίκτυα 26
3.3 Μέθοδοι βασισμένες σε Μηχανές Διανυσμάτων Υποστήριξης (SVM) 27
3.4 Μέθοδοι βασισμένες σε Κανόνες 30
3.5 Πλεονεκτήματα και Μειονεκτήματα Μεθόδων Βασισμένες στην Κατηγοριοποίηση 32
ΚΕΦΑΛΑΙΟ 4: ΜΕΘΟΔΟΙ AD ΒΑΣΙΣΜΕΝΕΣ ΣΤΟΝ ΠΛΗΣΙΕΣΤΕΡΟ ΓΕΙΤΟΝΑ 34

4.1 Περιγραφή 34
4.2 Τεχνικές που αξιοποιούν τον k-οστό πλησιέστερο γείτονα 35
4.3 Τεχνικές που βασίζονται στη σχετική πυκνότητα 37
4.4 Πλεονεκτήματα και Μειονεκτήματα Μεθόδων Βασισμένες στον Πλησιέστερο Γείτονα. . 42
ΚΕΦΑΛΑΙΟ 5: ΜΕΘΟΔΟΙ AD ΒΑΣΙΣΜΕΝΕΣ ΣΤΗΝ ΣΥΣΤΑΔΟΠΟΙΗΣΗ 43
5.1 Περιγραφή 43
5.2 Τεχνικές Ανίχνευσης Ανωμαλιών 44
5.3 Πλεονεκτήματα και Μειονεκτήματα Μεθόδων Βασισμένες στη Συσταδοποίηση 50
ΚΕΦΑΛΑΙΟ 6: ΜΕΘΟΔΟΙ AD ΓΙΑ ΔΕΔΟΜΕΝΑ ΜΕΓΑΛΗΣ ΔΙΑΣΤΑΣΗΣ 52
6.1 Περιγραφή 52
6.2 Η Κατάρα της Διάστασης 53
6.3 Μέθοδοι Ανίχνευσης Ανωμαλιών 55
6.4 Πλεονεκτήματα και Μειονεκτήματα Μεθόδων AD για Δεδομένα Μεγάλης Διάστασης ... 62
ΚΕΦΑΛΑΙΟ 7: ΕΦΑΡΜΟΓΗ 63
7.1 Περιγραφή 63
7.1.1 Μέθοδος Αξιολόγησης Αλγορίθμων 63
7.1.2 Διαδικασία Παραλληλοποίησης και Εργαλεία Εφαρμογής 64
7.2 Αποτελέσματα Αλγορίθμων 65
7.3 Σύνοψη – Συμπεράσματα 67
ΠΑΡΑΡΤΗΜΑΤΑ 68
Παράρτημα Α: Μεγάλο Ο 68
Παράρτημα Β: Κώδικας Εφαρμογής Αλγορίθμων 70
ΒΙΒΛΙΟΓΡΑΦΊΑ 77

Κατάλογος Σχημάτων

Σχήμα 1.1: *Κατηγοριοποίηση μεθόδων ανίχνευσης ασυνήθιστης συμπεριφοράς με βάση τη διαθεσιμότητα των ετικετών των δεδομένων* 3

Σχήμα 1.2: *Ένα απλό διδιάστατο παράδειγμα. Απεικονίζει δύο ολικές ανωμαλίες* (x_1, x_2)*, μια τοπική ανωμαλία* (x_3) *και την μικρό-συστάδα* C_3 5

Σχήμα 1.3: *Παράδειγμα συλλογικής ανωμαλίας σε ανθρώπινο εγκεφαλογράφημα*.....7

Σχήμα 2.1: *Παράδειγμα διαγράμματος Box-plot* 11

Σχήμα 2.2: *Παράδειγμα κλασσικής μεθόδου OLS (κόκκινη γραμμή) και ανθεκτικής μεθόδου LTS (μπλε γραμμή)*.....14

Σχήμα 2.3: *Παράδειγμα χρήσης ιστογράμματος για την ανίχνευση ανωμαλιών* 18

Σχήμα 3.1: *Είδη τεχνικών κατηγοριοποίησης για την ανίχνευση ανωμαλιών* 24

Σχήμα 3.2: *Γραφική απεικόνιση ενός πλήρως συνδεδεμένου RNN, με τρία κρυμμένα επίπεδα*.....25

Σχήμα 3.3: *Γραφική αναπαράσταση Μπεϋζιανού δικτύου* 26

Σχήμα 3.4: *Γραφική αναπαράσταση απεικόνιση του χώρου εισόδου σε ένα χώρο χαρακτηριστικών μεγαλύτερης διάσταση* 28

Σχήμα 4.1: *Απόσταση προσβασιμότητας και* k*-distance, για* k*=4*.....38

Σχήμα 5.1: *Συσταδοποίηση με τον αλγόριθμο DBSCAN για* $MinPts = 4$ 45

Σχήμα 6.1: *Παράδειγμα ανωμαλίας η οποία εντοπίζεται μόνο στον πλήρη χώρο*.....62

Σχήμα 7.1: *Τιμές A.U.C. των Αλγορίθμων, για* $10 \leq k \leq 50$ 66

Κεφάλαιο 1

Εισαγωγή

1.1 Ορισμός και χρήσεις

Η ανίχνευση ασυνήθιστης συμπεριφοράς ή ανωμαλιών αναφέρεται στο πρόβλημα εύρεσης μοτίβων ή αντικειμένων τα οποία δεν συνάδουν με την αναμενόμενη ή συνήθη συμπεριφορά. Αυτές οι μη συμμορφούμενες παρατηρήσεις συχνά αναφέρονται ως ανωμαλίες, εξωκείμενες, νεοτερισμοί, ασύμβατες, εκπλήξεις, ιδιομορφίες, αποκλίσεις, ή ρύποι, ανάλογα με το εκάστοτε πεδίο εφαρμογής. Οι πιο συνηθισμένοι όροι που χρησιμοποιούνται στο πλαίσιο της ανίχνευσης ασυνήθιστης συμπεριφοράς είναι ανωμαλίες (*anomalies*) και εξωκείμενες παρατηρήσεις (*outliers*).

Αρκετοί ερευνητές έχουν δώσει το δικό τους ορισμό για το τι είναι μια εξωκείμενη παρατήρηση και φαινομενικά δεν υπάρχει κάποιος ορισμός παγκοσμίως αποδεκτός. Ο πρώτος ορισμός, πιθανότατα, δόθηκε από τον Grubbs (1969): *"Μια απόμακρη ή εξωκείμενη παρατήρηση, είναι κάποια η οποία φαίνεται να αποκλίνει σημαντικά από άλλα μέλη του δείγματος στο οποίο ανήκει"*, και γενικεύτηκε αργότερα από τους Barnett και Lewis (1994): *"Εξωκείμενη παρατήρηση είναι μια παρατήρηση (ή μια υποομάδα από παρατηρήσεις) η οποία φαίνεται να είναι ασύμφωνη με το υπόλοιπο σύνολο των δεδομένων"*. Ο Hawkins (1980) ορίζει ως εξωκείμενη παρατήρηση, "*την παρατήρηση που παρεκκλίνει τόσο πολύ από τις υπόλοιπες παρατηρήσεις ώστε να κινήσει την υποψία ότι έχει παραχθεί από διαφορετικό μηχανισμό"*. Ένας ακριβής ορισμός για το τι είναι εξωκείμενη παρατήρηση είναι δύσκολο να δοθεί, καθώς το τι θεωρείται ανωμαλία στα δεδομένα συχνά εξαρτιέται από την δομή των δεδομένων και την εκάστοτε εφαρμογή.

Η ανίχνευση ασυνήθιστης συμπεριφοράς (*Anomaly Detection – AD)* χρησιμοποιείται σε πολλούς εφαρμοσμένους κλάδους, όπως στην ανίχνευση ανωμαλιών σε δορυφορικές εικόνες, σε κλειστά συστήματα παρακολούθησης, σε αιτήσεις δανείων σε τράπεζες, σε ιατρικές διαγνώσεις, σε φαρμακευτικές έρευνες και στην παρακολούθηση χρονοσειρών. Μια

εξωκείμενη παρατήρηση μπορεί να σηματοδοτεί εισβολή σε ένα σύστημα με κακόβουλες προθέσεις, οπότε η γρήγορη ανίχνευση είναι αναγκαία. Με χρήση AD μπορεί να γίνει εντοπισμός σφαλμάτων στη γραμμή παραγωγής ενός εργοστασίου με συνεχή παρακολούθηση συγκεκριμένων χαρακτηριστικών των προϊόντων. Ακόμη, η εφαρμογή της AD είναι ιδιαίτερα χρήσιμη για τη παρακολούθηση της χρήσης πιστωτικών καρτών ή κινητών τηλεφώνων, αφού η ανίχνευση ξαφνικής αλλαγής στο μοτίβο της χρήσης, μπορεί να υποδεικνύει πιθανή κλοπή. Επενδυτές μπορούν να αξιοποιήσουν μεθόδους της AD για την παρακολούθηση μεμονωμένων μετοχών ή αγορών για τον εντοπισμό καινούργιων τάσεων που μπορεί να σημαίνουν ευκαιρίες αγοράς ή πώλησης. Σε μια βάση δεδομένων, ανωμαλίες μπορεί να υποδεικνύουν προσπάθεια εξαπάτησης, ανθρώπινο λάθος κατά τη καταχώριση δεδομένων ή παρερμήνευση μιας ελλιπούς τιμής. Σε κάθε περίπτωση ο έγκαιρος εντοπισμός πιθανών ανωμαλιών είναι ζωτικής σημασίας για τη συνέπεια και την αξιοπιστία μιας βάσης δεδομένων.

1.2 Κατηγοριοποίηση Τεχνικών

1.2.1 Κατηγοριοποίηση με βάση τις ετικέτες των δεδομένων

Οι ετικέτες (*labels*) των δεδομένων υποδηλώνουν αν το εκάστοτε δεδομένο θεωρείται κανονικό ή μη-κανονικό. Αξίζει να σημειωθεί ότι η απόκτηση δεδομένων με ακριβείς και αντιπροσωπευτικές ετικέτες από κάθε είδους συμπεριφορά είναι σπάνια εφικτή. Κατά κανόνα, η λήψη ενός συνόλου μη-κανονικών δεδομένων που να καλύπτει κάθε πιθανό είδος ασυνήθιστης συμπεριφοράς είναι πολύ πιο δύσκολη από τη λήψη ενός συνόλου δεδομένων που να περιγράφουν την κανονική συμπεριφορά. Επιπρόσθετα, η ασυνήθιστη συμπεριφορά συχνά είναι δυναμική στο πραγματικό κόσμο, για παράδειγμα μπορεί να δημιουργούνται καινούργια είδη ανωμαλιών για τα οποία να μην υπάρχει ετικέτα στα δεδομένα εκμάθησης (*training data*).

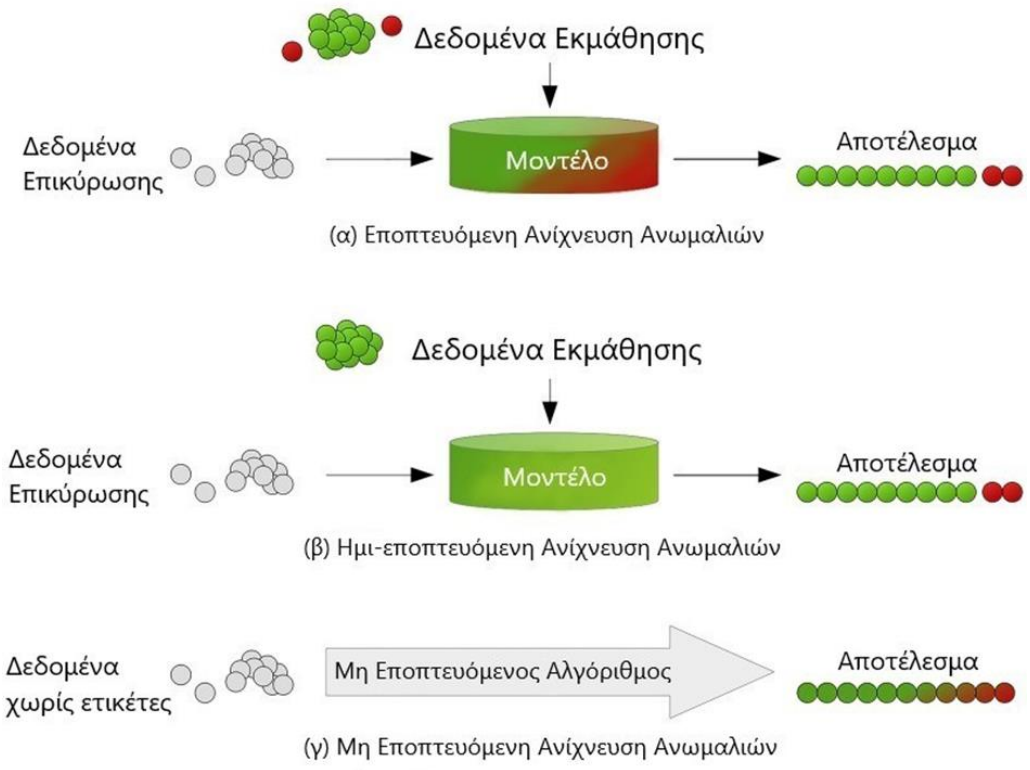
Ανάλογα με το αν οι ετικέτες των δεδομένων είναι διαθέσιμες, οι τεχνικές για την ανίχνευση ανωμαλιών μπορούν να κατηγοριοποιηθούν στις εξής τρεις κατηγορίες, όπως φαίνεται και από το Σχήμα 1.1:

1. Εποπτευόμενη Ανίχνευση Ανωμαλιών (*Supervised Anomaly Detection*): περιγράφει τη περίπτωση όπου στα δεδομένα εκμάθησης και επικύρωσης (*test data*) υπάρχουν διαθέσιμες ετικέτες. Αυτή η περίπτωση δεν είναι πολύ σημαντική από πρακτική άποψη λόγω των υποθέσεων ότι οι ανωμαλίες είναι γνωστές και έχουν επισημανθεί σωστά. Σε

πολλές εφαρμογές, οι ανωμαλίες δεν είναι εκ των προτέρων γνωστές ή μπορεί να παρουσιαστούν καινούργιες ανωμαλίες οι οποίες δεν ήταν γνωστές κατά τη διάρκεια της φάσης της επικύρωσης.

2. Ημι-εποπτευόμενη Ανίχνευση Ανωμαλιών (*Semi-supervised Anomaly Detection*): και σε αυτή την περίπτωση χρησιμοποιούνται δεδομένα εκμάθησης και επικύρωσης, όπου τα δεδομένα εκμάθησης αποτελούνται μόνο από κανονικά δεδομένα, χωρίς ανώμαλες παρατηρήσεις. Η βασική ιδέα είναι ότι δημιουργείται ένα μοντέλο που περιγράφει την συμπεριφορά των κανονικών παρατηρήσεων και στη συνέχεια οι ανωμαλίες ανιχνεύονται, καθώς θα αποκλίνουν από το μοντέλο.
3. Μη Εποπτευόμενη Ανίχνευση Ανωμαλιών (*Unsupervised Anomaly Detection*): είναι η πιο ευέλικτη περίπτωση, η οποία δεν απαιτεί καθόλου ετικέτες και δεν υπάρχει διάκριση μεταξύ δεδομένων εκμάθησης και επικύρωσης. Η ιδέα εδώ είναι ότι ένας μη εποπτευόμενος αλγόριθμος ανίχνευσης ανωμαλιών βαθμολογεί τα δεδομένα με βάση μόνο εγγενείς ιδιότητες του συνόλου των δεδομένων.

Σχήμα 1.1: *Κατηγοριοποίηση μεθόδων ανίχνευσης ασυνήθιστης συμπεριφοράς με βάση τη διαθεσιμότητα των ετικετών των δεδομένων.*



1.2.2 Κατηγοριοποίηση με βάση το αποτέλεσμα

Ένα σημαντικό χαρακτηριστικό μιας τεχνικής ανίχνευσης ασυνήθιστης συμπεριφοράς είναι ο τρόπος με τον οποίο παρουσιάζονται οι ανωμαλίες. Γενικά, τα αποτελέσματα που παράγονται από μια τεχνική ανίχνευσης συμπεριφοράς είναι τα εξής δύο:

1) Βαθμολογίες (*Scores*): Οι τεχνικές αναθέτουν σε κάθε δεδομένο μια βαθμολογία ανάλογα με το πόσο το κάθε δεδομένο μπορεί να θεωρηθεί ανωμαλία, δηλαδή το αποτέλεσμα αυτών των τεχνικών είναι μια λίστα κατάταξης ανωμαλιών. Ο αναλυτής μπορεί να επιλέξει είτε να αναλύσει μερικές ανωμαλίες από την κορυφή της λίστας είτε να διαλέγει τις υπό μελέτη ανωμαλίες με κάποιο όριο αποκοπής.
2) Ετικέτες (*Labels*): Οι τεχνικές αυτής της κατηγορίας αναθέτουν μια ετικέτα (κανονικό ή μη-κανονικό) σε κάθε υπό μελέτη δεδομένο.

Για τις τεχνικές εποπτευόμενης ανίχνευσης ανωμαλιών συχνά χρησιμοποιούνται ετικέτες, λόγω της διαθεσιμότητας αλγορίθμων κατηγοριοποίησης, ενώ για τις τεχνικές ημι-εποπτευόμενης και μη εποπτευόμενης ανίχνευσης ανωμαλιών είναι πιο συνηθισμένη η χρήση βαθμολογιών. Αυτό πρακτικά οφείλεται στο γεγονός ότι οι εφαρμογές συνήθως κατατάσσουν τις ανωμαλίες και αναφέρουν στον αναλυτή μόνο όσες βρίσκονται στη κορυφή. Φυσικά, με τη χρήση κατάλληλου ορίου, μια κατάταξη μπορεί να μετατραπεί σε διαχωρισμό με ετικέτες. [Goldstein and Uchida (2016)]

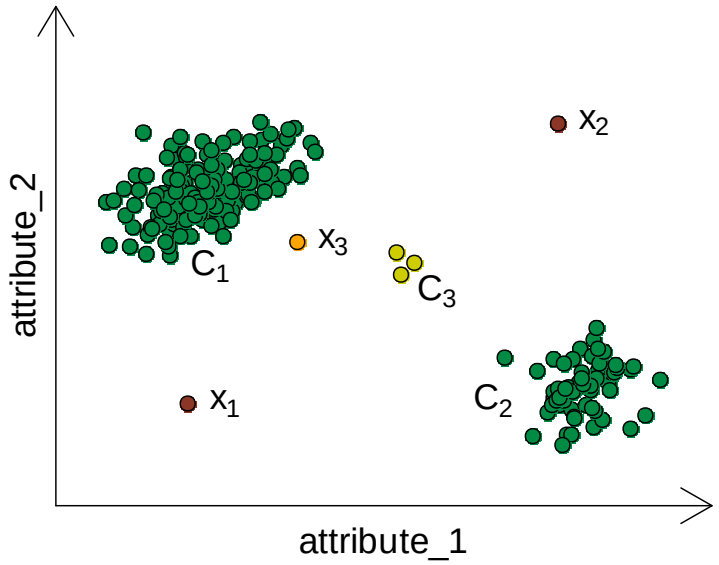
Σε προβλήματα εποπτευόμενης ανίχνευσης ανωμαλιών, όταν παράγεται για κάθε παρατήρηση βαθμός ανωμαλίας, το όριο, ή κατώφλι, που συνήθως επιλέγεται είναι ο μικρότερος βαθμός ανωμαλίας από τις ανώμαλες παρατηρήσεις του συνόλου εκμάθησης. Από την άλλη, στα προβλήματα ημι-εποπτευόμενης ανίχνευσης το κατώφλι, που συνήθως επιλέγεται είναι ο μεγαλύτερος βαθμός ανωμαλίας των κανονικών παρατηρήσεων στο σύνολο δεδομένων εκμάθησης. Τέλος, στα προβλήματα μη εποπτευόμενης ανίχνευσης ανωμαλιών, επειδή συνήθως ως ανωμαλίες θεωρείται ένας μικρός αριθμός παρατηρήσεων με τους μεγαλύτερους βαθμούς ανωμαλίας, το όριο που επιλέγεται είναι ο μικρότερος βαθμός ανωμαλίας από αυτές τις παρατηρήσεις.

1.3 Είδη Ανωμαλιών

Η βασική υπόθεση στις τεχνικές της μη εποπτευόμενης ανίχνευσης ανωμαλιών είναι η εύρεση δεδομένων που αποκλίνουν από το κανονικό. Ωστόσο, υπάρχουν αρκετές περιπτώσεις

όπου αυτή η υπόθεση είναι ασαφής. Στο Σχήμα 1.2 παρουσιάζονται τέτοιες περιπτώσεις με τη χρήση ενός απλού συνόλου διδιάστατων δεδομένων. Οι παρατηρήσεις x_1και x_2 μπορούν εύκολα να αναγνωριστούν ως ανωμαλίες, καθώς αποκλίνουν πολύ από τις πυκνές περιοχές που δημιουργούνται βάσει των δύο χαρακτηριστικών, και καλούνται ολικές ανωμαλίες (*global anomalies*). Όταν μελετάμε τα δεδομένα συνολικά η παρατήρηση x_3 μπορεί να χαρακτηριστεί ως κανονική παρατήρηση, καθώς δεν απέχει πολύ από την συστάδα C_1. Ωστόσο, επικεντρώνοντας την προσοχή μας μόνο στην C_1 και αγνοώντας τις υπόλοιπες παρατηρήσεις, η παρατήρηση x_3 μπορεί να θεωρηθεί ως ανωμαλία, και καλείται τοπική ανωμαλία (*local anomaly*), καθώς είναι ανωμαλία μόνο σε σύγκριση με την κοντινή της γειτονιά. Ένα ενδιαφέρον ερώτημα είναι αν οι παρατηρήσεις της συστάδας C_3 πρέπει να αντιμετωπιστούν ως τρείς ανωμαλίες ή ως μια μικρή κανονική συστάδα. Τέτοιες ομάδες καλούνται μικρό-συστάδες (*micro clusters*) και οι αλγόριθμοι ανίχνευσης ανωμαλιών θα πρέπει να αναθέτουν στα μέλη τους βαθμολογίες μεγαλύτερες των κανονικών παρατηρήσεων, αλλά μικρότερες των προφανών ανωμαλιών. Από το προηγούμενο παράδειγμα γίνεται φανερό ότι οι ανωμαλίες δεν είναι πάντα προφανείς και μια βαθμολογία είναι πολύ καλύτερη από μια απλή δίτιμη ανάθεση ετικετών.[Goldstein and Uchida (2016)]

Σχήμα 1.2: *Ένα απλό διδιάστατο παράδειγμα. Απεικονίζει δύο ολικές ανωμαλίες* (x_1, x_2)*, μια τοπική ανωμαλία* (x_3) *και την μικρό-συστάδα* C_3.



Αρκετά συχνά, ως ανωμαλία σε ένα σύνολο δεδομένων αναφέρεται μια μεμονωμένη παρατήρηση (*single instance*) που προκύπτει σπάνια. Στην πραγματικότητα, όμως, αυτό δεν ισχύει πάντα, για παράδειγμα, στην ανίχνευση εισβολής (*intrusion detection*) ανωμαλίες συχνά θεωρούνται πολλά ύποπτα μοτίβα πρόσβασης που μπορούν να παρατηρηθούν σε μεγαλύτερη συχνότητα από ό,τι η κανονική σύνδεση. Οι ανωμαλίες μπορούν να ομαδοποιηθούν, ανάλογα με την φύση τους, στις εξής τρεις κατηγορίες:

1) Σημειακές Ανωμαλίες (*Point Anomalies*): Αν μια μεμονωμένη παρατήρηση θεωρείται ως ανωμαλία, σε σχέση με τα υπόλοιπα δεδομένα, τότε ονομάζεται σημειακή ανωμαλία. Οι σημειακές ανωμαλίες είναι η πιο απλή κατηγορία και αντικείμενο μελέτης της πλειοψηφίας των ερευνητών.

   Στο Σχήμα 1.2, τα σημεία x_1, x_2, x_3, καθώς και τα σημεία της C_3, βρίσκονται εκτός των ορίων των κανονικών περιοχών και χαρακτηρίζονται ως σημειακές ανωμαλίες, καθώς διαφέρουν από τα κανονικά σημεία. Στην παρακολούθηση της χρήσης πιστωτικών καρτών, ας υποθέσουμε ότι τα δεδομένα προσδιορίζονται μόνο από ένα χαρακτηριστικό, το ποσό που ξοδεύεται. Μια συναλλαγή στην οποία το ποσό που ξοδεύεται είναι πολύ μεγαλύτερο σε σχέση με το εύρος των κανονικών συναλλαγών του κατόχου της πιστωτικής, θα θεωρείται σημειακή ανωμαλία.

2) Συναφείς Ανωμαλίες (*Contextual Anomalies*): Οι συναφείς ανωμαλίες ή υπό όρους ανωμαλίες (*conditional anomalies*) περιγράφουν το γεγονός που κάποιο δεδομένο μπορεί φαινομενικά να είναι κανονικό, αλλά λαμβάνοντας υπόψιν τα συμφραζόμενα να θεωρείται ανωμαλία. [Song et al. (2007)]

   Η έννοια του συμφραζόμενου προκαλείται από τη δομή του συνόλου των δεδομένων και πρέπει να έχει προσδιοριστεί ως κομμάτι της διατύπωσης του προβλήματος. Κάθε δεδομένο ορίζεται με τη χρήση των εξής δύο συνόλων από χαρακτηριστικά:

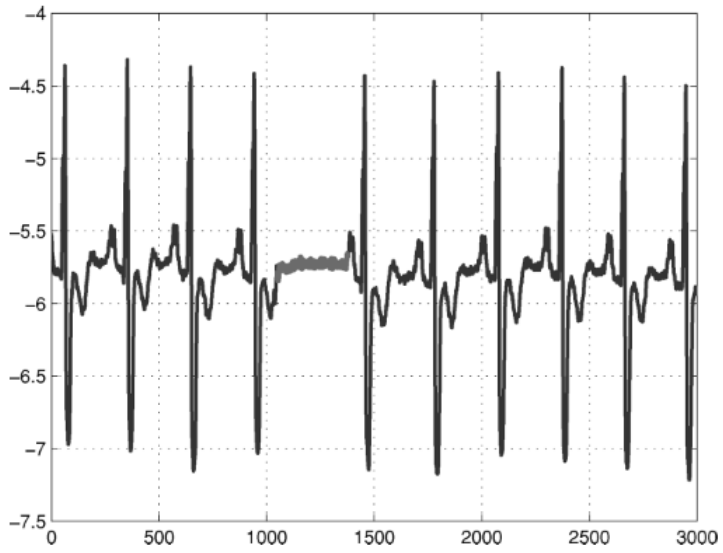
   (i) Συμφραζόμενα χαρακτηριστικά (*Contextual Attributes*). Χρησιμοποιούνται για να προσδιορίσουν το πλαίσιο (ή τη γειτονιά) της κάθε παρατήρησης, για παράδειγμα, στις χρονοσειρές ο χρόνος είναι ένα συμφραζόμενο χαρακτηριστικό που προσδιορίζει τη θέση της κάθε παρατήρησης στη χρονοσειρά.

   (ii) Χαρακτηριστικά συμπεριφοράς (*Behavioral Attributes*). Χρησιμοποιούνται για να προσδιορίσουν μη συμφραζόμενα χαρακτηριστικά μιας παρατήρησης, για παράδειγμα, σε ένα χωρικό σύνολο δεδομένων (*spatial dataset*) που περιγράφει τη

μέση βροχόπτωση σε όλη τη γη, το ποσό βροχόπτωσης σε κάθε περιοχή θεωρείται χαρακτηριστικό συμπεριφοράς.

Η ασυνήθιστη συμπεριφορά καθορίζεται χρησιμοποιώντας τις τιμές των χαρακτηριστικών συμπεριφοράς εντός ενός συγκεκριμένου πλαισίου. Το πιο συνηθισμένο πλαίσιο είναι ο χρόνος. Ας υποθέσουμε ότι μελετάμε τη θερμοκρασία κατά τη διάρκεια του έτους. Κατά μέσο όρο η θερμοκρασία κυμαίνεται από 0^{o} έως 35^{o}C, οπότε μια μέτρηση των 26^{o}C θεωρείται κανονική. Όμως, όταν συνυπολογίσουμε το πλαίσιο του χρόνου, μια τόσο υψηλή θερμοκρασία θα θεωρείται ως ανωμαλία τη χειμερινή περίοδο.

Σχήμα 1.3: *Παράδειγμα συλλογικής ανωμαλίας σε ανθρώπινο εγκεφαλογράφημα.* [Chandola et al. (2009)]



3) Συλλογικές Ανωμαλίες (*Collective Anomalies*): Αν μια ασυνήθιστη συμπεριφορά περιγράφεται από ένα σύνολο πολλών παρατηρήσεων, τότε καλείται συλλογική ανωμαλία. Στις συλλογικές ανωμαλίες, οι μεμονωμένες παρατηρήσεις μπορεί να μην θεωρούνται ως ανωμαλίες, αλλά η συνύπαρξή τους ως σύνολο να θεωρείται ασυνήθιστη συμπεριφορά. Στο Σχήμα 1.3 παρουσιάζεται το αποτέλεσμα ενός ανθρώπινου εγκεφαλογραφήματος, όπου η γκρίζα περιοχή υποδηλώνει μια ανωμαλία, καθώς η ίδια χαμηλή τιμή εμφανίζεται για ασυνήθιστα μεγάλη διάρκεια. Ωστόσο, η συγκεκριμένη χαμηλή τιμή από μόνη της, δεν θεωρείται ανωμαλία. [Chandola et al. (2009)]

Οι αλγόριθμοι ανίχνευσης ανωμαλιών επεξεργάζονται πάντοτε τα δεδομένα για την εύρεση σημειακών ανωμαλιών. Στα προβλήματα ανίχνευσης πιο περίπλοκων ανωμαλιών, όπως είναι οι συναφείς και οι συλλογικές ανωμαλίες, το υπό μελέτη σύνολο δεδομένων μετασχηματίζεται σε ένα νέο σύνολο δεδομένων, στο οποίο τα χαρακτηριστικά και τα δεδομένα μπορεί να είναι τελείως διαφορετικά από τα αρχικά. Ο μετασχηματισμός αυτός καλείται αναπαράσταση δεδομένων (*data view*) και επηρεάζει το είδος ανωμαλιών που ανιχνεύεται κάθε φορά, ειδικά στα προβλήματα μη εποπτευόμενης ανίχνευσης ανωμαλιών. [Goldstein and Uchida (2016)]

Το πρώτο βήμα είναι ο προσδιορισμός της οντότητας (*entity*) για την οποία ζητείται εντοπισμός ανωμαλιών. Σε αυτό το πλαίσιο, μια οντότητα είναι το συμβάν ή αντικείμενο, στο οποίο αναφέρεται μια ανωμαλία. Για παράδειγμα, ας υποθέσουμε ότι μετράμε την κατανάλωση ενέργειας ενός κτηρίου. Μια οντότητα θα μπορούσε να είναι μια μεμονωμένη συσκευή, αν θέλαμε τον εντοπισμό ασυνήθιστης κατανάλωσης ποσότητας ενέργειας, ενώ μια άλλη οντότητα είναι η συνάθροιση της συνολικής κατανάλωσης όλων των συσκευών εντός ενός χρονικού πλαισίου, για τον έλεγχο αν ολόκληρο το κτίριο χρησιμοποιεί ύποπτη ποσότητα ενέργειας. Ο προσδιορισμός της οντότητας απαντά σιωπηρά στο ερώτημα, σε ποιο είδος ανωμαλίας πρέπει να επικεντρωθούμε κάθε φορά.

Για την αντιμετώπιση προβλημάτων ανίχνευσης συναφών ανωμαλιών, το πλαίσιο (*context*) πρέπει να ενσωματωθεί με τη συνάθροιση (*aggregation*) και τη διακριτοποίηση (*discretization/ binning*) των δεδομένων. Συγκεκριμένα, για το πλαίσιο του χρόνου, το πιο σύνηθες είναι η συνάθροιση γεγονότων για ένα συγκεκριμένο χρονικό διάστημα, παραδείγματος χάριν την κατανάλωση ενέργειας μέσα σε μια ώρα. Φυσικά, πολλαπλά πλαίσια μπορεί να ληφθούν υπόψη. Για παράδειγμα, για τη συνάθροιση της συνολικής κατανάλωσης ενέργειας όλων των συσκευών μέσα σε ένα δωμάτιο για μια ώρα, απαιτούνται δύο πλαίσια, ο χρόνος και ο τόπος.

Όμοια με την ανίχνευση συναφών ανωμαλιών, για την μετατροπή του προβλήματος ανίχνευσης συλλογικών ανωμαλιών σε πρόβλημα ανίχνευσης σημειακών ανωμαλιών κατάλληλη αναπαράσταση δεδομένων πρέπει να δημιουργηθεί, συναθροίζοντας τα δεδομένα ώστε να ομαδοποιούνται οι πιθανές παρατηρήσεις που συσχετίζονται. Ο συγκεκριμένος μετασχηματισμός απαιτεί άριστη γνώση του πεδίου (*domain*) έτσι ώστε να συναθροίζονται τα γεγονότα που αντιστοιχούν σε μια συγκεκριμένη οντότητα. Για παράδειγμα, αν ελαττωματικές συσκευές προκαλούν κάποιο συγκεκριμένο πρότυπο (*pattern*) κατανάλωσης ενέργειας, αυτό το πρότυπο μπορεί να κωδικοποιηθεί ως νέο χαρακτηριστικό. [Goldstein and Uchida (2014)]

1.4 Δομή παρούσας εργασίας

Το υπόλοιπο της παρούσης διπλωματικής εργασίας οργανώνεται ως εξής. Στο Κεφάλαιο 2 περιγράφονται οι στατιστικές μέθοδοι AD, στο Κεφάλαιο 3 παρουσιάζονται οι μέθοδοι ανίχνευσης ανωμαλιών βασισμένες στην κατηγοριοποίηση, στο Κεφάλαιο 4 οι μέθοδοι AD βασισμένες στον πλησιέστερο γείτονα και στο Κεφάλαιο 5 οι μέθοδοι AD βασισμένες στην συσταδοποίηση. Στο Κεφάλαιο 6 παρουσιάζεται το πρόβλημα της κατάρας της διάστασης (*curse of dimensionality*) και μέθοδοι AD για δεδομένα μεγάλης διάστασης. Τέλος, στο Κεφάλαιο 7 γίνεται εφαρμογή και σύγκριση μερικών αλγορίθμων με τη χρήση του στατιστικού προγράμματος R.

Κεφάλαιο 2

Στατιστικές Μέθοδοι Ανίχνευσης Ανωμαλιών

2.1 Περιγραφή

Η βασική αρχή που διέπει τις στατιστικές τεχνικές ανίχνευσης ανωμαλιών είναι ότι "*ανωμαλία είναι μια παρατήρηση για την οποία υπάρχει υποψία ότι είναι εν μέρει ή εξολοκλήρου διαφορετική επειδή δεν έχει παραχθεί από το στοχαστικό μοντέλο που υποτίθεται*". Επίσης, οι στατιστικές τεχνικές ανίχνευσης ανωμαλιών στηρίζονται στην εξής βασική υπόθεση: Τα κανονικά δεδομένα εμφανίζονται σε περιοχές υψηλής πιθανότητας ενός στοχαστικού μοντέλου, ενώ οι ανωμαλίες εμφανίζονται στις περιοχές χαμηλής πιθανότητας.

Οι στατιστικές μέθοδοι προσαρμόζουν κάποιο στατιστικό μοντέλο στα δεδομένα, συνήθως για την περιγραφή της κανονικής συμπεριφοράς, και στη συνέχεια εφαρμόζουν κάποιο στατιστικό έλεγχο για να προσδιορίσουν αν κάποιο δεδομένο ανήκει στο μοντέλο ή όχι. Τα δεδομένα που παρουσιάζουν ιδιαίτερα χαμηλή πιθανότητα να έχουν παραχθεί από το μοντέλο θεωρούνται ως ανωμαλίες. Για την προσαρμογή του στατιστικού μοντέλου χρησιμοποιούνται παραμετρικές και μη-παραμετρικές μέθοδοι. Στις παραμετρικές μεθόδους υποθέτουμε ότι η κατανομή που ακολουθούν τα δεδομένα είναι γνωστή και οι παράμετροί της εκτιμώνται από τα ίδια τα δεδομένα, ενώ στις μη-παραμετρικές τεχνικές δεν γίνεται κάποια παραδοχή σχετικά με την κατανομή που ακολουθούν τα δεδομένα [Chandola et al. (2009)].

2.2 Παραμετρικές Μέθοδοι

Οι παραμετρικές μέθοδοι υποθέτουν ότι τα κανονικά δεδομένα έχουν παραχθεί από μια παραμετρική κατανομή με παραμέτρους $\boldsymbol{\theta}$ και συνάρτηση πιθανότητας πυκνότητας $f(\boldsymbol{x};\boldsymbol{\theta})$, όπου $\boldsymbol{x}$ είναι μια παρατήρηση. Οι παράμετροι $\boldsymbol{\theta}$ εκτιμώνται από τα ίδια τα δεδομένα και ο βαθμός ανωμαλίας μιας παρατήρησης $\boldsymbol{x}$ συνήθως θεωρείται ότι είναι το αντίστροφο της συνάρτησης πυκνότητας πιθανότητας $f(\boldsymbol{x};\boldsymbol{\theta})$.

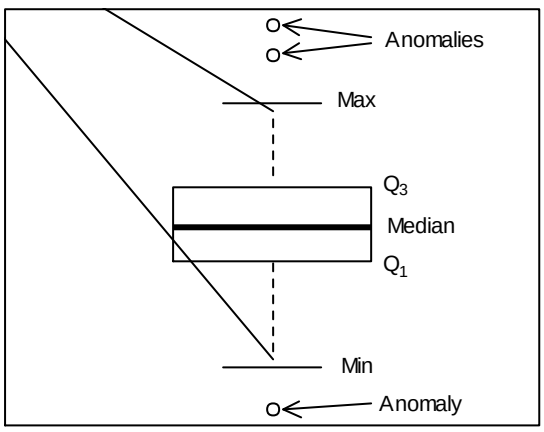
Εναλλακτικά, γίνεται η χρήση κάποιων στατιστικών ελέγχων υποθέσεων, γνωστοί και ως έλεγχοι ασυμφωνίας (*discordancy tests*). Η μηδενική υπόθεση H_0 τέτοιων ελέγχων είναι ότι η παρατήρηση $\boldsymbol{x}$ έχει παραχθεί από την υποτιθέμενη κατανομή (με παραμέτρους $\boldsymbol{\theta}$). Αν η μηδενική υπόθεση απορριφθεί από τον στατιστικό έλεγχο η παρατήρηση $\boldsymbol{x}$ θεωρείται ως ανωμαλία [Chandola et al. (2009)].

2.2.1 Μέθοδοι βασισμένες σε μοντέλα Κανονικών Κατανομών

Οι συγκεκριμένες τεχνικές υποθέτουν ότι τα δεδομένα έχουν παραχθεί από την Κανονική κατανομή (ή κατανομή του Gauss) και οι παράμετροι εκτιμώνται μέσω της μεθόδου της μεγίστης πιθανοφάνειας (*Maximum Likelihood Estimates - MLE*). Μια παρατήρηση θεωρείται ανωμαλία αν η απόστασή της από την εκτιμηθείσα μέση τιμή ξεπερνά κάποιο κατώφλι. Ο τρόπος υπολογισμού της απόστασης από τη μέση τιμή και του κατωφλιού διαφέρει σε κάθε τεχνική.

Ένας απλός κανόνας που συνήθως χρησιμοποιείται στον στατιστικό έλεγχο ποιότητας, γνωστός και ως "3σ – κανόνας", είναι ότι τα σημεία που αποκλίνουν περισσότερο από τρείς φορές την τυπική απόκλιση σ από τον μέσο μ της κανονικής κατανομής μπορούν να θεωρηθούν ως εξωκείμενες παρατηρήσεις. Η περιοχή $\mu \pm 3\sigma$ περιέχει το 99.7% των παρατηρήσεων, οπότε όσες παρατηρήσεις εμφανίζονται εκτός αυτής μπορούν να θεωρηθούν ως ανωμαλίες.

Σχήμα 2.1: *Παράδειγμα διαγράμματος Box-plot.*



Οι Laurikkala et al. (2000) χρησιμοποιούν διαγράμματα box-plots για τον εντοπισμό εξωκείμενων παρατηρήσεων σε μονομεταβλητά και πολυμεταβλητά σύνολα δεδομένων. Ένα διάγραμμα Box-plot (Σχήμα 2.1) είναι μια απλή μονοδιάστατη γραφική αναπαράσταση σε μορφή κουτιού πέντε τιμών: το άνω (*Max*) και κάτω (*Min*) όριο των ακραίων τιμών, το κάτω (Q_1) και το άνω (Q_3) τεταρτημόριο και τη διάμεσο (*Median*) των δεδομένων.

Η ποσότητα $Q_3 - Q_1$ καλείται διατεταρτημοριακό εύρος (*Inter Quartile Range -IQR*) και τα όρια των ακραίων τιμών υπολογίζονται από τους τύπους:

$$Max = Q_3 + 1.5 \cdot IQR$$

$$Min = Q_1 - 1.5 \cdot IQR$$

Με την επιλογή του συνόρου $1.5 \cdot IQR$ και με την υπόθεση ότι το σύνολο δεδομένων ακολουθεί την Κανονική κατανομή, η περιοχή μεταξύ των ποσοτήτων Max και Min περιέχει το 99.3% των παρατηρήσεων, και ως εξωκείμενες χαρακτηρίζονται οι παρατηρήσεις που ξεπερνούν τα όρια των ακραίων τιμών, όπως και με τον 3σ – κανόνα. Στα πολυμεταβλητά σύνολα, προτείνεται να γίνει πρώτα μετατροπή των παρατηρήσεων με την απόσταση Mahalanobis σε διατάξιμα δεδομένα και μετά η απεικόνισή τους σε box-plot [Laurikkala et al. (2000)].

Ο έλεγχος του Grubb χρησιμοποιείται για την ανίχνευση ανωμαλιών σε μονομεταβλητά σύνολα δεδομένων, κάτω από την υπόθεση ότι τα δεδομένα παράγονται από την Κανονική Κατανομή. Για κάθε υπό μελέτη παρατήρηση x υπολογίζεται ένα z σκορ από τον τύπο:

$$z = \frac{|x - \bar{x}|}{s}$$

όπου $\bar{x}$ και s είναι ο μέσος και η τυπική απόκλιση αντίστοιχα του συνόλου δεδομένων.

Στη συνέχεια ο έλεγχος για το αν μια παρατήρηση μπορεί να χαρακτηριστεί ως ανωμαλία γίνεται μέσω της ανισότητας:

$$z > \frac{N-1}{\sqrt{N}} \sqrt{\frac{t^2_{a/(2N),N-2}}{N - 2 + t^2_{a/(2N),N-2}}}$$

όπου N είναι το μέγεθος του συνόλου δεδομένων και $t^2_{a/(2N),N-2}$ είναι το αντίστοιχο ποσοστιαίο σημείο της *t* κατανομής με $N - 2$ βαθμούς ελευθερίας και επίπεδο σημαντικότητας

$\frac{\alpha}{2N}$. Η υπό μελέτη παρατήρηση χαρακτηρίζεται ως ανωμαλία όταν η ανισότητα είναι αληθής και εξαλείφεται από το σύνολο δεδομένων. Ο έλεγχος επαναλαμβάνεται μέχρι να απομακρυνθούν όλες οι πιθανές ανωμαλίες. [Chandola et al. (2009)].

Πάνω από εκατό έλεγχοι αυτής της κατηγορίας έχουν αναπτυχθεί από τους Barnett και Lewis (1994) και τους Rousseeuw και Leroy (1996) για διαφορετικά σενάρια, τόσο για μονομεταβλητά όσο και για πολυμεταβλητά σύνολα δεδομένων. Ο έλεγχος που χρησιμοποιείται κάθε φορά εξαρτάται από την κατανομή των δεδομένων, από το αν οι παράμετροι της κατανομής είναι γνωστές ή όχι, από τον αριθμό των αναμενόμενων εξωκείμενων παρατηρήσεων, καθώς επίσης και από το είδος τους [Knorr et al. (2000)].

2.2.2 Μέθοδοι βασισμένες σε μοντέλα παλινδρόμησης

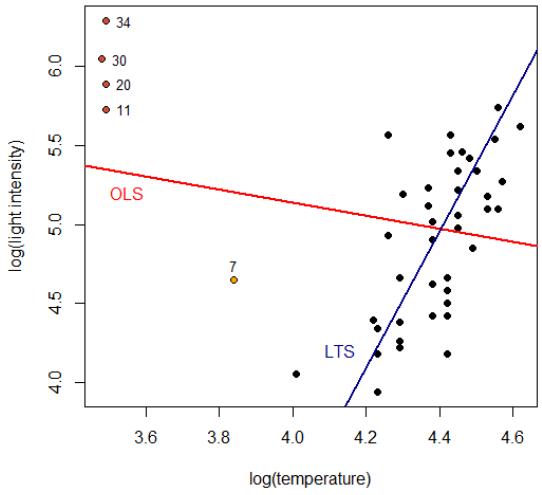
Οι βασικές τεχνικές ανίχνευσης ανωμαλιών που χρησιμοποιούν μοντέλα παλινδρόμησης αποτελούνται από δύο στάδια. Αρχικά, δημιουργείται ένα μοντέλο παλινδρόμησης από τα δεδομένα και στη συνέχεια το κατάλοιπο (*residual*) της κάθε παρατήρησης χρησιμοποιείται για τον προσδιορισμό του βαθμού ανωμαλίας. Το κατάλοιπο είναι το τμήμα της παρατήρησης που δεν εξηγείται από το μοντέλο παλινδρόμησης. Αν και το μέγεθος του καταλοίπου μπορεί να χρησιμοποιηθεί για τον προσδιορισμό των ανωμαλιών, έχουν προταθεί στατιστικοί έλεγχοι για συγκεκριμένα επίπεδα εμπιστοσύνης, ενώ ορισμένες τεχνικές ανιχνεύουν την ύπαρξη ανωμαλιών σε ένα σύνολο δεδομένων αναλύοντας το κριτήριο πληροφορίας Akaike (*Akaike Information Criterion - AIC*) κατά την εφαρμογή του μοντέλου [Chandola et al. (2009)].

Η παρουσία ανωμαλιών στα δεδομένα εκμάθησης μπορεί να επηρεάσει την κλασσική μέθοδο ελαχίστων τετραγώνων (*Ordinary Least Squares - OLS)* στην εκτίμηση των παραμέτρων του μοντέλου της παλινδρόμησης, και στα αποτελέσματα που παράγει. Μια δημοφιλής τεχνική αντιμετώπισης αυτού του προβλήματος καλείται ανθεκτική παλινδρόμηση, όπου η εκτίμηση των παραμέτρων γίνεται ταυτόχρονα με τη διαχείριση των ανωμαλιών. Ένα παράδειγμα ανθεκτικής παλινδρόμησης είναι η μέθοδος Least Trimmed Squares (LTS), όπου η εκτίμηση των παραμέτρων της παλινδρόμησης γίνεται χρησιμοποιώντας ένα υποσύνολο h παρατηρήσεων από τις συνολικές n παρατηρήσεις ($h < n$) και ελαχιστοποιώντας την εξίσωση:

$$\underset{\boldsymbol{\beta}}{minimize} \sum_{i=1}^{h} (r^2)_{(i)}$$

όπου $(r^2)_{(1)} \leq (r^2)_{(2)} \leq \cdots \leq (r^2)_{(n)}$ είναι διατεταγμένα τα τετράγωνα των καταλοίπων και $\boldsymbol{\beta}$ είναι οι παράμετροι του μοντέλου. Η τιμή του h στην παραπάνω εξίσωση επηρεάζει την ανθεκτικότητα του μοντέλου και προτείνεται $h \approx n/2$ [Rousseeuw and Hubert, (2017)]. Στο Σχήμα 2.2 δίνεται ένα απλό παράδειγμα, χρησιμοποιώντας τα δεδομένα του συμπλέγματος αστεριών "CYGOB1" το οποίο αποτελείται από 47 αστέρια. Για κάθε αστέρι, στον άξονα των x έχουμε τον λογάριθμο της θερμοκρασίας της επιφάνειάς του, και στον άξονα των y τον λογάριθμο της έντασης του φωτός του. Τα περισσότερα αστέρια ανήκουν στη λεγόμενη κύρια ακολουθία, ενώ οι παρατηρήσεις 11,20,30,34 είναι γιγαντιαία αστέρια και η παρατήρηση 7 είναι μεσαίο αστέρι. Στο Σχήμα 2.2 φαίνεται η επιρροή των ανωμαλιών στην κλασσική μέθοδο OLS (κόκκινη γραμμή), εν αντιθέσει με τη μέθοδο LTS (μπλε γραμμή), όπου το μοντέλο παλινδρόμησης δημιουργείται από την πλειοψηφία των κανονικών παρατηρήσεων, με αποτέλεσμα οι ανωμαλίες να τείνουν να έχουν μεγαλύτερα κατάλοιπα. Μια παρόμοια ανθεκτική τεχνική ανίχνευσης ανωμαλιών έχει εφαρμοστεί και σε ολοκληρωμένα αυτοπαλίνδρομα μοντέλα κινητού μέσου όρου (ARIMA) [Chandola et al. (2009)].

Σχήμα 2.2: *Παράδειγμα κλασσικής μεθόδου OLS (κόκκινη γραμμή) και ανθεκτικής μεθόδου LTS (μπλε γραμμή).*



Διάφορες παραλλαγές των τεχνικών παλινδρόμησης έχουν προταθεί για την αντιμετώπιση του προβλήματος ανίχνευσης ανωμαλιών σε πολυμεταβλητές χρονοσειρές. Οι Tsay et al. (2000) παρουσιάζουν την επιπρόσθετη πολυπλοκότητα στις πολυμεταβλητές χρονοσειρές έναντι των μονοπαραγοντικών και προτείνουν στατιστικές μεθόδους που εφαρμόζονται σε πολυμεταβλητά μοντέλα ARIMA για την ανίχνευση ανωμαλιών.

Οι Galeano et al. (2006) προτείνουν μια μέθοδο για τον εντοπισμό ανωμαλιών, η οποία δεν απαιτεί τον προσδιορισμό του πολυμεταβλητού μοντέλου και βασίζεται στην μονομεταβλητή ανίχνευση ανωμαλιών που εφαρμόζεται σε μερικές ενδιαφέρουσες προβολές (*interesting projections*) της πολυμεταβλητής χρονοσειράς. Η βασική ιδέα είναι ότι μια πολυμεταβλητή ανωμαλία παράγει τουλάχιστον μια μονομεταβλητή σε όλες σχεδόν τις προβαλλόμενες σειρές, και με την ανίχνευση των μονομεταβλητών ανωμαλιών μπορεί να προσδιοριστεί η πολυμεταβλητή. Οι ενδιαφέρουσες προβολές στο μονοδιάστατο χώρο, δηλαδή οι ενδιαφέροντες γραμμικοί συνδυασμοί, αποκτώνται χρησιμοποιώντας μια τεχνική προβολής που μεγιστοποιεί ή ελαχιστοποιεί το συντελεστή κύρτωσης (*Kurtosis coefficient*) των δεδομένων της χρονοσειράς. Η ανίχνευση ανωμαλιών σε κάθε προβολή γίνεται με τη χρήση μονομεταβλητών στατιστικών ελέγχων [Galeano et al. (2006)].

2.2.3 Μέθοδοι βασισμένες σε μίξεις παραμετρικών κατανομών

Οι τεχνικές αυτής της κατηγορίας συνδυάζουν παραμετρικές στατιστικές κατανομές για την μοντελοποίηση των δεδομένων και μπορούν να χωριστούν σε δύο υποκατηγορίες.

Στην πρώτη υποκατηγορία εντάσσονται οι τεχνικές που μοντελοποιούν τις κανονικές παρατηρήσεις και τις ανωμαλίες ξεχωριστά, και για τον προσδιορισμό των ανωμαλιών γίνεται έλεγχος αν η υπό μελέτη παρατήρηση έχει παραχθεί από την κατανομή που περιγράφει τις κανονικές παρατηρήσεις ή από την κατανομή που περιγράφει τις ανωμαλίες.

Οι Abraham and Box (1979) υποθέτουν ότι τα κανονικά δεδομένα παράγονται από μια κατανομή $N(0, \sigma^2)$ και οι ανωμαλίες από κατανομή με ίδια μέση τιμή αλλά με μεγαλύτερη διακύμανση $N(0, k\sigma^2)$. Μια παρατήρηση χαρακτηρίζεται ως κανονική ή ως ανωμαλία ανάλογα με το αποτέλεσμα του ελέγχου του Grubb που εφαρμόζεται και στις δύο κατανομές [Chandola et al. (2009)].

Ο Eskin (2000) πρότεινε μια προσέγγιση μείγματος μοντέλων (*mixture model*) για την ανίχνευση των εξωκείμενων παρατηρήσεων σε μονομεταβλητά δεδομένα. Ο συγγραφέας υποθέτει ότι το σύνολο δεδομένων αποτελείται κατά πλειοψηφία από κανονικές παρατηρήσεις, οι οποίες παράγονται από τη κατανομή $\boldsymbol{M}$ με πιθανότητα $(1-\lambda)$, ενώ οι ελάχιστες ανωμαλίες του συνόλου δεδομένων παράγονται από τη κατανομή $\boldsymbol{A}$, με μικρή πιθανότητα λ. Δηλαδή, η κατανομή $\boldsymbol{D}$ του συνόλου δεδομένων D δίνεται από τον τύπο:

$$\boldsymbol{D} = (1-\lambda)\boldsymbol{M} + \lambda\boldsymbol{A}$$

Το σύνολο δεδομένων D χωρίζεται σε δύο υποσύνολα, M και A, που αποτελούνται από τις παρατηρήσεις που έχουν παραχθεί από τις κατανομές $\boldsymbol{M}$ και $\boldsymbol{A}$, αντίστοιχα. Το πρόβλημα ανίχνευσης ανωμαλιών σε αυτό το πλαίσιο είναι αντίστοιχο του προβλήματος προσδιορισμού της κατανομής που έχει παραχθεί η κάθε παρατήρηση. Με M_t, A_t συμβολίζεται η κατάσταση των συνόλων μετά τον έλεγχο της παρατήρησης x_t, και αρχικά θεωρείται ότι όλες οι παρατηρήσεις είναι κανονικές και ανήκουν στο σύνολο M, δηλαδή ισχύει $M_0 = D$ και $A_0 = \emptyset$. Για κάθε παρατήρηση x_t γίνεται έλεγχος αν θα πρέπει να μετακινηθεί στο σύνολο A_t ή να παραμείνει στο σύνολο M_t, υπολογίζοντας την αλλαγή που προκαλεί η μετακίνηση στο λογάριθμο της συνάρτησης της πιθανοφάνειας L της κατανομής $\boldsymbol{D}$. Ο λογάριθμος της συνάρτησης της πιθανοφάνειας $LL_t(\boldsymbol{D})$ υπολογίζεται από τον τύπο:

$$LL_t(\boldsymbol{D}) = |M_t|\log(1-\lambda) + \sum_{x_i \in M_t} \log(P_{M_t}(x_i)) + |A_t|\log\lambda + \sum_{x_j \in A_t} \log(P_{A_t}(x_j))$$

όπου P_{M_t} και P_{A_t} είναι οι συναρτήσεις πυκνότητας πιθανότητας των κατανομών $\boldsymbol{M}$ και $\boldsymbol{A}$. Αν η διαφορά $(LL_t - LL_{t-1})$ είναι μεγαλύτερη ενός κατωφλιού (*threshold*) c η παρατήρηση x_t θεωρείται ανωμαλία και μετακινείται στο σύνολο A.

Για τον προσδιορισμό των P_{M_t} και P_{A_t} μπορεί να χρησιμοποιηθεί οποιαδήποτε τεχνική μοντελοποίησης, όπως *Naive Bayes, Maximum Entropy* ή *Markov chains,* και οι P_{M_t}, P_{A_t} επαναπροσδιορίζονται κάθε φορά που γίνεται κάποια μετακίνηση παρατήρησης. [Eskin et al. (2000)]

Οι τεχνικές της δεύτερης υποκατηγορίας μοντελοποιούν μόνο τις κανονικές παρατηρήσεις ως ένα μίγμα (*mixture*) παραμετρικών κατανομών και οι παρατηρήσεις που δεν υπάγονται σε κανένα από τα μοντέλα εκμάθησης θεωρούνται ανωμαλίες. Η GMM (*Gaussian Mixture Modelling)* είναι ίσως η πιο κλασσική παραμετρική τεχνική μοντελοποίησης, η οποία αποτελείται από K πολυμεταβλητές Κανονικές κατανομές, γνωστές ως συνιστώσες του μίγματος (*mixture components*), όπου κάθε συνιστώσα k έχει τη δική της παράμετρο $\theta_k = \{\boldsymbol{\mu}_k, \Sigma_k\}$, με $\boldsymbol{\mu}_k$ το διάνυσμα της μέσης τιμής και Σ_k τον πίνακα συνδιακύμανσης της αντίστοιχης πολυδιάστατης Κανονικής κατανομής . Η συνάρτηση πυκνότητας πιθανότητας του GMM δίνεται από τον τύπο:

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k f(\boldsymbol{x}; \theta_k)$$

όπου $f(\boldsymbol{x}; \theta_k)$ συνάρτηση πυκνότητας πιθανότητας της k συνιστώσας και π_κ οι μη αρνητικοί συντελεστές βάρους (*mixing weight*) της κάθε κατανομής, με $\sum \pi_\kappa = 1$. Οι παράμετροι $\boldsymbol{\theta} = \{\theta_1 \dots \theta_k\}$ του μοντέλου επιλέγονται μεγιστοποιώντας τον λογάριθμο της πιθανοφάνειας των δεδομένων εκπαίδευσης σε σχέση με το μοντέλο, με τη χρήση του αλγορίθμου EM (*Expectation-Maximization*). [Markou and Singh (2003); Falkman et al. (2009)]

Οι Yamanishi et al. (2004) παρουσιάζουν το σύστημα ανίχνευσης εξωκείμενων παρατηρήσεων SmartSifter (SS), το οποίο χρησιμοποιεί GMM για την αναπαράσταση της κανονικής συμπεριφοράς όταν το σύνολο δεδομένων αποτελείται από συνεχείς παρατηρήσεις. Σε κάθε δεδομένο ανατίθεται ένα σκορ βασισμένο στο μέγεθος της απόκλισης του δεδομένου από το μοντέλο. Όσο μεγαλύτερο είναι ένα σκορ τόσο πιο πιθανό είναι η παρατήρηση να είναι εξωκείμενη. [Yamanishi et al. (2004)].

2.3 Μη-Παραμετρικές Μέθοδοι

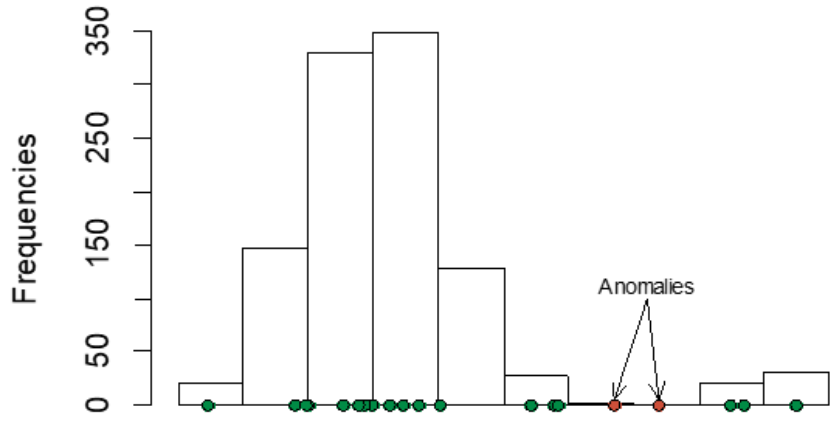
Οι τεχνικές ανίχνευσης ανωμαλιών σε αυτή την κατηγορία χρησιμοποιούν μη παραμετρικά στατιστικά μοντέλα, στα οποία η δομή του μοντέλου δεν καθορίζεται a priori, αλλά προσδιορίζεται από τα ίδια τα δεδομένα.

2.3.1 Μέθοδοι βασισμένες σε ιστογράμματα

Η πιο απλή μη παραμετρική, στατιστική τεχνική είναι η χρήση ιστογραμμάτων για την ανάδειξη του προφίλ των κανονικών δεδομένων. Η κύρια τεχνική βασισμένη σε ιστόγραμμα για μονομεταβλητά δεδομένα αποτελείται από δύο βήματα. Αρχικά, κατασκευάζεται το ιστόγραμμα βασισμένο στις διαφορετικές τιμές του χαρακτηριστικού των δεδομένων εκμάθησης. Στη συνέχεια, η τεχνική ελέγχει αν κάποια παρατήρηση του συνόλου επικύρωσης ανήκει σε κάποιο από τα κουτιά (*bins*) του ιστογράμματος και εκείνες που δεν ανήκουν σε κάποιο κουτί ή ανήκουν σε κουτί μικρής συχνότητας χαρακτηρίζονται ως ανωμαλίες, όπως φαίνεται στο Σχήμα 2.3. Μια παραλλαγή της τεχνικής είναι η ανάθεση σκορ ανωμαλίας σε κάθε παρατήρηση επικύρωσης βάσει του ύψους, δηλαδή της συχνότητας, του κουτιού στο οποίο ανήκει. Οι τεχνικές αυτής της κατηγορίας απαιτούν τα δεδομένα εκμάθησης να αποτελούνται από κανονικές παρατηρήσεις. [Chandola et al. (2009)].

Το πλάτος του κουτιού στην κατασκευή του ιστογράμματος είναι πολύ σημαντικό, καθώς αν επιλεγεί πολύ μικρό, κανονικές παρατηρήσεις θα πέφτουν σε κενά ή σπάνια κουτιά, με αποτέλεσμα να υπάρχει μεγάλη συχνότητα ψευδών θετικών σημάνσεων. Από την άλλη, αν το μέγεθος του κουτιού επιλεγεί πολύ μεγάλο αρκετές ανωμαλίες θα πέφτουν σε κουτιά με μεγάλη συχνότητα δημιουργώντας ψευδείς αρνητικές σημάνσεις. Συμπεραίνουμε ότι οι τεχνικές πρέπει να βρίσκουν το βέλτιστο πλάτος κουτιού ώστε να υπάρχει χαμηλός ρυθμός ψευδών αρνητικών και θετικών σημάνσεων.

Σχήμα 2.3: *Παράδειγμα χρήσης ιστογράμματος για την ανίχνευση ανωμαλιών.*



Στα πολυμεταβλητά δεδομένα, μια βασική τεχνική είναι η κατασκευή ιστογραμμάτων των χαρακτηριστικών. Στη φάση ελέγχου, για κάθε παρατήρηση επικύρωσης, υπολογίζεται το σκορ ανωμαλίας για κάθε χαρακτηριστικό τους ως το ύψος του κουτιού που το περιέχει. Στη συνέχεια συγκεντρώνονται τα σκορ για κάθε χαρακτηριστικό και δημιουργείται ένα συνολικό σκορ για την εκάστοτε παρατήρηση. Οι Goldstein και Dengel (2012) παρουσιάζουν τον μη-εποπτευόμενο αλγόριθμο HBOS (*Histogram Based Outlier Score*). Για κάθε χαρακτηριστικό, δηλαδή σε κάθε διάσταση d, κατασκευάζεται αρχικά ένα ιστόγραμμα, όπου το ύψος κάθε κουτιού αντιπροσωπεύει την εκτίμηση της πυκνότητας. Στη συνέχεια τα ιστογράμματα ομαλοποιούνται έτσι ώστε το μέγιστο ύψος να είναι 1.0. Έπειτα, ο βαθμός ανωμαλίας HBOS κάθε παρατήρησης p, υπολογίζεται από τον τύπο:

$$HBOS(p) = \sum_{i=0}^{d} \log\left(\frac{1}{hist_i(p)}\right)$$

όπου $hist_i(p)$ το ύψος του κουτιού όπου βρίσκεται το i χαρακτηριστικό της παρατήρησης στο αντίστοιχο ιστόγραμμα. [Goldstein and Dengel (2012)].

2.3.2 Μέθοδοι βασισμένες σε συνάρτηση πυρήνα (kernel function)

Μια μη παραμετρική μέθοδος για την εκτίμηση της πιθανότητας πυκνότητας είναι η εκτίμηση μέσω παραθύρων Parzen (*Parzen windows estimation*), γνωστή και ως εκτίμηση πυρήνα πυκνότητας (*Kernel Density Estimate*), όπου χρησιμοποιούνται συναρτήσεις πυρήνα για την προσέγγιση της πραγματικής πυκνότητας. Οι μέθοδοι ανίχνευσης ανωμαλιών βασισμένες σε συναρτήσεις πυρήνα είναι παρόμοιες με τις παραμετρικές τεχνικές που παρουσιάσαμε νωρίτερα, με μόνη διαφορά την τεχνική εκτίμησης της πυκνότητας. Αν η υπό μελέτη παρατήρηση εμφανίζει χαμηλή πιθανότητα πυκνότητα πιθανότητας, υπάρχει ένδειξη ότι έχει παραχθεί από διαφορετικό μηχανισμό, σε σχέση με το σύνολο δεδομένων, και χαρακτηρίζεται ως ανωμαλία [Chandola et al. (2009)].

Έστω $\boldsymbol{X}$ ένας $n \times d$ πίνακας που αντιπροσωπεύει το σύνολο n ανεξάρτητων και ομοιόμορφα κατανεμημένων δεδομένων $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n\}$ που παράγονται από κάποια άγνωστη συνάρτηση πιθανότητας πυκνότητας $p(x)$ σε έναν d-διάστατο Ευκλείδειο χώρο. Ο εκτιμητής πυκνότητας πυρήνα στο $\boldsymbol{x}$ δίνεται από τον τύπο:

$$\hat{p}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} {h_i}^{-d} K\left(\frac{\boldsymbol{x} - \boldsymbol{x}_i}{h_i}\right) \quad (1)$$

όπου $K(\cdot)$ μια συνάρτηση πυρήνα και h_i η παράμετρος πλάτους για τον έλεγχο της ομαλότητας (*smoothness*) του εκτιμητή. Οι συντελεστές $1/n$ και ${h_i}^{-d}$ κανονικοποιούν την εκτίμηση της πυκνότητας. Οι συναρτήσεις πυρήνα που χρησιμοποιούνται συνήθως είναι η *Gaussian*, η *Laplace*, η *Epanechnikov*, η *Uniform* και η *Tri-cube*. Για να επιτευχθεί ομαλότητα στην εκτίμηση της πυκνότητας απαιτείται ένας πυρήνα εξομάλυνσης, δηλαδή μια συνάρτηση που ικανοποιεί τις εξής συνθήκες: $\int K(x)dx = 1$, $\int xK(x)dx = 0$, και $\int x^2K(x)dx > 0$. Για την ανίχνευση ανωμαλιών στο σύνολο δεδομένων $\boldsymbol{X}$, υπολογίζεται η πυκνότητα κάθε παρατήρησης και στη συνέχεια θέτουμε ένα κατώφλι, ώστε οι παρατηρήσεις με χαμηλή πυκνότητα να χαρακτηρίζονται ως ανωμαλίες [Zhang et al. (2018)].

Στα κλασσικά προβλήματα εκτίμησης πυκνότητας μέσω της εκτίμησης παραθύρων Parzen η παράμετρος πλάτους h_i είναι σταθερή για όλες τις παρατηρήσεις, δηλαδή ισχύει $h_i = h$.

Ωστόσο, θέτοντας σταθερή τη παράμετρο πλάτους η μέθοδος μπορεί να αποτύχει στην ανίχνευση ανωμαλιών σε σύνολα δεδομένων που περιέχουν πολλές συστάδες με σημαντικές διαφορές στη πυκνότητά τους. Για τον λόγο αυτό, είναι απαραίτητο η μέθοδος να προσαρμόζεται κατάλληλα ώστε η h_i να παίρνει μεγάλες τιμές σε περιοχές μεγάλης πυκνότητας, για την εξομάλυνση της απόκλισης μεταξύ των κανονικών δειγμάτων, και χαμηλές τιμές σε περιοχές μικρής πυκνότητας, για να γίνει πιο έντονη η ανώμαλη συμπεριφορά των πιθανών ανωμαλιών. Αυτό επιτυγχάνεται αξιοποιώντας την μέση απόσταση $d_k(x_i)$ της παρατήρησης x_i από τους k κοντινότερους της γείτονες, δηλαδή η παράμετρος του πλάτους υπολογίζεται από τον τύπο:

$$h_i = cd_k(x_i) = \frac{c}{k} \sum_{j \in kNN(x_i)} d(x_i, x_j)$$

όπου $kNN(x_i)$ η γειτονιά της x_i, δηλαδή το σύνολο των k κοντινότερων παρατηρήσεων της x_i, $d(x_i, x_j)$ είναι ένα μέτρο απόστασης και c μια παράμετρος ορισμένη από τον χρήστη για τον έλεγχο της συνολικής ομαλότητας.

Οι Latecki et al. (2007) αξιοποιούν τις συναρτήσεις πυρήνα για να εκτιμήσουν την πυκνότητα του κάθε σημείου τοπικά και στη συνέχεια οι ανωμαλίες ανιχνεύονται συγκρίνοντας την τοπική πυκνότητα κάθε σημείου με την τοπική πυκνότητα των γειτονικών σημείων. Αρχικά, γίνεται εκτίμηση της τοπικής πυκνότητας LDE (*Local Density Estimate)* του κάθε σημείου x_j εφαρμόζοντας τον τύπο (1) στην γειτονιά $mNN(x_j)$, επιλέγοντας ως $K(\cdot)$ την πολυμεταβλητή Κανονική κατανομή διάστασης dim με μηδενική μέση τιμή και μοναδιαία τυπική απόκλιση, δηλαδή:

$$LDE(x_j) \propto \frac{1}{m} \sum_{x_i \in mNN(x_j)} \frac{1}{(2\pi)^{\frac{dim}{2}} \cdot {h_i}^{dim}} \exp\left(-\frac{rd_k(x_j, x_i)^2}{2{h_i}^2}\right)$$

όπου rd_k η απόσταση προσβασιμότητας (*reachability distance*) η οποία θα αναφερθεί αργότερα στο Κεφάλαιο 4. Στη συνέχεια, προσδιορίζεται ο παράγοντας τοπικής πυκνότητας LDF (*Local Density Factor*) κάθε παρατήρησης ως ο λόγος της μέσης LDE των πλησιέστερων γειτόνων της, προς την LDE της υπό μελέτη παρατήρησης:

$$LDF(x_j) \propto \frac{\sum_{x_i \in mNN(x_j)} \frac{LDE(x_i)}{m}}{LDE(x_j) + c \cdot \sum_{x_i \in mNN(x_j)} \frac{LDE(x_i)}{m}}$$

Εδώ η c είναι μια σταθερά κλίμακας ώστε οι τιμές του LDF να κυμαίνονται στο διάστημα $[0\,,1/c]$. Όσο πιο μεγάλη η τιμή του LDF τόσο πιο πιθανό το σημείο να είναι ανωμαλία. Σε κάθε πρόβλημα ο αναλυτής μπορεί να ορίσει ένα κατώφλι T, ώστε οι παρατηρήσεις για τις οποίες ισχύει $LDF > T$ να χαρακτηρίζονται ως ανωμαλίες.

Οι Zhang et al. (2018), ακολουθούν παρόμοια προσέγγιση, δηλαδή στη μέθοδό τους προτείνουν ένα μέτρο ανωμαλίας, το οποίο ορίζεται ως ένα σχετικό μέτρο της τοπικής πυκνότητας ενός δείγματος και του συνόλου των γειτονικών δειγμάτων του. Οι ερευνητές στη συνάρτηση πυρήνα θέτουν ως παράμετρο πλάτους r_i της x_i παρατήρησης:

$$r_i = c[d_{k-max} + d_{k-min} + \varepsilon - d_k(x_i)]$$

όπου $c > 0$ είναι πάλι ο συντελεστής κλίμακας που ελέγχει την συνολική ομαλότητα, και ε είναι μια μικρή θετική ποσότητα (π.χ. 10^{-5}) η οποία εξασφαλίζει ότι η παράμετρος πλάτους είναι μη μηδενική . Οι d_{k-max} και d_{k-min} είναι η μεγαλύτερη και μικρότερη τιμή του συνόλου $\{d_k(x_i)|i = 1,2,\dots,m\}$. Η τοπική πυκνότητα κάθε σημείου υπολογίζεται από τον τύπο (1), επιλέγοντας ως $K(\cdot)$ την Gaussian συνάρτηση πυρήνα, γνωστή και ως *Radial Basis Function* και αγνοώντας τον συντελεστή κανονικοποίησης,:

$$p(x_i) = \frac{1}{m-1} \sum_{j\in\{1,2,\dots,m\}\setminus\{i\}} \exp\left\{-\left(\frac{x_i - x_j}{r_i}\right)^2\right\}$$

Παρατηρούμε ότι ο τύπος εξαιρεί την συνεισφορά του ίδιου του σημείου x_i στο άθροισμα, καθώς $\exp\left\{-\frac{(x_i-x_i)^2}{{r_i}^2}\right\} = 1$, ώστε να επισημανθεί η σχετική διαφορά στην πυκνότητα μεταξύ διαφορετικών σημείων (για παράδειγμα η ποσότητα $0.1/0.3$ είναι πολύ πιο μικρή από την ποσότητα $1.1/1.3$). Ο βαθμός ανωμαλίας LOS (*Local Outlier Score)* της x_i παρατήρησης ορίζεται ως:

$$LOS(x_i) = \log\left[\frac{\frac{1}{k}\sum_{j\in mNN(x_i)}\rho(x_j)}{\rho(x_i)}\right]$$

Μια διαισθητική ερμηνεία της παραπάνω ποσότητας είναι ότι πρόκειται για μια σχετική σύγκριση της μέσης τοπικής πυκνότητας των πλησιέστερων γειτόνων ενός σημείου και της τοπικής πυκνότητας του ίδιου του σημείου. Όσο μεγαλύτερη είναι η τιμή της LOS, τόσο περισσότερο είμαστε σίγουροι για την ταξινόμηση του σημείου ως ανωμαλία, και αντίστροφα.

2.5 Πλεονεκτήματα και Μειονεκτήματα Στατιστικών Μεθόδων

Το βασικό πλεονέκτημα των στατιστικών μεθόδων ανίχνευσης ανωμαλιών είναι ότι αν οι υποθέσεις για την κατανομή των δεδομένων είναι αληθείς, η μέθοδος παρέχει στατιστικά δικαιολογημένα αποτελέσματα. Επίσης, ο βαθμός ανωμαλίας μιας στατιστικής μεθόδου συνδέεται με κάποιο διάστημα εμπιστοσύνης, το οποίο μπορεί να χρησιμοποιηθεί ως επιπρόσθετη πληροφορία κατά τη λήψη αποφάσεων σχετικά με οποιαδήποτε υπό μελέτη παρατήρηση. Τέλος, αν η μέθοδος είναι ανθεκτική στην παρουσία ανωμαλιών στα δεδομένα, μπορεί να εφαρμοστεί σε ένα μη-εποπτευόμενο περιβάλλον χωρίς να χρειάζονται ετικέτες στα δεδομένα εκμάθησης.

Από την άλλη, το βασικό μειονέκτημα των στατιστικών μεθόδων είναι ότι υποθέτουν ότι τα δεδομένα παράγονται από κάποια συγκεκριμένη κατανομή, γεγονός το οποίο συχνά δεν ισχύει και κυρίως όταν έχουμε σύνολα δεδομένων μεγάλης διάστασης (*high dimensional*). Ακόμη και αν η στατιστική παραδοχή μπορεί εύλογα να δικαιολογηθεί, υπάρχουν πολλοί στατιστικοί έλεγχοι υποθέσεων που μπορούν να χρησιμοποιηθούν για την ανίχνευση ανωμαλιών και η επιλογή του καλύτερου συχνά δεν είναι εύκολη. Συγκεκριμένα, η κατασκευή ελέγχων υποθέσεων για πολύπλοκες κατανομές που προσαρμόζονται σε δεδομένα μεγάλης διάστασης παρουσιάζει μεγάλες δυσκολίες.

Οι τεχνικές που βασίζονται σε ιστογράμματα αν και είναι σχετικά απλές στην εφαρμογή, έχουν το μειονέκτημα ότι για πολυμεταβλητά δεδομένα δεν λαμβάνουν υπόψιν τις αλληλεπιδράσεις μεταξύ διαφορετικών χαρακτηριστικών. Μια ανωμαλία μπορεί να έχει χαρακτηριστικά που από μόνα τους να εμφανίζονται πολύ συχνά, αλλά ο συνδυασμός τους να είναι πολύ σπάνιος. Η τεχνική κατασκευής ιστογραμμάτων ανά χαρακτηριστικό δεν θα μπορεί να εντοπίσει τέτοιου είδους ανωμαλίες. [Chandola et al. (2009)]

Κεφάλαιο 3

Μέθοδοι AD Βασισμένες στην Κατηγοριοποίηση

3.1 Περιγραφή

Στην κατηγοριοποίηση γίνεται εκπαίδευση ενός μοντέλου που καλείται *classifier*, από ένα σύνολο παρατηρήσεων εκμάθησης με ετικέτες, και στην συνέχεια δεδομένα επικύρωσης ταξινομούνται στις διάφορες κατηγορίες του μοντέλου. Οι τεχνικές ανίχνευσης ανωμαλιών που βασίζονται στην κατηγοριοποίηση λειτουργούν με παρόμοιο τρόπο. Οι τεχνικές αυτές λειτουργούν κάτω από την εξής γενική υπόθεση: Ένας classifier, που μπορεί να διακρίνει τις κανονικές και ανώμαλες κατηγορίες, μπορεί να δημιουργηθεί στον δεδομένο χώρο χαρακτηριστικών. Στην φάση ελέγχου τα δεδομένα χαρακτηρίζονται ως κανονικά ή ως ανωμαλίες μέσω του classifier.

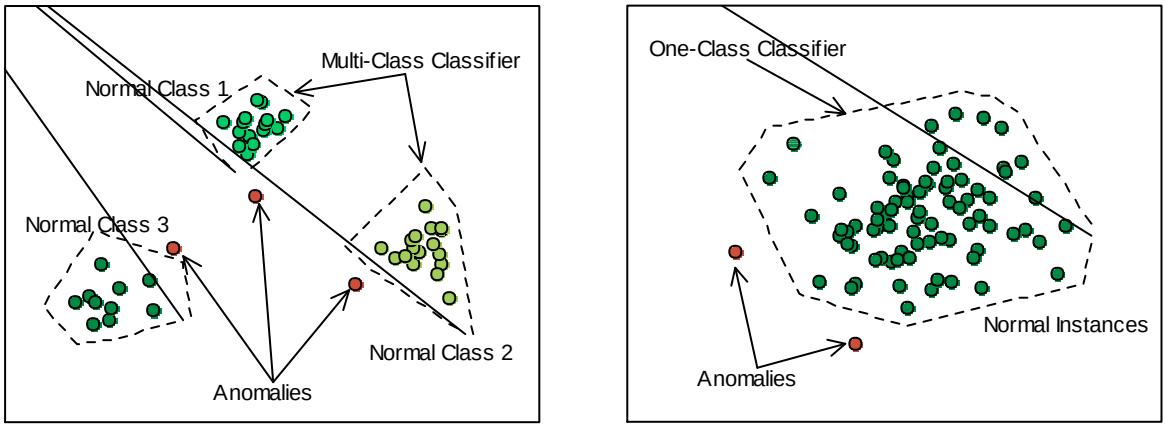
Ανάλογα με τις διαθέσιμες ετικέτες κατά την εκπαίδευση του μοντέλου, οι τεχνικές μπορούν να κατηγοριοποιηθούν σε δύο κατηγορίες: πολλών κλάσεων (*multi-class*) και μιας κλάσης (*one-class*).

Στις τεχνικές πολλαπλής κλάσης γίνεται η υπόθεση ότι τα δεδομένα εκμάθησης αποτελούνται από κανονικές παρατηρήσεις με ετικέτες, που ανήκουν σε πολλαπλές κατηγορίες και κατασκευάζουν έναν classifier, ο οποίος μπορεί να διακρίνει αυτές τις κανονικές κατηγορίες από τις υπόλοιπες. Μια παρατήρηση χαρακτηρίζεται ως ανωμαλία στην περίπτωση που δεν κατηγοριοποιηθεί ως κανονική μέσω του classifier, όπως παρουσιάζεται στο Σχήμα 3.1(a). Ορισμένες τεχνικές σε αυτήν την υποκατηγορία συσχετίζουν έναν βαθμό εμπιστοσύνης με την πρόβλεψη του classifier και στην περίπτωση που κάποια παρατήρηση κατηγοριοποιείται ως κανονική με χαμηλό βαθμό εμπιστοσύνης, θεωρείται ανωμαλία.

Στις τεχνικές ανίχνευσης ανωμαλιών που βασίζονται στην κατηγοριοποίηση μιας κλάσης, γίνεται η υπόθεση ότι τα δεδομένα εκμάθησης περιέχουν μια ετικέτα κατηγορίας, την κανονική συμπεριφορά. Αυτές οι τεχνικές δημιουργούν ένα διακριτό σύνορο γύρω από τις κανονικές παρατηρήσεις χρησιμοποιώντας αλγορίθμους κατηγοριοποίησης μιας κλάσης, όπως οι μηχανές υποστήριξης μιας κλάσης (*one-class SVMs*) και *one-class Kernel Fisher Διακρίνουσες (Discriminants).* Οι παρατηρήσεις που δεν περιέχονται εντός του συνόρου θεωρούνται ανωμαλίες, όπως παρουσιάζεται στο Σχήμα 3.1(b).

Σημαντικό ρόλο στις μεθόδους ανίχνευσης ανωμαλιών βασισμένες στην κατηγοριοποίηση έχει ο αλγόριθμος που επιλέγεται για την δημιουργία του classifier. Συνήθως χρησιμοποιούνται νευρωνικά δίκτυα (*neural networks*), Μπεϋζιανά δίκτυα (*Bayesian networks*), Μηχανές Διανυσμάτων Υποστήριξης (*Support Vector Machines- SVMs*) ή δημιουργούνται κάποιοι κανόνες για την περιγραφή της κανονικής συμπεριφοράς [Chandola et al. (2009)].

Σχήμα 3.1: *Είδη τεχνικών κατηγοριοποίησης για την ανίχνευση ανωμαλιών.*



(a) Multi-Class Anomaly Detection

(b) One-Class Anomaly Detection

3.2 Μέθοδοι βασισμένες σε Νευρωνικά Δίκτυα

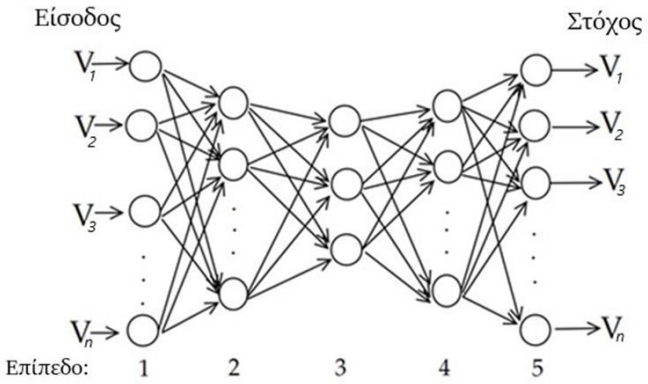
Τα νευρωνικά δίκτυα έχουν εφαρμοστεί σε μεθόδους ανίχνευσης ανωμαλιών τόσο σε προβλήματα πολλαπλής κλάσης όσο και σε προβλήματα μιας κλάσης.

Μια βασική τεχνική ανίχνευσης ανωμαλιών πολλαπλής κλάσης με τη χρήση νευρωνικών δικτύων αποτελείται από δύο βήματα. Αρχικά, ένα νευρωνικό δίκτυο εκπαιδεύεται από τις κανονικές παρατηρήσεις εκμάθησης ώστε να μπορεί να διακρίνει τις διαφορετικές κανονικές κατηγορίες. Στη συνέχεια, κάθε παρατήρηση ελέγχου εισάγεται στο νευρωνικό δίκτυο. Οι

παρατηρήσεις που γίνονται αποδεχτές από το νευρωνικό δίκτυο θεωρούνται κανονικές, αλλιώς θεωρούνται ανωμαλίες [Chandola et al (2009)].

Οι Hawkins et al. (2002) πρότειναν πρώτοι τη χρήση Replicator Neural Network (RNN) για την ανίχνευση ανωμαλιών. Η βασική ιδέα ενός RNN είναι ότι οι μεταβλητές εισόδου (*inputs*) του νευρωνικού δικτύου είναι επίσης και οι μεταβλητές εξόδου (*outputs/targets*). Δηλαδή, αντί να εκπαιδεύεται ένα δίκτυο να προβλέπει την τιμή στόχο y δεδομένου μιας τιμή εισόδου x, το δίκτυο εκπαιδεύεται να ανακατασκευάσει την ίδια την τιμή εισόδου x. Οι Hawkins et al. (2002) για την μοντελοποίηση ενός συνόλου δεδομένων n χαρακτηριστικών, χρησιμοποίησαν RNN, με τρία κρυμμένα επίπεδα (*hidden layers*), και n νευρώνες εισόδου και εξόδου, όπως φαίνεται στο Σχήμα 3.2.

Σχήμα 3.2: *Γραφική απεικόνιση ενός πλήρως συνδεδεμένου RNN, με τρία κρυμμένα επίπεδα.*



Κατά την εκπαίδευση, τα βάρη (*weights*) του RNN προσαρμόζονται ώστε να ελαχιστοποιούν το μέσο τετραγωνικό σφάλμα (*Mean Square Error*), γνωστό και ως μέσο σφάλμα ανακατασκευής (*Average Reconstruction Error*), το οποίο χρησιμοποιείται ως μέτρο ανωμαλίας OF_i κάθε παρατήρησης $\boldsymbol{x}_i$:

$$OF_i = \frac{1}{n}\sum_{j=1}^{n}\left(x_{ij} - o_{ij}\right)^2$$

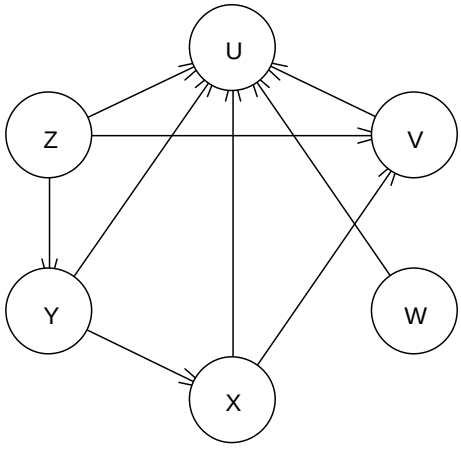
όπου $\boldsymbol{o}_i$ είναι η ανακατασκευή της παρατήρησης $\boldsymbol{x}_i$ από το μοντέλο. Το δίκτυο ανακατασκευάζει τις κανονικές παρατηρήσεις με μικρό σφάλμα ανακατασκευής και μεγάλες τιμές του OF υποδηλώνουν ότι η υπό μελέτη παρατήρηση είναι πιθανή ανωμαλία. [Hawkins et al. (2002)]

Οι Dau et al. (2014) έδειξαν ότι η αποτελεσματικότητα ενός RNN με ένα κρυμμένο επίπεδο είναι εξίσου καλή με το μοντέλο των Hawkins et al. (2002), και εκτός του βαθμού ανωμαλίας προτείνουν ένα κατώφλι για την διάκριση των κανονικών παρατηρήσεων από τις ανωμαλίες. Το εκπαιδευμένο δίκτυο ανακατασκευάζοντας κανονικές παρατηρήσεις θα παρουσιάζει μικρά σφάλματα ανακατασκευής, εν αντιθέσει με τις ανωμαλίες που υποδεικνύονται από υψηλά σφάλματα ανακατασκευής. Το μέγιστο σφάλμα στο τέλος της φάσης εκπαίδευσης του δικτύου επιλέγεται ως το κατώφλι για τον προσδιορισμό αν οι παρατηρήσεις στη φάση ελέγχου είναι κανονικές ή όχι. Οι παρατηρήσεις με βαθμό ανωμαλίας μεγαλύτερο του κατωφλιού χαρακτηρίζονται ως ανωμαλίες [Dau et al. (2014)].

3.2 Μέθοδοι βασισμένες σε Μπεϋζιανά Δίκτυα

Τα Μπεϋζιανά δίκτυα είναι κατευθυνόμενα ακυκλικά γραφήματα, όπου κάθε κόμβος περιέχει πιθανοτικές πληροφορίες σχετικά με όλες τις πιθανές τιμές μιας μεταβλητής. Συγκεκριμένα, κάθε κόμβος αντιστοιχεί σε μια τυχαία μεταβλητή, η οποία μπορεί να είναι διακριτή ή συνεχής, και οι κόμβοι μπορεί να συνδέονται μεταξύ τους με κατευθυνόμενους συνδέσμους ή βέλη. Αν υπάρχει βέλος από τον κόμβο Y στον κόμβο X, ο Y καλείται γονέας (*parent*) του X.

Σχήμα 3.3: *Γραφική αναπαράσταση Μπεϋζιανού δικτύου.*



Κάθε κόμβος X_i έχει μια υπό όρους κατανομή πιθανότητας $\boldsymbol{P}(X_i|Parents(X_i))$ η οποία ποσοτικοποιεί την επίδραση των γονέων στον κόμβο [Russel and Norvig (2003)]. Στην πραγματικότητα ένα Μπεϋζιανό Δίκτυο είναι μια παραγοντοποίηση της από κοινού πιθανότητας, και κάθε κόμβος του δικτύου έχει έναν υπό όρους πίνακα πιθανοτήτων

(*Conditional Probability Table*) που ποσοτικοποιεί την επίδραση των γονικών κόμβων. [Maes et al. (2002)]. Στο Σχήμα 3.2 παρουσιάζεται η δομή ενός απλού Μπεϋζιανού δικτύου, όπου οι κόμβοι αντιπροσωπεύουν τις πιθανότητες $P(W)$, $P(Z)$, $P(Y)$, $P(X|Y)$, $P(V|Z,X)$, $P(U|V,W,X,Y,Z)$.

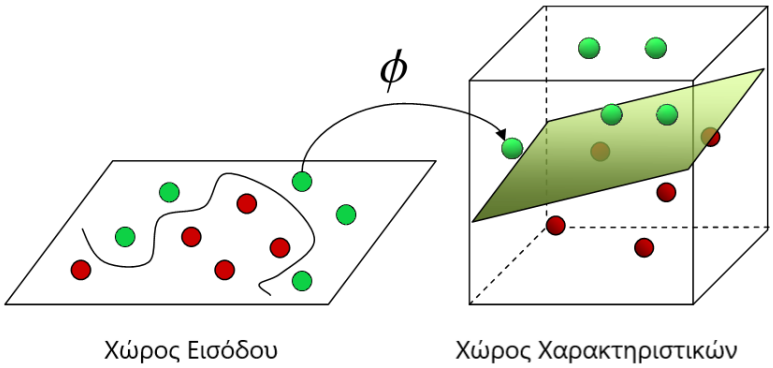
Τα Μπεϋζιανά δίκτυα χρησιμοποιούνται σε προβλήματα πολλαπλής κλάσης, όπου το γραφικό μοντέλο εκπαιδεύεται να ταξινομεί τις υπό μελέτη παρατηρήσεις σε κάποια από τις κατηγορίες, κανονική ή ανώμαλη. Η κατηγορία που επιλέγεται είναι εκείνη που παρουσιάζει μεγαλύτερη πιθανότητα εμφάνισης της παρατήρησης. Ένα από τα πλεονεκτήματα των Μπεϋζιανών δικτύων είναι ότι μπορούν να εξελίσσονται και να δημιουργούν νέους κόμβους για κατηγορίες που δεν υπάρχουν πληροφορίες στο σύνολο δεδομένων εκπαίδευσης.

Διαφορετικοί τύποι Μπεϋζιανών δικτύων έχουν χρησιμοποιηθεί για την ανίχνευση ανωμαλιών σε αρκετές εφαρμογές, όπως στον εντοπισμό εισβολής σε δίκτυα [Vlades and Skinner (2000), Siaterlis and Maglaris (2004)], στην ανίχνευση απάτης σε πιστωτικές κάρτες [Maes et al. (2002)], στην ανίχνευση ανωμαλιών στην κίνηση σκαφών [Korb et al. (2014)], σε δίκτυα αισθητήρων [Janakiram et al. (2006), Hill et al. (2007)], ενώ οι Biscarri et al. (2012) με τη χρήση Μπεϋζιανών δικτυών προσπαθούν να ανιχνεύσουν μοτίβα μη-τεχνικών απωλειών σε μια εταιρεία παροχής ηλεκτρικής ενέργειας.

3.3 Μέθοδοι βασισμένες σε Μηχανές Διανυσμάτων Υποστήριξης (SVM)

Η βασική ιδέα των SVM, για ένα σύνολο δεδομένων με παρατηρήσεις από δύο κατηγορίες, είναι η δημιουργία ενός γραμμικού υπερεπιπέδου (*hyperplane*) το οποίο θα διαχωρίζει τις δύο κατηγορίες, με το μέγιστο δυνατό περιθώριο. Ως περιθώριο ορίζεται η ελάχιστη απόσταση των σημείων από το υπερεπίπεδο. Στην περίπτωση που τα δεδομένα δεν είναι γραμμικά διαχωρίσιμα, οι SVM χρησιμοποιούν συναρτήσεις πυρήνων για τη απεικόνιση του χώρου εισόδου σε ένα χώρο χαρακτηριστικών (*feature space*) μεγαλύτερης διάστασης, στον οποίο μπορεί να κατασκευαστεί το βέλτιστο υπερεπίπεδο, όπως φαίνεται στο Σχήμα 3.3. Για να χρησιμοποιηθεί η SVM για την ανίχνευση ανωμαλιών θα πρέπει να είναι διαθέσιμες εκτός από κανονικές παρατηρήσεις, και παρατηρήσεις από όλες τις πιθανές ανώμαλες κατηγορίες, το οποίο όπως έχουμε αναφέρει σε πολλές περιπτώσεις είναι αρκετά δύσκολο και ασύμφορο.

Σχήμα 3.4: *Γραφική αναπαράσταση απεικόνιση του χώρου εισόδου σε ένα χώρο χαρακτηριστικών μεγαλύτερης διάστασης.*



Εν αντιθέσει με τις γενικές SVM, οι SVM μιας κλάσης, που προτάθηκαν από τους Scholkopf et al. (2000), κατασκευάζουν ένα σύνορο απόφασης το οποίο έχει το μέγιστο περιθώριο μεταξύ του συνόλου των κανονικών παρατηρήσεων και της αρχής των αξόνων. Έστω το σύνολο δεδομένων $\boldsymbol{X} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_N]^T \in \boldsymbol{\mathcal{R}}^{N \times M}$ από κανονικές παρατηρήσεις. Για την εύρεση του συνόρου λύνεται το εξής μοντέλο βελτιστοποίησης:

$$\min_{\boldsymbol{w} \in F, \boldsymbol{\xi} \in \mathcal{R}^N, \rho \in \mathcal{R}} \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{1}{N\nu}\sum_{i=1}^{N} \xi_i - \rho \text{ , υπό τους περιορισμούς } \boldsymbol{w} \cdot \boldsymbol{\Phi}(\boldsymbol{x}_i) \geq \rho - \xi_i, \xi_i \geq 0$$

όπου N το πλήθος των σημείων, $\nu \in (0,1)$ μια παράμετρος κανονικοποίησης και ξ_i η μεταβλητή περιθωρίου (*slack variable*), η οποία επιτρέπει στην παρατήρηση $\boldsymbol{x}_i$, να τοποθετηθεί εκτός του ορίου απόφασης και $\boldsymbol{\xi} = [\xi_1, \dots, \xi_N]$. Η ποσότητα $1/(N\nu)$ καλείται παράγοντας ποινής (*penalty factor*) και επηρεάζει τον αριθμό των παρατηρήσεων που θα τοποθετηθούν εκτός συνόρου κατά την εκπαίδευση του μοντέλου. Όσο πιο μικρή είναι η τιμή του παράγοντα ποινής, τόσο μεγαλύτερη είναι η πιθανότητα για μια παρατήρηση να τοποθετηθεί εκτός του συνόρου απόφασης.

Οι παράμετροι $\boldsymbol{w}$ και ρ καθορίζουν το σύνορο απόφασης και είναι οι μεταβλητές στόχος στο πρόβλημα βελτιστοποίησης. Το σύνορο απόφασης μπορεί να διατυπωθεί από την εξίσωση $f(\boldsymbol{x}) = sgn((\boldsymbol{w} \cdot \boldsymbol{\Phi}(\boldsymbol{x})) - \rho)$, όπου $x \in \mathcal{R}^M$ και $\boldsymbol{\Phi}$ μια συνάρτηση χαρτογράφησης. Η $f(\boldsymbol{x})$ δίνει τιμές $+1$ για τις παρατηρήσεις που είναι εντός του συνόρου, δηλαδή τις κανονικές, και τιμές -1 για τις παρατηρήσεις εκτός του συνόρου, δηλαδή τις ανωμαλίες.

Η $\boldsymbol{\Phi}$ προβάλλει το αρχικό σύνολο δεδομένων στον χώρο χαρακτηριστικών F και πρακτικά μόνο το σημείο $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{\Phi}(\boldsymbol{x}_i)\boldsymbol{\Phi}(\boldsymbol{x}_j)$ που παράγεται είναι απαραίτητο καθώς το παραπάνω πρόβλημα βελτιστοποίησης είναι ισοδύναμο με το εξής πρόβλημα:

$$\min_{\boldsymbol{a}} \boldsymbol{a}^T \boldsymbol{H} \boldsymbol{a}\,, με\ 0 \le \alpha_i \le \frac{1}{Nv}, \sum_{i=1}^{N} a_i = 1$$

όπου $\boldsymbol{a} = [a_1, \dots, a_N]^T$ και $\boldsymbol{H}$ είναι ο πίνακας πυρήνα (*kernel matrix*) με στοιχεία $H_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$, όπου K μια συνάρτηση πυρήνα. Οι συναρτήσεις πυρήνα που χρησιμοποιούνται συνήθως είναι η γραμμική, η πολυωνυμική, η Radial basis function και η Sigmoidal.

Σύμφωνα με τις Kuhn-Tucker συνθήκες, τα σημεία με $\alpha_i = 0$ βρίσκονται εντός του συνόρου, τα σημεία με $0 < \alpha_i < \frac{1}{Nv}$ βρίσκονται πάνω στο σύνορο και τα αντίστοιχα ξ_i είναι μηδενικά, ενώ τα σημεία για τα οποία ισχύει $\alpha_i = \frac{1}{Nv}$ βρίσκονται εκτός του συνόρου και θεωρούνται ανωμαλίες. [Yin et al. (2014)]

Στην περίπτωση που τα δεδομένα εκμάθησης περιέχουν θόρυβο, το σύνορο απόφασης που δημιουργείται από τις SVM αποκλίνει σημαντικά από το βέλτιστο υπερεπίπεδο, καθώς οι SVM είναι ευαίσθητες στον θόρυβο ,και ειδικά όταν οι εξωκείμενες βρίσκονται κοντά στο σύνορο απόφασης. Ένα μέτρο περιορισμού της επιρροής των εξωκείμενων είναι να θέσουμε χαμηλή τιμή στον παράγοντα ποινής, όμως η χαμηλή τιμή κάνει και ένα μέρος από τις κανονικές παρατηρήσεις να τοποθετηθούν εκτός συνόρου επηρεάζοντας την αποτελεσματικότητα του classifier.

Οι Yin et al. (2014) προτείνουν μια ανθεκτική μέθοδο SVM μιας κλάσης (*Robust 1-class SVM*) στην οποία τροποποιούν τον παράγοντα ποινής και το πρόβλημα βελτιστοποίησης γίνεται:

$$\min_{\boldsymbol{w}\in F, \boldsymbol{\xi}\in\mathcal{R}^N, \rho\in\mathcal{R}} \frac{1}{2}\|\boldsymbol{w}\|^2 + \theta \sum_{i=1}^{N} \hat{d}_i \xi_i - \rho\ , με \quad \boldsymbol{w}\cdot\boldsymbol{\Phi}(\boldsymbol{x}_i) \ge \rho - \xi_i, \qquad \xi_i \ge 0$$

όπου θ είναι μια σταθερά και $\hat{d}_i$ είναι οι προσαρμοστικοί παράγοντες ποινής που σχετίζονται με την απόσταση της x_i και το κέντρο C του συνόλου δεδομένων, το οποίο υπολογίζεται με την τεχνική *total Square Loss center* (*tSL-center*). Ο παράγοντας ποινής δίνεται από τον τύπο:

$$\hat{d}_i = \frac{d_{i,max}}{d_i}$$

όπου $d_i = \|x_i - C\|^2$ και $d_{i,max}$ είναι η μέγιστη απόσταση μεταξύ του σημείου x_i και του κέντρου του συνόλου δεδομένων. Με τη χρήση του $\hat{d}_i$ μικροί παράγοντες ποινής θα δίνονται στις παρατηρήσεις που απέχουν πολύ από το κέντρο, με αποτέλεσμα οι τυχόν εξωκείμενες να τοποθετηθούν εκτός συνόρου.

Οι Tax και Duin (2004) προτείνουν μια άλλη μορφή των SVM μιας κλάσης η οποία καλείται Support Vector Data Description (SVDD). Η βασική ιδέα των SVDD είναι η κατασκευή της υπερσφαίρας με την ελάχιστη πυκνότητα (*minimum-volume hypersphere*) σε ένα χώρο χαρακτηριστικών μεγαλύτερης διάστασης, η οποία υπερσφαίρα θα περικλείει όσο το δυνατό περισσότερα από τα κανονικά σημεία ,και όσα σημεία βρίσκονται εκτός της υπερσφαίρας θεωρούνται ανωμαλίες.

Οι μέθοδοι ανίχνευσης ανωμαλιών βασισμένες σε SVM έχουν εφαρμοστεί, μεταξύ άλλων, για ανίχνευση κυκλοφορίας δικτύου (*Network Traffic*) [Duan et al. (2004)], για ανίχνευση ανωμαλιών σε ακουστικά σήματα [Davy and Godsill (2002)] και για ανίχνευση σφαλμάτων ρουλεμάν κινητήρα επαγωγής [Chattopadhyay and Konar (2011)].

3.4 Μέθοδοι βασισμένες σε Κανόνες

Οι μέθοδοι ανίχνευσης ανωμαλιών βασισμένες σε κανόνες (*Rule-Based*) μαθαίνουν κανόνες που συλλαμβάνουν την κανονική συμπεριφορά ενός συστήματος. Στη φάση ελέγχου, οι παρατηρήσεις που δεν καλύπτονται από κάποιον κανόνα θεωρούνται ως ανωμαλίες. Οι τεχνικές βασισμένες σε κανόνες έχουν εφαρμοστεί τόσο σε προβλήματα πολλαπλής κλάσης όσο και σε προβλήματα ανίχνευσης ανωμαλιών μιας κλάσης.

Στα προβλήματα πολλαπλής κλάσης χρησιμοποιούνται αλγόριθμοι εκμάθησης κανόνων, όπως τα δέντρα αποφάσεων. Ένα δέντρο αποφάσεων αποτελείται από κόμβους απόφασης (*decision nodes*) και κόμβους φύλλα (*leaf nodes*). Κάθε κόμβος απόφασης αντιστοιχεί σε έναν έλεγχο *X* ενός μόνο χαρακτηριστικού των δεδομένων εισόδου και έχει τόσα κλαδιά (*branches*) όσα τα πιθανά αποτελέσματα του ελέγχου *X*. Κάθε κόμβος φύλλων αντιπροσωπεύει μια κλάση που είναι το αποτέλεσμα της απόφασης για κάποια περίπτωση.

Η διαδικασία κατασκευής ενός δέντρου αποφάσεων είναι μια βασική διαδικασία διαιρείν και βασίλευε (*divide-and-conquer*). Έστω ένα σύνολο δεδομένων εκμάθησης T το οποίο αποτελείται από k κλάσεις $(c_1, \dots, c_k)$. Αν το T αποτελείται από παρατηρήσεις μιας μόνο κλάσης, το T γίνεται κόμβος φύλλο. Αν το T αποτελείται από παρατηρήσεις από πολλές κλάσεις, εκτελείται ένας έλεγχος βασισμένος σε ένα χαρακτηριστικό του συνόλου εκπαίδευσης και το T θα διαχωριστεί σε n υποσύνολα $(T_1, \dots, T_n)$, όπου n είναι ο αριθμός των αποτελεσμάτων του ελέγχου. Η ίδια διαδικασία κατασκευής δέντρου αποφάσεων πραγματοποιείται αναδρομικά σε κάθε ένα υποσύνολο T_j, όπου $1 \leq j < n$, μέχρι κάθε υποσύνολο να ανήκει σε μια μόνο κλάση. Μια υπό μελέτη παρατήρηση θεωρείται ανωμαλία όταν οι διαδοχικοί έλεγχοι στα χαρακτηριστικά της από τους κόμβους αποφάσεων οδηγούν σε έναν κόμβο φύλλο ο οποίος αντιπροσωπεύει μια κλάση ανωμαλιών.

Η εξόρυξη κανόνων συσχέτισης (*association rule mining*) έχει χρησιμοποιηθεί για ανίχνευση ανωμαλιών σε προβλήματα μιας κλάσης, δημιουργώντας κανόνες από τα δεδομένα με μη-εποπτευόμενο τρόπο. Οι κανόνες συσχέτισης δημιουργούνται από ένα σύνολο κατηγορικών δεδομένων και για να διασφαλιστεί ότι οι κανόνες αντιστοιχούν σε ισχυρά μοτίβα, χρησιμοποιείται ένα κατώφλι υποστήριξης (*support threshold*) για τον περιορισμό των κανόνων με χαμηλή υποστήριξη.

Δεδομένου ότι τα συχνά πρότυπα (*frequent patterns*) που ανακαλύφθηκαν από τον αλγόριθμο κανόνων συσχέτισης αντικατοπτρίζουν τα «κοινά πρότυπα» στο σύνολο δεδομένων, είναι λογικό οι παρατηρήσεις που περιέχουν λιγότερο συχνά πρότυπα να θεωρούνται ανωμαλίες. Με άλλα λόγια, αν ένα αντικείμενο περιέχει κατά πλειοψηφία συνηθισμένα πρότυπα είναι απίθανο να είναι εξωκείμενο καθώς διαθέτει τα «κοινά χαρακτηριστικά» του συνόλου των δεδομένων.

Οι Agrawal και Srikant (1995) διατύπωσαν το πρόβλημα ανακάλυψης συχνών προτύπων σε βάσεις δεδομένων για το καλάθι αγορών ως εξής:

Έστω $I = \{i_1, i_2, \dots, i_m\}$ ένα σύνολο m στοιχείων (*literals*) και η βάση δεδομένων $D = \{t_1, t_2, \dots, t_n\}$ ένα σύνολο n συναλλαγών, καθεμία από τις οποίες αποτελείται από ένα σύνολο στοιχείων από το I. Ένα *k*-στοιχειοσύνολο (*k-itemset*) X είναι ένα μη κενό υποσύνολο του I, αποτελούμενο από *k* στοιχεία. Μια συναλλαγή $t \in D$ λέμε ότι περιέχει ένα στοιχειοσύνολο X αν $X \subseteq t$. Η υποστήριξη του στοιχειοσυνόλου X ορίζεται ως:

$$support(X) = \frac{|\{t \in D | X \subseteq t\}|}{|\{t \in D\}|}$$

Το πρόβλημα της εύρεσης όλων των συχνών στοιχειοσυνόλων (προτύπων) στο D ορίζεται ως εξής: Δεδομένου ενός κατωφλιού υποστήριξης $minisupport$ από τον αναλυτή, βρείτε όλα τα στοιχειοσύνολα με υποστήριξη μεγαλύτερη ή ίση του $minisupport$.

Οι He et al. (2004) προτείνουν έναν αλγόριθμο ανίχνευση ανωμαλιών για σύνολα κατηγορικών δεδομένων στα οποία ο βαθμός ανωμαλίας μιας παρατήρησης επικύρωσης σχετίζεται με τον αριθμό των συχνών στοιχειοσυνόλων που περιέχει. Για κάθε $t \in D$, ο βαθμός ανωμαλίας $FPOF$ (*Frequent Pattern Outlier Factor*) ορίζεται ως:

$$FPOF(t) = \frac{\sum_X support(X)}{|FPS(D, minisupport)|}$$

όπου $FPS(D, minisupport)$ είναι το σύνολο όλων των συχνών προτύπων στο D, βάση του κατωφλιού $minisupport$ και $X \in FPS(D, minisupport)$. Ο $FPOF$ παίρνει τιμές μεταξύ του 0 και 1. Όσο περισσότερα συχνά πρότυπα περιέχει μια συναλλαγή t τόσο μεγαλύτερη είναι η τιμή του $FPOF$. Δηλαδή οι παρατηρήσεις με μικρές τιμές $FPOF$ θα χαρακτηρίζονται ως εξωκείμενες.

Οι μέθοδοι ανίχνευσης ανωμαλιών βασισμένες στην εξόρυξη κανόνων συσχέτισης έχουν εφαρμοστεί στην ανίχνευση εισβολής σε δίκτυα [Mahoney and Chan (2003); Tandon and Chan (2007)], στην ανίχνευση απάτης πιστωτικών καρτών [Brause et al. (1999)] και στην ανίχνευση απάτης σε δεδομένα διατήρησης διαστημικών σκαφών [Yairi et al. (2001)].

3.5 Πλεονεκτήματα και Μειονεκτήματα Μεθόδων Βασισμένες στην Κατηγοριοποίηση

Ένα πλεονέκτημα που έχουν οι μέθοδοι βασισμένες στην κατηγοριοποίηση, και ειδικά οι τεχνικές πολλαπλής κλάσης, είναι η χρήση ισχυρών αλγορίθμων, οι οποίοι είναι ικανοί να διακρίνουν παρατηρήσεις που ανήκουν σε διαφορετικές κλάσεις. Ακόμη, η φάση ελέγχου των συγκεκριμένων μεθόδων είναι αρκετά γρήγορη, καθώς κάθε υπό μελέτη παρατήρηση συγκρίνεται με ένα προ-υπολογισμένο μοντέλο. Τέλος, οι SVM τεχνικές δεν απαιτούν κάποιο ρητό στατιστικό μοντέλο και αποφεύγουν το πρόβλημα της κατάρας της διάστασης (*curse of dimensionality* - βλ. Κεφάλαιο 6).

Το βασικό μειονέκτημα των τεχνικών πολλαπλής κλάσης βασισμένων στην κατηγοριοποίηση είναι ότι βασίζονται στην διαθεσιμότητα ακριβών ετικετών για διάφορες κανονικές κλάσεις, το οποίο δεν είναι πάντα εφικτό. Επίσης, οι μέθοδοι AD βασισμένες στην κατηγοριοποίηση αποδίδουν μια ετικέτα σε κάθε παρατήρηση ελέγχου, το οποίο μπορεί να είναι μειονέκτημα όταν ένας βαθμός ανωμαλίας είναι επιθυμητός για τις υπό μελέτη παρατηρήσεις.

Κεφάλαιο 4

Μέθοδοι AD Βασισμένες στον Πλησιέστερο Γείτονα

4.1 Περιγραφή

Η έννοια της ανάλυσης του πλησιέστερου γείτονα έχει χρησιμοποιηθεί σε πολλές τεχνικές ανίχνευσης ανωμαλιών. Τέτοιες τεχνικές βασίζονται στην ακόλουθη βασική υπόθεση: Οι παρατηρήσεις κανονικών δεδομένων εμφανίζονται σε πυκνές γειτονιές, ενώ οι ανωμαλίες εμφανίζονται μακριά από τους πλησιέστερους γείτονές τους.

Οι τεχνικές ανίχνευσης ανωμαλιών βασισμένες στον πλησιέστερο γείτονα απαιτούν κάποιο μέτρο απόστασης (*distance*) ή ομοιότητας (*similarity*), το οποίο ορίζεται μεταξύ δύο δεδομένων. Στη περίπτωση των συνεχών χαρακτηριστικών, η πιο δημοφιλής επιλογή ως μέτρο απόστασης είναι η Ευκλείδεια απόσταση, ενώ ο απλούστερος τρόπος εύρεσης της ομοιότητας μεταξύ δύο κατηγορικών χαρακτηριστικών είναι ο ορισμός ενός συντελεστή ομοιότητας ως 1, εάν οι τιμές είναι ίδιες, και 0, εάν οι τιμές είναι διαφορετικές.

Στα πολυμεταβλητά δεδομένα υπολογίζεται η απόσταση ή η ομοιότητα για κάθε χαρακτηριστικό, και στη συνέχεια τα αποτελέσματα συνδυάζονται. Αρκετές τεχνικές δεν απαιτούν το μέτρο απόστασης που θα χρησιμοποιηθεί να είναι αυστηρά μετρικό (*metric*), αρκεί να είναι θετικά ορισμένο και συμμετρικό, αλλά όχι απαραίτητα να ισχύει η τριγωνική ανισότητα.

Οι τεχνικές ανίχνευσης ανωμαλιών βασισμένες στον πλησιέστερο γείτονα μπορούν να ομαδοποιηθούν ευρέως σε δύο κατηγορίες. Η πρώτη κατηγορία περιλαμβάνει τις τεχνικές που αξιοποιούν την απόσταση μιας παρατήρησης από τον k-οστό πλησιέστερο γείτονα (*kth-NN – kth-Nearest Neighbor*) ως μέτρο ανωμαλίας, και η δεύτερη κατηγορία τις τεχνικές που

υπολογίζουν και χρησιμοποιούν ως μέτρο ανωμαλίας την σχετική πυκνότητα (*relative density*).[Chandola et al. (2009)]

4.2 Τεχνικές που αξιοποιούν τον k-οστό πλησιέστερο γείτονα

Η κύρια τεχνική ανίχνευσης ανωμαλιών βασισμένη στον k-οστό πλησιέστερο γείτονα ορίζει το βαθμό ανωμαλίας μιας παρατήρησης ως την απόστασή της από τον *kth-NN* της σε ένα σύνολο δεδομένων. Αν ο βαθμός ανωμαλίας είναι μεγαλύτερος κάποιου ορίου η παρατήρηση χαρακτηρίζεται ανωμαλία. Εναλλακτικά, οι n παρατηρήσεις με τον μεγαλύτερο βαθμό ανωμαλίας επιλέγονται ως εξωκείμενες.

Η βασική τεχνική αυτής της κατηγορίας έχει επεκταθεί από τους ερευνητές με τρεις διαφορετικούς τρόπους. Το πρώτο σύνολο παραλλαγών τροποποιεί τον ορισμό του βαθμού ανωμαλίας μιας παρατήρησης, ενώ το δεύτερο σύνολο παραλλαγών χρησιμοποιεί διαφορετικά μέτρα απόστασης για τη διαχείριση διαφορετικών τύπων δεδομένων. Τέλος, το τρίτο σύνολο παραλλαγών επικεντρώνεται στη βελτίωση της αποδοτικότητας της βασικής τεχνικής, καθώς η πολυπλοκότητά της είναι $O(N^2)$, όπου N το πλήθος των παρατηρήσεων.

Οι Knorr et al. (2000) ορίζουν την έννοια της ανώμαλης παρατήρησης ως: "*Ένα αντικείμενο o ενός συνόλου δεδομένων T είναι μια $DB(p, D)$ ανωμαλία (Distance Based - DB), αν και μόνο αν, τουλάχιστον ένα ποσοστό p των αντικειμένων του T βρίσκονται σε μεγαλύτερη απόσταση D από το o*". Πιο συγκεκριμένα, ένα αντικείμενο o είναι μια DB-ανωμαλία, αν και μόνο αν ισχύει:

$$|\{q \in T \mid d(o, q) > D\}| \geq Np$$

όπου, d μια συνάρτηση απόστασης και N το πλήθος του συνόλου δεδομένων T.

Με άλλα λόγια, ορίζοντας ως D-γειτονιά μιας παρατήρησης o, το σύνολο των παρατηρήσεων $q \in T$ τα οποία απέχουν το πολύ D από το o, δηλαδή το σύνολο $\{q \in T \mid d(o, q) \leq D\}$, ως ανωμαλίες χαρακτηρίζονται οι παρατηρήσεις που έχουν το πολύ $M = N(1 - p)$ παρατηρήσεις εντός της D-γειτονιάς τους.

Στη πρώτη μέθοδο που παρουσιάζουν για την ανίχνευση των DB-ανωμαλιών, οι Knorr et al. (2000) προτείνουν τη χρήση ενός αλγορίθμου ένθετου βρόχου (*Nested Loop*), ο οποίος έχει υπολογιστική πολυπλοκότητα (Παράρτημα Α) $O(\delta N^2)$, όπου δ είναι η διάσταση των δεδομένων. Προκειμένου να ξεπεραστεί η τετραγωνική πολυπλοκότητα του αλγορίθμου

ένθετου βρόχου, οι συγγραφείς πρότειναν μια προσέγγιση βασισμένη σε κελιά (*cell-based*). Η πολυπλοκότητα του δεύτερου αλγορίθμου είναι $O(c^{\delta} + N)$, όπου c είναι μια σταθερά αντιστρόφως ανάλογη της D. Η πολυπλοκότητα αυτή είναι γραμμική του πλήθους N, αλλά εκθετική της διάστασης των δεδομένων, με αποτέλεσμα η δεύτερη μέθοδος υπερέχει του αλγορίθμου ένθετου βρόχου για $\delta \leq 3$. [Knorr et al. (2000)]

Οι Ramaswamy et al. (2000), συμβολίζοντας με $D^k(p)$ την απόσταση ενός σημείου p από τον *kth-NN* του, κατατάσσουν τις παρατηρήσεις ανάλογα με την $D^k(p)$ και χαρακτηρίζουν ως ανωμαλίες τα r πρώτα σημεία αυτής της κατάταξης. Δεδομένου ότι ενδιαφερόμαστε μόνο για τις r πρώτες εξωκείμενες παρατηρήσεις, και συνήθως το r είναι πολύ μικρό, οι υπολογισμοί των αποστάσεων για τις περισσότερες από τις υπόλοιπες παρατηρήσεις είναι ελάχιστα χρήσιμοι και μπορούν να αποφευχθούν εντελώς. Για τον λόγο αυτό, οι ερευνητές προτείνουν μια τεχνική που βασίζεται στον διαμερισμό (*partition*).

Αρχικά, τα δεδομένα διαχωρίζονται με έναν αλγόριθμο συσταδοποίησης σε P υποσύνολα, στα οποία υπολογίζεται το κάτω όριο $P.lower$ και το άνω όριο $P.upper$ της D^k, έτσι ώστε για κάθε $p \in P$ να ισχύει $D^k(p) \geq P.lower$ και $D^k(p) \leq P.upper$. Στη συνέχεια, τα υποσύνολα διατάσσονται σε φθίνουσα σειρά βάσει του $P.lower$, και έστω $P_1, \dots, P_l$ τα υποσύνολα με τις μέγιστες τιμές του $P.lower$ έτσι ώστε ο αριθμός των σημείων στα υποσύνολα να είναι τουλάχιστον k, τότε ορίζεται ως το κατώτατο όριο $minDkDist$ της D^k για τις r ανωμαλίες ως:

$$minDkDist = \min\{P_i.lower : 1 \leq i \leq l\}$$

Έτσι, αν το $P.upper$ ενός υποσυνόλου P είναι μικρότερο του $minDkDist$ κανένα σημείο του δεν μπορεί να είναι στις r πρώτες ανωμαλίες, με αποτέλεσμα μόνο τα υποσύνολα για τα οποία ισχύει $P.upper \geq minDkDist$ είναι υποψήφια να περιέχουν τις r πρώτες ανωμαλίες. Τέλος, για τον υπολογισμό της D^k για κάθε σημείο ενός υποψήφιου υποσυνόλου P, πρέπει να εξετασθούν τα σημεία του P και τα σημεία των γειτονικών υποσυνόλων $P.neighbors$. Ως $P.neighbors$ ενός υποψήφιου υποσυνόλου P θεωρούνται εκείνα τα υποσύνολα που απέχουν το πολύ $P.upper$ από το P. Οι δοκιμές, μέχρι και για 10 διαστάσεις, δείχνουν ότι η μέθοδος κλιμακώνεται καλά τόσο με το μέγεθος των δεδομένων, όσο και με την διάσταση. [Ramaswamy et al. (2000)]

Οι Wu και Jermaine (2006) προτείνουν μια μέθοδο δειγματοληψίας για την βελτίωση της αποδοτικότητας της βασικής τεχνικής ανίχνευσης ανωμαλιών βασισμένη στον *kth-NN*. Για κάθε παρατήρηση $\boldsymbol{x} \in X$, ορίζεται ως βαθμός ανωμαλίας:

$$q_{kthSp}(\boldsymbol{x}) \coloneqq d^k(\boldsymbol{x}\,;\, S_{\boldsymbol{x}}(X))$$

όπου $S_{\boldsymbol{x}}(X)$ ένα υποσύνολο του X μεγέθους M $(M > k)$, το οποίο επιλέγεται τυχαία και επαναληπτικά για κάθε παρατήρηση $\boldsymbol{x}$ και $d^k(\boldsymbol{x}\,;\, S_{\boldsymbol{x}}(X))$ η απόσταση της $\boldsymbol{x}$ από τον *kth-NN* της στο $S_{\boldsymbol{x}}$. Μόλις ολοκληρωθεί η διαδικασία για κάθε παρατήρηση $\boldsymbol{x}$, τα γ πρώτα σημεία με την μεγαλύτερη δειγματική απόσταση από τον kth-*NN* επιλέγονται ως ανωμαλίες. Εκτός από την προφανή απλότητα του αλγορίθμου, το μεγαλύτερο όφελός του είναι ότι επιτρέπει στον χρήστη να ελέγχει τον χρόνο εκτέλεσης, καθώς ο αλγόριθμος είναι υπολογιστικής πολυπλοκότητας $O(MN)$, όπου N το πλήθος των δεδομένων του X. [Wu and Jermaine (2006)]

4.3 Τεχνικές που βασίζονται στη σχετική πυκνότητα

Οι τεχνικές αυτής της κατηγορίας υπολογίζουν και αξιοποιούν την πυκνότητα της γειτονιάς κάθε παρατήρησης. Οι παρατηρήσεις που βρίσκονται σε γειτονιές χαμηλής πυκνότητας χαρακτηρίζονται ανωμαλίες, ενώ οι παρατηρήσεις που βρίσκονται σε πυκνές γειτονιές θεωρούνται κανονικές.

Η απόσταση μιας παρατήρησης από τον *kth-NN* της είναι ισοδύναμη με την ακτίνα μιας υπερσφαίρας, κεντραρισμένη στην δεδομένη παρατήρηση, η οποία περιέχει k άλλες παρατηρήσεις. Συνεπώς, η απόσταση μιας παρατήρησης από τον *kth-NN* της μπορεί να θεωρηθεί ως εκτίμηση του αντιστρόφου της πυκνότητας της παρατήρησης, και η βασική τεχνική της Ενότητας 4.2 μπορεί να χαρακτηριστεί ως μέθοδος ανίχνευσης ανωμαλιών βασισμένη στην πυκνότητα. [Chandola et al. (2009)]

Η βασική τεχνική που στηρίζεται στην πυκνότητα αποτυγχάνει όταν τα δεδομένα αποτελούνται από περιοχές διαφορετικής πυκνότητας, και για την αντιμετώπιση αυτού του προβλήματος οι Breunig et al. (2000) αναθέτουν σε κάθε παρατήρηση ένα βαθμό ανωμαλίας γνωστό ως *Local Outlier Factor (LOF)*. Ο LOF είναι ο πιο γνωστός αλγόριθμος ανίχνευσης τοπικών ανωμαλιών και εκτελείται σε τρία βήματα.

Αρχικά, για κάθε παρατήρηση $\boldsymbol{x}$ ενός συνόλου δεδομένων X επιλέγονται οι k-κοντινότεροι γείτονες:

$$NN_k(\boldsymbol{x}) = \{\boldsymbol{y} \in X \backslash \{\boldsymbol{x}\} | d(\boldsymbol{y}, \boldsymbol{x}) \leq k\text{-}distance(\boldsymbol{x})\}$$

και σε περίπτωση ισοψηφίας του k-οστού γείτονα, χρησιμοποιούνται παραπάνω από k γείτονες. Για οποιοδήποτε θετικό ακέραιο k, η $k\text{-}distance(p)$ ενός σημείου $p \in D$, ορίζεται ως η απόσταση $d(p,q)$ μεταξύ του p και ενός σημείου $q \in D$, τέτοιο ώστε:

(i) Για τουλάχιστον k σημεία $q' \in D \setminus \{p\}$ ισχύει $d(p,q') \leq d(p,q)$ και

(ii) Το πολύ για $k-1$ σημεία $q' \in D \setminus \{p\}$ ισχύει $d(p,q') < d(p,q)$.

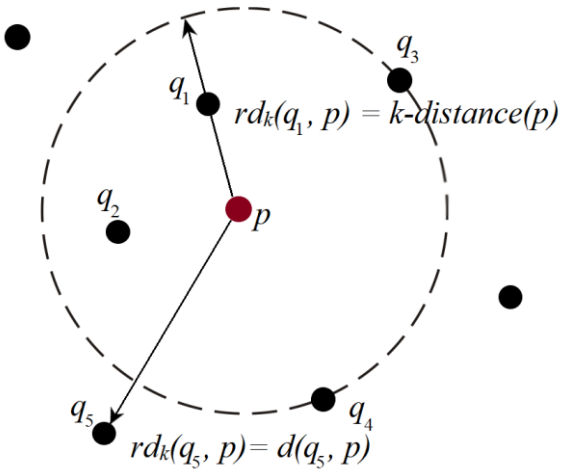
Στη συνέχεια, χρησιμοποιώντας τους k-κοντινότερους γείτονες NN_k, για κάθε παρατήρηση υπολογίζεται η τοπική πυκνότητα προσβασιμότητας (*Local Reachability Density - LRD*) ως εξής:

$$LRD_k(x) = 1/\left(\frac{\sum_{\boldsymbol{o} \in NN_k(\boldsymbol{x})} rd_k(x,o)}{|NN_k(x)|}\right)$$

όπου $rd_k(x,o) = \max\{k\text{-}distance(o), d(x,o)\}$, η απόσταση προσβασιμότητας (*reachability distance*).

Στο Σχήμα 4.1 παρουσιάζονται οι έννοιες της απόστασης προσβασιμότητας και της $k\text{-}distance$ για $k = 4$. Διαισθητικά, αν κάποιο σημείο q_5, είναι μακριά από το p τότε η απόσταση προσβασιμότητας μεταξύ των δύο είναι απλώς η πραγματική τους απόσταση. Ωστόσο, αν είναι αρκετά κοντά, όπως το σημείο q_1, η απόσταση προσβασιμότητας γίνεται η $k\text{-}distance$ του p. Με αυτόν τον τρόπο, οι στατιστικές διακυμάνσεις (*fluctuations*) της $d(p,q)$ για όλα τα q κοντά στο p μπορούν να μειωθούν σημαντικά. [Breuning et al. (2000)]

Σχήμα 4.1: *Απόσταση προσβασιμότητας και k-distance, για k=4.*



Τέλος, ο βαθμός ανωμαλίας LOF υπολογίζεται συνδυάζοντας την LRD μιας παρατήρησης με τις LRD των k κοντινότερων γειτόνων της:

$$LOF(x) = \frac{\sum_{o \epsilon NN_k(x)} \frac{LRD_k(o)}{LRD_k(x)}}{|NN_k(x)|}$$

Ο βαθμός ανωμαλίας LOF ουσιαστικά είναι ένας λόγος τοπικών πυκνοτήτων, με αποτέλεσμα οι κανονικές παρατηρήσεις, των οποίων η πυκνότητα είναι τόσο μεγάλη όσο των γειτόνων τους, να έχουν $LOF \cong 1$. Οι ανωμαλίες, από την άλλη, οι οποίες έχουν χαμηλή τοπική πυκνότητα, θα έχουν υψηλότερο βαθμό ανωμαλίας LOF. Επειδή ο LOF επηρεάζεται από την τιμή της παραμέτρου k, οι συγγραφείς προτείνουν τον υπολογισμό του LOF για ένα εύρος τιμών του k και ως τελικός βαθμός ανωμαλίας μιας παρατήρησης να θεωρείται η μέγιστη τιμή LOF. Ακόμη, προτείνεται η τιμή του k να είναι τουλάχιστον 10 και για το άνω όριο του εύρους των τιμών να επιλέγεται ο μέγιστος αριθμός των κοντινών αντικειμένων που ενδεχομένως είναι ανωμαλίες, ανάλογα με το εκάστοτε πρόβλημα. [Breuning et al. (2000), Goldstein and Uchida (2016)]

Όταν ένα σύνολο δεδομένων περιέχει συστάδες με διαφορετικές πυκνότητες, οι οποίες είναι κοντά η μια στην άλλη, ο LOF αποτυγχάνει να βαθμολογήσει σωστά τις παρατηρήσεις οι οποίες βρίσκονται στα σύνορα των συστάδων. Για τον λόγο αυτό, οι Jin et al. (2006) προτείνουν τον αλγόριθμο $INFLO$ (*INFLuenced Outlierness*) που για την εκτίμηση της πυκνότητας μιας παρατήρησης αξιοποιεί εκτός από το σύνολο NN_k και το σύνολο RNN_k των αντίστροφων γειτόνων της (*Reverse Nearest Neighbors- RNNs*). Οι *RNNs* ενός αντικειμένου p είναι αντικείμενα για τα οποία ο p ανήκει στους k-κοντινότερους γείτονές τους, δηλαδή ισχύει:

$$RNN_k(p) = \{q | q \in D, p \in NN_k(q)\}$$

Για οποιαδήποτε παρατήρηση $p \in D$, το σύνολο NN_k αποτελείται πάντα από τουλάχιστον k παρατηρήσεις, ενώ το σύνολο RNN_k μπορεί να είναι και κενό. Οι ερευνητές ορίζουν ως χώρο k-επιρροής (k-*Influence Space for p*) μιας παρατήρησης p, $IS_k(p)$, την ένωση των συνόλων $NN_k(p)$ και $RNN_k(p)$. Δηλαδή, το σύνολο $IS_k(p)$ είναι ένας τοπικός χώρος γειτονιάς που χρησιμοποιείται για την εκτίμηση της πυκνότητας γύρω από το σημείο p και του βαθμού ανωμαλίας $INFLO$:

$$INFLO_k(p) = \frac{den_{avg}(IS_k(p))}{den(p)}$$

όπου $den_{avg}(IS_k(p)) = \frac{\sum_{o \in IS_k(p)} den(o)}{|IS_k(p)|}$ και $den(p) = 1/k\text{-}\,distance(p)$ η τοπική πυκνότητα της p.

Ο βαθμός ανωμαλίας $INFLO$ μιας παρατήρησης p είναι η αναλογία της μέσης πυκνότητας των αντικειμένων του $IS_k(p)$ προς την τοπική πυκνότητα του p. Ο βαθμός ανωμαλίας $INFLO$ μιας παρατήρησης θα είναι αρκετά μεγάλος όταν η τοπική πυκνότητά της είναι πολύ μικρότερη από εκείνη των σημείων του χώρου k-επιρροής της, και η υπό μελέτη παρατήρηση χαρακτηρίζεται ως ανωμαλία. Γενικά, μια παρατήρηση p θεωρείται τοπική ανωμαλία αν ισχύει $INFLO_k > t$, όπου $t \gg 1$, ενώ για τις παρατηρήσεις με τοπική πυκνότητα κοντά στην τοπική πυκνότητα των αντικειμένων του χώρου k-επιρροής της θα ισχύει $INFLO \approx 1$ και θα θεωρούνται κανονικές. [Jin et al. (2006)]

Στον LOF δεν είναι σαφές πάνω από ποιο όριο μπορούμε να χαρακτηρίζουμε μια ανωμαλία με σιγουριά. Οι Kriegel et al. (2009) προσπάθησαν να αντιμετωπίσουν αυτό το ζήτημα προτείνοντας αντί για ένα βαθμό ανωμαλίας, μια πιθανότητα που ονομάζουν τοπική πιθανότητα ανωμαλίας (*Local Outlier Probability - LoOP*), η οποία μπορεί επίσης να οδηγήσει σε καλύτερη σύγκριση των ανώμαλων παρατηρήσεων μεταξύ διαφορετικών συνόλων δεδομένων. Θεωρώντας ως D ένα σύνολο n παρατηρήσεων και d τη συνάρτηση απόστασης που χρησιμοποιείται για τον εντοπισμό των ανωμαλιών, οι ερευνητές εισάγουν την έννοια της πιθανοτικής απόστασης (*probabilistic distance*) μιας παρατήρησης $o \in D$ σε ένα σύνολο αναφοράς (*context set*) S$\subseteq D$, η οποία συμβολίζεται ως $pdist(o, S)$ και έχει την εξής ιδιότητα:

$$\forall s \in S\colon P[d(o,s) \leq pdist(o,S)] \geq \varphi$$

Διαισθητικά, μια σφαίρα γύρω από το o με ακτίνα $pdist$ καλύπτει οποιαδήποτε στοιχείο στο σύνολο αναφοράς S με πιθανότητα φ. Το αντίστροφο της πιθανοτικής απόστασης μπορεί να θεωρηθεί ως εκτίμηση της πυκνότητας του S, και κατά συνέπεια της τοπικής πυκνότητας του o. Επιλέγοντας ως S το σύνολο των κοντινότερων γειτόνων μιας παρατήρησης, οι ερευνητές υποθέτουν ότι οι αποστάσεις των σημείων του S από την παρατήρηση o ακολουθούν κατά προσέγγιση την περικεκομένη Κανονική κατανομή . Έτσι, υπολογίζεται η τυπική απόσταση (*standard distance*) των αντικειμένων του S από το o, όπως η τυπική απόκλιση:

$$\sigma(o,S) = \sqrt{\frac{\sum_{s \in S} d(o,s)^2}{|S|}}$$

Χρησιμοποιώντας το $\lambda = \sqrt{2} \cdot erf^{-1}(\varphi)$, αντί του φ, όπου erf η συνάρτηση σφάλματος της Κανονικής κατανομής, στην εκτίμηση της πυκνότητας του S, μπορούμε να προσομοιώσουμε την κλασσική στατιστική έννοια των ανωμαλιών ως τις παρατηρήσεις που αποκλίνουν περισσότερο από λ φορές την τυπική απόκλιση σ, από την μέση τιμή. Οι εμπειρικές τιμές του λ είναι εκείνες του 3σ κανόνα, δηλαδή $\lambda = 1 \Leftrightarrow \varphi \approx 68\%$, $\lambda = 2 \Leftrightarrow \varphi \approx 95\%$ και $\lambda = 3 \Leftrightarrow \varphi \approx 99.7\%$. Επομένως, η πιθανοτική απόσταση του o στο S με σημαντικότητα λ ορίζεται ως:

$$pdist(\lambda, o, S) \coloneqq \lambda \cdot \sigma(o, S)$$

Η παράμετρος λ είναι απλώς ένας συντελεστής κανονικοποίησης (*normalization*) που επηρεάζει μόνο την σύγκριση στους βαθμούς ανωμαλίας που προκύπτουν και η κατάταξη των ανωμαλιών δεν θα επηρεάζεται από το λ. Στη συνέχεια, ορίζεται ο πιθανοτικός τοπικός παράγοντας ανωμαλίας (*Probabilistic Local Outlier Factor- PLOF*) ενός αντικειμένου $o \in D$, σε σχέση μιας σημαντικότητας λ και ενός συνόλου αναφοράς $S(o) \subseteq D$:

$$PLOF_{\lambda,S}(o) \coloneqq \frac{pdist\big(\lambda, o, S(o)\big)}{E_{s \in S(o)}[pdist(\lambda, s, S(s))]} - 1$$

Η PLOF τιμή μιας παρατήρησης $o \in D$ ορίζεται ως η αναλογία της εκτίμησης για την πυκνότητα γύρω από το o που βασίζεται στο $S(o)$ και την αναμενόμενη τιμή των εκτιμήσεων για τις πυκνότητες όλων των αντικειμένων στο σύνολο αναφοράς $S(o)$. Αν ο PLOF έχει αρνητική τιμή, τότε η παρατήρηση δεν είναι ανωμαλία, ενώ υψηλότερες τιμές είναι ένδειξη αυξανόμενης απόκλισης.

Για την επίτευξη της κανονικοποίησης ώστε η κλιμάκωση του PLOF να είναι ανεξάρτητη από τη συγκεκριμένη κατανομή των δεδομένων, γίνεται η υπόθεση ότι οι τιμές του PLOF ακολουθούν Κανονική κατανομή με μέση τιμή 1 και τυπική απόκλιση $nPLOF \coloneqq \lambda \cdot \sqrt{E[(PLOF)^2]}$.

Τέλος, εφαρμόζοντας την συνάρτηση σφάλματος της Κανονικής κατανομής, υπολογίζεται ο $LoOP$, που υποδεικνύει την πιθανότητα ένα σημείο $o \in D$ να είναι ανωμαλία:

$$LoOP_S(o) \coloneqq max\left\{0, erf\left(\frac{PLOF_{\lambda,s}(o)}{nPLOF \cdot \sqrt{2}}\right)\right\}$$

Ο $LoOP$ θα παίρνει τιμές κοντά στο 0 για τα σημεία εντός πυκνών περιοχών και κοντά στο 1 για τις ανωμαλίες. Επομένως, ενώ οι παραδοσιακοί βαθμοί ανωμαλίας των μεθόδων που βασίζονται στη σχετική πυκνότητα δεν είναι άμεσα συγκρίσιμοι μεταξύ τους, ακόμη και μέσα σε ένα ενιαίο σύνολο δεδομένων, με το $LoOP$ γίνεται άμεση εξαγωγή της πιθανότητας μια παρατήρηση του συνόλου δεδομένων να είναι ανωμαλία. [Kriegel et al. (2009)]

4.4 Πλεονεκτήματα και Μειονεκτήματα Μεθόδων Βασισμένες στον Πλησιέστερο Γείτονα.

Το βασικό πλεονέκτημα των μεθόδων που βασίζονται στον πλησιέστερο γείτονα είναι ότι από τη φύση τους ανήκουν στην κατηγορία των μη-εποπτευόμενων τεχνικών ανίχνευσης ανωμαλιών και δεν γίνεται κάποια υπόθεση για την κατανομή των δεδομένων. Ακόμη, οι ημι-εποπτευόμενες τεχνικές αποδίδουν ακόμα καλύτερα στον εντοπισμό των ανωμαλιών από τις μη-εποπτευόμενες τεχνικές, καθώς η πιθανότητα μια ανωμαλία να δημιουργήσει πυκνή γειτονιά στο σύνολο εκμάθησης είναι πολύ χαμηλή. Τέλος, η προσαρμογή των μεθόδων βασισμένων στον πλησιέστερο γείτονα σε διαφορετικά είδη συνόλων δεδομένων είναι απλή, και απαιτεί κυρίως τον καθορισμό ενός κατάλληλου μέτρου απόστασης για τα υπό μελέτη δεδομένα.

Από την άλλη, το βασικό μειονέκτημα των μη-εποπτευόμενων τεχνικών βασισμένων στον πλησιέστερο γείτονα είναι ότι αν στο υπό μελέτη σύνολο δεδομένων υπάρχουν κανονικές παρατηρήσεις που δεν έχουν αρκετούς κοντινούς γείτονες ή υπάρχουν ανωμαλίες με αρκετούς κοντινούς γείτονες, τότε η τεχνική αποτυγχάνει με αποτέλεσμα να υπάρχουν ανωμαλίες που να μην ανιχνεύονται. Ακόμη, στις ημι-εποπτευόμενες τεχνικές, αν για τις κανονικές παρατηρήσεις στο σύνολο εκπαίδευσης δεν υπάρχουν και αρκετές παρόμοιες, αυξάνεται το ποσοστό των ψευδών θετικών ανωμαλιών. Επίσης, η υπολογιστική πολυπλοκότητα της φάσης ελέγχου αποτελεί πρόκληση, καθώς απαιτείται ο υπολογισμός των αποστάσεων για κάθε παρατήρηση στο σύνολο ελέγχου, για την εύρεση των κοντινότερων γειτόνων. Τέλος, η απόδοση μιας μεθόδου βασισμένης στον πλησιέστερο γείτονα εξαρτάται σε μεγάλο βαθμό από το μέτρο απόστασης που ορίζεται μεταξύ ενός ζεύγους παρατηρήσεων, το οποίο μπορεί να διακρίνει αποτελεσματικά τις κανονικές από τις ανώμαλες παρατηρήσεις. [Chandola et al.(2009)]

Κεφάλαιο 5

Μέθοδοι AD βασισμένες στην Συσταδοποίηση

5.1 Περιγραφή

Η συσταδοποίηση είναι η διαδικασία διαχωρισμού ενός συνόλου δεδομένων σε ομάδες ή συστάδες όμοιων αντικειμένων. Κάθε συστάδα αποτελείται από αντικείμενα τα οποία είναι όμοια μεταξύ τους και διαφέρουν από τα αντικείμενα των άλλων συστάδων. Οι αλγόριθμοι συσταδοποίησης μπορούν να χωριστούν σε δύο βασικές κατηγορίες, ανάλογα με τη τεχνική εύρεσης και διαμόρφωσης των συστάδων: στους αλγορίθμους διαμερισμού (*partitioning*) και στους ιεραρχικούς (*hierarchical*) αλγορίθμους.

Οι αλγόριθμοι διαμερισμού διαχωρίζουν ένα σύνολο D, n αντικειμένων σε ένα σύνολο k συστάδων, με την τιμή της παραμέτρου k να εισάγεται από τον αναλυτή. Οι αλγόριθμοι διαμερισμού αρχίζουν συνήθως με έναν αρχικό διαχωρισμό του D, και στη συνέχεια χρησιμοποιείται μια επαναληπτική στρατηγική ελέγχου για τη βελτιστοποίηση μιας αντικειμενικής συνάρτησης (*objective function*). Κάθε συστάδα αντιπροσωπεύεται από το βάρος κέντρου της συστάδας (*k-means algorithms*) ή από ένα αντικείμενο της συστάδας που βρίσκεται κοντά στο κέντρο της (*k-medoids algorithms*). Κατά συνέπεια, οι αλγόριθμοι διαμερισμού, αρχικά, καθορίζουν τους k αντιπροσώπους ελαχιστοποιώντας την αντικειμενική συνάρτηση, και στη συνέχεια εισάγουν το κάθε αντικείμενο στη συστάδα, στην οποία ανήκει ο κοντινότερος, προς το υπό μελέτη αντικείμενο, αντιπρόσωπος.

Οι ιεραρχικοί αλγόριθμοι δημιουργούν μια ιεραρχική αποσύνθεση του D. Η ιεραρχική αποσύνθεση αντιπροσωπεύεται από ένα δενδρόγραμμα, δηλαδή ένα δέντρο που διαιρεί διαδοχικά το D σε μικρότερα υποσύνολα, έως ότου κάθε υποσύνολο να αποτελείται από ένα μόνο αντικείμενο. Σε μια τέτοια ιεραρχία, κάθε κόμβος του δέντρου αντιπροσωπεύει μια συστάδα του D. Το δενδρόγραμμα μπορεί είτε να δημιουργηθεί από τα φύλλα προς τη ρίζα (*agglomerative approach*) είτε από τη ρίζα προς τα κάτω στα φύλλα (*divisive approach*) συγχωνεύοντας ή διαιρώντας σε κάθε βήμα τις συστάδες.

Έχουν αναπτυχθεί αρκετές τεχνικές ανίχνευσης ανωμαλιών βασισμένες στην συσταδοποίηση, στις οποίες, κατά κύριο λόγο, ως ανωμαλίες χαρακτηρίζονται οι παρατηρήσεις που δεν ανήκουν σε κάποια συστάδα ή η συστάδα που ανήκουν είναι σημαντικά μικρότερη από τις υπόλοιπες.

5.2 Τεχνικές Ανίχνευσης Ανωμαλιών

Οι τεχνικές ανίχνευσης ανωμαλιών βασισμένες στη συσταδοποίηση μπορούν να ταξινομηθούν σε τρείς κατηγορίες. Η πρώτη κατηγορία αποτελείται από τεχνικές που στηρίζονται στην εξής υπόθεση: Οι κανονικές παρατηρήσεις ανήκουν σε κάποια συστάδα, ενώ οι ανωμαλίες δεν ανήκουν σε κάποια.

Οι τεχνικές της πρώτης κατηγορίας εφαρμόζουν κάποιον αλγόριθμο συσταδοποίησης στο σύνολο δεδομένων και χαρακτηρίζουν ως ανωμαλίες τις παρατηρήσεις που δεν ταξινομήθηκαν σε κάποια συστάδα. Αρκετοί αλγόριθμοι συσταδοποίησης έχουν αναπτυχθεί ώστε να μην αναγκάζουν κάθε παρατήρηση σε ένα σύνολο δεδομένων να ανήκει σε κάποια συστάδα και μπορούν να αξιοποιηθούν για την ανίχνευση ανωμαλιών.

Οι Ester et al. (1996) παρουσιάζουν τον αλγόριθμο συσταδοποίησης DBSCAN, ο οποίος έχει σχεδιαστεί για να ανακαλύπτει συστάδες αυθαίρετου σχήματος. Η βασική ιδέα στον DBSCAN είναι ότι η γειτονιά, μιας δεδομένης ακτίνας Eps, κάθε σημείου μιας συστάδας πρέπει να περιέχει τουλάχιστον έναν ελάχιστο αριθμό σημείων $MinPts$, δηλαδή η πυκνότητα της γειτονιάς πρέπει να υπερβαίνει κάποιο όριο. Οι συγγραφείς θεωρούν ότι μια συστάδα αποτελείται από δύο ειδών σημεία, τα σημεία εντός της συστάδας (σημεία πυρήνες – *core points*) και τα σημεία στο σύνορο της συστάδας (συνοριακά σημεία – *border points*). Η Eps-γειτονιά (Eps-*neighborhood*) ενός σημείου p ορίζεται ως:

$$N_{Eps}(p) = \{q \in D | dist(p,q) \leq Eps\}$$

όπου $dist(p,q)$ μια συνάρτηση απόστασης δύο σημείων p και q. Γενικά, η Eps-γειτονιά ενός συνοριακού σημείου θα περιέχει σημαντικά λιγότερα σημεία από την Eps-γειτονιά ενός σημείου πυρήνα. Επομένως, για να εισαχθούν και οι δύο τύποι σημείων σε μια συστάδα οι ερευνητές ορίζουν τρεις έννοιες: την άμεση προσβασιμότητα μέσω πυκνότητας (*directly density-reachable - DDR*), την προσβασιμότητα μέσω πυκνότητας (*density-reachable - DR*) και τη σύνδεση μέσω πυκνότητας (*density-connected - DC*).

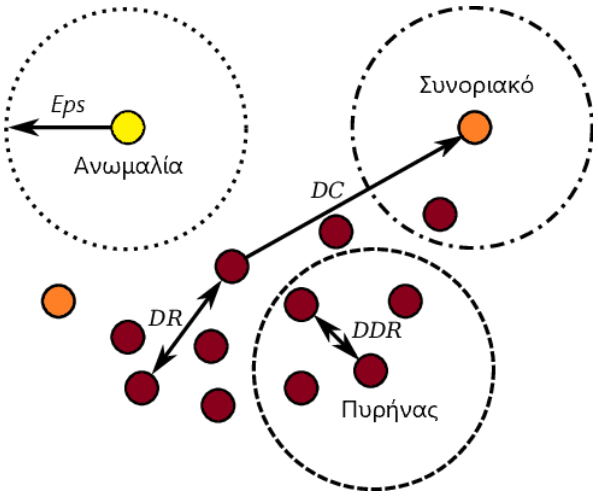
Ένα σημείο $p \in D$ είναι άμεσα προσβάσιμο μέσω πυκνότητας σε ένα σημείο $q \in D$ όταν ισχύει: $p \in N_{Eps}(q)$ και $\left|N_{Eps}(q)\right| \geq MinPts$ (συνθήκη σημείου πυρήνα) και ένα σημείο p είναι προσβάσιμο μέσω πυκνότητας από ένα σημείο q αν υπάρχει μια αλυσίδα από σημεία $p_1, \dots, p_n, με\ p_1 = q, p_n = p$ τέτοια ώστε το p_{i+1} είναι άμεσα προσβάσιμο μέσω πυκνότητας από το p_i, για $i = 1, \dots, n-1$. Τέλος, ένα σημείο p συνδέεται μέσω πυκνότητας με ένα σημείο q, αν υπάρχει ένα σημείο o, ώστε το p και το q να είναι προσβάσιμα μέσω πυκνότητας από το o. Αξιοποιώντας τους παραπάνω ορισμούς, ως συστάδα C ορίζεται ένα μη κενό υποσύνολο του D, το οποίο ικανοποιεί τις εξής συνθήκες:

1. $\forall\ p, q$: αν $p \in C$ και q είναι προσβάσιμο μέσω πυκνότητας από το p, τότε $q \in C$
2. $\forall\ p, q \in C$: p συνδέεται μέσω πυκνότητας στο q

Με την πρώτη συνθήκη, απαιτείται για κάθε σημείο p μιας συστάδας C, να υπάρχει ένα σημείο $q \in C$ τέτοιο ώστε το p να ανήκει στην Eps-γειτονιά του q και η $N_{Eps}(q)$ θα περιέχει τουλάχιστον $MinPts$ σημεία. Δύο συνοριακά σημεία μιας συστάδας είναι πιθανό να μην είναι προσβάσιμα μέσω πυκνότητας μεταξύ τους, αλλά θα υπάρχει ένα σημείο πυρήνας μέσα στη C από το οποίο και τα δύο θα είναι προσβάσιμα μέσω πυκνότητας. Επομένως, με τη δεύτερη συνθήκη, εξασφαλίζεται η επιλογή όλων των συνοριακών σημείων μιας συστάδας. [Ester et al. (1996)]

Στο Σχήμα 5.1, φαίνεται η συσταδοποίηση με τη χρήση του αλγορίθμου DBSCAN με $MinPts = 4$. Τα μπορντό σημεία είναι σημεία πυρήνες, ενώ τα πορτοκαλί είναι συνοριακά σημεία και αποτελούν την συστάδα που ανακαλύπτει ο αλγόριθμος DBSCAN. Από την άλλη, το κίτρινο σημείο θα θεωρείται ανωμαλία καθώς δεν εντάσσεται στην συστάδα.

Σχήμα 5.1: *Συσταδοποίηση με τον αλγόριθμο DBSCAN για* $MinPts = 4$.



Οι Hinneburg και Keim (1998) εισάγουν έναν αλγόριθμο συσταδοποίησης για μεγάλες βάσεις δεδομένων, που ονομάζεται DENCLUE. Ο αλγόριθμος DENCLUE βασίζεται στην ιδέα ότι η επιρροή (*influence*) κάθε δεδομένου σημείου μπορεί να διαμορφωθεί τυπικά χρησιμοποιώντας μια μαθηματική συνάρτηση, η οποία ονομάζεται συνάρτηση επιρροής (*influence function*). Η συνάρτηση επιρροής μπορεί να θεωρηθεί ως μια συνάρτηση που περιγράφει την επίδραση ενός δεδομένου σημείου εντός της γειτονιάς του. Η συνάρτηση επιρροής ενός δεδομένου $y \in F^d$ είναι μια συνάρτηση $f_B^y : F^d \to R_0^+$, η οποία ορίζεται γενικά ως $f_B^y(x) = f_B(x, y)$.

Για τον ορισμό συγκεκριμένων συναρτήσεων επιρροής, χρησιμοποιείται μια συνάρτηση απόστασης $d : F^d \times F^d \to \mathcal{R}_0^+$, η οποία ορίζει την απόσταση δύο d-διάστατων διανυσμάτων. Οι ερευνητές επιλέγουν ως συνάρτηση απόστασης την Ευκλείδεια και ως συνάρτηση επιρροής την Gaussian με τύπο:

$$f_{Gauss}(x, y) = e^{-\frac{d(x,y)^2}{2\sigma^2}}$$

Η συνολική πυκνότητα του χώρου δεδομένων μπορεί να υπολογιστεί ως το άθροισμα της συνάρτησης επιρροής όλων των δεδομένων σημείων. Δεδομένου ενός συνόλου δεδομένων $D = \{x_1, \dots, x_N\} \subset F^d$, η συνάρτηση πυκνότητας ορίζεται ως:

$$f_{Gauss}^D(x) = \sum_{i=1}^{N} f_{Gauss}^{x_i}(x)$$

Στη συνέχεια, οι συστάδες μπορούν να προσδιοριστούν μαθηματικά, αναγνωρίζοντας τους πόλους έλξης πυκνότητας (*density-attractors*). Οι πόλοι έλξης πυκνότητας είναι τοπικά μέγιστα της συνάρτησης της συνολικής πυκνότητας και επομένως, απαιτείται ο ορισμός της κλίσης της συνάρτησης πυκνότητας ως:

$$\nabla f_{Gauss}^D(x) = \sum_{i=1}^{N} (x_i - x) \cdot e^{-\frac{d(x,x_i)^2}{2\sigma^2}}$$

Ένα σημείο $x^* \in F^d$ καλείται πόλος έλξης πυκνότητας, δεδομένου μια συνάρτησης επιρροής, ανν το x^* είναι τοπικό μέγιστο της συνάρτησης πυκνότητας f_B^D. Ακόμη, ένα σημείο $x \in F^d$ ελκύεται μέσω πυκνότητας (*density-attracted*) σε έναν πόλο έλξης πυκνότητας x^*, ανν $\exists\, k \in N : d(x^k, x^*) \le \varepsilon$, με $x^0 = x, x^i = x^{i-1} + \delta \cdot \frac{\nabla f_B^D(x^{i-1})}{\|\nabla f_B^D(x^{i-1})\|}$. Τέλος, μια συστάδα

ακανόνιστου σχήματος, δεδομένου των παραμέτρων σ και ξ, για ένα σύνολο πόλων έλξης πυκνότητας X είναι ένα υποσύνολο $C \subseteq D$, όπου:

1. $\forall\, x \in C\ \exists\, x^* \in X$: $f_B^D(x^*) \geq \xi$, το x ελκύεται μέσω πυκνότητας στο x^* και
2. $\forall\, x_1^*, x_2^* \in X$: $\exists$ ένα μονοπάτι $P \subset F^d$ από το x_1^* στο x_2^* τ.ω. $\forall\, p \in P$: $f_B^D(p) \geq \xi$

Η παράμετρος σ περιγράφει την επιρροή ενός σημείου στη γειτονιά του και η παράμετρος ξ περιγράφει πότε ένας πόλος έλξης πυκνότητας είναι σημαντικός, επιτρέποντας μείωση του αριθμού των πόλων έλξης πυκνότητας και συμβάλλοντας με αυτόν τον τρόπο στη βελτίωση της απόδοσης. Τέλος, ως ανωμαλίες θεωρούνται τα σημεία $x \in D$ τα οποία ελκύονται μέσω πυκνότητας σε ένα τοπικό μέγιστο x_0^* με $f_B^D(x_0^*) < \xi$. [Hinneburg and Keim (1998)]

Η δεύτερη κατηγορία τεχνικών ανίχνευσης ανωμαλιών, αποτελείται από μεθόδους που στηρίζονται στην εξής υπόθεση: Οι κανονικές παρατηρήσεις βρίσκονται κοντά στο κέντρο βάρους (*centroid*) της πλησιέστερης συστάδας, ενώ οι ανωμαλίες βρίσκονται μακριά. Οι τεχνικές που βασίζονται στην υπόθεση αποτελούνται από δύο βήματα. Αρχικά, τα δεδομένα ομαδοποιούνται χρησιμοποιώντας έναν αλγόριθμο συσταδοποίησης, και στη συνέχεια για κάθε δεδομένο η απόσταση του ως προς το πλησιέστερο κέντρο βάρους συστάδας θεωρείται ως βαθμός ανωμαλίας.

Οι Münz et al. (2007) εφαρμόζουν τον αλγόριθμο συσταδοποίησης K-means για την ανίχνευση ανωμαλιών στην κίνηση ενός δικτύου (*network traffic*). Ο αλγόριθμος K-means είναι ένας αλγόριθμος συσταδοποίησης, ο οποίος ομαδοποιεί τα δεδομένα βάσει των τιμών των χαρακτηριστικών τους σε K διακεκριμένες συστάδες. Η παράμετρος K είναι ένας θετικός ακέραιος αριθμός που υποδηλώνει τον ακριβή αριθμό των συστάδων.

Στο πρώτο βήμα του αλγορίθμου, καθορίζεται ο αριθμός K από τον αναλυτή και επιλέγονται K αρχικά κέντρα βάρους. Η επιλογή των κέντρων βάρους γίνεται είτε χωρίζοντας αυθαίρετα τα δεδομένα σε K συστάδες, υπολογίζοντας στη συνέχεια τα κέντρα βάρους, είτε επιλέγοντας αυθαίρετα K δεδομένα. Έπειτα, για κάθε δεδομένο, υπολογίζονται οι αποστάσεις από τα κέντρα βάρους των συστάδων και το κάθε δεδομένο κατατάσσεται στην ομάδα του κοντινότερου κέντρου βάρους. Μόλις γίνει η ανάθεση των δεδομένων στις αντίστοιχες συστάδες, υπολογίζονται τα νέα κέντρα βάρους, και η διαδικασία επαναλαμβάνεται έως ότου τα κέντρα βάρους να παραμένουν ίδια.

Ο αλγόριθμος συσταδοποίησης χωρίζει το σύνολο δεδομένων εκπαίδευσης σε K ομάδες, αλλά δεν προσδιορίζει, ποιες συστάδες αντιπροσωπεύουν τις κανονικές παρατηρήσεις και ποιες τις ανώμαλες. Ο προσδιορισμός γίνεται από τον αναλυτή βάσει των χαρακτηριστικών των δεδομένων της κάθε συστάδας. Για παράδειγμα, η συστάδα με τον μεγαλύτερο μέσο αριθμό πακέτων (*packets*) μπορεί να θεωρηθεί ως ένδειξη ανώμαλης συμπεριφοράς στην κίνηση ενός δικτύου. Οι Münz et al. (2007) επιλέγουν $K = 2$, υποθέτοντας ότι οι κανονικές και ανώμαλες εγγραφές στο σύνολο εκπαίδευσης δημιουργούν δύο διαφορετικές συστάδες. Μια παρατήρηση χαρακτηρίζεται ως ανωμαλία αν είναι πιο κοντά στο κέντρο βάρους της συστάδας με τις ανώμαλες παρατηρήσεις, από ό,τι στο κέντρο βάρους της κανονικής συστάδας. Ακόμη, ορίζεται ένα όριο από τον αναλυτή ώστε, αν η απόσταση μιας παρατήρησης από το κέντρο βάρους της κανονικής συστάδας είναι μεγαλύτερη από αυτό το όριο, να θεωρείται ανωμαλία. Με αυτόν τον τρόπο γίνεται ανίχνευση καινούργιων ανωμαλιών που τυχόν να μην υπήρχαν στο σύνολο εκπαίδευσης.

Η τρίτη κατηγορία τεχνικών ανίχνευσης ανωμαλιών, αποτελείται από μεθόδους που στηρίζονται στην εξής υπόθεση: Οι κανονικές παρατηρήσεις ανήκουν σε μεγάλες και πυκνές συστάδες, ενώ οι ανωμαλίες ανήκουν σε μικρές ή αραιές συστάδες.

Οι He et al. (2003) παρουσιάζουν τον αλγόριθμο FindCBLOF, ο οποίος αναθέτει στα δεδομένα το βαθμό ανωμαλίας CBLOF (*Cluster-Based Local Outlier Factor*). Αρχικά, χρησιμοποιείται ένας αλγόριθμος συσταδοποίησης στο σύνολο δεδομένων D, και οι συστάδες που δημιουργούνται κατηγοριοποιούνται σε μεγάλες και μικρές. Στη συνέχεια, τα σημεία τα οποία δεν ανήκουν σε κάποια μεγάλη συστάδα χαρακτηρίζονται ως ανωμαλίες.

Θεωρητικά, οποιαδήποτε μέθοδος συσταδοποίησης μπορεί να χρησιμοποιηθεί στο πρώτο βήμα του αλγορίθμου. Ωστόσο στην πράξη χρησιμοποιείται συνήθως ο K-means επειδή έχει χαμηλή υπολογιστική πολυπλοκότητα [Goldstein and Uchida (2016)]. Έστω D ένα σύνολο δεδομένων, και έστω $C = \{C_1, C_2, \dots, C_k\}$ το σύνολο των συστάδων με σειρά τέτοια ώστε $|C_1| \geq |C_2| \geq \cdots \geq |C_k|$. Δεδομένων δύο παραμέτρων a και β, ορίζεται το b ως σύνορο των μεγάλων και μικρών συστάδων, αν ισχύει μια από τις παρακάτω ανισώσεις:

$$|C_1| + |C_2| + \cdots + |C_b| \geq |D| \cdot a \qquad (5.1)$$

$$|C_b|/|C_{b+1}| \geq \beta \qquad (5.2)$$

Έπειτα, το σύνολο των μεγάλων συστάδων ορίζεται ως: $LC = \{C_i | i \leq b\}$ και το σύνολο των μικρών συστάδων ορίζεται ως: $SC = \{C_j | j > b\}$. Οι παραπάνω ορισμοί παρέχουν ποσοτικά μέτρα για τη διάκριση των μεγάλων και των μικρών συστάδων. Η εξίσωση (5.1) θεωρεί το γεγονός ότι οι περισσότερες παρατηρήσεις στο σύνολο D δεν είναι ανωμαλίες. Επομένως, οι ομάδες που κατέχουν το μεγαλύτερο μέρος των παρατηρήσεων θα πρέπει να θεωρούνται ως μεγάλες συστάδες. Για παράδειγμα, εάν το a έχει οριστεί στο 90%, σκοπεύουμε να θεωρήσουμε ως μεγάλες συστάδες εκείνες που περιέχουν το 90% των παρατηρήσεων. Η εξίσωση (5.2) θεωρεί το γεγονός ότι οι μεγάλες και οι μικρές συστάδες θα πρέπει να έχουν σημαντικές διαφορές στο μέγεθος. Για παράδειγμα, αν θέσουμε $\beta = 5$, το μέγεθος οποιασδήποτε συστάδας του LC είναι τουλάχιστον πέντε φορές το μέγεθος των συστάδων του SC.

Για κάθε παρατήρηση $o \in D$, ο βαθμός ανωμαλίας CBLOF ορίζεται ως:

$$CBLOF(o) = \begin{cases} |C_i| \cdot \min\left(distance(o, C_j)\right) \text{ όπου } o \in C_i, C_i \in SC \text{ και } C_j \in LC \text{ για } j = 1, .., b \\ |C_i| \cdot distance(o, C_i) \text{ όπου } o \in C_i \text{ και } C_i \in LC \end{cases}$$

όπου $distance(o, C_i)$ η απόσταση της παρατήρησης o από το κέντρο της συστάδας C_i. Αν μια παρατήρηση o ανήκει σε μικρή συστάδα, τότε ο CBLOF της o καθορίζεται από το μέγεθος της συστάδας και την απόσταση μεταξύ της o και της κοντινότερης σε αυτή μεγάλη συστάδα. Αν η παρατήρηση o ανήκει σε μεγάλη συστάδα, τότε ο CBLOF καθορίζεται από το μέγεθος της συστάδας και της απόστασης της o από την συστάδα στην οποία ανήκει. [He et al. (2003)]

Οι Campello et al. (2015) προτείνουν τον αλγόριθμο HDBSCAN* (*Hierarchical DBSCAN**), ο οποίος δημιουργεί μια πλήρη ιεραρχική συσταδοποίηση βασισμένη στην πυκνότητα και για κάθε παρατήρηση μπορεί να υπολογιστεί ο βαθμός ανωμαλίας GLOSH (*Global-Local Outlier Score from Hierarchies*).

Αρχικά παρουσιάζουν μια παραλλαγή του αλγορίθμου DBSCAN, τον DBSCAN*, στον οποίο οι συστάδες ορίζονται μόνο από αντικείμενα πυρήνες. Όπως έχουμε αναφέρει, ένα αντικείμενο x_p είναι αντικείμενο πυρήνας, δεδομένων των παραμέτρων ε και m_{pts}, αν η ε-γειτονιά του περιέχει τουλάχιστον m_{pts} αντικείμενα δηλαδή, αν $\left|N_\varepsilon(x_p)\right| \geq m_{pts}$, όπου $N_\varepsilon(x_p) = \{x \in X | d(x, x_p) \leq \varepsilon\}$. Δύο αντικείμενα πυρήνες x_p και x_q είναι ε-προσβάσιμα, δεδομένων των ε και m_{pts}, αν $x_p \in N_\varepsilon(x_q)$ και $x_q \in N_\varepsilon(x_p)$, ενώ τα x_p και x_q συνδέονται μέσω πυκνότητας αν είναι άμεσα ή μεταβατικά ε-προσβάσιμα. Μια συστάδα C, δεδομένων

των ε και m_{pts}, είναι ένα μη-κενό μέγιστο (*maximal*) υποσύνολο του συνόλου αντικειμένων X, τέτοιο ώστε κάθε ζευγάρι αντικειμένων του C να συνδέονται μέσω πυκνότητας. Τα αντικείμενα που δεν είναι πυρήνες θεωρούνται θόρυβος.

Βασισμένοι στους παραπάνω ορισμούς, οι συγγραφείς, επινόησαν τον αλγόριθμο DBSCAN* που εννοιολογικά βρίσκει συστάδες ως τους συνδεδεμένους συντελεστές ενός γραφήματος, στο οποίο τα αντικείμενα του X είναι κορυφές, και κάθε ζεύγος κορυφών είναι γειτονικά αν και μόνο αν τα αντίστοιχα αντικείμενα είναι ε-προσβάσιμα, δεδομένου των παραμέτρων ε και m_{pts}.

Στον αλγόριθμο HDBSCAN* που προτείνεται εισάγεται μόνο η παράμετρος m_{pts} και η μέθοδος εφαρμόζει τον DBSCAN* για ένα εύρος τιμών του ε. Η βασική ιδέα του αλγορίθμου είναι ότι μειώνοντας σταδιακά την παράμετρο ε, δηλαδή αυξάνοντας το απαιτούμενο επίπεδο πυκνότητας μιας συστάδας, τότε η συστάδα συρρικνώνεται, αλλά παραμένει ενωμένη, μέχρι ένα επίπεδο πυκνότητας στο οποίο είτε η συστάδα χωρίζεται σε μικρότερες είτε εξαφανίζεται τελείως. Με αποτέλεσμα, ο αλγόριθμος HDBSCAN* να παράγει ένα δέντρο συσταδοποίησης που περιέχει όλες τις συστάδες που ανακαλύπτει ο αλγόριθμος DBSCAN*.

Αφού υπολογιστεί η ιεραρχική συσταδοποίηση βασισμένη στην πυκνότητα για όλο το σύνολο δεδομένων, ο βαθμός ανωμαλίας GLOSH για κάθε αντικείμενο x_i μπορεί να υπολογιστεί με βάση τη διαφορά της πυκνότητας γύρω από το x_i και την υψηλότερη πυκνότητα στο εσωτερικό της συστάδας που είναι πιο κοντά στο x_i στην ιεραρχία HDBSCAN*, ως εξής:

$$GLOSH(x_i) = 1 - \frac{\varepsilon_{max}(x_i)}{\varepsilon(x_i)}$$

όπου $\varepsilon(x_i)$ η τιμή της παραμέτρου ε, στην οποία το x_i εισέρχεται σε μια συστάδα C και $\varepsilon_{max}(x_i)$ η τιμή της παραμέτρου ε στην οποία όλα τα αντικείμενα της C χαρακτηρίζονται ως θόρυβος. Όσο πιο κοντά στη μονάδα είναι ο βαθμός ανωμαλίας GLOSH, τόσο πιο πιθανό το αντικείμενο να είναι ανωμαλία. [Campello et al. (2015)]

5.3 Πλεονεκτήματα και Μειονεκτήματα Μεθόδων Βασισμένες στη Συσταδοποίηση

Το βασικό πλεονέκτημα των μεθόδων που βασίζονται στη συσταδοποίηση είναι ότι η φάση ελέγχου είναι αρκετά γρήγορη καθώς περιλαμβάνει τη σύγκριση του υπό μελέτη αντικειμένου

με ένα μικρό αριθμό συστάδων. Ακόμη, οι μέθοδοι ανήκουν στην κατηγορία των μη-εποπτευόμενων τεχνικών ανίχνευσης ανωμαλιών, και μπορούν να προσαρμοστούν σε πολύπλοκους τύπους δεδομένων, επιλέγοντας απλά τον κατάλληλο αλγόριθμο συσταδοποίησης που μπορεί να χειριστεί τον συγκεκριμένο τύπο δεδομένων.

Από την άλλη, η απόδοση των τεχνικών που βασίζονται στη συσταδοποίηση εξαρτάται σε μεγάλο βαθμό από την αποτελεσματικότητα του αλγορίθμου που επιλέγεται για την εύρεση των συστάδων των κανονικών παρατηρήσεων. Αρκετοί αλγόριθμοι συσταδοποίησης εξαναγκάζουν την ανάθεση κάθε παρατήρησης σε κάποια συστάδα. Αυτό μπορεί να οδηγήσει στην ανάθεση ανωμαλιών σε κάποια μεγάλη συστάδα, με αποτέλεσμα να χαρακτηριστούν ως κανονικές από τους αλγορίθμους που λειτουργούν υπό την υπόθεση ότι οι ανωμαλίες δεν ανήκουν σε καμία συστάδα. Ακόμη, αρκετές μέθοδοι είναι αποτελεσματικές μόνο όταν οι ανωμαλίες δεν σχηματίζουν κάποια σημαντική συστάδα μεταξύ τους. Τέλος, αρκετές τεχνικές ανιχνεύουν τις ανωμαλίες ως παραπροϊόν της ομαδοποίησης και συνεπώς δεν έχουν βελτιστοποιηθεί για το σκοπό της ανίχνευσης ανωμαλιών. [Chandola et al. (2009)]

Κεφάλαιο 6

Μέθοδοι AD για Δεδομένα Μεγάλης Διάστασης

6.1 Περιγραφή

Οι κλασσικές μέθοδοι ανίχνευσης ανωμαλιών εξαρτώνται κατά κάποιον τρόπο από τον πλήρους διάστασης (*full-dimensional*) Ευκλείδειο χώρο δεδομένων, προκειμένου να εξεταστούν οι ιδιότητες κάθε αντικειμένου για τον εντοπισμό ανωμαλιών. Ωστόσο, οι σημερινές εφαρμογές χαρακτηρίζονται από την παραγωγή δεδομένων μεγάλης διάστασης (*high dimensional data*).

Γενικά, η εξόρυξη συνόλων δεδομένων μεγάλης διάστασης παρουσιάζει ιδιαιτερότητες και δυσκολίες λόγω της κατάρας της διάστασης (*curse of dimensionality*). Η κατάρα της διάστασης αναφέρεται σε διάφορα φαινόμενα και προβλήματα που προκύπτουν όταν γίνεται ανάλυση δεδομένων σε χώρους μεγάλης διάστασης, τα οποία δεν εμφανίζονται σε μικρούς διάστασης χώρους.

Οι μέθοδοι ανίχνευσης ανωμαλιών για δεδομένα μεγάλης διάστασης, συνήθως, προσπαθούν να δημιουργήσουν μια προσέγγιση των δεδομένων χρησιμοποιώντας ένα συνδυασμό από χαρακτηριστικά, τα οποία καταγράφουν το μεγαλύτερο μέρος της μεταβλητότητας των δεδομένων. Αυτές οι τεχνικές στηρίζονται στην εξής σημαντική υπόθεση: Τα μεγάλης διάστασης δεδομένα μπορούν να ενσωματωθούν σε έναν υπόχωρο μικρότερης διάστασης, στον οποίο οι κανονικές και ανώμαλες παρατηρήσεις εμφανίζονται σημαντικά διαφορετικές.

Κατ’ αυτόν τον τρόπο, η γενική προσέγγιση που υιοθετείται από αυτές τις τεχνικές ανίχνευσης ανωμαλιών είναι ο καθορισμός τέτοιων υποχώρων, στους οποίους οι ανώμαλες παρατηρήσεις μπορούν να εντοπιστούν εύκολα. [Chandola et al (2009)]

6.2 Η Κατάρα της Διάστασης

Η κατάρα της διάστασης έχει παρακινήσει πολλές έρευνες στον τομέα των βάσεων δεδομένων λόγω των επιπτώσεών της στην εξόρυξη γνώσης σε δεδομένα μεγάλης διάστασης. Στην παρούσα ενότητα παρουσιάζονται η επιρροή και η επίδραση της κατάρας της διάστασης, και οι προκλήσεις που δημιουργούνται και πρέπει να αντιμετωπίσουν οι αλγόριθμοι ανίχνευσης ανωμαλιών.

Οι αποστάσεις των χαρακτηριστικών ανεξαρτήτων και ταυτόσημων κατανεμημένων (*i.i.d.*) δεδομένων μεγάλης κλίμακας, λόγω του Κεντρικού Οριακού Θεωρήματος, συγκλίνουν σε μια κανονική κατανομή με χαμηλή διακύμανση, προκαλώντας αριθμητικά ζητήματα και προβλήματα παραμετροποίησης. Πιο συγκεκριμένα, η αναλογία της διακύμανσης του μήκους, οποιουδήποτε σημείου, $\|X_d\|$, με το μήκος του μέσου των σημείων, $E[\|X_d\|]$, συγκλίνει στο μηδέν με την αύξηση της διάστασης των δεδομένων. Η συνέπεια αυτού είναι ότι η αναλογική διαφορά μεταξύ της απόστασης του μακρινότερου σημείου, D_{max}, και της απόστασης του κοντινότερου σημείου, D_{min}, εξαλείφεται. Επίσημα:

$$Αν \lim_{d\to\infty} var\left(\frac{\|X_d\|}{E[\|X_d\|]}\right) = 0, τότε \frac{D_{max} - D_{min}}{D_{min}} \to 0$$

Διαισθητικά, κάτω από τη παραπάνω υπόθεση η σχετική αντίθεση ανάμεσα σε κοντινούς και μακρινούς γείτονες μειώνεται, όσο αυξάνεται η διάσταση. Συνεπώς, αυτό το φαινόμενο συγκέντρωσης (*concentration effect*) των μέτρων απόστασης μειώνει τη χρησιμότητα των μέτρων απόστασης για τη διάκριση μεταξύ κοντινών και μακρινών γειτόνων. Όλες οι μελέτες σχετικά με το φαινόμενο συγκέντρωσης, γενικά, υποθέτουν ότι το σύνολο δεδομένων ακολουθεί μια ενιαία κατανομή, με την επιφύλαξη κάποιων περιορισμών. Στην πραγματικότητα, όταν τα δεδομένα ακολουθούν ένα συνδυασμό κατανομών, το φαινόμενο της συγκέντρωσης δεν παρατηρείται πάντοτε. Αντίθετα, σε τέτοιες περιπτώσεις, οι ασυμμετρίες μεταξύ των μελών διαφορετικών κατανομών ίσως να μην τείνουν στην γενική μέση τιμή (*global mean*) καθώς αυξάνεται η διάσταση.

Στα δεδομένα μεγάλης διάστασης, ένα υψηλό ποσοστό άσχετων (*irrelevant*), προσεγγιστικά i.i.d. κατανεμημένων, χαρακτηριστικών μπορεί να καλύψει (*mask*) τις σχετικές αποστάσεις, δηλαδή τις αποστάσεις που αποκαλύπτουν την διαφορά μεταξύ κανονικών και ανώμαλων σημείων. Εάν έχουμε αρκετές άσχετες, θορυβώδεις, διαστάσεις στο σύνολο δεδομένων, οι ανωμαλίες μπορούν εύκολα να καλυφθούν. Μόλις όμως επιλεγούν ως επί το πλείστον σχετικά

χαρακτηριστικά ή προβολές, οι ανωμαλίες γίνονται πολύ πιο διακριτές. Η πρόκληση στην επίλυση αυτού του προβλήματος είναι η σωστή επιλογή του υποχώρου.

Οι κοινές έννοιες του τοπικού (*local*), στην ανίχνευση τοπικών ανωμαλιών, βασίζονται σε γειτονιές που δημιουργούνται μέσω των αποστάσεων, καταλήγοντας συχνά στον φαύλο κύκλο της ανάγκης να είναι γνωστοί οι γείτονες για να επιλεχθεί ο σωστός υπόχωρος, και να είναι γνωστός ο υπόχωρος ώστε να επιλεχθούν οι κατάλληλοι γείτονες.

Οι βαθμοί ανωμαλίας, που βασίζονται σε L_p νόρμες, μεροληπτούν προς τους υπόχωρους μεγάλης διάστασης, εάν δεν έχει προηγηθεί κατάλληλη κανονικοποίηση (*normalization*). Συγκεκριμένα, οι αποστάσεις σε διαφορετικές διαστάσεις, και επομένως οι αποστάσεις που μετρούνται σε διαφορετικούς υπόχωρους, δεν είναι άμεσα συγκρίσιμες. Εάν οι βαθμοί ανωμαλίας, επηρεάζονται από τις τιμές των αποστάσεων, και οι τιμές αυτές ποικίλλουν αρκετά σε σχέση με τις διαφορετικές διαστάσεις, οι βαθμοί ανωμαλίας που προέρχονται από υπόχωρους διαφορετικών διαστάσεων δεν μπορούν να συγκριθούν.

Ωστόσο, οι αποστάσεις και οι βαθμοί ανωμαλίας που αξιοποιούν τις αποστάσεις μπορούν να παρέχουν μια εύλογη κατάταξη (*ranking*), παρόλο που, λόγω του φαινομένου της συγκέντρωσης, οι βαθμοί δεν διαφέρουν αρκετά. Όμως, η επιλογή ενός ορίου μεταξύ των κανονικών παρατηρήσεων και των ανωμαλιών, βάσει αυτών των μέτρων, μπορεί να είναι σχεδόν αδύνατη.

Ένα ακόμα πρόβλημα, λόγω της αυξημένης διάστασης, είναι ότι ο αριθμός των πιθανών υποχώρων αυξάνεται εκθετικά με τη διάσταση, καθιστώντας όλο και πιο δύσκολο τη συστηματική σάρωση του αρχικού χώρου. Ακόμη, δεδομένου πολλών υποχώρων, για ένα σημείο μπορεί να βρεθεί τουλάχιστον ένας υπόχωρος, τέτοιος ώστε το σημείο να θεωρείται ανωμαλία. Πρέπει να χρησιμοποιηθούν στατιστικές αρχές ελέγχων υποθέσεων σε διαφορετικά σύνολα αντικειμένων. Το συγκεκριμένο πρόβλημα ίσως είναι το πιο λεπτό και επίμονο. Μπορούμε να το δούμε και ως πρόβλημα υπερπροσαρμογής (*overfitting)*, καθώς το μοντέλο ανίχνευσης ανωμαλιών προσαρμόζεται υπερβολικά σε κάθε δεδομένο σημείο. Αυτή η άποψη καθιστά προφανές ότι αυτό το πρόβλημα είναι το πιο θεμελιώδες που πρέπει να αποφευχθεί σε οποιαδήποτε διαδικασία εκμάθησης.

Όπως έχουμε δει, τα συνδυαστικά ζητήματα των δεδομένων μεγάλης διάστασης μπορεί να είναι προβληματικά σε διάφορες πτυχές. Αφενός, ο χώρος αναζήτησης του μοντέλου μπορεί να εκτοξευθεί, καθιστώντας πολλές μεθόδους αναζήτησης άχρηστες. Από την άλλη πλευρά, η

αξιολόγηση ενός αντικειμένου έναντι πολλών πιθανών υποχώρων, μπορεί να εισάγει μια στατιστική μεροληψία και να μας κοστίσει στατιστική εγκυρότητα και σημαντικότητα.

Τέλος, έχει μελετηθεί η επίδραση της αποκαλούμενης κομβικότητας (*hubness*) στα δεδομένα μεγάλης διάστασης, σχετικά με τις διάφορες εργασίες εξόρυξης δεδομένων και μηχανικής μάθησης. Η κομβικότητα είναι ένα φαινόμενο γνωστό στην ανάλυση δεδομένων γράφων. Σε διανυσματικούς χώρους, οι κόμβοι (*hubs*) είναι σημεία σχετικά κοντά σε πολλά άλλα σημεία, δηλαδή, εντοπίζονται πολύ συχνά σε k-γειτονιές άλλων σημείων, ενώ άλλα σημεία μπορεί να εμφανίζονται σπάνια ή και ποτέ σε k-γειτονιές. Η κομβικότητα, ή πιο συγκεκριμένα η k-κομβικότητα ενός αντικειμένου o, είναι επομένως ο αριθμός των φορών που το o θεωρείται ως ένας από τους πλησιέστερους γείτονες οποιουδήποτε άλλου σημείου ενός συνόλου δεδομένων. Αποδεικνύεται ότι, με την αύξηση της διάστασης, πολλά σημεία δείχνουν μικρή ή μέτρια κομβικότητα ενώ μερικά σημεία παρουσιάζουν μια πολύ υψηλή κομβικότητα. Αν και οι αντι-κόμβοι φαίνεται λογικό να χαρακτηρίζονται ως ανωμαλίες βασισμένες στην απόσταση, και οι κόμβοι είναι επίσης σπάνιοι και ασυνήθιστοι και, επομένως, πιθανώς να είναι ανωμαλίες από πιθανοτική σκοπιά. Συνολικά, η σχέση της κομβικότητας και του βαθμού ανωμαλίας φαίνεται να παραμένει ένα ανοιχτό ζήτημα για την ανίχνευση ανωμαλιών σε δεδομένα μεγάλης διάστασης. [Zimek et al. (2012)]

6.3 Μέθοδοι Ανίχνευσης Ανωμαλιών

Αρκετές τεχνικές χρησιμοποιούν την ανάλυση κύριων συνιστωσών (*Principal Component Analysis - PCA*) για την προβολή των δεδομένων σε ένα χώρο μικρότερης διάστασης. Αυτές οι τεχνικές αναλύουν την προβολή κάθε δεδομένου σημείου κατά μήκος των κύριων συνιστωσών με χαμηλή διακύμανση. Μια κανονική παρατήρηση που ικανοποιεί την δομή της συσχέτισης του συνόλου δεδομένων θα έχει χαμηλή τιμή σε αυτές τις προβολές, ενώ μια ανώμαλη παρατήρηση που αποκλίνει από τη δομή συσχετισμού θα έχει μεγάλη τιμή.

Στα δεδομένα μεγάλης διάστασης, για την μείωση της διάστασης συχνά χρησιμοποιείται η ανάλυση κύριων συνιστωσών. Η PCA αφορά την εξήγηση της δομής της διακύμανσης – συνδιακύμανσης (*variance - covariance structure*) ενός συνόλου μεταβλητών, μέσω μερικών νέων μεταβλητών που είναι συναρτήσεις των αρχικών μεταβλητών. Οι κύριες συνιστώσες, συγκεκριμένα, είναι γραμμικοί συνδυασμοί των τυχαίων μεταβλητών $\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_p$ με τρείς σημαντικές ιδιότητες: Οι κύριες συνιστώσες είναι ασυσχέτιστες, η πρώτη κύρια συνιστώσα

έχει την μεγαλύτερη διακύμανση, η δεύτερη κύρια συνιστώσα έχει τη δεύτερη μεγαλύτερη διακύμανση κ.ο.κ., και η συνολική μεταβλητότητα όλων των κύριων συνιστωσών είναι ίση με τη συνολική μεταβλητότητα των αρχικών μεταβλητών $\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_p$. Οι κύριες συνιστώσες λαμβάνονται εύκολα με μια ανάλυση των ιδιοτιμών και των ιδιοδιανυσμάτων του πίνακα συνδιακύμανσης ή του πίνακα συσχέτισης των μεταβλητών $\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_p$.

Οι κύριες συνιστώσες από τον πίνακα συνδιακύμανσης και τον πίνακα συσχέτισης δεν είναι συνήθως οι ίδιες. Όταν κάποιες μεταβλητές έχουν πολύ μεγαλύτερες τιμές από άλλες, θα λάβουν μεγαλύτερα βάρη (*weights*) στις πρώτες κύριες συνιστώσες. Για το λόγο αυτό, αν οι μεταβλητές μετρούνται σε κλίμακες με πολύ διαφορετικό εύρος, ή εάν οι μονάδες μέτρησης δεν είναι ανάλογες, είναι προτιμότερο να εκτελείται PCA στον πίνακα συσχέτισης.

Έστω $\boldsymbol{R}$ ο $p \times p$ δειγματικός πίνακας συσχέτισης των p τυχαίων μεταβλητών $\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_p$, n παρατηρήσεων. Αν $(\lambda_1, \boldsymbol{e}_1)$, $(\lambda_2, \boldsymbol{e}_2)$, …,$(\lambda_p, \boldsymbol{e}_p)$ είναι τα p ζεύγη ιδιοτιμών-ιδιοδιανυσμάτων του $\boldsymbol{R}$, με $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$, τότε η i-οστή κύρια συνιστώσα μιας παρατήρησης $\boldsymbol{x} = (x_1, x_2, \dots, x_p)'$ είναι:

$$y_i = \boldsymbol{e}_i'\boldsymbol{z} = e_{i1}z_1 + e_{i2}z_2 + \cdots + e_{ip}z_p, \qquad i = 1,2, \dots, p$$

όπου $\boldsymbol{e}_i' = (e_{i1}, e_{i2}, \dots, e_{ip})'$ είναι το i-οστό ιδιοδιάνυσμα και $\boldsymbol{z} = (z_1, z_2, \dots, z_p)'$ είναι το τυποποιημένο διάνυσμα της παρατήρησης, το οποίο ορίζεται ως:

$$z_k = \frac{x_k - \bar{x}_k}{\sqrt{s_{kk}}}, \qquad k = 1,2, \dots, p$$

όπου $\bar{x}_k$ και s_{kk} είναι ο δειγματικός μέσος και η δειγματική διακύμανση της μεταβλητής $\boldsymbol{X}_k$.

Η i-οστή κύρια συνιστώσα έχει δειγματική διακύμανση λ_i και η δειγματική συνδιακύμανση οποιουδήποτε ανά δύο συνδυασμού κυρίων συνιστωσών είναι 0. Επιπλέον, η συνολική δειγματική διακύμανση όλων των κυρίων συνιστωσών ισούται με την συνολική δειγματική διακύμανση όλων των τυποποιημένων μεταβλητών $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \dots, \boldsymbol{Z}_p$, δηλαδή:

$$\lambda_1 + \lambda_2 + \lambda_p = p$$

Η παραπάνω ισότητα σημαίνει ότι όλη η μεταβλητότητα των αρχικών δεδομένων συνυπολογίζεται από τις κύριες συνιστώσες. [Shyu et al. (2003)]

Οι πρώτες q κύριες συνιστώσες έχουν μεγάλες διακυμάνσεις και εξηγούν αθροιστικά το μεγαλύτερο ποσοστό της συνολικής διακύμανσης. Αυτές οι κύριες συνιστώσες τείνουν να σχετίζονται με τα χαρακτηριστικά που έχουν σχετικά μεγάλες διακυμάνσεις και συνδιακυμάνσεις. Συνεπώς, οι παρατηρήσεις που αποκλίνουν σε σχέση με τις πρώτες q κύριες συνιστώσες, συνήθως αντιστοιχούν σε ανωμαλίες σε μια ή περισσότερες από τις αρχικές μεταβλητές. Συνεπώς, αν η προβολή της παρατήρησης $\boldsymbol{x}$ στις κύριες συνιστώσες είναι $y_1, y_2, \dots, y_p$ και οι αντίστοιχες ιδιοτιμές είναι $\lambda_1, \lambda_2, \dots, \lambda_p$, τότε το άθροισμα:

$$\sum_{i=1}^{q} \frac{y_i^2}{\lambda_i} = \frac{y_1^2}{\lambda_1} + \frac{y_2^2}{\lambda_2} + \cdots + \frac{y_q^2}{\lambda_q}, q \leq p$$

ακολουθεί κατανομή χ-τετράγωνο, με q βαθμούς ελευθερίας. Με αποτέλεσμα, η υπό μελέτη παρατήρηση να χαρακτηρίζεται ως ανωμαλία, δεδομένου ενός επίπεδου σημαντικότητας α, όταν ισχύει:

$$\sum_{i=1}^{q} \frac{y_i^2}{\lambda_i} > \chi_q^2(a)$$

Οι Shyu et al. (2003) παρουσιάζουν μια τεχνική ανίχνευσης ανωμαλιών στην οποία αξιοποιούν την ανθεκτική (*robust*) PCA για την εκτίμηση των κύριων συνιστωσών. Το μοντέλο που δημιουργείται αξιοποιεί εκτός από τις μεγαλύτερες κύριες συνιστώσες, και ένα ποσοστό των μικρότερων κύριων συνιστωσών.

Όταν χρησιμοποιείται PCA στην ανίχνευση ανωμαλιών είναι σημαντικό τα δεδομένα εκμάθησης να αποτελούνται μόνο από κανονικές παρατηρήσεις, καθώς οι ανωμαλίες μπορεί να προκαλέσουν μεγάλη αύξηση στις διακυμάνσεις, συνδιακυμάνσεις και στις συσχετίσεις. Οι Shyu et al. (2003), για να πραγματοποιήσουν ανθεκτική εκτίμηση του πίνακα συσχέτισης, προτείνουν μια μέθοδο κλαδέματος (*trimming*). Αρχικά, αφαιρούνται τυχόν ανώμαλες παρατηρήσεις από το σύνολο δεδομένων εκμάθησης αξιοποιώντας την απόσταση Mahalanobis. Για κάθε παρατήρηση $\boldsymbol{x}_i$, $i = 1,2, \dots, n$, υπολογίζεται η απόσταση Mahalanobis:

$$d_i^2 = (\boldsymbol{x}_i - \overline{\boldsymbol{x}})' \boldsymbol{S}^{-1} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})$$

όπου $\overline{\boldsymbol{x}}$ ο δειγματικές μέσος και $\boldsymbol{S}$ ο δειγματικός πίνακας συνδιακύμανσης.

Στη συνέχεια, δεδομένου ενός ποσοστού $\gamma\%$, οι παρατηρήσεις που αντιστοιχούν στις $\gamma \cdot n$ μεγαλύτερες τιμές του συνόλου $\{d_i^2, i = 1,2, \dots, n\}$ αφαιρούνται από το σύνολο εκμάθησης.

Υπολογίζονται εκ νέου οι εκτιμητές $\bar{\boldsymbol{x}}'$ και $\boldsymbol{S}'$,και ο ανθεκτικός εκτιμητής του πίνακα συσχέτισης λαμβάνεται από τα στοιχεία του $\boldsymbol{S}'$. Η διαδικασία κλαδέματος μπορεί να επαναληφθεί για να διασφαλιστεί ότι οι εκτιμητές $\bar{\boldsymbol{x}}'$ και $\boldsymbol{S}'$ είναι ανθεκτικοί στις ανωμαλίες, αρκεί ο αριθμός των παρατηρήσεων που απομένουν μετά το κλάδεμα να υπερβαίνει το p, ώστε ο εκτιμητής $\boldsymbol{S}'$ να είναι θετικά ορισμένος. Υποθέτοντας ότι στο σύνολο εκμάθησης το ποσοστό των ανωμαλιών είναι μικρό, προτείνεται $\gamma = 0.5\%$.

Τέλος, το προτεινόμενο μοντέλο αποτελείται από δύο συναρτήσεις βαθμών των κύριων συνιστωσών, μια από τις q μεγαλύτερες συνιστώσες, $\sum_{i=1}^{q} \frac{y_i^2}{\lambda_i}$, και μια από τις r μικρότερες συνιστώσες, $\sum_{i=p-r+1}^{p} \frac{y_i^2}{\lambda_i}$. Ένα σημείο $\boldsymbol{x}$ χαρακτηρίζεται ως ανωμαλία αν:

$$\sum_{i=1}^{q} \frac{y_i^2}{\lambda_i} > c_1 \quad \text{ή} \quad \sum_{i=p-r+1}^{p} \frac{y_i^2}{\lambda_i} > c_2$$

Ενώ το σημείο $\boldsymbol{x}$ χαρακτηρίζεται ως κανονικό αν:

$$\sum_{i=1}^{q} \frac{y_i^2}{\lambda_i} \leq c_1 \quad \textit{και} \quad \sum_{i=p-r+1}^{p} \frac{y_i^2}{\lambda_i} \leq c_2$$

όπου c_1,c_2 είναι τα όρια ανώμαλης συμπεριφοράς τέτοια ώστε το μοντέλο να παράγει ένα συγκεκριμένο ποσοστό ψευδών ανιχνεύσεων (*false alarm rate*). Θεωρώντας ότι τα δεδομένα ακολουθούν πολυμεταβλητή κανονική κατανομή, το ποσοστό ψευδών ανιχνεύσεων του μοντέλου είναι:

$$\alpha = \alpha_1 + \alpha_2 - \alpha_1\alpha_2$$

όπου $\alpha_1 = P\left(\sum_{i=1}^{q} \frac{y_i^2}{\lambda_i} > c_1 \middle| \boldsymbol{x}\ normal\right)$ και $\alpha_1 = P\left(\sum_{i=p-r+1}^{p} \frac{y_i^2}{\lambda_i} > c_2 \middle| \boldsymbol{x}\ normal\right)$. Οι τιμές των α_1 και α_2 επιλέγονται έτσι ώστε να αντικατοπτρίζουν τη σχετική σημαντικότητα (*relative importance*) των τύπων των ανωμαλιών που θέλουμε να ανιχνευτούν, και οι ερευνητές χρησιμοποιούν $\alpha_1 = \alpha_2$. Για παράδειγμα, για την επίτευξη 2% ποσοστού ψευδών ανιχνεύσεων, πρέπει $\alpha_1 = \alpha_2 = 0.0101$. Καθώς, η υπόθεση της κανονικότητας είναι πιθανό να παραβιάζεται, τα όρια ανώμαλης συμπεριφοράς c_1,c_2 υπολογίζονται από τα αντίστοιχα ποσοστημόρια, δηλαδή τα 0.9899 ποσοστημόρια των εμπειρικών κατανομών των $\sum_{i=1}^{q} \frac{y_i^2}{\lambda_i}$ και $\sum_{i=p-r+1}^{p} \frac{y_i^2}{\lambda_i}$, αντίστοιχα. [Shyu et al. (2003)]

Οι Lazarevic και Kumar (2005) προτείνουν μια τεχνική που συνδυάζει αποτελέσματα από πολλαπλούς αλγορίθμους ανίχνευσης ανωμαλιών που εφαρμόζονται σε διαφορετικά σύνολα χαρακτηριστικών. Κάθε αλγόριθμος ανίχνευσης ανωμαλιών χρησιμοποιεί ένα μικρό υποσύνολο χαρακτηριστικών που επιλέγεται τυχαία από το αρχικό σύνολο χαρακτηριστικών. Ως αποτέλεσμα, κάθε ανιχνευτής ανωμαλιών εντοπίζει διαφορετικές ανωμαλίες, και επομένως αναθέτει σε κάθε παρατήρηση έναν βαθμό ανωμαλίας, που αντιστοιχεί στην πιθανότητα η παρατήρηση να είναι ανωμαλία. Οι βαθμοί ανωμαλίας στη συνέχεια συνδυάζονται, ώστε να βρεθούν πιο αξιόπιστες ανωμαλίες.

Η διαδικασία συνδυασμού των τεχνικών ανίχνευσης ανωμαλιών αποτελείται από T επαναλήψεις. Σε κάθε t επανάληψη, επιλέγεται τυχαία ένα υποσύνολο χαρακτηριστικών F_t τέτοιο ώστε, ο αριθμός των χαρακτηριστικών του F_t επιλέγεται τυχαία μεταξύ του $\lfloor d/2 \rfloor$ και $(d-1)$, όπου d ο αριθμός των χαρακτηριστικών του αρχικού χώρου του συνόλου δεδομένων. Μόλις επιλεχθεί ο N_t αριθμός των χαρακτηριστικών για το υποσύνολο F_t, N_t χαρακτηριστικά επιλέγονται τυχαία, χωρίς επανατοποθέτηση, από τον αρχικό χώρο χαρακτηριστικών.

Στη συνέχεια, εφαρμόζεται ο αλγόριθμος ανίχνευσης ανωμαλιών και δημιουργείται ένα διάνυσμα, AS_t, με ταξινομημένους τους βαθμούς ανωμαλίας των παρατηρήσεων, από το μεγαλύτερο στο μικρότερο βαθμό. Για παράδειγμα, αν $AS_{t,1}(i) > AS_{t,2}(j)$, η παρατήρηση x_i έχει μεγαλύτερη πιθανότητα να είναι ανωμαλία από τη παρατήρηση x_j, βάσει του αλγορίθμου ανίχνευσης ανωμαλιών που εφαρμόστηκε στη t επανάληψη. Στο τέλος της διαδικασίας, μετά τις T επαναλήψεις, έχουν δημιουργηθεί T διανύσματα βαθμών ανωμαλίας καθένα από τα οποία αντιστοιχεί σε έναν αλγόριθμο ανίχνευσης ανωμαλιών.

Οι Lazarevic και Kumar (2005) προτείνουν δύο μεθόδους συνδυασμού των αποτελεσμάτων των T διανυσμάτων για την επιλογή των ανωμαλιών. Στη πρώτη μέθοδο, η κατάταξη των ανωμαλιών δημιουργείται ως εξής: Αρχικά επιλέγονται όλες οι παρατηρήσεις με τους μεγαλύτερους βαθμούς ανωμαλιών από όλες τις επαναλήψεις, δηλαδή οι παρατηρήσεις που αντιστοιχούν στους βαθμούς $AS_{1,1}, AS_{2,1}, AS_{3,1},\ldots, AS_{t,1}$, και εισέρχονται στο τελικό διάνυσμα. Στη συνέχεια επιλέγονται οι παρατηρήσεις με τους δεύτερους μεγαλύτερους βαθμούς ανωμαλίας, δηλαδή οι παρατηρήσεις με σκορ $AS_{1,2}, AS_{2,2}, AS_{3,2},\ldots, AS_{t,2}$, και προστίθενται στο τελικό διάνυσμα. Σε κάθε βήμα επιλογής, αν μια παρατήρηση έχει ήδη εισαχθεί στο τελικό διάνυσμα, τότε αυτή παραλείπεται. Τελικά, έχει δημιουργηθεί ένα

διάνυσμα με την κατάταξη των παρατηρήσεων ανάλογα με τον βαθμό ανωμαλίας τους, λαμβάνοντας υπόψιν το αποτέλεσμα κάθε επανάληψης t.

Στη δεύτερη μέθοδο που προτείνεται, προστίθενται οι βαθμοί ανωμαλίας κάθε παρατήρησης, από κάθε επανάληψη t, και στη συνέχεια η κατάταξη γίνεται με τους νέους βαθμούς ανωμαλίας. Πιο συγκεκριμένα, ο τελικός βαθμός ανωμαλίας της x_i παρατήρησης υπολογίζεται από τον εξής τύπο:

$$AS_{FINAL}(i) = \sum_{t=1}^{T} AS_t(i)$$

Αθροίζοντας τους βαθμούς ανωμαλίας, εντοπίζονται οι ανωμαλίες που μπορεί να είναι ορατές μόνο σε μερικές διαστάσεις. Σε αυτή τη περίπτωση, αρκεί η επιλογή σχετικών χαρακτηριστικών μόνο σε ένα μικρό αριθμό επαναλήψεων, ώστε να υπολογιστούν μεγάλοι βαθμοί ανωμαλίας και συνεπώς να προκαλέσουν μεγάλη κατάταξη στον AS_{FINAL}. [Lazarevic and Kumar (2005)]

Οι Ngugen et al. (2011) προτείνουν τον αλγόριθμο HighDOD (*High-dimensional Distance-based Outlier Detection*), ο οποίος εξετάζει υπόχωρους διάστασης έως ένα κατώφλι για την εύρεση των n μεγαλύτερων ανωμαλιών. Έστω ένα σύνολο δεδομένων DS με N δεδομένα d διάστασης. Κάθε διάσταση έχει ομαλοποιηθεί έτσι ώστε όλες να έχουν την ίδια κλίμακα, για παράδειγμα οι τιμές κάθε διάστασης κυμαίνονται στο διάστημα $[0,1]$. Η απόσταση μεταξύ οποιωνδήποτε δύο σημείων $p = (p_1, p_2, \dots, p_d)$ και $q = (q_1, q_2, \dots, q_d)$ σε ένα υπόχωρο $S = \{s_1, s_2, \dots, s_{\dim(S)}\} \subset \{1,2, \dots, d\}$ ορίζεται ως:

$$D(p^S, q^S) = \left(\sum_{i \in S} |p_i - q_i|^l\right)^{1/l}$$

όπου l είναι ένας θετικός ακέραιος. Ο βαθμός ανωμαλίας ενός σημείου p σε σχέση με τους k κοντινότερους γείτονές του σε έναν υπόχωρο S διάστασης $\dim(S)$, ορίζεται από την αθροιστική απόσταση γειτονιάς. Η αθροιστική απόσταση γειτονιάς ενός σημείου ορίζεται ως το άθροισμα των αποστάσεων του p από τους k κοντινότερους γείτονές του στο DS, έχοντας πάρει τις προβολές τους στο S, και κανονικοποιηθεί από τη διάσταση $\dim(S)$, δηλαδή:

$$FS_{out}(p, S) = \frac{1}{[\dim(S)]^{1/l}} \sum_{m \in kNN_p(S)} D(p^S, m^S)$$

όπου q^S είναι η προβολή ενός σημείου $q \in DS$ στον υπόχωρο S και $kNN_p(S)$ το σύνολο των k γειτόνων του p στον υπόχωρο S. Ο βαθμός ανωμαλίας FS_{out} εκτός του ότι αναθέτει ένα βαθμό ανά υπόχωρο σε κάθε δεδομένο, είναι επίσης αμερόληπτος της διάστασης (*dimensionality unbiased*) και ολικά συγκρίσιμος (*globally comparable*). Με άλλα λόγια, οι βαθμοί ανωμαλίας μεταξύ διαφορετικής διάστασης υπόχωρων έχουν την ίδια κλίμακα, καθιστώντας τον FS_{out} συγκρίσιμο ανεξαρτήτως της διάστασης του υπόχωρου.

Αρχικά, πραγματοποιείται μια εξολοκλήρου εξερεύνηση όλων των υπόχωρων διάστασης μέχρι και m. Πιο συγκεκριμένα, αρχικά εξετάζονται οι υπόχωροι μιας διάστασης, στη συνέχεια οι διδιάστατοι υπόχωροι και ούτω καθεξής. Προτείνεται το κατώφλι της μέγιστης διάστασης m να είναι ίσο με τον λογάριθμο του μεγέθους του συνόλου δεδομένων N, δηλαδή $m = \lfloor \log_{10} N \rfloor$, καθώς η απώλεια στην ακρίβεια δεν είναι τόσο σοβαρή, σε σχέση με το κέρδος στην ταχύτητα εκτέλεσης. Οι n παρατηρήσεις με τους μεγαλύτερους βαθμούς ανωμαλίας FS_{out} διατηρούνται σε κάθε επανάληψη και το κατώφλι επιλογής για το αν μια παρατήρηση ανήκει στις μεγαλύτερες n ανωμαλίες είναι ίσο με τον βαθμό της n-οστής ανωμαλίας που έχει βρεθεί μέχρι στιγμής.

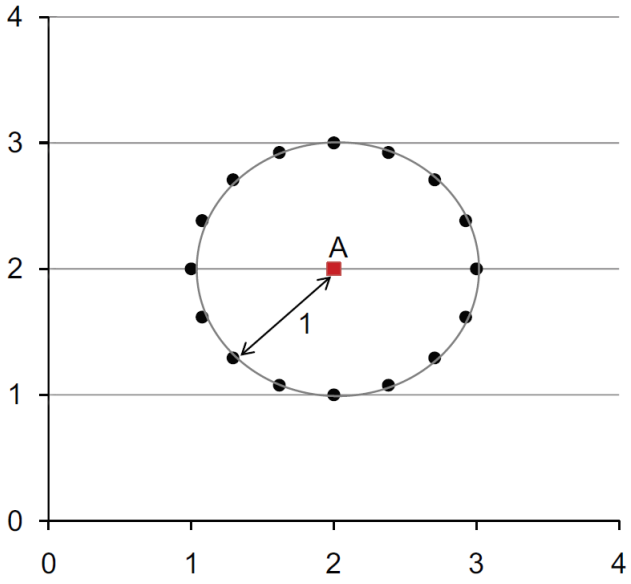
Για την βελτιστοποίηση του χρόνου εκτέλεσης του αλγορίθμου, οι Ngugen et al., προτείνουν την αξιοποίηση ενός εκτιμητή πυρήνα πυκνότητας (*kernel density estimator*) και συγκεκριμένα τη *Gaussian* συνάρτηση πυρήνα. Η βασική ιδέα είναι να γίνει η εκτίμηση της τοπικής πυκνότητας των δεδομένων χρησιμοποιώντας έναν εκτιμητή πυρήνα πυκνότητας, και να εξεταστούν ως υποψήφιες ανωμαλίες μόνο οι βn παρατηρήσεις με τη χαμηλότερη πυκνότητα. Αυτό προκύπτει από τη παραδοχή ότι οι ανωμαλίες είναι σπάνια φαινόμενα, δηλαδή δεν θα περιβάλλονται από πολλά αντικείμενα, με αποτέλεσμα να έχουν μικρή πυκνότητα. Σημειώνεται, ότι παρότι χρειάζεται η εξόρυξη n ανωμαλιών, οι υποψήφιες παρατηρήσεις πρέπει να είναι περισσότερες ώστε να αντιμετωπιστεί το σφάλμα του εκτιμητή πυρήνα πυκνότητας, δηλαδή πρέπει $\beta > 1$ και προτείνεται $\beta = 2$. Με άλλα λόγια, σε κάθε υπόχωρο επιλέγονται $2n$ παρατηρήσεις και εξετάζεται αν ανήκουν στις n μεγαλύτερες ανωμαλίες. [Ngugen et al. (2011)]

6.4 Πλεονεκτήματα και Μειονεκτήματα Μεθόδων AD για Δεδομένα Μεγάλης Διάστασης

Το βασικό πλεονέκτημα των μεθόδων ανίχνευσης ανωμαλιών για δεδομένα μεγάλης διάστασης είναι ότι μπορούν να εφαρμοστούν σε προβλήματα ημι-εποπτευόμενης και μη εποπτευόμενης ανίχνευσης. Ακόμη, επειδή εκτελούν αυτόματα τη μείωση της διάστασης, μπορούν να χρησιμοποιηθούν ως ένα στάδιο προ-επεξεργασίας και στη συνέχεια να γίνει εφαρμογή οποιασδήποτε μεθόδου ανίχνευσης ανωμαλιών στον μετασχηματισμένο διανυσματικό χώρο.

Το βασικό μειονέκτημα των μεθόδων ανίχνευσης ανωμαλιών για δεδομένα μεγάλης διάστασης είναι ότι είναι χρήσιμες μόνο όταν οι κανονικές και ανώμαλες παρατηρήσεις είναι διακριτές σε χαμηλής διάστασης υπόχωρους. Για παράδειγμα, στο Σχήμα 6.1 το σημείο *A* εντοπίζεται ως ανωμαλία στον διδιάστατο χώρο, ενώ αν μελετήσουμε τις τιμές των σημείων σε κάθε άξονα ξεχωριστά το σημείο *A* θα θεωρούνταν κανονικό. Ένα ακόμη μειονέκτημα των τεχνικών ανίχνευσης ανωμαλιών για δεδομένα μεγάλης διάστασης είναι ότι έχουν αναπόφευκτα μεγάλη υπολογιστική πολυπλοκότητα.

Σχήμα 6.1: *Παράδειγμα ανωμαλίας η οποία εντοπίζεται μόνο στον πλήρη χώρο.* [Ngugen et al. (2011)]



Κεφάλαιο 7

Εφαρμογή Αλγορίθμων Ανίχνευσης Ανωμαλιών

7.1 Περιγραφή

Στο παρόν κεφάλαιο γίνεται εφαρμογή και σύγκριση αλγορίθμων, με τη χρήση του στατιστικού προγράμματος R, στο σύνολο δεδομένων ALOI (*Amsterdam Library of Object Images*), στο πλαίσιο του προβλήματος μη-εποπτευόμενης ανίχνευσης ανωμαλιών. Το σύνολο δεδομένων ALOI προέρχεται από τη βιβλιοθήκη εικόνων αντικειμένων του Άμστερνταμ και είναι μια συλλογή έγχρωμων εικόνων από χιλιάδες μικρά αντικείμενα. Στο αρχικό σύνολο δεδομένων έγινε καταγραφή περίπου 110 εικόνων, από κάθε αντικείμενο, τραβηγμένων υπό διαφορετικές συνθήκες φωτισμού και γωνίες θέασης, αποδίδοντας συνολικά 110,000 εικόνες για τη συλλογή. Από τις πρωτότυπες εικόνες, δημιουργήθηκε ένα διάνυσμα 27 χαρακτηριστικών χρησιμοποιώντας ιστογράμματα χρώματος HSB. Ορισμένα αντικείμενα επιλέχθηκαν ως ανωμαλίες και τα δεδομένα υποβλήθηκαν σε δειγματοληψία έτσι ώστε το σύνολο δεδομένων που προέκυψε να περιέχει 50,000 παρατηρήσεις που περιλαμβάνουν 3.02% ανωμαλίες. [Goldstein and Uchiha (2016)]

7.1.1 Μέθοδος Αξιολόγησης Αλγορίθμων

Η σύγκριση της απόδοσης των μη εποπτευόμενων αλγορίθμων ανίχνευσης ανωμαλιών δεν είναι τόσο απλή, όπως για παράδειγμα στην περίπτωση της κλασσικής εποπτευόμενης κατηγοριοποίησης. Σε αντίθεση με την απλή σύγκριση της τιμής ακρίβειας θα πρέπει να ληφθεί υπόψη η σειρά των ανωμαλιών. Στην κατηγοριοποίηση, μια κακώς ταξινομούμενη περίπτωση είναι σίγουρα λάθος, ενώ στην μη εποπτευόμενη ανίχνευση ανωμαλιών αυτό είναι διαφορετικό. Για παράδειγμα, αν ένα μεγάλο σύνολο δεδομένων περιέχει δέκα ανωμαλίες και ταξινομηθούν μεταξύ των κορυφαίων δεκαπέντε ανωμαλιών, αυτό εξακολουθεί να είναι ένα καλό αποτέλεσμα, ακόμα και αν δεν είναι τέλειο.

Για το σκοπό αυτό, μια κοινή στρατηγική αξιολόγησης για τους μη εποπτευόμενους αλγορίθμους ανίχνευσης ανωμαλιών είναι η ταξινόμηση των αποτελεσμάτων σύμφωνα με το βαθμό ανωμαλίας, και στη συνέχεια η εφαρμογή ενός ορίου από την πρώτη έως την τελευταία τάξη. Αυτό οδηγεί σε ζεύγη των ποσοστών αληθώς θετικών (*true positive*) και ψευδών θετικών (*false positive*), τα οποία δημιουργούν μια καμπύλη λειτουργίας δέκτη ROC (*Receiver Operating Characteristic*). Έπειτα, η περιοχή κάτω από την καμπύλη AUC (*Area Under the Curve*), το ολοκλήρωμα της καμπύλης ROC, μπορεί να χρησιμοποιηθεί ως μέτρο απόδοσης. Η AUC ερμηνεύεται ως η πιθανότητα ενός αλγορίθμου ανίχνευσης ανωμαλιών να αποδώσει σε μια τυχαία επιλεγμένη κανονική παρατήρηση χαμηλότερο βαθμό ανωμαλίας από μια τυχαία επιλεγμένη ανωμαλία.

Κατά την αξιολόγησή μας, εφαρμόζουμε τους αλγορίθμους για τιμές της παραμέτρου k μεταξύ 10 και 50, και στο τέλος αναφέρουμε τη μέση AUC και το μέσο χρόνο εκτέλεσης, καθώς και τις αντίστοιχες τυπικές αποκλίσεις. Πιο συγκεκριμένα, για κάθε k, υπολογίζεται πρώτα η AUC μεταβάλλοντας ένα όριο μεταξύ των βαθμών ανωμαλίας, όπως περιγράψαμε παραπάνω, και ο χρόνος εκτέλεσης του αλγορίθμου. Στη συνέχεια, αναφέρεται η μέση τιμή για όλες τις τιμές AUC και για όλους τους χρόνους εκτέλεσης, και οι αντίστοιχες τυπικές αποκλίσεις. Αυτή η διαδικασία βασικά αντιστοιχεί σε μια στρατηγική τυχαίας επιλογής της τιμής του k στο συγκεκριμένο διάστημα, η οποία χρησιμοποιείται συχνά στη πράξη, όπου το k επιλέγεται αυθαίρετα.

7.1.2 Διαδικασία Παραλληλοποίησης και Εργαλεία Εφαρμογής

Για την επιτάχυνση της διαδικασίας σε χρόνο εκτέλεσης, αντί για σειριακή επεξεργασία (*serial processing*) εκτελούμε παράλληλη επεξεργασία (*parallel processing*). Έστω ότι πρέπει να εκτελεστεί μια σειρά $f_1, f_2, f_3, \ldots$ λειτουργιών. Στη σειριακή επεξεργασία, τρέχει πρώτα η f_1 και μέχρι να ολοκληρωθεί καμία άλλη f δεν μπορεί να τρέξει. Μόλις ολοκληρωθεί η f_1, αρχίζει η f_2 και η διαδικασία επαναλαμβάνεται, έως ότου να ολοκληρωθεί και η τελευταία f. Εν αντιθέσει, στην παράλληλη επεξεργασία όλες οι διαδικασίες f ξεκινούν ταυτόχρονα και εκτελούνται ανεξάρτητα.

Εάν διαθέτουμε έναν μόνο υπολογιστή και πρέπει να τρέξουμε n μοντέλα καθ' ένα από τα οποία χρειάζεται s δευτερόλεπτα για να εκτελεστεί, ο συνολικός χρόνος εκτέλεσης θα είναι $n \cdot s$. Εάν, ωστόσο, έχουμε $k < n$ υπολογιστές που μπορούμε να τρέξουμε τα μοντέλα μας, ο

συνολικός χρόνος εκτέλεσης θα είναι $n \cdot s/k$. Κατά την παλιά εποχή, αυτός ήταν ο τρόπος εκτέλεσης παράλληλου κώδικα, και εξακολουθεί να λειτουργεί σε μεγάλους διακομιστές (*servers*). Ωστόσο, οι σύγχρονοι υπολογιστές διαθέτουν επεξεργαστές πολλών πυρήνων (*multicore*), οι οποίοι μπορούν να ισοδυναμούν με την εκτέλεση πολλών υπολογιστών τη φορά.

Η R είναι ένα από τα πιο δημοφιλή στατιστικά λογισμικά, καθώς έχει πολλά πλεονεκτήματα, όπως πληθώρα στατιστικών μοντέλων, εργαλεία επεξεργασίας δεδομένων και ισχυρές δυνατότητες απεικόνισης δεδομένων. Ωστόσο, η R είναι ένα πρόγραμμα που από προεπιλογή εκτελεί τον κώδικα σε ένα πυρήνα. Έτσι, στους σύγχρονους επεξεργαστές πολλαπλών πυρήνων, η R δεν τους χρησιμοποιεί αποτελεσματικά όλους. Αυτό το πρόβλημα επιλύεται χρησιμοποιώντας το πακέτο '*parallel'*, στο οποίο περιλαμβάνονται οι εντολές που επιτρέπουν την παραλληλοποίηση στην R.

Ο ηλεκτρονικός υπολογιστής που χρησιμοποιήθηκε για την εφαρμογή έχει λειτουργικό σύστημα Windows 10 Pro 64bit και για την επίτευξη της παραλληλοποίησης αξιοποιήθηκαν δύο επεξεργαστές Inter® Xeon® E5-2660v4 2.00GHz, με σύνολο 28 πυρήνων και 56 threads. Ο κάθε επεξεργαστής έχει μνήμη Cache 35MB, και η μνήμη RAM του υπολογιστή είναι 128GB. Τέλος, το λειτουργικό σύστημα τρέχει σε ένα σκληρό δίσκο Samsung 960 Pro NVMe M.2 1TB, και για την αποθήκευση των δεδομένων χρησιμοποιήθηκαν 8 Seagate Barracuda Pro 8TB σκληροί δίσκοι χωρισμένοι σε δύο συστάδες των τεσσάρων δίσκων σε διαμόρφωση Raid 10 συνολικής χωρητικότητας 32TB.

7.2 Αποτελέσματα Αλγορίθμων

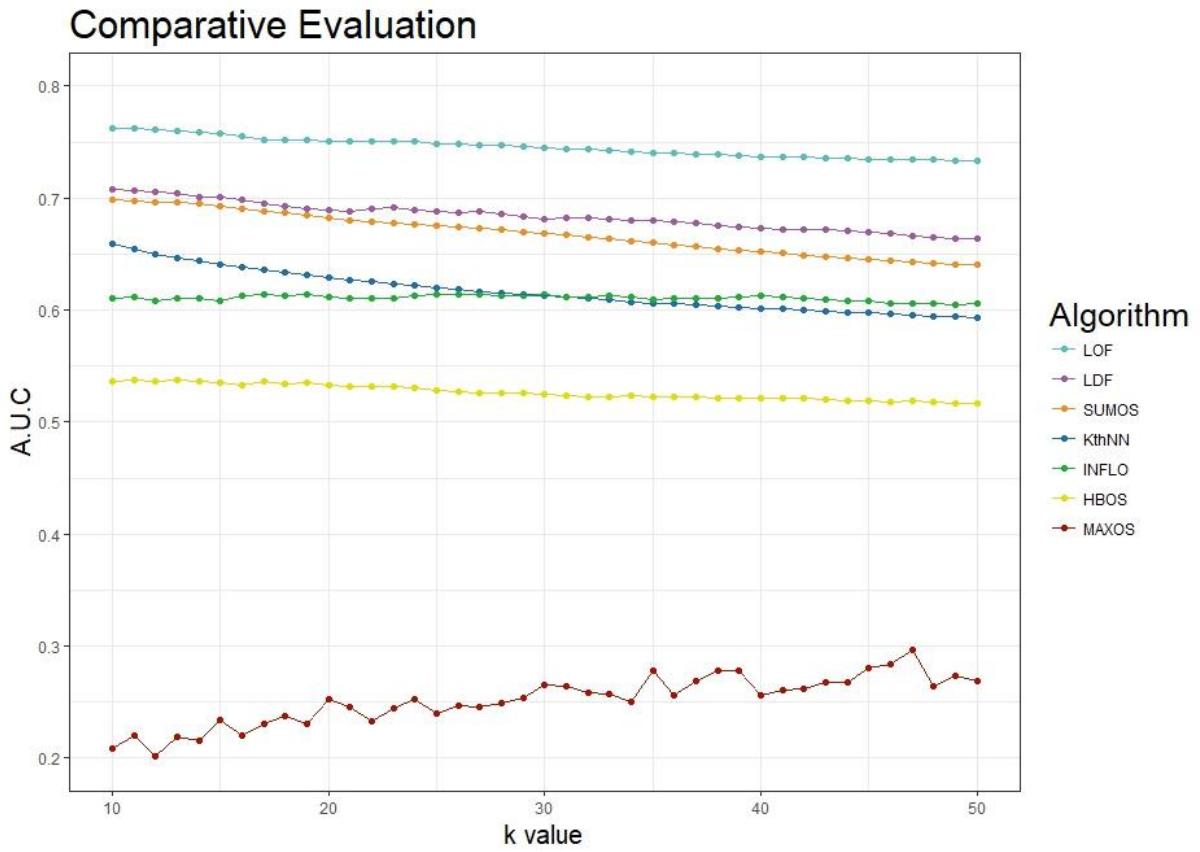
Στο σύνολο δεδομένων ALOI εφαρμόστηκαν από το Κεφάλαιο 2 οι αλγόριθμοι LDF και $HBOS$, από το Κεφάλαιο 3 οι αλγόριθμοι $D^k(p)$, LOF, $INFLO$ και από το Κεφάλαιο 6, ακολουθήθηκαν οι μέθοδοι ανίχνευσης ανωμαλιών των Lazarevic και Kumar (2005). Στις τυχόν απαιτούμενες παραμέτρους, εκτός της παραμέτρου k, χρησιμοποιήθηκαν οι προεπιλεγμένες τιμές, και για τις μεθόδους των Lazarevic και Kumar (2005) ο αριθμός των επαναλήψεων επιλέχθηκε $T = 15$, και σε κάθε επανάληψη εφαρμόζεται ο αλγόριθμος LOF.

Τα αποτελέσματα των αλγορίθμων παρουσιάζονται στον Πίνακα 7.1, όπου ο χρόνος εκτέλεσης (Χ.Ε.) μετριέται σε δευτερόλεπτα, ενώ στο Σχήμα 7.2 εμφανίζονται οι τιμές της AUC των αλγορίθμων για τιμές της παραμέτρου k μεταξύ 10 και 50.

Πίνακας 7.1: Αποτελέσματα Αλγορίθμων, για $10 \leq k \leq 50$.

Αλγόριθμος	Βέλτιστη τιμή k	Μέγιστη τιμή AUC	Μέση τιμή AUC	Τυπική Απόκλιση AUC	Μέση τιμή Χ. Ε.	Τυπική Απόκλιση Χ.Ε.
LOF	10	$\mathbf{0.7622}$	$\mathbf{0.7455}$	$\pm$ 0.0088	15.2624	$\pm$ 2.4098
LDF	10	0.7079	0.6838	$\pm$ 0.0124	16.2156	$\pm$ 2.7152
$SUM(OS)$	10	0.6987	0.6683	$\pm$ 0.0184	190.4290	$\pm$ 26.6731
$D^k(p)$	10	0.6590	0.6168	$\pm$0.0183	7.4490	$\pm$ 1.3554
$INFLO$	19	0.6143	0.6198	$\pm$ 0.0026	1,282.5090	$\pm$398.2038
$HBOS$	13	0.5375	0.5266	$\pm$ 0.0067	2.4341	$\pm$ 0.4569
$MAX(OS)$	47	0.2967	0.2516	$\pm$ 0.0218	5,744.4678	$\pm$ 415.5396

Σχήμα 7.1: *Τιμές A.U.C. των Αλγορίθμων, για* $10 \leq k \leq 50$.



Από τον Πίνακα 7.1 παρατηρούμε ότι στο σύνολο δεδομένων ALOI καλύτερη απόδοση στην μη εποπτευόμενη ανίχνευση ανωμαλιών έχει ο αλγόριθμος LOF με μέση τιμή AUC 0.7455. Αυτό σημαίνει, ότι ο αλγόριθμος LOF κατά μέσο όρο με πιθανότητα περίπου 75%

αποδίδει σε μια τυχαία επιλεγμένη κανονική παρατήρηση χαμηλότερο βαθμό ανωμαλίας από μια τυχαία επιλεγμένη ανωμαλία. Τον καλύτερο μέσο χρόνο εκτέλεσης έχει ο αλγόριθμος $HBOS$, με μέσο χρόνο εκτέλεσης 2.4 δευτερόλεπτα, ο οποίος όμως αποτυγχάνει στην σωστή ανίχνευση ανωμαλιών, στο σύνολο δεδομένων ALOI. Τη χειρότερη απόδοση ως προς τη μέση τιμή AUC, αλλά και ως προς τον χρόνο εκτέλεσης έχει η μέθοδος των Lazarevic και Kumar (2005), όπου ως βαθμός ανωμαλίας κάθε παρατήρησης επιλέγεται ο μεγαλύτερος βαθμός, από όλες τις T επαναλήψεις. Αναλυτικά τα πακέτα της R και ο κώδικας που χρησιμοποιήθηκε παρουσιάζονται στο Παράρτημα Β.

7.3 Σύνοψη – Συμπεράσματα

Στη παρούσα διπλωματική εργασία παρουσιάσαμε διάφορους τρόπους με τους οποίους το πρόβλημα της ανίχνευσης ανωμαλιών έχει διατυπωθεί και αντιμετωπιστεί. Για κάθε κατηγορία τεχνικών ανίχνευσης ανωμαλιών, ορίστηκε μια μοναδική υπόθεση σχετικά με την έννοια των κανονικών και των ανώμαλων δεδομένων. Κατά την εφαρμογή μιας συγκεκριμένης τεχνικής σε ένα συγκεκριμένο τομέα, αυτές οι υποθέσεις μπορούν να χρησιμοποιηθούν ως κατευθυντήριες γραμμές για την εκτίμηση της αποτελεσματικότητας της τεχνικής σε αυτόν τον τομέα.

Δεν υπάρχει κάποια μέθοδος ή τεχνική, η οποία να εφαρμόζεται σε οποιοδήποτε πρόβλημα ανίχνευσης ασυνήθιστης συμπεριφοράς. Στην ανίχνευση ασυνήθιστης συμπεριφοράς, ο αναλυτής θα πρέπει να εξετάσει σε ποια κατηγορία ανίχνευσης ανωμαλιών ανήκει το πρόβλημα που καλείται να αντιμετωπίσει. Δηλαδή, ανάλογα με τον τομέα εφαρμογής και τις διαθέσιμες ετικέτες, αν πρόκειται για πρόβλημα εποπτευόμενης, ημι-εποπτευόμενη ή μη-εποπτευόμενης ανίχνευσης ανωμαλιών. Θα πρέπει επίσης να επιλέξει έναν αλγόριθμο, ο οποίος να είναι κατάλληλος για το σύνολο των δεδομένων του, ανάλογα με την υποτιθέμενη κατανομή, τον τύπο των χαρακτηριστικών, την ταχύτητα, το μέγεθος των δεδομένων, καθώς και ανάλογα με το είδος των ανωμαλιών που καλείται να ανιχνεύσει.

Προσπαθήσαμε να παρουσιάσουμε ένα ευρύ δείγμα των υπαρχουσών μεθόδων, αλλά προφανώς δεν είναι δυνατό να περιγράψουμε όλες τις προσεγγίσεις. Ελπίζουμε να παρέχουμε στον αναγνώστη μια αίσθηση της ποικιλομορφίας και της πληθώρας των διαθέσιμων τεχνικών, καθώς και των προβλημάτων που πρέπει να ληφθούν υπόψη κατά την ανίχνευση ανωμαλιών.

Παραρτήματα

Παράρτημα Α: Μεγάλο O

Η σημειογραφία του μεγάλου O (*Big O*) χρησιμοποιείται στην Πληροφορική για να περιγράψει την απόδοση ή την πολυπλοκότητα ενός αλγορίθμου. Συγκεκριμένα, το *Big O* περιγράφει το χειρότερο σενάριο, και μπορεί να χρησιμοποιηθεί για να περιγράψει τον απαιτούμενο χρόνο εκτέλεσης ή τον χρησιμοποιούμενο χώρο (π.χ. στη μνήμη ή στο δίσκο) ενός αλγορίθμου.

Η αποτελεσματικότητα ενός αλγορίθμου μπορεί να εκτιμηθεί από το πόσο χρόνο χρειάζεται για να λειτουργήσει ως συνάρτηση του μεγέθους εισόδου n. Δύο συναρτήσεις $f(n)$ και $g(n)$ είναι της ίδιας τάξης (*order*) αν συμπεριφέρονται όμοια για μεγάλες τιμές του n, δηλαδή $f(n) \approx g(n)$ για μεγάλο n. Για παράδειγμα, ας θεωρήσουμε τις συναρτήσεις $f(n) = n^3$, $g(n) = n^3 + 3n + 4$ και $h(n) = 3^n$. Από τον Πίνακα Π1.1, παρατηρούμε ότι οι συναρτήσεις f και g είναι ίδιας τάξης, ενώ η h είναι μεγαλύτερης τάξης.

Πίνακας Π1.1: Αποτελέσματα συναρτήσεων $f(n)$, $g(n)$ και $h(n)$ για διάφορες τιμές του n.

n	1	10	50	100	300	500
$f(n)$	1	1000	125000	1000000	27000000	125000000
$g(n)$	8	1034	125154	1000304	27000904	125001504
$h(n)$	3	59049	$717898 \cdot 10^{18}$	$5153775 \cdot 10^{41}$	$1368915 \cdot 10^{137}$	$\approx 3636 \cdot 10^{235}$

Έστω μια συνάρτηση $f: \mathbb{R} \to \mathbb{R}$, ή η αντίστοιχη ακολουθία $a_n = f(n)$, τότε ορίζονται τα εξής:

1. $\Omega(f) = \left\{ g(x) \middle| \lim_{x \to \infty} \left| \frac{g(x)}{f(x)} \right| > 0 \right\}$ και αν $g \in \Omega(f)$ τότε η g είναι ίσης ή μεγαλύτερης τάξης από την f.
2. $O(f) = \left\{ g(x) \middle| \lim_{x \to \infty} \left| \frac{g(x)}{f(x)} \right| < \infty \right\}$ και αν $g \in O(f)$ τότε η g είναι μικρότερης ή ίσης τάξης από την f.

3. $\Theta(f) = \left\{g(x) \middle| \lim_{x\to\infty} \left|\frac{g(x)}{f(x)}\right| = L, 0 < L < \infty\right\}$ και αν $g \in \Theta(f)$, τότε οι f και g είναι περίπου ίδιας τάξης.
4. $o(f) = \left\{g(x) \middle| \lim_{x\to\infty} \left|\frac{g(x)}{f(x)}\right| = 0\right\}$ και αν $g \in o(f)$, τότε η g είναι μικρότερης τάξης από την f.

Η κύρια χρήση της σημειογραφίας της τάξης στην επιστήμη των υπολογιστών είναι η σύγκριση της αποτελεσματικότητας των αλγορίθμων. Στην περίπτωση αυτή, το n είναι το μέγεθος της εισόδου και το $f(n)$ είναι ο χρόνος λειτουργίας του αλγορίθμου σε σχέση με το μέγεθος εισόδου. Κάποιες συνηθισμένες τάξεις πολυπλοκότητας και τα ονόματά τους παρουσιάζονται στον Πίνακα Π1.2.

Πίνακας Π1.2: *Συμβολισμός και ονομασία τάξεων πολυπλοκότητας.*

Συμβολισμός	Ονομασία
$O(1)$	Σταθερή (*Constant*)
$O(\log(n))$	Λογαριθμική (*Logarithmic*)
$o(n)$	Υπο-γραμμική (*Sublinear*)
$O(n)$	Γραμμική (Linear)
$O(nlog(n))$	Λογαριθμογραμμική (*Loglinear*)
$O(n^2)$	Τετραγωνική (Quadratic)
$O(n^3)$	Κυβική (*Cubic*)
$O(n^c), c > 1$	Πολυωνυμική (*Polynomial*)
$O(c^n), c > 1$	Εκθετική (*Exponential*)
$O(n!)$	Παραγοντική (*Factorial*)

Παράρτημα Β: Κώδικας Εφαρμογής Αλγορίθμων

Στο συγκεκριμένο παράρτημα παρουσιάζεται ο κώδικας που χρησιμοποιήθηκε για την εφαρμογή των αλγορίθμων ανίχνευσης ανωμαλιών στο σύνολο δεδομένων ALOI. Αρχικά, φορτώνονται οι κατάλληλες βιβλιοθήκες του στατιστικού προγράμματος R, οι οποίες αποθηκεύονται στο αντικείμενο packages.

```
######################## LIBRARIES ########################

packages<-c("readr","parallel","pROC","DDoutlier","tidyverse","dbscan","rstudioapi")

for(package in packages){
  if(package %in% rownames(installed.packages()))  # if package is installed locally,
    do.call('library', list(package))              # load
  else {                                           # if package is not installed locally,
    install.packages(package)                      # download,
    do.call("library", list(package))              # then load
  }
}
```

Στη συνέχεια φορτώνεται το σύνολο δεδομένων, χωρίς τις ετικέτες στο αντικείμενο x, και τις ετικέτες τις αποθηκεύουμε στο αντικείμενο y για την αξιολόγηση της απόδοσης των αλγορίθμων. Ακόμη στο αντικείμενο k_range αποθηκεύεται το διάνυσμα με τις τιμές της παραμέτρους k και στο αντικείμενο n ο αριθμός των επαναλήψεων στην εφαρμογή των μεθόδων των Lazarevic και Kumar (2005).

Παρακάτω παρουσιάζεται ο κώδικας για την παράλληλη εκτέλεση των αλγορίθμων για τις διάφορες τιμές της παραμέτρου k.

- **LOF**

```
######################## LOF ########################

no_cores<-detectCores()-1                  #define the no of the pc's cores to run parallel

cl<-makeCluster(no_cores)                  #create parallel clusters

clusterEvalQ(cl,library("DDoutlier"))        #load needed library to clusters
clusterEvalQ(cl,library("tidyverse"))        #load needed library to clusters
clusterEvalQ(cl,library("pROC"))             #load needed library to clusters

clusterExport(cl,"x")                    #load needed variables to clusters
clusterExport(cl,"y")                    #load needed variables to clusters
```

```
fit_LOF<-parLapply(cl,k_range, function(k){
 ptm <- proc.time()                          # Start the timing the process

 z<-LOF(x,k) %>%
  as.tibble() %>%
  cbind(Type=y) %>%
  arrange(desc(value))

 roc_obj <- roc(z$Type, z$value,plot=T)  # create the ROC

 t<-proc.time() - ptm                        # Stop the timing

 c(AUC=auc(roc_obj),Time=t,k_value=k) # Calculate AUC and present results
})
stopCluster(cl)

LOF_results<-do.call(rbind,fit_LOF) %>% as.tibble()
```

- **LDF**

```
######################## LDF ########################

no_cores<-detectCores()-1                    # define the no of the pc's cores to run parallel

cl<-makeCluster(no_cores)                    # create parallel clusters

clusterEvalQ(cl,library("DDoutlier"))          # load needed library to clusters
clusterEvalQ(cl,library("tidyverse"))          # load needed library to clusters
clusterEvalQ(cl,library("pROC"))               # load needed library to clusters

clusterExport(cl,"x")                       # load needed variables to clusters
clusterExport(cl,"y")                       # load needed variables to clusters

fit_LDF<-parLapply(cl,k_range, function(k){
 ptm <- proc.time()                         # Start the clock

 z<-LDF(x,k)%>%
  as.tibble() %>%
  cbind(Type=y) %>%
  arrange(desc(LDF))
 roc_obj <- roc(z$Type, z$LDF,plot=T)

 t<-proc.time() - ptm                        # Stop the clock
 c(AUC=auc(roc_obj),Time=t,k_value=k)
})
stopCluster(cl)
LDF_results<-do.call(rbind,fit_LDF) %>% as.tibble()
```

- **INFLO**

```
######################## INFLO ########################

no_cores<-detectCores()-1                    #define the no of the pc's cores to run parallel

cl<-makeCluster(no_cores)                    #create parallel clusters

clusterEvalQ(cl,library("DDoutlier"))          #load needed library to clusters
clusterEvalQ(cl,library("tidyverse"))          #load needed library to clusters
clusterEvalQ(cl,library("pROC"))               #load needed library to clusters

clusterExport(cl,"x")                        #load needed variables to clusters
clusterExport(cl,"y")                        #load needed variables to clusters


fit_INFLO<-parLapply(cl,k_range, function(k){
 ptm <- proc.time()                          #Start the clock

 z<-INFLO(x,k) %>%
  as.tibble() %>%
  cbind(Type=y) %>%
  arrange(desc(value))

 roc_obj <- roc(z$Type, z$value,plot=T)

 t<-proc.time() - ptm                        #Stop the clock

 c(AUC=auc(roc_obj),Time=t,k_value=k)
})
stopCluster(cl)

INFLO_results<-do.call(rbind,fit_INFLO) %>% as.tibble()
```

- ***kth*-NN**

```
######################## kth-NN (D^k(p)) ########################
no_cores<-detectCores()-1                    #define the no of the pc's cores to run parallel
cl<-makeCluster(no_cores)                    #create parallel clusters

clusterEvalQ(cl,library("dbscan"))             #load needed library to clusters
clusterEvalQ(cl,library("tidyverse"))          #load needed library to clusters
clusterEvalQ(cl,library("pROC"))              #load needed library to clusters

clusterExport(cl,"x")                        #load needed variables to clusters
clusterExport(cl,"y")                        #load needed variables to clusters
```

```
fit_KthNN<-parLapply(cl,k_range, function(k){
 ptm <- proc.time()                          # Start the clock
 d<-kNN(x,k)                                 # find the k-neighbors for each data point
 z <-
  cbind(distance=d$dist[,k],Type=y) %>%      # select the distance to the kth neighbor
  arrange(desc(distance))                    # arrange distances from max to min
 roc_obj <- roc(z$Type, z$distance,plot=T)
 t<-proc.time() - ptm                        # Stop the clock
 c(AUC=auc(roc_obj),Time=t,k_value=k)
})
stopCluster(cl)

knn_results<-do.call(rbind,fit_KthNN) %>% as.tibble()
```

- **HBOS**

```
######################## HBOS ########################
no_cores<-detectCores()-1                   #define the no of the pc's cores to run parallel

cl<-makeCluster(no_cores)                   #create parallel clusters

clusterEvalQ(cl,library("tidyverse"))        #load needed library to clusters
clusterEvalQ(cl,library("pROC"))             #load needed library to clusters

clusterExport(cl,"x")                       # load needed variables to clusters
clusterExport(cl,"y")                       # load needed variables to clusters

fit_HBOS <- parLapply(cl,k_range, function(k){
 ptm <- proc.time()                         #Start the clock

 gg<-lapply(1:ncol(x), function(e){
  s<-hist(as.matrix(x[,e]),                                    #create Histogramms for each feature
       breaks = seq(min(as.matrix(x[,e])), max(as.matrix(x[,e])),length.out = k+1))

normlized_bin_dens<-s$counts/max(s$counts)                  #normalize the bin height/density

  os<-rep(1,nrow(x))
  for (g in 1:k){
   ind<-which(x[,e]>=s$breaks[g] & x[,e]<=s$breaks[g+1])
   os[ind]<-normlized_bin_dens[g]
  }
  os
 })
 os_matrix<-do.call(cbind,gg) %>% as.tibble()                 # create os matrix

 final_results<-
  cbind(HBOS=rowSums(log(1/os_matrix)),Type=y) %>%
  as.tibble() %>%
  arrange(desc(HBOS))
```

```
  roc_obj <- roc(final_results$Type, final_results$HBOS,plot=T)

  t<-proc.time() - ptm                    #Stop the clock

  c(AUC=auc(roc_obj),Time=t,k_value=k)
})
stopCluster(cl)

HBOS_results<-do.call(rbind,fit_HBOS) %>% as.tibble()
```

- **SUM(OS)**

```
no_cores<-detectCores()-1                 #define the no of the pc's cores to run parallel

cl<-makeCluster(no_cores)                 #create parallel clusters

clusterEvalQ(cl,library("DDoutlier"))      #load needed library to clusters
clusterEvalQ(cl,library("tidyverse"))      #load needed library to clusters
clusterEvalQ(cl,library("pROC"))           #load needed library to clusters

clusterExport(cl,"x")                     #load needed variables to clusters
clusterExport(cl,"y")                     #load needed variables to clusters
clusterExport(cl,"n")                     #load needed variables to clusters

fit_SUMOS<-parLapply(cl,k_range, function(k){
 ptm <- proc.time()                       #Start the clock

 r<-lapply(n, function(e){
  set.seed(1)
  no_f<-sample(floor(ncol(x)/2):27,1) #number of features
  f<-sample(1:ncol(x),no_f) %>% sort()

  z<-LOF(x[,f],k) %>%
   as.tibble() %>%
   cbind(Observation=c(1:nrow(x))) %>%
   arrange(desc(value))
  colnames(z)<-c(paste("OS",sep=""),
                 paste("Observation",sep="_"))
  z
 })
 interaction_results<-do.call(rbind,r) %>% as.tibble()
 final_results <-
  interaction_results %>%
  group_by(Observation) %>%
  summarise(FOS=sum(OS)) %>%
  cbind(Type=y)%>%
  arrange(desc(FOS)) %>%
  select(FOS,Type)
```

```
 roc_obj <- roc(final_results$Type, final_results$FOS,plot=T)

 t<-proc.time() - ptm                    #Stop the clock

 c(AUC=auc(roc_obj),Time=t,k_value=k)
})
stopCluster(cl)

SUMOS_results<-do.call(rbind,fit_SUMOS) %>% as.tibble()
```

- **MAX(OS)**

```
y_test<-y %>% mutate(Type=ifelse(Type==1,1,0),OBS=seq(1,nrow(y)))    # process y value

no_cores<-detectCores()-1                  #define the no of the pc's cores to run parallel

cl<-makeCluster(no_cores)                  #create parallel clusters

clusterEvalQ(cl,library("DDoutlier"))          #load needed library to clusters
clusterEvalQ(cl,library("tidyverse"))          #load needed library to clusters
clusterEvalQ(cl,library("pROC"))               #load needed library to clusters

clusterExport(cl,"x")                      #load needed variables to clusters
clusterExport(cl,"y_test")                     #load needed variables to clusters
clusterExport(cl,"n")

fit_MAXOS<-parLapply(cl,k_range, function(k){
 ptm <- proc.time()                        #Start the clock

 r<-lapply(n, function(e){
  no_f<-sample(floor(ncol(x)/2):27,1) #number of features
  f<-sample(1:ncol(x),no_f) %>% sort()

  z<-LOF(x[,f],k) %>%
   as.tibble() %>%
   cbind(Observation=c(1:nrow(x))) %>%
   arrange(desc(value))
  colnames(z)<-c(paste("OS",e,sep="_"),
           paste("OB",e,sep="_"))
  z
 })
 interaction_results<-do.call(cbind,r)

 observation_ranking <-
  interaction_results[,seq(2,2*length(n),2)] %>%
  t() %>%
  as.vector() %>%
  unique()
```

```
  l<-
    lapply(observation_ranking,function(f){
      a<-which(interaction_results==f,arr.ind = T)
      b<-a[which(a[,"col"] %% 2 ==0),]
      ind<-cbind(b[,"row"],seq(2,2*length(n),2)-1)
      max(interaction_results[ind],na.rm = T)
    })

  final_results <-
  do.call(rbind,l) %>%
    cbind(OBS=observation_ranking) %>%
    as.tibble() %>%
    left_join(y_test,by="OBS") %>% select(OS=V1,Type)

  pos.scores<-final_results %>% filter(Type==0)
  neg.scores<-final_results %>% filter(Type==1)

  t<-proc.time() - ptm                    #Stop the clock
  c(AUC=mean(
      sample(pos.scores$OS,100000,replace=T)>sample(neg.scores$OS,100000,replace=T)
           ),
     Time=t,
     k_value=k)
})
stopCluster(cl)

MAXOS_results<-do.call(rbind,fit_MAXOS) %>% as.tibble()
```

Για την απόκτηση των αποτελεσμάτων εφαρμόζουμε την παρακάτω συνάρτηση σε οποιοδήποτε αντικείμενο *_results.

```
best.results<-function(x){
  mean_auc <-mean(x$AUC)
  sd_auc   <-sd(x$AUC)
  mean_time<-mean(x$Time.elapsed)
  sd_time  <-sd(x$Time.elapsed)
  max_auc  <-max(x$AUC)
  best_k  <-(x$k_value[which(x$AUC==max_auc)])

  c(BEST_k  =round(best_k,4),
    MAX_AUC =round(max_auc,4),
    MEAN_AUC=round(mean_auc,4),
    SD_AUC  =round(sd_auc,4),
    MEAN_T  =round(mean_time,4),
    SD_T    =round(sd_time,4)
    )
}
```

Βιβλιογραφία

Agrawal, R., and Srikant, R. (1995). Mining sequential patterns. In Proceedings of the 11th International Conference on Data Engineering. IEEE Computer Society, 3–14.

Barnett, V., and Lewis, T. (1994). Outliers in Statistical Data. John Wiley & Sons., 3rd edition

Biscarri, F., Biscarri, J., Guerrero, J., León, C., Millán, R., and Monedero, I. (2012). Detection of frauds and other non-technical losses in a power utility using Pearson coefficient, Bayesian networks and decision trees, International Journal of Electrical Power & Energy Systems,Volume 34, Issue 1, Pages 90-98.

Brause, R., Langsdorf, T., and Hepp, M. 1999. Neural data mining for credit card fraud detection. In Proceedings of the IEEE International Conference on Tools with Artificial Intelligence. 103–106.

Breunig, M., Kriegel, H.-P., Ng, R. T., And Sander, J. (2000). LOF: Identifying density-based local outliers. In Proceedings of the ACM SIGMOD International Conference on Management of Data. ACM Press, 93–104.

Campello, R., J., G., B., Moulavi, D., Zimek, A., and Sander, J. (2015). "Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection". ACM Transactions on Knowledge Discovery from Data. 10 (1): 5:1–51.

Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly Detection: A Survey. ACM Computing Surveys. 41 (3):1–58

Chattopadhyay, P., and Konar, P. (2011). Bearing fault d`etection of induction motor using wavelet and Support Vector Machines (SVMs). Appl. Soft Comput. 11, 6 (September 2011), 4203-4211.

Dau, H., A., Ciesielski, V., and Song, A. (2014). Anomaly Detection Using Replicator Neural Networks Trained on Examples of One Class. In: Dick G. et al. (eds) Simulated Evolution and Learning. SEAL 2014. Lecture Notes in Computer Science, vol 8886. Springer, Cham

Davy, M., and Godsill, S.J. (2002). Detection of abrupt spectral changes using support vector machines an application to audio signal segmentation. 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2, II-1313-II-1316.

Duan, H., Li, X., and Tran, Q., (2004). One-class Support Vector Machine for Anomaly Network Traffic Detection.

Eskin, E. (2000). Anomaly Detection over Noisy Data Using Learned Probability Distributions. In Proceedings of the 17th International Conference on Machine Learning. Morgan Kaufmann Publishers Inc., 255–262.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, E. Simoudis, J. Han, and U. Fayyad, Eds. AAAI Press, 226–231

Galeano, P., Peña, D., and Tsay, S. (2006). Outlier Detection in Multivariate Time Series by Projection Pursuit. Journal of the American Statistical Association. 101. 654-669.

Goldstein, M., and Uchida, S. (2014). Behavior Analysis Using Unsupervised Anomaly Detection. In: The 10th Joint Workshop on Machine Perception and Robotics.

Goldstein, M., and Uchida, S. (2016). A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. PLoS ONE, 11(4):1–31

Grubbs, F., E. (1969). Procedures for detecting outlying observations in samples. Technometrics 11, 1–2

Hawkins D. (1980). Identification of Outliers. Chapman and Hall.

Hawkins, S., He, H., Williams, G., and Baxter, R. (2002). Outlier Detection Using Replicator Neural Networks. In: Kambayashi Y., Winiwarter W., Arikawa M. (eds) Data Warehousing and Knowledge Discovery. DaWaK 2002. Lecture Notes in Computer Science, vol 2454. Springer, Berlin, Heidelberg

He, Z., Xu, X., and Deng, S. (2003). Discovering Cluster-based Local Outliers. Pattern Recognition Letters. 24 (9–10):1641–1650.

He, Z., Xu, X., Huang, J.Z., and Deng, S. (2004). A Frequent Pattern Discovery Method for Outlier Detection. In: Li Q., Wang G., Feng L. (eds) Advances in Web-Age Information

Management. WAIM 2004. Lecture Notes in Computer Science, vol 3129. Springer, Berlin, Heidelberg

Hill, D., Minsker, B., and Amir, E. (2007). Real-time Bayesian anomaly detection for environmental sensor data. 32.

Hinneburg, A. and Keim, D. (1998) An Efficient Approach to Clustering in Large Multimedia Databases with Noise. Proceeding 4th International Conference on Knowledge Discovery & Data Mining, 58-65.

Janakiram, D., Reddy, V., and Kumar, A. (2006). Outlier detection in wireless sensor networks using Bayesian belief networks. In Proceedings of the 1st International Conference on Communication System Software and Middleware. 1–6.

Jin W., Tung A.K.H., Han J., and Wang W. (2006) Ranking Outliers Using Symmetric Neighborhood Relationship. In: Ng WK., Kitsuregawa M., Li J., Chang K. (eds) Advances in Knowledge Discovery and Data Mining. PAKDD 2006. Lecture Notes in Computer Science, vol 3918. Springer, Berlin, Heidelberg. 577–593.

Knorr, E., Ng, R., and Tucakov, V. (2000). Distance-Based Outliers: Algorithms and Applications. The VLDB Journal. 8. 237-253.

Korb, K., Mascaro, S., and Nicholso, A. (2014). Anomaly detection in vessel tracks using Bayesian networks, International Journal of Approximate Reasoning, Volume 55, Issue 1, Part 1, Pages 84-98,

Kriegel, H., P., Kröger, P., Schubert, E., and Zimek, A. (2009). LoOP: Local Outlier Probabilities. In: Proceeding of the 18th ACM Conference on Information and Knowledge Management (CIKM'09). New York, NY, USA: ACM Press.1649–1652

Lazarevic, A., and Kumar, V. (2005). Feature bagging for outlier detection. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 21. 157-166.

Laurikkala, J., Juhola, M., and Kentala, E. (2000). Informal Identification of Outliers in Medical Data. Fifth International Workshop on Intelligent Data Analysis in Medicine and Pharmacology.

Maes, S., Tuyls, K., Vanschoenwinkel., B., and Manderick., B. (2002). Credit Card Fraud Detection Using Bayesian and Neural Networks.

Mahoney, M. V., Chan, P. K., and Arshad, M. H. (2003). A machine learning approach to anomaly detection. Tech. rep. CS–2003–06, Department of Computer Science, Florida Institute of Technology Melbourne.

Münz, G., Li, S., and Carle, G. (2007). Traffic Anomaly Detection Using k-Means Clustering. In: In GI/ ITG Workshop MMBnet.

Nguyen H.V., Gopalkrishnan V., and Assent I. (2011) An Unbiased Distance-Based Outlier Detection Approach for High-Dimensional Data. In: Yu J.X., Kim M.H., Unland R. (eds) Database Systems for Advanced Applications. DASFAA 2011. Lecture Notes in Computer Science, vol 6587. Springer, Berlin, Heidelberg

Ramaswamy, S., Rastogi, R., and Shim, K. (2000). Efficient algorithms for mining outliers from large data sets. In Proceedings of the ACMSIGMOD International Conference on Management of Data. ACM Press, 427–438.

Russell, S., J., and Norvig, P. (2016). Artificial Intelligence: A Modern Approach (3rd ed.). Prentice Hall Press, Upper Saddle River, NJ, USA.

Scholkopf, B., Platt, J. C., Shawe-Taylor, J. C., Smola, A. J., and Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. Neural Comput. 13, 7, 1443–1471.

Shyu, M., Chen, S., Sarinnapakorn, K., and Chang, L. (2003). A Novel Anomaly Detection Scheme Based on Principal Component Classifier. Proceedings of International Conference on Data Mining.

Song, X., Wu, M., Jermaine, C., and Ranka, S. (2007). Conditional anomaly detection. IEEE Trans. Knowl. Data Eng. 19, 5, 631–645

Tandon, G., and Chan., P. (2007). Weighting versus pruning in rule validation for detecting network and host anomalies. In Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM Press.

Tax, D.M., and Duin, R.P. (2004). Machine Learning. 54: 45.

Tsay, S., and Peña, D., and Pankratz, A. (2000). Outliers in Multivariate Time Series. Biometrika.

Valdes, A., and Skinner, K. (2000). Adaptive, Model-Based Monitoring for Cyber Attack Detection. In: Debar H., Mé L., Wu S.F. (eds) Recent Advances in Intrusion Detection. RAID 2000. Lecture Notes in Computer Science, vol 1907. Springer, Berlin, Heidelberg

Wu, M. and Jermaine, C. (2006). Outlier detection by sampling with accuracy guarantees. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 767–772.

Yairi, T., Kato, Y., And Hori, K. (2001). Fault detection by mining association rules from housekeeping data. In Proceedings of the International Symposium on Artificial Intelligence, Robotics and Automation in Space.

Yin, S., Zhu, X., and Jing, C. (2014). Fault detection based on a robust one class support vector machine. Neurocomputing. 145. 263–268.

Zhang, L., Lin, J., and Karim, R. (2018). Adaptive Kernel Density-based Anomaly Detection for Nonlinear Systems. Knowledge-Based Systems, 139(1), 50–63.

Zimek, A. , Schubert, E. and Kriegel, H. (2012), A survey on unsupervised outlier detection in high-dimensional numerical data. Statistical Analy Data Mining, 5: 363-387.