ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΙΑΤΡΙΚΗ

Ομαδοποίηση και Κατηγοριοποίηση Δεδομένων Μεγάλης Διάστασης και Εφαρμογές

ΑΛΜΠΙΟΝΑ ΣΕΧΟΥ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ
Υπεύθυνος
Τασουλής Σωτήριος
Αναπληρωτής Καθηγητής

Λαμία, Οκτώβριος 2025

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΙΑΤΡΙΚΗ

Ομαδοποίηση και Κατηγοριοποίηση Δεδομένων Μεγάλης Διάστασης και Εφαρμογές

Αλμπιόνα Σέχου

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Επιβλέπων
Τασουλής Σωτήριος
Αναπληρωτής Καθηγητής

Λαμία, Οκτώβριος 2025

UNIVERSITY OF THESSALY
SCHOOL OF SCIENCE
DEPARTMENT OF COMPUTER SCIENCE AND BIOMEDICAL INFORMATICS

Clustering and Classification of High Dimensional Data and Applications

Almpiona Sechou

THESIS

Supervisor
Tasoulis Sotirios
Associate Professor

Lamia, October 2025

Με ατομική μου ευθύνη και γνωρίζοντας τις κυρώσεις [1], που προβλέπονται από της διατάξεις της παρ. 6 του άρθρου 22 του Ν. 1599/1986, δηλώνω ότι:

1. *Δεν παραθέτω κομμάτια βιβλίων ή άρθρων ή εργασιών άλλων αυτολεξεί* ***χωρίς να τα περικλείω σε εισαγωγικά*** *και χωρίς να αναφέρω το συγγραφέα, τη χρονολογία, τη σελίδα. Η αυτολεξεί παράθεση χωρίς εισαγωγικά χωρίς αναφορά στην πηγή, είναι λογοκλοπή. Πέραν της αυτολεξεί παράθεσης, λογοκλοπή θεωρείται και η παράφραση εδαφίων από έργα άλλων, συμπεριλαμβανομένων και έργων συμφοιτητών μου, καθώς και η παράθεση στοιχείων που άλλοι συνέλεξαν ή επεξεργάσθηκαν, χωρίς αναφορά στην πηγή. Αναφέρω πάντοτε με πληρότητα την πηγή κάτω από τον πίνακα ή σχέδιο, όπως στα παραθέματα.*

2. *Δέχομαι ότι η αυτολεξεί* ***παράθεση χωρίς εισαγωγικά****, ακόμα κι αν συνοδεύεται από αναφορά στην πηγή σε κάποιο άλλο σημείο του κειμένου ή στο τέλος του, είναι αντιγραφή. Η αναφορά στην πηγή στο τέλος π.χ. μιας παραγράφου ή μιας σελίδας, δεν δικαιολογεί συρραφή εδαφίων έργου άλλου συγγραφέα, έστω και παραφρασμένων, και παρουσίασή τους ως δική μου εργασία.*

3. *Δέχομαι ότι υπάρχει επίσης περιορισμός στο μέγεθος και στη συχνότητα των παραθεμάτων που μπορώ να εντάξω στην εργασία μου εντός εισαγωγικών. Κάθε μεγάλο παράθεμα (π.χ. σε πίνακα ή πλαίσιο, κλπ), προϋποθέτει ειδικές ρυθμίσεις, και όταν δημοσιεύεται προϋποθέτει την άδεια του συγγραφέα ή του εκδότη. Το ίδιο και οι πίνακες και τα σχέδια*

4. *Δέχομαι όλες τις συνέπειες σε περίπτωση λογοκλοπής ή αντιγραφής.*

Ημερομηνία: ……/…../20……

Η Δηλ.

Αλμπιόνα Σέχου

(1) «Όποιος εν γνώσει του δηλώνει ψευδή γεγονότα ή αρνείται ή αποκρύπτει τα αληθινά με έγγραφη υπεύθυνη δήλωση του άρθρου 8 παρ. 4 Ν. 1599/1986 τιμωρείται με φυλάκιση τουλάχιστον τριών μηνών. Εάν ο υπαίτιος αυτών των πράξεων σκόπευε να προσπορίσει στον εαυτόν του ή σε άλλον περιουσιακό όφελος βλάπτοντας τρίτον ή σκόπευε να βλάψει άλλον, τιμωρείται με κάθειρξη μέχρι 10 ετών.

Ομαδοποίηση και Κατηγοριοποίηση Δεδομένων Μεγάλης Διάστασης και Εφαρμογές

Αλμπιόνα Σέχου

Τριμελής Επιτροπή:

1. Τασουλής Σωτήριος, Αναπληρωτής Καθηγητής (επιβλέπων)
2. Δελήμπασης Κωνσταντίνος, Καθηγητής
3. Πλαγιανάκος Βασίλειος, Καθηγητής

ΠΕΡΙΛΗΨΗ

Η ομαδοποίηση ή συσταδοποίηση ενός συνόλου δεδομένων αποτελεί μια από τις πιο ενδιαφέρον προκλήσεις της μηχανικής μάθησης. Η ομαδοποίηση επιτυγχάνεται με τέτοιον τρόπο ώστε τα δεδομένα που ανήκουν στην ίδια συστάδα να παρουσιάζουν τη μέγιστη δυνατή ομοιότητα, ενώ ταυτόχρονα να έχουν τη μέγιστη δυνατή διαφορά από τα δεδομένα των άλλων συστάδων. Μία από τις σημαντικότερες προκλήσεις σε αυτή τη διαδικασία είναι η υψηλή διαστατικότητα (curse of dimensionality), η οποία εμφανίζεται σε σύνολα δεδομένων με μεγάλο αριθμό χαρακτηριστικών και καθιστά τις παραδοσιακές μεθόδους ομαδοποίησης αναποτελεσματικές. Για την αντιμετώπιση του προβλήματος αυτού, έχουν αναπτυχθεί τεχνικές μείωσης της διαστατικότητας. Ενώ κλασικές μέθοδοι όπως η Ανάλυση Κύριων Συνιστωσών (PCA) είναι ευρέως διαδεδομένες, ο στόχος τους δεν ταυτίζεται πάντα με την ανάδειξη της δομής των συστάδων. Μια πιο στοχευμένη προσέγγιση προσφέρουν οι μέθοδοι Αναζήτησης Προβολών (Projection Pursuit), οι οποίες αναζητούν «ενδιαφέρουσες» προβολές των δεδομένων σε χώρους χαμηλότερης διάστασης. Ωστόσο, η υπολογιστική πολυπλοκότητα των μεθόδων Projection Pursuit αποτελεί συχνά εμπόδιο. Αξιοποιώντας λοιπόν, τη μέθοδο των Τυχαίων Προβολών (Random Projections) αποφεύγετε το υπολογιστικό κόστος. Η παρούσα εργασία, βασιζόμενη στον αλγόριθμο Random Line Approximation Clustering (RLAC), προτείνει μια αποδοτική λύση. Χρησιμοποιεί τη μέθοδο των Τυχαίων Προβολών (Random Projections) με σκοπό τη μείωση της αρχικής διαστατικότητας του συνόλου δεδομένων, εφαρμόζει τη μέθοδο των Projection Pursuit έτσι ώστε να επιλέξει την πιο ενδιαφέρων προβολή από αυτές που δημιουργήθηκαν και ενσωματώνει όλα αυτά σε ένα ιεραρχικό διαιρετικό πλαίσιο συσταδοποίησης. Σε κάθε βήμα, τα δεδομένα διαχωρίζονται δυαδικά, επιλέγοντας την καλύτερη προβολή-διαχωρισμό μέσα από ένα πλήθος τυχαίων κατευθύνσεων, καθιστώντας τη διαδικασία υπολογιστικά εφικτή και ταυτόχρονα αποτελεσματική.

ABSTRACT

Clustering is a fundamental and challenging task in the field of machine learning. The primary goal of clustering is to partition a dataset into distinct groups, or clusters, such that data points within the same cluster exhibit high similarity, while being as dissimilar as possible from data points of other clusters. One of the most significant challenges in this process is the "curse of dimensionality," which arises in high-dimensional datasets and often renders traditional clustering algorithms ineffective. To mitigate this issue, dimensionality reduction techniques have been developed. Although classic methods like Principal Component Analysis (PCA) are widely used, their objective is not always aligned with revealing the underlying structure of the cluster. A more targeted approach is offered by Projection Pursuit methods, which seek to find "interesting" low-dimensional projections of the data. However, the computational complexity associated with Projection Pursuit methods often poses a significant barrier. This thesis, based on the Random Line Approximation Clustering (RLAC) algorithm, presents a lightweight and efficient solution. It utilizes the Random Projections method to reduce the initial dimensionality of the dataset, applies the Projection Pursuit method to select the most "interesting" projection from those created, and integrates all of this into a hierarchical divisive clustering framework. At each step, the data is split in a binary fashion by selecting the best projection for the split from a multitude of random directions, making the process both computationally feasible and effective.

ΕΥΧΑΡΙΣΤΙΕΣ

Στο σημείο αυτό, θα επιθυμούσα να ευχαριστήσω τα άτομα τα οποία με βοήθησαν και με υποστήριξαν να πραγματοποιήσω την παρούσα πτυχιακή εργασία.

Αρχικά, θα ήθελα να ευχαριστήσω τον επιβλέποντα της πτυχιακής εργασίας, Δρ. Τασουλή Σωτήριο, αναπληρωτή καθηγητή του τμήματος Πληροφορικής με Εφαρμογές στη Βιοϊατρική του Πανεπιστημίου Θεσσαλίας, για την ευκαιρία να συνεργαστούμε, την καθοδήγηση και την έμπνευση που παρείχε, και το χρόνο που διέθεσε για την πραγματοποίηση της εργασίας.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένειά και τους φίλους μου για την βοήθεια και στήριξη τους, τόσο στις όμορφες στιγμές όσο και στις δύσκολες.

Περιεχόμενα

Κατάλογος Εικόνων-Πινάκων

1. ΕΙΣΑΣΩΓΗ

Η συσταδοποίηση (clustering) αποτελεί μια θεμελιώδη διαδικασία στον τομέα της μη επιβλεπόμενης μηχανικής μάθησης, με κύριο στόχο την ανακάλυψη δομών και την ομαδοποίηση παρόμοιων δεδομένων σε διακριτές ομάδες, ή συστάδες. Η ικανότητα αυτόματης οργάνωσης μεγάλων συνόλων δεδομένων καθιστά τη συσταδοποίηση ένα πολύτιμο εργαλείο σε ένα ευρύ φάσμα επιστημονικών πεδίων, από τη βιοπληροφορική, στην ανάλυση εικόνας και πολλά άλλα.

Ωστόσο, η αποτελεσματικότητα πολλών κλασικών αλγορίθμων φθίνει δραματικά όταν εφαρμόζονται σε δεδομένα υψηλών διαστάσεων, ένα φαινόμενο γνωστό ως η «κατάρα της διαστατικότητας» (curse of dimensionality). Σε χώρους με πολλά χαρακτηριστικά, οι μετρικές απόστασης χάνουν το νόημά τους, τα δεδομένα γίνονται αραιά και ο εντοπισμός συμπαγών και καλά διαχωρισμένων συστάδων καθίσταται εξαιρετικά δύσκολος.

Μια ισχυρή στρατηγική για την αντιμετώπιση αυτής της πρόκλησης είναι η Αναζήτηση Προβολών (Projection Pursuit), η οποία αντί να υπολογίζει αποστάσεις στον πολυδιάστατο χώρο, αναζητά «ενδιαφέρουσες» προβολές των δεδομένων σε χώρους χαμηλότερης, συνήθως μίας, διάστασης. Σε αυτές τις προβολές, η δομή των συστάδων μπορεί να αποκαλυφθεί ως ένα μείγμα από διακριτές κατανομές. Ο αλγόριθμος Random Line Approximation Clustering (RLAC) αποτελεί μια υλοποίηση αυτής της φιλοσοφίας, χρησιμοποιώντας Τυχαίες Προβολές (Random Projections) για να εντοπίσει αποδοτικά τέτοιες κατευθύνσεις και εφαρμόζοντας μια ιεραρχική διαιρετική προσέγγιση για τον διαχωρισμό των δεδομένων.

Η παρούσα πτυχιακή εργασία ξεκίνησε με την υλοποίηση του αλγορίθμου RLAC, όπως αυτός περιγράφεται στην επιστημονική βιβλιογραφία [1]. Μια ισχυρή στρατηγική για την αντιμετώπιση αυτής της πρόκλησης είναι η Αναζήτηση Προβολών (Projection Pursuit), η οποία αντί να λειτουργεί στον αρχικό πολυδιάστατο χώρο, αναζητά «ενδιαφέρουσες» προβολές των δεδομένων σε χώρους χαμηλότερης, συνήθως μίας, διάστασης. Σε αυτές τις μονοδιάστατες προβολές, η πολύπλοκη δομή των συστάδων μπορεί να αποκαλυφθεί ως ένα μείγμα από διακριτές, συχνά πολυτροπικές (multimodal), κατανομές. Ο αλγόριθμος Random Line Approximation Clustering (RLAC) αποτελεί μια σύγχρονη υλοποίηση αυτής της φιλοσοφίας, αξιοποιώντας Τυχαίες Προβολές (Random Projections) για να εξερευνήσει αποδοτικά

τον χώρο των πιθανών κατευθύνσεων, ενσωματώνοντας τη διαδικασία σε ένα ιεραρχικό διαιρετικό πλαίσιο συσταδοποίησης.

Η εργασία αυτή εξερευνά και την επέκταση των δυνατοτήτων αυτού του πλαισίου, με κύριο στόχο τη διερεύνηση της επίδρασης που έχουν διαφορετικοί ορισμοί της «ενδιαφέρουσας» προβολής στην ποιότητα της τελικής ομαδοποίησης.

Για τον σκοπό αυτό, σχεδιάστηκε και ενσωματώθηκε ένα σύνολο διαφορετικών Δεικτών Αναζήτησης Προβολών (PPIs), πέραν αυτών που προτείνονται στην αρχική βιβλιογραφία. Συγκεκριμένα, η εργασία αυτή εισάγει και αναλύει τους εξής δείκτες τον δείκτη Νεοτροπίας (Negentropy), ο οποίος βασίζεται στη θεωρία της πληροφορίας για να μετρήσει την απόκλιση από την Γκαουσιανή (μη ενδιαφέρουσα) κατανομή. Το προσαρμοσμένο Κριτήριο του Fisher (Adapted Fisher's Discriminant PPI) για μη επιβλεπόμενη μάθηση, το οποίο αναζητά τον μέγιστο διαχωρισμό μεταξύ των σχηματιζόμενων ψευδο-ομάδων. Ένα δείκτη βασισμένο στα πολυώνυμα Hermite, που ανιχνεύουν δομές μέσω της εκτίμησης της ασυμμετρίας και της κύρτωσης. Τέλος τον δείκτη Friedman-Tukey, ο οποίος επιβραβεύει προβολές που ταυτόχρονα μεγιστοποιούν τη συνολική διασπορά και την τοπική πυκνότητα, αποκαλύπτοντας συμπαγείς, καλά διαχωρισμένες ομάδες.

Παράλληλα, η υλοποίηση ακολούθησε το άρθρο αναφοράς με την καθιέρωση ενός μηχανισμού για τον προσδιορισμό του σημείου διαχωρισμού, διασφαλίζοντας τη σταθερότητα και την αναπαραγωγιμότητα των πειραμάτων.

Στα κεφάλαια που ακολουθούν, παρουσιάζεται το θεωρητικό υπόβαθρο της συσταδοποίησης και της Αναζήτησης Προβολών, αναλύεται η υλοποίηση κάθε νέου δείκτη, παρατίθεται η πειραματική μεθοδολογία και τα συγκριτικά αποτελέσματα σε ποικίλα σύνολα δεδομένων, και, τέλος, συνάγονται τα τελικά συμπεράσματα.

2. ΥΠΟΒΑΘΡΟ

Το κεφάλαιο αυτό αποσκοπεί στην παρουσίαση του θεωρητικού πλαισίου πάνω στο οποίο βασίζεται η παρούσα πτυχιακή εργασία. Αρχικά, αναλύεται η αρχιτεκτονική και η θεμελιώδης λογική του αλγορίθμου Random Line Approximation Clustering (RLAC). Στη συνέχεια, εξετάζεται λεπτομερώς η ιδέα της Αναζήτησης Προβολών (Projection Pursuit) και παρουσιάζεται το σύνολο των δεικτών (PPIs) που υλοποιήθηκαν και αξιολογήθηκαν, τόσο εκείνοι από την αρχική δημοσίευση όσο και οι εναλλακτικοί που διερευνήθηκαν στο πλαίσιο της παρούσας μελέτης. Τέλος, παρατίθεται μια σύντομη περιγραφή των καθιερωμένων αλγορίθμων συσταδοποίησης που χρησιμοποιήθηκαν ως μέτρο σύγκρισης για την αξιολόγηση των αποτελεσμάτων.

2.1) Ο Αλγόριθμος Random Line Approximation Clustering (RLAC)

Ο αλγόριθμος Random Line Approximation Clustering (RLAC) είναι ένας ιεραρχικός διαιρετικός αλγόριθμος συσταδοποίησης, σχεδιασμένος ειδικά για να αντιμετωπίζει τις προκλήσεις των δεδομένων υψηλών διαστάσεων. Η κεντρική του φιλοσοφία δεν βασίζεται στον άμεσο υπολογισμό αποστάσεων στον αρχικό, πολυδιάστατο χώρο, αλλά στην προσέγγιση της Αναζήτησης Προβολών (Projection Pursuit). Σύμφωνα με αυτήν, η δομή των συστάδων, η οποία μπορεί να είναι κρυμμένη ή ασαφής στον αρχικό χώρο, συχνά αποκαλύπτεται σε χαμηλότερων διαστάσεων προβολές των δεδομένων.

Ο RLAC, αντί να αναζητά εξαντλητικά την βέλτιστη προβολή, προσεγγίζει τη λύση χρησιμοποιώντας Τυχαίες Προβολές (Random Projections). Αναζητά δηλαδή μονοδιάστατες τυχαίες γραμμές πάνω στις οποίες τα δεδομένα ενδέχεται να είναι διαχωρίσιμα. Ο αλγόριθμος λειτουργεί επαναληπτικά, ξεκινώντας με την υπόθεση ότι όλα τα δεδομένα αποτελούν μία ενιαία συστάδα. Σε κάθε βήμα, επιλέγει την καλύτερη τυχαία προβολή με βάση ένα συγκεκριμένο κριτήριο «ενδιαφέροντος» (PPI) και διαχωρίζει μια υπάρχουσα συστάδα σε δύο νέες υπο-συστάδες. Η διαδικασία αυτή συνεχίζεται μέχρι να επιτευχθεί ο προκαθορισμένος αριθμός συστάδων.

Η αρχιτεκτονική του RLAC είναι από τη φύση της διαιρετική (divisive), ακολουθώντας μια «top-down» προσέγγιση. Η διαδικασία ξεκινά θεωρώντας

ολόκληρο το σύνολο δεδομένων D ως μία μοναδική συστάδα C_0, η οποία αποτελεί τη ρίζα της ιεραρχίας.

Η λειτουργία του αλγορίθμου εξελίσσεται ως εξής:

1. *Εκκίνηση*: Το σύνολο των ενεργών συστάδων περιέχει μόνο τη συστάδα C_0.
2. *Επιλογή προς Διαχωρισμό*: Σε κάθε επανάληψη, ο αλγόριθμος επιλέγει μία συστάδα C_i από το σύνολο των ενεργών συστάδων ως υποψήφια προς διαχωρισμό.
3. *Διαχωρισμός*: Ο αλγόριθμος εφαρμόζει τον μηχανισμό διαχωρισμού (που θα αναλυθεί παρακάτω) στη συστάδα C_i, διαιρώντας την σε δύο νέες, διακριτές υπο-συστάδες, C_j και C_k.
4. *Ενημέρωση*: Η αρχική συστάδα C_i αφαιρείται από το σύνολο των ενεργών συστάδων και αντικαθίσταται από τις δύο νέες υπο-συστάδες C_j και C_k.
5. *Τερματισμός*: Η διαδικασία επαναλαμβάνεται μέχρι να ικανοποιηθεί ένα κριτήριο τερματισμού, το οποίο είναι η επίτευξη του επιθυμητού αριθμού συστάδων k.

Αυτή η ιεραρχική προσέγγιση είναι ιδιαίτερα διαισθητική και επιτρέπει στον αλγόριθμο να αποκαλύπτει δομές σε διαφορετικά επίπεδα ανάλυσης.

2.1.1) Μείωση Διαστατικότητας μέσω Τυχαίων Προβολών

Ο θεμελιώδης μηχανισμός που επιτρέπει στον αλγόριθμο RLAC να λειτουργεί αποδοτικά σε δεδομένα υψηλών διαστάσεων είναι η χρήση Τυχαίων Προβολών (Random Projections). Αντί να προσπαθεί να αναλύσει τις πολύπλοκες σχέσεις στον αρχικό, πολυδιάστατο χώρο, ο RLAC μετασχηματίζει τα δεδομένα σε έναν νέο χώρο πολύ χαμηλότερης διάστασης, όπου η αναζήτηση για διαχωρίσιμες δομές είναι υπολογιστικά εφικτή και πιο αποτελεσματική.

Η διαδικασία αυτή υλοποιείται μέσω ενός απλού πολλαπλασιασμού πινάκων. Έστω ότι το αρχικό σύνολο δεδομένων αναπαρίσταται από έναν πίνακα D διαστάσεων $n \times d$, όπου n είναι ο αριθμός των δειγμάτων και d ο αριθμός των αρχικών χαρακτηριστικών (διαστάσεων). Για να μειωθεί η διαστατικότητα, δημιουργείται ένας τυχαίος πίνακας προβολής RP διαστάσεων $d \times r$, όπου $r << d$ (r είναι ο αριθμός των νέων διαστάσεων). Ο νέος πίνακας προβαλλόμενων δεδομένων D προκύπτει από τον πολλαπλασιασμό:

$$D^P_{n\times r} = D_{n\times d} RP_{d\times r}$$

Κάθε μία από τις r στήλες του νέου πίνακα D αποτελεί πλέον μια μονοδιάστατη προβολή των αρχικών δεδομένων. Ουσιαστικά, κάθε στήλη αντιστοιχεί σε μια «τυχαία γραμμή-κατεύθυνση» πάνω στην οποία έχουν προβληθεί όλα τα αρχικά σημεία. Αυτό το σύνολο των r τυχαίων γραμμών αποτελεί τον χώρο αναζήτησης μέσα στον οποίο ο αλγόριθμος θα ψάξει για τον καλύτερο δυνατό διαχωρισμό.

Η φαινομενικά αυθαίρετη διαδικασία της τυχαίας προβολής έχει ισχυρή θεωρητική θεμελίωση στο λήμμα Johnson-Lindenstrauss (JL). Το λήμμα αυτό εγγυάται ότι οι αποστάσεις μεταξύ των σημείων σε έναν χώρο υψηλών διαστάσεων διατηρούνται προσεγγιστικά όταν τα σημεία προβάλλονται σε έναν τυχαία επιλεγμένο υποχώρο κατάλληλα χαμηλότερης διάστασης.

Πιο συγκεκριμένα, για ένα σύνολο n σημείων και για οποιοδήποτε $0 < \varepsilon < 1$ (όπου ε είναι ένας μικρός παράγοντας παραμόρφωσης), υπάρχει ένας γραμμικός μετασχηματισμός που μπορεί να προβάλει τα σημεία σε έναν χώρο r διαστάσεων, έτσι ώστε οι αποστάσεις μεταξύ των νέων σημείων να μην αποκλίνουν περισσότερο από έναν παράγοντα $(1 \pm \varepsilon)$ από τις αρχικές. Το πιο αξιοσημείωτο είναι ότι ο απαιτούμενος αριθμός διαστάσεων r εξαρτάται λογαριθμικά από τον αριθμό των σημείων n και τον παράγοντα παραμόρφωσης ε [$r \geq (4 * \log(n)) / (\varepsilon^2 / 2 - \varepsilon^3 / 3))$], αλλά είναι ανεξάρτητος από την αρχική διάσταση d. Αυτό καθιστά τη μέθοδο εξαιρετικά ισχυρή για δεδομένα με χιλιάδες χαρακτηριστικά. Για τον RLAC, αυτό σημαίνει ότι μπορούμε να μειώσουμε δραματικά τη διαστατικότητα του προβλήματος, διατηρώντας ταυτόχρονα τις σχετικές αποστάσεις και, κατ' επέκταση, τη δομή των συστάδων, με υψηλή πιθανότητα.

Η χρήση τυχαίων προβολών προσφέρει δύο καθοριστικά πλεονεκτήματα: υπολογιστική αποδοτικότητα και αποφυγή τοπικών βέλτιστων. Η διαδικασία είναι εξαιρετικά γρήγορη, καθώς απαιτεί έναν μόνο πολλαπλασιασμό πινάκων, σε αντίθεση με άλλες μεθόδους μείωσης διαστατικότητας όπως η Ανάλυση Κύριων Συνιστωσών (PCA), που απαιτεί τον υπολογισμό ιδιοτιμών και ιδιοδιανυσμάτων ενός πίνακα συνδιακύμανσης. Μέθοδοι που αναζητούν την "καλύτερη" προβολή με άπληστο τρόπο μπορεί να εγκλωβιστούν σε τοπικά βέλτιστα. Η τυχαία φύση του RLAC του επιτρέπει να εξερευνήσει ένα ευρύ φάσμα διαφορετικών, μη συσχετισμένων κατευθύνσεων, αυξάνοντας την πιθανότητα να εντοπίσει μια προβολή που αποκαλύπτει μια παγκοσμίως ενδιαφέρουσα δομή.

2.1.2) Εκτίμηση Πυκνότητας Πυρήνα (KDE) ως Μηχανισμός Διαχωρισμού

Ο αλγόριθμος RLAC προβάλει τα δεδομένα μιας υποψήφιας προς διαχωρισμό συστάδας πάνω σε ένα σύνολο r τυχαίων γραμμών, το επόμενο κρίσιμο βήμα είναι να αξιολογήσει αυτές τις μονοδιάστατες προβολές και να βρει την "καλύτερη" για διαχωρισμό. Ως "καλύτερη" προβολή ορίζεται εκείνη στην οποία η δομή των συστάδων είναι πιο εμφανής, πιο διαχωρίσιμη. Στο μονοδιάστατο χώρο, η ύπαρξη δύο ή περισσότερων συστάδων αντιστοιχεί στην παρουσία πολλαπλών κορυφών (modes) στην κατανομή πυκνότητας των δεδομένων.

Για να μοντελοποιήσει την κατανομή πυκνότητας κατά μήκος κάθε τυχαίας γραμμής, ο RLAC χρησιμοποιεί μια μη-παραμετρική τεχνική, την Εκτίμηση Πυκνότητας Πυρήνα (Kernel Density Estimation - KDE).

Η KDE είναι μια μέθοδος για την εκτίμηση της συνάρτησης πυκνότητας πιθανότητας (Probability Density Function - PDF) ενός συνόλου δεδομένων, χωρίς να γίνονται υποθέσεις για την κατανομή του (π.χ., ότι είναι Γκαουσιανή). Η ιδέα είναι απλή: για κάθε σημείο δεδομένων x_i, τοποθετείται μια συνάρτηση πυρήνα (kernel function) - συνήθως μια συμμετρική συνάρτηση όπως η Γκαουσιανή - κεντραρισμένη σε αυτό το σημείο. Η τελική εκτίμηση της πυκνότητας σε οποιοδήποτε σημείο x του χώρου προκύπτει από το άθροισμα των συνεισφορών όλων αυτών των συναρτήσεων πυρήνα. Η μαθηματική της μορφή είναι:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=0}^{n} K\left(\frac{x - Xi}{h}\right)$$

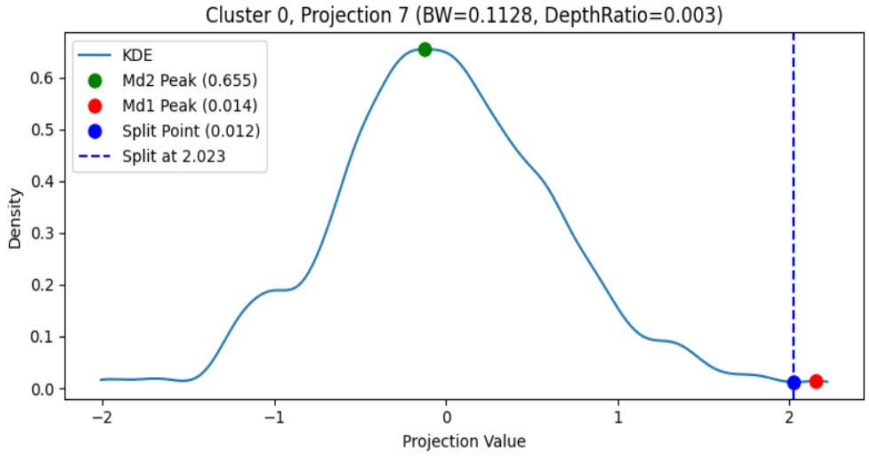
όπου:

- $f(x)$: η εκτιμώμενη πυκνότητα στο σημείο x
- n: είναι ο αριθμός των δειγμάτων
- h: είναι η παράμετρος εύρους ζώνης (bandwidth), η οποία ελέγχει την «ομαλότητα» της εκτιμώμενης καμπύλης.
- $K(\cdot)$: είναι η συνάρτηση πυρήνα.
- Xi: ένα από τα σημεία του συνόλου δεδομένων

Μέσα από τα πειράματα που πραγματοποιήθηκαν παρατηρείται ότι η επιλογή της παραμέτρου εύρους ζώνης h είναι ιδιαίτερα κρίσιμη για την ορθή εφαρμογή του

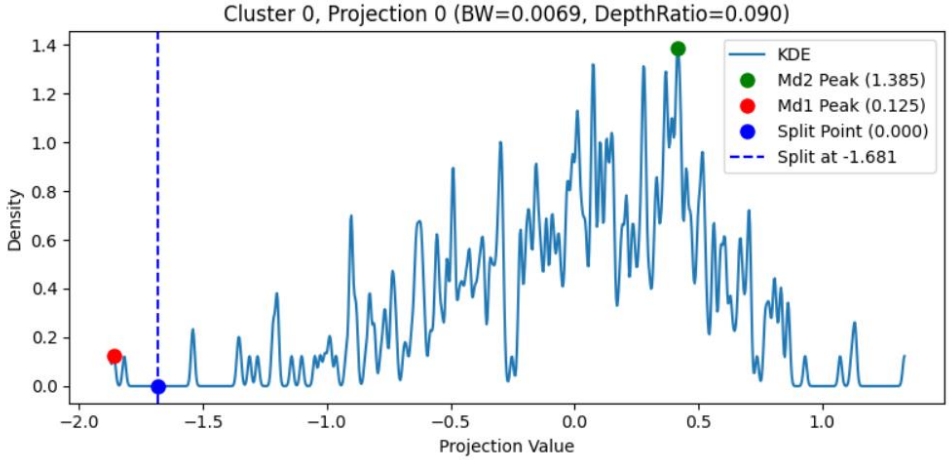
αλγορίθμου. Όπως αποδείχθηκε, αυτή η παράμετρος επηρεάζει δραματικά την ικανότητα του αλγορίθμου να ανιχνεύει δομές:

- Μεγάλο h (Over-smoothing): Μια μεγάλη τιμή για το εύρος ζώνης θα δημιουργήσει μια υπερβολικά «λεία» καμπύλη πυκνότητας. Αυτό μπορεί να «κρύψει» τις λεπτομερείς δομές, συγχωνεύοντας πολλαπλές κορυφές σε μία, καθιστώντας τον εντοπισμό συστάδων αδύνατο.



Εικόνα 1. Breast Cancer Dataset RLAC: Εφαρμογή μεγάλης τιμής εύρους ζώνης οδηγεί την κατανομή της πυκνότητας της προβολής σε απόκρυψη της δομής της και δημιουργία μιας «λείας» καμπύλης

- Μικρό h (Under-smoothing): Μια πολύ μικρή τιμή θα δημιουργήσει μια «τραχιά» καμπύλη, όπου κάθε μεμονωμένο σημείο μπορεί να δημιουργήσει τη δική του μικρή κορυφή, οδηγώντας σε μια θορυβώδη εκτίμηση που δεν αποκαλύπτει την πραγματική υποκείμενη δομή.



Εικόνα 2. Breast Cancer Dataset RLAC: Πολύ μικρή τιμή εύρους ζώνης οδηγεί σε μια «τραχιά» καμπύλη και πιθανώς σε μια θορυβώδη εκτίμηση της πυκνότητας

Η βέλτιστη ρύθμιση του h είναι επομένως μια πράξη εξισορρόπησης. Στην υλοποίηση του RLAC, χρησιμοποιείται ο κανόνας του Silverman για τον υπολογισμό του εύρους ζώνης.

Αφού υπολογιστεί η καμπύλη KDE για μια δεδομένη προβολή, ο αλγόριθμος την αναλύει για να βρει το καλύτερο σημείο διαχωρισμού. Η λογική είναι εξής: ένα καλό σημείο διαχωρισμού βρίσκεται σε μια περιοχή χαμηλής πυκνότητας (μια «κοιλάδα») και πιο συγκεκριμένα στο σημείο με την ολική ελάχιστη τιμή πυκνότητας (global minimum desnity point) στο γράφημα της κατανομής της πυκνότητας της μονοδιάστατης προβολής.
Έτσι λοιπόν, για κάθε μία από τις r τυχαίες προβολές, υπολογίζεται η καμπύλη KDE. Κάθε καμπύλη αξιολογείται με έναν Δείκτη Αναζήτησης Προβολών (PPI). Αυτός ο δείκτης ποσοτικοποιεί πόσο "ενδιαφέρουσα" είναι η εκάστοτε προβολή, μετρώντας ουσιαστικά το βάθος και την ευκρίνεια της πιο σημαντικής κοιλάδας. Η προβολή με το υψηλότερο σκορ PPI επιλέγεται ως η καλύτερη για τον διαχωρισμό. Το σημείο διαχωρισμού (split point) ορίζεται ως το σημείο της ολικής ελάχιστης πυκνότητας (global minimum) σε αυτήν την επιλεγμένη προβολή. Η συστάδα διαιρείται σε δύο νέες υπο-συστάδες, με βάση το αν τα σημεία της βρίσκονται αριστερά ή δεξιά από το σημείο διαχωρισμού. Αυτός ο μηχανισμός, που συνδυάζει τις τυχαίες προβολές με την ανάλυση πυκνότητας μέσω KDE, επιτρέπει στον RLAC να εντοπίζει γραμμικά διαχωρίσιμες συστάδες με έναν υπολογιστικά αποδοτικό τρόπο.

2.2) Δείκτες Αναζήτησης Προβολών

Η κεντρική φιλοσοφία πίσω από τον αλγόριθμο RLAC είναι η Αναζήτηση Προβολών (Projection Pursuit), μια ισχυρή τεχνική για την ανάλυση δεδομένων υψηλών διαστάσεων. Η θεμελιώδης υπόθεση της μεθόδου είναι ότι ενώ τα δεδομένα στον αρχικό τους, πολυδιάστατο χώρο μπορεί να φαίνονται ως ένα άμορφο «σύννεφο» σημείων, υπάρχουν συγκεκριμένες προβολές σε χώρους χαμηλότερης διάστασης (συνήθως σε μία γραμμή) που αποκαλύπτουν ενδιαφέρουσες δομές.
Στο πλαίσιο της συσταδοποίησης, «ενδιαφέρουσα» δομή σημαίνει η ύπαρξη σαφών ενδείξεων για την παρουσία δύο ή περισσότερων διακριτών ομάδων. Για να ποσοτικοποιηθεί αυτό το «ενδιαφέρον», χρησιμοποιούνται οι Δείκτες Αναζήτησης

Προβολών (Projection Pursuit Indexes - PPIs). Κάθε PPI είναι μια μαθηματική συνάρτηση που δέχεται ως είσοδο μια μονοδιάστατη προβολή των δεδομένων και επιστρέφει μια βαθμολογία (score) που αντικατοπτρίζει πόσο «καλή» είναι αυτή η προβολή για την αποκάλυψη δομής.

Η συντριπτική πλειοψηφία των δεικτών PPI βασίζεται σε μια κοινή αρχή: η πιο μη-ενδιαφέρουσα δομή είναι η κανονική (Γκαουσιανή) κατανομή. Σύμφωνα με το Κεντρικό Οριακό Θεώρημα, το άθροισμα πολλών ανεξάρτητων τυχαίων μεταβλητών τείνει να ακολουθεί την κανονική κατανομή. Κατ' αναλογία, οι περισσότερες τυχαίες προβολές πολυδιάστατων δεδομένων θα παράγουν κατανομές που μοιάζουν με Γκαουσιανές. Μια τέτοια κατανομή, με το χαρακτηριστικό σχήμα της καμπάνας, υποδηλώνει την ύπαρξη μιας ενιαίας, ομοιογενούς ομάδας δεδομένων.

Επομένως, ο στόχος της Αναζήτησης Προβολών είναι ο εντοπισμός εκείνων των προβολών που αποκλίνουν όσο το δυνατόν περισσότερο από την κανονική κατανομή. Αυτές οι μη-Γκαουσιανές προβολές είναι που αποκαλύπτουν την υποκείμενη δομή των συστάδων.

Ο αλγόριθμος RLAC, σε κάθε έναρξη της διαδικασίας διαχωρισμού, υπολογίζει τη τιμή ενός επιλεγμένου PPI για κάθε μία από τις r τυχαίες προβολές που έχει δημιουργήσει. Η προβολή που επιτυγχάνει την υψηλότερη βαθμολογία θεωρείται η πιο «ενδιαφέρουσα» και επιλέγεται για να πραγματοποιηθεί ο διαχωρισμός της τρέχουσας συστάδας. Η παρούσα εργασία διερεύνησε ένα ευρύ φάσμα τέτοιων δεικτών.

2.2.1) Δείκτες από την Αρχική Δημοσίευση (Depth Ratio, Dip-Test, Holes, Min and Max Kurtosis)

Στην εργασία που παρουσίασε τον αλγόριθμο Random Line Approximation Clustering (RLAC), προτάθηκε ένα πλήθος από δείκτες PPI, καθένας από τους οποίους στοχεύει στον εντοπισμό διαφορετικών τύπων δομής στα δεδομένα. Αυτοί οι δείκτες αποτέλεσαν το σημείο εκκίνησης για την έρευνα της παρούσας πτυχιακής.

Ο βασικότερος δείκτης, σχεδιασμένος ειδικά για το πλαίσιο του RLAC, είναι ο Depth Ratio PPI (Δείκτης Λόγου Βάθους). Η λειτουργία του βασίζεται εξ ολοκλήρου στην ανάλυση της καμπύλης πυκνότητας που παράγεται από την KDE, με κύριο στόχο να ποσοτικοποιήσει την ποιότητα μιας «κοιλάδας» ως διαχωριστή μεταξύ δύο «κορυφών». Μια ιδανική προβολή, σύμφωνα με αυτόν, είναι αυτή όπου τα δεδομένα

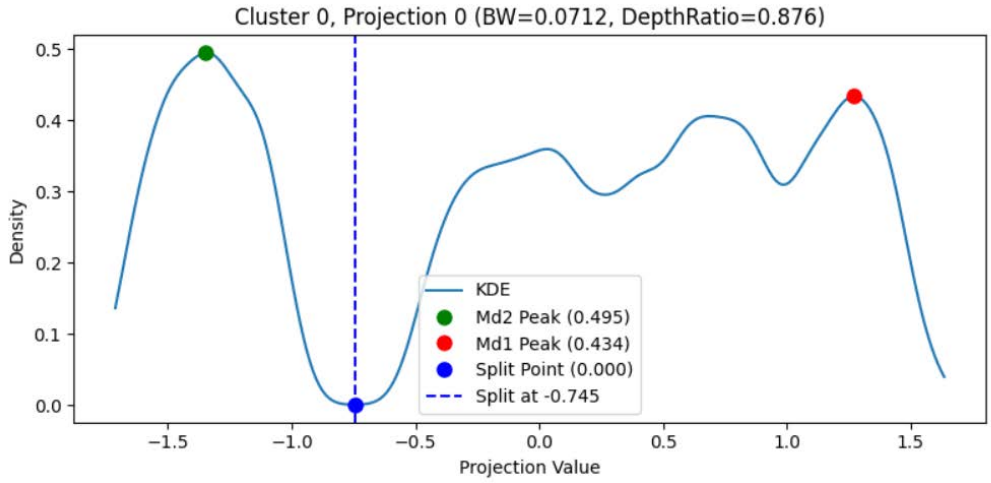
σχηματίζουν δύο ευδιάκριτες ομάδες, οι οποίες στην καμπύλη πυκνότητας μεταφράζονται σε δύο κορυφές (modes) που χωρίζονται από μια βαθιά κοιλάδα. Ο μηχανισμός του δείκτη πρώτα εντοπίζει το σημείο της ολικής ελάχιστης πυκνότητας της μονοδιάστατης προβολής, ορίζοντάς το ως υποψήφιο σημείο διαχωρισμού. Στη συνέχεια, εντοπίζει τις δύο υψηλότερες κορυφές που περιβάλλουν αυτή την κοιλάδα και υπολογίζει έναν λόγο που εκφράζει το βάθος της κοιλάδας σε σχέση με το ύψος των κορυφών. Μια υψηλή βαθμολογία δείχνει ότι η προβολή φανερώνει σαφείς διαχωρισμούς των δεδομένων.

Ο μαθηματικός τύπος του δείκτη αυτού είναι:

$$DepthRatio(rpj) = \frac{\hat{f}'(\mathrm{Md1};\mathrm{h}') - \hat{f}'(\mathrm{x}*;\mathrm{h})}{\hat{f}'(\mathrm{Md2};\mathrm{h}')}$$

όπου:

- DepthRatio(rp_j): Η τελική βαθμολογία (score) για τη συγκεκριμένη προβολή.
- $\hat{f}(...)$ = Η συνάρτηση Εκτίμησης Πυκνότητας Πυρήνα (KDE), η οποία δίνει την τιμή της πυκνότητας (το ύψος της καμπύλης) σε ένα δεδομένο σημείο.
- x* = Η θέση (τιμή στον άξονα x) του πυθμένα της κοιλάδας (το σημείο με την ολική χαμηλότερη πυκνότητα). Η θέση όπου θα πραγματοποιηθεί και ο διαχωρισμός.
- Md_1 = Η θέση της χαμηλότερης από τις δύο κορυφές που περιβάλλουν την κοιλάδα.
- Md_2 = Η θέση της υψηλότερης από τις δύο κορυφές που περιβάλλουν την κοιλάδα.



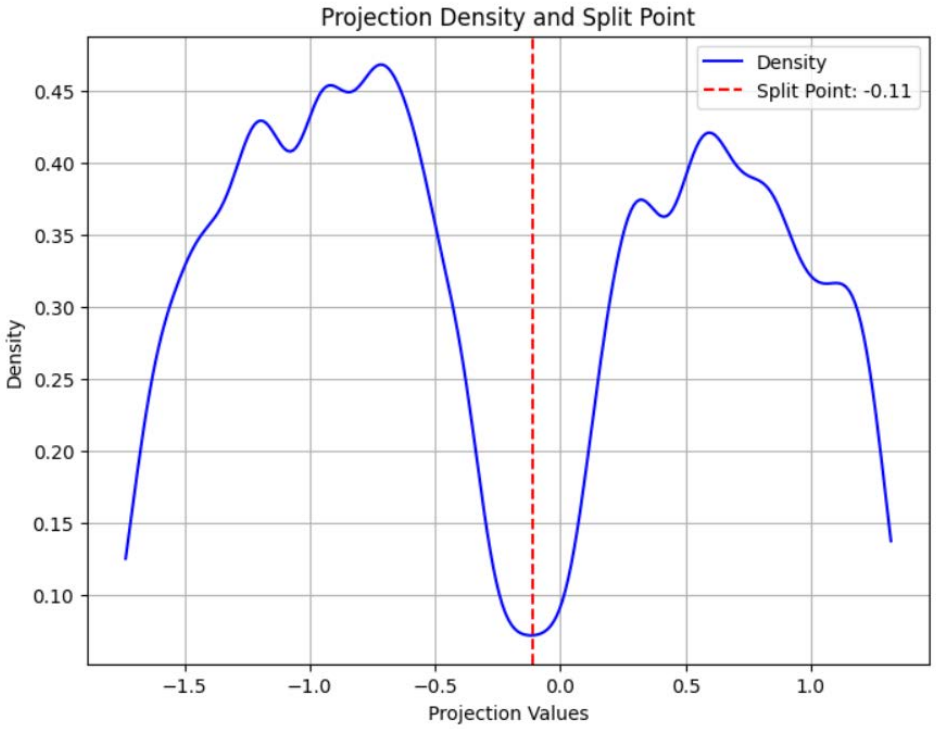
Εικόνα 3. Aggregation Dataset RLAC-Depth Ratio (Εύρος Ζώνης = 0.3, Πλήθος Προβολών = 50, Random_State = 44)

Επόμενος δείκτης που αναφέρεται στο άρθρο είναι ο Hartigan's Dip Test PPI, ο οποίος αξιοποιεί έναν καθιερωμένο, μη-παραμετρικό στατιστικό έλεγχο για την αξιολόγηση της υπόθεσης της μονοτροπικότητας (unimodality). Σκοπός του είναι να ανιχνεύσει στατιστικά σημαντικές αποκλίσεις από μια κατανομή με μία μόνο κορυφή, καθώς μια τέτοια απόκλιση αποτελεί ισχυρή ένδειξη πολυτροπικότητας και, κατ' επέκταση, ύπαρξης πολλαπλών συστάδων. Ο έλεγχος λειτουργεί πάνω στην Αθροιστική Συνάρτηση Κατανομής (CDF) και ορίζει την τιμή "dip" ως τη μέγιστη κάθετη απόσταση μεταξύ της εμπειρικής CDF των δεδομένων και της πλησιέστερης unimodal CDF. Μια προβολή με υψηλό σκορ Dip Test είναι πιθανότατα πολυτροπική. Σε αντίθεση με το Depth Ratio, ο Dip Test δεν εντοπίζει από μόνος του το σημείο διαχωρισμού, το οποίο στην πράξη εντοπίζεται εκ των υστέρων βρίσκοντας το ελάχιστο σημείο της KDE για λόγους συνέπειας. Ο τύπος του δείκτη αυτού είναι ο εξής:

$$dip(F) = d_{\infty}(F, U) = min_{UeU}||F - U||_{\infty}$$

όπου:

- dip(F): Το σκορ (dip statistic) για μια δεδομένη κατανομή F.
- F: Η εμπειρική Αθροιστική Συνάρτηση Κατανομής (Empirical Cumulative Distribution Function - CDF) των δεδομένων.
- U: Το U είναι μια οποιαδήποτε μονοτροπική (unimodal) Αθροιστική Συνάρτηση Κατανομής. Το U είναι το σύνολο όλων των πιθανών unimodal CDFs.
- min_{U ∈ $\mathcal{U}$}:Ο τελεστής ελαχίστου που εφαρμόζεται σε όλες τις πιθανές unimodal κατανομές U που ανήκουν στο σύνολο $\mathcal{U}$.
- || F - U ||$_{\infty}$: είναι η διαφορά μεταξύ της πραγματικής CDF των δεδομένων σου (F) και μιας υποψήφιας unimodal CDF (U). Με την νόρμα άπειρο σημαίνει ότι από όλες τις κάθετες αποστάσεις μεταξύ των δύο καμπυλών F και U, κρατάμε μόνο τη μέγιστη απόσταση.



Εικόνα 4. Aggregation Dataset RLAC-Dip_Test (Εύρος Ζώνης = 0.3, Πλήθος Προβολών = 50, Random_State = 44)

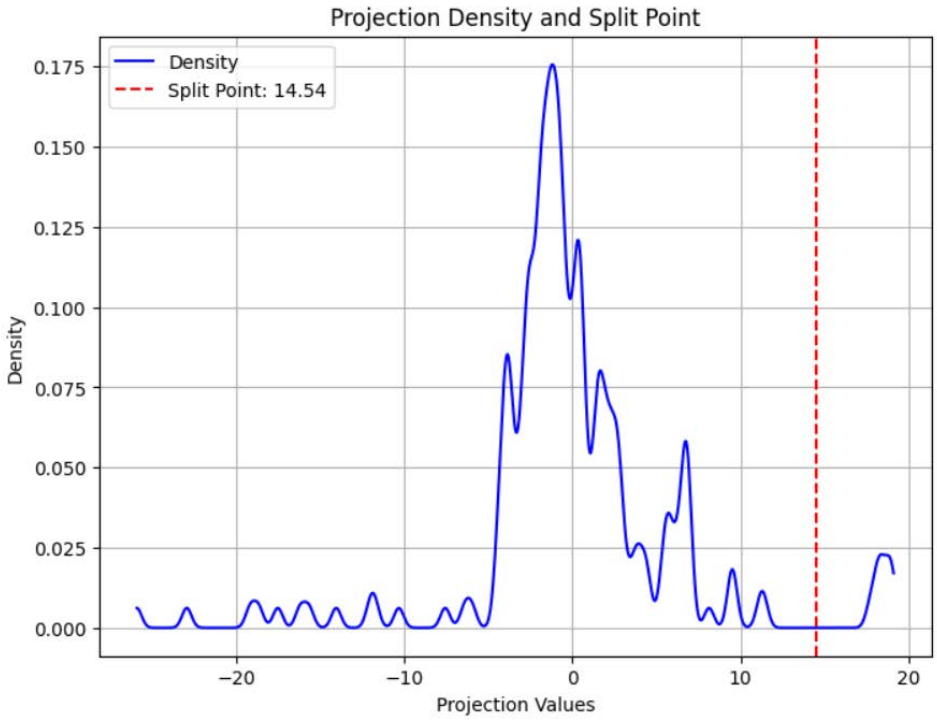
Ο Holes Index PPI (Δείκτης Κενών) αναζητά μια διαφορετική μορφή απόκλισης από την κανονική κατανομή: την ύπαρξη «κενών» ή «τρυπών» στην πυκνότητα, ιδίως στην κεντρική περιοχή της κατανομής. Στόχος του είναι να εντοπίσει προβολές όπου τα δεδομένα χωρίζονται σε δύο ομάδες, αφήνοντας μια περιοχή χαμηλής πυκνότητας ανάμεσά τους. Αυτό επιτυγχάνεται συγκρίνοντας την εκτιμώμενη πυκνότητα των δεδομένων με αυτή μιας ιδανικής κανονικής κατανομής που έχει τα ίδια στατιστικά χαρακτηριστικά. Μια υψηλή θετική τιμή στη μέση διαφορά τους υποδηλώνει ότι η πυκνότητα των δεδομένων είναι χαμηλότερη από την αναμενόμενη, κάτι που είναι ιδιαίτερα έντονο όταν υπάρχουν δύο διακριτές ομάδες εκατέρωθεν του μέσου όρου.
Ο τρόπος με τον οποίο υπολογίζεται ο παραπάνω δείκτης είναι:

Holes_Index = mean [pdf_normal(x) - pdf_kde(x)]

όπου:

- Holes_Index: Η βαθμολογία του δείκτη για την εκάστοτε προβολή.
- pdf_kde(x): Η τιμή της πυκνότητας ενός σημείου x των δεδομένων με βάση τον υπολογισμό KDE.

- pdf_normal(x): Η ιδανική γκαουσιανή τιμή πυκνότητας του σημείου x στα δεδομένα της προβολής.
- mean: Η μέση τιμή της διαφοράς των δυο παραπάνω τιμών.



Εικόνα 5. GSE45719 Dataset RLAC-Holes Index (Εύρος Ζώνης = 0.3, Πλήθος Προβολών = 50, Random_State = 44)

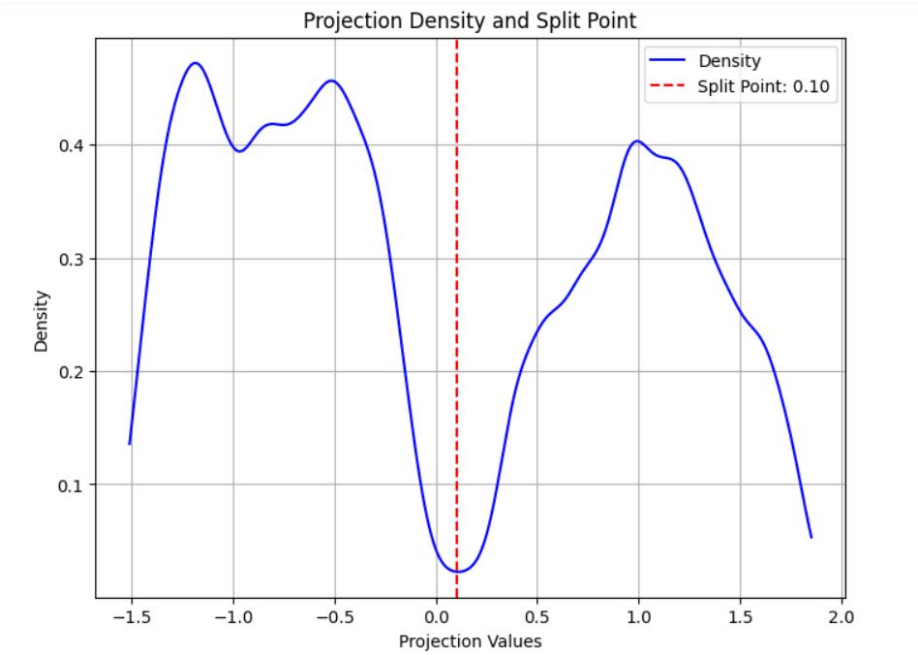
Τέλος, ο Kurtosis Index (Δείκτης Κύρτωσης) χρησιμοποιεί την τέταρτη στατιστική ροπή, η οποία μετρά το «βάρος» των ουρών μιας κατανομής σε σύγκριση με την κανονική. Αξιοποιείται με δύο τρόπους για να εντοπίσει δύο διαφορετικά είδη μη-κανονικότητας. Από τη μία πλευρά, η ελαχιστοποίηση της κύρτωσης (Min Kurtosis) αναζητά προβολές με την πιο αρνητική κύρτωση, καθώς μια κατανομή που είναι πιο «επίπεδη» από την κανονική, όπως μια διτροπική, τείνει να έχει αρνητική κύρτωση. Είναι ιδανικός για την εύρεση διτροπικών συστάδων. Από την άλλη πλευρά, η μεγιστοποίηση της κύρτωσης (Max Kurtosis) αναζητά προβολές με την πιο θετική κύρτωση, καθώς μια τέτοια κατανομή με «βαριές» ουρές συχνά υποδηλώνει την παρουσία ακραίων τιμών (outliers), οι οποίες μπορεί να αποτελούν μια ξεχωριστή μικρή συστάδα.

Ο μαθηματικός τύπος υπολογισμού του δείκτη της κύρτωσης είναι:

$$Kurt = \frac{\frac{1}{n}\sum_{i=1}^{n}[(x_i - \bar{x})^T v]^4}{\{\frac{1}{n}\sum_{i=1}^{n}[(x_i - \bar{x})^T v]^2\}^2}$$

όπου:

- Kurt: Η τιμή της Κύρτωσης για τη συγκεκριμένη προβολή.
- n: Ο συνολικός αριθμός των σημείων δεδομένων.
- x_i: Ένα μεμονωμένο σημείο δεδομένων (το οποίο είναι ένα διάνυσμα).
- $\bar{x}$: Το διάνυσμα του μέσου όρου όλων των σημείων δεδομένων.
- v: Το διάνυσμα της τυχαίας προβολής (η «τυχαία γραμμή»).
- $(x_i - \bar{x})^T v$: Η τιμή (συντεταγμένη) του σημείου i πάνω στην προβολή v.



Εικόνα 6. Aggregation Dataset RLAC-Min_Kurtosis (Εύρος ζώνης= 0.3, Πλήθος Προβολών= 50, Random_State=44)

2.3) Διερεύνηση Εναλλακτικών Δεικτών (Friedman-Tukey, Negentropy, Skewness, Fisher, Hermite)

Με σκοπό την περαιτέρω της συμπεριφοράς του αλγορίθμου RLAC και να καλυφθεί ένα ευρύτερο φάσμα στρατηγικών αναζήτησης προβολών, στην παρούσα εργασία υλοποιήθηκε και αξιολογήθηκε μια σειρά από επιπλέον κλασικούς και προχωρημένους δείκτες PPI. Αυτή η επέκταση αποσκοπούσε στην κατανόηση του

κατά πόσον η επιλογή του δείκτη επηρεάζει την ικανότητα του μοντέλου να ανακαλύπτει διαφορετικά είδη δομών.

Ένας από τους πιο θεμελιώδεις και ιστορικά σημαντικούς δείκτες που εξετάστηκαν είναι ο Friedman-Tukey Index. Ο δείκτης αυτός είναι ένας από τους θεμελιώδεις δείκτες στην αναζήτηση προβολών, σχεδιασμένος για να εντοπίζει μονοδιάστατες προβολές που παρουσιάζουν έναν επιθυμητό συνδυασμό συνολικής διασποράς και τοπικής συσσώρευσης. Η κεντρική ιδέα είναι ότι μια ενδιαφέρουσα προβολή για ομαδοποίηση δεν είναι ούτε απόλυτα ομοιόμορφη, ούτε υπερβολικά συμπαγής, αλλά μια προβολή όπου τα δεδομένα απλώνονται ικανοποιητικά, σχηματίζοντας ταυτόχρονα πυκνές τοπικές συγκεντρώσεις. Αυτή η δομή υποδηλώνει την πιθανή παρουσία διακριτών συστάδων που έχουν διαχωριστεί κατά μήκος της γραμμής προβολής.

Ο δείκτης, I(v), ορίζεται ως το γινόμενο δύο συνιστωσών. Η πρώτη συνιστώσα, s(v), είναι ένα μέτρο της συνολικής διασποράς και στην υλοποίησή μας υπολογίζεται ως η δειγματική τυπική απόκλιση των σημείων στην προβολή. Μια υψηλή τιμή για το s(v) υποδεικνύει ότι η προβολή καλύπτει ένα ευρύ φάσμα τιμών, παρέχοντας τον απαραίτητο "χώρο" για να αναδειχθούν οι διαφορετικές ομάδες.

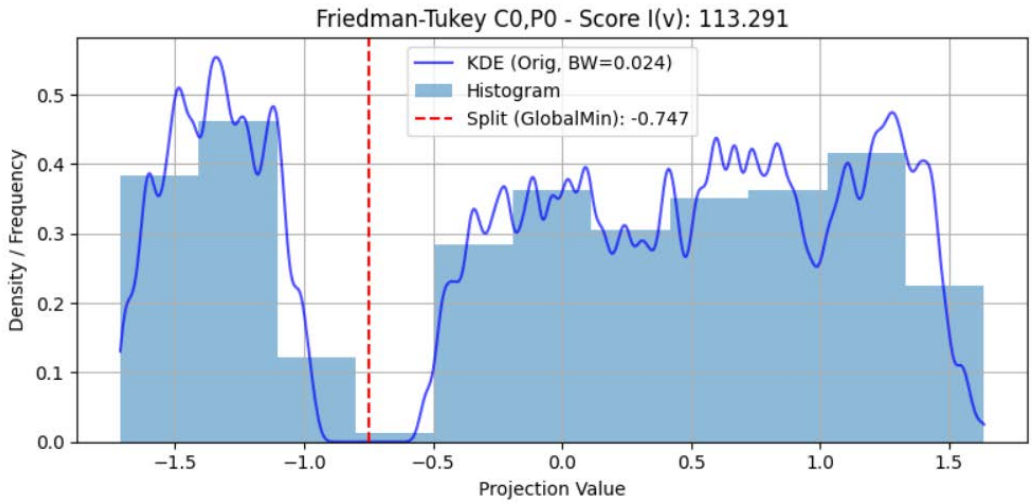
Η δεύτερη συνιστώσα, d(v), ποσοτικοποιεί την τοπική πυκνότητα ή την "συσσωρευσιμότητα" (clumpiness) των δεδομένων. Για τον υπολογισμό της, για κάθε σημείο της προβολής, βρίσκουμε τους k πλησιέστερους γείτονές του. Η τιμή του k προσδιορίζεται ευριστικά και προσαρμοστικά με βάση το πλήθος n των σημείων, ώστε να παραμένει σε λογικά πλαίσια. Στη συνέχεια, για κάθε σημείο, υπολογίζεται η μέση απόσταση από αυτούς τους k γείτονες. Το d(v) ορίζεται ως ο μέσος όρος των αντίστροφων αυτών των μέσων αποστάσεων (με την προσθήκη μιας μικρής σταθεράς για αριθμητική σταθερότητα). Έτσι, μια υψηλή τιμή για το d(v) σηματοδοτεί ότι τα σημεία, κατά μέσο όρο, βρίσκονται πολύ κοντά στους γείτονές τους, κάτι που είναι χαρακτηριστικό έντονης τοπικής συσσώρευσης.

Το τελικό σκορ, μεγιστοποιείται, ευνοώντας προβολές που είναι ταυτόχρονα απλωμένες και τοπικά πυκνές. Το σημείο διαίρεσης για μια τέτοια προβολή προσδιορίζεται, όπως και σε άλλους δείκτες, από το σημείο ολικού ελαχίστου της Εκτίμησης Πυρήνα Πυκνότητας (KDE). Ο μαθηματικός τύπος του δείκτη είναι ο εξής:

$$I(v) = s(v) * d(v)$$

όπου:

- I(v): Η βαθμολογία (score) του δείκτη Friedman-Tukey για την προβολή v.
- s(v): Το μέτρο της συνολικής διασποράς (spread) των προβαλλόμενων σημείων. Συνήθως, υλοποιείται ως η τυπική απόκλιση (standard deviation) των τιμών της προβολής.
- d(v): Το μέτρο της τοπικής πυκνότητας (clumping) των προβαλλόμενων σημείων. Υπολογίζεται ως ο μέσος όρος των αντίστροφων αποστάσεων κάθε σημείου από τους k-πλησιέστερους γείτονές του.



Εικόνα 7. Aggregation Dataset RLAC-Friedman Tukey (Εύρος Ζώνης= 0,1, Πλήθος Τυχαίων Προβολών=50, Random_State = 44)

Ο Δείκτης Αρνητικής Εντροπίας (Negentropy) είναι ένα μέτρο που προέρχεται από τη θεωρία πληροφοριών και χρησιμοποιείται για την ποσοτικοποίηση της απόκλισης μιας κατανομής από την Γκαουσιανή (κανονική) μορφή. Η κεντρική ιδέα είναι ότι η Γκαουσιανή κατανομή, για μια δεδομένη διακύμανση, είναι η πιο "τυχαία" ή "απρόβλεπτη" κατανομή, κατέχοντας τη μέγιστη δυνατή εντροπία. Συνεπώς, οποιαδήποτε απόκλιση από αυτήν, όπως η εμφάνιση διακριτών κορυφών (multimodality) ή έντονης ασυμμετρίας, υποδηλώνει την ύπαρξη "δομής" ή "πληροφορίας" και αντιστοιχεί σε χαμηλότερη εντροπία. Η Αρνητική Εντροπία J(Y) μιας τυχαίας μεταβλητής Y ορίζεται ως J(Y) = H(Y_gauss) - H(Y), όπου H(Y) είναι η διαφορική εντροπία της Y και H(Y_gauss) είναι η εντροπία μιας Γκαουσιανής μεταβλητής με την ίδια διακύμανση. Μια υψηλή τιμή Αρνητικής Εντροπίας

σηματοδοτεί μια έντονα μη-Γκαουσιανή προβολή, η οποία είναι επιθυμητή καθώς είναι πιθανό να αποκαλύπτει τη διαχωρισμένη δομή των συστάδων.

Η υλοποίηση του δείκτη ξεκινά με την τυποποίηση (standardization) των δεδομένων της μονοδιάστατης προβολής, ώστε να αποκτήσουν μέση τιμή μηδέν και μοναδιαία διακύμανση. Αυτό το βήμα εξασφαλίζει ότι η τιμή της Αρνητικής Εντροπίας που θα υπολογιστεί θα είναι ένα αμερόληπτο ως προς την κλίμακα μέτρο του "σχήματος" της κατανομής, επιτρέποντας τη δίκαιη σύγκριση μεταξύ διαφορετικών προβολών. Στη συνέχεια, για αυτή την τυποποιημένη προβολή, υπολογίζεται η διαφορική της εντροπία H(Y). Αυτό επιτυγχάνεται μέσω της Εκτίμησης Πυρήνα Πυκνότητας (KDE. Το ολοκλήρωμα H(Y) υπολογίζεται χρησιμοποιώντας τον κανόνα του Simpson για μεγαλύτερη ακρίβεια.

Το τελικό σκορ του δείκτη υπολογίζεται ως η διαφορά της σταθερής εντροπίας μιας πρότυπης Γκαουσιανής κατανομής (H(Y_gauss) ≈ 1.4189) μείον την υπολογισμένη εντροπία της προβολής (H(Y)). Το αποτέλεσμα διασφαλίζεται ότι είναι μη αρνητικό, καθώς θεωρητικά η Αρνητική Εντροπία δεν μπορεί να είναι αρνητική. Ενώ το σκορ υπολογίζεται στα τυποποιημένα δεδομένα, το σημείο διαίρεσης της προβολής προσδιορίζεται στην αρχική της κλίμακα. Αυτό πραγματοποιείται στα μη τυποποιημένα δεδομένα, βρίσκοντας το σημείο ολικού ελαχίστου της πυκνότητας (την "κοιλάδα"). Η προβολή με το υψηλότερο σκορ Αρνητικής Εντροπίας που συνοδεύεται από ένα έγκυρο σημείο διαίρεσης επιλέγεται από τον αλγόριθμο RLAC..

Πιο αναλυτικά ο τύπος με τον οποίο υπολογίζεται ο παραπάνω δείκτης είναι ο εξής:

$$J(Y) = H(Y_{gauss}) - H(Y)$$

όπου ο όρος H(Y) υπολογίζεται αριθμητικά:

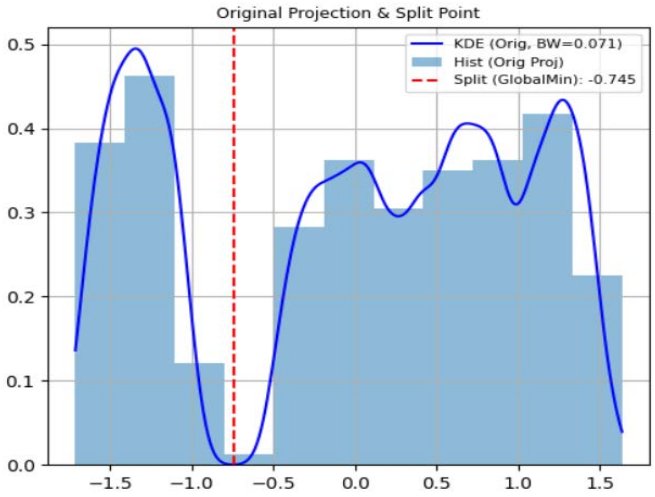
$$J(Y) = \left[\frac{1}{2}\log(2\pi e)\right] + \int p(y) * log(p(y))\, dy$$

και:

- J(Y): Είναι η Αρνητική Εντρόπια (Negentropy) της προβολής Y. Αποτελεί την τελική τιμή του δείκτη (PPI score), με υψηλότερες τιμές να υποδηλώνουν πιο «ενδιαφέρουσα», μη-Γκαουσιανή δομή.
- [1/2 * log(2 * π * e)]: Είναι μια σταθερά που αντιστοιχεί στη διαφορική εντροπία H(Y_gauss) μιας τυπικής κανονικής (Gaussian) κατανομής με μέση τιμή 0 και διακύμανση 1. Αποτελεί το θεωρητικό μέγιστο της εντροπίας για

οποιαδήποτε κατανομή με μοναδιαία διακύμανση και χρησιμεύει ως σημείο αναφοράς για τη «μέγιστη αταξία».

- ∫ p(y) * log(p(y)) dy: Είναι το αρνητικό της διαφορικής εντροπίας H(Y) της προβολής. Στην υλοποίηση, αυτό το ολοκλήρωμα δεν λύνεται αναλυτικά, αλλά υπολογίζεται αριθμητικά χρησιμοποιώντας την εκτιμώμενη πυκνότητα p(y) και μεθόδους όπως ο κανόνας του Simpson.
- Y: Είναι η κανονικοποιημένη μονοδιάστατη προβολή των δεδομένων, δηλαδή η προβολή αφού έχει μετασχηματιστεί ώστε να έχει μέση τιμή 0 και διακύμανση 1. Η κανονικοποίηση αυτή είναι απαραίτητη για τη μαθηματική ορθότητα του τύπου.
- p(y): Είναι η Συνάρτηση Πυκνότητας Πιθανότητας (Probability Density Function - PDF) της κανονικοποιημένης προβολής Y. Στην πράξη, επειδή η πραγματική κατανομή είναι άγνωστη, η p(y) εκτιμάται από τα δεδομένα μέσω της μεθόδου Εκτίμησης Πυκνότητας Πυρήνα (Kernel Density Estimation - KDE).



Εικόνα 8. Aggregation Dataset RLAC-NegEntropy (Εύρος Ζώνης= 0,3, Πλήθος Τυχαίων Προβολών=50, Random_State = 44)

Ο Δείκτης Ασυμμετρίας (Skewness Index) εστιάζει στον εντοπισμό προβολών που παραβιάζουν μια βασική ιδιότητα της Γκαουσιανής κατανομής: την τέλεια συμμετρία της γύρω από τη μέση τιμή. Μια προβολή που παράγει μια έντονα ασύμμετρη κατανομή θεωρείται "ενδιαφέρουσα", διότι αυτό συχνά υποδηλώνει την ύπαρξη κάποιας υποκείμενης δομής. Για παράδειγμα, μια τέτοια ασυμμετρία μπορεί να

προκύψει όταν μια μικρή αλλά πυκνή συστάδα διαχωρίζεται από μια μεγαλύτερη, δημιουργώντας μια μακριά "ουρά" προς τη μία κατεύθυνση, ή όταν δύο συστάδες έχουν σημαντικά διαφορετικές πυκνότητες.

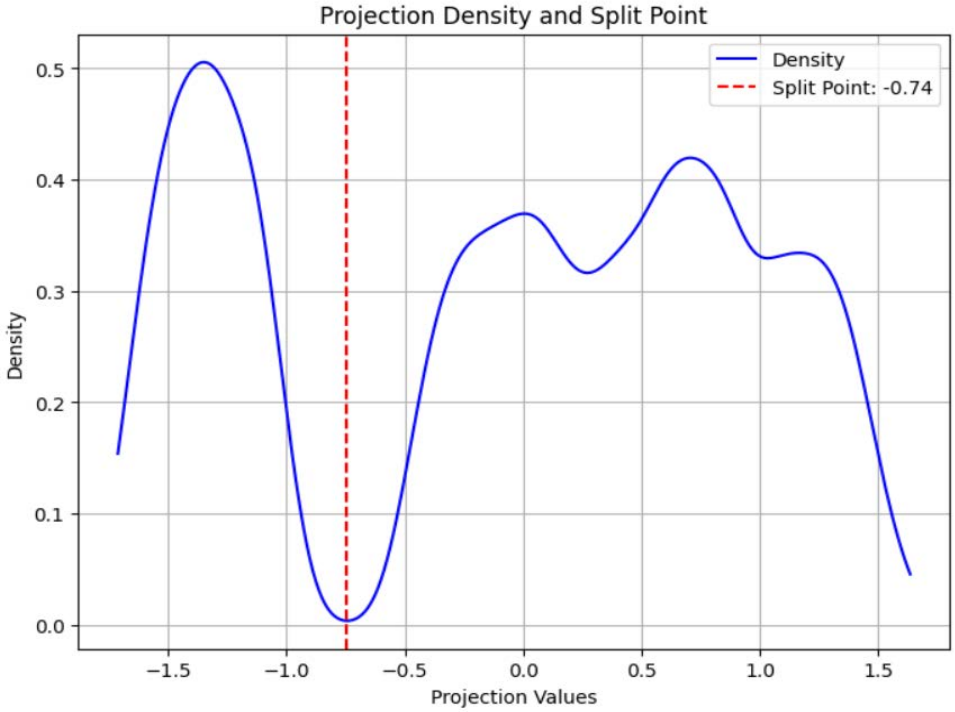
Ανεξαρτήτως της κατεύθυνσης της ασυμμετρίας (αν η "ουρά" είναι προς τα δεξιά ή προς τα αριστερά), ο δείκτης στην παρούσα υλοποίηση ορίζεται ως η απόλυτη τιμή της υπολογισμένης ασυμμετρίας. Με τον τρόπο αυτό, ο αλγόριθμος επιβραβεύει οποιαδήποτε σημαντική απόκλιση από τη συμμετρία, μεγιστοποιώντας την πιθανότητα εντοπισμού προβολών που διαχωρίζουν άνισες σε μέγεθος ή πυκνότητα συστάδες.

$$Skewness\_Score = \left| \frac{\sqrt{\left(n * (n-1)\right)}}{(n-2)} * \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \mu)^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \mu)^2\right]^{\frac{3}{2}}} \right|$$

όπου:

- y_i : Αντιπροσωπεύει την τιμή του κάθε μεμονωμένου σημείου δεδομένων μέσα στη μονοδιάστατη projection (προβολή).
- n: Είναι το συνολικό πλήθος των σημείων δεδομένων στην προβολή (δηλαδή, το len(projection)).
- μ : Είναι η μέση τιμή (ο μέσος όρος) όλων των τιμών y_i στην προβολή.
- Σ (σίγμα): Είναι το σύμβολο της άθροισης. Σημαίνει ότι εκτελούμε τον υπολογισμό που ακολουθεί (π.χ., (y_i - μ)^3) για κάθε σημείο i από 1 έως n και στη συνέχεια προσθέτουμε όλα τα αποτελέσματα.
- Σ(y_i - μ)^3: Αυτό το τμήμα μετρά το άθροισμα των κύβων των αποκλίσεων κάθε σημείου από τη μέση τιμή. Είναι ο πυρήνας της μέτρησης της ασυμμετρίας, καθώς διατηρεί το πρόσημο των αποκλίσεων.
- Σ(y_i - μ)^2: Αυτό είναι το άθροισμα των τετραγώνων των αποκλίσεων, το οποίο σχετίζεται άμεσα με τη διακύμανση (variance) των δεδομένων.
- (√(n * (n - 1)) / (n - 2)): Αυτός είναι ένας διορθωτικός παράγοντας που εφαρμόζεται για να γίνει η εκτίμηση της ασυμμετρίας "αμερόληπτη" (unbiased), δηλαδή πιο ακριβής, ειδικά όταν έχουμε μικρό αριθμό δειγμάτων n.
- | ... |: Το σύμβολο της απόλυτης τιμής. Το χρησιμοποιούμε στο τέλος επειδή για την εύρεση μιας "ενδιαφέρουσας" προβολής, μας ενδιαφέρει το μέγεθος

της ασυμμετρίας και όχι η κατεύθυνσή της (αν είναι λοξή προς τα δεξιά ή προς τα αριστερά).



Εικόνα 9. Aggregation Dataset RLAC-Skewness (Εύρος Ζώνης= 0,4, Πλήθος Τυχαίων Προβολών=50, Random_State = 44)

Μια ιδιαίτερη και ισχυρή προσέγγιση που υλοποιήθηκε είναι ο Προσαρμοσμένος Διακριτικός Δείκτης του Fisher. Αυτός ο δείκτης αξιοποιεί την κεντρική ιδέα του κλασικού κριτηρίου του Fisher, το οποίο στην αρχική του μορφή είναι ένα εργαλείο επιβλεπόμενης μάθησης που απαιτεί προκαθορισμένες ετικέτες κλάσεων. Για την εφαρμογή του στο μη-επιβλεπόμενο πλαίσιο του RLAC, ο δείκτης προσαρμόστηκε ώστε να αξιολογεί την ποιότητα ενός πιθανού διαχωρισμού, ο οποίος προτείνεται από τη δομή των ίδιων των δεδομένων.

Η διαδικασία για κάθε υποψήφια προβολή ξεκινά με τον προσδιορισμό ενός σημείου διαίρεσης. Όπως και σε άλλους δείκτες, αυτό το σημείο (split_point_final) εντοπίζεται ως το ολικό ελάχιστο της καμπύλης πυκνότητας (KDE), η οποία υπολογίζεται πάνω στα αρχικά, μη τυποποιημένα δεδομένα της προβολής. Αυτό το σημείο, που αντιπροσωπεύει τη βαθύτερη "κοιλάδα" μεταξύ περιοχών υψηλότερης πυκνότητας, αποτελεί τον προτεινόμενο "φυσικό" διαχωρισμό για τη συγκεκριμένη προβολή, υπό την προϋπόθεση ότι είναι αυστηρά εσωτερικό στα δεδομένα.

Στη συνέχεια, για να υπολογιστεί ένα σκορ που θα είναι συγκρίσιμο μεταξύ διαφορετικών προβολών με διαφορετικές κλίμακες, ολόκληρη η προβολή τυποποιείται ώστε να έχει μέση τιμή μηδέν και διακύμανση ένα (projection_std). Το σημείο διαίρεσης που βρέθηκε στην αρχική κλίμακα μετατρέπεται και αυτό στην τυποποιημένη κλίμακα (split_point_std). Χρησιμοποιώντας αυτό το τυποποιημένο σημείο, τα τυποποιημένα δεδομένα χωρίζονται σε δύο προσωρινές "ψευδο-ομάδες", την αριστερή και τη δεξιά.

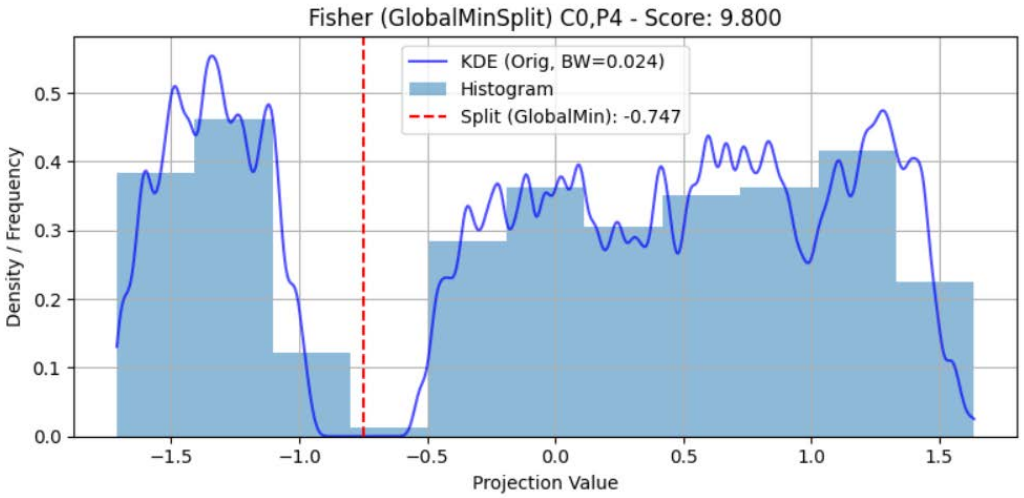
Σε αυτό το σημείο εφαρμόζεται το κριτήριο του Fisher. Για τις δύο αυτές τυποποιημένες ψευδο-ομάδες, υπολογίζονται οι μέσες τιμές τους και οι δειγματικές διακυμάνσεις τους. Ουσιαστικά, ο δείκτης επιβραβεύει τις προβολές όπου η κοιλάδα της KDE οδηγεί σε έναν διαχωρισμό που παράγει δύο ομάδες οι οποίες είναι ταυτόχρονα όσο το δυνατόν πιο απομακρυσμένες μεταξύ τους (μεγάλη διαφορά μέσων τιμών) και όσο το δυνατόν πιο συμπαγείς εσωτερικά (μικρές διακυμάνσεις). Η προβολή που επιτυγχάνει το υψηλότερο σκορ σε αυτόν τον δείκτη θεωρείται η καλύτερη για τη διαίρεση των δεδομένων σε εκείνη την επανάληψη του αλγορίθμου.

$$Fisher\_Score = \frac{(\text{mean\_left} - \text{mean\_right})^2}{var\_left^2 + var\_right^2 + e}$$

όπου:

- mean_left: Είναι η μέση τιμή της "αριστερής" ψευδο-ομάδας. Αυτή η ομάδα περιλαμβάνει όλα τα (κανονικοποιημένα) σημεία της προβολής που βρίσκονται αριστερά ή πάνω στο σημείο διαίρεσης.
- mean_right: Είναι η μέση τιμή της "δεξιάς" ψευδο-ομάδας (group_right_std). Αυτή η ομάδα περιλαμβάνει όλα τα (κανονικοποιημένα) σημεία της προβολής που βρίσκονται αυστηρά δεξιά του σημείου διαίρεσης.
- (mean_left - mean_right)^2: Αυτός είναι ο αριθμητής του κλάσματος. Είναι το τετράγωνο της απόστασης μεταξύ των μέσων τιμών των δύο ψευδο-ομάδων. Μια μεγάλη τιμή εδώ σημαίνει ότι οι δύο ομάδες είναι καλά διαχωρισμένες και τα κέντρα τους απέχουν πολύ μεταξύ τους πάνω στη γραμμή προβολής. Αυτό είναι το "σήμα" που θέλουμε να μεγιστοποιήσουμε.
- var_left: Είναι η δειγματική διακύμανση (sample variance) της "αριστερής" ψευδο-ομάδας. Μετρά πόσο "απλωμένα" είναι τα σημεία μέσα στην αριστερή ομάδα, γύρω από τη μέση τιμή τους (var_left).

- var_right: Είναι η δειγματική διακύμανση της "δεξιάς" ψευδο-ομάδας. Μετρά πόσο "απλωμένα" είναι τα σημεία μέσα στη δεξιά ομάδα.
- ((var_left)^2+(var_right)^2): Αυτό είναι το άθροισμα των ενδο-ομαδικών διακυμάνσεων. Αντιπροσωπεύει τη συνολική "σύγχυση" ή "θόρυβο" μέσα στις δύο ομάδες. Μια μικρή τιμή εδώ σημαίνει ότι και οι δύο ομάδες είναι πολύ συμπαγείς και τα σημεία τους είναι μαζεμένα γύρω από τις μέσες τιμές τους. Αυτό είναι το "κόστος" που θέλουμε να ελαχιστοποιήσουμε.
- + e : Είναι ένας πολύ μικρός αριθμός (στον κώδικά είναι 1e-10) που προστίθεται στον παρονομαστή για να αποφευχθεί η διαίρεση με το μηδέν, στην περίπτωση που και οι δύο ομάδες έχουν μηδενική διακύμανση.



Εικόνα 10. Aggregation Dataset RLAC-Fisher (Εύρος Ζώνης= 0,1, Πλήθος Τυχαίων Προβολών=50, Random_State = 44)

Τέλος, διερευνήθηκε ο Hermite Polynomial Index, που βασίζεται στα πολυώνυμα Hermite προσφέρει έναν τρόπο για την ποσοτικοποίηση της απόκλισης μιας προβολής από την Γκαουσιανή (κανονική) κατανομή, θεωρώντας ότι οποιαδήποτε τέτοια απόκλιση μπορεί να υποδηλώνει την ύπαρξη "ενδιαφέρουσας" δομής, όπως συστάδες ή ασυμμετρία. Τα πολυώνυμα Hermite είναι κατάλληλα για αυτόν τον σκοπό, καθώς αποτελούν μια μαθηματική βάση που μπορεί να περιγράψει αποτελεσματικά τις αποκλίσεις από την κανονικότητα, ειδικά αυτές που σχετίζονται με τις στατιστικές ροπές υψηλότερης τάξης.

Η διαδικασία για τον υπολογισμό του δείκτη ξεκινά με την τυποποίηση (standardization) των δεδομένων της μονοδιάστατης προβολής, ώστε να αποκτήσουν

μέση τιμή μηδέν και μοναδιαία διακύμανση. Αυτό το βήμα είναι απαραίτητο για τη διασφάλιση ότι το σκορ του δείκτη θα είναι αμετάβλητο ως προς την κλίμακα, επιτρέποντας τη δίκαιη σύγκριση μεταξύ προβολών με διαφορετικά εύρη τιμών. Στη συνέχεια, εκτιμώνται οι συντελεστές c_j του αναπτύγματος της κατανομής σε σειρά πολυωνύμων Hermite. Ο κάθε συντελεστής c_j εκτιμάται ως η μέση τιμή του j-οστού πολυωνύμου Hermite (He_j(Y)) υπολογισμένου πάνω στα τυποποιημένα δεδομένα Y της προβολής. Οι συντελεστές αυτοί έχουν άμεση στατιστική ερμηνεία: ο συντελεστής τρίτης τάξης (c_3) σχετίζεται με την ασυμμετρία (skewness) της κατανομής, ενώ ο συντελεστής τέταρτης τάξης (c_4) με την κύρτωση Fisher (excess kurtosis).

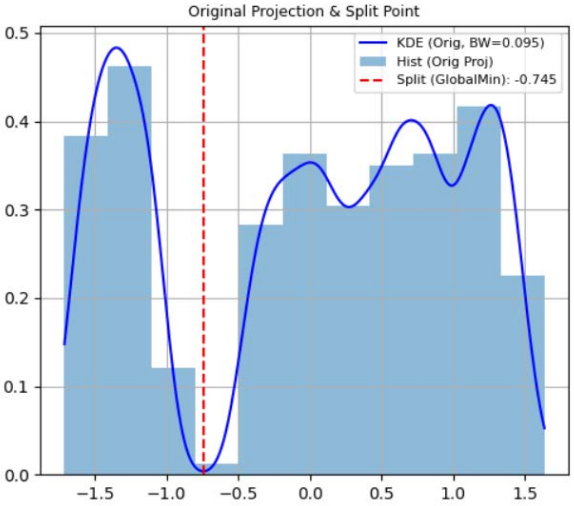
Ως τελικό σκορ για τον δείκτη προβολής, χρησιμοποιείται το άθροισμα των τετραγώνων των επιλεγμένων συντελεστών. Στην υλοποίηση, εστιάζουμε στους συντελεστές τρίτης και τέταρτης τάξης, οπότε το σκορ υπολογίζεται ως Score = c_3^2 + c_4^2. Μια μεγάλη τιμή για αυτό το άθροισμα υποδηλώνει σημαντική απόκλιση από την κανονικότητα είτε ως προς την ασυμμετρία είτε ως προς την κύρτωση (ή και τα δύο), καθιστώντας την προβολή "ενδιαφέρουσα". Το σημείο διαίρεσης για μια τέτοια προβολή προσδιορίζεται, όπως και με άλλους δείκτες, από το σημείο ολικού ελαχίστου της Εκτίμησης Πυρήνα Πυκνότητας (KDE) στην προβολή. Ο τύπος με τον οποίο υπολογίζεται ο δείκτης αυτός είναι ο εξής:

$$\text{Hermite\_Score} = (E[H_{e_3}(Y)])^2 + (E[H_{e_4}(Y)])^2$$

όπου:

- Y: Αντιπροσωπεύει τα δεδομένα της προβολής αφού έχουν τυποποιηθεί (standardized.
- He_3(Y): Είναι το τρίτης τάξης πολυώνυμο Hermite (y^3 - 3y) υπολογισμένο για κάθε σημείο y της τυποποιημένης προβολής Y.
- He_4(Y): Είναι το τέταρτης τάξης πολυώνυμο Hermite (y^4 - $6y^2$ + 3) υπολογισμένο για κάθε σημείο y της τυποποιημένης προβολής Y.
- E[...]: Συμβολίζει την αναμενόμενη τιμή, η οποία στην πράξη εκτιμάται ως η μέση τιμή των τιμών μέσα στην παρένθεση.
- E[He_3(Y)] είναι ο συντελεστής c_3 και μετρά την ασυμμετρία (skewness) της προβολής.

- E[He_4(Y)] είναι ο συντελεστής c_4 και μετρά την κύρτωση Fisher (excess kurtosis) της προβολής.



Εικόνα 11. Aggregation Dataset RLAC-Hermite (Εύρος Ζώνης= 0,4, Πλήθος Τυχαίων Προβολών=50, Random_State = 44)

2.4) Αλγόριθμοι Συσταδοποίησης προς Σύγκριση

Για την αντικειμενική αξιολόγηση της απόδοσης του αλγορίθμου RLAC και των παραλλαγών του, είναι απαραίτητη η σύγκρισή τους με καθιερωμένες και ευρέως αποδεκτές μεθόδους συσταδοποίησης.

Ο KMeans είναι αναμφισβήτητα ο πιο γνωστός και ευρέως χρησιμοποιούμενος αλγόριθμος συσταδοποίησης. Ανήκει στην κατηγορία των αλγορίθμων που βασίζονται σε κεντροειδή (centroid-based) και λειτουργεί με μια απλή, επαναληπτική διαδικασία. Αρχικά, ορίζονται τυχαία k σημεία ως αρχικά κέντρα των συστάδων. Στη συνέχεια, σε κάθε επανάληψη, κάθε σημείο δεδομένων ανατίθεται στην πλησιέστερη συστάδα (βήμα ανάθεσης) και ακολούθως, τα κέντρα των συστάδων επαναϋπολογίζονται ως ο μέσος όρος των σημείων που ανήκουν σε αυτές (βήμα ενημέρωσης). Η διαδικασία τερματίζει όταν οι αναθέσεις των σημείων σταθεροποιηθούν. Η ισχύς του KMeans έγκειται στην ταχύτητα και την απλότητά του, αλλά η απόδοσή του περιορίζεται από τη θεμελιώδη υπόθεσή του ότι οι συστάδες είναι σφαιρικές, ισομεγέθεις και καλά διαχωρισμένες, καθιστώντας τον αναποτελεσματικό σε δεδομένα με πολύπλοκα, μη κυρτά σχήματα.

Σε αντίθεση με την «επίπεδη» προσέγγιση του KMeans, ο Agglomerative Hierarchical Clustering (HClust) ακολουθεί μια ιεραρχική στρατηγική. Η διαδικασία ξεκινά θεωρώντας κάθε σημείο δεδομένων ως μια ξεχωριστή συστάδα. Σε κάθε επόμενο βήμα, οι δύο πλησιέστερες συστάδες συγχωνεύονται σε μία, μειώνοντας τον συνολικό αριθμό των συστάδων κατά ένα. Η διαδικασία συνεχίζεται μέχρι όλα τα σημεία να ανήκουν σε μία μόνο συστάδα. Ο τρόπος με τον οποίο ορίζεται η «απόσταση» μεταξύ δύο συστάδων καθορίζεται από το κριτήριο σύνδεσης (linkage criterion). Στην παρούσα εργασία, χρησιμοποιήθηκε το κριτήριο της απλής σύνδεσης (single linkage), όπου η απόσταση μεταξύ δύο συστάδων ορίζεται ως η απόσταση μεταξύ των δύο πλησιέστερων σημείων τους. Αυτό το κριτήριο επιτρέπει στον αλγόριθμο να ανιχνεύει επιμήκεις και μη κυρτές δομές, αλλά τον καθιστά ευαίσθητο στον θόρυβο.

Μια πιο προηγμένη προσέγγιση που μπορεί να διαχειριστεί πολύπλοκες γεωμετρίες είναι ο Spectral Clustering (NCut). Ο αλγόριθμος αυτός δεν λειτουργεί απευθείας στον αρχικό χώρο των χαρακτηριστικών, αλλά μετασχηματίζει το πρόβλημα σε ένα πρόβλημα διαμέρισης γράφου. Αρχικά, κατασκευάζει έναν γράφο ομοιότητας, όπου οι κόμβοι είναι τα σημεία δεδομένων και οι ακμές (με βάρη) αναπαριστούν την ομοιότητα μεταξύ τους. Στη συνέχεια, χρησιμοποιώντας τις ιδιότητες του Λαπλασιανού πίνακα του γράφου, προβάλλει τα δεδομένα σε έναν νέο χώρο χαμηλότερης διάστασης, όπου οι συστάδες είναι πλέον γραμμικά διαχωρίσιμες. Τέλος, σε αυτόν τον νέο χώρο εφαρμόζεται ένας απλός αλγόριθμος, όπως ο KMeans, για τον τελικό διαχωρισμό. Η ιδιότητά του αυτή το καθιστά ιδιαίτερα ισχυρό σε σύνολα δεδομένων με μη κυρτές δομές, όπως οι ομόκεντροι δακτύλιοι, παρέχοντας ένα ισχυρό μέτρο σύγκρισης για τις επιδόσεις του RLAC σε τέτοιες προκλήσεις.

Τέλος, εξετάστηκε και ο αλγόριθμος Minimum Density Hyperplanes (Mdh), ο οποίος είναι εννοιολογικά ο πιο κοντινός στον RLAC. Όπως και ο RLAC, ο Mdh είναι ένας διαιρετικός αλγόριθμος που βασίζεται στην αναζήτηση προβολών. Ωστόσο, αντί να βασίζεται σε τυχαίες προβολές, ο Mdh εκτελεί μια διαδικασία βελτιστοποίησης για να βρει το υπερεπίπεδο (π.χ., μια γραμμή σε 2D) που διέρχεται από την περιοχή με τη χαμηλότερη πυκνότητα δεδομένων, ελαχιστοποιώντας το ολοκλήρωμα της πυκνότητας κατά μήκος του. Ουσιαστικά, αναζητά ενεργά την καλύτερη γραμμική τομή που διαχωρίζει δύο συστάδες. Επειδή η φιλοσοφία του είναι παρόμοια με αυτή του RLAC αλλά η στρατηγική αναζήτησής του είναι ντετερμινιστική και όχι τυχαία, αποτελεί ένα εξαιρετικά σημαντικό σημείο αναφοράς για την αξιολόγηση της

αποτελεσματικότητας της τυχαίας προσέγγισης. Ο αλγόριθμος αυτός υλοποιήθηκε με βάση το άρθρο, σε γλώσσα python στο πλαίσιο της πτυχιακής εργασίας, με σκοπό την σύγκριση της απόδοσης του με αυτή του RLAC μοντέλου.

3. Υλοποίηση του RLAC

Το παρόν κεφάλαιο περιγράφει τη διαδικασία της υλοποίησης, της διάγνωσης των προβλημάτων που προέκυψαν και, βελτίωση τους. Ένας επαναληπτικός κύκλος πειραματισμού και βελτίωσης, ο οποίος οδήγησε σε μια βαθύτερη κατανόηση της συμπεριφοράς του μοντέλου και στη δημιουργία ενός ουσιαστικά πιο αποτελεσματικού πλαισίου.

3.1) Αρχική Υλοποίηση και Προσδιορισμός Προκλήσεων

Το πρώτο βήμα της εργασίας ήταν η πιστή υλοποίηση του αλγορίθμου RLAC σε γλώσσα προγραμματισμού Python, ακολουθώντας τη μεθοδολογία που περιγράφεται στο άρθρο-αναφοράς. Η αρχική αυτή υλοποίηση ενσωμάτωνε τους τέσσερις βασικούς δείκτες PPI (Depth Ratio, Dip-Test, Holes, Min/Max Kurtosis) και χρησιμοποιούσε τυχαίες προβολές που παράγονταν από μια Γκαουσιανή κατανομή. Σκοπός ήταν η δημιουργία μιας βάσης για την εμπειρική αξιολόγηση του μοντέλου και η επαλήθευση της απόδοσής του έναντι των καθιερωμένων αλγορίθμων (KMeans, HClust, NCut, Mdh).

Κατά την πειραματική αξιολόγηση αυτής της αρχικής έκδοσης, αναδύθηκαν δύο θεμελιώδεις προκλήσεις. Πρώτον, ενώ το μοντέλο απέδιδε ικανοποιητικά σε ορισμένα απλά συνθετικά σύνολα δεδομένων, η απόδοσή του σε πιο πολύπλοκα σύνολα πραγματικού κόσμου και στα απαιτητικά μονοκυτταρικά δεδομένα ήταν απογοητευτική, με τις μετρικές αξιολόγησης να κυμαίνονται κοντά στο μηδέν κάτι που υποδείκνυε αποτελέσματα ισοδύναμα με μια τυχαία ομαδοποίηση. Δεύτερον, και ίσως το πιο κρίσιμο πρόβλημα, το μοντέλο παρουσίαζε έντονη έλλειψη αναπαραγωγιμότητας (reproducibility). Επαναλαμβανόμενες εκτελέσεις του αλγορίθμου στα ίδια δεδομένα με τις ίδιες παραμέτρους παρήγαγαν διαφορετικά, συχνά αντιφατικά, αποτελέσματα. Αυτή η αστάθεια καθιστούσε αδύνατη την εξαγωγή ασφαλών συμπερασμάτων και αποτελούσε σοβαρό εμπόδιο για τη συστηματική αξιολόγηση του αλγορίθμου.

3.2) Βελτίωση της υλοποίησης του RLAC

Έχοντας εντοπίσει τις παραπάνω αδυναμίες, η έρευνα επικεντρώθηκε στην επίλυση τόσο του προβλήματος της χαμηλής απόδοσης όσο και αυτού της αστάθειας. Η διαδικασία αυτή βασίστηκε σε τρεις στρατηγικούς άξονες.

Η πρώτη βελτίωση αφορούσε τον μηχανισμό δημιουργίας των τυχαίων προβολών. Η αρχική μέθοδος, που βασιζόταν σε δειγματοληψία από Γκαουσιανή κατανομή, αντικαταστάθηκε από την υπολογιστικά αποδοτικότερη προσέγγιση του Achlioptas. Η μέθοδος αυτή παράγει "αραιούς" (sparse) πίνακες προβολών, των οποίων τα στοιχεία λαμβάνουν μόνο τρεις τιμές (-1, 0, +1) με συγκεκριμένες πιθανότητες. Αυτή η αλλαγή οδήγησε σε σημαντική επιτάχυνση των υπολογισμών, ειδικά σε δεδομένα με μεγάλο αριθμό χαρακτηριστικών, χωρίς να επηρεάζει τις θεωρητικές εγγυήσεις του λήμματος Johnson-Lindenstrauss.

Η πιο σημαντική ανακάλυψη, ωστόσο, ήταν ο καθοριστικός ρόλος της παραμέτρου του εύρους ζώνης (bandwidth) στην Εκτίμηση Πυκνότητας Πυρήνα (KDE). Οι αρχικές ευριστικές μέθοδοι, όπως ο κανόνας του Silverman, έτειναν να παράγουν υπερβολικά μεγάλες τιμές, οδηγώντας σε «υπερ-εξομαλυμένες» (over-smoothed) καμπύλες πυκνότητας. Αυτές οι λείες καμπύλες απέκρυπταν τις λεπτές δομές των δεδομένων, «σβήνοντας» τις κοιλάδες ανάμεσα στις συστάδες και εμποδίζοντας τον αλγόριθμο να εντοπίσει σημεία διαχωρισμού. Μέσω εκτεταμένων πειραματισμών, διαπιστώθηκε ότι η προσεκτική ρύθμιση αυτής της παραμέτρου, και συγκεκριμένα η χρήση μικρότερων τιμών εύρους ζώνης σε πραγματικά σύνολα δεδομένων, ήταν θεμελιώδης για την ικανότητα του μοντέλου να ανακαλύπτει δομές.

Τέλος, για την αντιμετώπιση του κρίσιμου ζητήματος της έλλειψης αναπαραγωγιμότητας, εισήχθη η παράμετρος random_state. Αυτή η παράμετρος λειτουργεί ως «σπόρος» (seed) για τη γεννήτρια τυχαίων αριθμών που δημιουργεί τον πίνακα προβολών. Ορίζοντας μια συγκεκριμένη τιμή για το random_state, διασφαλίζεται ότι ο ίδιος ακριβώς πίνακας προβολών παράγεται σε κάθε εκτέλεση.

3.3) Αρχιτεκτονική και Ροή

Το τελικό, μοντέλο RLAC που προέκυψε από την παραπάνω διαδικασία, ακολουθεί μια ορθή και αποτελεσματική ροή λειτουργίας.

Η διαδικασία ξεκινά με την προετοιμασία των δεδομένων και τον καθορισμό των βασικών παραμέτρων από τον χρήστη. Αυτές περιλαμβάνουν τον επιθυμητό αριθμό συστάδων (k), τον αριθμό των τυχαίων προβολών (r), τον συντελεστή προσαρμογής του εύρους ζώνης για την KDE, και τον κρίσιμο "σπόρο τυχαιότητας" (random_state) για την εξασφάλιση της αναπαραγωγιμότητας. Αξιοποιώντας το θεώρημα Johnson-Lindenstrauss, εάν ο χρήστης δεν ορίσει τον αριθμό των προβολών r, ο αλγόριθμος τον υπολογίζει αυτόματα βάσει του πλήθους των δειγμάτων, παρέχοντας μια θεωρητικά θεμελιωμένη προεπιλογή. Ο ελάχιστος απαιτούμενος αριθμός διαστάσεων r συνδέεται με τον αριθμό των δειγμάτων n και μια παράμετρο παραμόρφωσης ε (epsilon), η οποία καθορίζει το επιτρεπτό σφάλμα στη διατήρηση των αποστάσεων. Στην υλοποίησή, χρησιμοποιείται η συνάρτηση johnson_lindenstrauss_min_dim από τη βιβλιοθήκη scikit-learn, με μια προεπιλεγμένη τιμή ε = 0.1. Αυτή η προσέγγιση επιτρέπει στον αλγόριθμο να προσαρμόζει αυτόματα τον αριθμό των προβολών ανάλογα με το μέγεθος του συνόλου δεδομένων, παρέχοντας μια ισορροπημένη λύση μεταξύ υπολογιστικής απόδοσης και της ικανότητας διατήρησης της αρχικής γεωμετρικής δομής.

Μια σημαντική βελτίωση σε αυτό το στάδιο είναι η υπολογιστικά αποδοτική προβολή των δεδομένων. Αντί να επαναλαμβάνεται η δαπανηρή διαδικασία της προβολής των δεδομένων για κάθε συστάδα σε κάθε επανάληψη, ο αλγόριθμος αρχικά δημιουργεί τον αραιό πίνακα προβολής Achlioptas και προβάλλει ολόκληρο το σύνολο δεδομένων μια μόνο φορά. Το αποτέλεσμα, ένας πίνακας με τις μονοδιάστατες συντεταγμένες όλων των σημείων για όλες τις r προβολές, αποθηκεύεται στη μνήμη. Αυτή η στρατηγική μειώνει δραματικά τον υπολογιστικό φόρτο, καθιστώντας τον αλγόριθμο κατάλληλο για την ανάλυση πολύ μεγαλύτερων συνόλων δεδομένων.

Στη συνέχεια, ο αλγόριθμος εισέρχεται στον κεντρικό, ιεραρχικό διαιρετικό βρόχο, ο οποίος εκτελείται μέχρι να επιτευχθεί ο επιθυμητός αριθμός συστάδων. Για κάθε υπάρχουσα συστάδα που πληροί ένα ελάχιστο όριο μεγέθους προσαρμοσμένο στο μέγεθος του συνόλου δεδομένων, ο αλγόριθμος αναζητά την απολύτως καλύτερη δυνατή προβολή για τον διαχωρισμό της. Αυτό γίνεται αξιολογώντας και τις r προβολές όπου για κάθε μία από αυτές τις προβολές, υπολογίζεται η καμπύλη πυκνότητας KDE με το προσαρμοσμένο εύρος ζώνης και, στη συνέχεια, ο επιλεγμένος δείκτης PPI βαθμολογεί την "ενδιαφερότητά" της.

Αφού αξιολογηθούν όλες οι προβολές για τη συγκεκριμένη συστάδα, αποθηκεύεται η πληροφορία της καλύτερης προβολής για τη συστάδα αυτή (δηλαδή, η προβολή με το

υψηλότερο σκορ PPI και το αντίστοιχο σημείο διαχωρισμού). Αφού ολοκληρωθεί αυτή η διαδικασία για όλες τις υποψήφιες συστάδες, ο αλγόριθμος συγκρίνει τις "πρωταθλήτριες" προβολές μεταξύ τους και επιλέγει αυτή με το συνολικά υψηλότερο σκορ PPI από ολόκληρο το σύνολο. Η συστάδα που αντιστοιχεί σε αυτή την τελική νικήτρια προβολή είναι αυτή που τελικά θα διαχωριστεί, χρησιμοποιώντας ως κατώφλι το σημείο της ελάχιστης πυκνότητάς της (το ολικό ελάχιστο της KDE). Αυτή η διαδικασία εύρεσης του βέλτιστου δείκτη και διαχωρισμού των δεδομένων σε συστάδες επαναλαμβάνεται έως ότου να επιτευχθεί ο επιθυμητός αριθμός συστάδων.

4. Πειραματικός Σχεδιασμός και Αποτελέσματα

Το κεφάλαιο αυτό επικεντρώνεται στην εμπειρική αξιολόγηση του βελτιστοποιημένου αλγορίθμου RLAC και όλων των παραλλαγών του. Σκοπός της εκτενούς πειραματικής διαδικασίας ήταν να απαντηθούν κρίσιμα ερευνητικά ερωτήματα: Πώς αποδίδει το μοντέλο; Ποια είναι η επίδραση των βασικών παραμέτρων, όπως ο αριθμός των προβολών και το εύρος ζώνης του KDE, στην απόδοσή του; Πώς συγκρίνεται η απόδοσή του με αυτή των καθιερωμένων αλγορίθμων σε διαφορετικούς τύπους δεδομένων; Και τέλος, ποιοι δείκτες PPI αποδεικνύονται πιο αποτελεσματικοί υπό διαφορετικές συνθήκες;

4.1) Πλαίσιο Αξιολόγησης

Για να διασφαλιστεί η αντικειμενικότητα και η εγκυρότητα των συμπερασμάτων, ακολουθήθηκε ένα συστηματικό πλαίσιο αξιολόγησης, το οποίο βασίζεται στη μεθοδολογία που προτείνεται στο αρχικό άρθρο-αναφοράς του RLAC. Το πλαίσιο αυτό περιλαμβάνει την επιλογή κατάλληλων συνόλων δεδομένων, την εφαρμογή συγκεκριμένων μετρικών απόδοσης και μια συστηματική προσέγγιση για την εξερεύνηση των παραμέτρων.

Για την πολύπλευρη αξιολόγηση των αλγορίθμων, χρησιμοποιήθηκε μια ποικιλία συνόλων δεδομένων που καλύπτουν τρεις διακριτές κατηγορίες, κάθε μία με τα δικά της μοναδικά χαρακτηριστικά και προκλήσεις:

- Συνθετικά Σύνολα Δεδομένων: Χρησιμοποιήθηκαν τέσσερα συνθετικά σύνολα (Aggregation, Cluto-t4-8k, Dartboard1, Pathbased) τα οποία είναι ειδικά σχεδιασμένα για τη συγκριτική αξιολόγηση αλγορίθμων

συσταδοποίησης. Αυτά τα σύνολα περιλαμβάνουν συστάδες με διαφορετικά σχήματα (σφαιρικά, επιμήκη), μεγέθη, πυκνότητες και βαθμό διαχωρισμού, καθώς και δεδομένα με μη-γραμμικές δομές (ομόκεντροι δακτύλιοι) και παρουσία θορύβου.

- Σύνολα Δεδομένων Πραγματικού Κόσμου: Για να αξιολογηθεί η απόδοση σε ρεαλιστικά σενάρια, επιλέχθηκαν τρία ευρέως γνωστά σύνολα δεδομένων (Breast Cancer Wisconsin, Vertebral Column, Anuran Calls MFCCs). Αυτά τα σύνολα παρουσιάζουν προκλήσεις όπως η επικάλυψη των κλάσεων και η ύπαρξη μη γραμμικών σχέσεων, ελέγχοντας την ικανότητα των αλγορίθμων να ανακαλύπτουν δομές σε πραγματικά, θορυβώδη δεδομένα.
- Μονοκυτταρικά Σύνολα Δεδομένων Γονιδιακής Έκφρασης (GSE Series): Αυτή η κατηγορία αποτελεί την πιο απαιτητική πρόκληση, καθώς τα δεδομένα αυτά χαρακτηρίζονται από εξαιρετικά υψηλή διαστατικότητα (χιλιάδες γονίδια), έντονη αραιότητα (sparsity) και σημαντική βιολογική και τεχνική μεταβλητότητα. Η αξιολόγηση σε αυτά τα σύνολα είναι κρίσιμη για την κατανόηση της πρακτικής χρησιμότητας του RLAC σε σύγχρονα βιοπληροφορικά προβλήματα.

Δεδομένου ότι για όλα τα επιλεγμένα σύνολα δεδομένων οι πραγματικές ετικέτες των κλάσεων (ground truth) είναι γνωστές, η αξιολόγηση της ποιότητας της συσταδοποίησης έγινε με τη χρήση δύο καθιερωμένων εξωτερικών μετρικών:

- Adjusted Mutual Information (AMI): Η μετρική αυτή προέρχεται από τη θεωρία πληροφοριών και ποσοτικοποιεί την αμοιβαία εξάρτηση μεταξύ της ομαδοποίησης που παράγει ο αλγόριθμος και της πραγματικής ομαδοποίησης. Με απλά λόγια, μετρά πόση πληροφορία μοιράζονται οι δύο διαμερίσεις. Η "προσαρμοσμένη" (adjusted) εκδοχή της διορθώνει την τιμή ώστε να λαμβάνεται υπόψη η πιθανότητα τυχαίας συμφωνίας, κάνοντας τη σύγκριση πιο δίκαιη. Μια τιμή AMI κοντά στο 1 σημαίνει ότι η γνώση της μιας διαμέρισης μας δίνει σχεδόν πλήρη πληροφορία για την άλλη, ενώ μια τιμή κοντά στο 0 σημαίνει ότι είναι ανεξάρτητες (όπως θα περίμενε κανείς από μια τυχαία ομαδοποίηση).
- Adjusted Rand Index (ARI): Αυτή η μετρική αξιολογεί την ομαδοποίηση εξετάζοντας όλα τα ζεύγη σημείων στο σύνολο δεδομένων. Μετρά το

ποσοστό των ζευγών που έχουν τοποθετηθεί "σωστά": δηλαδή, ζεύγη που είναι στην ίδια πραγματική κλάση και τοποθετήθηκαν στην ίδια συστάδα από τον αλγόριθμο, καθώς και ζεύγη που είναι σε διαφορετικές πραγματικές κλάσεις και τοποθετήθηκαν σε διαφορετικές συστάδες. Και αυτή η μετρική είναι "προσαρμοσμένη" για την τύχη, ώστε μια τιμή κοντά στο 0 να αντιστοιχεί σε τυχαία απόδοση. Μια τιμή ARI 1 υποδηλώνει τέλεια συμφωνία.

Και οι δύο μετρικές λαμβάνουν τιμές στο διάστημα [-1, 1] ή, όπου η τιμή 1 υποδηλώνει τέλεια αντιστοιχία, ενώ μια τιμή κοντά στο 0 υποδηλώνει αποτέλεσμα ισοδύναμο με τυχαία ομαδοποίηση.

Για την αντιμετώπιση της ευαισθησίας του RLAC στις παραμέτρους του και για τον εντοπισμό των βέλτιστων συνθηκών λειτουργίας του για κάθε σύνολο δεδομένων, υιοθετήθηκε η μεθοδολογία Grid Search. Δημιουργήθηκε ένα πλέγμα τιμών για τις τρεις κρισιμότερες παραμέτρους: τον αριθμό των τυχαίων προβολών (r), τον συντελεστή εύρους ζώνης (bw_adjust) και τον σπόρο τυχαιότητας (random_state). Ο αλγόριθμος εκτελέστηκε για κάθε πιθανό συνδυασμό αυτών των παραμέτρων, επιτρέποντας μια συστηματική εξερεύνηση του χώρου των παραμέτρων και την ανάλυση της επίδρασης καθεμιάς από αυτές στην τελική απόδοση.

4.2) Ανάλυση Αποτελεσμάτων

Τα αποτελέσματα που προέκυψαν από την εκτενή αυτή πειραματική διαδικασία παρουσιάζονται και αναλύονται ξεχωριστά για κάθε κατηγορία συνόλου δεδομένων, εστιάζοντας στη σύγκριση της απόδοσης του βελτιστοποιημένου RLAC τόσο με τους καθιερωμένους αλγορίθμους όσο και μεταξύ των διαφορετικών παραλλαγών του.

4.2.1) Συνθετικά Σύνολα

Στα συνθετικά σύνολα δεδομένων, το βελτιστοποιημένο πλαίσιο RLAC έδειξε εξαιρετική απόδοση, ιδίως σε δεδομένα με σαφώς διαχωρισμένες και γραμμικά διαχωρίσιμες συστάδες όπως το Aggregation, όπου σχεδόν όλες οι παραλλαγές πέτυχαν σχεδόν τέλεια αποτελέσματα, συχνά ξεπερνώντας τους καθιερωμένους αλγορίθμους. Η ανάλυση μέσω Grid Search έδειξε ότι η απόδοση ήταν σταθερή σε ένα ευρύ φάσμα παραμέτρων, αποδεικνύοντας τη αποτελεσματικότητα του μοντέλου.

Η απόδοση των μοντέλων συσταδοποίησης στα τεχνητά σύνολα δεδομένων συνοψίζεται στον Πίνακα 1. Κάθε σύνολο δεδομένων προ-επεξεργάστηκε χρησιμοποιώντας το StandardScaler για την κανονικοποίηση των δεδομένων και στη συνέχεια εφαρμόστηκε Ανάλυση Κύριων Συνιστωσών (PCA) για τη μείωση της διαστατικότητας. Εάν η αρχική διαστατικότητα ήταν πολύ υψηλή, δηλαδή εάν οι διαστάσεις ξεπερνούν τις 50 εφαρμόζεται η PCA και διατηρούνται μόνο οι 50 πρώτες κύριες συνιστώσες (όπως αναφέρεται και στο άρθρο).

Το πειραματικό πλαίσιο σχεδιάστηκε για να διερευνήσει τρεις βασικές παραμέτρους:

- Αριθμός Τυχαίων Προβολών (r): Εξετάστηκαν οι τιμές [50, 100, 200, 300, 400] για να μελετηθεί η επίδραση της διαστατικότητας του προβαλλόμενου χώρου.
- Συντελεστής Εύρους Ζώνης : Δοκιμάστηκαν οι τιμές [0.1, 0.2, 0.3, 0.4, 0.5] για την εύρεση της βέλτιστης κατανομής πυκνότητας.
- Αρχικοποιητής Τυχαιότητας (random_state): Χρησιμοποιήθηκαν οι τιμές [42, 43, 44, 45] για τον έλεγχο της σταθερότητας των αποτελεσμάτων.

Πίνακας 1. Αποτελέσματα μετρήσεων από τα συνθετικά σύνολα δεδομένων

Datasets / ML Models	Aggregation	Cluto-t4-8k	Dartboard1	Pathbased
Evaluation Metrics	**AMI/ARI**	**AMI/ARI**	**AMI/ARI**	**AMI/ARI**
KMeans	0.83/ 0.73	0.68/ 0.53	0.0/ 0.0	0.55/ 0.47
HClust	0.88/ 0.80	**0.95/ 0.92**	**1.0/ 1.0**	0.0/ 0.0
NCut	0.72/ 0.47	0.74/ 0.51	1.0/ 1.0	**0.55/ 0.50**
Mdh	0.82/ 0.73	0.73/ 0.61	0.17/ 0.10	0.51/ 0.43
RLAC	0.99/ 0.99	0.0/ 0.0	0.23/ 0.06	0.51/ 0.42
Dip-RLAC	0.99/ 0.99	0.0/ 0.0	0.23/ 0.06	0.51/ 0.42
Holes-RLAC	0.0/ 0.0	0.0/ 0.0	0.0/ 0.0	0.0/ 0.0
Min- Kurt-RLAC	0.99/ 0.99	0.0/ 0.0	0.23/ 0.06	0.48/ 0.39
Max- Kurt-RLAC	0.0/ 0.0	0.0/ 0.0	0.0/ 0.0	0.18/ 0.04
Negentropy-RLAC	0.99/ 0.99	0.0/ 0.0	0.23/ 0.06	0.51/ 0.42
Skewness-RLAC	0.99/ 0.99	0.0/ 0.0	0.19/ 0.04	0.55/ 0.47
Fisher- RLAC	**0.99/ 0.99**	0.0/ 0.0	0.19/ 0.04	0.51/ 0.42
Hermit- RLAC	0.99/ 0.99	0.0/ 0.0	0.23/ 0.06	0.51/ 0.42
Friedman-Tukey-RLAC	**0.99/ 0.99**	0.0/ 0.0	0.23/ 0.06	0.55/ 0.47

Τα αποτελέσματα επιβεβαίωσαν την αποτελεσματικότητα του RLAC, ιδίως σε δεδομένα με σαφώς διαχωρισμένες και κατά βάση γραμμικά διαχωρίσιμες συστάδες. Στο σύνολο "Aggregation", σχεδόν όλες οι παραλλαγές του RLAC, πέτυχαν σχεδόν τέλεια ομαδοποίηση (AMI/ARI ≈ 0.99/0.99), επιδεικνύοντας απόδοση ανώτερη των καθιερωμένων αλγορίθμων όπως ο KMeans και ο HClust. Στο πιο πολύπλοκο σύνολο "Pathbased", που περιέχει επιμήκεις δομές, οι περισσότεροι δείκτες RLAC (όπως οι Negentropy, Skewness, Fisher, Friedman-Tukey και ο αρχικός RLAC) είχαν επίσης πολύ καλή απόδοση (AMI ≈ 0.51-0.55), ανταγωνιζόμενοι επιτυχώς τους KMeans, NCut και Mdh. Αντίθετα, στο σύνολο "Cluto-t4-8k", με τις πολύπλοκες και επικαλυπτόμενες συστάδες του, καμία παραλλαγή του RLAC δεν κατάφερε να βρει ουσιαστική δομή, σε αντίθεση με τον HClust που πέτυχε (0.95/0.92), υποδεικνύοντας ότι η ιεραρχική προσέγγιση στον αρχικό χώρο ήταν πιο κατάλληλη. Τέλος, στο μη-γραμμικό σύνολο "Dartboard", ο RLAC, όπως ήταν αναμενόμενο, υστέρησε σημαντικά έναντι των HClust και NCut που σχεδιάστηκαν για τέτοιες δομές. Οι δείκτες Holes-RLAC και Max-Kurt-RLAC παρουσίασαν σταθερά χαμηλές επιδόσεις σε όλα τα συνθετικά σύνολα, υποδηλώνοντας ότι οι συγκεκριμένες παραλλαγές είναι λιγότερο γενικεύσιμες για αυτό το είδος συνόλου δεδομένων.

4.2.2) Σύνολα Πραγματικού Κόσμου

Στα σύνολα δεδομένων του πραγματικού κόσμου, τα αποτελέσματα του βελτιστοποιημένου RLAC ήταν δραματικά βελτιωμένα. Ενώ προηγουμένως τα σκορ ήταν μηδενικά, το μοντέλο πλέον κατάφερε να παράγει μη-τυχαίες και συχνά μέτριες βαθμολογίες, αποδεικνύοντας ότι με τη σωστή παραμετροποίηση μπορεί να ανακαλύψει μερική δομή. Παρατηρήθηκε ότι η επιτυχία εξαρτιόταν σε μεγάλο βαθμό από τον σωστό συνδυασμό παραμέτρων, με τις μικρότερες τιμές εύρους ζώνης και έναν επαρκή αριθμό προβολών να παίζουν καθοριστικό ρόλο, επιβεβαιώνοντας τα ευρήματα της φάσης βελτιστοποίησης.

Με σκοπό την προεπεξεργασία των δεδομένων για συσταδοποίηση, αφαιρέθηκαν ακραίες τιμές όπου ήταν απαραίτητο, τα δεδομένα κανονικοποιήθηκαν με τη χρήση StandardScaler, και εφαρμόστηκε Ανάλυση Κυρίων Συνιστωσών (PCA) για μείωση διάστασης όπου κρίθηκε απαραίτητο, ακολουθώντας την ίδια μεθοδολογία με τα συνθετικά σύνολα δεδομένων. Η ανάλυση δείχνει τις δυνατότητες και τους

περιορισμούς όλων των αλγορίθμων σε υψηλής διάστασης δεδομένα με πιθανές επικαλύψεις κλάσεων και μη γραμμικές δομές.

Παρακάτω είναι το πειραματικό πλαίσιο που σχεδιάστηκε για να διερευνήσει τις τρεις βασικές παραμέτρους και τη σχέση τους με την επίδοση των μοντέλων:

- Αριθμός Τυχαίων Προβολών (r): Εξετάστηκαν οι τιμές [None, 50, 100, 200, 500, r όταν e=0.1] για να μελετηθεί η επίδραση της διαστατικότητας του προβαλλόμενου χώρου. Η τιμή του e είναι ίση με 0.2 ως προεπιλογή σε όλα τα μοντέλα RLAC. Η τιμή r όταν e=0.1 προστέθηκε για να παρατηρήσουμε την επίδοση των μοντέλων όταν e = 0.1.
- Συντελεστής Εύρους Ζώνης (bw_adjust): Δοκιμάστηκαν οι τιμές [0.05, 0.1, 0.15, 0.2, 0.3] για την εύρεση της βέλτιστης κατανομής πυκνότητας. Οι τιμές του εύρους σε αντίθεση με τα προηγούμενα σύνολα δεδομένων είναι μικρότερες με σκοπό να αναδείξουν καλύτερα (με πιο λεπτομερή τρόπο) την κατανομή πυκνότητας των δεδομένων της προβολής και να αποφύγουμε πολύ λείες κατανομές που δεν οδηγούν σε ομαδοποίηση.
- Αρχικοποιητής Τυχαιότητας (random_state): Χρησιμοποιήθηκαν οι τιμές [32, 42, 43, 44, 45] για τον έλεγχο της σταθερότητας των αποτελεσμάτων.

Πίνακας 2. Αποτελέσματα μετρήσεων από πραγματικά σύνολα δεδομένων

Datasets / ML Models	**Breast Cancer Wisconsin (Diagnostic)**	**Vertebral Column**	**Anuran Calls MFCCs**
Evaluation Metrics	**AMI/ARI**	**AMI/ARI**	**AMI/ARI**
KMeans	0.61/ 0.74	**0.35/ 0.27**	0.56/ 0.28
HClust	0.00/ 0.00	-0.0/ -0.0	0.00/ 0.00
NCut	**0.63/ 0.74**	0.31/ 0.23	**0.77/ 0.80**
Mdh	0.59/ 0.68	0.29/ 0.31	0.52/ 0.19
RLAC	0.31/ 0.31	0.20/ 0.00	0.35/ 0.24
Dip-RLAC	0.27/ 0.27	0.17/ -0.01	0.31/ 0.21
Holes-RLAC	0.0 0.0	0.0/ 0.0	0.0/ 0.0
Min- Kurt- RLAC	0.15/ 0.13	0.20/ 0.00	0.0/ 0.0
Max- Kurt- RLAC	0.27/ 0.27	0.0/ 0.0	0.33/ 0.23

Negentropy-RLAC	0.27/ 0.27	0.17/ -0.01	0.31/ 0.21
Skewness-RLAC	0.27/ 0.27	0.20/ 0.00	0.33/ 0.23
Fisher- RLAC	0.27/ 0.27	0.17/ -0.01	0.31/ 0.21
Hermit- RLAC	0.27/ 0.27	0.17/ -0.01	0.33/ 0.23
Friedman-Tukey-RLAC	0.27/ 0.27	0.17/ -0.01	0.31/ 0.21

Όπως φαίνεται στον Πίνακα 2, στο σύνολο "Anuran Calls (MFCCs)", οι περισσότερες παραλλαγές του RLAC (RLAC-DepthRatio, Max-Kurt, Negentropy, Skewness, Fisher, Hermit, Friedman-Tukey) κατάφεραν να επιτύχουν μέτρια απόδοση (AMI ≈ 0.31-0.35), αν και υστέρησαν έναντι του NCut που κυριάρχησε (0.77/0.80). Στο σύνολο "Breast Cancer Wisconsin", παρόλο που η βελτίωση ήταν εμφανής (π.χ., RLAC-DepthRatio με AMI 0.31), η απόδοση παρέμεινε χαμηλή σε σύγκριση με τους KMeans, NCut και Mdh (AMI > 0.59), υποδεικνύοντας ότι η γραμμική δομή του συνόλου ευνοεί τους αλγορίθμους που λειτουργούν στον αρχικό πολυδιάστατο χώρο. Στο "Vertebral Column", οι επιδόσεις του RLAC ήταν γενικά χαμηλότερες, με τους Skewness-RLAC και Min-Kurt-RLAC να έχουν την καλύτερη, αν και μέτρια, απόδοση μεταξύ των παραλλαγών του, ενώ ο KMeans (0.35/0.27) ήταν ο πιο αποτελεσματικός συνολικά. Ο HClust και ο Holes-RLAC παρουσίασαν πολύ χαμηλές επιδόσεις σε όλα τα πραγματικά σύνολα.

4.2.3) Μονοκυτταρικά Σύνολα Γονιδιακής Έκφρασης

Η πιο εντυπωσιακή βελτίωση παρατηρήθηκε στα μονοκυτταρικά σύνολα δεδομένων. Σε σύνολα όπου προηγουμένως αποτύγχανε πλήρως, το βελτιστοποιημένο μοντέλο RLAC, με τον κατάλληλο συνδυασμό παραμέτρων, κατάφερε να επιτύχει από πολύ καλές έως και τέλειες επιδόσεις (π.χ., AMI=1.0 στο GSE42268), ξεπερνώντας σε ορισμένες περιπτώσεις ακόμη και τους καθιερωμένους αλγορίθμους.

Τέλος, η συγκριτική αξιολόγηση των δεικτών PPI αποκάλυψε ότι δεν υπάρχει ένας μοναδικός «καλύτερος» δείκτης για όλα τα σενάρια. Δείκτες όπως ο Negentropy και ο Fisher αποδείχθηκαν ιδιαίτερα ισχυροί σε δεδομένα με καλά διαχωρισμένες συστάδες, ενώ ο Skewness και ο Max-Kurtosis έδειξαν τη χρησιμότητά τους σε περιπτώσεις με άνισες συστάδες ή ακραίες τιμές. Αυτό υποδηλώνει ότι η επιλογή του

κατάλληλου PPI μπορεί να εξαρτάται από την υποκείμενη φύση των δεδομένων και την αναμενόμενη δομή των συστάδων.

Για την προετοιμασία αυτών των δεδομένων για συσταδοποίηση, εφαρμόστηκε ένας αγωγός (pipeline) επεξεργασίας που ξεκινούσε με φιλτράρισμα ποιοτικού ελέγχου των ακατέργαστων πινάκων μετρήσεων. Τα γονίδια διατηρήθηκαν μόνο εάν εκφράζονταν σε έναν ελάχιστο αριθμό κυττάρων, και ομοίως, τα κύτταρα διατηρήθηκαν εάν εξέφραζαν έναν ελάχιστο αριθμό γονιδίων. Αυτό το βήμα στοχεύει στην απομάκρυνση κυττάρων χαμηλής ποιότητας και γονιδίων που δεν παρέχουν ουσιαστική πληροφορία. Στη συνέχεια, ακολούθησε κανονικοποίηση μέσω λογαριθμικού μετασχηματισμού (π.χ., log(μετρήσεις + 1)), σταθεροποιώντας τη διακύμανση και καθιστώντας τις τιμές έκφρασης πιο συγκρίσιμες μεταξύ των κυττάρων. Με σκοπό να εστιάσουμε σε σημαντικά βιολογικά σήματα και τη μείωση του υπολογιστικού φόρτου, εντοπίστηκαν τα γονίδια υψηλής μεταβλητότητας (Highly Variable Genes - HVGs) υπολογίζοντας μια κανονικοποιημένη διασπορά (λόγος διακύμανσης προς μέσο όρο) για κάθε γονίδιο σε όλα τα κύτταρα, και επιλέγοντας τα κορυφαία γονίδια που εμφάνιζαν την υψηλότερη διασπορά. Ο πίνακας δεδομένων που προκύπτει, που πλέον αποτελείται από κύτταρα (δείγματα) έναντι επιλεγμένων γονιδίων υψηλής μεταβλητότητας (χαρακτηριστικά), κλιμακώθηκε χρησιμοποιώντας StandardScaler. Αυτό διασφαλίζει ότι κάθε γονίδιο συμβάλλει εξίσου στις επόμενες αναλύσεις, ανεξάρτητα από το απόλυτο μέγεθος της έκφρασής του. Ως τελικό βήμα προεπεξεργασίας, εφαρμόστηκε PCA στον πίνακα έκφρασης HVG για τη μείωση της διαστατικότητας του, διατηρώντας ένα σημαντικό ποσοστό της διακύμανσης.

Παρακάτω είναι το πειραματικό πλαίσιο που σχεδιάστηκε για να διερευνήσει τις τρεις βασικές παραμέτρους και τη σχέση τους με την επίδοση των μοντέλων:

- Αριθμός Τυχαίων Προβολών (r): Εξετάστηκαν οι τιμές [None, 50, 100, 200, 300, 500] για να μελετηθεί η επίδραση της διαστατικότητας του προβαλλόμενου χώρου.
- Συντελεστής Εύρους Ζώνης (bw_adjust): Δοκιμάστηκαν οι τιμές [0.05, 0.1, 0.2, 0.3, 0.4, 0.5] για την εύρεση της βέλτιστης κατανομής πυκνότητας.
- Αρχικοποιητής Τυχαιότητας (random_state): Χρησιμοποιήθηκαν οι τιμές [32, 42, 43, 44, 45] για τον έλεγχο της σταθερότητας των αποτελεσμάτων.

Πίνακας 3. Αποτελέσματα μετρήσεων από τα γονιδιακά σύνολα δεδομένων

Datasets / ML Models	**GSE42268**	**GSE55291**	**GSE65528**	**GSE74596**	**GSE41265**	**GSE45719**	**GSE52583**
Evaluation Metrics	**AMI/ARI**	**AMI/ARI**	**AMI/ARI**	**AMI/ARI**	**AMI/ARI**	**AMI/ARI**	**AMI/ARI**
KMeans	**1.0/ 1.0**	**0.69/ 0.59**	0.54/ 0.47	0.01/ -0.00	0.77/ 0.78	0.31/ 0.29	0.41/ 0.34
HClust	0.05/ 0.07	0.04/ 0.03	0.00/ 0.00	-0.00/ -0.00	0.14/ 0.12	0.51/ 0.60	-0.00/ -0.01
NCut	1.0/ 1.0	0.59/ 0.53	**0.59/ 0.47**	**0.54/ 0.55**	**0.77/ 0.78**	0.19/ 0.15	**0.66/ 0.57**
Mdh	0.04/ 0.10	0.08/ 0.05	0.10/ 0.09	0.06/ 0.01	0.17/ 0.13	0.04/ 0.00	0.40/ 0.35
RLAC	1.0/ 1.0	0.69/ 0.59	0.23/ 0.13	0.07/ 0.00	0.58/ 0.42	0.77/ 0.82	0.21/ 0.07
Dip-RLAC	1.0/ 1/0	0.55/ 0.56	0.13/ 0.06	0.08/ -0.00	0.0/ 0.0	0.70/ 0.79	0.10/ 0.04
Holes-RLAC	0.41/ 0.48	0.34/ 0.26	0.11/ 0.04	0.09/ 0.01	0.58/ 0.42	0.77/ 0.82	0.10/ 0.07
Min- Kurt-RLAC	1.0/ 1/0	0.55/ 0.56	0.18/ 0.10	0.0/ 0.0	0.58/ 0.42	0.37/ 0.39	0.15/ 0.06
Max- Kurt-RLAC	0.60/ 0.74	0.46/ 0.32	0.11/ 0.05	0.06/ -0.00	0.0/ 0.0	0.55/ 0.62	0.12/ 0.05
Negentropy-RLAC	1.0/ 1.0	0.57/ 0.38	0.18/ 0.13	0.05/ 0.00	0.58/ 0.42	0.77/ 0.82	0.12/ 0.06
Skewness-RLAC	1.0/ 1.0	0.57/ 0.38	0.25/ 0.18	0.05/ 0.00	0.58/ 0.42	**0.85/ 0.91**	0.10/ 0.04
Fisher-RLAC	1.0/ 1.0	0.57/ 0.38	0.18/ 0.12	0.06/ 0.00	0.58/ 0.42	0.77/ 0.82	0.11/ 0.07
Hermit-RLAC	1.0/ 1.0	0.55/ 0.56	0.19/ 0.13	0.06/ -0.00	0.58/ 0.42	0.55/ 0.62	0.10/ 0.07
Friedman-Tukey-RLAC	1.0/ 1.0	0.57/ 0.38	0.20/ 0.14	0.05/ 0.00	0.58/ 0.42	0.77/ 0.82	0.10/ 0.03

Σε σύνολα όπου η αρχική υλοποίηση αποτύγχανε πλήρως, το βελτιστοποιημένο μοντέλο RLAC, με τον κατάλληλο συνδυασμό παραμέτρων από το Grid Search, κατάφερε να επιτύχει από πολύ καλές έως και τέλειες επιδόσεις, όπως φαίνεται στον

Πίνακα 3. Στο σύνολο "GSE42268", οι περισσότεροι δείκτες RLAC (RLAC-DepthRatio, Dip-RLAC, Min-Kurt, Negentropy, Skewness, Fisher, Hermit, Friedman-Tukey) πέτυχαν τέλεια ομαδοποίηση (AMI/ARI = 1.0/1.0), πλησιάζοντας την απόδοση των KMeans και NCut. Στο σύνολο "GSE45719", ο Skewness-RLAC ξεχώρισε με εξαιρετική απόδοση (0.85/0.91), ξεπερνώντας όλους τους άλλους αλγορίθμους, συμπεριλαμβανομένων των καθιερωμένων. Ακόμα και σε άλλα απαιτητικά σύνολα όπως τα "GSE55291" και "GSE41265", πολλές παραλλαγές του RLAC πέτυχαν υψηλές βαθμολογίες (π.χ., AMI > 0.55), αποδεικνύοντας την ανταγωνιστικότητά τους. Η συγκριτική αξιολόγηση των δεικτών PPI επιβεβαίωσε ότι δεν υπάρχει ένας μοναδικός "καλύτερος" δείκτης, αλλά η επιλογή εξαρτάται από τη φύση των δεδομένων, με δείκτες όπως οι Negentropy, Skewness και ο αρχικός RLAC να αποδεικνύονται συχνά πολύ ισχυροί.

5. Συμπεράσματα

Η πορεία της παρούσας εργασίας ακολούθησε μια διαδρομή από τον εντοπισμό των προκλήσεων κατά την αρχική υλοποίηση του RLAC, προς την ανάπτυξη ενός υψηλής απόδοσης πλαισίου ομαδοποίησης, το οποίο είναι ικανό να αντανακλά τις δυνατότητες που περιγράφονται στο άρθρο-αναφοράς. Η αρχική προσπάθεια υλοποίησης του αλγορίθμου, παρουσίαζε σημαντικά προβλήματα στην πράξη. Εμφάνιζε χαμηλή απόδοση σε πολύπλοκα σύνολα δεδομένων και, το πιο κρίσιμο, μια σημαντική έλλειψη αναπαραγωγιμότητας που καθιστούσε αδύνατη την εξαγωγή αξιόπιστων συμπερασμάτων. Αυτό το αρχικό αποτέλεσμα έθετε υπό αμφισβήτηση την ορθότητα της αρχικής υλοποίησης, και αποτέλεσε το έναυσμα για τον αποτελεσματικότερο ανασχεδιασμό του μοντέλου.

Η λύση προήλθε από τρεις στοχευμένες παρεμβάσεις που διόρθωσαν τις αρχικές αδυναμίες. Πρώτον, η εισαγωγή της παραμέτρου random_state επέλυσε οριστικά το πρόβλημα της αναπαραγωγιμότητας, μετατρέποντας την υλοποίηση σε ένα έγκυρο εργαλείο ομαδοποίησης, όπου τα αποτελέσματα είναι πλέον σταθερά και συγκρίσιμα. Δεύτερον, η μετάβαση από τις πυκνές Γκαουσιανές προβολές σε αραιές προβολές Achlioptas αύξησαν δραματικά την υπολογιστική αποδοτικότητα. Τρίτον, και ίσως το πιο σημαντικό, η προσεκτική ρύθμιση του εύρους ζώνης του KDE αναδείχθηκε ως ο

πιο κρίσιμος παράγοντας. Διαπιστώθηκε ότι οι μεγάλες τιμές εύρους ζώνης οδηγούσαν σε υπερ-εξομάλυνση, αποκρύπτοντας τις δομές των συστάδων και με χρήση μικρότερων, προσαρμοσμένων τιμών ο αλγόριθμος μπορούσε να διακρίνει τις κοιλάδες ανάμεσα στις συστάδες.

Τα αποτελέσματα του βελτιωμένου μοντέλου ήταν ενθαρυντικά και επιβεβαίωσαν ότι οι παρεμβάσεις ήταν ορθές και απαραίτητες. Σε ιδανικές συνθήκες, όπως στα συνθετικά δεδομένα, ο RLAC επέδειξε εξαιρετική και, πλέον, σταθερή απόδοση, ευθυγραμμιζόμενος με τα ευρήματα της αρχικής δημοσίευσης. Το πιο σημαντικό εύρημα, ωστόσο, ήταν η δραματική βελτίωση της απόδοσής του σε πολύπλοκα, πραγματικά και σε μονοκυτταρικά δεδομένα υψηλής διαστατικότητας. Σε περιπτώσεις όπου η αρχική, προβληματική υλοποίηση αποτύγχανε πλήρως, το βελτιστοποιημένο μοντέλο, με τον κατάλληλο συνδυασμό παραμέτρων που εντοπίστηκε μέσω Grid Search, πέτυχε εξαιρετικές επιδόσεις (π.χ., AMI=1.0 στο GSE42268), αποδεικνύοντας ότι ο αλγόριθμος RLAC, όταν υλοποιηθεί σωστά και παραμετροποιηθεί μεθοδικά, είναι πράγματι ένα ισχυρό εργαλείο ομαδοποίησης.

Οι νέοι δείκτες αναζήτησης προβολών που υλοποιήθηκαν στο πλαίσιο της παρούσας εργασίας έδειξαν ότι μπορούσαν να συγκριθούν με τις αρχικούς δείκτες του άρθρου και να παράγουν ουσιαστικές ομαδοποιήσεις των δεδομένων. Η συγκριτική ανάλυση αποκάλυψε ότι δεν υπάρχει ένας μοναδικός, παγκοσμίως καλύτερος δείκτης, αλλά η απόδοσή του εξαρτάται από τα χαρακτηριστικά του εκάστοτε συνόλου δεδομένων. Για παράδειγμα, ο Negentropy PPI και ο αρχικός Depth Ratio PPI αναδείχθηκαν ως ιδιαίτερα ισχυροί και σταθεροί σε ποικίλες συνθήκες, ικανοί να ανιχνεύουν γενική απόκλιση από την κανονικότητα και σαφείς διαχωρισμούς. Ο Skewness PPI και ο Max-Kurtosis PPI αποδείχθηκαν ιδιαίτερα αποτελεσματικοί σε περιπτώσεις με άνισες σε μέγεθος συστάδες ή στην απομόνωση μικρών, πυκνών ομάδων. Ο προσαρμοσμένος Fisher PPI έδειξε τη δύναμή του σε προβολές όπου οι συστάδες ήταν καλά διαχωρισμένες ως προς τις μέσες τιμές τους, ενώ ο Hermite PPI λειτούργησε ως ένας καλός γενικός δείκτης μη-κανονικότητας. Αντίθετα, ο Friedman-Tukey PPI, παρά τη στιβαρή θεωρητική βάση, αποδείχθηκαν πιο ευαίσθητοι στην παραμετροποίηση και λιγότερο αποτελεσματικοί στα σύνολα δεδομένων που εξετάστηκαν. Αυτό το εύρημα υποδηλώνει ότι η επιλογή του κατάλληλου PPI μπορεί να είναι μια μορφή "feature engineering" σε επίπεδο αλγορίθμου, όπου η γνώση της αναμενόμενης δομής των συστάδων μπορεί να καθοδηγήσει την επιλογή του πιο κατάλληλου δείκτη για την ανάλυση.

Συμπερασματικά, η εργασία αυτή ενισχύει την αξία του αλγορίθμου RLAC, αλλά ταυτόχρονα αναδεικνύει και τις προκλήσεις της πρακτικής του εφαρμογής. Απέδειξε ότι η επιτυχία των μεθόδων Projection Pursuit εξαρτάται σε κρίσιμο βαθμό από τη σωστή παραμετροποίηση, την υπολογιστική αποδοτικότητα και, πάνω απ' όλα, από τη διασφάλιση της αναπαραγωγιμότητας, στοιχεία που είναι θεμελιώδη για την εξαγωγή έγκυρων επιστημονικών συμπερασμάτων.

Η απόδοση του βελτιστοποιημένου πλαισίου RLAC, ιδίως στα μονοκυτταρικά δεδομένα, μπορεί να ερμηνευθεί από τη συμβατότητα της φιλοσοφίας του με τη φύση αυτών των προβλημάτων. Σε χώρους εξαιρετικά υψηλών διαστάσεων, όπως αυτοί της γονιδιακής έκφρασης, οι παραδοσιακές μετρικές απόστασης συχνά αποτυγχάνουν, καθώς κυριαρχούνται από τον θόρυβο και την "κατάρα της διαστατικότητας". Ο RLAC, αντίθετα, δεν προσπαθεί να υπολογίσει αποστάσεις σε αυτόν τον θορυβώδη χώρο. Αντ' αυτού, υιοθετεί μια στρατηγική "διαίρει και βασίλευε", αναζητώντας την κατάλληλη μονοδιάστατη προβολή που αποκαλύπτει την πιο σαφή και διαχωρίσιμη δομή των δεδομένων. Η τυχαιότητα της αναζήτησης, σε συνδυασμό με την αξιολόγηση εκατοντάδων τέτοιων προβολών, αυξάνει εκθετικά την πιθανότητα να εντοπιστεί μια τέτοια διαχωριστική κατεύθυνση, ακόμα και αν αυτή είναι κρυμμένη μέσα σε χιλιάδες άλλες, ασήμαντες διαστάσεις. Αυτό εξηγεί γιατί, με τη σωστή παραμετροποίηση, κατάφερε να επιτύχει αποδοτική ομαδοποίηση στα σύνολα δεδομένων.

Παρόλα αυτά, η μελέτη ανέδειξε και περιορισμούς του αλγορίθμου με την κύρια αδυναμία του παραμένει η δυσκολία του να διαχειριστεί αποτελεσματικά μη-γραμμικές δομές, όπως φάνηκε καθαρά στην περίπτωση του "Dartboard". Επειδή ο μηχανισμός του βασίζεται σε γραμμικές προβολές (ευθείες γραμμές) και γραμμικούς διαχωρισμούς (ένα σημείο-κατώφλι σε αυτή τη γραμμή, που αντιστοιχεί σε ένα υπερεπίπεδο στον αρχικό χώρο), είναι εκ φύσεως ακατάλληλος για συστάδες που δεν είναι γραμμικά διαχωρίσιμες, όπως οι ομόκεντροι δακτύλιοι. Σε αυτές τις περιπτώσεις, η ανωτερότητα μεθόδων όπως ο NCut (Normalized Cut), που μετασχηματίζει τα δεδομένα σε έναν φασματικό χώρο όπου οι δομές γίνονται γραμμικά διαχωρίσιμες, ήταν αναμενόμενη και επιβεβαιώθηκε από τα αποτελέσματα.

Τέλος, παρά τις σημαντικές βελτιώσεις στη σταθερότητα και την απόδοση, ο RLAC παραμένει ένας αλγόριθμος ευαίσθητος στην παραμετροποίηση. Η επιτυχία του εξαρτάται σε μεγάλο βαθμό από την εύρεση του σωστού συνδυασμού των παραμέτρων r (αριθμός προβολών) και, κυρίως του εύρους ζώνης. Όπως έδειξε η

διαδικασία του Grid Search, δεν υπάρχει ένας μαγικός συνδυασμός που να λειτουργεί για όλα τα είδη δεδομένων. Αυτό καθιστά αναγκαία τη ρύθμιση των παραμέτρων για κάθε νέο σύνολο. Η παραμετροποίηση λοιπόν, απαιτεί προσαρμογή στις ανάγκες των δεδομένων και ρύθμιση με σκοπό την αποτελεσματικότερη εφαρμογή του αλγορίθμου ομαδοποίησης στο σύνολο δεδομένων. Όπως αποδείχτηκε όταν η ρύθμιση αυτή μπορεί αν οδηγήσει σε ανταγωνιστικά αποτελέσματα με τους καθιερωμένους αλγορίθμους.

6. Βιβλιογραφία

[1] (P. Barbas, A. G. Vrahatis and S. K. Tasoulis, "RLAC: Random Line Approximation Clustering," 2021 IEEE International Conference on Big Data (Big Data), Orlando, FL, USA, 2021, pp. 985-993, doi: 10.1109/BigData52589.2021.9671596. keywords: {Dimensionality r)

[2] (Hartigan, J. A., & Hartigan, P. M. (1985). The Dip Test of Unimodality. The Annals of Statistics, 13(1), 70–84. http://www.jstor.org/stable/2241144)

[3] (Cook, D., Buja, A., & Cabrera, J. (1993). Projection Pursuit Indexes Based on Orthonormal Function Expansions. Journal of Computational and Graphical Statistics, 2(3), 225–250. https://doi.org/10.1080/10618600.1993.10474610)

[4] (Peña, D., & Prieto, F. J. (2001). Multivariate Outlier Detection and Robust Covariance Matrix Estimation. Technometrics, 43(3), 286–310. https://doi.org/10.1198/004017001316975899)

[5] (J. H. Friedman and J. W. Tukey, "A Projection Pursuit Algorithm for Exploratory Data Analysis," in IEEE Transactions on Computers, vol. C-23, no. 9, pp. 881-890, Sept. 1974, doi: 10.1109/T-C.1974.224051.)

[6] (C. E. Shannon, "A mathematical theory of communication," in The Bell System Technical Journal, vol. 27, no. 3, pp. 379-423, July 1948, doi: 10.1002/j.1538-7305.1948.tb01338.x.)

[7] (Pierre Comon. 1994. Independent component analysis, a new concept? Signal Process. 36, 3 (April 1994), 287–314. https://doi.org/10.1016/0165-1684(94)90029-9)

[8] (Hyvärinen A, Oja E. Independent component analysis: algorithms and applications. Neural Netw. 2000 May-Jun;13(4-5):411-30. doi: 10.1016/s0893-6080(00)00026-5. PMID: 10946390.)

[9] (Friedman, J. H. (1987). Exploratory Projection Pursuit. Journal of the American Statistical Association, 82(397), 249–266. https://doi.org/10.2307/2289161)

[10] (Huber, P. J. (1985). Projection Pursuit. The Annals of Statistics, 13(2), 435–475. http://www.jstor.org/stable/2241175)

[11] (M.C. Jones, Robin Sibson, What is Projection Pursuit?, Royal Statistical Society. Journal. Series A: General, Volume 150, Issue 1, January 1987, Pages 1–18, https://doi.org/10.2307/2981662)

[12] (Fisher, R.A. (1936) The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics, 7, 179-188.)

[13] (Pavlidis, NG, Hofmeyr, DP and Tasoulis, SK (2016) Minimum Density Hyperplanes. Journal of Machine Learning Research, 17 (156). pp. 1-33. ISSN 1532-4435)

[14] (P. Fänti and S. Sieranoja K-means properties on six clustering benchmark datasets Applied Intelligence, 48 (12), 4743-4759, December 2018 https://doi.org/10.1007/s10489-018-1238-7)

[15] (G. Karypis, "CLUTO A Clustering Toolkit," Dept. of Computer Science, University of Minnesota,Tech. Rep. 02-017, 2002, available at http://www.cs.umn.edu/~cluto)

[16] (Pallas, D. (2019). Dartboard1 dataset [Data set file]. In Clustering-benchmark. GitHub. Retrieved [Month Day, Year], from https://github.com/deric/clustering-benchmark/blob/master/src/main/resources/datasets/artificial/dartboard1.arff)

[17] (Wolberg, W., Mangasarian, O., Street, N., & Street, W. (1993). Breast Cancer Wisconsin (Diagnostic) [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5DW2B.)

[18] (Barreto, G. & Neto, A. (2005). Vertebral Column [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5K89B.)

[19] (Colonna, J., Nakamura, E., Cristo, M., & Gordo, M. (2015). Anuran Calls (MFCCs) [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5CC9H.)

[20] (Sasagawa, Y., Nikaido, I., Hayashi, T., Danno, H., Uno, K. D., Imai, T., & Ueda, H. R. (2013). Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. Genome biology, 14(4), R31)

[21] (Kim, D. H., Marinov, G. K., Pepke, S., Singer, Z. S., He, P., Williams, B., Schroth, G. P., Elowitz, M. B., & Wold, B. J. (2015). Single-cell transcriptome analysis reveals dynamic changes in lncRNA expression during reprogramming. Cell stem cell, 16(1),)

[22] (Avraham, R., Haseley, N., Brown, D., Penaranda, C., Jijon, H. B., Trombetta, J. J., Satija, R., Shalek, A. K., Xavier, R. J., Regev, A., & Hung, D. T. (2015). Pathogen Cell-to-Cell Variability Drives Heterogeneity in Host Immune Responses. Cell, 162(6), 1)

[23] (Engel, I., Seumois, G., Chavez, L., Samaniego-Castruita, D., White, B., Chawla, A., Mock, D., Vijayanand, P., & Kronenberg, M. (2016). Innate-like functions of natural killer T cell subsets result from highly divergent gene programs. Nature immunology, 17)

[24] (Engel, I., Seumois, G., Chavez, L., Samaniego-Castruita, D., White, B., Chawla, A., Mock, D., Vijayanand, P., & Kronenberg, M. (2016). Innate-like functions of natural killer T cell subsets result from highly divergent gene programs. Nature immunology, 17)

[25] (Deng, Q., Ramsköld, D., Reinius, B., & Sandberg, R. (2014). Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. Science (New York, N.Y.), 343(6167), 193–196. https://doi.org/10.1126/science.1245316)

[26] (Treutlein, B., Brownfield, D. G., Wu, A. R., Neff, N. F., Mantalas, G. L., Espinoza, F. H., Desai, T. J., Krasnow, M. A., & Quake, S. R. (2014). Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. Nature, 509(7500),)