

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
Σχολή Χρηματοοικονομικής και Στατιστικής



Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

ΜΕΤΡΑ ΑΠΟΣΤΑΣΗΣ ΚΑΙ ΟΜΟΙΟΤΗΤΑΣ
ΣΕ ΠΟΛΥΔΙΑΣΤΑΤΟΥΣ ΧΩΡΟΥΣ ΚΑΙ Η
ΧΡΗΣΗ ΤΟΥΣ ΣΕ ΠΡΟΒΛΗΜΑΤΑ
ΟΜΑΔΟΠΟΙΗΣΗΣ ΠΟΛΥΔΙΑΣΤΑΤΩΝ
ΔΕΔΟΜΕΝΩΝ

Τσάλας Ανδρέας

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,
Σεπτέμβριος 2026

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
Σχολή Χρηματοοικονομικής και Στατιστικής



Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

ΜΕΤΡΑ ΑΠΟΣΤΑΣΗΣ ΚΑΙ ΟΜΟΙΟΤΗΤΑΣ
ΣΕ ΠΟΛΥΔΙΑΣΤΑΤΟΥΣ ΧΩΡΟΥΣ ΚΑΙ Η
ΧΡΗΣΗ ΤΟΥΣ ΣΕ ΠΡΟΒΛΗΜΑΤΑ
ΟΜΑΔΟΠΟΙΗΣΗΣ ΠΟΛΥΔΙΑΣΤΑΤΩΝ
ΔΕΔΟΜΕΝΩΝ

Τσάλας Ανδρέας

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,
Σεπτέμβριος 2026

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

Τα μέλη της Επιτροπής ήταν:

- Καθηγητής Μ. Κούτρας
- Καθηγητής Γ. Τζαβελάς
- Αναπληρωτής Καθηγητής Αθ. Ρακιτζής

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

UNIVERSITY OF PIRAEUS
School of Finance and Statistics



Department of Statistics and Insurance Science

**POSTGRADUATE PROGRAM IN
APPLIED STATISTICS**

**DISTANCE AND SIMILARITY MEASURES
IN MULTIDIMENSIONAL SPACES AND
THEIR USE IN MULTIDIMENSIONAL DATA
CLUSTERING PROBLEMS**

By
Andreas G. Tsalas

MSc Dissertation

submitted to the Department of Statistics and Insurance Science, University of Piraeus, in
partial fulfillment of the requirements for the award of the Master of Science (M.Sc.) in
Applied Statistics.

Piraeus, Greece
September 2026

*Στους γονείς μου
Γεώργιο και Ελευθερία*

Ευχαριστίες

Θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες στον επιβλέποντα καθηγητή μου **ΜΑΡΚΟ ΚΟΥΤΡΑ** για την πολύτιμη καθοδήγηση, τις εύστοχες παρατηρήσεις και τη συνεχή υποστήριξη που μου παρείχε καθ' όλη τη διάρκεια της εκπόνησης της παρούσας διπλωματικής εργασίας.

Επιπλέον, θα ήθελα να ευχαριστήσω όλα τα μέλη του διδακτικού προσωπικού του Τμήματος Μαθηματικών του Πανεπιστημίου Θεσσαλίας στη Λαμία για τις γνώσεις, την επιστημονική κατάρτιση και τα πολύτιμα εφόδια που μου προσέφεραν κατά τη διάρκεια των προπτυχιακών μου σπουδών.

Επίσης, ευχαριστώ όλους όσους συνέβαλαν, άμεσα ή έμμεσα, στην ολοκλήρωση της εργασίας αυτής, προσφέροντας τη βοήθεια και τη στήριξή τους.

Τέλος, εκφράζω την ειλικρινή μου ευγνωμοσύνη προς την οικογένειά μου και τους φίλους μου για την αμέριστη αγάπη, την υπομονή, την κατανόηση και τη συνεχή ενθάρρυνσή τους. Η στήριξή τους αποτέλεσε πολύτιμη δύναμη και πηγή έμπνευσης σε όλη τη διάρκεια των σπουδών μου και της εκπόνησης της παρούσας διπλωματικής εργασίας.

Περίληψη

Η ανάλυση κατά συστάδες αποτελεί έναν από τους σημαντικότερους κλάδους της Στατιστικής και της Επιστήμης των Δεδομένων, με εφαρμογές σε πλήθος επιστημονικών και τεχνολογικών πεδίων. Στόχος της είναι η ανακάλυψη υποκείμενης δομής και προτύπων, μέσα σε σύνολα δεδομένων, μέσω της ομαδοποίησης παρατηρήσεων που παρουσιάζουν ομοιότητες και της διάκρισης εκείνων που εμφανίζουν διαφοροποιήσεις. Καθοριστικό ρόλο σε αυτή τη διαδικασία διαδραματίζει η έννοια της απόστασης ή της ομοιότητας, καθώς το μέτρο που επιλέγεται για τον υπολογισμό της εγγύτητας μεταξύ αντικειμένων καθορίζει άμεσα το αποτέλεσμα και την ποιότητα της ομαδοποίησης.

Η επιλογή διαφορετικών μετρικών αποστάσεων δύναται να οδηγήσει σε ουσιωδώς διαφορετικά αποτελέσματα, ακόμη και όταν εφαρμόζεται ο ίδιος αλγόριθμος clustering στο ίδιο σύνολο δεδομένων. Η διεθνής βιβλιογραφία περιγράφει μια πληθώρα εναλλακτικών μετρικών, από τις κλασικές Ευκλείδειες αποστάσεις έως πιο εξειδικευμένα μέτρα, όπως η απόσταση Mahalanobis ή οι δείκτες ομοιότητας τύπου Jaccard και Cosine. Κάθε μέτρο παρουσιάζει θεωρητικά πλεονεκτήματα και μειονεκτήματα, καθώς και διαφορετικό πεδίο εφαρμογής, γεγονός που καθιστά τη μελέτη και σύγκρισή τους αναγκαία για την κατανόηση της συμπεριφοράς τους σε πολυδιάστατους χώρους δεδομένων.

Μέσα στο πλαίσιο αυτό, η παρούσα διπλωματική εργασία εστιάζει σε τρεις άξονες. Στη συστηματική παρουσίαση των σημαντικότερων μετρικών αποστάσεων και στα Μέτρα ομοιότητας που χρησιμοποιούνται στη βιβλιογραφία, με αναφορά στις μαθηματικές τους ιδιότητες και στις θεωρητικές σχέσεις που τις συνδέουν. Στην ανάλυση της σχέσης μεταξύ απόστασης και ομοιότητας, καθώς και της τεχνικής μετάβασης από το ένα μέτρο στο άλλο, με ιδιαίτερη έμφαση στις εφαρμογές που συναντώνται σε τεχνικές μηχανικής μάθησης και εξόρυξης δεδομένων. Στην πειραματική αξιολόγηση διαφορετικών μετρικών μέσω της εφαρμογής αλγορίθμων clustering σε πολυδιάστατα σύνολα δεδομένων, ώστε να αποτιμηθεί η αποτελεσματικότητά τους σε πραγματικά σενάρια.

Η συνδυαστική θεώρηση θεωρίας και πράξης αναμένεται να προσφέρει μια ολοκληρωμένη εικόνα για τον ρόλο που διαδραματίζει η επιλογή μετρικής απόστασης στη διαδικασία της ομαδοποίησης. Με τον τρόπο αυτό, η εργασία στοχεύει όχι μόνο να φωτίσει τις θεωρητικές πτυχές του προβλήματος, αλλά και να προτείνει πρακτικές κατευθύνσεις για την ορθολογική επιλογή μετρικών ανάλογα με τα χαρακτηριστικά των δεδομένων και τις απαιτήσεις της εκάστοτε εφαρμογής.

Abstract

Cluster Analysis is one of the most important branches of Statistics and Data Science, with applications across a wide range of scientific and technological fields. Its primary objective is to discover underlying structures or patterns within datasets by grouping together observations that exhibit similarities while separating those that display significant differences. A central concept in this process is the notion of distance or similarity, as the measure selected to quantify the proximity between objects directly determines the outcome and quality of the clustering process.

The choice of different distance metrics may lead to substantially different clustering results, even when the same clustering algorithm is applied to the same dataset. The literature presents a wide variety of alternative metrics, ranging from classical Euclidean distances to more specialized measures such as the Mahalanobis distance and similarity indices such as the Jaccard and Cosine measures. Each metric possesses distinct theoretical advantages and limitations, as well as different areas of applicability, making their study and comparison essential for understanding their behavior in high-dimensional data spaces.

Within this context, the present master's thesis focuses on three main objectives. First, it provides a systematic presentation of the most widely used distance metrics and similarity measures found in the literature, including their mathematical properties and the theoretical relationships that connect them. Second, it investigates the relationship between distance and similarity, as well as the methodologies for transforming one type of measure into the other, with particular emphasis on applications in machine learning and data mining techniques. Third, it experimentally evaluates different distance metrics by applying clustering algorithms to multidimensional datasets to assess their effectiveness under realistic scenarios.

The combined consideration of theoretical foundations and practical applications is expected to provide a comprehensive understanding of the role that the choice of a distance metric plays in the clustering process. In this way, the thesis aims not only to highlight the theoretical aspects of the problem but also to propose practical guidelines for the appropriate selection of distance metrics according to the characteristics of the data and the specific requirements of each application.

Περιεχόμενα

Περίληψη	X
Abstract.....	XII
ΚΕΦΑΛΑΙΟ 1.....	1
Εισαγωγή.....	1
1.1 Μέτρα ομοιότητας και απόστασης: βασικές έννοιες και ορισμοί	1
1.2 Σκοπός και αντικείμενο της εργασίας	2
1.3 Η σημασία της Διπλωματικής εργασίας, των μέτρων απόστασης και ομοιότητας	3
1.4 Ιστορική εξέλιξη.....	4
ΚΕΦΑΛΑΙΟ 2.....	7
Μέτρα απόστασης	7
2.1 Ευκλείδεια απόσταση	7
2.2 Τυποποιημένη ευκλείδεια απόσταση.....	8
2.3 Απόσταση Mahalanobis	9
2.4 Απόσταση Manhattan.....	9
2.5 Απόσταση Minkowski.....	10
2.6 L_1 απόσταση και L_∞ απόσταση ή Chebyshev απόσταση	12
2.7 Απόσταση Gower	13
2.8 Απόσταση Soergel	15
2.9 Απόσταση Kulczynski.....	17
2.10 Απόσταση Bhattacharyya.....	18
2.11 Απόσταση Hellinger	19
2.12 Απόσταση Canberra	19
2.13 Απόσταση με ελλείπουσες τιμές.....	21
ΚΕΦΑΛΑΙΟ 3.....	23
Μέτρα ομοιότητας	23
3.1 Εισαγωγή.....	23
3.2 Ο δειγματικός συντελεστής συσχέτισης	24
3.3 Ο συντελεστής Spearman	26
3.4 Μέτρα ομοιότητας.....	29
3.4.1. Μέτρο ομοιότητας Jaccard	30
3.4.2. Μέτρο ομοιότητας Rogers - Tanimoto	31
3.4.3. Μέτρο ομοιότητας Sokal and Sneath.....	32
3.4.4. Μέτρο ομοιότητας Simple Matching	33

3.4.5 Μέτρο ομοιότητας Sørensen - Dice.....	34
3.4.6. Μέτρο ομοιότητας Russell-Rao	35
3.4.7. Μέτρο ομοιότητας Kulczynski	37
3.5 Συμπεράσματα	38
ΚΕΦΑΛΑΙΟ 4.....	39
Μέθοδοι ομαδοποίησης και σύγχρονες προσεγγίσεις στη μηχανική μάθηση	39
4.1 Ιεραρχικές μέθοδοι	39
4.1.1 Συσσωρευτικές μέθοδοι.....	39
4.1.2 Διαιρετικές μέθοδοι.....	42
4.2 Μη Ιεραρχικές μέθοδοι.....	44
4.2.1 K-Means και K-Medoids	45
4.2.2 Self - Organizing Maps (SOM).....	45
4.3 Πυκνωτικές και βασισμένες σε πυρήνες μέθοδοι	46
4.3.1 DBSCAN και OPTICS.....	47
4.3.2 Mean Shift και Gaussian Mixture Models	47
4.4 Αλγόριθμοι ομαδοποίησης στη μηχανική μάθηση	48
4.4.1 Χρήση αποστάσεων σε Supervised και Unsupervised Learning	49
4.4.2 Clustering με Embeddings και νευρωνικά δίκτυα	49
4.5 Σύνοψη και θεωρητικές συγκρίσεις τεχνικών	50
ΚΕΦΑΛΑΙΟ 5.....	53
Εισαγωγή στο πειραματικό μέρος με σύνθετα δεδομένα	53
5.1 Δημιουργία συνθετικών δεδομένων	53
5. 2 Εφαρμογή αλγορίθμων ομαδοποίησης με διάφορες αποστάσεις	54
5.2.1 Σενάριο 1: τμηματοποίησης νοικοκυριών με δεδομένα μικτού τύπου.....	57
5.2.2 Σενάριο 2: ομαδοποίηση δυαδικών δεδομένων με απόσταση Jaccard	58
5.2.3 Σενάριο 3: Ομαδοποίηση χωρών βάσει χρονικών σειρών με απόσταση Pearson	59
5.2.4 Σενάριο 4: μικτά δεδομένα με ελλείπουσες τιμές και απόσταση Gower (Ιεραρχικό clustering).....	60
5. 3 Αποτελέσματα σύγκρισης και ανάλυση	63
5.4 Εφαρμογές σε οικονομικά σενάρια, επιλογή μέτρου απόστασης και αξιολόγηση clustering	66
5.5 Συγκεντρωτική ανάλυση και σύγκριση μεθόδων.....	68
5.6 Ευαισθησία και ανθεκτικότητα των μεθόδων Clustering	69
5.7 Οπτικοποίηση και ερμηνεία σενάριο 6 (Investors / Portfolios)	71
5.8 Έλεγχος σταθερότητας με επαναλαμβανόμενη υποδειγματοληψία	72

5.9 Συμπεράσματα	74
ΚΕΦΑΛΑΙΟ 6.....	77
Συμπεράσματα	77
6.1 Ανακεφαλαίωση της εργασίας	77
6.2 Κύρια Ερευνητικά Ευρήματα	77
6.3 Θεωρητική συνεισφορά της εργασίας.....	78
6.4 Πρακτικές κατευθυντήριες γραμμές για την επιλογή μέτρου απόστασης.....	79
6.5 Μελλοντικές κατευθύνσεις έρευνας	79
6.6 Τελικό συμπέρασμα	80
7. Παράρτημα	83
8. Βιβλιογραφία.....	99

ΚΕΦΑΛΑΙΟ 1

Εισαγωγή

1.1 Μέτρα ομοιότητας και απόστασης: βασικές έννοιες και ορισμοί

Οι μετρήσεις ομοιότητας και απόστασης αποτελούν βασικό εργαλείο σε πολλούς τομείς της επιστήμης δεδομένων, όπως η μηχανική μάθηση, η τεχνητή νοημοσύνη, η εξόρυξη δεδομένων, η αναγνώριση προτύπων, η διαχείριση βάσεων δεδομένων και ιδιαίτερα η στατιστική. Χρησιμοποιούνται σε πληθώρα εφαρμογών, όπως η πρόβλεψη, η ανίχνευση συσχετίσεων, η αναγνώριση και η ομαδοποίηση δεδομένων.

Η επιλογή ενός κατάλληλου μέτρου ομοιότητας ή απόστασης για μια συγκεκριμένη εφαρμογή είναι ιδιαίτερα σημαντική. Για τον λόγο αυτό παρουσιάζεται ένα σύνολο ευρέως χρησιμοποιούμενων μέτρων ομοιότητας που εφαρμόζονται σε διάφορους τομείς της επιστήμης των δεδομένων. Προκειμένου να διαμορφωθεί μια ολοκληρωμένη και συστηματική ταξινόμηση, μελετήθηκαν σχετικές επιστημονικές έρευνες και βιβλιογραφικές πηγές. Από τη μελέτη αυτή προκύπτει ότι τα μέτρα απόστασης σε πολυδιάστατους χώρους είναι πολλά και διαφοροποιούνται ως προς τη χρήση και τις ιδιότητές τους.

Τα μέτρα ομοιότητας και τα μέτρα απόστασης αποτελούν στενά συνδεδεμένες έννοιες. Στα μέτρα ομοιότητας, όσο μεγαλύτερη είναι η ομοιότητα μεταξύ δύο αντικειμένων, τόσο υψηλότερη είναι η τιμή του μέτρου. Αντίθετα, στα μέτρα απόστασης η τιμή μειώνεται όσο αυξάνεται η ομοιότητα, ενώ τιμή μηδέν υποδηλώνει πλήρη ταύτιση των αντικειμένων. Στη μαθηματική βιβλιογραφία, ο όρος μετρική (metric) χρησιμοποιείται με συγκεκριμένο μαθηματικό ορισμό, ο οποίος περιγράφει τις ιδιότητες που πρέπει να ικανοποιεί μια συνάρτηση ώστε να θεωρείται μέτρο απόστασης.

Ορισμός 1.1. Μια μη αρνητική συνάρτηση απόστασης d στο σύνολο X , $d: X \times X \rightarrow \mathbb{R}$, ονομάζεται *μετρική* αν για όλα τα $x, y, z \in X$ ικανοποιεί τις ακόλουθες συνθήκες:

- i. $d(x, y) = 0$ αν και μόνο αν $x = y$
- ii. $d(x, y) \geq 0$
- iii. $d(x, y) = d(y, x)$ (Συμμετρία)
- iv. $d(x, y) \leq d(x, z) + d(z, y)$ (Τριγωνική ανισότητα)

Η έννοια της μετρικής μέσω αυτών των βασικών ιδιοτήτων επιτρέπει την περιγραφή διαφορετικών μορφών απόστασης μεταξύ αντικειμένων, ενώ παράλληλα, παρέχει το θεωρητικό πλαίσιο για τη μελέτη των μετρικών χώρων. Ωστόσο, αρκετά από τα μέτρα απόστασης που χρησιμοποιούνται σε πρακτικές εφαρμογές δεν ικανοποιούν πλήρως τις ιδιότητες μιας μετρικής.

1.2 Σκοπός και αντικείμενο της εργασίας

Η παρούσα διπλωματική εργασία εντάσσεται στο πεδίο της πολυμεταβλητής στατιστικής ανάλυσης και της επιστήμης δεδομένων, με έμφαση στη μελέτη των μέτρων απόστασης και ομοιότητας σε πολυδιάστατους χώρους. Η αποτίμηση της εγγύτητας μεταξύ παρατηρήσεων αποτελεί βασικό στοιχείο της στατιστικής ανάλυσης, καθώς επηρεάζει άμεσα διαδικασίες όπως η ταξινόμηση, η ομαδοποίηση και η αναγνώριση προτύπων.

Σκοπός της εργασίας είναι η παρουσίαση, η ανάλυση και σύγκριση βασικών μέτρων απόστασης και ομοιότητας, προκειμένου να αναδειχθεί ο ρόλος τους στη διαμόρφωση της δομής των δεδομένων και στη λήψη στατιστικών αποφάσεων. Οι αποστάσεις δεν αποτελούν απλώς υπολογιστικά εργαλεία, αλλά ενσωματώνουν υποθέσεις σχετικά με τη γεωμετρία του χώρου των δεδομένων.

Ιδιαίτερη έμφαση δίνεται στη σχέση μεταξύ απόστασης και ομοιότητας, οι οποίες αποτελούν αλληλοσυμπληρούμενες έννοιες. Η απόσταση εκφράζει τον βαθμό διαφοροποίησης μεταξύ δύο αντικειμένων, ενώ η ομοιότητα αποτυπώνει το επίπεδο κοινών χαρακτηριστικών τους. Οι δύο έννοιες συνδέονται με αντίστροφη σχέση και μπορούν να μετασχηματίζονται μεταξύ τους μέσω κατάλληλων μαθηματικών προσεγγίσεων. Παράλληλα, εξετάζεται ο ρόλος των μέτρων απόστασης σε διαδικασίες ομαδοποίησης (clustering), μία από τις βασικές τεχνικές ανάλυσης χωρίς επίβλεψη στη μηχανική μάθηση. Στις μεθόδους αυτές, το μέτρο απόστασης καθορίζει ποια αντικείμενα θεωρούνται εγγύτερα και συνεπώς ανήκουν στην ίδια ομάδα. Η επιλογή διαφορετικής μετρικής μπορεί να επηρεάσει σημαντικά τη μορφή των ομάδων και τα συμπεράσματα που προκύπτουν από την ανάλυση. Η εργασία λαμβάνει επίσης υπόψη τη φύση των δεδομένων, τα οποία σε πολλές περιπτώσεις περιλαμβάνουν ποσοτικές, κατηγορικές ή μεικτές μεταβλητές. Για τον λόγο αυτό εξετάζονται μετρικές που μπορούν να προσαρμοστούν σε διαφορετικούς τύπους δεδομένων καθώς και η σημασία της κανονικοποίησης και της στάθμισης των μεταβλητών. Η προεπεξεργασία των δεδομένων είναι ιδιαίτερα σημαντική, διότι η απόσταση επηρεάζεται από την κλίμακα και τη διακύμανση των χαρακτηριστικών.

Επιπλέον, αναλύεται η καταλληλότητα διαφορετικών μετρικών ανάλογα με τη δομή των δεδομένων. Για παράδειγμα, σε περιπτώσεις όπου οι μεταβλητές παρουσιάζουν ισχυρές συσχετίσεις, μέτρα όπως η Mahalanobis μπορούν να αποδώσουν πιο ακριβή γεωμετρική αναπαράσταση σε σύγκριση με την ευκλείδεια απόσταση. Αντίθετα, όταν τα δεδομένα είναι κανονικοποιημένα και σχετικά ανεξάρτητα, η ευκλείδεια απόσταση συχνά αποτελεί επαρκή και απλή επιλογή.

Συνολικά, η εργασία επιδιώκει να αναδείξει ότι τα μέτρα απόστασης και ομοιότητας δεν αποτελούν μόνο τεχνικά εργαλεία υπολογισμού, αλλά θεμελιώδεις έννοιες για την κατανόηση της δομής των δεδομένων. Η σωστή επιλογή τους συμβάλλει στην αξιόπιστη ανάλυση, στην ερμηνεία των σχέσεων μεταξύ παρατηρήσεων και στην εξαγωγή έγκυρων στατιστικών συμπερασμάτων.

1.3 Η σημασία της Διπλωματικής εργασίας, των μέτρων απόστασης και ομοιότητας

Ως προς τον χαρακτήρα της, η παρούσα Διπλωματική Εργασία είναι θεωρητικού και εφαρμοσμένου-πειραματικού χαρακτήρα. Το θεωρητικό μέρος περιλαμβάνει τη συστηματική παρουσίαση και συγκριτική εξέταση βασικών μέτρων απόστασης και ομοιότητας, καθώς και της σχέσης τους με μεθόδους ομαδοποίησης, ενώ το εφαρμοσμένο μέρος βασίζεται σε ελεγχόμενα πειράματα με συνθετικά δεδομένα και σενάρια εμπνευσμένα από οικονομικές εφαρμογές. Η ιδιαίτερη συμβολή της εργασίας έγκειται στη συνδυαστική διερεύνηση της αλληλεπίδρασης μεταξύ της φύσης των δεδομένων, της επιλεγμένης μετρικής και του αλγορίθμου ομαδοποίησης, καθώς και στην εμπειρική αξιολόγηση της επίδρασης των επιλογών αυτών στην ποιότητα και τη σταθερότητα των παραγόμενων ομαδοποιήσεων. Μέσω της συγκριτικής αυτής προσέγγισης επιδιώκεται η εξαγωγή τεκμηριωμένων συμπερασμάτων και πρακτικών κατευθύνσεων για την επιλογή κατάλληλου μέτρου απόστασης ή ομοιότητας ανάλογα με τα χαρακτηριστικά και τη δομή των υπό ανάλυση δεδομένων.

Η έννοια της απόστασης και της ομοιότητας κατέχει κεντρική θέση στη στατιστική ανάλυση και στην επιστήμη των δεδομένων, καθώς επιτρέπει την ποσοτικοποίηση των σχέσεων μεταξύ αντικειμένων, παρατηρήσεων ή μεταβλητών.

Στα μαθηματικά και στη στατιστική, η απόσταση δεν περιορίζεται στη φυσική ή γεωμετρική της έννοια, αλλά αποτελεί εργαλείο αναπαράστασης αφηρημένων σχέσεων σε πολυδιάστατους χώρους. Μέσω κατάλληλων μετρικών, οι πληροφορίες μετατρέπονται σε συγκρίσιμες ποσότητες ώστε να καθίσταται δυνατή η ανάλυση της εγγύτητας, της διαφοράς και της δομής των δεδομένων. Η σημασία των μέτρων απόστασης έγκειται ακριβώς στο ότι οργανώνουν τον χώρο των παρατηρήσεων και επιτρέπουν τη συστηματική μελέτη του.

Η επιλογή της μετρικής δεν είναι ουδέτερη, αλλά εξαρτάται από τη φύση των δεδομένων, την κλίμακα μέτρησης και το εκάστοτε ερευνητικό πλαίσιο. Διαφορετικά μέτρα απόστασης αποδίδουν διαφορετική γεωμετρία στον χώρο των δεδομένων και συνεπώς οδηγούν σε διαφορετική ερμηνεία των σχέσεων μεταξύ των παρατηρήσεων. Για τον λόγο αυτό, η απόσταση δεν πρέπει να θεωρείται απόλυτο μέγεθος, αλλά μέγεθος σχετικό, το οποίο αποκτά νόημα μόνο μέσα στο πλαίσιο εφαρμογής του.

Η μαθηματική θεμελίωση της απόστασης βασίζεται στην έννοια της μετρικής. Μια συνάρτηση απόστασης θεωρείται μετρική όταν ικανοποιεί βασικές ιδιότητες, όπως η μη αρνητικότητα, η ταυτότητα του αδιάκριτου, η συμμετρία και η ανισότητα του τριγώνου. Οι ιδιότητες αυτές εξασφαλίζουν ότι η μέτρηση της διαφοροποίησης είναι λογικά συνεπής και ερμηνεύσιμη. Όταν ένα σύνολο δεδομένων εφοδιάζεται με μια τέτοια συνάρτηση, συγκροτεί έναν μετρικό χώρο, μέσα στον οποίο η έννοια της εγγύτητας αποκτά σαφές μαθηματικό περιεχόμενο.

Η ομοιότητα αποτελεί τη συμπληρωματική όψη της απόστασης. Ενώ η απόσταση αναδεικνύει τη διαφορά, η ομοιότητα αποτυπώνει το κοινό στοιχείο μεταξύ δύο αντικειμένων. Στη στατιστική ανάλυση, η ομοιότητα είναι ιδιαίτερα σημαντική, επειδή επιτρέπει την

αναγνώριση προτύπων, τη συγκρότηση κατηγοριών και την αποκάλυψη υποκείμενων δομών μέσα στα δεδομένα.

Η θεωρητική τους σημασία υπερβαίνει τον καθαρά υπολογιστικό τους ρόλο. Τα μέτρα απόστασης και ομοιότητας λειτουργούν ως βασικά εργαλεία οργάνωσης και ερμηνείας της πληροφορίας. Δεν αποτυπώνουν απλώς σχέσεις που ήδη υπάρχουν στα δεδομένα. Υπό αυτή την έννοια, κάθε μετρική ενσωματώνει συγκεκριμένες θεωρητικές παραδοχές και συνιστά ταυτόχρονα εργαλείο μέτρησης και ερμηνευτικό φίλτρο.

Συνολικά, τα μέτρα απόστασης και ομοιότητας αποτελούν θεμελιώδεις έννοιες της στατιστικής. Σε μεθοδολογικό επίπεδο οργανώνουν και αναπαριστούν την πληροφορία, σε ερμηνευτικό επίπεδο καθορίζουν τον τρόπο με τον οποίο προσεγγίζονται οι σχέσεις μεταξύ των δεδομένων, και σε γνωσιολογικό επίπεδο συμβάλλουν στη μετατροπή της εμπειρικής παρατήρησης σε συστηματική γνώση. Επομένως, η κατανόησή τους είναι αναγκαία προϋπόθεση για κάθε αξιόπιστη ανάλυση δεδομένων.

1.4 Ιστορική εξέλιξη

α. Η συμβολή του Fisher (1936) στη θεμελίωση των μέτρων απόστασης και ομοιότητας.

Ένα από τα πρώτα και πιο θεμελιώδη έργα που ανέδειξαν τη σημασία των μέτρων απόστασης και ομοιότητας στην πολυδιάστατη στατιστική ήταν το άρθρο του Ronald A. Fisher (1936), “*The Use of Multiple Measurements in Taxonomic Problems*”. Στην εργασία αυτή, ο Fisher εισήγαγε τη μέθοδο της διακριτικής ανάλυσης (Discriminant Analysis), με σκοπό τη διάκριση και ταξινόμηση ειδών φυτών του γένους *Iris*, χρησιμοποιώντας τέσσερις ποσοτικές μετρήσεις των ανθών τους. Η καινοτομία της μελέτης έγκειται στο ότι για πρώτη φορά εφαρμόστηκε η ιδέα ότι ένας γραμμικός συνδυασμός πολλαπλών χαρακτηριστικών μπορεί να αξιοποιηθεί για να μεγιστοποιήσει τη διακριτική ικανότητα μεταξύ διαφορετικών πληθυσμών. Ο Fisher έδειξε ότι η χρήση μιας συνάρτησης, η οποία υπολογίζεται ως γραμμικός συνδυασμός των μετρήσεων, μπορεί να οδηγήσει σε βέλτιστο διαχωρισμό μεταξύ των ομάδων, ελαχιστοποιώντας ταυτόχρονα τα σφάλματα ταξινόμησης.

Το παράδειγμα με τα δεδομένα *Iris*, αποτέλεσε ορόσημο όχι μόνο για τη στατιστική, αλλά και για την επιστήμη των δεδομένων ευρύτερα. Μέχρι σήμερα, το συγκεκριμένο σύνολο δεδομένων χρησιμοποιείται εκτενώς ως κλασικό benchmark για την αξιολόγηση αλγορίθμων μηχανικής μάθησης και τεχνικών ομαδοποίησης.

Ο Fisher, στο κλασικό του άρθρο, χρησιμοποίησε ως παράδειγμα τρία είδη φυτών του γένους *Iris* (*setosa*, *versicolor*, *virginica*). Από κάθε είδος συλλέχθηκαν 50 παρατηρήσεις, καταγράφοντας τέσσερις διαστάσεις: μήκος και πλάτος σεφάλου, καθώς και μήκος και πλάτος πετάλου. Στόχος ήταν να βρεθεί μια γραμμική συνάρτηση της μορφής

$$Z = a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4$$

η οποία θα μεγιστοποιεί τον λόγο της μεταξύ ομάδων διακύμανσης προς την εντός ομάδων διακύμανση. Με αυτήν τη διαδικασία, ο Fisher κατέληξε σε μια συνάρτηση που παρείχε σαφή διάκριση μεταξύ του *Iris setosa* και των άλλων δύο ειδών, ενώ μεταξύ *versicolor* και *virginica* η διάκριση ήταν πιο δύσκολη, γεγονός που αντικατοπτρίζει και τη βιολογική τους συγγένεια.

Από θεωρητικής άποψης, η μεθοδολογία του Fisher συνδέεται άμεσα με τα μέτρα απόστασης. Η διακριτική συνάρτηση που πρότεινε στηρίζεται στην ελαχιστοποίηση της διακύμανσης εντός των ομάδων και στη μεγιστοποίηση της διακύμανσης μεταξύ των ομάδων, κάτι που ουσιαστικά συνεπάγεται την επιλογή ενός κατάλληλου μέτρου εγγύτητας για τη βέλτιστη διάκριση. Με τον τρόπο αυτό, το έργο του Fisher έθεσε τις βάσεις για τη σύγχρονη αντίληψη ότι η επιλογή μετρικής δεν είναι απλώς τεχνικό ζήτημα, αλλά καθορίζει σε μεγάλο βαθμό την ποιότητα των αποτελεσμάτων σε ταξινομητικές και ομαδοποιητικές διαδικασίες.

Συνεπώς, το έργο του Fisher μπορεί να θεωρηθεί ως πρώτο θεμέλιο στην κατανόηση της σχέσης ανάμεσα σε πολυδιάστατες μετρήσεις και μέτρα ομοιότητας - απόστασης, και ως προάγγελος των σύγχρονων μεθόδων μηχανικής μάθησης και εξόρυξης δεδομένων.

β. Η συμβολή του Mahalanobis (1936) και η Απόσταση Mahalanobis

Την ίδια χρονιά με τη θεμελιώδη εργασία του Fisher, ο Prasanta Chandra Mahalanobis παρουσίασε στο άρθρο του “*On the Generalized Distance in Statistics*” (1936) μία νέα μετρική που έμελλε να αποτελέσει σημείο αναφοράς για την πολυμεταβλητή στατιστική ανάλυση. Η απόσταση Mahalanobis εισάγει έναν γενικευμένο τρόπο μέτρησης της απόστασης μεταξύ ενός διανύσματος παρατηρήσεων και του μέσου όρου ενός πληθυσμού, λαμβάνοντας υπόψη τόσο την διακύμανση κάθε μεταβλητής όσο και τις συσχετίσεις μεταξύ τους.

Η βασική ιδέα ήταν ότι η κλασική ευκλείδεια απόσταση δεν είναι επαρκής όταν τα δεδομένα είναι πολυδιάστατα και παρουσιάζουν ετερογένεια στις κλίμακες ή σημαντική συσχέτιση μεταξύ μεταβλητών. Για τον λόγο αυτό, ο Mahalanobis πρότεινε τον ακόλουθο τύπο:

$$D(x, \mu) = \sqrt{(x - \mu)^T S^{-1} (x - \mu)}$$

όπου S ο πίνακας συνδιακύμανσης, όπου x το διάνυσμα παρατήρησης και όπου μ διάνυσμα των μέσων τιμών.

Η συγκεκριμένη μετρική έχει τρία καίρια πλεονεκτήματα. Πρώτον, η κανονικοποίηση ως προς τη διακύμανση, δηλαδή οι μεταβλητές με μεγάλη κλίμακα δεν κυριαρχούν τεχνητά στον υπολογισμό της απόστασης. Δεύτερον, η λήψη υπόψη συσχετίσεων, πιο συγκεκριμένα όταν δύο μεταβλητές είναι συσχετισμένες, η συνεισφορά τους στην απόσταση προσαρμόζεται αναλόγως, ώστε να μην υπάρχει διπλή μέτρηση της ίδιας πληροφορίας. Τρίτον, και κυρίως γενικά η μετρική είναι ανεξάρτητη από τη μονάδα μέτρησης και μπορεί να εφαρμοστεί σε οποιονδήποτε πολυδιάστατο χώρο.

Αρχικά, ο Mahalanobis χρησιμοποίησε την απόστασή του για την ταξινόμηση ανθρωπομετρικών δεδομένων, κυρίως κρανιομετρικών μετρήσεων, επιχειρώντας να συγκρίνει πληθυσμούς διαφορετικής προέλευσης. Ωστόσο, η μετρική βρήκε γρήγορα εφαρμογή σε ευρύ φάσμα επιστημών, από τη βιολογία και την κοινωνιολογία έως την οικονομία και αργότερα την αναγνώριση προτύπων και σήμερα στην τεχνητή νοημοσύνη.

Στη σύγχρονη εποχή, η απόσταση Mahalanobis χρησιμοποιείται κυρίως, στην ανίχνευση ακραίων τιμών (outliers), καθώς παρατηρήσεις με μεγάλη Mahalanobis απόσταση θεωρούνται ύποπτες, στην ταξινόμηση (π.χ. μέθοδος του πλησιέστερου κέντρου με Mahalanobis αντί για ευκλείδεια απόσταση), στην ομαδοποίηση (clustering), όπου προσφέρει πιο αξιόπιστη

αποτίμηση ομοιότητας σε δεδομένα με συσχετισμένες μεταβλητές, στη μηχανική μάθηση, για μετρικές σε χώρους χαρακτηριστικών με διαφορετικές κλίμακες.

Η σημασία του άρθρου του Mahalanobis (1936) είναι ανάλογη με εκείνη του Fisher την ίδια χρονιά. Ενώ ο Fisher εισήγαγε τη λογική του βέλτιστου γραμμικού διαχωρισμού πληθυσμών, ο Mahalanobis πρόσφερε ένα γενικευμένο μέτρο απόστασης που επέτρεπε την αξιόπιστη σύγκριση πολυδιάστατων δεδομένων. Και οι δύο συνεισφορές αποτέλεσαν ακρογωνιαίους λίθους για την ανάπτυξη της πολυμεταβλητής στατιστικής και εξακολουθούν να επηρεάζουν σε μεγάλο βαθμό τις εφαρμογές στη στατιστική μάθηση και την εξόρυξη δεδομένων.

ΚΕΦΑΛΑΙΟ 2

Μέτρα απόστασης

2.1 Ευκλείδεια απόσταση

Στο προηγούμενο κεφάλαιο παρουσιάστηκε το θεωρητικό πλαίσιο των μέτρων απόστασης και ομοιότητας. Στην ενότητα αυτή παρουσιάζονται ορισμένα από τα βασικότερα μέτρα απόστασης που χρησιμοποιούνται στην ανάλυση δεδομένων. Η εφαρμογή μεθόδων ομαδοποίησης προϋποθέτει τον καθορισμό κατάλληλων μετρικών, μέσω των οποίων είναι δυνατόν να εκτιμηθεί ο βαθμός ομοιότητας ή διαφοροποίησης μεταξύ των παρατηρήσεων. Η εισαγωγή τέτοιων μέτρων καθιστά εφικτή την ποσοτική αποτίμηση της σχέσης μεταξύ παρατηρήσεων και κατ' επέκταση, τη διάκρισή τους σε κοινές ή διακριτές ομάδες. Συνεπώς, ο στόχος της ανάλυσης κατά συστάδες εστιάζει στη διαμόρφωση συστάδων με υψηλό εσωτερικό βαθμό ομοιογένειας (μικρές αποστάσεις εντός ομάδας) και σαφή διαχωρισμό μεταξύ διαφορετικών συστάδων (μεγάλες αποστάσεις μεταξύ ομάδων). Το πρώτο μέτρο απόστασης που εξετάζεται είναι η ευκλείδεια απόσταση, η οποία αποτελεί το πιο γνωστό και ευρέως χρησιμοποιούμενο μέτρο απόστασης.

Η ευκλείδεια απόσταση δίνεται από τον τύπο:

$$d_{ij} = d(x_i, x_j) = \sqrt{\sum_{r=1}^p (x_{ir} - x_{jr})^2}$$

Η ευκλείδεια απόσταση ικανοποιεί τις βασικές ιδιότητες μιας μετρικής, όπως παρουσιάστηκαν στο προηγούμενο κεφάλαιο. Δηλαδή τις:

- i. $d_{ij} \geq 0$ για κάθε i, j & $d_{ij} = 0$ για κάθε $i = j$
- ii. $d_{ij} \leq d_{is} + d_{sj}$
- iii. $d_{ij} = d_{ji}$

Παρότι αποτελεί το απλούστερο και συχνότερα χρησιμοποιούμενο μέτρο, παρουσιάζει τα ακόλουθα μειονεκτήματα:

- Εξάρτηση από την κλίμακα μέτρησης, δηλαδή αλλαγή μονάδων ή εύρους μιας μεταβλητής μεταβάλλει τις αποστάσεις και μπορεί να αλλοιώσει τη δομή των συστάδων.
- Κυριαρχία μεταβλητών με μεγάλη διασπορά ή απόλυτα μεγέθη. Δηλαδή, οι μεταβλητές μεγάλης κλίμακας υπερκαλύπτουν μεταβλητές μικρής κλίμακας στο άθροισμα των τετραγώνων, με αποτέλεσμα η απόσταση να καθορίζεται σχεδόν αποκλειστικά από αυτές.
- Και τέλος η ευκλείδεια απόσταση είναι κατάλληλη για συνεχείς ποσοτικές μεταβλητές, ενώ για κατηγορικές - δυαδικές μεταβλητές απαιτούνται διαφορετικά μέτρα.

2.2 Τυποποιημένη ευκλείδεια απόσταση

Η ευκλείδεια απόσταση δεν μοιάζει να είναι ένα ιδανικό και κατάλληλο μέτρο απόστασης, αφού είναι ευαίσθητη στην κλίμακα των μεταβλητών. Συνεπώς, όταν οι μεταβλητές μετρώνται σε διαφορετικές μονάδες, ενδείκνυται η τυποποίηση (z-scores). Για κάθε μεταβλητή, αν συμβολίσουμε με s_r η δειγματική τυπική απόκλιση της r μεταβλητής και έχουμε ότι:

$$\bar{X}_r = \frac{1}{n} \sum_{i=1}^n x_{ir},$$

$$S_r = \left[\frac{1}{n-1} \sum_{i=1}^n (x_{ir} - \bar{x}_r)^2 \right]^{1/2}$$

Με βάση τις παραπάνω ποσότητες, η τυποποιημένη ευκλείδεια απόσταση μεταξύ δύο αντικειμένων x_i και x_j ορίζεται ως,

$$d_{ij} = d(x_i, x_j) = \sqrt{\sum_{r=1}^p \frac{(x_{ir} - x_{jr})^2}{s_r^2}} = \sqrt{\sum_{r=1}^p \left(\frac{x_{ir} - x_{jr}}{s_r} \right)^2}$$

Από την πλευρά της στατιστικής και της εξόρυξης δεδομένων, αυτή η απόσταση είναι αρκετά ενδιαφέρουσα γιατί διαθέτει στατιστικό πλεονέκτημα κάθε μεταβλητή συνεισφέρει στην απόσταση σε μονάδες τυπικής απόκλισης, άρα οι συγκρίσεις δεν επηρεάζονται από τις διαφορετικές κλίμακες μέτρησης. Η τυποποίηση ανά μεταβλητή συνεπάγεται ότι κάθε διάσταση συμβάλλει σε μονάδες τυπικής απόκλισης. Υπό αυτό το πρίσμα αυτό, η απόσταση με το παραπάνω τύπο δεν είναι πάρα ειδική περίπτωση της σταθμισμένης ευκλείδειας, με πίνακα βαρών:

$$W = \text{diag} \left(1/s_1^2, \dots, 1/s_p^2 \right)$$

όπου,

$$w_r = \frac{1}{s_r^2}, \quad r = 1, 2, \dots, p,$$

τότε η σταθμισμένη ευκλείδεια απόσταση γράφεται ως

$$d_W(x_i, x_j) = \sqrt{(x_i - x_j)^T W (x_i - x_j)} = \left(\sum_{r=1}^p w_r (x_{ir} - x_{jr})^2 \right)^{1/2}$$

Θέτοντας $w_r = 1/s_r^2$, προκύπτει αμέσως η τυποποιημένη ευκλείδεια απόσταση. Στην περίπτωση ενός συνόλου n αντικειμένων $x_1, x_2, \dots, x_n$, οι αποστάσεις ανά δύο αντικειμένων υπολογίζονται ως:

$$d_{ij} = d(x_i, x_j), \quad i, j = 1, 2, \dots, n,$$

και οργανώνονται στον πίνακα αποστάσεων

$$D = [d_{ij}]_{n \times n}.$$

Ο πίνακας αποστάσεων έχει τις ακόλουθες βασικές ιδιότητες:

- $d_{ii} = 0$, για κάθε i .
- $d_{ij} \geq 0$, για κάθε i, j .
- $d_{ij} = d_{ji}$, δηλαδή ο πίνακας είναι συμμετρικός.
- Η τυποποιημένη ευκλείδεια απόσταση ικανοποιεί επίσης την ανισότητα του τριγώνου και επομένως αποτελεί μετρική απόσταση.

2.3 Απόσταση Mahalanobis

Παρά τα πλεονεκτήματά τους, οι αποστάσεις που εξετάστηκαν ανωτέρω η ευκλείδεια, τυποποιημένη - σταθμισμένη ευκλείδεια, δεν λαμβάνουν υπόψη τη συσχέτιση μεταξύ μεταβλητών, χειρίζονται την κλίμακα δηλαδή τις διασπορές, αλλά αγνοούν τις συνδιακυμάνσεις. Όταν δύο μεταβλητές είναι ισχυρά συσχετισμένες, η αποτελεσματική διάσταση του χώρου μειώνεται και η απλή ευκλείδεια απόσταση μετρά παρόμοια πληροφορία. Για να ενσωματωθεί η δομή συσχέτισης, χρησιμοποιείται η απόσταση Mahalanobis, η οποία για δύο παρατηρήσεις $x_i, x_j \in \mathbb{R}^p$, ορίζεται ως:

$$d_M(x_i, x_j) = \left[(x_i - x_j)^T \Sigma^{-1} (x_i - x_j) \right]^{1/2},$$

όπου Σ είναι ο συμμετρικός θετικά ορισμένος πίνακας διακύμανσης - συνδιακύμανσης των μεταβλητών, που αντιστοιχεί στα διανύσματα $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ και $x_j = (x_{j1}, x_{j2}, \dots, x_{jp})$. Αν ο πίνακας συνδιακύμανσης είναι διαγώνιος με διαγώνια στοιχεία s_p η απόσταση του Mahalanobis γίνεται ίδια με τη σταθμισμένη ευκλείδεια απόσταση με $w_p = 1/s_p^2$.

2.4 Απόσταση Manhattan

Η απόσταση Manhattan συνιστά θεμελιώδη εναλλακτική της ευκλείδειας απόστασης. Αντί να αθροίζει τετράγωνα αποκλίσεων, αθροίζει απόλυτες αποκλίσεις, είναι συχνά καταλληλότερη σε υψηλές διαστάσεις ή όταν αναμένουμε Αραιές διαφορές μεταξύ παρατηρήσεων. Επιπλέον, μέσω κατάλληλων σταθμίσεων ή τυποποίησης, η Manhattan μπορεί να εξουδετερώσει διαφορές κλίμακας ανά μεταβλητή, παρέχοντας συγκρίσεις σε ενιαίες μονάδες. Στη συνέχεια δίνουμε τον ακριβή ορισμό, αποδεικνύουμε ότι αποτελεί μετρική, και παρουσιάζουμε τις σταθμισμένες - τυποποιημένες εκδοχές της και τη σχέση της με άλλες νόρμες της οικογένειας Minkowski.

$$d_{ij} = \sum_{r=1}^p |x_{ir} - x_{jr}|$$

Πράγματι, η απόσταση Manhattan όπως συνιστά θεμελιώδη εναλλακτική της ευκλείδειας απόστασης, δηλαδή μοιάζει πολύ με την ευκλείδεια απόσταση με τη διαφορά ότι αντί για τετραγωνικές αποκλίσεις χρησιμοποιούμε απόλυτες αποκλίσεις.

Σε αντιστοιχία με τη σταθμισμένη ευκλείδεια απόσταση, μπορεί να οριστεί και η σταθμισμένη απόσταση Manhattan μέσω του τύπου:

$$d_{ij} = \sum_{r=1}^p w_r |x_{ir} - x_{jr}|$$

όπου, w_r , $r = 1, 2, \dots, p$ είναι μη αρνητικά βάρη.

Σύμφωνα με τους Witten, Frank, Hall και Pal (2017), στη μάθηση βασισμένη σε παραδείγματα η επιλογή της συνάρτησης απόστασης καθορίζει την αποτελεσματικότητα της ταξινόμησης. Αν και η ευκλείδεια απόσταση είναι η πιο συνηθισμένη, η απόσταση Manhattan (city-block distance) αποτελεί μία εξίσου θεμελιώδη εναλλακτική. Σε αυτήν, οι διαφορές μεταξύ των τιμών των γνωρισμάτων δεν υψώνονται στο τετράγωνο, αλλά προστίθενται ως απόλυτες τιμές, γεγονός που την καθιστά πιο ανθεκτική σε ακραίες αποκλίσεις. Οι συγγραφείς επισημαίνουν επίσης ότι διαφορετικά γνωρίσματα ενδέχεται να βρίσκονται σε διαφορετικές κλίμακες, γι' αυτό προτείνεται η κανονικοποίηση όλων των τιμών στο διάστημα $[0, 1]$ μέσω του τύπου:

$$u_{ir} = \frac{x_{ir} - x_{min,r}}{x_{max,r} - x_{min,r}} \quad r = 1, 2, \dots, n$$

Με τον τρόπο αυτό, εξασφαλίζεται ότι κανένα γνώρισμα δεν κυριαρχεί στην απόσταση λόγω διαφορετικής μονάδας μέτρησης. Επιπλέον, για ονομαστικά γνωρίσματα (μη αριθμητικά), οι διαφορές λαμβάνονται ως 0 αν οι τιμές είναι ίδιες ή 1 αν διαφέρουν, ενώ για ελλείπουσες τιμές η διαφορά θεωρείται μέγιστη.

2.5 Απόσταση Minkowski

Σύμφωνα με τους Witten, Frank, Hall και Pal (2017), η απόσταση Minkowski αποτελεί γενίκευση της ευκλείδειας απόστασης, καθώς επιτρέπει την ύψωση των διαφορών σε εκθέτη p , με αποτέλεσμα οι μεγαλύτερες τιμές του p να αποδίδουν μεγαλύτερη βαρύτητα στις μεγάλες αποκλίσεις σε σχέση με τις μικρότερες (Witten et al., 2017, σ. 135–136). Η επιλογή του κατάλληλου μέτρου απόστασης είναι ιδιαίτερα σημαντική σε αλγόριθμους μηχανικής μάθησης που βασίζονται στη γειτνίαση, όπως οι k-Nearest Neighbors και οι μέθοδοι ομαδοποίησης, καθώς επηρεάζει άμεσα τον τρόπο σύγκρισης των παρατηρήσεων. Η απόσταση Minkowski προσφέρει ένα ενιαίο και ευέλικτο πλαίσιο μέτρησης αποστάσεων σε πολυδιάστατους χώρους, ιδίως στην περίπτωση συνεχών δεδομένων. Μέσω της παραμέτρου p , ενοποιεί γνωστές μετρικές, όπως την απόσταση Manhattan για $p = 1$ και την ευκλείδεια απόσταση για $p = 2$. Η σημασία της έγκειται στο ότι επιτρέπει την προσαρμογή της μετρικής στις ιδιότητες των δεδομένων, διαμορφώνοντας ένα γενικό πλαίσιο ανάλυσης αποστάσεων. Αντίθετα, η απόσταση Hamming, αν και σχετίζεται με τη μέτρηση διαφορών μεταξύ διανυσμάτων, εφαρμόζεται κυρίως σε δυαδικά ή κατηγορικά δεδομένα και δεν αποτελεί άμεση ειδική περίπτωση της Minkowski. Ο κλασικός ορισμός της απόστασης Minkowski δίνεται ως εξής:

$$D_q(x, y) = \left(\sum_{i=1}^p |x_i - y_i|^q \right)^{\frac{1}{q}}, \quad q \geq 1$$

όπου $x = (x_1, x_2, \dots, x_p)$ και $y = (y_1, y_2, \dots, y_p)$ είναι διανύσματα στον $\mathbb{R}^p$

Η επιλογή της παραμέτρου q καθορίζει τη μορφή της απόστασης. Ειδικότερα,

- για $q = 1$ προκύπτει η απόσταση Manhattan,
- για $q = 2$ προκύπτει η ευκλείδεια απόσταση,
- ενώ όταν $\lim_{q \rightarrow \infty} (\sum_{i=1}^p |x_i - y_i|^q)^{1/q} = \max |x_i - y_i|$ προκύπτει η απόσταση Chebyshev.

Σε ορισμένες εφαρμογές χρησιμοποιείται κανονικοποίηση της απόστασης ώστε το αποτέλεσμα να μην εξαρτάται έντονα από την κλίμακα των χαρακτηριστικών ή από τον αριθμό των διαστάσεων. Αυτή η διαδικασία ονομάζεται κανονικοποίηση. Κατά την εφαρμογή του μέτρου Minkowski σε συνεχή δεδομένα, καθίσταται συχνά απαραίτητη η διαδικασία της κανονικοποίησης, ώστε να αντιμετωπιστεί το πρόβλημα της ετερογένειας κλίμακας μεταξύ των μεταβλητών. Συγκεκριμένα, εάν οι μεταβλητές μετρούνται σε διαφορετικές μονάδες π.χ. μέτρα, κιλά, δευτερόλεπτα, τότε οι διαστάσεις με μεγαλύτερες τιμές τείνουν να κυριαρχούν στον υπολογισμό της απόστασης, αλλοιώνοντας την πραγματική γεωμετρία του χώρου. Η κανονικοποίηση, η οποία συνίσταται σε κατάλληλη κλιμάκωση ή τυποποίηση των τιμών κάθε μεταβλητής (π.χ. με διαίρεση δια της τυπικής απόκλισης ή του εύρους τιμών), εξασφαλίζει ότι όλες οι διαστάσεις συνεισφέρουν ισόρροπα στην εκτίμηση της απόστασης. Στην περίπτωση αυτή, εισάγεται ένας συντελεστής κανονικοποίησης s_i για κάθε διάσταση, οπότε ο τύπος γράφεται ως εξής:

$$D_q^{(norm)}(x, y) = \left(\sum_{i=1}^p \left| \frac{x_i - y_i}{s_i} \right|^q \right)^{1/q}, \quad q \geq 1$$

όπου η επιλογή του s_i μπορεί να γίνει με διάφορους τρόπους για παράδειγμα η τυπική απόκλιση σ_i , εύρος τιμών $\max - \min$, ή βάρη w_i που εκφράζουν τη σχετική σημασία κάθε χαρακτηριστικού. Με τον τρόπο αυτό εξασφαλίζεται ότι όλες οι διαστάσεις συμβάλλουν ισόρροπα στον υπολογισμό της απόστασης.

Μια τρίτη εκδοχή που απαντάται στη βιβλιογραφία είναι η κανονικοποίηση ως προς τον αριθμό των διαστάσεων. Στην περίπτωση αυτή, η απόσταση ορίζεται ως:

$$d(x, y) = \left(\frac{1}{p} \sum_{i=1}^p |x_i - y_i|^q \right)^{1/q}, \quad q \geq 1$$

Ο παράγοντας $\frac{1}{p}$ εξισορροπεί την επίδραση του πλήθους των διαστάσεων, επιτρέποντας την αποτίμηση της απόστασης ανεξάρτητα από το μέγεθος του διανυσματικού χώρου.

Η τιμή του παραμέτρου q μπορεί να προσαρμόζεται ανάλογα με τη φύση των δεδομένων. Μικρές τιμές οδηγούν σε μεγαλύτερη ανθεκτικότητα σε θόρυβο, ενώ υψηλότερες τιμές ενισχύουν την ευαισθησία σε απομακρυσμένα σημεία (outliers).

Συνοψίζοντας, η Minkowski απόσταση εμφανίζεται σε τρεις κύριες μορφές, στον κλασικό ορισμό που δώσαμε αρχικά, στην κανονικοποίηση ανά χαρακτηριστικό μέσω συντελεστών s_i , και στην κανονικοποίηση ως προς το πλήθος των διαστάσεων. Η επιλογή της κατάλληλης μορφής εξαρτάται από τη φύση των δεδομένων, τη διαστατικότητα του χώρου και τις απαιτήσεις της εκάστοτε ανάλυσης.

2.6 L_1 απόσταση και L_∞ απόσταση ή Chebyshev απόσταση

Η απόσταση Minkowski αποτελεί μια γενίκευση της ευκλείδειας απόστασης και ορίζεται για δύο διανύσματα $x, y \in \mathbb{R}^p$ ως:

$$d_p(x, y) = \left(\sum_{i=1}^p |x_i - y_i|^p \right)^{1/p}, \quad p \geq 1$$

Ανάλογα με την τιμή του εκθέτη p , προκύπτουν διαφορετικές μετρικές της ίδιας οικογένειας. Αρχικά όταν $p = 1$, προκύπτει η μετρική L_1 , γνωστή και ως απόσταση Manhattan ή City-Block, επειδή υπολογίζει την απόσταση ακολουθώντας κάθετους και οριζόντιους άξονες, παρόμοια με τη διαδρομή που θα ακολουθούσε κανείς σε ένα ορθογώνιο πλέγμα δρόμων. Η σχέση αυτή δίνεται από τον παρακάτω τύπο:

$$d_{L_1}(x, y) = \sum_{i=1}^d |x_i - y_i|.$$

Για $p = 2$, η Minkowski απόσταση ισοδυναμεί με την ευκλείδεια απόσταση, ενώ για $p \rightarrow \infty$ προκύπτει η απόσταση Chebyshev ή λεγόμενη L_∞ , η οποία ορίζεται ως:

$$d_{L_\infty}(x, y) = \max\{|x_i - y_i|, 1 \leq i \leq p\}.$$

Η απόσταση Chebyshev εκφράζει τη μέγιστη επιμέρους διαφορά μεταξύ των συντεταγμένων δύο διανυσμάτων και χρησιμοποιείται συχνά σε εφαρμογές όπου η μέγιστη απόκλιση είναι πιο σημαντική από το συνολικό άθροισμα διαφορών. Όλες οι παραπάνω αποστάσεις ανήκουν στην οικογένεια των μετρικών Minkowski και ικανοποιούν τις θεμελιώδεις ιδιότητες της μετρικής, τη μη αρνητικότητα, την ταυτοσημία, την συμμετρία και την ανισότητα του τριγώνου.

Είναι προφανές ότι ισχύουν οι τρεις ιδιότητές της μετρικής. Για λόγους πληρότητας από γεωμετρική σκοπιά η μοναδιαία «σφαίρα» της L_1 είναι ρόμβος (2D) ή οκτάεδρο (3D). Οι αποστάσεις L_1 και L_∞ χρησιμοποιούνται εκτενώς σε εφαρμογές μηχανικής μάθησης, αναγνώρισης προτύπων και ανάλυσης δεδομένων, καθώς παρέχουν διαφορετικές οπτικές μέτρησης της διαφοράς μεταξύ παρατηρήσεων.

Έχοντας ορίσει τις μετρικές L_1 και L_∞ ως ειδικές περιπτώσεις της οικογένειας Minkowski, προχωρούμε σε ενδεικτικούς υπολογισμούς ώστε να αποσαφηνιστεί η πρακτική τους συμπεριφορά. Τα παρακάτω παραδείγματα δείχνουν πώς η επιλογή της μετρικής επηρεάζει το μετρούμενο μήκος μεταξύ των ίδιων σημείων και, κατ' επέκταση, την έννοια της γειτνίασης σε εφαρμογές μηχανικής μάθησης και ομαδοποίησης.

Παράδειγμα - εφαρμογή του παραπάνω τύπου:

Έστω ότι έχουμε δύο σημεία στον τρισδιάστατο χώρο:

$$x = (5, 3, 4) \text{ και } y = (2, 5, 7)$$

Η απόσταση L_1 (Manhattan) υπολογίζεται ως το άθροισμα των απόλυτων διαφορών κάθε συντεταγμένης:

$$d_{L_1}(x, y) = |5 - 2| + |3 - 5| + |4 - 7| = 3 + 2 + 3 = 8$$

Η απόσταση L_∞ ορίζεται ως η μέγιστη επιμέρους διαφορά:

$$d_{L_\infty}(x, y) = \max\{|5 - 2|, |3 - 5|, |4 - 7|\} = \max\{3, 2, 3\} = 3$$

Η μετρική L_1 λαμβάνει υπόψη το συνολικό μήκος, κατά μήκος των αξόνων, ενώ η L_∞ επικεντρώνεται στη μεγαλύτερη απόκλιση μεταξύ των δύο σημείων. Η διαφορά στις τιμές 8 έναντι 3, καταδεικνύει τον διαφορετικό τρόπο με τον οποίο κάθε μετρική εκφράζει την έννοια της απόστασης.

2.7 Απόσταση Gower

Τα κλασικά μέτρα απόστασης, όπως η ευκλείδεια ή η Manhattan, είναι κατάλληλα κυρίως για συνεχείς ποσοτικές μεταβλητές. Ωστόσο, σε πραγματικά σύνολα δεδομένων συχνά συνυπάρχουν μεταβλητές διαφορετικής φύσης, αριθμητικές, κατηγορικές, δυαδικές ή ακόμη και ιεραρχικές. Για να αντιμετωπιστεί η πολυπλοκότητα αυτή, ο J. C. Gower (1971) πρότεινε έναν γενικό συντελεστή ομοιότητας, ο οποίος επιτρέπει τον υπολογισμό της απόστασης μεταξύ παρατηρήσεων με ετερογενείς τύπους μεταβλητών. Το μέτρο αυτό είναι γνωστό ως απόσταση Gower (Gower distance). Έχει καθιερωθεί ως βασικό εργαλείο στην ανάλυση δεδομένων μικτού τύπου, με εφαρμογές στη βιοπληροφορική, την ταξινόμηση, τις κοινωνικές επιστήμες και την ανάλυση ιατρικών δεδομένων (Sokal & Sneath, 1963 Gower, 1971).

Η απόσταση Gower μεταξύ δύο παρατηρήσεων x και y , με d συνολικές μεταβλητές, ορίζεται από τον παρακάτω μαθηματικό τύπο.

$$d_{Gow}(x, y) = 1 - \frac{\sum_{i=1}^d w_i s_i(x, y)}{\sum_{i=1}^d w_i},$$

όπου $s_i(x, y)$ είναι η μερική ομοιότητα μεταξύ των παρατηρήσεων στις τιμές της μεταβλητής i , ορίζεται ανάλογα με τον τύπο της μεταβλητής δηλαδή,

Αν έχουμε ποσοτικές μεταβλητές στα δεδομένα μας η συνάρτηση είναι:

$$s_i(x, y) = 1 - \frac{|x_i - y_i|}{R_i},$$

Όπου R_i συμβολίζουμε το εύρος (max - min) της μεταβλητής i . Ενώ για τις κατηγορικές μεταβλητές χρησιμοποιείται η συνάρτηση:

$$s_i(x, y) = \begin{cases} 1, & \text{αν } x_i = y_i \\ 0, & \text{αν } x_i \neq y_i \end{cases}$$

Είναι πράγματι αλήθεια ότι όταν όλες οι μεταβλητές είναι ποσοτικές, ο παραπάνω σχετικός τύπος στον ορισμό απλοποιείται στη μορφή:

$$d_{Gow}(x, y) = \frac{1}{d} \sum_{i=1}^d \frac{|x_i - y_i|}{R_i}$$

Η σχέση αυτή υπολογίζει τη μέση κανονικοποιημένη απόλυτη διαφορά ανάμεσα στα αντίστοιχα στοιχεία. Το πλεονέκτημά της είναι ότι αποτρέπει τη δυσανάλογη επίδραση μεταβλητών με διαφορετικές κλίμακες μέτρησης, καθώς όλες οι αποστάσεις κανονικοποιούνται ως προς το εύρος R_i . Απλοποιώντας λαμβάνουμε την παρακάτω σχέση,

$$d_{Gow}(x, y) = \sum_{i=1}^d \frac{|x_i - y_i|}{R_i},$$

όπως είναι φυσικό παρατηρούμε ότι λείπει το πηλίκο, πράγματι πρόκειται για την ίδια λογική, αλλά οι αποστάσεις μπορεί να είναι έως d . Πιο κατανοητά, όπως είναι αναμενόμενο, η διαφορά μεταξύ των δύο τύπων έγκειται αποκλειστικά στην κανονικοποίηση του αθροίσματος. Συγκεκριμένα, όταν το πηλίκο $\frac{1}{d}$ δεν περιλαμβάνεται, η απόσταση προκύπτει ως άθροισμα κανονικοποιημένων διαφορών και συνεπώς, η μέγιστη δυνατή τιμή της μπορεί να φτάσει μέχρι το πλήθος των μεταβλητών d . Αντίθετα, όταν χρησιμοποιείται ο παράγοντας $\frac{1}{d}$, υπολογίζεται ο μέσος όρος των μερικών αποστάσεων, με αποτέλεσμα η συνολική απόσταση να κανονικοποιείται στο διάστημα $[0,1]$. Έτσι, οι δύο μορφές του τύπου αποτυπώνουν την ίδια λογική, αλλά διαφέρουν ως προς την κλίμακα αναφοράς, η πρώτη δίνει αθροιστική τιμή, ενώ η δεύτερη εξασφαλίζει συγκρισιμότητα ανεξαρτήτως του αριθμού των μεταβλητών.

Επίσης, η απόσταση αυτή από την τελευταία σχέση που αποτελεί ειδική περίπτωση της σταθμισμένης απόστασης cityblock για $r_i = \frac{1}{R_i}$, χρησιμοποιείται συνήθως όταν στα δεδομένα μας περιλαμβάνονται και δίτιμες τυχαίες μεταβλητές.

Ως επέκταση η απόσταση Gower μπορεί να προσαρμοστεί ώστε να περιλαμβάνει βαθμισμένες μεταβλητές, προσδίδοντας μεγαλύτερη σημασία σε ορισμένα χαρακτηριστικά (Burnaby, 1970). Ακόμη, έχουν προταθεί εκδοχές για ιεραρχικές μεταβλητές, όπου οι δευτερεύουσες ιδιότητες συνεισφέρουν στο σκορ ομοιότητας μόνο εφόσον συμπίπτουν οι κύριες (Kendrick & Proctor, 1964).

Η ευελιξία του μέτρου το καθιστά ιδανικό για ανάλυση σε κλάδους όπως, η βιοπληροφορική (ταξινόμηση ειδών, γονιδιωματικά δεδομένα), οι κοινωνικές επιστήμες (ανάλυση ερωτηματολογίων με ποσοτικά και κατηγορικά στοιχεία), η υγεία (συνδυασμός κλινικών, δημογραφικών και εργαστηριακών δεδομένων).

Συμπερασματικά, η απόσταση Gower αποτελεί μία από τις πιο ισχυρές και ευέλικτες μετρικές απόστασης για δεδομένα μικτού τύπου. Ο γενικός της ορισμός επιτρέπει την ομοιόμορφη αντιμετώπιση διαφορετικών τύπων μεταβλητών, ενώ η απλοποιημένη εκδοχή της για ποσοτικά δεδομένα εξασφαλίζει απλότητα και υπολογιστική αποδοτικότητα. Η ευρεία χρήση της σε διαφορετικά επιστημονικά πεδία καταδεικνύει τη σημασία της ως εργαλείου στη σύγχρονη στατιστική και επιστήμη δεδομένων.

2.8 Απόσταση Soergel

Η απόσταση Soergel αποτελεί κεντρικό μέτρο ανομοιότητας σε περιβάλλοντα υψηλής διάστασης με αραιές αναπαραστάσεις, επειδή αγνοεί τις κοινές απουσίες χαρακτηριστικών και εστιάζει στις κοινές παρουσίες και στις αποκλίσεις τους. Στην κλασική βιβλιογραφία της η απόσταση Soergel εμφανίζεται ως συμπλήρωμα του συντελεστή ομοιότητας Tanimoto/Jaccard, όσο μεγαλύτερος ο Tanimoto τόσο μικρότερη η Soergel και αντιστρόφως. Η πρόσφατη βιβλιογραφία στα μέτρα ομοιότητας επιβεβαιώνει τόσο τον ρόλο της οικογένειας Jaccard/Tanimoto όσο και τις μετρικές της ιδιότητες. Στο παρόν κεφάλαιο ορίζουμε τυπικά την απόσταση Soergel, αποδεικνύουμε βασικές της ιδιότητες, παρουσιάζουμε ισοδυναμίες με εναλλακτικές διατυπώσεις του Jaccard, αναλύουμε τη σχέση της με άλλα μέτρα (Hamming, ευκλείδεια, Cosine), και περιγράφουμε τις συνέπειες στην ομαδοποίηση - ανάλυση συστάδων που αποτελεί αντικείμενο της διπλωματικής.

Σε πλήθος εφαρμογών στατιστικής μάθησης και εξόρυξης δεδομένων, τα αντικείμενα αναπαρίστανται ως μη αρνητικά διανύσματα. Σε αυτά τα πλαίσια είναι επιθυμητό ένα μέτρο ανομοιότητας που αγνοεί τις κοινές απουσίες χαρακτηριστικών, που είναι αναλλοίωτο, και παίρνει τιμές στο $[0,1]$ με σαφή ερμηνεία. Η απόσταση Soergel ικανοποιεί τα παραπάνω και ειδικά στα δυαδικά δεδομένα, ταυτίζεται με την απόσταση Jaccard.

Ο μαθηματικός τύπος αυτής της απόστασης δίνεται παρακάτω:

$$d_s(x, y) = \frac{\sum_{i=1}^p |x_i - y_i|}{\sum_{i=1}^p \max\{x_i, y_i\}},$$

Όμως παρατηρούμε ότι στην περίπτωση που τα x_i, y_i ισούνται με μηδέν έχουμε απροσδιόριστη μορφή στο τύπο $\frac{\sum_{i=1}^p |x_i - y_i|}{\sum_{i=1}^p \max\{x_i, y_i\}}$, συνεπώς συμβολικά ορίζουμε συμβατικά $d_s(0,0) = 0$.

Ας απλοποιήσουμε λίγο την παραπάνω σχέση και σύμφωνα με γνώσεις απειροστικού λογισμού και μαθηματικής ανάλυσης ισχύει:

$$d_s(x, y) = \frac{\sum_{i=1}^p |x_i - y_i|}{\sum_{i=1}^p \max\{x_i, y_i\}} = 1 - \frac{\sum_{i=1}^p \min\{x_i, y_i\}}{\sum_{i=1}^p \max\{x_i, y_i\}}$$

Στη βιβλιογραφία η απόσταση αυτή ονομάζεται και απόσταση Ruzicka, με εύρος $[0,1]$, δηλαδή για κάθε $0 \leq |x_i - y_i| \leq \max\{x_i, y_i\}$

Αθροίζοντας σε όλα τα i , έχουμε την παρακάτω σχέση:

$$0 \leq \sum_i |x_i - y_i| \leq \sum_i \max\{x_i, y_i\}$$

Αν $\sum_i \max\{x_i, y_i\}$ είναι μη αρνητικό τότε διαιρούμε την σχέση $0 \leq \sum_i |x_i - y_i| \leq \sum_i \max\{x_i, y_i\}$ ως προς $\sum_i \max\{x_i, y_i\}$ και αυτό που λαμβάνουμε είναι ότι το κλάσμα είναι πάντα ανάμεσα σε 0 και 1, δηλαδή $[0,1]$.

Θεώρημα (Η απόσταση Soergel-Ruzicka είναι μετρική)

Έστω $x, y \in \mathbb{R}_+^p$, ορίζουμε την συνάρτηση $d_s: \mathbb{R}_+^p \times \mathbb{R}_+^p \rightarrow [0,1]$

$$d_s(x, y) = \begin{cases} 0, & x = y = 0 \\ \frac{\sum_{i=1}^p |x_i - y_i|}{\sum_{i=1}^p \max\{x_i, y_i\}} = 1 - \frac{\sum_{i=1}^p \min\{x_i, y_i\}}{\sum_{i=1}^p \max\{x_i, y_i\}}, & x \neq y \end{cases} \text{ Τότε η } d_s \text{ είναι μετρική.}$$

Στην περίπτωση που έχουμε στα συλλεγόμενα δεδομένα μας, δυαδικά χαρακτηριστικά ο αντίστοιχος μαθηματικός τύπος δίνεται σύμφωνα με την βιβλιογραφία από τους Jaccard/Tanimoto και είναι η παρακάτω σχέση.

Έστω A, B δύο σύνολα δεδομένων τότε έχουμε,

$$S_j(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{c}{a + b - c}$$

$$d_j(A, B) = 1 - S_j(A, B)$$

Είναι πράγματι αλήθεια πως στην περίπτωση που έχουμε στην συλλογή μας συνεχή δεδομένα τότε έχουμε τον μαθηματικό τύπο του Tanimoto συνεχούς μορφής, ο οποίος είναι:

Για κάθε $x, y \in \mathbb{R}^n$

$$S_T(x, y) = \frac{\langle x, y \rangle}{\|x\|_2^2 + \|y\|_2^2 - \langle x, y \rangle}$$

Όμως το συμπλήρωμά του $1 - S_T$ δεν είναι γενικά μετρική, εκτός της δυαδικής περίπτωσης όπου συμπίπτει με τον d_j .

Μια απλή εφαρμογή που μπορεί να μας επιβεβαιώσει τον παραπάνω ισχυρισμό είναι η παρακάτω:

Έστω ότι έχουμε τα διανύσματα $a = (0,1), b = (2,1), c = (1,1)$.

Υπολογίζουμε τα $S_T(a, b), S_T(a, c)$ & $S_T(b, c)$ που είναι 0.2, 0.5 και 0.75 αντίστοιχα. Συνεπώς, τα αντίστοιχα $d_T(a, b), d_T(a, c)$ & $d_T(b, c)$ είναι 0.8, 0.5 και 0.25

Καταλήγουμε στο συμπέρασμα ότι $d_T(a, b) = 0.8 > 0.5 + 0.25 = d_T(a, c) + d_T(b, c)$ που παραβιάζει την τριγωνική ανισότητα της μετρικής, άρα το $1 - S_T$ δεν είναι μετρική.

Βασικές προτάσεις της Soergel

Πρόταση 2.8.1(Εμβέλεια και ακραίες τιμές): Για κάθε $x, y \geq 0$ ισχύει $d_s(x, y) \in [0,1]$. Επιπλέον $d_s = 0 \Leftrightarrow x = y$ ενώ $d_s = 1$ όταν τα στηρίγματα των x και y έχουν μηδενική επικάλυψη, δηλαδή όταν $\min\{x_i, y_i\} = 0$ για κάθε i και τουλάχιστον ένα από τα x και y είναι μη μηδενικό.

Πρόταση 2.8.2: Για κάθε $x, y \geq 0$ η απόσταση Soergel είναι συμμετρική, δηλαδή ικανοποιεί τις σχέσεις $d_s(x, y) = 0 \Leftrightarrow x = y$ και ισχύει επίσης $d_s(x, y) = d_s(y, x)$.

Πρόταση 2.8.3: Για κάθε $x, y \in \mathbb{R}_+^n \geq 0$ και για κάθε $a > 0$ τότε ισχύει $d_s(ax, ay) = d_s(x, y)$.

Λήμμα 2.8.1: Για κάθε $x, y \geq 0$ ισχύει ότι $\frac{\|x-y\|}{\|x\|_1 + \|y\|_1} \leq d_s(x, y) \leq \frac{\|x-y\|_1}{\max(\|x\|_1, \|y\|_1)}$, διότι $\sum_i \max\{x_i, y_i\} \in [\max(\|x\|_1, \|y\|_1), \|x\|_1 + \|y\|_1]$

2.9 Απόσταση Kulczynski

Η απόσταση Kulczynski ανήκει στην κατηγορία των δεικτών ομοιότητας/απόστασης που χρησιμοποιούνται για τη μέτρηση της εγγύτητας μεταξύ αντικειμένων σε πολυδιάστατο χώρο, δηλαδή ανήκει στην οικογένεια L_1 . Αρχικά προτάθηκε για δυαδικά δεδομένα παρουσίας-απουσίας, όμως στη βιβλιογραφία έχουν διατυπωθεί επεκτάσεις που επιτρέπουν τη χρήση της και σε συνεχή ή αριθμητικά δεδομένα. Η απόσταση Kulczynski λοιπόν ομαλοποιεί το άθροισμα των απόλυτων διαφορών μεταξύ των αντίστοιχων στοιχείων των διανυσμάτων x, y με το άθροισμα των ελάχιστων τιμών των αντίστοιχων στοιχείων των διανυσμάτων.

Ορισμός 2.9: Έστω τα διανύσματα $x, y \geq 0$. Η απόσταση Kulczynski ορίζεται ως $d_{kul}(x, y) = \frac{\sum_{i=1}^d |x_i - y_i|}{\sum_{i=1}^d \min(x_i, y_i)}$, με $x, y \in R^d \geq 0$ και $\sum_{i=1}^d \min(x_i, y_i) > 0$

Ερμηνεία

- Ο αριθμητής $\sum_{i=1}^d |x_i - y_i|$ εκφράζει τη συνολική διαφορά ανάμεσα στα δύο διανύσματα.
- Ο παρονομαστής $\sum_{i=1}^d \min(x_i, y_i)$ μετράει το κοινό μέρος των δύο διανυσμάτων.
- Η απόσταση Kulczynski, επομένως, αποτυπώνει τη σχετική διαφορά των διανυσμάτων σε σχέση με το κοινό τους τμήμα.

Η παραπάνω απόσταση δεν θεωρείται αυστηρά ότι είναι μετρική. Πράγματι ισχύουν όλες οι υπόλοιπες ιδιότητες εκτός από την τριγωνική ανισότητα που δεν ικανοποιείται πάντα, για αυτό θα δώσουμε ένα αντιπαράδειγμα.

Έστω τα τρία διανύσματα: $x = (0, 0, 1)$, $y = (0, 0, 3)$ και $z = (0, 0, 2)$

Υπολογίζουμε τις αποστάσεις ανά δύο με το τύπο εύρεσης απόστασης του Kulczynski και έχουμε:

$$d_{kul}(x, z) = \frac{\sum_{i=1}^d |x_i - z_i|}{\sum_{i=1}^d \min(x_i, z_i)} = \frac{\sum_{i=1}^d |(0, 0, 1) - (0, 0, 2)|}{\sum_{i=1}^d \min((0, 0, 1), (0, 0, 2))} = \frac{|0 - 0| + |0 - 0| + |1 - 2|}{0 + 0 + 1} = 1$$

$$d_{kul}(z, y) = \frac{\sum_{i=1}^d |z_i - y_i|}{\sum_{i=1}^d \min(z_i, y_i)} = \frac{\sum_{i=1}^d |(0, 0, 2) - (0, 0, 3)|}{\sum_{i=1}^d \min((0, 0, 2), (0, 0, 3))} = \frac{|0 - 0| + |0 - 0| + |2 - 3|}{0 + 0 + 2} = 0.5$$

$$d_{kul}(x, y) = \frac{\sum_{i=1}^d |x_i - y_i|}{\sum_{i=1}^d \min(x_i, y_i)} = \frac{\sum_{i=1}^d |(0, 0, 1) - (0, 0, 3)|}{\sum_{i=1}^d \min((0, 0, 1), (0, 0, 3))} = \frac{|0 - 0| + |0 - 0| + |1 - 3|}{0 + 0 + 1} = 2$$

Ελέγχουμε λοιπόν τώρα αν όντως ισχύει η σχέση της τριγωνικής ανισότητας, δηλαδή:

$$d(x, y) \leq d(x, z) + d(z, y)$$

Με βάση όσα έχουμε υπολογίσει παραπάνω παρατηρούμαι ότι δεν μπορεί να ισχύει η σχέση της τριγωνικής ανισότητας, πράγματι καταλήξαμε σε άτοπο, κάτι που δεν ισχύει, δηλαδή:

$$2 \leq 1 + 0.5,$$

Συνεπώς, η Kulczynski απόσταση δεν είναι μετρική με την αυστηρή μαθηματική έννοια, αλλά χρησιμοποιείται ως δείκτης ασυμφωνίας.

Ολοκληρώνοντας, η απόσταση Kulczynski συνιστά ένα ευέλικτο μέτρο ασυμφωνίας που επεκτείνει τον κλασικό ορισμό του δείκτη σε συνεχή δεδομένα, προσφέροντας μια ισορροπημένη αποτίμηση της απόκλισης δύο αντικειμένων σε σχέση με το κοινό τους τμήμα. Πάρα το γεγονός ότι δεν αποτελεί αυστηρή μετρική καθώς δεν ικανοποιεί την τριγωνική ανισότητα, εντούτοις βρίσκει ευρεία εφαρμογή σε πεδία όπου η έμφαση δίνεται περισσότερο στην ποιοτική σύγκριση παρά στις αυστηρές μαθηματικές ιδιότητες. Χαρακτηριστικά παραδείγματα αποτελούν οι οικολογικές μελέτες με δεδομένα αφθονίας, η βιοπληροφορική, καθώς και οι διαδικασίες ομαδοποίησης και ταξινόμησης σε πολυδιάστατους χώρους με αριθμητικά γνωρίσματα. Ως εκ τούτου, η απόσταση Kulczynski αποτελεί ένα χρήσιμο εργαλείο για την παρούσα εργασία, συνδέοντας την θεωρητική τεκμηρίωση με τις πρακτικές απαιτήσεις της ανάλυσης δεδομένων.

2.10 Απόσταση Bhattacharyya

Η απόσταση Bhattacharyya αποτελεί ένα σημαντικό μέτρο στατιστικής ομοιότητας μεταξύ δύο κατανομών πιθανότητας. Η έννοια εισήχθη από τον στατιστικό Anil Kumar Bhattacharyya και βασίζεται στην ποσοτικοποίηση του βαθμού επικάλυψης μεταξύ δύο πιθανοτικών κατανομών. Όσο μεγαλύτερη είναι η επικάλυψη μεταξύ των κατανομών, τόσο μικρότερη είναι η απόσταση Bhattacharyya, ενώ μεγαλύτερες τιμές της απόστασης υποδηλώνουν μεγαλύτερη στατιστική διαφοροποίηση μεταξύ των πληθυσμών. Για δύο διακριτές κατανομές πιθανότητας P και Q , ο συντελεστής Bhattacharyya ορίζεται ως:

$$BC(P, Q) = \sum_x \sqrt{P(x)Q(x)}$$

και ικανοποιεί την σχέση:

$$0 \leq BC(p, q) \leq 1.$$

Όσο μεγαλύτερη είναι η τιμή του, τόσο μεγαλύτερη είναι η ομοιότητα των δύο κατανομών.

Συνεπώς, η απόσταση Bhattacharyya δίνεται από τη σχέση:

$$d_B(P, Q) = -\ln \left(\sum_x \sqrt{P(x)Q(x)} \right)$$

Στην περίπτωση συνεχών κατανομών με συναρτήσεις πυκνότητας πιθανότητας $p(x)$ και $q(x)$, ο ορισμός γενικεύεται μέσω ολοκληρώματος:

$$d_B(p, q) = -\ln \left(\int \sqrt{p(x)q(x)} dx \right)$$

Η απόσταση Bhattacharyya χρησιμοποιείται συχνά στην ανάλυση ταξινόμησης. Συγκεκριμένα, όσο μεγαλύτερη είναι η τιμή της απόστασης μεταξύ δύο κατανομών, τόσο μεγαλύτερη είναι η διαχωρισιμότητα των αντίστοιχων κλάσεων και κατά συνέπεια, τόσο μικρότερη είναι η πιθανότητα σφάλματος ταξινόμησης.

2.11 Απόσταση Hellinger

Η απόσταση Hellinger αποτελεί ένα σημαντικό μέτρο διαφοροποίησης μεταξύ δύο πιθανοτικών κατανομών. Η έννοια εισήχθη από τον Νορβηγό μαθηματικό Ernst Hellinger και βασίζεται στη σύγκριση των τετραγωνικών ριζών των πιθανοτήτων δύο κατανομών. Σε αντίθεση με άλλα μέτρα απόκλισης, η απόσταση Hellinger αποτελεί πραγματική μετρική στον χώρο των πιθανοτικών κατανομών, καθώς ικανοποιεί τις ιδιότητες της μη αρνητικότητας, της συμμετρίας και της τριγωνικής ανισότητας.

Η απόσταση Hellinger μεταξύ δύο διακριτών κατανομών πιθανότητας P και Q ορίζεται ως:

$$d_H(P, Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_x (\sqrt{P(x)} - \sqrt{Q(x)})^2}$$

Για συνεχείς κατανομές με συναρτήσεις πυκνότητας πιθανότητας $p(x)$ και $q(x)$, ο ορισμός επεκτείνεται ως:

$$d_H(p, q) = \frac{1}{\sqrt{2}} \sqrt{\int (\sqrt{p(x)} - \sqrt{q(x)})^2 dx}$$

Η απόσταση Hellinger λαμβάνει τιμές στο διάστημα $[0, 1]$, όπου η τιμή 0 υποδηλώνει ταύτιση των δύο κατανομών, ενώ η τιμή 1 αντιστοιχεί σε πλήρη διαφοροποίηση. Λόγω των ιδιοτήτων της και της αριθμητικής σταθερότητάς της, χρησιμοποιείται συχνά σε προβλήματα στατιστικής εκτίμησης και συγκρίσεων πιθανοτικών μοντέλων.

Αξίζει να σημειωθεί ότι η απόσταση Hellinger συνδέεται στενά με την απόσταση Bhattacharyya, καθώς βασίζονται στην ίδια έννοια επικάλυψης μεταξύ πιθανοτικών κατανομών. Συγκεκριμένα, ο συντελεστής Bhattacharyya χρησιμοποιείται για τον υπολογισμό τόσο της απόστασης Bhattacharyya όσο και της απόστασης Hellinger, γεγονός που καθιστά τα δύο μέτρα θεωρητικά συναφή. Ωστόσο, η απόσταση Hellinger παρουσιάζει το πλεονέκτημα ότι αποτελεί κανονικοποιημένη μετρική, γεγονός που διευκολύνει την ερμηνεία και τη σύγκριση των αποτελεσμάτων σε πρακτικές εφαρμογές στατιστικής ανάλυσης.

2.12 Απόσταση Canberra

Η απόσταση Canberra, που εισήχθη από τους Lance και Williams στα τέλη της δεκαετίας του 1960 ως ένα μέτρο λογισμικού, αποτελεί μια σταθμισμένη εκδοχή της κλασικής απόστασης L_1 ή του κανόνα του Spearman, η οποία επεκτείνεται φυσικά σε μετρική επί των συμμετρικών ομάδων. Η συγκεκριμένη μετρική είναι ιδιαίτερα χρήσιμη κατά τη σύγκριση ταξινομημένων λιστών. Η απόσταση Canberra αποτελεί ένα σημαντικό μέτρο απόστασης που χρησιμοποιείται ευρέως στην ανάλυση πολυδιάστατων δεδομένων, ιδιαίτερα σε περιπτώσεις όπου οι μεταβλητές έχουν διαφορετικές κλίμακες ή περιλαμβάνουν μικρές ή μηδενικές τιμές. Εισήχθη από τον Lance και τον Williams (1967) και θεωρείται παραλλαγή της απόστασης Manhattan, με τη διαφορά ότι λαμβάνει υπόψη το σχετικό μέγεθος των διαφορών μεταξύ των στοιχείων των διανυσμάτων. Το Canberra υπολογίζει τη διαφορά μεταξύ δύο σημείων λαμβάνοντας υπόψη τόσο τις σχετικές όσο και τις απόλυτες αποκλίσεις των χαρακτηριστικών

τους. Σε αντίθεση με την ευκλείδεια απόσταση, είναι πιο ευαίσθητο σε μικρές μεταβολές τιμών, γεγονός που το καθιστά αποτελεσματικό για αραιά ή ετερογενή σύνολα δεδομένων. Το μέτρο Canberra, παρουσιάζει ιδιαίτερο ενδιαφέρον καθώς συνδυάζει τις απόλυτες και σχετικές διαφορές μεταξύ των τιμών των χαρακτηριστικών δύο σημείων. Ορίζεται μαθηματικά ως:

$$d_{ij} = \sum_{r=1}^p \frac{|x_{ir} - x_{jr}|}{|x_{ir}| + |x_{jr}|}$$

Όπου x_i, x_j είναι δύο διανύσματα στο p - διαστατό χώρο. Το χαρακτηριστικό που τη διαφοροποιεί από άλλα μέτρα είναι ότι σταθμίζει τις απόλυτες διαφορές των τιμών των μεταβλητών σε σχέση με το άθροισμα των απόλυτων τιμών τους. Με τον τρόπο αυτόν, αποδίδει μεγαλύτερη σημασία στις σχετικές διαφορές όταν οι τιμές είναι μικρές, καθιστώντας την ιδιαίτερα ευαίσθητη κοντά στο μηδέν και επομένως κατάλληλη για αραιά δεδομένα ή δεδομένα με διαφορετικές μονάδες μέτρησης. Η απόσταση Canberra χρησιμοποιείται συχνά σε προβλήματα ομαδοποίησης (clustering) και ταξινόμησης (classification), καθώς προσφέρει μια πιο ισορροπημένη εκτίμηση της διαφοράς μεταξύ σημείων σε σχέση με την ευκλείδεια ή τη Manhattan απόσταση. Ειδικά σε εφαρμογές μηχανικής μάθησης, όπως ο αλγόριθμος k-Nearest Neighbors (KNN), έχει αποδειχθεί ότι η χρήση του Canberra metric. Η ευαισθησία της απόστασης Canberra σε μικρές μεταβολές τιμών, σε συνδυασμό με τη σταθερότητά της απέναντι σε μεγάλες διαφορές, την καθιστά ιδιαίτερα χρήσιμη σε πολυδιάστατα στατιστικά προβλήματα, όπου η σωστή εκτίμηση της απόστασης μεταξύ σημείων επηρεάζει άμεσα την ποιότητα των αποτελεσμάτων της ομαδοποίησης.

Η απόσταση Canberra ανήκει στην κατηγορία των μετρικών αποστάσεων, καθώς πληροί τις τρεις θεμελιώδεις ιδιότητες που απαιτούνται για να θεωρηθεί μια συνάρτηση $d(x, y)$ μετρική σε έναν χώρο.

Πρόταση 2.12: Η συνάρτηση $d: R^p \times R^p \rightarrow R, d(x, y) = \sum_{i=1}^n \delta(x_i, y_i)$ είναι μετρική,

Όπου

$$\delta(x_i, y_i) = \begin{cases} \frac{|x_i - y_i|}{|x_i| + |y_i|}, & (x_i, y_i) \neq 0 \\ 0, & (x_i, y_i) = 0 \end{cases}$$

Η απόσταση Canberra αποτελεί μια ευέλικτη και ερμηνεύσιμη μετρική για την ανάλυση πολυδιάστατων δεδομένων. Σε αντίθεση με άλλες αποστάσεις που βασίζονται αποκλειστικά στις απόλυτες διαφορές μεταξύ των στοιχείων δύο διανυσμάτων, η Canberra λαμβάνει υπόψη και τη σχετική σημασία των διαφορών, αποδίδοντας μεγαλύτερο βάρος σε αποκλίσεις που εμφανίζονται σε μικρές τιμές.

Η απόσταση Canberra ικανοποιεί τις βασικές ιδιότητες μιας μετρικής, γεγονός που επιβεβαιώνει τη θεωρητική της εγκυρότητα και δικαιολογεί τη χρήση της σε στατιστικές και υπολογιστικές εφαρμογές. Σε πρακτικό επίπεδο, εφαρμόζεται συχνά σε προβλήματα ομαδοποίησης, ταξινόμησης και κατάταξης, όπου η ανθεκτικότητα σε ακραίες τιμές. Συνεπώς, η μετρική Canberra συνδυάζει θεωρητική συνέπεια και πρακτική αποτελεσματικότητα, συμβάλλοντας στη βελτίωση της αξιοπιστίας των μεθόδων ανάλυσης δεδομένων που βασίζονται στη μέτρηση απόστασης.

2.13 Απόσταση με ελλείπουσες τιμές

Στη μηχανική μάθηση και την εξόρυξη δεδομένων, ένα από τα πιο συχνά προβλήματα που συναντώνται κατά την ανάλυση είναι η παρουσία ελλειπουσών τιμών (missing values). Η ύπαρξη τέτοιων τιμών μπορεί να προκαλέσει σημαντικές στρεβλώσεις στα αποτελέσματα των αλγορίθμων μάθησης, ιδίως εκείνων που βασίζονται σε μετρικές απόστασης, όπως οι k-Nearest Neighbors (k-NN), οι αλγόριθμοι ομαδοποίησης (clustering) και οι μέθοδοι αναγνώρισης προτύπων (pattern recognition). Για τον λόγο αυτό, η αντιμετώπιση των ελλειπουσών τιμών κατά τον υπολογισμό της απόστασης μεταξύ παραδειγμάτων αποτελεί κρίσιμο στάδιο στην προεπεξεργασία και ανάλυση των δεδομένων. Οι Witten, Frank, Hall και Pal (2017) αφιερώνουν ειδική αναφορά στο ζήτημα αυτό, προτείνοντας μεθόδους που διασφαλίζουν τη συνέπεια του υπολογισμού χωρίς να αλλοιώνεται η φυσική ερμηνεία της απόστασης.

Όταν μία τιμή ενός γνωρίσματος λείπει, δεν μπορεί να υπολογιστεί αριθμητική διαφορά και συνεπώς δεν είναι δυνατό να εκτιμηθεί με βεβαιότητα αν οι δύο τιμές είναι όμοιες ή διαφορετικές. Έτσι, η μη κατάλληλη αντιμετώπιση των ελλειπουσών τιμών μπορεί να οδηγήσει, είτε σε υποεκτίμηση της απόστασης αν αγνοηθούν τα πεδία, είτε σε υπερεκτίμηση αν θεωρηθούν όλες οι ελλείπουσες τιμές ως μέγιστα διαφορετικές. Η επιλογή της στρατηγικής επηρεάζει άμεσα τη συμπεριφορά του αλγορίθμου και την αξιοπιστία των αποτελεσμάτων.

Οι προσεγγίσεις αντιμετώπισης στη βιβλιογραφία διακρίνονται σε τρεις βασικές. Για τον υπολογισμό απόστασης όταν υπάρχουν ελλείπουσες τιμές έχουμε :

1. Αγνόηση του γνωρίσματος (Attribute Omission)

Η απλούστερη προσέγγιση είναι να αγνοηθεί κάθε γνώρισμα για το οποίο λείπει η τιμή, μειώνοντας έτσι τη διάσταση του χώρου απόστασης. Ο συνολικός υπολογισμός προσαρμόζεται ως εξής:

$$D(x_i, x_j) = \frac{\sum_{k=1}^p u_k d_k(x_i, x_j)}{\sum_{k=1}^p u_k}$$

όπου:

$$u_k = \begin{cases} 0, & \text{αν οι τιμές } a_{ik}, a_{jk} \text{ είναι διαθέσιμες} \\ 1, & \text{αν οι τιμές } a_{ik}, a_{jk} \text{ δεν είναι διαθέσιμες} \end{cases}$$

Έτσι, η απόσταση υπολογίζεται μόνο για τα διαθέσιμα γνωρίσματα.

2. Αντικατάσταση ελλειπουσών τιμών (Imputation)

Μία άλλη μέθοδος είναι η εκτίμηση των ελλειπουσών τιμών βάσει των διαθέσιμων δεδομένων, πριν από τον υπολογισμό της απόστασης. Οι πιο συνηθισμένες τεχνικές είναι:

- Μέσος όρος (Mean Imputation) για συνεχείς μεταβλητές δηλαδή, αντικατάσταση με τον μέσο ή διάμεσο όρο του γνωρίσματος.
- Πιο συχνή τιμή (Mode Imputation) για κατηγορικές μεταβλητές δηλαδή, αντικατάσταση με την επικρατούσα κατηγορία.
- Προσέγγιση με k-NN Imputation. Η ελλείπουσα τιμή εκτιμάται βάσει των πιο όμοιων τιμών.

Αν και οι τεχνικές αυτές μειώνουν την απώλεια πληροφορίας, εισάγουν τεχνητές τιμές που μπορεί να οδηγήσουν σε προκατάληψη (bias) αν οι ελλείπουσες τιμές δεν είναι τυχαίες.

3. Υπόθεση Μέγιστης Διαφοράς (Maximum Distance Assumption)

Η μέθοδος που προτείνεται από τους Witten et al. (2017) είναι να θεωρείται η μέγιστη δυνατή διαφορά για κάθε ελλείπουσα τιμή. Αυτό σημαίνει ότι:

- αν τα δεδομένα είναι κανονικοποιημένα στο διάστημα $[0,1]$, τότε $d(a_i, a_j) = 1$ όταν λείπει μία τιμή.
- αν δεν είναι κανονικοποιημένα, τότε η διαφορά τίθεται ίση με $\max(A) - \min(A)$ δηλαδή το εύρος του γνωρίσματος.

Η προσέγγιση αυτή θεωρεί συντηρητικά ότι η έλλειψη πληροφορίας υποδηλώνει μέγιστη ανομοιότητα μεταξύ παραδειγμάτων. Στην πράξη, αυτό αποτρέπει τα παραδείγματα με ελλείπουσες τιμές από το να θεωρηθούν ψευδώς όμοια με άλλα.

Συμπερασματικά, η αντιμετώπιση των ελλειπουσών τιμών κατά τον υπολογισμό αποστάσεων είναι καθοριστικής σημασίας για την ορθότητα των αποτελεσμάτων στη μηχανική μάθηση. Η μέθοδος της μέγιστης διαφοράς, όπως περιγράφεται από τους Witten et al. (2017), προσφέρει έναν απλό τρόπο χειρισμού ελλειπουσών τιμών, διατηρώντας τη σταθερότητα των υπολογισμών. Ωστόσο, η προσέγγιση αυτή πρέπει να χρησιμοποιείται με προσοχή, ιδιαίτερα όταν τα δεδομένα είναι πολύ ελλιπή. Σε τέτοιες περιπτώσεις, προτείνεται η χρήση στατιστικών τεχνικών εκτίμησης ή πιο προηγμένων μετρικών, όπως η Gower Distance όπως έχουμε προαναφέρει παραπάνω, που λαμβάνουν υπόψη τη μερική πληρότητα των γνωρισμάτων. Σε κάθε περίπτωση, η κατανόηση της επίδρασης των ελλειπουσών τιμών στον υπολογισμό της απόστασης αποτελεί προϋπόθεση για τη σωστή σχεδίαση και εφαρμογή μοντέλων μηχανικής μάθησης σε πραγματικά δεδομένα.

ΚΕΦΑΛΑΙΟ 3

Μέτρα ομοιότητας

3.1 Εισαγωγή

Η έννοια της απόστασης αποτελεί βασικό στοιχείο στη μηχανική μάθηση και τη στατιστική, καθώς καθορίζει τη μέτρηση ομοιότητας μεταξύ παρατηρήσεων και επηρεάζει άμεσα μεθόδους όπως η ταξινόμηση και η ομαδοποίηση. Σε διακριτά δεδομένα (δυαδικά, ονομαστικά, διατεταγμένα ή μικτού τύπου), η επιλογή κατάλληλου μέτρου είναι κρίσιμη για την ακρίβεια, την ερμηνευσιμότητα και την υπολογιστική αποδοτικότητα. Στα δυαδικά δεδομένα διακρίνονται συμμετρικά και ασύμμετρα μέτρα, ανάλογα με τη σημασία της κοινής απουσίας, ενώ στις ονομαστικές μεταβλητές η ομοιότητα βασίζεται σε ισότητα κατηγοριών, συχνά με στάθμιση σπανιότητας. Στις διατεταγμένες μεταβλητές χρησιμοποιούνται μετασχηματισμοί που διατηρούν τη διάταξη χωρίς να υποθέτουν γραμμικότητα. Επιπλέον, η διαχείριση ελλειπουσών τιμών απαιτεί προσαρμοσμένες προσεγγίσεις, όπως στον συντελεστή Gower για μικτά δεδομένα. Η επιλογή μέτρου εξαρτάται από τη σημασιολογία της σύμπτωσης, τη σχετική σημασία των κατηγοριών και την πληρότητα των δεδομένων, ενώ στην πράξη δίνεται έμφαση κυρίως στη συγκριτική κατάταξη αποστάσεων πάρα στις αυστηρές μετρικές ιδιότητες.

Στις υποενότητες που ακολουθούν, παρουσιάζονται συστηματικά τα βασικότερα μέτρα για δυαδικά, ονομαστικά και διατεταγμένα δεδομένα. Για τα δυαδικά, υιοθετούμε την καθιερωμένη συμβολική αναπαράσταση με τον πίνακα συμφωνιών - ασυμφωνιών (a, b, c, d):

a: πλήθος θέσεων με 1 - 1 (κοινές παρουσίες),

b: 1 - 0 ή 0 - 1 (ασυμφωνίες),

c: 0 - 0 (κοινές απουσίες),

d: $p = a + b + c + d$: συνολικός αριθμός συγκρινόμενων γνωρισμάτων.

Με αυτήν τη συμβολική, τα περισσότερα μέτρα ορίζονται ως συναρτήσεις των a, b, c, d, διευκολύνοντας τόσο τη θεωρητική ανάλυση για παράδειγμα αν πληρούνται οι αξιωματικές συνθήκες μετρικής, όσο και την αποδοτική υλοποίηση. Θα δούμε επίσης ποια μέτρα είναι μετρικές δηλαδή αν ικανοποιούν την μη αρνητικότητα, ταυτοσημία, συμμετρία, ανισότητα τριγώνου και ποια είναι απλώς συντελεστές ομοιότητας, χωρίς βέβαια απαίτηση για την τριγωνική ανισότητα, καθώς και αυτό έχει σημασία στην πράξη το αν ένα μέτρο είναι αυστηρά μετρικό ή απλώς συντελεστής ομοιότητας σπάνια αλλάζει το αποτέλεσμα, οι περισσότεροι αλγόριθμοι στηρίζονται μόνο στη σειρά των αποστάσεων.

Τέλος, θα εντάξουμε τις επιλογές αυτές στη ροή μιας πραγματικής ανάλυσης επιλογή μέτρου με βάση τον στόχο ταξινόμηση, αναζήτηση γειτόνων, συμπερασματολογία, τον τύπο μεταβλητών συμμετρικές - ασύμμετρες, την πυκνότητα (sparsity) και την παρουσία ελλείψεων. Θα δοθούν παραδείγματα με αριθμητικούς υπολογισμούς και προτάσεις υλοποίησης.

3.2 Ο δειγματικός συντελεστής συσχέτισης

Αρχικά, λοιπόν θα δείξουμε την εναλλακτική κατηγορία μέτρων που μπορούν να χρησιμοποιηθούν για να μας δείξουν αν δύο άτομα ή παρατηρήσεις είναι όμοια ή ανόμοια μεταξύ τους, αυτά είναι τα λεγόμενα μέτρα ομοιότητας όπως αναφέραμε παραπάνω (similarity measures ή affinity measures). Αυτά έχουν το χαρακτηριστικό γνώρισμα ότι παρατηρήσεις που μοιάζουν πολύ μεταξύ τους, δίνουν πολύ μεγάλη τιμή στο μέτρο της ομοιότητας ενώ παρατηρήσεις που είναι ανόμοιες του δίνουν πολύ μικρή τιμή. Έχοντας εισάγει κατάλληλα τέτοια μέτρα, θα μπορούμε να τοποθετούμε ζεύγη παρατηρήσεων στην ίδια ή σε διαφορετική ομάδα ανάλογα με το αν η τιμή του μέτρου είναι μεγάλη ή μικρή αντίστοιχα.

Αν υποθέσουμε ότι για κάθε ζεύγος παρατηρήσεων $x_i = (x_{i1}, x_{i2}, \dots, x_{in})$ και $x_j = (x_{j1}, x_{j2}, \dots, x_{jn})$ ορίζεται ένας πραγματικός αριθμός $s_{ij} = s(x_i, x_j)$ έτσι ώστε να ισχύουν οι επόμενες τρεις ιδιότητες:

- $s_{ij} \geq 0$ για i, j και $i = j \Rightarrow s_{ij} = 1$
- $s_{ij} \leq 1$
- $s_{ij} = s_{ji}$

τότε θα λέμε ότι η συνάρτηση $s_{ij} = s(x_i, x_j)$ δίνει ένα μέτρο ομοιότητας.

Συνεπώς, ο δειγματικός συντελεστής συσχέτισης Pearson έχουν εύρος τιμών στο $[-1, 1]$.

Σε αυτές τις περιπτώσεις που δεν γίνει κανονικοποίηση, η κανονικοποιημένη μορφή δίνεται από την παρακάτω σχέση:

$$s'_{ij} = \frac{1 + s_{ij}}{2},$$

Έτσι, μεταφέρετε το εύρος τιμών στο $[0, 1]$, επιτρέποντας τη σύγκρισή του με άλλα μέτρα ομοιότητας.

Ο συντελεστής συσχέτισης Pearson, ή αλλιώς συντελεστής γραμμικής συσχέτισης του Pearson, αποτελεί το πιο ευρέως χρησιμοποιούμενο μέτρο στατιστικής συσχέτισης μεταξύ δύο ποσοτικών μεταβλητών. Εισήχθη από τον Karl Pearson (1896) και χρησιμοποιείται για τη μέτρηση του βαθμού γραμμικής σχέσης μεταξύ δύο μεταβλητών x και y .

Η τιμή του συντελεστή συμβολίζεται με r_p ή απλώς r και λαμβάνει όπως έχουμε αναφέρει τιμές στο διάστημα:

$$-1 \leq r \leq 1$$

- Όταν $r = 1$, η συσχέτιση είναι τέλεια θετική.
- Όταν $r = -1$, η συσχέτιση είναι τέλεια αρνητική.
- Όταν $r = 0$, δεν υπάρχει γραμμική σχέση μεταξύ των δύο μεταβλητών.

Το πιο γνωστό λοιπόν, μέτρο ομοιότητας για ποσοτικές παρατηρήσεις είναι ο δειγματικός συντελεστής συσχέτισης Pearson, ο οποίος ορίζεται ως ο λόγος της συνδιακύμανσης των δύο μεταβλητών προς το γινόμενο των τυπικών τους αποκλίσεων, δηλαδή δίνεται από το μαθηματικό τύπο:

$$r_p = \frac{cov(x, y)}{\sigma_x \sigma_y}$$

όπου:

$$cov(x, y) = \frac{1}{p} \sum_{i=1}^p (x_i - \bar{x})(y_i - \bar{y})$$

και:

$$\sigma_x = \sqrt{\frac{1}{p} \sum_{i=1}^p (x_i - \bar{x})^2}, \sigma_y = \sqrt{\frac{1}{p} \sum_{i=1}^p (y_i - \bar{y})^2}$$

Σε πιο γενική μορφή, όταν έχουμε πολλές μεταβλητές, η συσχέτιση μεταξύ των μεταβλητών x_i και x_j δίνεται από:

$$s_{ij} = \frac{\sum_{r=1}^p (x_{ir} - \bar{x}_i)(x_{jr} - \bar{x}_j)}{(\sum_{r=1}^p (x_{ir} - \bar{x}_i)^2 \sum_{k=1}^p (x_{jr} - \bar{x}_j)^2)^{1/2}}$$

όπου $\bar{x}_i = \frac{1}{p} \sum_{r=1}^p x_{ir}$, $\bar{x}_j = \frac{1}{p} \sum_{r=1}^p x_{jr}$, συνεπώς όπως προαναφέραμε και παραπάνω $s_{ij} \in [-1, 1]$ και αντιπροσωπεύει το στοιχείο της θέσης (i, j) στον πίνακα συσχετίσεων $S = [s_{ij}]$.

Ο συντελεστής συσχέτισης Pearson παρέχει μια αριθμητική εκτίμηση της γραμμικής σχέσης μεταξύ δύο μεταβλητών. Ωστόσο, η τιμή του r_p από μόνη της δεν επαρκεί για να εξαχθεί ασφαλές συμπέρασμα σχετικά με τη στατιστική σημαντικότητα της συσχέτισης. Είναι απαραίτητο να εξεταστεί εάν η τιμή αυτή θα μπορούσε να έχει προκύψει τυχαία, δηλαδή υπό την υπόθεση ότι οι δύο μεταβλητές είναι ασυσχέτιστες. Για τον σκοπό αυτό πραγματοποιείται έλεγχος στατιστικής σημαντικότητας του συντελεστή, ο οποίος παρουσιάζεται στη συνέχεια.

Για να ελεγχθεί αν η συσχέτιση είναι στατιστικά σημαντική, εξετάζεται η υπόθεση:

- $H_0: r_p = 0$ δεν υπάρχει γραμμική συσχέτιση.
- $H_1: r_p \neq 0$ υπάρχει γραμμική συσχέτιση.

Το στατιστικό ελέγχου είναι:

$$t = \frac{r_p \sqrt{p-2}}{\sqrt{1-r_p^2}}$$

όπου:

- r_p : συντελεστής συσχέτισης Pearson στο δείγμα,
- p : αριθμός παρατηρήσεων.

Το στατιστικό t ακολουθεί την κατανομή t - Student με $n - 2$ βαθμούς ελευθερίας. Απορρίπτουμε την H_0 αν $|t| > t_{crit}(n - 2, \alpha/2)$ ή ισοδύναμα απορρίπτουμε την H_0 αν $p\text{-value} < \alpha$

Εφαρμογή:

Για την πρακτική κατανόηση της εφαρμογής του συντελεστή συσχέτισης Pearson, παρουσιάζεται στη συνέχεια μια εφαρμογή με τη χρήση της γλώσσας R. Ο στόχος είναι να υπολογιστεί η τιμή του συντελεστή r_p και να ερμηνευθεί η σχέση μεταξύ δύο μεταβλητών X και Y .

Από την εφαρμογή του συντελεστή Pearson στα δεδομένα του Πίνακα 3.2.1 προέκυψε τιμή $r_p = 0.6744$ με $p - value = 0.2118$. Παρότι η σχέση είναι θετική και μέτρια, η τιμή p υπερβαίνει το επίπεδο σημαντικότητας 0.05, γεγονός που υποδεικνύει ότι δεν υπάρχει στατιστικά σημαντική γραμμική συσχέτιση. Το διάγραμμα 3.2.1 απεικονίζει τη γραμμική τάση κόκκινη γραμμή, η οποία επιβεβαιώνει την κατεύθυνση της συσχέτισης, χωρίς όμως να αποδεικνύει σημαντική εξάρτηση.

3.3 Ο συντελεστής Spearman

Η συσχέτιση αποτελεί μία από τις πιο θεμελιώδεις έννοιες της στατιστικής ανάλυσης. Εκφράζει το βαθμό συνάφειας μεταξύ δύο τυχαίων μεταβλητών δηλαδή, το κατά πόσο η μεταβολή της μίας συνοδεύεται από προβλέψιμη μεταβολή της άλλης. Η γραμμική συσχέτιση περιγράφεται παραδοσιακά από τον συντελεστή Pearson. Ωστόσο, σε περιπτώσεις όπου η σχέση δεν είναι γραμμική ή οι μεταβλητές δεν ακολουθούν κανονική κατανομή, το μέτρο αυτό μπορεί να οδηγήσει σε παραπλανητικά αποτελέσματα. Σε αυτές τις περιπτώσεις, χρησιμοποιείται ο συντελεστής βαθμολογικής συσχέτισης του Spearman, ο οποίος βασίζεται στην ταξινόμηση (ranking) των παρατηρήσεων και όχι στις πραγματικές τιμές τους. Ο Spearman επιτρέπει τη μελέτη το μη παραμετρικό μέτρο συσχέτισης.

Ας θεωρήσουμε δύο τυχαίες μεταβλητές X και Y , με δείγμα παρατηρήσεων:

$$(x_1, y_1), (x_2, y_2), \dots, (x_p, y_p)$$

Ο Spearman δεν εξετάζει τις πραγματικές τιμές των x_i και y_i , δηλαδή δεν μελετά τις ίδιες τις τιμές των δεδομένων, αλλά τη σειρά κατάταξης αυτών των τιμών. Τον συντελεστή Spearman δεν τον ενδιαφέρει πόσο απέχουν οι τιμές μεταξύ τους, αλλά ποια είναι η σειρά τους ποια είναι μεγαλύτερη, ποια μικρότερη.

Πρόταση 3.3:

$$R(x_i), R(y_i)$$

όπου $R(\cdot)$ συμβολίζει τη θέση της παρατήρησης στη σειρά ταξινόμησης από 1 έως p .

Ο συντελεστής Spearman ορίζεται ως ο συντελεστής Pearson μεταξύ των δύο μεταβλητών $R(X)$ και $R(Y)$:

$$\rho_s = \text{Corr}(R(X), R(Y))$$

Αυτό σημαίνει ότι παρόλο που είναι μη παραμετρικό μέτρο, το υπόβαθρό του βασίζεται στην ίδια θεωρητική λογική του Pearson με τη διαφορά ότι η μέτρηση πραγματοποιείται σε επίπεδο ταξινομημένων δεδομένων.

Όταν δεν υπάρχουν ισοβαθμίες, ο συντελεστής Spearman μπορεί να υπολογιστεί απευθείας από τις διαφορές των βαθμών $d_i = R(x_i) - R(y_i)$ ως εξής:

$$\rho_s = 1 - \frac{6 \sum_{i=1}^p d_i^2}{p(p^2 - 1)} \quad (3.3.1)$$

Η εξίσωση (3.3.1) προκύπτει από τον γενικό ορισμό του Pearson, μετά από αλγεβρική απλοποίηση των βαθμών $1, \dots, p$.

Αν υπάρχουν ισοβαθμίες, τότε χρησιμοποιείται ο γενικός μαθηματικός τύπος:

$$\rho_s = \frac{\sum_{i=1}^p (R(x_i) - \bar{R}_x)(R(y_i) - \bar{R}_y)}{\sqrt{\sum_{i=1}^p (R(x_i) - \bar{R}_x)^2 \sum_{i=1}^p (R(y_i) - \bar{R}_y)^2}}$$

όπου:

- $R(x_i), R(y_i)$: βαθμοί (ranks) των παρατηρήσεων,
- $\bar{R}_x = \bar{R}_y = \frac{p+1}{2}$: μέσος βαθμός κάθε μεταβλητής.

Ο συντελεστής Spearman λαμβάνει τιμές $[0,1]$ δηλαδή,

$$-1 \leq \rho_s \leq +1$$

- $\rho_s = +1$: τέλεια θετική συσχέτιση, οι βαθμοί των δύο μεταβλητών συμπίπτουν πλήρως.
- $\rho_s = -1$: τέλεια αρνητική συσχέτιση οι βαθμοί των δύο μεταβλητών είναι ακριβώς αντίστροφοι.
- $\rho_s = 0$: απουσία συσχέτισης.

Ο στατιστικός έλεγχος για το συντελεστή Spearman, δηλαδή η μηδενική υπόθεση H_0 ορίζει ότι δεν υπάρχει συσχέτιση μεταξύ των μεταβλητών, η οποία μηδενική υπόθεση είναι:

$$H_0: \rho_s = 0$$

ενώ η εναλλακτική υπόθεση ορίζει ότι υπάρχει συσχέτιση μεταξύ των μεταβλητών, η οποία εναλλακτική υπόθεση του στατιστικού ελέγχου είναι:

$$H_1: \rho_s \neq 0$$

Για μεγαλύτερα δείγματα συνήθως για $p > 30$, η κατανομή του ρ_s προσεγγίζει την κανονική κατανομή με μέση τιμή 0 και διακύμανση:

$$Var(\rho_s) = \frac{1}{n-1}$$

επομένως, ο έλεγχος σημαντικότητας μπορεί να γίνει με τη στατιστική:

$$t = \rho_s \sqrt{\frac{p-2}{1-\rho_s^2}}$$

η οποία ακολουθεί την κατανομή t - Student με $p-2$ βαθμούς ελευθερίας.

Ο κανόνας απόφασης, αν ισχύει η μηδενική υπόθεση ή εναλλακτική, διατυπώνεται με το γεγονός ότι εφόσον έχει υπολογιστεί η απόλυτη τιμή $|t|$ του παραπάνω μαθηματικού τύπου και

εφόσον έχει βρεθεί η τιμή της t_k , από τους πίνακες, πάντα στο εκάστοτε επίπεδο σημαντικότητας τότε ισχύουν τα παρακάτω :

Αν $|t| > t_k$, συνεπώς απορρίπτουμε την μηδενική υπόθεση H_0 , και υπάρχει σημαντική συσχέτιση.

Αν $|t| \leq t_k$, τότε δεν απορρίπτουμε την μηδενική υπόθεση, συνεπώς δεν υπάρχει στατιστική σημαντική συσχέτιση.

Ένας ακόμη εναλλακτικός τρόπος απόρριψης ή αποδοχής της μηδενικής υπόθεσης είναι με την τιμή p - *vaule*.

Αν $p < \alpha$, τότε απορρίπτεται η μηδενική υπόθεση H_0

Αν $p \geq \alpha$, Συνεπώς δεν απορρίπτεται η H_0 .

Η μελέτη της σχέσης μεταξύ δύο μεταβλητών μπορεί να πραγματοποιηθεί μέσω διαφορετικών στατιστικών συντελεστών, ανάλογα με τη φύση και την κατανομή των δεδομένων. Ο συντελεστής συσχέτισης του Pearson αποτελεί το πλέον διαδεδομένο μέτρο γραμμικής εξάρτησης, ενώ ο συντελεστής του Spearman βασίζεται στις βαθμολογίες (ranks) των δεδομένων. Όταν η σχέση μεταξύ των μεταβλητών X και Y είναι γραμμική, οι δύο συντελεστές συσχέτισης του Pearson (r_p) και του Spearman (ρ_s) παρουσιάζουν παρόμοιες αριθμητικές τιμές, καθώς και οι δύο εκφράζουν το ίδιο πρότυπο αύξησης ή μείωσης των παρατηρήσεων. Αντιθέτως, όταν η σχέση είναι μη γραμμική, ο συντελεστής του Spearman εξακολουθεί να αποδίδει υψηλές τιμές, καταγράφοντας με συνέπεια τη συσχέτιση των μεταβλητών, ενώ ο συντελεστής του Pearson υποεκτιμά τη δύναμη της σχέσης, λόγω της εξάρτησής του από τη γραμμικότητα. Η διαφοροποίηση αυτή καθιστά τον συντελεστή του Spearman ιδιαίτερα χρήσιμο εργαλείο για περιπτώσεις όπου οι παραδοχές της γραμμικότητας ή της κανονικότητας παραβιάζονται, όπως συμβαίνει συχνά σε εμπειρικά ή κοινωνικά δεδομένα.

Αν η σχέση είναι μη γραμμική, ο Spearman εξακολουθεί να αποδίδει υψηλές τιμές, ενώ ο Pearson υποεκτιμά τη συσχέτιση.

Όταν υπάρχουν ισοβαθμίες για παράδειγμα ίδιες τιμές, η ακρίβεια του υπολογισμού επηρεάζεται, αλλά ο Spearman μπορεί να διορθωθεί με παράγοντες διόρθωσης στους βαθμούς.

Η διορθωμένη μορφή είναι:

$$\rho_s = \frac{\text{cov}(R_X, R_Y)}{\sqrt{\text{var}(R_X) \text{var}(R_Y)}}$$

όπου οι βαθμοί υπολογίζονται ως μέσοι βαθμοί για κάθε ομάδα ισοβαθμιών. Από στατιστική πλευρά ο Spearman είναι λιγότερο ευαίσθητος σε υποθέσεις κατανομής, γεγονός που τον καθιστά ιδανικό σε περιπτώσεις, όπως μη κανονικών δεδομένων, μεταβλητών με διαφορετική κλίμακα, παρουσία ακραίων τιμών.

Από γεωμετρική άποψη, ο συντελεστής Spearman μπορεί να ερμηνευθεί ως το συνημίτονο της γωνίας που σχηματίζουν δύο διανύσματα βαθμών στον p - διαστατό χώρο των παρατηρήσεων. Αν θεωρήσουμε:

$$R_x = [R(x_1), R(x_2), \dots, R(x_p)], \quad R_y = [R(y_1), R(y_2), \dots, R(y_p)]$$

τότε:

$$\rho_s = \cos(\theta) = \frac{R_x R_y}{\|R_x\| \|R_y\|},$$

όπου θ είναι η γωνία μεταξύ των δύο διανυσμάτων στο p - διαστατό χώρο. Όσο μικρότερη είναι η γωνία, δηλαδή όσο πιο ευθυγραμμισμένα είναι τα διανύσματα των βαθμών, τόσο ισχυρότερη είναι η θετική συσχέτιση. Αντιστρόφως, η γωνία πλησιάζει τις 180° και ο συντελεστής λαμβάνει αρνητικές τιμές, υποδηλώνει αρνητική συσχέτιση. Υπό τις παραδοχές της διμεταβλητής κανονικότητας, ο συντελεστής Spearman παρουσιάζει ασυμπτωτική σχετική αποδοτικότητα (Pitman ARE) περίπου 0,912, δηλαδή 91% ως προς τον συντελεστή Pearson, γεγονός που υποδηλώνει μικρότερη ισχύ όταν οι προϋποθέσεις του παραμετρικού μέτρου ικανοποιούνται. Αντιθέτως, σε μη κανονικές κατανομές ή σε ακραίες τιμές, ο Spearman τείνει να είναι πιο ανθεκτικός και σταθερός, σε σχέση με τον Pearson που έχει μεγαλύτερη ευαισθησία σε outliers.

Για την πρακτική κατανόηση της εφαρμογής του συντελεστή συσχέτισης Spearman, παρουσιάζεται στη συνέχεια ένα παράδειγμα με τη χρήση της γλώσσας R. Ο στόχος είναι να υπολογιστεί η τιμή του συντελεστή ρ_s και να ερμηνευθεί η σχέση μεταξύ δύο μεταβλητών X και Y . Οι δύο μεταβλητές X και Y περιλαμβάνουν πέντε ζεύγη παρατηρήσεων, όπως φαίνονται στον Πίνακα 3.3.1.

Η εκτέλεση του κώδικα πίνακα 3.3.2 έδωσε τα εξής αποτελέσματα:

$$\rho_s = 0.7, \quad p\text{-value} = 0.2333$$

Η τιμή $\rho_s = 0.7$ υποδεικνύει ισχυρή θετική συσχέτιση μεταξύ των δύο μεταβλητών, δηλαδή όσο αυξάνονται οι τιμές της μεταβλητής X , τείνουν να αυξάνονται και οι τιμές της μεταβλητής Y . Ωστόσο, η τιμή $p = 0.2333 > 0.05$ δείχνει ότι η συσχέτιση δεν είναι στατιστικά σημαντική στο επίπεδο σημαντικότητας $\alpha = 0.05$, πιθανώς λόγω του μικρού μεγέθους δείγματος ($n = 5$).

Στο διάγραμμα 3.3.1 παρουσιάζεται το διάγραμμα διασποράς των δύο μεταβλητών με την προσθήκη της γραμμής της γραμμικής τάσης (κόκκινη γραμμή). Η βασική ερμηνεία είναι ότι η γραφική απεικόνιση επιβεβαιώνει ότι τα δεδομένα εμφανίζουν αυξητική τάση όσο αυξάνεται η μεταβλητή X , αυξάνεται και η Y . Η κόκκινη γραμμή αντιπροσωπεύει τη γενική γραμμική τάση των δεδομένων και αποτυπώνει την κατεύθυνση της συσχέτισης που καταγράφει ο Spearman. Το αποτέλεσμα συνάδει με τη θεωρητική ιδιότητα του συντελεστή Spearman να αποτυπώνει τη μονοτονικότητα της σχέσης, ανεξάρτητα από το εάν αυτή είναι γραμμική ή μη.

3.4 Μέτρα ομοιότητας

Οι αποστάσεις που αναφερθήκαμε στο παραπάνω κεφάλαιο, αφορούν συνεχή δεδομένα, στην συνέχεια θα δούμε πως ορίζονται οι αποστάσεις σε δίτιμα δεδομένα ή για δίτιμες μεταβλητές, οι οποίες μπορούν να πάρουν μια από τις τιμές 0-1, π.χ. απουσία - παρουσία κάποιο αντικειμένου ή χαρακτηριστικού..

Θα δούμε στη συνέχεια με πιο τρόπο ορίζονται μέτρα ομοιότητας μεταξύ ατόμων και αντικειμένων στα οποία παρατηρούνται δίτιμες μεταβλητές (Binary Variables), δηλαδή μεταβλητές που μπορούν να πάρουν μόνο μια από δύο τιμές, 0 και 1. Δημιουργούμε τον

παρακάτω πίνακα συνάφειας 2×2 όπου a, b, c, d είναι το πλήθος των συνδυασμών $(1,1)$ $(1,0)$ $(0,1)$, $(0,0)$ αντίστοιχα και $p = a + b + c + d$.

Άτομο j				
Άτομο i		1	0	
	1	a	b	a + b
	0	c	d	c + d
		a + c	b + d	

Τα μέτρα ομοιότητας (similarity measures) χρησιμοποιούνται για τις ανάγκες προβλημάτων ομαδοποίησης ατόμων - αντικειμένων, στα οποία παρατηρούνται δίτιμες μεταβλητές.

3.4.1. Μέτρο ομοιότητας Jaccard

Ο δείκτης ομοιότητας Jaccard αποτελεί ένα ευρέως χρησιμοποιούμενο μέτρο για την εκτίμηση της ομοιότητας μεταξύ συνόλων ή δυαδικών δεδομένων, ιδιαίτερα σε εφαρμογές στατιστικής ανάλυσης, μηχανικής μάθησης και εξόρυξης δεδομένων. Εισήχθη από τον Paul Jaccard (1901) στο πλαίσιο οικολογικών μελετών και βασίζεται στον λόγο των κοινών στοιχείων δύο συνόλων προς το σύνολο των διακριτών στοιχείων τους. Χρησιμοποιείται κυρίως όταν ενδιαφέρει η παρουσία ή απουσία χαρακτηριστικών και εφαρμόζεται ευρέως σε πεδία όπως η ομαδοποίηση και η ανάκτηση πληροφορίας.

Αν θεωρήσουμε δύο σύνολα A και B, ο συντελεστής ομοιότητας Jaccard ορίζεται ως:

$$S_J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

όπου $|A \cap B|$ είναι το πλήθος των κοινών στοιχείων και $|A \cup B|$ το πλήθος των στοιχείων που ανήκουν τουλάχιστον σε ένα από τα δύο σύνολα.

Αντίστοιχα, η απόσταση ή μετρική Jaccard (Jaccard distance) ορίζεται ως:

$$D_J(A, B) = 1 - S_J(A, B)$$

Ο συντελεστής παίρνει τιμές στο διάστημα $[0,1]$, όπου:

- $S_J = 1$ υποδηλώνει πλήρη ομοιότητα (τα σύνολα είναι ταυτόσημα),
- $S_J = 0$ υποδηλώνει πλήρη ανομοιότητα (κανένα κοινό στοιχείο).

Στην πράξη, τα σύνολα A και B αντιπροσωπεύονται συχνά από δυαδικά διανύσματα μήκους p:

$x = (x_1, x_2, \dots, x_p)$ και $y = (y_1, y_2, \dots, y_p)$ όπου $x_i, y_i \in \{0, 1\}$

Ορίζουμε:

a: πλήθος περιπτώσεων όπου $x_i = 1, y_i = 1$,

b: περιπτώσεις όπου $x_i = 0, y_i = 1$,

c: περιπτώσεις όπου $x_i = 1, y_i = 0$,

d: περιπτώσεις όπου $x_i = 0, y_i = 0$.

Ο συντελεστής αυτός δίνει βαρύτητα μόνο στα κοινά παρουσιαζόμενα γνωρίσματα και αγνοεί τις κοινές απουσίες συνεπώς εκφράζεται ως :

$$S_J = \frac{a}{a + b + c}$$

ενώ η αντίστοιχη απόσταση δίνεται από τον τύπο :

$$D_J = 1 - S_J = \frac{a + b}{a + b + c}$$

Ο δείκτης Jaccard είναι κατάλληλος για δυαδικά δεδομένα, ιδιαίτερα όταν τα μηδενικά δεν μεταφέρουν σημαντική πληροφορία. Όταν διαφέρει από τον συντελεστή απλής αντιστοιχίας Simple Matching Coefficient, ο οποίος λαμβάνει υπόψη και τις περιπτώσεις όπου και τα δύο αντικείμενα έχουν τιμή 0. Εκφράζει την ομοιότητα ως λόγο επικάλυψης, εστιάζοντας στα κοινά χαρακτηριστικά σε σχέση με το σύνολο των διακριτών χαρακτηριστικών δύο αντικειμένων και αγνοώντας τις ταυτόχρονες απουσίες. Σε αντίθεση με τον Simple Matching Coefficient, είναι καταλληλότερος για αραιά δεδομένα. Από θεωρητική άποψη, αποτελεί ημι-μετρικό μέτρο, καθώς ικανοποιεί συμμετρία και μη αρνητικότητα, ενώ για δυαδικά δεδομένα ικανοποιεί και την τριγωνική ανισότητα, επιτρέποντας τη χρήση του σε αλγορίθμους ομαδοποίησης. Η απλότητά του, η απουσία παραμέτρων και η ανεξαρτησία από την κλίμακα τον καθιστούν κατάλληλο για κατηγορικά δεδομένα, ενώ εντάσσεται σε ευρύτερη οικογένεια μέτρων επικάλυψης, όπως ο δείκτης Tanimoto και ο συντελεστής Dice. Επιπλέον, μπορεί να ερμηνευθεί ως ο λόγος της πιθανότητας ταυτόχρονης εμφάνισης δύο γεγονότων προς την πιθανότητα ένωσης τους, λειτουργώντας ως μη γραμμικό μέτρο συσχέτισης που παραμένει αξιόπιστο σε μη κανονικά και πολυδιάστατα δεδομένα.

3.4.2. Μέτρο ομοιότητας Rogers - Tanimoto

Ο δείκτης ομοιότητας Rogers-Tanimoto (RT) προτάθηκε από τους David J. Rogers και Takashi T. Tanimoto το 1960 στο άρθρο “*A Computer Program for Classifying Plants*” που δημοσιεύθηκε στο περιοδικό *Science*. Στόχος τους ήταν να αναπτύξουν έναν δείκτη που να λαμβάνει υπόψη τόσο τις συμφωνίες όσο και τις ασυμφωνίες ανάμεσα σε δύο αντικείμενα, αλλά να επιβάλλει μεγαλύτερη ποινή στις διαφορές από ό,τι στα σημεία συμφωνίας. Σε αντίθεση με άλλους δείκτες όπως ο Jaccard ή ο Dice που επικεντρώνονται αποκλειστικά στις κοινές παρουσίες (1-1), ο Rogers-Tanimoto εισάγει την έννοια του penalty factor, διπλασιάζοντας το βάρος των ασυμφωνιών (1-0 ή 0-1). Η λογική αυτή προκύπτει από την ανάγκη να μειωθεί η πιθανότητα δύο αντικείμενα να χαρακτηριστούν όμοια μόνο επειδή μοιράζονται μεγάλο αριθμό κοινών απουσιών (0-0).

Ο δείκτης ομοιότητας Rogers - Tanimoto προσφέρει μια προσέγγιση στη μέτρηση ομοιότητας για δυαδικά δεδομένα, λαμβάνοντας υπόψη τόσο τις συμφωνίες όσο και τις διαφορές μεταξύ αντικειμένων. Η διπλή στάθμιση των ασυμφωνιών τον καθιστά πιο ευαίσθητο σε αποκλίσεις, παρέχοντας πιο ρεαλιστική εκτίμηση της ομοιότητας. Ο Janowitz (1980) τονίζει ότι τέτοιου τύπου δείκτες είναι καταλληλότεροι για δεδομένα με πιθανά σφάλματα, καθώς παραμένουν σταθεροί και αποδοτικοί σε προσομοιωμένα πειράματα. Από μαθηματική σκοπιά

έχουμε ότι έστω δύο αντικείμενα x_i και x_j τα οποία περιγράφονται από p δυαδικά γνωρίσματα και αφού έχουμε ορίσει παραπάνω τα a, b, c, d , τότε ο δείκτης ορίζεται ως εξής:

$$S_{RT} = \frac{a + d}{a + d + 2(b + c)}, \quad (3.4.2.1)$$

Η ομοιότητα λαμβάνει τιμές στο διάστημα $0, 1$, δηλαδή $S_{RT} \in [0,1]$

- $S_{RT} = 1$: τα αντικείμενα είναι απόλυτα όμοια.
- $S_{RT} = 0$: καμία συμφωνία.

Η βασική διαφορά από το Simple Matching Coefficient (SMC) είναι ότι το RT διπλασιάζει τη συνεισφορά των διαφωνιών, καθιστώντας το αυστηρότερο μέτρο.

Από την σχέση 3.1.1 προκύπτει η αντίστοιχη απόσταση Rogers-Tanimoto:

$$d_{RT} = 1 - S_{RT} = \frac{2(b + c)}{a + d + 2(b + c)} \quad (3.4.2.2)$$

Ο δείκτης αποδεικνύεται ιδιαίτερα χρήσιμος όταν μας ενδιαφέρει όχι απλώς η ταύτιση, αλλά και η διαφορά των χαρακτηριστικών. Για παράδειγμα, σε ιατρικές βάσεις δεδομένων, όπου οι διαφορές στα συμπτώματα έχουν υψηλή σημασία για τη διάγνωση, ο Rogers-Tanimoto αποδίδει καλύτερα από συμμετρικά μέτρα όπως το SMC.

Συμπερασματικά, ο δείκτης Rogers-Tanimoto αποτελεί ένα ισορροπημένο αλλά αυστηρό μέτρο ομοιότητας για δυαδικά δεδομένα. Η εισαγωγή του συντελεστή στις ασυμφωνίες παρέχει μεγαλύτερη διακριτική ικανότητα όταν τα αντικείμενα παρουσιάζουν διαφορετικά πρότυπα, καθιστώντας τον κατάλληλο για εφαρμογές όπου η διαφοροποίηση είναι ουσιώδης. Η επιλογή του πρέπει να γίνεται λαμβάνοντας υπόψη τη δομή των δεδομένων σε ισορροπημένα σύνολα με περιορισμένες ασυμφωνίες αποδίδει σταθερά αποτελέσματα, ενώ σε εξαιρετικά αραιά δεδομένα μπορεί να υπερεκτιμήσει τη διαφοροποίηση. Πάρα τους περιορισμούς του, παραμένει ένα από τα πλέον αξιόπιστα και κλασικά μέτρα δυαδικής ομοιότητας που χρησιμοποιούνται μέχρι σήμερα στη στατιστική και στην επιστήμη δεδομένων.

3.4.3. Μέτρο ομοιότητας Sokal and Sneath

Ένα από τα κλασικά μέτρα ομοιότητας για δυαδικά δεδομένα είναι ο συντελεστής Sokal και Sneath (1963), ο οποίος αναπτύχθηκε στο πλαίσιο της θεωρίας της αριθμητικής ταξινόμησης (numerical taxonomy). Οι Sokal και Sneath πρότειναν ότι η ομοιότητα μεταξύ δύο αντικειμένων μπορεί να ποσοτικοποιηθεί με βάση το ποσοστό συμφωνίας και ασυμφωνίας στα χαρακτηριστικά τους, ανεξάρτητα από τη φύση ή τη σημασία αυτών των χαρακτηριστικών. Οι Sokal και Sneath πρότειναν τρεις παραλλαγές του δείκτη, γνωστές ως Sokal & Sneath I, II, και III, οι οποίες διαφοροποιούνται ως προς τη στάθμιση των συμφωνιών και ασυμφωνιών:

1. Sokal & Sneath I

$$S_{SS1} = \frac{a}{a + 2(b + c)}$$

Δίνει διπλή ποινή στις ασυμφωνίες $(b + c)$ και χρησιμοποιείται όταν δίνεται μεγαλύτερη έμφαση στην κοινή παρουσία χαρακτηριστικών $(1 - 1)$.

2. Sokal & Sneath II

$$S_{SS2} = \frac{2(a + d)}{2(a + d) + (b + c)}$$

Λαμβάνει υπόψη τόσο τις συμφωνίες στην παρουσία όσο και στην απουσία χαρακτηριστικών (1 - 1 και 0 - 0), επομένως θεωρεί την απουσία μιας ιδιότητας ως μορφή ομοιότητας.

3. Sokal & Sneath III

$$S_{SS3} = \frac{1}{4} \left(\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{b+d} + \frac{d}{c+d} \right)$$

Αντιμετωπίζει ισότιμα όλες τις περιπτώσεις συμφωνίας και παρέχει μια εξισορροπημένη εκτίμηση της συνολικής ομοιότητας.

Οι τιμές των δεικτών κυμαίνονται στο διάστημα [0,1], όπου:

- $S = 1$ δηλώνει πλήρη ομοιότητα μεταξύ των δύο αντικειμένων,
- $S = 0$ δηλώνει πλήρη διαφορά.

Ο δείκτης ικανοποιεί τα κριτήρια μη αρνητικότητας και συμμετρίας, αλλά δεν αποτελεί απόσταση με την αυστηρή μαθηματική έννοια δεν πληροί την τριγωνική ανισότητα. Ωστόσο, μπορεί να μετατραπεί σε μέτρο απόστασης μέσω της σχέσης:

$$D_{SS} = 1 - S_{SS}$$

και να χρησιμοποιηθεί σε αλγορίθμους ομαδοποίησης ή ανάλυσης πολυδιάστατης κλιμάκωσης.

Σε αντίθεση με μέτρα όπως του Jaccard, ο συντελεστής Sokal - Sneath λαμβάνει υπόψη και τις περιπτώσεις κοινής απουσίας χαρακτηριστικών (0 - 0), θεωρώντας ως ενδείξεις ομοιότητας. Με αυτόν τον τρόπο αποτυπώνει πληρέστερα την ολική συμφωνία μεταξύ δύο αντικειμένων, ιδιαίτερα σε περιπτώσεις όπου τα δεδομένα είναι δυαδικά και αραιά. Ο συντελεστής αυτός έχει χρησιμοποιηθεί ευρέως σε πεδία όπως η βιολογία, η ανάλυση προτύπων και η μηχανική μάθηση, για την ποσοτικοποίηση της ομοιότητας μεταξύ στοιχείων βάσει παρουσίας ή απουσίας χαρακτηριστικών.

3.4.4. Μέτρο ομοιότητας Simple Matching

Στο πλαίσιο αυτό, ένας από τους πιο κλασικούς και ευρέως χρησιμοποιούμενους συμμετρικούς δείκτες ομοιότητας είναι ο δείκτης απλής ταύτισης (Simple Matching Coefficient - SMC), γνωστός και ως δείκτης Sokal - Michener, από τα ονόματα των ερευνητών που τον πρότειναν πρώτοι το 1958. Ο δείκτης αυτός υπολογίζει το ποσοστό των περιπτώσεων στις οποίες δύο αντικείμενα συμφωνούν, είτε στην παρουσία τιμή 1 είτε στην απουσία τιμή 0 ενός χαρακτηριστικού. Λόγω της απλότητάς του και της συμμετρικής φύσης του.

Ας θεωρήσουμε δύο αντικείμενα A και B , τα οποία περιγράφονται από p δυαδικά χαρακτηριστικά. Κάθε χαρακτηριστικό μπορεί να λάβει τιμή 1 (παρουσία του χαρακτηριστικού.) ή 0 (απουσία του χαρακτηριστικού.). Για τη μέτρηση της ομοιότητας μεταξύ των δύο αντικειμένων δημιουργείται ο πίνακας συχνοτήτων τεσσάρων κελιών, γνωστός και ως contingency table ή πίνακας συνύπαρξης, ο οποίος παρουσιάζεται στην αρχή του κεφαλαίου.

Έτσι λοιπόν, ο μαθηματικός ορισμός του δείκτη Simple Matching δίνεται από τον παρακάτω τύπο:

$$SMC = \frac{a + d}{a + b + c + d}$$

όπου:

- $a + d$ αντιπροσωπεύει τον συνολικό αριθμό των συμφωνιών μεταξύ των δύο αντικειμένων,
- $b + c$ αντιπροσωπεύει τον συνολικό αριθμό των διαφωνιών.
- Ο δείκτης παίρνει τιμές στο διάστημα $[0,1]$ όπου:
- $SMC = 1$ σημαίνει απόλυτη ταύτιση (όλα τα χαρακτηριστικά συμφωνούν),
- $SMC = 0$ σημαίνει πλήρη διαφωνία.

Επιπλέον, η αντίστοιχη απόσταση Simple Matching δηλαδή το μέτρο ανομοιοότητας, ορίζεται ως:

$$D_{SMC} = 1 - SMC = \frac{b + c}{a + b + c + d}$$

και χρησιμοποιείται συχνά σε μεθόδους ιεραρχικής ανάλυσης συστάδων.

Η ερμηνεία είναι ευθεία, δηλαδή όσο μεγαλύτερη η τιμή του δείκτη, τόσο μεγαλύτερη η ομοιότητα μεταξύ των αντικειμένων. Σε αντίθεση με τον δείκτη Jaccard, ο οποίος υπολογίζει μόνο τη συμφωνία παρουσίας ($1 - 1$), ο SMC θεωρεί ότι και οι συμφωνίες απουσίας ($0 - 0$) είναι εξίσου σημαντικές. Γι' αυτό και χαρακτηρίζεται συμμετρικό μέτρο ομοιότητας.

Ο δείκτης Simple Matching (Sokal-Michener) αποτελεί ένα συμμετρικό μέτρο ομοιότητας για δυαδικά δεδομένα, καθώς λαμβάνει υπόψη τόσο τις συμφωνίες παρουσίας όσο και απουσίας χαρακτηριστικών. Είναι εύκολος στον υπολογισμό και κατάλληλος όταν η παρουσία και η απουσία έχουν ίση σημασία, όπως σε δεδομένα τύπου Ναι - Όχι. Ωστόσο, σε αραιά δεδομένα τείνει να υπερεκτιμά την ομοιότητα λόγω των πολλών κοινών απουσιών, ενώ δεν λαμβάνει υπόψη τη βαρύτητα των χαρακτηριστικών. Για τον λόγο αυτό είναι πιο αξιόπιστος σε ισορροπημένα σύνολα δεδομένων, ενώ σε άλλες περιπτώσεις προτιμώνται εναλλακτικά μέτρα όπως ο δείκτης Jaccard. Χρησιμοποιείται ευρέως στην ανάλυση συστάδων για τον υπολογισμό πινάκων ομοιότητας και την ομαδοποίηση δεδομένων.

3.4.5 Μέτρο ομοιότητας Sørensen - Dice

Ο δείκτης ομοιότητας Sørensen - Dice αποτελεί ένα από τα σημαντικότερα μέτρα για τη σύγκριση αντικειμένων που περιγράφονται μέσω δυαδικών χαρακτηριστικών, δηλαδή γνωρισμάτων που δηλώνονται ως παρόντα (1) ή απόντα (0). Σύμφωνα με τους Legendre και Legendre (1998), ο Dice (1945) εισήγαγε αρχικά τον δείκτη αυτόν με την ονομασία δείκτη σύμπτωσης (*coincidence index*), ενώ ο Sørensen (1948) πρότεινε ανεξάρτητα την ίδια μαθηματική μορφή ως συντελεστή ομοιότητας (*similarity coefficient*), δίνοντας διπλό βάρος στις κοινές παρουσίες στα δύο δείγματα. Η θεωρητική του βάση στηρίζεται στην αρχή ότι η ομοιότητα μεταξύ δύο αντικειμένων καθορίζεται κυρίως από τον αριθμό των χαρακτηριστικών που συνυπάρχουν και στα δύο. Για τον λόγο αυτό, ο δείκτης δίνει διπλό βάρος στα κοινά θετικά

χαρακτηριστικά (a), επειδή θεωρούνται περισσότερο πληροφοριακά από τις κοινές απουσίες, οι οποίες μπορεί να οφείλονται σε τυχαίους ή μη σημαντικούς λόγους. Η παράλειψη των κοινών απουσιών (d) καθιστά το μέτρο ασύμμετρο, εστιάζοντας αποκλειστικά στην παρουσία χαρακτηριστικών και όχι στην απουσία τους.

Ο μαθηματικός τύπος για το δείκτη Sørensen - Dice ορίζεται ως:

$$S_{DS} = \frac{2a}{2a + b + c}$$

εκφράζει αυτήν τη θεωρητική αρχή, δείχνοντας ότι η ομοιότητα αυξάνεται καθώς αυξάνεται ο αριθμός των κοινών θετικών γνωρισμάτων σε σχέση με τα μοναδικά γνωρίσματα του κάθε αντικειμένου. Ο δείκτης αυτός ανήκει στην κατηγορία των μέτρων συσχέτισης για δεδομένα παρουσίας και λαμβάνει τιμές από 0 έως 1.

- όπου η τιμή 0 υποδηλώνει πλήρη ανυπαρξία κοινών χαρακτηριστικών,
- ενώ η τιμή 1 δηλώνει πλήρη ταύτιση.

Ο δείκτης Sørensen-Dice μετρά την ομοιότητα δίνοντας έμφαση στις κοινές παρουσίες χαρακτηριστικών και αγνοώντας τις κοινές απουσίες, γεγονός που τον καθιστά κατάλληλο για δεδομένα παρουσίας-απουσίας όπου η απουσία δεν είναι πληροφοριακά σημαντική. Η απόσταση που προκύπτει από το συμπλήρωμά του μπορεί να χρησιμοποιηθεί σε μεθόδους ομαδοποίησης και ιεραρχικής ανάλυσης. Σε σύγκριση με δείκτες όπως ο Jaccard, δίνει μεγαλύτερη βαρύτητα στις κοινές παρουσίες και αποδίδει πιο ρεαλιστικές εκτιμήσεις σε αραιά δεδομένα, αν και ενδέχεται να υποεκτιμά τη συνολική συμφωνία. Πάρα τον περιορισμό αυτό, αποτελεί ένα απλό και αξιόπιστο μέτρο ομοιότητας με ευρεία εφαρμογή σε βιολογία, οικολογία και ανάλυση δεδομένων.

3.4.6. Μέτρο ομοιότητας Russell-Rao

Ανάμεσα στα κλασικά μέτρα ομοιότητας για δυαδικές μεταβλητές συγκαταλέγεται το μέτρο Russell - Rao, το οποίο εισήχθη ως ένας από τους πιο απλούς και διαφανείς τρόπους υπολογισμού της ομοιότητας μεταξύ δύο δυαδικών διανυσμάτων. Πάρα την απλότητά του, το μέτρο αυτό έχει σημαντικές εφαρμογές σε περιβάλλοντα όπου το πλήθος των κοινών χαρακτηριστικών (τιμές 1) έχει ουσιαστική σημασία και όπου η συνολική διάσταση των δεδομένων πρέπει να λαμβάνεται υπόψη.

Για να γίνει κατανοητό το μέτρο Russell - Rao, είναι απαραίτητο να παρουσιαστεί ο τρόπος με τον οποίο αναπαρίστανται συγκρίσεις δυαδικών δεδομένων. Έστω δύο δυαδικά διανύσματα μήκους p:

$$x = (x_1, x_2, \dots, x_p), y = (y_1, y_2, \dots, y_p)$$

Κάθε θέση σε αυτά τα διανύσματα μπορεί να λάβει τις τιμές 0 (απουσία) ή 1 (παρουσία). Η σύγκριση των δύο διανυσμάτων γίνεται μέσω ενός πίνακα συμφωνίας (contingency table), ο οποίος καταγράφει πόσες φορές εμφανίζεται κάθε δυνατός συνδυασμός τιμών. Αυτός, ο πίνακας αποτελεί τον πυρήνα των περισσότερων μέτρων ομοιότητας για δυαδικά δεδομένα.

Το μέτρο ομοιότητας Russell - Rao ορίζεται ως:

$$S_{RR} = \frac{a}{a+b+c+d}$$

Οι τιμές που λαμβάνει αυτός ο συντελεστής $S_{RR} \in [0, 1]$

- όπου η τιμή 0 υποδηλώνει πλήρη ανυπαρξία κοινών χαρακτηριστικών,
- ενώ η τιμή 1 υποδηλώνει ότι τα δύο διανύσματα έχουν απόλυτη ομοιότητα σε κάθε θέση ως προς την παρουσία της τιμής 1.

Η αντίστοιχη απόσταση ή μέτρο ανομοιότητας δίνεται με την ίδια φιλοσοφία όπως και στα προηγούμενα Μέτρα ομοιότητας, από τον μαθηματικό τύπο:

$$D_{RR} = 1 - S_{RR} = 1 - \frac{a}{a+b+c+d} = \frac{b+c+d}{a+b+c+d}$$

Η κεντρική λογική πίσω από τον δείκτη είναι ιδιαίτερα απλή, μετρά το ποσοστό των χαρακτηριστικών όπου και τα δύο αντικείμενα έχουν ενεργή ένδειξη. Επομένως, το μέτρο Russell - Rao εστιάζει αποκλειστικά στις περιπτώσεις όπου η τιμή είναι 1 και δεν λαμβάνει υπόψη τις συμφωνίες στην τιμή 0 ως στοιχείο ομοιότητας. Αυτό καθιστά το μέτρο κατάλληλο για εφαρμογές όπου η ένδειξη 1 έχει ουσιαστική σημασία (π.χ. παρουσία ενός χαρακτηριστικού., θετική ψήφος, ενεργή λειτουργία), ενώ η ταυτόχρονη απουσία του δεν θεωρείται εξίσου σημαντική.

Το μέτρο ομοιότητας Russell-Rao βασίζεται αποκλειστικά στο ποσοστό των κοινών θετικών ενδείξεων (τιμές 1) ως προς το σύνολο των χαρακτηριστικών, γεγονός που το καθιστά ιδιαίτερα απλό και υπολογιστικά αποδοτικό, ακόμη και για μεγάλα σύνολα δεδομένων. Είναι κατάλληλο όταν η παρουσία χαρακτηριστικών αποτελεί το βασικό κριτήριο σύγκρισης. Ωστόσο, αγνοεί πλήρως τις κοινές απουσίες, με αποτέλεσμα να υποεκτιμά την ομοιότητα σε αραιά δεδομένα ή όταν οι αρνητικές συμφωνίες έχουν σημασία. Επιπλέον, σε περιπτώσεις με λίγες θετικές τιμές τείνει να δίνει χαμηλές εκτιμήσεις ομοιότητας. Πάρα τους περιορισμούς αυτούς, παραμένει χρήσιμο μέτρο σε εφαρμογές όπου προέχει η σύμπτωση παρουσίας χαρακτηριστικών.

Ένα απλό παράδειγμα υπολογισμού για το μέτρο ομοιότητας είναι το παρακάτω.

Έστω δύο διανύσματα $x = (1, 0, 1, 1, 0)$ και $y = (1, 1, 0, 1, 0)$

Δίνεται λοιπόν ο παρακάτω 2×2 πίνακα συμφωνίας, σύμφωνα με τα παραπάνω δεδομένα, άρα έχουμε ότι:

Συμφωνίες / Διαφωνίες	1	0
1	2	1
0	1	1

Σύμφωνα, με το μαθηματικό τύπο $S_{RR} = \frac{a}{a+b+c+d}$ υπολογίζουμε το ποσοστό των κοινών ενεργών χαρακτηριστικών. Συνεπώς, $S_{RR} = \frac{a}{a+b+c+d} = \frac{2}{5} = 0.4$

Άρα τα διανύσματα έχουν 40% κοινά χαρακτηριστικά.

Ένα ακόμη παράδειγμα που έχει ενδιαφέρον και επεξηγεί την έκφραση «σε δεδομένα με λίγες θετικές ενδείξεις τείνει να παράγει χαμηλές τιμές ομοιότητας ακόμη και όταν υπάρχει υψηλός βαθμός σχετικής συμφωνίας»

Ας $x = (1, 0, 0, \dots, 0)$ και $y = (1, 0, 0, \dots, 0)$ το μήκος των διανυσμάτων είναι $p = 100$, τα διανύσματα $\tilde{x}$ και $\tilde{y}$ έχουν πλήρη συμφωνία. Φτιάχνοντας, έτσι το παρακάτω πίνακα 2×2 συμφωνίας.

Συμφωνίες Διαφωνίες	1	0
1	1	0
0	0	99

Σύμφωνα, με το παραπάνω μαθηματικό τύπο του μέτρο ομοιότητας Russell - Rao το ποσοστό που λαμβάνουμε είναι:

$$S_{RR} = \frac{a}{a + b + c + d} = \frac{1}{100} = 0.01$$

Παρότι υπάρχει πλήρης συμφωνία στα σημαντικά χαρακτηριστικά, η ομοιότητα βγαίνει μόλις 0.01 ή 1%, δηλαδή πολύ χαμηλή.

Από το παραπάνω παράδειγμα βγαίνει το συμπέρασμα ότι το μέτρο ομοιότητας Russell - Rao «τιμωρεί» περιπτώσεις όπου υπάρχουν πολύ λίγες τιμές 1, όπως αναφέραμε και στα μειονεκτήματα.

Συμπερασματικά, το μέτρο ομοιότητας Russell - Rao αποτελεί ένα από τα απλούστερα και πιο διαφανή μέτρα ομοιότητας για δυαδικά δεδομένα. Εστιάζει αποκλειστικά στην ταυτόχρονη παρουσία ενεργών χαρακτηριστικών, γεγονός που το καθιστά κατάλληλο για πλήθος εφαρμογών όπου η θετική ένδειξη είναι το σημαντικότερο στοιχείο πληροφορίας. Πάρα τους περιορισμούς του, ιδίως σε περιπτώσεις σπάνιων τιμών 1, παραμένει ένα χρήσιμο εργαλείο για αναλύσεις μεγάλης κλίμακας και για προβλήματα όπου η απλότητα και η υπολογιστική αποδοτικότητα είναι κρίσιμες. Το κεφάλαιο αυτό ανέδειξε την θεωρητική δομή του μέτρου, παρουσίασε τις ιδιότητές του και ανέλυσε παραδείγματα εφαρμογών, παρέχοντας ένα ολοκληρωμένο υπόβαθρο για τη χρήση του Russell - Rao στο πλαίσιο της διπλωματικής εργασίας.

3.4.7. Μέτρο ομοιότητας Kulczynski

Ο δείκτης ομοιότητας Kulczynski αποτελεί ένα ευρέως χρησιμοποιούμενο μέτρο για δυαδικά δεδομένα, ιδιαίτερα σε περιπτώσεις ασύμμετρων ή αραιών συνόλων. Εισήχθη από τον Stanisław Kulczynski στο πλαίσιο οικολογικών μελετών, με στόχο την αποτίμηση της ομοιότητας λαμβάνοντας υπόψη τη σχετική συχνότητα εμφάνισης των χαρακτηριστικών. Σε σύγκριση με δείκτες όπως ο Jaccard και ο Dice, παρέχει πιο σταθερές εκτιμήσεις σε δεδομένα με ανισορροπία, γεγονός που τον καθιστά κατάλληλο για εφαρμογές όπως η ομαδοποίηση, η ταξινόμηση, η αναγνώριση προτύπων και η ανάλυση δεδομένων.

Ο βασικός ορισμός του δείκτη Kulczynski για δεδομένα σε μορφή 2×2 δίνεται από τον μαθηματικό τύπο:

$$S_K = \frac{1}{2} \left(\frac{a}{a+b} + \frac{a}{a+c} \right)$$

Ο πρώτος όρος $\frac{a}{a+b}$ εκφράζει την πιθανότητα να εμφανιστεί το Β όταν εμφανίζεται το Α (conditional probability).

Ο δεύτερος όρος $\frac{a}{a+c}$ εκφράζει την πιθανότητα να εμφανιστεί το Α όταν εμφανίζεται το Β.

Το μέτρο Kulczynski είναι ο μέσος όρος αυτών των δύο επιμέρους πιθανοτήτων, προσφέροντας έτσι μια συμμετρική αλλά ταυτόχρονα ευαίσθητη στις ασυμμετρικές προσέγγιση ομοιότητας.

Ο δείκτης Kulczynski είναι ένα συμμετρικό μέτρο ομοιότητας για δυαδικά δεδομένα, το οποίο βασίζεται στον μέσο όρο δύο δεσμευμένων πιθανοτήτων και λαμβάνει τιμές στο διάστημα $[0,1]$, όπου 0 δηλώνει πλήρη ανομοιότητα και όπου 1 πλήρη ομοιότητα. Διακρίνεται για την ανθεκτικότητά του σε σπάνια και μη ισορροπημένα δεδομένα, καθώς δεν τιμωρεί έντονα τις περιπτώσεις χαμηλής κοινής εμφάνισης, σε αντίθεση με άλλους δείκτες όπως ο Jaccard ή ο Dice. Παρουσιάζει καλή ερμηνευσιμότητα και σταθερότητα σε δεδομένα υψηλής διάστασης, γεγονός που τον καθιστά κατάλληλο για εφαρμογές στη στατιστική ανάλυση, τη μηχανική μάθηση και την αναγνώριση προτύπων. Ωστόσο, αγνοεί την πληροφορία των κοινών απουσιών και δεν αποτελεί αυστηρό μέτρο απόστασης, περιορίζοντας τη χρήση του σε ορισμένες μεθόδους. Πάρα τους περιορισμούς αυτούς, χρησιμοποιείται ευρέως σε πεδία όπως η βιοστατιστική, η ομαδοποίηση δεδομένων και η ανίχνευση ανωμαλιών, προσφέροντας αξιόπιστη εκτίμηση ομοιότητας σε σύνθετα και ανομοιογενή δεδομένα.

3.5 Συμπεράσματα

Στο κεφάλαιο αυτό παρουσιάστηκαν βασικά μέτρα συσχέτισης και ομοιότητας που χρησιμοποιούνται στην πολυμεταβλητή ανάλυση, με έμφαση στις ιδιότητες και τις εφαρμογές τους. Οι συντελεστές Pearson και Spearman ανέδειξαν διαφορετικές προσεγγίσεις στη μέτρηση σχέσεων μεταξύ μεταβλητών, ανάλογα με τις υποθέσεις και τη μορφή της εξάρτησης. Παράλληλα, τα μέτρα ομοιότητας (όπως Jaccard, Rogers-Tanimoto και Sokal-Sneath) έδειξαν ότι η αποτίμηση ομοιότητας εξαρτάται από τη βαρύτητα που δίνεται σε παρουσίες και απουσίες χαρακτηριστικών. Τονίστηκε επίσης η σημασία της κανονικοποίησης και της σωστής επιλογής μέτρου, ανάλογα με τη φύση των δεδομένων και τον στόχο της ανάλυσης. Συνολικά, τα μέτρα αυτά αποτελούν βασικά εργαλεία για την κατανόηση σχέσεων και τη μετέπειτα εφαρμογή μεθόδων ομαδοποίησης.

ΚΕΦΑΛΑΙΟ 4

Μέθοδοι ομαδοποίησης και σύγχρονες προσεγγίσεις στη μηχανική μάθηση

Η ομαδοποίηση (clustering) αποτελεί βασική μέθοδο μη επιβλεπόμενης μάθησης για την αναγνώριση προτύπων και δομών σε δεδομένα χωρίς προκαθορισμένες κλάσεις. Η επιλογή κατάλληλου αλγορίθμου και μέτρου απόστασης ή ομοιότητας είναι κρίσιμη για την ποιότητα των αποτελεσμάτων. Στο κεφάλαιο αυτό παρουσιάζονται οι κύριες κατηγορίες αλγορίθμων ομαδοποίησης, όπως οι ιεραρχικές, οι μη ιεραρχικές (π.χ. K-means, K-medoids), καθώς και μέθοδοι βασισμένες στην πυκνότητα και σε πιθανοτικά μοντέλα (π.χ. DBSCAN, OPTICS, Mean Shift, Gaussian Mixture Models). Επιπλέον, εξετάζονται σύγχρονες προσεγγίσεις που βασίζονται σε embeddings και νευρωνικά δίκτυα για δεδομένα υψηλής διάστασης. Τέλος, συγκρίνονται τα βασικά χαρακτηριστικά των μεθόδων και αναδεικνύεται ο ρόλος των μετρικών στην απόδοσή τους, θέτοντας το υπόβαθρο για την εφαρμογή τους σε πραγματικά δεδομένα.

4.1 Ιεραρχικές μέθοδοι

Η ιεραρχική ομαδοποίηση (hierarchical clustering) αποτελεί μία από τις πιο θεμελιώδεις και ιστορικά σημαντικές τεχνικές. Η μέθοδος στοχεύει στη δημιουργία μίας δενδροειδούς δομής (dendrogram). Σε αντίθεση με τους μη ιεραρχικούς αλγορίθμους, όπως το k-means, η ιεραρχική ομαδοποίηση δεν απαιτεί εκ των προτέρων τον καθορισμό του αριθμού των clusters. Αντίθετα, δημιουργεί μία πλήρη αναπαράσταση των σχέσεων μεταξύ των δεδομένων, επιτρέποντας στον αναλυτή να επιλέξει το επιθυμητό επίπεδο τομής ανάλογα με τις ανάγκες της εκάστοτε εφαρμογής.

Η ιεραρχική ομαδοποίηση διακρίνεται σε δύο κύριες προσεγγίσεις (α) τη συσσωρευτική (agglomerative), κατά την οποία τα clusters χτίζονται από κάτω προς τα πάνω και (β) τη διαιρετική (divisive), κατά την οποία η διαδικασία ξεκινά από ένα ενιαίο cluster και διαχωρίζεται σταδιακά σε υποσύνολα. Επιπλέον, η επιλογή του κριτηρίου σύνδεσης (linkage criterion) επηρεάζει καθοριστικά τη γεωμετρία, τη σταθερότητα και τη μορφή των παραγόμενων ομάδων. Στις ακόλουθες ενότητες αναλύονται διεξοδικά οι δύο προσεγγίσεις καθώς και τα κυριότερα κριτήρια σύνδεσης.

4.1.1 Συσσωρευτικές μέθοδοι

Σε αυτό το υπό κεφάλαιο θα αναφερθούν οι σημαντικότερες συσσωρευτικοί μέθοδοι ιεραρχικής ομαδοποίησης. Η μέθοδος αυτή έχει το χαρακτηριστικό ότι πηγαίνει από πάνω προς τα κάτω δηλαδή, κάθε παρατήρηση αποτελεί την δική της σύσταση. Η συσσωρευτική μέθοδος είναι διαδομένη λόγω της απλότητας και της δυνατότητας αποτίμησης του αποτελέσματος μέσω δέντροδιαγράμματος. Θεωρούμε ότι υπάρχουν διαθέσιμες p παρατηρήσεις για p διαφορετικές μεταβλητές, οι οποίες σχηματίζουν τα διανύσματα

$$x_i = (x_{i1}, x_{i2}, \dots, x_{ip}), \quad i = 1, 2, \dots, p$$

και ότι για κάθε ζεύγος παρατηρήσεων (i, j) είναι γνωστή η απόσταση

$$d_{ij} = d(x_i, x_j),$$

η οποία συγκεντρώνεται σε ένα πίνακα αποστάσεων $D = [d_{ij}]$

Η βασική ιδέα των συγκολλητικών αλγορίθμων συσταδοποίησης στηρίζεται στο ότι κάθε παρατήρηση ξεκινά ως ξεχωριστή ομάδα, η οποία στη συνέχεια συνενώνεται σταδιακά με άλλες ομάδες. Στο αρχικό στάδιο, λοιπόν, υπάρχουν p ομάδες μία για κάθε παρατήρηση σε κάθε βήμα, αναζητείται η ελάχιστη απόσταση μεταξύ δύο ομάδων και πραγματοποιείται συγχώνευση. Η διαδικασία συνεχίζεται μέχρι όλες οι παρατηρήσεις να ενωθούν σε μία τελική ομάδα, σχηματίζοντας μια πλήρη ιεραρχική δομή.

Η γενική μορφή του αλγορίθμου έχει ως εξής:

Βήμα 1°: Ξεκινάμε με p ομάδες (clusters) του ενός ατόμου η καθεμιά και με τον $p \times p$ πίνακα των αποστάσεων $D = [d_{ij}]$ (εναλλακτικά θα μπορούσε να χρησιμοποιηθεί ένας πίνακας μέτρων ομοιότητας).

Βήμα 2°: Εντοπίζουμε στον πίνακα D το ζεύγος των πλησιέστερων πιο όμοιων ομάδων έστω A και B .

Βήμα 3°: Συγχώνευσε τις ομάδες A και B σε μια ομάδα, την $P = (AB)$ μειώνοντας έτσι τον αριθμό των ομάδων κατά ένα. Ανανέωσε τον πίνακα αποστάσεων D διαγράφοντας τις γραμμές και στήλες που αντιστοιχούσαν στις ομάδες A και B , και προσθέτοντας μια γραμμή και μια στήλη που περιέχει τις αποστάσεις της ομάδας $P = (AB)$ από τις υπόλοιπες ομάδες.

Βήμα 4°: Επανάλαβε τα βήματα 2 και 3 συνολικά $1 - p$ φορές έτσι ώστε με τη λήξη του αλγορίθμου, όλα τα άτομα να αποτελούν μία μόνο ομάδα. Σε κάθε βήμα κατάγραψε τις λεπτομέρειες συνένωσης, την ταυτότητα των ομάδων και τα επίπεδα (distances ή similarities) στα οποία πραγματοποιούνται οι συγχωνεύσεις.

Η μέθοδος Single Linkage (ή Μέθοδος της Απλής Συνένωσης) αποτελεί λοιπόν, μία από τις παλαιότερες και πιο θεμελιώδεις τεχνικές ιεραρχικής συσταδοποίησης. Ανήκει στις συγκολλητικές (agglomerative) ιεραρχικές μεθόδους και βασίζεται στη σταδιακή συνένωση των πιο κοντινών μεταξύ τους παρατηρήσεων ή ομάδων, χρησιμοποιώντας ως κριτήριο την ελάχιστη απόσταση μεταξύ στοιχείων δύο συστάδων. Σε αντίθεση με άλλες τεχνικές που επιδιώκουν τη βελτιστοποίηση κάποιας συνολικής μετρικής ομοιογένειας, η Single Linkage επικεντρώνεται στις τοπικές δομικές σχέσεις, εξετάζοντας τον μικρότερο δυνατό δεσμό μεταξύ δύο υποομάδων.

Η συγκολλητική ιεραρχική συσταδοποίηση αποτελεί μια διαδικασία όπου κάθε παρατήρηση ξεκινά ως ξεχωριστή μονάδα, και σε κάθε βήμα συνενώνονται οι δύο πιο κοντινές συστάδες. Η μέθοδος Single Linkage ορίζει την απόσταση μεταξύ δύο συστάδων A και B ως:

$$D(A, B) = \min\{d(x, y)\},$$

όπου $x \in A, y \in B$ και όπου $d(x, y)$ είναι η απόσταση ανάμεσα σε δύο σημεία.

Η πιο συνηθισμένη επιλογή για το $d(x, y)$ είναι η ευκλείδεια απόσταση, ωστόσο οποιαδήποτε συνάρτηση απόστασης που ικανοποιεί τις βασικές ιδιότητες μετρικής μπορεί να

χρησιμοποιηθεί. Η διαδικασία επαναλαμβάνεται μέχρι όλες οι παρατηρήσεις να ενωθούν σε μία συστάδα, παράγοντας έτσι ένα δενδρόγραμμα. Το κριτήριο της ελάχιστης απόστασης επιτρέπει τη δημιουργία συστάδων που χαρακτηρίζονται από «μονοπάτια» κοντινών σημείων.

Η κυριότερη αδυναμία της μεθόδους του πλησιέστερου γείτονα έγκειται στο γεγονός ότι όταν δύο διαφορετικές ομάδες έχουν κάποιο στοιχείο ή σημείο ή ένα σύνολο που τις συνδέει, τότε ο αλγόριθμος δεν είναι ικανοποιητικός γιατί δεν έχει την ιδιότητα να τις κρατήσει ξεχωριστά και τις συγχωνεύει. Έτσι, παρότι η μέθοδος έχει ορισμένες χρήσιμες μαθηματικές ιδιότητες, συνήθως παράγει ομάδες που δεν είναι συμπαγείς και συνήθως δημιουργεί μερικές πολύ μεγάλες ομάδες και κάποιες άλλες πολύ μικρές. Σύμφωνα με το γενικό αλγόριθμο που περιγράψαμε για τις συσσωρευτικές μεθόδους, στο Βήμα 3, μετά τη συγχώνευση δύο ομάδων, απαιτείται η ανανέωση του πίνακα αποστάσεων D διαγράφοντας έτσι τις γραμμές και στήλες που αντιστοιχούσαν στις ομάδες που συγχωνεύτηκαν, και προσθέτοντας μια γραμμή και μια στήλη που περιέχει τις αποστάσεις της νέας ομάδας από τις υπόλοιπες. Όταν δύο ομάδες A και B ενωθούν για να δημιουργήσουν μια νέα ομάδα, έστω $R = (AB)$, τότε όλες οι αποστάσεις μεταξύ της R και των άλλων ομάδων, έστω Q η άλλη ομάδα τότε μπορούν να προκύψουν από την παρακάτω έκφραση η οποία είναι ιδιαίτερα χρήσιμη για την ανανέωση του πίνακα αποστάσεων.

$$d(R, Q) = \min\{d(A, Q), d(B, Q)\} = \frac{1}{2}(d(A, Q) + d(B, Q)) - \frac{1}{2}|d(A, Q) - d(B, Q)|$$

Για την κατανόηση της μεθόδους δίνεται το παρακάτω παράδειγμα:

Ας θεωρήσουμε ένα σύνολο 4 ατόμων x_1, x_2, x_3, x_4 και ότι μετά τον υπολογισμό των αποστάσεων μεταξύ τους προέκυψε ο επόμενος πίνακας (έχουν παραληφθεί τα στοιχεία που βρίσκονται επάνω από τον κύριο διαγώνιο).

Ατομο	1	2	3	4
1	0			
2	8	0		
3	12	14	0	
4	13	11	10	0

Ξεκινάμε με 4 ομάδες, κάθε ομάδα αποτελούμενη από ένα άτομο δηλαδή τις (1), (2), (3), (4). Προκειμένου να ενώσουμε δύο ομάδες υπολογίζουμε την ελάχιστη απόσταση (μικρότερο στοιχείο του πίνακα)

$$\min d_{ij} = 8 = d_{12}$$

οπότε τα άτομα 1 και 2 θα ενωθούν για να αποτελέσουν την ομάδα (1,2). Άρα θα έχουμε τις επόμενες 3 ομάδες, (1, 2), 3, 4

$$d((1, 2), 3) = \min(d_{13}, d_{23}) = \min(12, 14) = 12$$

$$d((1, 2), 4) = \min(d_{14}, d_{24}) = \min(13, 11) = 11$$

Άτομο	(1, 2)	3	4
(1, 2)	0		
3	12	0	
4	11	10	0

$$\min d_{ij} = 10 = d_{34}$$

Άτομο	(1, 2)	(3, 4)
(1, 2)	0	
(3, 4)	10	0

Προχωρώντας λοιπόν στην επόμενη μέθοδο της πλήρους συνένωσης θα αναλυθεί πλήρως η λογική της. Η μέθοδος αυτή, σε σχέση με την προηγούμενη μέθοδο βασίζεται ακριβώς στην αντίθετη έκφραση της απόστασης μεταξύ δύο ομάδων. Πιο αναλυτικά η απόσταση αναμεσά σε δύο ομάδες και ως η μεγαλύτερη απόσταση από μια παρατήρηση μέσα στην μια ομάδα με κάποια άλλη παρατήρηση στην άλλη ομάδα, δηλαδή

$$d(R, Q) = \max d_{ij}, i \in R, j \in Q$$

Ο τελευταίος τύπος δικαιολογεί και τη χρήση της εναλλακτικής ορολογίας «μέθοδος του μακρινότερου γείτονα» (furthest neighbor method). Η έκφραση του τύπου ανανέωσης των αποστάσεων για τη μέθοδο αυτή έχει τη μορφή:

$$d(R, Q) = \max\{d(A, Q), d(B, Q)\} = \frac{1}{2}(d(A, Q) + d(B, Q)) + \frac{1}{2}|d(A, Q) - d(B, Q)|$$

Η μέθοδος του μακρινότερου γείτονα παρουσιάζει πρόβλημα όταν σε μία ομάδα υπάρχει έστω και ένα ζεύγος σημείων με μεγάλη απόσταση, καθώς αυτό μπορεί να αποτρέψει τη συγχώνευση ομάδων που κατά τα άλλα είναι ομοιογενείς. Σε αντίθεση με τον κοντινότερο γείτονα, που συχνά αποδίδει καλύτερα, η μέθοδος αυτή μπορεί να αποτύχει στη συνένωση στοιχείων της ίδιας εξωτερικής ομάδας. Αν και παράγει πιο συμπαγείς και μεγάλες συστάδες, δυσκολεύεται να εντοπίσει μικρές και πολύ πυκνές ομάδες, γεγονός που περιορίζει την αποτελεσματικότητά της σε δεδομένα με έντονη τοπική πυκνότητα.

4.1.2 Διαιρετικές μέθοδοι

Οι διαιρετικές μέθοδοι (Divisive Methods) αποτελούν τη δεύτερη βασική κατηγορία της ιεραρχικής ομαδοποίησης (hierarchical clustering). Σε αντίθεση με τις συσσωρευτικές μεθόδους (agglomerative methods), οι οποίες ξεκινούν από ένα σύνολο μεμονωμένων παρατηρήσεων και δημιουργούν σταδιακά μεγαλύτερες ομάδες μέσω διαδοχικών συγχωνεύσεων, οι διαιρετικές μέθοδοι ακολουθούν την αντίστροφη διαδικασία. Συγκεκριμένα, αρχικά θεωρούν ότι όλες οι παρατηρήσεις ανήκουν σε μία ενιαία ομάδα και στη συνέχεια εφαρμόζουν επαναλαμβανόμενες διασπάσεις, μέχρι να προκύψει η επιθυμητή ιεραρχική δομή ή μέχρι κάθε παρατήρηση να αποτελεί ξεχωριστή ομάδα.

Η φιλοσοφία των διαιρετικών μεθόδων βασίζεται στην αναζήτηση των σημαντικότερων διαφορών μεταξύ των παρατηρήσεων. Αντί να εξετάζεται ποια αντικείμενα είναι περισσότερο όμοια ώστε να συγχωνευθούν, εξετάζεται ποια ομάδα εμφανίζει τη μεγαλύτερη εσωτερική ανομοιογένεια και μπορεί να διαχωριστεί σε δύο περισσότερο ομοιογενείς υποομάδες. Η διαδικασία αυτή επαναλαμβάνεται αναδρομικά για κάθε νέα ομάδα που δημιουργείται, έως ότου ικανοποιηθεί κάποιο προκαθορισμένο κριτήριο τερματισμού.

Η γενική διαδικασία μιας διαιρετικής μεθόδου μπορεί να συνοψιστεί στα ακόλουθα βήματα:

Βήμα 1^ο: Όλες οι παρατηρήσεις τοποθετούνται σε μία ενιαία ομάδα.

Βήμα 2^ο: Υπολογίζεται η εσωτερική ανομοιογένεια της ομάδας και επιλέγεται εκείνη που παρουσιάζει τη μεγαλύτερη διασπορά.

Βήμα 3^ο: Η επιλεγμένη ομάδα διαχωρίζεται σε δύο νέες ομάδες σύμφωνα με κάποιο κριτήριο βελτιστοποίησης, όπως η μεγιστοποίηση της μεταξύ τους απόστασης ή η ελαχιστοποίηση της ενδοομαδικής διακύμανσης.

Βήμα 4^ο: Η ίδια διαδικασία εφαρμόζεται επαναληπτικά στις ομάδες που προκύπτουν, μέχρι να επιτευχθεί ο επιθυμητός αριθμός ομάδων ή να ικανοποιηθεί κάποιο κριτήριο τερματισμού.

Το αποτέλεσμα της διαδικασίας, όπως και στις συσσωρευτικές μεθόδους, αναπαρίσταται μέσω ενός δενδρογράμματος (dendrogram). Η διαφορά έγκειται στο γεγονός ότι το δενδρόγραμμα κατασκευάζεται από την κορυφή προς τα κάτω, ενώ στις συσσωρευτικές μεθόδους η κατασκευή πραγματοποιείται από τα κάτω προς τα πάνω. Το ύψος κάθε διαχωρισμού στο δενδρόγραμμα αντιστοιχεί στο επίπεδο ανομοιογένειας που απαιτήθηκε για τη διάσπαση της αντίστοιχης ομάδας.

Ένα από τα σημαντικότερα πλεονεκτήματα των διαιρετικών μεθόδων είναι η δυνατότητα εντοπισμού μεγάλων φυσικών διαχωρισμών που ενδέχεται να υπάρχουν στα δεδομένα ήδη από τα πρώτα στάδια της διαδικασίας. Εφόσον η αρχική διάσπαση πραγματοποιηθεί σωστά, οι επόμενες διασπάσεις τείνουν να παράγουν ομάδες με υψηλή εσωτερική συνοχή και σαφή διαχωρισμό μεταξύ τους. Για τον λόγο αυτό οι διαιρετικές τεχνικές μπορούν να αποδώσουν ιδιαίτερα ικανοποιητικά σε σύνολα δεδομένων που παρουσιάζουν έντονη φυσική ομαδοποίηση.

Ωστόσο, οι διαιρετικές μέθοδοι παρουσιάζουν και ορισμένα μειονεκτήματα. Η αναζήτηση της βέλτιστης διάσπασης σε κάθε στάδιο αποτελεί υπολογιστικά απαιτητικό πρόβλημα, καθώς υπάρχουν πολλές πιθανές διασπάσεις μιας ομάδας σε δύο υποομάδες. Για ένα σύνολο με p παρατηρήσεις, ο αριθμός των πιθανών διαχωρισμών αυξάνεται εκθετικά όσο αυξάνεται το μέγεθος του συνόλου, γεγονός που καθιστά απαγορευτική την εξαντλητική αναζήτηση της βέλτιστης λύσης για μεγάλα δεδομένα. Ως αποτέλεσμα, οι περισσότεροι διαιρετικοί αλγόριθμοι βασίζονται σε ευρετικές (heuristic) τεχνικές, οι οποίες παρέχουν ικανοποιητικές αλλά όχι πάντοτε βέλτιστες λύσεις.

Ο πιο γνωστός εκπρόσωπος των διαιρετικών μεθόδων είναι ο αλγόριθμος DIANA (Divisive Analysis), ο οποίος προτάθηκε από τους Kaufman και Rousseeuw (1990). Ο αλγόριθμος DIANA ξεκινά θεωρώντας όλες τις παρατηρήσεις ως μία ενιαία ομάδα. Στη συνέχεια επιλέγει την ομάδα με τη μεγαλύτερη εσωτερική ανομοιογένεια και εντοπίζει την παρατήρηση που

διαφέρει περισσότερο από τις υπόλοιπες. Η παρατήρηση αυτή χρησιμοποιείται ως αρχικός πυρήνας μιας νέας ομάδας, ενώ οι υπόλοιπες παρατηρήσεις εξετάζονται διαδοχικά ώστε να αποφασιστεί σε ποια από τις δύο ομάδες παρουσιάζουν μεγαλύτερη ομοιότητα. Η διαδικασία επαναλαμβάνεται μέχρι να μην απαιτούνται περαιτέρω μετακινήσεις στοιχείων μεταξύ των δύο ομάδων.

Η λειτουργία του DIANA μπορεί να περιγράφει συνοπτικά ως εξής:

- Δημιουργείται μία ομάδα που περιλαμβάνει όλες τις παρατηρήσεις.
- Επιλέγεται η ομάδα με τη μεγαλύτερη ανομοιογένεια.
- Εντοπίζεται η παρατήρηση που εμφανίζει τη μεγαλύτερη μέση απόσταση από τις υπόλοιπες και δημιουργεί μία νέα ομάδα.
- Οι υπόλοιπες παρατηρήσεις εξετάζονται μία προς μία και μεταφέρονται στη νέα ομάδα όταν παρουσιάζουν μεγαλύτερη ομοιότητα με αυτήν παρά με την αρχική ομάδα.
- Η διαδικασία συνεχίζεται μέχρι να ολοκληρωθεί η διάσπαση της ομάδας και εφαρμόζεται εκ νέου στις υπόλοιπες ομάδες.

Παρότι ο DIANA παράγει συχνά υψηλής ποιότητας ιεραρχικές δομές, η υπολογιστική του πολυπλοκότητα είναι σημαντικά μεγαλύτερη σε σχέση με εκείνη των συσσωρευτικών αλγορίθμων. Για τον λόγο αυτό η χρήση του περιορίζεται κυρίως σε σύνολα δεδομένων μικρού ή μεσαίου μεγέθους ή σε εφαρμογές όπου η ποιότητα της ομαδοποίησης είναι σημαντικότερη από τον χρόνο εκτέλεσης.

Συνοψίζοντας, οι διαιρετικές μέθοδοι αποτελούν μία εναλλακτική προσέγγιση της ιεραρχικής ομαδοποίησης, βασισμένη στη σταδιακή διάσπαση των ομάδων αντί της συγχώνευσής τους. Αν και είναι λιγότερο διαδεδομένες στην πράξη λόγω της αυξημένης υπολογιστικής πολυπλοκότητας, μπορούν να προσφέρουν ιδιαίτερα αξιόπιστα αποτελέσματα σε προβλήματα όπου η αρχική δομή των δεδομένων ευνοεί τη διαδοχική διάσπαση των ομάδων. Για τον λόγο αυτό αποτελούν σημαντικό τμήμα της θεωρίας της ιεραρχικής ομαδοποίησης και συμπληρώνουν τις συσσωρευτικές μεθόδους, παρέχοντας μία διαφορετική αλλά εξίσου ουσιαστική προσέγγιση στην ανάλυση δεδομένων.

4.2 Μη Ιεραρχικές μέθοδοι

Οι μη ιεραρχικές μέθοδοι ομαδοποίησης στοχεύουν άμεσα στον διαμερισμό των δεδομένων σε K συστάδες (όπου K είτε δίνεται εξ αρχής είτε εκτιμάται), βελτιστοποιώντας κάποιο κριτήριο. Σε αντίθεση με τις ιεραρχικές, δεν παράγουν απαραίτητα μια δομή δένδρογράμματος, αλλά τυπικά αποδίδουν έναν μοναδικό διαμερισμό σε K ομάδες. Οι πιο γνωστοί αλγόριθμοι σε αυτή την κατηγορία είναι ο k -means και ο k -medoids (PAM), καθώς και παραλλαγές τους. Επίσης, εδώ μπορούμε να συμπεριλάβουμε αλγορίθμους όπως οι Self-Organizing Maps (SOM), που δημιουργούν ένα προσχεδιασμένο δίκτυο “υποδοχέων” (neurons) ως κέντρα συστάδων.

4.2.1 K-Means και K-Medoids

Ο αλγόριθμος k-means είναι ίσως η πιο δημοφιλής μέθοδος clustering. Δέχεται ως παράμετρο τον αριθμό συστάδων K και προσπαθεί να τοποθετήσει K κέντρα (centroids) έτσι ώστε να ελαχιστοποιηθεί η ενδοσυσταδική διασπορά - συνήθως μετρώντας την με το άθροισμα τετραγώνων των αποστάσεων κάθε σημείου από το πλησιέστερο κέντρο (within-cluster SSE). Η διαδικασία συνήθως ξεκινά με τυχαία αρχικοποίηση των K κέντρων και εναλλάσσει δύο βήματα:

(α) Assignment step - απόδοση κάθε σημείου στο εγγύτερο κέντρο,

(β) Update step - ανακαθορισμός της θέσης των κέντρων (υπολογίζοντας τον μέσο όρο των σημείων κάθε συστάδας).

Τα βήματα επαναλαμβάνονται έως ότου δεν αλλάζει η απόδοση σημείων ή έως κάποιο μέγιστο αριθμό επαναλήψεων. Ο αλγόριθμος εγγυάται ότι θα συγκλίνει σε ένα τοπικό ελάχιστο της συνάρτησης κόστους (συνήθως μέσα σε λογικό αριθμό επαναλήψεων), αν και το αποτέλεσμα μπορεί να εξαρτάται από την αρχική τυχαία τοποθέτηση των κέντρων (γι' αυτό συχνά τρέχουμε το k-means πολλαπλές φορές και επιλέγουμε το καλύτερο αποτέλεσμα).

Ο αλγόριθμος k-means βασίζεται στην ευκλείδεια απόσταση, καθώς ο αριθμητικός μέσος ελαχιστοποιεί το άθροισμα των τετραγώνων των αποστάσεων. Κατά συνέπεια, οι συστάδες που προκύπτουν έχουν σφαιρικό σχήμα και ορίζονται ως περιοχές Voronoi γύρω από τα κεντροειδή. Ωστόσο, όταν χρησιμοποιούνται άλλες μετρικές (π.χ. Manhattan), ο μέσος όρος δεν αποτελεί κατάλληλο εκπρόσωπο, γεγονός που περιορίζει την εφαρμογή του κλασικού k-means πέρα από την ευκλείδεια απόσταση.

Για την αντιμετώπιση αυτού του περιορισμού χρησιμοποιείται ο k-medoids, όπου ως κέντρο κάθε συστάδας επιλέγεται ένα πραγματικό σημείο (medoid) που ελαχιστοποιεί το άθροισμα αποστάσεων από τα υπόλοιπα. Η προσέγγιση αυτή επιτρέπει τη χρήση οποιουδήποτε μέτρου απόστασης, αυξάνοντας την ευελιξία και την ερμηνευσιμότητα των αποτελεσμάτων. Ο k-medoids υλοποιείται συνήθως με τον αλγόριθμο PAM, ο οποίος μέσω επαναληπτικών ανταλλαγών βελτιστοποιεί τη συνολική λύση.

Συγκριτικά, ο k-means είναι ταχύτερος και αποδοτικός σε δεδομένα με σφαιρικές και καλά διαχωρισμένες συστάδες, αλλά είναι ευαίσθητος σε ακραίες τιμές και περιορίζεται στην ευκλείδεια μετρική. Αντίθετα, ο k-medoids είναι πιο ανθεκτικός σε θόρυβο και επιτρέπει διαφορετικές μετρικές, αλλά έχει μεγαλύτερο υπολογιστικό κόστος και μπορεί να συγκλίνει σε διαφορετικά τοπικά βέλτιστα. Επομένως, ο k-means προτιμάται για μεγάλα και “καθαρά” δεδομένα, ενώ ο k-medoids είναι καταλληλότερος όταν υπάρχουν outliers ή απαιτείται εναλλακτική έννοια απόστασης.

4.2.2 Self - Organizing Maps (SOM)

Οι αυτό-οργανωμένοι χάρτες (Self-Organizing Maps - SOM) ή χάρτες Kohonen, αποτελούν μέθοδο ομαδοποίησης βασισμένη σε νευρωνικά δίκτυα μη επιβλεπόμενης μάθησης. Ο SOM επιτελεί ταυτόχρονα μείωση διαστασιμότητας και ομαδοποίηση, προβάλλοντας πολυδιάστατα δεδομένα σε έναν συνήθως δισδιάστατο χάρτη, διατηρώντας την δομή.

Κάθε νευρώνας του πλέγματος αντιστοιχεί σε ένα διάνυσμα βαρών και λειτουργεί ως εκπρόσωπος μιας ομάδας. Η εκπαίδευση γίνεται με ανταγωνιστική μάθηση για κάθε παρατήρηση εντοπίζεται ο κοντινότερος νευρώνας (Best Matching Unit - BMU), και στη συνέχεια ενημερώνονται τόσο αυτός όσο και οι γειτονικοί του νευρώνες, ώστε να προσεγγίσουν το δεδομένο εισόδου. Η διαδικασία επαναλαμβάνεται με σταδιακή μείωση του ρυθμού μάθησης και της γειτονιάς, οδηγώντας στη δημιουργία οργανωμένων συστάδων στον χώρο.

Οι SOM είναι ιδιαίτερα χρήσιμοι για οπτικοποίηση και ανάλυση πολυδιάστατων δεδομένων, καθώς αποδίδουν όχι μόνο τις ομάδες αλλά και τις μεταξύ τους σχέσεις. Σε αντίθεση με άλλους αλγόριθμους, παρέχουν μια δομημένη αναπαράσταση όπου οι συστάδες έχουν χωρική εγγύτητα, επιτρέποντας την κατανόηση σταδιακών μεταβάσεων μεταξύ ομάδων.

Η επιλογή του BMU βασίζεται συνήθως στην ευκλείδεια απόσταση, αν και μπορούν να χρησιμοποιηθούν και άλλες μετρικές. Ο αλγόριθμος απαιτεί τον καθορισμό παραμέτρων, όπως το μέγεθος του πλέγματος, ο ρυθμός μάθησης και η ακτίνα γειτονιάς. Δεν απαιτεί προκαθορισμένο αριθμό συστάδων, αν και συχνά εφαρμόζονται επιπλέον τεχνικές όπως k-means στα βάρη των νευρώνων για περαιτέρω ομαδοποίηση.

Συνολικά, οι SOM αποτελούν μια ευέλικτη μέθοδο που συνδυάζει clustering και μείωση διαστασιμότητας, με εφαρμογές σε πεδία όπως η βιοπληροφορική και η οικονομική ανάλυση. Ωστόσο, είναι πιο σύνθετοι σε σχέση με κλασικούς αλγόριθμους και απαιτούν προσεκτική ρύθμιση των παραμέτρων.

4.3 Πυκνωτικές και βασισμένες σε πυρήνες μέθοδοι

Οι μέθοδοι ομαδοποίησης που παρουσιάστηκαν έως τώρα, όπως οι ιεραρχικές τεχνικές, οι αλγόριθμοι k-means / k-medoids και οι Self-Organizing Maps, τείνουν να εντοπίζουν κυρίως συστάδες με σχετικά κανονικό ή σφαιρικό σχήμα ή να εξαρτώνται από παραμέτρους που πρέπει να καθοριστούν εκ των προτέρων, όπως ο αριθμός των ομάδων. Αντίθετα, οι πυκνωτικές μέθοδοι προσεγγίζουν την ομαδοποίηση με διαφορετική λογική, ορίζουν τις συστάδες ως περιοχές υψηλής πυκνότητας, οι οποίες διαχωρίζονται μεταξύ τους από περιοχές χαμηλής πυκνότητας.

Η προσέγγιση αυτή επιτρέπει την ανίχνευση συστάδων αυθαίρετου σχήματος, τη διαχείριση θορύβου και ακραίων τιμών, καθώς και την εξαγωγή ομάδων χωρίς να απαιτείται προκαθορισμός του πλήθους τους. Παράλληλα, οι μέθοδοι που βασίζονται σε πυρήνες ή πιθανοτικά μοντέλα προσφέρουν μια πιο ευέλικτη ή πιο στατιστικά θεμελιωμένη αντιμετώπιση της δομής των δεδομένων, ιδιαίτερα όταν αυτή δεν είναι γραμμική ή δεν μπορεί να περιγραφεί ικανοποιητικά από απλά γεωμετρικά σχήματα.

Στην ενότητα αυτή παρουσιάζονται οι βασικότεροι αλγόριθμοι των παραπάνω κατηγοριών, με έμφαση στη λογική τους, στα πλεονεκτήματα και στους περιορισμούς τους, καθώς και στον ρόλο που διαδραματίζει το μέτρο απόστασης ή η έννοια της πυκνότητας στην απόδοσή τους.

4.3.1 DBSCAN και OPTICS

Ο DBSCAN (Density-Based Spatial Clustering of Applications with Noise) αποτελεί έναν από τους σημαντικότερους αλγορίθμους πυκνωτικής ομαδοποίησης. Η βασική του ιδέα είναι ότι μία συστάδα αντιστοιχεί σε συνδεδεμένη περιοχή του χώρου όπου τα σημεία είναι αρκετά πυκνά. Για τη λειτουργία του απαιτούνται δύο παράμετροι η ακτίνα γειτονιάς ϵ και ο ελάχιστος αριθμός σημείων $minPts$. Ένα σημείο θεωρείται κεντρικό (core point) όταν διαθέτει τουλάχιστον $minPts$ γείτονες μέσα στην ακτίνα ϵ . Από τέτοια σημεία ξεκινά η επέκταση των συστάδων, ενώ σημεία που δεν ανήκουν σε καμία πυκνή περιοχή χαρακτηρίζονται ως θόρυβος.

Το βασικό πλεονέκτημα του DBSCAN είναι ότι μπορεί να εντοπίζει συστάδες αυθαίρετου σχήματος και να απομονώνει ακραίες παρατηρήσεις χωρίς να απαιτεί εκ των προτέρων αριθμό ομάδων. Ωστόσο, η αποτελεσματικότητα του εξαρτάται έντονα από την επιλογή των παραμέτρων. Αν η τιμή του ϵ είναι πολύ μικρή, πολλά σημεία μπορεί να χαρακτηριστούν ως θόρυβος, αν είναι πολύ μεγάλη, διαφορετικές ομάδες ενδέχεται να συγχωνευθούν. Το πρόβλημα γίνεται εντονότερο όταν τα δεδομένα παρουσιάζουν μεταβαλλόμενη πυκνότητα.

Για την αντιμετώπιση αυτού του περιορισμού προτάθηκε ο αλγόριθμος OPTICS (Ordering Points To Identify the Clustering Structure). Ο OPTICS δημιουργεί μία ταξινομημένη διάταξη των σημείων και υπολογίζει για καθένα μία τιμή προσβασιμότητας (reachability distance). Με τον τρόπο αυτό αποτυπώνει τη δομή πυκνότητας των δεδομένων σε πολλαπλές κλίμακες. Οι πυκνές περιοχές εμφανίζονται ως χαμηλές τιμές στο αντίστοιχο διάγραμμα, ενώ οι μεταβάσεις μεταξύ συστάδων αποδίδονται με μεγαλύτερες τιμές.

Σε σχέση με τον DBSCAN, ο OPTICS είναι πιο κατάλληλος όταν οι συστάδες δεν έχουν ενιαία πυκνότητα, καθώς παρέχει πιο πλούσια πληροφορία για τη δομή των δεδομένων. Παρότι χρησιμοποιεί παρόμοιες παραμέτρους, επιτρέπει μεγαλύτερη ευελιξία στην ερμηνεία του αποτελέσματος, και στις δύο μεθόδους, η επιλογή της μετρικής απόστασης είναι κρίσιμη, αφού καθορίζει ποια σημεία θεωρούνται γειτονικά. Συνολικά, ο DBSCAN είναι ιδιαίτερα αποτελεσματικός όταν υπάρχουν σαφείς πυκνές περιοχές και θόρυβος, ενώ ο OPTICS υπερέχει όταν η πυκνότητα μεταβάλλεται και απαιτείται ανάλυση της ομαδοποίησης σε περισσότερες από μία κλίμακες.

4.3.2 Mean Shift και Gaussian Mixture Models

Ο Mean Shift αποτελεί μη παραμετρική μέθοδο ομαδοποίησης που βασίζεται στην εκτίμηση της πυκνότητας των δεδομένων. Η βασική του ιδέα είναι ότι κάθε παρατήρηση μετακινείται επαναληπτικά προς την τοπική μέση τιμή της γειτονιάς της, όπως αυτή προκύπτει μέσω μίας συνάρτησης πυρήνα και μιας παραμέτρου εύρους h (bandwidth). Η διαδικασία αυτή συνεχίζεται έως ότου το σημείο συγκλίνει σε μία περιοχή τοπικού μεγίστου της πυκνότητας. Όσα σημεία συγκλίνουν στο ίδιο μέγιστο θεωρούνται ότι ανήκουν στην ίδια συστάδα.

Το κύριο πλεονέκτημα του Mean Shift είναι ότι δεν απαιτεί προκαθορισμένο αριθμό συστάδων και μπορεί να ανιχνεύσει ομάδες αυθαίρετου σχήματος. Παράλληλα, προσφέρει μία πιο “συνεχή” προσέγγιση της ομαδοποίησης, καθώς βασίζεται άμεσα στο τοπίο της πυκνότητας. Ωστόσο, η απόδοσή του εξαρτάται σημαντικά από την επιλογή του bandwidth,

μικρές τιμές οδηγούν σε πολλές μικρές συστάδες, ενώ μεγάλες τιμές τείνουν να συγχωνεύουν ομάδες. Επιπλέον, ο αλγόριθμος είναι υπολογιστικά απαιτητικός, ιδίως σε μεγάλα σύνολα δεδομένων. Συνήθως εφαρμόζεται σε ευκλείδειο χώρο, αλλά μπορεί να προσαρμοστεί και σε άλλες έννοιες απόστασης.

Τα Gaussian Mixture Models (GMM) ακολουθούν διαφορετική λογική, βασισμένη στην πιθανοτική μοντελοποίηση. Υποθέτουν ότι τα δεδομένα προέρχονται από μείγμα K πολυδιάστατων κανονικών κατανομών, όπου κάθε κατανομή αντιστοιχεί σε μία συστάδα. Σε αντίθεση με το k-means, το GMM δεν αποδίδει αυστηρά κάθε σημείο σε μία μόνο ομάδα, αλλά υπολογίζει πιθανότητες συμμετοχής σε κάθε συστάδα. Η εκτίμηση των παραμέτρων του μοντέλου πραγματοποιείται συνήθως μέσω του αλγορίθμου Expectation-Maximization (EM), ο οποίος επαναλαμβάνει δύο στάδια, υπολογισμό των πιθανοτήτων συμμετοχής και ανανέωση των παραμέτρων των κατανομών.

Ένα σημαντικό πλεονέκτημα των GMM είναι ότι μπορούν να περιγράψουν ελλειψοειδείς και γενικά πιο σύνθετες δομές από αυτές που εντοπίζει ο k-means, επειδή κάθε συστάδα διαθέτει δικό της διάνυσμα μέσων τιμών και πίνακα συνδιακύμανσης. Έτσι, προσφέρουν μεγαλύτερη ευελιξία σε δεδομένα όπου οι ομάδες δεν είναι σφαιρικές. Επιπλέον, η πιθανοκρατική φύση τους επιτρέπει την αποτύπωση αβεβαιότητας στις αναθέσεις, κάτι ιδιαίτερα χρήσιμο όταν υπάρχουν επικαλύψεις μεταξύ συστάδων.

Παρά τα πλεονεκτήματά τους, τα GMM παρουσιάζουν και περιορισμούς. Η βασική παραδοχή της κανονικότητας δεν είναι πάντοτε ρεαλιστική, ενώ ο αλγόριθμος EM μπορεί να συγκλίνει σε τοπικά μέγιστα, καθιστώντας την αρχικοποίηση σημαντικό παράγοντα. Επίσης, η επιλογή του αριθμού K παραμένει κρίσιμο ζήτημα, αν και μπορούν να χρησιμοποιηθούν κριτήρια όπως τα AIC και BIC. Από υπολογιστική άποψη, η μέθοδος είναι πιο απαιτητική από τον k-means, ιδιαίτερα σε υψηλές διαστάσεις.

Αξίζει να σημειωθεί ότι το GMM μπορεί να θεωρηθεί γενίκευση του k-means. Στην ειδική περίπτωση όπου όλες οι συνδιακυμάνσεις είναι ίδιες και σφαιρικές, το κριτήριο του μοντέλου προσεγγίζει το κριτήριο της ευκλείδειας απόστασης που χρησιμοποιεί ο k-means. Στη γενική του μορφή, όμως, το GMM βασίζεται ουσιαστικά στη Mahalanobis απόσταση, επιτρέποντας πιο ρεαλιστική αποτύπωση της γεωμετρίας των δεδομένων.

Συνολικά, ο Mean Shift είναι κατάλληλος όταν επιδιώκεται μη παραμετρική ανίχνευση πυκνών περιοχών χωρίς προκαθορισμό αριθμού ομάδων, ενώ τα GMM είναι προτιμότερα όταν απαιτείται πιθανοκρατική ερμηνεία και ευελιξία ως προς το σχήμα των συστάδων. Και οι δύο μέθοδοι επεκτείνουν ουσιαστικά το πλαίσιο της ομαδοποίησης πέρα από τις κλασικές διαμεριστικές τεχνικές, επιτρέποντας πιο σύνθετη και ρεαλιστική μοντελοποίηση των δεδομένων.

4.4 Αλγόριθμοι ομαδοποίησης στη μηχανική μάθηση

Στην ενότητα αυτή εξετάζονται δύο βασικές διαστάσεις της ομαδοποίησης στη μηχανική μάθηση. Η πρώτη αφορά τον ρόλο των μέτρων απόστασης και ομοιότητας τόσο σε επιβλεπόμενα όσο και σε μη επιβλεπόμενα μοντέλα. Η δεύτερη σχετίζεται με σύγχρονες

προσεγγίσεις, στις οποίες η ομαδοποίηση συνδυάζεται με embeddings και νευρωνικά δίκτυα, ώστε να αντιμετωπίζονται αποτελεσματικότερα δεδομένα υψηλής διάστασης και σύνθετης δομής.

4.4.1 Χρήση αποστάσεων σε Supervised και Unsupervised Learning

Τα μέτρα απόστασης και ομοιότητας αποτελούν θεμελιώδες στοιχείο της μηχανικής μάθησης, καθώς καθορίζουν τον τρόπο με τον οποίο αποτιμάται η εγγύτητα ή η διαφοροποίηση μεταξύ παρατηρήσεων. Στα επιβλεπόμενα μοντέλα, χαρακτηριστικό παράδειγμα αποτελεί ο αλγόριθμος k-Nearest Neighbors (k-NN), όπου η πρόβλεψη βασίζεται στους πλησιέστερους γείτονες μιας νέας παρατήρησης. Στην περίπτωση αυτή, η επιλογή της μετρικής, όπως η ευκλείδεια ή η Manhattan απόσταση, επηρεάζει άμεσα το αποτέλεσμα της ταξινόμησης. Αντίστοιχα, σε μεθόδους πυρήνων, όπως τα Support Vector Machines, η συνάρτηση πυρήνα λειτουργεί ως μέτρο ομοιότητας, συνδέοντας την έννοια της απόστασης με τον μετασχηματισμό των δεδομένων σε νέο χώρο χαρακτηριστικών.

Στο μη επιβλεπόμενο πλαίσιο, οι αποστάσεις διαδραματίζουν εξίσου σημαντικό ρόλο. Στην ομαδοποίηση καθορίζουν τη σύνθεση των συστάδων, ενώ σε μεθόδους όπως η πολυδιάστατη κλιμάκωση (MDS) και η ανάλυση κύριων συνιστωσών (PCA) συνδέονται με τη γεωμετρική αναπαράσταση και τη διασπορά των δεδομένων. Παράλληλα, χρησιμοποιούνται και στην αξιολόγηση της ποιότητας των συστάδων, μέσω δεικτών όπως ο silhouette, ο οποίος βασίζεται στην ενδοομαδική συνοχή και τη μεταξύ ομάδων διαχωριστικότητα.

Ιδιαίτερη σημασία έχει και η μάθηση μετρικής (metric learning), όπου στόχος δεν είναι απλώς η χρήση μιας υπάρχουσας απόστασης, αλλά η εκμάθηση μιας νέας συνάρτησης απόστασης που αποτυπώνει καλύτερα τις πραγματικές σχέσεις μεταξύ των δεδομένων. Η προσέγγιση αυτή εφαρμόζεται σε προβλήματα όπως η αναγνώριση προσώπου, τα συστήματα σύστασης και η ανάκτηση πληροφορίας, όπου η απλή ευκλείδεια απόσταση συχνά δεν επαρκεί.

Η επιλογή κατάλληλης μετρικής εξαρτάται από τη φύση των δεδομένων. Για παράδειγμα, σε δεδομένα κειμένου προτιμώνται συχνά μέτρα όπως η cosine similarity, ενώ σε δεδομένα γονιδιακής έκφρασης χρησιμοποιούνται συχνά μετρικές βασισμένες στη συσχέτιση, όταν ενδιαφέρει περισσότερο η μορφή του προφίλ παρά οι απόλυτες τιμές. Συνεπώς, η μετρική δεν αποτελεί ουδέτερο τεχνικό στοιχείο, αλλά κρίσιμο παράγοντα που επηρεάζει την απόδοση, την ερμηνεία και την εγκυρότητα των αποτελεσμάτων.

Συνολικά, οι αποστάσεις και οι ομοιότητες αποτελούν κοινό υπόβαθρο για πλήθος μεθόδων της μηχανικής μάθησης. Εμφανίζονται ως κριτήρια γειννίας, ως βάση μετασχηματισμών και ως αντικείμενο βελτιστοποίησης, γεγονός που καθιστά την ορθή επιλογή τους απαραίτητη για κάθε αξιόπιστη εφαρμογή.

4.4.2 Clustering με Embeddings και νευρωνικά δίκτυα

Η εξέλιξη των νευρωνικών δικτύων και των τεχνικών αναπαράστασης δεδομένων έχει οδηγήσει στην ανάπτυξη μεθόδων ομαδοποίησης που βασίζονται σε embeddings. Η βασική ιδέα είναι ότι τα αρχικά δεδομένα, ιδιαίτερα όταν είναι υψηλής διάστασης ή παρουσιάζουν

σύνθετες μη γραμμικές σχέσεις, δεν είναι πάντοτε κατάλληλα για άμεση εφαρμογή κλασικών αλγορίθμων clustering. Για τον λόγο αυτό, χρησιμοποιούνται νευρωνικά δίκτυα ώστε να μάθουν μια πιο συμπίεσμένη και ουσιαστική αναπαράσταση των δεδομένων, μέσα στην οποία η ομαδοποίηση γίνεται αποτελεσματικότερη.

Μία χαρακτηριστική προσέγγιση είναι το Deep Embedded Clustering (DEC), όπου αρχικά εκπαιδεύεται ένας autoencoder για να μάθει μια συμπίεσμένη αναπαράσταση των δεδομένων και στη συνέχεια ο χώρος αυτός βελτιστοποιείται με βάση έναν στόχο ομαδοποίησης. Έτσι, η διαδικασία δεν περιορίζεται στην εφαρμογή clustering στα αρχικά χαρακτηριστικά, αλλά επιτρέπει στο ίδιο το δίκτυο να διαμορφώσει έναν χώρο όπου οι ομάδες γίνονται πιο διακριτές.

Παρόμοια λογική ακολουθούν και άλλες σύγχρονες μέθοδοι, όπως τα Variational Deep Embedding και Deep Clustering Networks, όπου συνδυάζονται στόχοι ανακατασκευής και ομαδοποίησης. Επιπλέον, σε πολλές εφαρμογές η ομαδοποίηση δεν γίνεται απευθείας με νευρωνικό μοντέλο αλλά πάνω σε embeddings που έχουν ήδη εξαχθεί από άλλα δίκτυα. Για παράδειγμα, σε δεδομένα εικόνας μπορούν να χρησιμοποιηθούν embeddings από convolutional neural networks, ενώ σε δεδομένα κειμένου αναπαραστάσεις από γλωσσικά μοντέλα. Στη συνέχεια, κλασικοί αλγόριθμοι όπως ο k-means εφαρμόζονται πάνω σε αυτές τις νέες αναπαραστάσεις με πολύ καλύτερα αποτελέσματα.

Η σημασία αυτών των μεθόδων είναι ιδιαίτερα μεγάλη σε περιβάλλοντα υψηλής διαστατικότητας, όπου η κλασική έννοια της απόστασης συχνά αποδυναμώνεται. Τα embeddings επιτρέπουν τη συγκράτηση της ουσιαστικής πληροφορίας και τη μείωση του θορύβου, ώστε η απόσταση στον νέο χώρο να αντανakλά περισσότερο τη σημασιολογική ομοιότητα των παρατηρήσεων.

Παρά τα πλεονεκτήματά τους, οι μέθοδοι αυτές είναι πιο σύνθετες και απαιτούν περισσότερες υπερπαραμέτρους και υπολογιστικούς πόρους σε σχέση με τις κλασικές τεχνικές clustering. Για τον λόγο αυτό δεν είναι πάντα απαραίτητες σε μικρά ή μεσαία σύνολα δεδομένων. Ωστόσο, σε πολύπλοκα δεδομένα, όπως εικόνες, κείμενα ή δίκτυα, αποτελούν ιδιαίτερα ισχυρή κατεύθυνση για σύγχρονες εφαρμογές ομαδοποίησης.

Συνολικά, οι προσεγγίσεις clustering με embeddings και νευρωνικά δίκτυα δεν αναιρούν τη σημασία των αποστάσεων και των ομοιοτήτων αντίθετα, επιδιώκουν να κατασκευάσουν έναν χώρο αναπαράστασης στον οποίο οι κλασικές έννοιες της εγγύτητας αποκτούν πιο ουσιαστικό και αξιόπιστο περιεχόμενο.

4.5 Σύνοψη και θεωρητικές συγκρίσεις τεχνικών

Στην ενότητα αυτή συνοψίζονται τα βασικά χαρακτηριστικά των κυριότερων μεθόδων ομαδοποίησης και αναδεικνύονται οι θεωρητικές τους διαφορές. Δεν υπάρχει ένας καθολικά βέλτιστος αλγόριθμος, η επιλογή εξαρτάται από τη φύση των δεδομένων, τη μορφή των συστάδων και τις απαιτήσεις της ανάλυσης.

Οι ιεραρχικές μέθοδοι δεν απαιτούν εκ των προτέρων καθορισμό του αριθμού συστάδων και προσφέρουν ευελιξία μέσω του δενδρογράμματος. Ωστόσο, είναι υπολογιστικά απαιτητικές και ευαίσθητες σε πρώιμες λανθασμένες συγχωνεύσεις. Ο k-means αποτελεί

ιδιαίτερα αποδοτική και απλή μέθοδο, κατάλληλη για μεγάλα σύνολα δεδομένων με σφαιρικές και καλά διαχωρισμένες συστάδες, αλλά απαιτεί καθορισμό του K , είναι ευαίσθητος σε outliers και εξαρτάται από την ευκλείδεια απόσταση. Ο k -medoids αποτελεί πιο ανθεκτική εναλλακτική, επιτρέποντας γενικότερες μετρικές και μειωμένη ευαισθησία σε ακραίες τιμές, με τίμημα αυξημένο υπολογιστικό κόστος.

Τα Self-Organizing Maps συνδυάζουν ομαδοποίηση και οπτικοποίηση, αποδίδοντας τη δομή των δεδομένων σε χαμηλότερη διάσταση. Αν και προσφέρουν πλούσια ερμηνεία, είναι πιο σύνθετα και δεν βασίζονται σε σαφές κριτήριο βελτιστοποίησης. Οι πυκνωτικές μέθοδοι, όπως ο DBSCAN, επιτρέπουν την ανίχνευση συστάδων αυθαίρετου σχήματος και τον εντοπισμό θορύβου χωρίς ανάγκη καθορισμού του K , αλλά εξαρτώνται έντονα από τις παραμέτρους πυκνότητας και δυσκολεύονται σε υψηλές διαστάσεις. Ο OPTICS επεκτείνει αυτή την προσέγγιση, αποτυπώνοντας τη δομή των δεδομένων σε πολλαπλές κλίμακες πυκνότητας.

Ο Mean Shift βασίζεται στην εκτίμηση πυκνότητας και δεν απαιτεί αριθμό συστάδων, αλλά είναι υπολογιστικά απαιτητικός και ευαίσθητος στην επιλογή του bandwidth. Τα Gaussian Mixture Models προσφέρουν πιθανοκρατική και πιο ευέλικτη προσέγγιση, επιτρέποντας ελλειψοειδή σχήματα και soft αναθέσεις, αλλά απαιτούν επιλογή του αριθμού συστατικών και είναι πιο απαιτητικά υπολογιστικά. Τέλος, οι μέθοδοι που βασίζονται σε embeddings και νευρωνικά δίκτυα επιτρέπουν την ανάλυση πολύπλοκων και υψηλοδιάστατων δεδομένων, βελτιώνοντας την ποιότητα των συστάδων, αλλά συνοδεύονται από αυξημένη πολυπλοκότητα και ανάγκη ρύθμισης παραμέτρων.

Συγκριτικά, οι μέθοδοι διαφοροποιούνται ως προς το σχήμα των συστάδων που μπορούν να ανιχνεύσουν, την ανθεκτικότητα σε outliers, τις υπολογιστικές απαιτήσεις και τις παραμέτρους που απαιτούν. Ο k -means και οι παραλλαγές του ευνοούν σφαιρικές δομές, ενώ μέθοδοι όπως ο DBSCAN και ο Mean Shift υποστηρίζουν αυθαίρετα σχήματα. Η ανθεκτικότητα σε θόρυβο είναι μεγαλύτερη σε μεθόδους όπως ο DBSCAN και ο k -medoids, ενώ οι ιεραρχικές και ο k -means είναι πιο ευαίσθητοι. Από υπολογιστική άποψη, ο k -means είναι από τις πιο αποδοτικές μεθόδους, ενώ οι ιεραρχικές, ο Mean Shift και τα GMM είναι πιο απαιτητικά.

Η επιλογή της κατάλληλης μεθόδου εξαρτάται από το πρόβλημα. Για μεγάλα και σχετικά κανονικά δεδομένα, ο k -means αποτελεί πρακτική επιλογή. Σε δεδομένα με θόρυβο ή μη κανονικά σχήματα, προτιμώνται πυκνωτικές μέθοδοι. Για στατιστική ερμηνεία και ευελιξία, τα GMM είναι κατάλληλα, ενώ σε υψηλή διαστατικότητα συχνά απαιτείται συνδυασμός με τεχνικές μείωσης διαστάσεων ή embeddings.

Συνολικά, η επιλογή αλγορίθμου και μετρικής απόστασης αποτελεί καθοριστικό στάδιο της ανάλυσης, καθώς επηρεάζει άμεσα την ποιότητα και την ερμηνεία των αποτελεσμάτων. Στα επόμενα κεφάλαια, οι παραπάνω μέθοδοι θα εφαρμοστούν σε πραγματικά δεδομένα, ώστε να αξιολογηθεί εμπειρικά η απόδοσή τους και να εξαχθούν συγκριτικά συμπεράσματα.

ΚΕΦΑΛΑΙΟ 5

Εισαγωγή στο πειραματικό μέρος με σύνθετα δεδομένα

Στο πέμπτο κεφάλαιο, περνάμε σε ένα πειραματικό μέρος, χρησιμοποιώντας πλήρως συνθετικά δεδομένα βάσει κάποιων σεναρίων εμπνευσμένων από οικονομικά δεδομένα. Θα πραγματοποιήσουμε ομαδοποιήσεις με διάφορες μεθόδους και θα συγκρίνουμε την αποτελεσματικότητά τους όταν χρησιμοποιούνται διαφορετικά μέτρα απόστασης. Ο στόχος είναι διπλός, αφενός να αναδείξουμε στην πράξη πώς η επιλογή μετρικής μπορεί να επηρεάσει το clustering, αφετέρου να συγκρίνουμε τις ίδιες τις μεθόδους μεταξύ τους ως προς την ικανότητα ανακάλυψης των πραγματικών ομάδων υπό διαφορετικές συνθήκες.

Θα εξετάσουμε ένα βασικό σενάριο όπου τα δεδομένα σχηματίζουν τρεις διακριτές συστάδες στο επίπεδο των 2 διαστάσεων αυτό προσομοιώνει τρεις κατηγορίες οικονομιών ή τρεις τύπους επενδυτών με βάση δύο χαρακτηριστικά. Γνωρίζουμε εκ των προτέρων την πραγματική κατάταξη των σημείων σε clusters (ground truth), γεγονός που μας επιτρέπει να αξιολογήσουμε ποσοτικά την απόδοση των αλγορίθμων. Ως βασικό μέτρο αξιολόγησης χρησιμοποιούμε τον Adjusted Rand Index (ARI), ο οποίος είναι ανεξάρτητος από την αυθαίρετη αρίθμηση των ετικετών των clusters. Επιπλέον, όπου χρησιμοποιείται το ποσοστό σωστής ταξινόμησης, προηγείται βέλτιστη αντιστοίχιση των ετικετών των εκτιμώμενων clusters με τις πραγματικές ομάδες, ώστε η αξιολόγηση να μην επηρεάζεται από διαφορετικές μεταθέσεις των ετικετών

Στη συνέχεια, θα αυξήσουμε την πολυπλοκότητα εισάγοντας ελαφρύ θόρυβο (outliers) και διαφοροποιώντας λίγο τα σχήματα των clusters κάνοντας τη μία συστάδα πιο επιμήκη ώστε να δούμε πώς ανταποκρίνονται οι μέθοδοι. Τέλος, θα πειραματιστούμε με διαφορετικά μέτρα απόστασης, όπως είναι ευκλείδεια απόσταση, Manhattan, και μια metric βασισμένη σε συσχέτιση (απόσταση Pearson) για να δούμε αν αλλάζει κάτι στον εντοπισμό των clusters.

5.1 Δημιουργία συνθετικών δεδομένων

Θα δημιουργήσουμε με τη βοήθεια της γλώσσας R ένα dataset 300 παρατηρήσεων με 2 μεταβλητές x_1, x_2 . Οι συστάδες θα διαμορφωθούν ως εξής, Cluster 1:100 σημεία περίπου γύρω από (0, 0), με κατανομή normal και με τυπική απόκλιση 0.5 και στα δύο axes. Cluster 2:100 σημεία γύρω από (5, 5), ίδια τυπική απόκλιση, Cluster 3:100 σημεία γύρω από (0, 5), ίδια τυπική απόκλιση.

Έτσι, οι τρεις ομάδες είναι σχετικά καλά διαχωρισμένες, cluster1 - cluster3 απέχουν, 5 μονάδες στο x_2 , cluster2 απέχει, 5 - 7 μονάδες από τις άλλες (διαγώνια από cluster1, οριζόντια από cluster3). Η τυπική απόκλιση 0.5 σημαίνει ότι, για κάθε άξονα ξεχωριστά, περίπου το 99,7% των παρατηρήσεων αναμένεται να βρίσκεται εντός τριών τυπικών αποκλίσεων () από την αντίστοιχη συντεταγμένη του κέντρου της συστάδας. Καθώς τα κέντρα των συστάδων

διαφέρουν κατά 5 μονάδες σε τουλάχιστον έναν από τους δύο άξονες, δηλαδή κατά 10σ, αναμένεται πολύ μικρή επικάλυψη μεταξύ των συστάδων. Αυτό το διαλέγουμε σκόπιμα ώστε οι μέθοδοι να μην έχουν δυσκολία να βρουν τα clusters.

Το διάγραμμα διασποράς 5.1. απεικονίζει με σαφήνεια την τρισυσταδική δομή των δεδομένων στον χώρο (x_1, x_2) . Παρατηρούνται τρεις συμπαγείς συγκεντρώσεις σημείων, καθεμία γύρω από το αντίστοιχο κέντρο της με σχετικά μικρή εσωτερική διασπορά λόγω της σχετικά μικρής τυπικής απόκλισης $\sigma = 0.5$. Η χωρική απομάκρυνση των ομάδων 5 μονάδες οριζόντια - κατακόρυφα και περίπου 7.07 διαγώνια, οδηγεί σε περιορισμένη επικάλυψη, γεγονός που καθιστά το παράδειγμα κατάλληλο ως σχετικά εύκολο σημείο εκκίνησης για την αξιολόγηση αλγορίθμων clustering.

5. 2 Εφαρμογή αλγορίθμων ομαδοποίησης με διάφορες αποστάσεις

Στο παρόν στάδιο, η διερεύνηση εστιάζει στην επίδραση της επιλογής μετρικής απόστασης ή ανομοιότητας στην ποιότητα της ομαδοποίησης, με βασικό κριτήριο αξιολόγησης τον δείκτη Adjusted Rand Index (ARI). Ο ARI αποτελεί καθιερωμένο μέτρο εξωτερικής εγκυρότητας (external validation), καθώς αποτιμά τον βαθμό συμφωνίας μεταξύ της παραγόμενης κατάτμησης (clustering partition) και μιας διαθέσιμης κατάτμησης αναφοράς (ground truth), διορθώνοντας για την τυχαία συμφωνία. Η μέγιστη τιμή του είναι 1, η οποία αντιστοιχεί σε πλήρη συμφωνία μεταξύ των δύο κατατμήσεων, ενώ τιμές κοντά στο 0 υποδηλώνουν συμφωνία αντίστοιχη με εκείνη που θα αναμενόταν τυχαία. Αρνητικές τιμές είναι επίσης δυνατές και υποδηλώνουν συμφωνία μικρότερη από την αναμενόμενη υπό τυχαία ανάθεση.

Για τη διερεύνηση της επίδρασης της μετρικής απόστασης εφαρμόστηκαν δύο διαδικασίες ομαδοποίησης στις εξεταζόμενες αποστάσεις ιεραρχική ομαδοποίηση με complete linkage και k-medoids μέσω του αλγορίθμου PAM, με αριθμό συστάδων $k = 3$. Για κάθε μετρική κατασκευάστηκε ο αντίστοιχος πίνακας αποστάσεων και στη συνέχεια εφαρμόστηκαν οι δύο μέθοδοι ομαδοποίησης. Για τη Minkowski χρησιμοποιήθηκε $p = 3$, η Chebyshev αντιστοιχεί στη νόρμα L_∞ , ενώ η L_1 είναι ισοδύναμη με την Manhattan. Παράλληλα εξετάστηκαν οι k-means, DBSCAN και Gaussian Mixture Model (GMM), ώστε να υπάρχει ευρύτερη εικόνα της δυνατότητας ανάκτησης της γνωστής δομής των συνθετικών δεδομένων. Η ποιότητα κάθε τελικής κατάτμησης αξιολογήθηκε μέσω του ARI, συγκρίνοντας τις παραγόμενες συστάδες με τις πραγματικές ετικέτες των δεδομένων.

Ιδιαίτερη προσοχή απαιτήθηκε για τις αποστάσεις Soergel, Kulczynski και Bhattacharyya λόγω της παρουσίας αρνητικών τιμών στο αρχικό συνθετικό σύνολο δεδομένων. Πριν από τον υπολογισμό των συγκεκριμένων αποστάσεων εφαρμόστηκε μετατόπιση (shift) ανά χαρακτηριστικό. Συγκεκριμένα, για κάθε μεταβλητή υπολογίστηκε η ελάχιστη τιμή της και όταν αυτή ήταν αρνητική, προστέθηκε σε όλες τις παρατηρήσεις της μεταβλητής η απόλυτη τιμή του ελαχίστου. Με τον τρόπο αυτό η ελάχιστη τιμή κάθε χαρακτηριστικού μεταφέρθηκε στο μηδέν, διατηρώντας παράλληλα τις μεταξύ των παρατηρήσεων διαφορές εντός του ίδιου χαρακτηριστικού. Οι Soergel και Kulczynski υπολογίστηκαν στη συνέχεια πάνω στα μετασχηματισμένα μη αρνητικά δεδομένα. Στην περίπτωση της Bhattacharyya, μετά τη μετατόπιση, κάθε συγκρινόμενο διάνυσμα κανονικοποιήθηκε επιπλέον ως προς το άθροισμα των στοιχείων του κατά τον υπολογισμό της απόστασης, ώστε να σχηματίζονται μη αρνητικά

κανονικοποιημένα διανύσματα. Με αυτόν τον τρόπο αντιμετωπίζεται ρητά ο περιορισμός που προκύπτει από τις αρνητικές τιμές του αρχικού συνθετικού δείγματος.

Τα αποτελέσματα του Πίνακα 5.1 παρουσιάζουν αρχικά μια ιδιαίτερα σαφή εικόνα για τις κλασικές μετρικές απόστασης. Η Ευκλείδεια, Manhattan, Minkowski ($p = 3$), Chebyshev (L_∞), L_1 , Mahalanobis και Gower επιτυγχάνουν $ARI = 1.000$ τόσο με την ιεραρχική ομαδοποίηση complete linkage όσο και με τον PAM. Πρόκειται για πλήρη συμφωνία της παραγόμενης κατάτμησης με τις πραγματικές ετικέτες των δεδομένων. Το εύρημα αυτό αποτελεί ισχυρή ένδειξη ότι, για το συγκεκριμένο συνθετικό σύνολο δεδομένων, η βασική γεωμετρική δομή των τριών συστάδων είναι έντονα διακριτή και μπορεί να ανακτηθεί πλήρως από διαφορετικές έννοιες απόστασης.

Ιδιαίτερο ενδιαφέρον παρουσιάζει η Mahalanobis, η οποία επίσης επιτυγχάνει $ARI = 1.000$ και με τις δύο διαδικασίες ομαδοποίησης. Σε αντίθεση με τις κλασικές αποστάσεις που βασίζονται άμεσα στις διαφορές μεταξύ των συντεταγμένων, η Mahalanobis ενσωματώνει τη δομή συνδιακύμανσης των χαρακτηριστικών και λαμβάνει υπόψη τόσο τις διαφορετικές διασπορές όσο και τις μεταξύ τους συσχετίσεις. Η πλήρης συμφωνία που παρατηρείται δείχνει ότι η προσαρμογή αυτή της γεωμετρίας του χώρου δεν μεταβάλλει την ανακτησιμότητα των πραγματικών ομάδων στη συγκεκριμένη πειραματική συνθήκη. Αντίστοιχα, η Gower επιτυγχάνει επίσης $ARI = 1.000$, γεγονός που υποδεικνύει ότι η κανονικοποίηση που ενσωματώνει η συγκεκριμένη ανομοιογένεια δεν εμποδίζει την πλήρη ανάκτηση των συστάδων.

Τα αποτελέσματα των Soergel και Kulczynski είναι ιδιαίτερα σημαντικά μετά την εφαρμογή του κατάλληλου μετασχηματισμού των δεδομένων σε μη αρνητικές τιμές. Για τη Soergel, η ιεραρχική ομαδοποίηση επιτυγχάνει $ARI = 0.734$, ενώ ο PAM εμφανίζει σαφώς υψηλότερη επίδοση, με $ARI = 0.951$. Επομένως, η συγκεκριμένη απόσταση διατηρεί σε μεγάλο βαθμό τη δομή των πραγματικών συστάδων μετά τον μετασχηματισμό, ιδιαίτερα όταν συνδυάζεται με τον PAM. Το αποτέλεσμα δείχνει ταυτόχρονα ότι η τελική ποιότητα της ομαδοποίησης δεν εξαρτάται αποκλειστικά από την επιλεγμένη απόσταση, αλλά και από τον τρόπο με τον οποίο αυτή συνδυάζεται με τον αλγόριθμο ομαδοποίησης.

Ακόμη εντονότερη είναι η διαφοροποίηση για την Kulczynski. Με complete linkage προκύπτει $ARI = 0.454$, γεγονός που αντιστοιχεί σε σαφώς χαμηλότερη συμφωνία με την κατάτμηση αναφοράς. Αντίθετα, με τον PAM ο ARI ανέρχεται σε 0.980, δηλαδή σε σχεδόν πλήρη συμφωνία. Η πολύ μεγάλη διαφορά μεταξύ των δύο αποτελεσμάτων είναι ιδιαίτερα σημαντική, καθώς αποδεικνύει ότι δεν είναι ορθό να αξιολογείται η καταλληλότητα μιας μετρικής ανεξάρτητα από τον αλγόριθμο στον οποίο ενσωματώνεται. Στην προκειμένη περίπτωση, η Kulczynski σε συνδυασμό με τον PAM διατηρεί σχεδόν εξ ολοκλήρου την πραγματική δομή των δεδομένων, ενώ η ίδια ανομοιογένεια σε συνδυασμό με complete linkage οδηγεί σε σημαντικά διαφορετική κατάτμηση.

Η Bhattacharyya παρουσιάζει περισσότερο σταθερή αλλά σαφώς χαμηλότερη συμπεριφορά μεταξύ των δύο μεθόδων, με $ARI = 0.525$ για την ιεραρχική ομαδοποίηση και $ARI = 0.544$ για τον PAM. Η Bhattacharyya διαφοροποιείται εννοιολογικά από τις κλασικές γεωμετρικές αποστάσεις, καθώς βασίζεται στον βαθμό επικάλυψης μεταξύ κανονικοποιημένων μη αρνητικών διανυσμάτων. Στην παρούσα εφαρμογή οι αρχικές παρατηρήσεις μετατοπίστηκαν αρχικά σε μη αρνητικές τιμές και στη συνέχεια κανονικοποιήθηκαν κατά τον υπολογισμό της

απόστασης. Συνεπώς, τα αποτελέσματα πρέπει να ερμηνευθούν ως αποτελέσματα της Bhattacharyya πάνω σε αυτή τη μετασχηματισμένη αναπαράσταση και όχι απευθείας πάνω στον αρχικό χώρο των χαρακτηριστικών. Οι τιμές ARI δείχνουν ότι ανακτάται μέρος της πραγματικής ομαδοποιητικής δομής, αλλά όχι στον βαθμό που παρατηρείται στις κλασικές γεωμετρικές αποστάσεις.

Η Canberra παρουσιάζει επίσης αισθητή εξάρτηση από τη μέθοδο ομαδοποίησης. Με complete linkage προκύπτει $ARI = 0.245$, ενώ με PAM η επίδοση αυξάνεται σε $ARI = 0.502$. Η συγκεκριμένη απόσταση βασίζεται σε σχετικές διαφορές μεταξύ των τιμών, κανονικοποιώντας τη διαφορά κάθε χαρακτηριστικού με βάση το άθροισμα των απόλυτων τιμών. Ως αποτέλεσμα, μπορεί να αποδίδει αυξημένη σχετική βαρύτητα σε διαφορές που εμφανίζονται κοντά στο μηδέν. Στο συγκεκριμένο σύνολο δεδομένων η συμπεριφορά αυτή δεν οδηγεί σε πλήρη ανάκτηση των πραγματικών ομάδων, αν και ο PAM αξιοποιεί τη συγκεκριμένη απόσταση αποτελεσματικότερα από το complete linkage.

Η χαμηλότερη επίδοση παρατηρείται στην απόσταση Pearson. Ο ARI ανέρχεται μόλις σε 0.143 για την ιεραρχική ομαδοποίηση και 0.142 για τον PAM. Στην περίπτωση αυτή, η απόσταση ορίστηκε με βάση το $1 - r$, όπου r είναι ο συντελεστής συσχέτισης Pearson μεταξύ δύο παρατηρήσεων. Επομένως, η σύγκριση εστιάζει περισσότερο στην ομοιότητα του μοτίβου των χαρακτηριστικών παρά στις απόλυτες γεωμετρικές διαφορές μεταξύ των σημείων. Η σχεδόν ταυτόσημη και χαμηλή επίδοση και στους δύο αλγόριθμους αποτελεί ένδειξη ότι η πληροφορία που διαχωρίζει τις πραγματικές συστάδες στο συγκεκριμένο σύνολο δεδομένων δεν αποτυπώνεται αποτελεσματικά μέσω συσχετιστικής ομοιότητας.

Παράλληλα, οι k-means, DBSCAN και GMM επιτυγχάνουν όλοι $ARI = 1.000$. Σε συνδυασμό με την πλήρη συμφωνία που παρατηρείται για μεγάλο αριθμό μετρικών τόσο στο complete linkage όσο και στον PAM, το αποτέλεσμα ενισχύει την εικόνα ότι το συγκεκριμένο συνθετικό σύνολο δεδομένων διαθέτει ισχυρή και σαφώς ανακτήσιμη ομαδοποιητική δομή. Ωστόσο, οι αποκλίσεις που εμφανίζονται στις Soergel, Kulczynski, Bhattacharyya, Canberra και Pearson δείχνουν ότι ακόμη και σε ένα τέτοιο σύνολο δεδομένων η επιλογή της αναπαράστασης της ανομοιότητας και ο συνδυασμός της με συγκεκριμένο αλγόριθμο μπορούν να επηρεάσουν σημαντικά την τελική κατάτμηση.

Συνολικά, τα αποτελέσματα οδηγούν σε δύο βασικά συμπεράσματα. Πρώτον, οι κλασικές γεωμετρικές μετρικές, καθώς και οι Mahalanobis και Gower, παρουσιάζουν εξαιρετική ευρωστία στο συγκεκριμένο πρόβλημα, επιτυγχάνοντας πλήρη συμφωνία με την κατάτμηση αναφοράς τόσο στο complete linkage όσο και στον PAM. Δεύτερον, τα αποτελέσματα καταδεικνύουν ότι η επίδοση μιας μετρικής δεν πρέπει να εξετάζεται απομονωμένα από τη μέθοδο ομαδοποίησης. Αυτό γίνεται ιδιαίτερα εμφανές στην Kulczynski, όπου ο ARI μεταβάλλεται από 0.454 με complete linkage σε 0.980 με PAM, αλλά και στη Soergel, όπου αυξάνεται από 0.734 σε 0.951. Αντίθετα, η Pearson παραμένει σταθερά χαμηλή ανεξάρτητα από τον αλγόριθμο.

Με βάση τα παραπάνω, για το υπό εξέταση συνθετικό σύνολο δεδομένων οι κλασικές μετρικές που αποτυπώνουν άμεσες γεωμετρικές διαφορές μεταξύ των παρατηρήσεων παρέχουν την πιο σταθερή συμπεριφορά. Παράλληλα, η πολύ υψηλή επίδοση των Soergel και Kulczynski όταν συνδυάζονται με τον PAM καταδεικνύει ότι δεν είναι ορθό να χαρακτηριστούν γενικά οι

συγκεκριμένες μετρικές ως ακατάλληλες. Αντιθέτως, η αποτελεσματικότητά τους εξαρτάται τόσο από την κατάλληλη προεπεξεργασία των δεδομένων όσο και από τον αλγόριθμο ομαδοποίησης με τον οποίο συνδυάζονται. Επομένως, η επιλογή μετρικής και αλγορίθμου πρέπει να αντιμετωπίζεται ως ενιαία πειραματική επιλογή και να αξιολογείται εμπειρικά με βάση το συγκεκριμένο σύνολο δεδομένων.

5.2.1 Σενάριο 1: τμηματοποίησης νοικοκυριών με δεδομένα μικτού τύπου

Στο πρώτο σενάριο εξετάζεται ένα πρόβλημα τμηματοποίησης νοικοκυριών με χρήση μικτών δεδομένων, δηλαδή δεδομένων που περιλαμβάνουν τόσο αριθμητικές όσο και κατηγορικές μεταβλητές. Στόχος είναι να διερευνηθεί κατά πόσο η επιλογή κατάλληλης μετρικής απόστασης μπορεί να αποτυπώσει αποτελεσματικά τη δομή των δεδομένων και να οδηγήσει σε ερμηνεύσιμα clusters.

Προσομοιώθηκαν 150 νοικοκυριά με τέσσερις μεταβλητές:

- δύο αριθμητικές (ετήσιο εισόδημα και αποταμίευση),
- δύο κατηγορικές (περιοχή κατοικίας και επίπεδο επενδυτικού ρίσκου).

Δεδομένου ότι το σύνολο περιλαμβάνει μικτές μεταβλητές, η άμεση εφαρμογή της Ευκλείδειας απόστασης δεν θεωρήθηκε κατάλληλη, καθώς προϋποθέτει αριθμητική αναπαράσταση των χαρακτηριστικών και, μετά από κωδικοποίηση των κατηγορικών μεταβλητών, μπορεί να επιβάλει γεωμετρικές σχέσεις που δεν αντανakλούν τη φύση των κατηγοριών. Για τον λόγο αυτό επιλέχθηκε η απόσταση Gower, η οποία επιτρέπει τον ταυτόχρονο χειρισμό αριθμητικών και κατηγορικών χαρακτηριστικών. Η ομαδοποίηση πραγματοποιήθηκε με τον αλγόριθμο k-medoids (PAM), ο οποίος λειτουργεί απευθείας πάνω σε πίνακα αποστάσεων και θεωρείται κατάλληλος για μη ευκλείδειες μετρικές. Φυσικά, ο αριθμός των clusters ορίστηκε σε $K = 3$, σύμφωνα με τη γνωστή δομή της προσομοίωσης.

Η εφαρμογή του PAM με απόσταση Gower παρήγαγε τρεις συστάδες ίσου μεγέθους (50 παρατηρήσεις ανά cluster), πίνακας 5.2.1. Η ισοκατανομή αυτή περιγράφει αποκλειστικά το μέγεθος των παραγόμενων συστάδων και δεν αποτελεί από μόνη της τεκμήριο ορθού διαχωρισμού. Η συμφωνία της παραγόμενης ομαδοποίησης με τις γνωστές ομάδες αναφοράς αξιολογείται χωριστά μέσω του Adjusted Rand Index (ARI).

Οι μέσες τιμές εισοδήματος και αποταμίευσης ανά cluster ήταν οι εξής, πίνακας 5.2.2:

Παρατηρείται σαφής κλιμάκωση από το πρώτο προς το τρίτο cluster, με τα αποτελέσματα να αντικατοπτρίζουν με συνέπεια τη σχεδιασμένη δομή των δεδομένων.

- Το **Cluster 1** αντιστοιχεί σε χαμηλό εισόδημα και χαμηλή αποταμίευση.
- Το **Cluster 2** σε μεσαίο οικονομικό προφίλ.
- Το **Cluster 3** σε υψηλό εισόδημα και σημαντικά υψηλότερη αποταμίευση.

Συνεπώς η διάκριση μεταξύ των ομάδων είναι καθαρή και οικονομικά ερμηνεύσιμη.

Όσο για τις κατηγορικές μεταβλητές έχουμε:

Η κατανομή της μεταβλητής περιοχή κατοικίας έδειξε πλήρη διαχωρισμό πίνακας 5.2.3:

- Το **Cluster 1** περιλαμβάνει αποκλειστικά Αστικές περιοχές.
- Το **Cluster 2** περιλαμβάνει αποκλειστικά Ημιαστικές περιοχές.
- Το **Cluster 3** περιλαμβάνει αποκλειστικά Αγροτικές περιοχές.

Ο διαχωρισμός αυτός επιβεβαιώνει ότι η απόσταση Gower ενσωμάτωσε αποτελεσματικά την κατηγορική πληροφορία στη διαδικασία ομαδοποίησης.

Αναφορικά με το επίπεδο επενδυτικού ρίσκου, παρατηρείται σαφής διαφοροποίηση πίνακας 5.2.4:

- Στο **Cluster 1** κυριαρχεί το χαμηλό ρίσκο.
- Στο **Cluster 2** επικρατεί το μεσαίο ρίσκο, με μικρή παρουσία υψηλού.
- Στο **Cluster 3** εμφανίζεται έντονη παρουσία υψηλού ρίσκου και απουσία χαμηλού.

Η παρατηρούμενη διαφοροποίηση του επενδυτικού κινδύνου μεταξύ των συστάδων είναι συνεπής με τον τρόπο κατασκευής των συνθετικών δεδομένων, στον οποίο το υψηλότερο οικονομικό προφίλ συνδέθηκε με αυξημένη πιθανότητα εμφάνισης υψηλότερου επιπέδου επενδυτικού κινδύνου. Συνεπώς, το συγκεκριμένο αποτέλεσμα αποτελεί ιδιότητα του σχεδιασμένου σεναρίου προσομοίωσης και δεν πρέπει να ερμηνεύεται ως εμπειρικό συμπέρασμα ή ως γενικεύσιμη οικονομική συμπεριφορά πραγματικών νοικοκυριών.

Συμπερασματικά, στο συγκεκριμένο συνθετικό σενάριο η χρήση της απόστασης Gower επέτρεψε την ενιαία διαχείριση αριθμητικών και κατηγορικών χαρακτηριστικών χωρίς να απαιτείται η επιβολή κοινής αριθμητικής γεωμετρίας στις κατηγορικές μεταβλητές. Τα αποτελέσματα αφορούν τη συγκεκριμένη κατασκευή και πειραματική ρύθμιση και δεν συνιστούν από μόνα τους τεκμήριο αντίστοιχης οικονομικής συμπεριφοράς σε πραγματικά δεδομένα νοικοκυριών.

5.2.2 Σενάριο 2: ομαδοποίηση δυαδικών δεδομένων με απόσταση Jaccard

Στο δεύτερο σενάριο εξετάζεται η ομαδοποίηση πελατών βάσει της κατοχής χρηματοοικονομικών προϊόντων. Σε αντίθεση με το προηγούμενο παράδειγμα μικτών δεδομένων, εδώ οι μεταβλητές είναι αποκλειστικά δυαδικές (0-1), γεγονός που απαιτεί διαφορετική προσέγγιση στη μέτρηση της ομοιότητας.

Προσομοιώθηκαν 80 πελάτες και για κάθε έναν καταγράφηκε αν διαθέτει τα εξής προϊόντα:

- Πιστωτική κάρτα
- Στεγαστικό δάνειο
- Καταναλωτικό δάνειο
- Επενδυτικό λογαριασμό
- Ασφαλιστικό προϊόν

Η τιμή 1 δηλώνει κατοχή του προϊόντος και η τιμή 0 απουσία.

Σε δυαδικά δεδομένα τύπου κατοχή - μη κατοχή, η κοινή απουσία ενός προϊόντος (0-0) δεν είναι ιδιαίτερα πληροφοριακή. Για τον λόγο αυτό χρησιμοποιήθηκε η απόσταση Jaccard, η οποία εστιάζει μόνο στις περιπτώσεις όπου τουλάχιστον ένας από τους δύο πελάτες διαθέτει το

προϊόν. Με τον τρόπο αυτό, η ομοιότητα βασίζεται ουσιαστικά στις κοινές κατοχές και όχι στις κοινές απουσίες.

Η ομαδοποίηση πραγματοποιήθηκε με ιεραρχικό clustering και κριτήριο σύνδεσης complete linkage. Ο αριθμός των συστάδων δεν επιλέχθηκε αυθαίρετα, αλλά προσδιορίστηκε με βάση τη δομή του δένδρογράμματος. Ειδικότερα, εξετάστηκαν τα ύψη στα οποία πραγματοποιούνται οι διαδοχικές συγχωνεύσεις των παρατηρήσεων και επιλέχθηκε οριζόντια τομή σε απόσταση Jaccard περίπου 0,70. Στο συγκεκριμένο επίπεδο διακρίνονται τρεις διαφορετικοί κλάδοι και επομένως, επιλέχθηκε ($K=3$), δηλαδή η δημιουργία τριών συστάδων πελατών. Η επιλογή αυτή επιτρέπει τη διατήρηση διακριτών προτύπων κατοχής χρηματοοικονομικών προϊόντων πριν από τη συγχώνευσή τους σε μεγαλύτερες και περισσότερο ετερογενείς ομάδες.

Η κατανομή των πελατών στις τρεις συστάδες και τα αντίστοιχα ποσοστά κατοχής των πέντε χρηματοοικονομικών προϊόντων μπορούν στη συνέχεια να χρησιμοποιηθούν για την ερμηνεία και την περιγραφή του προφίλ κάθε ομάδας. Με αυτόν τον τρόπο, κάθε cluster χαρακτηρίζεται με βάση το συγκεκριμένο πρότυπο προϊόντων που εμφανίζουν οι πελάτες που ανήκουν σε αυτό.

Ιδιαίτερη περίπτωση αποτελούν οι πελάτες που δεν διαθέτουν κανένα από τα εξεταζόμενα προϊόντα. Στην παρούσα υλοποίηση της απόστασης Jaccard, για δύο μηδενικά διανύσματα υιοθετείται η σύμβαση ότι η μεταξύ τους απόσταση ισούται με 0, ώστε δύο πελάτες χωρίς καμία κατοχή προϊόντος να θεωρούνται ταυτόσημοι ως προς τα εξεταζόμενα χαρακτηριστικά. Αντίθετα, η απόσταση ενός μηδενικού διανύσματος από οποιοδήποτε μη μηδενικό διάνυσμα ισούται με 1, καθώς δεν υπάρχει κοινή κατοχή προϊόντος. Η συγκεκριμένη σύμβαση συμβάλλει στον σαφή διαχωρισμό των πελατών αυτών από τις υπόλοιπες ομάδες. Συνεπώς, η επιλογή των οκτώ συστάδων βασίζεται στην πραγματική δομή του δένδρογράμματος και στις αποστάσεις Jaccard μεταξύ των πελατών και δεν αποτελεί έναν τυχαία επιλεγμένο αριθμό ομάδων.

Από τον πίνακα 5.2.6 προκύπτει ότι το Cluster 1 συγκεντρώνει τη μεγάλη πλειονότητα των πελατών και χαρακτηρίζεται από σχετικά υψηλή διείσδυση βασικών τραπεζικών προϊόντων, όπως πιστωτική κάρτα (82%) και ασφάλιση (58%), καθώς και μέτρια παρουσία δανειακών και επενδυτικών προϊόντων. Το Cluster 2 περιλαμβάνει πελάτες χωρίς κατοχή κανενός από τα εξεταζόμενα προϊόντα, γεγονός που το καθιστά σαφώς διακριτό από τις υπόλοιπες ομάδες. Το Cluster 3 εμφανίζει ιδιαίτερο επενδυτικό προσανατολισμό, καθώς το 100% των πελατών διαθέτει επενδυτικό λογαριασμό και το 67% ασφαλιστικό προϊόν, χωρίς όμως συμμετοχή σε δανειακά προϊόντα. Πρόκειται για μικρή αλλά σαφώς διαφοροποιημένη ομάδα.

Συμπερασματικά, το παράδειγμα αυτό αναδεικνύει τη σημασία επιλογής κατάλληλης μετρικής σε δυαδικά δεδομένα. Η χρήση της απόστασης Jaccard επιτρέπει την ομαδοποίηση βάσει πραγματικών κοινών χαρακτηριστικών και όχι βάσει κοινών απουσιών, οι οποίες συχνά δεν έχουν ερμηνευτική αξία.

5.2.3 Σενάριο 3: Ομαδοποίηση χωρών βάσει χρονικών σειρών με απόσταση Pearson

Στο τρίτο σενάριο εξετάζεται η ομαδοποίηση χωρών βάσει της δυναμικής των ρυθμών οικονομικής ανάπτυξης. Σε αντίθεση με τα προηγούμενα παραδείγματα, όπου οι μεταβλητές

ήταν στατικές, εδώ η ανάλυση αφορά χρονικές σειρές, και επομένως το ενδιαφέρον εστιάζεται περισσότερο στο μοτίβο κίνησης παρά στο απόλυτο επίπεδο των τιμών.

Προσομοιώθηκαν 10 χώρες για 50 χρονικές περιόδους. Η εξέλιξη του ρυθμού ανάπτυξης κάθε χώρας προέκυψε από:

- έναν κοινό παράγοντα (Common factor) που αποτυπώνει τη γενική διεθνή οικονομική συγκυρία,
- έναν τυχαίο ιδιοσυγκρασιακό θόρυβο,
- και μία μικρή μετατόπιση επιπέδου (shift) για κάθε χώρα.

Με αυτόν τον τρόπο, όλες οι χώρες ακολουθούν σε γενικές γραμμές παρόμοια πορεία, αλλά με διαφοροποιήσεις στην ένταση και στη μεταβλητότητα.

Για τη σύγκριση των χρονικών σειρών χρησιμοποιήθηκε η συσχετιστική ανομοιότητα $d_{ij} = 1 - \rho_{ij}$, όπου είναι ο συντελεστής συσχέτισης Pearson μεταξύ των χρονικών σειρών των χωρών και . Η ποσότητα αυτή εκφράζει τον βαθμό ανομοιότητας ως προς των δύο σειρών. Παρότι χρησιμοποιείται ως μέτρο ανομοιότητας σε εφαρμογές ομαδοποίησης, δεν αποτελεί μετρική με την αυστηρή μαθηματική έννοια, καθώς δεν ικανοποιεί γενικά την τριγωνική ανισότητα.

Η επιλογή αυτή βασίζεται στο γεγονός ότι σε οικονομικές εφαρμογές ενδιαφέρει κυρίως ο βαθμός συγχρονισμού των διακυμάνσεων αν οι χώρες ανεβαίνουν και πέφτουν μαζί και όχι οι διαφορές στο απόλυτο επίπεδο ανάπτυξης. Δύο χώρες με υψηλή συσχέτιση θα έχουν απόσταση κοντά στο μηδέν, ακόμη και αν διαφέρουν ελαφρώς ως προς το μέσο επίπεδο.

Η ομαδοποίηση πραγματοποιήθηκε με ιεραρχικό clustering (complete linkage) και το δένδrogramma κόπηκε σε τρεις συστάδες, διάγραμμα 5.2.2.

Η κατανομή των χωρών στα clusters ήταν η εξής, πίνακας 5.2.7:

Παρατηρείται ότι οι περισσότερες χώρες συγκεντρώνονται σε δύο βασικές ομάδες με παρόμοια δυναμική ανάπτυξης, γεγονός αναμενόμενο δεδομένης της ύπαρξης κοινού παράγοντα. Η Χώρα_9 σχηματίζει ξεχωριστό cluster, γεγονός που υποδηλώνει ότι παρουσιάζει διαφοροποιημένο μοτίβο συσχέτισης με τις υπόλοιπες. Η ομαδοποίηση συνεπώς δεν βασίζεται στο επίπεδο του ρυθμού ανάπτυξης, αλλά στη χρονική κίνηση των σειρών.

Συμπερασματικά, το σενάριο αυτό αναδεικνύει τη σημασία της επιλογής κατάλληλου μέτρου απόστασης σε δεδομένα χρονικών σειρών. Η χρήση της ανομοιότητας συσχέτισης Pearson ($1 - \rho$) επιτρέπει την ομαδοποίηση χωρών με βάση τον βαθμό συγχρονισμού των οικονομικών τους κύκλων, παρακάμπτοντας διαφορές στο απόλυτο επίπεδο.

5.2.4 Σενάριο 4: μικτά δεδομένα με ελλείπουσες τιμές και απόσταση Gower (Ιεραρχικό clustering)

Στο τέταρτο σενάριο εξετάζεται η ομαδοποίηση χωρών όταν το σύνολο δεδομένων περιλαμβάνει μεταβλητές διαφορετικού τύπου, δηλαδή τόσο αριθμητικές όσο και κατηγορικές, καθώς και ελλείπουσες τιμές. Για τον σκοπό αυτό προσομοιώθηκε ένα σύνολο δεδομένων αποτελούμενο από 60 χώρες και τέσσερις μεταβλητές. Οι αριθμητικές μεταβλητές είναι το

κατά κεφαλήν ΑΕΠ (gdp_pc) και το ποσοστό ανεργίας (unemployment), ενώ οι κατηγορικές μεταβλητές αφορούν το αναπτυξιακό καθεστώς (dev_status), με κατηγορίες «Αναπτυγμένη» και «Αναπτυσσόμενη», και τη γεωγραφική περιοχή (region), με κατηγορίες «Ευρώπη», «Ασία», «Αμερική» και «Αφρική».

Στη συνέχεια εισήχθησαν τεχνητά ελλείπουσες τιμές στις δύο αριθμητικές μεταβλητές, προκειμένου να διερευνηθεί η συμπεριφορά της μεθόδου ομαδοποίησης υπό συνθήκες ελλιπών δεδομένων. Συγκεκριμένα, ορίστηκε ως ελλιπές περίπου το 10% των παρατηρήσεων κάθε αριθμητικής μεταβλητής, δηλαδή 6 από τις 60 τιμές τόσο για το κατά κεφαλήν ΑΕΠ όσο και για την ανεργία. Οι θέσεις των ελλείψεων επιλέχθηκαν τυχαία και ανεξάρτητα από τις τιμές των υπόλοιπων μεταβλητών, προσομοιώνοντας έτσι έναν μηχανισμό ελλείψεων τύπου MCAR (Missing Completely At Random).

Για τον υπολογισμό της ανομοιότητας μεταξύ των χωρών χρησιμοποιήθηκε η απόσταση Gower μέσω της συνάρτησης daisy() του πακέτου cluster. Η συγκεκριμένη απόσταση αποτελεί κατάλληλη επιλογή για δεδομένα μικτού τύπου, καθώς επιτρέπει τον ταυτόχρονο συνυπολογισμό αριθμητικών και κατηγορικών χαρακτηριστικών σε ένα ενιαίο μέτρο ανομοιότητας. Επιπλέον, η daisy() μπορεί να χειριστεί τις ελλείπουσες τιμές κατά τον υπολογισμό των επιμέρους αποστάσεων μεταξύ ζευγών παρατηρήσεων, χρησιμοποιώντας τις μεταβλητές για τις οποίες υπάρχουν διαθέσιμες τιμές στο εκάστοτε ζεύγος. Στην παρούσα εφαρμογή ο υπολογισμός της απόστασης Gower πραγματοποιήθηκε επιτυχώς χωρίς να απαιτηθεί συμπλήρωση των ελλείψεων. Παρ' όλα αυτά, στον κώδικα προβλέφθηκε, για λόγους υπολογιστικής ασφάλειας, εναλλακτική διαδικασία συμπλήρωσης των ελλειπουσών αριθμητικών τιμών με τη διάμεσο της αντίστοιχης μεταβλητής, σε περίπτωση αποτυχίας της daisy().

Αφού κατασκευάστηκε ο πίνακας ανομοιοτήτων Gower, εφαρμόστηκε ιεραρχική ομαδοποίηση με τη μέθοδο Average Linkage. Στη μέθοδο αυτή, η απόσταση μεταξύ δύο συστάδων προσδιορίζεται από τον μέσο όρο των αποστάσεων όλων των δυνατών ζευγών παρατηρήσεων που ανήκουν στις δύο συστάδες. Η επιλογή του Average Linkage είναι κατάλληλη για την παρούσα εφαρμογή με απόσταση Gower, καθώς η μέθοδος μπορεί να εφαρμοστεί άμεσα σε έναν πίνακα ανομοιοτήτων και δεν βασίζεται στην υπόθεση ευκλείδειας γεωμετρίας που απαιτείται από μεθόδους όπως η Ward.D2.

Το αντίστοιχο δενδρόγραμμα παρουσιάζεται στο Διάγραμμα 5.2.3. Η δενδρογραμμαμική δομή καταδεικνύει έναν σαφή διαχωρισμό των παρατηρήσεων σε ανώτερο επίπεδο ανομοιότητας, ενώ σε χαμηλότερα επίπεδα διακρίνονται επιμέρους περισσότερο συνεκτικές ομάδες. Για τις ανάγκες της ανάλυσης της εργασίας και για λόγους απλότητας και καλύτερης παρουσίας του παραδείγματος το δενδρόγραμμα κόπηκε σε $K = 3$ συστάδες, ώστε να εξεταστεί αναλυτικότερα η σύνθεση και τα χαρακτηριστικά τους. Οφείλουμε να αναφέρουμε ότι η βέλτιστη τιμή είναι $K = 8$, υστέρη από έλεγχο που αξιοποιήσαμε με την χρήση της R. Η κατανομή των 60 χωρών στις τρεις συστάδες ήταν άνιση. Συγκεκριμένα, το Cluster 1 περιλαμβάνει 29 χώρες, το Cluster 2 24 χώρες και το Cluster 3 7 χώρες. Επομένως, τα δύο πρώτα clusters συγκεντρώνουν το μεγαλύτερο μέρος των παρατηρήσεων, ενώ το τρίτο αποτελεί μια σαφώς μικρότερη και περισσότερο εξειδικευμένη ομάδα.

Οι μέσες τιμές των αριθμητικών μεταβλητών ανά cluster, με αγνόηση των ελλειπουσών τιμών όπου αυτό ήταν απαραίτητο, παρουσιάζουν επίσης αξιοσημείωτες διαφοροποιήσεις. Το Cluster 1 εμφανίζει το υψηλότερο μέσο κατά κεφαλήν ΑΕΠ, ίσο με περίπου 26,16, και μέσο ποσοστό ανεργίας περίπου 7,74. Το Cluster 2 παρουσιάζει ελαφρώς χαμηλότερο μέσο κατά κεφαλήν ΑΕΠ, περίπου 23,95, και ταυτόχρονα το υψηλότερο μέσο ποσοστό ανεργίας μεταξύ των τριών ομάδων, περίπου 8,37. Αντίθετα, το Cluster 3 διαφοροποιείται ιδιαίτερα ως προς το οικονομικό επίπεδο, καθώς παρουσιάζει το χαμηλότερο μέσο κατά κεφαλήν ΑΕΠ, περίπου 14,39, ενώ η μέση ανεργία του ανέρχεται σε περίπου 7,67. Συνεπώς, ο σημαντικότερος αριθμητικός παράγοντας διαφοροποίησης του τρίτου cluster φαίνεται να είναι το αισθητά χαμηλότερο κατά κεφαλήν ΑΕΠ και όχι το ποσοστό ανεργίας.

Ακόμη εντονότερος είναι ο διαχωρισμός των συστάδων ως προς το αναπτυξιακό καθεστώς. Το Cluster 1 αποτελείται αποκλειστικά από Αναπτυγμένες χώρες, καθώς και οι 29 παρατηρήσεις του ανήκουν σε αυτή την κατηγορία. Αντίθετα, τα Clusters 2 και 3 αποτελούνται αποκλειστικά από Αναπτυσσόμενες χώρες, με 24 και 7 παρατηρήσεις αντίστοιχα. Το αποτέλεσμα αυτό δείχνει ότι η μεταβλητή του αναπτυξιακού καθεστώτος συνδέεται πολύ ισχυρά με τη δομή των συστάδων που προέκυψαν μέσω της Gower απόστασης και του Average Linkage.

Ενδιαφέρουσα διαφοροποίηση παρατηρείται και ως προς τη γεωγραφική περιοχή. Το Cluster 1 παρουσιάζει γεωγραφική ετερογένεια, καθώς περιλαμβάνει 4 χώρες από την Αμερική, 8 από την Ασία, 9 από την Αφρική και 8 από την Ευρώπη. Αντίστοιχα, το Cluster 2 περιλαμβάνει 9 χώρες από την Ασία, 9 από την Αφρική και 6 από την Ευρώπη, χωρίς όμως καμία χώρα από την Αμερική. Ιδιαίτερη περίπτωση αποτελεί το Cluster 3, καθώς και οι 7 χώρες που το απαρτίζουν προέρχονται αποκλειστικά από την Αμερική. Επομένως, το τρίτο cluster εμφανίζει πλήρη γεωγραφική ομοιογένεια, σε αντίθεση με τα δύο μεγαλύτερα clusters.

Ο συνδυασμός των παραπάνω αποτελεσμάτων επιτρέπει μια σαφέστερη ερμηνεία των τριών ομάδων. Το Cluster 1 μπορεί να χαρακτηριστεί ως ομάδα αναπτυγμένων χωρών με σχετικά υψηλότερο κατά κεφαλήν ΑΕΠ και γεωγραφική ποικιλομορφία. Το Cluster 2 συγκεντρώνει αναπτυσσόμενες χώρες από την Ασία, την Αφρική και την Ευρώπη, με ενδιάμεσο επίπεδο κατά κεφαλήν ΑΕΠ και ελαφρώς υψηλότερη μέση ανεργία. Τέλος, το Cluster 3 αποτελεί μια μικρή και ιδιαίτερα διακριτή ομάδα αναπτυσσόμενων χωρών της Αμερικής, η οποία χαρακτηρίζεται κυρίως από αισθητά χαμηλότερο μέσο κατά κεφαλήν ΑΕΠ.

Συμπερασματικά, το συγκεκριμένο σενάριο δείχνει ότι ο συνδυασμός της απόστασης Gower με ιεραρχικό clustering μέσω Average Linkage μπορεί να χρησιμοποιηθεί αποτελεσματικά για την ανάλυση δεδομένων μικτού τύπου, ακόμη και παρουσία ελλειπουσών τιμών. Η απόσταση Gower επιτρέπει την ενσωμάτωση διαφορετικών τύπων μεταβλητών στο ίδιο μέτρο ανομοιότητας, ενώ το Average Linkage παρέχει μια κατάλληλη διαδικασία συνένωσης των παρατηρήσεων χωρίς την απαίτηση ευκλείδειας δομής των αποστάσεων. Στην παρούσα εφαρμογή, η ομαδοποίηση ανέδειξε έναν ιδιαίτερα ισχυρό διαχωρισμό μεταξύ αναπτυγμένων και αναπτυσσόμενων χωρών και, παράλληλα, εντόπισε μια μικρότερη και γεωγραφικά πλήρως ομοιογενή ομάδα αναπτυσσόμενων χωρών της Αμερικής με σημαντικά χαμηλότερο κατά κεφαλήν ΑΕΠ. Το παράδειγμα υπογραμμίζει, επομένως, τη σημασία της κατάλληλης επιλογής

τόσο του μέτρου ανομοιογένειας όσο και της μεθόδου σύνδεσης στο ιεραρχικό clustering όταν τα δεδομένα είναι ετερογενή και περιλαμβάνουν ελλείψεις.

5.3 Αποτελέσματα σύγκρισης και ανάλυση

Από το παραπάνω πείραμα, είδαμε ότι σε ένα ιδανικά διαχωρίσιμο σύνολο δεδομένων, όλες οι μέθοδοι απέδωσαν άριστα και μάλιστα η επιλογή μετρικής (ευκλείδεια vs Manhattan) δεν έκανε διαφορά. Αυτό ήταν αναμενόμενο, όταν οι clusters είναι καλά χωρισμένες, οι περισσότερες αποστάσεις θα συμφωνήσουν ως προς το ποια σημεία είναι κοντά. Γενικά, σε δύο διαστάσεις ισχύει $L_2 \leq L_1 \leq \sqrt{2}L_2$, επομένως δεν υπάρχει σταθερός λόγος μεταξύ των δύο μετρικών. Ειδικά όταν οι απόλυτες διαφορές στους δύο άξονες είναι ίσες, ισχύει $L_1 = \sqrt{2}L_2$.

Για να κάνουμε το πρόβλημα πιο ενδιαφέρον και να αναδείξουμε διαφορές, ας εξετάσουμε δύο μεταβολές:

1. Περίπτωση outliers: Θα προσθέσουμε μερικά σημεία πολύ μακριά, που δεν ανήκουν πραγματικά σε καμία από τις τρεις cluster. Για παράδειγμα πέντε σημεία διάσπαρτα τυχαία σε ένα μεγάλο τετράγωνο $[10,15] \times [10,15]$ πολύ μακριά από τα άλλα. Αυτό προσομοιώνει οικονομίες που είναι εντελώς εκτός ομάδων ίσως λάθη καταγραφής ή ακραίες περιπτώσεις.
2. Περίπτωση επικαλυπτόμενων clusters: Θα αυξήσουμε τη διασπορά σε μία από τις τρεις cluster ώστε να γίνει πιο "spread" και να αρχίσει να επικαλύπτει μια γειτονική. Συγκεκριμένα, ας πούμε η cluster3 με κέντρο (0,5) θα πάρει $\sigma = 1.0$ αντί για 0.5. Αυτό σημαίνει ότι τα σημεία της cluster3 θα απλώνονται περισσότερο, ενδεχομένως μερικά να πλησιάσουν την cluster1 αρκετά.

Στην ενότητα αυτή διερευνάται συστηματικά πόσο ανθεκτικοί είναι διαφορετικοί αλγόριθμοι ομαδοποίησης όταν τα δεδομένα παύουν να είναι ιδανικά και εμφανίζονται κάποιες δυσκολίες. Συγκεκριμένα, υλοποιούμε ένα κοινό πλαίσιο πειραμάτων στο R, όπου εφαρμόζουμε ιεραρχική ομαδοποίηση (complete linkage) με τρεις ορισμούς απόστασης Euclidean, Manhattan και συσχετιστική απόσταση Pearson με $d = 1 - \text{cor}$, k-means (Euclidean), PAM (με τις ίδιες τρεις αποστάσεις), DBSCAN (Euclidean) και Gaussian Mixture Models (GMM). Για να είναι η σύγκριση δίκαιη και κατανοητή, αξιολογούμε τα αποτελέσματα με δύο συμπληρωματικά κριτήρια:

- **Adjusted Rand Index (ARI)** όταν υπάρχει πραγματική ετικέτα κλάσης (ground truth), ώστε να μετρηθεί άμεσα η συμφωνία με τη γνωστή δομή, και
- **μέσο Silhouette**, ως εσωτερικό μέτρο ποιότητας των clusters, με τη διευκρίνιση ότι όταν υπολογίζεται με διαφορετική μετρική απόστασης ανά περίπτωση, οι τιμές του δεν είναι πλήρως συγκρίσιμες μεταξύ διαφορετικών μετρικών, καθώς μεταβάλλεται ταυτόχρονα και το κριτήριο αξιολόγησης, με εξαίρεση το DBSCAN, όπου η παρουσία θορύβου με ετικέτα 0 καθιστά το silhouette μη ορισμένο στο συγκεκριμένο setup.

Επιπλέον, παραθέτουμε πίνακες με τα μεγέθη των συστάδων (cluster Size), συμπεριλαμβανομένου του θορύβου για τον DBSCAN, ώστε να φαίνεται όχι μόνο η ποιότητα

αλλά και το πώς κατανέμει κάθε μέθοδος τις παρατηρήσεις. Τα παραπάνω εφαρμόζονται σε τρία στενευμένα σενάρια δυσκολίας:

- προσθήκη outliers,
- παρουσία επιμήκους ή μη-σφαιρικής συστάδας, και
- ανισοκλίμακες μεταβλητές, όπου εξετάζεται ρητά και η επίδραση της τυποποίησης (z-scores).

Στο πείραμα 1 με την προσθήκη (Outliers) εξετάζουμε πόσο ανθεκτικοί είναι οι αλγόριθμοι ομαδοποίησης όταν στο σύνολο δεδομένων εισάγονται τεχνητά ακραίες παρατηρήσεις (outliers). Προσθέτουμε 10 σημεία σε απομακρυσμένη περιοχή του χώρου και στη συνέχεια αξιολογούμε την ποιότητα των clusters μόνο στα 300 σημεία, ώστε η σύγκριση να είναι δίκαιη και να μην αλλοιώνεται από τον τεχνητό θόρυβο. Η απόδοση μετριέται με δύο συμπληρωματικούς δείκτες, ARI και μέσο Silhouette (συνοχή και διαχωρισμός των clusters).

Με αυτόν τον τρόπο ελέγχουμε αν οι μέθοδοι διατηρούν τη σωστή δομή των δεδομένων πάρα την παρουσία outliers, αλλά και αν οι εσωτερικές μετρικές παραμένουν σταθερές. Τέλος, για τον DBSCAN καταγράφουμε επιπλέον τα μεγέθη των ομάδων-συστάδων, ώστε να διαπιστώσουμε αν ο αλγόριθμος απορροφά τα outliers ή τα απομονώνει ως θόρυβο.

Τα αποτελέσματα στο πίνακα 5.3.1 δείχνουν ότι οι περισσότερες μέθοδοι (hclust με ευκλείδεια και Manhattan απόσταση, k-means, PAM, GMM και DBSCAN) διατηρούν τέλεια συμφωνία με τις πραγματικές κλάσεις ($ARI = 1.000$), ως προς τα 300 μη ακραία σημεία στα οποία υπολογίστηκε ο ARI. Επομένως, οι μέθοδοι διατήρησαν την ορθή ανάθεση των αρχικών σημείων στις τρεις συστάδες. Ωστόσο, από την τιμή του ARI δεν μπορεί να εξαχθεί άμεσο συμπέρασμα για τον τρόπο με τον οποίο οι μέθοδοι χειρίστηκαν τα outliers. Οι τιμές του μέσου Silhouette παραμένουν υψηλές για τις περισσότερες μεθόδους, επιβεβαιώνοντας καλή συνοχή εντός των ομάδων και σαφή διαχωρισμό μεταξύ τους. Αντίθετα, η χρήση correlation distance (hclust_cor και pam_cor) οδηγεί σε χαμηλό ARI και αισθητά χαμηλότερο Silhouette, δείχνοντας ότι η συγκεκριμένη μετρική δεν είναι κατάλληλη για τη γεωμετρία του συγκεκριμένου dataset. Για τον DBSCAN, δεν υπολογίζεται Silhouette (NA), αλλά τα μεγέθη των clusters (100-100-100) δείχνουν ότι αναγνώρισε σωστά τη δομή χωρίς να επηρεαστεί από τα outliers, τουλάχιστον ως προς τα σημεία που αξιολογήθηκαν.

Στο διάγραμμα 5.3.1 παρατηρούμαι τρεις συμπαγείς και καλά διαχωρισμένες συστάδες στο κάτω και κεντρικό τμήμα του χώρου, ενώ τα outliers εμφανίζονται απομακρυσμένα στο άνω δεξί τμήμα. Η γεωμετρία των clusters είναι σχεδόν σφαιρική και με παρόμοια διασπορά, γεγονός που ευνοεί ιδιαίτερα μεθόδους όπως k-means και GMM. Συνολικά, το σχήμα επιβεβαιώνει ότι το dataset είναι εύκολο ως προς τη διαχωρισιμότητα, ακόμη και μετά την προσθήκη ακραίων τιμών.

Στο πείραμα 2 διερευνάται η συμπεριφορά κλασικών αλγορίθμων ομαδοποίησης όταν μία από τις ομάδες δεν είναι σφαιρική, αλλά επιμήκης δηλαδή έχει μεγαλύτερη διασπορά στον άξονα x_1 . Δημιουργούνται τρεις συστάδες ίδιου μεγέθους, όπου οι δύο είναι συμπαγείς και η τρίτη είναι εκτεταμένη οριζόντια. Η αξιολόγηση γίνεται με ARI και μέσο Silhouette. Επιπλέον, παρακολουθείται η συμπεριφορά του DBSCAN μέσω των μεγεθών των clusters που παράγει.

Η παρουσία επιμήκους cluster επηρεάζει αισθητά μεθόδους που ευνοούν σφαιρικές ομάδες. Από τον πίνακα 5.3.2 το *hclust* με ευκλείδεια απόσταση εμφανίζει μέτρια επίδοση $ARI = 0.593$, ένδειξη ότι η επιμήκης γεωμετρία οδηγεί σε συγχωνεύσεις ή διαχωρισμούς που δεν συμφωνούν πλήρως με την πραγματική δομή. Αντίθετα, το *hclust* με Manhattan απόσταση αποδίδει εξαιρετικά αφού $ARI = 0.990$, υποδηλώνοντας ότι η L_1 μετρική χειρίζεται καλύτερα την ανισοτροπία. Οι *k-means* και *PAM* παρουσιάζουν πολύ υψηλά ARI και ταυτόχρονα υψηλό *Silhouette*, γεγονός που δείχνει ότι πάρα την επιμήκη μορφή, οι ομάδες παραμένουν αρκετά διαχωρίσιμες ώστε οι μέθοδοι κεντροειδούς να δίνουν σταθερές λύσεις. Ο *GMM* επίσης είναι ισχυρός αφού $ARI = 0.970$, κάτι αναμενόμενο αφού μπορεί να προσαρμόσει ελλειπτικές κατανομές. Τέλος, οι εκδοχές με *correlation distance* (*hclust_cor*, *pam_cor*) αποτυγχάνουν αφού $ARI = 0.14$, επιβεβαιώνοντας ότι η συσχέτιση δεν είναι κατάλληλη μετρική όταν η διάκριση βασίζεται κυρίως σε απόλυτες γεωμετρικές αποστάσεις στο επίπεδο. Για τον *DBSCAN*, το ARI είναι χαμηλότερο και τα μεγέθη clusters (100, 200) δείχνουν ότι ο αλγόριθμος ουσιαστικά συγχώνευσε δύο ομάδες.

Το διάγραμμα 5.3.2 αποτυπώνει καθαρά τη δυσκολία του πειράματος, δύο ομάδες κάτω και δεξιά είναι συμπαγείς και περίπου κυκλικές, ενώ η τρίτη επάνω αριστερά είναι εμφανώς επιμήκης κατά τον άξονα x_1 . Αυτή η μορφή αυξάνει μοιάζει λιγότερο συνεκτική με βάση ευκλείδειες αντιλήψεις σφαιρικότητας, εξηγώντας την πτώση σε μεθόδους όπως το *hclust_euc*. Παράλληλα, η επιμήκης ομάδα παραμένει αρκετά απομακρυσμένη από την κάτω συστάδα, άρα οι αλγόριθμοι που προσαρμόζονται σε ελλειπτικές δομές για παράδειγμα οι *GMM* ή L_1 αποστάσεις έχουν σαφές πλεονέκτημα.

Στο πείραμα 3 εξετάζουμε πώς επηρεάζεται η ομαδοποίηση όταν οι μεταβλητές δεν βρίσκονται στην ίδια κλίμακα μέτρησης. Συγκεκριμένα, πολλαπλασιάζουμε τη δεύτερη μεταβλητή x_2 επί 20, ώστε να κυριαρχεί αριθμητικά στις αποστάσεις. Έπειτα εφαρμόζουμε όλους τους αλγόριθμους σε δύο συνθήκες: (α) χωρίς τυποποίηση και (β) με τυποποίηση *z-scores*. Στόχος είναι να φανεί καθαρά ότι η κλίμακα μπορεί να αλλοιώσει τις αποστάσεις και άρα την ομαδοποίηση, ενώ η τυποποίηση αποκαθιστά τη συγκρισιμότητα των χαρακτηριστικών. Η αξιολόγηση γίνεται με ARI και μέσο *Silhouette*, ώστε να αποτυπώνεται τόσο η συμφωνία με τις πραγματικές κλάσεις όσο και η εσωτερική ποιότητα των clusters.

Χωρίς τυποποίηση, παρατηρείται γενικευμένη πτώση του ARI πίνακα 5.3.3 για τις περισσότερες μεθόδους που βασίζονται σε αποστάσεις (*hclust*, *k-means*, *PAM*) με τιμές γύρω στο 0.50. Αυτό δείχνει ότι η μεγέθυνση του x_2 τραβά τις αποστάσεις σχεδόν αποκλειστικά προς τη δεύτερη διάσταση, με αποτέλεσμα οι αλγόριθμοι να ομαδοποιούν κυρίως με βάση το x_2 και να χάνουν τη σωστή δομή που προκύπτει από τον συνδυασμό x_1 και x_2 . Ενδιαφέρον είναι ότι το *Silhouette* μπορεί να παραμείνει σχετικά υψηλό κάτι που υπενθυμίζει ότι ένας εσωτερικός δείκτης μπορεί να επιβραβεύσει συνεκτικές ομάδες ακόμη και όταν αυτές δεν αντιστοιχούν στις πραγματικές κλάσεις.

Με την τυποποίηση (*z-scores*), η εικόνα αλλάζει δραστικά, πίνακας 5.3.4 οι περισσότερες μέθοδοι επανέρχονται σε $ARI = 1.000$ και υψηλό *Silhouette*, γεγονός που επιβεβαιώνει ότι η τυποποίηση αποκαθιστά την ισορροπημένη συνεισφορά των χαρακτηριστικών στις αποστάσεις και οδηγεί σε ορθή ανάκτηση των clusters. Οι εκδοχές με *correlation distance* βελτιώνονται

αλλά παραμένουν μέτριες, υποδηλώνοντας ότι η συσχέτιση δεν ευθυγραμμίζεται πλήρως με τη γεωμετρία αυτού του προβλήματος. Τέλος, η άμεση σύγκριση της επίδοσης του DBSCAN πριν και μετά την τυποποίηση δεν είναι έγκυρη όταν διατηρείται η ίδια τιμή της παραμέτρου, καθώς η τυποποίηση μεταβάλλει την κλίμακα των αποστάσεων. Επομένως, η τιμή του πρέπει να επαναπροσδιοριστεί στη νέα κλίμακα, για παράδειγμα μέσω k-distance plot, πριν αξιολογηθεί η επίδοση του DBSCAN στα τυποποιημένα δεδομένα.

Στο διάγραμμα 5.3.3 με άξονα x_2 πολλαπλασιασμένο επί 20 οι συστάδες εμφανίζονται πολύ πιο απομακρυσμένες κατακόρυφα σε σχέση με οριζόντια, επειδή ο άξονας x_2 έχει πλέον πολύ μεγαλύτερο εύρος τιμών. Οπτικά γίνεται σαφές ότι οι διαφορές στο x_2 κυριαρχούν στη διάταξη των σημείων και κατά συνέπεια, σε κάθε μέτρο απόστασης που δεν είναι ανεξάρτητο. Το πείραμα αναδεικνύει έτσι με πρακτικό τρόπο μια κρίσιμη αρχή στην ανάλυση δεδομένων όταν οι μεταβλητές έχουν διαφορετικές κλίμακες, η τυποποίηση δεν είναι απλώς προαιρετική βελτίωση, αλλά συχνά απαραίτητη για την αξιόπιστη ομαδοποίηση.

5.4 Εφαρμογές σε οικονομικά σενάρια, επιλογή μέτρου απόστασης και αξιολόγηση clustering

Στην παρούσα ενότητα εξετάζεται η εφαρμογή των μετρικών απόστασης σε ρεαλιστικά οικονομικά δεδομένα, με στόχο τη διερεύνηση της καταλληλότητας τους ανάλογα με τη δομή των παρατηρήσεων. Αρχικά, αναλύεται ένα πολυμεταβλητό σύνολο επιχειρήσεων όπου η ύπαρξη συσχετίσεων μεταξύ μεταβλητών καθιστά κρίσιμη τη χρήση μετρικών όπως η Mahalanobis και η robust εκδοχή της. Στη συνέχεια, μελετάται ένα σενάριο χρονοσειρών χαρτοφυλακίων, όπου η ομοιότητα βασίζεται στο μοτίβο κίνησης και όχι στα επίπεδα τιμών, καθιστώντας κατάλληλη την απόσταση Pearson ($1 - \rho$). Τα αποτελέσματα αξιολογούνται μέσω δεικτών ARI και Silhouette, ενώ η οπτικοποίηση πραγματοποιείται με PCA. Με τον τρόπο αυτό, διερευνάται η αντιστοίχιση μεταξύ τύπου δεδομένων και κατάλληλης μετρικής ομοιότητας.

Εκτελώντας το σχετικό κώδικα στην R για το σενάριο 5 λαμβάνουμε τα αποτελέσματα που είναι αποτυπωμένα στον πίνακα 5.4.1

Στο σενάριο 5 εξετάστηκε clustering επιχειρήσεων σε τετραδιάστατο χώρο (profitability, leverage, liquidity, risk) με τρία προφίλ. Τα αποτελέσματα του πίνακα δείχνουν ότι όλες οι προσεγγίσεις πετυχαίνουν πολύ χαμηλές τιμές ARI, γεγονός που υποδηλώνει ότι η ανάκτηση των πραγματικών ομάδων είναι ασθενής. Παράλληλα, οι τιμές Silhouette κυμαίνονται περίπου 0.13-0.17, δηλαδή σε επίπεδα που συνήθως αντιστοιχούν σε μέτρια έως χαμηλή εσωτερική συνοχή και σημαντική επικάλυψη μεταξύ ομάδων. Η robust Mahalanobis απόσταση με PAM εμφανίζει στο συγκεκριμένο πείραμα ARI = 0.038 και Silhouette = 0.165. Ωστόσο, δεδομένου του πολύ χαμηλού ARI και της μικρής διαφοροποίησης μεταξύ των μεθόδων, δεν μπορεί να τεκμηριωθεί υπεροχή κάποιας προσέγγισης από μία μόνο εκτέλεση, καθώς οι παρατηρούμενες διαφορές ενδέχεται να οφείλονται σε τυχαία διακύμανση. Παρόλα αυτά, η συνολικά χαμηλή επίδοση υποδεικνύει ότι, στο συγκεκριμένο σχεδιασμό δεδομένων, οι ομάδες είναι δύσκολα

διαχωρίσιμες, οπότε η επιλογή απόστασης βελτιώνει οριακά την εικόνα χωρίς να οδηγεί σε ξεκάθαρό διαχωρισμό.

Στο διάγραμμα 5.4.1 απεικονίζεται η προβολή των επιχειρήσεων στους δύο πρώτους κύριους άξονες (PC1-PC2). Η οπτική εικόνα επιβεβαιώνει τα ευρήματα του πίνακα, παρατηρείται σημαντική επικάλυψη μεταξύ των χρωματισμένων clusters, χωρίς σαφή κενά ή ξεκάθαρα συμπαγείς, απομονωμένες ομάδες. Επιπλέον, είναι σημαντικό ότι η PCA είναι τεχνική απεικόνισης διακύμανσης και όχι βελτιστοποίησης διαχωρισμού ομάδων επομένως, πιθανός διαχωρισμός που βρίσκεται σε υψηλότερες διαστάσεις μπορεί να μην φαίνεται πλήρως στο επίπεδο (PC1, PC2). Ωστόσο, ως εργαλείο ερμηνείας, το σχήμα προσφέρει μια συνεπή οπτική τεκμηρίωση ότι το clustering στο συγκεκριμένο σενάριο δεν οδηγεί σε έντονα διακριτές συστάδες.

Απόδοση μεθόδων clustering στο σενάριο 6 για διαφορετικά μέτρα απόστασης δίνεται στο πίνακα 5.4.2. Σε αντίθεση με το σενάριο 5 παρουσιάζει πολύ υψηλή διαχωριστικότητα μεταξύ των ομάδων επενδυτών. Δύο προσεγγίσεις επιτυγχάνουν τέλειο $ARI = 1.000$, δηλαδή πλήρη ανάκτηση των πραγματικών clusters:

- Pearson distance με PAM
- Euclidean distance (levels) με ιεραρχικό clustering (complete)

Η υψηλή τιμή Silhouette έως 0.798 υποδηλώνει έντονη εσωτερική συνοχή και καθαρό διαχωρισμό μεταξύ συστάδων. Το γεγονός ότι ακόμη και η Pearson απόσταση με hclust επιτυγχάνει ικανοποιητικό $ARI = 0.457$ και υψηλό Silhouette = 0.646 επιβεβαιώνει ότι το σενάριο περιέχει ισχυρή δομική διαφοροποίηση μεταξύ των ομάδων. Ενδιαφέρον παρουσιάζει ότι η Euclidean απόσταση στα επίπεδα οδηγεί επίσης σε τέλεια ανάκτηση των ομάδων. Αυτό οφείλεται στον τρόπο κατασκευής των δεδομένων τα clusters διαφοροποιούνται όχι μόνο ως προς το μοτίβο (shape), αλλά και ως προς το επίπεδο (shift) των χρονοσειρών. Συνεπώς, η απόσταση Euclidean συλλαμβάνει αποτελεσματικά αυτή τη μετατόπιση.

Συμπέρασμα, η δομή των δεδομένων είναι έντονα διακριτή, επιτρέποντας σε κατάλληλες μετρικές (Pearson ή Euclidean) να ανακτήσουν πλήρως τις πραγματικές ομάδες.

Στο διάγραμμα 5.4.2 απεικονίζεται η προβολή των επενδυτών στους δύο πρώτους κύριους άξονες. Η οπτική εικόνα παρουσιάζει τρεις σαφώς διαχωρισμένες ομάδες, με μεγάλο κενό μεταξύ τους στον χώρο (PC1, PC2). Δεν παρατηρείται επικάλυψη μεταξύ clusters, γεγονός που εναρμονίζεται πλήρως με τις υψηλές τιμές ARI και Silhouette. Η έντονη διάκριση κατά μήκος του πρώτου κύριου άξονα (PC1) υποδηλώνει ότι η κύρια πηγή διακύμανσης στο σύστημα σχετίζεται με τη διαφοροποίηση των underlying factor και των shift που ενσωματώθηκαν στη δημιουργία των χρονοσειρών. Σε αντίθεση με το σενάριο 5, εδώ η PCA επιβεβαιώνει οπτικά ότι η δομή των δεδομένων είναι φυσικά τμηματοποιημένη, γεγονός που εξηγεί την υψηλή επιτυχία των αλγορίθμων clustering.

Συμπερασματικά, ενώ στο σενάριο 5 οι ομάδες παρουσίαζαν σημαντική επικάλυψη, καθιστώντας το clustering εγγενώς δύσκολο, στο σενάριο 6 η δομή των δεδομένων ήταν έντονα διακριτή. Η σύγκριση αυτή αναδεικνύει ότι η επιτυχία ενός μέτρου απόστασης δεν εξαρτάται

μόνο από τη θεωρητική του καταλληλότητα, αλλά κυρίως από τη γεωμετρική και στατιστική δομή των δεδομένων.

5.5 Συγκεντρωτική ανάλυση και σύγκριση μεθόδων

Στην ενότητα αυτή πραγματοποιείται συγκεντρωτική σύγκριση των αποτελεσμάτων που προέκυψαν στην προηγούμενη παράγραφο, με στόχο να αποτιμηθεί συνολικά η επίδραση τόσο του μέτρου απόστασης όσο και του αλγορίθμου clustering στην ποιότητα ομαδοποίησης. Για λόγους πληρότητας παρουσιάζονται ενιαία όλα τα αποτελέσματα πίνακας 5.5.1, ενώ στη συνέχεια υπολογίζονται οι μέσες επιδόσεις (Mean ARI) ανά απόσταση πίνακας 5.5.2 και ανά αλγόριθμο πίνακας 5.5.3. Τέλος, η εικόνα συμπληρώνεται με γραφήματα που διευκολύνουν την ερμηνεία, αναδεικνύοντας τη διαφορά συμπεριφοράς μεταξύ των δύο σεναρίων.

Ο πίνακας 5.5.1 αναδεικνύει άμεσα μια βασική διαφοροποίηση στο S5_firms οι τιμές ARI είναι πολύ χαμηλές, ενώ στο S6_portfolios οι τιμές είναι σαφώς υψηλότερες και φτάνουν μέχρι το 1.000. Αυτό σημαίνει ότι η δυσκολία του clustering δεν είναι σταθερή ιδιότητα της μεθόδου, αλλά εξαρτάται κυρίως από το πόσο καθαρά διαχωρισμένη είναι η δομή των δεδομένων. Παράλληλα, οι τιμές Silhouette κινούνται σε μέτρια επίπεδα στο S5_firms, ενώ στο S6_portfolios είναι σημαντικά υψηλότερες, κάτι που συνάδει με ισχυρότερη συνοχή και διαχωρισσιμότητα των clusters.

Οι τιμές Mean ARI του πίνακα 5.5.2 δεν είναι άμεσα συγκρίσιμες μεταξύ όλων των αποστάσεων, καθώς δεν έχουν υπολογιστεί σε κοινό σύνολο σεναρίων. Ειδικότερα, ορισμένες αποστάσεις εφαρμόζονται μόνο σε συγκεκριμένα σενάρια, με αποτέλεσμα ο αντίστοιχος μέσος ARI να επηρεάζεται και από τον βαθμό δυσκολίας των σεναρίων στα οποία αξιολογήθηκαν. Συνεπώς, οι συγκεκριμένοι μέσοι όροι δεν μπορούν να χρησιμοποιηθούν για συνολική κατάταξη των μετρικών. Η σύγκριση είναι ασφαλέστερο να πραγματοποιείται εντός του ίδιου σεναρίου ή, εναλλακτικά, σε κοινό σύνολο σεναρίων για όλες τις μετρικές.

Από τον πίνακα 5.5.3 προκύπτει ότι ο hclust_complete έχει την υψηλότερη μέση απόδοση, Mean ARI = 0.729 ενώ ακολουθεί το PAM με 0.519. Αντίθετα, ο hclust_Ward.D2 και το GMM (VWV) εμφανίζουν χαμηλότερους μέσους όρους, επειδή εφαρμόζονται στο σενάριο 5 όπου ο διαχωρισμός είναι εγγενώς δυσκολότερος. Συνολικά, το αποτέλεσμα υποστηρίζει ότι η ιδανική μέθοδος δεν είναι ανεξάρτητη από το context. Ο ίδιος αλγόριθμος μπορεί να αποδώσει εξαιρετικά σε ένα σενάριο και πολύ πιο μέτρια σε ένα άλλο, όταν η δομή των δεδομένων αλλάζει.

Το διάγραμμα 5.5.1 παρουσιάζει τη σύγκριση των τιμών Mean ARI μεταξύ των εξεταζόμενων σεναρίων για τις διαφορετικές αποστάσεις. Παρατηρείται σαφής διαφοροποίηση της απόδοσης μεταξύ των σεναρίων, με το σενάριο 6 να εμφανίζει συνολικά υψηλότερες τιμές ARI, ενώ στο σενάριο 5 οι τιμές παραμένουν χαμηλές. Το εύρημα αυτό υποδηλώνει ότι η αποτελεσματικότητα του clustering επηρεάζεται έντονα από τα χαρακτηριστικά και τη δομή του εκάστοτε σεναρίου, ενώ η επίδραση της επιλογής απόστασης πρέπει να αξιολογείται εντός κάθε σεναρίου.

Το boxplot διάγραμμα 5.5.2 κάνει πολύ καθαρό το βασικό εύρημα, το S5_firms συγκεντρώνεται κοντά στο μηδέν, ενώ το S6_portfolios έχει σημαντικά υψηλότερες τιμές και φτάνει μέχρι το 1. Αυτό σημαίνει ότι η δυσκολία του clustering καθορίζεται κυρίως από τη δομή του προβλήματος και δευτερευόντως από το ποια μέθοδος χρησιμοποιείται.

Τέλος, το διάγραμμα 5.5.3 απεικονίζει πώς αλλάζει η επίδοση της ίδιας λογικής απόστασης ανάλογα με το σενάριο. Το interaction plot δείχνει ότι οι επιδόσεις δεν είναι παράλληλες μεταξύ σεναρίων, δηλαδή η επίδραση της απόστασης εξαρτάται από το σενάριο. Στο S6_portfolios οι αποστάσεις που αποτυπώνουν επίπεδα ή συσχετίσεις οδηγούν σε πολύ υψηλό ARI, ενώ στο S5_firms ακόμη και πιο εξειδικευμένες μετρικές (Mahalanobis, robust) δεν οδηγούν σε δραστική αύξηση. Έτσι, το διάγραμμα τεκμηριώνει ξεκάθαρα ότι η επιλογή απόστασης είναι context-dependent και πρέπει να γίνεται με βάση τη φύση των δεδομένων.

Συνολικά, τα συγκεντρωτικά αποτελέσματα δείχνουν ότι δεν υπάρχει καθολικά βέλτιστη μέθοδος, η επιτυχία εξαρτάται τόσο από το μέτρο απόστασης όσο και από τη δομή του σεναρίου. Η ενότητα αυτή επιβεβαιώνει ότι η πιο ασφαλής πρακτική είναι η επιλογή μέτρου ομοιότητας που αντανακλά τη φύση των δεδομένων και η αξιολόγηση με συνδυασμό εξωτερικών (ARI) και εσωτερικών (Silhouette) κριτηρίων.

5.6 Ευαισθησία και ανθεκτικότητα των μεθόδων Clustering

Στην παρούσα ενότητα εξετάζεται η ευαισθησία και η ανθεκτικότητα των μεθόδων clustering υπό συνθήκες μη ιδανικών δεδομένων. Ενώ στις προηγούμενες ενότητες τα δεδομένα είχαν ελεγχόμενη δομή, στην πράξη τα σύνολα δεδομένων συχνά περιλαμβάνουν θόρυβο, ακραίες τιμές (outliers), ελλειπείς παρατηρήσεις και μεταβλητές διαφορετικής κλίμακας. Στόχος της ανάλυσης είναι να αξιολογηθεί κατά πόσο οι διάφοροι συνδυασμοί απόστασης και αλγορίθμου διατηρούν σταθερή απόδοση όταν μεταβάλλονται οι συνθήκες. Για τον σκοπό αυτό δημιουργήθηκε πλέγμα πειραματικών ρυθμίσεων που περιλάμβανε, προσθήκη θορύβου, προσθήκη outliers, εισαγωγή ελλειπών τιμών, εφαρμογή ή μη τυποποίησης (scaling), διαφορετικά κριτήρια σύνδεσης (complete, Average, ward.D2) και διαφορετικές παραμέτρους DBSCAN (eps, minPts).

Για κάθε συνδυασμό πειραματικών συνθηκών πραγματοποιήθηκαν 30 επαναλήψεις. Σε κάθε επανάληψη δημιουργήθηκε νέο βασικό συνθετικό σύνολο δεδομένων, χρησιμοποιώντας ως τυχαίο σπόρο την τιμή $2026 + r$, όπου $r = 1, \dots, 30$, δηλαδή seeds από 2027 έως 2056. Το πλήρες πλέγμα περιλάμβανε 1.458 συνδυασμούς πειραματικών ρυθμίσεων ανά επανάληψη και, συνεπώς, 43.740 πειραματικές συγκρίσεις συνολικά. Σε κάθε σύγκριση αξιολογήθηκαν επτά εφαρμογές μεθόδων ομαδοποίησης, με αποτέλεσμα 306.180 επιμέρους τιμές ARI. Για την k-means χρησιμοποιήθηκε επιπλέον σταθερό seed 2026 για την τυχαία αρχικοποίησή της (nstart = 10). Για τον προσδιορισμό της «νίκης» σε κάθε πειραματική σύγκριση επιλεγόταν αρχικά η μέθοδος με τη μεγαλύτερη τιμή ARI. Σε περίπτωση ισοβαθμίας ως δεύτερο κριτήριο χρησιμοποιήθηκε η μεγαλύτερη τιμή Silhouette και καταγράφηκε μία μόνο νικήτρια μέθοδος ανά πειραματική σύγκριση.

Ο Πίνακας 5.6.1 δείχνει ότι οι περισσότερες μέθοδοι που βασίζονται σε ευκλείδεια ή Manhattan απόσταση, σε συνδυασμό με ιεραρχικό clustering (Complete, Average, ward.D2),

καθώς και οι k-means και PAM παρουσιάζουν ιδιαίτερα υψηλή μέση τιμή. Αυτό υποδηλώνει ότι οι μέθοδοι αυτές παραμένουν εξαιρετικά σταθερές ακόμη και όταν προστίθενται θόρυβος, ακραίες τιμές και ελλιπή δεδομένα. Αντίθετα, ο DBSCAN εμφανίζει σημαντικά χαμηλότερη μέση απόδοση, γεγονός που υποδηλώνει αυξημένη ευαισθησία στις παραμέτρους ϵ ps και minPts, καθώς και στις μεταβολές της πυκνότητας του χώρου. Το αποτέλεσμα αυτό είναι αναμενόμενο, καθώς ο DBSCAN εξαρτάται έντονα από την επιλογή υπερπαραμέτρων. Συνολικά, τα αποτελέσματα υποδεικνύουν ότι σε δεδομένα με σχετικά σαφή γεωμετρική δομή, οι κλασικές μέθοδοι (hierarchical, k-means, PAM, GMM) εμφανίζουν υψηλή ανθεκτικότητα σε λογικά επίπεδα διαταραχής.

Ο Πίνακας 5.6.2 παρουσιάζει τη συχνότητα με την οποία κάθε συνδυασμός απόστασης και αλγορίθμου πέτυχε την υψηλότερη επίδοση στις διαφορετικές πειραματικές συνθήκες. Τα αποτελέσματα δείχνουν σαφή υπεροχή της ιεραρχικής ομαδοποίησης με ευκλείδεια απόσταση. Ειδικότερα, η μέθοδος Ward.D2 καταγράφει τη μεγαλύτερη συχνότητα επικράτησης με 14.211 νίκες, ακολουθούμενη πολύ στενά από την Complete linkage με 14.199 και την Average linkage με 14.121 νίκες. Αντίθετα, η k-means επικράτησε σε 1.092 περιπτώσεις και η PAM σε μόλις 53 περιπτώσεις. Οι παραλλαγές της ιεραρχικής ομαδοποίησης με Manhattan απόσταση παρουσίασαν πολύ μικρότερη συχνότητα επικράτησης, με 27, 19 και 18 νίκες για Average, Complete και Ward.D2 αντίστοιχα. Συνεπώς, παρότι αρκετές μέθοδοι εμφανίζουν παρόμοια μέση τιμή ARI, η ανάλυση της συχνότητας επικράτησης αναδεικνύει την ιεραρχική ομαδοποίηση με ευκλείδεια απόσταση ως την πλέον συνεπή επιλογή, ιδιαίτερα στις εκδοχές Ward.D2, Complete και Average.

Ο Πίνακας 5.6.3 παρουσιάζει τη μέση τιμή του Adjusted Rand Index (ARI), την τυπική απόκλιση (SD), το τυπικό σφάλμα (SE) και τα αντίστοιχα 95% διαστήματα εμπιστοσύνης για κάθε συνδυασμό αλγορίθμου και μέτρου απόστασης. Με τον τρόπο αυτό, η αξιολόγηση δεν περιορίζεται αποκλειστικά στη μέση επίδοση των μεθόδων, αλλά λαμβάνει υπόψη και τη μεταβλητότητα των αποτελεσμάτων στις διαφορετικές πειραματικές συνθήκες. Τα αποτελέσματα δείχνουν ότι το GMM, οι παραλλαγές της ιεραρχικής ομαδοποίησης (Average, Complete και Ward.D2), η k-means και η PAM παρουσιάζουν ουσιαστικά ταυτόσημη μέση επίδοση, με $\text{Mean ARI} = 0,903$. Η τυπική απόκλιση για τις συγκεκριμένες μεθόδους κυμαίνεται μεταξύ 0,084 και 0,085, γεγονός που υποδηλώνει παρόμοιο επίπεδο διασποράς των επιδόσεών τους στο σύνολο των εξεταζόμενων πειραματικών συνθηκών. Παράλληλα, τα 95% διαστήματα εμπιστοσύνης των παραπάνω μεθόδων είναι ιδιαίτερα στενά και εντοπίζονται περίπου μεταξύ 0,902 και 0,904. Η μικρή έκταση των διαστημάτων εμπιστοσύνης υποδηλώνει υψηλή ακρίβεια στην εκτίμηση της μέσης τιμής του ARI, η οποία συνδέεται και με τον μεγάλο αριθμό αξιολογήσεων που περιλαμβάνονται στην ανάλυση. Συνεπώς, τα στενά διαστήματα εμπιστοσύνης δεν πρέπει να ερμηνεύονται από μόνα τους ως απουσία μεταβλητότητας των επιμέρους αποτελεσμάτων, η οποία αποτυπώνεται πληρέστερα μέσω της τυπικής απόκλισης.

Αντίθετα, ο DBSCAN διαφοροποιείται έντονα από τις υπόλοιπες μεθόδους. Η μέση τιμή ARI διαμορφώνεται μόλις στο 0,296, ενώ η τυπική απόκλιση ανέρχεται σε 0,399, σημαντικά υψηλότερα από την αντίστοιχη τιμή των υπόλοιπων αλγορίθμων. Το 95% διάστημα εμπιστοσύνης της μέσης τιμής του DBSCAN κυμαίνεται από 0,292 έως 0,299. Ο συνδυασμός της χαμηλότερης μέσης επίδοσης και της ιδιαίτερα υψηλής διασποράς καταδεικνύει ότι ο

DBSCAN παρουσιάζει μεγαλύτερη ευαισθησία στις διαφορετικές πειραματικές συνθήκες και στις επιλογές των υπερπαραμέτρων ϵ ps και minPts.

Συνολικά, η εξέταση της διασποράς και των διαστημάτων εμπιστοσύνης ενισχύει τα αποτελέσματα της ανάλυσης ευαισθησίας. Οι ιεραρχικές μέθοδοι, η k-means, η PAM και το GMM εμφανίζουν υψηλή μέση επίδοση και συγκρίσιμη μεταβλητότητα, ενώ ο DBSCAN χαρακτηρίζεται τόσο από χαμηλότερη μέση απόδοση όσο και από σαφώς μεγαλύτερη διασπορά. Τα ευρήματα αυτά υποδεικνύουν ότι, στο συγκεκριμένο πειραματικό πλαίσιο, οι κλασικές μέθοδοι ομαδοποίησης παρουσιάζουν μεγαλύτερη ανθεκτικότητα στις εξεταζόμενες διαταραχές των δεδομένων, ενώ η απόδοση του DBSCAN εξαρτάται εντονότερα από τις συνθήκες εφαρμογής και την επιλογή των παραμέτρων του.

Το Διάγραμμα 5.6.1 απεικονίζει τη μέση τιμή του Adjusted Rand Index (ARI) για τους διαφορετικούς συνδυασμούς αλγορίθμων και μέτρων απόστασης. Παρατηρείται ότι το GMM, οι παραλλαγές της ιεραρχικής ομαδοποίησης (Average, Complete και Ward.D2), η k-means και η PAM παρουσιάζουν ουσιαστικά ταυτόσημη μέση επίδοση, με $\text{Mean ARI} = 0,903$. Αντίθετα, ο DBSCAN διαφοροποιείται έντονα από τις υπόλοιπες μεθόδους, παρουσιάζοντας σημαντικά χαμηλότερη μέση τιμή ARI, ίση με 0,296. Η γραφική απεικόνιση αναδεικνύει επομένως τη σταθερά υψηλή μέση απόδοση των κλασικών μεθόδων ομαδοποίησης στις εξεταζόμενες πειραματικές συνθήκες, καθώς και τη σαφώς χαμηλότερη συνολική επίδοση του DBSCAN.

Συνολικά, η ανάλυση ευαισθησίας και ανθεκτικότητας ανέδειξε σαφείς διαφοροποιήσεις μεταξύ των εξεταζόμενων μεθόδων ομαδοποίησης. Οι ιεραρχικές μέθοδοι, η k-means, η PAM και το GMM διατήρησαν υψηλή μέση επίδοση $\text{Mean ARI} = 0,903$ και παρόμοια επίπεδα διασποράς στις διαφορετικές πειραματικές συνθήκες. Ιδιαίτερα οι ιεραρχικές μέθοδοι με ευκλείδεια απόσταση εμφάνισαν τη μεγαλύτερη συχνότητα επικράτησης, με τη Ward.D2 να καταγράφει τις περισσότερες νίκες. Αντίθετα, ο DBSCAN παρουσίασε σημαντικά χαμηλότερη μέση επίδοση $\text{Mean ARI} = 0,296$ και αισθητά μεγαλύτερη μεταβλητότητα $SD = 0,399$, γεγονός που καταδεικνύει αυξημένη ευαισθησία στις συνθήκες εφαρμογής και στις υπερπαραμέτρους του. Επομένως, τα αποτελέσματα επιβεβαιώνουν ότι η ανθεκτικότητα μιας μεθόδου ομαδοποίησης δεν εξαρτάται μόνο από τη μέση απόδοσή της, αλλά και από τη σταθερότητά της όταν μεταβάλλονται τα χαρακτηριστικά των δεδομένων και οι παράμετροι της ανάλυσης.

5.7 Οπτικοποίηση και ερμηνεία σενάριο 6 (Investors / Portfolios)

Στην παρούσα ενότητα παρουσιάζεται η οπτική διερεύνηση των αποτελεσμάτων του σεναρίου 6 με στόχο την καλύτερη κατανόηση της δομής των clusters που προέκυψαν από τη διαδικασία clustering. Ενώ οι προηγούμενες ενότητες αξιολόγησαν ποσοτικά την ποιότητα των μεθόδων μέσω δεικτών όπως ο Adjusted Rand Index (ARI) και ο δείκτης Silhouette, εδώ δίνεται έμφαση στη γραφική ερμηνεία των αποτελεσμάτων. Μέσω ενός δένδρουγράμματος, ενός heatmap συσχέτισης και της απεικόνισης των μέσων χρονοσειρών ανά cluster, επιχειρείται να αναδειχθεί η εσωτερική δομή των ομάδων επενδυτών και να επιβεβαιωθεί ο διαχωρισμός που εντοπίστηκε από τους αλγορίθμους.

Το διάγραμμα 5.7.1 παρουσιάζει το δενδρόγραμμα που προκύπτει από την εφαρμογή ιεραρχικού clustering με βάση την απόσταση Pearson (1-correlation). Η συγκεκριμένη μετρική χρησιμοποιείται συχνά σε χρονοσειρές καθώς αποτυπώνει τη ομοιότητα των σειρών δηλαδή το κατά πόσο κινούνται με παρόμοιο τρόπο, ανεξάρτητα από το απόλυτο επίπεδο τιμών. Στο διάγραμμα διακρίνονται τρεις βασικές ομάδες επενδυτών οι οποίες σχηματίζονται σε σχετικά χαμηλό επίπεδο απόστασης. Η κόκκινη γραμμή υποδεικνύει το επίπεδο στο οποίο πραγματοποιείται η τομή του δέντρου ώστε να προκύψουν τρία clusters. Παρατηρείται ότι οι ομάδες εμφανίζουν σαφή διαχωρισμό, γεγονός που επιβεβαιώνει την υψηλή απόδοση των μεθόδων clustering που καταγράφηκε στην προηγούμενη ανάλυση (ARI κοντά στο 1). Η δομή του δενδρογράμματος υποδηλώνει ότι οι επενδυτές εντός του ίδιου cluster παρουσιάζουν υψηλό βαθμό ομοιότητας ως προς την εξέλιξη των αποδόσεών τους.

Το διάγραμμα 5.7.2 παρουσιάζει έναν heatmap του πίνακα συσχέτισης μεταξύ των επενδυτών αφού πρώτα ταξινομήθηκαν με βάση τα clusters που προέκυψαν από το clustering. Οι χρωματικές εντάσεις αντιπροσωπεύουν το επίπεδο συσχέτισης μεταξύ των χρονοσειρών. Η εικόνα αποκαλύπτει τρεις διακριτές τετραγωνικές περιοχές κατά μήκος της κύριας διαγωνίου οι οποίες αντιστοιχούν στα clusters επενδυτών. Οι περιοχές αυτές χαρακτηρίζονται από υψηλή εσωτερική συσχέτιση, γεγονός που υποδηλώνει ότι οι επενδυτές εντός της ίδιας ομάδας παρουσιάζουν παρόμοια δυναμική συμπεριφορά. Αντίθετα, οι περιοχές εκτός διαγωνίου εμφανίζουν χαμηλότερες τιμές συσχέτισης, γεγονός που επιβεβαιώνει τη διαφοροποίηση μεταξύ των clusters. Η συγκεκριμένη οπτικοποίηση παρέχει μια ιδιαίτερα σαφή εικόνα της δομής των δεδομένων και επιβεβαιώνει ότι το clustering δεν είναι αποτέλεσμα τυχαίας ομαδοποίησης αλλά αντανακλά πραγματικές ομοιότητες μεταξύ των χρονοσειρών.

Το διάγραμμα 5.7.3 απεικονίζει τη μέση χρονοσειρά για κάθε cluster επενδυτών, επιτρέποντας την ποιοτική σύγκριση της δυναμικής συμπεριφοράς των ομάδων. Κάθε γραμμή αντιστοιχεί στη μέση εξέλιξη των χρονοσειρών που ανήκουν στο αντίστοιχο cluster. Παρατηρείται ότι τα τρία clusters εμφανίζουν διαφορετικά πρότυπα εξέλιξης. Ένα από τα clusters παρουσιάζει έντονα ανοδική πορεία, υποδηλώνοντας μια ομάδα επενδυτών με συστηματικά αυξανόμενες αποδόσεις. Ένα δεύτερο cluster εμφανίζει πιο σταθερή και ήπια μεταβολή, ενώ ένα τρίτο cluster χαρακτηρίζεται από χαμηλότερα επίπεδα και μεγαλύτερη μεταβλητότητα. Οι διαφορές αυτές επιβεβαιώνουν ότι τα clusters δεν διαχωρίζονται μόνο στατιστικά, αλλά αντανακλούν διαφορετικά μοτίβα συμπεριφοράς των επενδυτών.

Συνολικά, τα γραφήματα της ενότητας αυτής επιβεβαιώνουν οπτικά τα ποσοτικά αποτελέσματα που παρουσιάστηκαν στις προηγούμενες ενότητες. Το δενδρόγραμμα αποκαλύπτει καθαρή ιεραρχική δομή, ο heatmap συσχέτισης δείχνει ισχυρή εσωτερική συνοχή εντός των clusters, ενώ η απεικόνιση των μέσων χρονοσειρών αναδεικνύει ουσιαστικές διαφορές στη δυναμική των ομάδων επενδυτών.

5.8 Έλεγχος σταθερότητας με επαναλαμβανόμενη υποδειγματοληψία

Στην παρούσα ενότητα εξετάζεται η σταθερότητα των αποτελεσμάτων clustering του σεναρίου 6 μέσω επαναλαμβανόμενης υπό- δειγματοληψίας. Σκοπός της διαδικασίας είναι να αξιολογηθεί κατά πόσο τα clusters που προέκυψαν από την αρχική ανάλυση παραμένουν

συνεπή όταν το σύνολο των δεδομένων μεταβάλλεται ελαφρώς. Η σταθερότητα αποτελεί σημαντικό κριτήριο αξιολόγησης ενός αποτελέσματος clustering, καθώς επιτρέπει να εκτιμηθεί ο βαθμός στον οποίο η δομή των ομάδων διατηρείται υπό μεταβολές του δείγματος. Για τον έλεγχο της σταθερότητας πραγματοποιήθηκαν 200 επαναλήψεις επαναλαμβανόμενης υπό-δειγματοληψίας. Σε κάθε επανάληψη επιλέχθηκε τυχαία το 80% των επενδυτών χωρίς επανάθεση και εφαρμόστηκε εκ νέου η διαδικασία clustering. Η συνολική συμφωνία της νέας ομαδοποίησης με την αρχική αξιολογήθηκε μέσω του Adjusted Rand Index (ARI).

Ο Πίνακας 5.8.1 παρουσιάζει τα βασικά στατιστικά χαρακτηριστικά της κατανομής των τιμών ARI που προέκυψαν από τις επαναλαμβανόμενες υπό - δειγματοληψίες. Η μέση τιμή του ARI υποδηλώνει μέτριο επίπεδο συμφωνίας μεταξύ των ομαδοποιήσεων που προκύπτουν από τα επιμέρους υποδείγματα και της αρχικής ομαδοποίησης. Η σχετικά μικρή τυπική απόκλιση δείχνει ότι το επίπεδο αυτό συμφωνίας εμφανίζεται με συνέπεια μεταξύ των επαναλήψεων. Επιπλέον, οι τιμές του δείκτη κυμαίνονται από 0.409 έως 0.556, γεγονός που υποδηλώνει σχετικά περιορισμένη διακύμανση. Συνεπώς, τα αποτελέσματα υποδεικνύουν επαναλαμβανόμενη μέτρια συμφωνία και όχι ισχυρή συνολική σταθερότητα της ομαδοποίησης.

Το Διάγραμμα 5.8.1 απεικονίζει την κατανομή των τιμών ARI που προέκυψαν από τις επαναλαμβανόμενες υπό - δειγματοληψίες. Παρατηρείται ότι η πλειονότητα των τιμών συγκεντρώνεται σε σχετικά στενό εύρος, ενώ εμφανίζονται και ορισμένες υψηλότερες τιμές. Η περιορισμένη διασπορά δείχνει ότι η μέτρια συμφωνία με την αρχική ομαδοποίηση εμφανίζεται με συνέπεια στις επαναλήψεις, χωρίς όμως να αποτελεί ένδειξη ισχυρής σταθερότητας. Για την πληρέστερη διερεύνηση της σταθερότητας εξετάστηκε επιπλέον η επαναληψιμότητα της σύνθεσης κάθε cluster ξεχωριστά μέσω του δείκτη Jaccard. Για κάθε αρχικό cluster υπολογίστηκε, σε κάθε επανάληψη, η μέγιστη επικάλυψη με τα clusters της νέας ομαδοποίησης, ώστε να αντιμετωπιστεί η ενδεχόμενη αλλαγή των αριθμητικών ετικετών των clusters. Τα αποτελέσματα παρουσιάζονται στον Πίνακα 5.8.2. Το Cluster 1 παρουσίασε μέση τιμή Jaccard 0.502, με διάμεσο 0.500 και 95% εμπειρικό διάστημα 0.447-0.565. Το Cluster 2 εμφάνισε την υψηλότερη μέση τιμή Jaccard, ίση με 0.662, με διάμεσο 0.654 και 95% εμπειρικό διάστημα 0.500-0.917. Τέλος, το Cluster 3 παρουσίασε μέση τιμή Jaccard 0.498, με διάμεσο 0.500 και 95% εμπειρικό διάστημα 0.435-0.553. Επομένως, το Cluster 2 παρουσιάζει υψηλότερο βαθμό επαναληψιμότητας ως προς τη σύνθεση των μελών του σε σύγκριση με τα Clusters 1 και 3, όπως υποδηλώνει η υψηλότερη μέση τιμή του δείκτη Jaccard. Ωστόσο, το ευρύτερο 95% εμπειρικό διάστημα υποδηλώνει μεγαλύτερη μεταβλητότητα μεταξύ των επαναλήψεων. Το Διάγραμμα 5.8.2 παρουσιάζει την κατανομή των τιμών Jaccard για καθένα από τα τρία clusters. Η διαφοροποίηση μεταξύ των κατανομών επιβεβαιώνει ότι η σταθερότητα δεν είναι ομοιόμορφη μεταξύ των ομάδων. Ειδικότερα, το Cluster 2 εμφανίζει υψηλότερες τιμές Jaccard, ενώ τα Clusters 1 και 3 παρουσιάζουν χαμηλότερες και αρκετά παρόμοιες τιμές.

Συμπερασματικά, η ανάλυση επαναλαμβανόμενης υπό - δειγματοληψίας δείχνει ότι η ομαδοποίηση του σεναρίου 6 παρουσιάζει επαναλαμβανόμενη μέτρια συνολική συμφωνία με την αρχική λύση, όπως προκύπτει από τις τιμές του ARI. Η περιορισμένη διακύμανση του ARI

δείχνει ότι το επίπεδο αυτό συμφωνίας εμφανίζεται με συνέπεια στις επαναλήψεις, χωρίς ωστόσο να τεκμηριώνει ισχυρή συνολική σταθερότητα. Παράλληλα, η ανάλυση μέσω του δείκτη Jaccard ανέδειξε διαφοροποιήσεις στη σταθερότητα των επιμέρους clusters, με το Cluster 2 να παρουσιάζει μεγαλύτερη επαναληψιμότητα και τα Clusters 1 και 3 χαμηλότερη σταθερότητα ως προς τη σύνθεση των μελών τους. Συνεπώς, η σταθερότητα της συγκεκριμένης λύσης clustering μπορεί να χαρακτηριστεί ως μέτρια και ετερογενής μεταξύ των επιμέρους clusters.

5.9 Συμπεράσματα

Το παρόν κεφάλαιο διερεύνησε εμπειρικά τον ρόλο που διαδραματίζουν τα μέτρα απόστασης και οι αλγόριθμοι ομαδοποίησης στη διαδικασία του clustering μέσα από μια σειρά ελεγχόμενων πειραμάτων με συνθετικά αλλά οικονομικά εμπνευσμένα δεδομένα. Η βασική επιδίωξη δεν ήταν απλώς η σύγκριση επιδόσεων μεταξύ διαφορετικών τεχνικών, αλλά κυρίως η κατανόηση των συνθηκών υπό τις οποίες συγκεκριμένες επιλογές μετρικής και αλγορίθμου αποδεικνύονται κατάλληλες ή ανεπαρκείς. Η ανάλυση έδειξε ότι η επιτυχία μιας μεθόδου clustering δεν εξαρτάται αποκλειστικά από τον αλγόριθμο, αλλά σε μεγάλο βαθμό από τη γεωμετρική και στατιστική δομή των δεδομένων. Στα αρχικά πειράματα, όπου οι συστάδες ήταν καθαρά διαχωρισμένες, οι περισσότερες κλασικές μετρικές αποστάσεων οδήγησαν σε σχεδόν τέλεια ανάκτηση των πραγματικών ομάδων. Το αποτέλεσμα αυτό υποδηλώνει ότι όταν η δομή των δεδομένων είναι ισχυρά διακριτή, διαφορετικές μορφές γεωμετρικής απόστασης όπως οι νόρμες L_1 , L_2 και συναφείς γενικεύσεις συγκλίνουν ουσιαστικά στο ίδιο αποτέλεσμα. Ωστόσο, η εικόνα διαφοροποιήθηκε σημαντικά όταν χρησιμοποιήθηκαν μετρικές που βασίζονται σε διαφορετική έννοια ομοιότητας, όπως η Pearson ή άλλες κανονικοποιημένες αποστάσεις, οι οποίες σε αρκετές περιπτώσεις υποβάθμισαν την ποιότητα της ομαδοποίησης. Το εύρημα αυτό αναδεικνύει ότι η καταλληλότητα μιας μετρικής δεν είναι γενική ιδιότητα αλλά εξαρτάται από το αν η πληροφορία που διαχωρίζει τις ομάδες σχετίζεται με απόλυτες διαφορές τιμών ή με μοτίβα συσχέτισης.

Η εισαγωγή πιο απαιτητικών συνθηκών στα δεδομένα κατέδειξε ακόμη πιο καθαρά τη σημασία της δομής του προβλήματος. Η παρουσία ακραίων τιμών δεν επηρέασε ουσιαστικά την απόδοση των περισσότερων αλγορίθμων όταν οι βασικές συστάδες παρέμεναν σαφώς διαχωρισμένες. Αντίθετα, η εμφάνιση μη σφαιρικών συστάδων και η μεταβολή της κλίμακας των μεταβλητών επηρέασαν αισθητά την ποιότητα της ομαδοποίησης. Ιδιαίτερα η περίπτωση των ανισόκλιμακων μεταβλητών κατέδειξε ότι η τυποποίηση των δεδομένων αποτελεί συχνά αναγκαίο βήμα προεπεξεργασίας, καθώς διαφορετικά οι αποστάσεις μπορεί να κυριαρχούνται από μία μόνο διάσταση, οδηγώντας σε παραπλανητικές καταταμήσεις.

Η εφαρμογή των μεθόδων σε πιο ρεαλιστικά οικονομικά σενάρια ανέδειξε μια ακόμη σημαντική διάσταση. Στο σενάριο των επιχειρήσεων, όπου οι μεταβλητές εμφανίζουν συσχετίσεις και σημαντική επικάλυψη μεταξύ των ομάδων, η ανάκτηση των πραγματικών clusters αποδείχθηκε δύσκολη ανεξαρτήτως μετρικής. Το αποτέλεσμα αυτό δείχνει ότι ακόμη και εξελεγχόμενες αποστάσεις όπως η Mahalanobis ή η robust εκδοχή της μπορούν να βελτιώσουν

μόνο οριακά την εικόνα όταν οι ομάδες δεν είναι εγγενώς διακριτές. Αντίθετα, στο σενάριο των χαρτοφυλακίων επενδυτών, όπου η δομή των δεδομένων ήταν σαφώς διαχωρισμένη, τόσο η Pearson όσο και η Euclidean απόσταση κατάφεραν να ανακτήσουν πλήρως τις πραγματικές ομάδες. Η σύγκριση των δύο περιπτώσεων υπογραμμίζει ότι η αποτελεσματικότητα μιας μεθόδου clustering καθορίζεται πρωτίστως από το πόσο έντονα διαφοροποιούνται οι υποκείμενες δομές των δεδομένων.

Συνολικά, τα ευρήματα του κεφαλαίου οδηγούν σε ένα γενικό συμπέρασμα ότι δεν υπάρχει μία καθολικά βέλτιστη μετρική ή ένας αλγόριθμος clustering που να υπερέχει σε κάθε περίπτωση. Η επιλογή πρέπει να γίνεται με βάση τη φύση των δεδομένων, τον τύπο των μεταβλητών και το είδος της ομοιότητας που επιθυμεί να αποτυπώσει ο αναλυτής. Σε προβλήματα όπου οι ομάδες διαφοροποιούνται κυρίως ως προς επίπεδα τιμών, οι κλασικές γεωμετρικές αποστάσεις αποδεικνύονται επαρκείς. Αντίθετα, σε περιπτώσεις όπου το ενδιαφέρον εστιάζεται σε μοτίβα μεταβολής ή σε χρονικές δυναμικές, αποστάσεις βασισμένες στη συσχέτιση μπορεί να αποδειχθούν καταλληλότερες.

Καταληκτικά, η παρούσα μελέτη αναδεικνύει ότι το clustering αποτελεί διαδικασία που απαιτεί προσεκτική μεθοδολογική επιλογή και κατανόηση του πλαισίου εφαρμογής. Η σωστή αντιστοίχιση μεταξύ τύπου δεδομένων, μετρικής απόστασης και αλγορίθμου ομαδοποίησης αποτελεί κρίσιμο παράγοντα για την παραγωγή αξιόπιστων και ερμηνεύσιμων αποτελεσμάτων στην εμπειρική ανάλυση οικονομικών δεδομένων.

ΚΕΦΑΛΑΙΟ 6

Συμπεράσματα

6.1 Ανακεφαλαίωση της εργασίας

Η παρούσα διπλωματική εργασία εστίασε στη μελέτη των μέτρων απόστασης και ομοιότητας σε πολυδιάστατους χώρους, καθώς και στον ρόλο που αυτά διαδραματίζουν στη διαδικασία της ομαδοποίησης δεδομένων. Η έννοια της απόστασης αποτελεί θεμελιώδες εργαλείο στη στατιστική ανάλυση και στη μηχανική μάθηση. Στο θεωρητικό μέρος της εργασίας παρουσιάστηκαν τα βασικότερα μέτρα απόστασης, καθώς και εξειδικευμένα μέτρα που χρησιμοποιούνται σε ειδικές περιπτώσεις, όπως η απόσταση Gower για μικτά δεδομένα και η απόσταση Canberra. Παράλληλα, αναλύθηκαν σημαντικά μέτρα ομοιότητας, όπως οι δείκτες Jaccard, Cosine και Sorensen-Dice, τα οποία βρίσκουν εφαρμογή σε δυαδικά και κατηγορικά δεδομένα.

Στη συνέχεια, εξετάστηκαν οι βασικές μέθοδοι ομαδοποίησης (clustering), τόσο ιεραρχικές όσο και μη ιεραρχικές, καθώς και πιο σύγχρονες προσεγγίσεις που βασίζονται στην πυκνότητα και σε μοντέλα πιθανοτήτων. Αναλύθηκε ο τρόπος με τον οποίο τα μέτρα απόστασης ενσωματώνονται στους αλγορίθμους αυτούς και επηρεάζουν τον σχηματισμό των συστάδων. Το πρακτικό μέρος της εργασίας επικεντρώθηκε στην εφαρμογή διαφορετικών μετρικών αποστάσεων σε ποικίλα σενάρια δεδομένων, συμπεριλαμβανομένων αριθμητικών, δυαδικών, χρονικών και μικτών δεδομένων. Μέσω της χρήσης συνθετικών δεδομένων, πραγματοποιήθηκε συγκριτική αξιολόγηση των αποτελεσμάτων ομαδοποίησης. Η ανάλυση περιλάμβανε διαφορετικούς αλγορίθμους clustering, καθώς και τεχνικές αξιολόγησης της ποιότητας των συστάδων, με στόχο την αποτίμηση της αποτελεσματικότητας κάθε μετρικής.

Συνολικά, η εργασία συνδύασε θεωρητική ανάλυση και πειραματική διερεύνηση, επιδιώκοντας να προσφέρει μια ολοκληρωμένη εικόνα για τον ρόλο των μέτρων απόστασης και ομοιότητας στην ανάλυση πολυδιάστατων δεδομένων. Μέσα από αυτή τη συνδυαστική προσέγγιση αναδείχθηκε η σημασία της σωστής επιλογής μετρικής.

6.2 Κύρια Ερευνητικά Ευρήματα

Τα αποτελέσματα της παρούσας μελέτης ανέδειξαν την επιλογή του μέτρου απόστασης που αποτελεί καθοριστικό παράγοντα των μεθόδων ομαδοποίησης. Δεν πρόκειται για μια απλή τεχνική επιλογή, αλλά για μια απόφαση που επηρεάζει τη γεωμετρία του χώρου των δεδομένων. Ένα από τα βασικά ευρήματα είναι ότι η ευκλείδεια απόσταση, αν και ευρέως χρησιμοποιούμενη λόγω της απλότητάς της, δεν είναι πάντοτε η καταλληλότερη επιλογή. Σε περιπτώσεις όπου οι μεταβλητές έχουν διαφορετικές κλίμακες ή παρουσιάζουν ισχυρές συσχετίσεις, η χρήση της μπορεί να οδηγήσει σε παραπλανητικά αποτελέσματα. Σε περιπτώσεις όπου οι μεταβλητές έχουν διαφορετικές κλίμακες ή παρουσιάζουν ισχυρές

συσχετίσεις, η απόσταση Mahalanobis αποτελεί θεωρητικά κατάλληλη επιλογή, καθώς λαμβάνει υπόψη τη δομή συνδιακύμανσης των δεδομένων. Ωστόσο, στην παρούσα μελέτη η καταλληλότητα αυτή δεν μεταφράστηκε σε σαφή εμπειρική υπεροχή ως προς τον δείκτη ARI, όπου οι διαφορές μεταξύ των εξεταζόμενων μέτρων απόστασης ήταν περιορισμένες.

Παράλληλα, για δεδομένα δυαδικού χαρακτήρα, το μέτρο ομοιότητας Jaccard κρίθηκε ιδιαίτερα κατάλληλο, καθώς εστιάζει στην παρουσία κοινών χαρακτηριστικών και αγνοεί τις κοινές απουσίες. Σημαντικά αποτελέσματα προέκυψαν επίσης για τα μικτά δεδομένα, όπου η απόσταση Gower αποδείχθηκε ιδιαίτερα αποτελεσματική, καθώς μπορεί να συνδυάσει διαφορετικούς τύπους μεταβλητών σε ένα ενιαίο πλαίσιο. Επιπλέον, διαπιστώθηκε ότι η απόδοση των μετρικών εξαρτάται σε μεγάλο βαθμό από τη φύση των δεδομένων και τον αλγόριθμο ομαδοποίησης που χρησιμοποιείται. Δεν υπάρχει ένα βέλτιστο μέτρο απόστασης για όλες τις περιπτώσεις. Αντίθετα, η καταλληλότητα κάθε μετρικής είναι συνάρτηση των χαρακτηριστικών των δεδομένων, όπως η κλίμακα, η κατανομή, η παρουσία θορύβου και οι μεταξύ τους συσχετίσεις. Ένα ακόμη σημαντικό εύρημα αφορά την ευαισθησία των αποτελεσμάτων σε αλλαγές της μετρικής. Μικρές διαφοροποιήσεις στον τρόπο υπολογισμού της απόστασης μπορούν να οδηγήσουν σε σημαντικά διαφορετικές ομαδοποιήσεις.

Τέλος, η πειραματική ανάλυση ανέδειξε τη σημασία της προεπεξεργασίας των δεδομένων, όπως η κανονικοποίηση και η στάθμιση των μεταβλητών. Οι διαδικασίες αυτές επηρεάζουν άμεσα τις αποστάσεις και συνεπώς τη δομή των συστάδων. Η σωστή προετοιμασία των δεδομένων αποτελεί, επομένως, αναπόσπαστο μέρος της διαδικασίας ομαδοποίησης. Συνοψίζοντας, τα ευρήματα της εργασίας επιβεβαιώνουν ότι τα μέτρα απόστασης και ομοιότητας δεν αποτελούν απλώς υπολογιστικά εργαλεία, αλλά ενσωματώνουν ουσιαστικές υποθέσεις για τη δομή των δεδομένων. Η κατανόηση και η ορθή επιλογή τους είναι κρίσιμη για την εξαγωγή αξιόπιστων και ερμηνεύσιμων αποτελεσμάτων στην ανάλυση δεδομένων.

6.3 Θεωρητική συνεισφορά της εργασίας

Στην παρούσα εργασία αλλά και γενικώς κάθε μετρική ορίζει ένα διαφορετικό γεωμετρικό πλαίσιο μέσα στο οποίο αναλύονται τα δεδομένα. Στο πλαίσιο αυτό, η εργασία αναδεικνύει τη στενή σχέση μεταξύ μετρικής απόστασης και διαδικασίας ομαδοποίησης. Οι αλγόριθμοι clustering δεν λειτουργούν ανεξάρτητα από τη μετρική που χρησιμοποιείται αντίθετα, η μετρική αποτελεί ενσωματωμένο μέρος της λειτουργίας τους και επηρεάζει άμεσα το αποτέλεσμα. Συνεπώς, η ομαδοποίηση μπορεί να θεωρηθεί ως προβολή της δομής των δεδομένων υπό ένα συγκεκριμένο μέτρο απόστασης. Επιπλέον, αναδεικνύεται η έννοια της σχετικότητας της απόστασης. Η απόσταση δεν αποτελεί απόλυτο μέγεθος, αλλά εξαρτάται από την κλίμακα των δεδομένων, τη φύση των μεταβλητών και τον τρόπο προεπεξεργασίας. Η ίδια ομάδα δεδομένων μπορεί να παρουσιάζει διαφορετική δομή όταν μετράται με διαφορετικά μέτρα απόστασης, γεγονός που υπογραμμίζει τη σημασία της επιλογής της κατάλληλης μετρικής.

Ιδιαίτερη έμφαση δίνεται επίσης στη δυαδικότητα μεταξύ απόστασης και ομοιότητας. Η εργασία δείχνει ότι οι δύο έννοιες αποτελούν διαφορετικές εκφράσεις της ίδιας βασικής ιδέας και ότι η μετάβαση από τη μία στην άλλη δεν είναι μόνο μαθηματικά εφικτή αλλά και πρακτικά

χρήσιμη, ανάλογα με το είδος των δεδομένων και τον στόχο της ανάλυσης. Συνολικά, η θεωρητική συνεισφορά της εργασίας έγκειται στο ότι προσεγγίζει τα μέτρα απόστασης και ομοιότητας ως θεμελιώδεις έννοιες που διαμορφώνουν τον τρόπο με τον οποίο αναπαρίσταται και ερμηνεύεται η πληροφορία.

6.4 Πρακτικές κατευθυντήριες γραμμές για την επιλογή μέτρου απόστασης

Με βάση όσα είδαμε και αναφέραμε, καταλήγουμε για δεδομένα που αποτελούνται αποκλειστικά από ποσοτικές μεταβλητές με παρόμοια κλίμακα και χωρίς έντονες συσχετίσεις, η ευκλείδεια απόσταση αποτελεί μια απλή και αποτελεσματική επιλογή. Ωστόσο, όταν οι μεταβλητές έχουν διαφορετικά εύρη τιμών, είναι απαραίτητη η κανονικοποίηση των δεδομένων πριν από την εφαρμογή της, ώστε να αποφεύγεται η υπερεκτίμηση της σημασίας ορισμένων μεταβλητών.

Σε περιπτώσεις όπου οι μεταβλητές παρουσιάζουν σημαντικές συσχετίσεις, η απόσταση Mahalanobis είναι προτιμότερη, καθώς λαμβάνει υπόψη τη δομή συνδιακύμανσης των δεδομένων και προσφέρει πιο αξιόπιστη εκτίμηση.

Για δεδομένα δυαδικής φύσης ή δεδομένα παρουσίας-απουσίας χαρακτηριστικών, τα μέτρα ομοιότητας όπως ο δείκτης Jaccard αποτελούν καταλληλότερη επιλογή, καθώς εστιάζουν στα κοινά χαρακτηριστικά μεταξύ των παρατηρήσεων και αγνοούν τις κοινές απουσίες, οι οποίες συχνά δεν έχουν ουσιαστική πληροφορία.

Σε περιπτώσεις μικτών δεδομένων, δηλαδή όταν συνυπάρχουν ποσοτικές και κατηγορικές μεταβλητές, η απόσταση Gower προσφέρει σημαντικά πλεονεκτήματα, καθώς επιτρέπει την ταυτόχρονη επεξεργασία διαφορετικών τύπων δεδομένων σε ένα ενιαίο πλαίσιο. Η ευελιξία αυτή την καθιστά ιδιαίτερα χρήσιμη σε πραγματικά δεδομένα, όπου η ομοιογένεια των μεταβλητών είναι σπάνια. Επιπλέον, για δεδομένα υψηλής διάστασης ή αραιά δεδομένα (sparse data), όπως αυτά που προκύπτουν από κείμενα ή embeddings, μέτρα ομοιότητας όπως η cosine similarity συχνά αποδίδουν καλύτερα, καθώς βασίζονται στη γωνιακή σχέση μεταξύ διανυσμάτων και όχι στο απόλυτο μέγεθος των τιμών.

Τέλος, η εργασία αναδεικνύει τη σημασία της πειραματικής αξιολόγησης. Η δοκιμή διαφορετικών μετρικών και η σύγκριση των αποτελεσμάτων μέσω κατάλληλων δεικτών αξιολόγησης αποτελεί αναγκαίο βήμα για την επιλογή της βέλτιστης προσέγγισης σε κάθε συγκεκριμένο πρόβλημα. Συμπερασματικά, η επιλογή του μέτρου απόστασης δεν αποτελεί δευτερεύουσα τεχνική λεπτομέρεια, αλλά βασικό στοιχείο της αναλυτικής διαδικασίας, το οποίο καθορίζει την ποιότητα, την αξιοπιστία και την ερμηνευσιμότητα των αποτελεσμάτων.

6.5 Μελλοντικές κατευθύνσεις έρευνας

Η παρούσα εργασία ανοίγει τον δρόμο για περαιτέρω ερευνητική διερεύνηση στον τομέα των μέτρων απόστασης και ομοιότητας, ιδιαίτερα στο πλαίσιο της ανάλυσης πολυδιάστατων δεδομένων και της μηχανικής μάθησης.

Μία σημαντική κατεύθυνση αφορά την ανάπτυξη και μελέτη προσαρμοστικών ή μαθησιακών μετρικών αποστάσεων (metric learning). Σε αυτές τις προσεγγίσεις, το ίδιο το

μέτρο απόστασης δεν είναι προκαθορισμένο, αλλά προκύπτει μέσω διαδικασιών μάθησης από τα δεδομένα, με στόχο τη βελτιστοποίηση της απόδοσης σε συγκεκριμένες εφαρμογές, όπως η ταξινόμηση και η ομαδοποίηση.

Επιπλέον, ιδιαίτερο ενδιαφέρον παρουσιάζει η ενσωμάτωση των μέτρων απόστασης σε σύγχρονες τεχνικές αναπαράστασης δεδομένων, όπως τα embeddings που προκύπτουν από νευρωνικά δίκτυα. Στους χώρους αυτούς, η έννοια της απόστασης αποκτά διαφορετικό νόημα, καθώς τα δεδομένα μετασχηματίζονται σε νέες αναπαραστάσεις όπου η ομοιότητα αντανακλά πιο σύνθετες σχέσεις. Μία ακόμη κατεύθυνση αφορά τη μελέτη υβριδικών μετρικών, οι οποίες συνδυάζουν διαφορετικούς τύπους αποστάσεων ή ομοιοτήτων, με στόχο την καλύτερη προσαρμογή σε σύνθετα και ετερογενή δεδομένα. Τέτοιες προσεγγίσεις μπορούν να αποδειχθούν ιδιαίτερα χρήσιμες σε εφαρμογές όπου συνυπάρχουν πολλαπλοί τύποι πληροφορίας.

Παράλληλα, η διερεύνηση της συμπεριφοράς των μετρικών σε χώρους υψηλής διάστασης αποτελεί ένα σημαντικό πεδίο μελλοντικής έρευνας. Η κατανόηση των επιπτώσεων της διάστασης στην έννοια της απόστασης μπορεί να οδηγήσει στην ανάπτυξη πιο ανθεκτικών και αποτελεσματικών μεθόδων ανάλυσης. Επιπρόσθετα, θα μπορούσε να εξεταστεί η εφαρμογή των μετρικών αποστάσεων σε πιο εξειδικευμένα πεδία, όπως η ανάλυση οικονομικών δεδομένων, όπου η φύση των δεδομένων παρουσιάζει ιδιαίτερες προκλήσεις.

Τέλος, σημαντική προοπτική αποτελεί η ανάπτυξη αυτοματοποιημένων διαδικασιών επιλογής μέτρου απόστασης, οι οποίες θα μπορούν να προτείνουν την καταλληλότερη μετρική με βάση τα χαρακτηριστικά των δεδομένων και τον στόχο της ανάλυσης. Τέτοιες προσεγγίσεις μπορούν να συμβάλουν στη βελτίωση της αποδοτικότητας και της αξιοπιστίας των μεθόδων ανάλυσης δεδομένων.

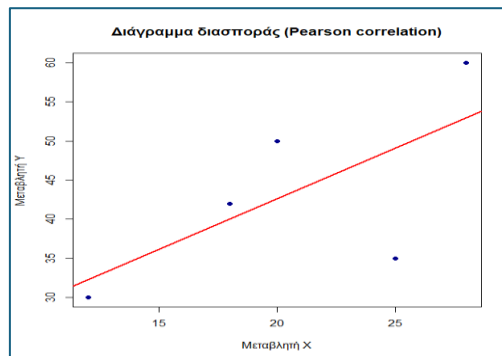
6.6 Τελικό συμπέρασμα

Η παρούσα διπλωματική εργασία ανέδειξε τη θεμελιώδη σημασία των μέτρων απόστασης και ομοιότητας στην ανάλυση πολυδιάστατων δεδομένων και στη διαδικασία της ομαδοποίησης. Μέσα από τη συνδυαστική μελέτη θεωρίας και εφαρμογών κατέστη σαφές ότι η επιλογή της κατάλληλης μετρικής δεν αποτελεί δευτερεύουσα τεχνική απόφαση, αλλά βασικό παράγοντα που επηρεάζει καθοριστικά τη δομή, την ποιότητα και την ερμηνευσιμότητα των αποτελεσμάτων.

Τα ευρήματα της εργασίας υποδεικνύουν ότι δεν υπάρχει μία ενιαία βέλτιστη μετρική για όλες τις περιπτώσεις. Αντίθετα, η καταλληλότητα κάθε μέτρου εξαρτάται από τη φύση των δεδομένων, τις ιδιότητές τους και το εκάστοτε ερευνητικό πλαίσιο. Η κατανόηση αυτής της εξάρτησης αποτελεί κρίσιμο στοιχείο για την επιτυχή εφαρμογή μεθόδων ανάλυσης δεδομένων. Παράλληλα, αναδείχθηκε ότι τα μέτρα απόστασης λειτουργούν ως μηχανισμοί που διαμορφώνουν τον ίδιο τον χώρο των δεδομένων, επηρεάζοντας τον τρόπο με τον οποίο αποτυπώνονται οι σχέσεις μεταξύ των παρατηρήσεων. Υπό αυτή την έννοια, η επιλογή της μετρικής αποτελεί ταυτόχρονα και επιλογή τρόπου ερμηνείας της πληροφορίας.

Συνοψίζοντας, η εργασία καταδεικνύει ότι η ουσιαστική κατανόηση και η ορθολογική επιλογή των μέτρων απόστασης και ομοιότητας είναι απαραίτητη προϋπόθεση για την εξαγωγή αξιόπιστων και ουσιαστικών συμπερασμάτων στην ανάλυση δεδομένων. Η συμβολή της έγκειται τόσο στη συστηματική παρουσίαση των μετρικών όσο και στην ανάδειξη της σημασίας τους ως θεμελιώδους στοιχείου της στατιστικής και της επιστήμης δεδομένων.

7. Παράρτημα



Διάγραμμα 3.2.1

Πίνακας 3.2.1

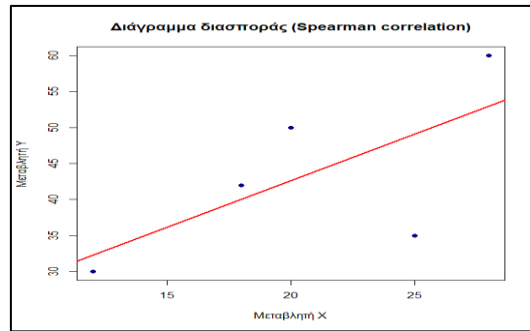
Παρατήρηση **X** **Y**

1	12	30
2	18	42
3	25	35
4	20	50
5	28	60

Πίνακας 3.3.1

Παρατήρηση **X** **Y**

1	12	30
2	18	42
3	25	35
4	20	50
5	28	60



Διάγραμμα 3.3.1

Πίνακας 3.3.2

```
# -----
# Παράδειγμα υπολογισμού του συντελεστή Spearman
# -----
# Δημιουργία δύο μεταβλητών X και Y
X <- c(12, 18, 25, 20, 28)
Y <- c(30, 42, 35, 50, 60)
# Υπολογισμός συντελεστή Spearman
spearman_result <- cor(X, Y, method = "spearman")
spearman_result
# Έλεγχος στατιστικής σημαντικότητας
corr.test(X, Y, method = "spearman")
# Διάγραμμα διασποράς με γραμμή τάσης
plot(X, Y,
     main = "Διάγραμμα διασποράς (Spearman correlation)",
     xlab = "Μεταβλητή X",
     ylab = "Μεταβλητή Y",
     pch = 19, col = "darkblue")
abline(lm(Y ~ X), col = "red", lwd = 2)
```

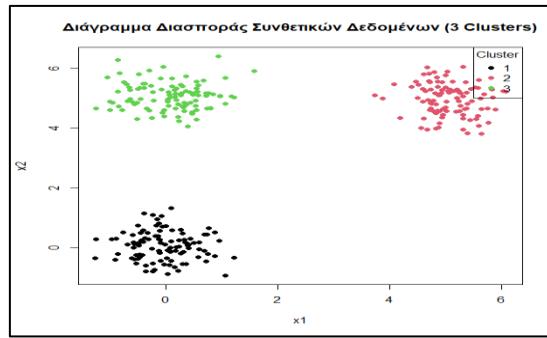
***** HIERARCHICAL CLUSTER ANALYSIS *****

Dendrogram using Single Linkage

Rescaled Distance Cluster Combine

CASE	0	3	6	7	8	9	10
Label Num	+-----+-----+-----+-----+-----+-----+						
1	-----+-----+						
2	-----+-----+					+-----+	
3	-----+				+-----+		
4	-----+			+-----+			

Διάγραμμα 4.4.1



Διάγραμμα 5.1

Πίνακας 5.1

Μετρική απόστασης	Hierarchical (complete) ARI	PAM ARI
Euclidean	1.000	1.000
Manhattan	1.000	1.000
Chebyshev (L_{∞})	1.000	1.000
Minkowski ($p = 3$)	1.000	1.000
L_1	1.000	1.000
Mahalanobis	1.000	1.000
Gower	1.000	1.000
Soergel	0.734	0.951
Kulczynski	0.454	0.980
Bhattacharyya	0.525	0.544
Canberra	0.245	0.502
Pearson	0.143	0.142

Πίνακας 5.2.1

Cluster Πλήθος Νοικοκυριών

1	50
2	50
3	50
Σύνολο	150

Πίνακας 5.2.2

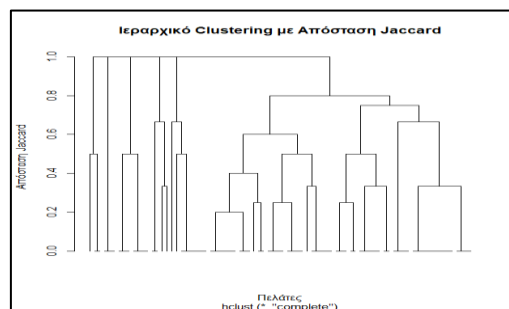
Cluster	Μέσο Εισόδημα	Μέση Αποταμίευση
1	14.95	2.12
2	29.10	8.21
3	50.88	20.40

Πίνακας 5.2.3

Cluster	Αστική	Ημιαστική	Αγροτική
1	50	0	0
2	0	50	0
3	0	0	50

Πίνακας 5.2.4

Cluster	Χαμηλό	Μεσαίο	Υψηλό
1	34	16	0
2	16	25	9
3	0	19	31



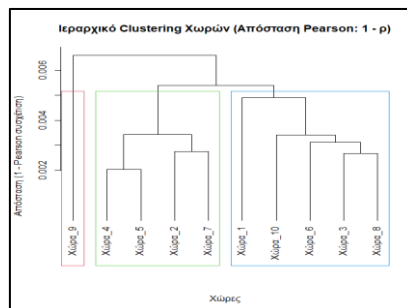
Διάγραμμα 5.2

Πίνακας 5.2.5

Cluster	Αριθμός Πελατών
1	74
2	3
3	3

Πίνακας 5.2.6

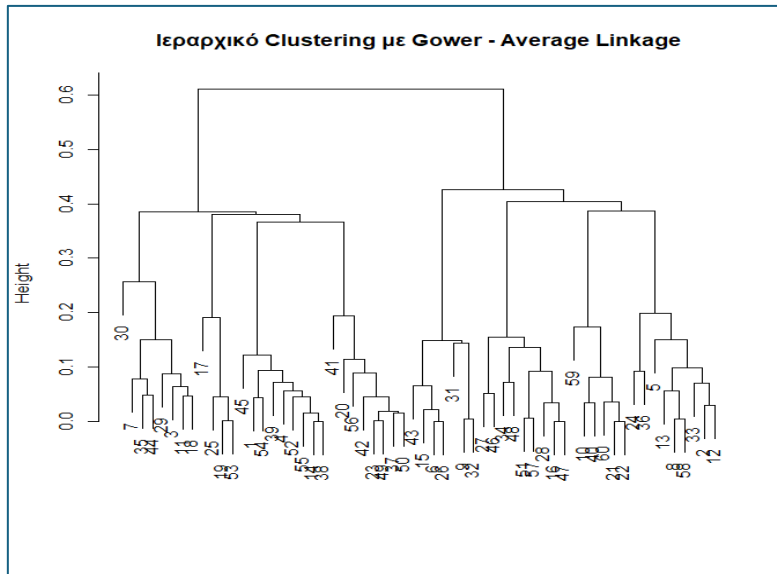
Cluster	Card	Mortgage	Consume r Loan	Investment	Insurance
1	0.82	0.39	0.46	0.32	0.58
2	0.00	0.00	0.00	0.00	0.00
3	0.00	0.00	0.00	1.00	0.67



Διάγραμμα 5.2.2

Πίνακας 5.2.7

Cluster	Χώρα
1	Χώρα_1, Χώρα_3, Χώρα_6, Χώρα_8, Χώρα_10
2	Χώρα_2, Χώρα_4, Χώρα_5, Χώρα_7
3	Χώρα_9



Διάγραμμα 5.2.3

Πίνακας 5.2.8

Cluster	Χώρες
1	29
2	24
3	7

Πίνακας 5.2.9

Cluster	gdp_pc	Unemployment
1	26.16	7.74
2	23.9	8.37
3	14.39	7.67

Πίνακας 5.2.10

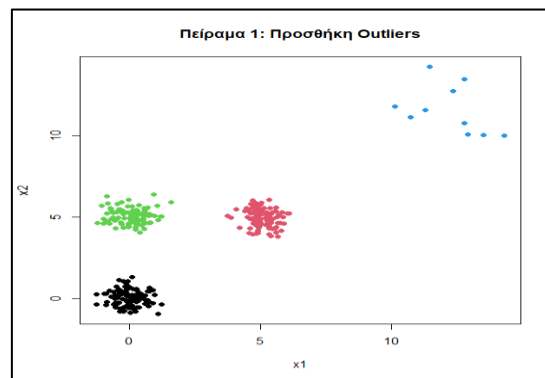
cluster	Αναπτυγμένη	Αναπτυσσόμενη
1	29	0
2	0	24
3	0	7

Πίνακας 5.2.11

cluster	Αμερική	Ασία	Αφρική	Ευρώπη
1	4	8	9	8
2	7	9	0	6
3	0	0	9	0

Πίνακας 5.3.1

Μέθοδος	ARI	Silhouette (Mean)
hclust_euc	1.000	0.818
hclust_man	1.000	0.792
hclust_cor	0.143	0.663
k-means	1.000	0.818
pam_euc	1.000	0.818
pam_man	1.000	0.792
pam_cor	0.142	0.337
dbscan	1.000	NA
gmm	1.000	0.818

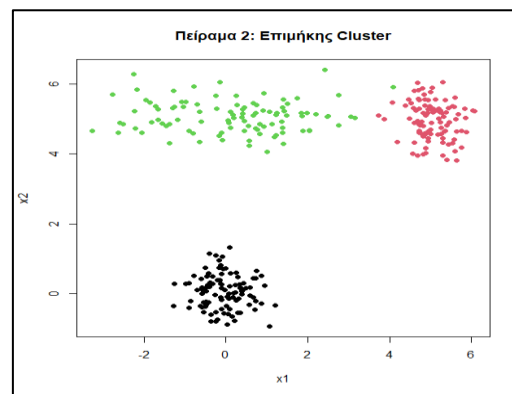


Διάγραμμα 5.3.1

Πίνακας 5.3.2

Μέθοδος	ARI	Silhouette (Mean)
hclust_euc	0.593	0.587
hclust_man	0.990	0.721
hclust_cor	0.143	0.663

k-means	0.932	0.739
pam_euc	0.951	0.740
pam_man	0.951	0.725
pam_cor	0.142	0.337
dbscan	0.570	NA
gmm	0.970	0.725



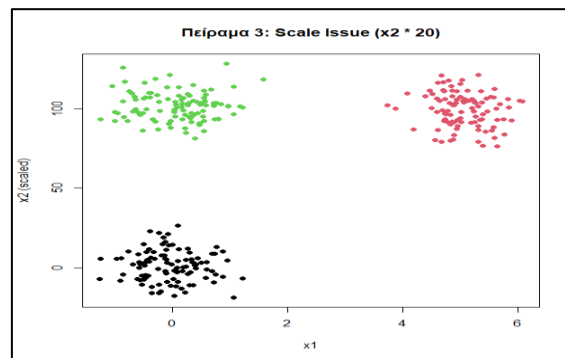
Διάγραμμα 5.3.2

Πίνακας 5.3.3

Μέθοδος	ARI	Silhouette (Mean)
hclust_euc	0.506	0.586
hclust_man	0.500	0.607
hclust_cor	0.099	0.850
k-means	0.498	0.633
pam_euc	0.497	0.633
pam_man	0.502	0.594
pam_cor	0.101	0.150
dbscan	0.570	NA
gmm	1.000	0.385

Πίνακας 5.3.4

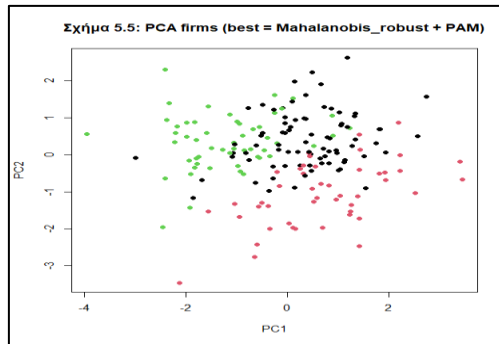
Μέθοδος	ARI	Silhouette (Mean)
hclust_euc	1.000	0.818
hclust_man	1.000	0.792
hclust_cor	0.565	0.667
k-means	1.000	0.818
pam_euc	1.000	0.818
pam_man	1.000	0.792
pam_cor	0.565	0.667
dbscan	0.000	NA
gmm	1.000	0.818



Διάγραμμα 5.3.3

Πίνακας 5.4.1

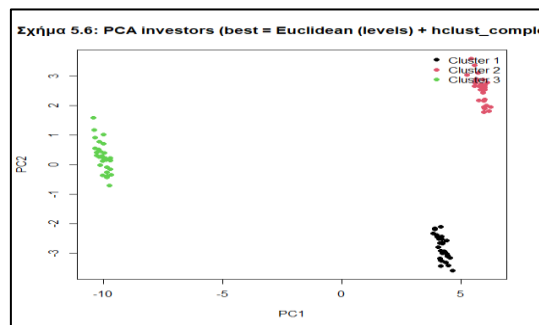
Distance	Algorithms	RI	Silhouette	Clusters	NoisePct
Euclidean	hclust_Ward.D2	- 0.001	0.171	3	0
Mahalanobis	hclust_Ward.D2	0.000	0.137	3	0
Mahalanobis_robust	hclust_Ward.D2	0.024	0.159	3	0
Mahalanobis_robust	PAM	0.038	0.165	3	0
Euclidean (GMM)	GaussianMixtures_VVV	0.014	0.130	3	0



Διάγραμμα 5.4.1

Πίνακας 5.4.2

Distance	Algorithm	ARI	Silhouette	Clusters	NoisePct
Pearson (1-corr)	hclust_complete	0.457	0.646	3	0
Pearson (1-corr)	PAM	1.000	0.693	3	0
Euclidean (levels)	hclust_complete	1.000	0.798	3	0



Διάγραμμα 5.4.2

Πίνακας 5.5.1

Σενάριο	Μέτρο Απόστασης	Αλγόριθμος	ARI	Silhouette	Clusters	Noise %
S5_firms	Euclidean	hclust_Ward.D2	- 0.001	0.171	3	0
S5_firms	Mahalanobis	hclust_Ward.D2	0.000	0.137	3	0
S5_firms	Mahalanobis (robust)	hclust_Ward.D2	0.024	0.159	3	0

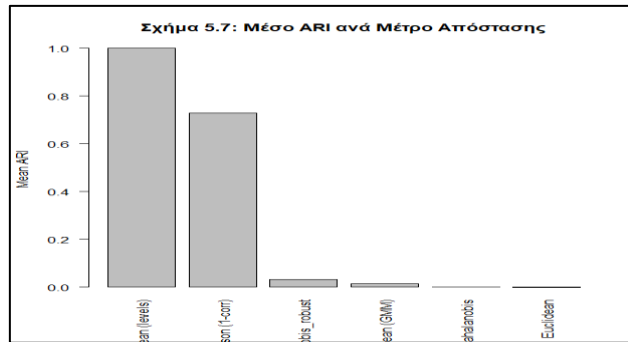
S5_firms	Mahalanobis (robust)	PAM	0.038	0.165	3	0
S5_firms	Euclidean	Gaussian Mixtures (VWV)	0.014	0.130	3	0
S6_portfolios	Pearson (1-corr)	hclust_complete	0.457	0.646	3	0
S6_portfolios	Pearson (1-corr)	PAM	1.000	0.693	3	0
S6_portfolios	Euclidean (levels)	hclust_complete	1.000	0.798	3	0

Πίνακας 5.5.2

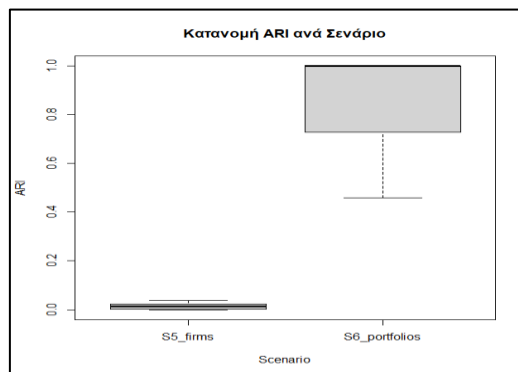
Μέτρο Απόστασης	Mean ARI
Euclidean (levels)	1.000
Pearson (1-corr)	0.729
Mahalanobis (robust)	0.031
Euclidean (GMM)	0.014
Mahalanobis	0.000
Euclidean	-0.001

Πίνακας 5.5.3

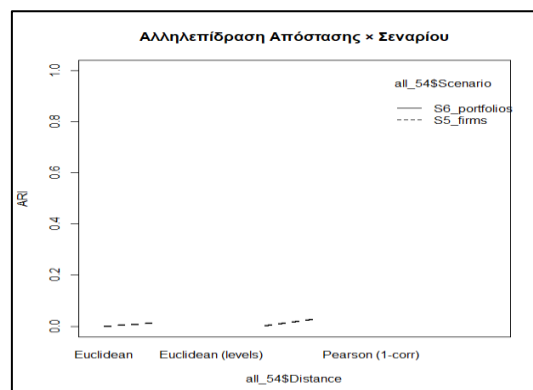
Αλγόριθμος	Mean ARI
hclust_complete	0.729
PAM	0.519
Gaussian Mixtures (VWV)	0.014
hclust_Ward.D2	0.008



Διάγραμμα 5.5.1



Διάγραμμα 5.5.2



Διάγραμμα 5.5.3

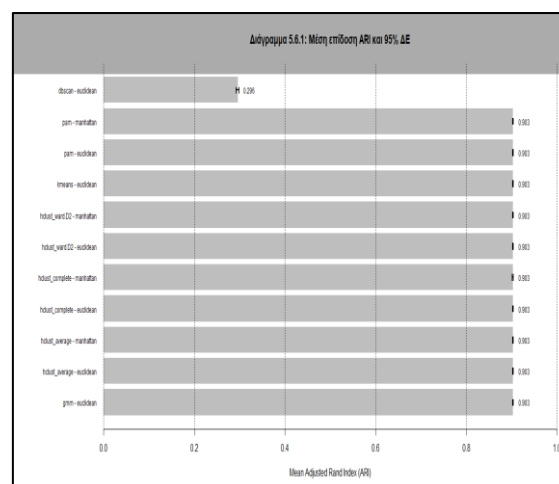
Πίνακας 5.6.1

Μέτρο Απόστασης	Αλγόριθμος	Mean ARI
Euclidean	GMM	0.903
Euclidean	hclust_average	0.903
Manhattan	hclust_average	0.903
Euclidean	hclust_complete	0.903
Manhattan	hclust_complete	0.903

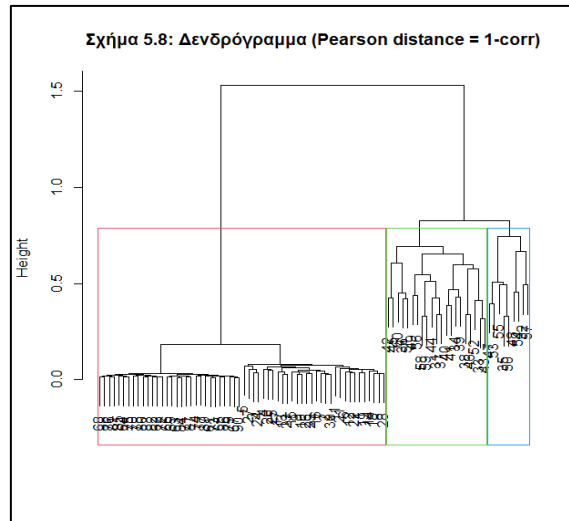
Euclidean	hclust_ward.D2	0.903
Manhattan	hclust_ward.D2	0.903
Euclidean	k-means	0.903
Euclidean	PAM	0.903
Manhattan	PAM	0.903
Euclidean	DBSCAN	0.296

Πίνακας 5.6.2

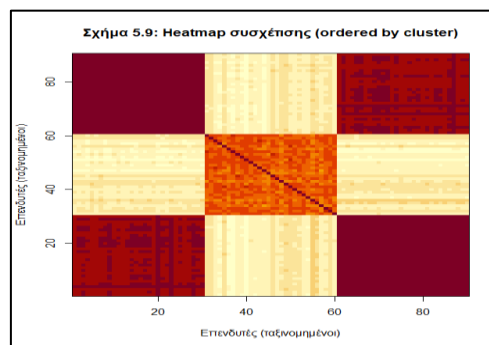
Απόσταση	Αλγόριθμος	Νίκες (Wins)
Euclidean	Hclust - Ward.D2	14.211
Euclidean	Hclust - Complete	14.199
Euclidean	Hclust - Average	14.121
Euclidean	K-means	1.092
Euclidean	PAM	53
Manhattan	Hclust - Average	27
Manhattan	Hclust - Complete	19
Manhattan	Hclust -Ward.D2	18
Manhattan	K-means	0
Manhattan	PAM	0



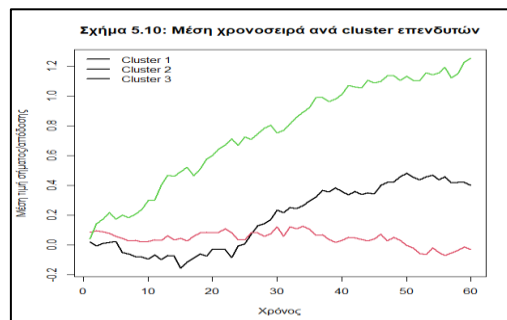
Διάγραμμα 5.6.1



Διάγραμμα 5.7.1



Διάγραμμα 5.7.2



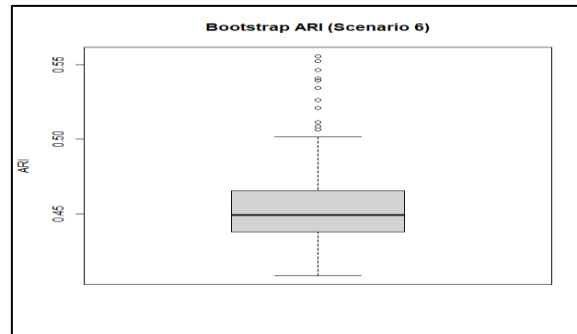
Διάγραμμα 5.7.3

Πίνακας 5.8.1

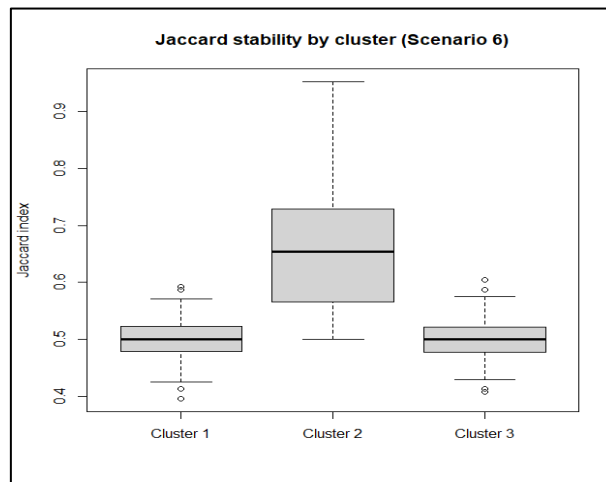
B	Mean ARI	SD	Min	Q25	Median	Q75	Max
200	0.457	0.027	0.409	0.438	0.449	0.465	0.556

Πίνακας 5.8.2

Cluster	Mean Jaccard	SD	Median	95% εμπειρικό διάστημα
Cluster 1	0.502	0.033	0.500	0.447–0.565
Cluster 2	0.662	0.117	0.654	0.500–0.917
Cluster 3	0.498	0.033	0.500	0.435–0.553



Διάγραμμα 5.8.1



Διάγραμμα 5.8.2

8. Βιβλιογραφία

Ελληνική Βιβλιογραφία

Χαλκιάς, Ι., & Σκαρβέλης, Ε. (2011). Ανάλυση Συστάδων (Cluster Analysis): Θεωρία και Εφαρμογές με το SPSS. Εκδόσεις Τζιόλα, Θεσσαλονίκη.

Κουτσογιάννης, Δ. (2009). Στατιστική Ανάλυση και Πολυμεταβλητές Τεχνικές. Εκδόσεις Τζιόλα, Θεσσαλονίκη.

Μάργαρης, Π. (2016). Στατιστική και Ανάλυση Δεδομένων με το R. Εκδόσεις Κλειδάριθμος, Αθήνα.

Μαυρίδου, Σ. (2008). Στατιστικές μέθοδοι στην οικολογία. Εκδόσεις Ζήτη.

Δημητρίου, Ι. (2010). Βιοστατιστική με εφαρμογές. Εκδόσεις Συμεών.

Ξενόγλωσση Βιβλιογραφία

Aggarwal, C. C. (2015). *Data mining: The textbook*. Springer. <https://doi.org/10.1007/978-3-319-14142-8>

Ali Abd Al-Hameed, K. (2022). Spearman's correlation coefficient in statistical analysis. *International Journal of Nonlinear Analysis and Applications*, 13(1), 3249–3255. <https://doi.org/10.22075/IJNAA.2022.6079>

Arif, M., & Alim, M. A. (2019). Similarity measures and dimensionality reduction techniques in pattern recognition: A review. *International Journal of Computer Applications*, 178(41), 1–7. <https://doi.org/10.5120/ijca2019918864>

Bădică, A., Bădică, C., Logofătu, D., & Neremzoïu, I.-D. (2025). Generalizations and properties of normalized similarity measures for Boolean models. *Mathematics*, 13(3), 384. <https://doi.org/10.3390/math13030384>

Batista, G. E. A. P. A., & Monard, M. C. (2003). An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence*, 17(5–6), 519–533.

Burnaby, T. P. (1970). On a method for character weighting a similarity coefficient, employing the concept of information. *Mathematical Geology*, 2(1), 25–38.

Choi, S. S., Cha, S. H., & Tappert, C. C. (2010). A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics*, 8(1), 43–48.

Conover, W. J., & Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician*, 35(3), 124–129. <https://doi.org/10.1080/00031305.1981.10479327>

de Carvalho, F. A. T., & Freitas, C. O. A. (2004). A fuzzy extension of the Jaccard index and its application to clustering. *Pattern Recognition Letters*, 25(12), 1503–1513.

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P., & Andersen, G. L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology*, 72(7), 5069–5072. <https://doi.org/10.1128/AEM.03006-05>

de Winter, J. C. F., & Dodou, D. (2022). Comparing the Pearson and Spearman correlation coefficients across distributions and sample sizes: A tutorial. *Frontiers in Psychology*, 13, 894484. <https://doi.org/10.3389/fpsyg.2022.894484>

Emran, S. M., & Ye, N. (2001). Robustness of Canberra metric in computer intrusion detection. In *Proceedings of the IEEE Workshop on Information Assurance and Security*.

Everitt, B. S. (1992). *The analysis of contingency tables* (2nd ed.). Chapman & Hall.

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis* (5th ed.). Wiley.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2), 179–188.

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53(3/4), 325–338.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 27(4), 857–871.

Gower, J. C., & Warrens, M. J. (2010). Measures of similarity, dissimilarity, and distance. In *Encyclopedia of statistics in behavioral science*. Wiley. <https://doi.org/10.1002/9781118445112.stat06551>

Han, J., Pei, J., & Tong, H. (2022). *Data mining: Concepts and techniques* (4th ed.). Morgan Kaufmann.

Hubálek, Z. (1982). Coefficients of association and similarity, based on binary (presence-absence) data: An evaluation. *Biological Reviews*, 57(4), 669–689. <https://doi.org/10.1111/j.1469-185X.1982.tb00376.x>

Jaccard, P. (1901). Étude comparative de la distribution florale. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37, 547–579.

Jaccard, P. (1912). The distribution of the flora in the alpine zone. *New Phytologist*, 11(2), 37–50.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys*, 31(3), 264–323.

Janani, K., Mohanrasu, S. S., Kashkynbayev, A., & Rakkiyappan, R. (2024). Minkowski distance measure in fuzzy PROMETHEE for ensemble feature selection. *Mathematics and Computers in Simulation*, 222, 264–295. <https://doi.org/10.1016/j.matcom.2023.08.027>

Jarrous, A., & Pinkas, B. (2009). Secure Hamming distance based computation and its applications. In M. Abdalla, D. Pointcheval, P.-A. Fouque, & D. Vergnaud (Eds.), *Applied cryptography and network security* (Lecture Notes in Computer Science, Vol. 5536, pp. 107–124). Springer. https://doi.org/10.1007/978-3-642-01957-9_7

Jurman, G., Riccadonna, S., Visintainer, R., & Furlanello, C. (2009). Canberra distance on ranked lists. In *Proceedings of Advances in Ranking NIPS 09 Workshop* (pp. 22–27).

Kacprzyk, J., Filev, D., & Beliakov, G. (Eds.). (2017). *Granular, soft and fuzzy approaches for intelligent systems* (Studies in Fuzziness and Soft Computing, Vol. 344). Springer. <https://doi.org/10.1007/978-3-319-40314-4>

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. Wiley.

Kendrick, W. B., & Proctor, J. R. (1964). Computer taxonomy in the fungi imperfecti. *Canadian Journal of Botany*, 42(1), 65–88.

Kulczyński, S. (1927). Die Pflanzenassoziationen der Pieninen. *Bulletin International de l'Académie Polonaise des Sciences et des Lettres*, 57–203.

Legendre, P., & Legendre, L. (1998). *Numerical ecology* (2nd English ed.). Elsevier.

Legendre, P., & Legendre, L. (2012). *Numerical ecology* (3rd English ed.). Elsevier.

Levandowsky, M., & Winter, D. (1971). Distance between sets. *Nature*, 234(5323), 34–35.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Wiley.

Liu, H., & Motoda, H. (2007). *Computational methods of feature selection*. Chapman & Hall/CRC.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India*, 2(1), 49–55.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

Mielke, P. W., Jr., & Berry, K. J. (2001). Euclidean distance and related permutation methods in atmospheric science. In T. J. Brown & P. W. Mielke, Jr. (Eds.), *Statistical methods in the atmospheric sciences* (pp. 7–27).

Pandey, R., & Gupta, S. (2020). A comprehensive survey on similarity measures in machine learning and data mining. *International Journal of Advanced Computer Science and Applications*, 11(3), 241–250. <https://doi.org/10.14569/IJACSA.2020.0110330>

Pearson, K. (1896). Mathematical contributions to the theory of evolution. III. Regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London A*, 187, 253–318.

Rahman Hera, M., & Koslicki, D. (2025). Estimating similarity and distance using FracMinHash. *Algorithms for Molecular Biology*, 20, Article 8. <https://doi.org/10.1186/s13015-025-00276-8>

Rebeka, A., Lončarić, Z., Petrović, S., & Marić, S. (2015). Pearson's or Spearman's correlation coefficient—which one to use? *Poljoprivreda*, 21(2), 54–58. <https://doi.org/10.18047/poljo.21.2.8>

Rodgers, J. L., & Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1), 59–66.

Rogers, D. J., & Tanimoto, T. T. (1960). A computer program for classifying plants. *Science*, 132(3434), 1115–1118. <https://doi.org/10.1126/science.132.3434.1115>

Rokach, L., & Maimon, O. (2005). Clustering methods. In *Data mining and knowledge discovery handbook*. Springer.

Russell, P. F., & Rao, T. R. (1940). On habitat and association of species of Anopheline larvae in southeastern Madras. *Journal of the Malaria Institute of India*, 3, 153–178.

Salkind, N. J. (Ed.). (2007). Spearman's rho. In *Encyclopedia of measurement and statistics* (pp. 932–934). Sage.

Sandhu, G., Singh, A., Lamba, P. S., Virmani, D., & Chaudhary, G. (2023). Modified Euclidean-Canberra blend distance metric for kNN classifier. *Intelligent Decision Technologies*, 17(2), 527–541. <https://doi.org/10.3233/IDT-220233>

Sharma, R., & Kumar, V. (2021). Network intrusion detection using hybrid similarity-based methods: A review. *Journal of Network and Computer Applications*, 180, 102985. <https://doi.org/10.1016/j.jnca.2021.102985>

Singh, J., & Thomas, G. (2013). A comparative study on distance and similarity measures in clustering. *International Journal of Computer Science and Information Technologies*, 4(1), 251–256.

Sneath, P. H. A., & Sokal, R. R. (1973). *Numerical taxonomy: The principles and practice of numerical classification*. W. H. Freeman.

Sokal, R. R., & Michener, C. D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletin*, 38(22), 1409–1438.

- Sokal, R. R., & Sneath, P. H. A. (1963). *Principles of numerical taxonomy*. W. H. Freeman.
- Tan, P.-N., Steinbach, M., Karpatne, A., & Kumar, V. (2019). *Introduction to data mining* (2nd ed.). Pearson.
- Todeschini, R., & Consonni, V. (2009). *Molecular descriptors for chemoinformatics*. Wiley-VCH.
- Todeschini, R., Consonni, V., Xiang, H., Holliday, J., Buscema, M., & Willett, P. (2012). Similarity coefficients for binary chemoinformatics data: Overview and extended comparison using simulated and real data sets. *Journal of Chemical Information and Modeling*, 52(11), 2884–2901.
- Warrens, M. J. (2008). On association coefficients for 2×2 tables and properties that do not depend on the marginal distributions. *Psychometrika*, 73(4), 777–789. <https://doi.org/10.1007/s11336-008-9070-3>
- Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2017). *Data mining: Practical machine learning tools and techniques* (4th ed.). Morgan Kaufmann.
- Xiao, C., Ye, J., Esteves, R. M., & Rong, C. (2016). Using Spearman's correlation coefficients for exploratory data analysis on big dataset. *Concurrency and Computation: Practice and Experience*, 28(14), 3866–3878.
- Zezula, P., Amato, G., Dohnal, V., & Batko, M. (2006). *Similarity search: The metric space approach*. Springer.
- Zhou, Y., Cheng, G., Jiang, S., & Dai, M. (2020). Building an efficient intrusion detection system based on feature selection and similarity measures. *Computers & Security*, 95, 101867. <https://doi.org/10.1016/j.cose.2020.101867>
- Zhu, Y., Xiang, Y., Sen, S., Dagois, E., Nehorai, A., & Akçakaya, M. (2020). Clutter identification based on sparse recovery and L1-type probabilistic distance measures. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (4772–4776). IEEE.

