Πρόγραμμα Μεταπτυχιακών Σπουδών
στην Αναλυτική των Επιχειρήσεων και Επιστήμη των Δεδομένων

Τμήμα Οργάνωσης και Διοίκησης Επιχειρήσεων

<u>Διπλωματική εργασία</u>

Σύγκριση μεθόδων ανάλυσης κατά συστάδες σε μεικτού τύπου δεδομένα ιστού του

Χρήστου-Σπυρίδωνα Μοσχοφίδη του Λεοντίου

Υποβλήθηκε ως απαιτούμενο για την απόκτηση του μεταπτυχιακού διπλώματος στην Αναλυτική των Επιχειρήσεων και Επιστήμη των Δεδομένων

Επιβλέπων Καθηγητής: Άγγελος Μάρκος

Νοέμβριος 2022

Αφιερώσεις

Η εργασία αυτή είναι αφιερωμένη στην οικογένειά μου, που έχει στηρίξει χωρίς ενδοιασμούς όλα τα εγχειρήματά μου και με βοήθησαν να έχω τις ευκαιρίες που χρειαζόμουν ώστε να είμαι ικανός να εξελιχθώ προσωπικά και επαγγελματικά.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή της παρούσας εργασίας για την απλόχερη και καθοριστικής σημασίας βοήθεια και καθοδήγηση που μου παρείχε καθ' όλη την διάρκειά και επιτυχή ολοκλήρωσή της.

Περίληψη

Η Ανάλυση Συστάδων (Cluster Analysis) είναι μια οικογένεια στατιστικών μεθόδων που αποσκοπούν στην ομαδοποίηση αντικειμένων ή παρατηρήσεων σε διακριτές ομάδες ή συστάδες (clusters), με βάση το πόσο «όμοια» είναι ως προς τις τιμές που λαμβάνουν σε ορισμένες μεταβλητές. Έτσι κάθε ομάδα αποτελείται από παρατηρήσεις που μοιάζουν όσο το δυνατόν περισσότερο μεταξύ τους και όσο το δυνατόν λιγότερο με τις παρατηρήσεις των άλλων ομάδων. Επιπλέον, πρόκειται για μια οικογένεια μεθόδων μη επιβλεπόμενης μάθησης (unsupervised learning), που σημαίνει ότι δεν γνωρίζουμε πόσες ομάδες υπάρχουν στα δεδομένα πριν εφαρμοστεί μια μέθοδος. Σε αντίθεση με πολλές άλλες στατιστικές μεθόδους, η Α.Σ. χρησιμοποιείται συνήθως όταν δεν υπάρχει καμία υπόθεση σχετικά με τις πιθανές σχέσεις μεταξύ των μεταβλητών. Παράλληλα, η Α.Σ μπορεί να αποτελέσει ένα ισχυρό εργαλείο ανάλυσης δεδομένων για κάθε οργανισμό ή επιχείρηση που χρειάζεται να εντοπίσει π.χ. διακριτές ομάδες πελατών, ομάδες συναλλαγών πωλήσεων ή άλλα είδη συμπεριφορών. Για παράδειγμα, οι ασφαλιστικές εταιρίες χρησιμοποιούν μεθόδους της Α.Σ. για να ανιχνεύσουν απάτες και οι τράπεζες τις χρησιμοποιούν για πιστοληπτική βαθμολόγηση των πελατών. Σκοπός της παρούσας εργασίας είναι η περιγραφή και η σύγκριση μεθόδων συσταδοποίησης για μεικτού τύπου δεδομένα. Τα μεικτού τύπου δεδομένα περιγράφονται τόσο από ποσοτικές όσο και από ποιοτικές μεταβλητές και η ανάλυσή τους με μεθόδους της Α.Σ. παρουσιάζει αρκετές προκλήσεις. Στο 2^{ο} κεφάλαιο γίνεται μια ανασκόπηση μεθόδων Α.Σ. για μεικτού τύπου δεδομένα και παρουσιάζεται αναλυτικά το μαθηματικό υπόβαθρό τους. Στο 3^{ο} κεφάλαιο, οι μέθοδοι εφαρμόζονται σε ένα πραγματικό σύνολο δεδομένων ιστού και γίνεται η σύγκριση των αποτελεσμάτων τους. Τα αποτελέσματα ανέδειξαν τις ομοιότητες αλλά και τις διαφορές μεταξύ των μεθόδων, χωρίς ωστόσο κάποια μέθοδος να ξεχωρίζει σημαντικά έναντι των υπολοίπων.

Περιεχόμενα

1. Εισαγωγή

Μία από τις έμφυτες ικανότητες του ανθρώπου είναι αυτή της ταξινόμησης αντικειμένων ή παρατηρήσεων, σε διαφορετικές ομάδες ή κατηγορίες, ανάλογα με τα χαρακτηριστικά που τα προσδιορίζουν. Η διαδικασία αυτή συναντάται από το πλέον βασικό, την ταξινόμηση της τροφής σε βλαβερή ή όχι (π.χ. με «μεταβλητές» όπως η μυρωδιά, γεύση, εμπειρία). Σε διαφορετικά επιστημονικά πεδία, αρκετές μέθοδοι χρησιμοποιούνται για την επίτευξη του σκοπού αυτού. Οι μέθοδοι αυτές μπορούν να διακριθούν σε δύο μεγάλες κατηγορίες, στις μεθόδους επιβλεπόμενης και τις μεθόδους μη επιβλεπόμενης μάθησης (James, Witten, Hastie & Tibshirani, 2013). Στην πρώτη περίπτωση οι ομάδες στις οποίες μπορεί να αντιστοιχιστεί η κάθε παρατήρηση ή το κάθε αντικείμενο είναι γνωστές εκ των προτέρων. Ένα υποσύνολο του αρχικού συνόλου δεδομένων χρησιμοποιείται για την «εκπαίδευση» ενός αλγορίθμου, ώστε να είναι σε θέση να ταξινομήσει αντικείμενα εκτός του συνόλου εκπαίδευσης με την μεγαλύτερη δυνατή ακρίβεια. Στην περίπτωση της μη επιβλεπόμενης μάθησης δεν γνωρίζουμε από πριν τις ομάδες και αντιθέτως, το ζητούμενο είναι να εντοπιστούν οι ομάδες και να ενταχθούν τα αντικείμενα στην κατάλληλη ομάδα. Τέτοιες τεχνικές αποσκοπούν περισσότερο σε μια προσπάθεια αποκάλυψης μοτίβων στα δεδομένα, τα οποία με διαφορετικό τρόπο θα ήταν πολύ δύσκολο να φανερωθούν, αν ληφθεί υπόψη ότι πολλές φορές τα δεδομένα είναι μεγάλου όγκου.

Η Ανάλυση Συστάδων ή Ανάλυση Ομαδοποίησης (Cluster Analysis) αποτελεί μία από τις βασικότερες οικογένειες μεθόδων μη επιβλεπόμενης μάθησης. Στόχος τους είναι η τοποθέτηση των αντικειμένων σε διακριτές ομάδες, που αποκαλούνται «συστάδες», με βάση ένα συγκεκριμένο κριτήριο ομοιότητας μεταξύ των αντικειμένων. Η Ανάλυση Συστάδων χρησιμοποιείται ευρέως στις μέρες μας, σε πεδία όπως η έρευνα αγοράς, η κατηγοριοποίηση ανεπιθύμητων μηνυμάτων ηλεκτρονικού ταχυδρομείου, η ανάλυση εικόνας η ανίχνευση ακραίων παρατηρήσεων και πολλά άλλα (βλ. James et al., 2013). Παρά το γεγονός ότι οι πρώτοι αλγόριθμοι ομαδοποίησης έχουν ήδη αναπτυχθεί τη δεκαετία του 1950, η μεγάλη πλειοψηφία τους μπορούσε να διαχειριστεί μόνο ποσοτικά δεδομένα. Πολλές από αυτές τις ιδέες αυτές στη συνέχεια επεκτάθηκαν σε σύνολα δεδομένων που αποτελούνταν μόνο από κατηγορικές μεταβλητές. Ωστόσο, σε πραγματικές εφαρμογές συναντάμε συχνά δεδομένα μεικτού τύπου, δηλαδή αντικείμενα που περιγράφονται τόσο από ποσοτικές όσο και από κατηγορικές μεταβλητές. Ο Πίνακας

1 είναι ένα παράδειγμα πίνακα δεδομένων με *n* αντικείμενα όπου οι δύο πρώτες μεταβλητές, X_1, X_2 είναι ποσοτικές συνεχείς και οι τρεις τελευταίες X_3, X_4 και X_5 είναι κατηγορικές (ποιοτικές).

Πίνακας 1. Παράδειγμα πίνακα δεδομένων μεικτού τύπου

Αντικείμενα	X_1	X_2	X_3	X_4	X_5
1	30.4	10	1	1	3
2	20.3	9.5	4	2	2
3	0	4	2	1	1
…	…	…	…	…	…
n	10.8	1.2	2	2	2

Στην εργασία αυτή, παρουσιάζουμε μερικούς από τους πιο γνωστούς αλγορίθμους ομαδοποίησης για μεικτά σύνολα δεδομένων (Κεφάλαιο 2) και εφαρμόζουμε τις μεθόδους αυτές σε πραγματικά δεδομένα κοινωνικών δικτύων (Κεφάλαιο 3).

2. Αλγόριθμοι συσταδοποίησης για μεικτού τύπου δεδομένα

2.1 Ο αλγόριθμος k-prototypes

Αρχικά, θα παρουσιαστεί ο αλγόριθμος k-prototypes. Ο αλγόριθμος προτάθηκε για την ομαδοποίηση δεδομένων μεικτού τύπου από τον Huang (1998) και συνδυάζει τις ιδέες του αλγορίθμου k-means και του αλγορίθμου k-modes. Ο αλγόριθμος διαμερίζει το σύνολο των παρατηρήσεων σε k ($k \in N^+$) διαφορετικά υποσύνολα – συστάδες μέσω της ελαχιστοποίησης μιας συνάρτησης κόστους. Ο αλγόριθμος βασίζεται στην έννοια του κεντροειδούς μιας συστάδας, το οποίο ορίζεται από τους μέσους όρους (means) των ποσοτικών μεταβλητών και τις επικρατούσες τιμές (modes) των κατηγορικών μεταβλητών, δημιουργώντας ένα νέο υβριδικό κεντροειδές ή πρωτότυπο (prototype). Η απόσταση κάθε παρατήρησης από το πρωτότυπο υπολογίζεται με βάση έναν κατάλληλο συντελεστή ανομοιότητας. Μια επιπλέον παράμετρος, γ, εισάγεται για τον έλεγχο της επιρροής των ποσοτικών και κυρίως των κατηγορικών μεταβλητών στη διαδικασία της συσταδοποίησης. Ο αλγόριθμος βασίζεται σε τρία βήματα: 1) Αρχική επιλογή πρωτοτύπων, 2) Αρχική τοποθέτηση σε συστάδες, 3) Επανατοποθέτηση.

Στη συνέχεια παρουσιάζουμε την απαραίτητη σημειογραφία για την καλύτερη κατανόηση της μαθηματικής περιγραφής του αλγορίθμου:

n	Αριθμός αντικειμένων - παρατηρήσεων
p	Αριθμός μεταβλητών
$\boldsymbol{X}$	$(n \times p)$ αρχικός πίνακας δεδομένων
Q_l	$\{q_{l1}, \ldots, q_{lp}\}$ – Κεντροειδές ή πρωτότυπο της συστάδας l, όπου $1 \leq l \leq k$
k	Αριθμός συστάδων
$d(-,-)$	Μέτρο απόστασης ή Μετρική απόστασης

2.1.1 Περιγραφή αλγορίθμου

Αρχική επιλογή πρωτοτύπων

Ας υποθέσουμε ότι γνωρίζουμε εκ των προτέρων τον αριθμό των συστάδων. Σύμφωνα με τον Huang (1998), η αρχική επιλογή των πρωτοτύπων μπορεί να γίνει εντοπίζοντας τα k πιο διακριτά αντικείμενα του συνόλου δεδομένων. Για παράδειγμα, οι πιο συχνά εμφανιζόμενες κατηγορίες μπορούν να αποτελέσουν τα k αρχικά πρωτότυπα.

Αρχική τοποθέτηση

Μετά την επιλογή των αρχικών πρωτοτύπων, υπολογίζεται η απόσταση κάθε αντικειμένου από κάθε πρωτότυπο και τοποθετείται στη συστάδα με της οποίας το πρωτότυπο έχει την μικρότερη απόσταση. Μετά την ανάθεση, υπολογίζονται εκ νέου τα πρωτότυπα της κάθε συστάδας.

Επανατοποθέτηση

Αφού έχουν κατανεμηθεί όλα τα αντικείμενα σε συστάδες, υπολογίζεται ξανά η απόστασή τους από τα πρωτότυπα που έχουν υπολογιστεί στο προηγούμενο βήμα. Εάν ένα αντικείμενο βρίσκεται πιο κοντά στο πρωτότυπο άλλης συστάδας από αυτή στην οποία είχε τοποθετηθεί, τότε τοποθετείται στη νέα συστάδα και επανυπολογίζονται τα πρωτότυπα των συστάδων. Το στάδιο αυτό επαναλαμβάνεται έως ότου κανένα αντικείμενο να μην αλλάζει συστάδα, αφού ελεγχθεί όλο το σύνολο των αντικειμένων.

2.1.2 Μαθηματική περιγραφή του αλγορίθμου

Ορίζουμε την συνάρτηση κόστους την οποία ο αλγόριθμος αποσκοπεί να ελαχιστοποιήσει. Αλγεβρικά αποτελεί τη συνάρτηση κόστους του πίνακα διασποράς της κάθε συστάδας:

$$E = \sum_{l=1}^{k} u_{ik} \sum_{i=1}^{n} y_{il}\, d(X_i, Q_l) \qquad (1)$$

Συμπληρωματικά, να σημειωθεί πως το X_i αποτελεί την i-στη γραμμή του $\boldsymbol{X}$ και συνεπώς δίνεται από το διάνυσμα με p στοιχεία $(x_{i1} \ldots, x_{ip})$. Το εσωτερικό μέρος της (1) δίνεται από:

$$E_l = \sum_{i=1}^{n} y_{il}\, d(X_i, Q_l)\,, 1 \leq l \leq k\,, \tag{2}$$

το οποίο αντιστοιχεί στο συνολικό κόστος ανάθεσης των αντικειμένων στη συστάδα *l*. Στόχος μας είναι να ελαχιστοποιήσουμε την (2), όπου θα ελαχιστοποιήσει και κατ' επέκταση την (1). Για να το καταφέρουμε όμως, πρέπει να είμαστε σε θέση να ορίσουμε ένα μέτρο ανομοιότητας τόσο για τις ποσοτικές όσο και για τις κατηγορικές μεταβλητές. Μπορούμε να χρησιμοποιήσουμε ένα συνδυασμό του τετραγώνου της Ευκλείδειας απόστασης για τις ποσοτικές μεταβλητές και μιας απόστασης για δίτιμες μεταβλητές, στην περίπτωση των κατηγορικών μεταβλητών. Πιο συγκεκριμένα η απόσταση $d(X_i, Q_l)$ δίνεται από:

$$d(X_i\,, Q_l) = \sum_{j=1}^{p_r} (x_{ij} - q_{lj})^2 + \gamma_l \sum_{j\,=\,p_r\,+\,1}^{p} \delta(x_{ij}\,, q_{lj}) \tag{3}$$

Στην περίπτωση αυτή έχουμε υποθέσει ότι οι πρώτες p_r στήλες του X_i αποτελούν ποσοτικές μεταβλητές ενώ οι υπόλοιπες $p - p_r$ είναι κατηγορικές. Το πρώτο άθροισμα αποτελεί το τετράγωνο της Ευκλείδειας απόσταση, ενώ το $\delta(x_{ij}\,, q_{lj})$ ισούται με το 0 εάν και εφόσον $x_{ij} = q_{lj}$, αλλιώς ισούται με το 1. Τέλος το γ_l ορίζεται ως ένας συντελεστής βαρύτητας των κατηγορικών μεταβλητών στη συστάδα *l*.

Η τιμή του συντελεστή γ_l έχει μεγάλη σημασία στον αλγόριθμο, συνεπώς πρέπει να είμαστε σίγουροι ότι έχουμε επιλέξει μια κατάλληλη τιμή. Συνήθως, ο συντελεστής σχετίζεται με τις υποκείμενες κατανομές συνεχών μεταβλητών, αλλά υπάρχουν αλγόριθμοι ομαδοποίησης που καθορίζουν αυτόματα την τιμή του, όπως ο αλγόριθμος Modha-Spangler που αναφέρεται στην Ενότητα 2.4. Αξιοσημείωτο είναι πως αν ορίσουμε γ_l = 0, που υποδηλώνει την παρουσία μόνο ποσοτικών μεταβλητών στα δεδομένα, τότε το αποτέλεσμα θα είναι ίδιο με αυτό του αλγόριθμου K-means.

Επιστρέφοντας στην συνάρτηση (2), και αντικαθιστώντας την έκφραση $d(X_i, Q_l)$ με την (3), παρατηρούμε πως η συνάρτηση κόστους πλέον δίνεται από:

$$E_l = \underbrace{\sum_{i=1}^{n} y_{il} \sum_{j=1}^{p_r} (x_{ij} - q_{lj})^2}_{\boldsymbol{E_l^r}} + \gamma_l \underbrace{\sum_{i=1}^{n} y_{il} \sum_{j=p_r+1}^{p} \delta(x_{ij}, q_{lj})}_{\boldsymbol{E_l^c}} \quad (4)$$

Ελαχιστοποιώντας το πρώτο τμήμα της (4), δηλαδή το $\boldsymbol{E_l^r}$, βρίσκουμε τις βέλτιστες τιμές των πρωτοτύπων που αντιστοιχούν στις πρώτες p_r μεταβλητές, δηλαδή τις ποσοτικές μεταβλητές. Αυτές δίνονται από τον τύπο $q_{lj} = \frac{1}{n_l}\sum_{i=1}^{n} y_{il}\ x_{ij}$, όπου $\sum_{i=1}^{n} y_{il}$ είναι ο συνολικός αριθμός των αντικειμένων στη συστάδα l , για $1 \leq j \leq p_r$.

Τέλος, για να ελαχιστοποιηθεί το $\boldsymbol{E_l^c}$, πρέπει να ορίσουμε ως C_j το σύνολο όλων των μοναδικών τιμών που παίρνει η κατηγορική μεταβλητή j. Υποθέτουμε ότι c_j είναι μία από αυτές τις μοναδικές τιμές στο σύνολο C_j και $\mathbb{P}(c_j \in C_j \backslash l)$ η πιθανότητα η κατηγορική μεταβλητή j να λάβει την τιμή c_j στη συστάδα l. Έτσι μπορούμε να επαναδιατυπώσουμε την $\boldsymbol{E_l^c}$ από την (4) ως εξής :

$$E_l{}^c = \gamma_l \sum_{i=1}^{n} y_{il} \sum_{j=p_r+1}^{p} \delta(x_{ij}, q_{lj}) = \gamma_l \sum_{j=p_r+1}^{p} n_l \left(1 - \mathbb{P}(q_{lj} \in C_j \backslash l)\right).$$

Συνεπώς, για να επιτευχθεί η ελαχιστοποίηση του $E_l{}^c$, πρέπει να μεγιστοποιηθεί η $\mathbb{P}(q_{lj} \in C_j \backslash l)$. Έτσι, πρέπει να ισχύει ότι $\mathbb{P}(q_{lj} \in C_j \backslash l) \geq \mathbb{P}(c_j \in C_j \backslash l)$ για κάθε $c_j \neq q_{lj}$, που σημαίνει πως θέλουμε να μεγιστοποιήσουμε την πιθανότητα η κατηγορική τιμή q_{lj} να περιλαμβάνεται στην κατηγορική μεταβλητή j στη συστάδα l . Αυτό πρακτικά σημαίνει πως η τιμή q_{lj} πρέπει να είναι η πιο «συχνή» τιμή του κατηγορικού γνωρίσματος j στη συστάδα l – με απλά λόγια, η επικρατούσα τιμή του γνωρίσματος j.

2.2 Ο συντελεστής ανομοιότητας του Gower και η Μέθοδος Partitioning Around Medoids

2.2.1 Ο συντελεστής ανομοιότητας του Gower

Ο δείκτης παρουσιάστηκε αρχικά από τον Gower (1971) και αποτελεί ένα μέτρο ανομοιότητας μεταξύ δύο αντικειμένων X_i και X_j που ορίζεται ως εξής:

$$d_G(X_i, X_j) = 1 - \frac{\sum_{k=1}^{p} w_k(X_i, X_j) s_k(X_i, X_j)}{w_k(X_i, X_j)}, 1 \leq i, j \leq n, i \neq j \quad (5)$$

Εδώ s_k είναι ο συντελεστής ομοιότητας μεταξύ της τιμής των δύο αντικειμένων x_{ij} $και$ x_{jk} της μεταβλητής k, ενώ w_k είναι το βάρος της μεταβλητής k. Η τιμή του w_k συνήθως ορίζεται ως 1, εκτός εάν κάποια από τις τιμές των x_{ij} $και$ x_{jk} απουσιάζει, κάτιπου σημαίνει ότι $w_k(X_i, X_j)$ = 0. Επομένως, η απόσταση του Gower μπορεί επίσης να διαχειριστεί ελλείπουσες τιμές (missing values).
Η τιμή του συντελεστή ομοιότητας s_k διαφέρει με βάση το είδος της μεταβλητής k. Πιο συγκεκριμένα:

- Αν η μεταβλητή είναι **ποσοτική**, αυτή ισούται με την απόλυτη διαφορά των τιμών των αντικειμένων x_{ij} $και$ x_{jk} (απόσταση Manhattan), κανονικοποιημένη (normalized) με το εύρος (R) των τιμών της μεταβλητής.

$$s_k(X_i, Y_j) = 1 - \frac{|x_{ik} - y_{jk}|}{R_k}$$

- Αν η μεταβλητή είναι **κατηγορική**, τότε:
  - $s_k(X_i, Y_j) = 0, αν \neq y_{jk}$
  - $s_k(X_i, Y_j) = 1,\ αν\ x_{ik} = y_{jk}$

  που σημαίνει ότι η τιμή που θα πάρει η $s_k(X_i, Y_j)$ θα είναι 1 αν οι παρατηρήσεις ταυτίζονται και 0 αν όχι.

- Αν η μεταβλητή είναι **συμμετρική δυαδική,** τότε :

  - $s_k(X_i, Y_j) = 0\ αν\ x_{ik} \neq y_{jk}$
  - $s_k(X_i, Y_j) = 1\ αν\ x_{ik} = y_{jk}$ = 1 ή $x_{ik} = y_{jk} = 0$

- Αν είναι μη **συμμετρική δυαδική**, τότε :

  Ισχύουν οι περιπτώσεις για τη συμμετρική δυαδική με τη διαφορά ότι,

  - $w_k(X_i, X_j) = 0 \; \alpha\nu \; x_i = y_i = 0$
  - $w_k(X_i, X_j) = 1$

Συνεπώς, αφού υπολογιστούν οι αποστάσεις Gower, μπορούμε να εφαρμόσουμε στον πίνακα αποστάσεων έναν αλγόριθμο συσταδοποίησης. Μία καλή και ίσως περισσότερο «ανθεκτική» επιλογή από τον ευρέως διαδεδομένο αλγόριθμο k-means, αποτελεί ο αλγόριθμος PAM.

2.2.2 Μέθοδος Partitioning Around Medoids

Η Partitioning Around Medoids (PAM) είναι μία μέθοδος ομαδοποίησης που ενδείκνυται για μεικτά δεδομένα. Προτάθηκε από τους Kaufman και Rousseeuw (2009) και χρησιμοποιεί τον δείκτη ανομοιότητας του Gower. Η PAM είναι μια μέθοδος ομαδοποίησης για μεικτά δεδομένα που παρουσιάζει πολλές ομοιότητες με τον αλγόριθμο k-means. Η ιδέα της PAM είναι να ορίσει ομάδες με βάση τις ομοιότητες που παρουσιάζουν τα αντικείμενα του συνόλου δεδομένων. Οι ομοιότητες μπορούν να υπολογιστούν είτε χρησιμοποιώντας την d_G από την (5), το οποίο είναι πολύ χρήσιμο όταν εργαζόμαστε με δεδομένα μεικτού τύπου ή χρησιμοποιώντας οποιαδήποτε άλλη μετρική ανομοιότητας ή απόστασης. Ο στόχος της PAM είναι να εντοπίσει k "αντιπροσωπευτικά" αντικείμενα - αυτά είναι αντικείμενα που ελαχιστοποιούν τη μέση ανομοιότητα εντός της κάθε συστάδας και ονομάζονται medoids. Τα medoids, χωρίς απαραίτητα να αντιστοιχούν στο μέσο της συστάδας, αποτελούν τα αντικείμενα που βρίσκονται πιο κοντά στο κέντρο της. Μόλις βρεθούν τα k medoids, κάθε αντικείμενο εκχωρείται στη συστάδα με της οποίας το medoid βρίσκεται πιο κοντά.

<u>Περιγραφή της μεθόδου:</u>

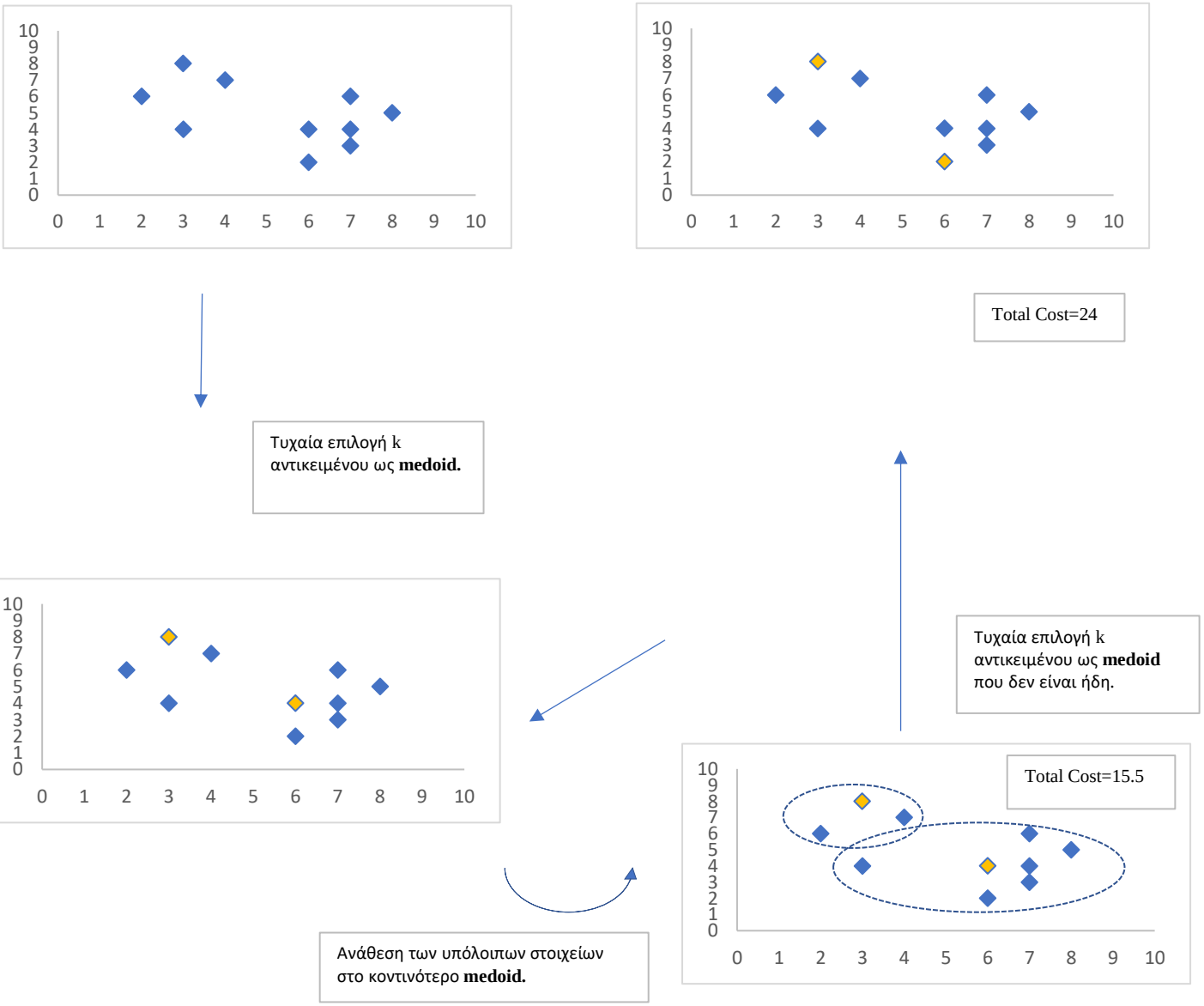
Αρχικά επιλέγονται τυχαία τα αντικείμενα που θα αποτελέσουν τα πρώτα κεντροειδή και εκτελούνται οι δύο ακόλουθες φάσεις:

1. ***Φάση οικοδόμησης:***
    - Υπολογισμός πίνακα αποστάσεων κάθε στοιχείου με όλα τα medoids.
    - Ανάθεση κάθε στοιχείου στο πλησιέστερο medoid (με κριτήριο τη μετρική της Ευκλείδειας Απόστασης).
2. ***Φάση ανταλλαγής:***

❖ Για κάθε συστάδα, ελέγχεται αν κάποιο αντικείμενο ως medoid, ελαχιστοποιεί την απόσταση αντικειμένου-medoid και αν ναι, ορίζεται ως το νέο medoid.

Σε περίπτωση αλλαγής του medoid, επαναλαμβάνονται οι φάσεις της οικοδόμησης και ανταλλαγής, διαφορετικά ο αλγόριθμος τερματίζει.

Για καλύτερη κατανόηση της μεθόδου, παρατίθεται ένα οπτικό παράδειγμα βασισμένο σε ένα τεχνητό σύνολο δεδομένων με δύο διακριτές ομάδες αντικειμένων.



Στη συγκεκριμένη περίπτωση βλέπουμε πως Total Cost_{new} - Total $\text{Cost}_{\text{old}} < 0$. Συνεπώς, τα medoids δεν θα αλλάξουν. Η παραπάνω διαδικασία θα συνεχιστεί έως ότου να μην μπορεί να γίνει κάποια ανταλλαγή medoids.

2.3 Η μέθοδος Mixed K-means

Ο αλγόριθμος K-means for mixed data ή Mixed K-Means παρουσιάστηκε αρχικά από τους Ahmad και Dey (2007). Είναι παρόμοιος με τον αλγόριθμο k-prototypes και προτάθηκε με στόχο να ξεπεραστούν ορισμένες από τις αδυναμίες του. Πιο συγκεκριμένα, $d(X_i, Q_l)$ η απόσταση που χρησιμοποιείται στη συνάρτηση κόστους του k-prototypes, η οποία δίνεται στην (3). Θα υποθέσουμε ότι το Q_l δεν είναι το $l^{\text{στο}}$ πρωτότυπο αλλά το κέντρο της συστάδας l (αν και εννοιολογικά το πρωτότυπο για μία συστάδα και το κέντρο της είναι ισοδύναμα). Όπως αναφέρθηκε στην ενότητα της παρουσίασης του k-prototypes, για κάθε κατηγορική μεταβλητή j, το $j^{\text{στο}}$ στοιχείο του $l^{\text{στου}}$ πρωτοτύπου Q_l θα είναι ίσο με την επικρατούσα τιμή της j. Ωστόσο, με την επικρατούσα τιμή ως κέντρο θέσης, χάνουμε πληροφορία, μιας και είναι πιθανό να υπάρχουν άλλες τιμές που μπορεί να λαμβάνει συχνά η j, αλλά δεν τις λαμβάνουμε υπόψιν.

Ένα άλλο μειονέκτημα της συνάρτησης απόστασης του k-prototypes είναι πως το βάρος όλων των ποσοτικών μεταβλητών θεωρείται ίσο με 1, ενώ των κατηγορικών δίνεται από μία παράμετρο γ_l που ορίζεται από τον χρήστη. Αυτό σημαίνει ότι σε ένα πραγματικό σύνολο δεδομένων, όλες οι ποσοτικές μεταβλητές θα έχουν το ίδιο βάρος, κάτι που δεν είναι απαραίτητα σωστό, ενώ το βάρος για τις κατηγορικές μεταβλητές πρέπει να ορίζεται με λογικό τρόπο, διαφορετικά μπορεί να οδηγήσει σε ανακριβή ομαδοποίηση. Τέλος, ο διχοτομικός δείκτης $\delta(x_{ij}, q_{lj})$ που χρησιμοποιείται θα μπορούσε στην πραγματικότητα να είναι ακατάλληλος για χρήση ως «απόσταση» μεταξύ δύο διακριτών τιμών μιας κατηγορικής μεταβλητής, καθώς δεν χρειάζεται πάντα να είναι 1, αλλά θα πρέπει να εξαρτάται από τη συνύπαρξη (συνάφεια) στα δεδομένα μεταξύ των τιμών των κατηγορικών μεταβλητών του συνόλου δεδομένων.

Η απόσταση, λοιπόν, που παρουσιάζεται ως η «βελτιωμένη» εκδοχή του (3), η οποία λαμβάνει υπόψιν τα προαναφερθέντα μειονεκτήματα δίνεται ως εξής:

$$d(X_i, Q_l) = \sum_{j=1}^{p_r} \left(w_j\left(x_{ij} - q_{lj}\right)\right)^2 + \sum_{j=p_r+1}^{p} \Omega\left(x_{ij}, q_{lj}\right)^2 \quad (6)$$

όπου w_j αποτελεί την σημαντικότητα της j στης ποσοτικής μεταβλητής και η τιμή της εκτιμάται από τα δεδομένα. Το $\Omega\left(x_{ij}, q_{lj}\right)^2$ αντικαθιστά τον διχοτομικό δείκτη $\delta\left(x_{ij}, q_{lj}\right)$ και υπολογίζεται ως συνάρτηση των κατανομών των x_{ij} και q_{lj} και της συνάφειάς τους με τις υπόλοιπες κατηγορικές μεταβλητές.

Πιο συγκεκριμένα, έστω δύο κατηγορικές μεταβλητές i και j όπου $i \neq j$. Ας υποθέσουμε πως η i λαμβάνει τις τιμές α και β και θέλουμε να υπολογίσουμε την απόσταση μεταξύ τους και σ ένα υποσύνολο των τιμών που μπορεί να πάρει η j.
Ας υποθέσουμε επίσης πως $\mathbb{P}_i(\sigma \backslash \alpha)$ αποτελεί την δεσμευμένη πιθανότητα, ένα αντικείμενο X_i για το οποίο ισχύει $x_{ij} = \alpha$, να λάβει μία τιμή της j που ανήκει στο σ.
Αντίστοιχα, η $\mathbb{P}_i(\sigma \backslash \beta)$ είναι η δεσμευμένη πιθανότητα, ένα αντικείμενο X_i για το οποίο ισχύει $x_{ij} = \beta$, να λάβει μία τιμή της j που ανήκει στο σ. Μπορούμε να ορίσουμε συνεπώς την απόσταση μεταξύ του ζεύγους τιμών α, β της i σε σχέση με τη j και ένα υποσύνολο σ ως εξής:

$$\delta_\sigma(\alpha, \beta) = \mathbb{P}_i(\sigma \backslash \alpha) + \mathbb{P}_i(\sigma \backslash \beta) \quad (7)$$

Η (7) αποτελεί την απόσταση μεταξύ δύο κατηγορικών μεταβλητών και ενός υποσυνόλου σ ενός γνωρίσματος j. Ωστόσο, ενδιαφερόμαστε κυρίως για την απόστασή τους με την ολότητα της μεταβλητής j. Αυτό μπορεί να αποδοθεί ως εξής:

$$\delta^{ij}(\alpha, \beta) = \mathbb{P}_i(\sigma * \backslash \alpha) + \mathbb{P}_i(\sigma * \backslash \beta) - 1 \quad (8)$$

όπου σ* είναι το υποσύνολο των τιμών του j που μεγιστοποιεί την (7). Ο λόγος που αφαιρούμε την μονάδα στην (8) είναι για να είμαστε σίγουροι πως $0 \leq \delta^{ij}(\alpha, \beta) \leq 1$. Τέλος, υπάρχει περίπτωση οι i και j να μην είναι οι μοναδικές μεταβλητές του συνόλου δεδομένων, οπότε πρέπει να λάβουμε υπόψιν και όλες τις υπόλοιπες, κατηγορικές ή ποσοτικές. Μπορούμε λοιπόν να λάβουμε τη μέση απόσταση για όλα τις $i \neq j$, ως εξής :

$$\delta\ (\alpha, \beta) = \frac{1}{p\ -\ 1} \sum_{j=1}^{p} \delta^{ij}(\alpha, \beta) \quad (9)$$

όπου p ο αριθμός των μεταβλητών στο σύνολο δεδομένων μεικτού τύπου.

Σε περίπτωση που η μεταβλητή j είναι ποσοτική, πρέπει να μετατραπεί σε διακριτή. Αυτό μπορεί να επιτευχθεί χωρίζοντας τις τιμές της σε N μέρη (διαστήματα). Η ίδια τιμή N πρέπει να χρησιμοποιηθεί για όλες τις ποσοτικές μεταβλητές. Ύστερα από αυτή τη διαδικασία μπορεί να υπολογιστεί το βάρος (ή σημαντικότητα) μιας ποσοτικής μεταβλητής j, όπως φαίνεται στην (6). Πιο συγκεκριμένα το w_j υπολογίζεται ως ο μέσος όρος της απόστασης όλων τω συνδυασμών των N τιμών που μπορεί να πάρει η κωδικοποιημένη ποσοτική μεταβλητή. Έχουμε $\binom{N}{2}$ διαφορετικούς συνδυασμούς, έτσι το βάρος μπορεί να αποδοθεί ως εξής:

$$w_j = \frac{\sum_{s=1}^{N} \sum_{r>s} \delta^j(r,s)}{\binom{N}{2}} \tag{10}$$

Όπου $\delta^j(r,s)$ η απόσταση μεταξύ των τιμών r και s της ομαδοποιημένης συνεχούς μεταβλητής j, υπολογισμένη σύμφωνα με τον τύπο (9). Πρακτικά η ομαδοποίηση γίνεται για την εκτίμηση της τιμής του w_j.

2.4 Ο αλγόριθμος Modha-Spangler K-means

Ένας εναλλακτικός αλγόριθμος ομαδοποίησης, στηριζόμενος στις αρχές του αλγορίθμου k-prototypes, είναι ο Modha-Spangler K-means (Modha & Spangler, 2003). Αποτελεί έναν κυρτό (convex) αλγόριθμο K-means ο οποίος υπολογίζει το βάρος των κατηγορικών μεταβλητών στη συστάδα *l* με αυτόματο τρόπο. Η κυρτότητα (convexity) του αλγορίθμου οφείλεται στο γεγονός πως η απόσταση ή ανομοιότητα μεταξύ του αντικειμένου X_i και του κεντροειδούς Q_l είναι ο σταθμισμένος μέσος της τετραγωνικής Ευκλείδειας απόστασης και της απόστασης συνημιτόνου (Cosine Distance) για κατηγορικές και συνεχείς μεταβλητές, αντίστοιχα. Και οι δύο αποστάσεις είναι κυρτές, έτσι το πρόβλημα φαίνεται να είναι πρόβλημα κυρτής βελτιστοποίησης. Πιο συγκεκριμένα, η απόσταση μεταξύ των X_i και Q_l μπορεί να αποδοθεί ως εξής:

$$d(X_i, Q_l) = \underbrace{\sum_{j=1}^{p_r} (x_{ij} - q_{lj})^2}_{d^{con}(X_i, Q_l)} + \underbrace{\gamma_l \left(1 - \frac{\sum_{j=p_r+1}^{p} x_{ij} q_{lj}}{\sqrt{\sum_{j=p_r+1}^{p} x^2{}_{ij}} \sqrt{\sum_{j=p_r+1}^{p} q^2{}_{lj}}} \right)}_{d^{cat}(X_i, Q_l)} \tag{11}$$

υποθέτοντας ξανά πως οι πρώτες p_r μεταβλητές είναι συνεχείς ποσοτικές και οι υπόλοιπες κατηγορικές. Αξίζει να σημειωθεί πως το δεύτερο μέρος της (11) λαμβάνει υπόψιν την υπόθεση πως οι μεταβλητές είναι κωδικοποιημένες σε 0-1 (dummy-coded). Όπως αναφέρθηκε, το βάρος γ_l των κατηγορικών μεταβλητών υπολογίζεται με αυτόματο τρόπο. Αυτό επιτυγχάνεται λαμβάνοντας υπόψιν το γινόμενο των αναλογιών της μέσης διασποράς στο εσωτερικό της συστάδας και της μέσης διασποράς μεταξύ των συστάδων για συνεχείς και κατηγορικές μεταβλητές, αντίστοιχα.

Η διασπορά εντός της συστάδας μπορεί να εκφραστεί ως:

$$\begin{aligned} \Gamma_{con} &= \sum_{m=1}^{l} \sum_{i:y_{im}=1} d^{con}\, d(X_i, Q_l) \\ \Gamma_{cat} &= \sum_{m=1}^{l} \sum_{i:y_{im}=1} d^{cat}\, d(X_i, Q_l) \end{aligned} \tag{12}$$

όπου d^{con} και d^{cat} είναι η τετραγωνική Ευκλείδεια απόσταση και η απόσταση συνημιτόνου (Cosine Distance) για συνεχείς και κατηγορικές μεταβλητές, αντίστοιχα, εκφρασμένες με τον ίδιο τρόπο όπως και στην (11). Το κέντρο της συστάδας l για συνεχείς μεταβλητές υπολογίζεται με τον γνωστό τρόπο, δηλαδή ως η μέση τιμή για κάθε μία μεταβλητή μεταξύ όλων των αντικειμένων στη συστάδα l, ενώ για τις κωδικοποιημένες (dummy) κατηγορικές μεταβλητές $\mathrm{j} = \mathrm{p_r} + 1, \ldots \mathrm{p}$, δίνεται από τον τύπο

$$\frac{\sum_{\mathrm{i:y_{il}=1}} \sum_{\mathrm{j=p_r+1}} \mathrm{x}_{ij}}{\left\| \sum_{\mathrm{i:y_{il}=1}} \sum_{\mathrm{j=p_r+1}} \mathrm{x}_{ij} \right\|}$$

Η μέση διασπορά μεταξύ των συστάδων για συνεχείς και κατηγορικές μεταβλητές δίνεται από τους τύπους:

$$\begin{aligned} \Delta_{con} &= \sum_{i=1}^{n} d^{con}\,(X_i, \bar{Q}_{con}) - \Gamma_{con} \\ \Delta_{cat} &= \sum_{i=1}^{n} d^{cat}\,(X_i, \bar{Q}_{cat}) - \Gamma_{cat} \end{aligned} \tag{13}$$

όπου τώρα $\bar{Q}_{con}$ και $\bar{Q}_{cat}$ είναι τα κέντρα όλων των συνεχών και κατηγορικών μεταβλητών αντίστοιχα. Αυτά υπολογίζονται με τον ίδιο τρόπο ακριβώς που περιγράφεται στην προηγούμενη παράγραφο, με τη διαφορά ότι δεν λαμβάνονται υπόψη μόνο τα αντικείμενα μίας συστάδας *l* , αλλά αντίθετα το σύνολο των αντικειμένων.
Έτσι, χρησιμοποιώντας τις (12), (13) λαμβάνουμε το γινόμενο των αναλογιών ως εξής:

$$\Pi_{cat} = \frac{\Gamma_{con}\Gamma_{cat}}{\Delta_{con}\Delta_{cat}} \quad (14)$$

Αυτή είναι η αντικειμενική συνάρτηση που πρέπει να ελαχιστοποιηθεί για να βρεθεί το βέλτιστο βάρος γ_l.

2.5 Μέθοδοι συσταδοποίησης και μείωσης διαστάσεων

Σε αυτή την ενότητα, θα αναφερθούμε στη χρήση μεθόδων μείωσης διαστάσεων (Dimensionality Reduction) για μεικτού τύπου δεδομένα. Πολλές φορές είναι προτιμότερο και ίσως απαραίτητο να διεξαχθεί η ανάλυση σε ένα χώρο με λιγότερες διαστάσεις, παρά στο αρχικό σύνολο των μεταβλητών. Θα παρουσιάσουμε αρχικά το συνδυασμό των μεθόδων PCAMIX και K-Means, ως μια προσέγγιση ομαδοποίησης μεικτού τύπου δεδομένων. Η συγκεκριμένη προσέγγιση περιλαμβάνει τη μείωση των διαστάσεων του συνόλου δεδομένων ως 1ο βήμα και ακολουθείται από έναν αλγόριθμο ομαδοποίησης (στη συγκεκριμένη περίπτωση τον αλγόριθμο K-Means) ως 2ο βήμα. Αυτού του είδους η προσέγγιση σε δύο βήματα (dimensionality Reduction and clustering) αναφέρεται στη βιβλιογραφία ως «Παράλληλη Ανάλυση» ή “Tandem Analysis” (Hubert & Arabie, 1985).

2.5.1 Η μέθοδος FAMD ή PCAMIX

Η Ανάλυση Παραγόντων για μεικτά δεδομένα (FAMD) ή αλλιώς Ανάλυση Κυρίων Συνιστωσών για μεικτά δεδομένα (PCAMIX) αποτελεί μία μέθοδο που ως πρώτο στάδιο έχει την μείωση των διαστάσεων και ως δεύτερο την ομαδοποίηση των αντικειμένων στο χώρο των λιγότερων διαστάσεων (Pagès, 2004). Ουσιαστικά αποτελεί μια ενδιάμεση

μέθοδο ανάμεσα στην Ανάλυση σε Κύριες Συνιστώσες – PCA και την Παραγοντική Ανάλυση των Πολλαπλών Αντιστοιχιών - MCA (Markos et al., 2020).

Η παραγοντική ανάλυση των πολλαπλών αντιστοιχιών (MCA) είναι μια επέκταση της απλής παραγοντικής ανάλυσης των αντιστοιχιών (CA) που επιτρέπει σε κάποιον να αναλύσει το μοτίβο των σχέσεων μεταξύ κατηγορικών μεταβλητών. Ως εκ τούτου, μπορεί επίσης να θεωρηθεί ως γενίκευση της μεθόδου PCA, όταν οι μεταβλητές που πρόκειται να αναλυθούν είναι κατηγορικές αντί ποσοτικές. Πρακτικά, όπως και στην PCA, η MCA δέχεται ως είσοδο έναν λογικό πίνακα (δηλαδή, έναν πίνακα του οποίου οι τιμές είναι 0 ή 1).

Παρατίθεται ένα παράδειγμα με ένα μικρό σύνολο δεδομένων.

	Τομέας απασχόλησης	Φύλο	Οικ. κατάσταση
1	Πωλήσεις	Άνδρας	Ελεύθερος/η
2	Πωλήσεις	Άνδρας	Διαζευγμένος/η
3	Πληροφορική	Γυναίκα	Ελεύθερος/η
4	Οικονομικά	Άνδρας	Παντρεμένος/η
5	Οικονομικά	Γυναίκα	Διαζευγμένος/η
6	Πωλήσεις	Γυναίκα	Παντρεμένος/η
7	Πληροφορική	Άνδρας	Ελεύθερος/η

Αρχικά λοιπόν ο παραπάνω πίνακας δεδομένων μετατρέπεται σε λογικό ή διαζευκτικό ή πίνακα 0-1. Έχοντας λοιπόν n παρατηρήσεις, p κατηγορικές μεταβλητές, και K_j κατηγορίες, ο πίνακας 0-1 θα είναι ένας πίνακας μεγέθους $n \times K_j$, όπως φαίνεται παρακάτω.

	Άνδρας	*Γυναίκα*	*Πωλήσεις*	*Πληροφορική*	*Οικονομικά*	Ελεύθερος/η	Παντρεμένος/η	Διαζευγμένος/η
1	1	0	1	0	0	1	0	0
2	1	0	1	0	0	0	0	1
3	0	1	0	1	0	1	0	0
4	1	0	0	0	1	0	1	0
5	0	1	0	0	1	0	0	1
6	0	1	1	0	0	0	1	0
7	1	0	0	1	0	1	0	0

Ορίζεται επίσης ως π_κ το βάρος ή το ποσοστό υποκειμένων τα οποία ανήκουν στην κατηγορία k. Αυτό ισούται με $\frac{1}{n\sum_{i=1}^{n} y_{ik}}$, όπου y_{ik} είναι η παρατήρηση στην i γραμμή

και k στήλη του διαζευκτικού πίνακα. Για παράδειγμα, το k αφορά την κατηγορία 'Γυναίκα'. Τότε το $\pi_{\kappa} = \frac{1}{7(1+1+1)} = \frac{1}{21}$.

Εδώ παρατηρούμε πως οι απαντήσεις κάθε ερωτώμενου/ης παρουσιάζονται σαν ξεχωριστές γραμμές στον διαζευκτικό πίνακα και είναι εύκολο να αντιληφθούμε τις διαφορές μεταξύ των υποκειμένων για κάθε κατηγορία Κ, αν παρατηρήσουμε με ποιο τρόπο διαφοροποιούνται οι τιμές για κάθε κατηγορία Κ στις αντίστοιχες γραμμές. Πρακτικά, όπως και στην PCA, όπου «εξάγονται» οι κύριες συνιστώσες που μας βοηθούν να ερμηνεύσουμε το μεγαλύτερο μέρος της μεταβλητότητας των ποσοτικών μεταβλητών, έτσι και η MCA «εξάγει» τους κύριους άξονες που μας βοηθούν να εξηγήσουμε το μεγαλύτερο μέρος της μεταβλητότητας των κατηγορικών μεταβλητών.

Όπως προαναφέρθηκε, ένας ακόμα στόχος της MCA είναι να εξετάσει τις πιθανές σχέσεις μεταξύ των κατηγοριών. Αυτό επιτυγχάνεται με τη δημιουργία «συνθετικών» μεταβλητών, από τις οποίες μπορούμε να εξάγουμε κάποιους μετρήσιμους δείκτες με βάση τις κατηγορικές μεταβλητές που «εξηγούν» αυτές τις κατηγορικές μεταβλητές. Αυτό πολλές φορές μπορεί να είναι περίπλοκο, έτσι χρησιμοποιείται γραφική αναπαράσταση ώστε να επιλεχθούν τα πιο σημαντικά στοιχεία.

Έστω X ένας πίνακας μεγέθους n × p. Διαχωρίζουμε τον πίνακα ως $X = [X_{con}, X_{cat}]$ όπου X_{con} ένας $n \times p_r$ πίνακας που περιλαμβάνει όλες τις συνεχείς μεταβλητές και X_{cat} ένας $n \times (p - p_r)$ πίνακας με όλες τις κατηγορικές μεταβλητές. Στη συνέχεια, ο X_{cat} μετατρέπεται στον $\widehat{Xcat}$ όπου αποτελεί τον διαζευκτικό πίνακα του X_{cat}. Κανονικοποιούμε σε επόμενη φάση τον X_{con} με τον γνωστό τρόπο, αφαιρώντας δηλαδή από κάθε στήλη τον μέσο και διαιρώντας με την τυπική απόκλιση της στήλης. Η κανονικοποίηση για τις κατηγορικές μεταβλητές που βρίσκονται X_{cat} γίνεται διαιρώντας τα στοιχεία κάθε στήλης με την ρίζα του «ποσοστού» των υποκειμένων που ανήκουν στην κατηγορία j, $\sqrt{\pi_j}$.

Αφού επιτευχθεί η διαδικασία της κανονικοποίησης, ενώνονται οι πίνακες και η PCA εφαρμόζεται στον τελικό ενιαίο πίνακα.

Παρατηρούμε πως η 1η κύρια συνιστώσα $\mathbf{f_1}$ μεγιστοποιεί τη σχέση μεταξύ συνεχών και κατηγορικών μεταβλητών, μεγιστοποιώντας το

$$\sum_{j=1}^{p_r} R^2(f_1, X_{con_j}) + \sum_{j=p_r+1}^{p} \eta^2(f_1, X_{cat_j}) \qquad (15)$$

όπου X_{con_j} αποτελεί την j-οστή συνεχή και X_{cat_j} την j-οστή κατηγορικό μεταβλητή αντίστοιχα. R^2 είναι ο Συντελεστής Προσδιορισμού, ενώ η^2 είναι ο Συντελεστής Ενδοσυσχέτισης. Πρακτικά ο η^2 είναι το πηλίκο της σταθμισμένης διακύμανσης των μέσων μεταξύ των ομάδων με την διακύμανση του δείγματος και παίρνει τιμές μεταξύ 0 και 1. Αυτό σημαίνει πως η τιμή $\eta^2 = 0$ αντιστοιχεί σε μηδενική διακύμανση μεταξύ των μέσων των ομάδων, ενώ η τιμή $\eta^2 = 1$ αντιστοιχεί σε μηδενική διασπορά εσωτερικά των ομάδων. Έτσι, η $\mathbf{f_1}$ μπορεί να ερμηνευθεί ως η «συνθετική» μεταβλητή η οποία συσχετίζεται περισσότερο με τις συνεχείς και κατηγορικές μεταβλητές, σε σχέση με τον συντελεστή προσδιορισμού και τον Συντελεστή Ενδοσυσχέτισης αντίστοιχα.

Εναλλακτικά, η Ανάλυση Παραγόντων για μεικτά δεδομένα (FAMD) μπορεί να ιδωθεί ως η Ανάλυση Κυρίων Συνιστωσών (PCA) του σταθμισμένου πίνακα $XD_{\Sigma}^{1/2}$, όπου D_{Σ} είναι ο πίνακας μεγέθους

$$\left(n \; x \; \left(p_r + \sum_{j=p_r+1}^{p} K_j \right) \right)$$

όπου K_j ο αριθμός των μοναδικών τιμών που μπορεί να πάρει η κατηγορική μεταβλητή j , ως εξής:

$$\boldsymbol{D_{\Sigma}} = \begin{pmatrix} \boldsymbol{s_1^2} & & & & & \\ & \ddots & & & & \\ & & \boldsymbol{s_{p_r}^2} & & & \\ & & & \boldsymbol{\pi_{pr+1}} & & \\ & & & & \ddots & \\ & & & & & \boldsymbol{\pi_j} \end{pmatrix}$$

όπου $J = p_r + \sum_{j=p_r+1}^{p} K_j$, π_κ το ποσοστό των υποκειμένων τα οποία ανήκουν στην κατηγορία k και $\boldsymbol{s_j}$ η τυπική απόκλιση της j-οστής συνεχούς μεταβλητής. Επίσης, η λύση της FAMD μπορεί να βρεθεί λαμβάνοντας μέσω της Διάσπασης των Ιδιοτιμών (ή Singular Value Decomposition, ή SVD) (Hastie et al., 2009) του $XD_{\Sigma}^{1/2}$ – M, με τον M έναν n × $\boldsymbol{J}$ πίνακα με όλες τις γραμμές του ίσες με τους μέσους των στηλών του σταθμισμένου πίνακα $XD_{\Sigma}^{1/2}$. Με την αφαίρεση του M, λοιπόν, καταφέρνουμε κάθε στήλη να έχει μέσο όρο, ίσο με το 0.

Έχει ενδιαφέρον ο λόγος που χρησιμοποιείται ο σταθμισμένος πίνακας $XD_{\Sigma}^{1/2}$ στο κομμάτι της μείωσης διαστάσεων στην FAMD. Αυτή η στάθμιση συνεπάγεται ότι οι αποστάσεις μεταξύ δύο αντικειμένων στον αρχικό χώρο (πριν την εφαρμογή της PCA), δίνονται από το συνδυασμό της τετραγωνικής Ευκλείδειας απόστασης και της σταθμισμένης απόστασης $\mathbf{X^2}$ για τις συνεχείς και τις κατηγορικές μεταβλητές, αντίστοιχα. Πρακτικά η PCA χρησιμοποιεί την τετραγωνική Ευκλείδεια απόσταση για τις πρώτες p_r συνεχείς μεταβλητές, ενώ η MCA την απόσταση $\mathbf{X^2}$ **(**Pagès, 2014) για τις υπόλοιπες J - p_r κατηγορικές μεταβλητές. Έτσι, η εφαρμογή της PCA στον $XD_{\Sigma}^{1/2}$, αποτελεί μια «ενδιάμεση» μέθοδο μεταξύ των PCA και MCA, όπως αναφέρθηκε παραπάνω. Η απόσταση μεταξύ δύο αντικειμένων $\boldsymbol{X_i}$ *και* $\boldsymbol{X_i}'$ δίνεται από:

$$d^2(Xi\,, Xi') \;=\; \sum_{j=1}^{p_r} \frac{(x_{ij} \;-\; x_{i'\,j})^2}{s_j^2} \;+\; \sum_{j=p_r+1}^{J} \frac{(x_{ij} \;-\; x_{i'\,j})^2}{\pi_j} \qquad (16)$$

Πράγματι, το 1ο μέρος της (16) αντιστοιχεί στην τετραγωνική Ευκλείδεια απόσταση μεταξύ των συνεχών μεταβλητών, διαιρεμένη με την τυπική απόκλιση. Η στάθμιση στο 2ο μέρος της (16) επιβεβαιώνει πως δύο αντικείμενα τα οποία ανήκουν σε διαφορετικές κατηγορίες της ίδιας όμως κατηγορικής μεταβλητής, έχουν μεγαλύτερη απόσταση όταν ένα από αυτά ανήκει σε μία «σπάνια» κατηγορία σε σχέση με όταν και τα δύο ανήκουν σε κατηγορίες με μεγαλύτερη συχνότητα. Οι συχνότητες των κατηγοριών σχετίζονται με την διασπορά στον αρχικό χώρο, που ονομάζεται αδράνεια (inertia). Η αδράνεια για μια κατηγορία j δίνεται από τον τύπο $1 - \pi_j$, που δείχνει ότι οι περισσότερο συχνές κατηγορίες έχουν μικρή αδράνεια, ενώ η αδράνεια για τις πιο σπάνιες πλησιάζει το 1.

Στο τελικό στάδιο της FAMD, εφαρμόζεται μια μέθοδος ομαδοποίησης στα σκορ των υποκειμένων στις s < p διαστάσεις που παρέμειναν μετά την μείωση των διαστάσεων. Ωστόσο, η επιλογή του s στο πρώτο βήμα, μπορεί να είναι κρίσιμη για την τελική λύση ομαδοποίησης. Επιπλέον, προσέξτε ότι σε αυτή την προσέγγιση δύο βημάτων, βελτιστοποιούνται δύο διαφορετικά κριτήρια. Ενώ η μείωση διαστάσεων στοχεύει στον καθορισμό ενός μειωμένου συνόλου συνδυασμών των αρχικών μεταβλητών που μεγιστοποιούν την αρχική μεταβλητότητα, η ανάλυση συστάδων στοχεύει στην ελαχιστοποίηση της μεταβλητότητας εντός της ομάδας μεγιστοποιώντας παράλληλα την μεταβλητότητα μεταξύ των ομάδων. Αυτή η ασυμφωνία των στόχων

μπορεί να οδηγήσει στο λεγόμενο *πρόβλημα απόκρυψης των ομάδων* (cluster masking problem), που παρουσιάζεται αναλυτικά σε επόμενη ενότητα.

2.6 Ταυτόχρονη Μείωση Διαστάσεων & Ομαδοποίηση Μεικτών Δεδομένων

Σε αυτή την ενότητα, θα παρουσιαστεί μία μέθοδος ταυτόχρονης μείωσης των διαστάσεων και ομαδοποίησης των μεικτού τύπου δεδομένων. Η χρήση της μεθόδου αυτής είναι ικανή να ξεπεράσει το προαναφερθέν πρόβλημα της απόκρυψης των ομάδων (cluster masking problem). Πιο συγκεκριμένα θα αναφερθούμε στη μέθοδο Reduced K-means (RKM) και στον αλγόριθμο εναλλασσόμενων ελαχίστων τετραγώνων (Alternating Least Squares – ALS) που προτείνεται για την αποφυγή του προβλήματος.

2.6.1 Το πρόβλημα της απόκρυψης των ομάδων

Όπως αναφέρεται παραπάνω, η διαδοχική μέθοδος ανάλυσης (tandem analysis) αποτελείται από δύο βήματα, αρχικά το βήμα της μείωσης των διαστάσεων και έπειτα την εφαρμογή ενός αλγορίθμου ομαδοποίησης. Αυτή η διαδοχική προσέγγιση όμως, ενέχει πιθανούς κινδύνους, καθώς ενώ η μείωση διαστάσεων έχει σαν στόχο την διατήρηση μόνο των διαστάσεων που εξηγούν το μεγαλύτερο μέρος της μεταβλητότητας στο σύνολο δεδομένων, ενώ η ομαδοποίηση έχει σκοπό την ελαχιστοποίηση της μεταβλητότητας εντός της κάθε συστάδας, μεγιστοποιώντας τη μεταβλητότητα μεταξύ των συστάδων. Έτσι είναι φανερή η διαφορά μεταξύ των στόχων των δύο μεθόδων. Αυτό το πρόβλημα εμφανίζεται όταν το βήμα μείωσης των διαστάσεων αποκρύπτει την υποκείμενη δομή των ομάδων. Αυτό θα μπορούσε, για παράδειγμα, να συμβεί όταν μεταβλητές που δεν σχετίζονται με τη δομή ομαδοποίησης συσχετίζονται ισχυρά μεταξύ τους. Ενδεικτικό παράδειγμα του παραπάνω προβλήματος δίνεται στους Vichi & Kiers (2001) και συζητείται περαιτέρω στους Van Buuren & Heiser (1989) και De Soete & Carroll (1994). Μια λύση σε αυτό το πρόβλημα έχει δοθεί με την κοινή βελτιστοποίηση των δύο κριτηρίων που θα περιγραφεί παρακάτω.

2.6.2 Η μέθοδος Mixed Reduced K-Means

Ο αλγόριθμος **Mixed Reduced K-Means** είναι μία μέθοδος που επιτυγχάνει ταυτόχρονη μείωση των διαστάσεων και ανάλυση σε συστάδες, η οποία προτάθηκε αρχικά μόνο για

συνεχείς μεταβλητές από τους De Soete & Carroll (1994) και γενικεύτηκε στη συνέχεια για δεδομένα μεικτού τύπου από τους van de Velden, Iodice D'Enza & Markos (2019).

Για τη μαθηματική περιγραφή της μεθόδου ορίζουμε τα παρακάτω σύμβολα:

n	Αριθμός αντικειμένων
p	Αριθμός μεταβλητών
s	Αριθμός παραγόντων – διαστάσεων που θα διατηρήσουμε
k	Αριθμός συστάδων
$\|.\|_F$	Νόρμα Frobenius
X	πίνακας δεδομένων $n \times p$, όπου οι μεταβλητές έχουν κανονικοποιηθεί
B	πίνακας φορτίσεων $p \times s$ με $B^T B = \mathrm{I}_\sigma$
G	πίνακας $k \times s$ με τα κέντρα των συστάδων στο χώρο των s-διαστάσεων
Z_k	πίνακας διαμέρισης $n \times k$

Ο πίνακας διαμέρισης Z_k είναι ένας πίνακας 0-1 με στοιχεία $(Z_k)_{i,j}$ = 1, όταν το αντικείμενο i βρίσκεται στη συστάδα j , διαφορετικά $(Z_k)_{i,j}$ = 0. Η μεθόδος Mixed RKM επιδιώκει να κατανείμει τα αντικείμενα σε συστάδες και ταυτόχρονα να μειώσει τις διαστάσεις (μεταβλητές), με τρόπο τέτοιο ώστε να μεγιστοποιηθεί η διακύμανση μεταξύ των συστάδων στο μικρότερο χώρο των s-διαστάσεων. Η αντικειμενική συνάρτηση που πρέπει να βελτιστοποιηθεί είναι η:

$$f_{RKM}(B, Z_k, G) = \|X - Z_k G B^T\|_F^2 \quad (17)$$

Για την ελαχιστοποίηση της (17), αρχικά υπολογίζεται ο πίνακας με τα κέντρα των συστάδων, $G = (Z_\kappa^T Z_\kappa)^{-1} Z_k^T XB$ (van de Velden et al., 2019). Αν αντικαταστήσουμε το G στην (17) έχουμε:

$$min f_{RKM}(B, Z_k) = \left\|X - PXBB^{\mathrm{T}}\right\|_F^2 \quad (18)$$

όπου P = $Z_\kappa (Z_\kappa^T Z_\kappa)^{-1} Z_k^T$, είναι ένας πίνακας προβολής (projection matrix) μεγέθους $n \times n$. Η (18) μπορεί να απλοποιηθεί ως εξής:

$$min f_{RKM}(B, Z_k) = trace(X^{\mathrm{T}}X) - trace(B^{\mathrm{T}}X^{\mathrm{T}}PXB) \quad (19)$$

όπου trace το ίχνος του πίνακα. Αυτό σημαίνει ότι η ελαχιστοποίηση της (18) ισοδυναμεί με τη μεγιστοποίηση του 2ου μέρους της (19), δηλ. $trace(B^T X^T PXB)$. Ορίζουμε τον πίνακα Z, ως τον λογικό πίνακα 0-1 που κατασκευάζεται από τις κατηγορικές μεταβλητές. Αν υποθέσουμε ότι οι πρώτες $p_r < p$ μεταβλητές στο σύνολο δεδομένων είναι συνεχείς και οι υπόλοιπες $p - p_r$ είναι κατηγορικές με κάθε κατηγορική μεταβλητή j να έχει K_j διαφορετικές κατηγορίες, ο λογικός πίνακας 0-1 θα είναι ένας πίνακας μεγέθους $n \times Q$, όπου $Q = \sum_{j=p_r+1}^{p} K_j$. Επίσης, ορίζεται ο πίνακας $M = I_n - \frac{1}{n}1_n 1_n^T$, που αντιστοιχεί στην πράξη της αφαίρεσης του μέσου όρου από κάθε μεταβλητής ώστε ο μέσος όρος όλων των μεταβλητών να είναι 0.

Τελικά, οδηγούμαστε στη μεγιστοποίηση της συνάρτησης (van de Velden et al., 2017):

$$f_{CA}(B, Z_k) = \frac{1}{n(p - p_r)^2} trace(B^T Z^T MPMZB) \quad (20)$$

όπου $\frac{1}{n(p-p_r)} B^T D_z B = I_s$ **,** όπου D_z ο διαγώνιος πίνακας μεγέθους $Q \times Q$ που ικανοποιεί την εξίσωση $D_z 1_Q = Z^T I_n$ και B είναι τώρα ένας πίνακας φορτίσεων μεγέθους $Q \times s$. Μπορεί επίσης να οριστεί ο πίνακας $\boldsymbol{B^* = \frac{1}{\sqrt{n(p-p_r)}} D_z^{\frac{1}{2}} B}$, ώστε η (20) να πάρει τη μορφή:

$$f_{CA}(\boldsymbol{B^*, Z_k}) = \frac{1}{p - p_r} trace(\boldsymbol{B^{*T} D_z^{-\frac{1}{2}} Z^T MPMZ D_z^{-\frac{1}{2}} B^*}) \quad (21)$$

Ισχύει ότι $\boldsymbol{B^{*T} B^* = I_s}$. Αν υποθέσουμε ότι ο πίνακας διαμέρισης $\boldsymbol{Z_k}$ παραμένει σταθερός και η (21) μπορεί να γραφεί ως η παρακάτω διάσπαση σε ιδιοτιμές:

$$\frac{1}{p - p_r} \boldsymbol{D_z^{-\frac{1}{2}} Z^T MPMZ D_z^{-\frac{1}{2}} = B^* \Lambda^2 B^{*T}} \quad (22)$$

όπου πολλαπλασιάζουμε την (21) με $\boldsymbol{B^*}$ και $\boldsymbol{B^{*T}}$ στο αριστερό και δεξί μέρος, αντίστοιχα και τέλος θέτουμε

$\boldsymbol{\Lambda^2 = D_z^{-\frac{1}{2}} Z^T MPMZ D_z^{-\frac{1}{2}}}$**.** Έτσι μπορούμε πλέον να υπολογίσουμε το $\boldsymbol{B^*}$

και κατ΄ επέκταση το $\boldsymbol{B} = \sqrt{n(p - p_r)}\boldsymbol{D}_{\boldsymbol{z}}^{-\frac{\mathbf{1}}{\mathbf{2}}}\boldsymbol{B}^{*}$.

Αφού υπολογίσουμε το νέο $\boldsymbol{B}$, το κρατάμε σταθερό και επιδιώκουμε να υπολογίσουμε τον πίνακα διαμέρισης $\boldsymbol{Z_k}$ που θα μεγιστοποιήσει την (21). Σύμφωνα με τους van de Velden et al. (2017), αυτό το πρόβλημα μεγιστοποίησης μπορεί ναεκφραστεί ως ένα πρόβλημα συσταδοποίησης *K*-means. Πιο συγκεκριμένα, έχουμε στόχο να ελαχιστοποιήσουμε την:

$$f'_{CA}(\boldsymbol{Z_k}, \boldsymbol{G}) = \left\| \sqrt{\frac{n}{p - p_r}} \boldsymbol{M}\boldsymbol{Z}\boldsymbol{D}_{z}^{-\frac{1}{2}}\boldsymbol{B}^{*} - \boldsymbol{Z}_k\boldsymbol{G}) \right\|_F^2 \qquad (23)$$

Συγκρίνοντας τις (17) και (23), μπορούμε να παρατηρήσουμε μεγάλη ομοιότητα μεταξύ τους. Αν θέσουμε $\boldsymbol{Y} = \sqrt{\frac{n}{p - p_r}}\boldsymbol{M}\boldsymbol{Z}\boldsymbol{D}_{z}^{-\frac{1}{2}}\boldsymbol{B}^{*}$, η (23) γίνεται $\|\boldsymbol{Y} - \boldsymbol{Z}_k\boldsymbol{G})\|_F^2$, όπου είναι σχεδόν ίδια με την f_{RKM}. Έτσι, είναι εύκολο να συμπεράνουμε πως το ζητούμενο $\boldsymbol{G}^{*}$ που ελαχιστοποιεί την $f'_{CA}(\boldsymbol{Z_k}, \boldsymbol{G})$ είναι το $\boldsymbol{G}^{*} = (\boldsymbol{Z}_k^{\mathrm{T}}\boldsymbol{Z}_k)^{-\mathbf{1}}\boldsymbol{Z}_k^{\mathrm{T}}\boldsymbol{Y}$.

Για να εφαρμοστεί η μέθοδος **Mixed Reduced K-Means**, πρέπει να κατασκευαστεί ένας πίνακας 0-1 για τις κατηγορικές μεταβλητές, να ενωθεί κατά στήλες με τις συνεχείς μεταβλητές. Έπειτα γίνεται κανονικοποίηση των κατηγορικών αλλά και των συνεχών μεταβλητών και τότε ο πίνακας δεδομένων $\boldsymbol{X}$ είναι έτοιμος για εφαρμογή του παραπάνω αλγόριθμου. Ο αρχικός πίνακας διαμέρισης $\boldsymbol{Z_k}$, συνήθως αρχικοποιείται με τυχαίες τιμές. Με την επιλογή κατάλληλου αριθμού διαστάσεων $\boldsymbol{s}$, οι στήλες του $\boldsymbol{B}^{*}$ υπολογίζονται με την (22) και καταλήγουμε στον υπολογισμό του $\boldsymbol{B}$. Με σταθερό τονν $\boldsymbol{B}^{*}$, μπορούμε να υπολογίσουμε τον $\boldsymbol{Y}$ στον οποίο εφαρμόζουμε τον αλγόριθμο *k*-means για την ενημέρωση των τιμών του $\boldsymbol{Z}_k$, καθώς και για την ενημέρωση των κέντρων των συστάδων που βρίσκονται στον πίνακα $\boldsymbol{G}$. Αυτή η επαναληπτική διαδικασία επαναλαμβάνεται έως ότου ο πίνακας $\boldsymbol{Z}_k$, παραμείνει σταθερός (καθώς και οι $\boldsymbol{Y}$ και $\boldsymbol{G}$).

3. Εφαρμογή σε πραγματικό σύνολο δεδομένων ιστού

Σε αυτό το κεφάλαιο θα εφαρμόσουμε τις τεχνικές που περιγράφηκαν παραπάνω σε ένα πραγματικό σύνολο δεδομένων ιστού. Τα δεδομένα είναι ανωνυμοποιημένα και αφορούν σε μια αθλητική εκπομπή που μεταδόθηκε μέσω πλατφόρμας κοινωνικών δικτύων κατά τη διάρκεια ενός παγκοσμίου κυπέλλου ποδοσφαίρου, πριν ή μετά τη διεξαγωγή κάθε αγώνα. Οι γραμμές του συνόλου δεδομένων αναφέρονται σε 71 εκπομπές διάρκειας 25 έως 52 λεπτών η καθεμιά, για τις οποίες καταγράφηκαν τα παρακάτω χαρακτηριστικά – μεταβλητές (στήλες):

- *duration*: η διάρκεια της εκάστοτε εκπομπής
- *hostess*: αν στο πάνελ ήταν παρούσα η γυναίκα παρουσιάστρια (τιμή = 1) ή μόνο οι άντρες παρουσιαστές (τιμή = 0)
- *day*: ο αριθμός ημέρας διεξαγωγής του τουρνουά
- *weekday*: η ημέρα της εβδομάδας που προβλήθηκε η εκπομπή
- *match*: η φάση της ημέρας που προβλήθηκε η εκπομπή (π.χ. αργά το απόγευμα)
- *peakliveview*: ο μεγαλύτερος αριθμός θεατών που παρακολουθούσαν το βίντεο, ενώ μεταδιδόταν ζωντανά
- *minview*: τα συνολικά λεπτά του χρόνου παρακολούθησης του βίντεο
- *uniqueview*: ο αριθμός μοναδικών θεάσεων
- *videoviews*: ο αριθμός των βίντεο που προβλήθηκαν
- *tensecviews*: ο αριθμός των φορών που παρακολουθήθηκε το βίντεο για τουλάχιστον 10 δευτερόλεπτα ή σχεδόν για τη συνολική του διάρκεια, όποιο από τα δύο συνέβη πρώτο.
- *avgwatch*: Ο μέσος χρόνος παρακολούθησης ενός βίντεο. Η τιμή του υπολογίζεται ως ο συνολικός χρόνος παρακολούθησης του βίντεο, διαιρεμένος με τον συνολικό αριθμό των αναπαραγωγών βίντεο, συμπεριλαμβανομένων των επαναλήψεων.
- *followers*: αντιπροσωπεύει το % της διατήρησης κοινού μεταξύ των ακολούθων της σελίδας και των υπόλοιπων θεατών.
- *peoplereach*: ο αριθμός των μοναδικών ατόμων που έχουν δει το περιεχόμενο των βίντεο

- *reactions*: ο αριθμός αντιδράσεων στο βίντεο (π.χ. θυμωμένος, χαρούμενος κλπ.)
- *comments*: ο αριθμός σχολίων στο βίντεο
- *shares*: ο αριθμός κοινοποιήσεων του βίντεο
- *Brazil*: δίτιμη μεταβλητή που υποδεικνύει εάν αγωνιζόταν η ομάδα της Βραζιλίας (τιμή = 1) ή όχι (τιμή = 0).

Όλες οι αναλύσεις που ακολουθούν πραγματοποιήθηκαν στη γλώσσα R και ο κώδικας βρίσκεται στο Παράρτημα Α.

3.1 Συντελεστής ανομοιότητας του Gower & Μέθοδος Partitioning Around Medoids

Αρχικά, εφαρμόστηκε ο συνδυασμός του συντελεστή ανομοιότητας του Gower με τη μέθοδο PAM. Για την εφαρμογή αυτή, χρησιμοποιήθηκε το πακέτο ***cluster*** της **R** (Maechler et al., 2013), όπου με την βοήθεια της συνάρτησης ***daisy()*** κατασκευάζουμε τον πίνακα ανομοιότητας μεταξύ των 71 εκπομπών. Στη συνέχεια, με την συνάρτηση ***pam(),*** εφαρμόζουμε την μέθοδο PAM, εξετάζοντας τις λύσεις από 2 έως 6 ομάδες. Για την εξέταση της ποιότητας κάθε λύσης, μπορούμε να βασιστούμε σε κριτήρια αξιολόγησης της εσωτερικής εγκυρότητας της λύσης της ομαδοποίησης, όπως το Average Silhouette Width (ASW) (Rousseeuw, 1987), το οποίο υπολογίζεται σε επίπεδο παρατήρησης και λαμβάνει τιμές ανάμεσα στο –1 και +1. Τιμές κοντά στο +1 δείχνουν ότι μια παρατήρηση βρίσκεται πιο κοντά στην ομάδα όπου έχει τοποθετηθεί απ’ ό,τι στις άλλες ομάδες. Αντίθετα, τιμές κοντά στο 0 ή κάτω από το 0, δείχνουν ότι η παρατήρηση βρίσκεται πιο κοντά σε άλλες ομάδες απ’ ό,τι στην ομάδα που έχει τοποθετηθεί. Για τον υπολογισμό του δείκτη σε επίπεδο ομάδας, υπολογίζεται ο μέσος όρος των τιμών του ASW των παρατηρήσεων της ομάδας. Τα αποτελέσματα έδειξαν ότι η μικρότερη τιμή του ASW αντιστοιχεί στη λύση με 2 ομάδες (0.217). Για την ερμηνεία των δύο ομάδων, υπολογίστηκαν οι δείκτες Cla/Mod, Mod/Cla και Global μέσω του πακέτου FactoMineR της R και τη συνάρτηση catdes() (Lê, Josse, & Husson, 2008). Για κάθε κατηγορία μεταβλητής, ο δείκτης Cla/Mod δείχνει το ποσοστό των αντικειμένων της κατηγορίας που βρίσκονται εντός της ομάδας, ο δείκτης Mod/Cla δείχνει το ποσοστό των αντικειμένων της ομάδας που χαρακτηρίζονται από τη συγκεκριμένη κατηγορία και ο

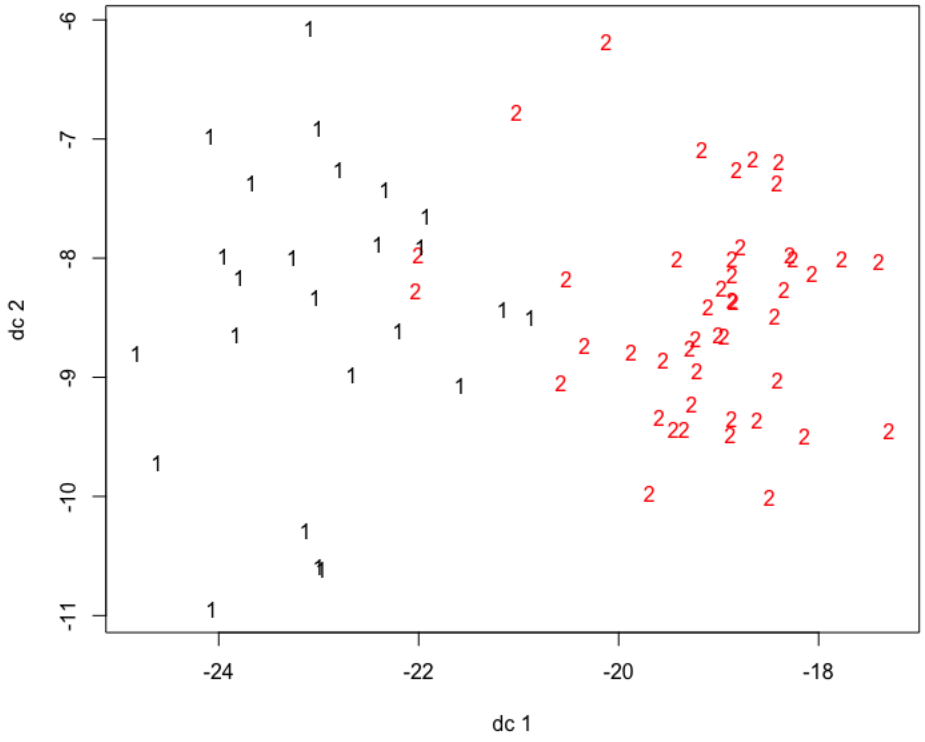
δείκτης Global δείχνει το ποσοστό της κατηγορίας στο σύνολο των αντικειμένων. Με βάση αυτούς τους δείκτες μπορούμε να ερμηνεύσουμε τις δύο ομάδες και τα χαρακτηριστικά τους.

```
Description of each cluster by the categories
=============================================
$`1`
                    Cla/Mod Mod/Cla    Global
hostess=1         84.615385      88 36.619718
weekday=Tuesday   90.909091      40 15.492958
day=14           100.000000      12  4.225352
day=13           100.000000      12  4.225352
day=6            100.000000      12  4.225352
weekday=Saturday   0.000000       0 21.126761
hostess=0          6.666667      12 63.380282

$`2`
                    Cla/Mod   Mod/Cla    Global
hostess=0         93.333333 91.304348 63.380282
weekday=Saturday 100.000000 32.608696 21.126761
day=14             0.000000  0.000000  4.225352
day=13             0.000000  0.000000  4.225352
day=6              0.000000  0.000000  4.225352
weekday=Tuesday    9.090909  2.173913 15.492958
hostess=1         15.384615  8.695652 36.619718
```

```
$`1`
               v.test Mean in category Overall mean
reactions     5.681458           132.68    91.859155
tensecviews   4.512016          2868.56  2427.464789
peakliveview  4.453820           120.44   100.380282
videoviews    4.299360          5698.32  4867.112676
uniqueview    4.285580          5555.76  4748.084507
minview       4.258737          7269.12  6179.647887
shares        3.875567             4.60     3.422535
comments      2.777499           393.08   274.788732
avgwatch      2.308307            23.64    21.901408
followers    -4.609342            86.96    89.943662
```

```
$`2`
                v.test Mean in category Overall mean
followers     4.609342         91.565217    89.943662
avgwatch     -2.308307         20.956522    21.901408
comments     -2.777499        210.500000   274.788732
shares       -3.875567          2.782609     3.422535
minview      -4.258737       5587.543478  6179.647887
uniqueview   -4.285580       4309.130435  4748.084507
videoviews   -4.299360       4415.369565  4867.112676
peakliveview -4.453820         89.478261   100.380282
tensecviews  -4.512016       2187.739130  2427.464789
reactions    -5.681458         69.673913    91.859155
```



Διάγραμμα 1. Οι παρατηρήσεις των δύο ομάδων σε χώρο δύο διαστάσεων (παραγοντικό επίπεδο της μεθόδου MDS στον πίνακα ανομοιότητας). Οι αριθμοί 1 και 2 αντιστοιχούν στις δύο ομάδες στις οποίες κατέληξε η μέθοδος Gower – PAM.

- **1η ομάδα (*n* = 46, 64,8%)**
  - Κατηγορικές μεταβλητές:
    - Η 1η ομάδα χαρακτηρίζεται κυρίως από εκπομπές όπου συμμετέχει και η γυναίκα παρουσιάστρια,
    - από εκπομπές που προβλήθηκαν ημέρα Τρίτη και

- από εκπομπές που προβλήθηκαν την 6η, 13η και 14η ημέρα της διοργάνωσης.

➢ Συνεχείς μεταβλητές:

- Η 1η ομάδα χαρακτηρίζεται κυρίως από εκπομπές με μεγαλύτερο μέσο όρο αντιδράσεων, φορών που παρακολουθήθηκε το βίντεο για τουλάχιστον 10 δευτερόλεπτα ή σχεδόν για τη συνολική του διάρκεια, όποιο από τα δύο συνέβη πρώτο και μέγιστου αριθμού θεατών όσο μεταδιδόταν ζωντανά.

❖ **2η ομάδα (*n* = 25, 35,2%)**

➢ Κατηγορικές μεταβλητές:

- Η 2η ομάδα χαρακτηρίζεται κυρίως από εκπομπές όπου δεν συμμετέχει η γυναίκα παρουσιάστρια και
- από εκπομπές που προβλήθηκαν ημέρα Σάββατο

➢ Συνεχείς μεταβλητές:

- Η 2η ομάδα χαρακτηρίζεται κυρίως από εκπομπές όπου το ποσοστό ακολούθων της σελίδας επί τον αριθμό των θεατών ήταν μεγαλύτερο του μέσου όρου.

Το Διάγραμμα 1 αντιστοιχεί στο παραγοντικό επίπεδο 1 x 2 της εφαρμογής της μεθόδου MDS (multidimensional scaling) στον πίνακα ανομοιότητας 71 x 71. Τα υποκείμενα που ανήκουν σε κάθε ομάδα έχουν αριθμηθεί (1 ή 2). Από το διάγραμμα διαπιστώνουμε ότι η μέθοδος μπορεί να διαχωρίσει τις δύο ομάδες, αλλά υπάρχει και κάποιος βαθμός επικάλυψης μεταξύ των δύο ομάδων.

3.2 Αλγόριθμος k-prototypes

Συνεχίζουμε με την εφαρμογή της μεθόδου **k-prototypes** με τη χρήση του πακέτου ***clustMixType*** της **R** (Szepannek, 2018) και τη συνάρτηση kproto().

```
Description of each cluster by the categories
=============================================
$`1`
                     Cla/Mod  Mod/Cla   Global
match=late_evening 71.42857 42.85714 29.57746
match=night        24.00000 17.14286 35.21127
                        p.value    v.test
match=late_evening 0.018256386  2.360377
match=night        0.001940467 -3.099196

$`2`
                     Cla/Mod  Mod/Cla   Global
match=night        76.00000 52.77778 35.21127
match=late_evening 28.57143 16.66667 29.57746
                        p.value    v.test
match=night        0.001940467  3.099196
match=late_evening 0.018256386 -2.360377
```

```
$`1`
                v.test Mean in category
avgwatch      3.320733         23.77143
peakliveview -2.318925         92.57143
reactions    -2.686293         77.42857
minview      -2.945249       5616.31429
comments     -3.095240        176.22857
tensecviews  -3.984514       2136.22857
videoviews   -5.130443       4125.51429
uniqueview   -5.177636       4018.51429
peoplereach  -7.113529      19234.80000
```

```
$`2`
                v.test Mean in category
peoplereach   7.113529      28799.86111
uniqueview    5.177636       5457.38889
videoviews    5.130443       5588.11111
tensecviews   3.984514       2710.61111
comments      3.095240        370.61111
minview       2.945249       6727.33333
reactions     2.686293        105.88889
peakliveview  2.318925        107.97222
avgwatch     -3.320733         20.08333
```

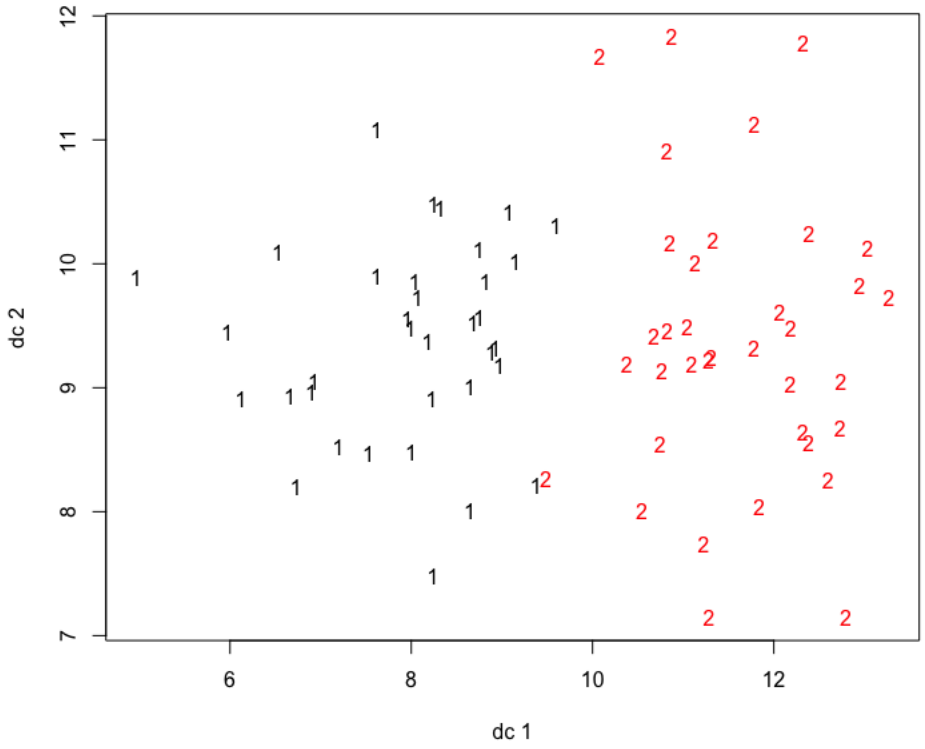
- **1η ομάδα**
  - Κατηγορικές μεταβλητές:
    - Η 1η ομάδα αποτελείται κυρίως από εκπομπές όπου προβλήθηκαν αργά το απόγευμα.
    - Οι υπόλοιπες μεταβλητές δεν φαίνεται να χαρακτηρίζουν την ομάδα.

  - Συνεχείς μεταβλητές:
    - Η 2η ομάδα χαρακτηρίζεται κυρίως από εκπομπές με μέσο χρόνο παρακολούθησης υψηλότερο του μέσου όρου.
    - Οι υπόλοιπες μεταβλητές δεν φαίνεται να χαρακτηρίζουν την ομάδα.

- **2η ομάδα**
  - Κατηγορικές μεταβλητές:
    - Η 2η ομάδα χαρακτηρίζεται κυρίως από εκπομπές όπου προβλήθηκαν τη νύχτα.
    - Οι υπόλοιπες μεταβλητές δεν φαίνεται να χαρακτηρίζουν την ομάδα.

  - Συνεχείς μεταβλητές:
    - Η 2η ομάδα χαρακτηρίζεται κυρίως από εκπομπές με αριθμό μοναδικών θεατών, αριθμό βίντεο που παρακολουθήθηκαν και αριθμό σχολίων μεγαλύτερο του μέσου όρου.

Το παρακάτω διάγραμμα (Διάγραμμα 2) αντιστοιχεί στο παραγοντικό επίπεδο 1 x 2 της εφαρμογής της μεθόδου MDS (multidimensional scaling) στα αρχικά δεδομένα, με αριθμημένα τα υποκείμενα που ανήκουν σε κάθε ομάδα (1 ή 2), σύμφωνα με τα αποτελέσματα της μεθόδου k-prototypes. Από το διάγραμμα διαπιστώνουμε ότι η μέθοδος διαχωρίζει αρκετά καλά τις δύο ομάδες, με μικρότερο βαθμό επικάλυψης σε σχέση με τη μέθοδο Gower/PAM. Ωστόσο, οι δύο λύσεις ομαδοποίησης δεν διαφέρουν σημαντικά.



Διάγραμμα 2. Οι παρατηρήσεις των δύο ομάδων σε χώρο δύο διαστάσεων (παραγοντικό επίπεδο της μεθόδου MDS). Οι αριθμοί 1 και 2 αντιστοιχούν στις δύο ομάδες στις οποίες κατέληξε η μέθοδος K-prototypes.

3.3 K-means for mixed data (ή Mixed K-means)

Συνεχίζουμε με την εφαρμογή της μεθόδου **Mixed K-means**. Θα χρησιμοποιήσουμε το πακέτο ***kmed*** της ***R.*** Μετά την εφαρμογή του αλγορίθμου, καταλήξαμε σε αποτελέσματα παρόμοια με αυτά του της μεθόδου k-prototypes, δηλ:

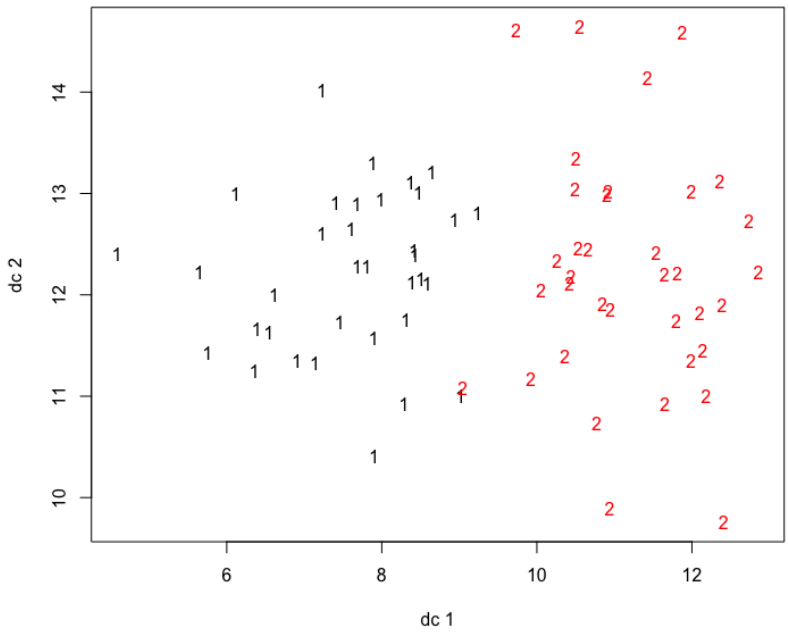
- **1η ομάδα**
  - Κατηγορικές μεταβλητές:
    - Η 1η ομάδα αποτελείται κυρίως από εκπομπές όπου προβλήθηκαν αργά το απόγευμα.
    - Οι υπόλοιπες μεταβλητές δεν φαίνεται να χαρακτηρίζουν την ομάδα.

- Συνεχείς μεταβλητές:
  - Η 2η ομάδα χαρακτηρίζεται κυρίως από εκπομπές με μέσο χρόνο παρακολούθησης υψηλότερο του μέσου όρου.
  - Οι υπόλοιπες μεταβλητές δεν φαίνεται να χαρακτηρίζουν την ομάδα.

❖ **2η ομάδα**

- Κατηγορικές μεταβλητές:
  - Η 2η ομάδα χαρακτηρίζεται κυρίως από εκπομπές όπου προβλήθηκαν τη νύχτα.
  - Οι υπόλοιπες μεταβλητές δεν φαίνεται να χαρακτηρίζουν την ομάδα.

- Συνεχείς μεταβλητές:
  - Η 2η ομάδα χαρακτηρίζεται κυρίως από εκπομπές με αριθμό μοναδικών θεατών, αριθμό βίντεο που παρακολουθήθηκαν και αριθμό σχολίων μεγαλύτερο του μέσου όρου.



Διάγραμμα 3. Οι παρατηρήσεις των δύο ομάδων σε χώρο δύο διαστάσεων (παραγοντικό επίπεδο της μεθόδου MDS). Οι αριθμοί 1 και 2 αντιστοιχούν στις δύο ομάδες στις οποίες κατέληξε η μέθοδος Mixed K-means.

Το Διάγραμμα 3 αντιστοιχεί στο παραγοντικό επίπεδο 1 x 2 της εφαρμογής της μεθόδου MDS (multidimensional scaling) στα αρχικά δεδομένα, με αριθμημένα τα υποκείμενα που ανήκουν σε κάθε ομάδα (1 ή 2), σύμφωνα με τα αποτελέσματα της μεθόδου Mixed K-means. Από το διάγραμμα διαπιστώνουμε ότι η μέθοδος διαχωρίζει αρκετά καλά τις δύο ομάδες, όπως και η μέθοδος K-prototypes.

3.4 Modha-Sprangler convex K-means

Έπειτα θα συνεχίσουμε με τον αλγόριθμο **Modha-Sprangler convex K-means.** Θα χρησιμοποιήσουμε το πακέτο **kamila** της **R**.

```
$`1`
               v.test Mean in category
followers     5.683338        92.068182
comments     -2.512256       213.000000
minview      -2.930857      5746.659091
videoviews   -3.190611      4510.886364
uniqueview   -3.204920      4399.272727
tensecviews  -3.334248      2239.227273
shares       -3.962684         2.727273
peakliveview -3.999633        89.977273
reactions    -5.525668        68.931818

$`1`
                    Cla/Mod    Mod/Cla   Global
hostess=0          97.77778 100.000000 63.38028
weekday=Saturday  100.00000  34.090909 21.12676
weekday=Sunday    100.00000  25.000000 15.49296
weekday=Wednesday  25.00000   4.545455 11.26761
weekday=Tuesday     0.00000   0.000000 15.49296
hostess=1           0.00000   0.000000 36.61972
```

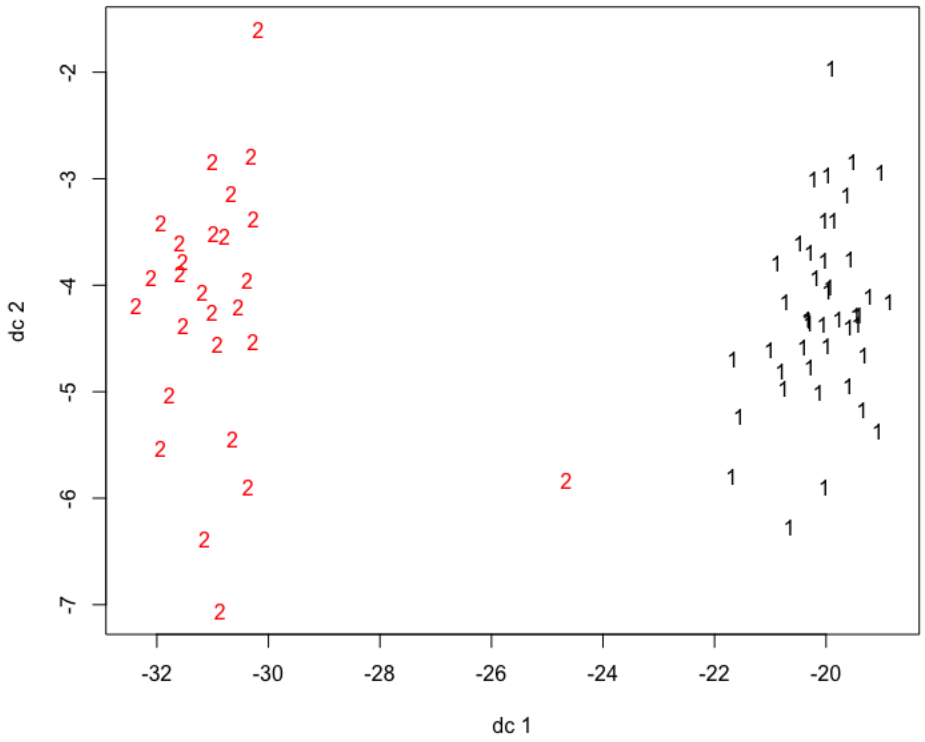
```
$`2`
                v.test Mean in category
reactions     5.525668       129.222222
peakliveview  3.999633       117.333333
shares        3.962684         4.555556
tensecviews   3.334248      2734.222222
uniqueview    3.204920      5316.518519
videoviews    3.190611      5447.629630
minview       2.930857      6885.259259
comments      2.512256       375.481481
followers    -5.683338        86.481481
```

```
$`2`
                    Cla/Mod   Mod/Cla   Global
hostess=1        100.000000 96.296296 36.61972
weekday=Tuesday  100.000000 40.740741 15.49296
weekday=Wednesday 75.000000 22.222222 11.26761
weekday=Sunday     0.000000  0.000000 15.49296
weekday=Saturday   0.000000  0.000000 21.12676
hostess=0          2.222222  3.703704 63.38028
```

- **1η ομάδα (n = 44, 62%)**
  - Κατηγορικές μεταβλητές:
    - Η 1η ομάδα χαρακτηρίζεται κυρίως από εκπομπές όπου δεν συμμετέχει και η γυναίκα παρουσιάστρια,
    - Από εκπομπές που προβλήθηκαν τις ημέρες Σάββατο και Κυριακή,
  - Συνεχείς μεταβλητές:
    - Η 1η ομάδα χαρακτηρίζεται κυρίως από εκπομπές όπου το ποσοστό ακολούθων της σελίδας επί τον αριθμό των θεατών ήταν μεγαλύτερο του μέσου όρου.

- **2η ομάδα (n = 27, 38%)**
  - Κατηγορικές μεταβλητές:
    - Η 2η ομάδα χαρακτηρίζεται κυρίως από εκπομπές όπου κυρίως συμμετέχει η γυναίκα παρουσιάστρια,
    - Από εκπομπές που προβλήθηκαν τις ημέρες Τρίτη και Τετάρτη.
  - Συνεχείς μεταβλητές:

Η 2η ομάδα χαρακτηρίζεται κυρίως από εκπομπές με αριθμό αντιδράσεων, αριθμό θεατών όσο μεταδιδόταν ζωντανά και κοινοποιήσεων μεγαλύτερο του μέσου όρου.



Διάγραμμα 4. Οι παρατηρήσεις των δύο ομάδων σε χώρο δύο διαστάσεων (παραγοντικό επίπεδο της μεθόδου MDS). Οι αριθμοί 1 και 2 αντιστοιχούν στις δύο ομάδες στις οποίες κατέληξε η μέθοδος Modha Spangler K-means.

Το Διάγραμμα 4 αντιστοιχεί στο παραγοντικό επίπεδο 1 x 2 της εφαρμογής της μεθόδου MDS (multidimensional scaling) στα αρχικά δεδομένα, με αριθμημένα τα υποκείμενα που ανήκουν σε κάθε ομάδα (1 ή 2), σύμφωνα με τα αποτελέσματα της μεθόδου Modha-Spangler K-means. Από το διάγραμμα διαπιστώνουμε ότι η μέθοδος διαχωρίζει πολύ καλά τις δύο ομάδες. Ωστόσο, η λύση της ομαδοποίησης δεν διαφέρει σημαντικά από αυτές των προηγούμενων μεθόδων.

3.5 Mixed Reduced K-means

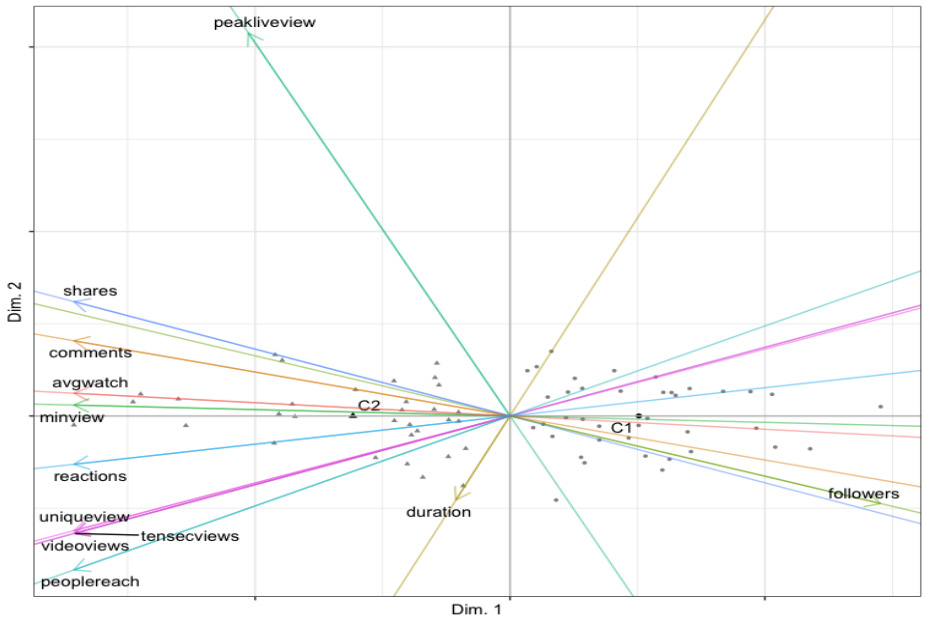
Πριν την εφαρμογή της μεθόδου Mixed Reduced K-means, ο χρήστης καλείται να επιλέξει τον αριθμό των ομάδων και τον αριθμό των διαστάσεων. Για τον σκοπό

αυτό, μπορούμε να βασιστούμε σε κριτήρια αξιολόγησης της εσωτερικής εγκυρότητας της λύσης της ομαδοποίησης, όπως το Average Silhouette Width (ASW). Για τον υπολογισμό του δείκτη σε επίπεδο ομάδας, υπολογίζεται ο μέσος όρος των τιμών του ASW των παρατηρήσεων της ομάδας.

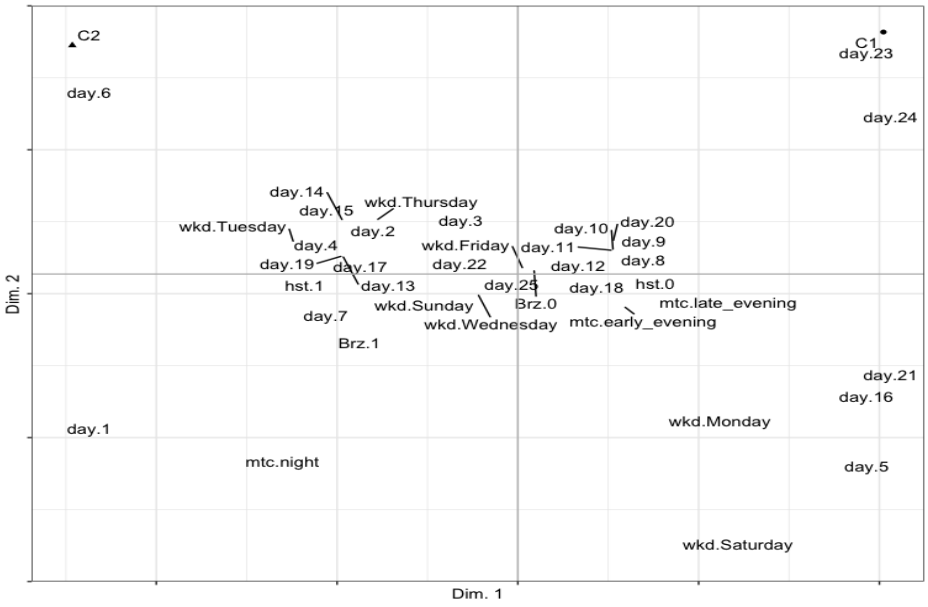
Η μέθοδος Reduced K-means εφαρμόστηκε με τη συνάρτηση cluspcamix() του πακέτου clustrd της R (Markos et al., 2019). Στο ίδιο πακέτο, υπάρχει η συνάρτηση tuneclus(), η οποία χρησιμοποιήθηκε για τον υπολογισμό του δείκτη ASW για λύσεις ομαδοποίησης από 2 έως 6 ομάδες και 1 έως 5 διαστάσεις. Η εφαρμογή της συνάρτησης έδειξε ότι η βέλτιστη λύση ήταν αυτή με τις 2 ομάδες σε 2 διαστάσεις, με τιμές του ASW ίσες με 0,27 και 0,13, αντίστοιχα, και είναι αυτή που περιγράφεται παρακάτω.

Η πρώτη ομάδα περιέχει 39 εκπομπές (54,9%) και η δεύτερη ομάδα τις υπόλοιπες 32 (45,1%). Στα διαγράμματα 1α και 1β παρουσιάζονται τα παραγοντικά επίπεδα 1x2 των συνεχών και των κατηγορικών μεταβλητών αντίστοιχα. Στο διάγραμμα 1α οι συνεχείς μεταβλητές παρουσιάζονται ως διανύσματα, όπου η γωνία μεταξύ δύο διανυσμάτων είναι ενδεικτική της συσχέτισης μεταξύ τους. Διανύσματα με γωνία μικρότερη των 90^{o} δείχνουν θετική συσχέτιση, διανύσματα με γωνία γύρω στις 90^{o} δείχνουν απουσία συσχέτισης και διανύσματα με γωνία μεγαλύτερη των 90^{o} δείχνουν αρνητική συσχέτιση. Επομένως, από το διάγραμμα 1α διαπιστώνουμε ότι οι περισσότερες συνεχείς μεταβλητές του πίνακα δεδομένων συσχετίζονται θετικά μεταξύ τους στα αριστερά του παραγοντικού επιπέδου (shares, comments, avgwatch, minview, reactions, uniqueview, videoview, tenseviews, peoplereach). Αυτές είναι μεταβλητές σχετικές με τον αριθμό των προβολών του κάθε βίντεο και τη συχνότητα αλληλεπίδρασης των χρηστών με τα βίντεο. Αντίθετα, η μεταβλητή followers συσχετίζεται αρνητικά με την προηγούμενη ομάδα μεταβλητών. Οι μεταβλητές duration (διάρκεια βίντεο) και peakliveview (μέγιστος αριθμός θεατών οποιαδήποτε στιγμή κατά τη διάρκεια του βίντεο) είναι αυτές που διαμορφώνουν τον 2^{o} παραγοντικό άξονα και συσχετίζονται λιγότερο με τις υπόλοιπες. Τα βίντεο της ομάδας 2 (C2) βρίσκονται

στα αριστερά του παραγοντικού επιπέδου, δηλαδή έχουν υψηλές τιμές στις μεταβλητές που βρίσκονται στα αριστερά. Αντίθετα, τα βίντεο της ομάδας 1 (C1) βρίσκονται στα δεξιά του παραγοντικού επιπέδου και είναι αυτά με χαμηλές τιμές στις παραπάνω μεταβλητές.



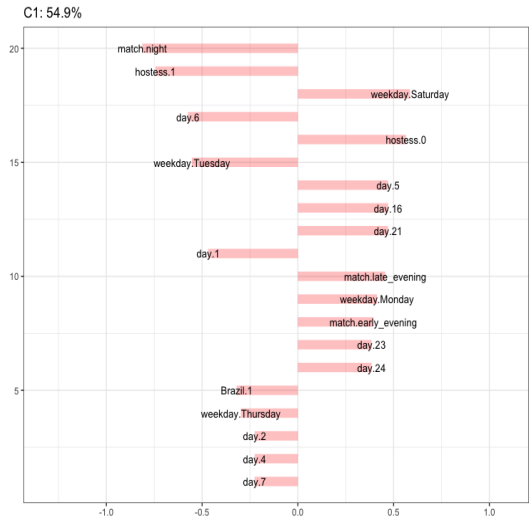
Διάγραμμα 5α. Παραγοντικό επίπεδο των συνεχών μεταβλητών



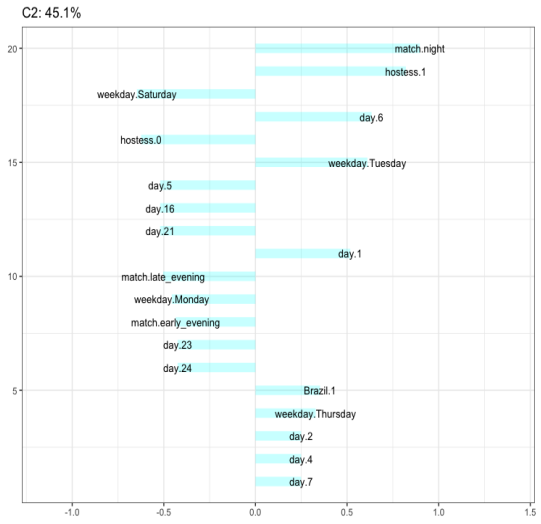
Διάγραμμα 5β. Παραγοντικό επίπεδο των κατηγορικών μεταβλητών

Στο διάγραμμα 5β προβάλλονται οι τιμές των κατηγορικών μεταβλητών. Το διάγραμμα ερμηνεύεται με αντίστοιχο τρόπο με αυτόν του διαγράμματος 5α,

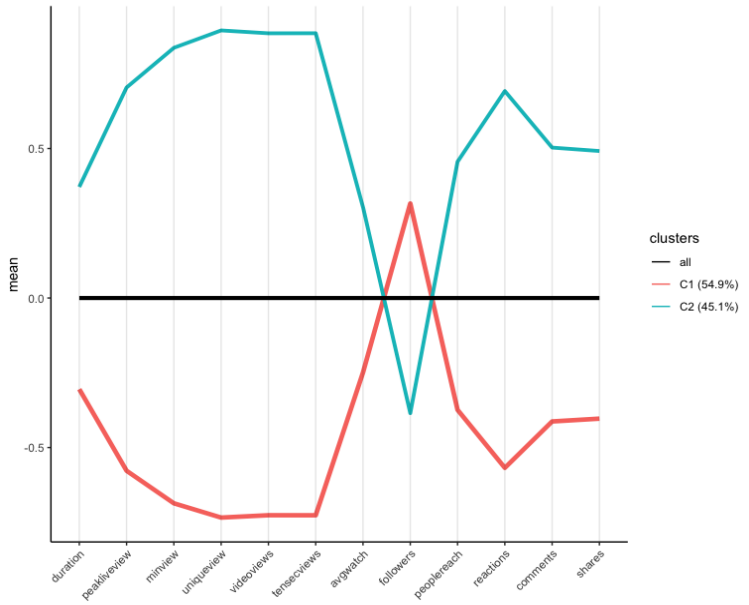
δηλαδή οι τιμές που βρίσκονται στα αριστερά του παραγοντικού επιπέδου χαρακτηρίζουν τα βίντεο της ομάδας C2. Επομένως, τα βίντεο αυτής της ομάδας έχουν προβληθεί συνήθως Τρίτη ή Πέμπτη, είναι σε αυτά παρούσα η κεντρική παρουσιάστρια, πραγματοποιήθηκαν βραδινές ώρες και ανάμεσα σε αυτά βρίσκονται βίντεο που αφορούν σε αγώνες της Βραζιλίας. Αντίθετα, στα δεξιά του παραγοντικού επιπέδου βρίσκονται οι τιμές που χαρακτηρίζουν τα βίντεο της ομάδας C1, με τα αντίθετα χαρακτηριστικά από αυτά της C2 (απουσία της κεντρικής παρουσιάστριας, αγώνες που πραγματοποιήθηκαν απογευματινές ώρες, αγώνες που δεν παίζει η Βραζιλία). Τα αποτελέσματα αυτά, επιβεβαιώνονται και από τα διαγράμματα 2α και 2β, δηλαδή ένα διάγραμμα παράλληλων συντεταγμένων (parallel coordinates plot) και τρία διαγράμματα με τα τυποποιημένα υπόλοιπα.



Διάγραμμα 6α. Τυποποιημένα υπόλοιπα ομάδας C1

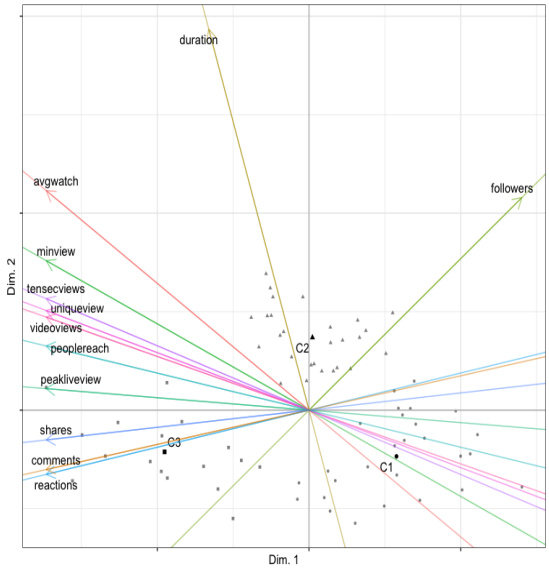


Διάγραμμα 6β. Τυποποιημένα υπόλοιπα ομάδας C2

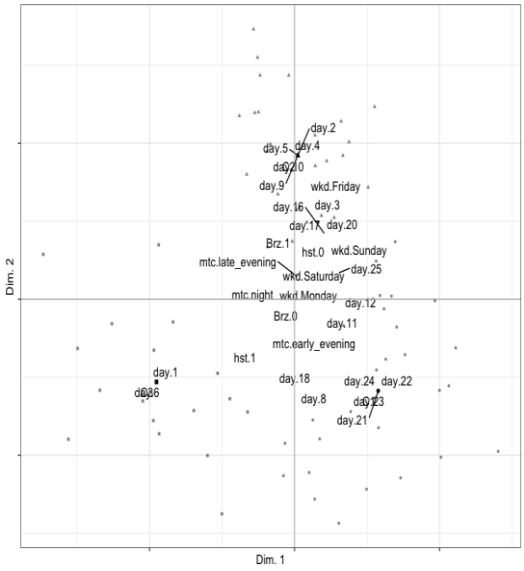


Διάγραμμα 6γ. Διάγραμμα παράλληλων συντεταγμένων

Αποτελέσματα με 3 ομάδες



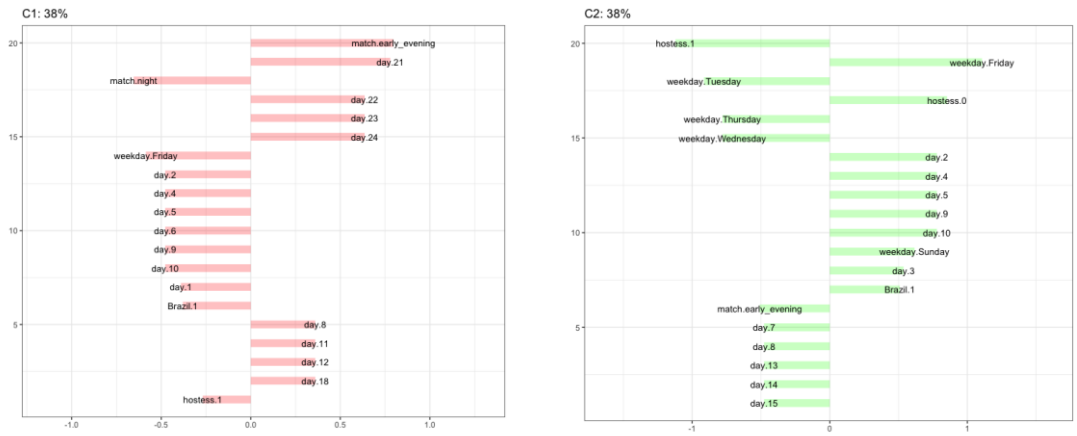
Διάγραμμα 7α. Παραγοντικό επίπεδο των συνεχών μεταβλητών

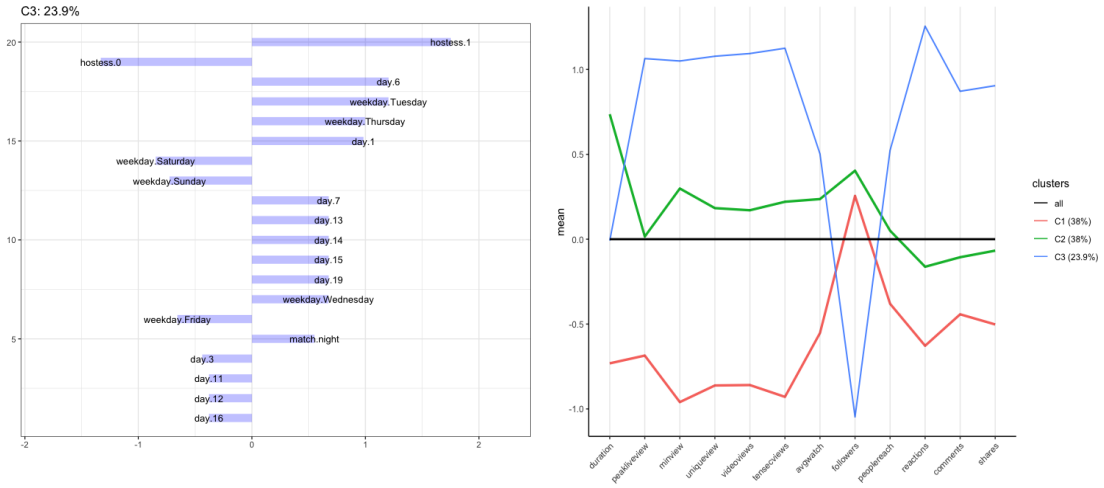


Διάγραμμα 7β. Παραγοντικό επίπεδο των κατηγορικών μεταβλητών

Η πρώτη ομάδα περιέχει 27 εκπομπές (38%), η δεύτερη ομάδα άλλες 27 (38%), ενώ η τρίτη περιέχει 17 (24%). Στα διαγράμματα 7α και 7β παρουσιάζονται τα παραγοντικά επίπεδα 1x2 των συνεχών και των κατηγορικών μεταβλητών αντίστοιχα. Στο διάγραμμα 7α οι συνεχείς μεταβλητές παρουσιάζονται ως

διανύσματα, όπου η γωνία μεταξύ δύο διανυσμάτων είναι ενδεικτική της συσχέτισης μεταξύ τους. Διανύσματα με γωνία μικρότερη των 90^{o} δείχνουν θετική συσχέτιση, διανύσματα με γωνία γύρω στις 90^{o} δείχνουν απουσία συσχέτισης και διανύσματα με γωνία μεγαλύτερη των 90^{o} δείχνουν αρνητική συσχέτιση. Επομένως, από το διάγραμμα 7α διαπιστώνουμε ότι οι περισσότερες συνεχείς μεταβλητές του πίνακα δεδομένων συσχετίζονται θετικά μεταξύ τους στα αριστερά του παραγοντικού επιπέδου (shares, comments, avgwatch, minview, reactions, uniqueview, videoview, tenseviews, peoplereach). Αυτές είναι μεταβλητές σχετικές με τον αριθμό των προβολών του κάθε βίντεο και τη συχνότητα αλληλεπίδρασης των χρηστών με τα βίντεο. Αντίθετα, η μεταβλητή followers συσχετίζεται αρνητικά με την προηγούμενη ομάδα μεταβλητών. Οι μεταβλητές duration (διάρκεια βίντεο) και peakliveview (μέγιστος αριθμός θεατών οποιαδήποτε στιγμή κατά τη διάρκεια του βίντεο) είναι αυτές που διαμορφώνουν τον 2^{o} παραγοντικό άξονα και συσχετίζονται λιγότερο με τις υπόλοιπες. Τα βίντεο της ομάδας 2 (C2) βρίσκονται στα αριστερά του παραγοντικού επιπέδου, δηλαδή έχουν υψηλές τιμές στις μεταβλητές που βρίσκονται στα αριστερά. Αντίθετα, τα βίντεο της ομάδας 1 (C1) βρίσκονται στα δεξιά του παραγοντικού επιπέδου και είναι αυτά με χαμηλές τιμές στις παραπάνω μεταβλητές.





Διάγραμμα 8. Τυποποιημένα υπόλοιπα των τριών ομάδων και διάγραμμα παράλληλων συντεταγμένων

4. Συμπεράσματα

Στην παρούσα εργασία παρουσιάστηκαν πέντε μέθοδοι ανάλυσης συστάδων για δεδομένα μεικτού τύπου. Αφού παρουσιάστηκε το μαθηματικό υπόβαθρο κάθε μεθόδου, οι μέθοδοι εφαρμόστηκαν σε ένα πραγματικό σύνολο δεδομένων ιστού. Από την μελέτη των αποτελεσμάτων κάθε μεθόδου, προέκυψε ως πιο αποτελεσματική, η μέθοδος Mixed Reduced K-means και αμέσως μετά η Modha-Spangler K-means. Αξίζει να σημειωθεί ότι με τον όρο «πιο αποτελεσματική» θεωρούμε τη μέθοδο που κατέληξε στην πιο ενδιαφέρουσα ερμηνεία των ομάδων. Μια πιο αμερόληπτη και ενδελεχής σύγκριση μεταξύ των μεθόδων μπορεί να επιτευχθεί μέσω της εφαρμογής τους σε προσομοιωμένα σύνολα δεδομένων.

Βιβλιογραφία

Ahmad, A., & Dey, L. (2007). A k-mean clustering algorithm for mixed numeric and categorical data. *Data & Knowledge Engineering*, *63*(2), 503-527

De Soete, G., & Carroll, J. D. (1994). K-means clustering in a low-dimensional Euclidean space. E. DIDAY, Y. LECHEVALLIER, M. SCHADER, P. BERTRAND & B. BURTSCHY (Eds.), *New approaches in classification and data analysis*, Berlin, Springer-Verlag, 212-219.

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

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data mining and knowledge discovery*, *2*(3), 283-304.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*(1), 193-218.

Iodice D'Enza, A., Markos, A. & van de Velden, M. Package 'clustrd' (2016).

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning (*Vol. 112, p. 18). New York: springer.

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons.

Lê, S., Josse, J., & Husson, F. (2008). FactoMineR: an R package for multivariate analysis. *Journal of Statistical Software*, *25*, 1-18.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K., & Studer, M. (2013). Package 'cluster'.

Markos, A., & Moschidis, O., & Chadjipadelis, T. (2020). Sequential dimension reduction and clustering of mixed-type data. *International Journal of Data Analysis Techniques and Strategies, 12*, 28-30

Modha, D. S., & Spangler, W. S. (2003). Feature weighting in k-means clustering. *Machine learning*, *52*(3)

Pagès, J. (2004) Analyse Factorielle de Données Mixtes. *Revue de Statistique Appliquée, 52,* 93-111

Pagès, J. *Multiple factor analysis by example using R*. CRC Press, 2014.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53-65

Szepannek, G. (2018). clustMixType: User-Friendly Clustering of Mixed-Type Data in R. *R Journal*, *10*(2), 200

Van Buuren, S., & Heiser, W. J. (1989). Clusteringn objects into k groups under optimal scaling of variables. *Psychometrika, 54*(4), 699-706

van de Velden, M. Iodice D'Enza, A. and Palumbo, F. (2017). Cluster correspondence analysis. *Psychometrika, 82*(1), 158–185.

van de Velden, M., Iodice D'Enza, A., & Markos, A. (2019). Distance-based clustering of mixed data. Wiley Interdisciplinary Reviews: Computational Statistics, 11(3), e1456

Vichi, M., & Kiers, H. A. (2001). Factorial k-means analysis for two-way data. *Computational Statistics & Data Analysis*, *37*(1), 49-64

Παράρτημα Α – Κώδικας R

```
require(FactoMineR)
load("telecom.Rdata")

## Apply Clustering to Mixed-Type Data
install.packages("mclust")
require(mclust)
install.packages("cluster")
require(cluster)
install.packages("clustMD")
require(clustMD)
install.packages("clustMixType")
require(clustMixType)
install.packages("kmed")
require(kmed)
install.packages("kamila")
require(kamila)
install.packages("FactoMineR")
require(FactoMineR)
require(fpc)

############# CLUSTERING METHODS ####################

#1. Gower's distance + PAM (Partitioning Around Medoids)

#' Compute Gower distance
gower_dist <- daisy(telecom, metric = "gower")
gower_mat <- as.matrix(gower_dist)
#' Print most similar objects
#telecom[which(gower_mat == min(gower_mat[gower_mat !=
min(gower_mat)]), arr.ind = TRUE)[1, ], ]
```

```
pam_fit <- pam(gower_dist, diss = TRUE, k = 2)
sil_width <- pam_fit$silinfo$avg.width
sil_width

table(pam_fit$clustering)

#Ερμηνεία των ομάδων
catdes(cbind(telecom,as.factor(pam_fit$clustering)),39
)

#2. K-prototypes
outk = kproto(telecom, 2)
plotcluster(data.matrix(telecom),outk$cluster)

#Ερμηνεία των ομάδων
catdes(cbind(telecom,as.factor(outk$cluster)),39)

# 3. Mixed K-means
# Distances for mixed variables data set
mix <- distmix(telecom, method = "ahmad", idcat =
c(3,5,6,7,10,11,16,31,32,33,34,38), idnum =
c(1,2,4,8,9,12:15,17:30,35:37))
kmedres <- fastkmed(mix, 2, iterate = 20, init = NULL)

#Ερμηνεία των ομάδων
catdes(cbind(telecom,as.factor(kmedres$cluster)),39)


# 4. Modha-Sprangler K-means
```

```
conDf <-
data.frame(scale(telecom[,c(1,2,4,8,9,12:15,17:30,35:3
7)]))
catDf <-
dummyCodeFactorDf(data.frame(telecom[,c(3,5,6,7,10,11,
16,31,32,33,34,38)]))
#Modha-Sprangler
msRes <- gmsClust(conDf, catDf, nclust = 2)

#Ερμηνεία των ομάδων
catdes(cbind(telecom,as.factor(msRes$results$cluster))
,39)

# 5. FAMD + K-means (Two-step)
howmany <- estim_ncp(data.matrix(telecom))$ncp
outpcamix <- FAMD(telecom, ncp = howmany)
outkm <- kmeans(outpcamix$ind$coord, 2, nstart = 100)

#Ερμηνεία των ομάδων
catdes(cbind(telecom,as.factor(outkm$cluster)),39)

# 6. Mixed Reduced K-means
install.packages("clustrd")
require(clustrd)

outc = cluspcamix(telecom, nclus = 2, ndim = 2, nstart
= 10)
plotcluster(data.matrix(telecom),outc$cluster)

#Ερμηνεία των ομάδων
catdes(cbind(telecom,as.factor(outc$cluster)),40)
```