Η ΜΕΘΟΔΟΣ KARAP ΩΣ ΕΡΓΑΛΕΙΟ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

Δρ. Δημήτριος Καραπιστόλης
τ. καθηγητής Στατιστικής-Ανάλυσης Δεδομένων
στο Αλεξάνδρειο Τεχνολογικό Ίδρυμα Θεσσαλονίκης

Περίληψη

Σε μία έρευνα αγοράς η μέθοδος KARAP χρησιμεύει κυρίως στη σύνδεση κάθε ερωτώμενου με ένα συγκεκριμένο κριτήριο από τα p κριτήρια που τέθηκαν στο ερωτηματολόγιο. Η σύνδεση αυτή πραγματοποιείται με τη χρήση της Παραγοντικής Ανάλυσης Δεδομένων και τον Ευκλείδειο διανυσματικό χώρο $R^{(p-1)}$ που δημιουργείται από τους p-1 παραγοντικούς άξονες.

Γενικά

Μία ανιούσα ιεραρχική ταξινόμηση των n «υποκειμένων» (γραμμές) ενός πίνακα δεδομένων T(n,p) είναι μία διαδικασία που παράγει μια ακολουθία διαμελισμών του αρχικού συνόλου σε υποσύνολα μη κενά και ξένα ανά δύο μεταξύ τους, τις λεγόμενες **κλάσεις**, τη μία μέσα στην άλλη, συνενώνοντας κάθε φορά δύο μόνο κλάσεις οι οποίες βάσει κάποιας μετρικής παρουσιάζουν σε κάθε βήμα ομαδοποίησης την μικρότερη απόσταση.

Απ' ότι γίνεται αντιληπτό στόχος της ανιούσας ιεραρχικής ταξινόμησης είναι να ομαδοποιήσει το σύνολο των στατιστικών μονάδων ενός πληθυσμού σ' ένα περιορισμένο πλήθος ομοιογενών κλάσεων, τις επονομαζόμενες **«συστάδες»** ως προς την συμπεριφορά ορισμένων μεταβλητών, λαμβάνοντας υπόψη το σύνολο των μεταβλητών, ώστε κάθε μία να διαφέρει από τις άλλες, όσο το δυνατόν περισσότερο.

Συγκεκριμένα στη ταξινόμηση με την διαδικασία VACOR οι κλάσεις δημιουργούνται βάσει ενός αντικειμενικού αλγορίθμου (αλγόριθμος του Ward), πέρα από τις υποκειμενικές μεθόδους που μπορεί να αναπτύξει κάθε ερευνητής. Λέμε αντικειμενικό αλγόριθμο γιατί η ομαδοποίηση των στατιστικών μονάδων γίνεται χωρίς καμιά a priori υπόθεση στον αρχικό πίνακα δεδομένων βάσει της μετρικής x^2.

Στις κλάσεις που δημιουργεί η μέθοδος VACOR εντοπίζεται το πλήθος των μεταβλητών που τις χαρακτηρίζουν. Αυτό γίνεται με την βοήθεια που παρέχει το πρόγραμμα MAD με τον πίνακα «Συμβολή των μεταβλητών στο χαρακτηρισμό των κλάσεων» σε συνδυασμό με τον πίνακα «Συμβολή των μεταβλητών στη διάσπαση των k υψηλότερων κόμβων». Κατά συνέπεια τα «υποκείμενα» που συμμετέχουν στη διαμόρφωση των κλάσεων συνδέονται επίσης με τις μεταβλητές που χαρακτηρίζουν κάθε κλάση.

Είναι γνωστό ότι χρησιμοποιώντας την μέθοδο VACOR με τα γνωστά προγράμματα που υλοποιούν την Ανιούσα Ιεραρχική Ταξινόμηση, δεν είναι εφικτή η σύνδεση κάθε «υποκειμένου» με μια συγκεκριμένη μεταβλητή, εκτός και αν χρησιμοποιηθεί ο στατιστικός έλεγχος της διαφοράς των αναλογιών σε επίπεδο σημαντικότητας 5%, μεταξύ της αναλογίας P_δ κάθε μιας από τις p τιμές που αντιστοιχεί στο «δείγμα» που αναφέρεται σε κάθε «υποκείμενο» και της αναλογίας P_M κάθε μεταβλητής ως προς το σύνολο των p μεταβλητών (Morineau A. 1984).

$$\begin{aligned} H_0 &: P_\delta - P_M = 0 \\ H_1 &: P_\delta - P_M > 0 \end{aligned} \qquad (1)$$

Στη συνέχεια υπολογίζεται η τιμή του z

$$z = \frac{P_\delta - P_M}{s_p} \qquad (2)$$

Όπου $s_p = \sqrt{\frac{P_M \cdot Q_M}{n}}$ $\quad \mu\varepsilon \quad$ $Q_M = 1 - P_M$

Όπως αναφέρει ο καθηγητής Θ. Μπεχράκης «Με βάση την τιμή του ελέγχου κατατάσσουμε τα προβλήματα για κάθε ομάδα. Η τιμή του ελέγχου αποτελεί κριτήριο που χρησιμοποιείται για την επιλογή των προβλημάτων τα οποία χαρακτηρίζουν την κάθε ομάδα. Όσο μεγαλύτερη είναι η τιμή του ελέγχου για μια συγκεκριμένη ομάδα και ένα συγκεκριμένο πρόβλημα , τόσο περισσότερο χαρακτηριστικό είναι το πρόβλημα αυτό για τη συγκεκριμένη ομάδα» (Μπεχράκης Θ. σελ. 74)

Όταν, λοιπόν, από την σχέση (2) προκύπτει η τιμή $z>1,645$ τότε ισχύει η εναλλακτική υπόθεση H_1, δηλαδή η μεταβλητή M με αναλογία P_M χαρακτηρίζει έντονα την κλάση ή το υποκείμενο με αναλογία P_{δ}. Σε διαφορετικές τιμές του z όπου $-1,645<z<1,645$ απλά η μεταβλητή παρουσιάζει μέτρια σύνδεση (θετική ή αρνητική) ανάλογα με την τιμή που παρουσιάζει), ενώ όταν προκύπτει τιμή $z<-1,645$ τότε η απουσία εξάρτησης της μεταβλητής με την κλάση (ή το αντικείμενο) θεωρείται έντονη.

Από την άλλη η σύνδεση των υποκειμένων με κάθε μεταβλητή μπορεί να πραγματοποιηθεί εφ' όσον αναλυθεί ο πίνακας δεδομένων T(n,p) με την Παραγοντική Ανάλυση των Αντιστοιχιών (-Π.Α.Α-), εξαχθούν οι συντεταγμένες των p μεταβλητών και των n αντικειμένων πάνω στους p-1 παραγοντικούς άξονες και στη συνέχεια να υπολογιστούν οι αποστάσεις κάθε αντικειμένου A από κάθε μεταβλητής B χρησιμοποιώντας την παρακάτω σχέση η οποία υπολογίζει την απόσταση μεταξύ δύο διανυσμάτων του διανυσματικού χώρου R^n (Serge Lang σελ. 16)

$$\left\| A - B \right\| = \sqrt{(A - B) \cdot (A - B)} \quad (3)$$

Αρχικά θα συγκριθούν τα αποτελέσματα της σύνδεσης των «υποκειμένων» με τις μεταβλητές που προέκυψαν αφενός μετά την ταξινόμηση με την μέθοδο VACOR, χρησιμοποιώντας τον προαναφερόμενο έλεγχο με την z κατανομή, αφετέρου με την τοποθέτηση όλων των σημείων (γραμμών και στηλών) του πίνακα δεδομένων T(n,p) στον Ευκλείδειο διανυσματικό χώρο $R^{(p-1)}$, βάσει των **παραγόντων** που προκύπτουν μετά την εφαρμογή της Παραγοντικής Ανάλυσης των Αντιστοιχιών. Η εργασία αυτή θα αποτελέσει επίσης μια νέα διαδικασία ταξινόμησης των «υποκειμένων».

Η επιλογή του Ευκλείδειου χώρου R^{p-1} έγινε αφού ως γνωστό οι παραγοντικοί άξονες δημιουργούν μία ορθοκανονική βάση στον χώρο $R^{(p-1)}$, όπου βάσει των συντεταγμένων τοποθετούνται οι p μεταβλητές και οι n γραμμές του πίνακα δεδομένων στις πραγματικές τους θέσεις, απ' όπου αντλείται το σύνολο της πληροφόρησης που παρέχει ο πίνακας δεδομένων που αναλύεται.

Στο σημείο αυτό θα παραθέσουμε ένα παράδειγμα προς επιβεβαίωση των προαναφερόμενων προτάσεων.

Σχέσεις μεταξύ παραγόντων

Με βάση ένα πίνακα σύμπτωσης δύο ποιοτικών μεταβλητών A και B δημιουργείται ο αντίστοιχος πίνακας σχετικών συχνοτήτων.

Πίνακας 1:Πίνακας σχετικών συχνοτήτων

A \ B	b_1 b_2 ... ... b_j ..... b_p	
a_1 a_2 . a_i . . a_n	..... f_{ij}	f_i
	f_j	

Η παραγοντική ανάλυση των αντιστοιχιών επιτρέπει όχι μόνο τη γεωμετρική και αλγεβρική διαπίστωση της απόκλισης από την κατάσταση της ανεξαρτησίας των δύο ποιοτικών μεταβλητών X και Y, αλλά και τη διερεύνηση των ομοιοτήτων που παρουσιάζουν μεταξύ τους οι σχετικές κατανομές (προφίλ) των γραμμών ή των στηλών του πίνακα, που αντιστοιχούν στο σύνολο των διαβαθμίσεων των δύο μεταβλητών X και Y.

Στη συνέχεια δημιουργείται ο πίνακας των προφίλ των γραμμών

Πίνακας 2: Πίνακας f_J^I των προφίλ των γραμμών

Διαβαθμίσεις	b_1……..b_j……b_p	
a_1	.	
.	.	
a_i	.....$f^i_j = f_{ij}/f_i$	1
.	.	
a_n	.	

Για να γίνει κατανοητό για πιο λόγο είναι προτιμότερο να χρησιμοποιείται το προφίλ μιας γραμμής ενός πίνακα, ως διανυσματική έκφραση της αντίστοιχης στατιστικής μονάδας i, παρά η γραμμή με τα αρχικά δεδομένα, η απάντηση έχει ως εξής: Εφόσον δύο γραμμές είναι ανάλογες μεταξύ τους, τα προφίλ τους θα είναι ταυτόσημα και όταν παρασταθούν σ' ένα γράφημα οι γραφικές παραστάσεις των διανυσμάτων στα οποία αντιστοιχούν θα συμπέσουν, ενώ αντιθέτως οι γραμμές με τα αρχικά δεδομένα θα παριστάνουν δύο συγγραμμικά διανύσματα.

Αυτή η διαπίστωση είναι πολύ σημαντική καθόσον στην Παραγοντική Ανάλυση των Αντιστοιχιών (-Π.Α.Α-) το ενδιαφέρον εστιάζεται στις αναλογίες των «υποκειμένων» μέσα στις διαβαθμίσεις των «μεταβλητών».

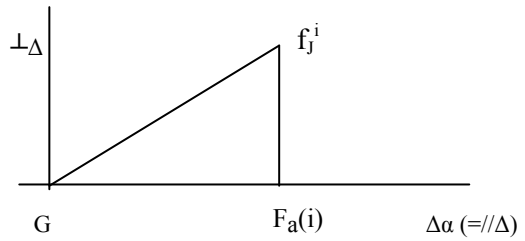
Οι προβολές των σημείων f_J^i του νέφους $N(I)_J$ των γραμμών του πίνακα δεδομένων πάνω στους παραγοντικούς άξονες Δ_a (a=1,...p-1),οι οποίες συμβολίζονται με $F_a(i)$ (όπου $i=f_j^i$ τυχόν προφίλ γραμμή), αποτελούν τις συντεταγμένες των σημείων αυτών πάνω στους άξονες Δ_a. Κάθε συντεταγμένη $F_a(i)$ σχετική με τον παραγοντικό άξονα a ονομάζεται **παράγοντας a του προφίλ i** (J-P & F. Benzecri 1980 σελ. 65).

Για να προσδιορίσουμε τους παραγοντικούς άξονες Δ_a σ' ένα επίπεδο (a=1,2) χρησιμοποιούμε το θεώρημα του Huyghens, το οποίο αναφέρει ότι η ολική αδράνεια $I_{O\lambda}$ του νέφους N(I) μπορεί να αναλυθεί σε δύο μέρη. Το ένα μέρος αφορά στην αδράνεια $I_{//\Delta}$ κατά μήκος μιας ευθείας Δ_a η οποία περνά από το βαρύκεντρο $G\{=f_J\}$ του νέφους και το άλλο στην αδράνεια $I\perp_\Delta$ κάθετα της ευθείας Δ_α.

Από τις άπειρες ευθείες που διέρχονται από το σημείο G, λαμβάνεται εκείνη η οποία καθιστά την αδράνεια $I_{//\Delta}$ μέγιστη και συνεπώς ελάχιστη την αδράνεια $I\perp_\Delta$. Ήτοι

$$I_{O\lambda} = I_{//\Delta} + I\perp_\Delta \qquad (4)$$

Σχηματικά έχουμε:



σχήμα 1: *Διάσπαση της ολικής αδράνειας*

Ως γνωστόν σε κάθε χαρακτηριστική ρίζα λ_a ενός τετραγωνικού πίνακα

$S_J^J = f_J^I \circ f_I^J$ αντιστοιχεί ένα χαρακτηριστικό διάνυσμα $u_a=\varphi_a^J$ που συνδέεται με το παραγοντικό άξονα Δ_a.

Για κάθε διάνυσμα φ_a^J (j=1,...p) οι συντεταγμένες του ικανοποιούν τις παρακάτω σχέσεις:

$$\sum_{j=1}^{p} f_j \cdot \varphi_a^J = 0 \qquad (5)$$

$$\sum_{j=1}^{p} f_j \cdot (\varphi_a^J)^2 = 1 \qquad (6)$$

Κάθε παράγοντας $F_a(i)$, ο οποίος είναι διάνυσμα, υπολογίζεται με την παρακάτω σχέση

$$F_a(i) = \sum_{j=1}^{p} f_j^i \cdot \varphi(j) \qquad (7)$$

Σε κάθε παραγοντικό άξονα Δ_a ισχύουν οι εξής σχέσεις:

$$\sum_{i=1}^{n} f_i \cdot F_a(i) = 0 \qquad (8)$$

$$\sum_{i=1}^{n} f_i \cdot (F_a(i))^2 = \lambda_a \qquad (9)$$

ενώ για δύο διαφορετικούς παραγοντικούς άξονες Δ_r και Δ_s ισχύει:

$$\sum_{i=1}^{n} f_i \cdot F_r(i) F_s(i) = 0 \qquad (10)$$

Ο αριθμός $F_a(i)$ σε απόλυτη τιμή μετρά την απόσταση που χωρίζει το κέντρο βάρους $G=\{f_J\}$ του νέφους $N(I)_J$ από τη προβολή του προφίλ f_{ij} (που παριστάνει η γραμμή i του πίνακα δεδομένων) πάνω στον άξονα Δ_a.

Γενικώς ισχύει

$$d^2\left(f_J^{\,i},\ G\right) = \sum_{a=1}^{p-1} F_a^2(i) \qquad (11)$$

Επομένως η απόσταση που χωρίζει το κέντρο βάρους $G=\{f_J\}$ από την προβολή του προφίλ γραμμή f_j^i λ.χ στο παραγοντικό επίπεδο $\Delta_1 x \Delta_2$ είναι η υποτείνουσα του ορθογωνίου τριγώνου με πλευρές τα $F_1(i)$ και $F_2(i)$. Δηλαδή για το παραγοντικό επίπεδο 1x2 ισχύει η σχέση

$$d^2\left(f_J^{\,i},\ G\right) = F_1^2(i) + F_2^2(i) \qquad (12)$$

Η απόσταση $d^2(f_J^i, f_J)$ υπολογίζεται επίσης με τη χρήση της παρακάτω σχέσης

$$d^2\left(f_J^{\,i},\ G\right) = \sum_{j=1}^{p} \frac{1}{f_i} (f_j^i - f_J)^2 \qquad (13)$$

Η σχέση 12 υποδεικνύει ότι οι παραγοντικοί άξονες πάνω στους οποίους υπολογίζονται οι παράγοντες $F_a(i)$ είναι ορθογώνιοι, οπότε το σύστημα των p-1 παραγοντικών αξόνων αποτελεί εκ κατασκευής ένα ορθοκανονικό σύστημα συντεταγμένων στον διανυσματικό χώρο R^{p-1}. Στη συνέχεια θα παραθέσουμε ένα αριθμητικό παράδειγμα προς επιβεβαίωση

της σχέσης 13.

Αριθμητικό παράδειγμα

Χρησιμοποιώντας ένα απλό αριθμητικό παράδειγμα διαπιστώνει κανείς εύκολα την ισχύει των σχέσεων 12 και 13.
Δίνεται ο παρακάτω πίνακας σύμπτωσης

Πίνακας 3:Πίνακας δεδομένων

	J_1	J_2	J_3	
I_1	0	1	0	1
I_2	1	0	1	2
I_3	1	1	0	2
I_4	0	0	1	1
	2	**2**	**2**	**6**

Βρίσκουμε καταρχήν τα προφίλ των γραμμών f_J^I και των στηλών f_I^J .Ήτοι

$$f_J^I = \begin{pmatrix} 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \\ 1/2 & 1/2 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{και} \quad f_I^J = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1/2 & 0 \\ 0 & 1/2 & 0 & 1/2 \end{pmatrix}$$

Τα προφίλ f_I^J προέκυψαν αφού προηγουμένως πήραμε τον ανάστροφο πίνακα του T(4,3)

Για την βασική εφαρμογή της Π.Α.Α απαιτείται η εύρεση του συμμετρικού τετραγωνικού πίνακας S_J^J ο οποίος υπολογίζεται με το παρακάτω γινόμενο

$$S_J^J = f_J^I \circ f_I^J = \frac{1}{2}\begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix} \cdot \frac{1}{2}\begin{pmatrix} 0 & 2 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix} = \frac{1}{4} \cdot \begin{pmatrix} 2 & 1 & 1 \\ 1 & 3 & 0 \\ 1 & 0 & 3 \end{pmatrix}$$

Στη συνέχει ακολουθεί η εύρεση των τριών χαρακτηριστικών ριζών του τετραγωνικού πίνακα S_J^J οι οποίες είναι οι εξής : $\lambda_0 = \frac{4}{4} = 1$ $\lambda_1 = \frac{3}{4}$ $\lambda_2 = \frac{1}{4}$

Ως γνωστόν σε κάθε χαρακτηριστική ρίζα λ_i (εκτός της τερτιμένης ρίζας λ_0) αντιστοιχεί ένα χαρακτηριστικό διάνυσμα $u_a=\varphi_a^J$ που συνδέεται με το παραγοντικό άξονα Δ_a.

Για κάθε χαρακτηριστικό διάνυσμα φ_a^J (j=1,...p) οι συντεταγμένες του ικανοποιούν όπως προαναφέραμε τις σχέσεις 5 και 6

Οι τιμές των χαρακτηριστικών διανυσμάτων παρουσιάζονται στις σχέσεις 14 και 15

$$\varphi_{11}=0 \text{ , } \varphi_{12}=\frac{\sqrt{6}}{2} \text{ και } \varphi_{13}= -\frac{\sqrt{6}}{2} \qquad (14)$$

$$\varphi_{21}=-\sqrt{2} \quad \varphi_{22}= \frac{\sqrt{2}}{2} \text{ και } \varphi_{23}=\frac{\sqrt{2}}{2} \qquad (15)$$

Χρησιμοποιώντας τη σχέση 7 οι παράγοντες F_a^I έχουν τις παρακάτω τιμές.

$$F_a^I=\begin{pmatrix}0 & 1 & 0\\ 1/2 & 0 & 1/2\\ 1/2 & 1/2 & 0\\ 0 & 0 & 1\end{pmatrix}\cdot\begin{pmatrix}0 & -\sqrt{2}\\ \frac{\sqrt{6}}{2} & \frac{\sqrt{2}}{2}\\ -\frac{\sqrt{6}}{2} & \frac{\sqrt{2}}{2}\end{pmatrix}=\begin{pmatrix}\frac{\sqrt{6}}{2} & \frac{\sqrt{2}}{2}\\ -\frac{\sqrt{6}}{4} & -\frac{\sqrt{2}}{4}\\ \frac{\sqrt{6}}{4} & -\frac{\sqrt{2}}{4}\\ -\frac{\sqrt{6}}{2} & \frac{\sqrt{2}}{2}\end{pmatrix}$$

Διαπιστώνουμε ότι επαληθεύεται η σχέση 9

$$\lambda_1=\frac{1}{6}\cdot\left(-\frac{\sqrt{6}}{2}\right)^2+\frac{2}{6}\cdot\left(\frac{\sqrt{6}}{4}\right)^2+\frac{2}{6}\cdot\left(-\frac{\sqrt{6}}{4}\right)^2+\frac{1}{6}\cdot\left(\frac{\sqrt{6}}{2}\right)^2=\frac{3}{4}$$

$$\lambda_2=\frac{1}{6}\cdot\left(\frac{\sqrt{2}}{2}\right)^2+\frac{2}{6}\cdot\left(-\frac{\sqrt{2}}{4}\right)^2+\frac{2}{6}\cdot\left(-\frac{\sqrt{2}}{4}\right)^2+\frac{1}{6}\cdot\left(\frac{\sqrt{2}}{2}\right)^2=\frac{1}{4}$$

Οι παράγοντες των μεταβλητών G_a^J βρίσκονται χρησιμοποιώντας την σχέση

$$G_a^J=\frac{1}{\sqrt{\lambda_\alpha}}F_a^J\circ f_I^J \qquad (16)$$

Απ' όπου προκύπτει

$$G_a^J=\begin{pmatrix}0 & -\frac{\sqrt{2}}{2}\\ \frac{3\sqrt{2}}{4} & \frac{\sqrt{2}}{4}\\ -\frac{3\sqrt{2}}{4} & \frac{\sqrt{2}}{4}\end{pmatrix}$$

Για τις τέσσερις γραμμές i_1,i_2,i_3 και i_4 οι τιμές των συντεταγμένων στους δύο παραγοντικούς άξονες Δ_1 και Δ_2 έχουν ως εξής:

$$F_1(i_1)=\frac{\sqrt{6}}{2},F_1(i_2)=-\frac{\sqrt{6}}{4},F_1(i_3)=\frac{\sqrt{6}}{4},F_1(i_4)=-\frac{\sqrt{6}}{2}$$

$$F_2(i_1)=\frac{\sqrt{2}}{2},F_2(i_2)=-\frac{\sqrt{2}}{4},F_2(i_3)=-\frac{\sqrt{2}}{4},F_2(i_4)=\frac{\sqrt{2}}{2}$$

ενώ οι αντίστοιχες μάζες των γραμμών ισούνται με

$$f_1=\frac{1}{6},f_2=\frac{2}{6},f_3=\frac{2}{6},f_4=\frac{1}{6}$$

Το ότι οι δύο παραγοντικοί άξονες Δ_1 και Δ_2 είναι κάθετοι, θα προκύψει, με την επαλήθευση της σχέσης 12, η οποία αποτελεί τη διατύπωση του Πυθαγόρειου Θεωρήματος στο επίπεδο.

Π.χ Για την γραμμή i_2 έχουμε

$$F_1(i_2)=-\frac{\sqrt{6}}{4}=-0,612,F_2(i_2)=-\frac{\sqrt{2}}{4}=-0,354$$

Χρησιμοποιώντας τη σχέση 13 έχουμε ότι

$$d^2(i_2,f_J)=\frac{1}{\frac{2}{6}}\left(\frac{1}{2}-\frac{2}{6}\right)^2+\frac{1}{\frac{2}{6}}\left(0-\frac{2}{6}\right)^2+\frac{1}{\frac{2}{6}}\left(\frac{1}{2}-\frac{2}{6}\right)^2=3\cdot\frac{1}{36}+3\cdot\frac{4}{36}+3\cdot\frac{1}{36}=\frac{18}{36}=\frac{1}{2}$$

Οπότε $$d^2(i_2, f_J) = \frac{1}{2} \rightarrow \quad d(i_2, f_J) = 0,707$$

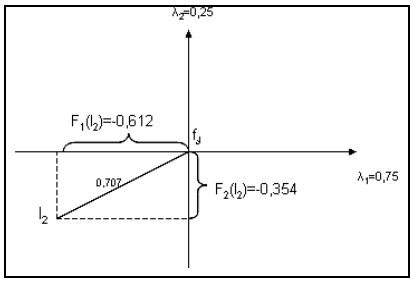
Χρησιμοποιώντας τις τιμές των $F_1(i_2)$ και $F_2(i_2)$ της γραμμής i_2 στους δύο παραγοντικούς άξονες Δ_1 και Δ_2 έχουμε

$$F_1^2(i_2) + F_2^2(i_2) = \left(-\frac{\sqrt{6}}{4}\right)^2 + \left(-\frac{\sqrt{2}}{4}\right)^2 = \frac{6}{16} + \frac{2}{16} = \frac{1}{2}$$

Επομένως επαληθεύεται η σχέση 12

$$d^2(i_2, f_J) = F_1^2(i_2) + F_2^2(i_2) \qquad (17)$$

Σχηματικά το παραγοντικό επίπεδο 1x2 είναι το ακόλουθο.



σχήμα 2: Επαλήθευση της σχέσης $d^2(i_2, f_J) = F_1^2(i_2) + F_2^2(i_2)$

Προφανώς η σχέση 17 ισχύει και στο πολυδιάστατο χώρο R^{p-1}

$d^2(i_k, f_J) = F_1^2(i_k) + F_2^2(i_k) + ... + F_s^2(i_k)$ όπου k=1,…n και s=1,…,p-1 (18)

Η σχέση 18 είναι η έκφραση του Πυθαγόρειου θεωρήματος στο πολυδιάστατο χώρο R^{p-1}

Όσον αφορά στην απόσταση μεταξύ του προφίλ μιας γραμμής f_J^i και του προφίλ μιας μεταβλητής f_I^j ισχύει η παρακάτω σχέση.

$$d^2\left(f_J^{\,i},\ f_I^j\right) = \sum_{p=1}^{p-1}\left[F_p(i) - G_p(j)\right]^2 \qquad (19)$$

Με τα δεδομένα του παραδείγματος για την γραμμή i_1 και την στήλη j_1 έχουμε

$$d^2(i_1, j_1) = \left[F_1(i_1) - G_1(j_1)\right]^2 + \left[F_2(i_1) - G_2(j_1)\right]^2 = \left[\frac{\sqrt{6}}{2} - 0\right]^2 + \left[\frac{\sqrt{2}}{2} - (-\frac{\sqrt{2}}{2})\right]^2 = 3,5$$

Οπότε $$d(i_1, j_1) = \sqrt{3,5} = 1,871$$

Σύνδεση μεταξύ των «γραμμών» και των «στηλών» ενός διδιάστατου πίνακα δεδομένων

Έστω σ' ένα πίνακα δεδομένων T(n,p) οι n γραμμές αντιστοιχούν σε n ερωτώμενους, ενώ στις p γραμμές του πίνακα οι τιμές των p ερωτήσεων που αντιστοιχούν σε p μεταβλητές. Στο ερώτημα του εντοπισμού ενός ερωτώμενου με ποια μεταβλητή συνδέεται περισσότερο, θα γίνει λεπτομερής αναφορά, χρησιμοποιώντας ένα συγκεκριμένο παράδειγμα έξι ποιοτικών μεταβλητών (για την μέτρηση των οποίων χρησιμοποιήθηκε η 5βάθμια κλίμα Likert, όπου το 5 αφορούσε την άριστη εντύπωση), στο οποίο απάντησαν 99 άτομα, τα οποία αποτελούν μία από τις πέντε κλάσεις, που δημιουργήθηκαν με την εφαρμογή της Ανιούσας Ιεραρχικής ταξινόμησης με την διαδικασία VACOR, σε 1721 ξένους επισκέπτες της Θεσσαλονίκης. Τα δεδομένα περιέχονται στην έρευνα που διεξήχθη στα πλαίσια του προγράμματος

ΑΡΧΙΜΗΔΗΣ ΙΙΙ με τίτλο «Τεχνολογίες Ανάλυσης Δεδομένων και Διαχείρισης Γνώσης στο σχεδιασμό τουριστικών προϊόντων»

Το ερωτηματολόγιο αφορούσε έξι ερωτήματα σχετικά με το πώς βαθμολόγησαν οι ξένοι επισκέπτες α) τα αξιοθέατα της πόλης της Θεσσαλονίκης β) την Ελληνική κουζίνα γ) την νυχτερινή ζωή της πόλης δ) το αρχιτεκτονικό της στυλ ε) την ασφάλειά της και στ) την φιλικότητα των ντόπιων.

Οι έξι μεταβλητές παρίστανται αντιστοίχως ως εξής : Δ4,Δ5,Δ6,Δ7,Δ8,Δ9. Με δεδομένη την ταξινόμηση των 99 ατόμων σε μία από τις ομάδες που δημιούργησε η ταξινόμηση των 1721 ατόμων με την μέθοδο VACOR ο πίνακας 4 παρουσιάζει τμήμα των απαντήσεών τους.

Πίνακας 4: Τμήμα του πίνακα δεδομένων

A/A	ind	Δ4	Δ5	Δ6	Δ7	Δ8	Δ9
1	11	5	3	3	5	5	5
2	20	2	2	1	2	3	3
3	60	4	2	3	5	3	5
.	.	.	.	.	.	.	.
21	315	4	3	4	5	3	5
.	.	.	.	.	.	.	.
67	1085	3	4	1	3	4	5
.	.	.	.	.	.	.	.
94	1623	4	0	3	5	4	3
.	.	.	.	.	.	.	.
99	1712	4	2	1	5	3	2

Με βάση τον στατιστικό έλεγχο της διαφοράς των αναλογιών (σχέσεις 1 και 2) σε επίπεδο σημαντικότητας 5%, προσδιορίζεται η σύνδεση κάθε κλάσης ή ερωτώμενου με μία ή περισσότερες μεταβλητές.

Έχοντας εφαρμόσει την Ανιούσα Ιεραρχική Ταξινόμηση με την μέθοδο VACOR στα δεδομένα του πίνακα 4, δημιουργήθηκε η συγκεκριμένη τυπολογία με τις εξής πέντε ομοιογενείς συστάδες :180,186,191,192 και 193 στις οποίες χρησιμοποιώντας τον προαναφερόμενο έλεγχο υποθέσεων προκύπτει ο πίνακας 5, όπου η συστάδα 186 φαίνεται να συνδέεται με δύο μεταβλητές τις Δ5 και Δ6, ενώ η συστάδα 192 συνδέεται περισσότερο με τις μεταβλητές Δ5, Δ7 και Δ9, με μεγαλύτερη ένταση όμως με την Δ5 ($Z_{\Delta 5}$=6,3422).

Πίνακας 5: Παρουσιάζει την σύνδεση μεταξύ των 5 συστάδων και των έξι μεταβλητών

ind	Δ4	Δ5	Δ6	Δ7	Δ8	Δ9
180	-0,4215	-14,4224	13,6994	0,9611	2,2438	-2,414
186	0,9345	1,8842	5,5692	-1,3116	-3,1378	-2,3919
191	12,4939	-15,3903	-15,8726	11,0566	-2,7345	3,528
192	-4,0938	6,3422	-6,8836	1,9318	0,1465	2,51
193	-0,2294	0,1392	-1,2727	-1,6656	2,2553	0,5183

Από την εφαρμογή του ελέγχου στις τιμές των έξι μεταβλητών για κάθε ερωτώμενο προκύπτει ο πίνακας 6,από τον οποίο διαπιστώνεται ότι ο ερωτώμενος 20 συνδέεται περισσότερο με τις μεταβλητές Δ5, Δ8 και Δ9 με μεγαλύτερη ένταση όμως με την Δ5 ($Z_{\Delta 5}$=7,1634), ενώ ο ερωτώμενος 315 φαίνεται να συνδέεται με τρεις μεταβλητές τις Δ5, Δ6 και Δ7 με μεγαλύτερη ένταση όμως με την Δ6 ($Z_{\Delta 6}$=7,9739).

Πίνακας 6: Παρουσιάζει την σύνδεση μεταξύ ερωτώμενων και μεταβλητών

ind	Δ4	Δ5	Δ6	Δ7	Δ8	Δ9
11	-0,5345	1,5176	0,631	0,0817	0,5959	-1,7833
20	-4,8848	7,1634	-4,8654	-4,3151	5,021	2,4658
60	-1,7209	-2,0604	3,6369	4,0674	-5,8574	2,0785
315	-3,4385	2,9401	7,9739	1,909	-7,1596	-0,0129
1185	21,1945	-15,3903	-15,8726	13,2493	-21,5643	10,9749
1623	1,522	-15,3903	6,7143	8,1673	2,6932	-5,6008

Βέβαια με την διαδικασία αυτή υπάρχουν και πολλοί ερωτώμενοι που δεν συνδέονται ιδιαίτερα με καμία μεταβλητή, όπως λ.χ ο ερωτώμενος 11, επειδή η τιμή της z κατανομής περιέχεται μεταξύ των τιμών $-1,645<z<1,645$, δηλαδή οφείλεται σε τυχαίους παράγοντες, αφού ισχύει η H_0.

Από τον πίνακα 6, λοιπόν, όπως προαναφέρθηκε θεωρούμε ότι η μεγαλύτερη τιμή του z που αντιστοιχεί σε κάποια μεταβλητή από την σειρά των έξι τιμών, προσδιορίζει την σύνδεσή του ερωτώμενου με την συγκεκριμένη μεταβλητή.

Χρησιμοποιώντας ως εναλλακτική προσέγγιση του προβλήματος της σύνδεσης των ερωτωμένων με τις μεταβλητές, σκόπιμο είναι να εφαρμοστεί στα δεδομένα του πίνακα 4 η Παραγοντική Ανάλυση των Αντιστοιχιών (-Π.Α.Α-) και στη συνέχεια βάσει των συντεταγμένων των ερωτωμένων και των μεταβλητών στους p-1 παραγοντικούς άξονες, να προσδιοριστεί με την χρήση της Ευκλείδειας μετρικής, ποιος ερωτώμενος συνδέεται κυρίως με ποια μεταβλητή.

Η εξαγωγή των αποτελεσμάτων με την χρήση της Π.Α.Α προϋποθέτει σε κάθε περίπτωση να ληφθεί υπόψη η ερμηνευτική ικανότητα των παραγοντικών αξόνων.

Προς τούτο υπάρχουν τρεις περιπτώσεις ερμηνείας, ανάλογα με την ερμηνευτική ικανότητα των παραγοντικών αξόνων. Δηλαδή να χρησιμοποιηθούν δύο παραγοντικοί άξονες ή τρεις ή να χρησιμοποιηθούν όλοι οι παραγοντικοί άξονες που προσδιορίζονται από το σύνολο των μεταβλητών

1. Με δύο παραγοντικούς άξονες.

Αναλύοντας τα δεδομένα του πίνακα 4 με την Παραγοντική Ανάλυση των Αντιστοιχιών, προκύπτουν τα εξής:

Α) Προβολή χαρακτηριστικών ριζών

Πίνακας 7:Ιστόγραμμα χαρακτηριστικών ριζών

ΑΞΟΝ	ΑΔΡΑΝΕΙΑ	%ΕΡΜΗΝΕΙΑΣ	ΑΘΡΟΙΣΜΑ	|ΙΣΤΟΓΡΑΜΜΑ ΧΑΡΑΚΤΗΡ.ΡΙΖΩΝ
1	0,0305713	33,07	33,07	|**********
2	0,0285268	30,86	63,93	|**********
3	0,0149623	16,19	80,11	|**********
4	0,0097252	10,52	90,63	|**********
5	0,0086587	9,37	100,00	|**********

Από τον πίνακα 7 προκύπτει ότι με τους δύο πρώτους παραγοντικούς άξονες αντλείται το 63,93% της συνολικής πληροφόρησης που προέρχεται από τα δεδομένα του πίνακα 4.

Πίνακας 8: Συντεταγμένες των ερωτώμενων και των μεταβλητών για το σύνολο των παραγοντικών αξόνων

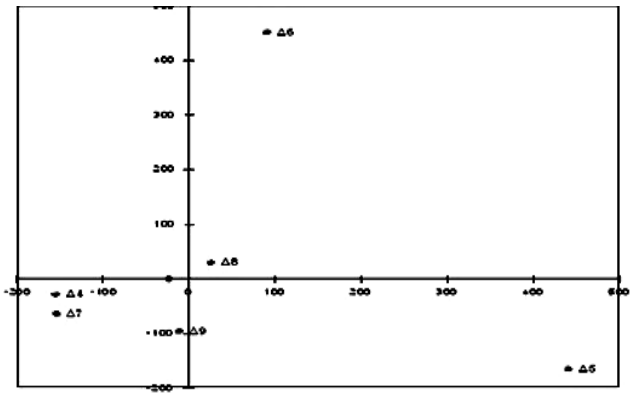
Συντεταγμένες των ερωτώμενων

ind	FA1	FA2	FA3	FA4	FA5
11	34	12	-3	-15	-34
20	182	-122	-111	46	12
60	-49	51	80	-107	90
73	-3	-111	-135	-103	-238
82	162	73	-101	-60	-33
88	-81	-35	-182	-103	-108
99	-7	104	92	-48	24

Συντεταγμένες των μεταβλητών

ind	GA1	GA2	GA3	GA4	GA5
Δ4	-153	-29	136	114	-78
Δ5	441	-165	97	-34	-63
Δ6	91	451	70	-34	37
Δ7	-152	-65	-1	-177	-29
Δ8	27	29	-229	57	-64
Δ9	-10	-97	-8	38	169

Στη συνέχεια παρατηρούμε το παραγοντικό επίπεδο 1x2

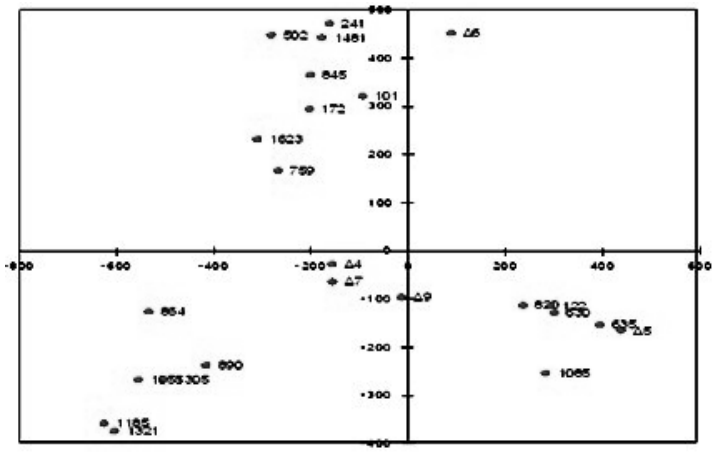


σχήμα 3: Παραγοντικό επίπεδο 1x2 των μεταβλητών

Με βάση τους δύο πρώτους παραγοντικούς άξονες το επίπεδο 1x2 χωρίζεται σε τέσσερις υποχώρους. Στον 1^{ο} υποχώρο τοποθετούνται τα σημεία τα οποία έχουν και τις δύο συντεταγμένες θετικές. Με την ίδια διαδικασία ανάλογα με τα πρόσημα των συντεταγμένων των σημείων τοποθετούνται στον 2^{ο}, στον 3^{ο}, είτε στον 4^{ο} υποχώρο.

Έτσι στον 1^{ο} υποχώρο βρίσκονται οι μεταβλητές Δ6 και Δ8, στον 2^{ο} δεν βρίσκεται καμία μεταβλητή, στον 3^{ο} υποχώρο βρίσκονται οι μεταβλητές Δ4,Δ7 και Δ9, ενώ στον 4^{ο} υποχώρο βρίσκεται η μεταβλητή Δ5.

Πάνω στο παραγοντικό επίπεδο 1x2 τοποθετούνται και τα προφίλ των 99 ερωτώμενων, ανάλογα με τα πρόσημα των συντεταγμένων τους, όπου μία ομάδα «υποκειμένων» στο 2^{ο} υποχώρο είναι «ορφανή» από μεταβλητή. (Η επιλογή των σημείων έγινε βάσει των κριτηρίων COR και CTR[1])



σχήμα 4: Παραγοντικό επίπεδο 1x2 των «αντικειμένων» και των μεταβλητών

Επειδή οι δύο πρώτοι παραγοντικοί άξονες δημιουργούν ένα ορθογώνιο σύστημα συντεταγμένων, για να μετρηθεί μία απόσταση μεταξύ μιας μεταβλητής και ενός «υποκειμένου» χρησιμοποιείται η Ευκλείδεια μετρική μεταξύ δύο σημείων κάνοντας χρήση των συντεταγμένων τους πάνω στους δύο άξονες. Με την χρήση του λογισμικού MAD υπολογίζεται η μικρότερη απόσταση κάθε «υποκειμένου» μεταξύ των έξι μεταβλητών, οπότε προκύπτει το παρακάτω αποτέλεσμα.

Πίνακας 9: Σύνδεση μεταξύ «υποκειμένων» και μεταβλητών βάσει δύο παραγόντων

ΜΕΤ	Δ4	Δ5	Δ6	Δ7	Δ8	Δ9
ΠΛΗΘΟΣ	8	7	5	14	45	20
	88	122	101	358	11	20
	172	238	241	389	60	73
	554	420	502	629	82	116
	644	630	845	645	99	209
	693	635	1481	864	176	213
	759	820		890	246	378
	1548	1085		1006	298	596
	1623			1055	301	641
				1185	312	643
				1279	315	678
				1305	355	703
				1321	368	704
				1433	399	705
				1712	452	879
					536	882
					553	1086
					555	1105
					567	1328
					702	1503
					745	1696
					914	
					922	
					990	
					1000	
					1019	
					1020	
					1114	
					1127	
					1140	
					1156	
					1172	
					1200	
					1269	
					1307	
					1383	
					1482	
					1530	
					1540	
					1546	
					1570	
					1580	
					1619	
					1626	
					1633	
					1694	

Από τον πίνακα 9 διαπιστώνεται ότι 7 ερωτώμενοι {122,238,420,630,635,820 και 1085} συνδέονται με την μεταβλητή Δ5, καθόσον βρίσκονται στη μικρότερη απόσταση, από τις υπόλοιπες πέντε μεταβλητές, ενώ με την μεταβλητή Δ4 συνδέονται 8 ερωτώμενοι.

2) Με τρεις παραγοντικούς άξονες

Εφόσον χρησιμοποιηθούν οι τρεις πρώτοι παραγοντικοί άξονες και με δεδομένο ότι αποτελούν στον χώρο των τριών διαστάσεων ένα τρισορθογώνιο σύστημα συντεταγμένων, τότε δημιουργούνται οκτώ υποχώροι, στους οποίους ανάλογα με τα πρόσημα των συντεταγμένων των σημείων τοποθετούνται σ' έναν απ' αυτούς.

Συνεπώς αν συμβολίσουμε τα σημεία που βρίσκονται στο 1° ,2° ,3° και 4° υποχώρο με το σύμβολο ↑ δίπλα από την ταυτότητα του σημείου, ενώ για τα σημεία στον 5°, 6°, 7° και 8° υποχώρο θέσουμε το σύμβολο ↓ ,έχουμε για πρώτη φορά στη παγκόσμια βιβλιογραφία, μια απεικόνιση του τριδιάστατου χώρου στο επίπεδο, (χωρίς να χρησιμοποιείται η μέθοδος της προοπτικής), οπότε τα διαγράμματα αυτά στο εξής θα αναφέρονται ως **διαγράμματα Καραπιστόλη**.

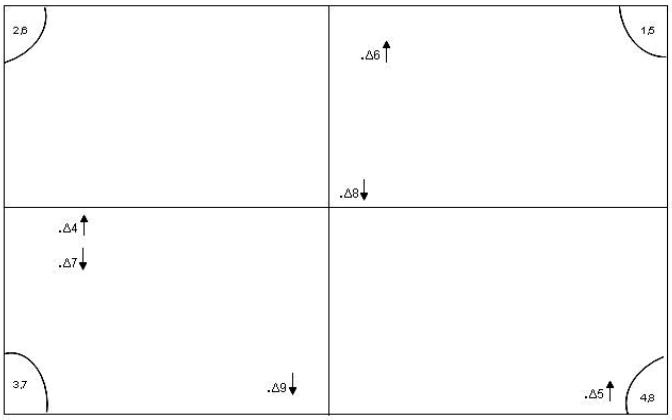
Εφαρμόζοντας την συγκεκριμένη διαδικασία. προκύπτουν αφ' ενός ο πίνακας 10 με την τοποθέτηση των μεταβλητών στους 8 υποχώρους, αφετέρου ο πίνακας 11 με την τοποθέτηση των 99 ερωτώμενων στους αντίστοιχους υποχώρους, καθώς και το αντίστοιχο διάγραμμα Καραπιστόλη.

Πίνακας 10: Οι οκτώ υποχώροι με τις μεταβλητές που εντοπίζονται σ' αυτούς

Α/Α	1ος	2ος	3ος	4ος	5ος	6ος	7ος	8ος
1	Δ6		Δ4	Δ5	Δ8		Δ7	
2							Δ9	

Πίνακας 11: Οι οκτώ υποχώροι με τους ερωτώμενους που εντοπίζονται σ' αυτούς

ΜΕΤ	1ος	2ος	3ος	4ος	5ος	6ος	7ος	8ος
ΠΛΗΘΟΣ	23	6	10	9	11	10	15	15
	298	60	358	116	11	101	73	20
	312	99	645	209	82	172	88	122
	315	301	879	630	246	241	213	176
	399	502	1055	635	355	554	389	238
	567	536	1105	643	452	759	596	368
	702	1546	1185	703	553	845	629	378
	745		1321	704	555	1383	644	420
	914		1328	705	1140	1481	693	641
	922		1433	882	1172	1548	864	678
	1019		1712		1307	1623	890	820
	1020				1694		1006	990
	1114						1086	1000
	1127						1279	1085
	1156						1305	1482
	1200						1503	1696
	1269							
	1530							
	1540							
	1570							
	1580							
	1619							
	1626							
	1633							



σχήμα 5: Παραγοντικός χώρος 1x2x3 των μεταβλητών

Από το διάγραμμα Καραπιστόλη διαπιστώνεται ότι οι μεταβλητές Δ4 και Δ7, καθώς και οι μεταβλητές Δ6 και Δ8 στον τριδιάστατο χώρο βρίσκονται σε διαφορετικούς υποχώρους, με ότι αυτό μπορεί να σημαίνει για την ερμηνεία τους σε ποσοστό 80,11%, έναντι της ερμηνείας τους σε ποσοστό 63,93% που παρέχει το παραγοντικό επίπεδο 1x2.

Από τον πίνακα 11 προκύπτει με βάση τα πρόσημα των τριών συντεταγμένων των σημείων, ότι με την τοποθέτησή τους στους οκτώ υποχώρους, η μεταβλητή Δ5 και οι 9 ερωτώμενοι που ανήκουν στον 4^{ο} υποχώρο όφειλαν με την χρήση των τριών παραγόντων να συνδέονται μεταξύ τους. Υπολογίζοντας όμως τις ελάχιστες αποστάσεις μεταξύ των ερωτώμενων και των μεταβλητών προκύπτει ο πίνακας 12, ο οποίος πληροφορεί ότι γενικώς μόνο 4 ερωτώμενοι (οι 122,630,635 και 1085) συνδέονται με την μεταβλητή Δ5.

Το ενδιαφέρον είναι ότι από τους εννέα ερωτώμενους που ανήκουν στο 4^{ο} υποχώρο μαζί με την μεταβλητή Δ5 (πίνακας 11), μόνο δύο εξ' αυτών ο 630 και ο 635 συνδέονται πράγματι με την μεταβλητή Δ5, επειδή βρίσκονται σε μικρότερη απόσταση, ενώ οι άλλοι 7 συνδέονται με άλλες μεταβλητές όπως λ.χ ο ερωτώμενος 116 συνδέεται με την μεταβλητή Δ9

που βρίσκεται στον 7^{ο} υποχώρο, λόγω μικρότερης απόστασης από εκείνη που έχει με την μεταβλητή Δ5.

Πίνακας 12: Σύνδεση των ερωτώμενων με τις μεταβλητές βάσει τριών παραγόντων

ΜΕΤ	Δ4	Δ5	Δ6	Δ7	Δ8	Δ9	
ΠΛΗΘΟΣ	15	4	6	16	11	47	
	60	122	101	389	82	11	705
	99	630	241	554	88	20	745
	301	635	502	629	172	73	820
	358	1085	845	644	452	116	879
	536		1269	693	553	176	882
	645		1481	759	555	209	914
	702			864	990	213	922
	1019			890	1000	238	1086
	1020			1006	1140	246	1114
	1105			1055	1503	298	1127
	1156			1279	1694	312	1172
	1185			1305		315	1200
	1321			1383		355	1307
	1433			1548		368	1328
	1546			1623		378	1482
				1712		399	1530
						420	1540
						567	1570
						596	1580
						641	1619
						643	1626
						678	1633
						703	1696
						704	

Αυτή η διαφορετική σύνδεση των ερωτώμενων με την μεταβλητή Δ5, οφείλεται σε ποσοστό 16,19% της πληροφορίας που παρέχει ο 3^{ος} παραγοντικός άξονας.

3.Με το σύνολο των παραγοντικών αξόνων

Χρησιμοποιώντας το σύνολο των πέντε παραγοντικών αξόνων δημιουργείται μία ορθοκανονική βάση στον R^5, όπου τοποθετούνται όλες οι μεταβλητές και όλοι οι ερωτώμενοι στις πραγματικές τους θέσεις, απ' όπου αντλείται το σύνολο της πληροφόρησης που παρέχει ο πίνακας δεδομένων. Η σύνδεση των ερωτώμενων με βάση τη ελάχιστη απόσταση μεταξύ μεταβλητών και «υποκειμένων» χρησιμοποιώντας όλες τις συντεταγμένες παρουσιάζεται στον πίνακα 13, σε αντιδιαστολή με τις συνδέσεις τους χρησιμοποιώντας την z κατανομή.

Από τον πίνακα 13 προκύπτει ότι 7 ερωτώμενοι {122,238,420,630,635,820,1085} συνδέονται με την μεταβλητή Δ5, (όπως και στη περίπτωση των δύο παραγόντων), ενώ όσον αφορά τους ερωτώμενους που συνδέονται με την μεταβλητή Δ4 διαπιστώνουμε ότι αυξάνοντας το ποσοστό πληροφόρησης από 63,11% στο 100%, διαφοροποιείται το πλήθος και οι ερωτώμενοι που συνδέονται με την μεταβλητή, παρατήρηση η οποία ισχύει και για κάποιες από τις υπόλοιπες μεταβλητές.

Έτσι με την μεταβλητή Δ4 με τους δύο πρώτους παράγοντες συνδέονται 8 ερωτώμενοι, με τους τρεις παράγοντες συνδέονται 15 και με όλους τους παράγοντες συνδέονται 16. Αξιοσημείωτο είναι ότι κανείς από τους 8 ερωτώμενους που φαίνεται να συνδέονται με την μεταβλητή Δ4,βάσει των δύο παραγόντων δεν φαίνεται να συνδέονται με τους 15 ερωτώμενους, όταν γίνεται χρήση των τριών παραγόντων, ενώ όταν γίνεται χρήση του συνόλου των παραγόντων μόνο 10 από τους 15 συνεχίζουν να συνδέονται με την μεταβλητή Δ4.

Αυτή η διαφοροποίηση στο πλήθος και στο ποιοι ερωτώμενοι συνδέονται με κάθε μεταβλητή, ανάλογα με το πλήθος των παραγόντων που λαμβάνονται υπόψη, σημαίνει ότι η μόνη ορθή διαδικασία εντοπισμού της σύνδεσης των ερωτώμενων με τις μεταβλητές ενός πίνακα δεδομένων, είναι να χρησιμοποιούνται όλοι οι παράγοντες που εξάγονται από την Παραγοντική Ανάλυση των Αντιστοιχιών, αφού σε αυτή την περίπτωση δημιουργείται ένας ν-διάστατος ορθοκανονικός χώρος, ο οποίος απεικονίζει την πραγματική εικόνα των σχέσεων μεταξύ των σημείων που απεικονίζουν τους ερωτώμενους και των σημείων που

απεικονίζουν τις μεταβλητές, προσφέροντας το 100% της πληροφορίας ανεξάρτητα από την συμμετοχή κάθε «υποκειμένου» στη διαμόρφωση των παραγοντικών αξόνων.

Η σύνδεση των υποκειμένων με τις μεταβλητές βάσει της Ευκλείδειας μετρικής και της μέγιστης τιμής της z κατανομής παρουσιάζονται στον παρακάτω πίνακα.

Πίνακας 13: Σύνδεση υποκειμένων και μεταβλητών με τις δύο διαφορετικές διαδικασίες (Ευκλείδεια μετρική και z κατανομής)

	Συνδεση με Ευκλείδεια μετρική							Συνδεση με z μετρική					
IND	**Δ4**	**Δ5**	**Δ6**	**Δ7**	**Δ8**	**Δ9**		**Δ4**	**Δ5**	**Δ6**	**Δ7**	**Δ8**	**Δ9**
Πλήθος	**16**	**7**	**9**	**20**	**26**	**21**		**6**	**36**	**21**	**11**	**22**	**3**
	99	122	101	60	11	116		536	11	99	60	73	213
	301	238	172	358	20	209		645	20	101	358	82	1433
	536	420	241	389	73	213		1055	116	172	629	88	1696
	645	630	502	567	82	298		1105	122	241	879	176	
	702	635	845	629	88	312		1185	209	301	1086	368	
	759	820	1269	644	176	315		1548	238	315	1279	378	
	1019	1085	1481	693	246	378			246	399	1305	389	
	1020		1570	864	355	596			298	502	1321	452	
	1055		1633	879	368	641			312	845	1328	554	
	1105			890	399	643			355	1114	1623	644	
	1127			914	452	678			420	1156	1712	693	
	1185			1006	553	703			553	1172		759	
	1200			1086	554	704			555	1269		864	
	1546			1279	555	705			567	1481		890	
	1548			1305	990	745			596	1530		990	
	1626			1321	1000	882			630	1540		1000	
				1328	1140	922			635	1546		1006	
				1383	1172	1114			641	1570		1140	
				1623	1307	1156			643	1580		1383	
				1712	1482	1433			678	1619		1482	
					1503	1696			702	1633		1503	
					1530				703			1694	
					1540				704				
					1580				705				
					1619				745				
					1694				820				
									882				
									914				
									922				
									1019				
									1020				
									1085				
									1127				
									1200				
									1307				
									1626				

Στόχος της εργασίας τελικά είναι να αποδειχθεί η πιο αποτελεσματική λύση εντοπισμού της σύνδεσης μεταξύ μεταβλητών και υποκειμένων, έχοντας χρησιμοποιήσει την z κατανομή και την Ευκλείδεια μετρική. Ο στόχος αυτός μπορεί να υλοποιηθεί εφόσον εκπαιδευτούν τα δεδομένα που απορρέουν από τον πίνακα 13 χρησιμοποιώντας ταξινομητές μηχανικής μάθησης και ειδικότερα την μηχανή μάθησης SVM

Γενικά περί εκπαίδευσης δεδομένων με ταξινομητές μηχανικής μάθησης

Η **μηχανική μάθηση** (machine learning) είναι μια περιοχή της τεχνικής νοημοσύνης η οποία αφορά αλγορίθμους και μεθόδους που επιτρέπουν στους υπολογιστές να «μαθαίνουν». Στόχος της Μηχανικής μάθησης είναι με την χρήση ενός υπολογιστικού συστήματος, η δημιουργία μοντέλων χρησιμοποιώντας ένα σύνολο δεδομένων.

Έχουν αναπτυχθεί πολλές τεχνικές μηχανικής μάθησης που χρησιμοποιούνται ανάλογα με τη φύση του προβλήματος και εμπίπτουν σε ένα από τα παρακάτω δυο είδη:

1.Μάθηση με επίβλεψη (supervised learning)

2.Μάθηση χωρίς επίβλεψη (unsupervised learning

Στη μάθηση με επίβλεψη το σύστημα καλείται να «μάθει» μια έννοια ή συνάρτηση από ένα σύνολο δεδομένων, η οποία αποτελεί περιγραφή ενός μοντέλου.

Στη μάθηση χωρίς επίβλεψη το σύστημα πρέπει μόνο του να ανακαλύψει συσχετίσεις ή ομάδες σε ένα σύνολο δεδομένων, δημιουργώντας πρότυπα, χωρίς να είναι γνωστό αν υπάρχουν, πόσα και ποια είναι.

Στη παρούσα εργασία θα χρησιμοποιηθεί η μάθηση με επίβλεψη, όπου το σύστημα πρέπει να «μάθει» επαγωγικά μια συνάρτηση που ονομάζεται **συνάρτηση στόχος** (target function) και αποτελεί έκφραση του μοντέλου που περιγράφει τα δεδομένα.

Η συνάρτηση στόχος χρησιμοποιείται για την πρόβλεψη της τιμής μίας μεταβλητής, που ονομάζεται **μεταβλητή εξόδου**, βάσει των τιμών ενός συνόλου μεταβλητών, που ονομάζονται **μεταβλητές εισόδου** ή **χαρακτηριστικά**

Στην μάθηση με επίβλεψη διακρίνονται δυο είδη προβλημάτων (learning tasks), τα προβλήματα ταξινόμησης και τα προβλήματα παρεμβολής.

Η ταξινόμηση (classification) αφορά στη δημιουργία μμοντέλων πρόβλεψης διακριτών τάξεων (κλάσεων/κατηγοριών).

Η παρεμβολή (regression) αφορά στη δημιουργία μμοντέλων πρόβλεψης αριθμητικών τιμών.

Μηχανές Διανυσμάτων Υποστήριξης (ΜΔΥ). Support Vector Machines (SVM)

Οι **Μηχανές Διανυσμάτων Υποστήριξης (ΜΔΥ)** χαρακτηρίζονται ως μμηχανές μμάθησης και στηρίζονται στη Θεωρία Στατιστικής Μάθησης (Statistical Learning Theory) και στα νευρωνικά δίκτυα τύπου Perceptron. Προτάθηκαν από τον Vladimir Vapnik.

Στην περίπτωση της ταξινόμησης, οι ΜΔΥ προσπαθούν να βρουν μια υπέρ-επιφάνεια (hypersurface) που να διαχωρίζει στο χώρο των παραδειγμάτων τα αρνητικά από τα θετικά παραδείγματα.

Τα ΜΔΥ χαρακτηρίζονται από τα ακόλουθα στάδια:

1 Εκπαίδευση: Στη φάση αυτή γίνονται οι υπολογισμοί των παραμέτρων του μοντέλου μμάθησης με χρήση κατάλληλου συνόλου δεδομένων μμάθησης.

2 Δοκιμή: Το μμοντέλο παραμέτρων (support vectors) που υπολογίστηκε δοκιμάζεται για τη δυνατότητα επιτυχημένης εκτίμησης αποτελέσματος σε ένα σετ δεδομένων που δεν έχει εκπαιδευθεί.

3 Εκτίμηση επίδοσης: Υπολογίζονται κατάλληλοι δείκτες επίδοσης του μοντέλου, κυρίως του ποσοστού σφάλματος, με στόχο την διερεύνηση της δυνατότητας γενίκευσης του μοντέλου.

Οι Μηχανές Διανυσμάτων Υποστήριξης (SVM, Support Vector Machines) ανήκουν στους αλγόριθμους Επιβλεπόμενης Μηχανικής Μάθησης με αξιοσημείωτη επιτυχία σε προβλήματα κατάταξης. Όπως και οι περισσότεροι αλγόριθμοι μηχανικής μάθησης, παριστάνουν τα προς κατάταξη αντικείμενα ως διανύσματα ιδιοτήτων.

Στην περίπτωσή μας, τα προς κατάταξη αντικείμενα είναι ερωτώμενοι και οι ιδιότητες παρέχουν πληροφορίες όπως αν ο προς κατάταξη ερωτώμενος συνδέεται με την Α ή Β μεταβλητή.

Προκειμένου να χρησιμοποιηθούν οι ΜΔΥ σε προβλήματα ταξινόμησης με περισσότερες από δύο τάξεις έχουν προταθεί δύο κατηγορίες προσεγγίσεων

- Άμεσες: Εύρεση των διαχωριζόντων υπέρ-επιπέδων σε ένα βήμα (Vapnik (1998), Crammer and Singer (2000))
- Έμμεσες: Συνδυασμός των αποτελεσμάτων ενός συνόλου δυαδικών ΜΔΥ: ένας-εναντίον-ενός, ένας-εναντίον-όλων (Vapnik (1998))

Οι έμμεσες προσεγγίσεις είναι απλούστερες και υλοποιούνται ευκολότερα, πάντως καμία από τις προσεγγίσεις δεν επιστρέφει πιθανότητες.

Εφαρμογή της μηχανής μάθησης SVM

Για την εφαρμογή της προτεινόμενης σύγκρισης χρησιμοποιούνται εκ νέου οι έξι ποιοτικές μεταβλητές Δ4 έως Δ9 οι τιμές των οποίων αφορούν στις απαντήσεις που δόθηκαν από τους 99 ξένους επισκέπτες της Θεσσαλονίκης. Ο πίνακας 14 παρουσιάζει την σύνδεση των αντικειμένων με τις αντίστοιχες μεταβλητές, αφενός βάσει της ελάχιστης απόστασης χρησιμοποιώντας την Ευκλείδεια μετρική, αφετέρου με την μέγιστη τιμή της z κατανομής.

Πίνακας 14: Σύνδεση των 99 αντικειμένων και των 6 μεταβλητών βάσει της Ευκλείδειας μετρικής και της μέγιστης τιμής της z κατανομής

ΕΝΔ	Δ4	.....	Δ9	MET1
11	5		5	5
20	2		3	5
60	4		5	4
73	3		2	5
82	3		4	5
88	2		2	5
99	4		4	1
101	5		5	3
116	3		5	6
122	3		5	2
172	4		5	3

ΕΝΔ	Δ4	.....	Δ9	MET2
11	5		5	2
20	2		3	2
60	4		5	4
73	3		2	5
82	3		4	5
88	2		2	5
99	4		4	3
101	5		5	3
116	3		5	2
122	3		5	2
172	4		5	3

Σημείωση 1:Οι τιμές των μεταβλητών MET1 και MET2 από 1 έως 6 αντιστοιχούν στις μεταβλητές Δ4 έως Δ9. Η σύμπτωση τιμών μεταξύ MET1 και MET2 ανέρχεται στο 49,5%

Εκπαίδευση των δεδομένων με την Μηχανή Διανυσμάτων Υποστήριξης ΜΔΥ

Εκπαιδεύοντας τα δεδομένα του πίνακα 13 με την χρήση της Μηχανής Διανυσμάτων Υποστήριξης (ΜΔΥ, Support Vector Machines) διαπιστώνεται ότι η διαδικασία σύνδεσης αντικειμένων και μεταβλητών με την Ευκλείδεια μετρική υπερέχει από την αντίστοιχη με την z κατανομή. Αυτή η διαπίστωση προκύπτει επειδή το ποσοστό επίδοσης στην εκμάθηση των δεδομένων του πίνακα 13, που αφορά στη σύνδεση των αντικειμένων με την Ευκλείδεια μετρική είναι υψηλότερο (78,89% πίνακας 16) από εκείνο που προκύπτει με την κατανομή z (71,11% Πίνακας 15).

Επί πλέον με την Ευκλείδειο μετρική στις 20 επαναλήψεις εκμάθησης των δεδομένων τα ποσοστά πάνω από την μέση τιμή είναι κατά πολύ υψηλότερα (7 στις 20 επαναλήψεις πάνω από 80% με μέγιστη τιμή 100% και κατώτερη 61,11%) από τα αντίστοιχα ποσοστά με την z κατανομή που παρουσιάζουν μέγιστη τιμή μόλις στο 88,89% και κατώτερη 44,44%.

Πίνακας 15: Εκπαίδευση με την ΜΔΥ βάσει της z κατανομής

MET_Z (20%)	
0,7778	
0,6667	
0,7778	
0,6667	
0,7778	
0,6667	
0,4444	
0,5556	71,11%
0,6111	
0,8333	
0,7778	
0,6667	
0,6667	
0,7778	
0,7222	
0,7778	
0,7778	
0,6667	
0,8889	
0,7222	

Πίνακας 16: Εκπαίδευση με την ΜΔΥ βάσει της Ευκλείδειας μετρικής

MET_EU (20%)	
0,8333	
0,7778	
0,7778	
0,7778	
0,8333	
0,8889	
0,7222	
0,7778	78,89%
0,9444	
0,7778	
1,0000	
0,7778	
0,6111	
0,7222	
0,7222	
0,8889	
0,7222	
0,8333	
0,6667	
0,7222	

Η νέα διαδικασία ταξινόμησης των γραμμών ενός πίνακα δεδομένων. Μέθοδος karap

Η προτεινόμενη διαδικασία ταξινόμησης απαντά στον προβληματισμό που υφίσταται σε κάθε Ανιούσα Ιεραρχική Ταξινόμηση με τη μέθοδο VACOR, ότι δηλαδή δεν εντοπίζονται με ακρίβεια τα αντικείμενα που συνδέονται με τις μεταβλητές των κλάσεων.

Για τον λόγο αυτό η προτεινόμενη νέα διαδικασία ταξινόμησης των n αντικειμένων ενός πίνακα δεδομένων T(n,p), επιτυγχάνει στο βαθμό που επιθυμεί ο ερευνητής, την ομοιογένεια των αντικειμένων ως προς την αποκλειστική σύνδεσή τους με κάθε μία μεταβλητή ή αν επιθυμεί ως προς ένα συνδυασμό μεταβλητών. Ο αλγόριθμος της προτεινόμενης ταξινόμησης με την ονομασία **KARAP**, ο οποίος υλοποιείται με το πρόγραμμα MAD είναι ο εξής:

1. Δημιουργείται ο λογικός πίνακας 0-1 ο προερχόμενος από τον πίνακα δεδομένων T(n,p), είτε χρησιμοποιώντας διαβαθμίσεις των μεταβλητών, είτε κλίμακες Likert.Η αρίθμηση κάθε αντικειμένου αντιστοιχεί από το 1 έως το n.
2. Ο λογικός πίνακας 0-1 αναλύεται με την -Π.Α.Α- για την εξαγωγή των συντεταγμένων των μεταβλητών και των αντικειμένων στους παραγοντικούς άξονες
3. Χρησιμοποιώντας τις συντεταγμένες Fa και Ga εντοπίζεται με βάση την Ευκλείδεια μετρική η σύνδεση κάθε αντικειμένου με κάθε μεταβλητή.

Εάν επιθυμεί ο ερευνητής ταξινόμηση των αντικειμένων ως προς ένα συνδυασμό μεταβλητών, λόγω του μεγάλου πλήθους των, συνεχίζει με τα παρακάτω βήματα.

4. Δημιουργείται ο πίνακας Burt που αντιστοιχεί στον λογικό πίνακα 0-1
5. Εφαρμόζεται η Ανιούσα Ιεραρχική Ταξινόμηση με την μέθοδο VACOR στα δεδομένα του πίνακα Burt
6. Βάσει της τυπολογίας της ιεραρχίας, η οποία προκύπτει από την τομή του δενδρογράμματος σε k συστάδες, κατατάσσονται τα αντικείμενα ανάλογα με τις μεταβλητές με τις οποίες συνδέονται, σύμφωνα με το βήμα 3.

Εφαρμόζοντας την προτεινόμενη ταξινόμηση με βάση τα βήματα 1 έως 3 στα 84 δεδομένα του πίνακα 4 (δεν ελήφθησαν σκόπιμα υπόψη όσοι επισκέπτες δεν απάντησαν σε κάποιο κριτήριο), προέκυψαν τα ακόλουθα αποτελέσματα.

Πίνακας 17: Ταξινόμηση των 84 ερωτώμενων βάσει της μεθόδου karap

	ΑΞΙΟΘΕΑΤΑ				
IND	Δ41	Δ42	Δ43	Δ44	Δ45
ΠΛΗΘΟΣ	0	2	4	0	5
1		20	420		554
2		1140	630		643
3			820		705
4			1085		1114
5					1626
6					
7					
8					
9					
10					
11					
12					
13					
14					
15					

ΚΟΥΖΙΝΑ				
Δ51	Δ52	Δ53	Δ54	Δ55
3	2	8	0	0
101	879	11		
629	1696	246		
1503		298		
		315		
		355		
		399		
		922		
		1482		

ΝΥΚΤΕΡΙΝΗ ΖΩΗ				
Δ61	Δ62	Δ63	Δ64	Δ65
6	4	5	0	0
73	122	567		
88	635	702		
358	990	914		
1279	1000	1020		
1433		1546		
1712				

ΑΡΧΙΤΕΚΤΟΝΙΚΟ ΣΤΥΛ				
Δ71	Δ72	Δ73	Δ74	Δ75
2	8	8	10	6
452			82	60
1105			381	312
			368	678
			1269	745
			1530	1307
			1540	1383
			1570	
			1580	
			1619	
			1633	

ΑΣΦΑΛΕΙΑ				
Δ81	Δ82	Δ83	Δ84	Δ85
0	1	1	1	3
	645	1156	238	553
				555
				693

ΦΙΛΙΚΟΤΗΤΑ				
Δ91	Δ92	Δ93	Δ94	Δ95
0	0	0	6	15
			99	116
			536	176
			1019	209
			1127	213
			1172	378
			1200	596
				641
				644
				703
				704
				882
				1006
				1086
				1548
				1694

Από τον πίνακα 17 προκύπτει με σαφήνεια η αντίληψη που έχουν οι 84 ερωτηθέντες της συγκεκριμένης κλάσης σχετικά με την εικόνα της Θεσσαλονίκης, ως προς τα έξι κριτήρια που χρησιμοποιήθηκαν στην έρευνα. Π.χ δεν τους άρεσε ή ήταν αδιάφοροι για την Ελληνική κουζίνα (Δ_{51},Δ_{52} και Δ_{53} ποσοστό 13/84=15,48%), ενώ υπερθεμάτιζαν για την φιλικότητα

των ντόπιων (Δ_{94} και Δ_{95} 21/84=25%), όπως επίσης ήταν τελείως αρνητική για την νυκτερινή ζωή της πόλης σε ποσοστό 17,86%

Συμπέρασμα

1. Με τα διαγράμματα Καραπιστόλη απεικονίζεται παραστατικά ο τριδιάστατος χώρος στο επίπεδο, ώστε ο ερευνητής να μην χρειάζεται να πραγματοποιήσει ταξινόμηση για να μην συγχέει γειτονικά σημεία τα οποία ανήκουν σε διαφορετικούς υποχώρους του R^3
2. Με την μηχανή μάθησης SVM αξιολογείται αντικειμενικά κάθε μορφή ταξινόμησης με οποιαδήποτε μετρική και αν δημιουργήθηκε, βοηθώντας τον ερευνητή στη σύγκριση των αποτελεσμάτων μεταξύ δύο διαφορετικών μεθόδων ταξινόμησης, αλλά και στην διαπίστωση της ομοιογένειας των κλάσεων κάθε ταξινόμησης.
3. Κρίνεται σκόπιμο, όταν πρόκειται να ερευνηθεί η σύνδεση των αντικειμένων με συγκεκριμένες μεταβλητές, αφενός να χρησιμοποιηθεί το σύνολο των παραγόντων που προκύπτουν μετά την Παραγοντική Ανάλυση των Αντιστοιχιών, αφετέρου να χρησιμοποιηθεί η προτεινόμενη διαδικασία ταξινόμησης με την ονομασία karap, αφού ο ερευνητής εντοπίζει την μοναδικότητα της σύνδεσης των αντικειμένων εντός των κλάσεων με κάθε διαβάθμιση των μεταβλητών, ερμηνεύοντας με ευκολότερο τρόπο την συμπεριφορά του συνόλου των αντικειμένων.
4.Η μέθοδος karap εξασφαλίζει στον ερευνητή συμπαγείς κλάσεις, ως προς την μοναδικότητα της συμπεριφοράς των αντικειμένων κάθε συστάδας μεταβλητών, αφού περιλαμβάνει μόνο αντικείμενα των οποίων το προφίλ συνδέεται με συγκεκριμένες μεταβλητές της κάθε συστάδας.

Βιβλιογραφία

Α. Ξένη

BENZECRI J.P., BENZECRI F., CHEUNG Y.L et MAIZA S. ,(1985) : «Aides a l'interprétation et étiquetage des arbres en classification ascendante hierarchique:listage FACOR, VACOR et ISUP». Les cahiers de l'analyse des données, n^{O}1, pp 101-123, Dunod, Paris

CRAMMER K, SINGER Y. (2001) On the algorithmique Implémentation of Multi class Kernel-based Vector Machines. Journal of Machine Learning Research 2 (2001) 265-292

KARAPISTOLIS D.,(1986):"Codage et analyse d'une épreuve logique".Les Cahiers de l'Analyse des Données. Vol XI,no 4 pp.482-484

LANG S. (1976) Algèbre Linéaire 1 InterEditions, Paris

MORINEAU A. (1984) «Note sur la caractérisation statistique d'une classe et les valeurs-tests » Bulletin Technique du Centre se Statistique et Informatique Appliquées. Vol. 2, CESIA Paris p.20-27

VAPNIK V., (1998), 'Statistical Learning Theory' , John Wiley&Sons,Inc.

Β. Ελληνική

ΒΛΑΧΟΣ Α. Ενεργητική Μάθηση Με Χρήση Μηχανών Διανυσμάτων Υποστήριξης. http://sdep.cs.aueb.gr/slides/slides_vlachos_26_10_2004.pdf

ΚΑΡΑΠΙΣΤΟΛΗΣ Δ. (2002) «Το λογισμικό MAD». Τετράδια Ανάλυσης Δεδομένων Τεύχος 2. σελ. 133.Θεσσαλονίκη.

ΚΑΡΑΠΙΣΤΟΛΗΣ Δ. (2011) «Πολυδιάστατη Στατιστική Ανάλυση». Εκδόσεις Αλτιντζής. Θεσσαλονίκη.

ΜΠΕΧΡΑΚΗΣ Θ. (1999) «Πολυδιάστατη Ανάλυση Δεδομένων. Μέθοδοι και Εφαρμογές»Αθήνα Εκδόσεις Νέα Σύνορα-Α.Α Λιβάνης.