«ΣΠ0ΥΔΑΙ», Τόμος 47, Τεύχος 3ο-4ο, Πανεπιστήμιο Πειραιώς / «SPOUDAI», Vol. 47, No 3-4, University of Piraeus

ΑΠΟΣΤΑΣΕΙΣ ΓΙΑ ΤΗΝ ΤΑΞΙΝΟΜΗΣΗ ΣΕ ΠΟΙΟΤΙΚΕΣ ΜΕΤΑΒΛΗΤΈΣ (ΤΑΞΙΝΟΜΗΣΗ ΣΕ ΛΟΓΙΚΑ ΔΕΔΟΜΕΝΑ)

Υπό

Γιάννης Παπαδημητρίου και *Γιαννούλα Φλώρου*
Πανεπιστήμιο Μακεδονίας
Τμήμα Εφαρμοσμένης Πληροφορικής

Abstract

In this paper we study how the distance between objects which are characterizewd by qualitative variables, is influenced by three differents forms of distances, the eucleidien, the X^2 and Jaccard.

Our aim is to compare the above distances and to find the best distance for a certain set of data. The best distance depends on the number of classes of variables and the repetition of objects. (JEL C60).

1. Εισαγωγή

Σκοπός της ταξινόμησης κατά αύξουσα ιεραρχία, είναι να επιτευχθεί διαμελισμός ενός συνόλου παρατηρήσεων σε ομογενείς ομάδες, ως προς το σύνολο των μεταβλητών, κάθε μια των οποίων να διαφέρει σημαντικά από τις υπόλοιπες. Μ' αυτό τον τρόπο επιτυγχάνεται μια ιεράρχηση των παρατηρήσεων, δηλαδή, μια σειρά διαμελισμών «ο ένας μέσα στον άλλο», που όσο προχωρά, τόσο πιο λεπτομερής γίνεται.

Η ταξινόμηση κατά αύξουσα ιεραρχία, διενεργείται σε δύο στάδια. Στο πρώτο υπολογίζεται η ομοιότητα μεταξύ των αντικειμένων, χρησιμοποιώντας μια μετρική απόστασης (ή κάποιο συντελεστής συσχέτισης), ενώ στο δεύτερο στάδιο η δημιουργία της ιεράρχησης συντελείται με την συνένωση ομάδων (από τις μικρότερες με ένα αντικείμενο μέχρι την ομάδα που τα περιέχει όλα). Οι μέθοδοι που χρησιμοποιούνται στο δεύτερο στάδιο είναι MIN, MAX, ΜΟΥΕΝΝΕ, κ.ά. Σε προηγούμενες εργασίες μας (Γ. Παπαδημητρίου, Γ. Φλώ-

ρου 1993, Γ. Παπαδημητρίου, Γ. Φλώρου 1994), ασχοληθήκαμε με τις μεθόδους που αφορούν το δεύτερο στάδιο.

Οι πιο γνωστές αποστάσεις που χρησιμοποιούνται στο πρώτο στάδιο της ταξινόμησης κατ' αύξουσα ιεραρχία, είναι του X^2, η ευκλείδεια, και ο δείκτης Jaccard.

Στην εργασία αυτή επιχειρούμε να συγκρίνουμε τις 3 αυτές αποστάσεις, όταν μελετάμε δεδομένα που χαρακτηρίζονται από ποιοτικές μεταβλητές ή ποσοτικές μεταβλητές που είναι χωρισμένες σε κλάσεις. Ο πίνακας δεδομένων A, nxk, παριστάνει ή αντικείμενα που χαρακτηρίζονται ή όχι από κάθε μία από τις k κλάσεις των μεταβλητών. Κάθε στήλη του πίνακα αντιστοιχεί σε μια κλάση, κι ένα αντικείμενο μπορεί να χαρακτηρίζεται ή όχι από την κλάση αυτή. Στη διασταύρωση, λοιπόν, της i γραμμής και j στήλης του πίνακα A(i, j), υπάρχει 1, όταν το i αντικείμενο χαρακτηρίζεται από την αντίστοιχη κλάση της j στήλης, και 0 διαφορετικά.

Ορισμός των 3 αποστάσεων

Η απόσταση μεταξύ δύο αντικειμένων i και i΄ είναι:

με την **ευκλείδια μετρική:** $d_e^2 (i, i') = \sum_{j=1}^{p} [A(i, j) - A(i', j)]^2$

με την x^2: $d_{x^2}^2(i, i') = \sum_{j=1}^{p} \frac{1}{A(.,j)} \left[\frac{A(i,j)}{A(i,.)} - \frac{A(i',j)}{A(i',.)} \right]^2$

και με τον **δείκτη Jaccard**: $d_J(i, i') = 1 - \frac{c}{p+q-c}$

όπου: p το πλήθος των στηλών με 1 του i αντικειμένου,
q το πλήθος των στηλών με 1 του i΄ αντικειμένου και
c το πλήθος των στηλών με κοινά 1 στα δύο αντικείμενα
A (i, .) το άθροισμα της i γραμμής
A (., j) το άθροισμα της j στήλης

Η ευκλείδεια μετρική (Benzecri, 1973) βασίζεται στον υπολογισμό των διαφορών μεταξύ όλων των συντεταγμένων των αντικειμένων. Το ίδιο ισχύει για τη X^2 (Benzecri, 1980), μόνο που κάθε διαφορά σταθμίζεται με το άθροισμα της στήλης για την αντίστοιχη συντεταγμένη.

Η ευκλείδεια μετρική χρησιμοποιείται και στον υπολογισμό των αποστάσεων για δεδομένα που χαρακτηρίζονται από ποσοτικές μεταβλητές, ενώ η X^2 κυρίως χρησιμοποιείται, για ποιοτικές μεταβλητές ή ποσοτικές χωρισμένες σε κλάσεις, όταν ο πίνακας δεδομένων περιέχει συχνότητες εμφάνισης κάποιας κλάσης.

Ο δείκτης Jaccard (Roux, 1985), υπολογίζεται μετρώντας τον αριθμό των συντεταγμένων με κοινά 1 και τον αριθμό αυτών που έχουν 1 μόνο στο ένα από τα δύο αντικείμενα. Το πλήθος των κοινών 0 δεν λαμβάνεται υπόψη. Ο δείκτης αυτός χρησιμοποιείται αποκλειστικά για πίνακες δεδομένων με στοιχεία μόνο Οή 1.

Όταν έχουμε πίνακα ποσοτικών μεταβλητών ή πίνακα συχνοτήτων, μπορούμε να βρούμε το κέντρο βάρους των αντικειμένων, να υπολογίσουμε την αδράνεια (διασπορά), και έτσι να μπορέσουμε να αξιολογήσουμε την ιεράρχηση που προκύπτει από μια ταξινόμηση κατ' αύξουσα ιεραρχία (Γ. Παπαδημητρίου, Γ. Φλώρου 1994). Όταν όμως τα δεδομένα αφορούν μόνο την απουσία ή παρουσία μιας κλάσης, δεν μπορούμε να εφαρμόσουμε την προηγούμενη διαδικασία. Για να πάρουμε αξιόπιστα αποτελέσματα, πρέπει να εφαρμόσουμε την ταξινόμηση κατ' αύξουσα ιεραρχία, επιλέγοντας την σωστή μετρική απόστασης.

2. Υπολογισμός Αποστάσεων

2α. Πλήρη πίνακας μιας μεταβλητής χωρίς επαναλήψεις (ανόμοια αντικείμενα)

Θεωρούμε κατ' αρχήν μια μόνο μεταβλητή χωρισμένη σε k κλάσεις και κάθε αντικείμενο, ανήκει οπωσδήποτε σε μία μόνο από αυτές. Εξετάζουμε όλα τα δυνατά διαφορετικά αντικείμενα, που είναι σε πλήθος k. Ο πίνακας δεδομένων είναι διαστάσεων kxk, το άθροισμα κάθε γραμμής του είναι 1 και το άθροισμα κάθε στήλης του είναι επίσης 1.

Δεν έχει νόημα η ταξινόμηση όλων των διαφορετικών αντικειμένων όταν τα δεδομένα αφορούν μια μεταβλητή. Για καθαρά εισαγωγικούς όμως λόγους μελετάμε αρχικά την περίπτωση αυτή.

Βρίσκουμε λοιπόν τις αποστάσεις ανά δύο, όλων των αντικειμένων, που είναι διαφορετικά μεταξύ τους, με τις 3 μετρικές.

Έστω $i, i' \leqslant k$ δύο διαφορετικές κλάσεις και δύο διαφορετικά αντικείμενα α_i και $\alpha_{i'}$, που χαρακτηρίζονται από τις κλάσεις αυτές αντίστοιχα.

α_i = (0, 0, 1 ..., 0, ..., 0) και α_i = (0, 0, ..., 1, ..., 0)
Οι αποστάσεις τους είναι:

$$d_e\,(a_i, a_{i'}) = \sqrt{(0-0^2 + ... + (1-0)^2 + ... + (0-1)^2 + ...(0-0)^2} = \sqrt{2}$$

$$d_{x^2}\,(a_i, a_{i'}) = \sqrt{\tfrac{1}{1}(\tfrac{1}{1}-0^2 + \tfrac{1}{1}(0-\tfrac{1}{1})^2} = \sqrt{2}$$

$$d_J\,(a_i, a_{i'}) = 1 - \frac{0}{2} = 1$$

Η απόσταση λοιπόν, ανά δύο όλων των διαφορετικών αντικειμένων, όπως είναι φυσικό, είναι η ίδια για την κάθε μετρική, και η ταξινόμηση δεν έχει νόημα.

2β. Πίνακας μια μεταβλητής με επαναλήψεις

Υποθέτουμε ότι έχουμε επαναλήψεις αντικειμένων στον προηγούμενο πίνακα δεδομένων (υπάρχουν κι όμοια αντικείμενα εκτός από τα k διαφορετικά).

΄Εστω ότι έχουμε ένα πίνακα δεδομένων nxk (n>k), τον οποίο φέρνουμε στην μορφή:

	1					k
1	1			.....		0
.	0	1		.....		0
.			1	.....		.
.			.	.....		.
.				.....		.
k				.....		1
.				1		.
n	0			.....		1

Αν το α_i ($i \leqslant K$) αντικείμενο εμφανίζεται p φορές και το α_i, ($i' \leqslant k$) εμφανίζεται q φορές, τότε το άθροισμα της i στήλης· είναι p, της i στήλης είναι q και οι αποστάσεις μεταξύ των δύο αντικειμένων α_i, $\alpha_{i'}$, που διαφέρουν σε μία μόνο κλάση της μεταβλητής, είναι:

$$d_e\,(a_i, a_{i'}) = \sqrt{(0-0)^2 + ... + (1-0)^2 + ... + (0-1)^2 + ... + (0-0)^2} = \sqrt{2}$$

ανεξάρτητη του αριθμού επαναλήψεων.

$$d_{X^2}(a_i, a_{i'}) = \sqrt{\frac{1}{p}(\frac{1}{1}-0)^2 + \frac{1}{q}(0-\frac{1}{1})2} = \sqrt{\frac{1}{p}+\frac{1}{q}} = d_{X^2}(p, q)$$

εξαρτάται από τα p και q.

$d_J(a_i, a_{i'}) = 1 - \frac{0}{2} = 1$ ανεξάρτητη του αριθμού επαναλήψεων.

Έστω τα αντικείμενα α_i, $\alpha_{i'}$ και $\alpha_{i''}$, που εμφανίζονται p, q και r φορές αντίστοιχα.

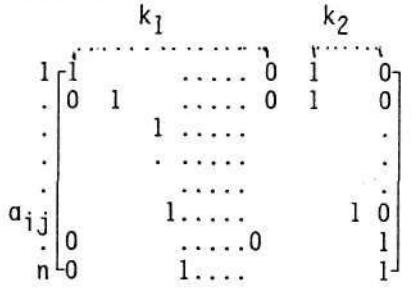
Παρατηρούμε ότι αν π< q, τότε: $\delta_x^2(r, p) > d_x^2(r, q) \Rightarrow$
$\Rightarrow \delta_x^2(\alpha_{i''}, \alpha_i) > d_x^2(\alpha_{i''}, \alpha_{i'})$

δηλαδή η απόσταση με την μετρική X^2 του $\alpha_{i''}$ από το α_i (αντικείμενο με p επαναλήψεις), είναι μεγαλύτερη από την απόστασή του με το $\alpha_{i'}$ (αντικείμενο με q επαναλήψεις), όταν p<q. Αυτό οφείλεται στη στάθμιση που χρησιμοποιείται στην X^2, με συνέπεια τα «βαριά» (με πολλές επαναλήψεις) αντικείμενο να ενώνονται ταχύτερα για τον σχηματισμό των κόμβων της ταξινόμησης.

Η X^2, λοιπόν, οδηγεί σε λάθος υπολογισμό των αποστάσεων, όταν δεν θέλουμε να παίζει σημαντικό ρόλο στην ταξινόμηση ο αριθμός των επαναλήψεων. Στην περίπτωση αυτή ενδείκνυται η χρησιμοποίηση της ευκλείδειας απόστασης ή του δείκτη Jaccard. Αν όμως επιθυμούμε να λαμβάνεται υπόψη η επανάληψη ενός αντικειμένου στην ταξινόμηση, πρέπει να χρησιμοποιήσουμε τη μετρική X^2.

2γ. Πίνακας δύο μεταβλητών, χωρίς επαναλήψεις

Έστω ότι μελετάμε διαφορετικά αντικείμενα που χαρακτηρίζονται από δύο μεταβλητές, η πρώτη με k_1 κλάσεις κι η δεύτερη με k_2. Όλα τα δυνατά διαφορετικά αντικείμενα είναι σε πλήθος k_1k_2, με συνέπεια ο πίνακας δεδομένων να έχει k_1k_2 γραμμές και k_1+k_2 στήλες. Το άθροισμα κάθε γραμμής είναι 2 (πλήθος μεταβλητών), ενώ το άθροισμα των στηλών της πρώτης μεταβλητής είναι k_2 και της δεύτερης k_1.



Έστω a_{ij} το αντικείμενο που χαρακτηρίζεται από τη i-οστη κλάση της πρώτης μεταβλητής και την j-οστή κλάση της δεύτερης μεταβλητής ($i \leqslant k_1$), $j \leqslant k_2$).

Αν δύο αντικείμενα a_{ij}, $a_{i'j}$, διαφέρουν μόνο ως προς μια κλάση της πρώτης μεταβλητής οι αποστάσεις τους είναι:

$$d_e(a_{ij}, a_{i'j}) = \sqrt{1^2 + 1^2} = \sqrt{2}$$

$$d_{x^2}(a_{jj}, a_{j'j}) = \sqrt{\frac{2}{k_2}(\frac{1}{2})^2 = d_x 2(k_2)}$$

$$d_J(a_{ij}, a_{i'j}) 1 - \frac{1}{2+2-1} = \frac{2}{3}$$

Αν δύο αντικείμενα a_{ij}, $a_{ij'}$, διαφέρουν μόνο ως προς μια κλάση της δεύτερης μεταβλητής οι αποστάσεις τους είναι:

$$d_e(a_{ij}, a_{ij'}) = \sqrt{1^2 + 1^2} = \sqrt{2}$$

$$d_{x^2}(a_{ij}, a_{ij'}) = \sqrt{\frac{2}{k_2}(\frac{1}{2})^2 = d_x 2(k_2)}$$

$$d_J(a_{ij}, a_{ij'}) = 1 - \frac{1}{2+2-1} = \frac{2}{3}$$

Τέλος όταν δύο αντικείμενα a_{ij}, $a_{i'j'}$, χαρακτηρίζονται από διαφορετικές κλάσεις τόσο για την πρώτη όσο και για την δεύτερη μεταβλητή, οι αποστάσεις τους είναι:

$$d_e(a_{ij}, a_{i'j'}) = \sqrt{1^2 + 1^2 + 1^2 + 1^2} = \sqrt{4}$$

$$d_{x^2}(a_{ij}, a_{i'j'}) = a\sqrt{\frac{2}{k_2}(\frac{1}{2})^2 + \frac{2}{k_1}(\frac{1}{2})^2} = d_x 2(k_1, k_2)$$

$$d_J(a_{ij}, a_{i'j'}) = 1 - \frac{0}{2+2} = 1$$

Παρατήρηση: Η απόσταση (με τη μετρική X^2) δύο αντικειμένων που διαφέρουν σε μία μόνο κλάση, εξαρτάται από την μεταβλητή στην οποία ανήκει η κλάση αυτή. *Όσο περισσότερες κλάσεις έχει μια μεταβλητή, τόσο μικρότερη*

είναι η απόσταση δυο σημείων που διαφέρουν ως προς μια κλάση της άλλης μεταβλητής. Δηλαδή αν την ίδια μεταβλητή την χωρίσουμε σε διαφορετικό αριθμό κλάσεων, η νέα ταξινόμηση των αντικειμένων θα διαφέρει στους χαμηλούς κόμβους (στους πρωτοσχηματιζόμενους).

Στην περίπτωση που δεν θέλουμε ο αριθμός κλάσεων κάθε μεταβλητής να επηρεάζει την ταξινόμηση, η χρήση της μετρικής X^2, οδηγεί σε εσφαλμένα αποτελέσματα, και ενδείκνυται η χρησιμοποίηση της ευκλείδειας μετρικής ή του δείκτη Jaccard.

Παράδειγμα

Θεωρούμε 6 αντικείμενα που χαρακτηρίζονται από δύο μεταβλητές. Στον πίνακα 1, η πρώτη μεταβλητή είναι χωρισμένη σε 4 κλάσεις και η δεύτερη σε 2 κλάσεις. Στον πίνακα 2 έχουμε πάλι τα ίδια αντικείμενα για τις ίδιες μεταβλητές, μόνο που τώρα η πρώτη μεταβλητή είναι χωρισμένη σε 3 κλάσεις.

Εφαρμόζουμε την ταξινόμηση κατ' αύξουσα ιεραρχία, με την μέθοδο MOYENNE, χρησιμοποιώντας τη μετρική του X . Τα αποτελέσματα φαίνονται στα αντίστοιχα δενδρογράμματα (σχήματα 1 και 2).

Παρατηρούμε σημαντικές διαφορές στην ταξινόμηση των αντικειμένων, σε σχέση με το πλήθος κλάσεων. Έτσι όταν η πρώτη μεταβλητή χωρίζεται σε 3 κλάσεις το αντικείμενο που διαφέρει από τα υπόλοιπα είναι το a3, ενώ όταν χωριστεί σε 4 κλάσεις διαχωρίζεται το a4 αντικείμενο, το οποίο όπως φαίνεται από τον πίνακα δεδομένων, είναι όντως τελείως διαφορετικό από τα υπόλοιπα.

2δ. Πίνακας m μεταβλητών, χωρίς επαναλήψεις

Γενικεύουμε τα παραπάνω για m μεταβλητές, όπου η μεταβλητή j ($1 \leqslant j \leqslant m$) έχει k_j κλάσεις και όλα τα δυνατά διαφορετικά αντικείμενα είναι σε πλήθος $k_1 k_2 \ldots k_m$.

Ο πίνακας δεδομένων έχει $k_1 k_2 \ldots k_m$ γραμμές και $k_1 + k_3 + \ldots + k_m$ στήλες Το άθροισμα των στοιχείων κάθε γραμμής είναι m, ενώ το άθροισμα μιας στήλης για την j μεταβλητή είναι $k_1 k_2 \ldots k_{j-1} k_{j+1} \ldots k_m$.

Έστω $\alpha_{i1i2 \ldots ij \ldots im}$, το αντικείμενο που χαρακτηρίζεται από την i1-οστή κλάση της 1ης μεταβλητής, την i2-οστή της 2ης, ..., την im-οστή κλάση της m μεταβλητής. Οι αποστάσεις δύο αντικειμένων $\alpha_{i1i2 \ldots ij \ldots im}$, $\alpha_{i1i2 \ldots ij' \ldots im}$, που διαφέρουν μόνο ως προς μια κλάση j μεταβλητής, είναι:

$$d_e\,(a_{i1i2...ij...im}, a_{i1i2...ij...im}) = \sqrt{1^2+1^2} = \sqrt{2}$$

$$d_{X^2}\,(a_{i1i2...ij...im'}, a_{i1i2...ij'...im}) = \sqrt{\frac{2}{k_1k_2...k_{j-1}k_{j+1}...k_m}(\frac{1}{m})^2} = d_x2(m,k_1,...,k_{j-1},k_{j+1},...,k_m)$$

$$d_j(a_{i1i2...ij..im'}, a_{i1i2...ij'...im}) = 1 - \frac{m-1}{m+m-(m-1)} = \frac{2}{m+1} = d_j(m)$$

Αν $\alpha_{i1i2\,...\,ij\,...\,im}$ και $\alpha'_{i1i2\,...\,ij\,...\,im}$ είναι δύο αντικείμενα που διαφέρουν σε h κλάσεις, h μεταβλητών, τότε η απόστασή τους είναι:

$$d_e(a_{i1i2...ij...im'}, a'_{i1i2...ij...im}) = \sqrt{2h} = d_e(h)$$

$$d_x2(a_{i1i2...ij...im'}, a'_{i1i2...ij...im}) = \sqrt{2(\frac{1}{m})^2\sum_{j=1}^{h}\frac{1}{k_1k_2...k_{j-1}k_{j+1}...k_m}} = d_x2(m,h,k_1,...,k_m)$$

$$d_j2(a_{ili2...ij..im'}, a'_{i1i2...ij...im}) = 1 - \frac{m-h}{m+m(m-h)} = \frac{2h}{m+h} = d_j(m,h)$$

Παρατήρηση: Όπως και προηγουμένως, παρατηρούμε ότι στην μετρική του X^2, συντελεί στον υπολογισμό των αποστάσεων και το πλήθος των κλάσεων κάθε μεταβλητής.

Όσο περισσότερες κλάσεις έχει μια μεταβλητή, τόσο μικρότερη είναι η απόσταση δύο σημείων που διαφέρουν ως προς μια κλάση κάποιας άλλης μεταβλητής. Δηλαδή αν μια μεταβλητή την χωρίσουμε σε διαφορετικό αριθμό κλάσεων, η νέα ταξινόμηση θα διαφέρει στους χαμηλούς κόμβους (στους πρωτοσχηματιζόμενους).

Στην περίπτωση που δεν θέλουμε ο αριθμός κλάσεων κάθε μεταβλητής να επηρεάζει την ταξινόμηση, η χρήση της μετρικής X^2, οδηγεί σε εσφαλμένα αποτελέσματα, και ενδείκνυται η χρησιμοποίηση της ευκλείδειας μετρικής ή του δείκτη Jaccard.

Όταν όμως έχει σημασία για την ταξινόμηση και το πλήθος κλάσεων των μεταβλητών, πρέπει να χρησιμοποιείται η μετρική του X^2.

2ε. Πίνακας m μεταβλητών, με επαναλήψεις

Αν τα αντικείμενα που χρακτηρίζονται από m μεταβλητές επαναλαμβάνονται, η απόσταση τους με την ευκλείδεια μετρική ή με τον δείκτη Jaccard δεν

αλλάζει, αλλά ο αριθμός επαναλήψεων επιδρά στον υπολογισμό των αποστάσεων τους με τη μετρική X^2.

Οι αποστάσεις δύο αντικειμένων που διαφέρουν σε h κλάσεις h μεταβλητών στην περίπτωση αυτή, δίνεται από τους τύπους:

$$d_e(a_{ili2...ij..im'}, a'_{i1i2...ij..im}) = \sqrt{2h} = d_e(h)$$

$$d_{X^2}(a_{ili2...ij..im'}, a'_{i1i2...ij..im}) = \sqrt{2(\frac{1}{m})^2 \sum_{j-1}^{h} \frac{1}{g_j k_1 k_2 ... k_{j-1} k_{j+1} ... k_m}} = d_{X^2}(m, h, g_j, k_1, ..., k_m)$$

όπου g_j ο αριθμός επαναλήψεων της j γραμμής.

$$d_J(a_{ili2...ij..im'}, a'_{i1i2...ij...im}) = 1 - \frac{m-h}{m+m(m-h)} = \frac{2h}{m+h} = d_J(m, h)$$

Στην περίπτωση που δεν θέλουμε ο αριθμός των κλάσεων των μεταβλητών ή η επανάληψη ενός αντικειμένου, να συντελεί στην ταξινόμηση του, δεν πρέπει να χρησιμοποιούμε την μετρική του X^2 (που οδηγεί σε εσφαλμένη ιεράρχηση) αλλά την ευκλείδεια μετρική ή τον δείκτη Jaccard.

Όταν όμως έχει σημασία για την ταξινόμηση η επανάληψη κάποιου αντικειμένου ή το πλήθος κλάσεων των μεταβλητών, πρέπει να χρησιμοποιείται η μετρική του X^2.

3. Σύγκριση Ευκλείδειας - Jaccard

Έστω ότι έχουμε πίνακα δεδομένων, όπου τα αντικείμενα χαρακτηρίζονται από ποιοτικές ή και ποσοτικές μεταβλητές χωρισμένες σε κλάσεις.

Όταν κάθε αντικείμενο ανήκει σε μία και μόνο κλάση κάθε μεταβλητής, η διάταξη των αντικειμένων είναι ίδια είτε χρησιμοποιούμε την ευκλείδεια μετρική είτε τον δείκτη Jaccard και οδηγούμαστε ε όμοια ταξινόμηση κατ' αύξουσα ιεραρχία (C.A.H.).

Περίπτωση που τα αντικείμενα μπορούν να ανήκουν σε περισσότερες από μία κλάσεις της ίδιας μεταβλητής.

Στις ειδικές περιπτώσεις που κάθε αντικείμενο είναι δυνατό να ανήκει σε περισσότερες από μία κλάσεις ή και σε καμία κλάση μιας μεταβλητής (π.χ. για

την μεταβλητή ξένη γλώσσα, ένα άτομο μπορεί να μην γνωρίζει καμία ή να γνωρίζει 1 και περισσότερες ξένες γλώσσες), η ευκλείδεια απόσταση μεταξύ δύο αντικειμένων που έχουν σε μια στήλη 1, είναι ίδια με την ευκλείδεια απόσταση αν έχουν στην στήλη αυτή 0. Με τον δείκτη Jaccard όμως, που λαμβάνει υπόψη τις κοινές εμφανίσεις μιας κλάσης (κοινά 1), οι δύο αποστάσεις είναι διαφορετικές.

Έστω ο πίνακας δεδομένων A:

	1 ..... k	άθροισμα γραμμής
1	1 0	A(1,.)
.	1 1	A(2,.)
.	....	.
.	..... .	.
n	0 1	A(n,.)
άθροισμα στήλης	A(.,1) ..... A(.,k)	

Οι αποστάσεις μεταξύ δύο αντικειμένων α_i και $\alpha_{i'}$ όταν c είναι το πλήθος των κοινών 1 στις δύο γραμμές i και i', είναι:

$$d_e(a_i, a_{i'}) = \sqrt{[A(i,.) + A(i',.) - 2c]^2}$$

$$d_J(a_i, a_{i'}) = 1 - \frac{c}{A(i,.) + A(i',.) - c} = \frac{A(i,.) + A(i',.) - 2c}{A(i,.) + A(i,.) - c}$$

Όταν λοιπόν θέλουμε η παρουσία κάποιας κλάσης (1) να χαρακτηρίζει ένα αντικείμενο περισσότερο από την απουσία της (0), στον υπολογισμό της απόστασης, πρέπει να χρησιμοποιούμε τον δείκτη Jaccard και όχι την ευκλείδεια μετρική που εξισώνει τις κοινές παρουσίες και απουσίες.

Παράδειγμα

Έστω τα δεδομένα του πίνακα 3, που αφορούν 5 άτομα που χαρακτηρίζονται από 1 μεταβλητή (ξένη γλώσσα), χωρισμένη σε 4 κλάδεις.

Ταξινομούμε τα άτομα αυτά, εφαρμόζοντας την ταξινόμηση κατ' αύξουσα ιεραρχία, με τη μέθοδο MOYENNE και χρησιμοποιώντας την πρώτη φορά την ευκλείδια μετρική, ενώ τη δεύτερη, τον δείκτη Jaccard.

Τα πρώτα άτομα που ενώνονται και με την ευκλείδεια μετρική και με τον δείκτη Jaccard είναι τα a2 και a4, που γνωρίζουν 2 κοινές ξένες. Κατόπιν με

τον δείκτη Jaccard, ενώνεται το a5 με την ομάδα των a2 και a4, γνωρίζοντας και τα 3 άτομα 2 κοινές ξένες γλώσσες. Με την ευκλείδεια απόσταση όμως, ενώνεται πρώτα το a3 με το a2, a4, γνωρίζοντας μόνο μια κοινή ξένη γλώσσα, και κατόπιν το a5.

Τέλος, τελευταίο στην ταξινόμηση, ενώνεται το a3 με τον δείκτη Jaccard, το οποίο γνωρίζει μόνο μια ξένη γλώσσα, ενώ με την ευκλείδεια απόσταση τελευταία ενώνεται το a1, αν και γνωρίζει 3 ξένες γλώσσες και έχει δύο κοινές με τα υπόλοιπα.

4. Πορεία μετά τον Υπολογισμό των Αποστάσεων

Οι 3 μετρικές απόστασης, χρησιμοποιούνται μόνο στην αρχή της ιεράρχησης, για τον υπολογισμό των αποστάσεων ανά δύο όλων των αντικειμένων. Για να είναι στατιστικά πιο αξιόπιστα τα συμπεράσματα, πρέπει να επιλέγεται η κατάλληλη απόσταση ανάλογα με το είδος των δεδομένων και τους σκοπούς της ανάλυσης.

Το δεύτερο στάδιο της ταξινόμησης, όπως προαναφέραμε, εξελίσσεται εφαρμόζοντας μια από τις μεθόδους MIN, MAX, MOYENEE και υπολογίζοντας την απόσταση μεταξύ ομάδων, με βάση τις ήδη υπάρχουσες αποστάσεις μεταξύ των αντικειμένων. (Η απόσταση δύο ομάδων με την επιλεγείσα μετρική, είναι η ελάχιστη, μέγιστη ή μέση απόσταση των αντικειμένων τους, όπως έχει υπολογισθεί στο προηγούμενο στάδιο).

Αρχικά βρίσκουμε τα δύο «πλησιέστερα» στοιχεία (αυτά που έχουν τη μικρότερη απόσταση μεταξύ τους), και στη θέση τους θέτουμε ένα νέο, με συντεταγμένες (τιμές για τις ρ μεταβλητές), τις συντεταγμένες του κέντρου βάρους των στοιχείων που αντικαθίστανται. Στη συνέχεια υπολογίζουμε τις αποστάσεις μεταξύ του στοιχείου αυτού και των υπολοίπων, με μία από τις ακόλουθες 3 μεθόδους: MIN, MAX, MOYENNE.

Αν i και i΄ είναι τα δύο συνενούμενα στοιχεία σε μια κλάση, που συμβολίζεται iUi και K ένα άλλο στοιχείο, η απόσταση του k από τη κλάση αυτή είναι:

MIN $\quad d_m(iUi', k) = \min \{ d(i, k), d(i', k)\}$

MAX $\quad d_M(iUi', k) = \max \{d(i, k), d(i', k)\}$

MOYENNE $\quad d_\mu (iUi', k) = \dfrac{p(i)d(i,k) + p(i')d(i,k)}{p(i) + p(i',)}$

(όπου p(x) είναι το πλήθος στοιχείων της κλάσης x)

Συνεχίζουμε έτσι, ενώνοντας διαδοχικά τις δύο πλησιέστερες κλάσεις, μέχρι να καταλήξουμε σε δύο μόνο, τις οποίες ενώνουμε σε μια τελική κλάση, που περιέχει όλα τα δεδομένα.

Η ταξινόμηση κατά αύξουσα ιεραρχία (C.A.H.), σχηματικά παριστάνεται με ένα δενδρόγραμμα, όπου κάθε κόμβος αντιστοιχεί σε μία κλάση (ομάδα παρατηρήσεων) και στο επόμενο επίπεδο, χωρίζεται σε δύο υποκλάσεις (υποομάδες) στις κλάσεις που ενώθηκαν για να σχηματίσουν τον κόμβο). Η κάθε μία από τις υποκλάσεις χωρίζεται με τη σειρά της σε δύο υποκλάσεις μέχρι να καταλήξουμε στους τερματικούς κόμβους οι οποίοι αντιστοιχούν στις αρχικές παρατηρήσεις και εμφανίζονται μόνο μία φορά.

5. Συμπεράσματα

Όταν η επανάληψη ενός αντικειμένου παίζει ή επιθυμούμε να παίζει σημαντικό ρόλο στην ταξινόμηση του, πρέπει να χρησιμοποιούμε την μετρική του X^2. Με την μετρική αυτή, αντικείμενα που επαναλαμβάνονται πολλές φορές, ταξινομούνται γρηγορότερα, και στα πρώτα επίπεδα στο δενδρόγραμμα ιεράρχησης.

Όταν δεν θέλουμε να λάβουμε υπ' όψιν την επανάληψη κάποιου αντικειμένου ή τον αριθμό κλάσεων που έχει κάθε μεταβλητή, πρέπει να χρησιμοποιήσουμε σαν μετρική απόστασης, είτε την ευκλείδεια μετρική, είτε τον δείκτη απόστασης του Jaccard, και όχι τη μετρική του X^2, γιατί όταν οι μεταβλητές έχουν διαφορετικό πλήθος κλάσεων, στον υπολογισμό των αποστάσεων με τη χρήση της μετρικής X^2, επομένως και στη δημιουργία της ιεράρχησης, επιδρά το πλήθος κλάσεων.

Η επιλογή μεταξύ ευκλείδειας απόστασης και δείκτη Jaccard, εξαρτάται από τον αριθμό των κλάσεων στις οποίες μπορεί να ανήκει κάθε αντικείμενο. Αν οποιοδήποτε αντικείμενο, ανήκει σε μόνο μία κλάση κάθε μεταβλητής, τότε η ιεράρχηση που παίρνουμε με τον δείκτη Jaccard, είναι όμοια μ' αυτή που λαμβάνουμε χρησιμοποιώντας την ευκλείδεια απόσταση. Όταν όμως ένα αντικείμενο μπορεί να χαρακτηρίζεται από περισσότερες κλάσεις της ίδιας μεταβλητής, και η παρουσία μιας κλάσης είναι σημαντικότερη (στον χαρακτηρισμό του αντικειμένου) απ' ότι η απουσία της, η χρήση της ευκλείδειας απόστασης οδηγεί σε λάθη, αφού θεωρεί τις κοινές παρουσίες και απουσίες, ισοδύναμες. Στην περίπτωση αυτή πρέπει να χρησιμοποιούμε τον δείκτη του Jaccard.

ΠΙΝΑΚΑΣ 1
Πίνακας δεδομένων 6x6

άτομο	μεταβλ.1				μεταβλ.2	
a1	0	0	1	0	1	0
a2	0	1	0	0	1	0
a3	1	0	0	0	1	0
a4	0	0	0	1	0	1
a5	0	0	0	1	1	0
a6	0	0	1	0	1	0

ΠΙΝΑΚΑΣ 2
Πίνακας δεδομένων 6x5

άτομο	μεταβλ.1			μεταβλ.2	
a1	0	0	1	1	0
a2	0	1	0	1	0
a3	1	0	0	1	0
a4	0	0	1	0	1
a5	0	0	1	1	0
a6	0	0	1	1	0

ΠΙΝΑΚΑΣ 3
Το σύνολο δεδομένων για τα 5 άτομα ως προς τις 4 κλάσεις της μεταβλητής Ξένης γλώσσης

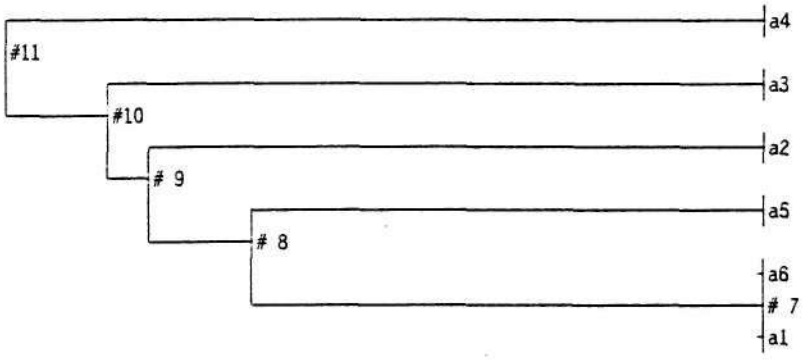
άτομο	Ιταλικά	Γερμανικά	Αγγλικά	Γαλλικά
a1	1	1	0	1
a2	1	0	1	1
a3	0	0	0	1
a4	0	0	1	1
a5	0	1	1	1

ΠΙΝΑΚΑΣ 4
Κάμβοι της ταξινόμησης των 5 ατόμων
με την ευκλείδια μετρική
(μέθοδος MOYENNE)

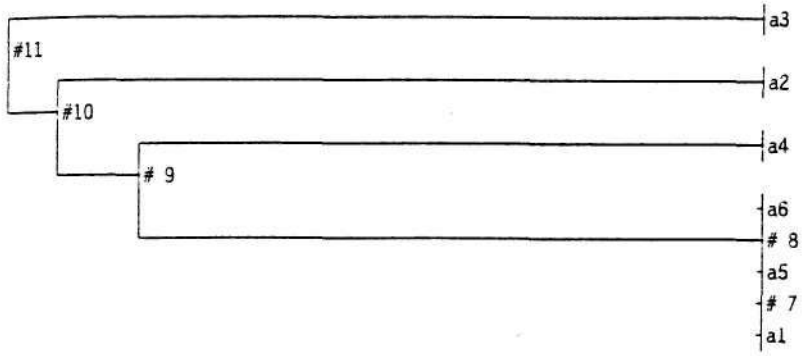
κόμβος	αριστερό	δεξιό	βάρος	απόσταση
6	a4	a2	2	1.000000
7	a3	6	3	1.207107
8	a5	7	4	1.276142
9	8	a1	5	1.493673

ΠΙΝΑΚΑΣ 5
Κάμβοι της ταξινόμησης των 5 ατόμων
με τον δείκτη Jaccard
(μέθοδος MOYENNE)

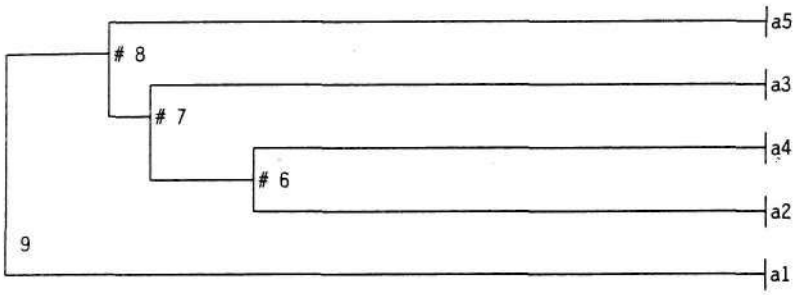
κόμβος	αριστερό	δεξιό	βάρος	απόσταση
6	a4	a2	2	0.333333
7	a5	6	3	0.416667
8	7	7	4	0.583333
9	a3	8	5	0.625000



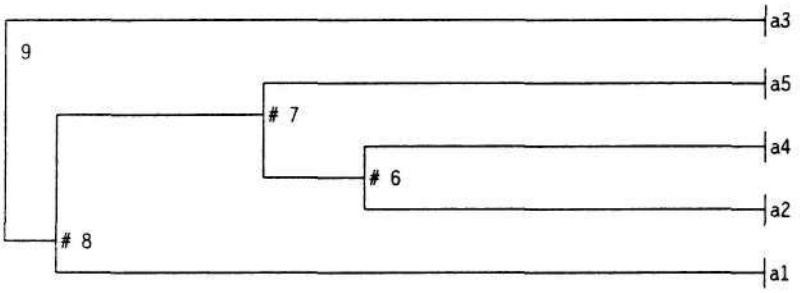
Σχήμα 1. Δενδρογράμματα ιεράρχησης των 6 ατόμων (για 6 κλάσεις) με την χ^2 μετρική (μέθοδος MOYENNE).



Σχήμα 2. Δενδρόγραμμα ιεράρχησης των 6 ατόμων (για 5 κλάσεις) με την χ^2 μετρική (μέθοδος MOYENNE).



Σχήμα 3. Δενδρόγραμμα ιεράρχησης των 7 ατόμων με την ευκλείδια μετρική (μέθοδος MOYENNE).



Σχήμα 4. Δενδρόγραμμα ιεράρχησης των 5 ατόμων με τον δείκτη Jaccard (μέθοδος MOYENNE).

Βιβλιογραφία

Benzecri J. P. et Collaborateurs, (1973), L'Analyse des donnees, Vol. 1, Dunod, Paris.

Benzecri J. P et Collaborateurs, (1980), Pratique de l'Analyse des Donnees, Vol. 1, Dunod, Paris.

Παπαδημητρίον Γ. Φλώρου Γ., (1993), Προσδιορισμός της ιδανικότερης μεθόδου ιεράρχησης μεταξύ των MIN, MAX, MOYENNE, για κάθε πίνακα δεδομένων, 6ο Πανελλήνιο Συνέδριο Στατιστικής, Θεσσαλονίκη.

Παπαδημητρίον Γ., Φλώρου Γ., (1994), Συμβολή της ευκλείδειας και X^2 μετρικής στον προσδιορισμό της ιδανικότερης ταξινόμησης κατ' αύξουσα ιεραρχία, Τιμητικός τόμος καθηγητή κ. Ι. Λιάκη, Θεσσαλονίκη.

Roux M., (1985), Algorithmes de classification, Masson, Paris.