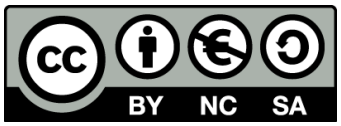
ΑΡΙΣΤΟΤΕΛΕΙΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΟΝΙΚΗΣ

ΑΝΟΙΚΤΑ
ΑΚΑΔΗΜΑΪΚΑ
ΜΑΘΗΜΑΤΑ

Αποθήκες Δεδομένων και Εξόρυξη Δεδομένων

Ενότητα 10: Ομαδοποίηση – Μέρος Δ΄

Αναστάσιος Γούναρης, Επίκουρος Καθηγητής
Τμήμα Πληροφορικής ΑΠΘ



Άδειες Χρήσης

- Το παρόν εκπαιδευτικό υλικό υπόκειται σε άδειες χρήσης Creative Commons.
- Για εκπαιδευτικό υλικό, όπως εικόνες, που υπόκειται σε άλλου τύπου άδειας χρήσης, η άδεια χρήσης αναφέρεται ρητώς.

Χρηματοδότηση

- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στα πλαίσια του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο «Ανοικτά Ακαδημαϊκά Μαθήματα στο Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης» έχει χρηματοδοτήσει μόνο τη αναδιαμόρφωση του εκπαιδευτικού υλικού.
- Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.

Ομαδοποίηση – Μέρος Δ΄

Έλεγχος εγκυρότητας, τάση ομαδοποίησης

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Περιεχόμενα ενότητας

1. Έλεγχος εγκυρότητας και τάση ομαδοποίησης.
2. Γενικά θέματα κλιμάκωσης.
3. Ποιός αλγόριθμος πρέπει να επιλέγεται.

Σκοποί ενότητας

- Εξέταση θεμάτων που αφορούν τον έλεγχο εγκυρότητας και της τάσης ομαδοποίησης.
- Παρουσιάζονται κριτήρια και παράγοντες που καθορίζουν την επιλογή του αλγορίθμου ομαδοποίησης.

Εγκυρότητα ομαδοποίησης

- Πώς αξιολογείται η ποιότητα των συστάδων;
- Πολλές φορές το πόσο καλό είναι το αποτέλεσμα είναι υποκειμενικό.
  - “Clusters are in the eye of the beholder”!
- Γιατί χρειάζεται η αξιολόγηση:
  - Για την αποφυγή εύρεσης προτύπων σε θόρυβο.
  - Για τη σύγκριση διαφορετικών αλγορίθμων.
  - Για τη σύγκριση δύο συνόλων από συστάδες.
  - Για τη σύγκριση δύο συστάδων.
- Αποτίμηση αποτελέσματος ομαδοποίησης:
  - Εσωτερικά κριτήρια (καμία γνώση κλάσεων).
  - Εξωτερικά κριτήρια (γνώση κλάσεων).
  - Σχετικά κριτήρια.

Εγκυρότητα αλγορίθμων τμηματοποίησης

- Συμπάγεια (Συνοχή)
  - Πόσο κοντά ή στενά συσχετισμένα είναι τα αντικείμενα μίας συστάδας.

$$SSE = \sum_{i} \sum_{x \in C_i} (x - m_i)^2$$

    - m_i είναι το κέντρο της συστάδας i
- Απομόνωση
  - Πόσο μακριά ή καλά διαχωρισμένες είναι οι συστάδες μεταξύ τους.

$$SSB = \sum_{i} |C_i| (m - m_i)^2$$

    - $|C_i|$ είναι το μέγεθος της συστάδας i
    - m το κέντρο όλων των συστάδων

Αλληλεξάρτηση SSE και SSB (1/2)

K=1 cluster:

$$SSE = (1-3)^2 + (2-3)^2 + (4-3)^2 + (5-3)^2 = 10$$

$$SSB = 4 \times (3-3)^2 = 0$$

$$Total = 10 + 0 = 10$$

K=2 clusters:

$$SSE = (1-1.5)^2 + (2-1.5)^2 + (4-4.5)^2 + (5-4.5)^2 = 1$$

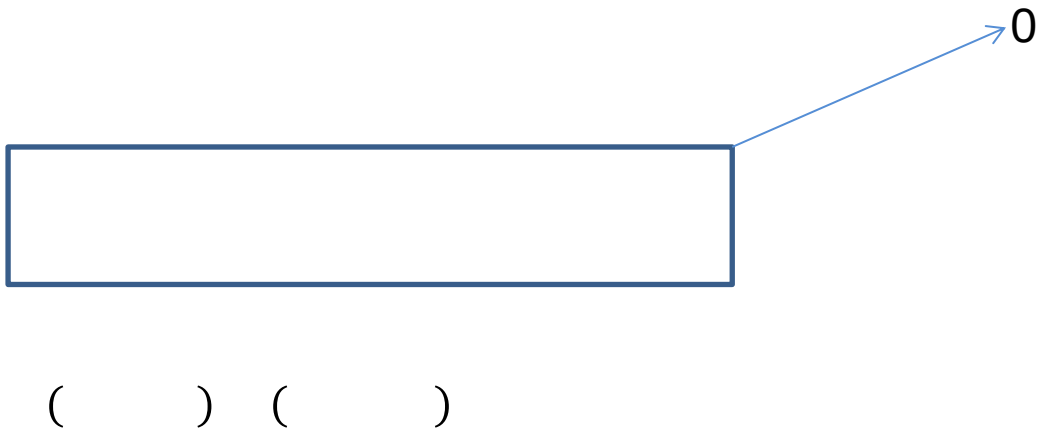
$$SSB = 2 \times (3-1.5)^2 + 2 \times (4.5-3)^2 = 9$$

$$Total = 1 + 9 = 10$$

- Αποδεικνύεται ότι πάντα το άθροισμα παραμένει σταθερό.
- Άρα η ελαχιστοποίηση του SSE σημαίνει μεγιστοποίηση του SSB.

Αλληλεξάρτηση SSE και SSB (2/2)

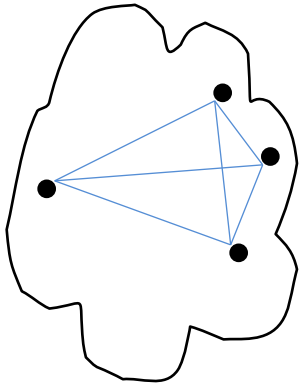
- Απόδειξη:



Περισσότερα για συνοχή-απομόνωση

- Η συνοχή μπορεί να υπολογιστεί αν πάρουμε την απόσταση των σημείων της συστάδας μεταξύ τους.

$$cluster-SSE=\sum_{x\in C_i}(x-m_i)^2=\frac{1}{2c_i}\sum_{x\in C_i}\sum_{y\in C_i}(x-y)^2$$



- Η απομόνωση μπορεί να υπολογιστεί αν πάρουμε την απόσταση μεταξύ των κέντρων των συστάδων.
  - Για ισομεγέθεις συστάδες:

$$ολικό-SSB=\sum_{x\in C_i}c_i\,(m_i-m)^2=\frac{1}{2K}\sum_{i=1}^{K}\sum_{j=1}^{K}\frac{n}{K}\,(m_i-m_j)^2$$

Έλεγχος σημαντικότητας SSE και SSB

- Οι τιμές SSE και SSB ερμηνεύονται μόνο συγκριτικά.
- Παράγουμε έναν αριθμό συνόλων με αντικείμενα που ακολουθούν ομοιόμορφη κατανομή και έχουν τον ίδιο αριθμό αντικειμένων
- Ομαδοποιούμε κάθε τέτοιο σύνολο και υπολογίζουμε τις τιμές SSE και SSB.
- Εξετάζουμε την κατανομή των τιμών SSE και SSB .
- Ελέγχουμε κατά πόσο μη αναμενόμενες είναι οι τιμές SSE και SSB που προέκυψαν από την υπό εξέταση ομαδοποίηση.

Συντελεστής σιλουέτας (αλγόριθμοι τμηματοποίησης)

- Συνδυασμός συμπάγειας και απομόνωσης
  - Απόσταση ενός αντικειμένου από τα αντικείμενα της ομάδας του συγκριτικά με την απόσταση από τα αντικείμενα άλλων ομάδων
- α_i = μέση απόσταση i-οστού αντικειμένου από τα αντικείμενα της ομάδας του
- b_i = η ελάχιστη μέση απόσταση i-οστού αντικειμένου από αντικείμενα άλλης ομάδας

$$S_i = \frac{b_i - a_i}{\max\{a_i, b_i\}}$$

- Τιμή στο [-1,1], επιθυμητές τιμές > 0

Εγκυρότητα για ιεραρχικούς αλγορίθμους

- Συντελεστής CPCC (cophenetic correlation coefficient) μετράει κατά πόσο τα αντικείμενα ταιριάζουν στο (πλήρες) δενδρόγραμμα.
- d(x,y): απόσταση x, y.
- $d_c(x,y)$: απόσταση των ομάδων που περιείχαν τα x και y, όταν αυτά τοποθετήθηκαν για πρώτη φορά στην ίδια ομάδα, έπειτα από συγχώνευση των ομάδων τους.

$$\frac{(\quad) \qquad (\quad)}{\sqrt{(\quad)} \quad \sqrt{(\quad)}}$$

- Οι τιμές είναι στο διάστημα [-1,1].
  - Οι τιμές πλησιέστερα στο 1 υποδηλώνουν καλύτερη συσταδοποίηση.

Γενική μέθοδος εγκυρότητας μέσω συσχέτισης (1/2)

- Δύο πίνακες:
  - Πίνακας Γειτνίασης (Ομοιότητας ή απόστασης).
  - Πίνακας Εμφάνισης (Incidence).
    - Μία γραμμή και μία στήλη για κάθε σημείο.
    - Τιμή 1 σημαίνει ότι τα αντίστοιχα σημεία βρίσκονται στην ίδια συστάδα.
    - Τιμή 0 σημαίνει ότι τα αντίστοιχα σημεία βρίσκονται σε διαφορετική συστάδα.

Γενική μέθοδος εγκυρότητας μέσω συσχέτισης (2/2)

- Υπολογισμός συσχέτισης μεταξύ των δύο πινάκων.
  - Καθώς οι πίνακες είναι συμμετρικοί, χρειάζεται ο υπολογισμός μόνο
    n(n-1) / 2 συνδυασμών.
- Υψηλή συσχέτιση σημαίνει ότι τα σημεία μιας συστάδας είναι κοντά το ένα στο άλλο.
- Δεν ενδείκνυται για συσταδοποίηση βάσει πυκνότητας ή συνέχειας.

Αποτίμηση μέσω οπτικοποίησης

- Ταξινομούμε τα αντικείμενα ως προς την ομάδα που ανήκουν και αναπαριστούμε τις αποστάσεις τους σε έναν πίνακα
- Οι συστάδες σε τυχαία δεδομένα δεν είναι ξεκάθαρες

Εγκυρότητα με εξωτερικά κριτήρια

- Γνωρίζουμε εκ των προτέρων την κλάση στην οποία ανήκουν.
- a: Το πλήθος των αντικειμένων που ανήκουν και στην ίδια ομάδα και στην ίδια κλάση.
- b: Το πλήθος των αντικειμένων που ανήκουν και στην ίδια ομάδα αλλά όχι στην ίδια κλάση.
- c: Το πλήθος των αντικειμένων που δεν ανήκουν στην ίδια ομάδα αλλά ανήκουν στην ίδια κλάση.
- Συντελεστής Jaccard: $$\frac{a}{a+b+c}$$
  - Ο συντελεστής Jaccard παίρνει τιμές στο διάστημα [0,1]. Τιμές πλησιέστερα στο 1 δηλώνουν καλύτερη ομαδοποίηση.

Τελευταίο σχόλιο πάνω στην εγκυρότητα

"The validation of clustering structures is the most difficult and frustrating part of cluster analysis. Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage."

Algorithms for Clustering Data, Jain and Dubes

Έλεγχος βάσει MST

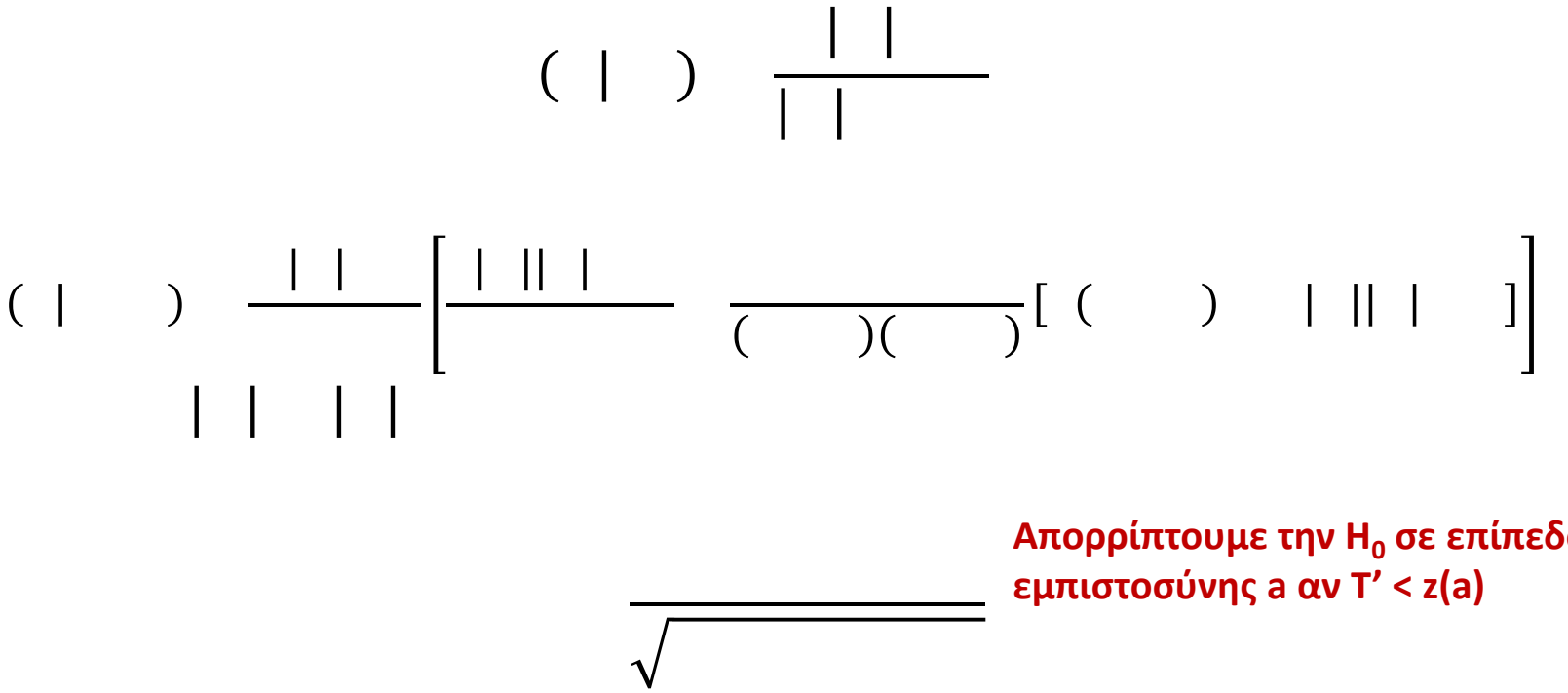
- Παράγουμε σημεία με τυχαία κατανομή
- Έστω η υπόθεση H_0 ότι τα σύνολα δεδομένων προέρχονται από τον ίδιο πληθυσμό.

Βήματα ελέγχου (1/2)

- Βρες την κυρτή (convex) περιοχή που περικλείει όλα τα σημεία του συνόλου δεδομένων X.
- Δημιούργησε σύνολο δεδομένων Y με |Y|=|X| σημεία, που βρίσκονται εντός της κυρτής περιοχής του βήματος 1.
- Βρες το MST του συνόλου X ∪ Y.
- T είναι ο αριθμός των ακμών του MST, των οποίων η μία τους κορυφή ανήκει στο X και η άλλη στο Y.
- Απόρριψε την υπόθεση H_0, αν ο αριθμός T είναι μικρός.

Βήματα ελέγχου (2/2)

- Εναλλακτικά:



Απορρίπτουμε την H_0 σε επίπεδο εμπιστοσύνης a αν T' < z(a)

Έλεγχος βάσει κοντινότερων γειτόνων

- Βρες την κυρτή (convex) περιοχή που περικλείει όλα τα σημεία του συνόλου δεδομένων X.
- Δημιούργησε το σύνολο δεδομένων Y με |Y|<<|X| σημεία που βρίσκονται εντός της κυρτής περιοχής του βήματος 1.
- Πάρε ένα τυχαίο δείγμα S από σημεία του X, όπου |S|=|Y|.
- Θέτουμε ως X'=X-S το σύνολο των σημείων που δεν επιλέχθηκαν στο δείγμα.
- Για κάθε σημείο y_i στο Y βρίσκουμε την απόσταση u_i από τον κοντινότερο γείτονα που ανήκει στο X. Ομοίως, για κάθε σημείο s_i στο S βρίσκουμε την απόσταση w_i από τον κοντινότερο γείτονα που ανήκει στο X'.
- Υπολόγισε το στατιστικό του Hopkins:
  - Υπάρχουν συστάδες όταν H > 0.5

$$H = \frac{\sum_{i=1}^{|S|} u_i}{\sum_{i=1}^{|S|} w_i + \sum_{i=1}^{|S|} u_i}$$

- Απόρριψε την υπόθεση H_0, αν το H είναι περίπου 0.5.

Θέματα που πρέπει να προσεχθούν εκτός της πολυπλοκότητας

- Απαιτήσεις σε μνήμη.
- Αποδοτικός υπολογισμός πλησιέστερου κέντρου, γειτόνων ή σημείων μέσα σε δεδομένη απόσταση.
- Όρια στην γειτονικότητα:
  - μείωση του αριθμού συγκρίσεων ενός σημείου με κεντρικά σημεία
- Δειγματοληψία.
- Τεχνικές διαίρει και βασίλευε.
- Περίληψη των δεδομένων.
- Παραλληλισμός.

Κριτήρια-παράγοντες επιλογής αλγορίθμου (1/2)

- Τύπος συσταδοποίησης:
  - ιεραρχίες, πληρότητα, μοναδικότητα
- Τύπος συστάδων:
  - Βάσει πρωτοτύπου-κεντρικού σημείου.
  - Βάσει γράφου.
  - Βάσει πυκνότητας.
  - Άλλα κριτήρια, όπως αποδοτικότητα σε συστάδες διαφορετικού μεγέθους, σχήματος, πυκνότητας, δυνατότητα υπολογισμού κεντρικού σημείου, συσταδοποίηση σε υποσύνολο χαρακτηριστικών.

Κριτήρια-παράγοντες επιλογής αλγορίθμου (2/2)

- Θόρυβος και ανώμαλες-ακραίες τιμές.
- Περιγραφή συστάδας.
- Ντετερμινιστική συμπεριφορά.
- Ευαισθησία στη σειρά των δεδομένων.
- Αυτόματος υπολογισμός αριθμού συστάδων.
- Ύπαρξη αντικειμενικής συνάρτησης.

Σημείωμα Αναφοράς

Σημείωμα Αδειοδότησης

Τέλος ενότητας

Επεξεργασία: Ανδρέας Κοσματόπουλος

Θεσσαλονίκη, Χειμερινό Εξάμηνο 2013-2014

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Σημειώματα

Διατήρηση Σημειωμάτων

Οποιαδήποτε αναπαραγωγή ή διασκευή του υλικού θα πρέπει να συμπεριλαμβάνει:

- το Σημείωμα Αναφοράς
- το Σημείωμα Αδειοδότησης
- τη δήλωση Διατήρησης Σημειωμάτων
- το Σημείωμα Χρήσης Έργων Τρίτων (εφόσον υπάρχει)

μαζί με τους συνοδευόμενους υπερσυνδέσμους.