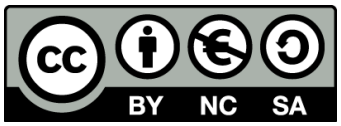
ΑΡΙΣΤΟΤΕΛΕΙΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΟΝΙΚΗΣ

ΑΝΟΙΚΤΑ
ΑΚΑΔΗΜΑΪΚΑ
ΜΑΘΗΜΑΤΑ

Αποθήκες Δεδομένων και Εξόρυξη Δεδομένων

Ενότητα 8: Ομαδοποίηση – Μέρος Β΄

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 την 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.

ΑΡΙΣΤΟΤΕΛΕΙΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΟΝΙΚΗΣ

ΑΝΟΙΚΤΑ
ΑΚΑΔΗΜΑΪΚΑ
ΜΑΘΗΜΑΤΑ

Ομαδοποίηση – Μέρος Β΄

Αλγόριθμοι τμηματοποίησης, ιεραρχικοί αλγόριθμοι

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. Αλγόριθμοι ομαδοποίησης:
   i. Αλγόριθμοι τμηματοποίησης.
   ii. Ιεραρχικοί αλγόριθμοι.

Σκοποί ενότητας

- Ανάλυση αλγορίθμων τμηματοποίησης.
- Ανάλυση ιεραρχικών αλγορίθμων.
- Παρουσίαση ομαδοποίησης πλησιέστερων γειτόνων.
- Παρουσίαση συναθροιστικής ιεραρχικής ομαδοποίησης.

Αλγόριθμοι Τμηματοποίησης (Partitioning Algorithms)

- **Μεθοδολογία**: Δημιουργία τμηματοποίησης μίας ΒΔ ***D*** με ***n*** αντικείμενα σε ένα σύνολο ***k*** συστάδων.
- Δεδομένου του *k*, πρέπει να βρεθεί η τμηματοποίηση που βελτιστοποιεί το κριτήριο τμηματοποίησης.
  - Βέλτιστη λύση: πρέπει να εξεταστούν όλες οι περιπτώσεις.
  - Ευρετικές μέθοδοι: *k-means, k-medoids , k-nn* .
  - ***k-means*** (MacQueen'67): Κάθε συστάδα αντιπροσωπεύεται από το κέντρο της.
  - ***k-medoids*** or PAM (Partition around medoids) (Kaufman & Rousseeuw'87): Κάθε συστάδα αντιπροσωπεύεται από ένα αντικείμενό της.

K-means

1. Διάλεξε k τυχαία κέντρα. (Τα κέντρα μπορεί να μην αντιστοιχούν σε ένα από τα δεδομένα αντικείμενα).
2. Ανάθεσε κάθε αντικείμενο στο πλησιέστερο προς αυτό κέντρο.
3. Για κάθε μία από τις k ομάδες, υπολόγισε το νέο κέντρο.
4. Αν όλα τα νέα κέντρα συμπίπτουν με τα προηγούμενα (δηλαδή, δεν υπήρξε μεταβολή), τότε τερμάτισε γιατί ο αλγόριθμος έχει συγκλίνει. Αλλιώς, επανάλαβε το βήμα 2.

Χαρακτηριστικά K-means

Πλεονεκτήματα

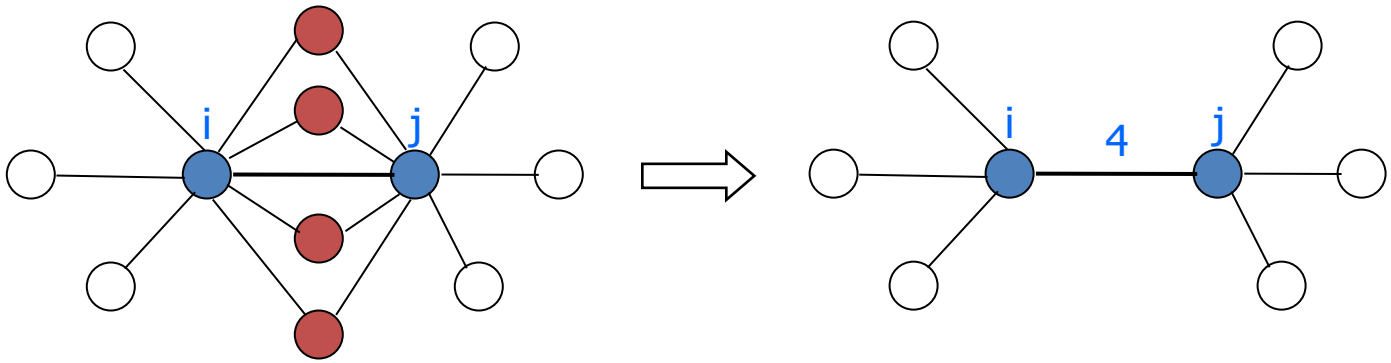
- Απλός.
- Γρήγορος (O(nd)).

Μειονεκτήματα

- Επιλογή k.
- Τοπικά ελάχιστα.
- Ευαισθησία σε θόρυβο, outliers.
- Χαμηλή αποτελεσματικότητα όταν οι πραγματικές συστάδες είναι διαφορετικού μεγέθους, πυκνότητας, ή δεν έχουν σφαιρικό σχήμα.
  - Χρειάζονται αυξημένο k.

Ομαδοποίηση πλησιέστερων γειτόνων (1/3)

- 2 σημεία στην ίδια ομάδα αν μοιράζονται αρκετούς πλησιέστερους γείτονες.



Ομαδοποίηση πλησιέστερων γειτόνων (2/3)

- Για κάθε αντικείμενο x_i υπολόγιζε τη λίστα $L(x_i)$ με τα k πλησιέστερα αντικείμενα.
- Για κάθε ζεύγος σημείων x_i και x_j, **αν** $L(x_i) \cap L(x_j) \geq k_t$:
  - τότε τοποθέτησε τα x_i και x_j στην ίδια ομάδα.

Ομαδοποίηση πλησιέστερων γειτόνων (3/3)

- Οι τιμές των παραμέτρων ***k*** και k_t προσδιορίζονται εμπειρικά.
  - Ο αλγόριθμος είναι ευαίσθητος στο ***k.***
  - Όταν το ***k*** παίρνει μεγάλες τιμές, ο αλγόριθμος τείνει να συνενώνει ομάδες,
  - ενώ αντιθέτως, για μικρές τιμές του ***k*** τείνει να προκαλεί διασπάσεις.
  - Σχετικά με το k_t όταν $k_t = 1$, τότε ο αλγόριθμος παράγει παρόμοιες ομάδες με αυτές του ιεραρχικού αλγορίθμου μονής σύνδεσης (δες επόμενες διαφάνειες), καθώς είναι ευκολότερο να γίνουν συνδέσεις αντικειμένων.
- Πολυπλοκλότητα $O(n^2)$.

Ομαδοποίηση αμοιβαίας γειτνίασης

1. Για κάθε αντικείμενο x_i υπολόγιζε τη λίστα $L(x_i)$ με τα k πλησιέστερα αντικείμενα.
2. Για κάθε ζεύγος αντικειμένων x_i και x_j, υπολόγισε την τιμή $ΑΓ(x_i, x_j)$.
3. Εντόπισε όλα τα ζεύγη αντικειμένων με $ΑΓ$ ίσο με 2. Ανάθεσε όλα αυτά τα ζεύγη αντικειμένων στην ίδια ομάδα.
4. Επανάλαβε το βήμα 3 για την επόμενη τιμή ΑΓ μέχρι την τιμή $2k$.

x_j είναι το είναι το p-οστό αντικείμενο προς το x_i.

x_i είναι το είναι το q-οστό αντικείμενο προς το x_j.

Αν $p,q < k$, $ΑΓ(x_i, x_j) = p + q$, αλλιώς άπειρο.

Βασικές Δομές

- Πίνακας Δεδομένων

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

- Πίνακας απόστασης

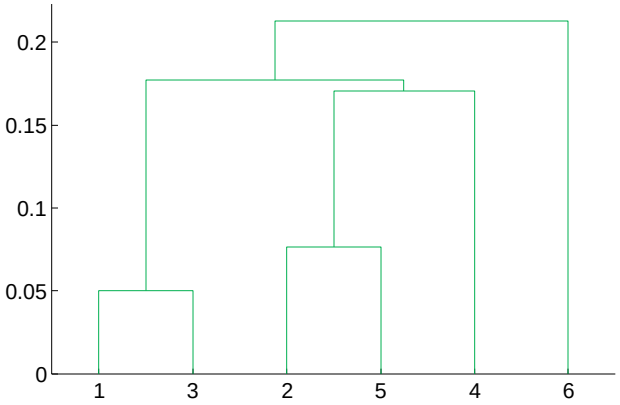
$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

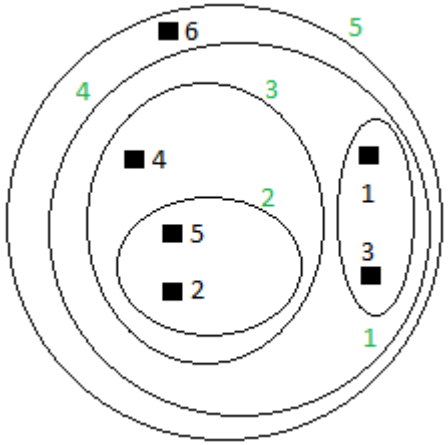
Ιεραρχική Συσταδοποίηση

- Δυο βασικοί τύποι ιεραρχικής συσταδοποίησης:
  - Συσσωρευτικός (Agglomerative):
    - Αρχίζει με τα σημεία ως ξεχωριστές συστάδες.
    - Σε κάθε βήμα, συγχωνεύει το πιο κοντινό ζευγάρι συστάδων μέχρι να μείνει μόνο μία (ή k) συστάδες.
  - Διαιρετικός (Divisive):
    - Αρχίζει με μία συστάδα που περιέχει όλα τα σημεία.
    - Σε κάθε βήμα, διαχωρίζει μία συστάδα, έως κάθε συστάδα να περιέχει μόνο ένα σημείο (ή να δημιουργηθούν k συστάδες.

Ιεραρχική Συσταδοποίηση: Εισαγωγικά

- Παράγει ένα σύνολο από εμφωλευμένες συστάδες οργανωμένες σε ένα ιεραρχικό δέντρο.
- Μπορεί να παρασταθεί με ένα δενδρόγραμμα.
  - Δηλ. ένα διάγραμμα που μοιάζει με δένδρο και καταγράφει τις ακολουθίες από συγχωνεύσεις (merges) και διαχωρισμούς (splits).

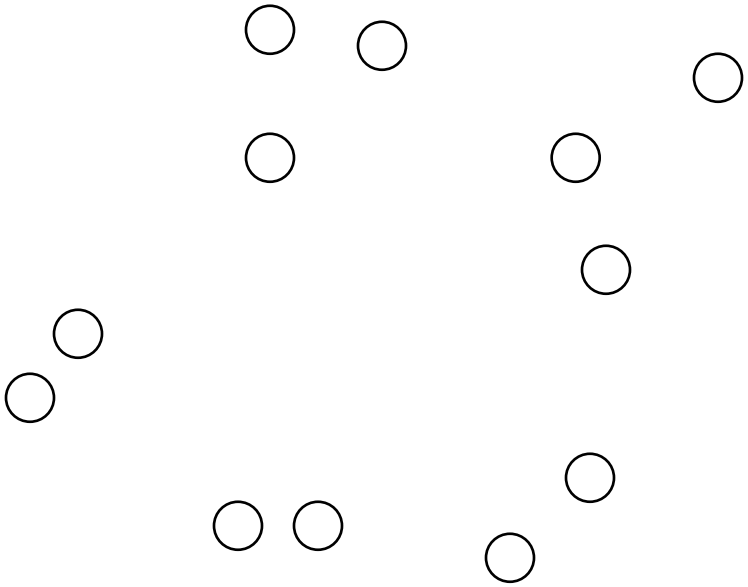




Συναθροιστική Ιεραρχική Ομαδοποίηση

1. Ξεκίνα από την αρχική διαμέριση $C_1 = \{\{x_1\},\ldots,\{x_n\}\}$. Επίσης $i=1$.
2. Βρες το ζεύγος ομάδων c_r και c_s με τη μικρότερη απόσταση $d(c_r, c_s)$.
3. Συνένωσε τις ομάδες c_r και c_s και δημιούργησε την ακολουθία C_{i+1}. Θέσε $i=i+1$.
4. Ενημέρωσε τον πίνακα D, διαγράφοντας τις γραμμές και τις στήλες που αντιστοιχούν στις ομάδες c_r και c_s . Θεώρησε μία γραμμή και στήλη, που να αντιστοιχούν στη συνενωμένη ομάδα $c_r \cup c_s$. Υπολόγισε την απόσταση της συνενωμένης ομάδα $c_r \cup c_s$ από όλες τις υπόλοιπες ομάδες, και συμπλήρωσε τις τιμές στη νέα στήλη και τη νέα γραμμή του D.
5. Αν στο βήμα 3 έχει παραχθεί διαμέριση με k ομάδες, τότε σταμάτησε. Διαφορετικά, επανάλαβε το βήμα 2.

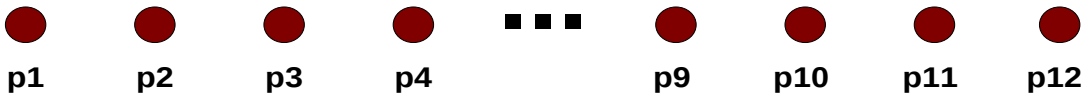
Αρχική κατάσταση



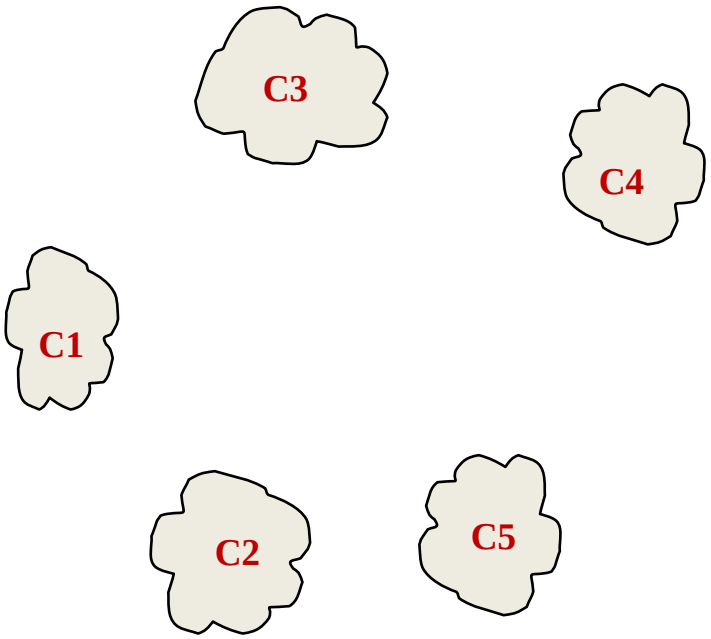
	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Πίνακας Αποστάσεων

- Κάθε σημείο είναι ένα cluster:

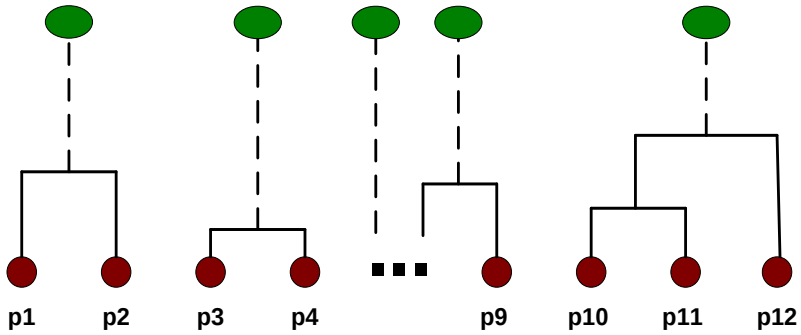


Ενδιάμεση κατάσταση

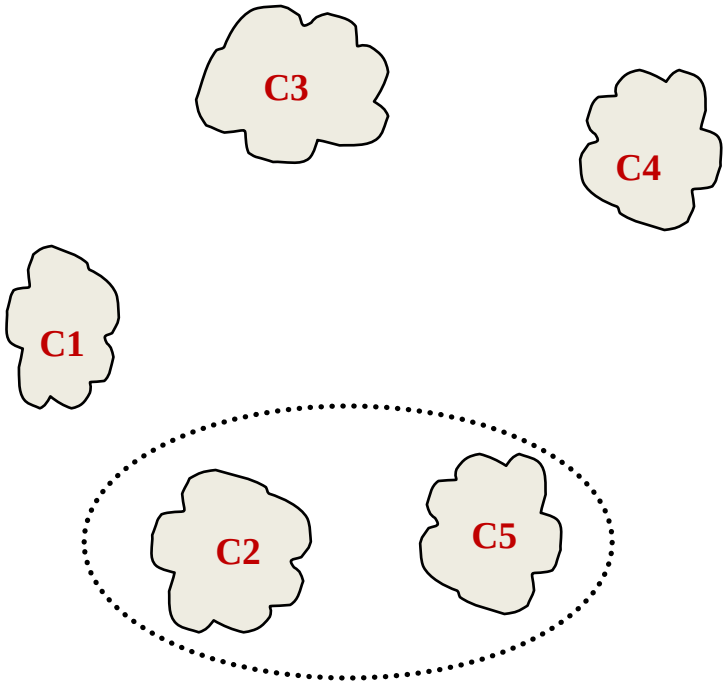


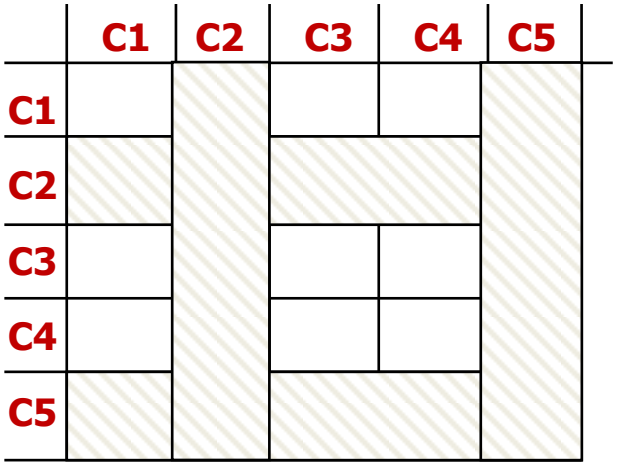
	C1	C2	C3	C4	C5	. . .
C1						
C2						
C3						
C4						
C5						
.						
.						
.						

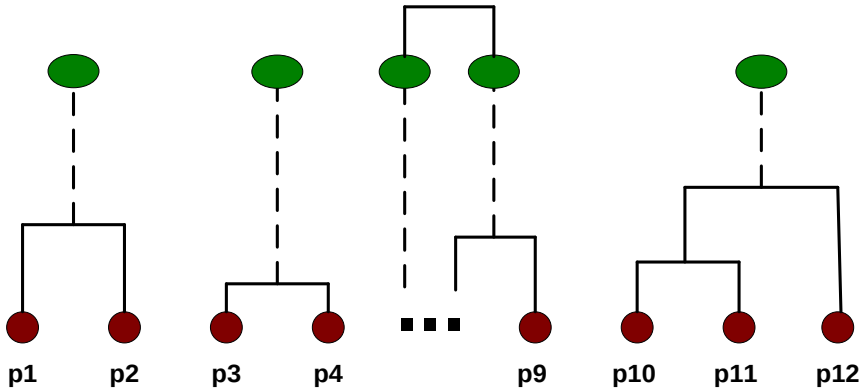
Πίνακας Αποστάσεων



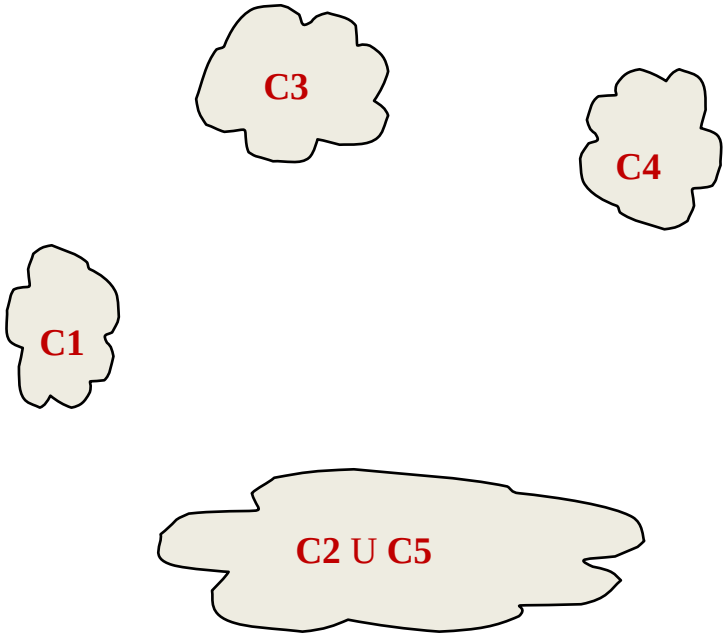
Συνένωση ομάδων



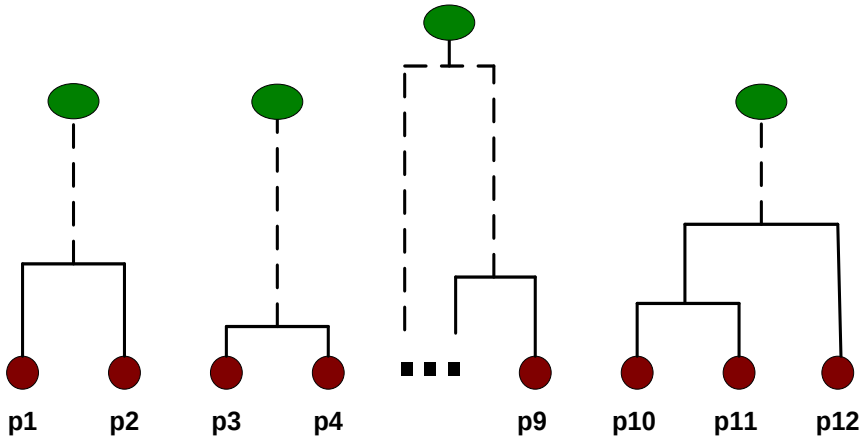




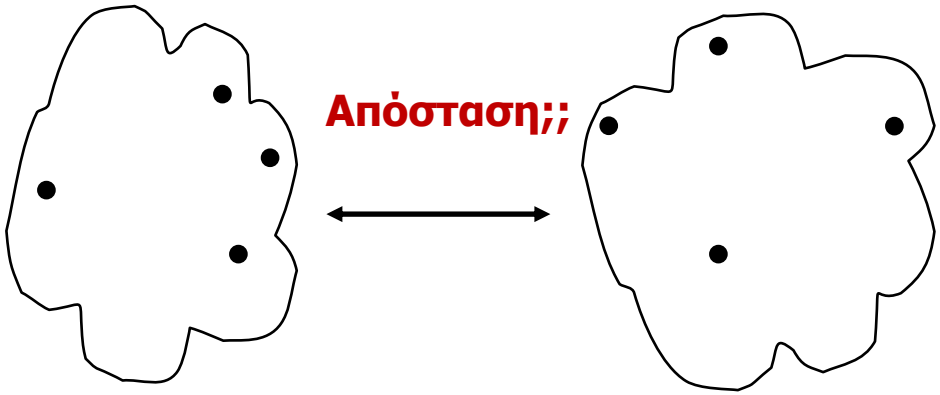
Μετά τη συνένωση



	C1	C2 U C5	C3	C4
C1		?		
C2 U C5	?	?	?	?
C3		?		
C4		?		



Απόσταση ομάδων (1/5)

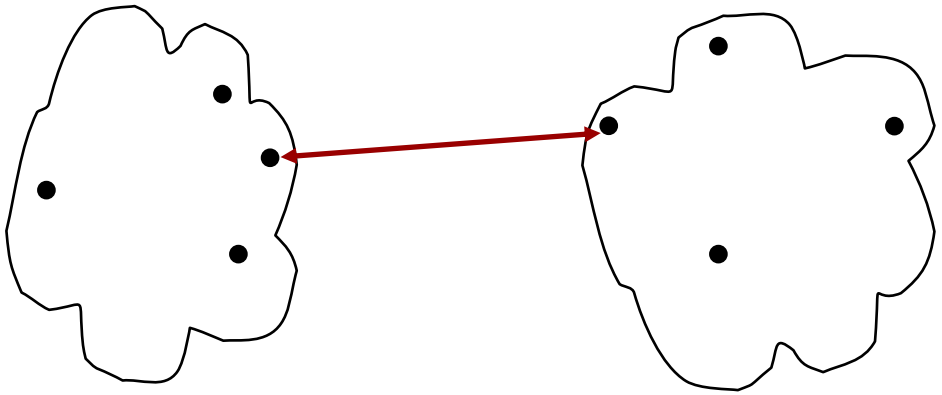


- MIN
- MAX
- Μέσος όρος
- Απόσταση μεταξύ κέντρων
- Μέθοδος Ward (τετραγωνικό σφάλμα)

	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Πίνακας Αποστάσεων

Απόσταση ομάδων (2/5)



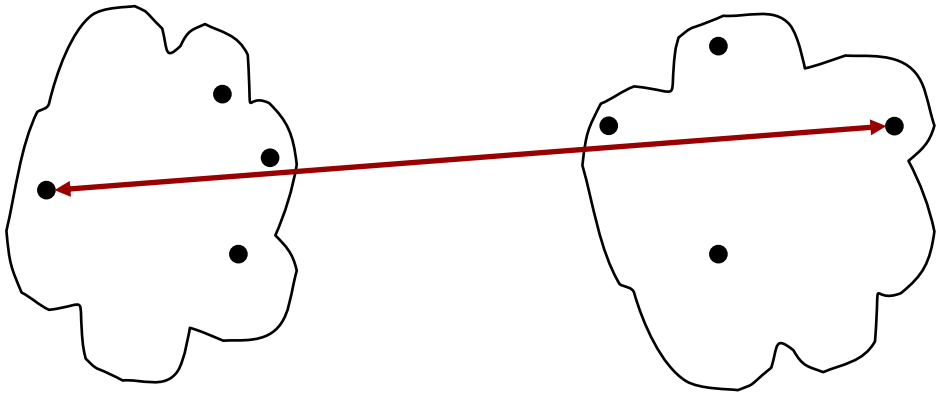
- MIN
- MAX
- Μέσος όρος
- Απόσταση μεταξύ κέντρων
- Μέθοδος Ward (τετραγωνικό σφάλμα)

	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Πίνακας Αποστάσεων

- **Ανάλογα , χειριζόμαστε και ομοιότητες αντί για αποστάσεις.**

Απόσταση ομάδων (3/5)

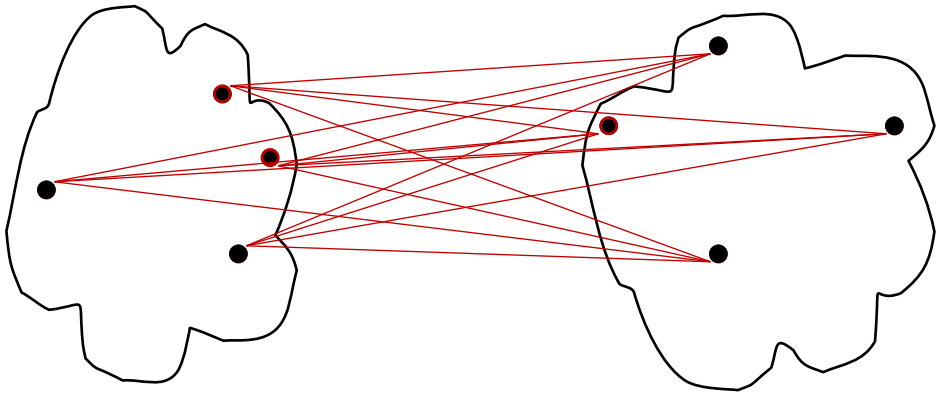


- MIN
- MAX
- Μέσος όρος
- Απόσταση μεταξύ κέντρων
- Μέθοδος Ward (τετραγωνικό σφάλμα)

	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Πίνακας Αποστάσεων

Απόσταση ομάδων (4/5)

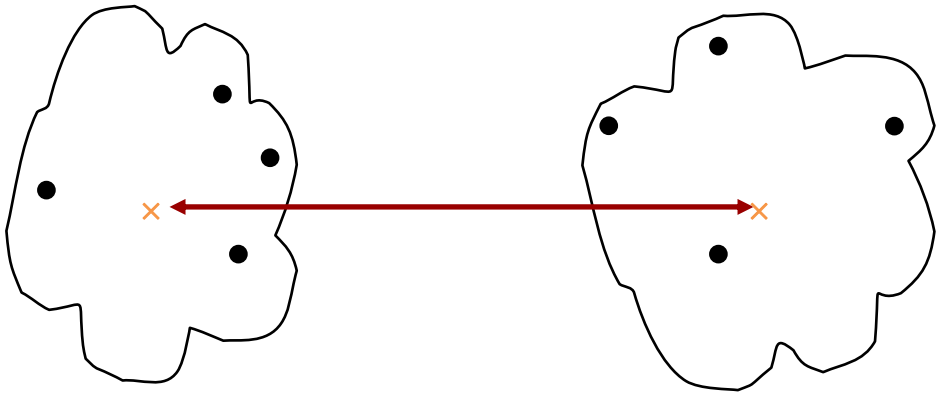


- MIN
- MAX
- Μέσος όρος
- Απόσταση μεταξύ κέντρων
- Μέθοδος Ward (τετραγωνικό σφάλμα)

	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Πίνακας Αποστάσεων

Απόσταση ομάδων (5/5)

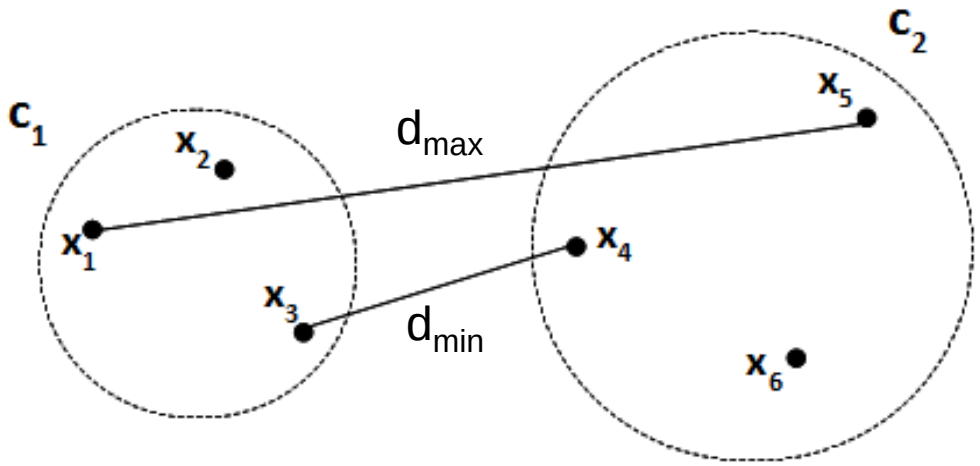


- MIN
- MAX
- Μέσος όρος
- Απόσταση μεταξύ κέντρων
- Μέθοδος Ward (τετραγωνικό σφάλμα)

	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

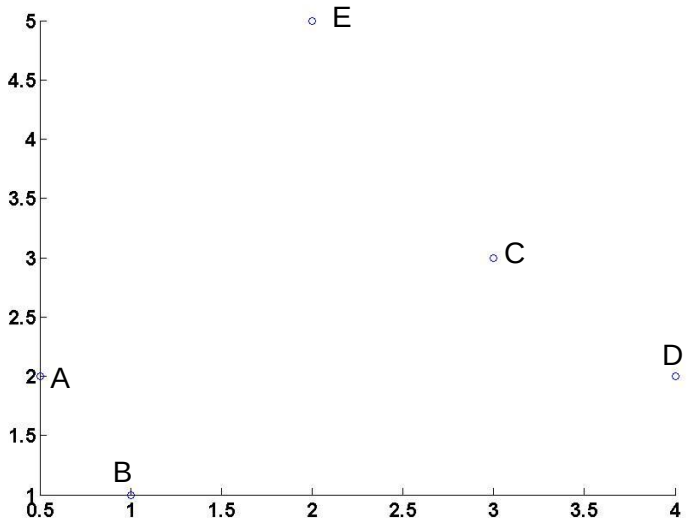
Πίνακας Αποστάσεων

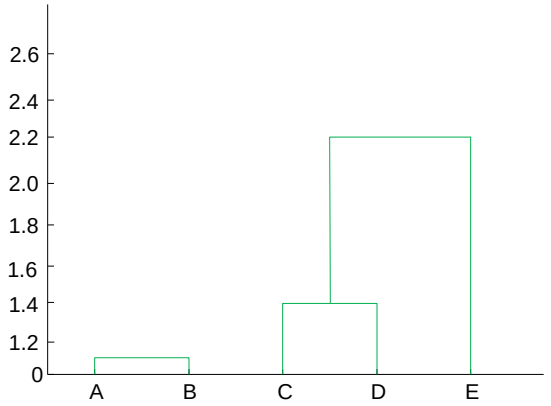
Σύγκριση MIN vs. MAX (1/2)



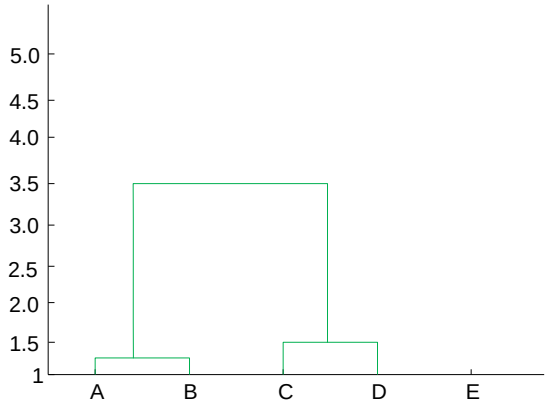
- MIN: αλγόριθμος μονής σύνδεσης (single-link).
- MAX: αλγόριθμος πλήρους σύνδεσης (complete-link).

Σύγκριση MIN vs. MAX (2/2)



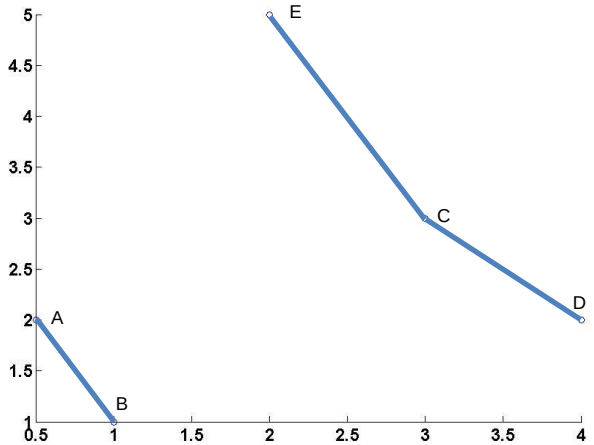


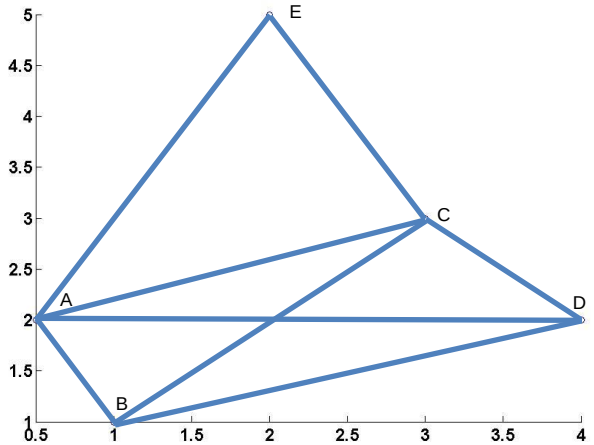
0	1.1180	2.6926	3.5	3.3541
1.1180	0	2.8282	3.1623	4.1231
2.6926	2.8284	0	1.4142	2.2361
3.5	3.1623	1.4142	0	3.6056
3.3541	4.1231	2.2361	3.6056	0



Ερμηνεία με θεωρία γράφων

- Ζυγισμένος γράφος G:
  - κορυφές = αντικείμενα
  - βάρη = αποστάσεις.
- d(i): η απόσταση των ομάδων που συνενώνονται στο i βήμα.
- H(d(i)): ο υπογράφος του G με ακμές με βάρη < d(i).
- Στο τελικό αποτέλεσμα:
  - Μονής σύνδεσης: κάθε ομάδα αντιστοιχεί σε μία συνιστώσα που είναι spanning tree (ζευγνύον δένδρο).
  - Πλήρους σύνδεσης: κάθε ομάδα, μία κλίκα (πλήρης υπογράφος).





Γενική εξίσωση απόστασης

- Μετά τη συνένωση r και s, βρίσκουμε την απόστασή τους από κάθε άλλη ομάδα k:

( ) ( ) ( ) ( ) ( ) ( )

Αλγόριθμος	α_r	α_s	**β**	**γ**
Μονής σύνδεσης	1/2	1/2	0	-1/2
Πλήρους σύνδεσης	1/2	1/2	0	1/2
Μέσου όρου	———	———	0	0
Κέντρων	———	———	———	0
Ward's	———	———	———	0

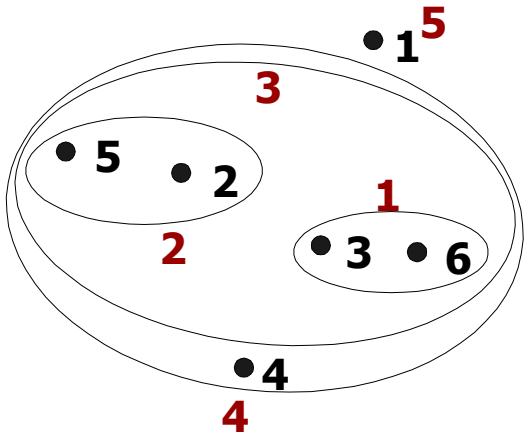
Αλγόριθμος Ward

- Το κέντρο της k ομάδας:
- Τετραγωνικό λάθος:
- Συνολικό λάθος (K ομάδες):
- Κατά τη συνένωση r, s → t:

$$\_\_\_\_\_$$

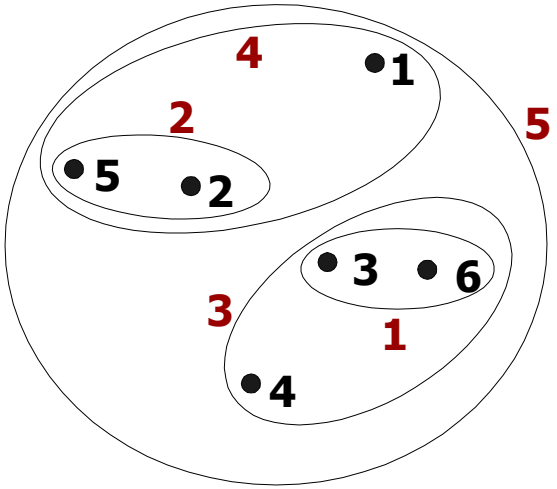
- Σε κάθε βήμα: συνένωση που προκαλεί μικρότερο ΔE^2

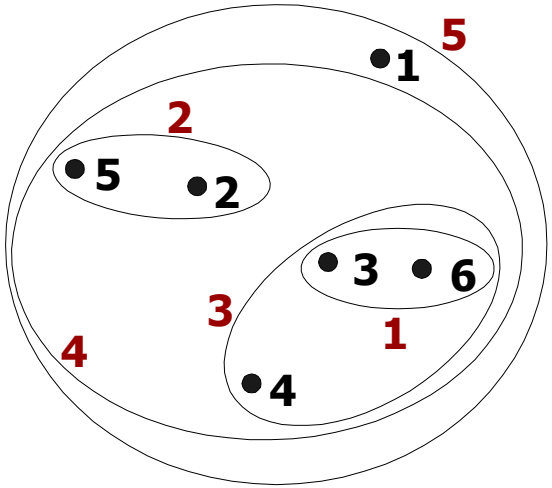
Σύγκριση



MIN

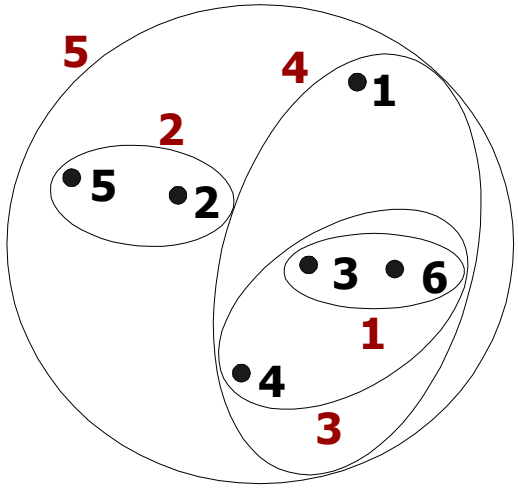
MAX





Group Average

Ward's Method



Ποιος είναι καταλληλότερος; (1/2)

- Εξαρτάται από τα δεδομένα.
- **MAX**:

  +λιγότερη εξάρτηση σε θόρυβο και outliers.

  - Τείνει να διασπά μεγάλες συστάδες. Οδηγεί συνήθως σε κυκλικά σχήματα.

- **MIN**:

  + Contiguity-based (συνεχόμενες συστάδες).

  + Μπορεί να χειριστεί μη ελλειπτικά (non-elliptical) σχήματα.

  - Ευαίσθητο σε θόρυβο και outliers.

Ποιος είναι καταλληλότερος; (2/2)

- **Μέσος όρος**:
  - Ανάμεσα σε MIN-MAX

    \+ Πλεονεκτήματα: μικρότερη ευαισθησία σε θόρυβο και outliers.

    \- Μειονεκτήματα: Ευνοεί κυκλικές συστάδες.

Πολυπλοκότητα Ιεραρχικών αλγορίθμων

- $O(N^2)$ χώρο, για τον πίνακα αποστάσεων.
- $O(N^3)$ χρόνο.
  - N βήματα, σε κάθε βήμα N^2, για τη συνένωση.
  - $O(N^2 \log(N))$ για μονής σύνδεσης σε χαμηλές διαστάσεις.

Σύγκριση με k-means

- Δεν απαιτεί τον ορισμό του αριθμού των clusters.
- Οι αποφάσεις σε κάθε βήμα δεν αναιρούνται.
- Όχι τοπικά βέλτιστα. Όμως…
- Υπάρχουν μελέτες που υποστηρίζουν ότι παράγουν καλύτερες συστάδες.
- Συνδυάζεται με k-means.

Σημείωμα Αναφοράς

Διαθέσιμο από τη δικτυακή διεύθυνση:http://eclass.auth.gr/courses/OCRS182/

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.