Αναγνώριση Προτύπων

Ομαδοποίηση – Επιπλέον θέματα

Ανδρέας Λ. Συμεωνίδης

Αν. Καθηγητής

Τμήμα Ηλεκτρολόγων Μηχ/κών

&

Μηχ/κών Υπολογιστών, Α.Π.Θ.

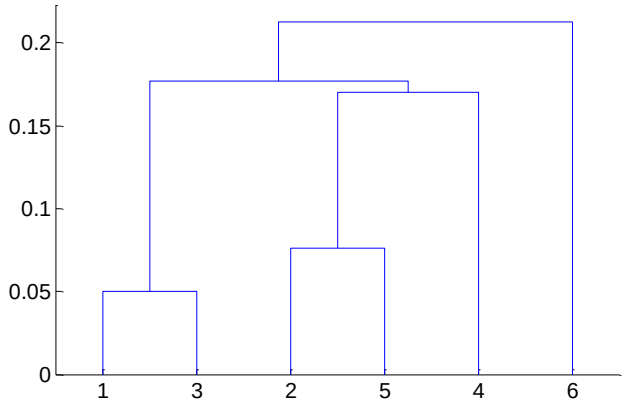
Email: asymeon@eng.auth.gr

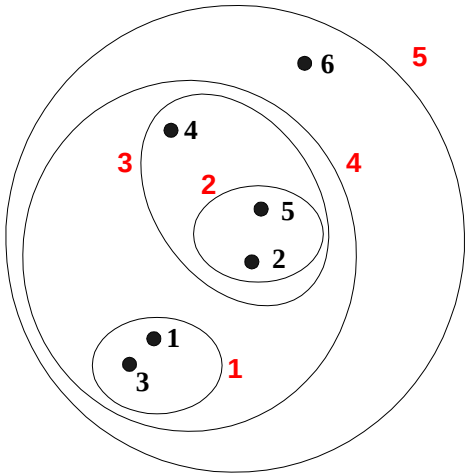
Διάρθρωση διάλεξης

- Ιεραρχική ομαδοποίηση
- Πυκνωτική ομαδοποίηση
- Self-organizing maps
- Σύνοψη ομαδοποίησης

Ιεραρχική ομαδοποίηση

- Παράγει ένα σετ από εμφωλευμένες ομάδες, οργανωμένες ως ιεραρχικό δένδρο
- Μπορούν να οπτικοποιηθούν σα δενδρόγραμμα
  - Μια δενδρική δομή που απεικονίζει τις αλληλουχίες των διαχωρισμών ή ενώσεων





Ιεραρχική ομαδοποίηση - πλεονεκτήματα

- Δε χρειάζεται να γίνει υπόθεση για τον αριθμό των ομάδων
  - Ο αριθμός των ομάδων καθορίζεται από το επίπεδο στο οποίο «κόβουμε» το δενδρόγραμμα
- Αυτό θα μπορούσε να συνεπάγεται χρήσιμες ταξονομίες (π.χ. ζωικό βασίλειο)

Τύποι ιεραρχικής ομαδοποίησης

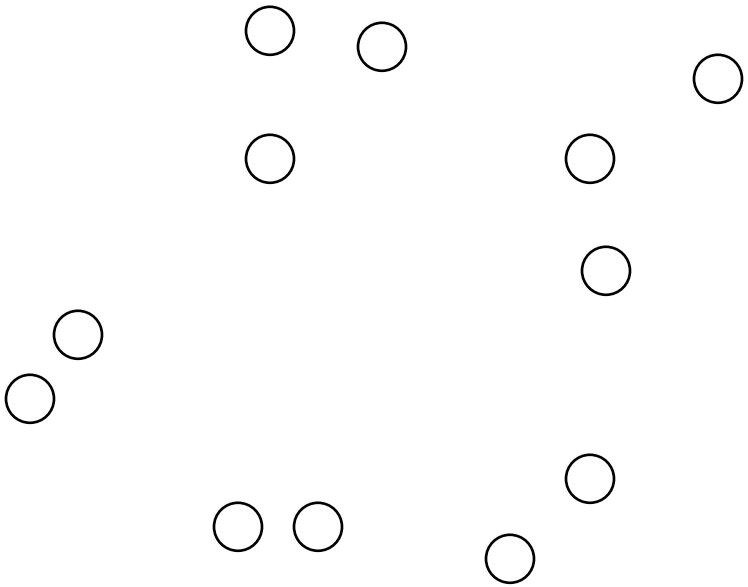
- Δυο βασικοί τύποι ιεραρχικής ομαδοποίησης
  - Συνάθροισης:
    - Στην αρχή κάθε σημείο είναι μια ομάδα
    - Σε κάθε βήμα, συνενώνεται το πλησιέστερο ζευγάρι ομάδων μέχρις ότου να μείνει μόνο μια (ή K) ομάδα(ες)
  - Διαχωρισμού:
    - Στην αρχή όλα τα σημεία ανήκουν σε μια ομάδα
    - Σε κάθε βήμα, αποσχίζεται ένα σημείο, μέχρις ότου να γίνουν όλα τα σημεία από μια (ή K) ομάδα(ες)
- Οι κλασικοί ιεραρχικοί αλγόριθμοι χρησιμοποιούν ως μετρική τον πίνακα ομοιότητας ή απόστασης

Ιεραρχικοί αλγόριθμοι συνάθροισης

- Οι πλέον διαδεδομένες τεχνικές
- Ψευδοκώδικας
  1. Υπολόγισε τον πίνακα απόστασης
  2. Όρισε σημείο να είναι μια ομάδα
  3. **Επανέλαβε**
  4. Συνένωσε τις δυο κοντινότερες ομάδες
  5. Ανανέωσε τον πίνακα απόστασης
  6. **Μέχρι** να μείνει μόνο μια ομάδα
- Βασικό σημείο στον αλγόριθμο είναι ο υπολογισμός του πίνακα απόστασης ανάμεσα σε δυο ομάδες
  - Διαφορετικές προσεγγίσεις του ορισμού της απόστασης ορίζουν διαφορετικούς αλγορίθμους.

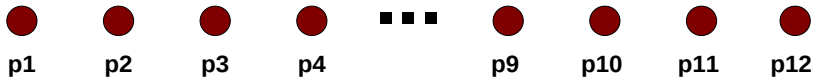
Αρχική κατάσταση

- Κάθε σημείο είναι μια ομάδα και κάθε ομάδα απέχει από την άλλη απόσταση όσ δηλώνεται στον πίνακα απόστασης



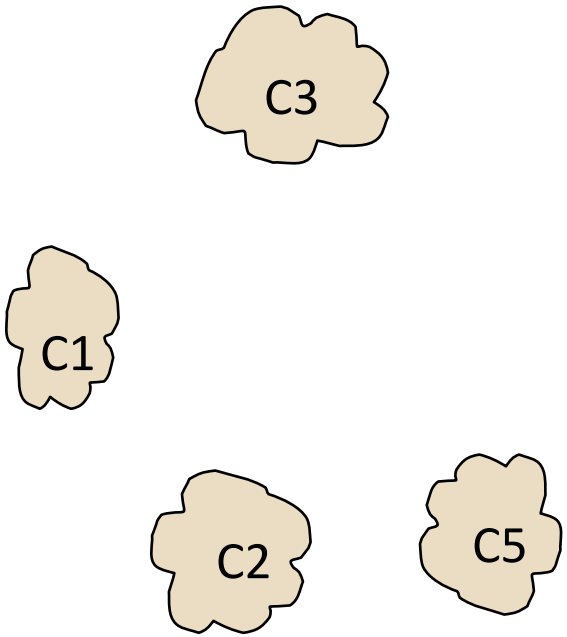
	p1	p2	p3	p4	p5	…
p1						
p2						
p3						
p4						
p5						
.						
.						

Proximity Matrix



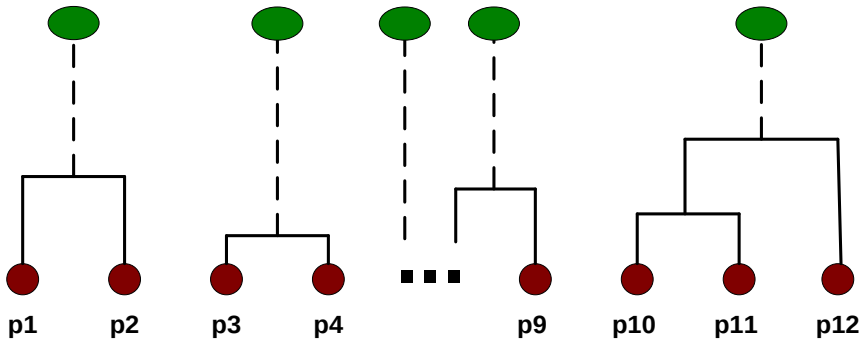
Ενδιάμεση κατάσταση

- Μετά από κάποια βήματα συνένωσης, υπάρχουν κάποιες συναθροισμένες ομάδες.



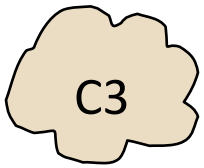
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

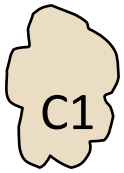
Proximity Matrix

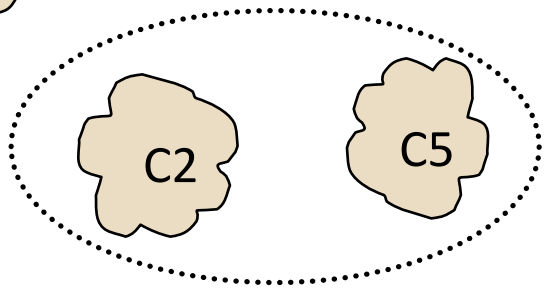


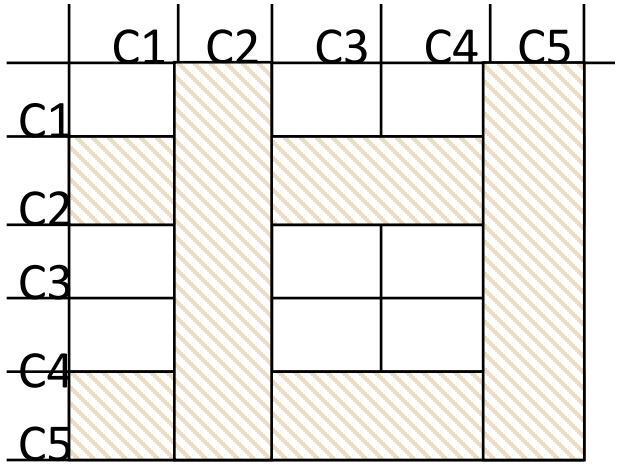
Ενδιάμεση κατάσταση

- Συνενώνουμε τις δυο πλησιέστερες ομάδες (C2 και C5) και ανανεώνουμε τον πίνακα απόστασης.

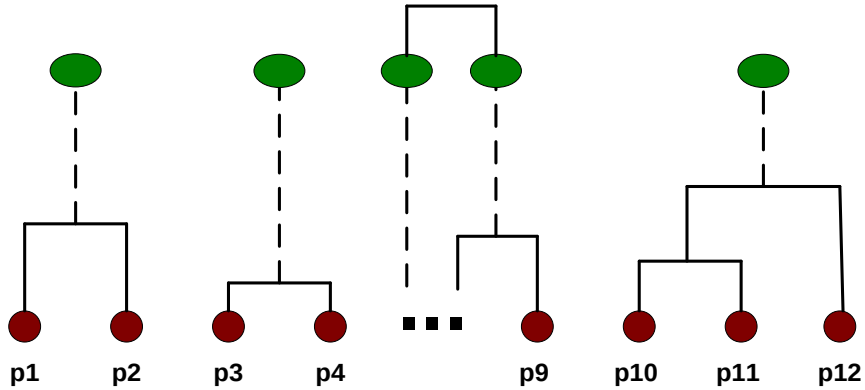






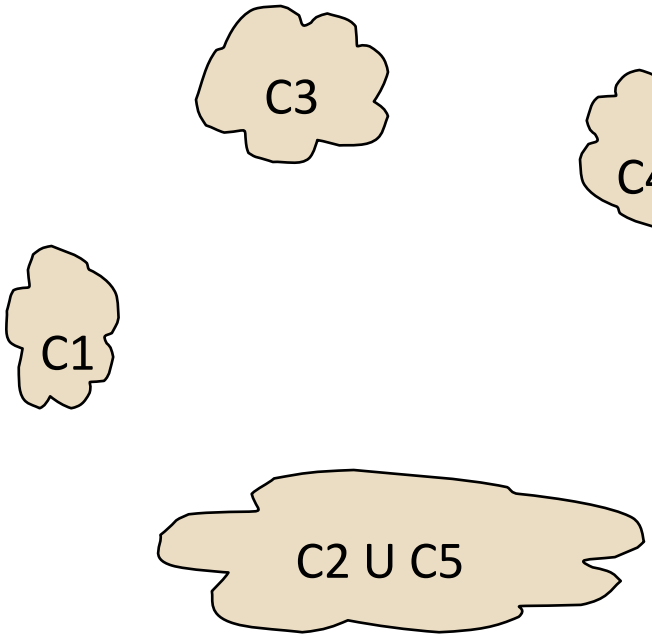


Proximity Matrix



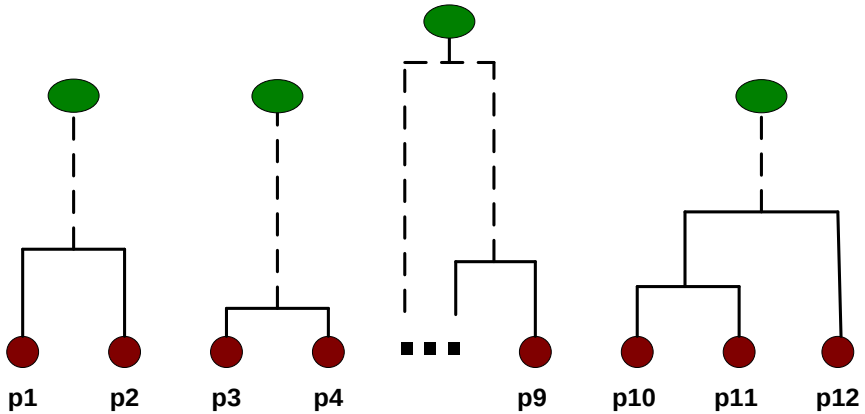
Μετά τη συνένωση

- Η ερώτηση είναι: “Με ποιον τρόπο ανανεώνουμε τον πίνακα;”

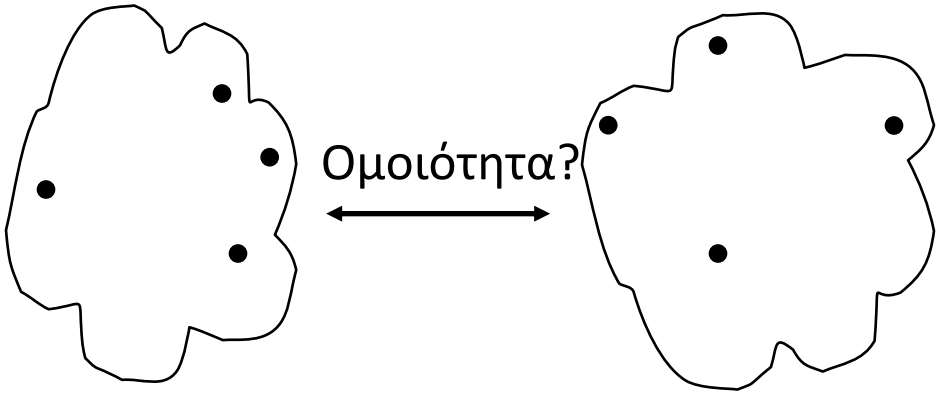


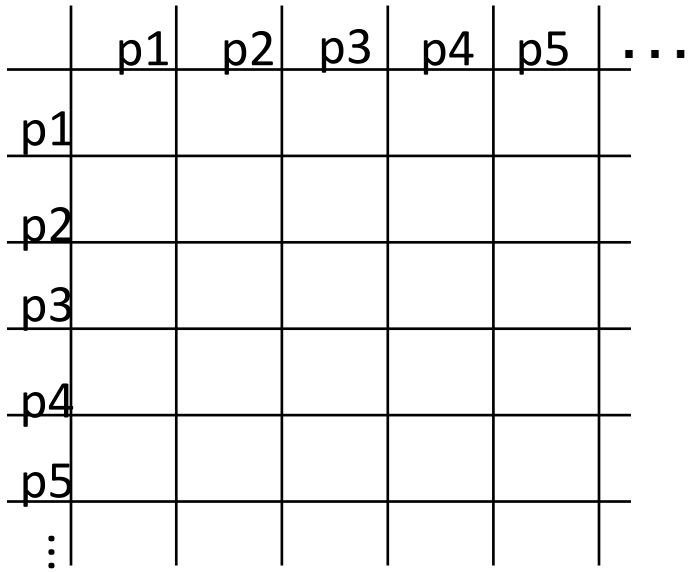
	C1	C2 U C5	C3	C4
C1		?		
C2 U C5	?	?	?	?
C3		?		
C4		?		

Proximity Matrix



Ορισμός της ενδο-ομαδικής ομοιότητας





Proximity Matrix

MIN

MAX

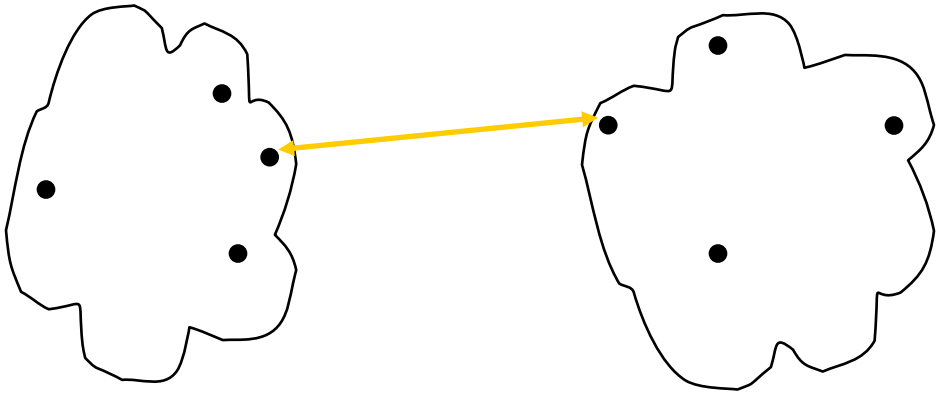
Group Average

Απόσταση ανάμεσα στα κέντρα

Ορισμός συνάρτησης στόχου

Μέθοδος του Ward (τετραγωνικό σφάλμα)

Ορισμός της ενδο-ομαδικής ομοιότητας



MIN

MAX

Group Average

Απόσταση ανάμεσα στα κέντρα

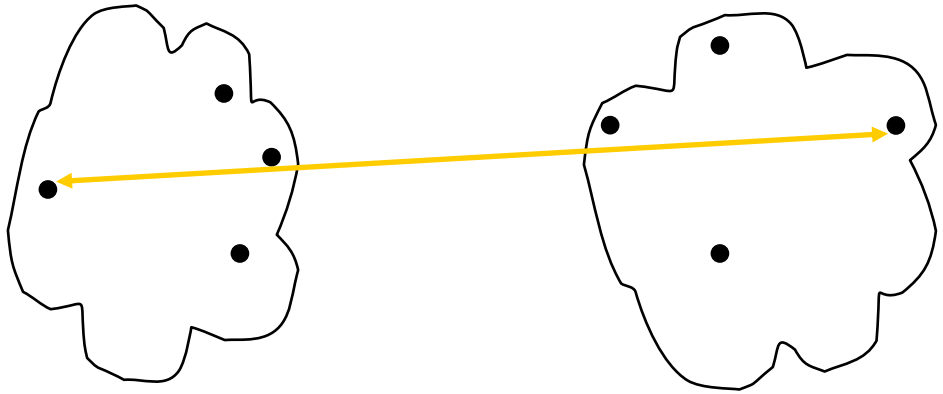
Ορισμός συνάρτησης στόχου

Μέθοδος του Ward (τετραγωνικό σφάλμα)

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

Proximity Matrix

Ορισμός της ενδο-ομαδικής ομοιότητας



MIN

MAX

Group Average

Απόσταση ανάμεσα στα κέντρα

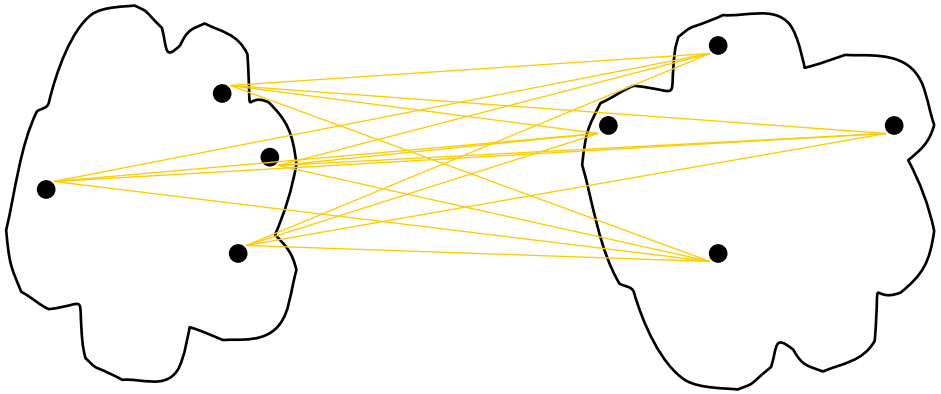
Ορισμός συνάρτησης στόχου

Μέθοδος του Ward (τετραγωνικό σφάλμα)

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

Proximity Matrix

Ορισμός της ενδο-ομαδικής ομοιότητας



MIN

MAX

Group Average

Απόσταση ανάμεσα στα κέντρα

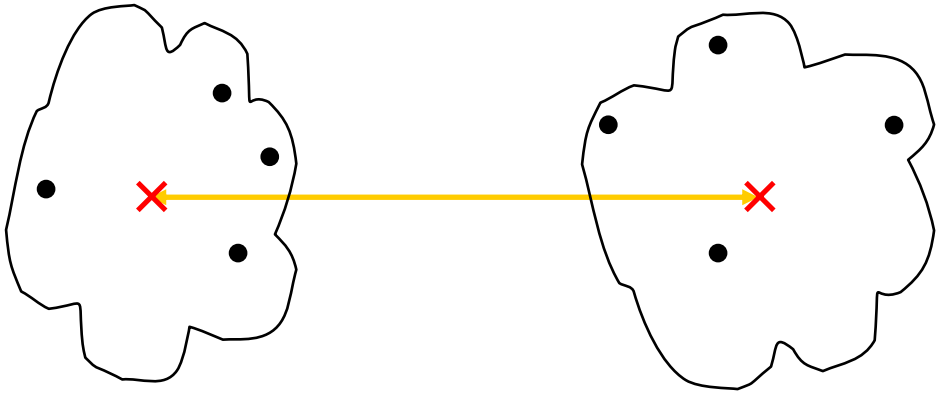
Ορισμός συνάρτησης στόχου

Μέθοδος του Ward (τετραγωνικό σφάλμα)

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

Proximity Matrix

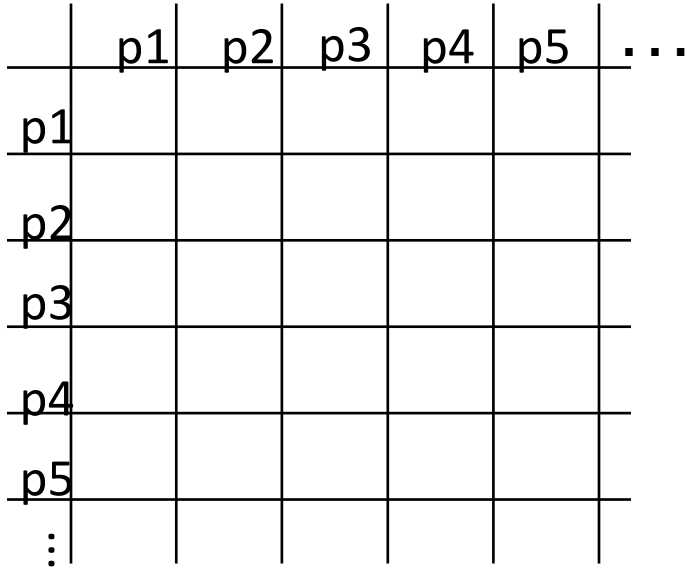
Ορισμός της ενδο-ομαδικής ομοιότητας



MIN

MAX

Group Average



Proximity Matrix

Απόσταση ανάμεσα στα κέντρα

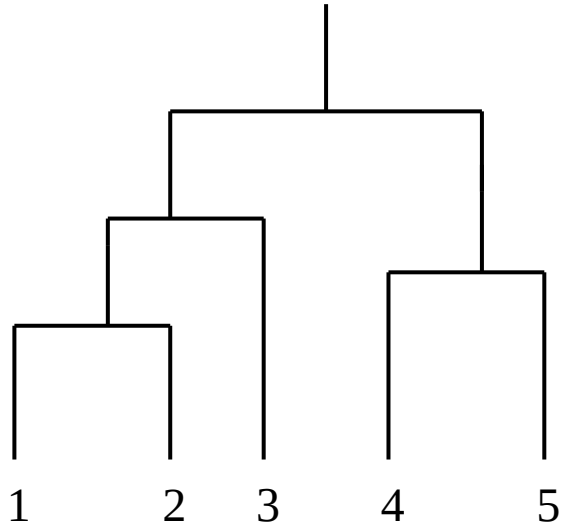
Ορισμός συνάρτησης στόχου

Μέθοδος του Ward (τετραγωνικό σφάλμα)

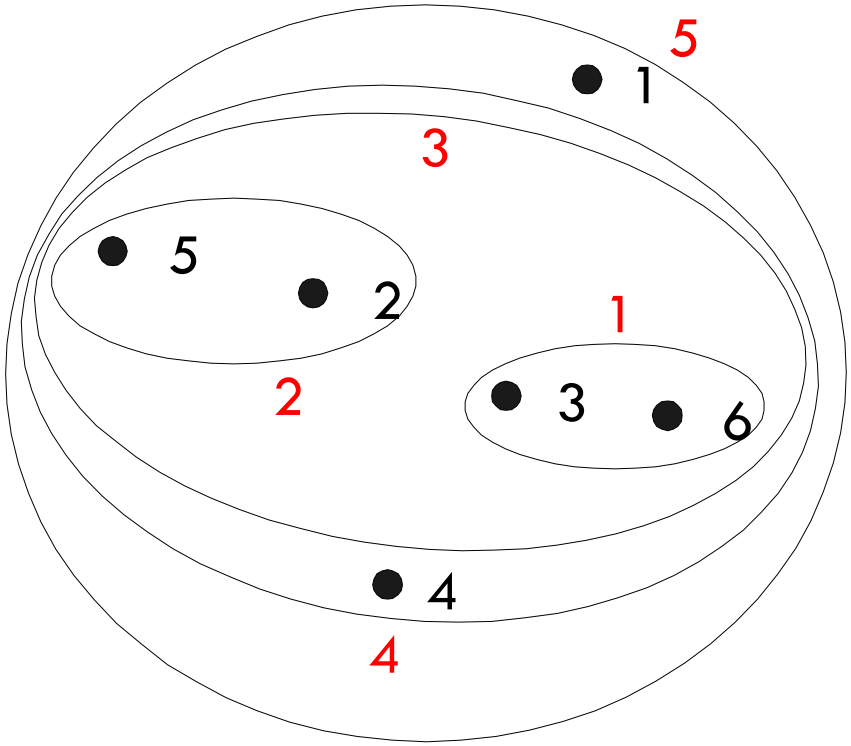
Ομοιότητα ομάδων: MIN ή Single Link

- Η ομοιότητα δυο ομάδων καθορίζεται από τα πιο όμοια (π.χ. κοντινά) σημεία των ομάδων
  - Ορίζεται από ένα ζεύγος σημείων, δηλ. μια σύνδεση στον γράφο γειτνίασης (proximity graph).

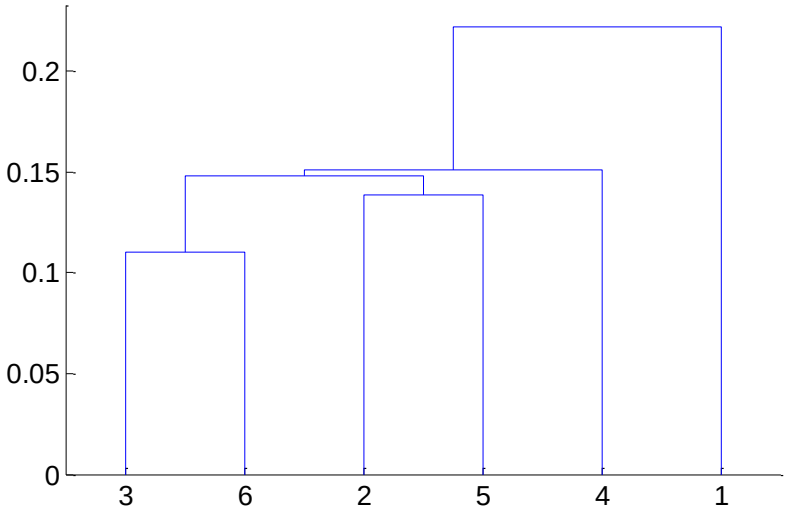
	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00



Ιεραρχική ομαδοποίηση: MIN

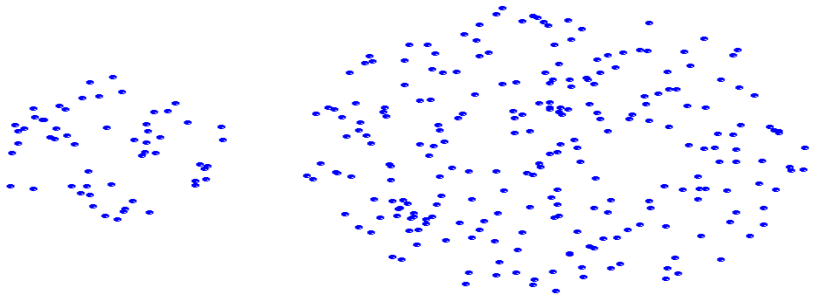


Εμφωλευμένες ομάδες

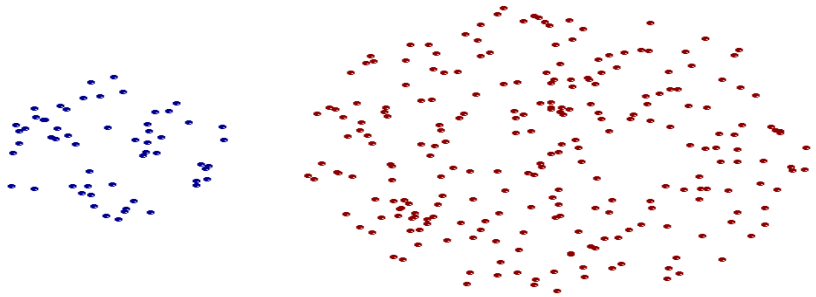


Δενδρόγραμμα

Πλεονεκτήματα του MIN



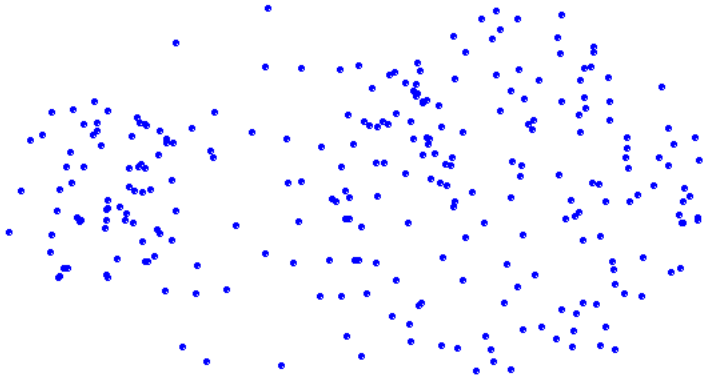
Αρχικά σημεία



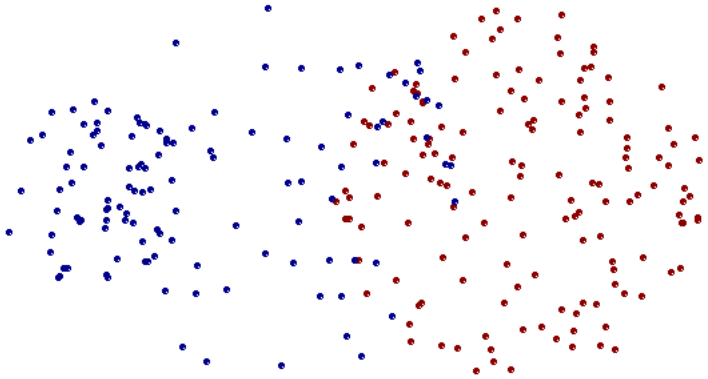
Δυο ομάδες

Μπορεί να χειριστεί μη κανονικά σχήματα

Μειονεκτήματα του MIN



Αρχικά σημεία



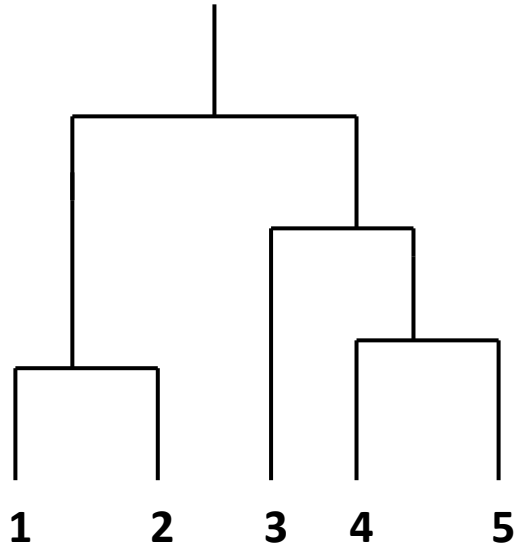
Δυο ομάδες

Ευαίσθητος σε θόρυβο και εξωκείμενες τιμές

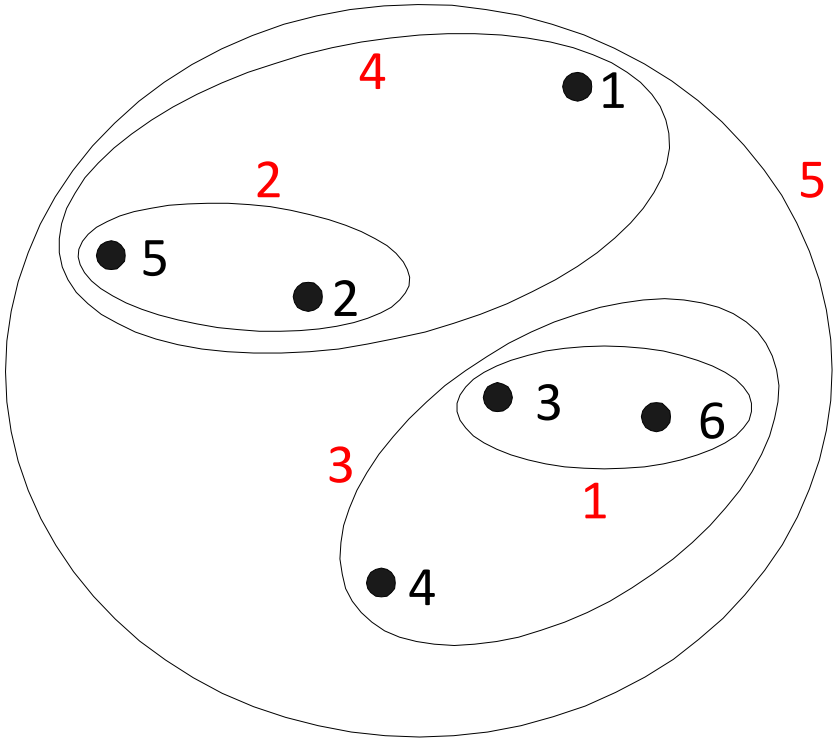
Ομοιότητα ομάδων: MAX ή Complete Linkage

- Η ομοιότητα δυο ομάδων καθορίζεται από τα πιο ανόμοια (π.χ. μακρινά) σημεία των ομάδων
  - Ορίζεται από όλα τα ζεύγη σημείων στις ομάδες

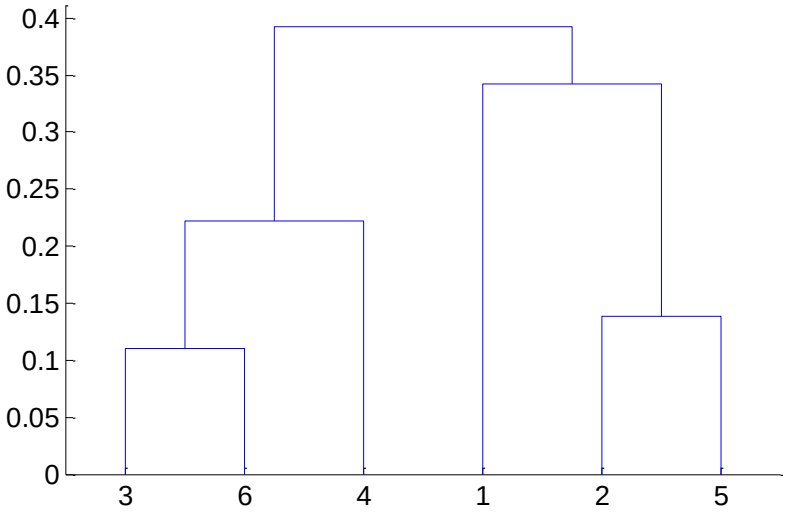
	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00



Ιεραρχική ομαδοποίηση: MAX

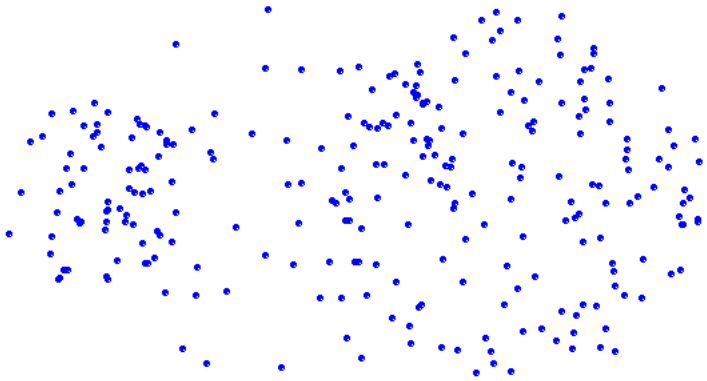


Εμφωλευμένες ομάδες

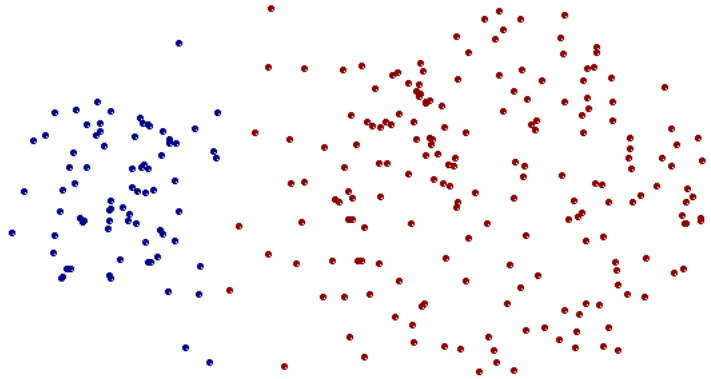


Δενδρόγραμμα

Πλεονεκτήματα του MAX



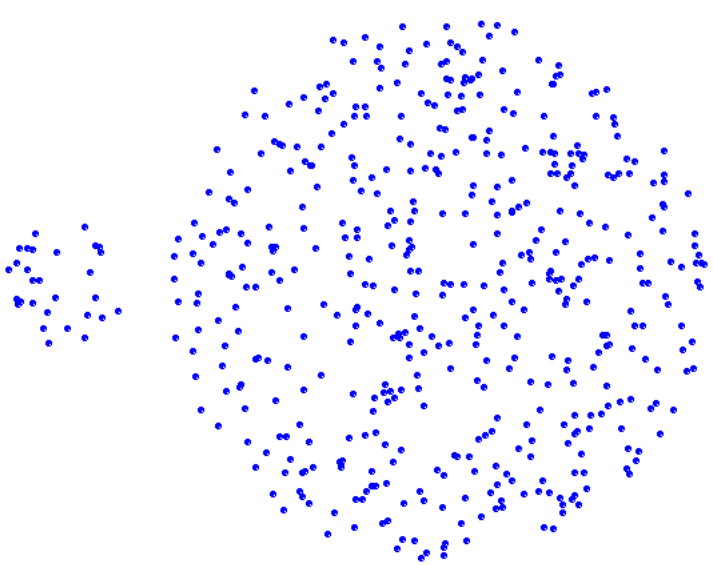
Αρχικά Σημεία



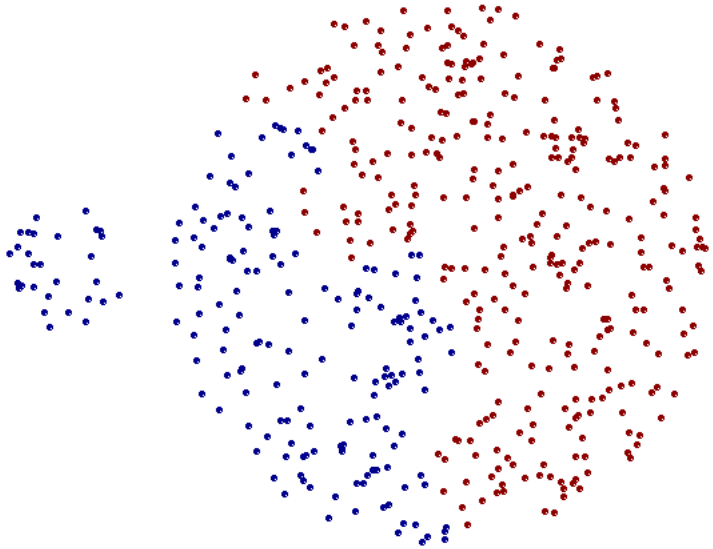
Δυο ομάδες

Λιγότερο ευαίσθητος σε θόρυβο και εξωκείμενες τιμές

Μειονεκτήματα του MAX



Αρχικά σημεία



Δυο ομάδες

Συνήθως «σπάει» τις μεγάλες ομάδες

Πολωμένος προς τη δημιουργία σφαιρικών ομάδων

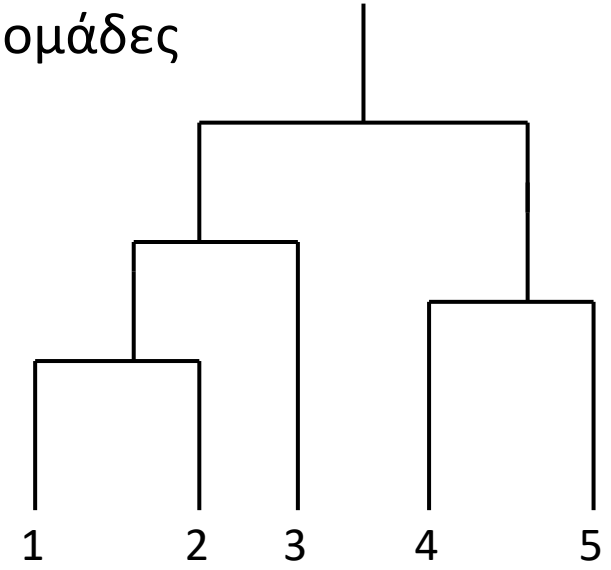
Ομοιότητα ομάδων: Group Average

- Η ομοιότητα δυο ομάδων είναι ο μέσος όρος των ομοιοτήτων (αποστάσεων) ανάμεσα σε όλα τα ζεύγη των ομάδων

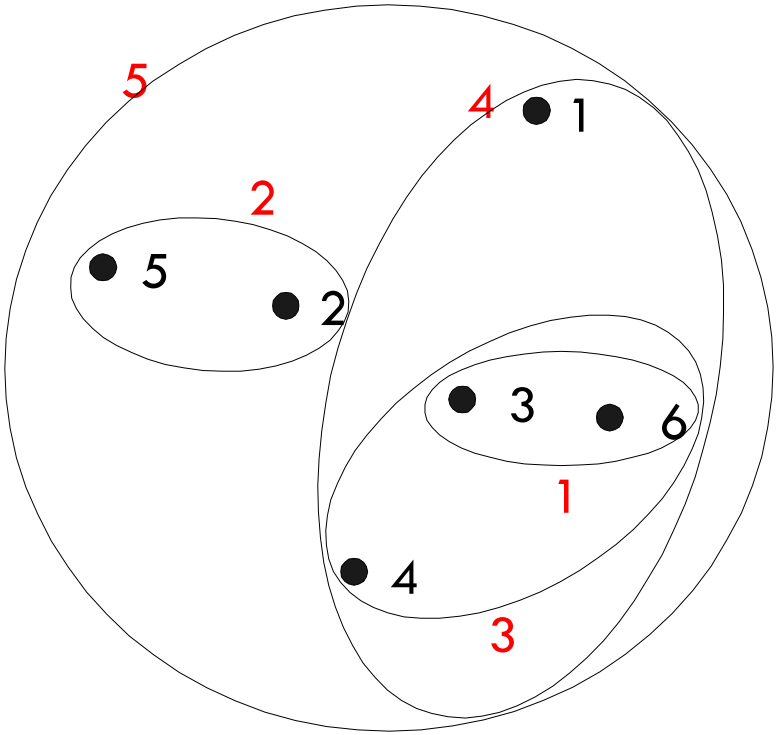
$$\text{proximity}(\text{Cluster}_i, \text{Cluster}_j) = \frac{\sum_{\substack{p_i \in \text{Cluster}_i \\ p_j \in \text{Cluster}_j}} \text{proximity}(p_i, p_j)}{|\text{Cluster}_i| * |\text{Cluster}_j|}$$

- Ανάγκη για χρήση της μέσης συνδεσιμότητας (κλιμάκωση), καθώς η συνολική ομοιότητα ευνοεί τις μεγάλες ομάδες

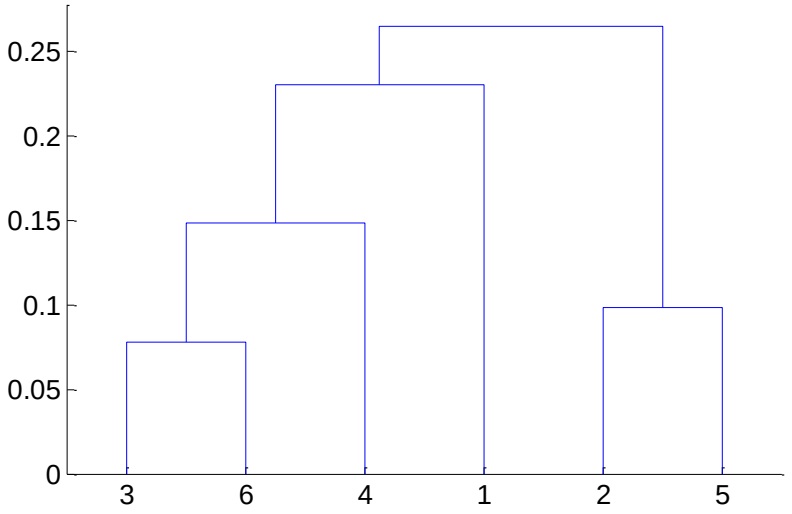
	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00



Ιεραρχική ομαδοποίηση: Group Average



Εμφωλευμένες ομάδες



Δενδρόγραμμα

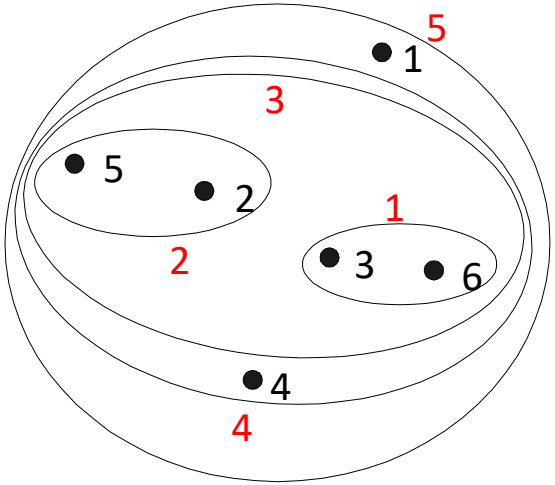
Ιεραρχική ομαδοποίηση: Group Average

- Συμβιβασμός ανάμεσα στον Single και τον Complete Link αλγόριθμο
  - Λιγότερο ευαίσθητος σε θόρυβο και εξωκείμενες τιμές
  - Πολωμένος προς τη δημιουργία σφαιρικών ομάδων

Ιεραρχική ομαδοποίηση: μέθοδος Ward

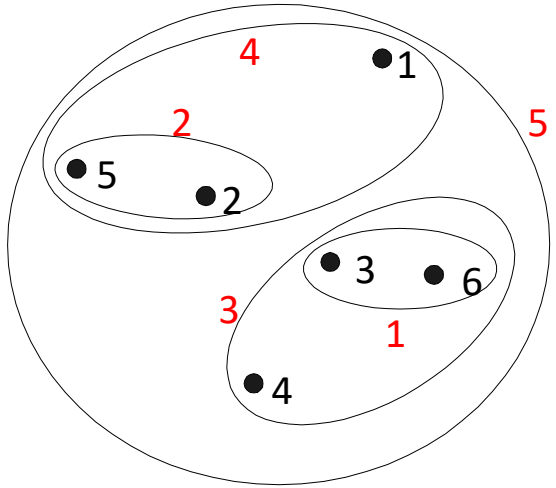
- Η ομοιότητα των ομάδων βασίζεται στην αύξηση του τετραγωνικού σφάλματος όταν ενώνονται δυο ομάδες
  - Λιγότερο ευαίσθητος σε θόρυβο και εξωκείμενες τιμές
  - Πολωμένος προς τη δημιουργία σφαιρικών ομάδων
- Ιεραρχικό ανάλογο του K-means
  - Μπορεί να χρησιμοποιηθεί για την αρχικοποίηση του K-means

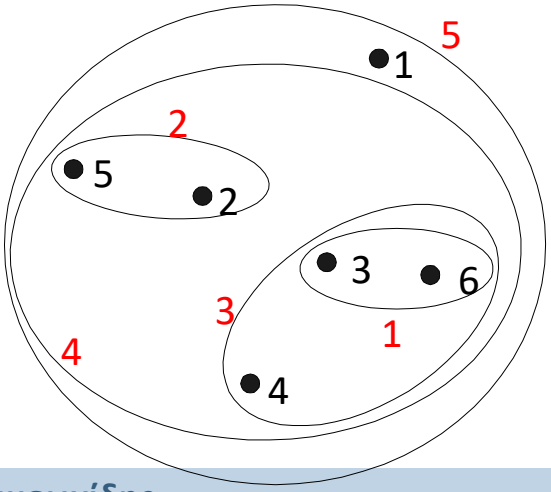
Σύγκριση ιεραρχικών αλγορίθμων



MIN

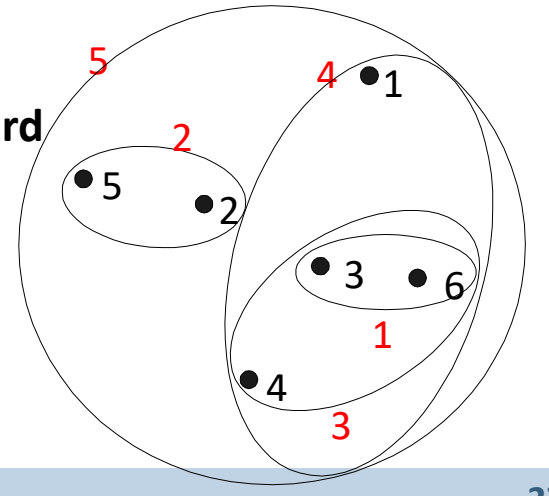
MAX





Group Average

Μέθοδος Ward



Ιεραρχική ομαδοποίηση: Απαιτήσεις σε χρόνο και χώρο

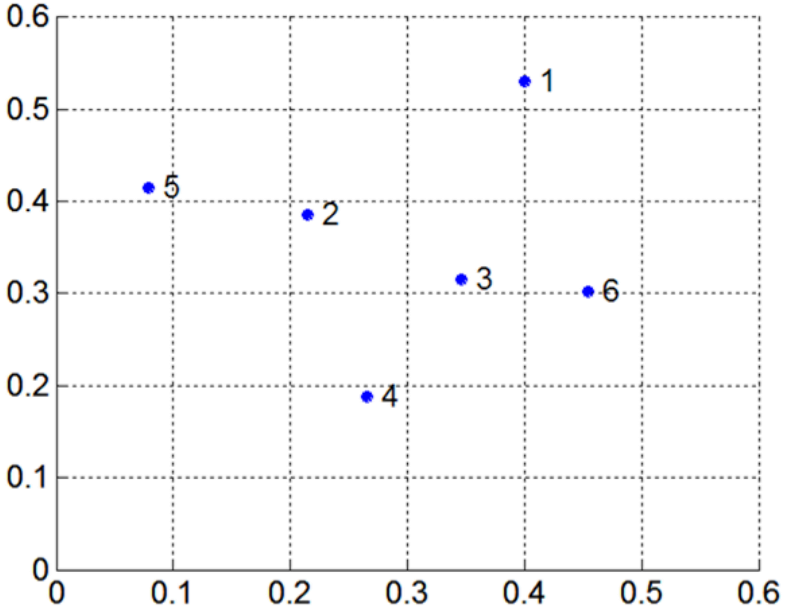
- $O(N^2)$ χώρος, λόγω του πίνακα απόστασης (N ο αριθμός των σημείων)
- $O(N^3)$ χρόνος σε αρκετές περιπτώσεις
  - N βήματα και σε κάθε βήμα ο χώρος, N^2, του πίνακα απόστασης πρέπει να διασχιστεί και να ανανεωθεί
  - Μπορεί να μειωθεί σε κάποιες περιπτώσεις σε $O(N^2 \log(N))$

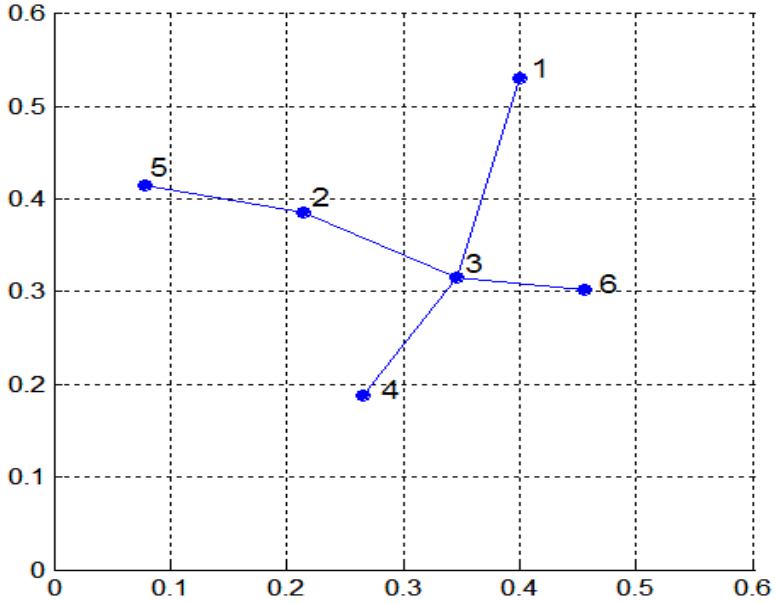
Προβλήματα και περιορισμοί

- Μετά την απόφαση συνένωσης δυο ομάδων, αυτή θεωρείται δεδομένη (άπληστη επιλογή)
- Δεν ελαχιστοποιείται καμία συνάρτηση στόχου

MST: Ιεραρχική ομαδοποίηση διαχωρισμού

- Κατασκευή ενός MST (Minimum Spanning Tree, ΕΣΔ)
  - Ξεκινήστε το δένδρο από οποιοδήποτε σημείο
  - Στα ακόλουθα βήματα, βρείτε το πιο όμοιο ζεύγος σημείων (p, q), όπου το (p) ανήκει στο δένδρο αλλά το (q) όχι
  - Προσθέστε το q στο δένδρο και δημιουργήστε μια ακμή ανάμεσα στα p και q





MST: Ιεραρχική ομαδοποίηση διαχωρισμού

- Χρησιμοποιήστε το MST για την κατασκευή μιας ιεραρχίας ομάδων

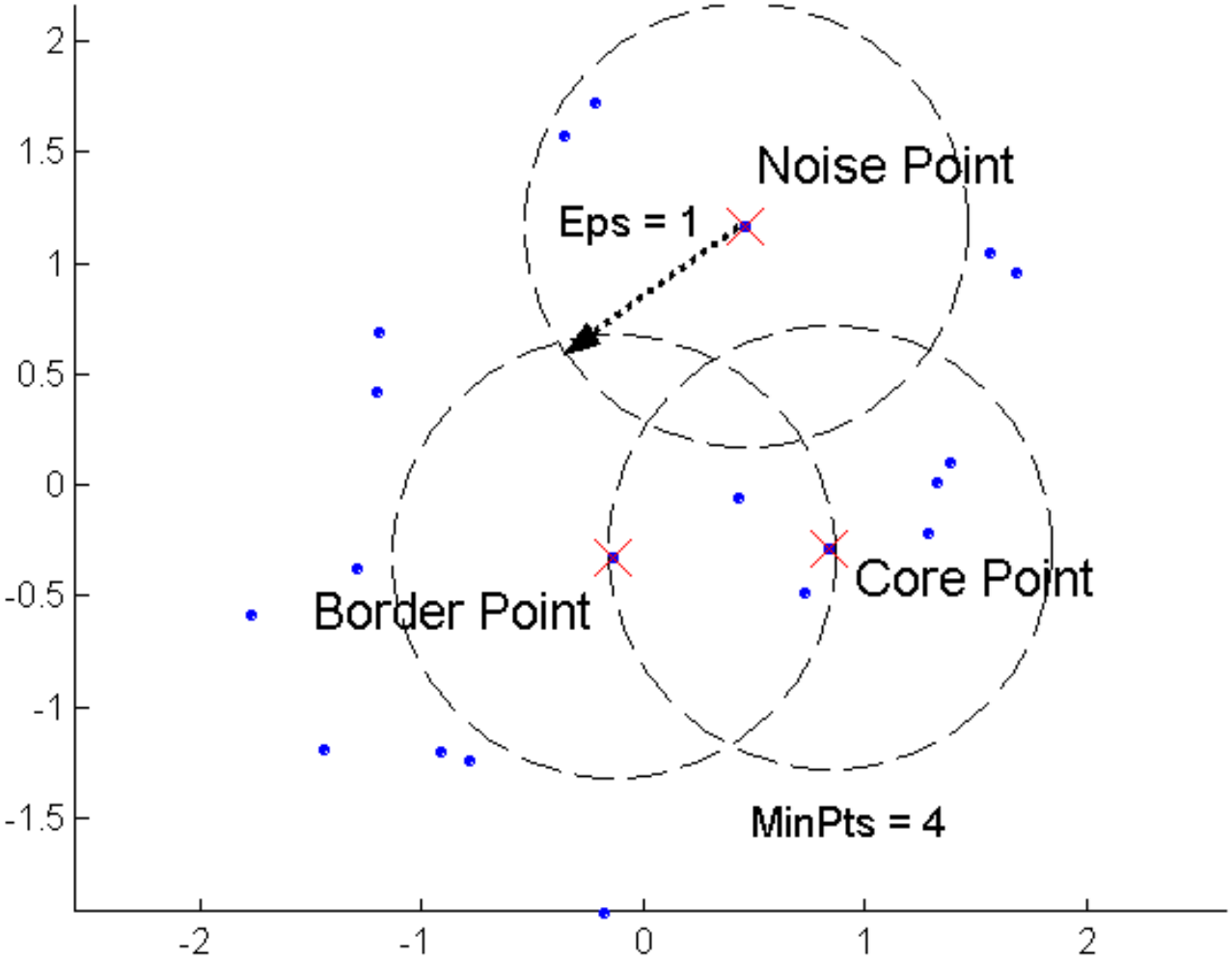
Algorithm 7.5 MST Divisive Hierarchical Clustering Algorithm

1: Compute a minimum spanning tree for the proximity graph.
2: **repeat**
3: Create a new cluster by breaking the link corresponding to the largest distance (smallest similarity).
4: **until** Only singleton clusters remain

Πυκνωτικοί αλγόριθμοι: DBSCAN

- Density = ο αριθμός των σημείων μέσα σε μια περιοχή ακτίνας Eps
- Ένα σημείο είναι κύριο σημείο αν έχει πυκνότητα ≥ από MinPts μέσα στην Eps του
  - Αυτά είναι εσωτερικά σημεία στις ομάδες
- Ένα συνοριακό σημείο έχει πυκνότητα ≤ από MinPts μέσα στην Eps του, βρίσκεται στη γειτονιά ενός κύριου σημείου
- Ένα σημείο θορύβου είναι ένα σημείο που δεν είναι ούτε κύριο, ούτε συνοριακό

DBSCAN: Σημεία κύρια, συνοριακά και θορύβου



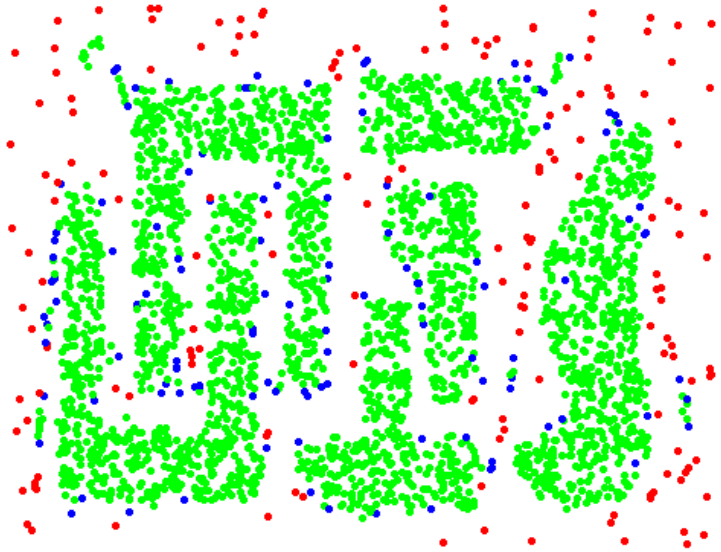
DBSCAN: ψευδοκώδικας

- Απαλοιφή των σημείων θορύβου
- Εφαρμογή ομαδοποίησης στα υπόλοιπα σημεία

current_cluster_label ← 1
for all core points **do**
  if the core point has no cluster label **then**
    current_cluster_label ← *current_cluster_label* + 1
    Label the current core point with cluster label *current_cluster_label*
  end if
  for all points in the *Eps*-neighborhood, except i^{th} the point itself **do**
    if the point does not have a cluster label **then**
      Label the point with cluster label *current_cluster_label*
    end if
  end for
end for

DBSCAN: Σημεία κύρια, συνοριακά και θορύβου

Αρχικά σημεία

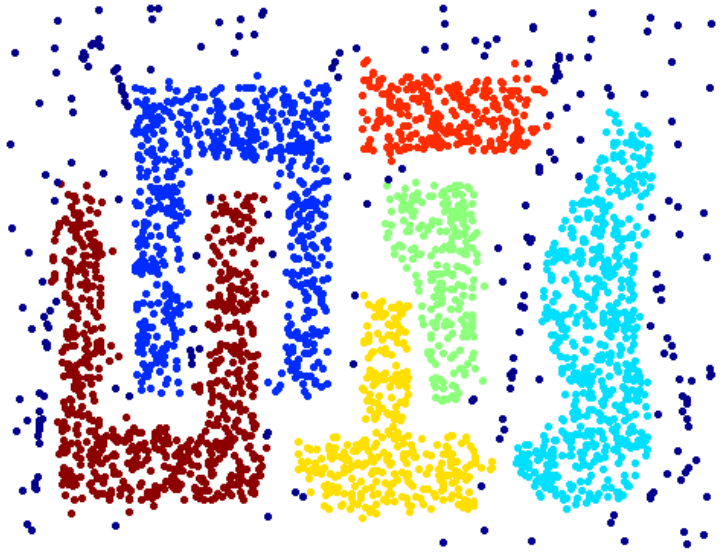


Σημεία κύρια,
συνοριακά και θορύβου

Eps = 10, MinPts = 4

Όταν ο DBSCAN δουλεύει καλά

Αρχικά σημεία

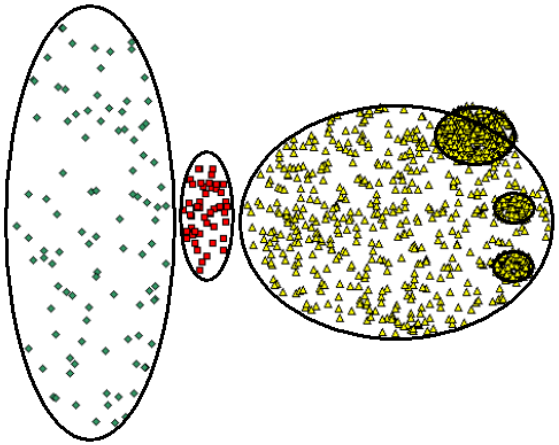


Ομάδες

Ανθεκτικός σε θόρυβο

Μπορεί να διαχειριστεί διαφορετικά σχήματα και μεγέθη ομάδων

Όταν ο DBSCAN δε δουλεύει καλά

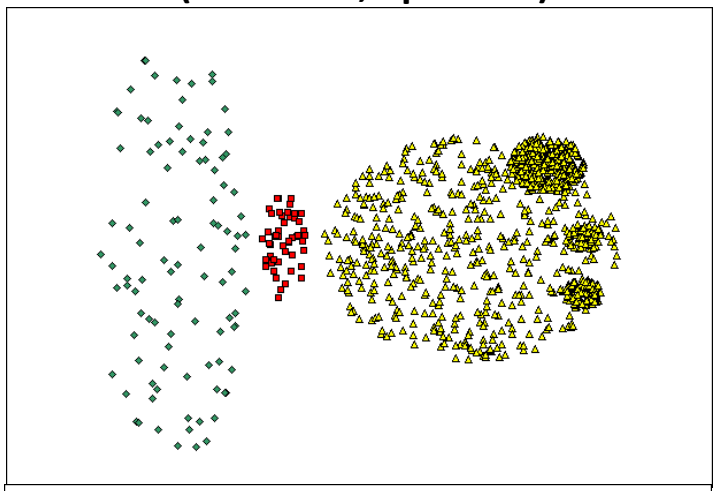


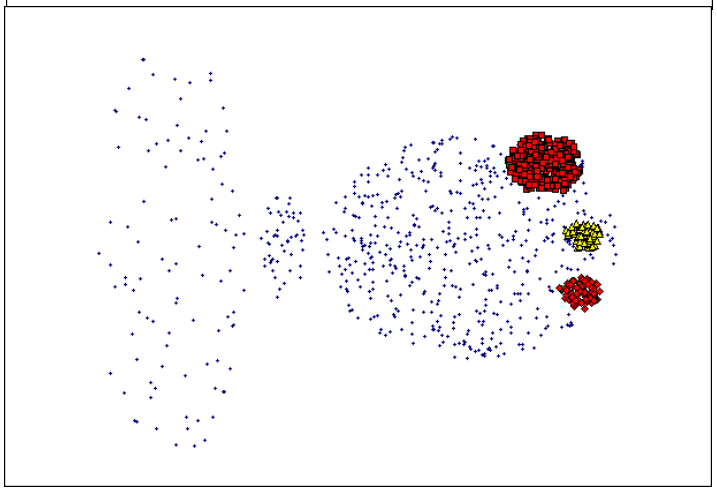
Αρχικά σημεία

Διαφορετικές πυκνότητες

Πολυδιάστατα δεδομένα

(MinPts=4, Eps=9.75)



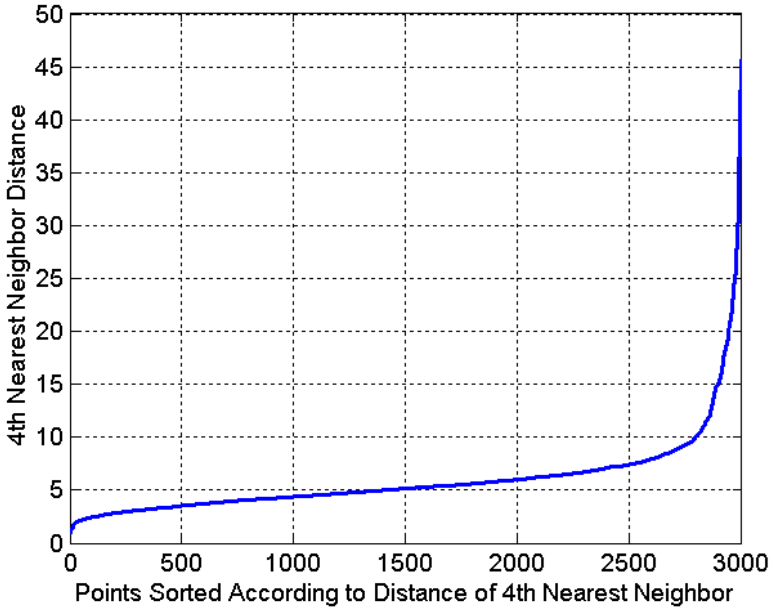


(MinPts=4, Eps=9.92)

DBSCAN: Ορίζοντας το EPS και το MinPts

- Η ιδέα είναι ότι, για τα σημεία μιας ομάδας, οι k^{th} κοντινότεροι γείτονές τους βρίσκονται στην ίδια περίπου απόσταση
- Τα σημεία θορύβου έχουν τον k^{th} κοντινότερο γείτονά τους σε μεγαλύτερη απόσταση

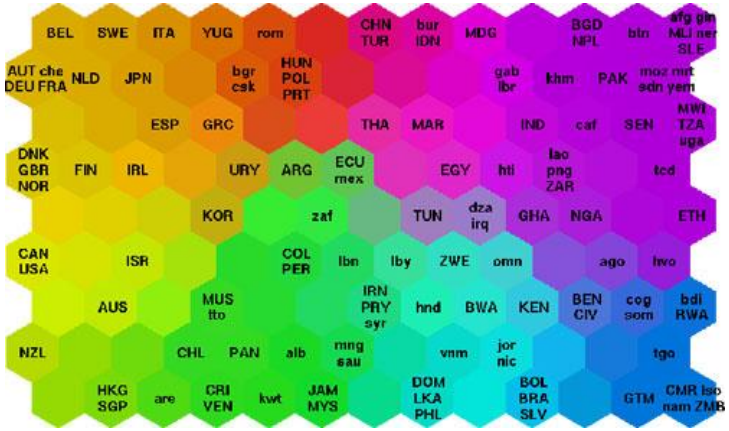
$\Rightarrow$ Δημιουργήστε ένα διατεταγμένο διάγραμμα της απόστασης κάθε σημείου από τον k^{th} κοντινότερο γείτονά τους

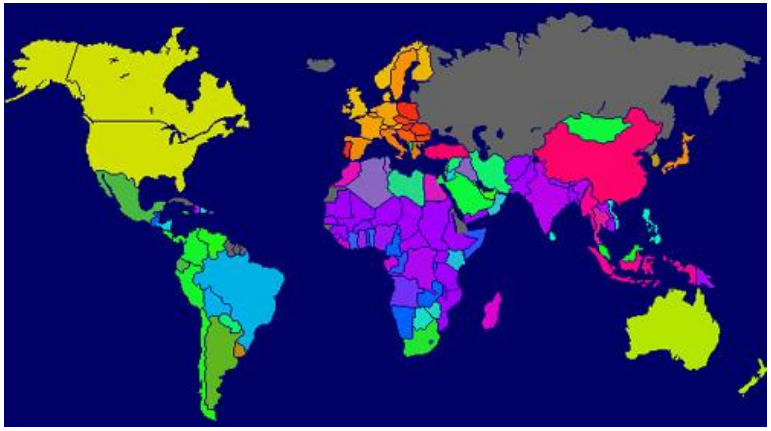


Self-organizing maps (SOMs)

- Ένα SOM αναπαριστά n-διάστατα δεδομένα σε 2-D ή 3-D μορφή, έτσι ώστε παρόμοια δεδομένα ομαδοποιούνται μαζί.
- Μετά τη σύγκλιση του SOM, δεν μπορεί να γίνει επανεκπαίδευση. Μπορεί απλά να ταξινομεί νέα δεδομένα στις ομάδες που έχει αναγνωρίσει.
- Το SOM έχει δυο φάσεις:
  - Φάση εκπαίδευσης: φτιάχνεται ο χάρτης, το δίκτυο οργανώνεται ακολουθώντας μια ανταγωνιστική διαδικασία
  - Φάση αντιστοίχισης: νέα δεδομένα (n-διάστατα διανύσματα) τοποθετούνται στον χάρτη και αντιστοιχίζονται σε ομάδα.

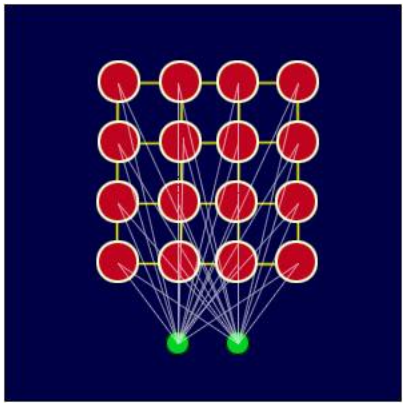
Χρήσεις των SOMs

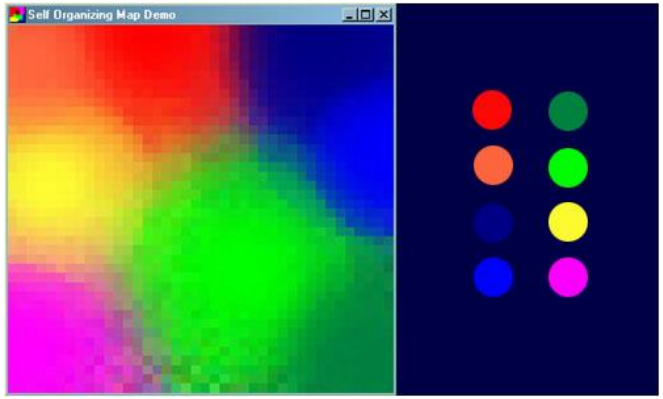


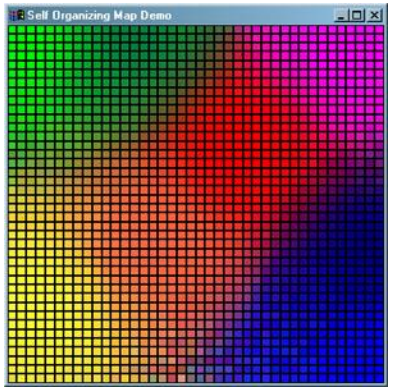


- Παράδειγμα: Σετ δεδομένων για το επίπεδο οικονομικής κατάστασης σε διάφορες χώρες.
  - Το σετ δεδομένων έχει πολλά στατιστικά στοιχεία για το επίπεδο οικονομικής κατάστασης κάθε χώρας.
  - Το SOM δε δείχνει μόνο την οικονομική κατάσταση, αλλά και τις ομοιότητες ανάμεσα στις καταστάσεις των χωρών (Παρόμοιο χρώμα = παρόμοια στοιχεία).

Η δομή του SOM







- Κάθε κόμβος συνδέεται με την είσοδο με τον ίδιο τρόπο και κανένας κόμβος δε συνδέεται με άλλον.
- Σε ένα SOM που προσπαθεί να ομαδοποιήσει παρόμοια χρώματα βάσει των τιμών RGB, κάθε “pixel” στην εικόνα είναι και κόμβος στο SOM.
- Θεωρήστε κάθε βάρος κόμβου ως ένα κέντρο ομάδας

Η διαδικασία δημιουργίας SOM

Ψευδοκώδικας

1. Αρχικοποίησε τα κέντρα (βάρη των κόμβων).
2. **Επανέλαβε**
3. Επέλεξε το επόμενο σημείο
4. Βρες ποιο είναι το πιο κοντινό κέντρο στο σημείο (Best Matching Unit - BMU)
5. Υπολόγισε την ακτίνα της γειτονιάς γύρω από το BMU. Το μέγεθος της γειτονιάς μειώνεται σε κάθε επανάληψη.
6. Ανανέωσε την τιμή του BMU και των γειτονικών κέντρων.
7. **Μέχρι** να συγκλίνει ο αλγόριθμος ή να φτάσει ο μέγιστος αριθμός επαναλήψεων
8. Τοποθέτησε κάθε σημείο στο πιο κοντινό του BMU και επέστρεψε τα κέντρα και τις ομάδες

Υπολογισμός της BMU

- Ο υπολογισμός της BMU γίνεται με τη χρήση της Ευκλείδειας απόστασης ανάμεσα στο κέντρο (W_1, W_2, … , W_n) και κάθε σημείο (V_1, V_2, … , V_n).

$$Dist = \sqrt{\sum_{i=0}^{i=n} (V_i - W_i)^2}$$

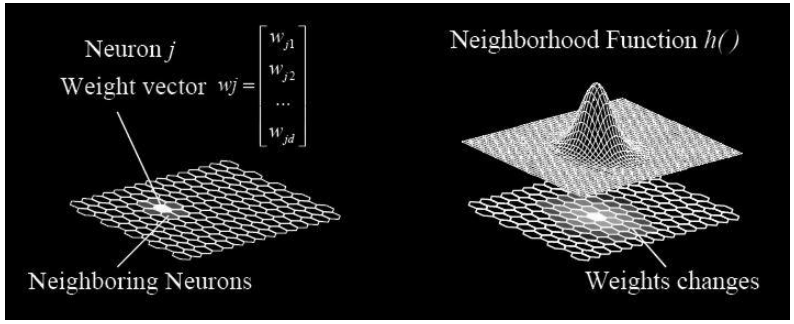
Υπολογισμός της γειτονιάς του BMU

- Μέγεθος της γειτονιάς: Χρήση μιας *εκθετικά φθίνουσας* συνάρτησης, η οποία μειώνεται σε κάθε επανάληψη, έτσι ώστε τελικά η γειτονιά του BMU να είναι το ίδιο το BMU

$$\sigma(t) = \sigma_0 \, exp\left(-\frac{t}{\lambda}\right)$$

- Η γειτονιά ορίζεται από μια καμπύλη gauss. Κόμβοι κοντά στην BMU επηρεάζονται περισσότερο.

$$\Theta(t) = exp\left(-\frac{dist^2}{2\sigma^2(t)}\right)$$



Ανανέωση των βαρών (κέντρων)

- Το βάρος ενός κόμβου:

$$W(t+1) = W(t) + \Theta(t) L(t) (V(t) - W(t))$$

όπου ο ρυθμός εκμάθησης, *L(t)*, είναι επίσης μια *εκθετικά φθίνουσα* συνάρτηση:

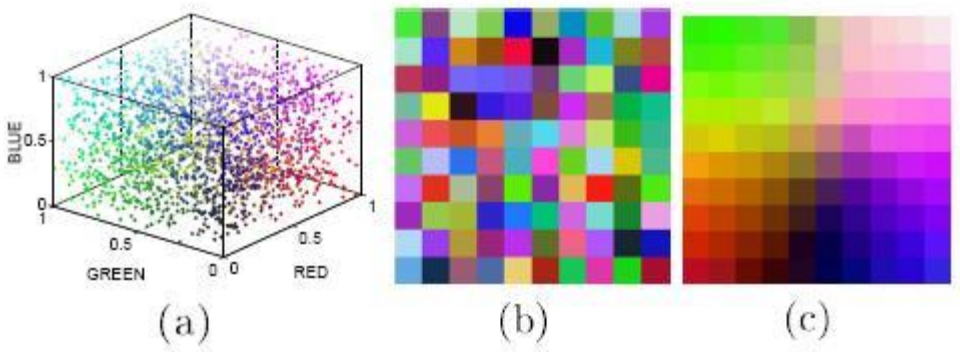
$$L(t) = L_0 \, exp\left(-\frac{t}{\lambda}\right)$$

  - λ είναι μια χρονική σταθερά
- Αυτό εξασφαλίζει σύγκλιση του SOM

Παράδειγμα εφαρμογής

- 2-D πλέγμα από κόμβους
- Οι είσοδοι είναι χρώματα
- Το SOM συγκλίνει ώστε παρόμοια χρώματα να ομαδοποιούνται μαζί
- Κώδικας και demo στο:

http://www.ai-junkie.com/files/SOMDemo.zip



a) Σετ δεδομένων
b) Αρχικά βάρη
c) Τελικά βάρη

Ανακεφαλαίωση

- Μη επιβλεπόμενη μάθηση
- Διάφοροι τύποι ομαδοποίησης
  - Διαχωρισμού
  - Ιεραρχική
  - Πυκνωτική
  - SOM
- Πλεονεκτήματα/μειονεκτήματα κάθε μεθόδου
- Αξιολόγηση ομάδων με βάση εξωτερικούς, εσωτερικούς και σχετικούς δείκτες
- Πηγές:
  - Introduction to Data Mining, Tan, Steinbach, Kumar.
  - Pattern recognition, Theodoridis & Koutroumbas.
  - Corby Ziesman slides on SOMs, 2007.