

Συσταδοποίηση III

Μέρος των διαφανειών είναι από το P.-N. Tan, M. Steinbach, V. Kumar, «Introduction to Data Mining», Addison Wesley, 2006

DBSCAN

DBSCAN: Γενικά

Ο DBSCAN είναι ένας αλγόριθμος βασισμένος στην πυκνότητα
Πυκνότητα = αριθμός σημείων (*MinPts*) μέσα σε ακτίνα *Eps*

Τα σημεία διαχωρίζονται σε:

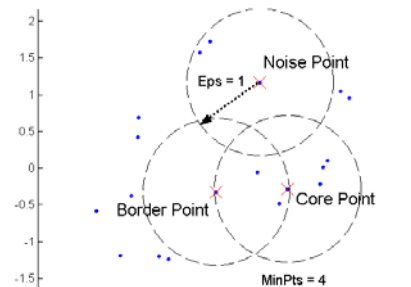
- **Βασικά (core):** ένα σημείο για το οποίο υπάρχουν περισσότερα από ένα προκαθορισμένο αριθμό (*MinPts*) σημεία σε ακτίνα *Eps*

Αυτά είναι τα σημεία που είναι στο εσωτερικό μιας συστάδας

- **Οριακά (border):** ένα σημείο για το οποίο υπάρχουν λιγότερα από ένα προκαθορισμένο αριθμό (*MinPts*) σημεία σε ακτίνα *Eps*, αλλά είναι στη γειτονιά ενός βασικού σημείου

- **Θορύβου (noise):** ένα σημείο που δεν είναι ούτε βασικό ούτε οριακό

DBSCAN: Γενικά



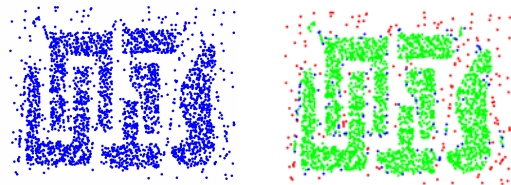
DBSCAN: Αλγόριθμος

Βασικός Αλγόριθμος

- 1: Χαρακτήρισε κάθε σημείο ως βασικό, οριακό ή θόρυβο
- 2: Διέγραψε τα σημεία θορύβου
- 3: Τοποθέτησε μια ακμή μεταξύ όλων των βασικών σημείων που είναι σε απόσταση έως *Eps* μεταξύ τους
- 4: Κάνε κάθε ομάδα συνδεδεμένων βασικών σημείων μια διαφορετική συστάδα
- 5: Ανάθεσε κάθε οριακό σημεία σε μία από τις συστάδες των συσχετιζόμενων του βασικών σημείων

DBSCAN: Αλγόριθμος

Βήμα 1&2




Αρχικά σημεία

Τύποι σημείων: core,
border και noise

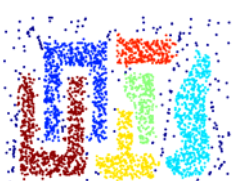
Eps = 10, MinPts = 4

DBSCAN: Πλεονεκτήματα

Βήμα 3&4



Αρχικά Σημεία



Συστάδες

- Δεν επηρεάζεται από το θόρυβο
- Μπορεί να χειριστεί συστάδες με διαφορετικά σχήματα και μεγέθη

Εξώφυλλο Δεδομένων: Ακ. Έτος 2007-2008 ΣΥΓΧΡΟΝΗ ΠΡΟΣΕΓΓΙΣΗ III 7

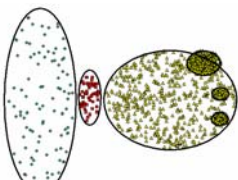
DBSCAN: Πολυπλοκότητα

$O(m \times \text{χρόνος εντοπισμού σημείων σε εψς-γειτονιά})$
 $O(m^2)$
 Για μικρό αριθμό διαστάσεων, υπάρχουν δομές που υποστηρίζουν την πράξη σε $O(m \log m)$

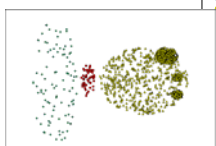
$O(m)$ χώρος (κρατάμε μόνο ένα label)

Εξώφυλλο Δεδομένων: Ακ. Έτος 2007-2008 ΣΥΓΧΡΟΝΗ ΠΡΟΣΕΓΓΙΣΗ III 8

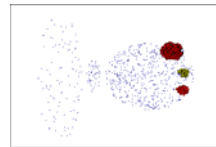
DBSCAN: Περιορισμοί



Αρχικά Σημεία



(MinPts=4, Eps=9.75)



(MinPts=4, Eps=9.92)

- Διαφορετικές πυκνότητες
- Πολυ-διάστατα δεδομένα – δύσκολος ορισμός πυκνότητας και δαπανηρός υπολογισμός γειτόνων

Εξώφυλλο Δεδομένων: Ακ. Έτος 2007-2008 ΣΥΓΧΡΟΝΗ ΠΡΟΣΕΓΓΙΣΗ III 9

DBSCAN: Καθορισμός των MinPts και Eps

Η ιδέα είναι να κοιτάζουμε την απόσταση ενός σημείου από τον k-στό κοντινότερο γείτονα του → k-dist

Γενικά, για τα σημεία που ανήκουν στην ίδια ομάδα, η τιμή του k-dist θα είναι μικρή (αν το k δεν είναι μεγαλύτερο από το μέγεθος της συστάδας)

Θα θέλαμε για τα σημεία μιας συστάδας, να έχουν περίπου την ίδια k-dist

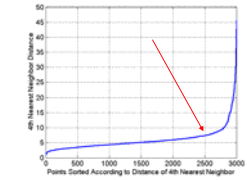
Τα σημεία θορύβου έχουν μεγαλύτερες k-dist

Υπολογίζουμε την k-dist για όλα τα σημεία, για κάποιο k

Ταξινομούμε τις αποστάσεις με φθίνουσα διάταξη

Περιμένουμε ξαφνική αλλαγή στο k-dist που αντιστοιχεί στο Eps

Οπότε $k = \text{MinPts}$ και $\text{Eps} = k\text{-dist}$



4th Nearest Neighbor Distance
Points Sorted According to Distance of 4th Nearest Neighbor

Εξώφυλλο Δεδομένων: Ακ. Έτος 2007-2008 ΣΥΓΧΡΟΝΗ ΠΡΟΣΕΓΓΙΣΗ III 10

BIRCH

T. Zhang, R. Ramakrishnan and M. Livny, BIRCH: An Efficient Data Clustering Method for Very Large Databases, SIGMOD 1996

Εξώφυλλο Δεδομένων: Ακ. Έτος 2007-2008 ΣΥΓΧΡΟΝΗ ΠΡΟΣΕΓΓΙΣΗ III 11

BIRCH

Μεγάλα Σύνολα Δεδομένων

Περιορισμένη μνήμη (πολύ μικρότερη από το μέγεθος των δεδομένων)

ΣΤΟΧΟΣ: μείωση του χρόνου εισόδου/εξόδου (I/O)

- Κόστος I/O γραμμικό στο μέγεθος του συνόλου δεδομένων
 - Αρκεί ένα πέρασμα (scan) των δεδομένων
 - Ένα ή περισσότερα επιπρόσθετα πέρασματα για βελτίωση της ποιότητας της συσταδοποίησης

BIRCH: Balanced Iterative Reducing and Clustering using Hierarchies

Εξώφυλλο Δεδομένων: Ακ. Έτος 2007-2008 ΣΥΓΧΡΟΝΗ ΠΡΟΣΕΓΓΙΣΗ III 12



Αντί να κρατάμε όλα τα σημεία μιας συστάδας κρατάμε κάποια «στατιστικά» για κάθε συστάδα και για τις σχέσεις μεταξύ των συστάδων



Έστω μια συστάδα σημείων: $\{\vec{X}_i\}$

Centroid(κεντρικό σημείο): $\vec{X}_0 = \frac{\sum_{i=1}^N \vec{X}_i}{N}$

Radius (ακτίνα) μέση απόσταση των σημείων της συστάδας από το κεντρικό σημείο

$$R = \left(\frac{\sum_{i=1}^N (\vec{X}_i - \vec{X}_0)^2}{N} \right)^{\frac{1}{2}}$$

Diameter (διάμετρος): μέση ανα-δύο απόσταση των σημείων της συστάδας

$$D = \left(\frac{\sum_{i=1}^N \sum_{j=1}^N (\vec{X}_i - \vec{X}_j)^2}{N(N-1)} \right)^{\frac{1}{2}}$$



Μας ενδιαφέρει και η απόσταση των κεντρικών σημείων δύο συστάδων

Μεταξύ δυο συστάδων

centroid Euclidean distance $D0 = ((\vec{X}_{01} - \vec{X}_{02})^2)^{\frac{1}{2}}$

centroid Manhattan distance $D1 = |\vec{X}_{01} - \vec{X}_{02}| = \sum_{k=1}^d |\vec{X}_{01}^{(k)} - \vec{X}_{02}^{(k)}|$



Συγχώνευση Συστάδων

Συστάδα $\{X_i\}$:
 $i = 1, 2, \dots, N_1$

X_i

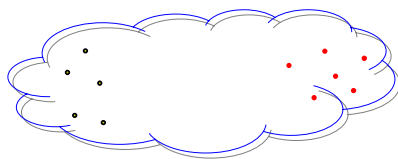


Συστάδα $\{X_j\}$:
 $j = N_1+1, N_1+2, \dots, N_1+N_2$

X_j



Συστάδα $X_k = \{X_i\} + \{X_j\}$:
 $l = 1, 2, \dots, N_1, N_1+1, N_1+2, \dots, N_1+N_2$



average inter-cluster (D2) μέση απόσταση των σημείων της μιας συστάδας από τα σημεία της άλλης

$$D2 = \left(\frac{\sum_{i=1}^{N_1} \sum_{j=N_1+1}^{N_1+N_2} (\vec{X}_i - \vec{X}_j)^2}{N_1 N_2} \right)^{\frac{1}{2}}$$

intra-cluster (D3) μέση απόσταση όλων των σημείων

$$D3 = \left(\frac{\sum_{i=1}^{N_1+N_2} \sum_{j=1}^{N_1+N_2} (\vec{X}_i - \vec{X}_j)^2}{(N_1 + N_2)(N_1 + N_2 - 1)} \right)^{\frac{1}{2}}$$

D της συγχωνευμένης συστάδας

variance increase (D4)

Νέα Απόσταση

$$D4 = \left(\frac{\sum_{i=1}^{N_1+N_2} (\vec{X}_i - \frac{\sum_{i=1}^{N_1+N_2} \vec{X}_i}{N_1+N_2})^2}{N_1+N_2} - \frac{\sum_{i=1}^{N_1} (\vec{X}_i - \frac{\sum_{i=1}^{N_1} \vec{X}_i}{N_1})^2}{N_1} - \frac{\sum_{j=N_1+1}^{N_1+N_2} (\vec{X}_j - \frac{\sum_{j=N_1+1}^{N_1+N_2} \vec{X}_j}{N_2})^2}{N_2} \right)^{\frac{1}{2}}$$

Απόσταση στο C_i

Απόσταση στο C_j

BIRCH: CF

Clustering Feature (CF): μια περίληψη μιας υπο-συστάδας δεδομένων - Μια τριάδα (αριθμός-σημείων, γραμμικό-άθροισμα-σημείων-συστάδας, άθροισμα-τετραγώνου-σημείων-συστάδας)

Given a cluster $\{\vec{X}_1, \vec{X}_2, \dots, \vec{X}_N\}$

$$CF = (N, \vec{LS}, SS)$$

N is the number of data points

$$\vec{LS} = \sum_{i=1}^N \vec{X}_i$$

$$SS = \sum_{i=1}^N \vec{X}_i^2$$

Σημαντική (προσθετική) ιδιότητα:

$$CF_1 + CF_2 = (N_1 + N_2, \vec{LS}_1 + \vec{LS}_2, SS_1 + SS_2)$$

BIRCH: CF

• CF εγγραφές είναι συνοπτικές - πολύ λιγότερη πληροφορία από ότι όλα τα σημεία της υπο-συστάδας

• Λόγω της προσθετικής ιδιότητας μπορούμε να συγχωνεύσουμε δυο υπο-συστάδες σταδιακά

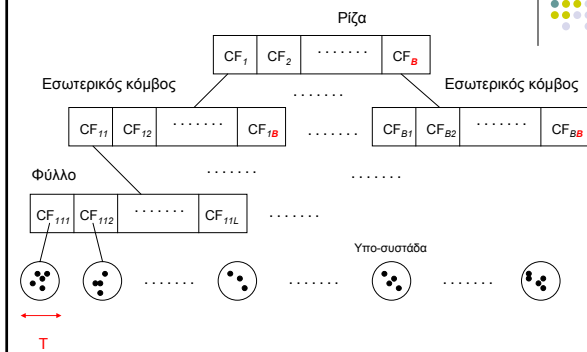
Μια εγγραφή CF έχει αρκετή πληροφορία για να υπολογίσουμε τα D0-D4

BIRCH

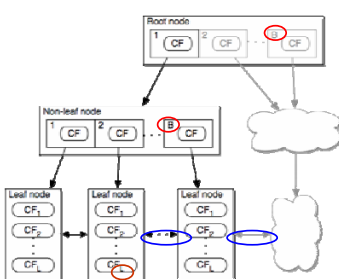
Ιεραρχικός αλγόριθμος

Χτίζει σταδιακά καθώς διαβάζει τα δεδομένα ένα δεντρογράμμα του οποίου κόμβοι είναι οι τιμές CF που περιγράφουν τα δεδομένα κάθε υπο-συστάδας

BIRCH: CF-δέντρο



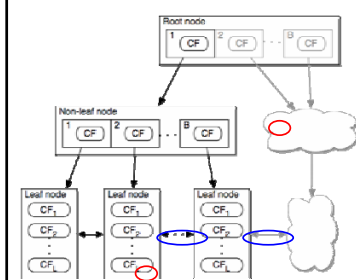
BIRCH: CF-δέντρο



- Κάθε εσωτερικός κόμβος περιέχει έναν αριθμό από παιδιά - B (παράγοντας διακλάδωσης) εγγραφές $\langle CF_i, \text{παιδί} \rangle$
- Κάθε φύλλο περιέχει έναν αριθμό από υπο-συστάδες το πολύ L CF εγγραφές $[CF_i]$ και $\langle \text{prev}, \langle \text{next} \rangle$ pointers

Κάθε εσωτερικός κόμβος μια υποσυστάδα που αποτελείται από τις υποσυστάδες των παιδιών του

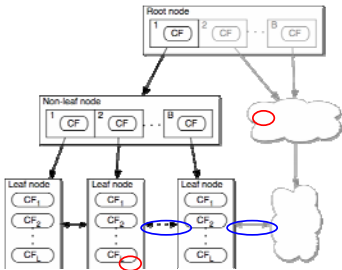
BIRCH: CF-δέντρο



- Όπως σε όλες τις σχετικές δομές απαιτούμε κάθε κόμβος του δέντρου να χωρά σε ένα block

Το μέγεθος των κόμβων (B, L) καθορίζεται από τη διάσταση των δεδομένων και το μέγεθος της σελίδας P (που δίνεται ως είσοδος)

BIRCH: CF-δέντρο



Κάθε υποσυστάδα ενός φύλλου πρέπει να έχει

διάμετρο μικρότερη από κάποιο κατώφλι **T**

Το μέγεθος του **T** καθορίζει το μέγεθος του δέντρου

Όσο πιο μεγάλο είναι το **T**, τόσο μικρότερο είναι το δέντρο

BIRCH: CF δέντρο

Συνοπτικά, το CF-δέντρο είναι ένα ισοζυγισμένο δέντρο με δυο παραμέτρους

- Παράγοντα διακλάδωσης **B** (που καθορίζεται από το μέγεθος του block)
- Κατώφλι **T** (που καθορίζει την ποιότητα της συσταδοποίησης)

BIRCH: CF-δέντρο

Για ένα φύλλο:

$$LS = \sum_{i \in N} \bar{P}_i$$

$$SS = \sum_{i \in N} |\bar{P}_i|^2$$

Για κάθε εσωτερικό κόμβο που έχει παιδιά τα N_1, N_2, \dots, N_k

$$\bar{LS} = \sum_{i=1}^k \bar{LS} \text{ of } N_i$$

$$SS = \sum_{i=1}^k SS \text{ of } N_i$$

BIRCH: CF-δέντρο εισαγωγή στοιχείου

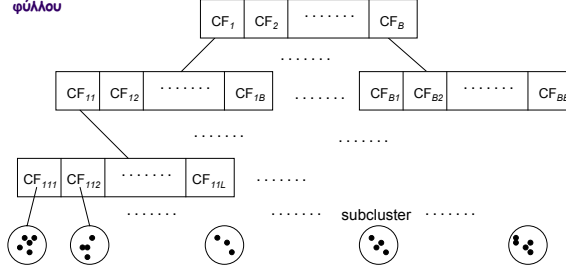
- Ο αλγόριθμος διαβάζει (scan) τα δεδομένα και τα εισάγει στο CF δέντρο ένα-ένα
- Η εισαγωγή ενός στοιχείου στο CF-δέντρο γίνεται με top-down διάσχιση ξεκινώντας από τη ρίζα με βάση μια συνάρτηση απόστασης Distance(σημείο, cluster)
 - Χρήση της D0, D1, D2, D3 ή D4
- Κάθε σημείο εισάγεται στην κοντινότερη υπο-συστάδα που υπάρχει σε κάποιο από τα φύλλα

BIRCH: CF-δέντρο εισαγωγή στοιχείου

1. Εύρεση κατάλληλου φύλλου
αν το φύλλο μπορεί να το απορροφήσει (διάμετρος παραμένει $\leq T$) οκ, Αλλιώς 3
2. Ενημέρωση του φύλλου
3. Διάσπαση φύλλου
4. Ενημέρωση τιμής CF

BIRCH: CF-δέντρο εισαγωγή στοιχείου

1. Εύρεση φύλλου



BIRCH: CF-δέντρο εισαγωγή στοιχείου

4. Τροποποίηση CF τιμών

31

BIRCH: CF-δέντρο εισαγωγή στοιχείου

2. Μετατροπή φύλλου

32

BIRCH: CF-δέντρο εισαγωγή στοιχείου

3-4 Διάσπαση και ενημέρωση του μονοπατιού από τη ρίζα

33

BIRCH: CF-δέντρο εισαγωγή στοιχείου

3-4 Διάσπαση και ενημέρωση του μονοπατιού από τη ρίζα

34

BIRCH: CF-δέντρο

- Κάθε σημείο εισάγεται στο κοντινότερη υπο-συστάδα που υπάρχει σε κάποιο από τα φύλλα
 - Αν η εισαγωγή ενός σημείου **μεγαλώσει τη διάμετρο της υποσυστάδας πάνω από T**, τότε έχουμε δημιουργία νέας υποσυστάδας
 - Αν η νέα συστάδα χωρά στο φύλλο, ok -> ενημέρωση προγόνων
 - Αν η νέα συστάδα δε χωρά -> υπερχείλιση στο φύλλο

35

BIRCH: CF-δέντρο

- Διάσπαση φύλλου (split)**
 - Δημιουργία νέου φύλλου και μοίρασμα των συστάδων, **πως?**

Εύρεση των δύο υπο-συστάδων του φύλλου που έχουν τη μεγαλύτερη απόσταση μεταξύ τους, έστω C_i και C_j

Αυτές οι δύο αποτελούν το κριτήριο διάσπασης των υπο-συστάδων του φύλλου - κάθε μια από αυτές σε ένα από τα δύο νέα φύλλα

όλες οι άλλες υπο-συστάδες C ανατίθενται στο φύλλο της C_i ή στο φύλλο της C_j με βάση ποια από τις δύο είναι πιο όμοια της

36

BIRCH: CF-δέντρο

Διάσπαση φύλλου μπορεί να οδηγήσει σε υπερχειλίση εσωτερικού κόμβου (όταν περιέχει περισσότερα παιδιά από ότι ο παράγοντας διακλάδωσης)

Διάσπαση εσωτερικού κόμβου

- Οι εσωτερικοί κόμβοι διασπώνται αναδρομικά με βάση μια μέτρηση της απόστασης των συστάδων τους
- Διάσπαση της ρίζας, οδηγεί σε αύξηση του ύψους του δέντρου κατά 1

BIRCH: CF-δέντρο

Οι διασπάσεις οφείλονται στο ότι ξεπερνιέται το όριο της σελίδας - μπορούν να οδηγήσουν σε κακές διασπάσεις!

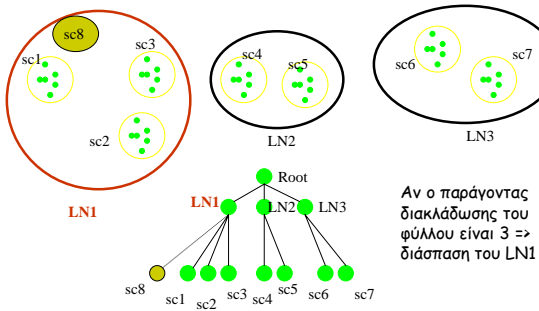
Μια μικρή διόρθωση:

- Όταν η διάσπαση κάποιων κόμβων τελειώνει (χωρούν σε ένα κόμβο) έστω στον κόμβο N_i κοιτάμε τον κόμβο N_j και προσπαθούμε να συγχωνεύσουμε τις δύο πιο κοντινές συστάδες - αν αυτές δε προέκυψαν από την πιο πρόσφατη διάσπαση
- Αυτό σημαίνει ότι πρέπει να συγχωνεύσουμε και τα αντίστοιχα 2 παιδιά
- Αν χωρούν σε μια σελίδα -> ελάττωση χώρου,
- Αλλιώς ανακατανέμουμε τις εγγραφές - Πως; κάνουμε πάλι διάσπαση

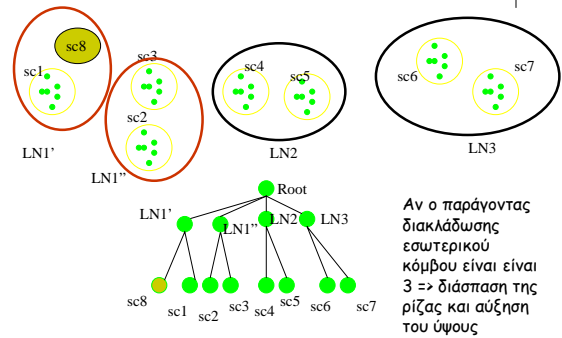
Τελικά ή συγχώνευση και ελευθέρωση χώρου ή καλύτερη ανακατανομή των εγγραφών σε κάποιο από τα παιδιά

BIRCH: CF-δέντρο

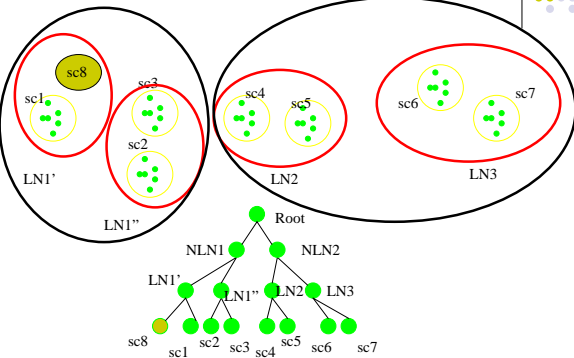
Νέα υπο-συστάδα



BIRCH: CF-δέντρο

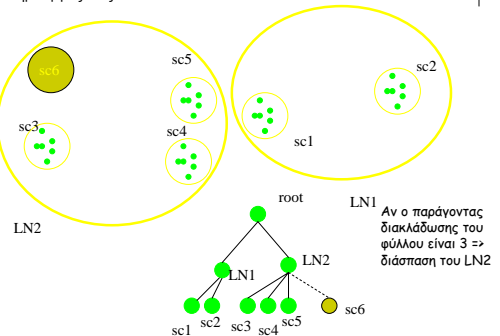


BIRCH: CF-δέντρο



BIRCH: CF-δέντρο

Έστω ότι η αρίθμηση των υποσυστάδων αντιστοιχεί στη σειρά δημιουργίας τους



BIRCH: CF-δέντρο

root Merge?

LN1 LN3'' LN2''

sc1 sc2 sc3 sc4 sc5 sc6

Εξώφυλλο Διδασκόντων: Ακ. Έτος 2007-2008 ΣΥΓΓΡΑΜΜΑΤΟΓΡΑΦΙΑ III 43

BIRCH: CF-δέντρο

LN3' LN3'' LN2''

sc2 sc3 sc4 sc5 sc6

Συγχώνευση LN2' και LN1 και ο καινούργιος κόμβος θα διασπαστεί πάλι

Εξώφυλλο Διδασκόντων: Ακ. Έτος 2007-2008 ΣΥΓΓΡΑΜΜΑΤΟΓΡΑΦΙΑ III 44

BIRCH: αλγόριθμος

Επειδή η κατασκευή επηρεάζεται από το μέγεθος της σελίδας:

- Οι συστάδες που δημιουργούνται μπορεί να μην είναι πραγματικές
- ανάλογα με το skew (κατανομή) και τη σειρά που έρχονται τα δεδομένα

Επίσης, αν ξανά-εισαγάμε ένα σημείο μπορεί να εισαχθεί σε διαφορετική συστάδα

Αριθμός αντιστοιχεί στη σειρά εισαγωγής. Έστω $dist(1, 2) > T$

Εξώφυλλο Διδασκόντων: Ακ. Έτος 2007-2008 ΣΥΓΓΡΑΜΜΑΤΟΓΡΑΦΙΑ III 45

BIRCH: αλγόριθμος

Αυτό αντιμετωπίζεται με προαιρετικές επιπρόσθετες φάσεις

Εξώφυλλο Διδασκόντων: Ακ. Έτος 2007-2008 ΣΥΓΓΡΑΜΜΑΤΟΓΡΑΦΙΑ III 46

BIRCH-αλγόριθμος

Φάση 1: Μια δομή κώδικα μνήμης που συνοψίζει τα δεδομένα

Φάση 2: Κοιτά τα φύλλα και προσπαθεί να διώξει τους outliers και να εντοπίσει «όμοιες» συστάδες που αντιστοιχούν σε περιοχές με πολλά σημεία. Χρειάζεται για να βελτιώσει τη Φάση 3

Εξώφυλλο Διδασκόντων: Ακ. Έτος 2007-2008 ΣΥΓΓΡΑΜΜΑΤΟΓΡΑΦΙΑ III 47

BIRCH

Φάση 3

Ξανα-συσταδοποιεί τα φύλλα του δέντρου

Γιατί:
Πχ κοντινές συστάδες που (έτυχε να) είναι σε διαφορετικά φύλλα

Πως:

- Για κάθε συστάδα που εμφανίζεται στα φύλλα, υπολογίζουμε το κεντρικό της σημείο (centroid) και τα θεωρούμε ως αρχικά σημεία - αυτά τα αρχικά σημεία μπορούμε να τα συσταδοποιήσουμε χρησιμοποιώντας έναν οποιαδήποτε αλγόριθμο συσταδοποίησης
- Μπορούμε αντί ένα σημείο ανά συστάδα, κάθε συστάδα τόσες φορές όσες τα σημεία της
- Εναλλακτικά, μπορούμε να συσταδοποιήσουμε τις συστάδες ως έχουν - πχ με έναν ιεραρχικό συγκεντρωτικό αλγόριθμο

Εξώφυλλο Διδασκόντων: Ακ. Έτος 2007-2008 ΣΥΓΓΡΑΜΜΑΤΟΓΡΑΦΙΑ III 48



Φάση 4 (προαιρετική)

Χρησιμοποιεί τα κεντρικά σημεία των συστάδων που παράγει η Φάση 3 ως seeds, και αναδιανέμει όλα τα στοιχεία εισόδου (δευτερο πέρασμα!)

Μπορεί να έχουμε και παραπάνω από ένα επιπρόσθετα περάσματα (έχει αποδειχτεί σύγκλιση)

- Εξασφαλίζει ότι όλα τα αντίγραφα ενός σημείου πάνε στην ίδια συστάδα
- Μπορούμε επίσης να βάλουμε ως ετικέτα σε κάθε σημείο, τη συστάδα που ανήκει
- Μπορούμε να απαλλαγούμε από outliers (πχ σημεία πολύ μακριά από όλα τα seeds)



Λίγα ακόμα για τη Φάση 1

Ξεκίνα με κάποια αρχική τιμή για το threshold (T)

Διαβάζει τα δεδομένα και τα εισάγει στο δέντρο

Αν ξεπεράσει το διαθέσιμο χώρο πριν διαβάσει όλα τα δεδομένα:

Αύξηση του threshold
Κτίσιμο νέου (μικρότερου) δέντρου ξανα-
εισάγοντας τις τιμές από το παλιό δέντρο

Μόλις εισαχθούν όλες οι τιμές από το παλιό στο νέο δέντρο,
Συνεχίζεται η ανάγνωση των δεδομένων από εκεί που
είχε σταματήσει

Πώς γίνεται η ανα-κατασκευή



Μονοπάτι-Μονοπάτι

Ανακατασκευάζουμε κάθε μονοπάτι από τη ρίζα στο φύλλο, ξεκινώντας από το πιο αριστερό μονοπάτι (old-current path)

Δημιουργούμε το new-current path

Κάθε φύλλο είτε στο new είτε στο newclosest

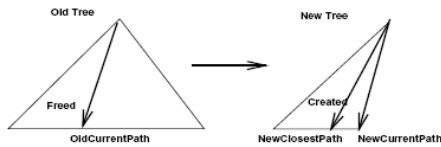


Figure 3: Rebuilding CF Tree



1. Create the corresponding "NewCurrentPath" in the new tree
2. Insert leaf entries in "OldCurrentPath" to the new tree
 - ⓐ NewClosestPath
 - ⓑ NewCurrentPath
3. Free space in "OldCurrentPath" and "NewCurrentPath"
4. Set "OldCurrentPath" to the next path if there exists one



Και άλλες βελτιώσεις όπως:

Έλεγχος για outliers
Delay-split



- Τοπικότητα: κάθε απόφαση σχετικά με συσταδοποίηση παίρνεται χωρίς να χρειάζεται να διαβαστούν όλα τα σημεία ή όλες οι υπάρχουσες συστάδες
- Σημεία σε αραιές περιοχές θεωρούνται οριακά (outliers) και (προαιρετικά) αφαιρούνται
- Λαμβάνει υπ' όψιν τη διαθέσιμη μνήμη

Διαχείριση Ποιότητας Cluster validity

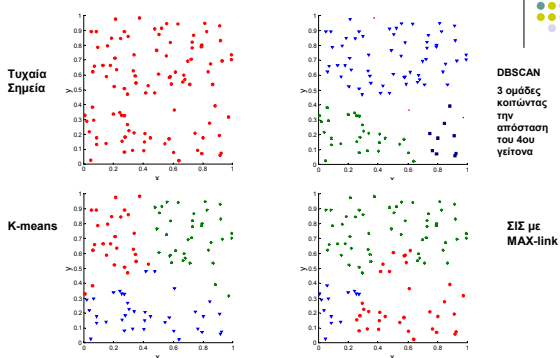
Ποιότητα Συσταδοποίησης

Πόσο καλή είναι συσταδοποίηση που επιτύχαμε;

Οι αλγόριθμοι που είδαμε παράγουν κάποιες συστάδες ακόμα και όταν τα δεδομένα παράγονται τυχαία

Δύσκολη η αξιολόγηση, ιδιαίτερα σε πολλές διαστάσεις

Συστάδες σε Τυχαία Δεδομένα



Κριτήρια Ορθότητας Συσταδοποίησης

1. Υπάρχει τάση ομαδοποίησης (clustering tendency), δηλαδή μη τυχαία δομή στο σύνολο των δεδομένων;
 2. Σύγκριση των αποτελεσμάτων της ανάλυσης της ομαδοποίησης με κάποια ήδη γνωστά αποτελέσματα, πχ κάποια ετικέτα που ήδη έχει δοθεί για μια συστάδα
 3. Πόσο καλά τα αποτελέσματα της ανάλυσης ταιριάζουν με τα δεδομένα χωρίς αναφορά σε εξωτερική πληροφορία, χρησιμοποιώντας μόνο τα δεδομένα
 4. Σύγκριση των αποτελεσμάτων δυο διαφορετικών συσταδοποιήσεων για να αποφασιστεί ποια είναι καλύτερη.
 5. Καθορισμός του «σωστού» αριθμού συστάδων
- Τα 2, 3 και 4 μπορεί να αφορούν είτε την ολική συσταδοποίηση είτε τη κάθε συστάδα χωριστά

Μετρήσεις Ποιότητας Συσταδοποίησης

Οι μετρήσεις για την ποιότητα (το πόσο καλή) είναι μια συσταδοποίηση ανήκουν σε μία από τις παρακάτω τρεις κατηγορίες:

▪ **Με επίβλεψη (supervised) - Εξωτερικό Ευρετήριο (External Index):**

Υπάρχει εξωτερική πληροφορία (πληροφορία εκτός των δεδομένων), πχ ετικέτες για τις συστάδες
Μετράμε πόσο οι περιγραφές των συστάδων ταιριάζουν με τις ετικέτες των κλάσεων. - πχ Εντροπία

▪ **Χωρίς επίβλεψη (unsupervised) Εσωτερικό Ευρετήριο (Internal Index):**

Εκτιμάμε το πόσο καλή είναι μια συσταδοποίηση χωρίς παροχή εξωτερικής πληροφορίας

- Συνεκτικότητα (cohesion)
- Διακριτότητα ή διαχωρισμός (separation)

Μετρήσεις Ποιότητας Συσταδοποίησης

▪ **Συγκριτικοί -Σχετικό Ευρετήριο (Relative Index):**

Χρησιμοποιείται για τη σύγκριση δυο διαφορετικών συσταδοποιήσεων ή συστάδων - Συχνά για αυτό το σκοπό χρησιμοποιείται ένα εσωτερικό ή εξωτερικό ευρετήριο
Εσωτερικό, πχ δυο k-means συσταδοποιήσεις με βάση το SSE

Κριτήρια vs Ευρετήρια - κριτήριο: η γενική στρατηγική και ευρετήριο η αριθμητική μέτρηση που υλοποιεί το κριτήριο

Χαρακτηρισμός Ποιότητας Συσταδοποίησης χωρίς Επίβλεψη:

- Χρήση Συνεκτικότητας και Διαχωρισμού
- Χρήση Πίνακα Γειτνίασης

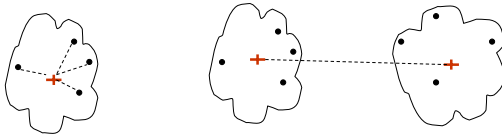
Χαρακτηρισμός Ποιότητας Συσταδοποίησης χωρίς Επίβλεψη:
Συνεκτικότητα και Διαχωρισμός

$$overall - validity = \sum_{i=1}^k w_i validity(C_i)$$

Όπου το βάρος (w_i) μπορεί να είναι πχ ανάλογο του μεγέθους της συστάδας ή η τετραγωνική ρίζα της συνεκτικότητας ή 1

Χαρακτηρισμός Ποιότητας Συσταδοποίησης χωρίς Επίβλεψη:
Συνεκτικότητα και Διαχωρισμός

Συσταδοποίηση βασισμένη σε κεντρικά σημεία - Centroid-based clustering (πχ k-means)



$$cohesion(C_i) = \sum_{x \in C_i} proximity(x, c_i)$$

Αν proximity = τετράγωνο της Ευκλείδειας, τότε ESS

Όπου c_i το κεντρικό σημείο (X_0) στον BIRCH
ακτίνα R στον BIRCH/k-means

$$separation(C_i, C_j) = proximity(c_i, c_j)$$

$$separation(C_i) = proximity(c_i, c)$$

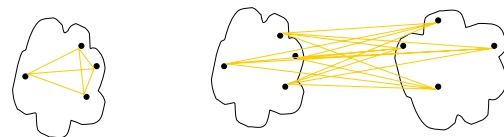
Όπου c το κέντρο όλων των σημείων

αντιστοιχεί στα D0 (D1) στον BIRCH

Χαρακτηρισμός Ποιότητας Συσταδοποίησης χωρίς Επίβλεψη:
Συνεκτικότητα και Διαχωρισμός

Συσταδοποίηση βασισμένη σε γραφήματα (ΣΓΣ)

- Η **συνεκτικότητα** μιας συστάδας (**cluster cohesion**) είναι το άθροισμα των βαρών (συνήθως απόσταση) μεταξύ όλων των συνδέσεων σε μια συστάδα.
- Ο **διαχωρισμός** (**cluster separation**) είναι το άθροισμα των βαρών (συνήθως απόσταση) μεταξύ κόμβων της συστάδας και των κόμβων εκτός συστάδας



$$cohesion(C_i) = \sum_{\substack{x \in C_i \\ y \in C_i}} proximity(x, y)$$

αντιστοιχεί στο D - διάμετρο στον BIRCH

$$separation(C_i, C_j) = \sum_{\substack{x \in C_i \\ y \in C_j}} proximity(x, y)$$

αντιστοιχεί στο D2 στον BIRCH

Χαρακτηρισμός Ποιότητας Συσταδοποίησης χωρίς Επίβλεψη:
Συνεκτικότητα και Διαχωρισμός

$$overall - cohesion = \sum_{i=1}^k w_i cohesion(C_i)$$

Άθροισμα συνεκτικότητας
κάθε συστάδας

$$overall - separation = \sum_{i=1}^k w_i separation(C_i)$$

Άθροισμα διαχωρισμού των
συστάδων

$$overall - validity = \sum_{i=1}^k \frac{separation(C_i)}{cohesion(C_i)}$$

Χαρακτηρισμός Ποιότητας Συσταδοποίησης χωρίς Επίβλεψη:
Συνεκτικότητα και Διαχωρισμός

Σχέση prototype και graph-based συνεκτικότητας και διαχωρισμού (για Ευκλείδειες αποστάσεις)

Χαρακτηρισμός Ποιότητας Συσταδοποίησης χωρίς Επίβλεψη:
Συνεκτικότητα και Διαχωρισμός

Σχέση prototype και graph-based συνεκτικότητας (για Ευκλείδειες αποστάσεις)

Έστω Ευκλείδεια απόσταση, **σχέση SSE με συνεκτικότητα** (πόσο στενά σχεπζόμενα είναι τα αντικείμενα μιας συστάδας);

$$cluster - SSE = \sum_{x \in C_i} dist^2(c_i, x)$$

$$Total - SSE = \sum_{i=1}^K \sum_{x \in C_i} dist^2(c_i, x)$$

Αποδεικνύεται ότι

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

Δηλαδή, είτε πάρουμε την απόσταση από το κέντρο είτε το μέσο όρο των ανά δύο αποστάσεων των σημείων είναι το ίδιο Σχέση διαμέτρου και ακτίνας

Χαρακτηρισμός Ποιότητας Συσταδοποίησης χωρίς Επίβλεψη:
Συνεκτικότητα και Διαχωρισμός

Σχέση δυο προσεγγίσεων διαχωρισμού (για Ευκλείδειες αποστάσεις)

Έστω Ευκλείδεια απόσταση, **σχέση SSB (group sum of squares) με διαχωρισμό** (πόσο μακριά είναι οι συστάδες);

$$cluster - SSB = dist(c_i, c)^2$$

$$(ολικό-)SSB = \sum_{i=1}^K m_i dist(c_i, c)^2$$

Αποδεικνύεται ότι

Ισομεγείς συστάδες

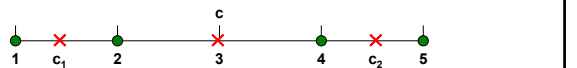
$$m_i = m/K$$

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

Δηλαδή, είτε πάρουμε την απόσταση των κέντρων κάθε συστάδας από το ολικό κέντρο είτε το μέσο όρο των ανά δύο αποστάσεων των κέντρων κάθε συστάδας είναι το ίδιο

Χαρακτηρισμός Ποιότητας Συσταδοποίησης χωρίς Επίβλεψη:
Συνεκτικότητα και Διαχωρισμός

Total-SSE + Total-SSB = constant



K=1 cluster: $total - SSE = (1-3)^2 + (2-3)^2 + (4-3)^2 + (5-3)^2 = 10$

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

$$Total = 10 + 0 = 10$$

K=2 clusters: $total - SSE = (1-1.5)^2 + (2-1.5)^2 + (4-4.5)^2 + (5-4.5)^2 = 1$

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

$$Total = 1 + 9 = 10$$

Χαρακτηρισμός Ποιότητας Συσταδοποίησης χωρίς Επίβλεψη:
Συνεκτικότητα και Διαχωρισμός

Αποδεικνύεται ότι

Total SSB + Total SSE = constant

$$TSS = \sum_{i=1}^K \sum_{x \in C_i} (x - c)^2$$

Ίσο με το τετράγωνο των αποστάσεων όλων των σημείων από το ολικό μέσο

Ελαχιστοποίηση της SSE (συνεκτικότητας) =>
Μεγιστοποίηση του SSB (διαχωρισμού)

Χαρακτηρισμός Ποιότητας Συσταδοποίησης χωρίς Επίβλεψη:
Συνεκτικότητα και Διαχωρισμός

Μπορούν να χρησιμοποιηθούν για τη βελτίωση της συσταδοποίησης

Πχ μια συστάδα με κακή συνεκτικότητα μπορεί να χρειαστεί να διασπαστεί

Δυο συστάδες όχι καλά διαχωρισμένες μπορεί να συγχωνευτούν

- Το πόσο καλή είναι μια συσταδοποίηση
- Το ποσό καλή είναι μια συστάδα
- Το ποσό καλό είναι ένα σημείο σε μια συστάδα

Χαρακτηρισμός Ποιότητας Συσταδοποίησης χωρίς Επίβλεψη:
Συντελεστής Σκιαγράφησης

Silhouette Coefficient (συντελεστής σκιαγράφησης)

Για κάθε σημείο, i

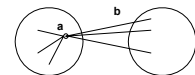
Υπολογισμός a = μέση απόσταση του i από τα σημεία της συστάδας

Υπολογισμός b = μέση απόσταση του i από όλα τα σημεία κάθε άλλης συστάδας - επιλογή του μικρότερου, δηλαδή μέση απόσταση από την κοντινότερη συστάδα

$$s = 1 - a/b \text{ if } a < b, \text{ (or } s = b/a - 1 \text{ if } a \geq b, \text{ not the usual case)}$$

Συνήθως μεταξύ του 0 και του 1

Όσο πιο κοντά στο 1, τόσο το καλύτερο



Μπορεί να χρησιμοποιηθεί και για μια συστάδα ή συσταδοποίηση θεωρώντας μέσες τιμές για όλα τα σημεία τους ή συστάδες

Χαρακτηρισμός Ποιότητας Συσταδοποίησης χωρίς Επίβλεψη:
Συντελεστής Σκιαγράφησης

Silhouette Coefficient (συντελεστής σκιαγράφησης)

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
 Silhouette Coefficient

Ο συντελεστής σκιαγράφησης για σημεία στις 10 συστάδες

Εξώφυλλο Διδακμίνων: Ακ. Έτος 2007-2008 ΣΥΣΤΑΔΟΠΟΙΗΣΗ ΙΙΙ 73

Χαρακτηρισμός Ποιότητας Συσταδοποίησης χωρίς Επίβλεψη:
Πίνακας Γειτνίασης

Δύο Πίνακες

Πίνακας Γειτνίασης (proximity matrix)
 ο πίνακας με την ομοιότητα των σημείων

Πίνακας Εμφάνισης ("incidence" matrix)
 Μια γραμμή και μια στήλη για κάθε σημείο
 Μια εγγραφή είναι 1 αν το αντίστοιχο ζευγάρι σημείων ανήκει στην ίδια συστάδα
 Μια εγγραφή είναι 0 αν το αντίστοιχο ζευγάρι σημείων ανήκει σε διαφορετική συστάδα

Υπολογισμός της **συσχέτισης (correlation)** των δύο πινάκων

Εξώφυλλο Διδακμίνων: Ακ. Έτος 2007-2008 ΣΥΣΤΑΔΟΠΟΙΗΣΗ ΙΙΙ 74

Χαρακτηρισμός Ποιότητας Συσταδοποίησης χωρίς Επίβλεψη:
Πίνακας Γειτνίασης

Υψηλή **συσχέτιση** σημαίνει ότι τα σημεία που ανήκουν στην ίδια συστάδα είναι κοντινά μεταξύ τους

- Δεν είναι καλή μέτρηση για κάποιες συστάδες που βασίζονται σε πυκνότητα και σε συνέχεια (contiguity)
- Επειδή, οι δύο πίνακες είναι συμμετρικοί, χρειάζεται ο υπολογισμός $n(n-1) / 2$ εγγραφών

Εξώφυλλο Διδακμίνων: Ακ. Έτος 2007-2008 ΣΥΣΤΑΔΟΠΟΙΗΣΗ ΙΙΙ 75

Χαρακτηρισμός Ποιότητας Συσταδοποίησης χωρίς Επίβλεψη:
Πίνακας Γειτνίασης

Υπολογισμός correlation των δύο πινάκων όταν χρησιμοποιείται ο K-means στα παρακάτω σύνολα

Corr = -0.9235 Corr = -0.5810

Εξώφυλλο Διδακμίνων: Ακ. Έτος 2007-2008 ΣΥΣΤΑΔΟΠΟΙΗΣΗ ΙΙΙ 76

Χαρακτηρισμός Ποιότητας Συσταδοποίησης χωρίς Επίβλεψη:
Πίνακας Γειτνίασης - Οπτικοποίηση

Αναδιατάσσουμε τα σημεία στον πίνακα έτσι ώστε τα σημεία που ανήκουν στην ίδια συστάδα να είναι γειτονικά

Συγκεκριμένα, τα διατάσσουμε με βάση τη συστάδα:
 Σημεία Συστάδας 1, Σημεία Συστάδας 2, Σημεία Συστάδας 3

Εξώφυλλο Διδακμίνων: Ακ. Έτος 2007-2008 ΣΥΣΤΑΔΟΠΟΙΗΣΗ ΙΙΙ 77

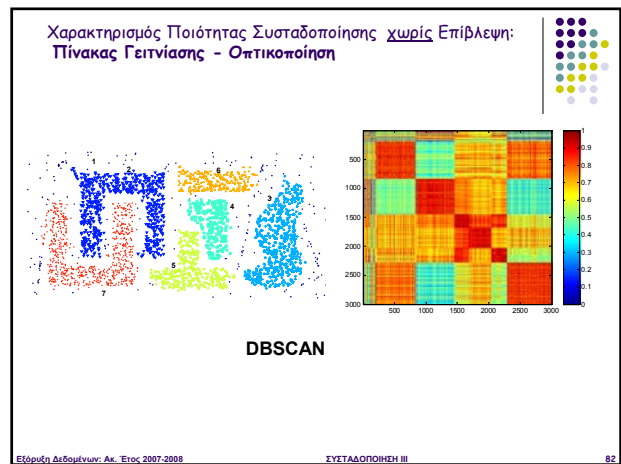
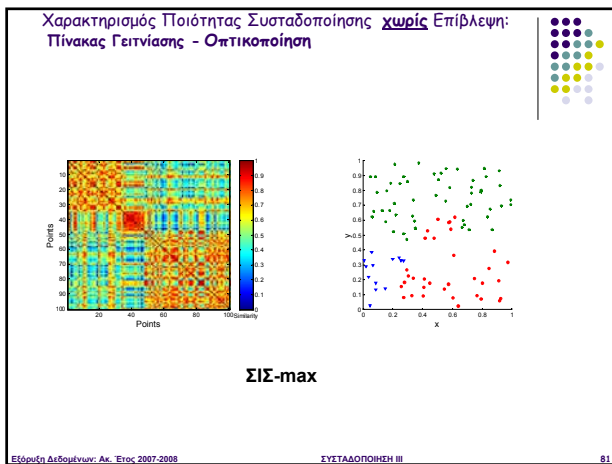
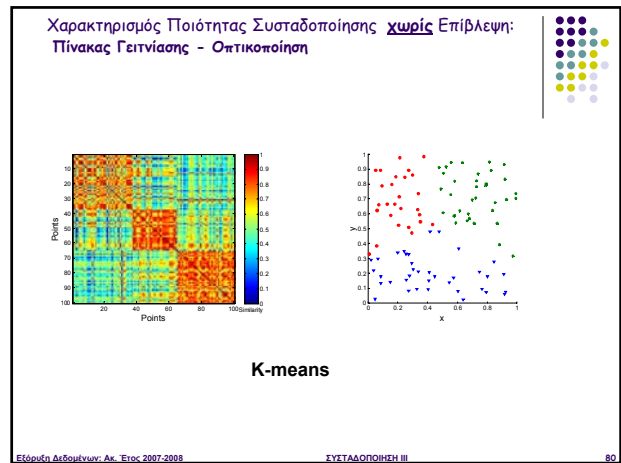
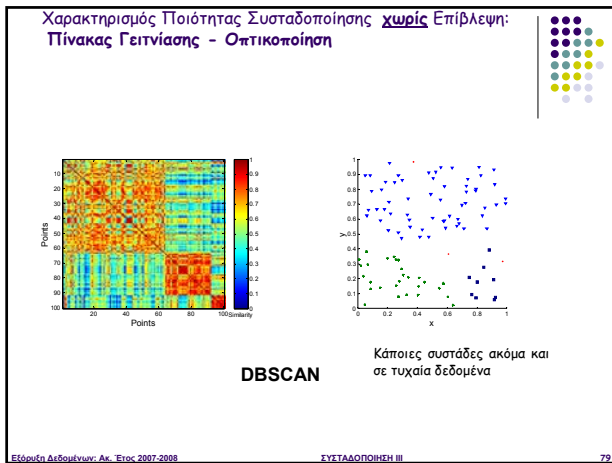
Χαρακτηρισμός Ποιότητας Συσταδοποίησης χωρίς Επίβλεψη:
Πίνακας Γειτνίασης - Οπτικοποίηση

Αναδιατάσσουμε τα σημεία στον πίνακα έτσι ώστε τα σημεία που ανήκουν στην ίδια συστάδα να είναι γειτονικά

Συγκεκριμένα, τα διατάσσουμε με βάση τη συστάδα:
 Σημεία Συστάδας 1, Σημεία Συστάδας 2, Σημεία Συστάδας 3

Σημείωση $s = 1 - (d - \min_d) / (\max_d - \min_d)$

Εξώφυλλο Διδακμίνων: Ακ. Έτος 2007-2008 ΣΥΣΤΑΔΟΠΟΙΗΣΗ ΙΙΙ 78



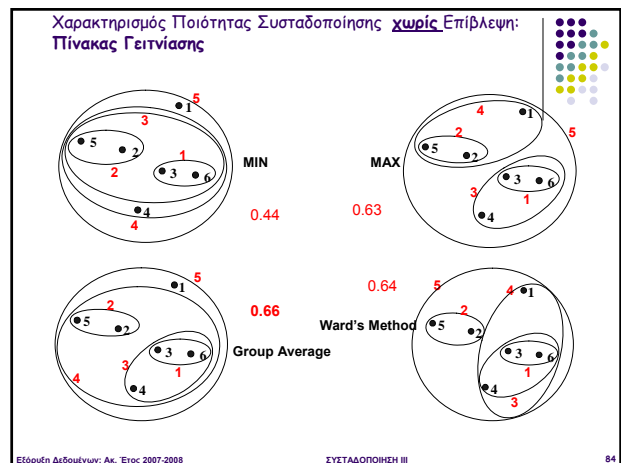
Χαρακτηρισμός Ποιότητας Συσταδοποίησης χωρίς Επίβλεψη:
Πίνακας Γειτνίασης

Ειδικά για ιεραρχικούς αλγόριθμους

Cophenetic distance: είναι η απόσταση (proximity) όταν ο αλγόριθμος τοποθετεί τα δυο σημεία στην ίδια συστάδα για πρώτη φορά
Πχ συγχωνεύω τα σημεία του C1 με τα σημεία του C2 σε απόσταση 0.1, όλα τα σημεία του C1 απέχουν από το C2 0.1

Cophenetic Correlation Coefficient (CPC)
Χρησιμοποιείται για επιλογή του είδους της ιεραρχικής συσταδοποίησης
Κατασκευάζω τον πίνακα των cophenetic αποστάσεων
Θεωρώ τη συνέλιξη του με τον αρχικό πίνακα αποστάσεων

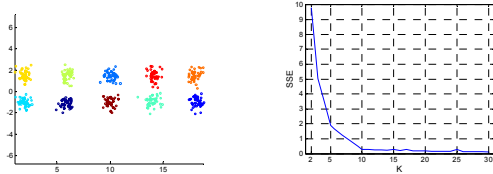
Εξόρυξη Δεδομένων: Ακ. Έτος 2007-2008 ΣΥΓΧΡΟΝΗ ΠΡΟΣΕΓΓΙΣΗ III 83



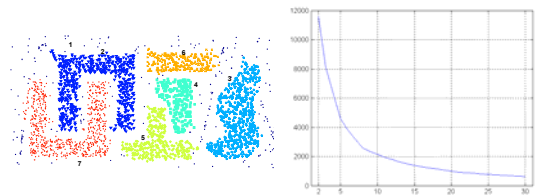
Χαρακτηρισμός Ποιότητας Συσταδοποίησης χωρίς Επίβλεψη: Συνεκτικότητα και Διαχωρισμός



Χρήση SSE για υπολογισμό του σωστού αριθμού συστάδων χρησιμοποιώντας τον K-means
(K = 5 και 10 φαίνονται καλές τιμές)



Χαρακτηρισμός Ποιότητας Συσταδοποίησης χωρίς Επίβλεψη: Συνεκτικότητα και Διαχωρισμός



Χαρακτηρισμός Ποιότητας Συσταδοποίησης με Επίβλεψη:



Μας δίνονται κάποιες ετικέτες κλάσεων και θέλουμε να δούμε πόσο καλά ταιριάζουν με τα δεδομένα

- **Classification-oriented** (μετρήσεις για ταξινόμηση): κατά πόσο μια συστάδα περιέχει αντικείμενα **μίας μόνο** κλάσης
- **Similarity-oriented**: κατά πόσο δύο αντικείμενα που ανήκουν στην ίδια κλάση, ανήκουν και στην ίδια συστάδα

Θα τα δούμε όταν μιλήσουμε για ταξινόμηση