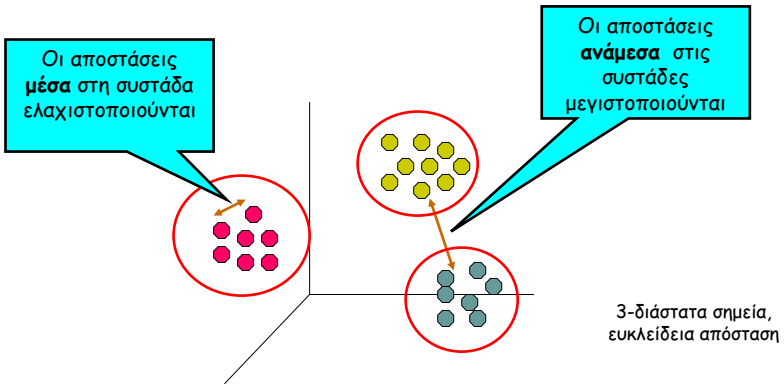
Συσταδοποίηση ΙΙ

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

Τι είναι συσταδοποίηση

Εύρεση συστάδων αντικειμένων έτσι ώστε τα αντικείμενα σε κάθε ομάδα να είναι όμοια (ή να σχετίζονται) και διαφορετικά (ή μη σχετιζόμενα) από τα αντικείμενα των άλλων ομάδων



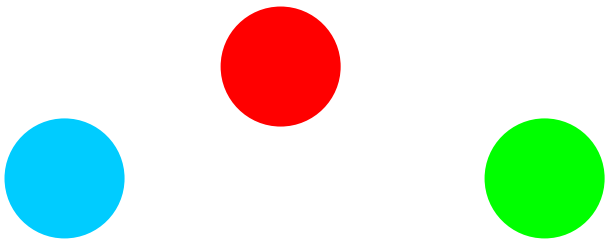
3-διάστατα σημεία, ευκλείδεια απόσταση

Είδη Συστάδων

- Καλώς διαχωρισμένες συστάδες
- Συστάδες βασισμένες σε κέντρο
- Συνεχής (contiguous) συστάδες
- Συστάδες Βασισμένες σε πυκνότητα
- Βασισμένα σε ιδιότητες ή έννοιες
- Περιγράφονται από μια αντικειμενική συνάρτηση (Objective Function)

Καλώς Διαχωρισμένες Συστάδες

Μια συστάδα είναι ένα σύνολο από σημεία τέτοια ώστε κάθε σημείο μιας ομάδας είναι κοντινότερο σε (ή πιο όμοιο με) όλα τα άλλα σημεία της ομάδας από ότι σε οποιοδήποτε άλλο σημείο που δεν ανήκει στη συστάδα.



3 καλώς-διαχωρισμένες συστάδες

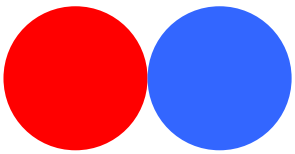
Συχνά υπάρχει η έννοια του κατωφλιού (threshold)

Όχι απαραίτητα κυκλικοί (οποιοδήποτε σχήμα)

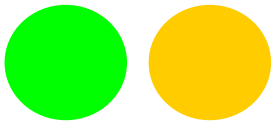
Συστάδες βασισμένες σε κέντρο ή πρότυπο

Μια συστάδα είναι ένα σύνολο από αντικείμενα τέτοιο ώστε ένα αντικείμενο στην ομάδα είναι κοντινότερο σε (ή πιο όμοιο με) το «κέντρο» ή πρότυπο της ομάδας από ότι από το κέντρο οποιασδήποτε άλλης ομάδας.
Το κέντρο της ομάδας είναι συχνά

- centroid, ο μέσος όρος των σημείων της συστάδας, ή
- a medoid, το πιο «αντιπροσωπευτικό» σημείο της συστάδας (πχ όταν κατηγορικά γνωρίσματα)



4 συστάδες βασισμένες σε κέντρο



Τείνουν στο να είναι κυκλικοί

Συνεχής Συστάδες

Συνεχής Συστάδες (Contiguous Cluster) (Κοντινότερος γείτονα ή μεταβατικά)

Μια συστάδα είναι ένα σύνολο σημείων τέτοιο ώστε κάθε σημείο είναι πιο κοντά σε ένα ή περισσότερα σημεία της συστάδας από ό,τι σε οποιοδήποτε σημείο εκτός συστάδας

Συχνά σε περιπτώσεις συστάδων με μη κανονικό σχήμα ή με αλληλοπλεκόμενα σχήματα – ή όταν έχουμε γραφήματα και θέλουμε να βρούμε συνεκτικά υπογραφήματα

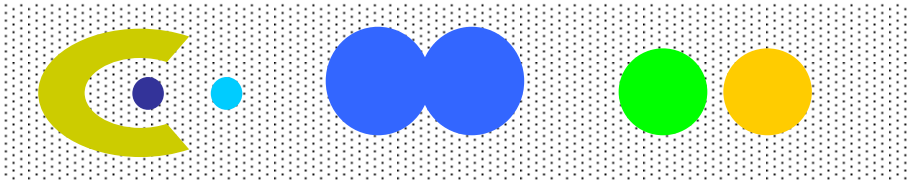
Πρόβλημα με θόρυβο

8 συνεχείς συστάδες

Συστάδες βασισμένες στην πυκνότητα

Μια συστάδα είναι μια πυκνή περιοχή από σημεία την οποία χωρίζουν από άλλες περιοχές μεγάλης πυκνότητας περιοχές χαμηλής πυκνότητας

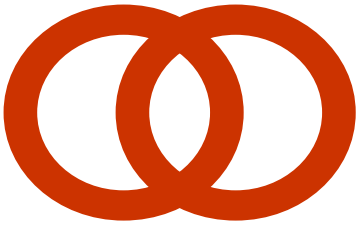
Συχνά σε περιπτώσεις συστάδων με μη κανονικό σχήμα ή με αλληλοπλεκόμενα σχήματα ή όταν θόρυβος ή outliers



6 συστάδες βασισμένες στην πυκνότητα

Εννοιολογική συσταδοποίηση

Συστάδες με κοινή ιδιότητα ή εννοιολογικές συστάδες.



2 αλληλοκαλυπτόμενοι κύκλοι

Συστάδες βασισμένες σε μια Αντικειμενική Συνάρτηση

Εύρεση συστάδων που ελαχιστοποιούν ή μεγιστοποιούν μια αντικειμενική συνάρτηση

Απαρίθμηση όλων των δυνατών τρόπων χωρισμού των σημείων σε συστάδες και υπολογισμού του «πόσο καλό» ("goodness") είναι κάθε πιθανό σύνολο από συστάδες χρησιμοποιώντας τη δοθείσα αντικειμενική συνάρτηση (NP-hard)

Οι στόχοι (objectives) μπορεί να είναι ολικοί (global) ή τοπικοί (local)
Οι ιεραρχικοί συνήθως τοπικού
Οι διαχωριστικοί ολικές

K-means: Βασικός Αλγόριθμος

Βασικός αλγόριθμος

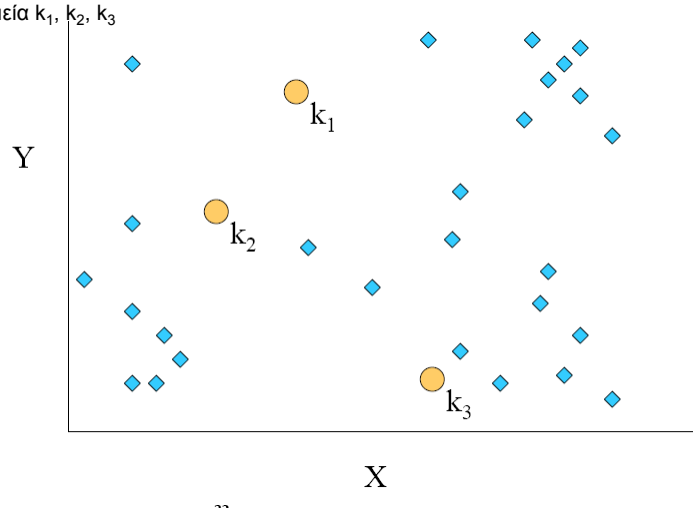
1: Επιλογή *K* σημείων ως τα αρχικά κεντρικά σημεία
2: **Repeat**
3: Ανάθεση όλων των αρχικών σημείων στο *κοντινότερο* τους από τα *K* κεντρικά σημεία
4: Επανα-υπολογισμός του *κεντρικού σημείου* κάθε συστάδας
5: **Until** τα κεντρικά σημεία να μην αλλάζουν

K-means: Βασικός Αλγόριθμος

Αρχική κατάσταση,

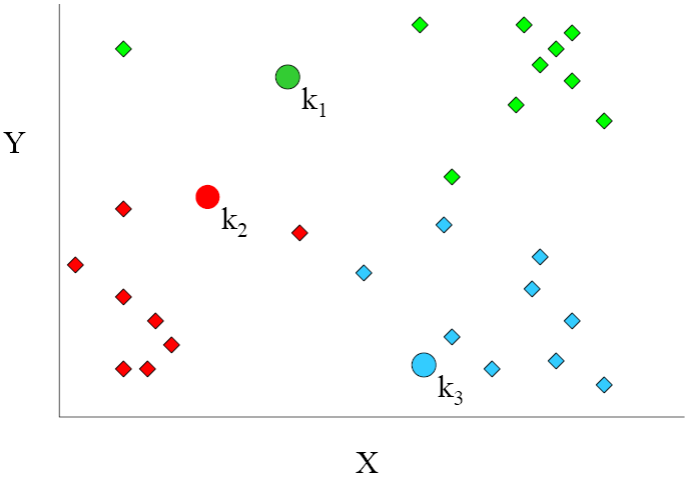
K = 3 συστάδες

Αρχικά σημεία k_1, k_2, k_3



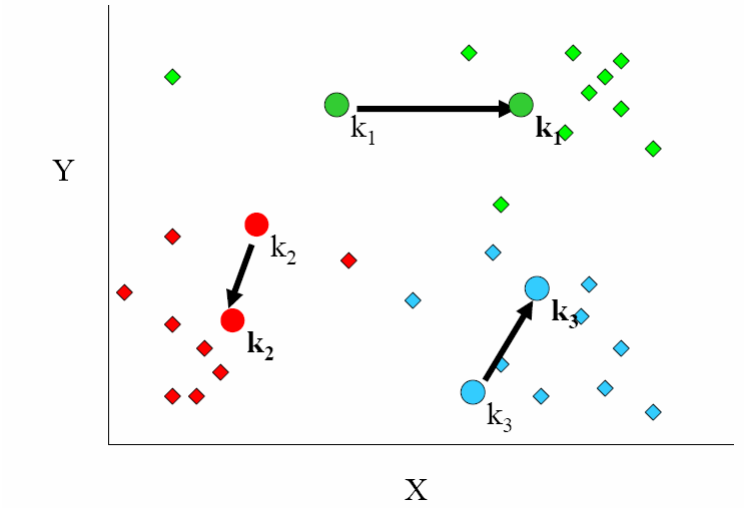
K-means: Βασικός Αλγόριθμος

Τα σημεία ανατίθενται στο πιο γειτονικό από τα 3 αρχικά σημεία



K-means: Βασικός Αλγόριθμος

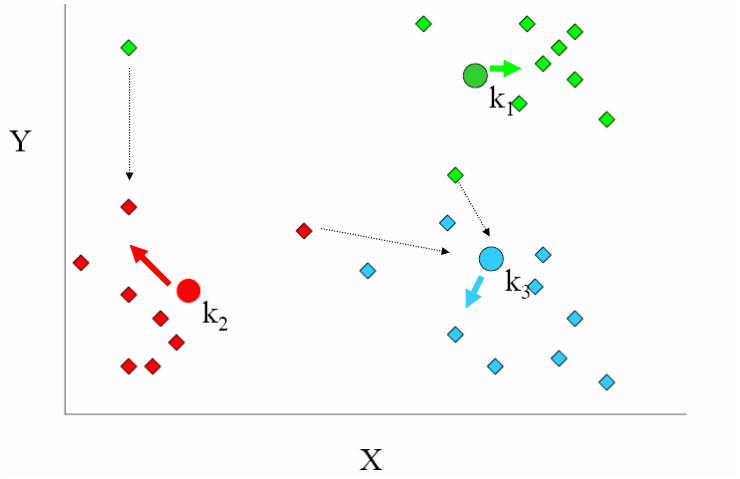
Επανα-υπολογισμός του κέντρου (κέντρου βάρους) κάθε σημείου



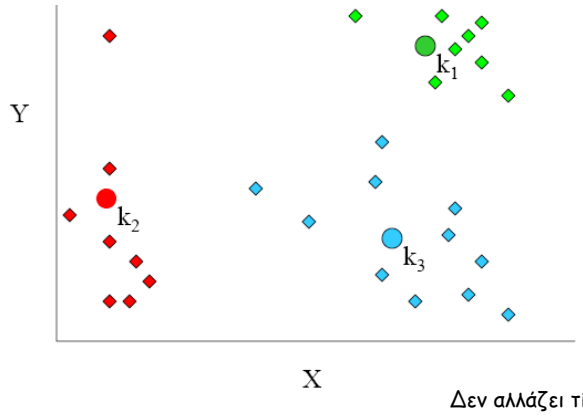
K-means: Βασικός Αλγόριθμος

Νέα ανάθεση των σημείων

Νέα κέντρα βάρους



K-means: Βασικός Αλγόριθμος



Δεν αλλάζει τίποτα -> ΤΕΛΟΣ

K-means: Επιλογή αρχικών σημείων

Αν υπάρχουν K «πραγματικές συστάδες» η πιθανότητα να επιλέξουμε ένα κέντρο από κάθε συστάδα είναι μικρή, συγκεκριμένα αν όλες οι συστάδες έχουν το ίδιο μέγεθος n, τότε:

$$P = \frac{\text{number of ways to select one centroid from each cluster}}{\text{number of ways to select } K \text{ centroids}} = \frac{K!n^K}{(Kn)^K} = \frac{K!}{K^K}$$

Για παράδειγμα, αν K = 10, η πιθανότητα είναι = 10!/1010 = 0.00036

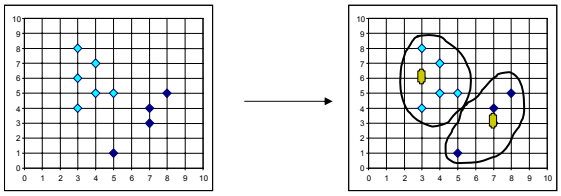
K-medoid

Συνήθως συνεχή d-διάστατο χώρο

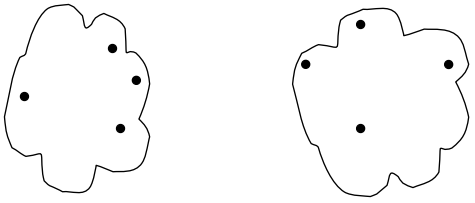
Διαλέγει ένα αντιπροσωπευτικό σημείο από τα δεδομένα και ελαχιστοποιεί την απόσταση από αυτό – Medoid: το πιο κεντρικό σημείο της συστάδας (αντί να χρησιμοποιεί το mean)

Μειώνει την ευαισθησία σε outliers

Μπορεί να εφαρμοστεί σε δεδομένα οποιουδήποτε τύπου (πχ και για κατηγορικά δεδομένα)



ΣΙΣ: Ορισμός απόστασης μεταξύ συστάδων



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

- MIN
- MAX
- Μέσος όρος της ομάδας
- Η απόσταση μεταξύ των κεντρικών σημείων
- Άλλες μέθοδοι βασισμένες σε μια αντικειμενική συνάρτηση
  - Η μέθοδος του Ward χρησιμοποιεί τετραγωνικά λάθη

· **Πίνακας Γειτνίασης**

·

DBSCAN

DBSCAN: Γενικά

Ο DBSCAN είναι ένας αλγόριθμος βασισμένος στην πυκνότητα
Πυκνότητα για ένα σημείο = αριθμός σημείων (MinPts) μέσα σε ποια προκαθορισμένη ακτίνα (Eps) από αυτό (συμπεριλαμβανομένου του σημείου)

Τα σημεία διαχωρίζονται σε:

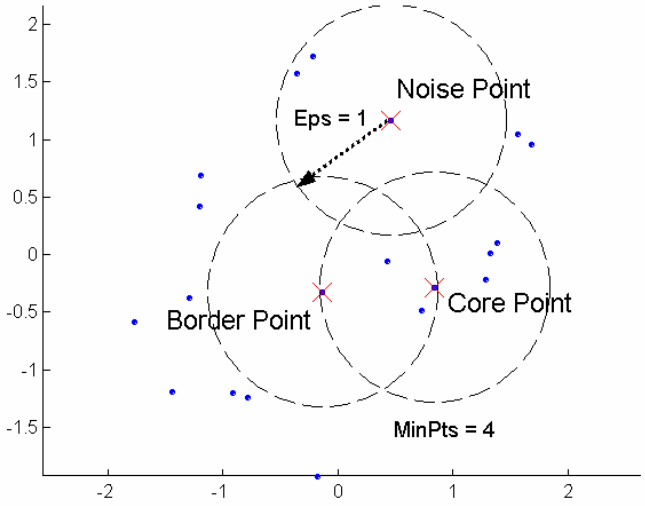
- Βασικά (core): ένα σημείο για το οποίο υπάρχουν περισσότερα από ένα προκαθορισμένο αριθμό (***MinPts***) σημεία σε ακτίνα ***Eps***

  Αυτά είναι τα σημεία που είναι στο εσωτερικό μιας συστάδας (ομάδας πυκνών σημείων)

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

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

DBSCAN: Γενικά



DBSCAN: Αλγόριθμος

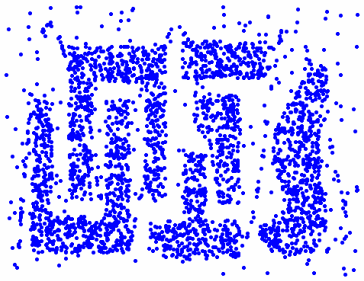
Βασικός Αλγόριθμος

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

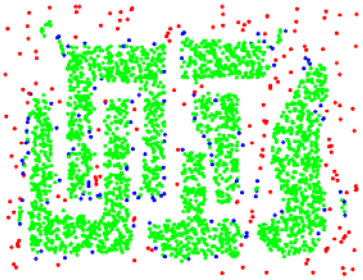
5: Ανάθεσε κάθε οριακό σημεία σε μία από τις συστάδες των συσχετιζόμενων του βασικών σημείων

DBSCAN: Αλγόριθμος



Αρχικά σημεία

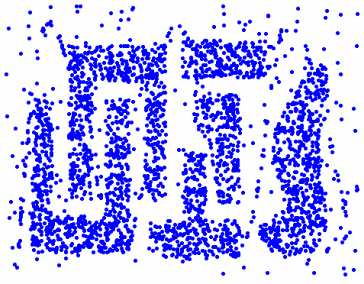
Βήμα 1&2



Τύποι σημείων: core, border και noise

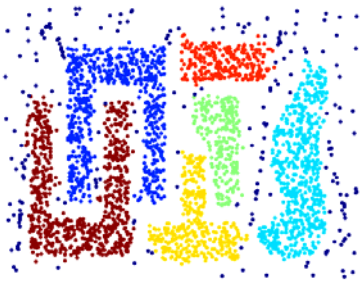
Eps = 10, MinPts = 4

DBSCAN: Πλεονεκτήματα



Αρχικά Σημεία

Βήμα 3&4



Συστάδες

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

DBSCAN: Πολυπλοκότητα

Για m σημεία εισόδου:

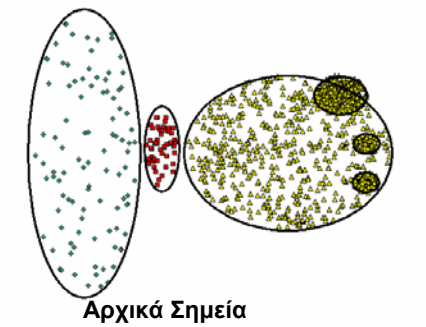
O(m x χρόνος εντοπισμού σημείων σε eps-γειτονιά)

O(m^2)

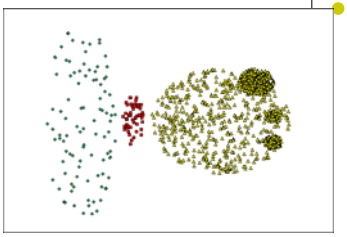
Για μικρό αριθμό διαστάσεων, υπάρχουν δομές που υποστηρίζουν την πράξη σε *O*(m logm)

O(m) χώρος (για κάθε σημείο κρατάμε μόνο ένα label σε μια συστάδα ανήκει και το είδος του (βασικό, οριακό, θόρυβος)

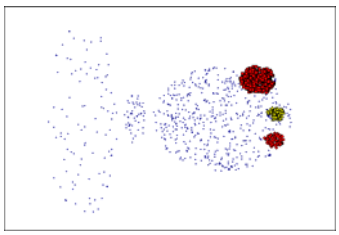
DBSCAN: Περιορισμοί



Αρχικά Σημεία



(MinPts=4, Eps=9.75).



(MinPts=4, Eps=9.92)

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

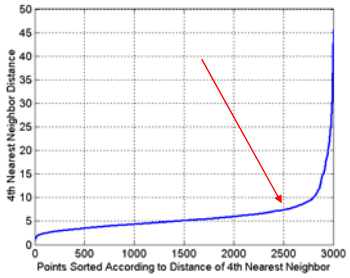
DBSCAN: Καθορισμός των MinPts και Eps

Η ιδέα είναι να κοιτάξουμε την απόσταση ενός σημείου από τον k-οστό κοντινότερο γείτονα του -> k-dist
Γενικά (κατά μέσο όρο), 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 = MinPts και Eps = k-dist

Eps ~ 7

MinPts = 4



Διαχείριση Ποιότητας
Cluster validity

Ποιότητα Συσταδοποίησης

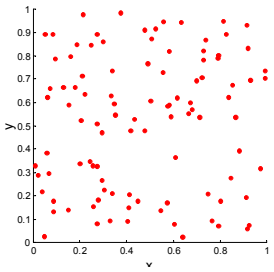
Πόσο καλή είναι η συσταδοποίηση που επιτύχαμε;

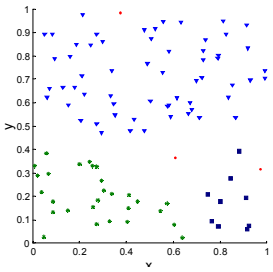
Οι αλγόριθμοι που είδαμε παράγουν κάποιες συστάδες ακόμα και όταν τα δεδομένα παράγονται τυχαία

Δύσκολη η αξιολόγηση, ιδιαίτερα σε πολλές διαστάσεις

Συστάδες σε Τυχαία Δεδομένα

Τυχαία Σημεία

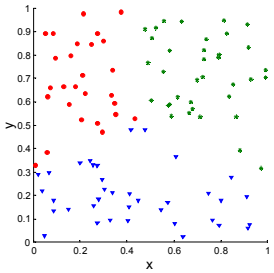


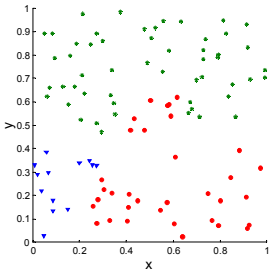


DBSCAN

3 ομάδες κοιτώντας την απόσταση του 4ου γείτονα

K-means





ΣΙΣ με MAX-link

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

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 ταιριάζουν τα αποτελέσματα της ανάλυσης 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

Χαρακτηρισμός Ποιότητας Συσταδοποίησης χωρίς Επίβλεψη

- Χρήση Συνεκτικότητας και Διαχωρισμού
- Χρήση Πίνακα Γειτνίασης

Χαρακτηρισμός Ποιότητας Συσταδοποίησης **χωρίς** Επίβλεψη: **Συνεκτικότητα και Διαχωρισμός**

Δύο μέτρα:

Ένα για να χαρακτηρίσουμε κάθε συστάδα ξεχωριστά (cohesion – συνεκτικότητα: πόσο κοντά (όμοια) είναι τα σημεία κάθε συστάδας)

Ένα για τις συστάδες μεταξύ τους (separation – διαχωρισμός: πόσο μακριά (ανόμοιες) είναι δύο συστάδες

Ορίζονται είτε

Prototype-based:

με βάση το «κεντρικό σημείο» κάθε συστάδας είτε

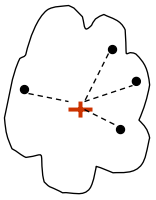
Graph-based:

με βάση τις ανά-δύο αποστάσεις των σημείων

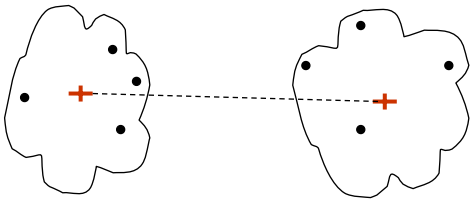
Χαρακτηρισμός Ποιότητας Συσταδοποίησης **χωρίς** Επίβλεψη: **Συνεκτικότητα και Διαχωρισμός**

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

Συνεκτικότητα (cohesion)



Διαχωρισμός (separation)



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

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

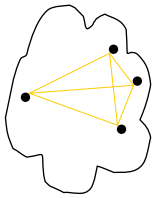
αντιστοιχεί στα D0 (D1) στον BIRCH

Χαρακτηρισμός Ποιότητας Συσταδοποίησης **χωρίς** Επίβλεψη: **Συνεκτικότητα και Διαχωρισμός**

Συσταδοποίηση βασισμένη σε γραφήματα (ΣΙΣ)

- 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

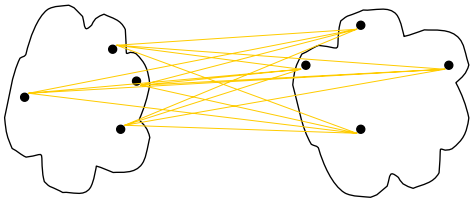
Συνεκτικότητα (cohesion)



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

αντιστοιχεί στο D – διάμετρο στον BIRCH

Διαχωρισμός (separation)



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

αντιστοιχεί στο D2 στον BIRCH

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**

Συνολική Συνεκτικότητα

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

Άθροισμα συνεκτικότητας κάθε συστάδας

Συνολικός Διαχωρισμός

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

Άθροισμα διαχωρισμού των συστάδων

Συνολικός Χαρακτηρισμός Ποιότητας για τη συσταδοποίηση

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

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 σχετιζόμενα 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

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

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);

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

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

Το ολικό κέντρο (σημείο c στους τύπους) είναι το σημείο που προκύπτει αν πάρουμε το μέσο (mean) των κέντρων όλων των συστάδων

Αποδεικνύεται ότι

Ισομεγέθεις συστάδες

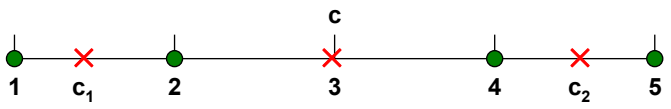
$m_i = m / K$

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

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 = σταθερά



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 = σταθερά

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

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**

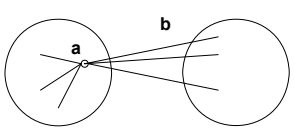
Silhouette Coefficient (συντελεστής σκιαγράφησης)

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

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

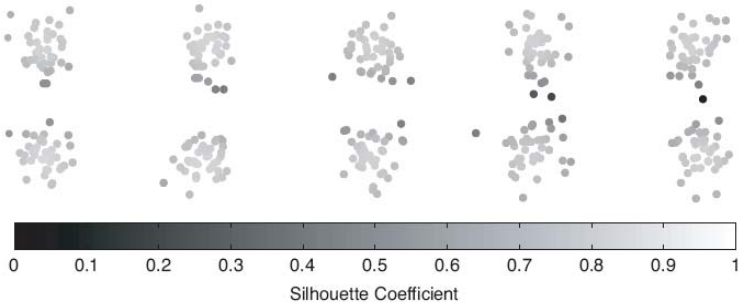
Συνήθως μεταξύ του 0 και του 1
Όσο πιο κοντά στο 1, τόσο το καλύτερο



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)



Ο συντελεστής σκιαγράφησης για σημεία στις 10 συστάδες

Πόσο «κεντρικό» είναι ένα σημείο για μία συστάδα

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

Χαρακτηρισμός Ποιότητας Συσταδοποίησης **χωρίς** Επίβλεψη: **Πίνακας Γειτνίασης**

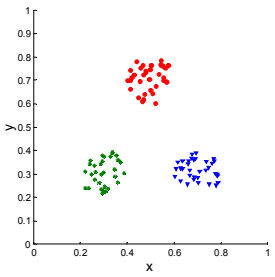
Υψηλή συσχέτιση σημαίνει ότι τα σημεία που ανήκουν στην ίδια συστάδα είναι κοντινά μεταξύ τους

- Δεν είναι καλή μέτρηση για κάποιες συστάδες που βασίζονται σε πυκνότητα και σε συνέχεια (contiguity)

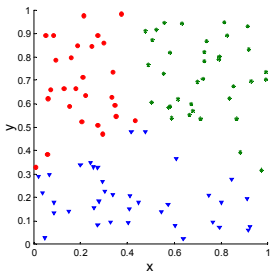
- Επειδή, οι δυο πίνακες είναι συμμετρικοί, χρειάζεται ο υπολογισμός n(n-1) / 2 εγγραφών

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

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 διατάσουμε με βάση τη συστάδα:

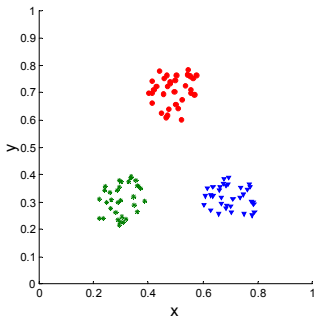
Σημεία Συστάδας 1, Σημεία Συστάδας 2, Σημεία Συστάδας 3

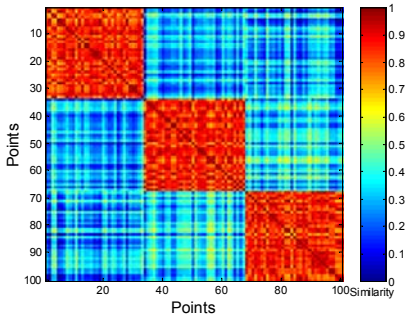
Χαρακτηρισμός Ποιότητας Συσταδοποίησης **χωρίς** Επίβλεψη: **Πίνακας Γειτνίασης - Οπτικοποίηση**

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

Συγκεκριμένα, τα διατάσουμε με βάση τη συστάδα:

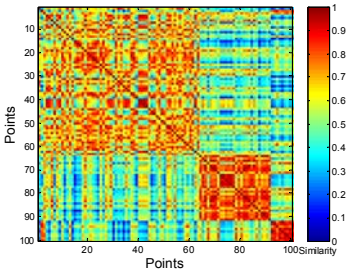
Σημεία Συστάδας 1, Σημεία Συστάδας 2, Σημεία Συστάδας 3

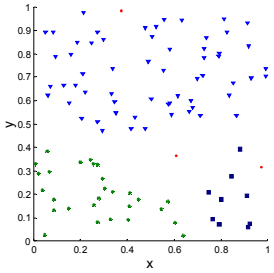




Σημείωση s = 1 – (d –min_d)/(max_d – min_d)

Χαρακτηρισμός Ποιότητας Συσταδοποίησης **χωρίς** Επίβλεψη: **Πίνακας Γειτνίασης - Οπτικοποίηση**

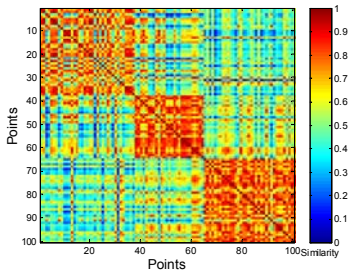


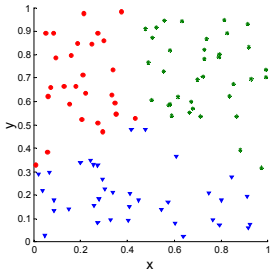


DBSCAN

Κάποιες συστάδες ακόμα και σε τυχαία δεδομένα

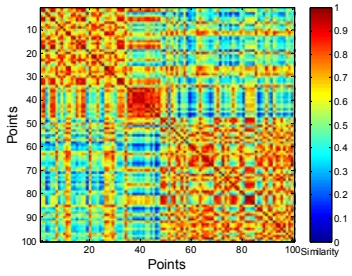
Χαρακτηρισμός Ποιότητας Συσταδοποίησης **χωρίς** Επίβλεψη: **Πίνακας Γειτνίασης - Οπτικοποίηση**

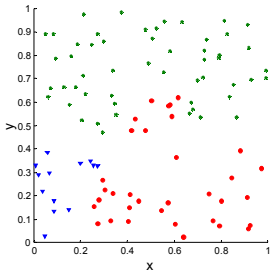




K-means

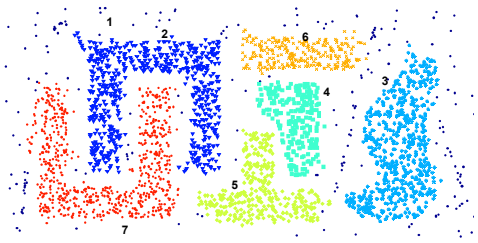
Χαρακτηρισμός Ποιότητας Συσταδοποίησης **χωρίς** Επίβλεψη:
Πίνακας Γειτνίασης - Οπτικοποίηση

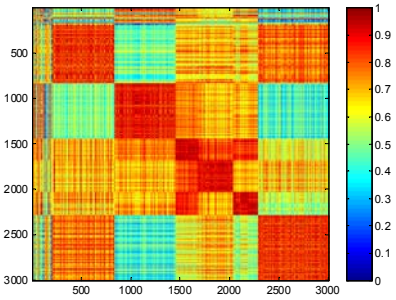




ΣΙΣ-max

Χαρακτηρισμός Ποιότητας Συσταδοποίησης χωρίς Επίβλεψη:
Πίνακας Γειτνίασης - Οπτικοποίηση





DBSCAN

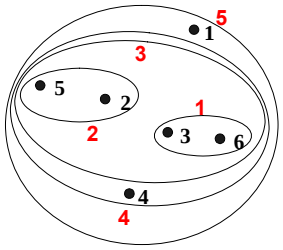
Χαρακτηρισμός Ποιότητας Συσταδοποίησης **χωρίς** Επίβλεψη: **Πίνακας Γειτνίασης**

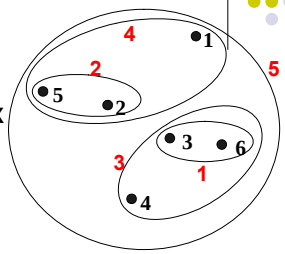
Ειδικά για ιεραρχικούς αλγόριθμους

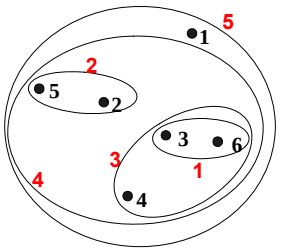
Cophenetic distance: είναι η απόσταση (proximity) όταν ο αλγόριθμος τοποθετεί τα δυο σημεία στην ίδια συστάδα για πρώτη φορά
Πχ συγχωνεύω τα σημεία του C1 με τα σημεία του C2 σε απόσταση 0.1, όλα τα σημεία του C1 απέχουν από το C2 0.1

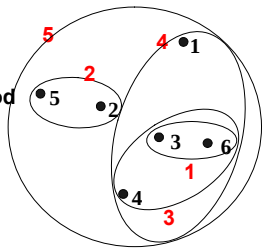
CoPhenetic Correlation Coefficient (CPCC)
Χρησιμοποείται για επιλογή του είδους της ιεραρχικής συσταδοποίησης
Κατασκευάζω τον πίνακα των cophenetic αποστάσεων
Θεωρώ τη συνέλιξη του με τον αρχικό πίνακα αποστάσεων

Χαρακτηρισμός Ποιότητας Συσταδοποίησης **χωρίς** Επίβλεψη: **Πίνακας Γειτνίασης**



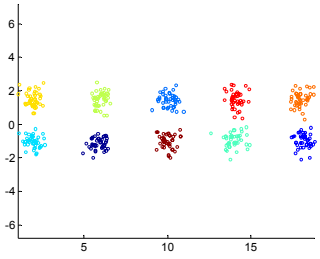


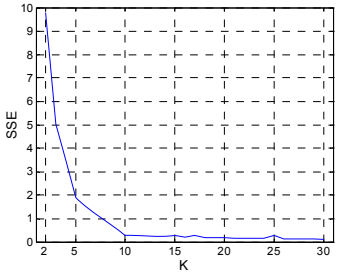




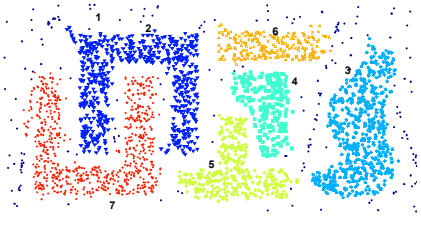
Χαρακτηρισμός Ποιότητας Συσταδοποίησης **χωρίς** Επίβλεψη: **Συνεκτικότητα και Διαχωρισμός**

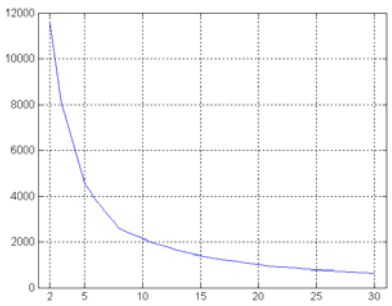
Χρήση SSE για υπολογισμό του σωστού αριθμού συστάδων χρησιμοποιώντας τον K-means
(K = 5 και 10 φαίνονται καλές τιμές)





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

BIRCH

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

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 περάσματα για βελτίωση της ποιότητας της συσταδοποίησης

BIRCH: Balanced Iterative Reducing and Clustering using Hierarchies

BIRCH

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

BIRCH

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

Centroid(κεντρικό σημείο): $\vec{X0} = \frac{\sum_{i=1}^{N} \vec{X}_i}{N}$

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

$$R = \left(\frac{\sum_{i=1}^{N} (\vec{X}_i - \vec{X0})^2}{N}\right)^{\frac{1}{2}}$$

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

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

BIRCH

Μας ενδιαφέρει και η απόσταση των κεντρικών σημείων δυο συστάδων

Μεταξύ δυο συστάδων

centroid Euclidean distance $D0 = ((\vec{X0}_1 - \vec{X0}_2)^2)^{\frac{1}{2}}$

centroid Manhattan distance

$$D1 = |\vec{X0}_1 - \vec{X0}_2| = \sum_{i=1}^{d} |\vec{X0}_1^{(i)} - \vec{X0}_1^{(i)}|$$

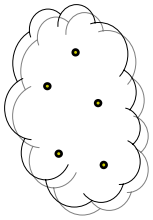
BIRCH

Συγχώνευση Συστάδων

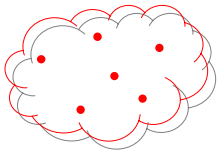
Συστάδα $\{X_i\}$:
$i = 1, 2, \ldots, N_1$

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

X_i

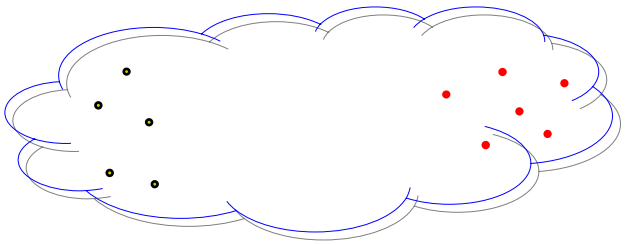


X_j



BIRCH

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



BIRCH

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

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

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 $\{\vec{X}_1, \vec{X}_2, \ldots, \vec{X}_N\}$

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

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:

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

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

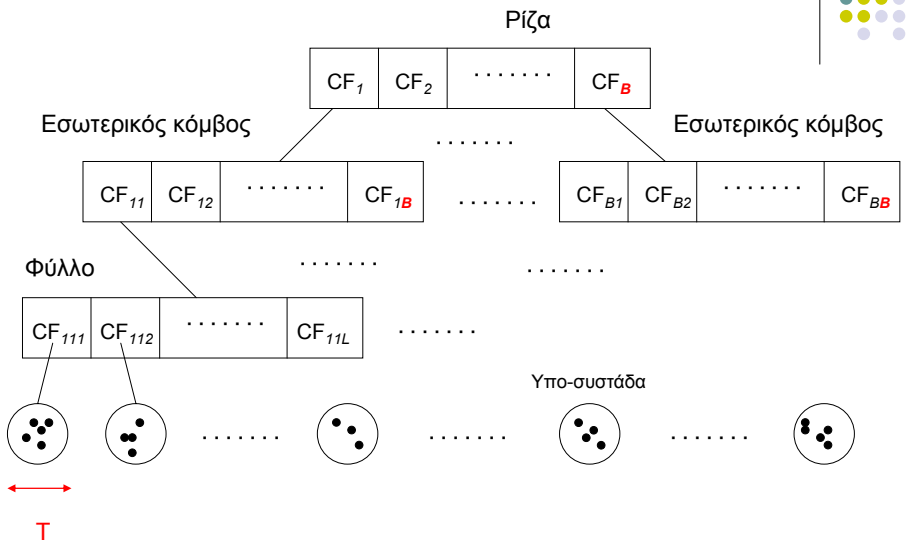
Μια εγγραφή CF έχει αρκετή πληροφορία για να υπολογίσουμε τα D0-D4

BIRCH

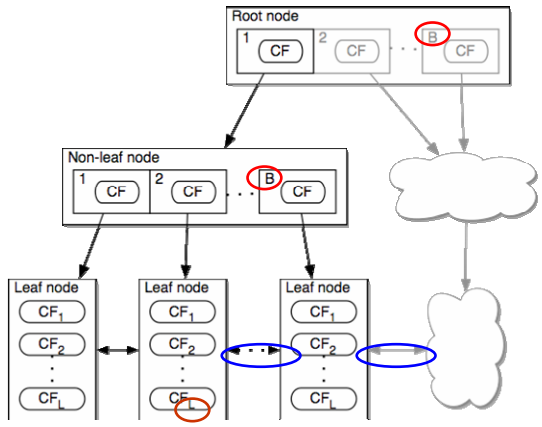
Ιεραρχικός αλγόριθμος

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

BIRCH: CF-δέντρο



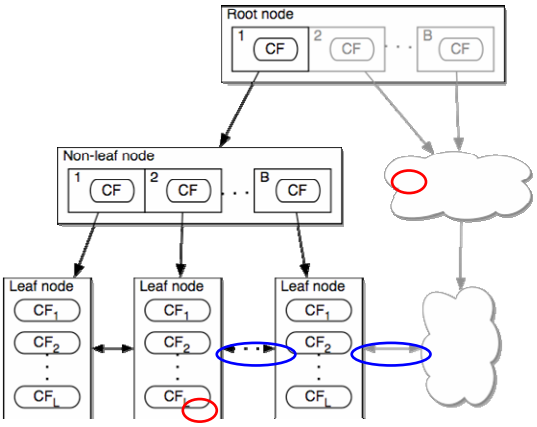
BIRCH: CF-δέντρο



- Κάθε εσωτερικός κόμβος περιέχει έναν αριθμό από παιδιά - B (παράγοντας διακλάδωσης) εγγραφές <CF_i, παιδί$_i$>
- Κάθε φύλλο περιέχει έναν αριθμό από υπο-συστάδες το πολύ L CF εγγραφές [CF_i] και <prev>, <next> pointers

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

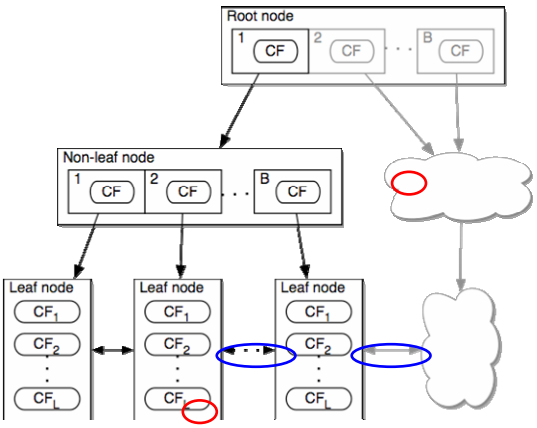
BIRCH: CF-δέντρο



- 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:

$$LS = \sum_{P_i \in N} \vec{P}_i$$

$$SS = \sum_{P_i \in N} \left|\vec{P}_i\right|^2$$

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

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

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

BIRCH: CF-δέντρο εισαγωγή στοιχείου

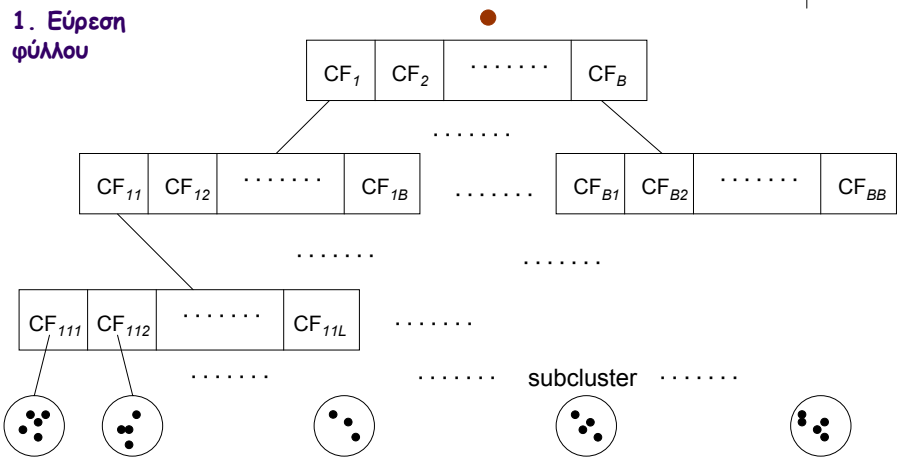
- Ο αλγόριθμος διαβάζει (scan) τα δεδομένα και τα εισάγει στο CF δέντρο ένα-ένα
- Η εισαγωγή ενός στοιχείου στο CF-δέντρο γίνεται με top-down διάσχιση ξεκινώντας από τη ρίζα με βάση μια συνάρτηση απόστασης Distance(σημείο, cluster)
  - Χρήση της D0, D1, D2, D3 ή D4
- Κάθε σημείο εισάγεται στην κοντινότερη 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 <= T) ok,
   Αλλιώς 3
2. Ενημέρωση του φύλλου
3. Διάσπαση φύλλου
4. Ενημέρωση τιμής CF

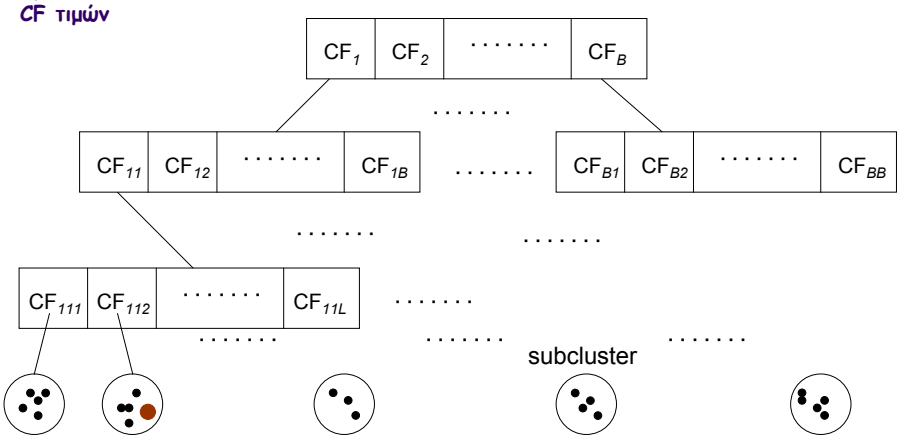
BIRCH: CF-δέντρο εισαγωγή στοιχείου

1. Εύρεση φύλλου

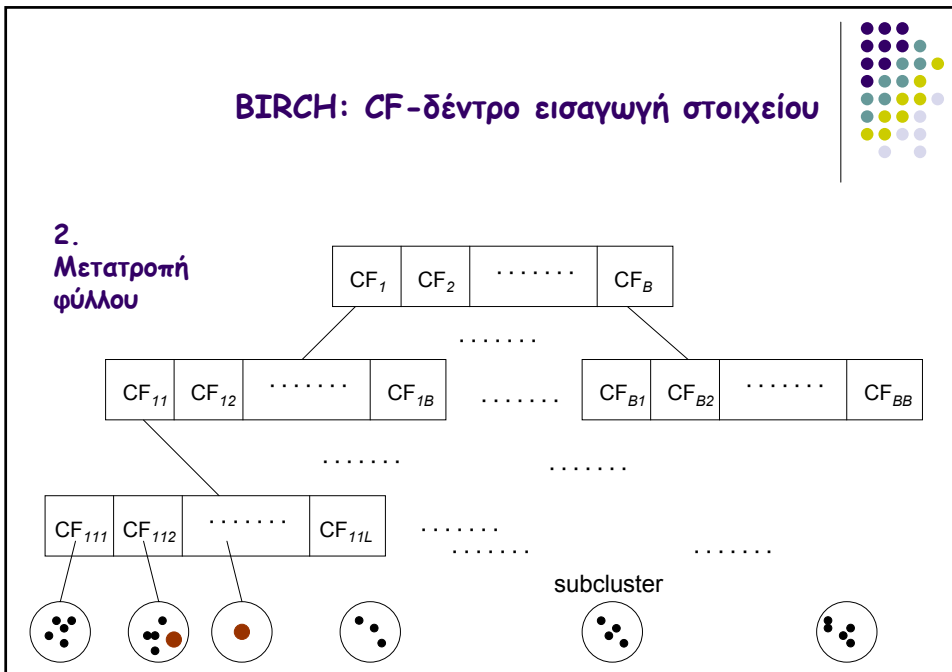


BIRCH: CF-δέντρο εισαγωγή στοιχείου

4. Τροποποίηση CF τιμών



BIRCH: CF-δέντρο εισαγωγή στοιχείου



2. Μετατροπή φύλλου

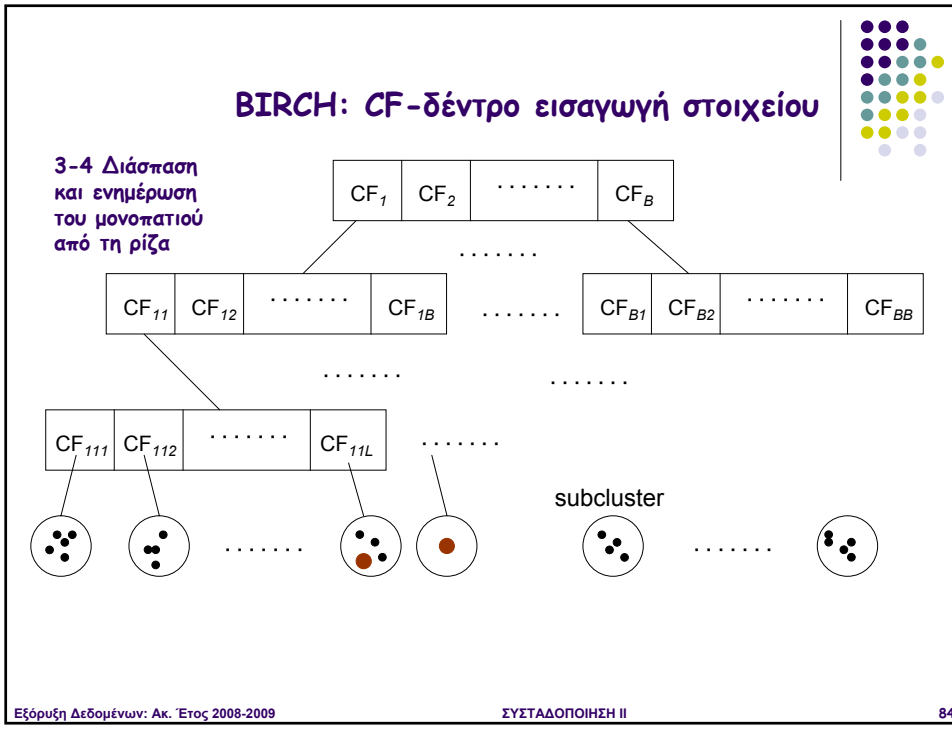
CF_1 CF_2 ..... CF_B

CF_{11} CF_{12} ..... CF_{1B} ..... CF_{B1} CF_{B2} ..... CF_{BB}

CF_{111} CF_{112} ..... CF_{11L}

subcluster

BIRCH: CF-δέντρο εισαγωγή στοιχείου



3-4 Διάσπαση και ενημέρωση του μονοπατιού από τη ρίζα

CF_1 CF_2 ..... CF_B

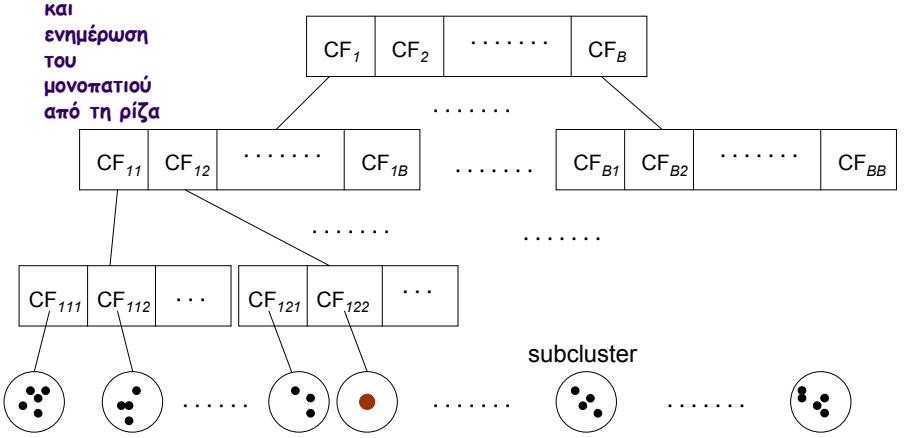
CF_{11} CF_{12} ..... CF_{1B} ..... CF_{B1} CF_{B2} ..... CF_{BB}

CF_{111} CF_{112} ..... CF_{11L}

subcluster

BIRCH: CF-δέντρο εισαγωγή στοιχείου

3-4 Διάσπαση και ενημέρωση του μονοπατιού από τη ρίζα



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 Τ**, τότε έχουμε δημιουργία νέας υποσυστάδας
    - Αν η νέα συστάδα χωρά στο φύλλο, ok -> ενημέρωση προγόνων
  - Αν η νέα συστάδα δε χωρά -> υπερχείλιση στο φύλλο

BIRCH: CF-δέντρο

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

Εύρεση των δύο υπο-συστάδων του φύλλου πού έχουν τη μεγαλύτερη απόσταση μεταξύ τους, έστω Ci και Cj

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 Ci ή στο φύλλο της Cj με βάση ποια από τις δύο είναι πιο όμοια της

BIRCH: CF-δέντρο

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

Διάσπαση εσωτερικού κόμβου

- Οι εσωτερικοί κόμβοι διασπούνται αναδρομικά με βάση μια μέτρηση της απόσταση 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!

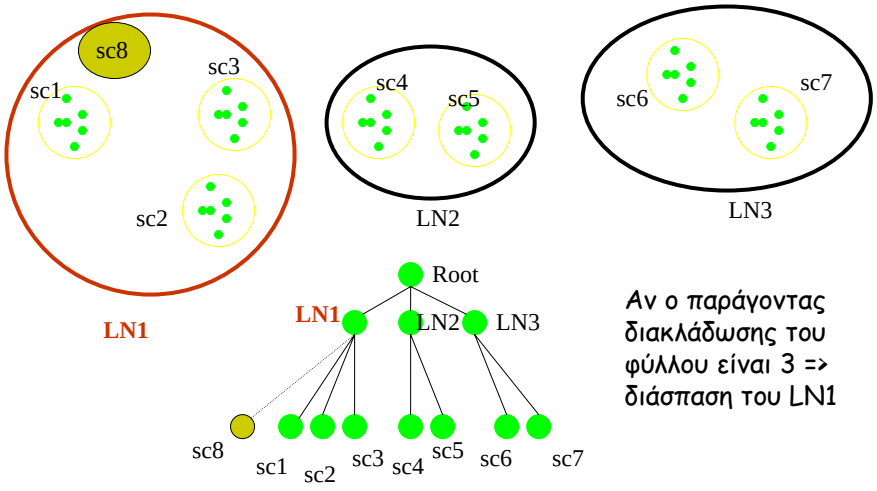
Μια μικρή διόρθωση:

- Όταν η διάσπαση κάποιων κόμβων τελειώνει (χωρούν σε ένα κόμβο) έστω στον κόμβο N_j 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

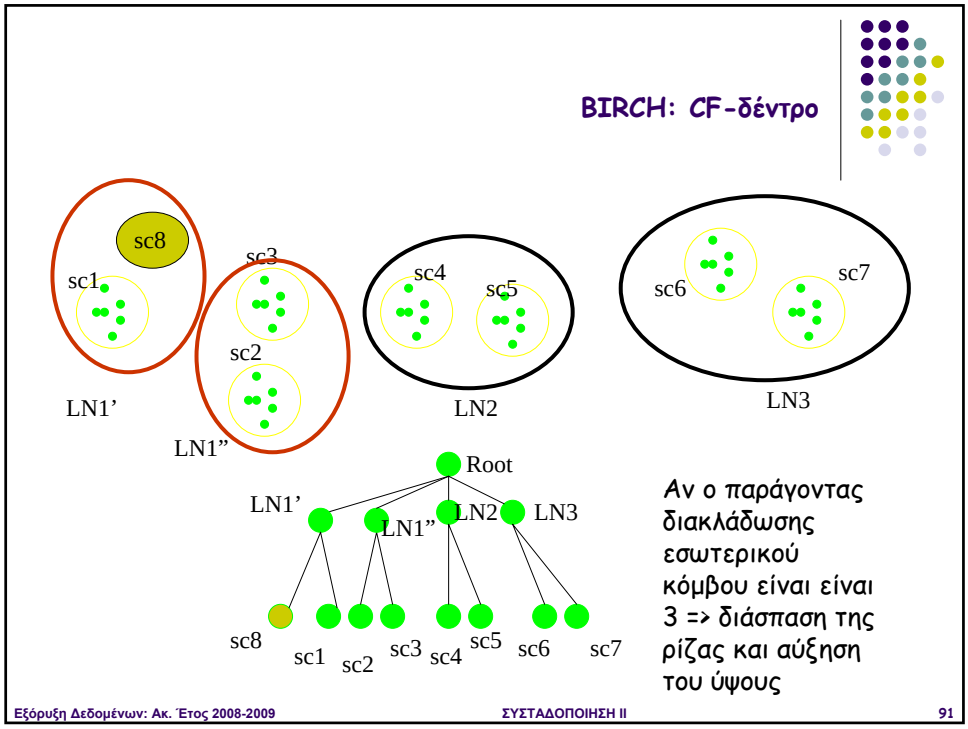
Τελικά ή συγχώνευση και ελευθέρωση χώρου ή καλύτερη ανακατανομή των εγγραφών σε κάποιο από τα παιδιά

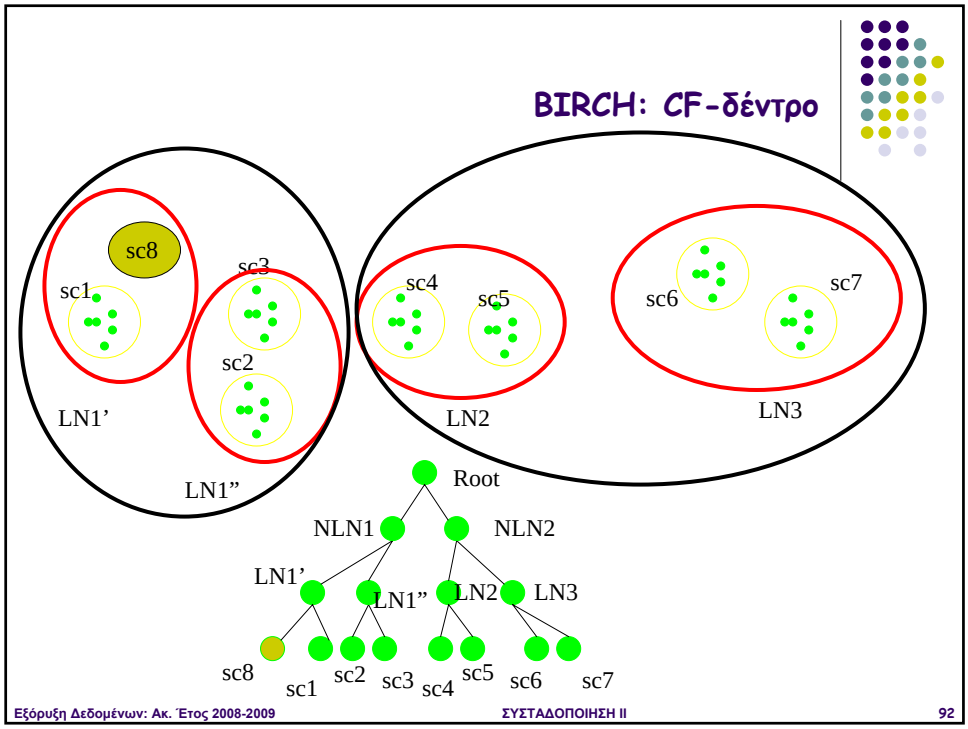
BIRCH: CF-δέντρο

Νέα υπο-συστάδα



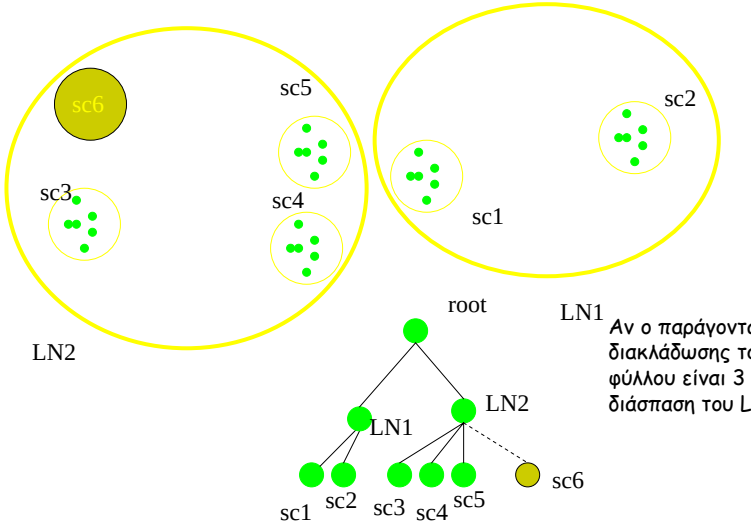
Αν ο παράγοντας διακλάδωσης του φύλλου είναι 3 => διάσπαση του LN1


BIRCH: CF-δέντρο
sc8
sc1
sc3
sc2
LN1'
LN1"
sc4
sc5
LN2
sc6
sc7
LN3
Root
LN1'
LN1"
LN2
LN3
sc8
sc1
sc2
sc3
sc4
sc5
sc6
sc7
Αν ο παράγοντας διακλάδωσης εσωτερικού κόμβου είναι είναι 3 => διάσπαση της ρίζας και αύξηση του ύψους
Εξόρυξη Δεδομένων: Ακ. Έτος 2008-2009
ΣΥΣΤΑΔΟΠΟΙΗΣΗ ΙΙ
91


BIRCH: CF-δέντρο
sc8
sc1
sc3
sc2
LN1'
LN1"
sc4
sc5
LN2
sc6
sc7
LN3
Root
NLN1
NLN2
LN1'
LN1"
LN2
LN3
sc8
sc1
sc2
sc3
sc4
sc5
sc6
sc7
Εξόρυξη Δεδομένων: Ακ. Έτος 2008-2009
ΣΥΣΤΑΔΟΠΟΙΗΣΗ ΙΙ
92

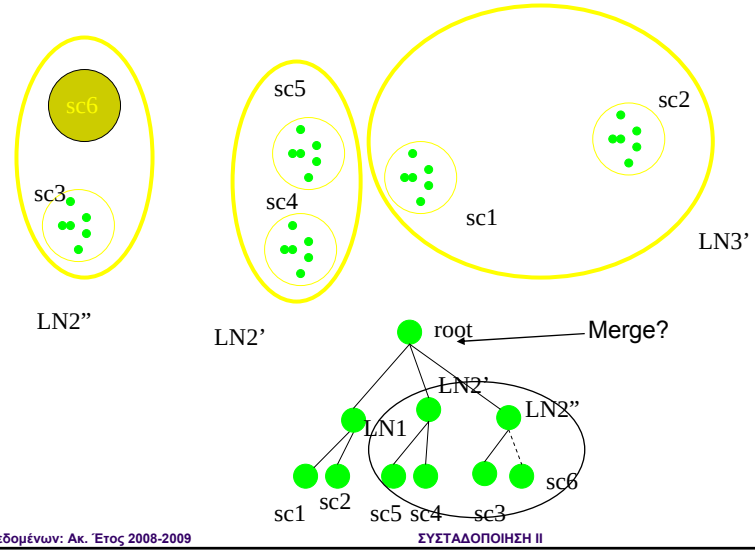
BIRCH: CF-δέντρο

Έστω ότι η αρίθμηση των υποσυστάδων αντιστοιχεί στη σειρά δημιουργίας τους

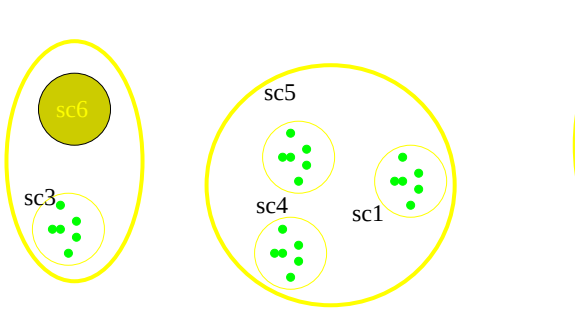


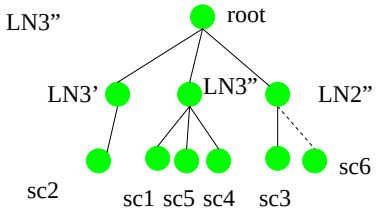
Αν ο παράγοντας διακλάδωσης του φύλλου είναι 3 => διάσπαση του LN2

BIRCH: CF-δέντρο



BIRCH: CF-δέντρο





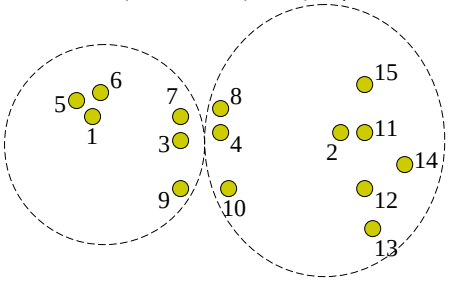
Συγχώνευση LN2’ και LN1 και ο καινούργιος κόμβος θα διασπαστεί πάλι

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-εισάγουμε ένα σημείο μπορεί να εισαχθεί σε διαφορετική συστάδα



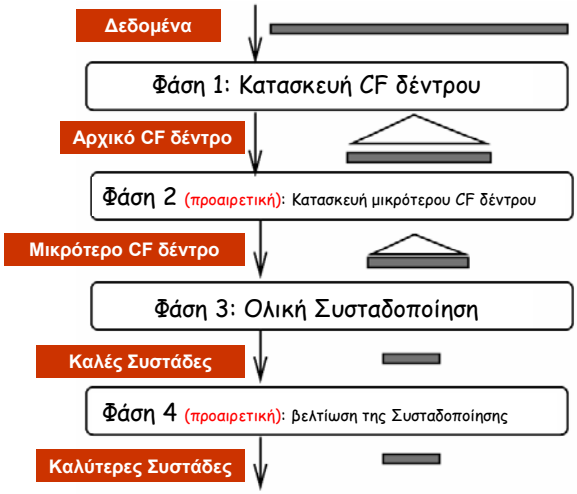
Αριθμός αντιστοιχεί στη σειρά εισαγωγής,

Έστω dist(1, 2) > T

BIRCH: αλγόριθμος

Αυτό αντιμετωπίζεται με προαιρετικές επιπρόσθετες φάσεις

BIRCH-αλγόριθμος



Φάση 1: Μια δομή κύριας μνήμης που συνοψίζει τα δεδομένα

Φάση 2: Κοιτά τα φύλλα και προσπαθεί να διώξει τους outliers και να ενοποιήσει «όμοιες» συστάδες που αντιστοιχούν σε περιοχές με πολλά σημεία

Χρειάζεται για να βελτιώσει τη Φάση 3

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

BIRCH

Φάση 4 (προαιρετική)

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

BIRCH-αλγόριθμος

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

BIRCH-αλγόριθμος

Πως 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

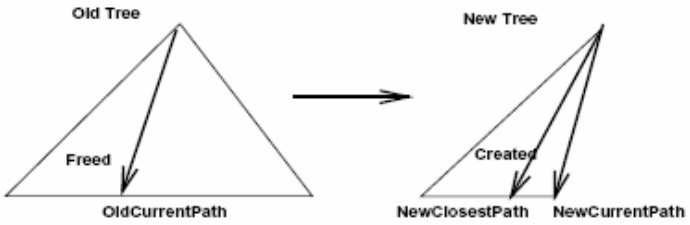


Figure 3: *Rebuilding CF Tree*

BIRCH-αλγόριθμος

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

BIRCH

Και άλλες βελτιώσεις όπως:

Έλεγχος για outliers
Delay-split

BIRCH

- Τοπικότητα: κάθε απόφαση σχετικά με συσταδοποίηση παίρνετε 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