Συσταδοποίηση ΙΙ

Οι διαφάνειες στηρίζονται στο **P.-N. Tan, M.Steinbach, V. Kumar, «Introduction to Data Mining»,** Addison Wesley, 2006

Διαχείριση Ποιότητας
Cluster validity

Ποιότητα Συσταδοποίησης

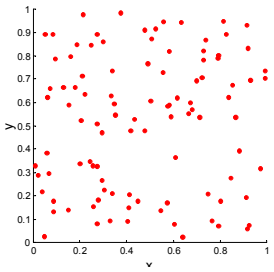
Πόσο καλή είναι συσταδοποίηση που επιτύχαμε;

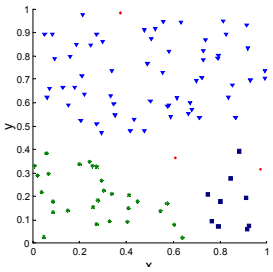
Οι αλγόριθμοι που είδαμε παράγουν κάποιες συστάδες ακόμα και όταν τα δεδομένα παράγονται τυχαία

Δύσκολη η αξιολόγηση, ιδιαίτερα σε πολλές διαστάσεις

Συστάδες σε Τυχαία Δεδομένα

Τυχαία Σημεία

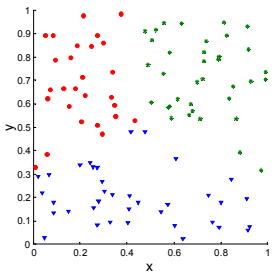


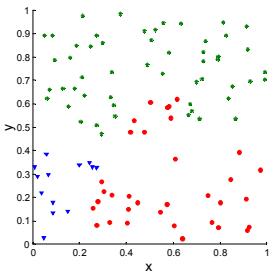


DBSCAN

3 ομάδες κοιτώντας την απόσταση του 4ου γείτονα

K-means





ΣΙΣ με MAX-link

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 Πίνακα Γειτνίασης
- Χρήση Συνεκτικότητας και Διαχωρισμού

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**

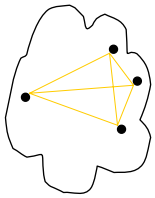
$$overall - validity = \sum_{i=1}^{k} w_i \, validity(C_i)$$

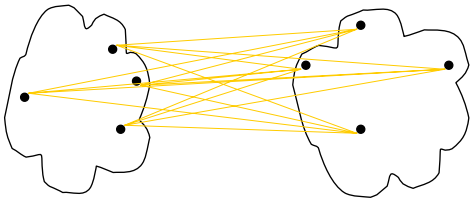
Όπου το βάρος (w_i) μπορεί να είναι πχ ανάλογο του μεγέθους της συστάδας ή η τετραγωνική ρίζα της συνεκτικότητας ή 1

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

Συσταδοποίηση βασισμένη σε γραφήματα (ΣΙΣ)

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



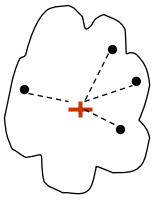


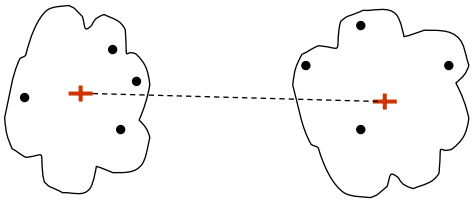
$$cohesion(C_i) = \sum_{\substack{x \in C_i \\ y \in C_i}}^{n} proximity(x, y)$$

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

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

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





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

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

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

Αν proximity = τετράγωνο της Ευκλείδειας, τότε ESS

Όπου c το κέντρο όλων των σημείων

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

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

Για prototype και graph

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

Για prototype

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

Για graph

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

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

Ισομεγέθεις συστάδες

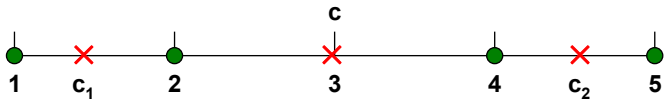
$m_i = m / K$

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

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 = 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**

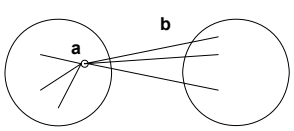
Silhouette Coefficient (συντελεστής σκιαγράφησης)

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 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: **Πίνακας Γειτνίασης**

Δύο Πίνακες

Πίνακας Γειτνίασης (proximity matrix)

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

Συσχέτιση

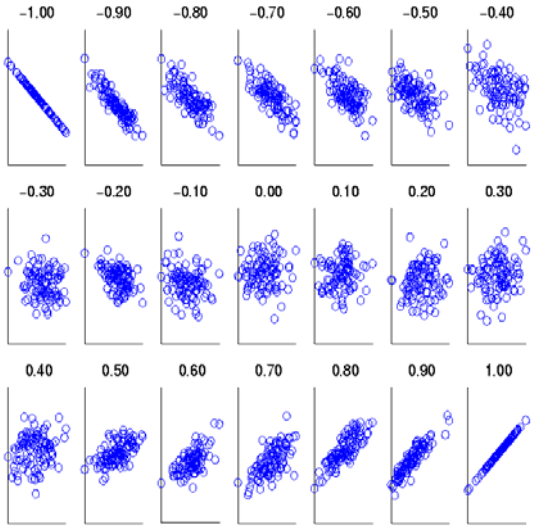
- Μετρά τη *γραμμική σχέση* μεταξύ αντικειμένων
- To compute correlation, we standardize data objects, p and q, and then take their *dot product*

$$p'_k = (p_k - mean(p)) / std(p)$$

$$q'_k = (q_k - mean(q)) / std(q)$$

$$correlation(p, q) = p' \bullet q'$$

Συσχέτιση



Θετική συσχέτιση

Μεγάλο x => μεγάλο y

1 (-1) σημαίνει τέλεια γραμμική συσχέτιση

0 όχι γραμμική συσχέτιση

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

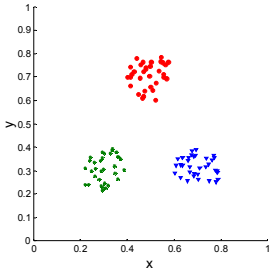
Υψηλή συσχέτιση σημαίνει ότι τα σημεία που ανήκουν στην ίδια συστάδα είναι κοντινά μεταξύ τους

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

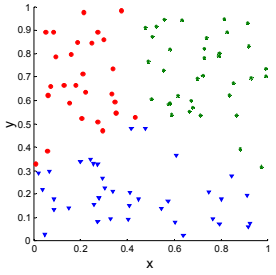
- Επειδή, οι δυο πίνακες είναι συμμετρικοί, χρειάζεται ο υπολογισμός n(n-1) / 2 εγγραφών

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:

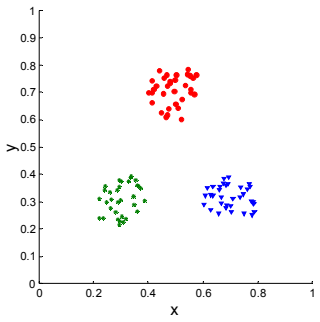
Σημεία Συστάδας 1, Σημεία Συστάδας 2, Σημεία Συστάδας 3

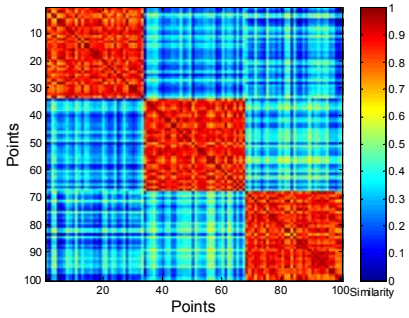
Χαρακτηρισμός Ποιότητας Συσταδοποίησης **χωρίς** Επίβλεψη: **Πίνακας Γειτνίασης - *Οπτικοποίηση***

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

Συγκεκριμένα, τα διατάσουμε με βάση τη συστάδα:

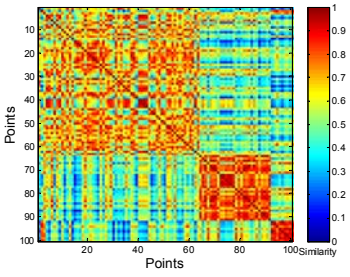
Σημεία Συστάδας 1, Σημεία Συστάδας 2, Σημεία Συστάδας 3

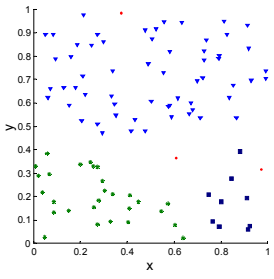




Σημείωση s = 1 – (d –min_d)/(max_d – min_d)

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

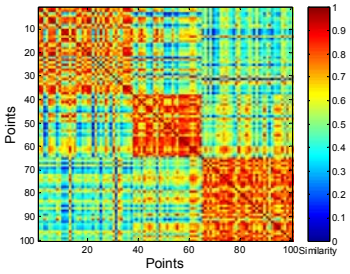


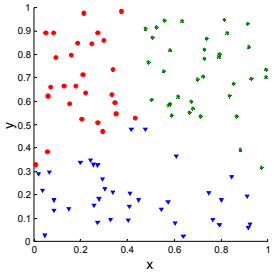


DBSCAN

Κάποιες συστάδες ακόμα και σε τυχαία δεδομένα

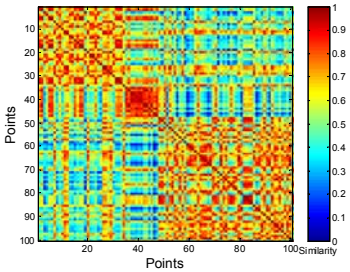
Χαρακτηρισμός Ποιότητας Συσταδοποίησης **χωρίς** Επίβλεψη:
Πίνακας Γειτνίασης - Οπτικοποίηση

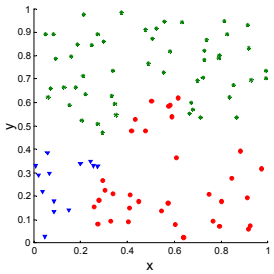




K-means

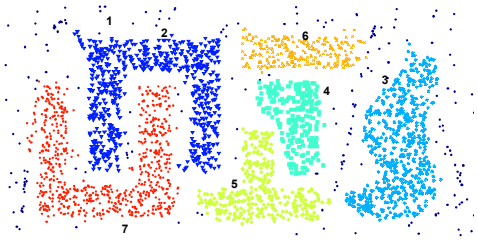
Χαρακτηρισμός Ποιότητας Συσταδοποίησης **χωρίς** Επίβλεψη:
Πίνακας Γειτνίασης - Οπτικοποίηση

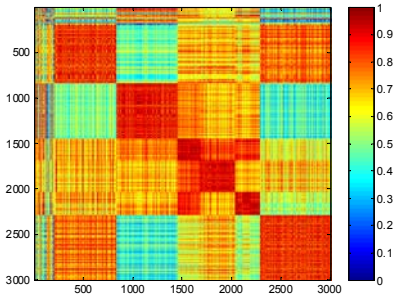




ΣΙΣ-max

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





DBSCAN

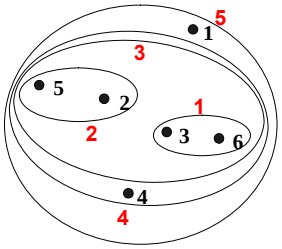
Χαρακτηρισμός Ποιότητας Συσταδοποίησης **χωρίς** Επίβλεψη: **Πίνακας Γειτνίασης**

Ειδικά για ιεραρχικούς αλγόριθμους

Cophenetic distance: είναι η απόσταση (proximity) όταν ο αλγόριθμος τοποθετεί τα δυο σήμεια 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

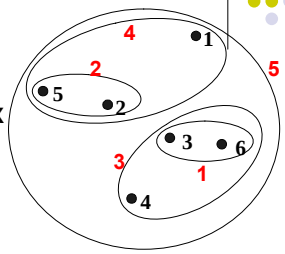
CoPhenetic Correlation Coefficient (CPCC)
Χρησιμοποείται για επιλογή του είδους της ιεραρχικής συσταδοποίησης

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



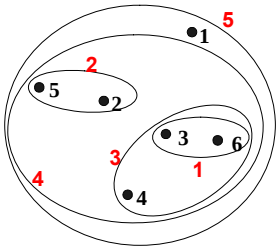
MIN

0.44



MAX

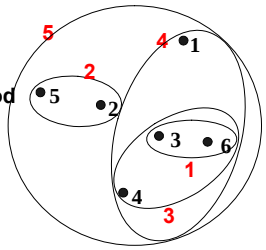
0.63



0.66

Group Average

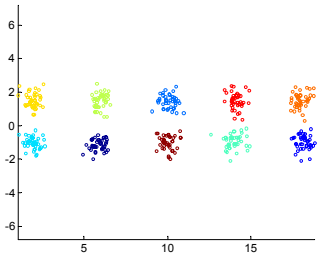
0.64

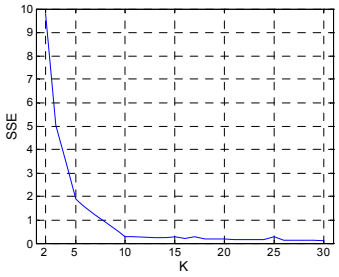


Ward's Method

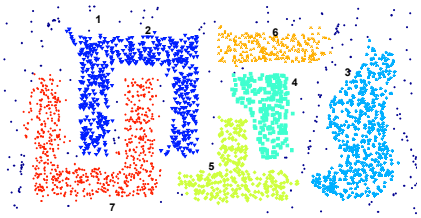
Χαρακτηρισμός Ποιότητας Συσταδοποίησης **χωρίς** Επίβλεψη: **Συνεκτικότητα και Διαχωρισμός**

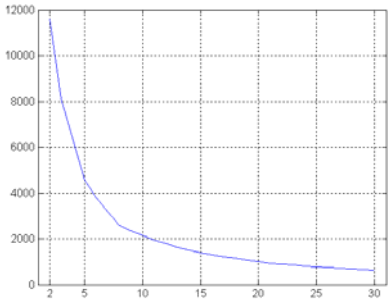
Χρήση SSE για υπολογισμό του σωστού αριθμού συστάδων χρησιμοποιώντας τον K-means
(K = 5 και 10 φαίνονται καλές τιμές)





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

Χαρακτηρισμός Ποιότητας Συσταδοποίησης με Επίβλεψη:

Table 5.9. K-means Clustering Results for LA Document Data Set

Cluster	Entertainment	Financial	Foreign	Metro	National	Sports	Entropy	Purity
1	3	5	40	506	96	27	1.2270	0.7474
2	4	7	280	29	39	2	1.1472	0.7756
3	1	1	1	7	4	671	0.1813	0.9796
4	10	162	3	119	73	2	1.7487	0.4390
5	331	22	5	70	13	23	1.3976	0.7134
6	5	358	12	212	48	13	1.5523	0.5525
Total	354	555	341	943	273	738	1.1450	0.7203

entropy For each cluster, the class distribution of the data is calculated first, i.e., for cluster j we compute p_{ij}, the 'probability' that a member of cluster j belongs to class i as follows: $p_{ij} = m_{ij}/m_j$, where m_j is the number of values in cluster j and m_{ij} is the number of values of class i in cluster j. Then using this class distribution, the entropy of each cluster j is calculated using the standard formula $e_j = \sum_{i=1}^{L} p_{ij} \log_2 p_{ij}$, where the L is the number of classes. The total entropy for a set of clusters is calculated as the sum of the entropies of each cluster weighted by the size of each cluster, i.e., $e = \sum_{i=1}^{K} \frac{m_i}{m} e_j$, where m_j is the size of cluster j, K is the number of clusters, and m is the total number of data points.

purity Using the terminology derived for entropy, the purity of cluster j, is given by $purity_j = \max p_{ij}$ and the overall purity of a clustering by $purity = \sum_{i=1}^{K} \frac{m_i}{m} purity_j$.

Χαρακτηρισμός Ποιότητας Συσταδοποίησης με Επίβλεψη:

- Ιδανικός πίνακας ομοιότητας συστάδων
- Ιδανικός πίνακας ομοιότητας κλάσεων

Αλγόριθμοι Συσταδοποίησης

BIRCH

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 απλό διάβασμα (scan) των δεδομένων
  - Ένα ή περισσότερα επιπρόσθετα περάσματα για βελτίωση της ποιότητας της συσταδοποίησης

BIRCH: **B**alanced **I**terative **R**educing and **C**lustering using **H**ierarchies

BIRCH

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

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

Radius: *average distance from member points to centroid*

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

Diameter: *average pair-wise distance within a cluster*

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

BIRCH

centroid Euclidean distance
centroid Manhattan distance

Μεταξύ δυο συστάδων

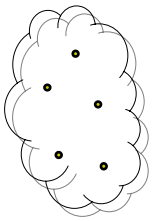
$$D0 = ((\vec{X0}_1 - \vec{X0}_2)^2)^{\frac{1}{2}}$$

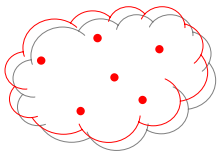
$$D1 = |\vec{X0}_1 - \vec{X0}_2| = \sum_{i=1}^{d} |\vec{X0}_1^{(i)} - \vec{X0}_1^{(i)}|$$

BIRCH

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

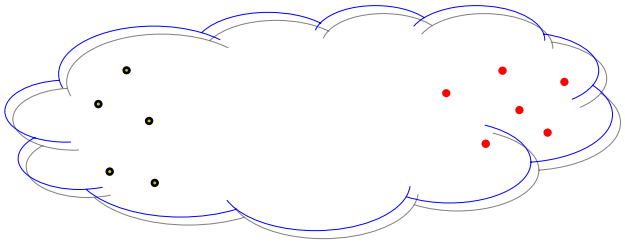
Cluster $\{X_j\}$:
$j = N_1+1, N_1+2, \ldots, N_1+N_2$





BIRCH

Cluster $X_l = \{X_i\} + \{X_j\}$:
$l = 1, 2, \ldots, N_1, N_1+1, N_1+2, \ldots, N_1+N_2$



BIRCH

average inter-cluster (D2)
average intra-cluster (D3)
variance increase (D4)

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

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

D της συγχωνευμένης συστάδας

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

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:

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

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

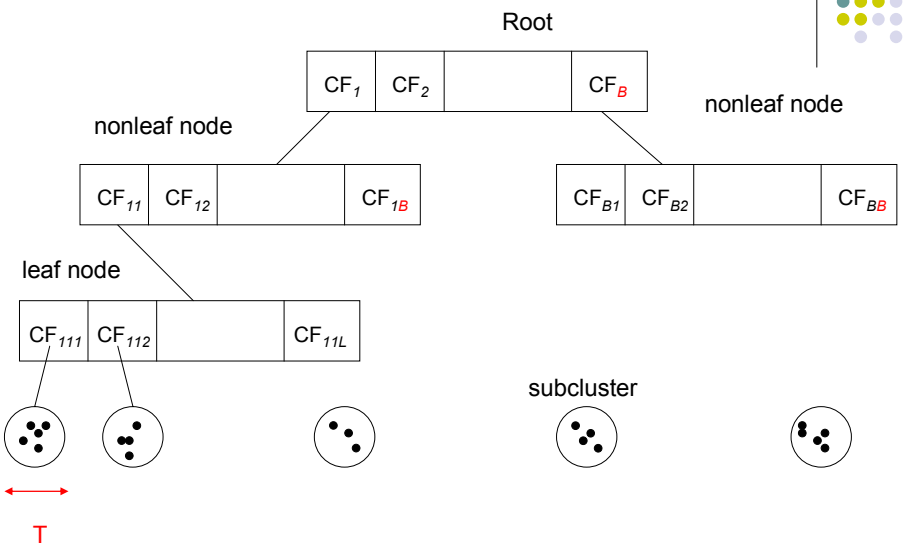
Χτίζει σταδιακά καθώς διαβάζει τα δεδομένα ένα δεντρό-γραμμα του οποίου κόμβοι είναι οι τιμές CF που περιγράφουν τα δεδομένα κάθε υπο-συστάδας

BIRCH: CF δέντρο

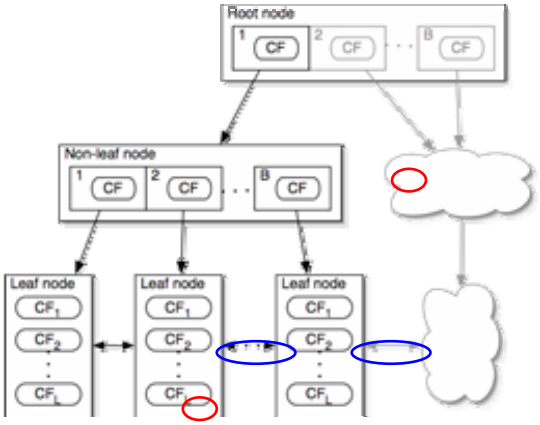
Το CF-δέντρο είναι ένα ισοζυγισμένο δέντρο με δυο παραμέτρους

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

BIRCH: CF-δέντρο



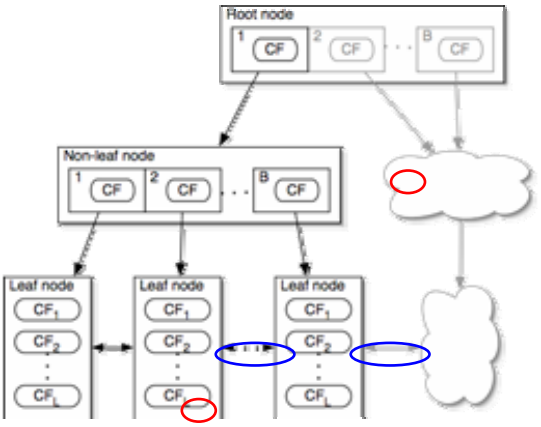
BIRCH: CF-δέντρο



- 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

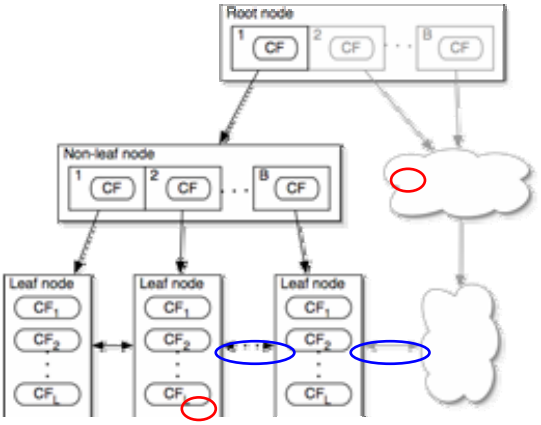
BIRCH: CF-δέντρο



- 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

Για ένα φύλλο:

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

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

For a non-leaf node, which has child nodes N_1, N_2, …, 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

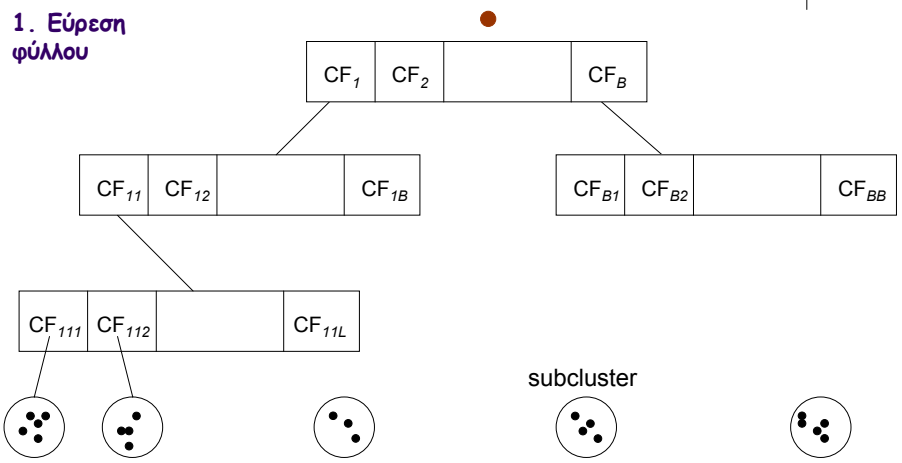
- Ο αλγόριθμος διαβάζει (scan) τα δεδομένα και τα εισάγει στο CF δέντρο ένα-ένα
- Η εισαγωγή ενός στοιχείου στο CF-δέντρο γίνεται με top-down διάσχιση ξεκινώντας από τη ρίζα με βάση μια συνάρτηση απόστασης Distance(σημείο, cluster)
  - Χρήση της D0, D1, D2, D3 ή D4
- Κάθε σημείο εισάγεται στο κοντινότερη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

BIRCH: CF-δέντρο εισαγωγή στοιχείου

1. Εύρεση κατάλληλου φύλου
   αν το φύλλο μπορεί να το απορροφήσει
   (διάμετρος παραμένει <= T) ok,
   Αλλιώς 3
2. Ενημέρωση του φύλλου
3. Διάσπαση φύλλου
4. Ενημέρωση τιμής CF

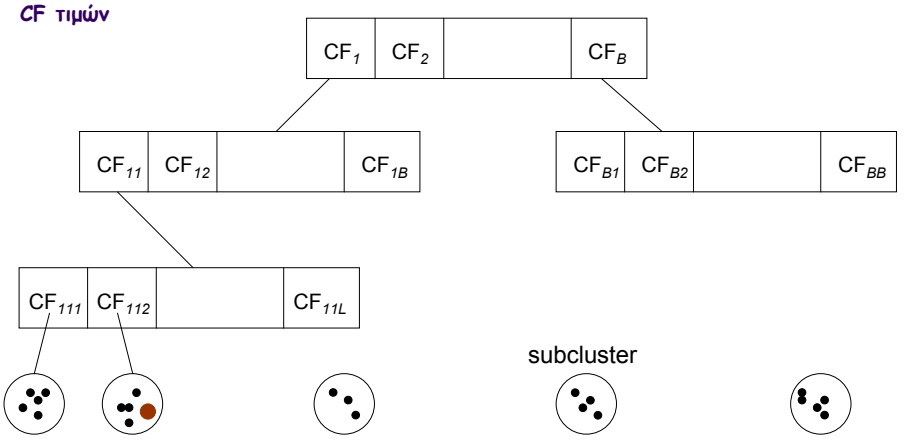
BIRCH: CF-δέντρο εισαγωγή στοιχείου

1. Εύρεση φύλλου



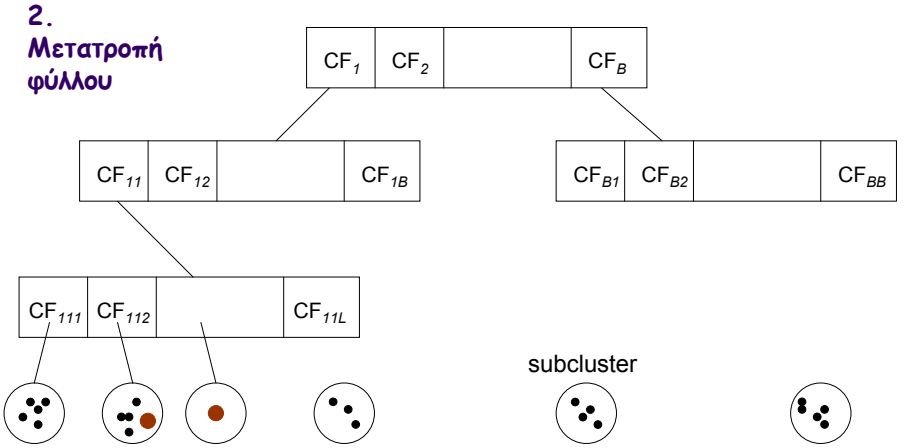
BIRCH: CF-δέντρο εισαγωγή στοιχείου

4.
Τροποποίηση
CF τιμών



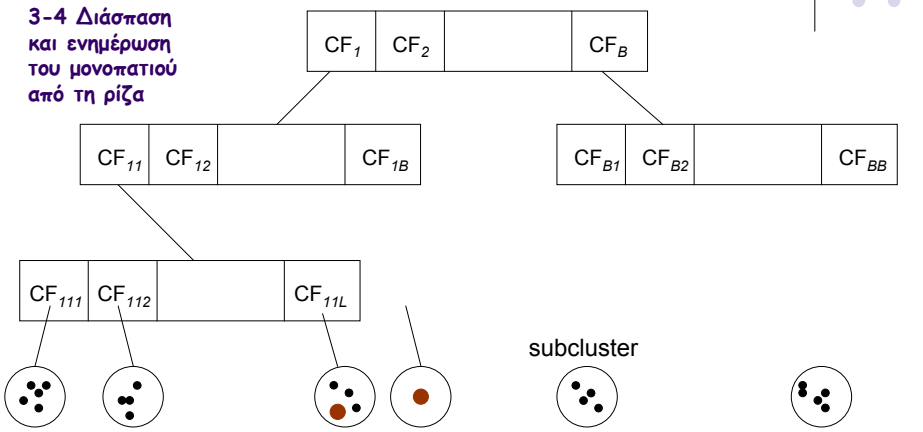
BIRCH: CF-δέντρο εισαγωγή στοιχείου

2.
Μετατροπή
φύλλου



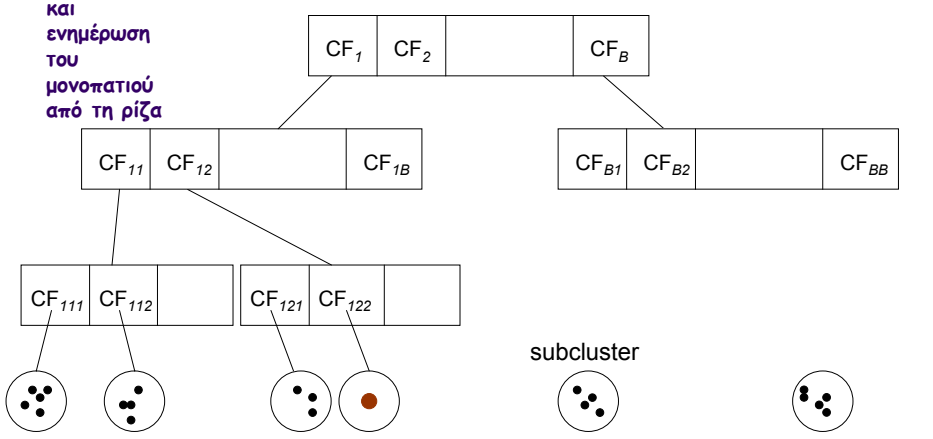
BIRCH: CF-δέντρο εισαγωγή στοιχείου

3-4 Διάσπαση και ενημέρωση του μονοπατιού από τη ρίζα



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 T**, τότε έχουμε δημιουργία νέας υποσυστάδας
  - Η δημιουργία μιας νέας υπο-συστάδας μπορεί να οδηγήσει το φύλλο που την περιέχει να υπερχειλίσει

BIRCH: CF-δέντρο

- Διάσπαση φύλλου (split of a leaf)

  Εύρεση των δύο υπο-συστάδων του φύλλου πού έχουν τη μεγαλύτερη απόσταση μεταξύ τους, έστω C1 και C2

  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 C1 ή στο φύλλο της C2 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!

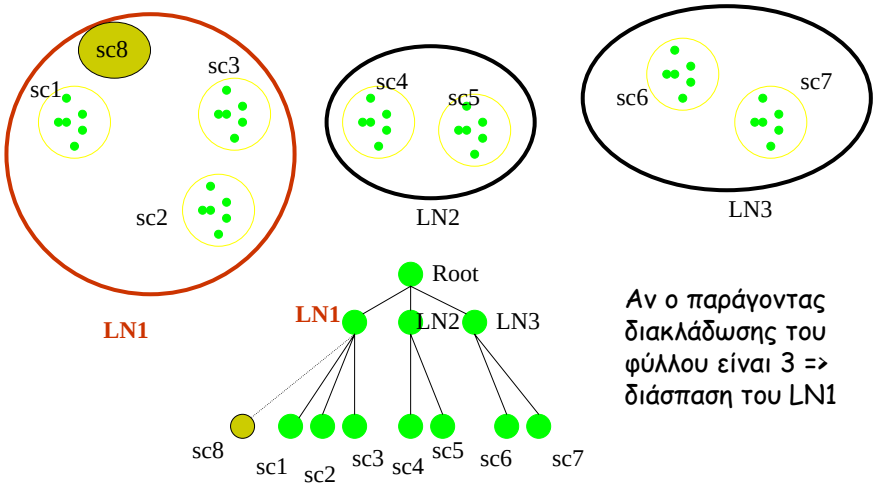
Μια μικρή διόρθωση:

- Όταν η διάσπαση κάποιων κόμβων τελειώνει (χωρούν σε ένα κόμβο) έστω στον κόμβο Nj κοιτάμε τον κόμβο Nj 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 δε χωρούν πρέπει να κάνουμε πάλι διάσπαση

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

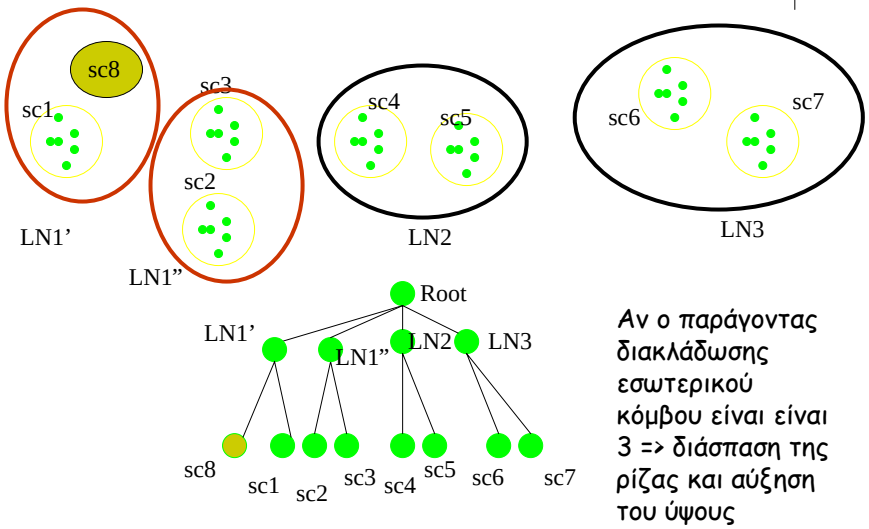
BIRCH: CF-δέντρο

Νέα υπο-συστάδα



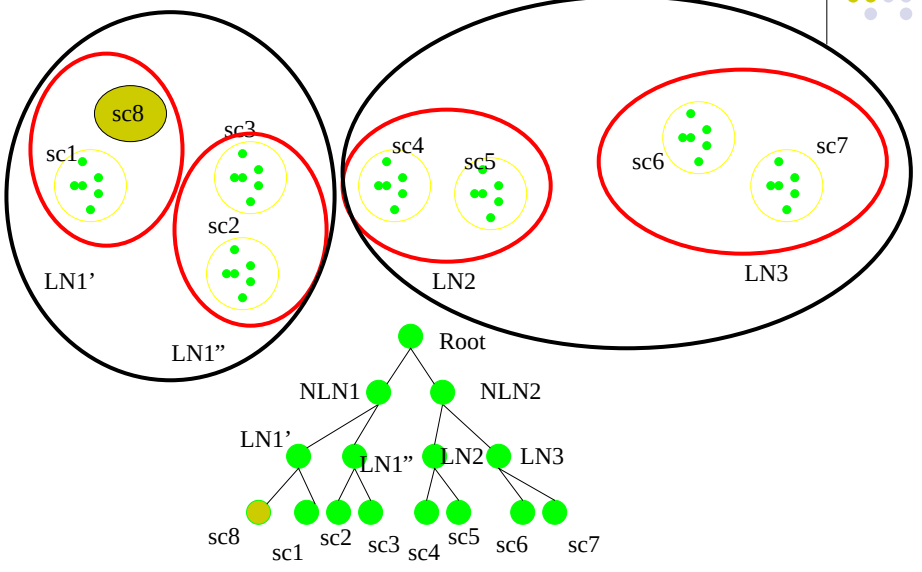
Αν ο παράγοντας διακλάδωσης του φύλλου είναι 3 => διάσπαση του LN1

BIRCH: CF-δέντρο



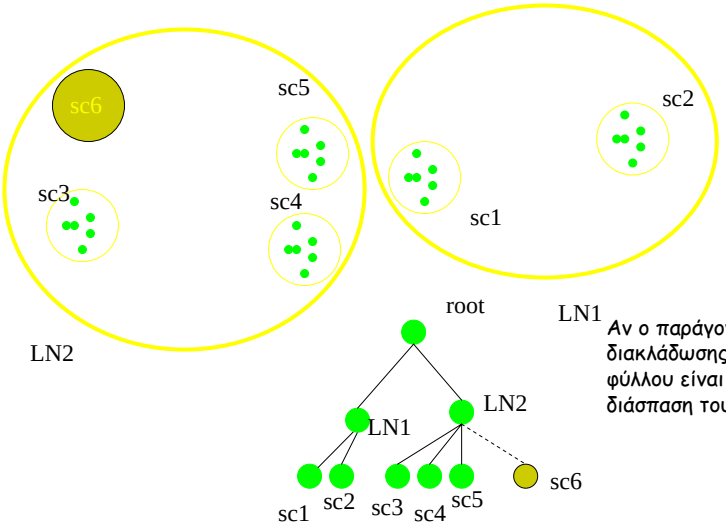
Αν ο παράγοντας διακλάδωσης εσωτερικού κόμβου είναι είναι 3 => διάσπαση της ρίζας και αύξηση του ύψους

BIRCH: CF-δέντρο



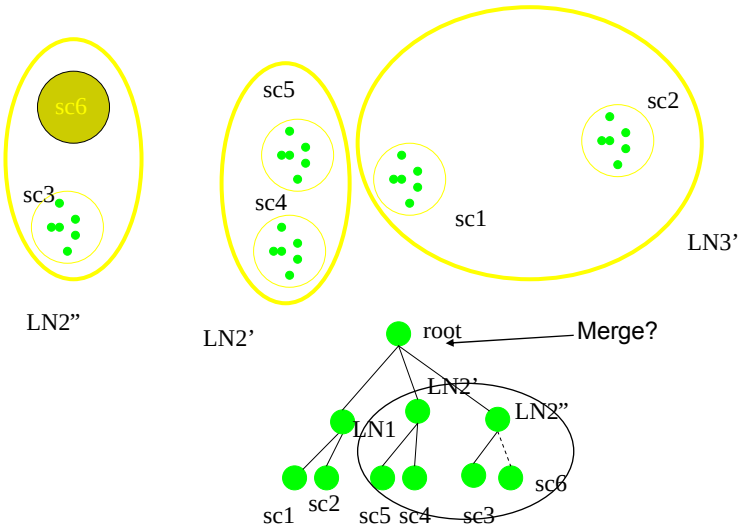
BIRCH: CF-δέντρο

Assume that the subclusters are numbered according to the order of formation.

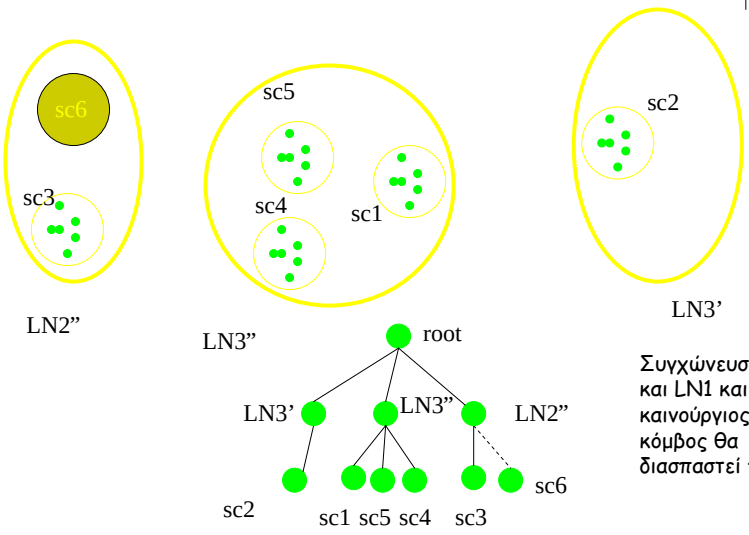


Αν ο παράγοντας διακλάδωσης του φύλλου είναι 3 => διάσπαση του LN2

BIRCH: CF-δέντρο



BIRCH: CF-δέντρο



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

BIRCH: αλγόριθμος

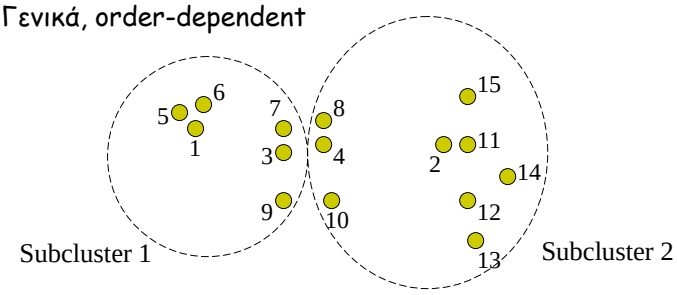
Επειδή η κατασκευή επηρεάζεται από το μέγεθος της σελίδας

Αφού κατασκευαστεί το δέντρο,

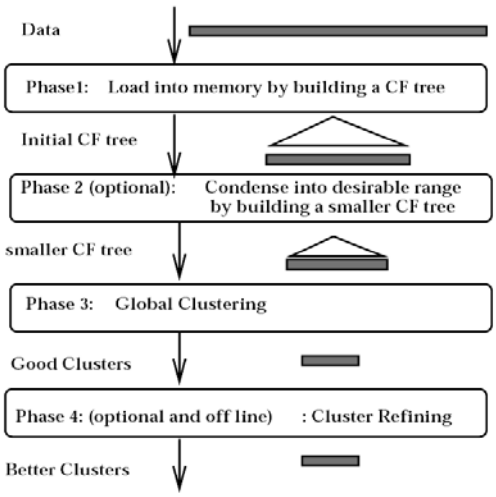
ο BIRCH χρησιμοποιεί έναν ιεραρχικό αλγόριθμο συσταδοποίησης για να συσταδοποιήσει τις συστάδες των φύλλων.

BIRCH: αλγόριθμος

- The objects are numbered by the incoming order and assume that the distance between objects 1 and 2 exceeds the diameter threshold.
- Επίσης, αν το ίδιο αντικείμενο ξανα-εισαχθεί μπορεί να μπει σε άλλο φύλλο
- Τέλος, πρόβλημα με skewed data
- Γενικά, order-dependent



BIRCH-αλγόριθμος



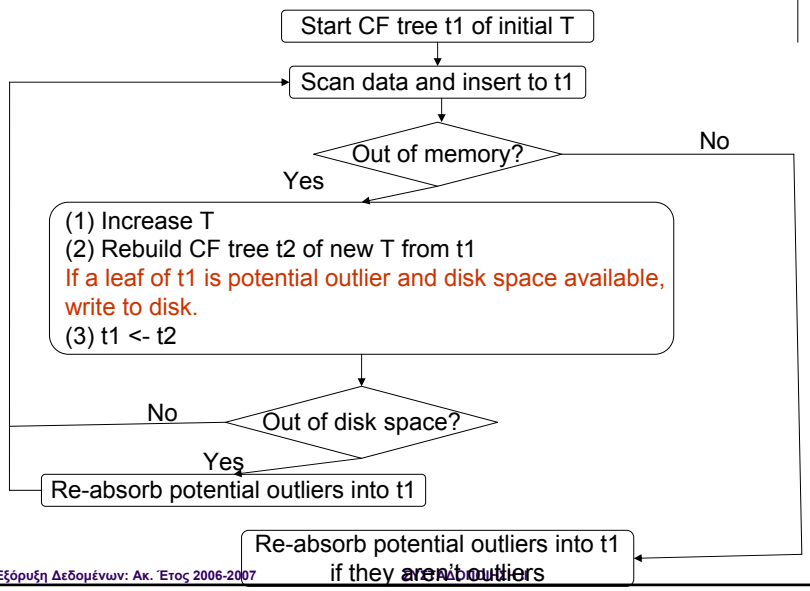
BIRCH-αλγόριθμος

Ξεκίνα με κάποια αρχική τιμή για το threshold (T)

Διάβασε τα δεδομένα και εισήγαγε τα 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 οι τιμές από το παλιό στο νέο δέντρο,
Συνεχίζεται η ανάγνωση των δεδομένων από εκεί που είχε σταματήσει

BIRCH-αλγόριθμος



BIRCH-αλγόριθμος

Βασίζεται σε μονοπάτια

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

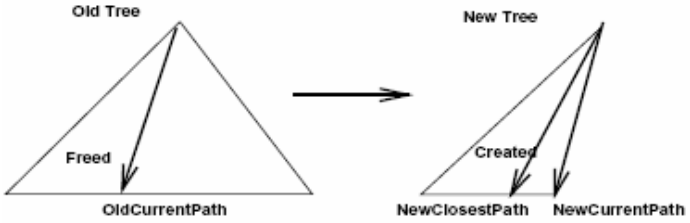


Figure 3: *Rebuilding CF Tree*

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: Επανακατασκευή του CF-δέντρου

Με την αύξηση του T απορροφά πιο πολλά δεδομένα

Rebuilding "pushes" CFs over
The larger T allows different CFs to group together

Reducibility theorem
Increasing T will result in a CF-tree as small or smaller then the original
Rebuilding needs at most h extra pages of memory

BIRCH

Phase 1: Φόρτωση δεδομένων στη μνήμη

Κτίσιμο αρχικού in-memory CF-δέντρου με τα δεδομένα (one scan)

Phase 2: Condense data
Rebuild the CF-tree with a larger T
Condensing is optional
Απομάκρυνση outliers

BIRCH

Phase 3: Global clustering
Use existing clustering algorithm on CF entries
Helps fix problem where natural clusters span nodes
D2 or D4

Phase 4: Cluster refining
Do additional passes over the dataset & reassign data points to the closest centroid from phase 3
Refining is optional

BIRCH

Why have optional phases?

Phase 2 allows us to resize the data set so Phase 3 runs on an optimally sized data set

Phase 4 fixes a problem with CF-trees where some data points may be assigned to different leaf entries

Phase 4 will always converge to a minimum

Phase 4 allows us to discard outliers

BIRCH: Delay-Split Option

When we run out of memory and some points require us to split a node, we can write them to disk until we run out of disk space.

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