Αλγόριθμος BIRCH

Αναπλ. Καθηγ. Στελιος Ζήμερας
Τμημα Στατιστικης και Αναλογιστικων – Χρηματοοικονομικων Μαθηματικων
Πανεπιστημιο Αιγαιου
Σαμος

2021

Εισαγωγή

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

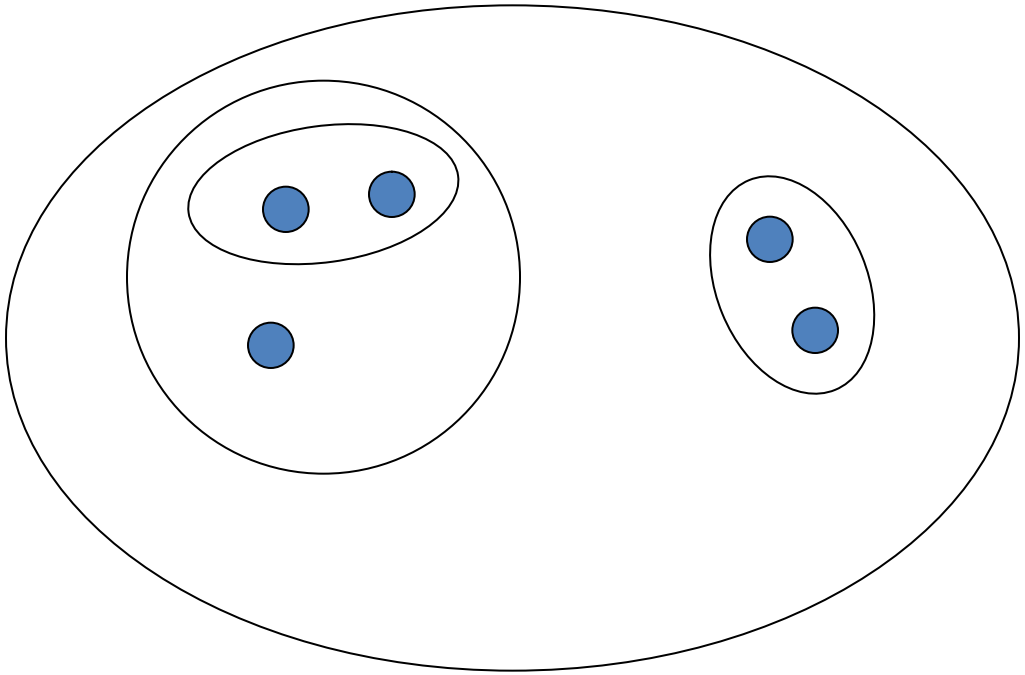
Ένας αλγόριθμος εξόρυξης δεδομένων (data mining) χρησιμοποιώντας την ιεραρχική ομαδοποίηση (hierarchical clustering) σε ιδιαίτερα μεγάλα σύνολα δεδομένων.

Εισαγωγή

- κάθε απόφαση για ομαδοποίηση γίνεται χωρίς σάρωση όλων των σημείων δεδομένων και των ήδη δημιουργημένων clusters.
- Δεν είναι όλα τα αντικείμενα το ίδιο σημαντικά για την ομαδοποίηση

Εισαγωγή

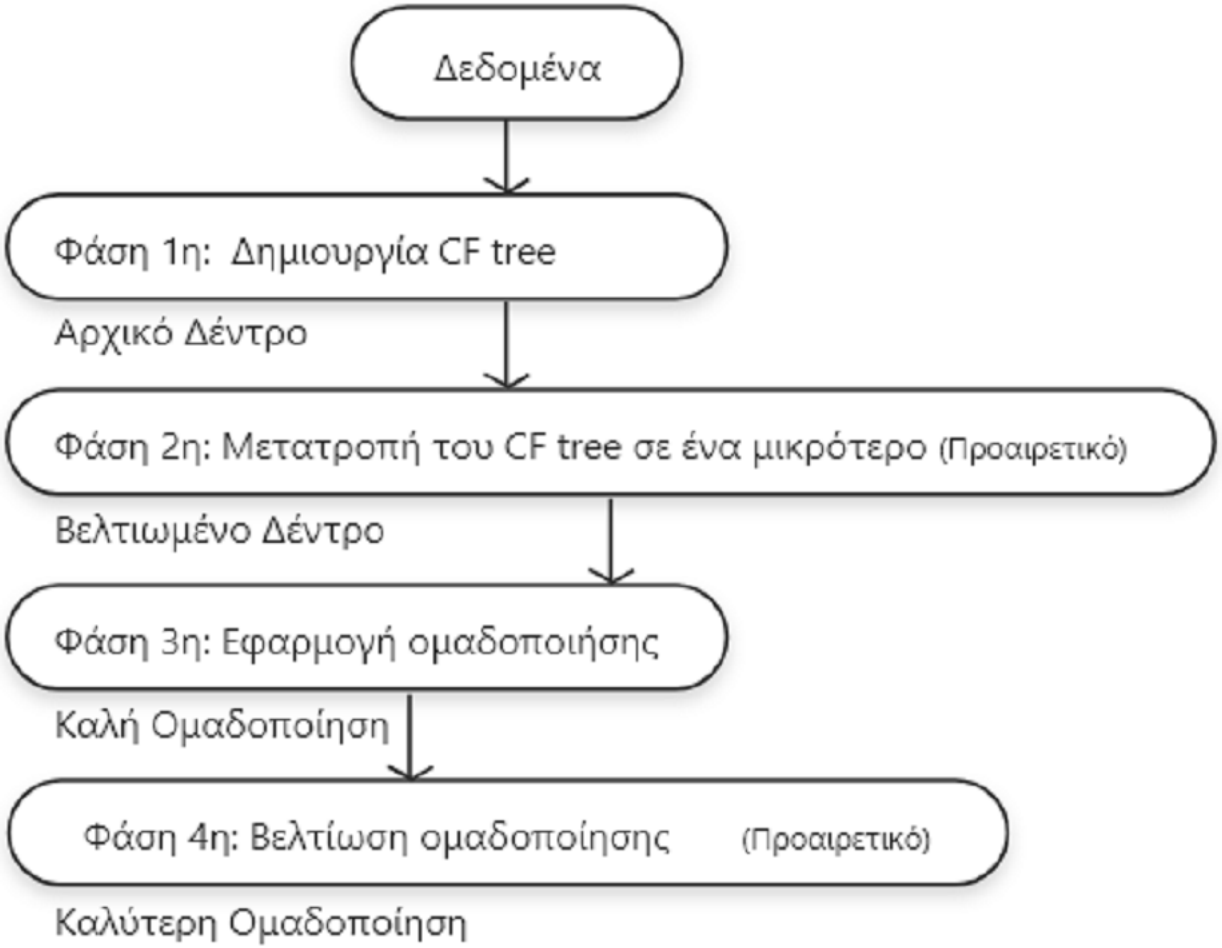
- Hierarchical clustering



Μέθοδοι

- Ο αλγόριθμος ξεκινά με ένα μόνο σημείο(κάθε σημείο στη βάση δεδομένων θεωρείται ένα cluster)
- Μετά ομαδοποιεί τα κοντινότερα σημεία στο ίδιο cluster και συνεχίζει μέχρι να έχουμε ένα μόνο cluster.
- Ο BIRCH δημιουργεί ένα ζυγισμένο δέντρο CF tree καθώς διατρέχει τα δεδομένα.

Μέθοδοι


Δεδομένα
Φάση 1η: Δημιουργία CF tree
Αρχικό Δέντρο
Φάση 2η: Μετατροπή του CF tree σε ένα μικρότερο (Προαιρετικό)
Βελτιωμένο Δέντρο
Φάση 3η: Εφαρμογή ομαδοποιήσης
Καλή Ομαδοποίηση
Φάση 4η: Βελτίωση ομαδοποίησης (Προαιρετικό)
Καλύτερη Ομαδοποίηση

Αλγόριθμος BIRCH

Δεδομένης μια ομάδας από αντικείμενα $\{\vec{X}_i\}$, ορίζουμε:

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

Ακτίνα Μέση απόσταση των σημείων από το κέντρο

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

Διάμετρος Μέση pair-wise απόσταση των σημείων μέσα στην ομάδα

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

Αλγόριθμος BIRCH

Clustering Feature- Χαρακτηριστικό Συσταδοποίησης

- Ο BIRCH δημιουργεί ένα ζυγισμένο δέντρο CF tree καθώς διατρέχει τα δεδομένα.

- Κάθε κόμβος στο CF tree αντιπροσωπεύει ένα cluster και χαρακτηρίζεται από μια CF τριάδα (N, LS, SS).

  - N - number of data points in the cluster
  - LS - linear sum of the N data points

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

  - SS - square sum of the N data points

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

Αλγόριθμος BIRCH

- ένα ισοζυγισμένο δέντρο με δύο παραμέτρους:
  - παράγοντα διακλάδωσης *B*
  - δοθέν κατώφλι *T*

- Κάθε εσωτερικός κόμβος αποτελείται από το πολύ B εγγραφές της μορφής $[CF_i, child_i]$, όπου $child_i$ είναι δείκτης στο *i-οστό* κόμβο-παιδί και CF_i είναι ένα CF της υποσυστάδας που αντιπροσωπεύεται από το παιδί.

- Κάθε εσωτερικός κόμβος αντιπροσωπεύει ένα cluster κατασκευασμένο από όλα τα subclusters που δημιουργούνται από αυτές τις εγγραφές.

Αλγόριθμος BIRCH

- Ένα φύλλο αποτελείται το πολύ από L εγγραφές της μορφής $[CF_i]$, όπου i = 1, 2, …, L .

- Αποτελείται από δύο δείκτες, *prev* and *next*, που ενώνουν όλα τα φύλλα μαζί για αποδοτικές σαρώσεις.

- Ένα φύλλο απεικονίζει ένα cluster φτιαγμένο από όλα τα subclusters που απεικονίζουν τις εγγραφές του.

- Όλες οι εγγραφές ενός φύλλου υπακούν σε ένα όριο που υποδεικνύεται από τη τιμή κατωφλίου *T*.

- Η διάμετρος ενός κόμβου φύλλου → μικρότερη από T.

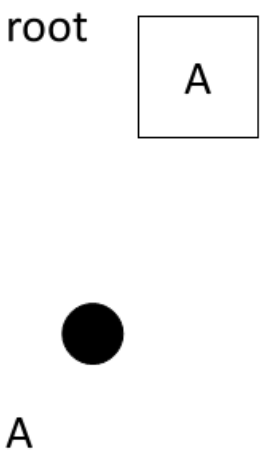
Αλγόριθμος BIRCH

- Το μέγεθος του CF Tree είναι μια συνάρτηση του*T* (όσο μεγαλύτερο το *T* , τόσο μικρότερο το δέντρο).
- **P** (page size σε bytes) είναι το μέγιστο μέγεθος ενός κόμβου
- *B* και *L* καθορίζονται από το *P* (*που μπορεί να είναι διαφορετικό για καλύτερη απόδοση*).
- Το κάθε φύλλο περιέχει ένα cluster.
- Το μέγεθος του κάθε cluster σε ένα φύλλο δεν είναι μεγαλύτερο από *T*.

Αλγόριθμος BIRCH

Παράδειγμα CF Tree

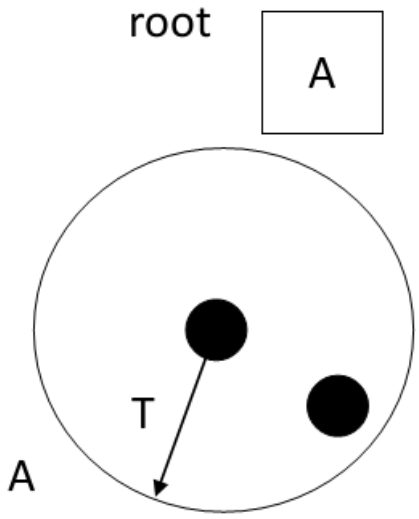
Αρχικά, τα δεδομένα αποτελούν ένα cluster



ΠΑΡΑΔΕΙΓΜΑΤΑ

Παράδειγμα CF Tree

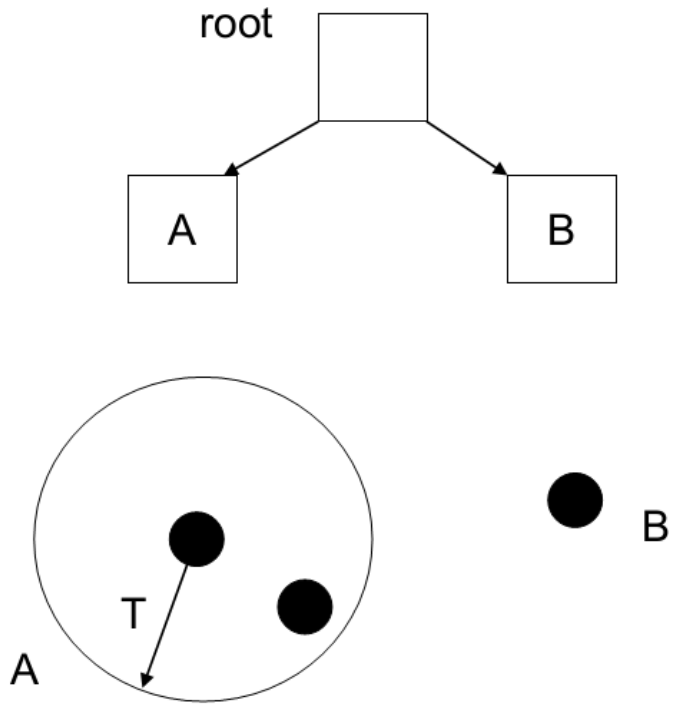
Όταν φτάνουν τα δεδομένα, ελέγχουμε αν το μέγεθος του cluster δεν ξεπερνά το κατώφλι T.



ΠΑΡΑΔΕΙΓΜΑΤΑ

Παράδειγμα CF Tree

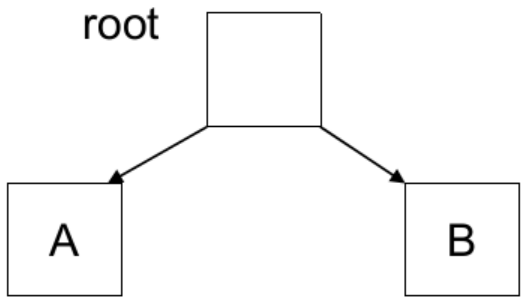
Αν το μέγεθος του cluster μεγαλώνει αρκετά, τότε το cluster «σπάει» σε δύο και τα σημεία αναδιανέμονται.

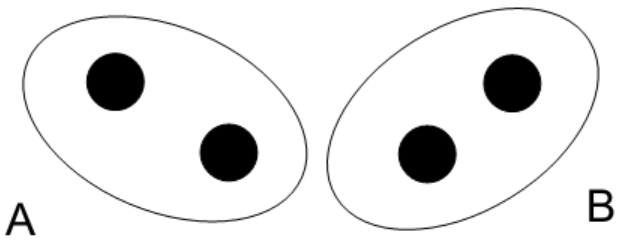


ΠΑΡΑΔΕΙΓΜΑΤΑ

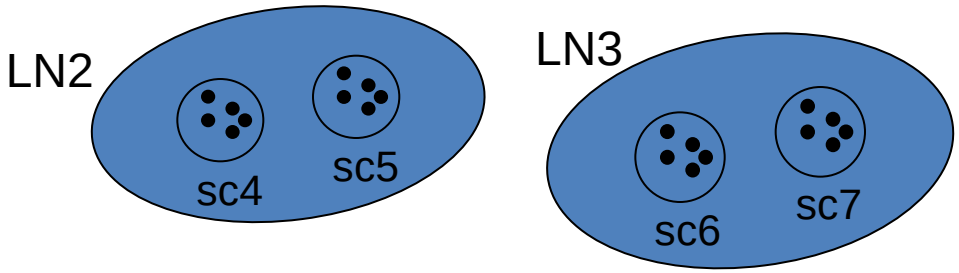
Παράδειγμα CF Tree

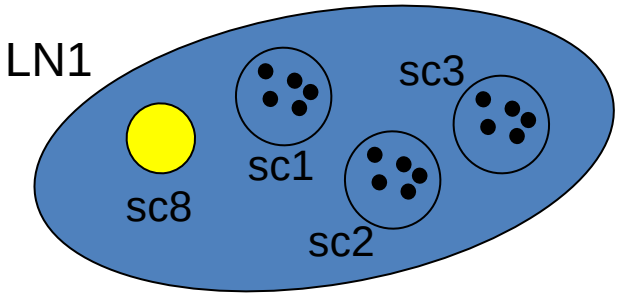
Σε κάθε κόμβο, το CF δέντρο κρατά ως πληροφορία: το κέντρο του cluster και το Μ.Ο του αθροίσματος των τετραγώνων για να υπολογίζει σωστά το μέγεθος του cluster

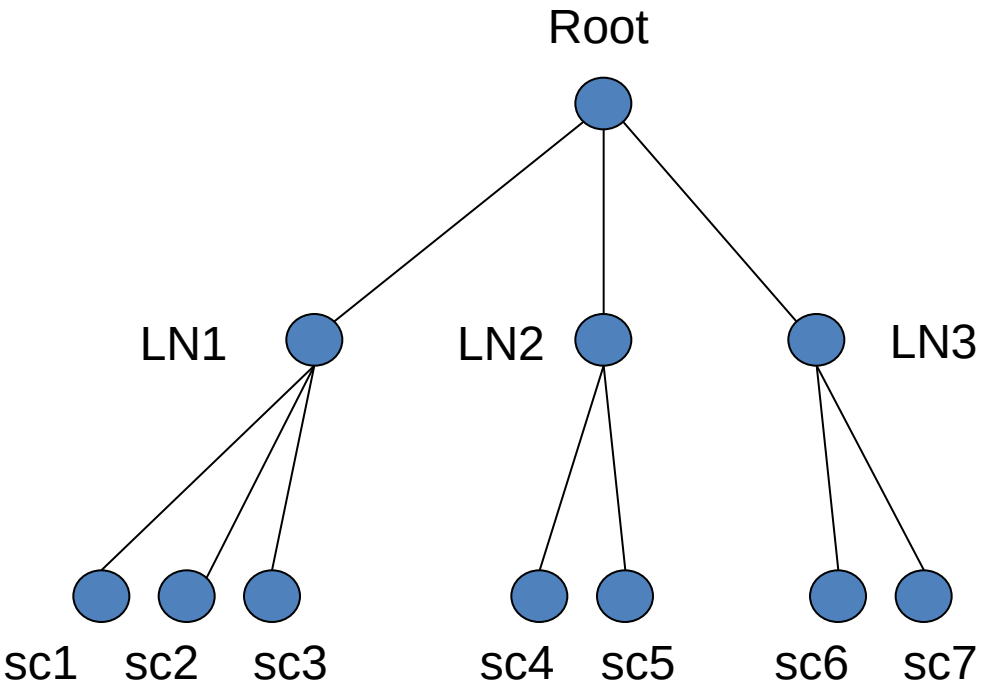




ΠΑΡΑΔΕΙΓΜΑΤΑ

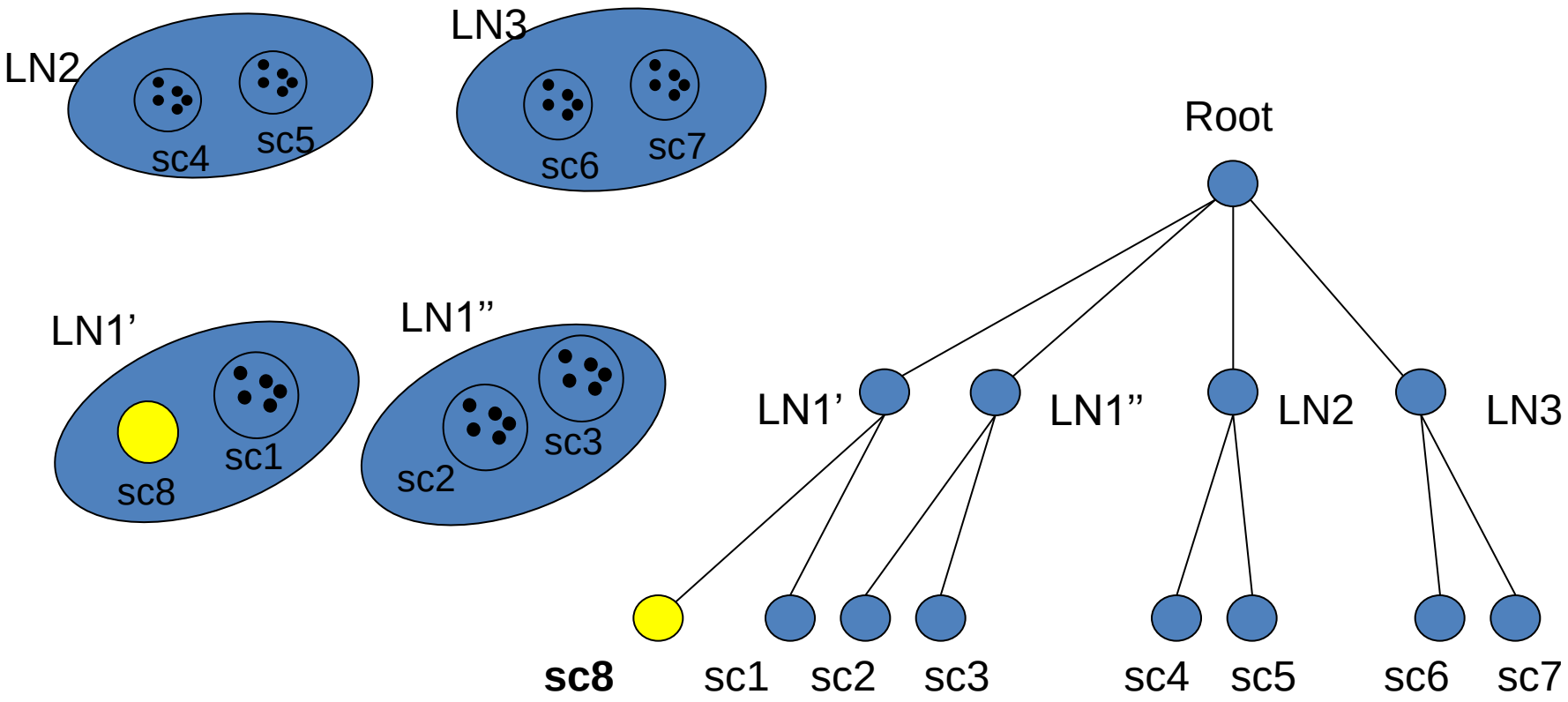






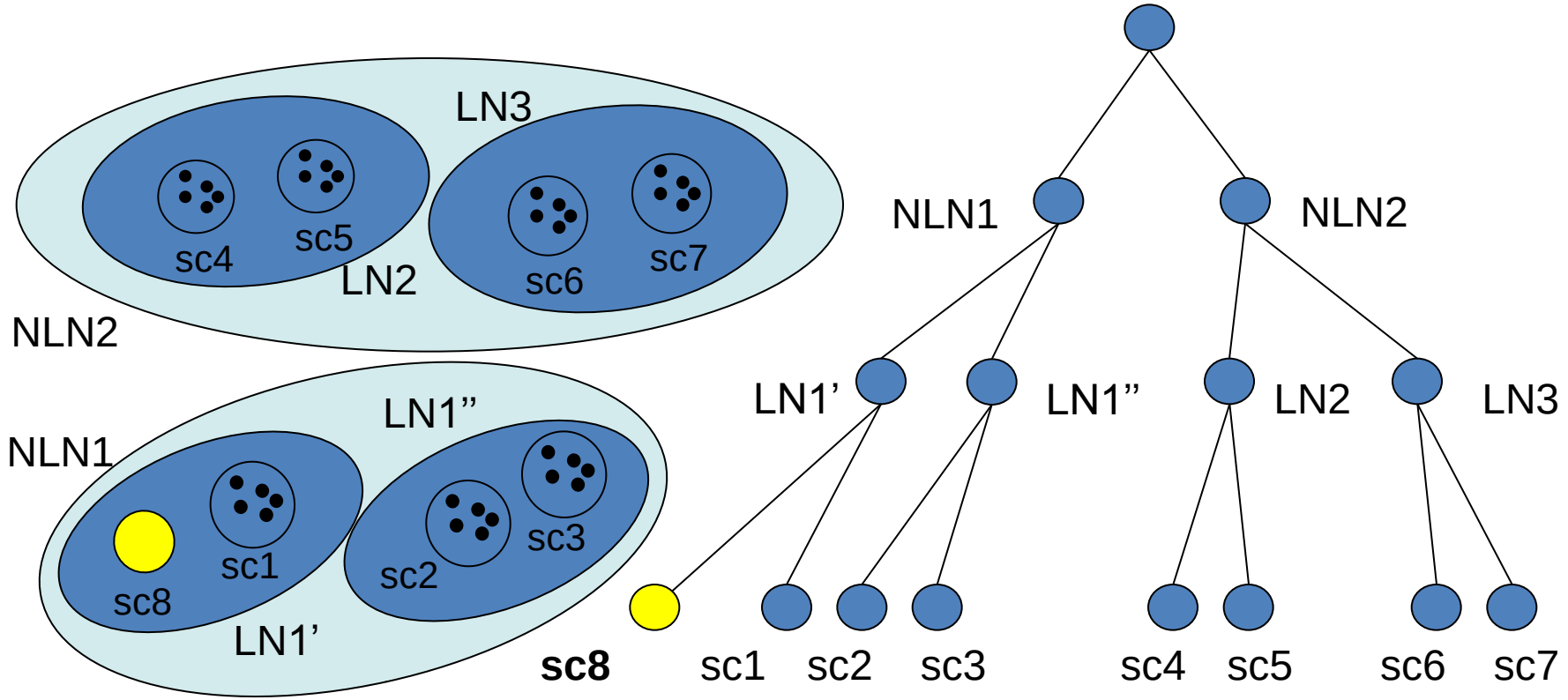
ΠΑΡΑΔΕΙΓΜΑΤΑ

Αν ο παράγοντας διακλάδωσης του φύλλου υπερβαίνει το 3, τότε το LN1 χωρίζεται.



ΠΑΡΑΔΕΙΓΜΑΤΑ

Αν ο παράγοντας διακλάδωσης ενός κόμβου υπερβαίνει το 3, τότε η ρίζα χωρίζεται και το ύψος του CF Δέντρου αυξάνεται κατά ένα.



Αλγόριθμος

- Φάση 1: Σάρωση όλων των δεδομένων και δημιουργία ενός αρχικού CF tree, χρησιμοποιώντας τη δεδομένη ποσότητα της μνήμης και το χώρο στο δίσκο.
- Φάση 2: Συμπύκνωση στο επιθυμητό μήκος δημιουργώντας ένα μικρότερο CF Tree
- Φάση 3: Global clustering.
- Φάση 4: Cluster refining – αυτό είναι προαιρετικό, και απαιτεί περισσότερες σαρώσεις στα δεδομένα για βελτίωση των αποτελεσμάτων.

Αλγόριθμος

Φάση 1 – Δημιουργία αρχικού CF tree

- Ξεκινά με ένα συγκεκριμένο κατώφλι, σκανάρει όλα τα δεδομένα και εισάγει τα σημεία στο δέντρο.
- Αν δεν υπάρχει αρκετή μνήμη, αυξάνουμε τη τιμή του κατωφλίου και κατασκευάζεται νέο μικρότερο CF tree εισάγοντας τα φύλλα του προηγούμενου στο μικρότερο.
- *Καλή τιμή κατωφλίου: σημαντικό αλλά δύσκολα προβλέψιμο.*
- Απομάκρυνση των outliers (όταν αναδομείται το δέντρο).

Αλγόριθμος

Φάση 2 – Συμπύκνωση δεδομένων

- Προετοιμασία για τη Φάση 3.
- Υπάρχει ένα κενό μεταξύ του μεγέθους της Φάσης 1 και της εισόδου της φάσης 3.
- Σαρώνει τα φύλλα στο αρχικό CF tree για να δομήσει το μικρότερο, ενώ αφαιρεί περισσότερα outliners και ομαδοποιεί τα subclusters σε μεγαλύτερα.

Αλγόριθμος

Φάση 3-Συσταδοποίηση

- Προβλήματα από τη Φάση 1:
  - Η σειρά εισόδου επηρεάζει αρνητικά το αποτέλεσμα.
  - Κάθε είσοδος περιορισμένο μέγεθος – πρόβλημα για συσταδοποίηση.
- Φάση 3:
  - Χρησιμοποιεί ένα αλγόριθμο συσταδοποίησης.
  - Κάθε φύλλο ως ξεχωριστό σημείο για τη συσταδοποίηση.

Αλγόριθμος

Φάση 4 – Προαιρετικά

- Πρόσθετο πέρασμα από τα δεδομένα για διόρθωση ανακριβειών και συσπειρώσεις clusters
- Χρησιμοποιεί τα centroids των clusters από τη Φάση 3 και συσταδοποιεί εκ νέου τα σημεία
- Συγκλίνει σε ελάχιστο (ανεξάρτητα από το πλήθος των επαναλήψεων).
- Απομακρύνει τα outliers που είναι μακριά από το centroid.

Πλεονεκτήματα

Ο Birch πιο γρήγορος από άλλους (KMEANS) σε μεγάλα σύνολα δεδομένων

Διαχειρίζεται τα outliers καλύτερα

Ανώτερος από τους άλλους ως προς σταθερότητα και επεκτασιμότητα

Μειονεκτήματα

- Κάθε κόμβος μπορεί να χωρέσει μόνο ένα συγκεκριμένο αριθμό από σημεία λόγω μεγέθους – περιορισμός στη φυσική έννοια του cluster.
- Αν τα clusters δεν έχουν σφαιρικό σχήμα, δεν αποδίδει καλά γιατί χρησιμοποιεί την ακτίνα και τη διάμετρο για να ελέγξει τα όρια του cluster.