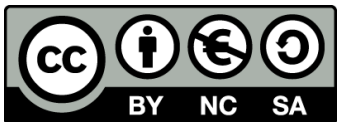
ΑΡΙΣΤΟΤΕΛΕΙΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΟΝΙΚΗΣ

ΑΝΟΙΚΤΑ
ΑΚΑΔΗΜΑΪΚΑ
ΜΑΘΗΜΑΤΑ

Αποθήκες Δεδομένων και Εξόρυξη Δεδομένων

Ενότητα 9: Ομαδοποίηση – Μέρος Γ΄

Αναστάσιος Γούναρης, Επίκουρος Καθηγητής
Τμήμα Πληροφορικής ΑΠΘ



Άδειες Χρήσης

- Το παρόν εκπαιδευτικό υλικό υπόκειται σε άδειες χρήσης Creative Commons.
- Για εκπαιδευτικό υλικό, όπως εικόνες, που υπόκειται σε άλλου τύπου άδειας χρήσης, η άδεια χρήσης αναφέρεται ρητώς.

Χρηματοδότηση

- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στα πλαίσια του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο «Ανοικτά Ακαδημαϊκά Μαθήματα στο Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης» έχει χρηματοδοτήσει μόνο την αναδιαμόρφωση του εκπαιδευτικού υλικού.
- Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.

Ομαδοποίηση – Μέρος Γ΄

Αλγόριθμοι βασισμένοι στην πυκνότητα, αλγόριθμοι βασισμένοι στη θεωρία γράφων

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Περιεχόμενα ενότητας

1. Αλγόριθμοι βασισμένοι στην πυκνότητα.
2. Αλγόριθμοι βασισμένοι στην θεωρία γράφων.

Σκοποί ενότητας

- Παρουσίαση αλγορίθμων που βασίζονται στην πυκνότητα, όπως ο DBScan.
- Παρουσίαση αλγορίθμων που βασίζονται στη θεωρία γράφων, όπως ο MST.
- Μελέτη ζητημάτων κλιμάκωσης στην ομαδοποίηση.

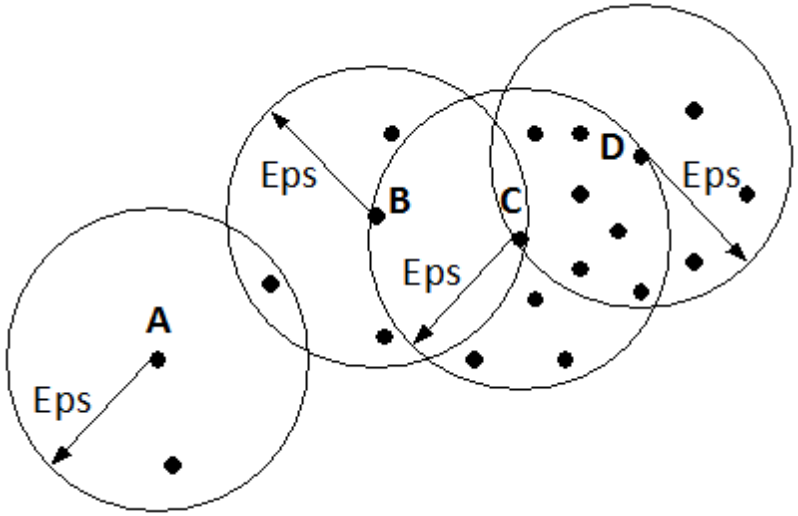
DBScan (1/2)

- Κατάλληλος για ομάδες που έχουν υψηλή πυκνότητα σημείων, οι οποίες μπορεί να είναι διαχωρισμένες από άλλα σημεία (θόρυβος) χαμηλότερης πυκνότητας.
- Προϋποθέτει ότι η πυκνότητα των ομάδων είναι παρόμοια, χωρίς μεγάλες διακυμάνσεις.
- πυκνότητα = #σημείων σε ακτίνα Eps

DBScan (2/2)

- **Κεντρικό σημείο**: έχει πυκνότητα μεγαλύτερη ή ίση από μία τιμή MinPts (ανήκουν στο εσωτερικό των ομάδων).
- **Συνοριακό σημείο**: έχει πυκνότητα μικρότερη από MinPts, αλλά απέχει από ένα κεντρικό σημείο απόσταση μικρότερη ή ίση από Eps (βρίσκονται στα όρια των ομάδων).
- **Θορυβώδες σημείο**: κάθε άλλο σημείο (ανήκουν στις περιοχές χαμηλής πυκνότητας).

- MinPts = 10
- C,D κεντρικά.
- B συνοριακό.
- A θορυβώδες.



DBSCAN: Αλγόριθμος

1. Χαρακτήρισε κάθε σημείο ως κεντρικό, συνοριακό ή θόρυβο.
2. Αγνόησε όλα τα σημεία θορύβου.
3. Δημιούργησε ένα γράφο με μια κορυφή για κάθε σημείο.
4. Τοποθέτησε μια ακμή μεταξύ όλων των κεντρικών σημείων που είναι σε απόσταση έως *Eps* μεταξύ τους.
5. Θέσε κάθε ομάδα συνδεδεμένων κεντρικών σημείων ως μια διαφορετική συστάδα.
6. Ανάθεσε κάθε συνοριακό σημείο σε μία από τις συστάδες που περιέχει το πιο κοντινό του κεντρικό σημείο.

Επιλογή Eps, MinPts

- Για κάθε σημείο, βρίσκουμε το *k* πλησιέστερο προς αυτό, καθώς και τη μεταξύ τους απόσταση.
- Ταξινομούμε τα σημεία ως προς την απόστασή τους από το *k*-οστό πλησιέστερο τους.
- Καθορίζουμε τις τιμές των Eps και MinPts έτσι ώστε να διαχωρίζονται τα σημεία που ανήκουν σε ομάδες από τα θορυβώδη σημεία.

Χαρακτηριστικά

- Δεν επηρεάζεται από το θόρυβο.
- Μπορεί να χειριστεί συστάδες με διαφορετικά σχήματα και μεγέθη.
- Αρνητικά:
  - Πρόβλημα με διαφορετικές πυκνότητες.
  - Ευαισθησία στις παραμέτρους.
  - Πολυ-διάστατα δεδομένα: δύσκολος ορισμός πυκνότητας και δαπανηρός υπολογισμός γειτόνων.
- Πολυπλοκότητα:
  - 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 (κρατάμε μόνο ένα label).

Σύγκριση με k-means (1/2)

- Και οι 2 τοποθετούν ένα σημείο σε μία μοναδική ομάδα. Αλλά ο dbscan μπορεί να μην ομαδοποιήσει όλα τα σημεία.
- Ο k-means χρησιμοποιεί την έννοια του κέντρου της ομάδας ενώ ο dbscan της πυκνότητας.
- Ο dbscan χειρίζεται σωστά ομάδες με διαφορετικό σχήμα και μέγεθος, σε αντίθεση με τον k-means.
    - Και οι 2 αλγόριθμοι δεν μπορούν να χειριστούν ομάδες με διαφορετική πυκνότητα.
- Ο k-means απαιτεί να μπορεί να οριστεί το κέντρο της ομάδας, ενώ ο dbscan απαιτεί να έχει νόημα η έννοια της πυκνότητας.
- Ο k-means έχει καλή απόδοση σε αραιά, πολυδιάστατα δεδομένα (πχ., κείμενα). Ο dbscan δεν έχει καλή απόδοση σε αυτήν την περίπτωση.

Σύγκριση με k-means (2/2)

- Και οι 2 μπορούν (με επεκτάσεις) να χειριστούν δεδομένα άλλων τύπων εκτός από αριθμητικά.
- Και οι 2 λαμβάνουν υπόψη όλα τα χαρακτηριστικά, δηλ. δεν δημιουργούν ομάδες βάσει μόνο κάποιων χαρακτηριστικών.
- Ο dbscan συνενώνει ομάδες που εφάπτονται ή επικαλύπτονται, σε αντίθεση με τον k-means.
- Η πολυπλοκότητα του k-means είναι μικρότερη από αυτή του dbscan.
- Ο k-means παράγει διαφορετικές συστάδες για τα ίδια δεδομένα, σε αντίθεση με τον dbscan.
- Ο dbscan δεν απαιτεί τον εκ των προτέρων ορισμό του αριθμού των ομάδων όπως κάνει ο k-means. Απαιτεί όμως τον ορισμό παραμέτρων όπως MinPts/Eps.
- Ο k-means μπορεί να εκφραστεί ως (NP-hard) πρόβλημα βελτιστοποίησης καθώς προσπαθεί να ελαχιστοποιήσει το SSE. Ο dbscan δεν μπορεί να εκφραστεί με ένα θεωρητικό μοντέλο.

Αλγόριθμοι γράφων: MST

1. Αντιστοίχισε κάθε αντικείμενο με μία κορυφή και την απόσταση μεταξύ δύο αντικειμένων με μία ακμή ίσου βάρους.
2. Βρες το ελάχιστο ζευγνύον δένδρο (MST) του γράφου που δημιουργείται στο βήμα 1.
3. Εντόπισε και αφαίρεσε τις «ασυνεπείς» ακμές.
4. Βρες τις συνδεδεμένες συνιστώσες του γράφου που προκύπτει, δημιούργησε για κάθε συνιστώσα μία ομάδα, και ανάθεσέ της τα αντικείμενα των αντίστοιχων κορυφών της.

Ασυνεπείς ακμές

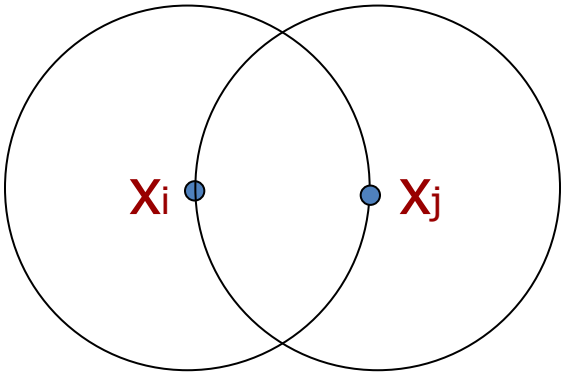
- Ο ορισμών των ασυνεπών ακμών μπορεί να γίνει με διάφορους τρόπους.
- Ο απλούστερος τρόπος είναι:
  - να εξετάσουμε την κατανομή των βαρών των ακμών του MST, και
  - να θεωρήσουμε ασυνεπείς αυτές με βάρος που απέχει από τη μέση τιμή του βάρους ποσότητα ίση με k φορές την τυπική απόκλιση της κατανομής των βαρών,
  - όπου k μικρή σταθερά.

Αλγόριθμοι γράφων: RNG (relative neighborhood graph)

- Ακμή αν:

$$d^2(x_i, x_j) \leq d^2(x_i, x_k) + d^2(x_j, x_k)\}, \forall k \neq i, j$$

- Γεωμετρικά, υπάρχει ακμή μεταξύ των αντικειμένων xi και xj, αν δεν υπάρχει κάποιο άλλο αντικείμενο εντός της τομής των δύο σφαιρών με κέντρα τα xi και xj, και ακτίνα ίση με d(xi , xj).

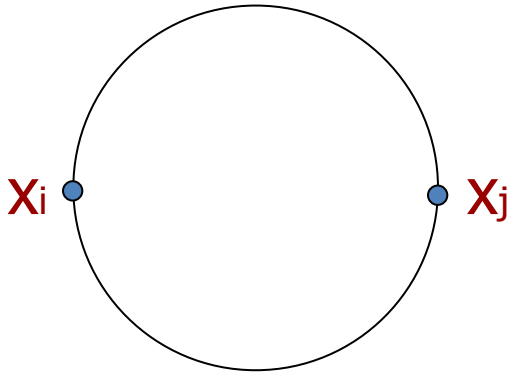


Αλγόριθμοι γράφων: GG (gabriel graph)

- Ακμή αν:

$$d^2(x_i, x_j) \leq d^2(x_i, x_k) + d^2(x_j, x_k)\}, \forall k \neq i, j$$

- Γεωμετρικά, υπάρχει ακμή μεταξύ των αντικειμένων xi και xj, αν δεν υπάρχει κάποιο άλλο αντικείμενο εντός της τομής των δύο σφαιρών με κέντρα τα xi και xj, και ακτίνα ίση με d(xi , xj).



Σύγκριση

- Αν Ε είναι το σύνολο ακμών ενός γράφου, τότε μπορεί να αποδειχθεί ότι:

$$E(MST) \subseteq E(RNG) \subseteq E(GG)$$

- Ενώ στο MST κάθε ζεύγος κορυφών συνδέεται με ένα μονοπάτι, στους RNG και GG, μπορεί να υπάρχουν περισσότερα από ένα μονοπάτια που να συνδέουν ένα ζεύγος αντικειμένων.
- Πολυπλοκότητα:
  - MST: $O(n^2)$. Μεγαλύτερη για RNG/GG.

BIRCH

- Balanced Iterative Reducing and Clustering using Hierarchies, by Zhang, Ramakrishnan, Livny (SIGMOD'96).
- Επιτυγχάνει μεγάλη κλιμάκωση καθώς είναι πολύ μικρότεροι οι χωροχρονικοί περιορισμοί.
- Χρησιμοποιεί μία δενδρική δομή CF-δένδρου που έχει ομοιότητες με το B^+ δένδρο.
- Αντί να κρατάμε όλα τα σημεία μιας συστάδας κρατάμε κάποια «στατιστικά» για κάθε συστάδα και για τις σχέσεις μεταξύ των συστάδων.

Χαρακτηριστικά συστάδας

- Έστω μια συστάδα σημείων: $\longrightarrow$
- Centroid(κεντρικό σημείο): $\longrightarrow$ $\longrightarrow$
- Radius (ακτίνα) - **μέση απόσταση των σημείων της συστάδας από το centroid**:

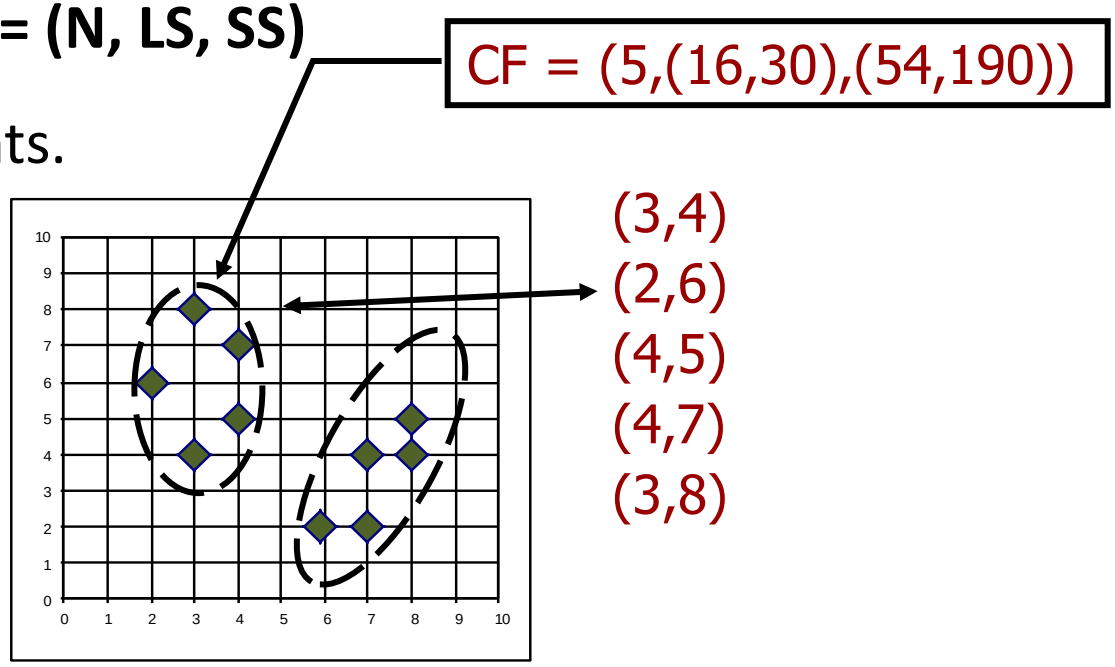
$$\longrightarrow \quad \longrightarrow \qquad -$$

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

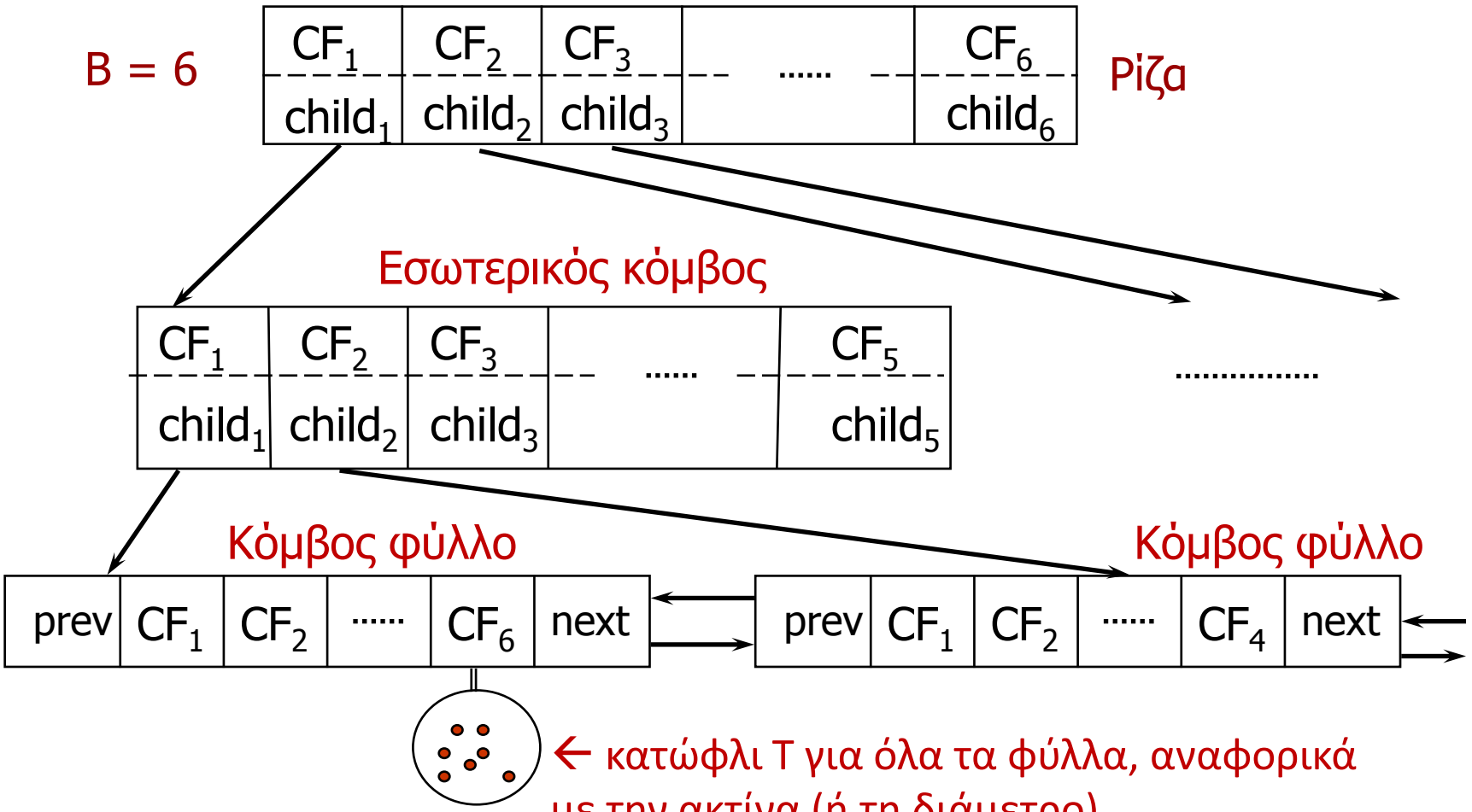
$$\longrightarrow \quad \longrightarrow \qquad -$$

Clustering Feature Vector

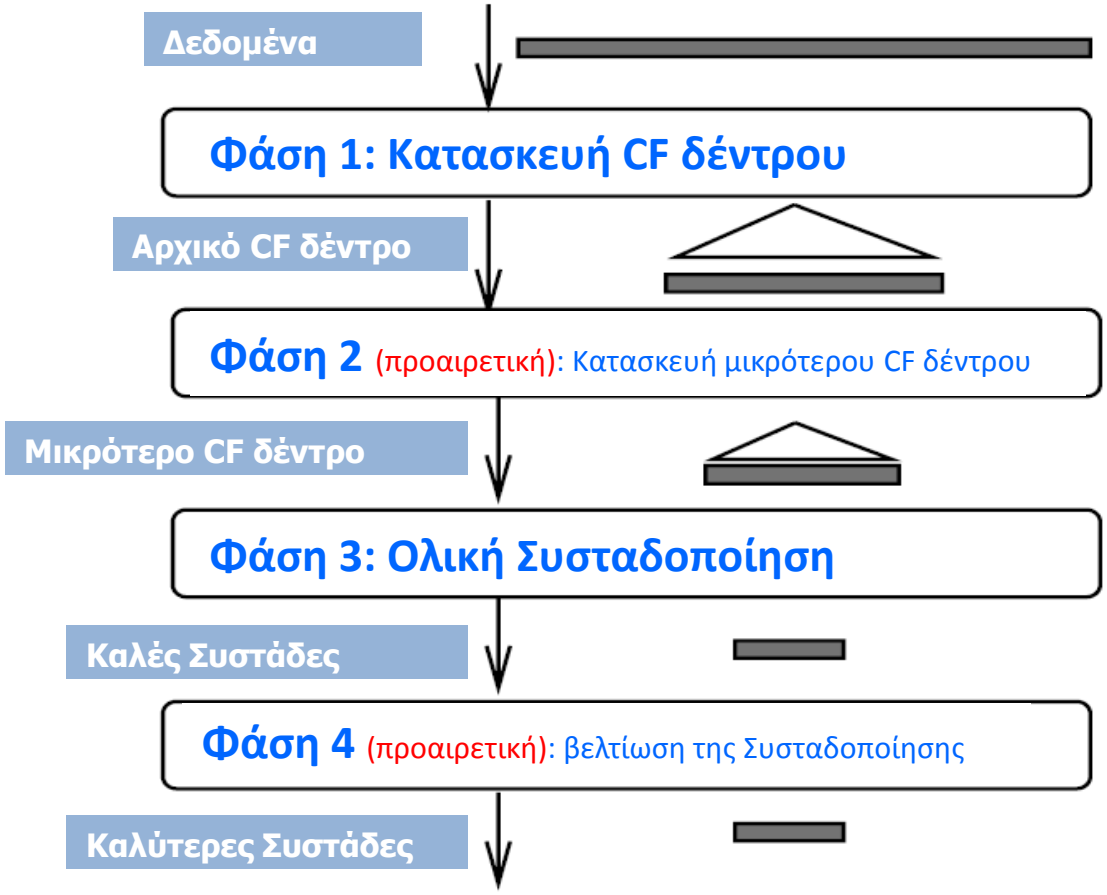
- Οι εγγραφές στο δένδρο είναι τριάδες, που περιέχουν αρκετή πληροφορία για τον υπολογισμό των (μετρικών) αποστάσεων μεταξύ συστάδων.
- **Clustering Feature: CF = (N, LS, SS)**
- N: Number of data points.
- LS:
- SS:



CF Tree


B = 6
CF1 CF2 CF3 ..... CF6
child1 child2 child3 child6
Ρίζα
Εσωτερικός κόμβος
CF1 CF2 CF3 ..... CF5
child1 child2 child3 child5
.....
Κόμβος φύλλο
Κόμβος φύλλο
prev CF1 CF2 ..... CF6 next
prev CF1 CF2 ..... CF4 next
← κατώφλι T για όλα τα φύλλα, αναφορικά με την ακτίνα (ή τη διάμετρο).

Φάσεις 1 - 2



- **Φάση 1:** Μια δομή κύριας μνήμης που συνοψίζει τα δεδομένα.
- Χτίζεται σε χρόνο O(n).

- **Φάση 2:** προσπαθεί να διώξει τους outliers και να ενοποιήσει «όμοιες» συστάδες που αντιστοιχούν σε περιοχές με πολλά σημεία (με αύξηση του *T*)).
- Χρειάζεται για να βελτιώσει τη Φάση 3.

Φάσεις 3 - 4

- **Φάση 3:**
  - Ξανα-συσταδοποιεί τα φύλλα του δέντρου, π.χ. κοντινές συστάδες που (έτυχε να) είναι σε διαφορετικά φύλλα.
  - 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.
- **Φάση 4:**
  - Ενοποίηση ομάδων που είναι σε διαφορετικά φύλλα.
  - Εξασφαλίζει όλα τα αντίγραφα να είναι σε ένα φύλλο.

Προβλήματα BIRCH

- Λόγω Β και σειράς εισαγωγής:
  - Διάσπαση σημείων της ίδιας ομάδας, επειδή ανήκουν σε διαφορετικά φύλλα.
  - Συγχώνευση σημείων διαφορετικών ομάδων, επειδή ανήκουν στο ίδιο φύλλο.
  - Εφαρμόζεται σε αριθμητικά δεδομένα.
  - Ευαισθησία στη σειρά των δεδομένων.

Χωρικά δεδομένα

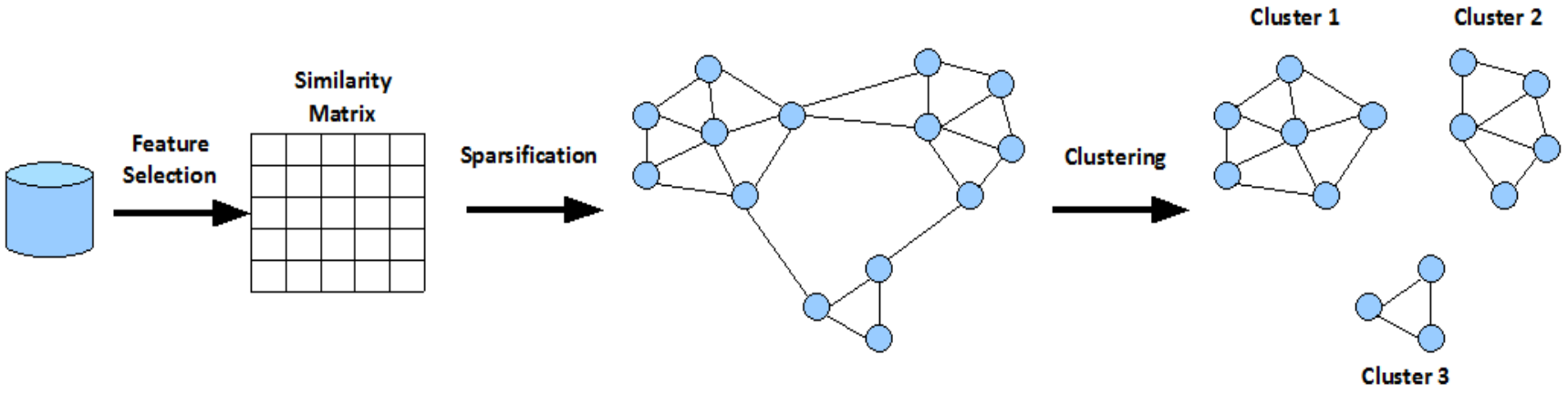
- Οι ομάδες ορίζονται ως πυκνές περιοχές στο χώρο.
- Έχουν οποιοδήποτε σχήμα, κατεύθυνση και μέγεθος.
- Διαφέρει η πυκνότητα και μέσα στις ομάδες αλλά και μέσα στην ίδια ομάδα.
- Θόρυβος.
- **Ο αλγόριθμος ομαδοποίησης πρέπει να μπορεί να αντιμετωπίζει όλα τα παραπάνω χωρίς να απαιτείται επίβλεψη (όσο γίνεται).**

CHAMELEON

- Πρόκειται για ένα συναθροιστικό ιεραρχικό αλγόριθμο που βασίζεται σε δυναμικό μοντέλο.
- Προεπεξεργασία:
  - Αναπαράσταση δεδομένων ως γράφο
    - Δημιουργούμε έναν γράφο k-πλησιέστερων γειτόνων που αναπαριστά τις συσχετίσεις ενός σημείου με τους k-πλησιέστερους γείτονες.
    - Η έννοια της γειτνίασης μεταβάλλεται δυναμικά.

Chameleon: Βήμα 1

- **Φάση 1**: αλγόριθμος graph partitioning → εύρεση ενός μεγάλου αριθμού ομάδων με πυκνά συνδεδεμένες κορυφές.
  - Κάθε μία από αυτές τις ομάδες περιμένουμε να περιέχει σημεία από την ίδια «πραγματική» ομάδα, δηλ. να είναι υποσύνολό της.
  - Ελάχιστο πλήθος ανά ομάδα: 1-5% του αρχικού συνόλου.



Chameleon: Βήμα 2

- **Φάση 2**: Χρήση συσσωρευτικού ιεραρχικού αλγορίθμου για να συνενώσει τις ομάδες της προηγούμενης φάσης.
  - Δύο ομάδες συνενώνονται αν η συστάδα που προκύπτει έχει κοινά χαρακτηριστικά με τις αρχικές συστάδες.

- Δύο βασικά χαρακτηριστικά για τον έλεγχο της ομοιότητας:
  - **Relative Interconnectivity-Σχετική συνοχή**: Απόλυτη συνοχή (άθροισμα βαρών των ακμών που συνδέουν τις δύο συστάδες) κανονικοποιημένη ως προς την εσωτερική συνοχή (άθροισμα των βαρών των ακμών που χωρίζουν μία συστάδα σε δυο σχεδόν ίσα μέρη).
  - **Relative Closeness-Σχετική εγγύτητα**: Απόλυτη εγγύτητα δύο συστάδων (μέσος όρος των βαρών των ακμών που συνδέουν τις δύο συστάδες) κανονικοποιημένη ως προς την εσωτερική εγγύτητα (μέσος όρος των βαρών των ακμών που ανήκουν στην τομή ελαχίστου κόστους μίας συστάδας).

Χαρακτηριστικά

- Μεγάλη αποτελεσματικότητα σε χωρικά δεδομένα με διαφορετικά σχήματα, μεγέθη και πυκνότητες.
- Στην αρχική τμηματοποίηση του γράφου, πρέπει τα σημεία όντως να ανήκουν στην ίδια ομάδα.
- Πολυπλοκότητα:
  - K-nn γράφος: O(mlogm) – $O(m^2)$
  - m σημεία σε p τμήματα γράφου: O(mlog(m/p))
  - Υπολογισμός συνοχής – εγγύτητας: O(mp)
  - Ιεραρχικός αλγόριθμος: $O(p^2 \log p)$

Σημείωμα Αναφοράς

Διαθέσιμο από τη δικτυακή διεύθυνση:http://eclass.auth.gr/courses/OCRS182/

Σημείωμα Αδειοδότησης

Τέλος ενότητας

Επεξεργασία: Ανδρέας Κοσματόπουλος

Θεσσαλονίκη, Χειμερινό Εξάμηνο 2013-2014

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Σημειώματα

Διατήρηση Σημειωμάτων

Οποιαδήποτε αναπαραγωγή ή διασκευή του υλικού θα πρέπει να συμπεριλαμβάνει:

- το Σημείωμα Αναφοράς
- το Σημείωμα Αδειοδότησης
- τη δήλωση Διατήρησης Σημειωμάτων
- το Σημείωμα Χρήσης Έργων Τρίτων (εφόσον υπάρχει)

μαζί με τους συνοδευόμενους υπερσυνδέσμους.