Πτυχιακή εργασία

«"ΕΦΑΡΜΟΓΗ ΤΗΣ ΤΕΧΝΙΚΗΣ ΣΥΣΤΑΔΟΠΟΙΗΣΗΣ ΔΕΔΟΜΕΝΩΝ ΓΙΑ ΤΗΝ ΕΠΙΤΑΧΥΝΣΗ ΤΗΣ ΑΝΑΖΗΤΗΣΗΣ ΟΜΟΙΟΤΗΤΑΣ ΑΠΟ ΜΕΓΑΛΕΣ ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ "»

ΘΕΣΣΑΛΟΝΙΚΗ 2011

Τι είναι το data mining ;

- "Η σύνθετη διαδικασία εξαγωγής συγκεκριμένης γνώσης, που προηγουμένως ήταν άγνωστη και δυνητικά ωφέλιμη, από δεδομένα ".

Τεχνικές Data Mining

- Clustering
- Classification
- Regression Analysis
- Association Rule Learning

Τι είναι classification και τι Clustering ;

classification

- είναι η διαδικασία η οποία απεικονίζει ένα σύνολο δεδομένων σε προκαθορισμένες ομάδες.

Clustering

- είναι μια μέθοδος ανάθεσης των στοιχείων ενός συνόλου σε υποσύνολα (συστάδες) έτσι ώστε οι συστάδες που θα δημιουργηθούν να είναι παρόμοιες ως προς κάποιο κριτήριο

Αλγόριθμοι Συσταδοποίησης

ιεραρχικοί

- Ήδη καθιερωμένες ομάδες
- Συγχώνευση συνόλου
- Συσσωρευτικοί (agglomerative)
- Διαχωριστικοί (divisive)

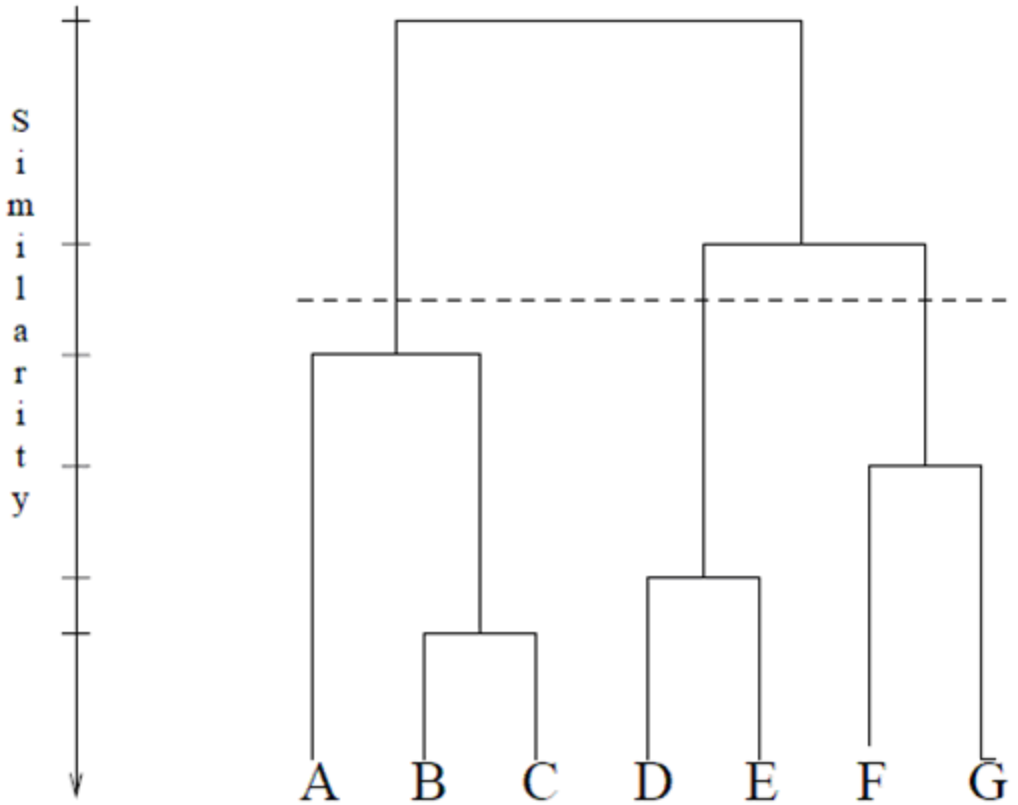
μη ιεραρχικοί

- Καθορισμός νέων ομάδων
- Διάσπαση συνόλου

Ιεραρχική Συσταδοποίηση

•Agglomerative (Συσσωρευτικοί):

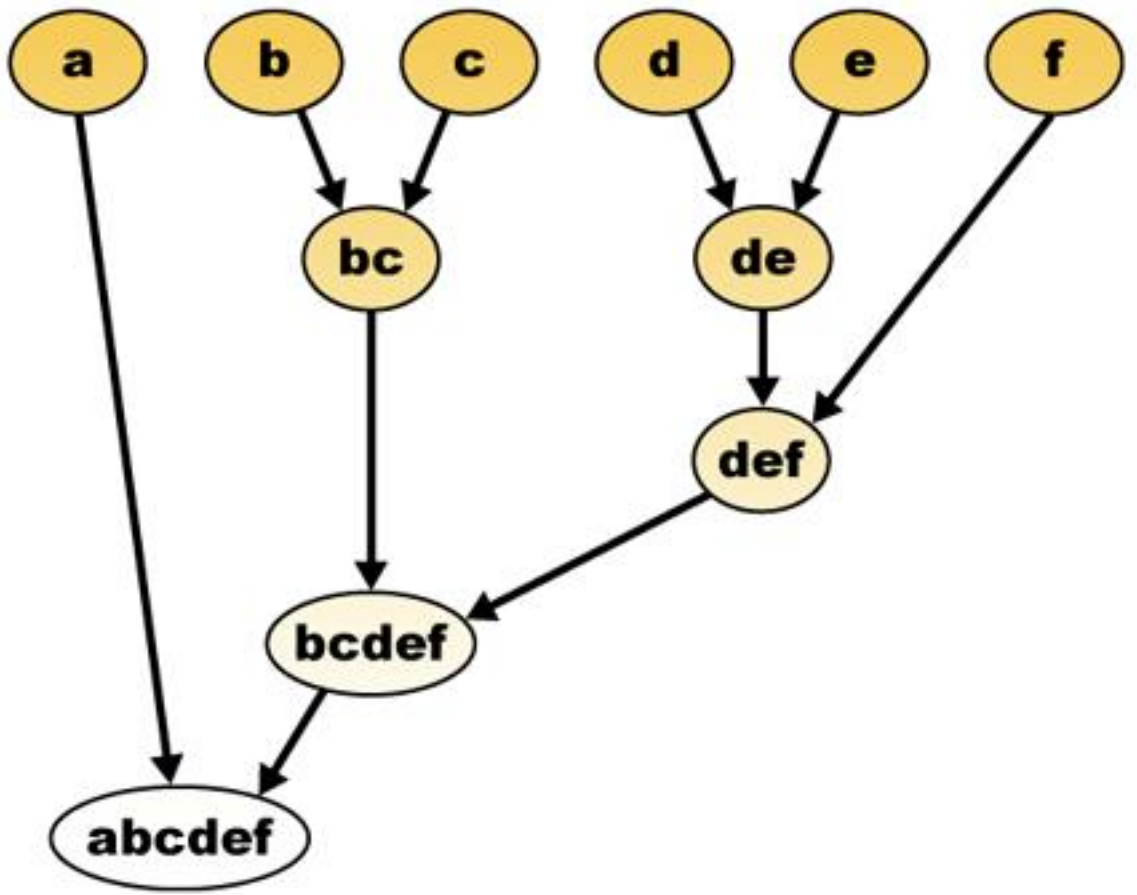
είναι μία προσέγγιση από κάτω προς τα πάνω



Ιεραρχική Συσταδοποίηση

- Divisive

(Διαχωριστικοί): είναι η προσέγγιση από πάνω προς τα κάτω



Παράδειγμα Hierarchical clustering(I)

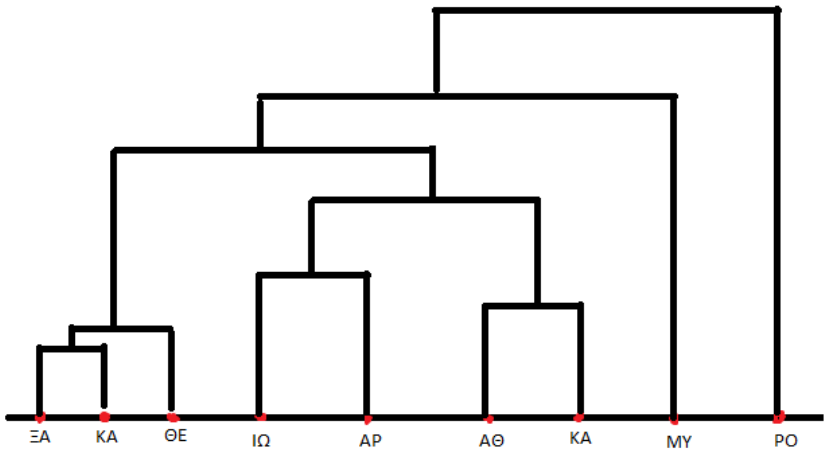
- Βλέπουμε έναν χάρτη της Ελλάδας στον οποίο έχουν τονιστεί με μια κόκκινη κουκίδα ορισμένες πόλεις (Καβάλα, Θεσσαλονίκη, Ξάνθη και άλλες).
- Αρχικά ενώνουμε με μια γραμμή τις δύο κοντινότερες πόλεις όπως εμφανίζονται στον χάρτη, (φυσικά ανάλογα με την χιλιομετρική τους απόσταση)

Παράδειγμα Hierarchical clustering(II)

- Οι πιο κοντινές πόλεις εμφανίζονται να είναι η Ξάνθη με την Καβάλα οπότε ενώνουμε αυτές πρώτα.
- Έτσι τώρα η Καβάλα και η Ξάνθη αποτελούν μαζί ένα καινούργιο cluster. Συνεχίζουμε ενώνοντας τις κοντινότερες πόλεις μεταξύ τους.

Ιεραρχικό δέντρο

- εφαρμογή του αλγόριθμου στις παραπάνω πόλεις τις Ελλάδας



Ο αλγόριθμος k-NN

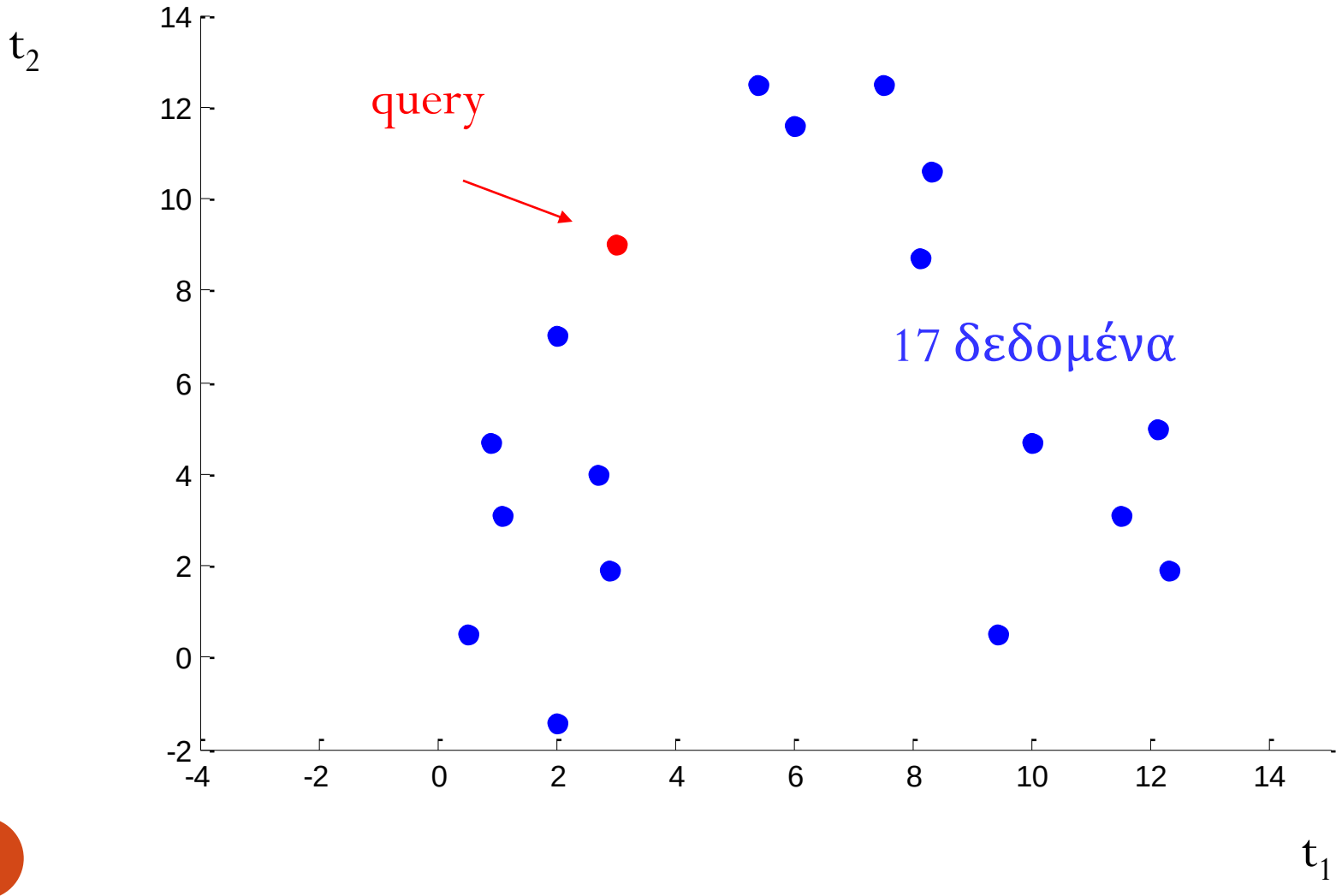
- Πιο συνηθισμένος αλγόριθμος ταξινόμησης
- Το σύνολο εκπαίδευσης περιλαμβάνει και τις κλάσεις
- Εξετάζουμε Κ αντικείμενα «κοντά» στο συγκεκριμένο που ταξινομούμε
- Το αντικείμενο τοποθετείται στην κλάση με το μεγαλύτερο αριθμό «κοντινών» αντικειμένων
- Πολυπλοκότητα $O(q)$: το μέγεθος του συνόλου εκπαίδευσης

Η προτεινόμενη μέθοδος CLUREP

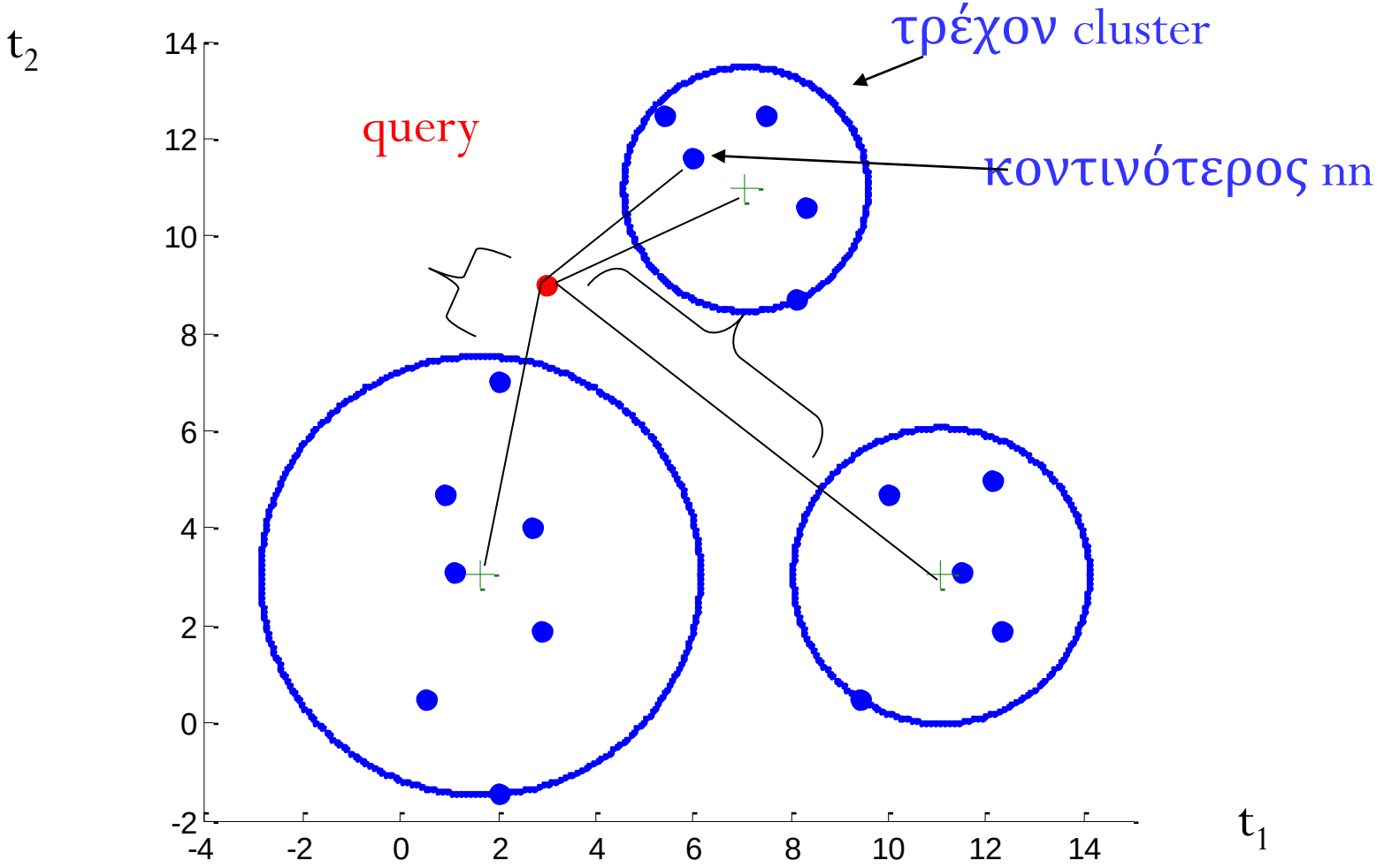
- Αντιπροσώπευση της μειώσεις της διαστατικότητας των δεδομένων(PAA)
- Συσταδοποίηση των δεδομένων με τη βοήθεια του K-means αλγόριθμου
- Διαδοχική αναζήτηση και περιορισμός του χώρου αναζήτησης με δύο τρόπους: (α) απόρριψη ενός αριθμού συστάδων (β) περιορισμό του χώρου αναζήτησης μέσα σε μια συστάδα

Η μέθοδος CLUREP μας εγγυάται ότι ο πλησιέστερος γείτονας, θα βρεθεί

I. Συσταδοποίηση (k-means)

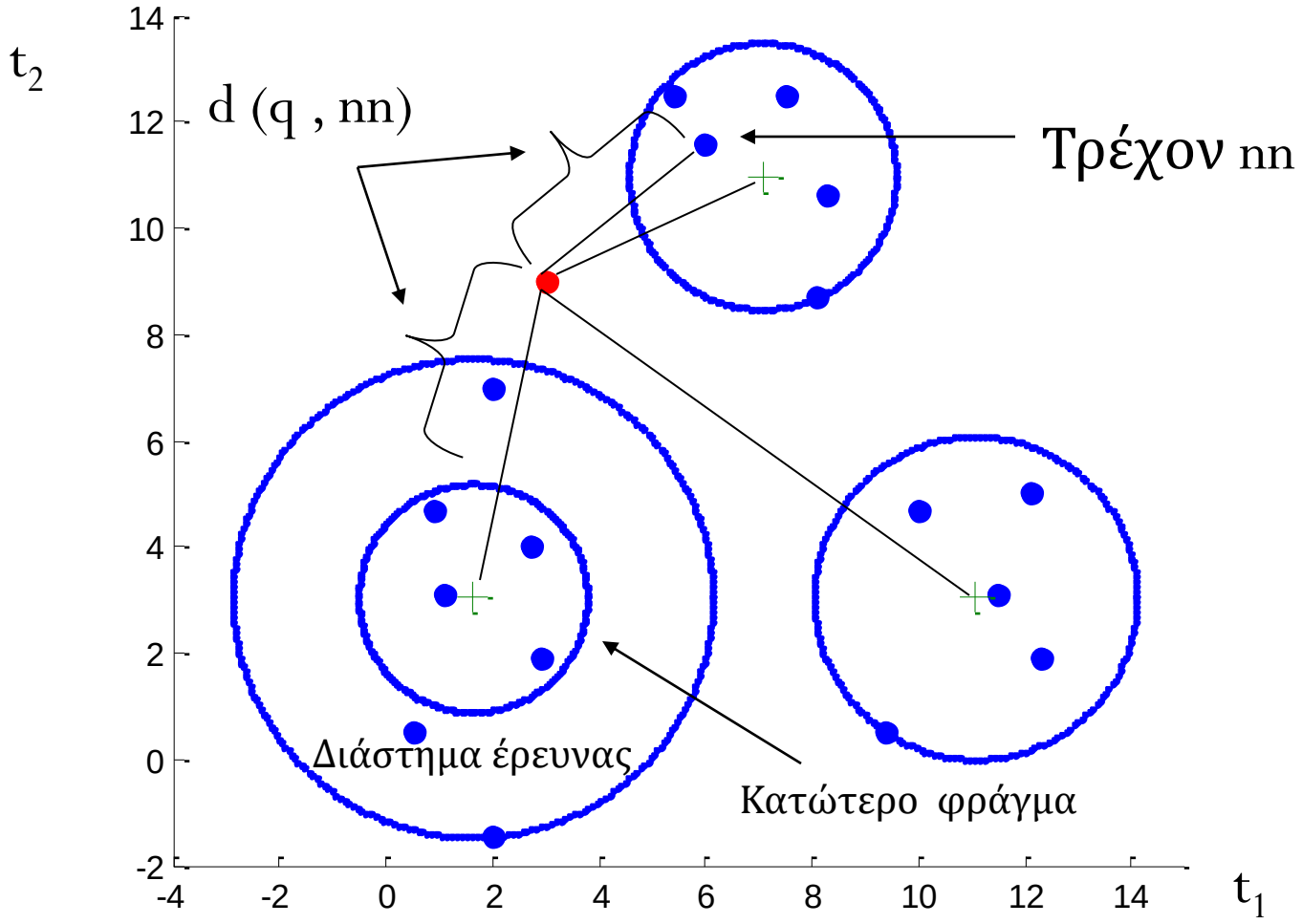


II. Αναζήτηση ομοιότητας(1)



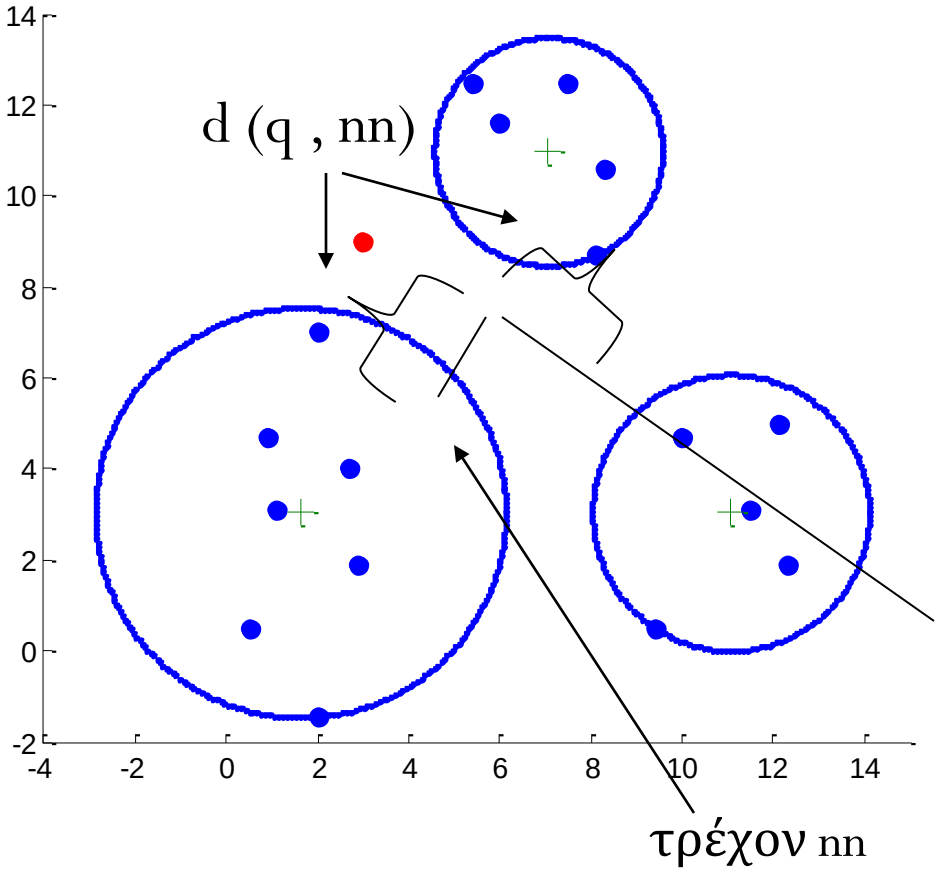
Καθορισμός κοντινότερης, συστάδας - βρείτε τον τρέχον πλησιέστερο γείτονα

II. Αναζήτηση ομοιότητας(2)



Καθορίσει το επόμενο κοντινότερο cluster - ελέγχει αν η αίτηση θα πρέπει να αναζητηθεί -Εάν ναι, να καθορίσει το χώρο αναζήτησης

II. Αναζήτηση ομοιότητας(3)



βρείτε τον τρέχον πλησιέστερο γείτονα - τον καθορισμό του επόμενου πλησιέστερου cluster - έλεγχος από το αν θα πρέπει να αναζητηθούν - εάν ναι, να καθορίσει το χώρο αναζήτησης.

Αποτελέσματα (I)

		Dimensionality						
ID	Datasets	4	6	8	10	12	14	16
1	LandSat-Satellite	89.38	85.62	84.20	86.81	82.06	82.06	82.06
2	CBF	10.54	0.43	0.00	0.11	0.00	0.00	0.00

1-NN classification error rates (%) (κ=10)

- Στο παραπάνω πείραμα κρατήσαμε σταθερό των αριθμό των συστάδων (k = 10) και αυξάνουμε σταδιακά ανά δυο την διαστατικότητα

Αποτελέσματα (II)

		Dimensionality						
ID	Dataset	4	6	8	10	12	14	16
1	LandSat-Satellite	13.87	16.14	16.73	18.14	18.99	19.02	18.96
2	CBF	12.12	13.31	13.88	14.73	16.21	15.57	17.18

1-NN classification average space (%) (k=10)

- όσο αυξάνεται η διαστατικότητα τόσο αυξάνεται και η τιμή του average space.

Αποτελέσματα (III)

		Number of clusters						
ID	Dataset	4	6	8	10	12	14	16
1	LandSat-Satellite	89.38	89.38	89.38	89.38	89.38	89.38	89.38
2	CBF	10.54	10.54	10.54	10.54	10.54	10.54	10.54

1-NN classification error rates (%) (i=4)

- Κρατάμε σταθερή την διαστατικότητα και αυξάνουμε σταδιακά τον αριθμό των clusters ανά 2.

Αποτελέσματα (IV)

		Number of Clusters						
ID	Dataset	4	6	8	10	12	14	16
1	LandSat-Satellite	28.67	20.33	15.25	12.36	10.31	8.75	7.90
2	CBF	27.13	20.05	16.15	13.88	12.25	10.96	9.95

1-NN classification average space (%) (i=4)

- όσο αυξάνεται ο αριθμός από τα cluster τόσο καλύτερα αποτελέσματα μας επιστρέφει κ το average space

Συμπεράσματα

- Η προτεινόμενη μέθοδος (CLUREP) είναι σαφώς ταχύτερη από ό, τι διαδοχική αναζήτηση στην πλειοψηφία των συνόλων των στοιχείων.
- Όταν ο αριθμός των συστάδων αυξάνεται, η αποτελεσματικότητα της προτεινόμενης μεθόδου βελτιώνεται , το ποσοστό της βελτίωσης μειώνεται όταν ο αριθμός των συστάδων που προέρχονται είναι 8 – 12.
- Όταν η διαστατικότητα των μετασχηματισμένων δεδομένων αυξάνεται, το ποσοστό του λάθους ταξινόμησης μειώνεται.

ΣΑΣ ΕΥΧΑΡΙΣΤΩ ΠΟΛΥ!!!!

ΜΑΥΡΕΑΣ ΓΕΩΡΓΙΟΣ 05/2856