



Αποθήκες Δεδομένων και Εξόρυξη Γνώσης (Data Warehousing & Data Mining)

Επανάληψη

Γιάννης Θεοδωρίδης, Νίκος Πελέκης

Εργαστήριο Πληροφοριακών Συστημάτων

<http://infolab.cs.unipi.gr>

14 Ιαν. '10



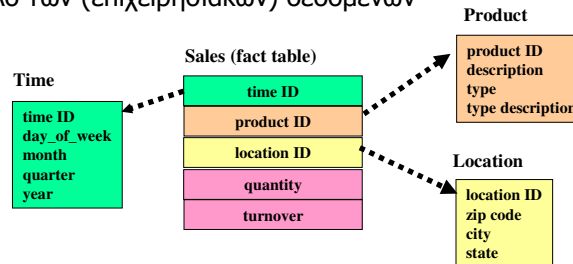
Αποθήκες Δεδομένων

05 Νοε. 08

Από τις ΒΔ στις ΑΔ



- Μια **αποθήκη δεδομένων** κρατά συνοπτική πληροφορία για κάποιο υποσύνολο των (επιχειρησιακών) δεδομένων

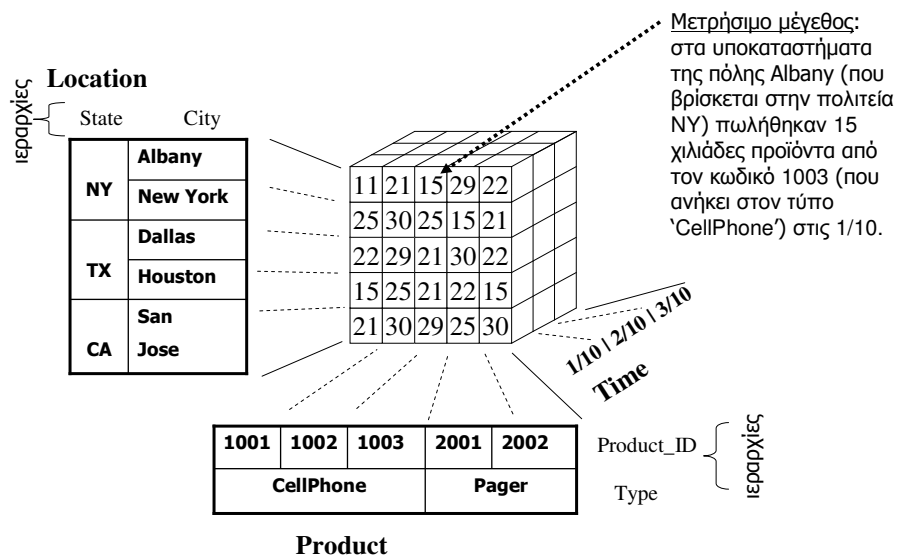


- Μια ΑΔ ορίζεται πάνω στα δεδομένα μιας (επιχειρησιακής) ΒΔ και η πλήρωσή της γίνεται με κατάλληλη διαδικασία 'φόρτωσης'
 - «διαστάσεις» (dimensions) της ΑΔ: οι πίνακες **Time**, **Product**, **Location**
 - «μετρήσιμα μεγέθη» (measures) της ΑΔ: οι συναθροιστικές τιμές **quantity**, **turnover**

3

ΠΑ.ΠΕΛ - Γιάννης Θεοδορίδης

Όψη κύβου δεδομένων



4

ΠΑ.ΠΕΛ - Γιάννης Θεοδορίδης

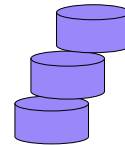
Παράδειγμα φόρτωσης ΑΔ



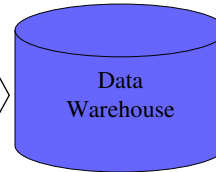
■ Παράδειγμα ΒΔ:

```
product (product_ID, description,
         type, type_description)
location (location_ID, zip code,
         city, state)
sales-transaction (timestamp,
                  product_ID, location_ID,
                  units_sold, unit_price)
```

Operational
DBs



Extract
Transform
Load
Refresh



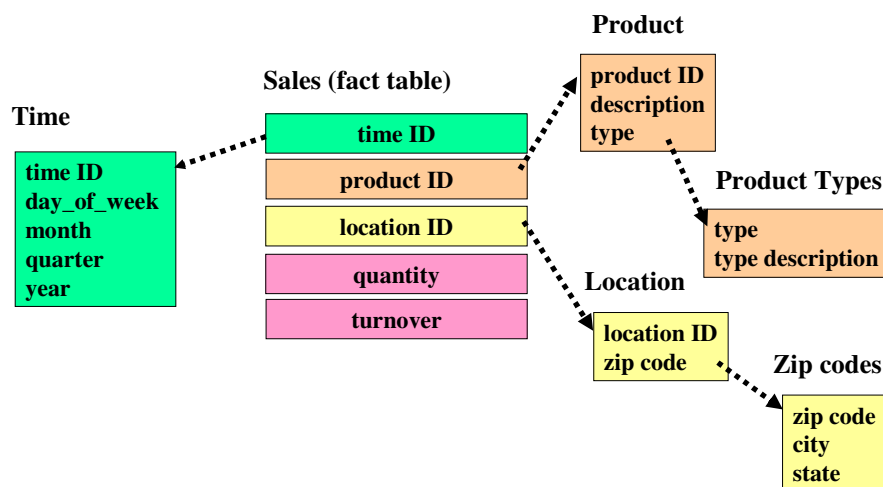
■ Παράδειγμα ETL:

```
INSERT INTO sales
( SELECT date(timestamp) AS time_id, product_ID, location_ID,
  sum(units_sold) AS quantity, sum(units_sold*unit_price) AS turnover
FROM sales-transaction
GROUP BY time_id, product_ID, location_ID HAVING quantity > 0 )
```

■ Παράδειγμα ΑΔ:

```
product (product_ID, description, type, type_description)
location (location_ID, zip code, state, city)
time (time_ID, day_of_week, month, quarter, year)
sales (time_ID, product_ID, location_ID, quantity, turnover)
```

Σχήμα χιονονιφάδας



Βασικές λειτουργίες OLAP



■ Συσώρευση (roll-up):

- άνοδος στο lattice των κυβοειδών (με αφαίρεση διάστασης) ή
- άνοδος στην ιεραρχία μιας διάστασης

■ Εμβάθυνση (drill-down):

το αντίστροφο της λειτουργίας roll-up

Quantity (in 1000's)		2008			
		Q1	Q2	Q3	Q4
TX	Dallas	78	98	34	56
	Houston	90	94	87	91

Roll Up
Dimension: time
(month → quarter)

Quantity (in 1000's)		Q1			Q2			...
		Jan	Feb	Mar	Apr	May	Jun	...
TX	Dallas	30	26	22	20	36	42	...
	Houston	28	30	32	38	30	26	...

7

ΠΑ.ΠΕΛ - Γιάννης Θεοδορίδης

Βασικές λειτουργίες OLAP (συν.)



■ Τεμαχισμός (slice) και

Κομμάτισμα (dice):

εξέταση υποσυνόλου του κύβου με επιβολή φίλτρων σε κάποιες διαστάσεις

Quantity (in 1000's)		2008 (any product)			
		Q1	Q2	Q3	Q4
TX	Dallas	78	45	34	56
	Houston	90	67	87	91

Slice
product.type = 'CellPhone'

Quantity (in 1000's)		2008 (CellPhones only)			
		Q1	Q2	Q3	Q4
TX	Dallas	53	35	20	48
	Houston	76	57	40	80

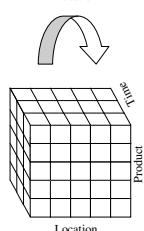
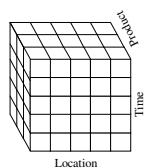
8

ΠΑ.ΠΕΛ - Γιάννης Θεοδορίδης

Βασικές λειτουργίες OLAP (συν.)



- **Περιστροφή (pivot):** εμφάνιση κύβου από διαφορετική οπτική γωνία



Quantity (in 1000's)		CellPhone			Pager	
		1001	1002	1003	2001	2002
TX	Dallas	20	33	12	8	12
	Houston	32	45	34	20	23

Pivot

Quantity (in 1000's)		2008 (any product)			
		Q1	Q2	Q3	Q4
TX	Dallas	78	45	34	56
	Houston	90	67	87	91

9

Π.Α.ΠΕΛ. – Γιάννης Θεοδορίδης

Θέμα 1



- Θεωρήστε την Αποθήκη Δεδομένων (ΑΔ) της διοργανώτριας αρχής του πρωταθλήματος καλαθοσφαίρισης στην οποία καταγράφονται στατιστικά στοιχεία για τους καλαθοσφαιριστές που αγωνίζονται. Ενδεικτικοί άξονες ανάλυσης είναι προσωπικά στοιχεία των αθλητών (η ηλικία, το ύψος, ο τόπος προέλευσης κλπ.) και αθλητικά στοιχεία (όπως οι θέσεις στις οποίες αγωνίζεται, τα λεπτά συμμετοχής σε κάθε θέση, οι πόντοι που έχει σημειώσει, τα κοψίματα, τα rebounds κλπ.).
 - (α) Σχεδιάστε ένα ενδεικτικό σχήμα αστέρα που να μοντελοποιεί την παραπάνω ΑΔ.
 - (β) Με βάση το παραπάνω σχήμα, δώστε από ένα παράδειγμα για τις λειτουργίες Rollup και Slice και τη συγκεκριμένη ανάλυση που εκτελούν.

10

Π.Α.ΠΕΛ. – Γιάννης Θεοδορίδης



Κατηγοριοποίηση (classification)

2 Δεκ. 08

Παράδειγμα: πρόγνωση καιρού

dataset

πιθανότητες εμφάνισης

Outlook	Temp.	Humidity	Windy	Play?
sunny	hot	high	false	No
sunny	hot	high	true	No
overcast	hot	high	false	Yes
rain	mild	high	false	Yes
rain	cool	normal	false	Yes
rain	cool	normal	true	No
overcast	cool	normal	true	Yes
sunny	mild	high	false	No
sunny	cool	normal	false	Yes
rain	mild	normal	false	Yes
sunny	mild	normal	true	Yes
overcast	mild	high	true	Yes
overcast	hot	normal	false	Yes

Outlook			Temperature			Humidity			Windy			Play		No
Yes No			Yes No			Yes No			Yes No			Yes	No	
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5	απορίθμης
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3			
Rainy	3	2	Cool	3	1									
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14	
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5			
Rainy	3/9	2/5	Cool	3/9	1/5									

Παράδειγμα: πρόγνωση καιρού (συν.)



Outlook			Temperature			Humidity			Windy			Play	
Yes No			Yes No			Yes No			Yes No			Yes No	
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9 5	
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

- Μια νέα ημέρα:

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

Πιθανοφάνειες για τις δύο κλάσεις

"yes" : $2/9 \times 3/9 \times 3/9 \times 3/9 \times 9/14 = 0.0053$

"no" : $3/5 \times 1/5 \times 4/5 \times 3/5 \times 5/14 = 0.0206$

Πιθανότητες (μετά την κανονικοποίηση):

$P(\text{"yes"}) = 0.0053 / (0.0053 + 0.0206) = 0.205$

$P(\text{"no"}) = 0.0206 / (0.0053 + 0.0206) = 0.795$

Θεοδωρίδης

Ο κανόνας του Bayes



- Η πιθανότητα να συμβεί ένα γεγονός H δοθείσης μιας μαρτυρίας E :

$$\Pr[H | E] = \frac{\Pr[E | H] \Pr[H]}{\Pr[E]}$$

- $A \text{ priori}$ πιθανότητα του H : $\Pr[H]$
 - Η πιθανότητα του γεγονότος *χωρίς* την επίκληση της μαρτυρίας
- $A \text{ posteriori}$ πιθανότητα του H : $\Pr[H | E]$
 - Η πιθανότητα του γεγονότος *με* την επίκληση της μαρτυρίας

Κατηγοριοποίηση Naïve Bayes



- Εκμάθηση κατηγοριοποίησης: ποια η πιθανότητα μιας κλάσης δοθείσης μιας μαρτυρίας;
 - Η μαρτυρία E είναι η εγγραφή στη ΒΔ
 - Το γεγονός H είναι η κλάση της εγγραφής
- Απλοϊκή (naïve) παραδοχή: η μαρτυρία διαιρείται σε μέρη (όσο και τα γνωρίσματα) που είναι ανεξάρτητα μεταξύ τους

$$\Pr[H | E] = \frac{\Pr[E_1 | H] \Pr[E_2 | H] \dots \Pr[E_n | H] \Pr[H]}{\Pr[E]}$$

15

ΠΑ.ΠΕΛ - Γιάννης Θεοδορίδης

Παράδειγμα πρόγνωσης καιρού



Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

μαρτυρία E

$$\Pr[\text{yes} | E] = \Pr[\text{Outlook} = \text{Sunny} | \text{yes}]$$

$$\times \Pr[\text{Temperature} = \text{Cool} | \text{yes}]$$

$$\times \Pr[\text{Humidity} = \text{High} | \text{yes}]$$

$$\times \Pr[\text{Windy} = \text{True} | \text{yes}]$$

$$\times \frac{\Pr[\text{yes}]}{\Pr[E]} = \frac{2/9 \times 3/9 \times 3/9 \times 3/9 \times 9/14}{0/14}$$

Πιθανότητα να συμβεί ένα γεγονός (δηλ. $\text{Play} = \text{"yes"}$) δοθείσης της μαρτυρίας E (δηλ. $\text{Outlook} = \text{"Sunny"}$ and $\text{Temp.} = \text{"Cool"}$ and $\text{Humidity} = \text{"High"}$ and $\text{Windy} = \text{"True"}$)

16

ΠΑ.ΠΕΛ - Γιάννης Θεοδορίδης

Αλγόριθμος ID3



- Δημιουργεί ΔΑ με χρήση στοιχείων από τη θεωρία πληροφορίας (εντροπία)
- Επιλέγει για διάσπαση το γνώρισμα με το μεγαλύτερο κέρδος πληροφορίας (information gain):

$$Gain(D, S) = H(D) - \sum_{i=1}^s P(D_i) H(D_i)$$

- $H(D)$ η εντροπία του D (πριν το διαχωρισμό)
- $H(D_i)$ η εντροπία των επιμέρους D_i (μετά το διαχωρισμό)
- Όσο μεγαλύτερη είναι η μείωση (το «άλμα» προς το 0), τόσο μεγαλύτερο είναι το κέρδος $Gain(D, S)$

17

ΠΑ.ΠΕΛ - Γιάννης Θεοδωρίδης

Παράδειγμα ID3

$$H(p_1, p_2, \dots, p_s) = - \sum_{i=1}^s \left(p_i \log \left(\frac{1}{p_i} \right) \right)$$



- Αρχική κατάσταση εντροπίας:
 $H(D) = 4/15 \log(15/4) + 8/15 \log(15/8) + 3/15 \log(15/3) = \mathbf{0.4384}$
- Κέρδος αν γίνει διάσπαση στο **gender**:
 - Gender='F':
 $H(D_1) = 3/9 \log(9/3) + 6/9 \log(9/6) = 0.2764$
 - Gender='M':
 $H(D_2) = 1/6 \log(6/1) + 2/6 \log(6/2) + 3/6 \log(6/3) = 0.4392$
 - Weighted sum = $(9/15)(0.2764) + (6/15)(0.4392) = 0.3415$
 - Gain = $0.4384 - 0.3415 = \mathbf{0.0969}$
- Κέρδος αν γίνει διάσπαση στο **height**:
 - Τμηματοποίηση: $(0, 1.6], (1.6, 1.7], (1.7, 1.8], (1.8, 1.9], (1.9, 2.0], (2.0, \infty)$
 - Weighted sum = $0 + 0 + \dots + (2/15)(0.301) = 0.0401$
 - Gain: $0.4384 - 0.0401 = \mathbf{0.3983} \rightarrow$ επιλέγουμε να γίνει διάσπαση ως προς **height**

Name	Gender	Height	Output1
Kristina	F	1.6m	Short
Jim	M	2m	Tall
Maggie	F	1.9m	Medium
Martha	F	1.88m	Medium
Stephanie	F	1.7m	Short
Bob	M	1.85m	Medium
Kathy	F	1.6m	Short
Dave	M	1.7m	Short
Worth	M	2.2m	Tall
Steven	M	2.1m	Tall
Debbie	F	1.8m	Medium
Todd	M	1.95m	Medium
Kim	F	1.9m	Medium
Amy	F	1.8m	Medium
Wynette	F	1.75m	Medium

18

ΠΑ.ΠΕΛ - Γιάννης Θεοδωρίδης

Αλγόριθμος C4.5



- Ο αλγόριθμος ID3 μεροληπτεί υπέρ των γνωρισμάτων με μεγάλο αριθμό διαιρέσεων
- Ο αλγόριθμος C4.5 αποτελεί βελτιωμένη έκδοχή του ID3. Συνάρτηση καταλληλότητας:

$$GainRatio(D, S) = \frac{Gain(D, S)}{H\left(\frac{|D_1|}{|D|}, \dots, \frac{|D_s|}{|D|}\right)}$$

- Πλεονεκτήματα:
 - Καλύτερη διαχείριση ελλিপών / συνεχών δεδομένων
 - Αποφυγή υπερπροσαρμογής

Θέμα



- Έστω το διπλανό σύνολο δεδομένων. Να κατασκευάσετε δέντρο απόφασης χρησιμοποιώντας τον αλγόριθμο ID3 και θεωρώντας ως σύνολο δεδομένων εκπαίδευσης (training set) τις δοσοληψίες t1-t8 και στη συνέχεια να βρείτε την απόδοση του δένδρου που κατασκευάσατε θεωρώντας ως σύνολο δεδομένων ελέγχου (testing set) τις δοσοληψίες t9-t11.
- (δίδονται οι λογάριθμοι με βάση το 2 που ίσως σας χρειαστούν:
 $\log(5/2)=1.325$, $\log(5/3)=0.735$,
 $\log(4/3)=0.42$, $\log 3=1.585$,
 $\log(3/2)=0.59$)

Trans.	A1	A2	Class
t ₁	Y	N	+
t ₂	N	Y	-
t ₃	N	N	+
t ₄	Y	Y	+
t ₅	N	N	+
t ₆	Y	N	-
t ₇	N	Y	-
t ₈	N	Y	-
t ₉	Y	Y	+
t ₁₀	N	Y	-
t ₁₁	N	N	+



Συσταδοποίηση (clustering)

2 Δεκ. 08

Συσταδοποίηση K-Means



- Δεδομένης μίας συστάδας $K_i = \{t_{i1}, t_{i2}, \dots, t_{im}\}$, ο **μέσος της συστάδας** είναι $m_i = (1/m)(t_{i1} + \dots + t_{im})$
 - Ο μέσος της συστάδας ταυτίζεται με το **κέντρο βάρους** (centroid)
- Βασική ιδέα:
 - Τυχαία επιλέγεται το αρχικό σύνολο των μέσων των συστάδων
 - Επαναληπτικά, τα στοιχεία μετακινούνται μεταξύ των συστάδων μέχρι να επιτύχουμε ισορροπία
- Μειονεκτήματα:
 - ο αριθμός k του πλήθους των συστάδων είναι προκαθορισμένος
 - η ποιότητα του αποτελέσματος επηρεάζεται από την αρχική επιλογή των μέσων

Παράδειγμα K-Means (σε 1 διάσταση)



- Δίνεται: $\{2, 4, 10, 12, 3, 20, 30, 11, 25\}$, $k=2$
- Τυχαία επιλέγουμε 2 μέσους, έστω $m_1=3$, $m_2=4$
- 1^η επανάληψη:
 $K_1=\{2, 3\}$, $K_2=\{4, 10, 12, 20, 30, 11, 25\}$, $m_1=2.5$, $m_2=16$
- 2^η επανάληψη:
 $K_1=\{2, 3, 4\}$, $K_2=\{10, 12, 20, 30, 11, 25\}$, $m_1=3$, $m_2=18$
- 3^η επανάληψη:
 $K_1=\{2, 3, 4, 10\}$, $K_2=\{12, 20, 30, 11, 25\}$, $m_1=4.75$, $m_2=19.6$
- 4^η επανάληψη:
 $K_1=\{2, 3, 4, 10, 11, 12\}$, $K_2=\{20, 30, 25\}$, $m_1=7$, $m_2=25$
- 5^η επανάληψη: δεν αλλάζει τίποτα. Τέλος

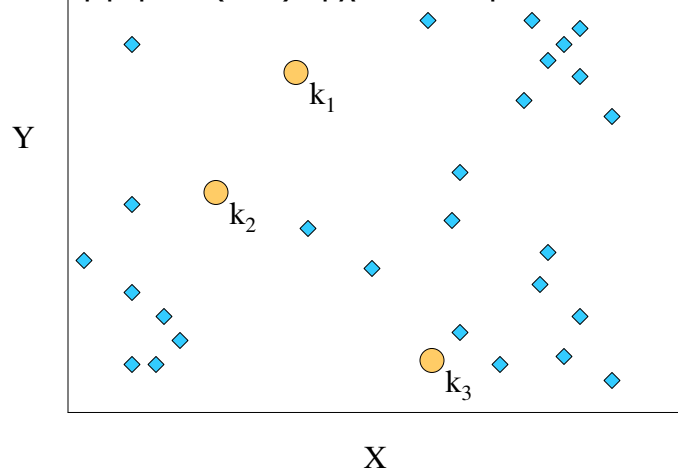
23

ΠΑ.ΠΕΛ - Γιάννης Θεοδωρίδης

Παράδειγμα K-Means (σε 2 διαστάσεις)



- Τυχαία επιλογή τριών ($k=3$) αρχικών κέντρων



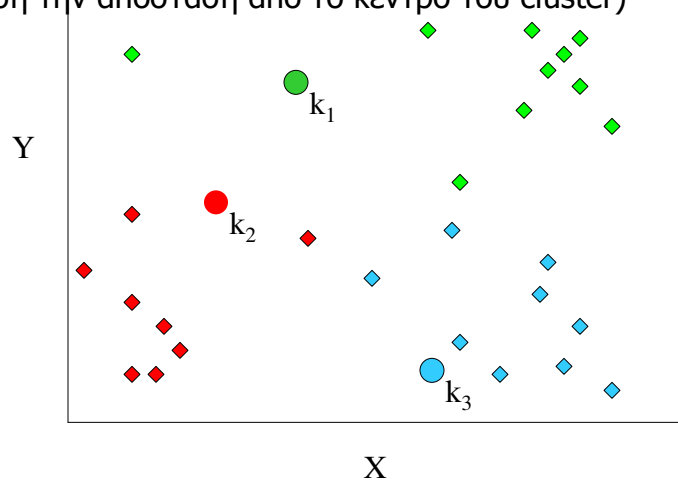
24

ΠΑ.ΠΕΛ - Γιάννης Θεοδωρίδης

Παράδειγμα K-means, 1^η επανάληψη



- Εκχώρηση κάθε στοιχείου στο πλησιέστερό του cluster (με βάση την απόσταση από το κέντρο του cluster)



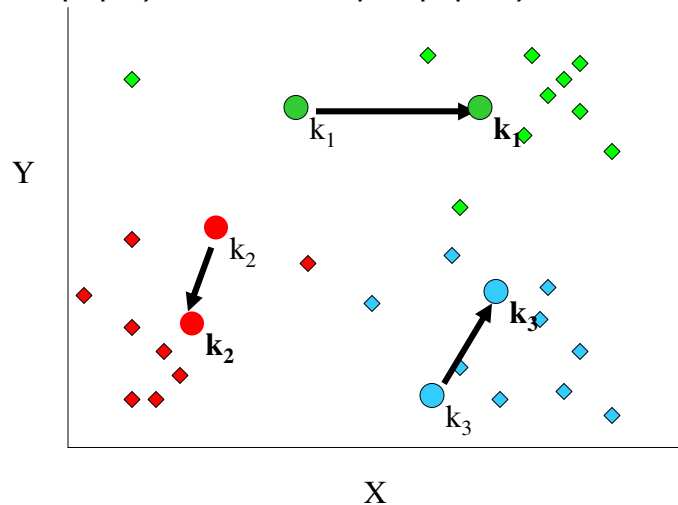
25

ΠΑ.ΠΕΛ - Γιάννης Θεοδωρίδης

Παράδειγμα K-means, 1^η επανάληψη



- Επανυπολογισμός του νέου κέντρου βάρους του κάθε cluster



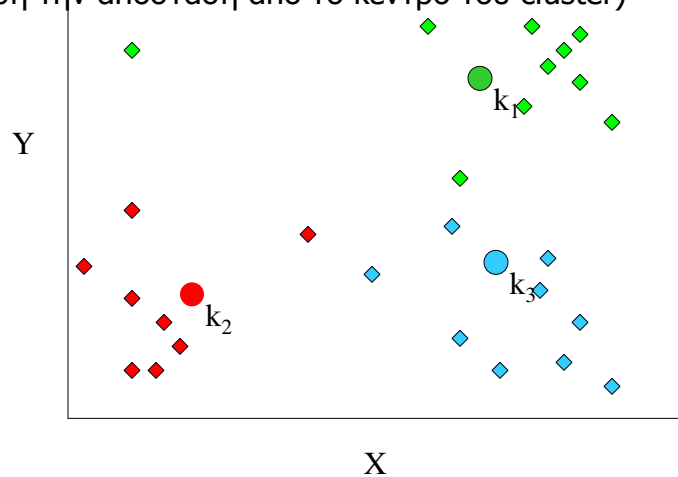
26

ΠΑ.ΠΕΛ - Γιάννης Θεοδωρίδης

Παράδειγμα K-means, 2^η επανάληψη



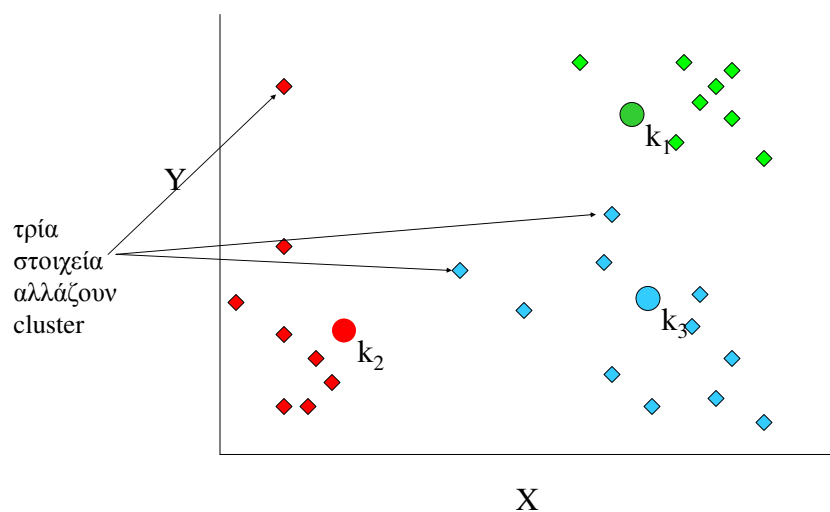
- Εκχώρηση κάθε στοιχείου στο πλησιέστερό του cluster (με βάση την απόσταση από το κέντρο του cluster)



27

ΠΑ.ΠΕΛ - Γιάννης Θεοδωρίδης

Παράδειγμα K-means, 2^η επανάληψη



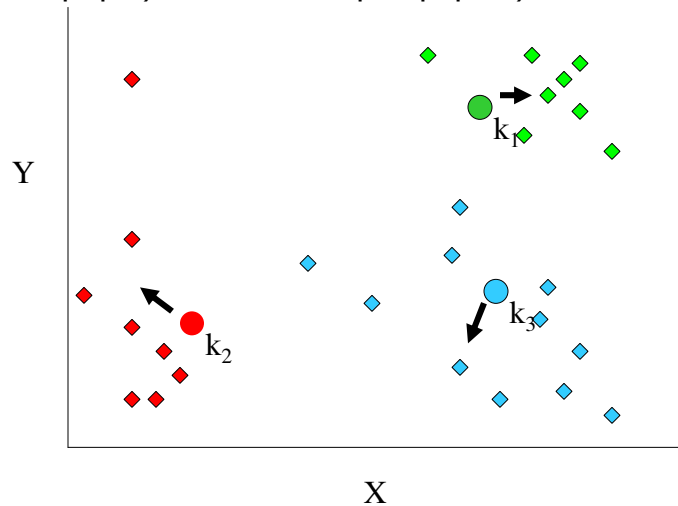
28

ΠΑ.ΠΕΛ - Γιάννης Θεοδωρίδης

Παράδειγμα K-means, 2^η επανάληψη



- Επανυπολογισμός του νέου κέντρου βάρους του κάθε cluster



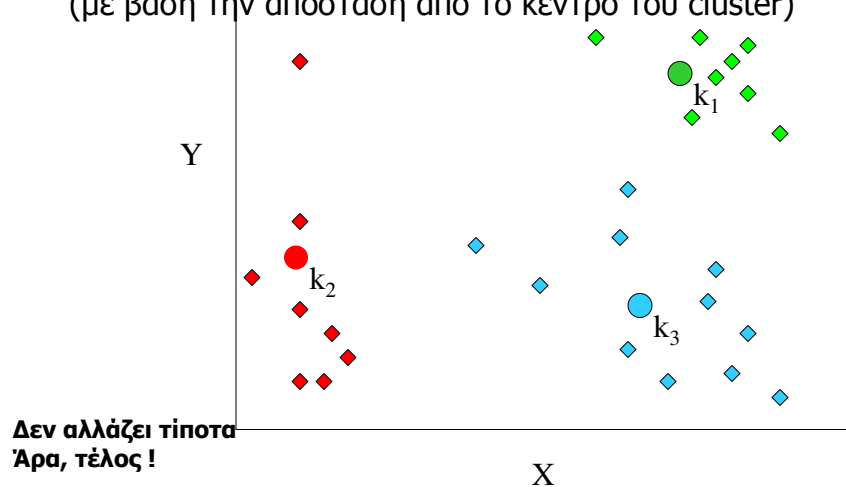
29

ΠΑ.ΠΕΛ - Γιάννης Θεοδωρίδης

Παράδειγμα K-means, 3^η επανάληψη



- Εκχώρηση κάθε στοιχείου στο πλησιέστερο του cluster (με βάση την απόσταση από το κέντρο του cluster)



Δεν αλλάζει τίποτα
Άρα, τέλος !

30

ΠΑ.ΠΕΛ - Γιάννης Θεοδωρίδης

Θέμα



- Αν θεωρήσουμε ότι κάθε καλαθοσφαιριστής στην παραπάνω ΒΔ/ΑΔ αποτελεί ένα 'διάνυσμα' σε έναν πολυδιάστατο χώρο με διαστάσεις τα χαρακτηριστικά που καταγράφουμε γι' αυτούς, τότε μπορούμε να εκτελέσουμε τεχνικές συσταδοποίησης για να βρούμε πιθανά προφίλ καλαθοσφαιριστών. Σχηματίστε ένα δικό σας παράδειγμα με 6 καλαθοσφαιριστές ($x_1, x_2, x_3, x_4, x_5, x_6$) και 2 αριθμητικές διαστάσεις: 'ηλικία' και 'λεπτά συμμετοχής'. Πάνω στο παράδειγμα αυτό, τρέξτε τον αλγόριθμο k-means δύο φορές, για $k=2$ και $k=3$. Κατόπιν, συγκρίνετε τα δύο αποτελέσματα συσταδοποίησης.

31

ΠΑ.ΠΕΙ. - Γιάννης Θεοδωρίδης

Θέμα



- Δεδομένων των πέντε 2-διάστατων σημείων $x_1 \dots x_5$ και υποθέτοντας ότι $k = 2$ και ότι αρχικά τα σημεία-σπόροι (seeds) είναι τα x_1 (για το 1ο cluster) και x_5 (για το 2ο cluster), να εφαρμόσετε τον αλγόριθμο συσταδοποίησης k-means μέχρι αυτός να συγκλίνει (δηλαδή, οι συστάδες να μην αλλάζουν) χρησιμοποιώντας Ευκλείδεια απόσταση. Να παραθέσετε όλα τα βήματα του αλγορίθμου.

	A_1	A_2
x_1	0	2
x_2	0	0
x_3	1.5	0
x_4	5	0
x_5	5	2

32

ΠΑ.ΠΕΙ. - Γιάννης Θεοδωρίδης



Εξόρυξη Κανόνων Συσχετίσεων

2 Δεκ. 08

Παράδειγμα (συν.)



Transaction	Items
t_1	Bread,Jelly,PeanutButter
t_2	Bread,PeanutButter
t_3	Bread,Milk,PeanutButter
t_4	Beer,Bread
t_5	Beer,Milk

$X \Rightarrow Y$	s	α
Bread $\Rightarrow$ PeanutButter	60%	75%
PeanutButter $\Rightarrow$ Bread	60%	100%
Beer $\Rightarrow$ Bread	20%	50%
PeanutButter $\Rightarrow$ Jelly	20%	33.3%
Jelly $\Rightarrow$ PeanutButter	20%	100%
Jelly $\Rightarrow$ Milk	0%	0%

Apriori



- **Συχνό** ονομάζεται το itemset που έχει υποστήριξη πάνω από ένα κατώφλι

- Παράδειγμα

(κατώφλι = 40%):

- {Beer}
- {Bread}
- {PeanutButter}
- {Bread, PeanutButter}

Transaction	Items
t_1	Bread,Jelly,PeanutButter
t_2	Bread,PeanutButter
t_3	Bread,Milk,PeanutButter
t_4	Beer,Bread
t_5	Beer,Milk

- Η **ιδιότητα των συχνών itemsets**:

- Κάθε υποσύνολο ενός συχνού itemset είναι συχνό.
- Αντιθέτως, αν ένα itemset δεν είναι συχνό, κανένα από τα υπερσύνολά του δεν μπορεί να είναι συχνό.

ΠΑ.ΠΕΛ - Γιάννης Θεοδορίδης

Παράδειγμα Apriori (συν.)



Pass	Candidates	Large Itemsets
1	{Beer},{Bread},{Jelly}, {Milk},{PeanutButter}	{Beer},{Bread}, {Milk},{PeanutButter}
2	{Beer,Bread},{Beer,Milk}, {Beer,PeanutButter},{Bread,Milk}, {Bread,PeanutButter},{Milk,PeanutButter}	{Bread,PeanutButter}

s=30% α = 50%

Transaction	Items
t_1	Bread,Jelly,PeanutButter
t_2	Bread,PeanutButter
t_3	Bread,Milk,PeanutButter
t_4	Beer,Bread
t_5	Beer,Milk

Θέμα



- Αν θεωρήσουμε ότι κάθε καλαθοσφαιριστής στην παραπάνω ΒΔ/ΑΔ σχηματίζει ένα 'καλάθι' μέσα στο οποίο βρίσκονται όλες οι θέσεις στις οποίες έχει αγωνιστεί κατά καιρούς, τότε θα μπορούσαμε να εκτελέσουμε τεχνικές εξόρυξης συχνών στοιχειοσυνόλων για να βρούμε ποιες θέσεις εμφανίζονται συχνά μαζί στην αθλητική ζωή ενός καλαθοσφαιριστή. Σχηματίστε ένα δικό σας παράδειγμα με 6 καλαθοσφαιριστές ($\kappa_1, \kappa_2, \kappa_3, \kappa_4, \kappa_5, \kappa_6$) και 3 διαφορετικές θέσεις ($\theta_1, \theta_2, \theta_3$) όπου οι θέσεις στις οποίες έχει αγωνιστεί ο καθένας είναι δική σας επιλογή. Πάνω στο παράδειγμα αυτό, τρέξτε τον αλγόριθμο Apriori και βρείτε τα συχνά στοιχειοσύνολα με βάση το κατώφλι $\text{minSupport} = 40\%$.

Θέμα



- Έστω το διπλανό σύνολο δοσοληψιών (καλαθιών). Με χρήση του αλγορίθμου Apriori, να βρείτε τους κανόνες συσχέτισης που προκύπτουν από αυτό το σύνολο για $\text{minSupport} = 30\%$ και $\text{minConfidence} = 50\%$. Σε κάθε βήμα του αλγορίθμου, να δείξετε τα επιμέρους αποτελέσματα που προκύπτουν.

<i>Trans.</i>	<i>Items</i>
t_1	a, b, d
t_2	a, d
t_3	a, d, e
t_4	c, d
t_5	c, e