Εξόρυξη Δεδομένων

Εισαγωγή

Εύρεση ενδιαφερόντων τάσεων ή προτύπων σε μεγάλα σύνολα δεδομένων

Στατιστική: Διερευνητική Ανάλυση Δεδομένων (exploratory data analysis)

Τεχνητή Νοημοσύνη: Ανακάλυψη γνώσης και μηχανική μάθηση

Δυνατότητα κλιμάκωσης σε σχέση με το μέγεθος του συνόλου των δεδομένων

Ένας αλγόριθμος **κλιμακώνεται** αν ο χρόνος εκτέλεσής του αυξάνεται ανάλογα (γραμμικά) με το μέγεθος του συνόλου δεδομένων για δοσμένους πόρους του συστήματος

Εισαγωγή

Τι σημαίνει ο ορισμός:

Εύρεση ενδιαφερόντων τάσεων ή προτύπων σε μεγάλα σύνολα δεδομένων

Ερωτήσεις SQL (βασίζονται στη σχεσιακή άλγεβρα)

Ερωτήσεις OLAP (υψηλότερου επιπέδου σύνταξη που βασίζεται στη χρήση του πολυδιάστατου μοντέλου δεδομένων)

Τεχνικές Εξόρυξης Δεδομένων

Η Διαδικασία Ανακάλυψης Γνώσης

The Knowledge Discovery Process (KDD) – Η Διαδικασία Ανακάλυψης Γνώσης σε Βάσεις Δεδομένων

Επιλογή Δεδομένων (Data Selection) — *Προσδιορισμός του συνόλου δεδομένων και των σχετικών γνωρισμάτων που μας ενδιαφέρουν – στα οποία θα γίνει η εξόρυξη*

↓

Καθαρισμός Δεδομένων Data Cleaning — *Απομάκρυνση θορύβου και των προς εξαίρεση τιμών (outliers), μετασχηματισμός των τιμών των πεδίων σε κοινές μονάδες μέτρησης, δημιουργία νέων πεδίων, αποθήκευση των δεδομένων σε σχεσιακό σχήμα*

↓

Εξόρυξη Δεδομένων

↓

Αξιολόγηση (Evaluation) — *Παρουσίαση των προτύπων με ένα τρόπο κατανοητό στον τελικό χρήστη (π.χ., μέσω τεχνικών οπτικοποίησης)*

Θέματα

- Απαρίθμηση Ταυτόχρονων Εμφανίσεων
  - Συχνά Στοιχειοσύνολα
  - Ερωτήσεις Παγόβουνου
- Εξόρυξη Κανόνων
  - Συνδυαστικοί Κανόνες
  - Ακολουθιακοί Κανόνες
- Κατηγοριοποίηση και Παλινδρόμηση
- Δενδρικοί Κανόνες
- Συγκρότηση (Clustering)
- Ομοιότητα Ακολουθιών

Απαρίθμηση Ταυτόχρονων Εμφανίσεων

Counting Co-Occurrence (Απαρίθμηση Ταυτόχρονων Εμφανίσεων)

Το καλάθι της νοικοκυράς (market basket) είναι μια συλλογή στοιχείων (collection of items) που αγοράστηκαν από ένα πελάτη κατά τη διάρκεια μιας μοναδικής συναλλαγής του (transaction) του

Στόχος: Εντοπισμός των αντικειμένων που εμφανίζονται μαζί σε μια συναλλαγή

Παράδειγμα

Transid	custid	date	item	qty
111	201	5/1/99	pen	2
111	201	5/1/99	ink	1
111	201	5/1/99	milk	3
111	201	5/1/99	juice	6
112	105	6/3/99	pen	1
112	105	6/3/99	ink	1
112	105	6/3/99	milk	1
113	106	5/10/99	pen	1
113	106	5/10/99	milk	1
114	201	6/1/99	pen	2
114	201	6/1/99	ink	2
114	201	6/1/99	juice	4

Μια συναλλαγή

Παρατηρείστε ότι υπάρχει επανάληψη πληροφορίας

Παρατηρήσεις του τύπου:

σε 75% των συναλλαγών αγοράστηκαν μαζί και μελάνι και πένες

Συχνά Στοιχειοσύνολα

Στοιχειοσύνολο (Itemset): είναι ένα σύνολο από αντικείμενα

Υποστήριξη ενός στοιχειοσυνόλου (Support of an itemset): το ποσοστό των συναλλαγών στης βάσης δεδομένων που περιέχουν όλα τα στοιχεία του στοιχειοσυνόλου

Παράδειγμα:

Στοιχειοσύνολο {pen, ink} Υποστήριξη 75%

Στοιχειοσύνολο {milk, juice} Υποστήριξη 25%

Μας ενδιαφέρουν στοιχειοσύνολα με μεγάλη υποστήριξη, γιατί;

Συχνά Στοιχειοσύνολα

Συχνά Στοιχειοσύνολα : Στοιχειοσύνολα των οποίων η υποστήριξη είναι μεγαλύτερη από κάποια ελάχιστη υποστήριξη (minimum support, minsup) που θέτουν οι χρήστες

Παράδειγμα:

Αν minsup = 70%,

Συχνά Στοιχειοσύνολα:

{pen}, {ink}, {milk}, {pen, ink}, {pen, milk}

Συχνά Στοιχειοσύνολα

Έναν αλγόριθμο που να εντοπίζει (όλα) τα συχνά στοιχειοσύνολα

Ο πιο απλός αλγόριθμος,

For k = 1 to n

Κατασκεύασε όλα τα δυνατά στοιχειοσύνολα με k στοιχεία

έλεγξε αν είναι συχνά

Συχνά Στοιχειοσύνολα

Αλγόριθμος

For each item,

check if it is a frequent itemset /* δηλαδή, αν υπάρχει > minsum συναλλαγές */

k = 1

repeat

for each new frequent itemset I_k with k items

Generate all itemsets I_{k+1} with k+1 items, $I_k \subset I_{k+1}$

Scan all transactions *once* and check if the generated k+1 itemsets are frequent

k = k + 1

Until n

Παράδειγμα

Transid	custid	date	item	qty
111	201	5/1/99	pen	2
111	201	5/1/99	ink	1
111	201	5/1/99	milk	3
111	201	5/1/99	juice	6
112	105	6/3/99	pen	1
112	105	6/3/99	ink	1
112	105	6/3/99	milk	1
113	106	5/10/99	pen	1
113	106	5/10/99	milk	1
114	201	6/1/99	pen	2
114	201	6/1/99	ink	2
114	201	6/1/99	juice	4

Έστω minsum = 70%

k=1

{pen} 4/4 ok

{ink} 3/4 ok

{milk} 3/4 ok

{juice} 2/4 όχι

k = 2

{pen, ink} 3/4 ok

{pen, milk} 3/4 ok

{pen, juice}

κλπ

Μπορούμε να κάνουμε κάτι καλύτερο;

Συχνά Στοιχειοσύνολα

Η ιδιότητα a priory:

Κάθε υποσύνολο ενός συχνού συνόλου πρέπει να αποτελεί επίσης ένα συχνό σύνολο

Γιατί;

Παράδειγμα

Transid	*custid*	*date*	*item*	*qty*
111	*201*	*5/1/99*	*pen*	*2*
111	*201*	*5/1/99*	*ink*	*1*
111	*201*	*5/1/99*	*milk*	*3*
111	*201*	*5/1/99*	*juice*	*6*
112	*105*	*6/3/99*	*pen*	*1*
112	*105*	*6/3/99*	*ink*	*1*
112	*105*	*6/3/99*	*milk*	*1*
113	*106*	*5/10/99*	*pen*	*1*
113	*106*	*5/10/99*	*milk*	*1*
114	*201*	*6/1/99*	*pen*	*2*
114	*201*	*6/1/99*	*ink*	*2*
114	*201*	*6/1/99*	*juice*	*4*

Έστω minsum = 70%

{pen} 4/4 ok

{ink} 3/4 ok

{milk} 3/4 ok

~~{juice} 2/4 όχι~~

{pen, ink} 3/4 ok

{pen, milk} 3/4 ok

~~{ink, milk} 2/4 όχι~~

Άρα τέλος

Αποτέλεσμα

{pen}, {ink}, {milk}, {pen, ink}
{pen, milk}

Συχνά Στοιχειοσύνολα

Βελτιωμένος Αλγόριθμος Εύρεσης Συχνών Στοιχειοσυνόλων

```
For each item,
        check if it is a frequent itemset
k = 1
repeat
        for each new frequent itemset I_k with k items
                Generate all itemsets I_{k+1} with k+1 items, I_k ⊂ I_{k+1} whose
                        subsets are frequent itemsets
                Scan all transactions once and check if the generated
                        k+1 itemsets are frequent
                k = k + 1
Until no new frequent itemsets are identified
```


Ερωτήσεις Τύπου Παγόβουνου

Υποθέστε ότι θέλουμε να βρούμε ζεύγη αγοραστών και στοιχείων τέτοιων ώστε ο αγοραστής τους να έχει αγοράσει το στοιχείο τουλάχιστον 5 φορές

select P.custid, P. item, **sum**(P.qty)
from Purchases P
group by P.custid, P.item
having sum (P.qty) > 5

Πως θα υπολογιζόταν αυτή ερώτηση σε ένα σχεσιακό ΣΔΒΔ;

Ο αριθμός των ομάδων είναι πολύ μεγάλος αλλά η απάντηση στην ερώτηση (η κορυφή του παγόβουνου) είναι πολύ μικρή

Ερωτήσεις Τύπου Παγόβουνου

Iceberg query (Ερώτηση τύπου παγόβουνου)

select R.A1, R.A2, …, R.Ak, **aggregate** (R.B)
from Relation R
group by R.A1, R.A2, …, R.Ak
having aggregate (R.B) > = constant

Μπορείτε να παρατηρείστε κάποια ιδιότητα a priory παρόμοια με αυτήν στην περίπτωση των συχνών στοιχειοσυνόλων;

Ερωτήσεις Τύπου Παγόβουνου

select P.custid, P. item, **sum**(P.qty)
from Purchases P
group by P.custid, P.item
having sum (P.qty) > 5

Αρκεί να εξετάσουμε μόνο εκείνες τις τιμές για το custid που αφορά πελάτες που έχουν αγοράσει τουλάχιστον 5 στοιχεία συνολικά (όχι απαραίτητα το ίδιο στοιχείο)

Q1:
select P.custid
from Purchases P
group by P.custid
having sum (P.qty) > 5

Ερωτήσεις Τύπου Παγόβουνου

Αντίστοιχα, αρκεί να εξετάσουμε μόνο εκείνες τις τιμές για το item που αφορούν στοιχεία που έχουν αγοραστεί 5 φορές συνολικά (όχι απαραίτητα από τον ίδιο πελάτη

Q2:

```
select P.item
from Purchases P
group by P.item
having sum (P.qty) > 5
```


Ερωτήσεις Τύπου Παγόβουνου

```
select P.custid, P. item, sum(P.qty)
from Purchases P
group by P.custid, P.item
having sum (P.qty) > 5
```

Q1

```
select P.custid
from Purchases P
group by P.custid
having sum (P.qty) > 5
```

Q2

```
select P.item
from Purchases P
group by P.item
having sum (P.qty) > 5
```

Δημιουργία (custid, item) ζευγών μόνο για custid από την Q1 και item από την Q2

Θέματα

Απαρίθμηση Ταυτόχρονων Εμφανίσεων
- Συχνά Στοιχειοσύνολα
- Ερωτήσεις Παγόβουνου

→ Εξόρυξη Κανόνων
- Συνδυαστικοί Κανόνες
- Ακολουθιακοί Κανόνες

Κατηγοριοποίηση και Παλινδρόμηση

Δενδρικοί Κανόνες

Συγκρότηση (Clustering)

Ομοιότητα Ακολουθιών

Συνδυαστικοί Κανόνες

Παράδειγμα

{pen} ⇒ {ink}

Αν ένα στοιχείο pen αγοράζεται σε μια συναλλαγή, τότε είναι πιθανό ότι αγοράζεται και το στοιχείο ink

Γενικά, συνδυαστικός κανόνας (association rule)

LHS ⇒ RHS

Όπου LHS και RHS είναι στοιχειοσύνολα

Συνδυαστικοί Κανόνες

LHS ⇒ RHS

Υποστήριξη (support): support(LHS ∪ RHS)

Το ποσοστό των συναλλαγών που περιέχουν όλα τα στοιχεία του (LHS ∪ RHS)

Εμπιστοσύνη (confidence): support(LHS ∪ RHS) / support(LHS)

Μια ένδειξη της ισχύς του κανόνα

P(RHS | LHS)

Παράδειγμα

Transid	*custid*	*date*	*item*	*qty*
111	*201*	*5/1/99*	*pen*	*2*
111	*201*	*5/1/99*	*ink*	*1*
111	*201*	*5/1/99*	*milk*	*3*
111	*201*	*5/1/99*	*juice*	*6*
112	*105*	*6/3/99*	*pen*	*1*
112	*105*	*6/3/99*	*ink*	*1*
112	*105*	*6/3/99*	*milk*	*1*
113	*106*	*5/10/99*	*pen*	*1*
113	*106*	*5/10/99*	*milk*	*1*
114	*201*	*6/1/99*	*pen*	*2*
114	*201*	*6/1/99*	*ink*	*2*
114	*201*	*6/1/99*	*juice*	*4*

pen ⇒ milk (K1)
support(K1) = 75%
confidence(K1) = 75%

milk ⇒ pen (K2)
support(K2) = 75%
confidence(K2) = 100%

Συνδυαστικοί Κανόνες

Έναν αλγόριθμο εύρεσης όλων των κανόνων με ελάχιστο minsup και ελάχιστο minconf

Ελάχιστο minsum, support(LHS ∪ RHS) ≥ minsum

Βήμα 1: Βρες όλα τα συχνά στοιχειοσύνολα με minsup

Βήμα 2: Παρήγαγε όλους τους κανόνες από το Βήμα 1

Παράδειγμα

Transid	*custid*	*date*	*item*	*qty*
111	*201*	*5/1/99*	*pen*	*2*
111	*201*	*5/1/99*	*ink*	*1*
111	*201*	*5/1/99*	*milk*	*3*
111	*201*	*5/1/99*	*juice*	*6*
112	*105*	*6/3/99*	*pen*	*1*
112	*105*	*6/3/99*	*ink*	*1*
112	*105*	*6/3/99*	*milk*	*1*
113	*106*	*5/10/99*	*pen*	*1*
113	*106*	*5/10/99*	*milk*	*1*
114	*201*	*6/1/99*	*pen*	*2*
114	*201*	*6/1/99*	*ink*	*2*
114	*201*	*6/1/99*	*juice*	*4*

Αποτέλεσμα

{pen}, {ink}, {milk}, {pen, ink}
{pen, milk}

Δυνατοί κανόνες

pen ⇒ ink

ink ⇒ pen

pen ⇒ milk

milk ⇒ pen

Συνδυαστικοί Κανόνες

Βήμα 2

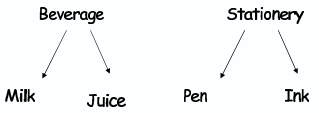
For each frequent itemset I with support support(I)
Divide I into LHS_I and RHS_I
confidence = support(I) / support(LHS_I)

Παρατήρηση

Τα support(I) και support(LHS_I) τα έχουμε ήδη υπολογίσει σε προηγούμενα βήματα του αλγορίθμου εύρεσης συχνών στοιχειοσυνόλων

Συνδυαστικοί Κανόνες και Ιεραρχίες ISA

Μια ISA ιεραρχία ή ιεραρχία κατηγοριών (category hierarchy) ανάμεσα στα σύνολα των στοιχείων: μια συναλλαγή εμμέσως περιέχει για κάθε στοιχείο και όλα τα στοιχεία που είναι προγονοί του στην ιεραρχία



- Επιτρέπει τον εντοπισμό σχέσεων μεταξύ στοιχείων που ανήκουν σε διαφορετικά επίπεδα της ιεραρχίας
- Γενικά, η υποστήριξη (support) ενός στοιχείου μπορεί μόνο να αυξηθεί με την αντικατάσταση του στοιχείου από κάποιο πρόγονο του στην ιεραρχία

Γενικευμένοι Συνδυαστικοί Κανόνες

Πιο γενικά: όχι μόνο συναλλαγές αγοραστών

Transid	*custid*	*date*	*item*	*qty*
111	*201*	*5/1/99*	*pen*	*2*
111	*201*	*5/1/99*	*ink*	*1*
111	*201*	*5/1/99*	*milk*	*3*
111	*201*	*5/1/99*	*juice*	*6*
112	*105*	*6/3/99*	*pen*	*1*
112	*105*	*6/3/99*	*ink*	*1*
112	*105*	*6/3/99*	*milk*	*1*
113	*106*	*5/10/99*	*pen*	*1*
113	*106*	*5/10/99*	*milk*	*1*
114	*201*	*6/1/99*	*pen*	*2*
114	*201*	*6/1/99*	*ink*	*2*
114	*201*	*6/1/99*	*juice*	*4*

π.χ., Ομαδοποίηση πλειάδων με βάση το custid

Ο κανόνας **{pen} ⇒ {milk}**: αν το στοιχείο pen αγοραστεί από κάποιο πελάτη είναι πιθανό ότι ο πελάτης θα αγοράσει και το στοιχείο milk (με υποστήριξη και εμπιστοσύνη 100%)

Γενικευμένοι Συνδυαστικοί Κανόνες

Ομαδοποίηση πλειάδων με βάση την ημερομηνία: Ημερολογιακή ανάλυση του καλαθιού της νοικοκυράς (Calendric market basket analysis)

Ένα ημερολόγιο (calendar) είναι μια οποιαδήποτε ομάδα ημερομηνιών (π.χ., κάθε πρώτη του μήνα)

Δοσμένου ενός ημερολογίου, υπολόγισε τους συνδυαστικούς κανόνες που αφορούν πλειάδες που έχουν πραγματοποιηθεί σε ημερομηνίες που ανήκουν στο ημερολόγιο

Γενικευμένοι Συνδυαστικοί Κανόνες

Ημερολόγιο: κάθε πρώτη του μήνα

Ο κανόνας **{pen} ⇒ {juice}**: έχει support 100%

Ενώ γενικά: 50%

Ο κανόνας **{pen} ⇒ {milk}**: έχει support 50%

Ενώ γενικά: 75%

Transid	custid	date	item	qty
111	201	5/1/99	pen	2
111	201	5/1/99	ink	1
111	201	5/1/99	milk	3
111	201	5/1/99	juice	6
112	105	6/3/99	pen	1
112	105	6/3/99	ink	1
112	105	6/3/99	milk	1
113	106	5/10/99	pen	1
113	106	5/10/99	milk	1
114	201	6/1/99	pen	2
114	201	6/1/99	ink	2
114	201	6/1/99	juice	4

Ακολουθιακά Πρότυπα

Ακολουθία (sequence) στοιχειοσυνόλων:

Η ακολουθία των στοιχείων που αγοράστηκαν από τον πελάτη:

Παράδειγμα custid 201: {pen, ink, milk, juice}, {pen, ink, juice}

(ordered by date)

Μια υπο-ακολουθία (subsequence) μιας ακολουθίας στοιχειοσυνόλων προκύπτει διαγράφοντας ένα ή περισσότερα στοιχειοσύνολα και αποτελεί επίσης μια ακολουθία στοιχειοσυνόλων

Ακολουθιακά Πρότυπα

Μια ακολουθία $a_1, a_2, .., a_n$ περιέχεται σε μια ακολουθία S αν η S έχει μια υπο-ακολουθία $b_q, .., b_m$ such that $a_i \subseteq b_i$ for $1 \le i \le m$

Παράδειγμα

{pen}, {ink, milk}, {pen, juice} περιέχεται στην {pen, ink}, {shirt}, {juice, ink, milk}, {juice, pen, milk}

Η σειρά των στοιχείων σε κάθε στοιχειοσύνολο δεν έχει σημασία αλλά η σειρά των στοχειοσυνόλων στην ακολουθία έχει

{pen}, {ink, milk}, {pen, juice} δεν περιέχεται στην {pen, ink}, {shirt}, {juice, pen, milk}, {juice, milk, ink}

Ακολουθιακά Πρότυπα

Η υποστήριξη μια ακολουθίας στοιχειοσυνόλων S είναι το ποσοστό των ακολουθιών του πελάτη των οποίων η S είναι υπο-ακολουθία

Βρες όλες τις ακολουθίες που έχουν μια ελάχιστη υποστήριξη

Θέματα

- Απαρίθμηση Ταυτόχρονων Εμφανίσεων
  - Συχνά Στοιχειοσύνολα
  - Ερωτήσεις Παγόβουνου
- Εξόρυξη Κανόνων
  - Συνδυαστικοί Κανόνες
  - Ακολουθιακοί Κανόνες
- → Κατηγοριοποίηση και Παλινδρόμηση
- Δενδρικοί Κανόνες
- Συγκρότηση (Clustering)
- Ομοιότητα Ακολουθιών

Κανόνες Κατηγοριοποίησης και Παλινδρόμησης

InsuranceInfo(age: integer, cartype: string, highrisk: boolean)

Υπάρχει ένα γνώρισμα (highrisk: ΥψηλούΚινδύνου) του οποίου την τιμή θα θέλαμε να προβλέψουμε: Εξαρτημένο Γνώρισμα

Τα άλλα γνωρίσματα ονομάζονται προβλέποντα γνωρίσματα (predictors)

Η γενική μορφή των κανόνων που θέλουμενα προβλέψουμε είναι:

$$P_1(X_1) \wedge P_2(X_2) \wedge \ ... \wedge P_k(X_k) \Rightarrow Y = c$$

Κανόνες Κατηγοριοποίησης και Παλινδρόμησης

$P_1(X_1) \wedge P_2(X_2) \wedge \ ... \wedge P_k(X_k) \Rightarrow Y = c$

$P_i(X_i)$ είναι κατηγορήματα (predicates)

Δύο τύποι γνωρισμάτων:

- αριθμητικά $P_i(X_i) : l_i \le X_i \le h_i$
- κατηγοριακά (categorical) $P_i(X_i) : X_i \in \{v_1, ..., v_j\}$

- αριθμητικά εξαρτημένα γνωρίσματα κανόνας παλινδρόμησης
- κατηγοριακά εξαρτημένα γνωρίσματα κανόνας κατηγοριοποίησης

$(16 \le age \le 25) \wedge (cartype \in \{Sports, Truck\}) \Rightarrow highrisk = true$

Κανόνες Κατηγοριοποίησης και Παλινδρόμησης

$P_1(X_1) \wedge P_2(X_2) \wedge \ ... \wedge P_k(X_k) \Rightarrow Y = c$

Υποστήριξη:

Η υποστήριξη για μια συνθήκη C είναι το ποσοστό των πλειάδων που ικανοποιούν την C.

Η υποστήριξη του κανόνα $C1 \Rightarrow C2$ είναι η υποστήριξη της συνθήκης $C1 \wedge C2$

Εμπιστοσύνη

Έστω όλες οι πλειάδες που ικανοποιούν τη συνθήκη $C1$. Η εμπιστοσύνη του κανόνα $C1 \Rightarrow C2$ είναι το ποσοστό των πλειάδων αυτών που ικανοποιούν και τη συνθήκη $C2$

Κανόνες Κατηγοριοποίησης και Παλινδρόμησης

Γενικεύοντας

$P_1(X_1) \wedge P_2(X_2) \wedge \ ... \wedge P_k(X_k) \Rightarrow Y = f(X_1, X_2, ..., X_k)$

Η διαφορά από τους συνδυαστικούς κανόνες είναι ότι θεωρούν συνεχή και κατηγοριακά γνωρίσματα και όχι μόνο ένα πεδίο με πολλαπλές καθορισμένες τιμές

Πολλές εφαρμογές, π.χ. σε επιστημονικά πειράματα, προώθηση προϊόντων μέσω ταχυδρομείου, προβλέψεις χρηματοοικονομικών μεγεθών, ιατρικές προγνώσεις

Θα δούμε στη συνέχεια έναν ειδικό τύπο τους

Δενδρικοί Κανόνες

Δέντρα αποφάσεων ή δέντρα κατηγοριοποίησης

Δέντρα Παλινδρόμησης

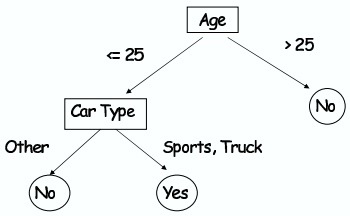
Συνήθως το *ίδιο το δέντρο* είναι το αποτέλεσμα της εξόρυξης δεδομένων

Εύκολο να κατανοηθεί

Αποδοτικοί αλγόριθμοι για την κατασκευή του

Δενδρικοί Κανόνες

Παράδειγμα δέντρου



Δενδρικοί Κανόνες

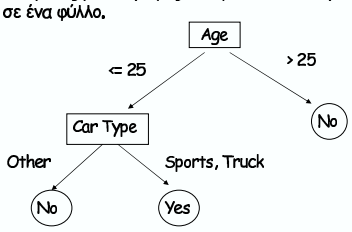
Γραφική απεικόνιση μια συλλογής κανόνων κατηγοριοποίησης.

- Εσωτερικοί κόμβοι: ετικέτα (labeled) με ένα προβλέπον γνώρισμα (που ονομάζεται και διαχωρίζον γνώρισμα (splitting attribute)
- Εξερχόμενες ακμές: έχουν ως ετικέτα το κατηγόρημα που περιλαμβάνει το διαχωρίζον γνώρισμα του κόμβου (κριτήριο διαχωρισμού του κόμβου - splitting criterion) - ξένα μεταξύ τους
- Φύλλα: έχουν ως ετικέτα το κατηγόρημα μια τιμή του εξαρτώμενου γνωρίσματος

Θα εξετάσουμε δυαδικά δέντρα αποφάσεων αλλά μπορεί να έχουμε και δέντρα υψηλότερου βαθμού

Δενδρικοί Κανόνες

Δοσμένης μια εγγραφής δεδομένων το δέντρο οδηγεί από τη ρίζα σε ένα φύλλο.



Μπορούμε να κατασκευάσουμε κανόνες κατηγοριοποίησης ακολουθώντας κάθε μονοπάτι από τη ρίζα σε φύλλο:

LHS σύζευξη των κατηγορημάτων, RHS η τιμή του φύλλου

Ένας κανόνας για κάθε φύλλο

$(16 \le age \le 25) \wedge (cartype \in \{Sports, Truck\}) \Rightarrow highrisk = true$

Δενδρικοί Κανόνες

Κατασκευή σε δύο φάσεις

ΦΑΣΗ 1: φάση ανάπτυξης

κατασκευή ενός πολύ μεγάλου δέντρου (π.χ., ένα φύλλο για κάθε εγγραφή της βάσης δεδομένων)

ΦΑΣΗ 2: φάση κλαδέματος

Κατασκευή του δέντρου greedily από πάνω προς τα κάτω:

Στη ρίζα εξέτασε τη βάση δεδομένων και επέλεξε το *καλύτερο κριτήριο διαχωρισμού (τοπικά βέλτιστο κριτήριο)*

Διαμέρισε τη βάση σε δύο μέρη

Εφάρμοσε αναδρομικά σε κάθε παιδί

Δενδρικοί Κανόνες

Input: node n partition D split selection method S

Output: decision tree for D rooted at node n

Top down Decision Tree Induction Schema

BuildTree(node n, partition D, method S)

Apply S to D to find the splitting criterion

If (a good splitting criterion is found)

create two children nodes n1 and n2 of n

partition D into D1 and D2

BuildTree(n1, D1, S)

Build Tree(n2, D2, S)

Δενδρικοί Κανόνες

Μέθοδος επιλογής διαχωρισμού

Ένας αλγόριθμος που παίρνει ως είσοδο μια σχέση (ή ένα τμήμα μιας σχέσης) και δίνει ως έξοδο το τοπικά βέλτιστο κριτήριο

Παράδειγμα: εξετάστε τα γνωρίσματα cartype και age, επέλεξε ένα από αυτό ως γνώρισμα διαχωρισμού και μετά επέλεξε το γνώρισμα διαχωρισμού

Δενδρικοί Κανόνες

Πως μπορούμε να κατασκευάσουμε δέντρα αποφάσεων που είναι μεγαλύτερα από τη κύρια μνήμη:

Αντί να φορτώσουμε όλη τη βάση δεδομένων στη μνήμη:

Δίνουμε στη μέθοδο επιλογής διαχωρισμού συανθροιστική πληροφορία για τα γνωρίσματα

Χρειαζόμαστε συναθροιστική πληροφορία για κάθε γνώρισμα πρόβλεψης

Σύνολο AVC του γνωρίσματος πρόβλεψης X στον κόμβο n είναι η προβολή του διαμερίσματος της βάσης δεδομένων του κόμβου n στο X και στο εξαρτημένο γνώρισμα όπου συναθροίζονται οι συχνότητες (counts) των διακριτών τιμών του εξαρτημένου γνωρίσματος

Δενδρικοί Κανόνες

age	cartype	highrisk
23	Sedan	false
30	Sports	false
36	Sedan	false
25	Truck	true
30	Sedan	false
23	Truck	true
30	Truck	false
25	Sports	true
18	Sedan	false

Σύνολο AVC του γνωρίσματος πρόβλεψης *age* στη *ρίζα*

```
select R.age, R.highrisk, count(*)
from InsuranceInfo R
group by R.age, R.highrisk
```

	true	false
18	0	1
23	1	1
25	2	0
30	0	3
36	0	1

Δενδρικοί Κανόνες

age	cartype	highrisk
23	Sedan	false
30	Sports	false
36	Sedan	false
25	Truck	true
30	Sedan	false
23	Truck	true
30	Truck	false
25	Sports	true
18	Sedan	false

Σύνολο AVC του γνωρίσματος πρόβλεψης *cartype* στη *ρίζα*

select R.cartype, R.highrisk, **count(*)**
from InsuranceInfo R
group by R.cartype, R.highrisk

	true	false
Sedan	0	4
Sports	1	1
Truck	2	1

Δενδρικοί Κανόνες

age	cartype	highrisk
23	Sedan	false
30	Sports	false
36	Sedan	false
25	Truck	true
30	Sedan	false
23	Truck	true
30	Truck	false
25	Sports	true
18	Sedan	false

Σύνολο AVC του γνωρίσματος πρόβλεψης *cartype* στο αριστερό παιδί της ρίζας

select R.cartype, R.highrisk, **count(*)**
from InsuranceInfo R
where R.age <=25
group by R.age, R.highrisk

	true	false
Sedan	0	2
Sports	1	0
Truck	2	0

Δενδρικοί Κανόνες

Ομάδα AVC ενός κόμβου n: το σύνολο των AVC συνόλων όλων των γνωρισμάτων πρόβλεψης στον κόμβο n

Ποιο είναι το μέγεθος του συνόλου AVC;

Δενδρικοί Κανόνες

Input: node n partition D split selection method S
Output: decision tree for D rooted at node n

Top down Decision Tree Induction Schema
BuildTree(node n, partition D, method S)
Make a scan over D and construct the AVC group of node n in memory
Apply S to AVC group to find the splitting criterion
If (a good splitting criterion is found)
    create two children nodes n1 and n2 of n
    partition D into D1 and D2
    BuildTree(n1, D1, S)
    Build Tree(n2, D2, S)

Θέματα

- Απαρίθμηση Ταυτόχρονων Εμφανίσεων
  - Συχνά Στοιχειοσύνολα
  - Ερωτήσεις Παγόβουνου
- Εξόρυξη Κανόνων
  - Συνδυαστικοί Κανόνες
  - Ακολουθιακοί Κανόνες
- Κατηγοριοποίηση και Παλινδρόμηση
- Δενδρικοί Κανόνες
- → Συγκρότηση (Clustering)
- Ομοιότητα Ακολουθιών

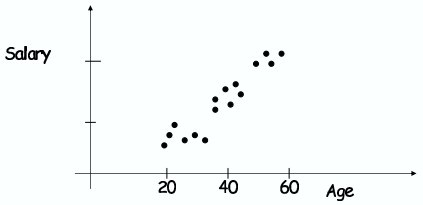
Συγκρότηση

Διαμέριση ενός συνόλου εγγραφών σε ομάδες - συγκροτήματα **(clusters)** έτσι ώστε όλες οι εγγραφές που ανήκουν σε μια ομάδα να είναι όμοιες μεταξύ τους και οι εγγραφές που ανήκουν σε διαφορετικές ομάδες να είναι ανόμοιες

Η ομοιότητα μεταξύ των εγγραφών υπολογίζεται μέσς μιας συνάρτησης απόστασης **distance function**.

Εξαρτάται από τον τύπο των δεδομένων και την εφαρμογή

CustomerInfo(age: integer, salary:real)



- Μπορούμε να εντοπίσουμε τρία συγκροτήματα (clusters) – το σχήμα τους σφαιρικό

Η έξοδος ενός αλγορίθμου συγκρότησης είναι μια **συνοπτική αναπαράσταση** κάθε συγκροτήματος

Η μορφή του αποτελέσματος εξαρτάται από τον τύπο και το σχήμα των συγκροτημάτων

Για παράδειγμα για σφαιρικά συγκροτήματα, έξοδος: κέντρο C (μέσο) and ακτίνα R:

δοσμένου μιας συλλογής εγγραφών $r_1, r_2, \ldots, r_n$

$$C = \frac{\sum r_i}{n} \qquad R = \frac{\sum (r_i - C)}{n}$$

Δύο κατηγορίες αλγορίθμων συγκρότησης:

- **Αλγόριθμος Διαχωρισμού**: διαμερίζει τα δεδομένα σε k συγκροτήματα έτσι ώστε να βελτιστοποιείται η τιμή κάποιου κριτηρίου – το k συνήθως προσδιορίζεται από τον χρήστη

- **Ιεραρχικός Αλγόριθμος** παράγει μια ακολουθία από διαμερίσεις των δεδομένων.

Ξεκινώντας από μια διαμέριση όπου κάθε εγγραφή αποτελεί ένα συγκρότημα, σε κάθε βήμα συγχωνεύει και δυο συγκροτήματα

Ο αλγόριθμος **BIRCH**:

Υποθέσεις

- Μεγάλος αριθμός εγγραφών, μόνο ένα πέρασμα
- Περιορισμένη μνήμη

Δύο παράμετροι

- *k:* όριο διαθεσιμότητας κύριας μνήμης: μέγιστος αριθμός συνοπτικών αναπαραστάσεων συγκροτημάτων που μπορούν αν καταχωρηθούν στην κύρια μνήμη
- *e:* αρχικό όριο της ακτίνας κάθε συγκροτήματος – καθορίζει και τον αριθμό των συγκροτημάτων. Ένα συγκρότημα είναι *συμπαγές* αν η ακτίνα του είναι μικρότερη από e.

Πάντα διατήρησε στην κύρια μνήμη k ή λιγότερα συμπαγή συγκροτήματα (Ci, Ri)

(Αν αυτό δεν είναι δυνατόν, τροποποίησε το e)

Ο αλγόριθμος **BIRCH**:

Read a record *r* from the database /* διάβασε τις εγγραφές σειριακά */
Compute the distance of *r* and each of the existing cluster centers
Let i be the cluster (index) such that the distance between *r* and C_i is the smallest /* i: το πιο "κοντινό" συγκρότημα */
Compute R'_i assuming *r* is inserted in the ith cluster /* υπολόγισε τη νέα ακτίνα */
If $R'_i \le e$, /* μικρότερη, το συγκρότημα παραμένει συμπαγές, επιτύχαμε το r */
insert r in the ith cluster
recompute R_i and C_i
else /* μεγαλύτερη δημιούργησε νέο συγκρότημα */
start a new cluster containing only r

Ο αλγόριθμος **BIRCH**:

- Μέγιστος αριθμός συγκροτημάτων -> τροποποίησε το e
- Αποδοτικοί τρόποι εντοπισμού του κοντινότερου συγκροτήματος, π.χ., χρήση B+ δέντρων

Θέματα

- Απαρίθμηση Ταυτόχρονων Εμφανίσεων
  - Συχνά Στοιχειοσύνολα
  - Ερωτήσεις Παγόβουνου
- Εξόρυξη Κανόνων
  - Συνδυαστικοί Κανόνες
  - Ακολουθιακοί Κανόνες
- Κατηγοριοποίηση και Παλινδρόμηση
- Δενδρικοί Κανόνες
- Συγκρότηση (Clustering)
- → Ομοιότητα Ακολουθιών

Ομοιότητα Ακολουθιών

Ο χρήστης καθορίζει μία ακολουθία αίτημα (query sequence) και θέλει να ανακτήσει όλες τις ακολουθίες δεδομένων που είναι όμοιες με αυτήν

Όχι απαραίτητα να ταιριάζουν ακριβώς (Not exact matches)

Μια ακολουθία δεδομένων (data sequence) X είναι μια σειρά αριθμών X = $<x_1, x_2, .., x_k>$

Συχνά ονομάζεται και χρονολογική σειρά (time series)

k μήκος (length) της ακολουθίας

Μια υποακολουθία (subsequence) Z = $<z_1, z_2, ..., z_j>$ προκύπτει από μια ακολουθία X με τη διαγραφή αριθμών από των αρχή και το τέλος της ακολουθίας

Ομοιότητα Ακολουθιών

Μπορούμε να ορίσουμε την απόσταση μεταξύ δυο ακολουθιών ως την Ευκλείδεια κανονική (**Euclidean norm**)

Δοθείσας μιας ακολουθίας-αίτημα και ενός ορίου ε, θέλουμε να ανακτήσουμε όλες τις ακολουθίες δεδομένων που είναι μέσα σε απόσταση ε

- Πλήρης ταύτιση ακολουθιών Complete sequence matching η ακολουθία-αίτημα και οι ακολουθίες δεδομένων έχουν το ίδιο μήκος
- Ταύτιση υποακολουθιών Subsequence matching (η ακολουθία αίτημα έχει μικρότερο μήκος)

Ομοιότητα Ακολουθιών

Δοθείσας μιας ακολουθίας-αίτημα και ενός ορίου ε, θέλουμε να ανακτήσουμε όλες τις ακολουθίες δεδομένων που είναι μέσα σε απόσταση ε

Brute-force μέθοδος

Αναπαράσταση τους ως ένα σημείο στον πολυδιάστατο (k-διάστατο) χώρο

Πολυδιάστατο ευρετήριο

Μη ακριβή ταιριάσματα; Ερώτηση υπερ-ορθογώνια περιοχή Query (hyper-rectangle) με πλευρά μήκους 2-ε και κέντρο την ακολουθία αίτημα

Θέματα

- Απαρίθμηση Ταυτόχρονων Εμφανίσεων
  - Συχνά Στοιχειοσύνολα
  - Ερωτήσεις Παγόβουνου
- Εξόρυξη Κανόνων
  - Συνδυαστικοί Κανόνες
  - Ακολουθιακοί Κανόνες
- Κατηγοριοποίηση και Παλινδρόμηση
- Δενδρικοί Κανόνες
- Συγκρότηση (Clustering)
- Ομοιότητα Ακολουθιών