Αποθήκες Δεδομένων και Εξόρυξη Γνώσης

Κανόνες Συσχέτισης

Association Rules

Ανακάλυψη κανόνων συσχετίσεων: Ορισμός

- Δεδομένου ενός συνόλου από εγγραφών, κάθε μια εκ των οποίων περιέχει ένα αριθμό αντικειμένων από μια δεδομένη συλλογή
  - Παραγωγή κανόνων εξάρτησης οι οποίοι προβλέπουν την εμφάνιση ενός αντικειμένου με βάση την εμφάνιση άλλων αντικειμένων

ID	*Αντικείμενα*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Εξαγωγή Κανόνων:
{Milk} --> {Coke}
{Diaper, Milk} --> {Beer}

Ανακάλυψη κανόνων συσχετίσεων: Εφαρμογή 1

- Προώθηση προϊόντων
  - Έστω ότι ο κανόνας που εξήχθηκε λέει ότι:
    - {παξιμάδια,…}→πατατάκια
  - Τα πατατάκια ως συνεπαγωγή → χρησιμοποιούνται για να εξακριβωθεί πως θα προωθήσουν την πώληση τους
  - Τα παξιμάδια ως προγενέστερο αντικείμενο της συνεπαγωγής → χρησιμοποιούνται για να εξακριβωθούν οι συνέπειες εάν σταματήσει η πώληση τους
  - Ο συνολικός κανόνας → μπορεί να χρησιμοποιηθεί για να εξακριβωθεί ποια προϊόντα πρέπει να πωλούνται μαζί με τα παξιμάδια για να πωλούνται και πατατάκια μαζί!

Ανακάλυψη κανόνων συσχετίσεων: Εφαρμογή 2

- Αυτοδιοίκηση σε πολυκατάστημα
  - Στόχος: να αναγνωρίσουν ποια προϊόντα αγοράζονται μαζί από πολλούς πελάτες
  - Προσέγγιση:
    - Επεξεργασία των κωδικών κάθε προϊόντος όπως αυτά περνάνε από το barcode του ταμείου
  - Ένας κλασικός κανόνας
    - Εάν ο πελάτης αγοράσει γάλα και πάνες, τότε πιθανό να αγοράσει και μπύρα
    - Επομένως, δεν είναι απίθανο να τοποθετηθούν οι μπύρες δίπλα στις πάνες!

Ανακάλυψη κανόνων συσχετίσεων: Εφαρμογή 3

- Διαχείριση εργαλειοθήκης:
  - Στόχος: μια εταιρία επισκευής ηλεκτρικών συσκευών θέλει να προβλέπει τη φύση της βλάβης και να έχει στα οχήματα τα κατάλληλα εργαλεία
- Προσέγγιση:
  - Επεξεργασία των δεδομένων για τα εργαλεία και τις βλάβες σε προγενέστερες περιπτώσεις βλαβών και ανακάλυψη προτύπων συνεμφάνισης

Table 1.2 The weather data.

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

```
If temperature = cool then humidity = normal
```

Coverage ή Support: Πόσες μέρες είναι κρύες ΚΑΙ έχουν κανονική υγρασία (4)

Accuracy ή Confidence: Πόσες από τις κρύες μέρες έχουν κανονική υγρασία (4/4)=100%

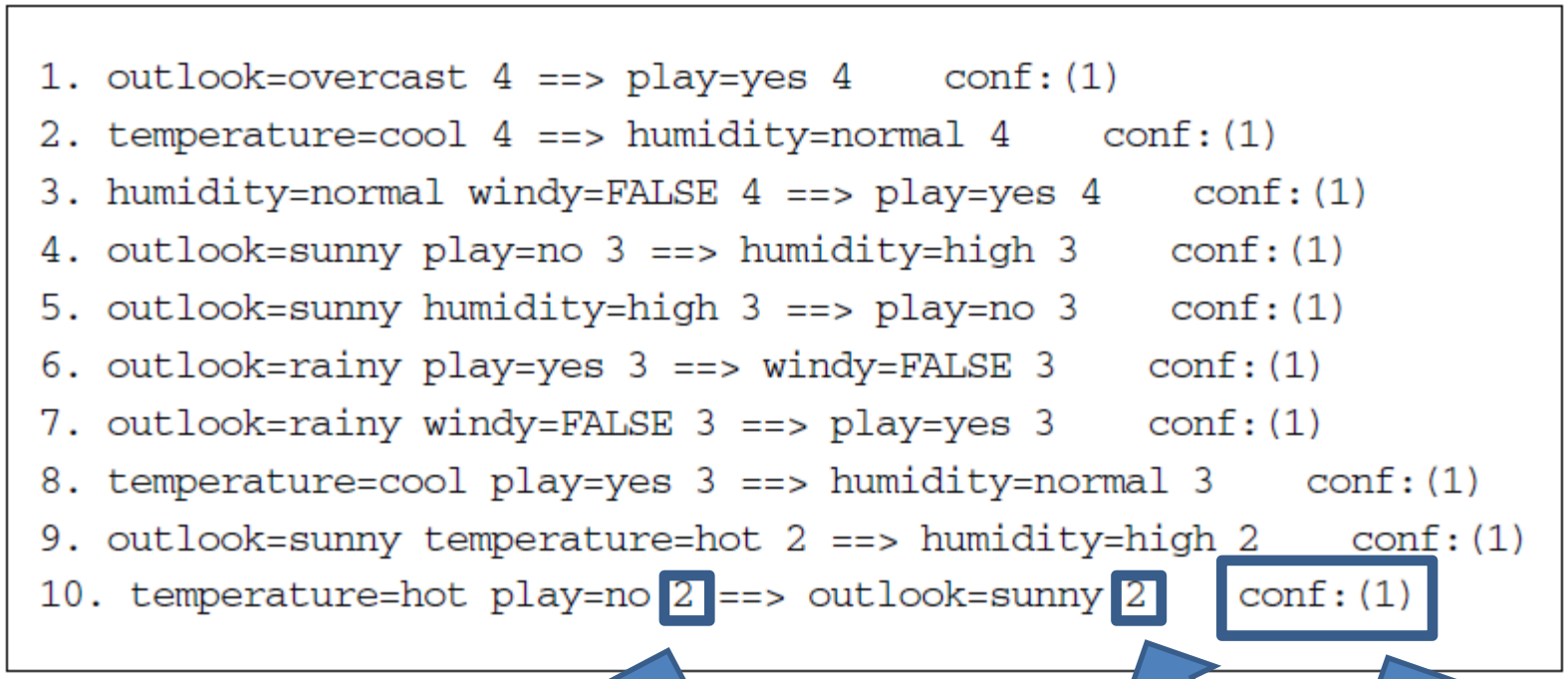

Figure 10.15 Output from the A iori program for associati rules.

Αριθμός παραδειγμάτων για τα οποία ισχύει το αριστερό μέρος του κανόνα

Αριθμός παραδειγμάτων για τα οποία ισχύει και το αριστερό και το δεξί μέρος του κανόνα

Confidence (Accuracy): Ο λόγος των δύο

Μετρικές για την ταξινόμηση των κανόνων (Y ⇒ A)

- **Confidence:** το ποσοστό των παραδειγμάτων που καλύπτονται από την υπόθεση, που καλύπτονται και από την απόδοση. Οι κανόνες συσχέτισης της κλάσης ταξινόμησης (Class association rules) μπορούν να ανιχνευθούν μόνο με χρήση του confidence. **Conf = P(Y,A) / P(Y)**
- **Lift:** το confidence διαιρεμένο με τον αριθμό όλων των παραδειγμάτων για τα οποία ισχύει η απόδοση. Μετράει την σημαντικότητα της συσχέτισης ανεξάρτητα από το support. **Lift = P(Y,A) / (P(Y)*P(A))**

Τιμή Lift=1 σημαίνει ότι η υπόθεση και η απόδοση είναι ανεξάρτητες μεταξύ τους.

- **Leverage:** το ποσοστό των επιπλέον παραδειγμάτων που καλύπτονται και από την υπόθεση και από την απόδοση πέρα από αυτά που θα αναμένονταν αν η υπόθεση και η απόδοση ήταν ανεξάρτητες μεταξύ τους. Ο συνολικός αριθμός αυτών των παραδειγμάτων εμφανίζεται σε παρενθέσεις μετά το leverage. **Lev = P(Y,A) - (P(Y)*P(A))**
- **Conviction:** επίσης ένα μέτρο απόστασης από την ανεξαρτησία. Λαμβάνει υπόψη του την επίδραση του να μην ισχύει η απόδοση.

Conv = P(Y)*P(NOT A) / P(Y,NOT A)