

Principles of Knowledge Discovery in Databases

Fall 1999

Chapter 4: Data Mining Operations

Dr. Osmar R. Zaiane



University of Alberta

Source:
Dr. Jiawei Han

© Dr. Osmar R. Zaiane, 1999

Principles of Knowledge Discovery in Databases

University of Alberta

1

Summary of Last Chapter

- What is the motivation behind data preprocessing?
- What is data cleaning and what is it for?
- What is data integration and what is it for?
- What is data transformation and what is it for?
- What is data reduction and what is it for?
- What is data discretization?
- How do we generate concept hierarchies?

© Dr. Osmar R. Zaiane, 1999

Principles of Knowledge Discovery in Databases

University of Alberta

2

Course Content

- Introduction to Data Mining
- Data warehousing and OLAP
- Data cleaning
- **Data mining operations**
- Data summarization
- Association analysis
- Classification and prediction
- Clustering
- Web Mining
- Similarity Search
- *Other topics if time permits*



© Dr. Osmar R. Zaiane, 1999

Principles of Knowledge Discovery in Databases

University of Alberta

3

Chapter 4 Objectives

Realize the difference between data mining operations and become aware of the process of specifying data mining tasks.

Get an brief introduction to a query language for data mining: DMQL.

© Dr. Osmar R. Zaiane, 1999

Principles of Knowledge Discovery in Databases

University of Alberta

4

Data Mining Operations Outline



- What is the motivation for ad-hoc mining process?
- What defines a data mining task?
- Can we define an ad-hoc mining language?

© Dr. Osmar R. Zaiane, 1999

Principles of Knowledge Discovery in Databases

University of Alberta

5

Motivation for ad-hoc Mining

- Data mining: an interactive process
 - user directs the mining to be performed
- Users must be provided with a set of **primitives** to be used to communicate with the data mining system.
- By incorporating these primitives in a **data mining query language**
 - User's interaction with the system becomes more flexible
 - A foundation for the design of graphical user interface
 - Standardization of data mining industry and practice

Source:JH

© Dr. Osmar R. Zaiane, 1999

Principles of Knowledge Discovery in Databases

University of Alberta

6

Data Mining Operations Outline



- What is the motivation for ad-hoc mining process?
- What defines a data mining task?
- Can we define an ad-hoc mining language?

What Defines a Data Mining Task ?

- Task-relevant data
- Type of knowledge to be mined
- Background knowledge
- Pattern interestingness measurements
- Visualization of discovered patterns

Task-Relevant Data

- Database or data warehouse name
- Database tables or data warehouse cubes
- Condition for data selection
- Relevant attributes or dimensions
- Data grouping criteria



Source:JH

Types of Knowledge to Be Mined

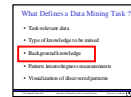
- Characterization
- Discrimination
- Association
- Classification/prediction
- Clustering
- Outlier analysis
- and so on ...



Source:JH

Background Knowledge

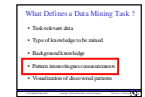
- Concept hierarchies
 - schema hierarchy
 - Ex. street < city < province_or_state < country
 - set-grouping hierarchy
 - Ex. {20-39} = young, {40-59} = middle_aged
 - operation-derived hierarchy
 - e-mail address, login-name < department < university < country
 - rule-based hierarchy
 - low_profit (X) <= price(X, P1) and cost (X, P2) and (P1 - P2) < \$50



Source:JH

Pattern Interestingness Measurements

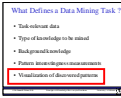
- Simplicity
 - Ex. rule length
- Certainty
 - Ex. confidence, $P(A|B) = \text{Card}(A \cap B) / \text{Card}(B)$
- Utility
 - potential usefulness
 - Ex. Support, $P(A \cup B) = \text{Card}(A \cap B) / \# \text{ tuples}$
- Novelty
 - not previously known, surprising



Source:JH

Visualization of Discovered Patterns

- Different background/purpose may require different form of representation
 - Ex., rules, tables, crosstabs, pie/bar chart, etc.
- Concept hierarchies is also important
 - discovered knowledge might be more understandable when represented at high concept level.
 - Interactive drill up/down, pivoting, slicing and dicing provide different perspective to data.
- Different knowledge requires different representation.



Data Mining Operations Outline



- What is the motivation for ad-hoc mining process?
- What defines a data mining task?
- Can we define an ad-hoc mining language?

A Data Mining Query Language (DMQL)

- Motivation
 - A DMQL can provide the ability to support ad-hoc and interactive data mining.
 - By providing a standardized language like SQL, we hope to achieve the same effect that SQL have on relational database.
- Design
 - DMQL is designed with the primitives described earlier.

Syntax for DMQL

- ❖ Syntax for specification of
 - task-relevant data
 - the kind of knowledge to be mined
 - concept hierarchy specification
 - interestingness measure
 - pattern presentation and visualization
- ❖ Putting it all together — a DMQL query

Syntax for Task-relevant Data Specification

- *use database* database_name,
or *use data warehouse* data_warehouse_name
- *from relation*(s)/cube(s) [*where* condition]
- *in relevance* to att_or_dim_list
- *order by* order_list
- *group by* grouping_list
- *having* condition

Syntax for Specifying the Kind of Knowledge to be Mined

➤ Characterization

mine characteristics [*as pattern_name*]
analyze measure(s)

➤ Discrimination

mine comparison [*as pattern_name*]
for target_class *where* target_condition
{*versus contrast_class_i* *where*
contrast_condition_i}
analyze measure(s)

Syntax for Specifying the Kind of Knowledge to be Mined

➤ Association

mine associations [*as pattern_name*]

Source:JH

© Dr. Omar R. Zaiane, 1999

Principles of Knowledge Discovery in Databases

University of Alberta

19

Syntax for Specifying the Kind of Knowledge to be Mined (Cont.)

➤ Classification

mine classification [*as pattern_name*]
analyze classifying_attribute_or_dimension

➤ Prediction

mine prediction [*as pattern_name*]
analyze prediction_attribute_or_dimension
{*set* {*attribute_or_dimension_i* = *value_i*}

Source:JH

© Dr. Omar R. Zaiane, 1999

Principles of Knowledge Discovery in Databases

University of Alberta

20

Syntax for Concept Hierarchy Specification

- To specify what concept hierarchies to use
use hierarchy <*hierarchy*> *for* <*attribute_or_dimension*>
- We use different syntax to define different type of hierarchies
 - schema hierarchies
define hierarchy *time_hierarchy* *on date* *as* [*date,month quarter,year*]
 - set-grouping hierarchies
define hierarchy *age_hierarchy* *for age* *on customer* *as*
 level1: {*young, middle_aged, senior*} < level0: all
 level2: {20, ..., 39} < level1: *young*
 level2: {40, ..., 59} < level1: *middle_aged*
 level2: {60, ..., 89} < level1: *senior*

Source:JH

© Dr. Omar R. Zaiane, 1999

Principles of Knowledge Discovery in Databases

University of Alberta

21

Syntax for Concept Hierarchy Specification (Cont.)

- operation-derived hierarchies
define hierarchy *age_hierarchy* *for age* *on customer* *as*
{*age_category*(1), ..., *age_category*(5)} := *cluster*(*default, age, 5*) < all(*age*)
- rule-based hierarchies
define hierarchy *profit_margin_hierarchy* *on item* *as*
 level_1: *low_profit_margin* < level_0: all
 if (*price* - *cost*) ≤ \$50
 level_1: *medium-profit_margin* < level_0: all
 if ((*price* - *cost*) > \$50) and ((*price* - *cost*) ≤ \$250)
 level_1: *high_profit_margin* < level_0: all
 if (*price* - *cost*) > \$250

Source:JH

© Dr. Omar R. Zaiane, 1999

Principles of Knowledge Discovery in Databases

University of Alberta

22

Syntax for Interestingness Measure Specification

- Interestingness measures and thresholds can be specified by the user with the statement:
with <*interest_measure_name*> *threshold* = *threshold_value*
- **Example:**
with **support** *threshold* = **0.05**
with **confidence** *threshold* = **0.7**

Source:JH

© Dr. Omar R. Zaiane, 1999

Principles of Knowledge Discovery in Databases

University of Alberta

23

Syntax for Pattern Presentation and Visualization Specification

- We have syntax which allows users to specify the display of discovered patterns in one or more forms.
display as <*result_form*>
- To facilitate interactive viewing at different concept levels, the following syntax is defined:

Multilevel_Manipulation ::= *roll up on* *attribute_or_dimension*
| *drill down on* *attribute_or_dimension*
| *add* *attribute_or_dimension*
| *drop* *attribute_or_dimension*

Source:JH

© Dr. Omar R. Zaiane, 1999

Principles of Knowledge Discovery in Databases

University of Alberta

24

Putting It All Together: the Full Specification of a DMQL Query

```
use database OurVideoStore_db
use hierarchy location_hierarchy for B.address
mine characteristics as customerRenting
analyze count%
in relevance to C.age, I.type, I.place_made
from customer C, item I, rentals R, items_rent S, works_at W, branch
where L.item_ID = S.item_ID and S.trans_ID = R.trans_ID
    and R.cust_ID = C.cust_ID and R.method_paid = "Visa"
    and R.empl_ID = W.empl_ID and W.branch_ID = B.branch_ID and
    B.address = "Alberta" and I.price >= 100
with noise threshold = 0.05
display as table
```

© Dr. Omar R. Zaiane, 1999

Principles of Knowledge Discovery in Databases

University of Alberta

25

Designing Graphical User Interfaces Based on a Data Mining Query Language

- ❖ Data collection and data mining query composition
- ❖ Presentation of discovered patterns
- ❖ Hierarchy specification and manipulation
- ❖ Manipulation of data mining primitives
- ❖ Interactive multi-level mining
- ❖ Other miscellaneous information

Source:JH

© Dr. Omar R. Zaiane, 1999

Principles of Knowledge Discovery in Databases

University of Alberta

26

Summary: Five Primitives for Specifying a Data Mining Task

- task-relevant data
 - database/date warehouse, relation/cube, selection criteria, relevant dimension, data grouping
- kind of knowledge to be mined
 - characterization, discrimination, association...
- background knowledge
 - concept hierarchies...
- interestingness measures
 - simplicity, certainty, utility, novelty
- knowledge presentation and visualization techniques to be used for displaying the discovered patterns
 - rules, table, reports, chart, graph, decision trees, cubes ...
 - drill-down, roll-up,....

Source:JH

© Dr. Omar R. Zaiane, 1999

Principles of Knowledge Discovery in Databases

University of Alberta

27