Mining Association Rules From Relational Data

Average Distance Based Method

Vladimír Bartík, Jaroslav Zendulka

Faculty of Information Technology

Brno University of Technology,
Czech Republic

Outline

- Basic terms
- Related work
  - Current state of the problem
- Average distance based method
  - Method overview
- Implementation, experimental results
- Conclusion and future works

An Association Rule (in Transactional Database)

- A transactional database
  - A set of transactions, where a transaction is a set of items
- **Association rule** is a rule of a form A⇒B, where A and B are sets of items
- Interpretation of an association rule
  - *"If a transaction contains a set of items A, it is likely to contain a set of items B"*

Association Rules - measures

- **Support**
  - Probability that items from $A \cup B$ occur together in a transaction
- **Confidence**
  - Conditional probability that transaction contains the set B provided that it contains the set A
- Minimal support and confidence threshold are used to eliminate uninteresting rules

Association Rules - terms

- **Frequent itemset**
  - An itemset, which satisfies the condition of minimal support
- **Strong association rule**
  - An association rule, which satisfies the condition of minimal support and minimal confidence

Association Rules in Relational Databases

- Two types of attributes
  - **Categorical attributes**
    - » e.g. town, job …
    - » Finite set of possible values
    - » Some of well-known methods can be used
  - **Quantitative attributes**
    - » e.g. age, price
    - » Infinite set of possible values
    - » Implicit ordering is defined

Related work - I

- Quantitative association rules
  - Contain predicates of a form *(Attr=val)* or *intervals*
  - A measure *"K-partial completeness"*: ensures that intervals are not too large and too small
  - Consecutive joining intervals into larger
  - Disadvantage: doesn't respect the semantics of data (needs initial equi-depth discretization)

Related work - II

- Distance based methods
  - Using of clustering methods to find intervals into association rules
  - 1st step: Find clusters of quantitative values
  - 2nd step: Create the association rules
  - The semantics of data is respected

Average Distance Based Method

- Average distance
  - Searching for a value *v*, which has a number of neighbors in a short distance defined by a measure
  - Definition:

$$\boxed{AD(v) = \sum_{i=1}^{n} \frac{(v - v_i)}{n}}$$

  » n…number of neighbors

  » v_i…neighbors

Average Distance Based Method

- Average Distance
  - Number of neighbors can be counted from the minimum support threshold

$$\boxed{n = minsupp * numrows}$$

    » numrows … number of rows in a table

  - Value of *maximal average distance* (MaxAD) must be entered for each quantitative attribute in a relational table as a parameter

Average Distance Based Method

- Precision (P)
  - Value used to choose the values *v*
  - Determines the precision of values in resultant association rules

    Ex.: If P=2, values of quantitative attributes contained in association rules will be even
  - Determines number of steps of the algorithm
  - Also must be entered for each quantitative attribute in a table

Average Distance Based Method

- Method overview
  - **Categorical attributes processing** – discovery of frequent itemsets from data in categorical columns
  - **Quantitative attributes processing** – process the attributes one by one …
  - **Association rules generating** – create association rules from frequent itemsets

Example – I
(Categorical attributes processing)

Age	Salary	Car	Country
19	15000	VW Golf	Czech Rep.
20	20000	Opel Astra	Germany
44	14000	Ferrari	Germany
20	21000	VW Golf	Czech Rep.
:	:	:	:

Quantitative attributes

Categorical attributes

Frequent itemsets after categorical attributes processing:

{ Car = ‘VW Golf’, Country = Czech Rep.}
{ Country = Germany}
…

Quantitative attributes processing

- Result: Set of frequent itemsets containing values of both categorical and quantitative attributes
- For each frequent itemset and each quantitative attribute, the values of an attribute are discretized and found interesting values are added to an itemset
- Steps:

1. Construction of a number sequence
2. Searching for interesting values
3. Adding the interesting values to the frequent itemset

Example – II
(Construction of a Number Sequence)

Age	Salary	Car	Country
19	15000	VW Golf	Czech Rep.
20	20000	Opel Astra	Germany
44	14000	Ferrari	Germany
20	21000	VW Golf	Czech Rep.
:	:	:	:

FI: {Car = ‘VW Golf’, Country = Czech Rep.}

Number sequence:

17	19	19	19	20	21	24	33	34

Added values

Example – III
(Searching for Interesting Values)

MaxAD=1
P=1; n=6

Number sequence:

17	**19**	**19**	**19**	**20**	**21**	**24**	**33**	**34**

v = 19 => AD=0.83 => is interesting

v = 20 => AD = 1.17 => is not interesting

New frequent itemset:

{Car = ‘VW Golf’, Country = Czech Rep., Age = 19}

Frequent Itemsets With Several Quantitative Attributes

1. Assume that we have a frequent itemset FI containing one quantitative value. We will denote c the cluster of values represented by a value v.

2. Choose a quantitative attribute, which has not been processed yet.

3. Values of a selected quantitative attribute from rows in which values of categorical attributes correspond to the values in FI are stored into an ordered number sequence

Frequent Itemsets With Several Quantitative Attributes

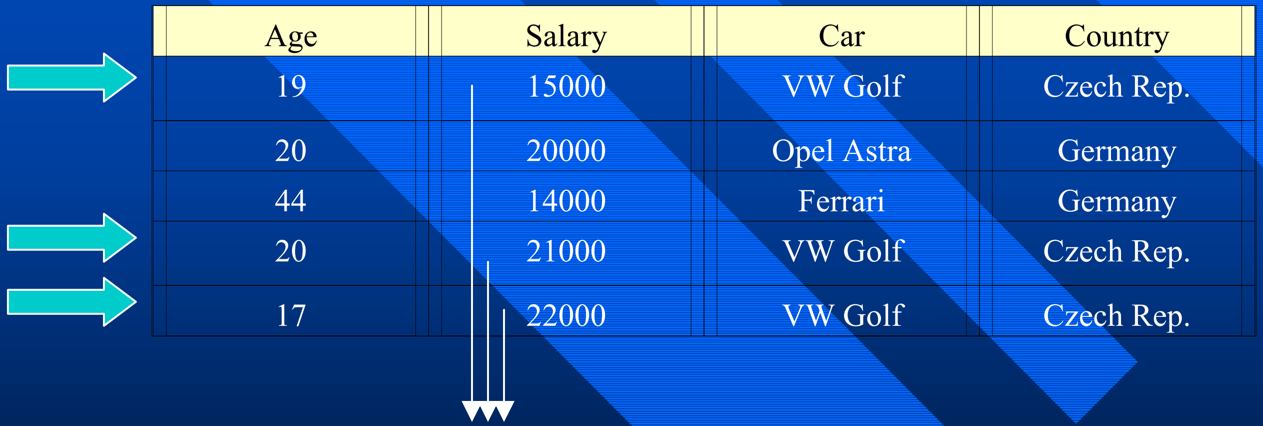
4. Find interesting values in this number sequence
5. Create new frequent itemsets by adding new interesting values to the actual frequent itemset FI.
6. Repeat the steps 2-5 for each quantitative attribute in the table, which has not been processed yet.

Example - IV

FI: {Car = 'VW Golf', Country = Czech Rep., Age = 19}

Cluster representing the item "Age = 19":

c=17..21



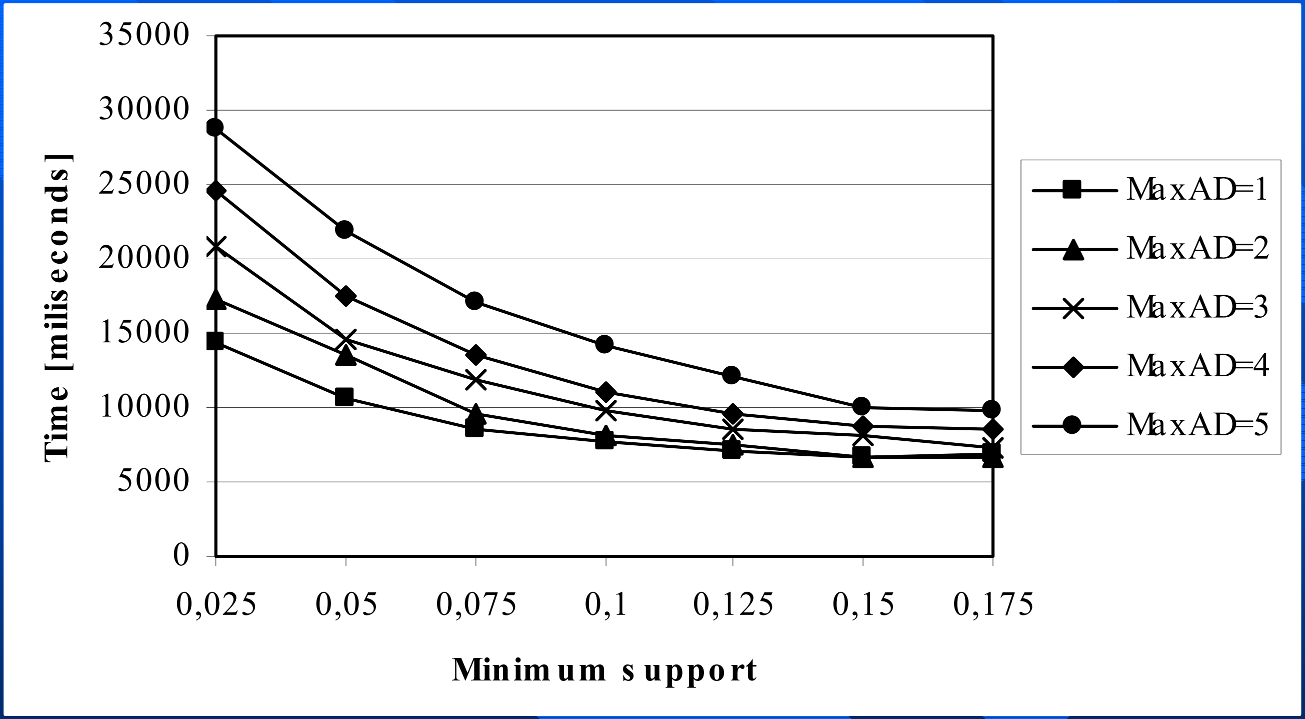
Age	Salary	Car	Country
19	15000	VW Golf	Czech Rep.
20	20000	Opel Astra	Germany
44	14000	Ferrari	Germany
20	21000	VW Golf	Czech Rep.
17	22000	VW Golf	Czech Rep.

Values stored to the number sequence

Experiments

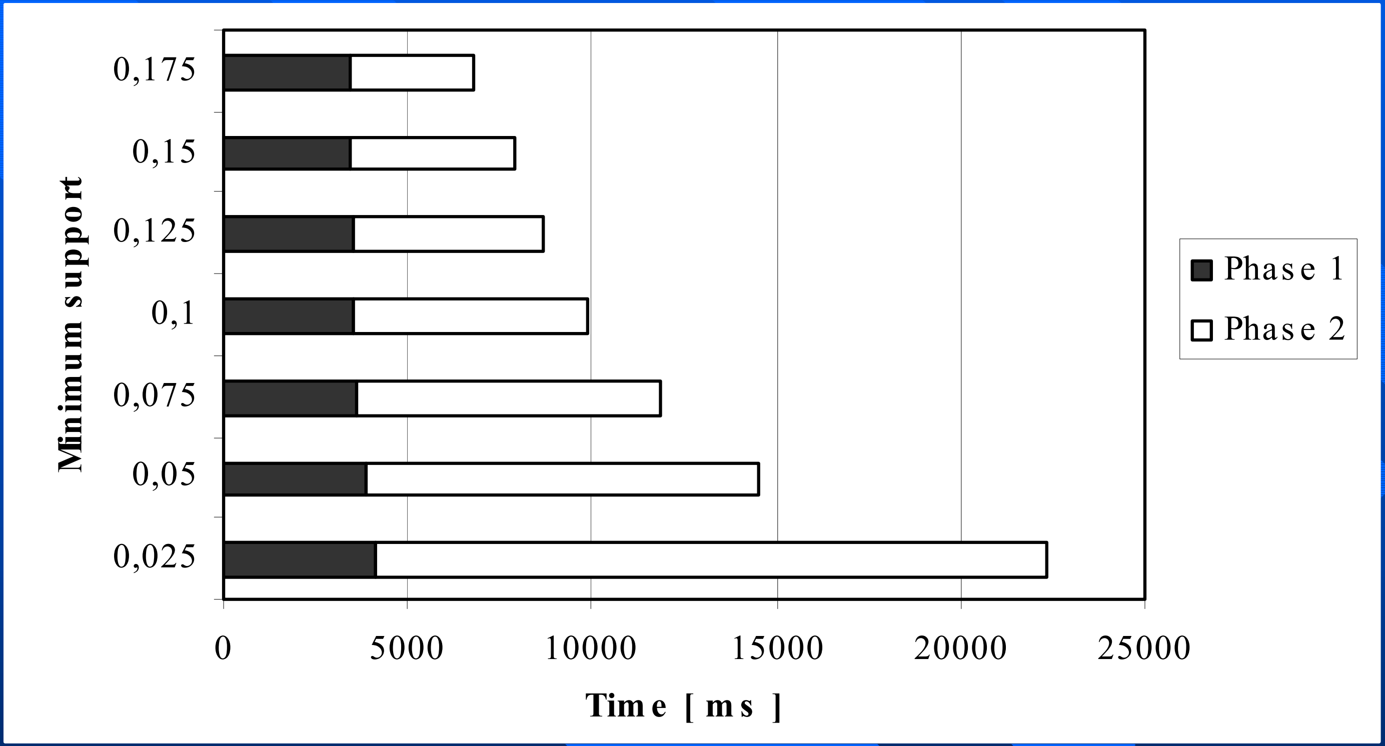
- Method was implemented in Java
- Experiments:
  - Data from a medical study
  - Data mining task contained:
    - 2 categorical attributes (frequency of drinking alcohol, physical activity in work)
    - 2 quantitative attributes (weight, systolic blood pressure)

Experimental results - I



Dependency between time, minimum support and maximal average distance

Experimental results - II



Dependency between time of phase 1 and phase 2

Phase 1 – categorical attributes processing
Phase 2 – quantitative attributes processing

Experimental results - III

- Order of quantitative attributes processing
  - It is better to process first:
    - » Attributes with lower precision
    - » Attributes with higher maximal average distance
- Solution: A heuristics (H)

$$\boxed{H = \frac{MaxAD}{P} \cdot \left(1 - \frac{missing}{numrows}\right)}$$

  - *missing:* number of missing values of the attributes
  - *numrows:* number of rows in the table

Conclusion

- Advantages of the method
  - Separation of categorical and quantitative attributes processing => may be more effective
  - Quantitative items in association rules can be in the form *(Attr=val)* => may be more useful information
- Future works
  - Find suitable data structures to store number sequences and effective algorithm for searching in them
  - Comparison with other methods