Mining Multi-Dimensional Quantitative Associations

Michał Okoniewski, Lukasz Gancarz, Piotr Gawrysiak

Institute of Computer Science, Warsaw University of Technology
ul. Nowowiejska 15/19, 00-665 Warsaw, Poland
{okoniews,lgancarz,gawrysia}@ii.pw.edu.pl

Abstract. The new form of quantitative and multi-dimensional association rules, unlike other approaches, does not require the discretization of real value attributes as a preprocessing step. Instead, associations are discovered with data-driven algorithms. Thus, such rules may be considered as a good tool to learn useful and precise knowledge from scientific, spatial or multimedia data, because data-driven algorithms work well with any sampling method. This paper presents the whole methodology of automatic discovery of new rules that includes theoretical background, algorithms, complexity analysis and postprocessing techniques. The methodology was designed for a specific telecom research problem, but it is expected to have a wide range of applications.
Keywords: Data Mining, Quantitative Association Rules, Automatic Analysis of Numeric Data, Knowledge Discovery Process, Multi Dimensional Database Indexing

1 Introduction

Problem statement. Association rules are a highly popular data mining method. However, most of the approaches are designed for .market basket analysis. and operate on categorical (qualitative) data. It renders them useless for learning from many common types of data based on numeric values. Special forms of association rules for quantitative attributes may be applicable here. There are only few algorithms and methodologies to deal with quantitative associations [8, 4, 5].
Most of them are based on some form of discretization labeled as partitioning, quantization or bucketing of numeric attributes, which means dividing the attribute domain into separate ranges. Such preprocessing method results in the loss of informational value of discovered rules, or often looses significant ones.
There are lots of examples of misleading discretization. The paper [9] gives a proof that the problem of optimal quantization of real value attributes is NP-hard.
Data-driven algorithms are expected to be competitive to those based on discretization. An example of such algorithm is Window algorithm proposed

in [4] for their new form of a quantitative rule. In Window, the boundaries of ranges in the antecedent of an association rule are determined by attribute values for specific tuples. A set of these ranges, called a profile, selects a subset of tuples. The antecedent consists of a statistical measure (usually the mean), which is based on values of another numeric attribute. The measure for the subset is compared with the same measure for the whole relation. The rule is significant if the difference between these two measures is high. An example of the rule discovered by this algorithm would be:

$$Age \in (35,45) \Rightarrow$$
$$Salary : mean = 38k \ (overall \ \ mean \ \ Salary = 29k)$$

In [4] only rules with single numeric attribute in the antecedent are presented. This paper describes a generalization of this solution to multiple attributes.

The main task of this methodology is the automatic discovery of or hyper-cuboid sub-spaces that have significantly different qualities from the whole space. It may be useful for intelligent analysis of maps, continuous processes or even multimedia. Consequently, this paper discusses various aspects of such multidimensional quantitative rules.

Contribution. Definitions of quantitative association rules of Aummann and Lindell type are extended in this paper to fit multiple attributes. This allows us to construct the algorithm for discovering such rules. The algorithm utilizes specific rule properties described in theorems. We have also proposed some variations of the algorithm enhanced by heuristic strategies and advanced database indexing. The whole methodology is completed with proposition of post-processing techniques with the use of similarity and significance measures. Finally, the paper justifies that the methodology may be applied to real-world databases, utilizing our experience from the telecom GIS mining research project [2].

Outline. The rest of the paper is organized as follows. Section 2 presents definition and properties of extended quantitative rule model. This properties are used for algorithms and strategies in Section 3, while their indispensable part: multi-attribute database indexes is described in Section 4. Section 5 presents a general outlook on the rule discovery cycle, which includes definitions of similarity and importance measures used for rule management. Section 6 discusses two examples of application in knowledge discovery from a telecom company GIS database. Section 7 contains conclusions and recommendations for future work.

2 Definitions

In [4] rules with single numeric attribute in both antecedent and consequent are presented. In this paper we consider their generalized forms. Thus,

definitions included in this section are multi-dimensional extensions of definitions for .Quantitative toQuantitative. rule from [4].

Notations. Let D be a relational table with a set of quantitative attributes $E = \{I_1, I_2,\ldots,I_k, J\}$. Letters A, B, ... mean single attributes from E, while X, Y, ... mean subsets of E. Table D may be viewed as a set of tuples $D = \{t_1, t_2,\ldots,t_n\}$. Notation $t_i.A$ indicates the value of attribute A for tuple i. A range (A, a, b) is defined by a single attribute $A \in E$ and two numbers $\{a,b\} \in domain(A) \subseteq \mathbf{R}$, $a \le b$. A profile Pr_X over $X \subseteq E$ is defined as a common part of ranges $\bigcap_{i \in X} ia_i, b_i$ - one range for each attribute in X. Notation $(A,a,b) \in \mathrm{Pr}_x$ means that range (A, a, b) is one of the ranges that delimit Pr_X. Basically, a profile may be simply viewed as a k-dimensional hyper-cuboid. $|\mathrm{Pr}_X|$ is a number of tuples from D that have all corresponding attribute values within profile Pr_X. A statistical measure M is defined over distribution of attribute J values. $M(Pr_X)$ is a value of this measure for distribution of J for tuples that have all corresponding attribute values within Pr_X. In addition, $M(D)$ is the measure value for distribution of J attribute values for the whole D. As in [4], the measure M is usually the mean of J values. With the use of above notations we can build up a definition of generalized "Quantitative to Quantitative" rule.

Definition 21 Multi-dimensional (mean based) quantitative association rule is a rule of the form:

$$\mathrm{Pr}_x \Rightarrow M(\mathrm{Pr}_x)(M(D))$$

where:

- $J \notin X$
- $M(\mathrm{Pr}_x) - M(D)) \ge mindif$
- $|\mathrm{Pr}_x| \ge minsup$

The antecedent of the rule is a profile that defines a sub-population of tuples that is significantly different from the whole D with regard to the attribute J. It is assured by the second condition (a difference condition) that holds if there is a minimal difference *mindif* between the measure for D and for the Pr_X. In [4] standard methods for statistical hypothesis testing were then applied (e.g. a Z-test for the mean) to check the significance of the difference. The third condition is a standard support requirement for an association rule. Constants *mindif* and *minsup* are user-defined parameters. There is no confidence parameter of the rule. The rule has the difference parameter $dif = M(Pr_X) - M(D)$ instead, to indicate its strength. Let us here specify minimal M for a rule by $\mu = M(D) + mindif$. The dimensionality of the rule is equal to the number of attributes in its profile.

Remark. Definition 21 describes a rule that has the mean above average $(M(PrX) > M(D))$. The work in this paper considers above-average rules that

follow this definition. All this may be also applied by the simple analogy for below-average rules.

Examples of quantitative rules are:

$$cigarettes\ daily \in (10,\ 20) \wedge overweight \in (10;\ 20) \Rightarrow life\ expectancy = 58\ (life\ expectancy = 72)$$

$$latitude \in (49N,\ 50N) \wedge longitude \in (19E,\ 21E) \Rightarrow AprAvgTemp = 3^{\circ}C\ (AprAvgTempPoland = 7^{\circ}C)$$

Now we can define rules with profiles contained in other rules. profiles and basic rules:

Definition 22 (Sub-rules) The rule $Pr_X \Rightarrow M(Pr_X)$ is contained in $Pr_Y \Rightarrow M(Pr_Y)$ (i.e. is a sub-rule), if $Y \subseteq X$, and for each attribute B with the range $(B,\ a,\ b) \in Pr_X$ exist such $c,\ d$, that $(B, c,\ d) \in PrY \wedge c \leq a \leq b \leq d$

Definition 23 (Basic rule) Basic rule is not contained in any other rule.

Other important notions are irreducible and maximal rules:

Definition 24 (Irreducible rule) The rule $Pr_X \Rightarrow M(Pr_X)$ is irreducible, if for every range $(A,\ a,\ b) \in Pr_X$ and every number c, a $<$ c $<$ b is true as follows: profiles Pr_{X1} and Pr_{X2} that are created by exchanging $(A,\ a,\ b)$ in Pr_X respectively with ranges $(A,\ a,\ c)$ and $(A,\ b,\ c)$ result in rules $Pr_{X1} \Rightarrow M(Pr_{X1})$ and $Pr_{X2} \Rightarrow M(Pr_{X2})$ that fulfill at least the difference condition from definition 21.

Definition 25 (Maximal rule) The rule $Pr_X \Rightarrow M(Pr_X)$ is a maximal rule, if for every range (A, a, b) $\in$ $\mathrm{Pr_X}$ and every c, c $>$ b(c $<$ a) the rule which is created by exchanging range (A, a, b) in the input rule with range (A, a, c) ((A, c, b)) does not fulfill the difference condition from definition 21 or is reducible.

Accordingly, irreducible rule profile may be divided by any hyperplane $A = c$ into two profiles, that maintain above-average difference condition. As it is pointed out in section 5 in multiple dimensions irreducibility is not good enough to fit the intuitive connotation of a homogeneous rule. However, irreducibility is a basic quality that makes the rule desired.

Maximal rule is one that can not be extended into a single dimension. Nonetheless, it may be extended into two or more dimensions by enlarging more than one range from $\mathrm{Pr_X}$. That is why definition of maximality is useful mainly for one-dimensional rules.
Let us present two theorems that describe properties of quantitative rules

and are essential for discovering them.

Theorem 21 If the quantitative association rule $Pr_X M \Rightarrow (Pr_Y)$ is irreducible, then

$$\forall_{(A,a,b)\in \Pr_X} \exists_{t_1,t_2\in D}$$
$$t_1.A = a \wedge t_2.A = b \wedge t_1.J \geq \mu \wedge t_2.J \geq \mu$$

Proof. This theorem states that on every profile boundary of irreducible rule is a tuple (called *μ-tuple*), that has J value above μ. Let us assume that, on the contrary, there is a plain, below-average tuple that is closer to profile boundary than a μ-tuple. Then we can draw a division line between the tuple and the rest of the profile along the boundary. As a result the part with this single tuple is below average, so the whole profile can not be irreducible rule. The practical consequence of this theorem is that μ-tuples with maximal and minimal $\Pr_X$ attribute values define the profile area of the rule.

Theorem 22 There are minimum *2*, maximum *2k* μ-tuples to define a profile of the irreducible rule.

Proof. The profile of the rule is a hyper-cuboid with *2k* faces and 2^k vertexes. A single μ -tuple defines maximum of k faces, if is in one of vertexes. If the μ-tuple is neither a vertex nor an edge, it defines only 1 face. Hence, a minimum of 2 μ-tuples is needed (in opposite vertexes) and maximum *2k* μ-tuples one in each face of the hyper-cuboid.
For example, a profile in two dimensions is defined by 2,3 or 4 μ-tuples (Fig. 1).

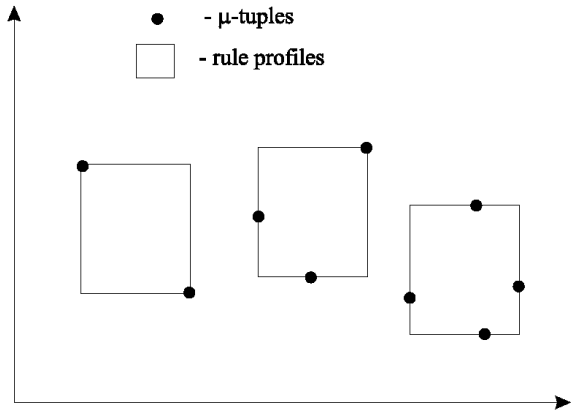

Fig. 1. μ-tuples that define a profile

3 Mining algorithm and cost reduction strategies

The general outline of the mining algorithm that utilizes theorems 21 and 22 is as follows. First, select all the μ-tuples. Then consider the μ-tuples in sets from minimum 2 to maximum $2k$ elements. For each set find minimal and maximal values for each attribute $I_1, \ldots, I_k$. In this way we obtain suspected profile boundaries. Then, check the irreducibility of the profile by incrementally checking divisions of the profile into two hyper-cuboids with all included tuples. All the hyper-cuboids have to be above average in terms of the mean J value. If only one of the checks fails, the profile does not form a rule and so may be rejected.
The algorithm may be sketched in a recursive form:

```
1 int CheckMiTuples (int level, int TupleNo)
2 {
3   if (level < 2*k)
4   {
5     for(int i=TupleNo+1; i<=MiTuplesQ,i++)
6     {
7       AddTuple(MiTupleSet,i);
8       if (level>1) CheckProfile(MiTupleSet);
9       CheckMiTuples(level+1,i);
10    }
11   }
12 }
13 FindQRules()
14 {
15    SelectAndSortMiTuples();
16    CheckMiTuples(1,0);
17 }
```

The computational complexity of above algorithm depends on the percentage p of μ-tuples in the database, and may be estimated [7] as $O(k(pn)^{2k})$. This assumes that the cost of selecting tuples inside a profile hyper-cuboid is small, because of effective indexing method for k attributes.
The complexity is polynomial, but may be still considered high. However if we divide the attribute space into r sub-spaces, the complexity is decreased to:

$$O\left(rk\left(\frac{pn}{r}\right)^{2k}\right)=O\left(k\frac{(pn)^{2k}}{r^{2k-1}}\right) \qquad (1)$$

There are several complexity reduction strategies that are based on this property. For example μ-tuples may be divided into r disjoint groups with some clustering algorithm in order to search for rules within clusters. Alternatively, if we assume that rules can not be found in regions where there are no μ-tuples, excluding one most spacious "empty" hyper-cuboid results in division into 2^k smaller search spaces.

Another cost reduction strategy is based on the construction of sorting function from line 15. If we sort μ-tuples by the descending value of attribute J we can expect most significant rules to be found first.

4 The use of multi-dimensional indexes

Presented in this paper discovery of quantitative associations in spatial data is related only to point objects what enables data to be efficiently stored in relational DBMS. The problem of fast selection of tuples within a rectangle that in.uences the complexity of the algorithm may be solved by the latest advances in the area of multi-dimensional indexing.

There have been many attempts to provide efficient method for management of multi-dimensional data as a result of increasing interest in GIS or VLSI CAD, etc. Broad research in the area resulted in proposal of many multidimensional access methods. [3] gives survey of almost all of these techniques, presenting requirements that such access methods should meet and upon which they are evaluated. A classification of point access methods for storing point objects and classification of spatial access methods for storing objects with spatial extensions are also presented.

Analysis of processing characteristics of the proposed algorithm implies that the most essential subjects to optimization of operations are exact match query and region search query. Since algorithm for mining quantitative associations deals only with point data, one of point access methods may be utilized to improve performance.

A very promising multi-dimensional point access method that could improve the proposed algorithm performance is UB-tree [1]. It uses a space filling curve to create a partitioning of the space while preserving multi-dimensional clustering. As a result of disjoint partitioning of a multi-dimensional space and utilization of B-tree to store Z-addresses, UB-tree provides logarithmic performance, guaranteed for insertion, deletion and exact match queries.

Answering a range query over a database, which is organized as a UB-Tree, requires time proportional to the size of the answer to the query. In fact, this is another data-driven solution in our methodology. The most crucial task of the range query algorithm is to calculate the next region in Z-order which intersects the query box after a Z-region (which also intersects the query box) that has been retrieved. [6] presents three variations of an algorithm solving that problem with complexity ranging from exponential to linear.

5 Rule management

After discovering rules, they have to be presented in understandable form to the user. It may happen that the number of rules is high or that there is a lot of rules that overlap one another. In such cases there is a demand for a rule

management system that will refine the result and give feedback for subsequent stages of knowledge discovery cycle.
Crucial problems with multi-dimensional quantitative associations is the need to determine the most significant rules and to distinguish between groups of rules with very similar profiles [4]. This problem may be solved by the application of specific quality measures in a rule management system [7].
One kind of measures determines the significance of the rule. Such measure may be support, difference (*diff(P)*), volume (*V(P)*), rule density ($\rho(P)$) or differential density ($\rho diff\ (P)$), defined as follows:

$$diff(P) = M(P) - M(R) \tag{2}$$

$$V(P) = \prod_{i \in \mathrm{Pr}_p} (v_i - u_i) \tag{3}$$

$$\rho(P) = \frac{spp(P)}{V(P)} \tag{4}$$

$$\rho_{diff}(P) = \frac{spp(P)\ diff(P)}{V(P)} \tag{5}$$

As it was noticed in Section 2, irreducibilty is not always enough to determine intuitively homogeneous rules. Figure 2 shows some examples of not uniform distribution of μ-tuples in irreducible *2D* rules. There are various formulas possible for measuring the consistency of the rule. For instance let us consider that the rule profile is divided into 2^k equal hyper-cuboid parts $(\mathrm{Pr}^1, ..., \mathrm{Pr}^{2^k})$ by splitting all attribute ranges in two. Consistence then may be expressed as:

$$cons(P) = \frac{2^k}{\sum_{1 \le i,j \le 2^k} |M(\mathrm{Pr}^i) - M(\mathrm{Pr}^j)|} \tag{6}$$

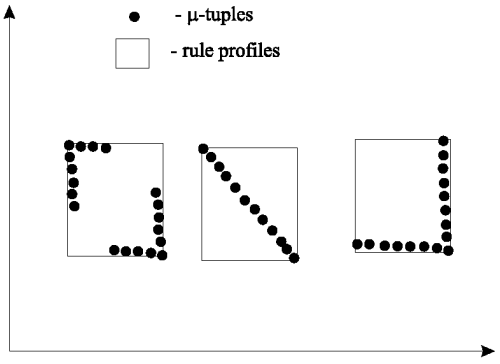

Fig. 2. Rules irreducible, but not consistent

Another kind of measures are used for comparison of rules - to determine if the profiles are close or distant. Such measure may be common support *Csupp* or common volume *CV* or mean intra-rule distance L_{ir} between tuples.

$$Csupp(P_1, P_2) = |\{t : t \in \Pr(P_1) \cap \Pr(P_2)\}| \tag{7}$$

$$CV(P_1, P_2) = V(\Pr(P_1) \cap \Pr(P_1)) \tag{8}$$

$$L_{ir}(P_1, P_2) = \frac{\sum_{i=1}^{N1} \sum_{j=1}^{N2} (t_i^{P1}, t_j^{P2})}{N_1 N_2} \tag{9}$$

6 Applications

The whole idea of quantitative rules research came from a problem of planning cellular radio network according to cellular traffic data from GIS database of a telecom provider [2]. Classic association rules failed here, due to problems with discretization. Thus, quantitative and data-driven associations are an obvious form of knowledge that may be automatically inferred from spatial data. In this section we present two examples of application of a new methodology to raw and preprocessed spatial data.

Raw spatial data. Such data may be sampled, even at random, and used as an input to rule mining algorithm. It is obvious that the frequency of sampling increases rule accuracy and consequently the algorithm running time. Antecedent attributes I_1, *ffi*, I_k are coordinates of points in 2D, 3D or even higher dimensionality space. Decisive attribute J describes the analyzed value (i.e. elevation, temperature, cellular traffic, etc.). As a result we obtain hyper-cuboid regions (squares in 2D) where the value is high above (or below) average for the whole space. As a post-processing step one can use the measures from Section 5 to find significant and representative rules.

Preprocessed spatial data. As described in [2], the space may be divided into regions, for example mobile telecom cells. For each region we can establish a number of numeric parameters (e.g. population or percentages of area types in the cell: forests, urban, water,etc.). For each regionwe obtain a tuple of attributes I_1, *ffi*, I_k that stand for parameters plus one analyzed attribute J. These tuples are the input to rule discovery algorithm. As a result we obtain a rule-based predictive model that may be used for classification of other regions in the space. The predictive model in the telecom research was utilized to determine cellular traffic in newly designed cells.

In a similar way the methodology may be applied to other kinds of numeric data, for example to sampled mutimedia or to readings from sensors in scientific or engineering data sets.

7 Conclusions

Quantitative association rules in the form presented in this paper are applicable to any form of numeric data and have clear advantages. Data-driven algorithms for rule discovery have polynomial complexity, and are additionally sped up by heuristic strategies. Thanks to latest advances in spatial indexing, the rule discovery can be now even more accelerated. Profile boundaries are determined by the data themselves, without errors in.icted by the static discretization. Input data may be sampled even at random. Output rules, especially mean based, are understandable and may be easily visualized because a square or hyper-cuboid is very intuitive in its perception. All the algorithms and strategies are currently under rigorous experimental examination that will be described in some follow-up papers. Other future work in this field includes discovery algorithms with dynamic changes of μ-level, improved performance strategies and new measures for rule management. The knowledge discovery methodology may be even closer linked to spatio-temporal databases by new preprocessing and visualization techniques. It is also expected that quantitative association rules will be applicable to other forms of numeric data.

References

[1] Bayer, R. The universal B-tree for multidimensional indexing, Institut fur Informatic, TUMunchen Technical Report (1996)

[2] Gawrysiak, P., Okoniewski, M. Applying data mining methods for cellular radio network planning, Intelligent Information Systems, Springer-Physica Verlag (2000)

[3] Geade, V., Gunther, O. Multidimensioanal Access Methods, ACM Computing Surveys, 30(2), (1997)

[4] Lindell, Y., Aumann, Y. Theory of Quantitative Association Rules with Statistical Validation, Proceedings of SIGKDD Conference, Boston, (1999)

[5] Miller, R.J., Yang, Y. Association Rules Over Interval Data, Proceedings of ACM SIGMOD 97 Conference (1997)

[6] Markl V. MISTRAL: Processing Relational Queries using a Multidimensional Access Techinque, Ph.D Thesis, Institut f.r Informatic, TUMunchen (1999)

[7] Okoniewski, M. Discovering Quantitative, Multi-dimensional Association Rules, Ph.D. Thesis (draft), Warsaw University of Technology (2001)

[8] Srikant, R., Agrawal, R. Mining Quantitative Association Rules in Large Relational Tables, Proceedings of VLDB-96 Conference (1996)

[9] Skowron, A., Nguyen, S.H. Quantization of Real Value Attributes: Rough Set and Boolean Reasoning Approach, Warsaw University ofTechnolgy Technical Report (1995)