Paper

The use of quantitative association rules in cellular network planning

Michał Okoniewski

Abstract — This paper describes the problem of planning cellular network base stations with optimization to traffic requirements. This research problem was a main incentive to add some development to the theory of association rules. 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. They are discovered with data driven algorithm that gives precise and complete results and has polynomial complexity for a given dimensionality.

Keywords — data mining, quantitative association rules, knowledge discovery process, cellular network planning.

1. Problem statement

One of the most important areas for a young cellular telecom provider is network expansion. This creates a need for traffic prediction, i.e. we would like to estimate the number of calls made during a certain time span, on an area where we want to build a new base station. Such information is crucial for station equipment design – there must be enough transceivers to ensure that every subscriber in the GSM cell created by this station is able to place or receive call. On the other hand, there should not be too much available – and unused – radio channels, because this would mean unnecessary costs.

Traffic prediction is a complex task, as the number of subscribers present on a certain area may vary. After all GSM is an abbreviation of name Global System for Mobile Telecommunications, and GSM subscribers travel between cells, for example moving into city centers at day, and going to suburbs (where their homes are) in the evening. Similar effect can be observed also for longer time periods. So called vacation traffic analysis shows that in the summer average traffic generally increases in popular resort areas – like mountains, seashore etc. Fortunately our analysis showed that these variations are periodic, and predictable, at least for regions with well developed GSM coverage. We can therefore try to predict traffic for a certain characteristic time period – say, for vacation time, using measurements of existing network elements, and than interpolate obtained values.

In this particular experiment we were able to extract two types of information from ERA GSM network monitoring system. First was the traffic information. For each cell we obtained average of weekly traffic measured at busy hour (usually around midday). Other type of information has been extracted from company's geographical information system (GIS). For each cell, the types of terrain occupied have been established. The GIS database contained information about nine terrain types (landuses) that may occur in particular cell. These were:

1) fields,
2) forests,
3) water,
4) rocks, seashores, swamps,
5) roads, concrete, parks,
6) suburbs,
7) urban,
8) dense urban,
9) industrial.

For each cell the amount of ground pixels occupied by every landuse have been defined. One ground pixel width and length are approximately 5 seconds of arc. Our initial data about existing network have been collected and recorded in a table with following attributes:

– cell identification number,
– landuse type 1..9,
– number of pixels occupied by above landuse,
– cell number that allows to determine region in which the cell is situated,
– traffic – the average weekly value in Erlangs for a given cell.

The main goal of the project was to build a predictive model of network traffic that would allow analysts to plan the power of base station in newly created radio network cells. Various methods such as neural networks, clustering, regression or regressional clustering were used [1, 2].

However there were almost no results with classic association rules. It was due to the fact that the quality of data, already not perfect, was made even worse with discretization. It turned out that it's difficult to find a discretization that does not loose informational value of discovered rules. After the research project in thesis [7] there was introduced a new theory of quantitative rules that could be able to cope with the problem using new form of rules and data driven approach instead of discretization.

2. New form of quantitative rules

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 to Quantitative" 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 $A, B, \ldots$ mean single attributes from E, while $X, Y, \ldots$ 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 i tuple. A range (A, a, b) is defined by a single attribute $A \in E$ and two numbers $\{a, b\} \in domain(A) \subseteq \mathbf{R}, a \leq b$. A profile Pr_X over $X \subseteq E$ is defined as a common part of ranges $\bigcap_{i \in X}(a_i, b_i)$ – one range for each attribute in X. Notation $(A, a, b) \in Pr_X$ 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. $|Pr_X|$ 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.
The antecedent of the rule is a profile that defines subpopulation of tuples that is significantly different from the whole D with regard to the attribute J. It is assured by 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. Another condition is a standard support requirement for an association rule.
With the use of above notations we can build up a definition of generalized "Quantitative to Quantitative" rule.

Definition 2.1. Multi-dimensional (mean based) quantitative association rule is a rule of the form:

$$Pr_X \Rightarrow M(Pr_X) \ \ (M(D)),$$

where:

- $X \cap Y = \emptyset$,
- $|M(Pr_X) - MD| \geq mindif$,
- $|Pr_X| \geq minsup$.

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 2.1 describes a rule that has the mean above average ($(M(Pr_X) > M(D))$. 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:

$$cigarettes\_daily \in (10, 20) \wedge overweight \in (10, 20) \Rightarrow life\_expectancy = 58 (life\_expectancy = 72)$$

$$latitude \in (49N, 50N) \wedge longitude \in (19E, 21E) \Rightarrow avg\ temp.April = 3C (avg\ temp\ April\ Poland = 7C)$$

Important notions are irreducible and maximal rules. These are rules that are intuitively "interesting" to discover:

Definition 2.2 (irreducible rule). Rule $Pr_X \Rightarrow M_J(T_{Pr_X})$ is irreducible, if for every range $(A, a, b) \in Pr_X$ and every number c, $a < c < b$ the following is true: profiles $Pr_X 1$ and $Pr_X 2$ that are created by exchanging (A, a, b) in Pr_X respectively with ranges (A, a, c) and (A, b, c) result in rules $Pr_X 1 \Rightarrow M_J(T_{Pr_X 1})$ i $Pr_X 2 \Rightarrow M_J(T_{Pr_X 2})$ that fulfill at least the difference condition from Definition 2.1.

Definition 2.3 (maximal rule). Rule $Pr_X \Rightarrow M_J(T_{Pr_X})$ is a maximal rule , if for every range $(A, a, b) \in Pr_X$ 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 2.1 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 the intuitive connotation of a rule with homogeneous distribution of tuples. However, the irreducibility is a basic quality that makes the rule desired.
Maximal rule is one that cannot be extended into a single dimension. Nonetheless, it may be extended into two or more dimensions by enlarging more than one range from Pr_X . That is why definition of maximization 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 2.1. If the quantitative association rule $(A, a, b) \Rightarrow M_J(T_{Pr_X})$ is irreducible, then

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

Theorem 2.2. There are minimum 2, maximum $2k$ μ-tuples to define a profile of the irreducible rule, where k is the dimensionality of the profile.

For example, a profile in two dimensions is defined by 2,3 or 4 μ-tuples (Fig. 1).

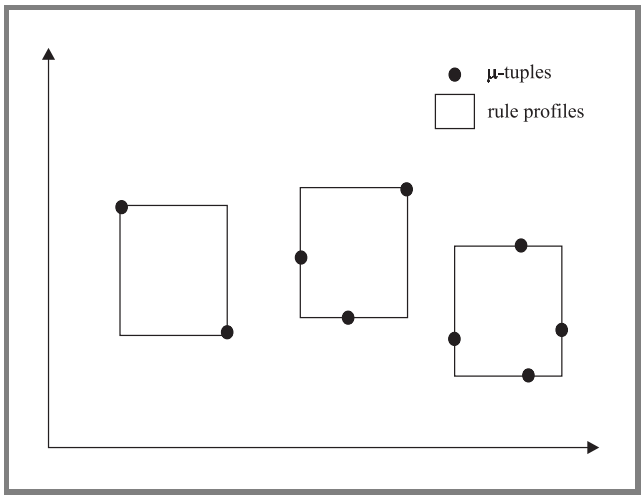

Fig. 1. The μ-tuples that define a profile.

3. Exploratory algorithm

The general outline of the mining algorithm that utilizes Theorems 2.1 and 2.2 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 that works on object variable MiTupleSet:

```
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 with the number of tuples, but may be still considered high. It may be decreased with clustering strategies for μ-tuples.
The efectiveness of the algorithm and comparisons with classic mining algorithms for quantitative rules [8] was described in detail in [7].

4. Application

As described in section with problem statement, 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 region we obtain a tuple of attributes $I_1, \ldots, I_k$ 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. Examples of such rules are

$$urban \in (10\%, 28\%) \wedge$$
$$\wedge roads \in (5\%, 11\%) \wedge forest \in (0\%, 8\%) \Rightarrow$$
$$traffic = 4.1erl (traffic = 1.2erl)$$

$$water \in (6\%, 25\%) \wedge forest \in (30\%, 80\%) \Rightarrow$$
$$traffic = 0.2erl (traffic = 1.2erl)$$

$$water \in (15\%, 35\%) \wedge suburbs \in (20\%, 40\%) \Rightarrow$$
$$traffic = 3.0erl (traffic = 1.2erl)$$

5. Other areas of application and future work

The new form of rules may be used also directly to 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 $I_1, \ldots, I_k$ 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 hole space.
Sampled graphic files are almost the same case as raw spatial data. In [7] there is described a series of experiments that prove usefulness of new methodology to discover patterns in graphic files.

Future work planned in the area of data driven discovery of quantitative rules is as follows:

- rule management systems to filter interesting and representative rules,
- enhanced mining algorithms,
- better use of database systems – some experiments with multidimensional indexing [3] were already done,
- intelligent management of knowledge discovery process,
- the use of new rules for various form of multimedia data.

References

[1] P. Gawrysiak and M. Okoniewski, *Applying Data Mining Methods for Cellular Network Planning, Intelligent Information Systems*. Springer-Physica Verlag, 2000.

[2] P. Gawrysiak, M. Okoniewski, and H. Rybiński, *Regression – Yet Another Clustering Method, Intelligent Information Systems*. Springer-Physica Verlag, 2001.

[3] V. Geade and O. Günther, "Multidimensioanal access methods", *ACM Comput. Surv.*, vol. 30, no. 2, 1997.

[4] Y. Lindell and Y. Aumann, "Theory of quantitative association rules with statistical validation" in *Proc. Conf. SIGKDD*, San Diego, USA, 1999.

[5] R. J. Miller and Y. Yang, "Association rules over interval data" in *Proc. Conf. ACM SIGMOD'97*, 1997.

[6] M. Okoniewski, "Discovery of multi-dimensional quantitative association rules". Ph.D. thesis. Warsaw University of Technology, Warsaw, 2001.

[7] M. Okoniewski, Ł. Gancarz, and P. Gawrysiak, "Mining multidimensional quantitative associations" in *Rule-based Data Mining Workshop, Conf. INAP 2001*, Tokio, Japan, 2001.

[8] R. Srikant and R. Agrawal, "Mining quantitative association rules in large relational tables" in *Proc. Conf. SIGMOD*, Montreal, Canada, 1996.

[9] A. Skowron and S. H. Nguyen, "Quantization of real value attributes: rough set and boolean reasoning approach". Tech. Rep., Warsaw University of Technology, Warsaw, 1995.

Michał Okoniewski received the Ph.D. in computer science from Warsaw University of Technology, Poland, in 2002. He is currently post-doc researcher at Intelligent Systems Laboratory (ISLab) at the University of Antwerp, Belgium. His research interests include knowledge discovery, information systems and bioinformatics.

okoniews@ii.pw.edu.pl
Institute of Computer Science
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland
e-mail: Michal.Okoniewski@ua.ac.be
Intelligent Systems Laboratory (ISLab)
Department of Mathematics and Computer Science
University of Antwerp, RUCA
Middelheimlaan 1
B-2020 Antwerpen, Belgium