MULTI-DATABASE MINING
– A Granular Computing Approach –

Ning Zhong
Dept. of Computer Science and Systems Engineering
Faculty of Engineering, Yamaguchi University
Tokiwa-Dai, 2557, Ube 755, Japan
E-mail: zhong@ai.csse.yamaguchi-u.ac.jp

Abstract: *The paper proposes a way of mining peculiarity rules from multiple statistical and transaction databases. We introduce the peculiarity rules as a new type of association rules, which can be discovered from a relatively small number of the peculiar data by searching the relevance among the peculiar data. We argue that the peculiarity rules represent a typically unexpected, interesting regularity hidden in statistical and transaction databases. We describe how to mine the peculiarity rules in the multi-database environment by using granular computing techniques.*

Keywords: Multi-Database Mining (MDM), Granular Computing (GrC), Peculiarity Oriented.

1 Introduction

So far the main stream in the data mining community is limited to rule discovery in single universal relation (or an information table) [10, 1]. Although theoretically, any multi-relational database can be transformed into a single universal relation, practically this can lead to many issues such as universal relations of unmanageable sizes, infiltration of uninteresting attributes, losing of useful relation names, unnecessary join operation, and inconveniences for distributed processing. On the other hand, the purpose of building databases in most organizations is usually that of managing information sources effectively. In other words, data are rarely specially collected/stored in a database for the purpose of mining knowledge. Hence some concepts, regularities, causal relationships, and rules cannot be discovered if we just search a single database [13].

Recently, it has been recognized in the data mining community that *multi-database mining* (MDM for short) is an important research topic [7, 13]. Multi-database mining is to mine knowledge in multiple information sources. Generally speaking, the task of multi-database mining can be divided into three levels:

1. Mining from multiple relations in a database.
2. Mining from multiple relational databases.
3. Mining from multiple mixed-media databases.

In particular, many datasets in real world contain more than just a single type of data [6]. For example, medical data sets often contain numeric data (e.g. test results), images (e.g. X-rays), nominal data (e.g. person smokes/does not smoke), and acoustic data (e.g. the recording of a doctor's voice). How to handle such multiple data sources is a new, challenging research issue.

Multi-database mining involves many related topics including interestingness checking, relevance, database reverse engineering, granular computing, and distributed data mining. Zhong, Yao and Ohsuga proposed a way of mining peculiarity rules from multiple statistical and transaction databases [13]. A challenge in multi-database mining is semantic heterogeneity among multiple databases since no explicit foreign key relationships exist among them usually. Hence, the key issue is how to find/create the relevance among different databases. Granular computing techniques provide a useful tool to find/create the relevance among different databases by changing information granularity [11, 8, 9, 10, 3].

Database reverse engineering is a research topic that is closely related to multi-database mining. The objective of database reverse engineering is to obtain the domain semantics of legacy databases in order to provide meaning of their executable schemas' structure [2]. Although database reverse engineering has been investigated recently, it was not researched in the context of multi-database mining. By taking a unified view of multi-database mining and database reverse engineering, Zhong, Yao and Ohsuga proposed the RVER (Reverse Variant Entity-Relationship) model to represent the result of multi-database mining [12, 13]. The RVER model can be regarded as a variant of semantic network that is a

kind of well-known method for knowledge representation. From this point of view, multi-database mining can be regarded as a kind of database reverse engineering.

2 Peculiarity

Peculiarity is a kind of interestingness. Peculiarity relationships/rules (with common sense) may be hidden in a relatively small number of data. Generally speaking, hypotheses (knowledge) generated from databases can be divided into the following three types:

- Incorrect hypotheses.
- Useless hypotheses.
- *New, surprising, interesting hypotheses.*

The purpose of data mining is to discover new, surprising, interesting knowledge hidden in databases. Hence, the evaluation of interestingness (including peculiarity, surprisingness, unexpectedness, usefulness, novelty) should be done in pre-processing and/or post-processing of knowledge discovery [4, 13].

Zhong, Yao and Ohsuga proposed *peculiarity rules* as a new type of association rules [13]. A *peculiarity rule* is discovered from the *peculiar* data by searching the relevance among the *peculiar* data. Roughly speaking, a data is *peculiar* if it represents a peculiar case described by a relatively small number of objects and is very different from other objects in a data set. Although it looks like the exception rule[5] from the viewpoint of describing a relatively small number of objects, the peculiarity rule represents the well-known fact with common sense, which is a feature of the general association rules [1].

We argue that the *peculiarity rules* are a typical regularity hidden in statistical and transaction databases. Sometimes, the general association rules that represent the well-known fact with common sense cannot be found from numerous statistical or transaction data, or although they can be found, the rules may be uninteresting ones to the user since data are rarely specially collected/stored in a database for the purpose of mining knowledge in most organizations. In particular, unexpected (common sense) relationships/rules may be hidden in a relatively small number of data. Thus, we may focus some interesting data (the peculiar data), and then we find more novel and interesting rules (peculiarity rules) from the data. For example, the following rules are the peculiarity ones that can be discovered from a relation called *Japan-Geography* (see Table 1) in a *Japan-Survey* database:

$rule_1$: *ArableLand(large)* & *Forest(large)*
$\rightarrow$ *PopulationDensity(low)*.

$rule_2$: *ArableLand(small)* & *Forest(small)*
$\rightarrow$ *PopulationDensity(high)*.

In order to discover the rules, we first need to search the peculiar data in the relation *Japanese-Geography*. From Table 1, we can see that the values of the attributes *ArableLand* and *Forest* for Hokkaido (i.e. 1209 Kha and 5355 Kha) and for Tokyo and Osaka (i.e. 12 Kha, 18 Kha, and 80 Kha, 59 Kha) are very different from other values in the attributes. Hence, the values are regarded as the peculiar data. Furthermore, $rule_1$ and $rule_2$ are generated by searching the relevance among the peculiar data. Note that we use the qualitative representation for the quantitative values in the above rules. The transformation of quantitative to qualitative values can be done by using the following background knowledge on information granularity:

Basic granules:
bg_1 = {*high, low*}; bg_2 = {*large, small*};
bg_3 = {*many, little*}; bg_4 = {*far, close*};
bg_5 = {*long, short*};
.....

Specific granules:
biggest-cities = {*Tokyo, Osaka*};
kanto-area = {*Tokyo, Tiba, Saitama, ...*};
kansei-area = {*Osaka, Kyoto, Nara, ...*};
.....

That is, *ArableLand = 1209, Forest = 5355* and *PopulationDensity = 67.8* for Hokkaido are replaced by the granules, "large" and "low", respectively. Furthermore, Tokyo and Osaka are regarded as a neighborhood (i.e. the biggest cities in Japan). Hence, $rule_2$ is generated by using the peculiar data for both Tokyo and Osaka as well as their granules (i.e. "small" for *ArableLand* and *Forest,* and "high" for *PopulationDensity*).

3 Mining Peculiarity Rules in Multi-Database

This section describes a way of mining peculiarity rules in multiple databases.

3.1 Finding the Peculiar Data

There are many ways of finding the peculiar data. In this section, we describe an attribute-oriented method.

Let $X = \{x_1, x_2, \ldots, x_n\}$ be a data set related to an attribute in a relation, and n be the number of different values in an attribute. The peculiarity of x_i can be evaluated by the *Peculiarity Factor*, $PF(x_i)$,

Table 1: Japan-Geography

Region	Area	Population	PopulationDensity	PeasantFamilyN	ArableLand	Forest	...
Hokkaido	82410.58	5656	**67.8**	93	**1209**	**5355**	...
Aomori	9605.45	1506	156.8	87	169	623	...
...	...	...	...	...	...	...	...
Tiba	5155.64	5673	1100.3	116	148	168	...
Tokyo	2183.42	11610	**5317.2**	21	**12**	**80**	...
...	...	...	...	...	...	...	...
Osaka	1886.49	8549	**4531.6**	39	**18**	**59**	...
...	...	...	...	...	...	...	...

$$PF(x_i) = \sum_{j=1}^{n} \sqrt{N(x_i, x_j)}. \quad (1)$$

It evaluates whether x_i occurs relatively small number and is very different from other data x_j by calculating the sum of the square root of the conceptual distance between x_i and x_j. The reason why the square root is used in Eq. 1 is that we prefer to evaluate more near distances for relatively large number of data so that the peculiar data can be found from relatively small number of data. Major merits of the method are

- It can handle both the continuous and symbolic attributes based on a unified semantic interpretation;
- Background knowledge represented by binary neighborhoods can be used to evaluate the peculiarity if such background knowledge is provided by a user.

If X is a data set of a continuous attribute and no background knowledge is available, in Eq. (1),

$$N(x_i, x_j) = |x_i - x_j|. \quad (2)$$

On the other hand, if X is a data set of a symbolic attribute and/or the background knowledge for representing the conceptual distances between x_i and x_j is provided by a user, the peculiarity factor is calculated by the conceptual distances, $N(x_i, x_j)$. However, all the conceptual distances are 1, as default, if background knowledge is not available.

After the evaluation for the peculiarity, the peculiar data are elicited by using a threshold value,

$$\begin{aligned} &threshold \\ &= mean\ of\ PF(x_i) + \alpha \times \\ &\quad standard\ deviation\ of\ PF(x_i), \end{aligned} \quad (3)$$

where α can be specified by a user. That is, if $PF(x_i)$ is over the threshold value, x_i is a peculiar data.

3.2 Relevance among the Peculiar Data

A peculiarity rule is discovered from the peculiar data by searching the relevance among the peculiar data. Let $X(x)$ and $Y(y)$ be the peculiar data found in two attributes X and Y respectively. We deal with the following two cases:

- If the $X(x)$ and $Y(y)$ are found in a relation, the relevance between $X(x)$ and $Y(y)$ is evaluated in the following equation:

  $$R_1 = P_1(X(x)|Y(y))P_2(Y(y)|X(x)). \quad (4)$$

  That is, the larger the product of the probabilities of P_1 and P_2, the stronger the relevance between $X(x)$ and $Y(y)$.

- If the $X(x)$ and $Y(y)$ are found in two different relations, we need to use a value (or its granule) in a key (or foreign key/link) as the relevance factor, $K(k)$, to find the relevance between $X(x)$ and $Y(y)$. Thus, the relevance between $X(x)$ and $Y(y)$ is evaluated in the following equation:

  $$R_2 = P_1(K(k)|X(x))P_2(K(k)|Y(y)). \quad (5)$$

Furthermore, Eq. (4) and Eq. (5) are suitable for handling more than two peculiar data found in more than two attributes if $X(x)$ (or $Y(y)$) is a granule of the peculiar data.

3.3 Mining in Multiple Relations

First, the concept of foreign key in the relational databases needs to be extended into foreign link because we are also interested in getting to non-key attributes for data mining from multiple relations in a database [7]. A major work is to find the peculiar data in multiple relations for a given discovery task while foreign link relationships exist. In other words, our task is to select n relations, which contain the peculiar data, among m relations ($m \geq n$) with foreign links.

We again use the *Japan-Survey* database as an example. There are many relations (tables) in

this database such as *Japan-Geography, Economy, Alcoholic-Sales, Crops, Livestock-Poultry, Forestry, Industry*, and so on. The method for selecting n relations among m relations can be briefly described as follows:

Step 1. Focus on a relation as the main table and find the peculiar data from this table. Then elicit the peculiarity rules from the peculiar data by using the methods stated in Section 3.1 and 3.2.

For example, if we select the relation called *Japan-Geography* shown in Table 1 as the main table, $rule_1$ and $rule_2$ stated in Section 2 are a result for the step.

Step 2. Find the value(s) of the focused key corresponding to the mined peculiarity rule(s) in *Step* 1 and change its granularity of the value(s) of the focused key if the background knowledge on information granularity is available.

For example, "Tokyo" and "Osaka" that are the values of the key attribute *region* can be changed into a granule, "biggest cities".

Step 3. Find the peculiar data in the other relations (or databases) corresponding to the value (or its granule) of the focused key.

Here we need to find the related relations by using foreign keys (or foreign links).

Step 4. Select n relations that contain the peculiar data, among m relations ($m \geq n$). In other words, we just select the relations that contain the peculiar data that are relevant to the peculiarity rules mined from the main table.

For example, as a result, the relations, *Economy* and *Alcoholic-Sales,* are selected corresponding to the value of the foreign key, *region = biggest cities.*

The methodology stated above can be extended for mining from multiple databases. For example, if we found that the turnover was a marked drop in some day from a supermarket transaction database, maybe we cannot understand why. However, if we search a weather database, we can find that there was a violent typhoon this day in which the turnover of the supermarket was a marked drop. Hence, we can discover the reason why the turnover was a marked drop.

A challenge in multi-database mining is semantic heterogeneity among multiple databases since no explicit foreign key relationships exist among them usually. Hence, the key issue is how to find/create the relevance among different databases. *Granular computing* techniques based on semantics, approximation, and abstraction are available for solving the issue [3, 11].

4 Conclusion

We presented a way of mining peculiarity rules from multiple statistical and transaction databases. The peculiarity rules are defined as a new type of association rules. We can change the granularity of the peculiar data dynamically in the discovery process. Some of databases such as Japan-survey, web-log, weather, supermarket have been tested or have been testing for our approach.

References

[1] Agrawal R. et al. "Fast Discovery of Association Rules", *Advances in Knowledge Discovery and Data Mining*, AAAI Press (1996) 307-328.

[2] Chiang, Roger H.L. et al "A Framework for the Design and Evaluation of Reverse Engineering Methods for Relational Databases", *Data & Knowledge Engineering*, Vol.21 (1997) 57-77.

[3] Lin, T.Y. "Granular Computing on Binary Relations 1: Data Mining and Neighborhood Systems", L. Polkowski and A. Skowron (eds.) *Rough Sets in Knowledge Discovery 1*, In Studies in Fuzziness and Soft Computing series, Vol. 18, Physica-Verlag (1998) 107-121.

[4] Liu, B., Hsu W., and Chen, S. "Using General Impressions to Analyze Discovered Classification Rules", *Proc. Third International Conference on Knowledge Discovery and Data Mining (KDD-97)*, AAAI Press (1997) 31-36.

[5] Suzuki E.. "Autonomous Discovery of Reliable Exception Rules ", *Proc Third Inter. Conf. on Knowledge Discovery and Data Mining (KDD-97)*, AAAI Press (1997) 259-262.

[6] Thrun, S. et al. "Automated Learning and Discovery", AI Magazine (Fall 1999) 78-82.

[7] Wrobel, S. "An Algorithm for Multi-relational Discovery of Subgroups", J. Komorowski and J. Zytkow (eds.) *Principles of Data Mining and Knowledge Discovery.* LNAI 1263, Springer-Verlag (1997) 367-375.

[8] Yao, Y.Y. "Granular Computing using Neighborhood Systems ", Roy, R., Furuhashi, T., and Chawdhry, P.K. (eds.) *Advances in Soft Computing: Engineering Design and Manufacturing*, Springer-Verlag (1999) 539-553.

[9] Yao, Y.Y. and Zhong, N. "Potential Applications of Granular Computing in Knowledge Discovery and Data Mining", *Proc. The 5th International Conference on Information Systems Analysis and Synthesis (IASA'99)*, edited in the invited session on Intelligent Data Mining and Knowledge Discovery (1999) 573-580.

[10] Yao, Y.Y. and Zhong, N. "Granular Computing using Information Tables", *Granular Computing and Data Mining*, Physica-Verlag (2000) (in press).

[11] Zadeh, L. A. "Toward a Theory of Fuzzy Information Granulation and Its Centrality in Human Reasoning and Fuzzy Logic", *Fuzzy Sets and Systems*, Elsevier Science Publishers, 90 (1997) 111-127.

[12] Zhong N. and Yamashita S. "A Way of Multi-Database Mining ", *Proc. the IASTED International Conference on Artificial Intelligence and Soft Computing (ASC'98)*, IASTED/ACTA Press (1998) 384-387.

[13] Zhong, N., Yao, Y.Y., and Ohsuga, S. "Peculiarity Oriented Multi-Database Mining", J. Zytkow and Jan Rauch (eds.) *Principles of Data Mining and Knowledge Discovery.* LNAI 1704, Springer-Verlag (1999) 136-146.