Mutlirelatonal GUHA Method and Genetic Data

Martin Ralbovský, Alexander Kuzmin, Jan Rauch

Department of Information and Knowledge Engineering,
University of Economics, Prague, W. Churchill Sq. 4, 130 67 Praha 3, Czech Republic
martin.ralbovsky@gmail.com, alexander.kuzmin@gmail.com, rauch@vse.cz

Abstract. The paper presents multirelational GUHA method, focusing on multirelational association rules. Background and principles of the method are introduced together with comparison with related methods. New implementation in the Ferda tool is presented and initial experiments in the genetic domain are shown.

Keywords: GUHA method, Multirelational GUHA, 4FT, virtual attribute, Ferda, genetic data

1 Introduction

The GUHA method is one of the first methods of exploratory data analysis, which has been in development since the mid-sixties. It is a general mainframe for retrieving interesting knowledge from data. The method has firm theoretical foundations based on logics, especially observational calculi and statistics [3]. Figure 1 shows the main principle of the method.

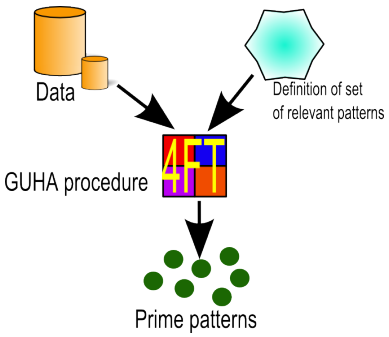

Fig. 1. The GUHA method

GUHA method is realized by GUHA procedures such as the 4FT procedure, located in the middle of the figure. A GUHA task consists of data and a simple definition of a possibly large set of relevant patterns defined with the

aid of observational calculi, which are inputs to the procedure. The procedure automatically generates all the relevant patterns and verifies them against the provided data. Patterns that are true and do not logically follow from the other true output and more simple patterns are called *prime patterns*. We call them also *hypotheses* as in [3]. The most known GUHA procedure is the ASSOC (*4ft-miner*, 4FT) procedure for mining generalized association rules [11, 12], based on different approach than the mainstream *apriori* algorithm [1].

In its initial form, procedures of the GUHA method were designed to mine over one relation only. In [10], proper theory for the multirelational form of the method was developed. However, until recently this form lacked suitable implementation and data to prove usability. We present in this paper recent implementation of the multirelational GUHA in the Ferda system [5] and we also start experiments with genetic data where the method seems to be perspective.

The paper is structured as follows: section 2 explains main principles of multirelational GUHA and briefly introduces the new implementation of the method. Section 3 compares our method to other mainly ILP methods and section 4 describes the initial experiments with genetic data. Finally section 5 concludes the paper.

2 Principles of Multirelational Mining with GUHA

Because of the short format of the paper, we explain only basics of the principles without going into detail. For more details, see [10, 4]. The multirelational GUHA method currently supports star-scheme of the database with one *master table* and several *detail tables*. The key term is *virtual attribute*, which is attribute from detail data table, that is created during the process of GUHA pattern verification and is treated as normal attribute of master table although not physically stored. The most interesting type of virtual attribute is the *hypotheses attribute*. Hypotheses attribute is defined by the GUHA (sub)task on the detail table. Value of the attribute corresponds to validity of a GUHA pattern for subset of records of the detail table that "belongs" to the master record. Validity can be expressed as boolean value or (in the future) generally as a real number. One GUHA (sub)task in on the detail table usually generates large amount of hypotheses attributes.

We present an example of the hypotheses attribute from the banking domain[1]. The used GUHA procedure is 4FT for association rules mining: there are two tables concerning clients of a bank. The master table contains information about clients' accounts and the detail table contains information about transactions of individual clients. One client in the master table can have several transactions in the detail table. Example of hypotheses attribute can be *client that often pays by credit card*, which can be formally written as

$$ClientID \approx Payment(CreditCard).$$

[1] We think that banking domain is more comprehensible than the genetic domain to the non-expert.

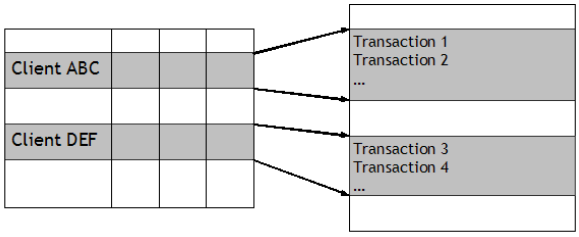

Fig. 2. Example of multirelational task setting

The situation is shown in figure 2. It is obvious, that this rule may be very useful as an attribute in the master table concerning clients' accounts. We name the virtual attribute $ClientPayingByCreditCard$. Then one can examine status of a client based on client's payments and address. Example of such generalized association rule is

$$District(SouthEast)\&ClientPayingByCreditCard$$

$$\approx Status(good).$$

There are two experimental implementation of relational GUHA method, one in the frame of LISp-Miner system [11] and the second one in the Rel-Miner system [4], but they are not used any more due to various reasons. The new implementation in the Ferda system [5] takes advantage of visual and modular environment, which makes the complex task setting comprehensible to the user. Figure 3 shows a sample multirelational task in Ferda. All the implementations of relational GUHA do not use *apriori*, they are based on representation of analyzed data by strings of bits [11].

3 Related Methods

The most known KDD technique for discovering knowledge from multirelational data is inductive logic programming (ILP). Principle of propositionalization approaches in ILP [6] is very close to principle of hypotheses attribute. Below is list of main differences:

- Propositioned attributes of ILP are conjunctions of (possibly negated) literals of predicate logic. In contrary, hypotheses attributes are formulas of observational calculus, enabling to represent i.e. implication or statistical significance of the attribute.
- In practical cases, multirelational GUHA is limited to star-scheme of the database. Relations in ILP does not have this restriction.

The *WARMR* algorithm [2] performs ILP propositionalization and then searches for association rules in *apriori*-like manner. Detailed comparison of *WARMR* and

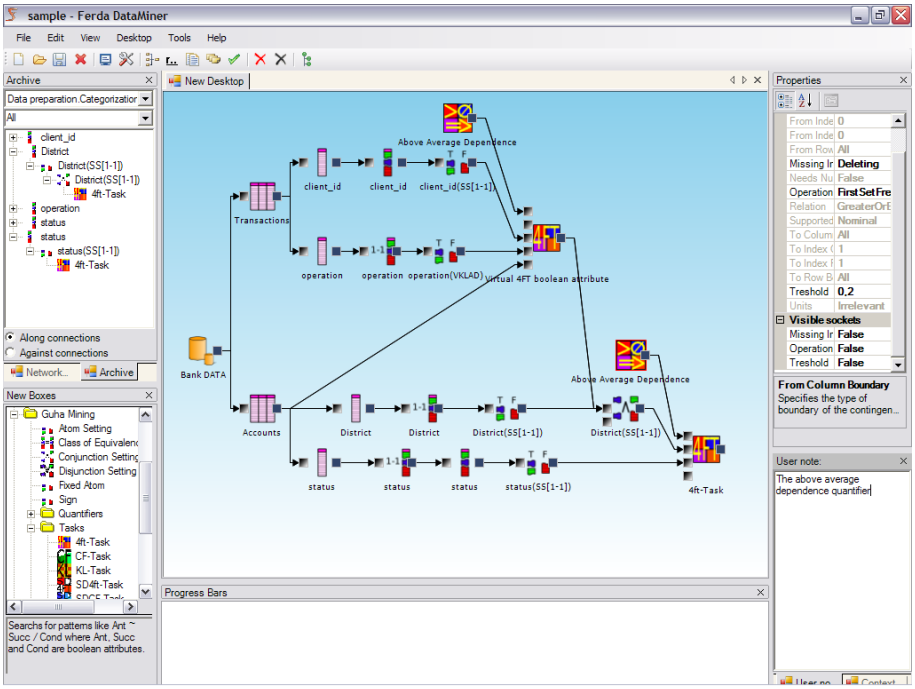

Fig. 3. The Ferda environment

multirelational GUHA can be found in [4]. Another approach based on *apriori* [9] adapts *support* and *confidence* measures and calculates them on multiple tables without joining them. The RELAGGS system [7] works in similar manner: it calculates aggregations of records of columns from tables bound by foreign keys. Because the method is based on database joins, it is not unable to calculate the expressive GUHA patterns on the detail tables.

4 Mining over Genetic Data

The big problem throughout the history of multirelational GUHA was to find suitable domain to prove the usefulness of the method. The first attempts were made in the banking domain - dataset Barbora was used[2]. These attempts were unsuccessful.

During this spring, we initiated cooperation with Czech Technical University (CTU) concerning mining genetic data. Team from CTU lead by F. Zelezny compiled genetic dataset from publicly available datasets. The dataset contains genetic measurements acquired from Affymetrix DNA microarrays[3] from hu-

[2] Used in the PKDD 1999 Discovery challenge, see `http://lisp.vse.cz/challenge`.

[3] National Center for Biotechnology Information GEO Datasets `http://www.ncbi.nlm.nih.gov`

man, mouse and rat for two different types of cells: hematopoetic and stromal, both of which are involved in blood cells production in bone marrow. The gene measurements were enriched with gene semantics involving information about pathways (maps representing molecular interaction and reaction networks) and fully-coupled-fluxes (FCF), linear pathway subgraphs[4].

The longterm scientific goal is to examine how does the expression of genes in FCF correlate with types of cells (and possibly other characteristics) [8]. Because the expressiveness of hypotheses attribute, multirelational GUHA is especially fit for this purpose. We have tried initial experiments with hypotheses attributes such as *high expression of genes in FCF* showing promising results: one of the experiments included examination of 500K gene measurements concerning 500 FCF's. With procedure 4FT we obtained 1394 *hypotheses* out of 3187 verifications. All the hypotheses were in form:

$$[FluxID(\alpha) \approx_1 GeneLevel(\beta)] \approx_1 CellType(\gamma)$$

Where $\approx_1$ stands for $Conf = 100\%$. However following work need to be done in order to obtain scientifically sound results:

- **Proper discretization:** GUHA has ways to handle numeric data. Yet these ways are unsuitable mainly because of the fact that expressions of different genes have different ranges, but are contained in one attribute. We need to build a new genome expression table based on discrete values directly from Affymetrics DNA microarrays.
- **Scaling:** It is the first time that multirational GUHA has been used with data of such a large size and we have experienced performance problems. Effective ways to handle results of queries, which by far exceed the capacity of operating memory need to by found and implemented.
- **Chip handling:** The probes measuring one gene are placed on several chips. It remains an open question how much does this fact influence the gene measurement.

5 Conclusion

We present multirelational extension of the GUHA method of exploratory analysis. Like other methods such as ILP propositionalization, the principle is to enrich a data table with attributes taken from other data tables. The advantage of multirelational GUHA lies in providing an expressive language for virtual attributes based on observational calculus.

We also present recent implementation of multirelational GUHA method in the Ferda system and possible and promising usage of the method in genetic experiments. At present, we do not yet have any scientifically sound genetic results, the paper states next steps to be made in order to achieve them. Nonetheless, usage of multirelational GUHA seems to be suitable for exploratory analysis of complex data over multiple tables such as genetic data.

[4] Taken from KEGG genome database, `http://www.genome.jp.kegg`

Acknowledgements

This work was supported by the project MSM6138439910 of the Ministry of Education of the Czech Republic and grant 201/08/0802 of the Czech Science Foundation. We thank and acknowledge contribution of our research colleagues Matěj Holec, Filip Železný and Jiří Kléma from Czech Technical University for providing the genetic data and guidance in the genetic domain.

References

1. Agrawal R., Mannila H., Srikant R., Toivonen H., Verkamo A.: *Fast discovery of association rules.* Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R., eds.: Advances in Knowledge Discovery and Data Mining. AAAI Press, Menlo Park (1996) p. 307 – 328
2. Dehaspe L., De Raedt L.: *Mining Association Rules in Multiple Relations.* In Proceedings of the 7th International Workshop on Inductive Logical Programming, Volume 1297, LNAI, pp. 125–132, Springer-Verlag, 1997
3. Hájek P., Havránek, T.: *Mechanising Hypothesis Formation - Mathematical Foundations for a General Theory.* Springer-Verlag: Berlin - Heidelberg - New York, 1978.
4. Karban T.: *Relational Data Mining and GUHA.* in Richta K., Snášel V., Pokorný J.(eds.): Proceedings of the 5th annual workshop DATESO 2005(Databases, Texts, Specifications and Objects), ISBN:80-01-03204-3, pp.103–112
5. Kováč M., Kuchař T., Kuzmin A., Ralbovský M.: *Ferda, New Visual Environment for Data Mining.* Znalosti 2006, Conference on Data Mining, Hradec Králové 2006, p. 118 – 129 (in Czech)
6. Kramer S., Lavrač N., Flach P.: *Proposionalization Approaches to Relational Data Mining.* In: Džeroski, Lavrač: Relational Data Mining, ISBN 3-540-42289-7, Springer Verlag 1998, pp. 262–291
7. Krogel M.-A., Rawles S., Železný F., Flach P.A., Lavrač N., Wrobel S.: *Comparative Evaluation of Approaches to Propositionalization* In: Horváth T., Yamamoto A. (Eds.) Proceedings of the 13th International Conference on Inductive Logic Programming. LNCS 2835, Springer-Verlag, 2003
8. Notebaar R.A., Teusink B., Siezen R.J., Papp B.: *Co-Regulation of Metabolic Genes is Better Explained by Flux Than Network Distance.* PLoS Computational Biology 4(1), 2008: e26 doi:10.1371/journal.pcbi.0040026
9. Pizzi, L.C., Ribeiro, M.X., Vieira, M.T.P.: *Analysis of Hepatitis Dataset using Multirelational Association Rules.* ECML/PKDD Discovery Challenge, 2005.
10. Rauch J.: *Many Sorted Observational Calculi for Multi-Relational Data Mining.* In: Data Mining Workshops. Piscataway: IEEE Computer Society, 2006 ISBN 0-7695-2702-7 p. 417–422
11. Rauch J., Šimůnek, M.: *An Alternative Approach to Mining Association Rules* Lin T Y, Ohsuga S, Liau C J, and Tsumoto S (eds): Foundations of Data Mining and Knowledge Discovery, Springer-Verlag, 2005 p. 219 – 239
12. Ralbovský M., Kuchař T.: *Using Disjunctions in Association Mining.* In: Perner P.: Advances in Data Mining - Theoretical Aspects and Applications, LNAI 4597, Springer Verlag, Heidelberg 2007