APPLICATION OF THE GUHA METHOD IN MEDICAL DIAGNOSIS

Goran Cvetkoski
Netcetera
Skopje, Macedonia

Ivan Chorbev
Faculty of Electrical Engineering and Information Technologies
Skopje, Macedonia

ABSTRACT

The large growth of medical databases has motivated researchers to use data mining on medical data for knowledge discovery. Many mining techniques have been implemented and demonstrated on databases from the medical domain.
This paper presents a use of the data mining method GUHA in the discovery process of association rules for the purpose of medical diagnostics. Experiments were performed and the results are presented.

I. INTRODUCTION

Data mining techniques can be widely utilized in medicine. The healthcare environment is usually information rich, but knowledge poor. However, data mining techniques can be applied to create a knowledge rich healthcare environment. Sophisticated equipment used in the practice of modern medicine generates huge amount of data. If the medical data is properly organized and integrated in a medical information system, patterns and structures in data can be explored and discovered. The medical data is usually stored in digital form, and considerable effort is being made to find automated methods of data analysis to generate knowledge.

The acquired knowledge can later be used for fast and better clinical decision-making. Further on in this paper we focus on the GUHA method and its capabilities in data mining and knowledge discovery in databases. In the second part of the paper we present a short overview of the theoretical basis of the GUHA method, and in the third and fourth part we present an experiment and its results. The concluding remarks and future work are presented in the fifth part.

II. GUHA METHOD

GUHA (General Unary Hypotheses Automaton) is a method of exploratory data analysis developed in Prague since the mid-sixties of the twentieth century. GUHA has several features found in contemporary systems of data mining and knowledge discovery in databases (KDD) and may be well considered as an early example of such systems.

Its main principle is to generate all possible hypotheses based on user task setting, verify them and output the valid ones. For example, the GUHA procedure 4ft-Miner mines for association rules from a single table, while other GUHA procedures mine for other types of patterns. Because GUHA is based on simple logical and statistical analysis, as a result it provides easily measurable results of hypothesis support and confidence. This is used as an advantage in many new studies for simple decision support system (DSS) advice explanation or in background knowledge confirmation.

The association rule is an expression $\phi \approx \psi$ where ϕ and ψ are derived Boolean attributes. The intuitive meaning of association rule $\phi \approx \psi$ is that Boolean attributes ϕ and ψ are associated in the way corresponding to the condition given by the symbol $\approx$. The symbol $\approx$ is called 4ft-quantifier. It denotes a condition concerning a four-fold contingency table of ϕ and ψ.

The four-fold table of ϕ and ψ in the data matrix M is the quadruple <a,b,c,d> see Figure 1. The quadruple is denoted as 4ft(ϕ, ψ, M).

M	ψ	$\neg\psi$	
φ	a	b	r
$\neg\varphi$	c	d	s
	k	l	n

Figure 1: Four-fold table 4ft(ϕ, ψ, M) of ϕ, ψ in M.

The association rule can be true or false in the given data matrix M. The true-value of $\phi \approx \psi$ in the data matrix M is denoted by Val($\phi \approx \psi$,M). If it is Val($\phi \approx \psi$,M) = true then the Boolean attributes ϕ and ψ are associated in the way corresponding to the 4ft-quantifier $\approx$ in the data matrix M. There is a condition concerning four-fold tables <a,b,c,d> associated to each 4ft quantifier $\approx$. This condition is understood as a {0,1}-function $\approx$(a,b,c,d). The association rule $\phi \approx \psi$ is true in the data matrix M if it is $\approx$(a,b,c,d) = 1 where <a,b,c,d> = 4ft(ϕ, ψ, M).
There are various 4ft quantifiers. Here is an example of two of them:

- **Founded implication** $\Rightarrow_{p;Base}$ with parameters $0 < p \leq 1$ and $Base > 0$. The condition $\frac{a}{a+b} \geq p \wedge a \geq Base$ is associated to 4ft quantifier $\Rightarrow_{p;Base}$. Association rule $\phi \Rightarrow_{p;Base} \psi$ can be interpreted as "100p per cent of objects satisfying ϕ satisfy also ψ" or "ϕ implies ψ on the level 100p per cent".
- **Double founded implication** $\Leftrightarrow_{p;Base}$ with parameters $0 < p \leq 1$ and $Base > 0$. The condition $\frac{a}{a+b+c} \geq p \wedge a \geq Base$ is associated to 4ft quantifier $\Leftrightarrow_{p;Base}$. Association rule $\phi \Leftrightarrow_{p;Base} \psi$ can be interpreted as "100p per cent of objects satisfying ϕ or ψ satisfy both ϕ and ψ" or "$\phi \wedge \psi$ implies $\phi \vee \psi$ on the level 100p per cent".

III. EXPERIMENT PREPARATION

There are several computer implementations of the GUHA method procedures. In our experiments, we used the academic software system for KDD, Lisp-Miner (http://lispminer.vse.cz/). The Lisp-Miner is a set of modules and several data mining procedures, all based on the GUHA principle. We are using the basic GUHA procedure *4ft* that mines association rules in a single-table database.

For the purpose of the experiment, we picked the Breast Cancer Ljubljana medical dataset. It contains data for breast cancer diagnosis, with 286 instances and 9 attributes (Table 1). The datasets contains both numeric and nominal attributes. Numeric attributes consist of integer valued numbers. Nominal attributes take on values from a finite set of possibilities. Because the Lisp-Miner native data storage is in the ODBC format, the dataset needed previous preparation in order to assure compatibility.

Table 1: Attributes of the Breast Cancer Ljubljana dataset.

Table 1: Margins, text width, etc. definitions.

attribute	type
age	discrete
menopause	discrete
tumor size	discrete
inv nodes	discrete
node caps	discrete
deg malig	discrete
breast	discrete
breast quad	discrete
irradiat	discrete
class	discrete
14 pt	discrete
8 pt	discrete

The experiment was carried out in four separate phases:

A. Data Preparation

In the data preparation process, the dataset is loaded into the Lisp-Miner. All the attributes from the table should be precisely defined with their categories. Some additional helpful features are possible while defining the attributes, like the simple data exploration, or the *Frequency Histogram* on some attribute.

B. Entering the cedents

After the data preparation part is finished, cedents must be defined in order to start the *4ft* procedure. Each cedent consists of one or more partial cedents that can have one or more attributes defined. Several parameters can be set for each added attribute, influencing in many ways the final association rules that will be generated by the algorithm.

C. Quantifier selection

The experiment was conducted using the simplest of quantifiers, that's the *founded implication* quantifier, with just two parameters to be set, parameters p and *Base*.

D. Association rules generation

The association rules are generated based on the defined task from the user.

IV. RESULTS AND DISCUSSION

In the experiment case, all parameters from the dataset needed for the diagnostics were included in the antecedent. Basically, we were trying to solve the question: What combinations of characteristics of the patient's condition lead to one of the two defined possible diagnostic outcomes? The goal is to extract rules that could be implemented to future medical cases in a diagnostics system.

Based on the results, we made a comparison with some other known data mining procedures - CN2, C4.5, And Miner and SA Tabu Miner – results are taken from our other research papers [8] and experiments available on the Internet. The comparison was made across two criteria, the predictive validity and the simplicity of the observed association rules.

Predictive accuracy was measured by a well-known ten-fold cross-validation procedure. The predictive validity was measured by separating the dataset in 10 equal pieces which were then again grouped in 9 parts for rule-generation and one part for testing. The parameters that define the used quantifier, founded implication, were set to values p = 80 % and *Base* = 50 %. The parameter values were selected from analysis of other experiments. The average values from all measurements are given in Table 2.

Table 2. Predictive validity comparison

KDD procedure	predictive validity
SA Tabu Miner	65,1
CN2	67,69
Ant Miner	75,28
C45	73,22
GUHA (*4ft*)	78,9

As we can see from the results, the *4ft* procedure achieved best predictive validity compared to all other KDD procedures. However, additional experiments are needed to come to conclusions about the overall performance of GUHA.

Very important property of the generated association rules is their simplicity. This is because the knowledge acquired from these rules should be understandable for the human, and easily applicable for future use. The extracted knowledge must be validated by human experts, especially in the medical domain. Simplicity of the generated association rules expressed by their number and number of conditions per rule (average values), are given in Table 3.

Table 3. Association rules simplicity comparison

KDD procedure	rule simplicity
SA Tabu Miner	8,55;1,7
CN2	55,4;2,21
Ant Miner	7,1;1,28
C45	9,7;2,56
GUHA (*4ft*)	6,4;2,26

The set of association rules that is generated from the *4ft* GUHA procedure contains far less members compared to the other KDD procedures. Ant Miner procedure for the same dataset generates on average of 7,1 rules, which is comparable with the 6,4 rules that are in average generated by the *4ft* GUHA procedure. Ant Miner and SA Tabu Miner are better when the number of conditions per rule is taken into account, but the *4ft* GUHA procedure gives satisfactory results here as well.

The *4ft* GUHA procedure is slightly better when we look at the predictive validity comparison, although for a good test of performance, other and various datasets should be also tested. In the rules simplicity part, *4ft* GUHA method showed very good results and it is among the best compared to the other KDD procedures. To sum up, from the experiment we gained some positive impression about the GUHA method in a whole. We are more than satisfied with the overall performance of it and our future work will expand the number and type of datasets tested.

V. FUTURE WORK

Our goal so far was to acquire initial knowledge about the possibilities and performance of the GUHA method possibly while working with specific medical datasets used in the clinical decision process. The dataset we chose for the experiment preliminary gave us good overall results. However, we have to keep in mind that the dataset is small and we can't fully rely on the results we got, especially because the GUHA method is more precise and generally recommended for bigger datasets.

In our future work, we will focus on testing some of the other GUHA procedures, mainly on multi-tabled databases. Another point of interest will for sure be the GUHA treatment of the missing database values.

REFERENCES

[1] J. Rauch, "Interesting Association Rules and Multi-relational Association Rules." Communications of Institute of Information and Computing Machinery, Taiwan. Vol. 5, No. 2, May 2002. pp. 77–82

[2] P. Hajek, "Logics for data mining (GUHA Reditiva)", Institute of Computer Science, Academy of Sciences 182 07 Prague, Chech Republic

[3] P. Hajek, T. Feglar, J. Rauch and D. Coufal, "The GUHA method, data preprocessing and mining", Technical Report No. 867

[4] T. Karban, "Relational Data Mining and GUHA" Department of Software Engineering, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

[5] P. Hajek, "The GUHA method and mining association rules" Institute of Computer Science, Academy of Sciences 182 07 Prague, Chech Republic

[6] T. Havranek, "The Statistical interpretation and modification of GUHA method" Kybernetika, Vol. 7 (1971), No. 1, (13)–21

[7] S. K. Wasan, V. Bhathagar, H. Kaur, "The Impact of Data Mining Techniques on medical diagnostics" Data Science Journal, Volume 5, 19 October 2006. pp 119-126

[8] Chorbev I., Mihajlov D., Jolevski I., Web Based Medical Expert System with a Self Training Heuristic Rule Induction Algorithm, Proc. of The First International Conference on Advances in Databases, Knowledge, and Data Applications, DBKDA 2009, Cancun, Mexico, March 2009, page 143-148.