Subgroup Discovery: On-line Data Mining Server and its Application

Dragan Gamberger[1], Tomislav Šmuc[1] & Nada Lavrač[2]
[1]*Rudjer Bošković Institute, Bijenička 54, 10000 Zagreb, Croatia*
[2]*Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia*

Abstract

Data Mining Server is an internet based data analysis tool available for public usage. It offers simple use of data preprocessing and machine learning algorithms for subgroup discovery. The paper presents the Data Mining Server, illustrates its usage and describes the results obtained by its application on an artherosclerotic coronary heart disease database, resulting in the identification and early detection of patient risk groups.

1 Introduction

Data mining is a novel approach to data analysis and knowledge extraction from databases. Data Mining Server (DMS) is an internet service which enables execution of some data mining tasks in a very simple way. Its goal is to make recent scientific developments in the field of knowledge discovery accessible to the general public, including scientists in the fields of medicine and biology.

The Data Mining Server (DMS) is realized as an interactive Web site and the user interface is based on internet browsers. Data analysis is performed at the server site so that users do not have problems with software downloads and maintenance. The site is available at the Web address http://dms.irb.hr and its home page is presented in Figure 1.

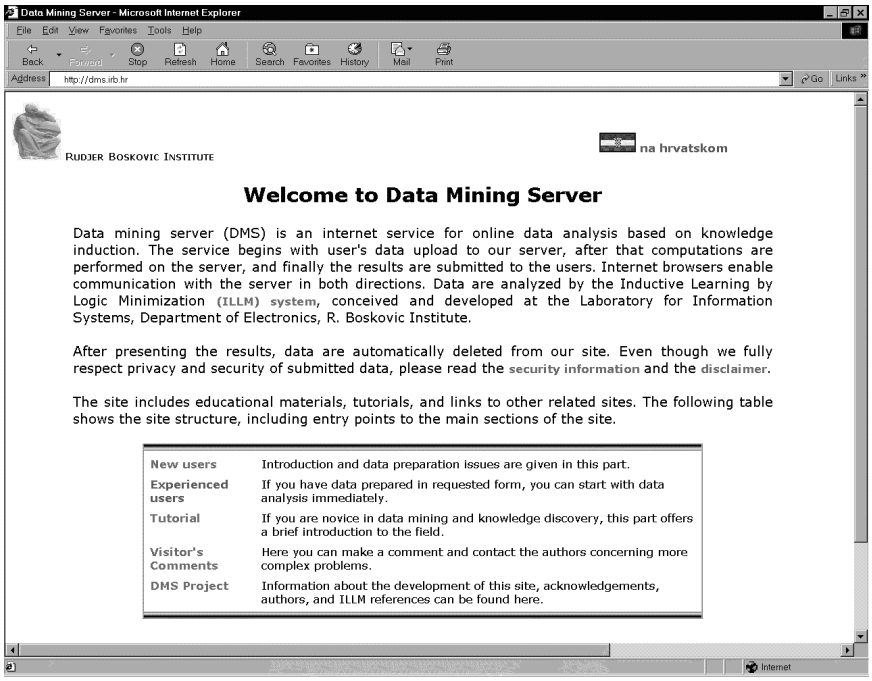

Figure 1: The first server page.

The site consists of two main parts. The first part contains educational materials and tutorials, including many links to other sites describing data mining techniques. In this part both experienced users and non-experts can learn a lot about data mining problem definition, data preparation, interpretation of the results obtained by the analysis, available data mining tools, available literature, lessons learned and similar topics. The second part enables on-line execution of some data analysis procedures. Its main task is induction of rules describing relevant subgroups. Each rule has the form of a conjunction of conditions (literals) that are automatically constructed from the user submitted data. Induction is based on heuristic search for one or more hypotheses that optimally describe the selected example class. Additionally, it is possible to detect noisy (erroneous) examples and outliers in the database. This option can be used independently of the rule induction results as a preprocessor for any data analysis procedure or as a data consistency test. Section 2 describes the Data Mining Server and shows how it can be used for medical data analysis, while Section 3 presents the results of risk group identification obtained by analysing the data about coronary heart disease (CHD)

patients, collected at the Institute for Cardiovascular Prevention and Rehabilitation, Zagreb, Croatia.

2 Data analysis by the Data Mining Server

The Data Mining Server (DMS) offers an implementation of machine learning algorithms for supervised learning. This means that data analysis uses knowledge discovery techniques which try to find descriptions of a given class of instances (examples) in contrast to instances that are not in this class. The induced rules can be used either for classification purposes or for understanding underlying connections among the data (Kukar etal.[5]).

The input data must have the form of a table in which every instance (patient in this case) is represented in a separate row. Patients are described by a fixed set of descriptors (attributes). Attribute values represent columns of the input data table. Some attribute values may be unknown. Attributes must be of one of three possible types: nominal, continuous, or discrete. Table 1 illustrates a small part of the input data file used in the experiments presented in this work. In its first row there are attribute names, while every consequent row represents a patient from the CHD database. Every patient is described by the attribute values corresponding to the names from the first row. In this table attributes Name and Sex are of type nominal, Stress and Diag are of type discrete (Stress value 1 corresponds to negative, 2 to positive, and 3 to very positive; Diag has levels 1-5 where 1 denotes no CHD and 5 corresponds to very ill CHD patients) while other attributes are of type continuous. The question mark denotes an unknown value.

Data analysis based on supervised inductive learning requires that the problem must have the target attribute and the target (selected) class. In Table 1 attribute Diag is selected as the target attribute (note the exclamation mark at the beginning of the attribute name) and attribute value 3 or higher is selected as the value discriminating the target class instances from the others. Again, the exclamation

Table 1: A small part of the input data table.

Name	SEX	Age	BMI	Stress	Tryglic.	Fibrinogen	HOL_ST_s.d.	!Diag
SB	male	64	27.30	2	1.74	4.0	0.5	!3
RK	male	57	25.30	1	?	3.5	0.2	2
IC	male	65	25.15	1	1.68	5.5	1.8	!4
AB	female	19	20.00	1	1.20	2.5	0.0	1
DK	male	46	32.95	3	2.99	3.1	0.2	2

mark is used to denote the target class. One and only one attribute must be used as the target attribute. In contrast to this, more than one attribute value can be selected for the target class definition but there must always remain some non target class examples. The definition of the target attribute and the target class are necessary in the supervised data analysis approach because the object of the analysis is the induction of rules which describe target class instances (examples) by attribute values other than the target attribute. In the concrete medical domain, the target attribute is diagnosis (Diag). Its value 3, 4, or 5 denote patients with confirmed CHD. The object of induction is the search for subgroups (rules) which describe CHD patients in contrast to not ill persons (non-CHD), i.e., non target class instances. Their diagnosis value is 1 or 2.

2.1 Subgroup discovery

The main result of the induction process implemented in the Data Mining Server is the detection and description of relevant subgroups of the target class examples (CHD patients in our experiments). A subgroup is described by its properties in the form of a rule which consists of a conjunction (logical AND operation) of conditions. If the rule is satisfied (true) for the concrete attribute values of a person then the person should be a patient with the CHD disease. The conditions can have only logical values true or false. The form of induced conditions (also called features or literals in the machine learning terminology) is determined by the attribute types. For nominal attributes conditions of the form $attr\_value$ **equal** (or **not equal**) $some\_value$ are constructed, for continuous attributes of the form $attr\_value$ **greater than** (or **less than**) $some\_value$ are constructed, and for discrete attributes conditions are constructed as if they were both nominal and continuous.

Figure 2 presents the induction result obtained for the available coronary heart disease database consisting of 238 instances: 111 of them have been classified by medical experts as CHD patients (their Diag attribute value is 3 or higher) and 127 of them represent either healthy persons or patients with non-CHD diagnosis. Every instance (patient) is in the database described by 41 attributes, including important disease risk factors described in Maron etal. [6]. A small part of these attributes is presented in Table 1. The attribute set includes anamnestic data (11 attributes like age and family anamnesis, referred as diagnostic stage A in Section 3), laboratory test results (stage B with 6 attributes like trygliceride and fibrinogen values), the resting ECG data (stage C with 5 attributes like detection of serious arrhythmias and conduction disorders), the exercise test data (8 attributes like ST segment depression and hypertensive reaction), echocardiogram results (2 attributes: left ventricular diameter and left ventricular ejection fraction),

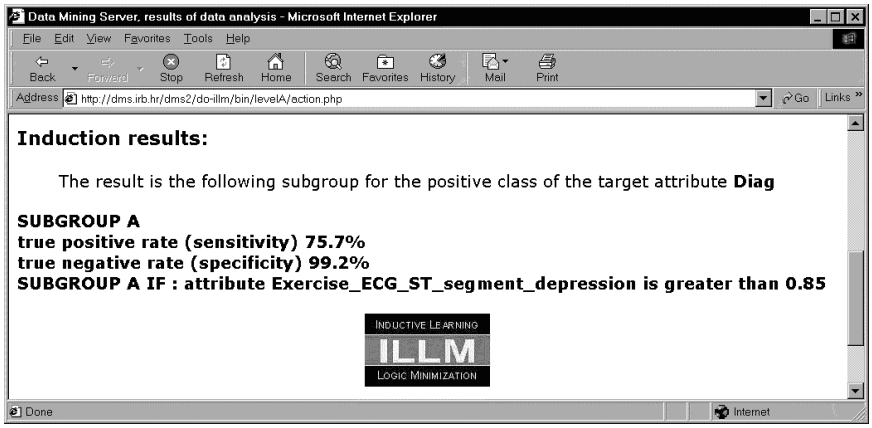

Figure 2: The subgroup induced for the CHD domain.

vectorcardigram results (2 attributes: transmural MI and left ventricular hypertrophy), and long term continuous ECG recording data (4 attributes like detection of arrhythmias and ST segment depression value). Additionally, there are two administrative attributes (like person's initials) and the classification attribute representing the diagnosis.

The result in Figure 2 (Subgroup A) has been obtained using the subgroup discovery algorithm using default parameters (default generality value equal to 1) which tends to construct rules that are correct for a relative small number of positive cases but which does not cover negative cases (high specificity rules). Details of the subgroup discovery algorithm can be found in Gamberger and Lavrač [3]. Induction of such a subgroup is very easy using DMS. Figure 3 shows the web page which is used to start the induction process. It can be reached directly from the first page through the link named *Experienced users* or after reading the instructions for the *New users*. In order to perform induction the user must supply the name of the file containing the data (in a flat ascii file, one instance per line) and select the requested generality of the subgroup. In Figure 2 it can be noticed that the induced rule has very good covering properties (it successfully detects about 75% of all CHD patients and only one of the 127 negative cases has been erroneously classified as having CHD). The medical experts are not surprised by this result because the condition is based on the ST segment depression value during a controlled exercise. Even the induced discrimination value of 0.85 millimeters is rather expected.

If the same database is used but induction is performed with a high requested generalization parameter value (50 or 100) again a very simple rule with only one condition: $Hol\_ECG\_ST\_segment\_depression > 0.65$ describing a group

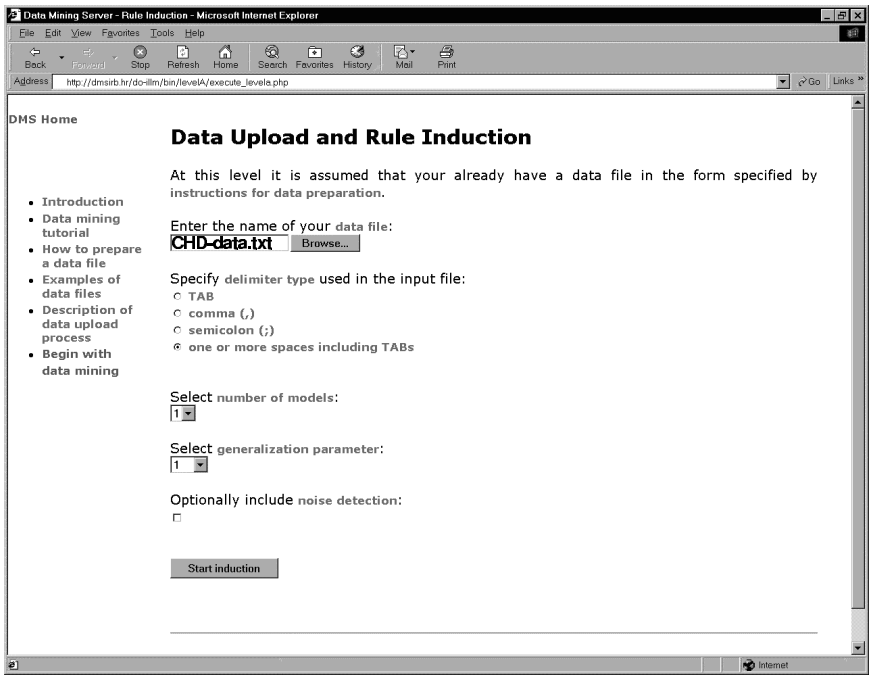

Figure 3: The data upload page.

of CHD patients is obtained. Its sensitivity and specificity are 95.5% and 96.9%, respectively. It can be noticed that this rule covers much more positive patients (even 95% of them) and that it is only slightly worse on negative instances (it erroneously classifies 4 negative cases into the positive class). But it can not be stated that the second subgroup is better than the first one generated with a low generalization parameter value. Any of the two may be preferred by the experts. The task of the DMS is to enable induction of subgroups that are potentially interesting while it is the task of the expert who uses the tool, to direct the search by selecting different generalization parameter values and to finally decide which of induced subgroups may be useful for disease description purposes.

2.2 Noise and outlier detection

The noise (error) and outlier detection procedure, described in detail in Gamberger etal. [2], can be optionally selected when starting the induction process (see Figure 3). Noise occurs if some attribute values (or even the classification attribute itself) have been incorrectly measured or incorrectly recorded in the database. In

this sense the noise detection procedure represents a consistency test that may improve the reliability of data in the database. The other type of detected examples are outliers, i.e., correct cases with some exceptional properties. Detection and expert analysis of such examples may be important for understanding the relations in the database.

The result of noise and outlier detection does not depend on the selected generalization parameter value and the outcome of noise/outlier detection does not influence the subgroup construction. The procedure results in a list with up to five instances (patients) that are detected as potential outliers in the uploaded database (Figure 4). The order of instances corresponds to the order in which they have been detected.

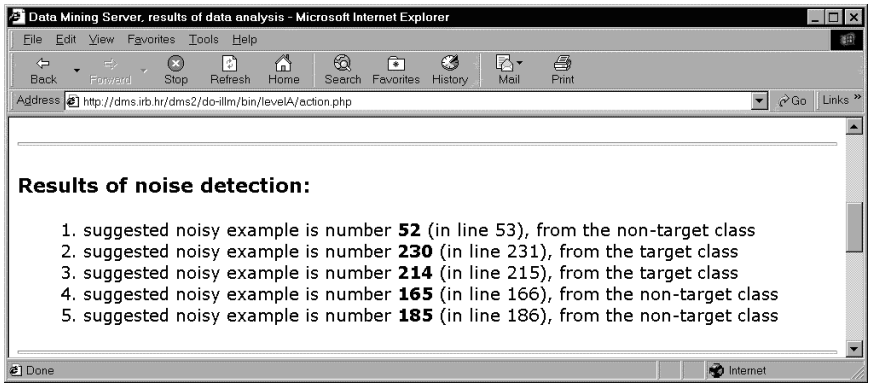

Figure 4: The results obtained by the noise detection procedure.

Note that the procedure can not guarantee that the detected examples indeed represent errors or outliers or that all such examples are detected. In any case detailed expert analysis is necessary to confirm whether the noise and outlier detection was successful and if so, what is the reason for their occurence.

In the case of the CHD database, the iterative usage of the noise detection procedure has enabled the detection of some mistakes and imprecisions that have been already corrected (Gamberger etal. [1]). The approach managed to detect a serious mistake in the data entry process and a few cases in which diagnostic classification was not systematically used throughout the database. At the current stage patients detected by the noise detection procedure are cases that can be accepted as outliers. Three of them are from the non target class (patients with not confirmed CHD) which do have relative high ST segment depression during exercise. This has been known as a very important disease indicator, as confirmed by our experiments for subgroup discovery. In all three cases echocardiography and vectorcardiography

results are negative, demonstrating their correct classification in spite of the bad exercise ECG results. By a more detailed analysis it was found that the analysed patients, although different in age and sex, were all rather fat (body mass index near to 30). The two remaining detected noisy cases are from the CHD class; one of them is a very heavy CHD patient with diagnosed cardiopathia dilatativa and the second one is a confirmed CHD patient with an atypical CHD that can be detected only by echocardiography.

3 Coronary heart disease risk group identification

Both CHD patient subgroups induced in Section 2.1 use attributes that are known as good diagnostic indicators. In this sense the induced knowledge only confirms the existing expert knowledge. Potentially interesting new knowledge can be obtained if we intentionally eliminate some important attributes from the database and force the system to search for other relations among the data. This approach has been tested for various subsets of the available attribute set. The goal was to try to identify CHD risk groups which might be useful for the early disease detection and to test relative importance of some attributes.

Data about potential CHD patients are in general practice available at three different diagnostic stages: A anamnestic data, B laboratory test, C ECG measurements at rest (Maron etal. [6]). The induction process has been repeated for every stage so that only data available at the respective stage could be used in rule conditions. Note that for this purpose the same database may be used but so that the names of attributes that should be excluded from the induction process are changed so that their name begins with a question mark.

Table 2 summarizes descriptions of the subgroups that have been detected at different stages. The process was performed according to the descriptive induction methodology proposed in Gamberger and Lavrač [3]. The process was iterative and explicitly driven by the medical expert. At different stages additional data refinements have been performed in order to obtain satisfactory results. For example, at stage A it was very difficult to get reasonably good rules for the complete domain. It has turned out that much better subgroups could be induced when the domain has been separated into the male and the female population. Also some attributes (like smoking status at stage A or heart rate at stage C) have been eliminated from the database because the collected data have been identified as unreliable or if rules without these attributes were preferred.

Table 2 includes in its second column the so called *supporting factors*. They are not obtained by the DMS application, but by the statistical significance tests. In this process, statistical values are computed for two populations. The target population consists of CHD patients included into the analyzed subgroup, whereas the refer-

Table 2: Induced subgroup conditions (principal factors) and their statistical characterisations (supporting factors). Subgroup A1 is for males, subgroup A2 for females, while subgroups B1, B2, and C1 are for both genders.

	Principal Factors	**Supporting Factors**
A1	positive family history age over 46 year	psychosocial stress cigarette smoking hypertension overweigth
A2	body mass index over 25 kgm^{-2} age over 63 years	positive family history hypertension slightly increased LDL cholesterol normal but decreased HDL cholesterol
B1	total cholesterol over 6.1 $mmolL^{-1}$ age over 53 years body mass index below 30 kgm^{-2}	increased triglycerides value
B2	total cholesterol over 5.6 $mmolL^{-1}$ fibrinogen over 3.7 $mmolL^{-1}$ body mass index below 30 kgm^{-2}	positive family history
C1	left ventricular hypertrophy	positive family history hypertension diabetes mellitus

ence population are all the healthy subjects. Statistical significance is computed for all available risk factors using the χ^2 test with 95% confidence level ($p = 0.05$).

Conclusion

This work presents an application of the Data Mining Server and its subgroup discovery tool for the coronary heart disease domain. It can be expected that this publicly available server will enable that the same approach can be applied to other medical, and not only medical domains.

The presented CHD risk groups obtained by expert guided iterative usage of the server seem to represent interesting medical knowledge, as confirmed by the expert involved in the described experiment. Given that medical experts dislike short rules and prefer descriptions including as much supportive evidence as possible, the detection of supporting factors and their inclusion in subgroup descriptions is very

important to achieve descriptions that are reasonably complete and acceptable for medical practice.

Acknowledgment

This work was supported by the Croatian Ministry of Science and Technology, the Slovenian Ministry of Education, Science and Sport, and the EU funded project Data Mining and Decision Support for Business Competitiveness: A European Virtual Enterprise (IST-1999-11495). We are grateful to Goran Krstačić from the Institute for Cardiovascular Prevention and Rehabilitation, Zagreb, Croatia for his collaboration in the experiments in CHD risk group detection.

References

[1] Gamberger, D., Lavrač, N., Krstačić, G. & Šmuc, T. (2000) Inconsistency tests for patients records in a coronary heart disease database. In *Proc. of International Symposium on Medical Data Analysis (ISMDA 2000)*, pp. 183–189.

[2] Gamberger, D., Lavrač, N. & Grošelj, C. (1999) Experiments with noise filtering in a medical domain. In *Proc. of the International Conference of Machine Learning (ICML'99)*, pp. 143–151.

[3] Gamberger, D. & Lavrač, N. (2002) Descriptive induction through subgroup discovery: a case study in a medical domain. In *Proc. of the International Conference of Machine Learning (ICML'2002)*, pp. 163–170.

[4] Goldman, L., Garber, A. M., Grover, S. A. & Hlatky, M. A. (1996). Cost-effectiveness of assessments and management of risk factors. *Journal of American College Cardiology*, *27*, 1020–1030.

[5] Kukar, M., Kononenko, I., Grošelj, C., Kralj, K. & Fettich, J. J. (1998). Analysing and improving the diagnosis of ischaemic heart disease with machine learning. *Artificial Intelligence in Medicine*, *16*, 25–50.

[6] Maron, D., Ridker, P. M. & Pearson, A. T. (1998). Risk factors and the prevention of coronary heart disease. In *Wayne A.R., Schlant R.C., Fuster V. : HURST'S: The Heart*, (pp. 1175–1195), McGraw–Hill, NY.