ELSEVIER

Artificial Intelligence in Medicine 28 (2003) 27–57

Artificial Intelligence in Medicine

www.elsevier.com/locate/artmed

Active subgroup mining: a case study in coronary heart disease risk group detection

Dragan Gamberger[a], Nada Lavrač[b,*], Goran Krstačić[c]

[a]*Rudjer Bošković Institute, Zagreb, Croatia*
[b]*Jožef Stefan Institute, Ljubljana, Slovenia*
[c]*Institute for Cardiovascular Prevention and Rehabilitation, Zagreb, Croatia*

Received 9 May 2002; received in revised form 11 January 2003; accepted 15 January 2003

Abstract

This paper presents an approach to active mining of patient records aimed at discovering patient groups at high risk for coronary heart disease (CHD). The approach proposes active expert involvement in the following steps of the knowledge discovery process: data gathering, cleaning and transformation, subgroup discovery, statistical characterization of induced subgroups, their interpretation, and the evaluation of results. As in the discovery and characterization of risk subgroups, the main risk factors are made explicit, the proposed methodology has high potential for patient screening and early detection of patient groups at risk for CHD.

Keywords: Coronary heart disease; Active mining; Machine learning; Subgroup discovery; Risk group detection; Non-invasive cardiovascular tests

1. Introduction

This paper presents an approach to active mining of patient records aimed at discovering patient groups at high risk for coronary heart disease (CHD). This section presents the backgrounds of this work: the problem of coronary heart disease risk group detection, the subgroup discovery task and the task of active mining of patient risk groups, followed by the outline of this paper.

* Corresponding author. Tel.: +386-61-177-3272; fax: +386-61-125-1038.
E-mail addresses: dragan.gamberger@irb.hr (D. Gamberger), nada.lavrac@ijs.si (N. Lavrač), goran.krstacic@zg.hinet.hr (G. Krstačić).

0933-3657/03/$ – see front matter
doi:10.1016/S0933-3657(03)00034-4

1.1. Coronary heart disease and the problem of risk group detection

Atherosclerotic coronary heart disease is one of the world's most frequent causes of mortality and an important problem in medical practice. Many extensive epidemiological studies have been performed with the intention to detect and evaluate factors that increase the risk of this cardiovascular disease. The well-known Framingham Heart Study, which began in 1948 with a sample of about 5000 people, followed up for the period of 40 years [2,40], is one of such studies. As a result, we know a lot about CHD risk factors including artherosclerotic attributes, living habits, hemostatic factors, blood pressure, and metabolic factors. Clinical studies have revealed plausible biological links between many risk factors and atherosclerosis [22]. In addition, it was detected that coexistence of risk factors increases the disease rate.

Risk factors can be classified into four categories, based on the evidence supporting their association with the disease, the usefulness of measuring them, and their responsiveness to intervention [32]. Category I consists of the most important risk factors for which high correlation with CHD rate has been proved (cigarette smoking, LDL cholesterol, high fat cholesterol diet, hypertension, left ventricular hypertrophy [10] and thrombogenic factors). Category II includes risk factors for which the correlation with CHD is likely (diabetes mellitus, physical inactivity, HDL cholesterol, triglycerides [3], obesity and postmenopausal status for women). Category III is formed of risk factors associated with increased CHD rate that, if modified, may decrease the risk (psychosocial factors, lipoprotein, homocystein, oxidative stress and no alcohol consumption). In contrast to the previous categories, Category IV consists of risk factors associated with the increased CHD rate which can not be influenced (age, male gender, low socioeconomic status and family history of early CHD onset).

Today's CHD prevention relies practically on two significantly different concepts.

1. General education of the whole population about known risk factors, especially about life-style factors. On the one hand, the results of this approach can be evaluated as very good, since a significant part of the population is now aware of CHD risk factors. However, its practical influence is estimated as small because people are not ready to accept the suggestions seriously before the occurrence of first actual signs of the disease.
2. Risk factor screening in general practice by data collection performed in three different stages.
   - 2.1. Collecting anamnestic information and physical examination results, including risk factors like age, positive family history, weight, height, cigarette smoking, alcohol consumption, blood pressure, and previous heart and vascular diseases.
   - 2.2. Collecting results of laboratory tests, including information about risk factors like lipid profile, glucose tolerance, and trombogenic factors.
   - 2.3. Collecting ECG at rest test results, including measurements of heart rate, left ventricular hypertrophy, ST segment depression, cardiac arhythmias and conduction disturbances.

The data collected in general practice screening can be used as a basis for detecting patients at risk for coronary heart disease. In many cases with significantly pathological test

values (especially, for example, left ventricular hypertrophy, increased LDL cholesterol, decreased HDL cholesterol, hypertension, and intolerance glucose) the decision is not difficult. However, the problem of disease prevention is to decide in cases with slightly abnormal values and in cases when combinations of different risk factors occur.

1.2. Active mining and subgroup discovery

The process of knowledge discovery in databases (KDD, [16]) consists of a sequence of steps, including problem understanding, data understanding and preparation, data mining, result interpretation and evaluation, and finally, the use of induced knowledge. This induction process is iterative and interactive. It is iterative, since many steps may need to be repeated before a satisfactory solution is found. It is also interactive, assuming expert's involvement in most of the phases of the knowledge discovery process.

In this work, the data mining step corresponds to subgroup mining. In the addressed patient risk group detection application, the expert's role is of ultimate importance for the success of the knowledge discovery process. Expert's involvement supports *active mining* of patient groups at high risk for coronary heart disease.

Active mining is an approach to data mining, emphasizing the importance of expert's involvement in the discovery process, as opposed to fully automated knowledge discovery approaches. This process, propagated in the large Japanese active mining project (2001–2005, [34]), is in line with the KDD process, which is also iterative and interactive.

The main ingredient of the proposed active mining methodology for risk group detection is an algorithm supporting expert-guided discovery of 'interesting' subgroups in a population of individuals. As in the MIDOS subgroup discovery approach [43], the addressed subgroup discovery task is defined as follows: given a population of individuals and a property of those individuals we are interested in, find population subgroups that are statistically 'most interesting', e.g. are as large as possible and have the most unusual distributional characteristics with respect to the property of interest.

The subgroups, discovered by an expert-guided rule induction process, are represented as `if-then` rules of the form Class ← Cond, where Class is the property of interest (e.g. the target class like 'coronary heart disease' in our case study), and Cond is a conjunction of conditions (e.g. a conjunction of features describing the illness).

1.3. Active mining of coronary heart disease risk groups

The database that was available for this study was collected at the Institute for Cardiovascular Prevention and Rehabilitation in Zagreb, Croatia. Given that the dataset has been collected at a specialized medical institution, its disadvantage is that it is not an appropriate epidemiological CHD database reflecting actual CHD occurrence in a general population, as about 50% of gathered patient records represent CHD patients. On the other hand, the advantage of this dataset is that it includes a sufficient number of records of different types of the disease. In order to have an adequately large number of CHD patient records from a general population, a very large data collection from general practice screening would need to be made available for the CHD risk group discovery experiments. Such a dataset has not been available. Consequently, we had to do our best to overcome this

deficiency by using a subgroup discovery system (system Data Mining Server (DMS) described in this paper) as a toolbox supporting the medical expert and the knowledge engineer in the interactive and iterative active mining process of subgroup discovery from the available biased dataset.

Semi-automated active mining may at a first glance seem inappropriate as means for knowledge discovery. This paper shows that, on the contrary, such an approach may be very productive and useful, overcoming the deficiencies of automated knowledge discovery from an inappropriate data collection in the cases when it is impossible to gather a sufficiently large unbiased data collection from general practice screening. This is achieved by the extensive use of the available expert knowledge and active involvement of the expert in all steps of the discovery process. In several iterative runs of the subgroup discovery algorithm, in which the expert can affect the attributes used for learning, as well as the generality and the complexity of induced rules, this active mining process can result in the discovery of relevant new rules. The main characteristic of the methodology is that the obtained results reflect the knowledge and experience of the medical expert.

The evaluation of the quality of rules on a separate validation set is a necessary part of the proposed methodology. One should be aware, however, that validation results will depend on the tested population as well as on the data collecting procedures (equipment, standards, and medical practice). In this work, the applicability of the achieved results has been evaluated on two independent test sets, one of them being a validation set of employees of two large Croatian companies. This validation set can be considered as a small epidemiological validation set, although the dataset is too small to give reliable epidemiological results, and includes only a part of the interesting population for CHD screening, as, for instance, unemployed or retired people as well as children had not been included in this set.

Given the limitations and specifics of the available data, and the biases of the expert involved in the experiments, the induced subgroup descriptions should not be considered as the ultimate CHD risk group descriptions to be used for CHD risk group detection worldwide. The limitations and biases of the actual experiments are described in sufficient detail for the reader to be able to judge the generality of the induced results and their applicability in general medical practice. In our view, the main contribution of this paper is a proposed subgroup discovery methodology, rather than the resulting CHD risk group descriptions: the paper should be viewed as a methodological study suggesting how to address the problem of patient risk group detection. Its main advantage is that a sequence of steps constituting the active subgroup mining process is proposed, most of which are supported by algorithms implemented in the Data Mining Server (DMS), available on-line for public use at the web site, http://dms.irb.hr.

1.4. Paper outline

The rest of the paper is organized as follows. Section 2 describes the available CHD dataset. Section 3 illustrates the individual steps of the proposed active subgroup mining methodology as implemented in the Data Mining Server: data selection, cleaning and transformation in the required format, followed by interactive and iterative subgroup discovery and statistical characterization of subgroups. The discovered risk groups are then

interpreted and formulated in natural language sentences (Section 4), and their applicability is evaluated on two independent datasets (Section 5). The paper concludes with a section on related work (Section 6).

2. The CHD dataset

The database consists of records of patients who entered the Institute for Cardiovascular Prevention and Rehabilitation in the period of a few months. The set of descriptors represents all potentially interesting and typically available information about patients. The descriptor set includes anamnestic parameters (stage A: 10 items, see Table 1), parameters describing laboratory test results (stage B: seven items, Table 2), ECG at rest (stage C: five items, Table 3), the exercise test data (five items), echocardiography results (two items), vectorcardiography results (two items), and long-term continuous ECG recording data (three items). In this study, only anamnestic, laboratory and ECG at rest data were used to form the risk group descriptions, since, for screening purposes, one needs to take into account only those parameters that can be observed and measured by general practitioners.

In this study, only patients with complete data were included into the dataset, resulting in the dataset with 238 patient records: 111 CHD patients (positive cases), and 127 people

Table 1
The names and characteristics of 10 anamnestic descriptors used at stage A

Descriptor	Abbreviation	Characteristics
Sex	SEX	Man, woman
Age	AGE	Continuous (years)
Height	H	Continuous (cm)
Weight	W	Continuous (kg)
Body mass index	BMI	Continuous (kg m^{-2})
Family anamnesis	F.A.	Negative, positive
Present smoking	P.S.	1: negative; 2: positive; 3: very positive
Systolic blood pressure	SBP	Continuous (mmHg)
Diastolic blood pressure	DBP	Continuous (mmHg)
Stress	STR	1: negative; 2: positive; 3: very positive

Table 2
The names and characteristics of seven laboratory test descriptors additionally used at stage B

Descriptor	Abbreviation	Characteristics
Total cholesterol	T.CH.	Continuous (mmol l^{-1})
Trygliceride	TR	Continuous (mmol l^{-1})
High-density lipoprotein	HDL/CH	Continuous (mmol l^{-1})
Low-density lipoprotein	LDL/CH	Continuous (mmol l^{-1})
Uric acid	U.A.	Continuous (μmol l^{-1})
Fibrinogen	FIB	Continuous (g l^{-1})
Glucose	GLU	Continuous (mmol l^{-1})

Table 3
The names and characteristics of five ECG at rest descriptors added to stages A and B descriptors at stage C

Descriptor	Abbreviation	Characteristics
Heart rate	HR	Continuous (beats min^{-1})
ST segment depression	ECGst	1 if <1 mm, 2 if 1–2 mm, 3 if $\geq$2 mm (1 mm corresponds to 0.1 mV)
Serious arrhythmias	ECGrhyt	Negative, positive
Conduction disorders	ECGcd	Negative, positive
Left ventricular hypertrophy	ECGhlv	Negative, positive

without CHD (negative cases). Among them there are 177 males (80 positive and 97 negative) and 61 females (31 positive and 30 negative). The database collected at the institute is not an epidemiological CHD database reflecting actual CHD occurrence in a general population, since about 50% of gathered patient records represent CHD patients. Moreover, the included negative cases (patients who do not have CHD) are not randomly selected people but individuals with some subjective problems or those considered by general practitioners as potential CHD patients, and hence sent for further investigations to the institute. However, the dataset is very valuable since it includes a number of records of different types of the disease. The consequences of the biased dataset are two-fold:

1. Features induced as main subgroup characteristics detected by the subgroup discovery algorithms may be influenced by the bias of this dataset. Among the patients that are in the dataset, there are many patients who have had the disease already for a long time and have been exposed to medicamental therapy which reduces important risk factors. Moreover, most patients have already changed their life styles concerning smoking and nutrition habits. On the other side, negative cases are people that do not have CHD but may be ill because of some other heart-related disease. Consequently, their test values may be different from typical values expected for a healthy person.
2. Subgroup statistics measured on this data are significantly different from those that can be expected in other populations (e.g. the general population). In order to enable comparisons of induced subgroups and to estimate their behaviour in different populations (e.g. the employee population, the general population, ...), the rules describing the discovered CHD subgroups should be accompanied with appropriate evaluation measures. The selected measures are the *Sensitivity* or the *true positive rate* (TPr) which represents the percentage of CHD patients described by the subgroup, and the second is *false alarm* or *false positive rate* (FPr) which is the percentage of healthy people incorrectly classified by this subgroup as patients with CHD.[1]

[1] Sensitivity measures the fraction of positive cases that are classified as positive, whereas specificity measures the fraction of negative cases classified as negative. Let TP denote true positives, TN true negatives, FP false positives and FN false negative answers, then $\text{Sensitivity} = \text{TPr} = \text{TP}/(\text{TP} + \text{FN}) = \text{TP}/\text{Pos}$, $\text{Specificity} = \text{TN}/(\text{TN} + \text{FP}) = \text{TN}/\text{Neg}$, and $\text{FalseAlarm} = 1 - \text{Specificity} = \text{FPr} = \text{FP}/(\text{TN} + \text{FP}) = \text{FP}/\text{Neg}$.

3. Active mining using the Data Mining Server

The main ingredients of the proposed subgroup discovery methodology are implemented in the interactive Data Mining Server available on-line for public use at the web site http://dms.irb.hr. The DMS home page is shown in Fig. 1. The DMS web site consists of two main parts.

1. The first part contains educational materials and tutorials, including links to other sites describing data mining techniques. In this part users can learn about the 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.
2. The second part enables on-line execution of selected data analysis procedures, including the induction of rules describing interesting subgroups. Conditions of induced `if-then` rules are in the form of conjunction of features that are automatically constructed from the data. In addition, DMS makes it possible to detect noisy (erroneous) examples and outliers in data preprocessing.

The algorithms implemented in DMS are applicable for demonstration purposes on datasets with up to 250 training examples. A more sophisticated implementation of the algorithms has been used in the experiments described in this paper.

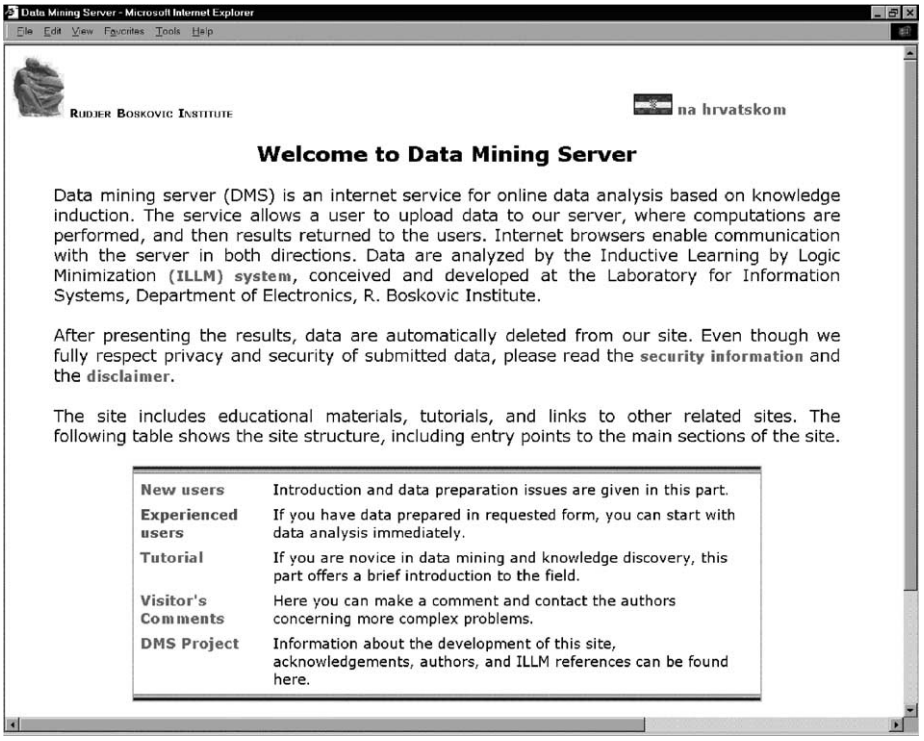

Fig. 1. The Data Mining Server home page.

Table 4
A sample input data table, including attributes from different CHD data collection stages

SEX	AGE (years)	BMI	Stress	Trygliceride	Fibrinogen	HOL_ST_s.d.	!Diag
Male	64	27.30	2	1.74	4.0	0.5	!3
Male	57	25.30	1	?	3.5	0.2	2
Male	65	25.15	1	1.68	5.5	1.8	!4
Female	19	20.00	1	1.20	2.5	0.0	1
Male	46	32.95	3	2.99	3.1	0.2	2

3.1. Data representation and preprocessing

3.1.1. Data representation

The input dataset, referred to as *training examples E*, has the form of a table in which every instance (e.g. each patient record) is represented in a separate row. Instances are described by a fixed set of descriptors (attributes). Attributes are of one of the three possible types: discrete, continuous, or integer. Table 4 illustrates a part of the input data file, consisting of attributes from all CHD data collection stages. In the first row, there are attribute names, while every consequent row represents a patient record, where a patient is described by the corresponding attribute values. In this table, attribute sex is of type discrete, stress is of type integer (stress value 1 corresponds to negative, 2 to positive, and 3 to very positive) and all other attributes are of type continuous. The question mark (?) denotes an unknown value.

In subgroup discovery, like in supervised inductive learning, a class attribute needs to be specified. In Table 4 attribute Diag is selected as the class attribute (note the exclamation mark in front of the attribute name). Diag has values 1–5, where 1 denotes no CHD and 5 corresponds to very ill CHD patients. Attribute value 3 or higher is selected as the value discriminating the target class instances from the others. Again, the exclamation mark is used to denote the target class. While only one attribute may be used as the class attribute, more than one attribute value may be selected to define the target class (examples belonging to the target class are called *positive* examples, and others are called *negative* examples). In the concrete medical domain, the object of induction is the search for subgroups (rules) which describe confirmed CHD patients with diagnosis values 3–5 in contrast to not ill people (non-CHD), i.e. non-target class instances with diagnosis values 1 or 2.

3.1.2. Data transformation

In the Data Mining Server, a training set that has initially the form of a table of attribute values (as in Table 4) is transformed into a table of truth values of features.

Features are logical conditions formed of attribute values describing examples E. Features have the form Attribute $=$ value, Attribute $\neq$ value, Attribute $>$ value, or Attribute $\leq$ value. Features are constructed by DMS from the dataset. To formalize feature construction, let values v_{ix} $(x = 1, \ldots, k_{ip})$ denote the k_{ip} different values of attribute A_i that appear in the positive examples and w_{iy} $(y = 1, \ldots, k_{in})$ the k_{in} different values of A_i appearing in the negative examples. Feature construction results in a set of features L generated as follows:

- For discrete attributes A_i, features of the form $A_i = v_{ix}$ and $A_i \neq w_{iy}$ are generated.
- For continuous attributes A_i, features of the form $A_i \leq (v_{ix} + w_{iy})/2$ are created for all neighboring value pairs (v_{ix}, w_{iy}), and features $A_i > (v_{ix} + w_{iy})/2$ for all neighboring pairs (w_{iy}, v_{ix}). The motivation is similar as in [15].
- For integer valued attributes A_i, features are generated as if A_i were both discrete and continuous, resulting in features of four different forms: $A_i \leq (v_{ix} + w_{iy})/2$, $A_i > (v_{ix} + w_{iy})/2$, $A_i = v_{ix}$, and $A_i \neq w_{iy}$.

3.1.3. Noise and outlier detection

Noise (random errors) and outlier detection can be used before starting the induction process.

- Noise occurs if some attribute values (or even the class values) have been incorrectly measured or incorrectly recorded in the database.
- Outliers are correctly recorded cases with some exceptional properties.

Detection and expert analysis of noisy examples and outliers may be important for understanding the data and the relations in the database.

The implemented noise and outlier detection procedure is based on the computation of the *complexity of the least complex correct hypothesis* (CLCH value) for the given dataset. The complexity of the hypothesis is measured by the number of different features used in the rule that is true for all target class examples and false for all non-target class examples. An example is detected as noisy if its elimination enables the reduction of the CLCH value. A detailed description of the noise and outlier detection procedure, which is applied to the transformed data table, is out of the main scope of this paper (the algorithm is described in detail in [18]).

Fig. 2 presents a list of patients that have been detected as noise/outliers for the complete available CHD dataset (the DMS implementation of the procedure can detect up to five noisy examples/outliers). The interpretation of the results of noise and outlier detection for the CHD domain is given in Section 3.4.2.

3.2. Output of subgroup discovery

The output of the induction process is the description of induced subgroups of a given target class (CHD patients) in `if-then` rule form, Class ← Cond, where Cond is a

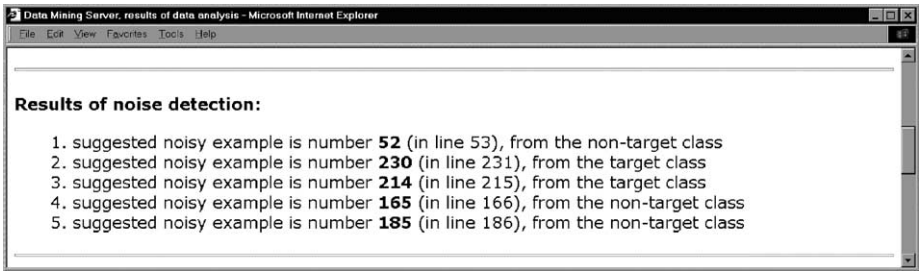

Fig. 2. A list of detected noisy examples and outliers, to be analysed by the expert.

conjunction of conditions (a conjunction of features describing the illness). The form of features depends on the attribute types, as defined in Section 3.1.2.

- For discrete attributes features have the form Attribute = value or Attribute ≠ value, e.g. ECGhlv = positive.
- For continuous attributes features have the form Attribute > value or Attribute ≤ value, e.g. AGE > 63.
- For integer valued attributes (e.g. P.S., STR, ECGst) features have the forms of both discrete and continuous attributes.

If a patient record satisfies the conditions of a rule, then it is classified as a patient with CHD disease.

Fig. 3 illustrates a result obtained for the coronary heart disease risk group detection problem, induced from a database consisting of 238 instances described by all the available attributes (including stages A–C, as well as exercise and long-term ECG, echocardiography and vectorcardiography tests). The rule in Fig. 3 was induced by the subgroup discovery algorithm using the generality parameter value $g = 3$, which tends to construct rules that are correct for a relatively small number of positive cases but which cover none or very few negative cases (rules with high *specificity*). This rule successfully detects about 75% of all CHD patients, while only one of the 127 negative cases has been erroneously classified as having CHD. The medical experts are not surprised by this good result because the rule condition involves the ST segment depression value during a controlled exercise. Even the induced discrimination value of 0.85 mm is rather expected. This rule is listed just for illustration and has no practical value for risk group screening in a general population; the reason is that the exercise ECG ST segment depression measurement is not performed in general medical practice and can therefore not be used for early risk group detection.

For illustration, let the same database be used for induction performed with a high value of the generalization parameter, e.g. $g = 50$ or 100. Induction again results in a very simple rule with only one condition: Holter ECG ST segment depression >0.65, describing a group of CHD patients. 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).

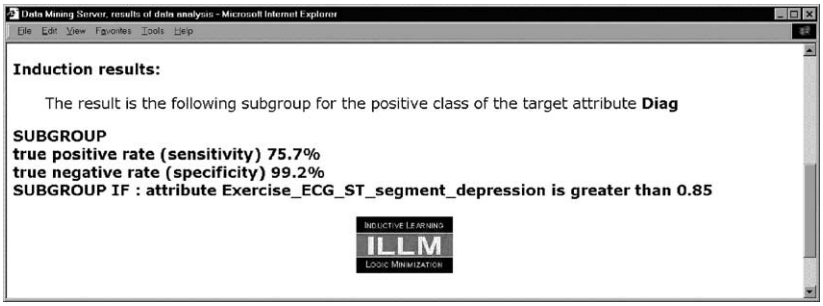

Fig. 3. A subgroup induced for the CHD domain.

An automated subgroup discovery system would give preference to the second rule, due to high sensitivity and specificity values. However, in the active mining approach both subgroups would be shown to the expert for evaluation. The above two examples show that the task of DMS is to enable the induction of subgroups that are potentially interesting while it is the task of the expert using the tool to direct the search by selecting different generalization parameter values and to finally decide which of the induced subgroups will be used for disease description and risk group detection purposes.

3.3. DMS subgroup discovery algorithms

To construct rules describing subgroups of CHD patients at risk, a new active mining method has been developed, which combines machine learning based induction of interesting subgroups and statistical analysis of the detected subgroups. Expert knowledge is included into the approach so that both the process of subgroup detection and the statistical analysis are guided by a domain expert. The final result of active mining are subgroup descriptions reformulated into natural language sentences in the interaction with the medical expert.

The subgroup discovery algorithms implemented in the Data Mining Server are:

- a subgroup discovery algorithm for individual rule construction (Algorithm SD, outlined in Section 3.3.1, with pseudo-code in Appendix A), and
- a covering algorithm involving example weighting for rule set construction (Algorithm DMS outlined in Section 3.3.2, with pseudo-code in Appendix B).

3.3.1. Algorithm SD for individual rule construction

The aim of the heuristic subgroup discovery algorithm (Algorithm SD) is the search for rules with a maximal q value, where q is defined as $q = \mathrm{TP}/(\mathrm{FP} + g)$. In the definition of q, TP are true positives (the number of CHD patients correctly classified by the rule as patients with CHD), FP are false positives (i.e. the number of non-CHD cases incorrectly classified as patients with CHD), and g is a *generalization parameter*. By searching for rules with high quality q, the heuristic confirmation rule induction algorithm tries to find rules that cover many target class examples (CHD cases) and a low number of non-target examples. The number of tolerated non-target examples, relative to the number of covered target class cases, is determined by the parameter g.

Variations of parameter g enable the expert to guide subgroup discovery by varying the TP/FP ratio. In Algorithm SD, increased generality results in more general subgroups discovered. If g value is low (1 or less) then covering of any non-target example (any non-CHD patient) is made relatively very expensive and the final result are rules that cover only few target cases but also nearly no non-target class cases. This results in rules with high specificity (low false alarm rate). If the value of g is high (10 or higher) then covering of few non-target class examples is not so expensive and more general rules can be generated. Rule quality measure q serves two purposes: first, rule evaluation, and second, evaluation of features and their conjunctions with high potential for the construction of high quality rules in subsequent iterations.

For the first purpose, a measure assigning different costs to false positives and false negatives could perform equally well, but for the purpose of guiding the search the used measure q is more appropriate. Details of this analysis can be found in [21].

Appendix A contains the pseudo-code of Algorithm SD. It shows that in addition to the generalization parameter g, the user needs to select the values of other parameters, including beam_width, defining the number of best rules induced in each step of rule induction, and min_support, defining the minimal number of target class examples covered by a rule (see a complete list of parameters in Section 3.4.3).

3.3.2. Algorithm DMS: a covering algorithm for rule set construction

Algorithm SD can generate many rules satisfying the requested condition of a minimal number of covered target class examples, defined by the min_support parameter. Acceptance of all these rules as interesting subgroups is not desired because experiments demonstrated that there are subsets of very similar rules which use almost the same attribute values and have similar prediction properties. A solution to this problem is to reduce generated rule sets so that they include only a relatively small number of rules with diverse covering properties. The weighted covering approach has been proposed for the selection of diverse rules in postprocessing [19,21].

Instead of rule postprocessing, the publicly available Data Mining Server directs rule set construction by Algorithm DMS, which is also based on the weighted covering approach. In its inner loop it calls the individual rule construction algorithm (Algorithm SD) and selects from its beam only one best rule which is included into the output rule set. To enable Algorithm SD to induce a different solution in each iteration, weights $c(e)$ for examples from the positive class are introduced and used in the quality measure which is defined as follows:

$$q = \frac{\sum_{\mathrm{TP}} 1/c(e)}{\mathrm{FP} + g}.$$

This is the same quality measure as in Algorithm SD except that the weights of true positive examples are not constant and equal to 1 but defined by the expression $1/c(e)$, changing from iteration to iteration. Implementation details can be found in [21], and the pseudo-code in Appendix B.

3.3.3. Sample run of the algorithms

Table 5 shows some of the rules induced for the data at stage C in the first loop of the DMS algorithm, with four algorithm SD iterations. The used generalization parameter value g is 10. The table includes the position of the rule in the beam after the iteration, description of the rule, computed rule quality value q and rule covering properties on the training set (TP and FP). It can be noticed that after the first SD iteration there are only rules with one feature, after the second iteration there are rules with up to two features, and so on. The iterative process stops after no further rule quality improvements are possible by adding new features to the rule conditions in the beam.

In the concrete experiment, the medical expert selected the simple rule ECGhlv = positive as the optimal final solution for level C although some rules with higher rule

Table 5
Some 'interesting' subgroups discovered from stage C data in first iteration of the DMS algorithm with generalization parameter value $g = 10$

Rule	Pos	q	TP	FP
Beam after first SD iteration				
ECGhlv = positive	1	1.529	26	7
ECGrhyt = positive	2	1.187	19	6
AGE > 63	3	1.166	42	26
FIB > 3.75	4	1.147	70	51
H < 171	10	1.098	56	41
Beam after second SD iteration				
ECGhlv = positive	1	1.529	26	7
FIB > 3.75 and H < 173	2	1.480	37	15
H < 171 and AGE > 55	8	1.400	42	20
Beam after third SD iteration				
AGE > 57 and H < 172 and TR < 1.96	1	1.857	26	4
AGE > 53 and H < 173 and FIB > 3.35	5	1.708	41	14
Beam after fourth SD iteration				
AGE > 53 and H < 173 and FIB > 3.22 and HR > 64	1	1.941	33	7

The second column is the rule position in the beam. Attribute abbreviations are described in Tables 1–3.

quality q have been detected. The reason is that the other rules were characterized by the expert as non-intuitive and too complex.

3.4. Supporting active subgroup mining in DMS

Expert involvement is desired in all phases of the subgroup discovery process.

3.4.1. Expert-guided dataset selection

Consider the data table shown in Table 4 of Section 3.1.1, where guiding subgroup discovery was presented. In Table 4 attribute *Diag* was selected as the target attribute, which was denoted by the exclamation mark at the beginning of the target attribute name, and attribute values 3 or higher were selected as the values discriminating the target class instances from the others, which is denoted by the exclamation mark in front of the target class values.

Attribute subset selection can be achieved by using a question mark (?) in front of the attribute name that should be eliminated. In our experiments, reporting on subgroup discovery at stages A–C, attributes of stages B and C were eliminated from the dataset in the stage A experiments, and attributes of stage C were eliminated in stage B experiments (in addition to all other non-stage A, B or C attributes): five exercise tests (including exercise ECG ST segment depression), two echocardiography attributes, two vectorcardiography attributes, and three long-term continuous ECG recording attributes (including Holter ECG ST segment depression).

Attribute subset selection was used for experiments at different stages A–C. In addition, the expert may decide to temporary or permanently exclude from the induction process some attributes that are already used in other rules, attributes which are expensive or hard

to measure, or attributes which are unreliable. In the available CHD domain, it has been found through experimentation that attribute *present smoking* is seldom used in the induced rules although smoking is known as an important CHD risk factor. It has been also observed that the negative *present smoking* status is more often correlated with very ill patients than with healthy people simply because ill people stopped smoking after the occurrence of serious disease symptoms. Practically, in the given dataset, the *present smoking* status does not say anything about person's smoking history; consequently, this attribute has been eliminated in all the experiments.

Instance subset selection is enabled as well. This is achieved by using a question mark (?) in front of the class value of the instance that should be eliminated. Instance subset selection can be applied in order to enable induction of more coherent subgroups. For example, at data stage A the patients have been partitioned according to sex in two subsets and a subgroup has been induced for each of them separately. Additionally, the expert may decide to permanently exclude a patient record from the dataset if it is an outlier, detected by the noise and outlier detection procedure. Examples of detected outliers are described in the following section.

3.4.2. Noise and outlier detection

The noise and outlier detection procedure can not guarantee that the detected examples indeed represent errors or outliers or that all such examples can be detected. Detailed expert analysis is necessary to confirm whether the noise and outlier detection has been successful and if so, what is the reason for the occurrence of such instances.

The iterative usage of the noise detection procedure has enabled the detection of some mistakes and imprecisions in the data [20]. One problem has been that the diagnostic classification was not systematically performed throughout the database. In a few cases the medical doctor decided to change the patient classification after the patient record has been detected as noisy. It was also shown that some attributes were inconsistently measured. For example, *exercise ST segment depression* has been measured as low also for patients that have been in such a bad condition that they could not sustain the exercise test. Such measurement values had to be transformed to value 'unknown'. We have managed also to detect a healthy person with incorrectly attributed data from an ill patient with the same name. The noise detection procedure was the one to notice this serious mistake which has been later corrected in the official hospital records.

All the patients detected by the noise detection procedure which are shown in Fig. 2 are cases that can be interpreted as outliers. Three of them are from the non-target class (patients with non-confirmed CHD) which have relatively high value of ECG ST segment depression during exercise. This is as a very important disease indicator, as confirmed by our experiments. In all three cases echocardiography and vectorcardiography results are negative, demonstrating their correct classification despite 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 cardiomyopathia dilatativa and the second one is a confirmed CHD patient with an a-typical CHD that can be detected by echocardiography and myocardial perfusion imaging.

These observations demonstrate how the analysis of the detected noisy examples or outliers may help the expert in data, disease, and diagnostic procedures understanding.

3.4.3. Parameter setting

In subgroup discovery, the expert is involved in guiding the search for 'interesting' subgroups based on the existing expert knowledge. For the coronary heart disease risk group detection problem used in this study, an ideal subgroup is described by a rule that is correct for many (or all) target class cases (CHD patients), and incorrect for all non-target class cases (healthy subjects). In practice, a good subgroup includes many target class cases, but also some healthy people (false positives). By allowing the number of false positives to increase, which can be achieved by increasing the value of generalization parameter g of the SD algorithm, the domain expert can guide the system to induce more general subgroups of patients, at a cost of covering an increased number of false positives. In this way, the generalization parameter enables the expert to induce different subgroups from the same dataset. As mentioned in Section 3.4.1, subgroup variation can be achieved also by selecting a subset of attributes to be used in rule induction. In the CHD problem, this corresponds to risk factor selection. By combining these two techniques the expert may interactively guide the inductive search process through many iterations until interesting subgroups have been detected.

In addition to the g parameter, there are other parameters whose values need to be adjusted by the user (or the default values are used). All the parameters, their meaning and the default values are listed below.[2]

- g: the generalization parameter ($0.1 < g < 100$, default value $= 1$),
- number: the maximum number of subgroups induced by the DMS algorithm (default value $= 1$),
- min_support: minimal support for rule acceptance (default value $= \sqrt{P}/E$, where P is the number of target class examples in E) which indirectly defines the minimal number of target class examples which must be covered by every subgroup,
- beam_width: number of rules in the 'beam' of rules that the algorithm evaluates as 'best' in each iteration of the beam search Algorithm SD (default value $= 20$),
- max_number_of_iterations: maximal number of iterations of Algorithm SD (default value $= 5$) which defines maximal number of features in the generated rule and which indirectly defines rule complexity, and
- covering_weight_value: number which is added to the $c(e)$ in the DMS algorithm if e is covered by the constructed rule (default value $= 1$) which indirectly affects the diversity of induced subgroups.

The process of expert-guided subgroup discovery was performed as follows. For every data stages A–C, the DMS algorithm was used a number of times with different g parameter values in the range 0.5–100, and a fixed number of selected output rules equal to 3. The rules induced in this iterative process were shown to the expert for selection and interpretation. The inspection of 15–20 rules for each data stage triggered further experiments. Concrete expert suggestions were to limit the number of features in rule body and to try to avoid the generation

[2] In the publicly available Data Mining Server, the users can change only the first two parameters.

of rules whose features would involve expensive and/or unreliable laboratory tests. Consequently, we have performed further experiments by intentionally limiting the feature space and the number of iterations in the main loop of Algorithm SD.

3.4.4. Expert-guided subgroup selection

The main selection criterion used by the expert was the intuitive understanding of the induced subgroups in terms of the logical connections of features in the induced rule. For example, the rules with attributes like patient's height and trygliceride value were not intuitive, in contrast to rules with attributes body mass index and trygliceride value. The expert also disliked the rules that included features which may be the result of medicamental therapy (like the low total cholesterol value) or changed life style (no stress, diet) of the known CHD patients.

Actionability of induced subgroups is also an important issue. In this sense it is important how easily and reliably the attributes included into the rule can be measured. Attributes sex and age are very favourable while stress and especially present smoking status were considered as unreliable, as their use could result in unactionable knowledge. Furthermore, the cost and time needed for measuring attribute values is also important. If the cost and complexity of exercise and long ECG measurement were low then they could be included in the standard general practice, resulting in significantly increased quality of early CHD detection.

Finally, estimated TPr and FPr values are decisive. For the purpose of early disease detection, subgroups with relatively high FPr are acceptable. Nevertheless, during the induction and selection process the intention was to keep it below 10% whenever possible. The necessary TPr of every subgroup is not very high. Values above 25% are acceptable. But the intention is to induce different subgroups to cover diverse population segments in order that the set of induced subgroups covers the intended 85–95% of the total CHD population.

4. Results: coronary heart disease risk groups, their interpretation and deployment

The process of expert-guided subgroup discovery was performed for every data stages A–C, using the DMS algorithm a number of times with different g parameter values. In this iterative active mining process, the expert has selected five interesting CHD risk groups. Their description, interpretation and deployment potential are described below.

4.1. Results of subgroup discovery

Five interesting subgroups A1, A2, B1, B2, C1 shown in Table 6 were selected from a set of subgroup descriptions induced by running the DMS algorithm with g parameter values in the range 0.5–100 and by intentionally limiting the feature space and the number of iterations in the main loop of the SD algorithm. Subgroups A1, A2, B1, B2 and C1 were selected using two main criteria: intuitive understanding of subgroups and the actionability of subgroup descriptions for targeting a population at risk for CHD which would need to be called to the institute for further medical testing. The features describing the induced subgroups are called the *principal risk factors*.

Table 6
Induced subgroups in the form of rules

Subgroup	Expert Selected Subgroups	g
A1	CHD $\leftarrow$ F.A. = positive and AGE $>$ 46	14
A2	CHD $\leftarrow$ BMI $>$ 25 and AGE $>$ 63	8
B1	CHD $\leftarrow$ T.CH. $>$ 6.1 and AGE $>$ 53 and BMI $\leq$ 30	10
B2	CHD $\leftarrow$ T.CH. $>$ 5.6 and FIB $>$ 3.7 and BMI $\leq$ 30	12
C1	CHD $\leftarrow$ ECGhlv = positive	10

Rule conditions are conjunctions of principal factors. Subgroup A1 is for male patients, subgroup A2 for female patients, while subgroups B1, B2, and C1 are for both male and female patients. The subgroups are induced from different attribute subsets with corresponding g parameter values indicated in column g.

4.2. Subgroup interpretation

The next step in the subgroup discovery process is expert description of discovered subgroups and the interpretation of potential connections among the detected subgroup characteristics. To support this process and provide further characterization of subgroups, statistical differences in distributions were computed for two populations: the target population consisting of true positive cases (CHD patients included into the analyzed subgroup), and the reference population consisting of all non-target class examples (all the healthy subjects). Statistical differences in distributions for all the descriptors (attributes) between these two populations were tested using the χ^2-test with 95% confidence level ($P = 0.05$). For this purpose, numerical attributes were partitioned in up to 30 intervals so that in every interval there were at least five instances. Among the attributes with significantly different distributions there were always those that form the features describing the subgroups (the principal risk factors), but usually there were also other attributes with significantly different value distributions. These attributes are called *supporting attributes*, and the features formed of their values that are characteristic for the discovered subgroups are called *supporting risk factors*. Supporting factors are very important to achieve pattern descriptions that are reasonably complete because medical decision process requires as much supportive evidence as possible [26].

The decision whether the supporting risk factors will be used to support user's confidence in the subgroup description is left to the expert, regardless of their actual statistical significance. In the CHD application, the expert has decided whether the proposed factors are indeed interesting, how reliable they are and how easily they can be measured in practice. The resulting supporting risk factors are listed in Table 7.

Table 7
Supporting risk factors for CHD patients belonging to subgroups A1, A2, B1, B2 and C1

Subgroup	Expert selected supporting risk factors for CHD patients
A1	Psychosocial stress, present smoking, hypertension, overweight
A2	Positive family history, slightly increased LDL cholesterol, hypertension, normal but decreased HDL cholecterol
B1	Increased triglycerides value
B2	Positive family history
C1	Positive family history, hypertension, diabetes mellitus

4.2.1. Interpretation of rules for stage A

At stage A, there are only anamnestic information and physical examination results available. At this stage it was rather difficult to find subgroups with a relatively small number of false positive predictions. The reason is a very restricted amount of available information. In order to make the problem easier, separate subgroups were developed for male and female patients.

4.2.1.1. Subgroup A1 for male patients (CHD ← positive family history and age over 46 years). The sensitivity of rule A1 for men is very good (47%) but its false positive rate is extremely high as well (27%), if measured on the training set. Supporting characteristics are psychosocial stress, present cigarette smoking, hypertension, and overweight. Both principal risk factors for this rule are non-modifiable. Positive family history is a well known and important risk factor, indicating the need for careful screening of other risk factors. The selected age margin in the second factor is rather low but it is in accordance with the existing medical experience. This low age margin is good for prevention and early CHD diagnosis, although typical patients in this subgroup are significantly older. The rule describes well the existing medical knowledge about CHD risk but its applicability is rather low because of the high estimated false positive rate.

4.2.1.2. Subgroup A2 for female patients (CHD ← body mass index over 25 kg m^{-2} and age over 63 years). This simple rule is very good for the female population. Its sensitivity is about 50% and false positive rate is below 10%. Supporting characteristics are positive family history, hypertension, increased LDL cholesterol values and normal but decreased HDL cholesterol values. Body mass index over 25 (the first principal risk factor) is exactly the generally excepted margin meaning overweight [32]. It is well known that obesity (high body mass index) strongly and positively correlates with the CHD rate. Typical values of the measured body mass index of CHD patients in this subgroup are significantly over the margin of 25. It is interesting to notice that the male population at the same age is significantly less sensitive to the overweight risk factor.

4.2.2. Interpretation of rules for stage B

At stage B, which includes anamnestic and physical examination results as well as basic laboratory tests, two rules were selected. The first rule has a high risk group detection potential: it includes total cholesterol as the only laboratory test result, and this risk factor can be easily and inexpensively measured. The second rule, describing a subgroup by a combination of two risk factors based on blood tests, demonstrates that also values close to the generally accepted normal values for these risk factors may be significant for early CHD risk detection and prevention.

4.2.2.1. Subgroup B1 (CHD ← total cholesterol over 6.1 mmol l^{-1} and age over 53 years and body mass index below 30 kg m^{-2}). This rule is characteristic for an older part of the population, especially for women (sensitivity about 30%, false positive rate about 10%, all estimated on the training set). For the male population, the sensitivity is about 25% and false positive rate 10%. Typical age of people in this subgroup is 65 years for women and 61 years for men. Further statistical analysis shows that, interestingly, typical patients in B1

do not have problems with overweight and hypertension. Moreover, very high body mass index is an important contraindication for this CHD patient group because the increased cholesterol value is mainly due to overweight. The only supporting risk factor is increased triglycerides value (typically about 2.5 mmol l^{-1}, reference values from 0.9 to 2.0 [32]) which is more often detected for men. The female patients in this subgroup have HDL cholesterol values about 1.3 mmol l^{-1} (reference values higher than 0.9) which is generally accepted as a normal value but which is significantly different from the typical values for the healthy part of the female population (2.2 mmol l^{-1}).

4.2.2.2. Subgroup B2 (CHD ← total cholesterol over 5.6 mmol l^{-1} and fibrinogen over 3.7 g l^{-1} and body mass index below 30 kg m^{-2}). This rule represents a group of CHD patient with similar properties for male and female patients. Its supporting risk factor is positive family history but one can also notice increased triglycerides values (typically about 2.5 mmol l^{-1}, reference values from 0.9 to 2.0), increased LDL cholesterol value (typically about 4.5 mmol l^{-1}, reference values less than 3.0), and borderline HDL cholesterol value (typically about 1 mmol l^{-1}, reference values higher than 0.9 [32]). Typical patients in B2 do not have problems with overweight, hypertension and cigarette smoking. Very high body mass index is a contraindication for this CHD patient group. Although the main subgroup properties are similar for both genders, a representative female in this subgroup is about 66 years old while a male is 10 years younger. The subgroup strongly correlates with subgroup B1 especially for women. The rule has an estimated sensitivity of about 30% and false positive rate about 15%. The discovered rule is important, demonstrating similar symptoms for male and female patients but at significantly different age.

4.2.3. Interpretation of rules for stage C

Stage C additionally includes test results of ECG at rest. At this stage there are many different acceptable rules and some of them have a relatively low false positive rate.

4.2.3.1. Subgroup C1 (CHD ← left ventricular hypertrophy). This rule covers male and female patients above the age of 55 years. Rule sensitivity is 25% and false positive rate about 5%. Left ventricular hypertrophy is a well-known risk factor which includes many other known CHD risk factors like hypertension and obesity. The supporting risk factor detected for this subgroup are positive family history, hypertension and diabetes mellitus. The practical importance of the discovered rule is that it has a relatively low error rate and that it does not correlate strongly with other rules. This means that its combination with any of the other rules can significantly increase the sensitivity of the CHD screening process.

4.2.4. Comparison of induced subgroups

By investigating all the five rules separately for men and women, some significant and interesting global differences among the subgroups can be observed; this may turn out to be important for the improved understanding of disease manifestations.

First well-known observation is that there are significant differences between male and female patients and that women are typically faced with CHD risk about 10–15 years later than men. From this perspective it is interesting to observe that subgroup C1 is the only one

which equally well describes patients of both genders without the difference in age. This rule is based on the heart patologic changes (left ventricular hypertrophy) and it particularly strongly correlates with decreased HDL cholesterol and increased fibrinogen values. In this rule other typical risk factors are also very similar for both genders.

Another important observation is that subgroups B2 and C1 are similar in the sense that they apply almost equally well to both genders but subgroup B2 has a significant delay of 10 years in favor of women. It seems that this effect is mainly due to the fibrinogen risk factor. When comparing rules induced separately for the two genders (A1 and A2) or which have different properties depending on the gender (B1), quite large differences in supporting risk factors can be observed: total cholesterol and overweight turn out to be characteristic for women, while positive family history and stress are particularly characteristic for men.

4.3. Deployment potential of induced subgroup descriptions

The induced subgroups can be used in the prevention process at an individual or at a population level.

- Individual level: If a person belongs to any of the induced risk groups, the person should come to a specialized medical institution for further medical testing.
- Population level: Induced risk group descriptions can be publicly announced with a call that people who recognize themselves as being at CHD risk should come for further tests to a specialized medical institution. This should be done with precaution because of the potential abundance of people, both those that really need help and those who did not understand the risk group descriptions. Alternatively, the risk group descriptions can be used to direct systematic prevention testing to subpopulations in which proportionally many CHD patients can be expected. For both purposes subgroups A1 and A2 seem especially appropriate.

Risk group descriptions can be, in some of their elements, considered also as new pieces of medical knowledge that can help medical experts in improving their decision making processes. For example, high total cholesterol is in accordance with medical knowledge a high risk factor for CHD and it is also known that high total cholesterol is characteristic for fat people. In this context it is interesting to notice that risk groups B1 and B2 describe people that have high total cholesterol and who are not very fat. In B1, this property is correlated with age and in B2 with high fibrinogen value. In medical literature and in normal medical practice, medical practitioners usually pay attention to individual risk factors, while combined risk factors are harder to observe, particularly when occurring for a non-typical population which is, in our case, relatively slim people with high total cholesterol value.

5. Risk group evaluation

The detected risk groups, described in Section 4, have estimated false positive rates between 5 and 27% on the training set. Therefore, the rules should not be considered as the ultimate prognostic rules to be used in decision making process. Instead, their primary function is in early detection of CHD from anamnestic data and routine laboratory tests of

risk factors and the definition of the risk population groups which should undertake non-invasive cardiovascular diagnostic procedures before the occurrence of first disease symptoms. But even for this purpose the induced rules can not be used unconditionally. The obtained classification quality may significantly depend on the properties of the tested population as well as on the data collection standards.

The evaluation tests were performed on two independent sets of people: the first one collected in the same medical institution but with a very different distribution of ill and healthy people and collected about 1 year after the training set, and the second one collected during medical prevention testing of employed people about 2 years after the training set collection.

5.1. Evaluation on an independent test set of patients from the same biased population

An independent set of 50 CHD patients and 20 people without CHD constitutes the first test set. The results show that the rules are successful in detecting CHD patients. About 90% of CHD patients were covered by at least one out of the five rules. The obtained results for CHD patients are summarized in Table 8, while Table 9 shows the summary TPr and FPr results for CHD and non-CHD patients, respectively. The measured sensitivity values on the test set for rules A1, B2, and C1 (85, 42 and 82%, respectively, presented in the test set TPr column of Table 8) are significantly higher than the values (47, 32 and 23%, respectively) computed on the set of patients used for subgroup discovery. For rules A2 and B1 the values are 41 and 36%, which is comparable to estimated values 48 and 29% on the test set.

The last column of Table 8 shows the percentages of satisfied supporting factors for subgroups A1, A2, B1, B2 and C1 on the test set (supporting risk factors for each rule are listed in Table 7). For subgroups A1, A2, and C1 which have more then one supporting factor the last column of Table 8 lists the mean values. The values between 60 and 93% demonstrate the relevance of selected supporting factors. Most of these factors have a higher rate for the specific subgroup than for the whole CHD population. For example, subgroup B1 has one supporting factor which is increased trygliceride value. Trygliceride value is known as an important risk factor and in our test group about 60% of all CHD patients have this value above 2.0 mmol l^{-1}. But patients described by subgroup B1 have

Table 8
Summary of results on an independent test set of 50 CHD patients from the institute

Subgroup	Sensitivity (TPr)		Supporting factors	
	Training set (%)	Test set (%)	Training set (no. of factors)	Test set (percentage satisfied)
A1	47	85	4	60
A2	48	41	4	79
B1	29	36	1	81
B2	32	42	1	93
C1	23	82	3	76

Table 9
Summary of sensitivity (TPr) and false positive rate (FPr) results for rules A1, A2, B1, B2 and C1 measured on the training and the test datasets from the specialized medical institution, and on an independent employee dataset

Subgroup	Training set (Pos = 111, Neg = 127)		Test set (Pos = 50, Neg = 20)		Employee set (Pos = 30, Neg = 170)	
	TPr (%)	FPr (%)	TPr (%)	FPr (%)	TPr (%)	FPr (%)
A1	47	27	85	78	95	28
A2	48	7	41	27	60	6
B1	29	9	36	20	37	5
B2	32	13	42	15	83	48
C1	23	5	82	40	27	8
Expert	–	–	–	–	97	18

In the last dataset, the positive (suspected CHD) cases are individuals which have ST segment depression of 1 mm or higher either during exercise or long-term ECG measurements.

increased trygliceride value above this limit in more than 80% of cases. A similar effect can be observed with positive family history for subgroup B2 and with HDL cholesterol values below 1.0 mmol l^{-1} for subgroup A2.

5.2. Evaluation on an independent validation set gathered from the general population

An independent set of 200 people was used to test the practical applicability of the induced rules for CHD risk group detection. The majority of this set are employees of two large Croatian companies. These companies paid for medical prevention testing of their employees, hence the available dataset can be assumed to represent a part of the general population. We call this set a small epidemiological study set although it is too small to give reliable epidemiological results, and includes only a part of the interesting population for CHD screening. For instance, unemployed or retired people as well as children have not been included in this set.

In this validation set there are 123 males and 77 females (age between 18 and 65 years, median value 46 years). By prevention tests, 30 (15%) suspect CHD patients were detected, but the final diagnoses for these people are still unknown (this is however not very relevant for this study because detection of CHD risk groups is our main goal, not the actual diagnosis). In this study, suspect CHD cases are defined as those individuals with ST segment depression of 1 mm or higher either during ECG exercise test or during 24 h ECG (Holter). In the male population, there are 20 (16%) suspect CHD cases and 10 (13%) cases in the female population.

Table 9 shows the sensitivity (true positive rate TPr) and the false alarm (false positive rate FPr) for the five subgroups. The column 'employee set' shows the results obtained on the independent set of 200 individuals, while the columns 'training set' and 'test set' summarize values measured on the training and the test set which were already analysed in Sections 4 and 5.1, respectively. Analysis of the 'employee set' column of Table 9 shows that rules A2, B1, and C1 have low false positive rates in contrast to rules A1 and especially B2. The comparison

of the results on the training set and the employee set shows good agreement of estimated and measured results, except for rule B2. The good news is that for most rules the measured results are better (higher sensitivity and lower false positive rates) than the corresponding estimated values. In this sense the measured results are very promising.

The last line of the table shows the results of domain expert classification for the employee data. Expert results are better than any of the results obtained by the induced rules, but such differences were expected. The experiment was designed favorably for the domain expert who had all three data levels available at the time of his classification. Moreover, the expert is not a general practitioner but the cardiologist working in the specialized medical institution. Nevertheless, the false positive rates of 5–8% of rules A2, B1, and C1 are significantly better compared to the 18% false positive rate achieved in expert classification.

The unexpectedly high false positive rate of rule B2 attracted our special attention. The high sensitivity of the rule seems to be the consequence of the fact that the rule covers more than 50% of the whole tested population. It is interesting to notice that on the training set the same rule covers only about 20% of the examples. This result demonstrates significant differences between the general population (employee dataset) and patients of the Cardiovascular Institute (training set). This means that we have to be very careful when estimating properties of rules induced from very biased datasets and that only relevant epidemiological studies may prove the usefulness of induced rules.

The dissatisfying results for rule B2 stimulated further result analysis, which showed that fibrinogen measurements on the employee dataset have caused the problem. Value of 3.7 g l^{-1} is typically accepted as normal upper limit for fibrinogen and for the employed people this or higher value has been measured for 60% of the validation set. The problem has been detected in the laboratory measurement procedure which has changed between collecting data for the training and the employee datasets. According to the suggestions of the International Federation of Clinical Chemistry (IFCC), temperature of laboratory testing has changed from 30 to 37 °C. Malfunction of rule B2 was the first evidence in the institute that the increase of temperature in laboratory measurements can have significant influence on the data values.

6. Related work

This section gives some links to related work in active mining, decision tree and rule learning, and subgroup discovery.

6.1. Active mining

The need of expert involvement in the knowledge discovery process has been recognized long ago, especially emphasizing the need of expert involvement in knowledge acquisition for expert systems [11,12]. For instance, in the knowledge acquisition research community, lots of attention was devoted to the development of ripple down rules (RDR) which allow expert-guided incremental rule learning by including exceptions to the current rule set [8,9].

Recent developments in data mining and knowledge discovery in databases also show awareness of the need for expert's support of the knowledge discovery process [16]. The need for user interactivity in subgroup discovery has been addressed in [42], describing a system developed in the KESO project (Knowledge Extraction for Statistical Offices, http://orgwis.gmd.de/projects/KESO/). More generally, the CRISP-DM methodology (Cross-Industry Standard Process for Data Mining, http://www.crisp-dm.org) [4] also emphasizes the need of a feedback loop at every stage of the knowledge discovery process.

The active mining approach to knowledge discovery has further emphasized user-centered mining and user interaction/reaction, to address the need for actively collecting relevant data sources, mining useful knowledge from different forms of data sources and promptly reacting to situation change. The term *active mining* [34,35], propagated by a large Japanese data mining project (2001–2005), represents a collection of activities each solving a part of this need, but collectively achieving the mining objective through the spiral effect of several interleaving data mining steps.

6.2. Related decision tree and rule learning approaches

In symbolic predictive induction, the two most common approaches are decision tree learning [1,36] and rule learning [33]. At a first glance it may seem that standard decision tree and rule learning algorithms can be used for the task of subgroup discovery and risk group detection. In this section we give arguments why the presented approach to subgroup discovery is advantageous to standard decision tree and rule learning as a solution to this task.

Decision tree learning algorithms like the ones implemented in CART [1], ID3 [36], its followers many others are inappropriate for the subgroup discovery task. The reason is that the rules which can be formed from paths leading from the root node to class labels in the leaves represent *discriminating descriptions*, formed from properties that best discriminate between the classes, and not from descriptions characterizing the individuals in the subgroup.

The goal of rule learning, on the other hand, is to generate models, one for each class, inducing class characteristics in terms of properties occurring in the descriptions of training examples. Classification rule learning results in *characteristic descriptions*, generated separately for each class by repeatedly applying the covering algorithm. Classical rule learning algorithms [33,5,6], as well as more sophisticated rule learners like RL [30], RIPPER [6], SLIPPER [7] have been designed to construct classification and prediction rules.

As opposed to model construction, subgroup discovery aims at discovering individual 'patterns' of interest, representing population subgroups. Standard classification rule learning algorithms can not appropriately address the task of subgroup discovery for two main reasons: first, they use inappropriate search heuristics optimizing rule accuracy, and second, they use the covering algorithm for rule set construction.

- Various rule evaluation measures and heuristics have been studied for subgroup discovery [25,43], aimed at balancing the size of a group (referred to as factor g) with its distributional unusualness (referred to as factor p). The properties of functions that combine these two factors have been extensively studied (the so-called 'p-g-space', [25]). One of the heuristics of this kind is the weighted relative accuracy heuristic, defined as $\mathrm{WRAcc}(\mathrm{Class} \leftarrow \mathrm{Cond}) = p(\mathrm{Cond})(p(\mathrm{Class}|\mathrm{Cond}) - p(\mathrm{Class}))$ and used

in [39,28], which also trades off the generality of a rule (p(Cond), i.e. rule coverage) and its relative accuracy (p(Class|Cond) − p(Class)). Besides these 'objective' measures of interestingness, some 'subjective' measure of interestingness of a discovered pattern can be taken into account, such as actionability ('a pattern is interesting if the user can do something with it to his or her advantage') and unexpectedness ('a pattern is interesting to the user if it is surprising to the user') [38].

- The main deficiency of the covering algorithm is that only the first few induced rules may be of interest as subgroup descriptors with sufficient coverage and significance. Subsequently induced rules are induced from biased example subsets, i.e. subsets including only positive examples not covered by previously induced rules, which inappropriately biases the subgroup discovery process. As a remedy to this problem we have proposed to use the weighted covering algorithm (Algorithm DMS), in which the subsequently induced rules allow for discovering interesting subgroup properties of the entire population. The weighted covering algorithm modifies the classical covering algorithm in such a way that covered positive examples are not deleted from the current training set. Instead, in each run of the covering loop, the algorithm stores with each example a count that shows how many times (with how many rules) the example has been covered so far. This allows the algorithm to discover less biased rules discovering interesting subgroup properties of the entire population, still covering different population segments. Instance weights play an important role also in boosting [17] and alternating decision trees [37]. Instance weights have been used also in variants of the covering algorithm implemented in rule learning approaches such as SLIPPER [7], RL [30] and DAIRY [23].

6.3. Related subgroup discovery approaches

Two most important systems in the field of subgroup discovery are, however, EXPLORA [25] and MIDOS [43,44]. The first system treats the learning task as a single relation problem, i.e. all the data are assumed to be available in one table (relation), while the second one extends this task to multi-relation databases, which is related to a number of other learning tasks [13,31,41], mostly in the field of Inductive Logic Programming [14,27]. The most important features of EXPLORA and MIDOS, related to this paper, concern the use of heuristics for subgroup discovery, briefly outlined in Section 6.2 above.

Note that some approaches to association rule induction can also be used for subgroup discovery. For instance, the APRIORI-C algorithm [24], adapting association rule induction to classification rule induction, outputs individual classification rules with guaranteed support and confidence with respect to a target class. If a rule satisfies also a user-defined significance threshold, an induced APRIORI-C rule is an independent 'chunk' of knowledge about the target class, which can be viewed as a subgroup description with guaranteed significance, support and confidence. Similarly, the confirmation rule concept, introduced in [19] and used as a basis for the subgroup discovery algorithm in this paper, utilizes the minimal support requirement as a measure which must be satisfied by every rule in order to be included in the induced confirmation rule set.

Recent approaches to subgroup discovery aim at overcoming the problem of the inappropriate bias of the standard covering algorithm discussed in Section 6.2. The recently developed subgroup discovery algorithms CN2-SD [28] and RSD [29] use the

so-called weighted covering algorithm, similar to the one implemented in Algorithm DMS described in this paper.

7. Conclusions

The paper shows that active subgroup mining through expert-guided subgroup discovery may lead to interesting results even in cases when relatively small and very biased datasets are available. The induced descriptions of coronary heart disease risk groups illustrate what results can be obtained by the novel subgroup mining methodology. Despite the biased data available, the presented risk group descriptions seem to be important both as chunks of novel medical knowledge about CHD as well as decision rules which can be used to help decision making at a level of one person (the person should be invited to a specialized medical institution for further medical testing) or at a global level to direct systematic CHD prevention.

There are indications that the active subgroup discovery methodology, applied in this work to CHD risk group detection, will be interesting also for other medical (and non-medical) data analysis and knowledge discovery applications. The proposed approach is based on a combination of machine learning subgroup detection and statistical risk factor analysis. For both steps the expert knowledge and experience are the main guiding factors. Putting the medical expert in the center of the knowledge discovery process is one of the main emphasis of this work. In this context publicly available tools on the internet are very important: in the future we plan to add additional options and functionality to the Data Mining Server

Acknowledgements

This work has been supported in part by the Croatian Ministry of Science and Technology, Slovenian Ministry of Education, Science and Sport, and the EU funded project Data Mining and Decision Support for Business Competitiveness: A European Virtual Enterprize (IST-1999-11495). We are grateful to Tomislav Šmuc for his work on the implementation of the Data Mining Server.

Appendix A. Rule construction by Algorithm SD

Algorithm SD takes as its input the complete training set E and the feature set L, where features $l \in L$ are logical conditions constructed from attribute values describing the examples in E (see Fig. 4). The main algorithm parameter is g. There are two additional parameters which are typically not adjusted by the user: min_support and beam_width. The output of the algorithm is set S of beam_width different rules with highest q values. The rules have the form of conjunctions of features from L.

The algorithm initializes all the rules in Beam and New_beam by empty rule conditions. Their quality values $q(i)$ are set to zero (step 1). Rule initialization is followed by an infinite

Algorithm SD: Subgroup Discovery
Input: $E = P \cup N$ (E training set, $|E|$ training set size,
P positive (target class) examples, N negative (non-target class) examples
L set of all defined features (attribute values), $l \in L$
Parameter: g (generalization parameter, $0.1 < g$, default value 1)
min_support (minimal support for rule acceptance)
beam_width (maximal number of rules in *Beam* and *New_Beam*)
Output: $S = \{TargetClass$ IF $Cond\}$ (set of rules formed of *beam_width*
best conditions *Cond*)

```
(1)   for all rules in Beam and New_Beam (i = 1 to beam_width) do
        initialize condition part of the rule to be empty, Cond(i) ← {}
        initialize rule quality, q(i) ← 0
(2)   while there are improvements in Beam do
(3)     for all rules in Beam (i = 1 to beam_width) do
(4)       for all l ∈ L do
(5)         form a new rule by forming a new condition as a conjunction of the
              condition from Beam and feature l, Cond(i) ← Cond(i) AND l
(6)         compute the quality of a new rule as q = TP/(FP+g)
(7)         if TP/|E| ≥ min_support and if q is larger than any q in New_Beam
                and if the new rule is relevant do
(8)           replace the worst rule in New_Beam with the new rule and
                reorder the rules in New_Beam with respect to their quality
(9)       end for features
(10)    end for rules from Beam
(11)    Beam ← New_Beam
(12)  end while
```

Fig. 4. Heuristic beam search rule construction algorithm for subgroup discovery.

loop (steps 2–12) that stops when, for all rules in the beam, it is no longer possible to further improve their quality. Rules can be improved only by conjunctively adding features from *L*. After the first iteration, a rule condition consists of a single feature, after the second iteration up to two features, and so forth. The search is systematic in the sense that for all rules in the beam (step 3) all features from *L* (step 4) are tested in each iteration. For every new rule, constructed by conjunctively adding a feature to rule body (step 5) quality q is computed (step 6). If the support of the new rule is greater than min_support and if its quality q is greater than the quality of any rule in New_beam, the worst rule in New_beam is replaced by the new rule. The rules are reordered in New_beam according to their quality q. At the end of each iteration, New_beam is copied into Beam (step 11). When the algorithm terminates, the first rule in Beam is the rule with maximum q.

A necessary condition (in step 7) for a rule to be included in New_beam is that it must be *relevant*. The new rule is irrelevant if there exists a rule *R* in New_beam such that true positives of the new rule are a subset of true positives of *R* and false positives of the new rule are a superset of false positives of *R*. After the new rule is included in New_beam it may happen that some of the existing rules in New_beam become irrelevant with respect to this new rule. Such rules are eliminated from New_beam during its reordering (in step 8). Relevance testing ensures that New_beam contains only different and relevant rules.

Algorithm DMS: Rule set construction
Input: $E = P \cup N$ (E training set, $|E|$ training set size,
P positive (target class) examples,
N negative (non-target class) examples)
L set of all defined features (attribute values), $l \in L$
Parameter: *number* (required number of selected rules in output set SS)
g (generalization parameter, $0.1 < g < 100$, default value 1)
min_support (minimal support for rule acceptance)
beam_width (number of rules in the beam)
Output: SS set of relatively independent rules for the target class
(1) **initialize** $SS \leftarrow \{\}$ (empty set of selected rules)
(2) **for** every $e \in P$ **do** $c(e) \leftarrow 1$
(3) **repeat** *number* times
(4) call **Algorithm SD** to construct a rule with maximal quality $q = \frac{\sum_{TP} \frac{1}{c(e)}}{FP+g}$
(5) **for** every $e \in P'$ covered by the constructed rule
do $c(e) \leftarrow c(e) + 1$
(6) **add** the constructed rule into set SS
(7) **end repeat**

Fig. 5. Weighted covering algorithm for iterative rule set construction.

Appendix B. Rule set construction by Algorithm DMS

Algorithm DMS iteratively calls Algorithm SD and selects from its beam the single best rule to be included into the output set *SS* (see Fig. 5). Parameter *number* determines the total number of induced rules. For every example $e \in P$ there is a counter $c(e)$. Initially, the output set of selected rules is empty (step 1) and all counter values are set to 1 (step 2). After rule selection $c(e)$ values for all target class examples covered by the selected rule are incremented by 1 (step 5). Weights of true positive examples used in the quality measure q are not constant and equal to 1 but defined by expression $1/c(e)$, changing from iteration to iteration (step 4). The main reason for the described implementation is to ensure the diversity of induced subgroups even though, because of the short execution time limit on the publicly available server, a low beam_width parameter value in Algorithm SD had to be set (the default value is 20).

References

[1] Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees. Boca Raton: CRC Press; 1984.
[2] Casteli WP. Epidemiology of coronary heart disease: the Framingham study. Am J Med 1984;76(2A): 4–12.
[3] Casteli WP. The triglycerides issue: a view from Framingham. Am Heart J 1986;112:432–7.
[4] Chapman P, et al. CRISP-DM 1.0 Step-by-step data mining guide, http://www.crisp-dm.org.
[5] Clark P, Niblett T. The CN2 induction algorithm. Mach Learn 1989;3(4):261–83.
[6] Cohen WW. Fast effective rule induction. In: Proceedings of the 12th International Conference on Machine Learning. Los Altos (CA): Morgan Kaufmann; 1995. p. 115–23.

[7] Cohen WW, Singer Y. A simple, fast, and effective rule learner. In: Proceedings of Annual Conference of American Association for Artificial Intelligence. American Association for Artificial Intelligence; 1995. p. 335–42.
[8] Compton P, Jansen R. Knowledge in context: a strategy for expert system maintenance. In: Proceedings of 2nd Australian Joint Artificial Intelligence Conference. Berlin: Springer LNAI 406; 1988. p. 292–306.
[9] Compton P, Horn R, Quinlan R, Lazarus L. Maintaining an expert system. In: Quinlan R, editor. Applications of expert systems. Reading: Addison-Wesley; 1989. p. 366–85.
[10] B. Dahlof, Pennert K, Hansson L. Reversal of left ventricular hypertrophy in hypertensive patients: a metaanalysis of 109 treatment studies. Am J Hypertens 1992;5:95–110.
[11] Davis R. Knowledge acquisition in rule-based systems: knowledge representation as a basis for system construction and maintainance. In: Shine ME, Coombs MJ, editors. Designing for human–computer communication. New York: Academic Press; 1978. p. 87–137.
[12] Davis R. TEIRESIAS: experiments in communicating with a knowledge-based systems. In: Waterman DA, Hayes-Roth F, editors. Pattern directed inference systems. New York: Academic Press; 1983. p. 99–134.
[13] De Raedt L, Dehaspe L. Clausal discovery. Mach Learn 1997;26:99–146.
[14] Džeroski S, Lavrač N, editors. Relational data mining. Berlin: Springer; 2001.
[15] Fayyad UM, Irani KB. On the handling of continuous-valued attributes in decision tree generation. Mach Learn 1992;8:87–102.
[16] Fayyad UM, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery: an overview. In: Fayyad UM, Piatetski-Shapiro G, Smyth P, Uthurusamy R, editors. Advances in knowledge discovery and data mining. AAAI Press; 1996. p. 1–34.
[17] Freund Y, Shapire RE. Experiments with a new boosting algorithm. In: Proceedings of Thirteenth International Machine Learning Conference ICML'96. Los Altos (CA): Morgan Kaufmann; 1996. p. 148–56.
[18] Gamberger D, Lavrač N, Grošelj C. Experiments with noise filtering in a medical domain. In: Proceedings of International Conference on Machine Learning ICML-99. Los Altos (CA): Morgan Kaufmann; 1999. p. 143–51.
[19] Gamberger D, Lavrač N. Confirmation rule sets. In: Proceedings of 4th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD2000). Berlin: Springer; 2000. p. 34–43.
[20] Gamberger D, Lavrač N, Krstačić G, Šmuc T. Inconsistency tests for patient records in a coronary heart disease database. In: Proceedings of International Symposium on Medical Data Analysis (ISMDA 2000). Berlin: Springer; 2000. p. 183–9.
[21] Gamberger D, Lavrač N. Expert-guided subgroup discovery: methodology and application. J Artif Intell Res 2002;17:501–27.
[22] Goldman L, Garber AM, Grover SA, Hlatky MA. Cost-effectiveness of assessments and management of risk factors. J Am Coll Cardiol 1996;27:1020–30.
[23] Hsu D, Soderland S, Etzioni O. A redundant covering algorithm applied to text classification. In: Proceedings of the AAAI Workshop on Learning from Text Categorization. AAAI Press; 1998.
[24] Jovanoski V, Lavrač N. Classification rule learning with APRIORI-C. In: Progress in Artificial Intelligence: Proceedings of the Tenth Portuguese Conference on Artificial Intelligence. Berlin: Springer; 2001. p. 44–51.
[25] Klösgen W. Explora: a multipattern and multistrategy discovery assistant. In: Fayyad UM, Piatetski-Shapiro G, Smyth P, Uthurusamy R, editors. Advances in knowledge discovery and data mining. Cambridge: MIT Press; 1996. p. 249–71.
[26] Kononenko I. Inductive and Bayesian learning in medical diagnosis. Appl Artif Intell 1993;7:317–37.
[27] Lavrač N, Džeroski S. Inductive logic programming: techniques and applications. Chichester: Ellis Horwood; 1994.
[28] Lavrač N, Flach P, Kavšek B, Todorovski L. Adapting classification rule induction to subgroup discovery. In: Proceedings of the IEEE International Conference on Data Mining. IEEE Computer Society; 2002. p. 266–73.
[29] Lavrač N, Železný F, Flach P. RSD: relational subgroup discovery through first-order feature construction. In: Proceedings of the Twelfth International Conferences on Inductive Logic Programming. Berlin: Springer LNAI 2583; 2003. p. 152–69.

[30] Lee Y, Buchanan BG, Aronis JM. Knowledge-based learning in exploratory science: learning rules to predict rodent carcinogenicity. Mach Learn 1998;30:217–40.
[31] Mannila H, Toivonen H. On an algorithm for finding all interesting sentences. In: Trappl R, editor. Proceedings of the Cybernetics and Systems'96, 1996. p. 973–8.
[32] Maron D, Ridker PM, Pearson AT. Risk factors and the prevention of coronary heart disease. In: Wayne AR, Schlant RC, Fuster V, editors. HURST'S: the heart. New York: McGraw-Hill; 1998. p. 1175–95.
[33] Michalski RS, Mozetič I, Hong J, Lavrač N. The multi-purpose incremental learning system AQ15 and its testing application on three medical domains. In: Proceedings of the Fifth National Conference on Artificial Intelligence. Los Altos (CA): Morgan Kaufmann; 1986. p. 1041–5.
[34] Motoda H, editor. Active mining: new directions of data mining. In: Frontiers in artificial intelligence and applications, vol. 79. IOS Press; 2002.
[35] Motoda H. Active mining: a spiral model for knowledge discovery. In: Proceedings of the Invited talk at IEEE International Conference on Data Mining, ICDM-2002, Maebashi, Japan, 2002.
[36] Quinlan JR. Induction of decision trees. Mach Learn 1986;1:81–106.
[37] Schapire RE, Singer Y. Improved boosting algorithms using confidence-rated predictions. In: Proceedings of the Eleventh Conference on Computational Learning Theory. ACM Press; 1998. p. 80–91.
[38] Silberschatz A, Tuzhilin A. On subjective measures of interestingness in knowledge discovery. In: Proceedings of the First International Conference on Knowledge Discovery and Data Mining. AAAI Press; 1995. p. 275–81.
[39] Todorovski L, Flach P, Lavrač N. Predictive performance of weighted relative accuracy. In: Zighed D, Komorowski J, Zytkow J, editors. Proceedings of the Fourth European Conference on Principles of Data Mining and Knowledge Discovery. Berlin: Springer; 2000. p. 255–64.
[40] Wilson PWF. Established risk factors and coronary artery disease: the Framingham Study. Am J Hypertension 1984;7:7S–12S.
[41] Wrobel S, Džeroski S. The ILP description learning problem: towards a general model-level definition of data mining in ILP. In: Morik K, Herrmann J, editors. Proceedings of the Fachgruppentreffen Maschinelles Lernen. University of Dortmund; 1995.
[42] Wrobel S, Wettschereck D, Verkamo AI, Siebes A, Mannila H, Kwakkel F, Klösgen W. User interactivity in very large scale data mining. In: Proceedings of the German Workshop on Machine Learning. Technical Report. University of Chemnitz; 1996. p. 125–30.
[43] Wrobel S. An algorithm for multi-relational discovery of subgroups. In: Proceedings of the First European Symposium on Principles of Data Mining and Knowledge Discovery. Berlin: Springer; 1997. p. 78–87.
[44] Wrobel S. Inductive Logic Programming for Knowledge Discovery in Databases. In: Džeroski S, Lavrač N, editors. Relational Data Mining. Berlin: Springer; 2001. p. 74–101.

Further Reading

[45] Agrawal R, Mannila H, Srikant R, Toivonen H, Verkamo AI. Fast discovery of association rules. In: Fayyad UM, Piatetski-Shapiro G, Smyth P, Uthurusamy R, editors. Advances in knowledge discovery and data mining. AAAI Press; 1996. p. 307–28.
[46] Ali KM, Pazzani MJ. Error reduction through learning multiple descriptions. Mach Learn 1996;24: 173–206.
[47] Boland PJ. Majority systems and the Concorcet jury theorem. Statistician 1989;38:181–9.
[48] Breiman L. Bagging predictors. Mach Learn 1996;24(2):123–40.
[49] Clark P, Boswell R. Rule induction with CN2: some recent improvements. In: Kodratoff Y, editor. Proceedings of the 5th European Working Session on Learning. Berlin: Springer; 1991. p. 151–63.
[50] Kagan T, Ghosh J. Error correlation and error reduction in ensemble classifiers. Connect Sci 1996;8: 385–404.
[51] Kittler J, Hatef M, Duin RPW, Matas J. On combining classifiers. IEEE Trans Pattern Anal Mach Intell 1998;20:226–39.
[52] Kukar M, Kononenko I, Grošselj C, Kralj K, Fettich JJ. Analysing and improving the diagnosis of ischaemic heart disease with machine learning (special issue on Data Mining Techniques and Applications in Medicine). Artif Intell Medicine 1999;16:25–50.

[53] Lavrač N, Gamberger D, Turney P. A relevancy filter for constructive induction. IEEE Intell Syst Their Appl 1997;13:50–6.
[54] Lavrač N, Zupan B, Keravnou E, editors. Intelligent data analysis in medicine and pharmacology. Dordrecht: Kluwer Academic Publishers; 1997.
[55] Lavrač N. Selected techniques for data mining in medicine (special issue on Data Mining Techniques and Applications in Medicine). Artif Intell Medicine 1999;16:3–23.
[56] Pazzani M, Murphy P, Ali K, Schulenburg D. Trading off coverage for accuracy in forecasts: applications to clinical data analysis. In: Proceedings of the AAAI Symposium on AI in Medicine, 1994. p. 106–10.
[57] Provost F, Fawcett T. Robust classification for imprecise environments. Mach Learn 2001;42(3):203–31.
[58] Quinlan JR. Boosting, bagging, and C4.5. In: Proceedings of the Thirteenth National Conference on Artificial Intelligence. AAAI Press; 1996. p. 725–30.
[59] Rivest RL, Sloan R. Learning complicated concepts reliably and usefully. In: Proceedings of the Workshop on Computational Learning Theory. Los Altos (CA): Morgan Kaufman; 1988. p. 69–79.