# Descriptive Induction through Subgroup Discovery: A Case Study in a Medical Domain

**Dragan Gamberger** DRAGAN.GAMBERGER@IRB.HR
Rudjer Bošković Institute, Bijenička 54, 10000 Zagreb, Croatia

**Nada Lavrač** NADA.LAVRAC@IJS.SI
J. Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia

## Abstract

This paper presents an approach to descriptive induction based on a combination of subgroup discovery, statistical characterization of induced subgroups, and their visualization. Subgroup discovery is performed using a novel heuristic confirmation rule induction algorithm. Induced subgroup descriptions are enriched by adding statistically significant properties of detected subgroups. Subgroup visualization can be used to present the distribution of detected subgroups in terms of selected properties. The approach is illustrated by the results obtained for the problem of early detection of patient groups with risk for atherosclerotic coronary heart disease.

## 1. Introduction

Classical rule learning algorithms were designed to construct classification and prediction rules (Clark and Niblett, 1989; Michalski et al., 1986). In addition to these *predictive induction* approaches, developments in *descriptive induction* have recently gained much attention: mining of association rules (e.g., association rule learning algorithm APRIORI (Agrawal et al., 1996), subgroup discovery (e.g., subgroup discovery algorithm MIDOS (Wrobel, 1997)), and other non-classificatory induction approaches.

The methodology presented in this paper can be applied to subgroup discovery. As in the MIDOS approach, a subgroup discovery task can be 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 statistical (distributional) characteristics with respect to the property of interest.

Notice that some approaches to association rule induction can be used for subgroup discovery. For instance, the APRIORI-C algorithm (Jovanoski and Lavrač, 2001), adapting association rule induction to classification rule induction, outputs classification rules with guaranteed high support and confidence. As such, each APRIORI-C rule represents a 'chunk' of knowledge about the problem, which is very important for knowledge discovery. Similarly, the confirmation rule concept, introduced by Gamberger and Lavrač (2000) 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.

Both above mentioned approaches to subgroup discovery exploit the information about class membership. One of the main reasons why these approaches are of interest for subgroup discovery is that, unlike the classical classification rule induction algorithms such as CN2 (Clark and Niblett, 1989) and AQ (Michalski et al., 1986), they do not use the covering algorithm. In covering algorithms only the first few induced rules may be of interest as subgroup descriptors with sufficient coverage. Subsequently induced rules are induced from biased example subsets, e.g., subsets including only positive examples not covered by previously induced rules. This bias constrains the population for subgroup discovery in a way that is unnatural for the subgroup discovery process which is, in general, aimed at discovering interesting properties of subgroups of the entire population.

In this paper, a new heuristic confirmation rule learning algorithm is proposed for subgroup discovery, and incorporated in the confirmation rule based decision making concept. One of the basic characteristics of this concept is that separate rule sets are built for every target class. As such, it is similar to human reasoning

processes. In a broader sense, the confirmation rule set concept introduces a decision model in which different rules can be incorporated: either rules induced by (one or more) inductive learning algorithms or even human encoded expert rules. If used for prediction, high predictive accuracy of such a decision model is expected only if, besides the high predictive accuracy of each individual rule, the whole set is as diverse as possible. This decision concept follows the paradigm of *reliable, probably almost always useful learning*, defined by Rivest and Sloan (1988). As opposed to confirmation rules which cover only target class examples, the property of the heuristic confirmation rule learning algorithm, proposed in this paper, is that it enables the construction of rules that cover also a limited number of examples of the non-target class. This algorithm is used as a basis of the proposed algorithm for subgroup discovery.

This paper presents an approach to descriptive induction based on a combination of subgroup discovery, statistical characterization of induced subgroups, and their visualization. The proposed algorithms used in the subgroup discovery process for subgroup detection (see Figure 1) and selection (see Figure 2) are implemented in the on-line Data Mining Server, available at http://dms.irb.hr.

The problem of population screening for early detection of atherosclerotic coronary heart disease (CHD) risk groups is used to illustrate the results obtained by applying the suggested subgroup discovery methodology. The problem of early CHD detection is very important because CHD is one of the world's most frequent causes of mortality and a common problem in medical practice. The problem is known as a difficult one. Clinical studies have revealed plausible biological links between many risk factors and atherosclerosis (Goldman et al., 1996). In addition, it was detected that coexistence of risk factors increases the disease rate. 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.

The paper organization is as follows. Subgroup discovery using heuristic confirmation rule induction is described in Section 2. Enriched subgroup description achieved by adding statistically significant properties of detected subgroups is presented in Section 3. The final step of the suggested approach to descriptive induction is model visualization, described in Section 4. Section 5 describes the available CHD data set, the resulting models formulated in collaboration with a medical expert, and visualized models showing the distributions of detected risk groups.

## 2. Subgroup Discovery

### 2.1 A Heuristic Confirmation Rule Construction Algorithm for Subgroup Discovery

The basic property of confirmation rules introduced by Gamberger and Lavrač (2000) is that they should cover (satisfy) only examples of a given target class. For subgroup detection, this requirement is not appropriate. Therefore we have developed a novel, heuristic beam search algorithm that allows for the induction of more general rules which may cover also a small number of non-target class examples. This concept, upgrading the confirmation rule induction framework to subgroup discovery, is implemented in **Algorithm SD** outlined in Figure 1.

The aim of the heuristic confirmation rule learning algorithm is the search for rules with a maximal $q$ value, where $q$ is defined as $q = TP/(FP + g)$. In the definition of $q$, $TP$ are true positives (the number of CHD patients correctly classified 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 parameter $g$.

Variations of parameter $g$ enable the expert to guide subgroup discovery in the $TP/FP$ space, in which $FP$ (plotted on the $X$-axis) needs to be minimized, and $TP$ (plotted on the $Y$-axis) needs to be maximized. The $TP/FP$ space is similar to the ROC (Receiver Operating Characteristic) space (see e.g., Kukar et al. (1998) and Provost and Fawcett (2001)) in which a point in the ROC space shows classifier performance in terms of false alarm or *false positive rate* $FPr = \frac{FP}{TN+FP}$ (plotted on the $X$-axis) that needs to be minimized, and sensitivity[1] or *true positive rate*

[1] *Sensitivity* measures the fraction of positive cases that are classified as positive, whereas *specificity* measures the fraction of negative cases classified as negative. If $TP$ denotes true positives, $TN$ true negatives, $FP$ false positives, $FN$ false negatives, $Pos$ all positives, and $Neg$ all neg-

**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, $0.1 < g < 100$, default value 10)
*min_support* (minimal support for rule acceptance)
*beam_depth* (number of rules in the beam)
**Output:** $S = \{TargetClass \leftarrow Cond\}$ (set of rules formed of *beam_depth* best conditions $Cond$)
(1) **for** all rules in the beam ($i = 1$ to *beam_depth*) **do**
initialize condition part of the rule to be empty, $Cond(i) \leftarrow \{\}$
initialize rule quality, $q(i) \leftarrow 0$
(2) **while** there are improvements in the beam **do**
(3) **for** all rules in the beam ($i = 1$ to *beam_depth*) **do**
(4) **for** all $l \in L$ **do**
(5) form a new rule by forming a new condition as a conjunction of the condition from the beam and feature $l$, $Cond(i) \leftarrow Cond(i) \wedge l$
(6) compute the quality of a new rule as $q = \frac{TP}{FP+g}$
(7) **if** $\frac{TP}{|E|} \geq min\_support$ **and** $q$ is larger than any $q(i)$ in the beam **do**
(8) replace the worst rule in the beam with the new rule and reorder the rules with respect to their quality
(9) **end for** features
(10) **end for** rules from the beam
(11) **end while**

*Figure 1.* Heuristic beam search confirmation rule construction algorithm for subgroup discovery.

$TPr = \frac{TP}{TP+FN}$ (plotted on the $Y$-axis) that needs to be maximized. In the ROC space, an appropriate tradeoff, determined by the expert, can be achieved by applying different algorithms, or by different parameter settings of a selected data mining algorithm.

In Algorithm SD, increased generality (increasing $g$ means moving to the right in the $TP/FP$ space) 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 are omitted due to space restrictions.

atives, then $Sensitivity = TPr = \frac{TP}{TP+FN} = \frac{TP}{Pos}$, and $Specificity = \frac{TN}{TN+FP} = \frac{TN}{Neg}$, and $FalseAlarm = FPr = 1 - Specificity = \frac{FP}{TN+FP} = \frac{FP}{Neg}$.

## 2.2 Rule Subset Selection

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. Inclusion of all these rules into the rule set is generally not desired because (a) it is difficult to make decisions based on a large sets of rules, and (b) experiments demonstrated that there are subsets of very similar rules which use almost the same attribute values and have similar prediction properties. The same effect occurs also when induced rules are used as subgroup descriptions induced in a descriptive induction process. A solution to this problem is to reduce generated rule sets so that they include only a relatively small number of rules which are as diverse as possible.

Selecting a subset of independent classifiers for the same target class is known as a complex task which occurs in most multiclassifier decision systems (Kagan and Ghosh, 1996). The problem is difficult because there are many circumstances which can make different rules statistically dependent, including (even for domain expert unknown) relations among attribute values reflecting inherent domain properties. The weighted covering approach proposed for confirmation rule subset selection (Gamberger and Lavrač, 2000) accepts as diverse those rules that cover diverse sets of target class examples. The approach, implemented in **Algorithm RSS** outlined in Figure 2, can not guarantee statistical independence of the selected rules, but it ensures the diversity of generated models.

**Algorithm RSS: Rule Subset Selection**
**Input:** $S$ set of rules for the target class
$P$ target class examples
**Parameter:** *number* (required number of selected rules in output set $SS$)
**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) **select** from $S$ the rule with the highest weight $\sum 1/c(e)$ where summation is over the set $P' \subseteq P$ of target class examples covered by the rule
(5) **for** every $e \in P'$ covered by the selected rule **do** $c(e) \leftarrow c(e) + 1$
(6) **eliminate** the selected rule from $S$
(7) **add** the selected rule into set $SS$
(8) **end repeat**

*Figure 2.* Heuristic rule subset selection algorithm.

### 2.3 Enabling Expert-Guided Induction

Expert involvement in the suggested process of descriptive induction is important and necessary in all of its phases. In subgroup discovery, the expert is involved in guiding the search for relevant and important subgroups based on the existing expert knowledge. For the coronary heart disease (CHD) risk group detection problem used in this paper, an ideal subgroup is described by a rule that is correct for many (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 persons (false positives). By allowing the number of false positives to increase, which can be achieved by increasing the value of generalization parameter $g$, the domain expert can guide the system to induce more general subgroups of patients. In this way, the generalization parameter enables the expert to induce different models from the same data set. In addition, subgroup variation can be achieved by selecting a subset of attributes to be used in rule induction. In the CHD medical domain this corresponds to risk factor selection. By combining these two techniques an expert may interactively guide the inductive search process through many iterations until interesting subgroups have been detected. Finally, the RSS algorithm may be used to select a small set of relatively independent subgroup models.

## 3. Statistical Characterization of Subgroups

The second step in inductive modeling starts from the induced subgroup descriptions. In this process, statistical values are computed for two populations. The target population consists of CHD patients included into the analyzed subgroup, whereas the reference population are all the healthy subjects. Statistical significance is computed for all available risk factors using the $\chi^2$ test with 95% confidence level ($p = 0.05$). The statistically significant factors, distinguishing between the CHD patients belonging to the subgroup and the healthy patients, always include the conditions of rules describing the subgroup. These are called the *principal risk factors.* Often there are some additional significant differences, which are called *supporting risk factors.*

It is well known, that medical experts dislike short rules and prefer rules including as much supportive evidence as possible (Kononenko, 1993). Therefore supporting factors are very important to achieve model descriptions that are reasonably complete and acceptable for medical practice.

By selecting risk factors as candidates for statistical analysis and characterization, the expert can get different model descriptions. Additionally, interesting information can be obtained by partitioning a detected subgroup in several parts (for example, differentiating between male and female patients in the CHD domain) and comparing subgroup descriptions. Finally, statistical analysis that compares a detected subgroup with other subgroups or the whole target population (CHD patients) may be useful for a differential model description.[2]

A more detailed description of the applied statistical analysis is out of the main scope of this paper. In this work, the role of statistical analysis is to detect meaningful supporting factors, whereas the decision whether these will be used 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 or how easily they can be measured.

## 4. Subgroup Visualization

Subgroup visualization, described in this section, allows us to compare distributions of different subgroups. The visualization approach assumes the existence of at least one numeric (or ordered discrete) attribute of expert's interest for subgroup analysis. The selected attribute is plotted on the $X$-axis of the diagram. The $Y$-axis usually represents the class, or

[2]This analysis is not used to define supporting risk factors but can be included in the subgroup description to achieve better understanding.

more precisely, the number of instances belonging to the given class. It must be noted that both directions of the $Y$-axis ($Y^+$ and $Y^-$) are used to indicate the number of instances. In Figure 3, for instance, the $X$-axis represents *age*, $Y^+$-axis denotes class CHD and $Y^-$ denotes class non-CHD (or 'healthy'). Out of four graphs at the $Y^+$ side, three represent induced subgroups ($A1$, $A2$ and $C1$) of CHD patients, and the fourth shows the age distribution of the entire population of CHD (all CHD) patients. The graph at the $Y^-$ side shows only the distribution of non-CHD (all healthy) subjects in the training set.

In general, it is not necessary that $Y^+$ and $Y^-$ denote two opposite classes. If appropriate, they may denote any two classes, or even any two different attribute values, which the expert would like to compare. In medical domains, however, we typically use the $Y^+$ side to represent the number of positive cases (CHD patients, in this paper) in order to reveal properties of induced models for subgroups of these patients. On the other hand, the $Y^-$ side is reserved to reveal properties of these same models (or other models) for the negative cases (individuals without CHD). For instance, in the graph of subgroup $B1$ shown in Figure 4, the dashed line at the $Y^-$ side represents the distribution of non-CHD subjects in subgroup $B1$.

One of the advantages of using $Y^+$ and $Y^-$ as proposed above is that in binary classification problems the comparison of the area under the graph of a subgroup and the graph of the entire population visualizes the fractions of $\frac{TP}{Pos} = \frac{TP}{TP+FN}$ at the $Y^+$ side (sensitivity $TPr$), and $\frac{FP}{Neg} = \frac{FP}{TN+FP}$ at the $Y^-$ side (false alarm $FPr$), where $Pos$ and $Neg$ stand for the numbers of positive and negative cases in the entire population, respectively. For instance, in the visualization of subgroup $B1$ in Figure 4 the area between the dashed line and the $X$-axis on the $Y^-$ side represents the numbers of misclassified training instances of subgroup $B1$.

On purpose, the graphs of subgroups $A1$, $A2$ and $C1$ in Figure 3 show only the coverage of positive cases (CHD patients), and in Figure 4 the graph of subgroup $B2$ shows only the coverage of positive cases, whereas the graph of $B1$ indicates that the description of subgroup $B1$ covers positive cases (CHD patients) as well as some negative cases (healthy individuals). Except for the correct visualization of subgroup $B1$ and of the entire CHD and non-CHD distribution, Figures 3 and 4 are simplified in order to enable better understanding of the visualization method, by showing just the coverage of positive cases. In fact, all subgroups cover both positive and negative cases; if on the $Y^-$ side we showed the actual dashed line counterparts of all subgroups, figures would be too hard to understand. Please bear this in mind when interpreting the discovered rules in Section 5.2.

## 5. An Application to CHD Risk Group Detection

Early detection of CHD is an important and difficult medical problem. CHD risk factors include atherosclerotic attributes, living habits, hemostatic factors, blood pressure, and metabolic factors (Goldman et al., 1996). Their screening is performed in general practice by data collection in three different stages.

**A** 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.

**B** Collecting results of laboratory tests, including information about risk factors like lipid profile, glucose tolerance, and trombogenic factors.

**C** Collecting ECG at rest test results, including measurements of heart rate, left ventricular hypertrophy, ST segment depression, cardiac arhythmias and conduction disturbances.

Our goal was to construct at least one model for each stage, **A**, **B**, and **C**, respectively.

### 5.1 The CHD Data Set

A database with 238 patients representing typical medical practice in CHD diagnosis, collected at the Institute for Cardiovascular Prevention and Rehabilitation, Zagreb, Croatia, was used for subgroup discovery.

This database is in no respect a good epidemiological CHD database reflecting actual CHD occurrence in a general population, since about 50% of gathered patient records represent CHD patients. Nevertheless, it is very valuable since the database includes records of different types of the disease. Moreover, the included negative cases (patients who do not have CHD) are not randomly selected persons 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. This biased data set is appropriate for CHD risk group discovery, but it is inappropriate for measuring

the success of CHD risk detection and for model performance estimation in general medical practice. For this purpose, an unbiased validation data set needs to be collected. This is left for further work when designing an actual CHD epidemiological study.

### 5.2 Models of CHD Risk Groups, Their Interpretation and Visualization

At stage **A**, subgroup discovery is based on anamnestic information and physical examination results. At this stage it was rather difficult to find models with a small number of false positive predictions. The reason is a very restricted amount of information available. In order to make the problem easier, separate models were developed for groups of male and female patients.

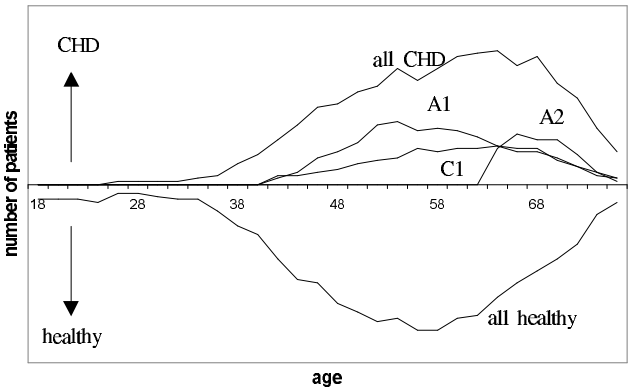


*Figure 3.* Distribution of CHD patients and healthy subjects in terms of age (in years). Graphs $A1$, $A2$, and $C1$ represent the distribution of CHD patients belonging to the corresponding subgroups: $A1$ is for male patients, $A2$ is for female, and $C1$ for patients with left ventricular hypertrophy. About 60% of CHD patients in $C1$ are also part of $A1$ and $A2$.

**Model A1 for male patients:**
CHD $\leftarrow$ *positive family history* **AND** *age over 46 years*
The main supporting characteristic is psychosocial stress, but cigarette smoking, hypertension, and overweight are also important.

The sensitivity of this model is high (45%) but its false positive rate is high as well (27%, measured on the male population in the training set). Both principal risk factors for this model 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 model are significantly older.

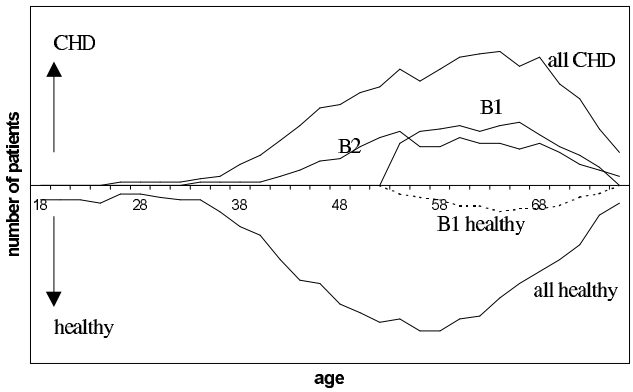


*Figure 4.* Distribution of CHD patients for subgroups $B1$ and $B2$. Subgroup $B1$ is formed of elderly people with increased total cholesterol values while subgroup $B2$ are patients with increased fibrinogen and total cholesterol values. A dashed line represents healthy subjects included in subgroup $B1$. Subgroups $B1$ and $B2$ have about 70% of patients in common.

**Model A2 for female patients:**
CHD $\leftarrow$ *body mass index over 25* $kgm^{-2}$ *(typically 29)* **AND** *age over 63 years*
Supporting characteristics are positive family history and hypertension. Women in this risk group typically have slightly increased LDL cholesterol values and normal but decreased HDL cholesterol values.

This simple model is very good for the female population. Its sensitivity is about 50% and false positive rate is only 10%, measured on the female population. Body mass index over 25 (first principal risk factor) is exactly the generally excepted margin meaning overweight (Maron, Ridker, and Pearson, 1998). 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 detected for CHD patients in this model are significantly over the margin of 25.

At stage **B**, which includes anamnestic and physical examination results as well as basic laboratory tests, two different subgroup models were induced. The first model 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 model, describing a subgroup by a combination of two risk factors, 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.

**Model B1:**
CHD ← *total cholesterol over 6.1 mmol$L^{-1}$ (typically 7.2, normal 3.6 to 5.0)*[3] **AND** *age over 53 years*
The only supporting risk factor is increased triglycerides value which is more often detected for men.

The model is characteristic for an older part of the population, especially for women (sensitivity 50%, false positive rate 20%, estimated on the training set). For the male population, the sensitivity is about 25% and false positive rate 10%. Graph B1 in Figure 4 shows the distribution of CHD patients correctly described by this model (line B1 at the $Y^+$ side) and healthy subjects erroneously included into this model (dashed line 'B1 healthy' at the $Y^-$ side). 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 model because the increased cholesterol value is mainly due to overweight.

**Model B2:**
CHD ← *total cholesterol over 5.6 mmol$L^{-1}$ (typically 6.6, normal 3.6 to 5.0)* **AND** *fibrinogen over 3.7 mmol$L^{-1}$ (typically 4.4, normal 2.0 to 3.7)*
The only supporting risk factor is positive family history.

This is a CHD patient model with similar properties for male and female patients. Further statistical analysis shows that 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 model. Although the main model properties are similar for both genders, a representative female in this model is about 66 years old while a typical male is 10 years younger.

The model has an estimated sensitivity of about 30% and the false positive rate of about 15%.

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

**Model C1:**
CHD ← *left ventricular hypertrophy*
The main supporting risk factor detected for this model is positive family history. Often the patients in this CHD group have problems with hypertension and diabetes mellitus.

The model is important both for male and female patients above the age of 55 years. Model sensitivity is 25% (see Figure 3) and false positive rate is about 5%. Left ventricular hypertrophy is a well known risk factor which includes many other known CHD risk factors like hypertension and obesity. Practical importance of the model is that it has a relatively low error rate and that it does not correlate strongly with other previously described models.

[3]Normal values between 3.6 and 5.0 are reported in Maron, Ridker, and Pearson, (1998).

### 5.3 Risk Group Evaluation

In order to evaluate the discovered risk groups, the medical expert has tested the induced subgroup models on an independent set of 50 CHD patients who entered the same Institute after having completed data collection for the needs of this study. The results for these patients, summarized in Table 1, show that the models are successful in detecting CHD patients. About 90% of CHD patients were included into at least one out of the five models. The detected sensitivity values (presented in the second column of Table 1) for models A1, B2, and C1 are significantly higher than the values computed on the set of patients used for subgroup discovery. For other two models the values do not differ significantly.

The last column in Table 1 shows mean values of the percentage of satisfied supporting factors of subgroup models (descriptions of supporting factors for each model are given in Section 5.2). The high average values between 60% and 93% demonstrate the relevance of selected supporting factors. Most of these factors have a higher rate for the specific model than for the whole CHD population. For example, model 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 model B1 have increased trygliceride value above this limit in more than 80% cases. A similar effect can be observed with *family anamnesis* for model B2 and with *HDL cholesterol values below 1.0 mmol$L^{-1}$* for model A2.

## Conclusion

This paper describes a novel approach to descriptive induction. The main part is a novel subgroup discovery algorithm incorporated into the confirmation rule set decision concept. The algorithm for selecting a small subset of confirmation rules is useful for the selection of diverse subgroups. In the second step of this process, statistical characterization adds supporting factors to the induced subgroup descriptions. They practically represent redundant informa-

*Table 1.* Summary of results obtained on an independent set of 50 CHD patients.

| Model | Percentage of CHD patients detected by the model | Percentage of supporting factors satisfied by the patients |
|---|---|---|
| A1 | 85% | 60% |
| A2 | 41% | 79% |
| B1 | 42% | 81% |
| B2 | 54% | 93% |
| C1 | 82% | 76% |

tion about subgroups, but, in our opinion, their function is extremely important in model description. On the one hand, they help the experts to get a more complete characterization and better understanding of subgroups but also they enable the expert to have increased confidence that the model is appropriate for the problem that he is trying to solve. In addition, subgroup visualization helps in understanding the relationships among models and gives visual insight into their sensitivity and false alarm rate.

The presented approach to descriptive induction uses expert knowledge at every step. Our intention was not to build a system that will substitute experts but a methodology which will help experts in model development. In our view, the possibility of influencing the induction process is an advantage of this approach.

## Acknowledgment

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

## References

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., & Verkamo, A. I. (1996). Fast discovery of association rules. In U.M. Fayyad, G. Piatetski-Shapiro, P. Smyth & R. Uthurusamy (Eds.) *Advances in Knowledge Discovery and Data Mining* (pp. 307–328). AAAI Press.

Clark, P. & Niblett, T. (1989). The CN2 induction algorithm. *Machine Learning*, *3(4)*, 261–283.

Gamberger, D. & Lavrač, N. (2000). Confirmation rule sets. In *Proc. of 4th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD2000)*, (pp. 34–43), Springer.

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

Jovanoski, V. & Lavrač, N. (2001). Classification Rule Learning with APRIORI-C. *Proceedings of the Tenth Portuguese Conference on Artificial Intelligence, EPIA-2001*, Porto, Portugal, 44–51.

Kagan T. & Ghosh, J. (1996). Error correlation and error reduction in ensemble classifiers. *Connection Science*, *8*, 385–404.

Kononenko, I. (1993). Inductive and Bayesian learning in medical diagnosis. *Applied Artificial Intelligence*, *7*, 317–337.

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

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

Michalski, R. S., Mozetič, I., Hong, J., & Lavrač, N. (1986). The multi-purpose incremental learning system AQ15 and its testing application on three medical domains. In *Proc. Fifth National Conference on Artificial Intelligence*, (pp. 1041–1045), Morgan Kaufmann.

Provost, F. & Fawcett, T. (2001). Robust classification for imprecise environments. *Machine Learning*, *42(3)*, 203–231.

Rivest, R. L. & Sloan, R. (1988). Learning complicated concepts reliably and usefully. In *Proc. Workshop on Computational Learning Theory*, (pp. 69–79), Morgan Kaufman.

Wrobel, S. (1997). An algorithm for multi-relational discovery of subgroups. *Proc. First European Symposium on Principles of Data Mining and Knowledge Discovery*, (pp. 78–87), Springer.