Machine Learning, 57, 115–143, 2004

Decision Support Through Subgroup Discovery: Three Case Studies and the Lessons Learned

NADA LAVRAČ nada.lavrac@ijs.si
Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia; Nova Gorica Polytechnic, Vipavska 13, 5001 Nova Gorica, Slovenia

BOJAN CESTNIK bojan.cestnik@ijs.si
Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia

DRAGAN GAMBERGER dragan.gamberger@irb.hr
Rudjer Bošković Institute, Bijenička 54, 10000 Zagreb, Croatia

PETER FLACH peter.flach@bristol.ac.uk
University of Bristol, The Merchant Venturers Building, Bristol BS8 1UB, UK

Editors: Nada Lavrač, Hiroshi Motoda, Tom Fawcett

Abstract. This paper presents ways to use subgroup discovery to generate actionable knowledge for decision support. Actionable knowledge is explicit symbolic knowledge, typically presented in the form of rules, that allows the decision maker to recognize some important relations and to perform an appropriate action, such as targeting a direct marketing campaign, or planning a population screening campaign aimed at detecting individuals with high disease risk. Different subgroup discovery approaches are outlined, and their advantages over using standard classification rule learning are discussed. Three case studies, a medical and two marketing ones, are used to present the lessons learned in solving problems requiring actionable knowledge generation for decision support.

Keywords: data mining, subgroup discovery, decision support, actionability, lessons learned

1. Introduction

Rule learning is an important form of *predictive* machine learning, aimed at inducing classification and prediction rules from examples (Michalski et al., 1986; Clark & Niblett, 1989; Cohen, 1995). Developments in *descriptive induction* (Wrobel & Džeroski, 1995) have recently also gained much attention of researchers interested in rule learning. These include mining of association rules (Agrawal et al., 1996), clausal discovery (De Raedt & Dehaspe, 1997; De Raedt et al., 2001), subgroup discovery (Wrobel, 1997, 2001; Kloesgen, 1996) and other approaches to non-classificatory induction.

This paper discusses actionable knowledge generation by means of subgroup discovery. The term *actionability* is described in Piatetsky-Shapiro and Matheus (1994) and Silberschatz and Tuzhilin (1995) as follows: "a pattern is interesting to the user if the user

can *do something with it* to his or her advantage." As such, actionability is a subjective measure of interestingness.

The lessons in actionable knowledge generation, described in this paper, were learned from three applications—a medical and two marketing ones—that motivated our research in actionable knowledge generation for decision support. In an ideal case, the induced knowledge should enable the decision maker to perform an action to his or her advantage, for instance, by appropriately selecting customers for a marketing campaign, or by appropriately selecting individuals for population screening concerning high disease risk.

In addition to actionable knowledge, this paper discusses also the importance of *operational knowledge*, which is in our view the most valuable form of induced knowledge. Operational knowledge enables performing an action which can operate on the target population. If an operational rule is effectively executed, this operation can change the rule coverage.

To clarify the terminology, let us discuss the actionability and operationality of three simplified rules below. Let B denote the class of people who recognize and use brand Br, $\bar{B}$ those who do not recognize the brand, and $\leftarrow$ the implication sign (used reversely, with rule conditions at the right- and the conclusion at the left-hand side of the implication sign).

(1) $B \leftarrow$ a person received the catalog of product Br
(2) $\bar{B} \leftarrow$ a person reads newspaper N & lives in area A
(3) $\bar{B} \leftarrow$ a person is younger than age A & does not read magazine M

The first rule is operational, since, by using this knowledge, the decision maker can perform an operation of sending out more catalogs, consequently increasing the rule coverage (i.e., the part of the whole population that is covered by the rule). The second rule is not operational, because the decision maker can hardly do anything to change peoples' reading habits. On the other hand, the rule is actionable as it enables performing an action: to communicate a message to $\bar{B}$ it is possible to advertise in newspaper N, or to attach a leaflet to the copies of N distributed in area A, or even send a mailing to the readers of newspaper N in area A if their addresses can be obtained from the publisher of the newspaper. In contrast to the second rule, the third rule is neither actionable nor operational, as it is much harder to get addresses of people who do not read the magazine. However, note that actionability of a rule depends on the access to the relevant databases. If one had access to the census data and newspaper readers' addresses, the third rule would become actionable as well.

The distinction between actionable and operational is particularly important in marketing, since an operational chunk of knowledge can be used not only for targeting but also for enlarging the target. It can be observed that operationality of induced descriptions is harder to achieve than their actionability. If, for example, the learned concept includes customers of a certain age and living in a certain area, this knowledge is actionable but not operational, as the rule includes attributes that cannot be manipulated. The only thing one can do is to take it into account when targeting the commercial message.

Consider another rule from the application of selecting individuals for population screening concerning high risk for coronary heart disease (CHD):

$$\text{CHD} \leftarrow \text{body mass index} > 25\,\text{kgm}^{-2}\ \&\ \text{age} > 63\,\text{years}$$

This rule is both operational and actionable. It is actionable as the general practitioner can select from his patients the overweight patients older than 63 years. The rule is also operational, as overweight can be manipulated by starting a diet, which can result in decreased rule coverage.

We provide arguments in favor of operational and/or actionable knowledge generation through recently developed subgroup discovery approaches, where a subgroup discovery task is defined as follows (Wrobel, 1997): 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 an important aspect of the above definition: there is a predefined property of interest, meaning that a subgroup discovery task aims at characterizing population subgroups of a given *target* class. This property suggests that standard classification rule learning could be used for solving the subgroup discovery task. However, the goal of standard rule learning is to generate models, one for each class, consisting of rulesets describing class characteristics in terms of properties occurring in the descriptions of training examples. In contrast, subgroup discovery aims at discovering individual rules or *patterns* of interest, which must be represented in explicit symbolic form and which must be relatively simple in order to be recognized as actionable by potential users.

This paper is organized as follows. In Section 2 three real-life problems providing the motivation for this work are outlined, followed by Sections 3 which presents the form of induced subgroups, measures of subgroup interestingness and subgroup evaluation in ROC space. Sections 4–6 present three recently developed subgroup discovery approaches, applied to three real-life problem domains in medicine and marketing. The main goal of this paper is to present the lessons learned from these applications and from the development of novel subgroup discovery algorithms; these lessons are described in detail in Sections 7 and 8, respectively. The paper concludes with a brief summary and directions for further work.

2. Motivation for subgroup discovery: Three case studies

For the work described in this paper the motivation comes from the need for decision support in targeting a marketing campaign, or planning a population screening campaign aimed at detecting individuals with high disease risk. For instance, the problem of targeting a marketing campaign for a brand can be addressed by finding population subgroups that will be interested in buying the brand product. Finding population subgroups of this kind can be viewed as a subgroup discovery task (Wrobel, 1997). In subgroup discovery we wish to obtain subgroups of the population that are sufficiently large and have a significantly different target class distribution than the entire population (e.g., subgroups of people for which brand recognition success rate is much higher than average).

Coronary heart disease risk group detection. In the problem of detection and description of coronary heart disease (CHD) risk groups (Gamberger & Lavrač, 2002), described in Section 4, data collected in general screening include medical history information and physical examination results, laboratory tests, and ECG at-rest test results. 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 hard problem in CHD prevention is to find endangered individuals with slightly abnormal values of risk factors and in cases when combinations of different risk factors occur. The induced risk group descriptions aim at helping the general practitioners to recognize CHD and/or detect the illness even before the first symptoms actually occur. The expert-guided discovery and use of induced risk group descriptions developed in our research is aimed at enabling easier detection of important risk groups in the population.

Decision support in a direct mailing campaign. The first marketing problem, described in Section 5, is the direct mailing problem (Flach & Gamberger, 2001). The starting point is a relational database obtained by interviewing potential customers. The customers are described by their answers concerning age, level of education, profession, place of living, preferences, and habits like what TV programs they watch and what newspapers they read regularly. The direct mailing problem is the problem of selecting potential customer subgroups that can be targeted by mailing advertising campaigns. In many respects, this problem is similar to the patient risk group detection problem; the main difference is that in this application the goal is clearly stated in the form of profit maximization.

Decision support in a public advertising campaign. The second marketing problem, described in Section 6, is the problem of targeting a public advertising campaign for a Slovenian natural non-alcoholic sparkling drink brand (Cestnik et al., 2002). The same customer database as in the first marketing problem was used for data analysis. The task is to find significant characteristics of customer subgroups who do not know a brand, relative to the characteristics of the population that recognizes the brand. Again, the nature of the problem is similar as in the above two applications; here, however, the emphasis is on high uncertainty of the domain (unreliable data and probabilistic nature of the decision making process). In marketing, for example, it makes no sense saying that all the readers of a specific newspaper will buy a certain product; however, it may be reasonable to conclude that the probability of them buying the product is higher than average.

Below we describe in some more detail the motivation for subgroup discovery in the two marketing problems, through a well-known marketing notion of *market segmentation*. This notion—if viewed in more general terms as *population segmentation* which is more appropriate for the medical application—explains the need for subgroup discovery in all the three problems studied in this paper.

The market targeting task is that of selecting potential customer subgroups of the population that can be specifically targeted by advertising campaigns. The core of a campaign usually consists of three phases: identification of the goal of the campaign, selection of a target population, and communication of a desired message to the target population. The goal of the campaign may be, for example, to gain new customers, or to reinforce existing customers. Setting the goal determines the targeting strategy. In this study we focus on the second phase of the marketing campaign: selecting a target population.

The key question is how to help marketing analysts decide which segments of customers are the most promising with respect to the expected results. One heuristic is to target the segment with the heaviest customers, i.e., the customers with the highest expected purchase

frequency and the largest amount of money spent. Although this makes sense intuitively, there are some strong indications that such an approach is not the most promising one (Myers, 1996): one or more major competitors may have already targeted this group successfully; the company product line is not well designed for this group; in reality there are no heavy customers; the company is too small to go after the heavy-customer segment; or the company wants to develop different brand marketing campaigns for each usage group.

For these reasons, a straightforward decision to target the segment with the heaviest customers is not always the best option. Consequently, the effort to better understand other niches of the market segmentation is well justified. Here, data mining tools can provide a crucial leverage.

3. Subgroup discovery

This section gives a brief introduction to subgroup discovery, and describes approaches to learning and evaluation.

3.1. Subgroup discovery as rule learning

The result of standard rule induction is a classification model consisting of a set of rules. In contrast with model induction, subgroup discovery aims at finding patterns in the data, described in the form of individual rules. As in classification rule learning, an induced subgroup description has the form of a (backwards) implication:

$$Class \leftarrow Cond$$

In terms of rule learning, the property of interest for subgroup discovery is the target class (*Class*) that appears in the rule consequent, and the rule antecedent (*Cond*) is a conjunction of features (attribute-value pairs) selected from the features describing the training instances.

Subgroup discovery is a task at the intersection of predictive and descriptive induction. By inducing rules from labeled training instances (labeled positive if the property of interest holds, and negative otherwise), the process of subgroup discovery is targeted at uncovering properties of a selected target population of individuals with the given property of interest. In this sense, subgroup discovery is a form of *supervised learning*. The fact that a subgroup discovery task aims at characterizing population subgroups of a given target class suggests that standard classification rule learning could be used for solving the task. However, in many respects subgroup discovery is a form of *descriptive induction* as the task is to uncover individual rules or *patterns*, which must be relatively simple in order to be recognized as actionable by potential users.

Standard classification rule learning algorithms cannot appropriately address the task of subgroup discovery for at least two other reasons: first, they usually apply the covering algorithm for ruleset construction, and second, they use search heuristics aimed at optimizing ruleset accuracy; as will be seen in this paper, these two reasons hinder the applicability of classification rule induction approaches to subgroup discovery. Moreover, in subgroup

discovery one can often tolerate many more false positives (negative examples incorrectly classified as positives) than in a classification task. However, both tasks, subgroup discovery and classification rule learning, can be unified under the umbrella of *cost-sensitive* classification. This is because when deciding which classifiers are optimal in a given context it does not matter whether we penalize false negatives as is typically the case in classification, or reward true positives as in subgroup discovery—it only matters for determining the expected profit in a given context.

Each rule describing a subgroup can be extended with the information about the rule *quality*. In this paper, a standard rule describing a subgroup has the following form:

$$Class \leftarrow Cond\ [TPr, FPr] \tag{1}$$

where *Class* is the target property of interest, *Cond* is a conjunction of features (attribute-values), *TPr* is the *true positive rate* or the *sensitivity*, computed as $p(Cond \mid Class) = \frac{n(Class \cdot Cond)}{Pos}$, and *FPr* is the *false alarm* or *false positive rate*, computed as $p(Cond \mid \overline{Class}) = \frac{n(\overline{Class} \cdot Cond)}{Neg}$. In these formulas $n(Class \cdot Cond)$ is the number of true positives *TP* (the number of covered instances belonging to *Class*), $n(\overline{Class} \cdot Cond)$ the number of false positives *FP* (the number of covered instances not belonging to *Class*), *Pos* is the number of positives (instances of the target class), *Neg* the number of negatives, and $N = Pos + Neg$ is the size of the entire population.

In addition to earlier approaches to subgroup discovery (Kloesgen, 1996; Wrobel, 1997, 2001), several novel subgroup discovery methods have been developed. The algorithms, outlined in this paper, output subgroup descriptions in different forms, attaching different quality measures.

- The CN2-SD algorithm (Lavrač et al., 2002, 2004), used in the direct mailing application outlined in Section 5, induces subgroups in the form of rules described in Eq. (1).
- The SD algorithm (Gamberger & Lavrač, 2002), used in the coronary heart disease risk group detection application outlined in Section 4, induces rules of the form $Class \leftarrow Cond\ [TP, FP]$, where the features forming rule conditions are called the *principal factors*. In subgroup descriptions, the principal factors are supplemented by a list of *supporting factors*, which denote features (attribute values) characterising the subgroup (target class instances covered by the rule) by being statistically significantly different from the values characterising the reference population (all non-target class instances). Notice that each supporting factor itself can be viewed as a simple rule with one condition only. In this respect, supporting factors are similar to decision stumps (Holte, 1993).
- In data analysis from highly uncertain data, it turns out that—instead of using a subgroup discovery approach resulting in subgroup descriptions in the standard rule form—it is beneficial to present a population subgroup by listing its supporting factors only. This claim is supported by the outcome of the marketing application outlined in Section 6 (Cestnik et al., 2002), where a population subgroup is presented by a list of *supporting factors*. In addition, the *opposing factors*, characteristic for the negation of the target concept, are also listed, which is in accordance with the basic principles of Bayesian analysis (Berger, 1985).

3.2. Subgroup evaluation measures

As shown in Section 3.1, each rule describing a subgroup can be extended with the information about the rule quality. While the basic information of rule quality is usually attached to the induced rule itself, as output of the learning algorithm, other quality measures are usually computed separately, in order to evaluate the output of the induction process as a whole, enabling the comparison of the performance of different algorithms.

One can distinguish between *objective* quality measures and *subjective* measures of interestingness (Silberschatz & Tuzhilin, 1995). Both the objective and subjective measures need to be considered in order to solve subgroup discovery tasks. Which of the quality criteria are most appropriate depends on the application. Obviously, for automated rule induction it is only the objective quality criteria that apply. However, for evaluating the quality of induced subgroup descriptions and their usefulness for decision support, the subjective criteria are more important, but also harder to evaluate.

Below is a list of *subjective* measures of interestingness:

- *Usefulness.* Usefulness is an aspect of rule interestingness which relates a finding to the goals of the user (Kloesgen, 1996).
- *Actionability.* "A rule is interesting if the user can do something with it to his or her advantage" (Piatetsky-Shapiro, & Matheus, 1994; Silberschatz & Tuzhilin, 1995).
- *Operationality.* In this paper we have introduced the notion of operationality, which is a special case of actionability. Operational knowledge is the most valuable form of induced knowledge, as it enables performing an action which can operate on the target population. If an operational rule is effectively executed, this operation can affect the target population and change the rule coverage.
- *Unexpectedness.* A rule is interesting if it is surprizing to the user (Silberschatz & Tuzhilin, 1995).
- *Novelty.* A finding is interesting if it deviates from prior knowledge of the user (Kloesgen, 1996).
- *Redundancy.* Redundancy amounts to the similarity of a finding with respect to other findings; it measures to what degree a finding follows from another one (Kloesgen, 1996), or to what degree multiple findings support the same claims.

In listing the *objective* quality measures—in line with the distinction between *predictive induction* and *descriptive induction* made in Section 1—we distinguish between the *predictive* and *descriptive* quality measures. A typical predictive quality measure, measuring the quality of a ruleset, is *predictive accuracy* of a ruleset, defined as the percentage of correctly predicted instances.[1]

In contrast with predictive quality measures, descriptive quality measures evaluate each individual subgroup and are thus appropriate for evaluating the success of subgroup discovery. The following measures turn out to be most appropriate for measuring the quality of individual rules: *rule size*, *coverage*, *support*, *accuracy* (in different contexts also called *precision* or *confidence*), *significance* and *unusualness*. The measures for evaluating each individual rule can be complemented by their variants that compute the average over the induced set of subgroup descriptions, which enables the comparison of

different subgroup discovery algorithms (see Lavrač et al., 2004) for the definitions of these measures).

To explain rule *significance* and *unusualness*, which are the most important subgroup discovery measures, some of the other measures for evaluating the quality of rules of the form $Class \leftarrow Cond$ need to be explained first. *Coverage* $p(Cond)$ is a measure of *generality*, computed as the relative frequency of all the examples covered by the rule. *Support* $p(Class \cdot Cond)$ is computed as the relative frequency of correctly classified covered examples. Rule *accuracy* $p(Class \mid Cond)$ (called *precision* in information retrieval and *confidence* in association rule learning) is the fraction of predicted positives that are true positives. Next, we define *accuracy gain*, $p(Class \mid Cond) - p(Class)$, as the difference between rule accuracy $p(Class \mid Cond)$ and default accuracy $p(Class)$ achieved by the trivial rule $Class \leftarrow true$.

– *Significance* (or *evidence* in the terminology of Kloesgen (1996) indicates how significant a finding is compared to a null hypothesis of statistical independence. In the CN2 algorithm (Clark & Niblett, 1989), significance is measured in terms of the likelihood ratio statistic of a rule:

$$Sig(Class \leftarrow Cond) = 2 \sum_i n(Class_i \cdot Cond) \cdot \log \frac{n(Class_i \cdot Cond)}{n(Class_i)p(Cond)} \tag{2}$$

  where $n(Class_i)p(Cond)$ is the expected number of $Class_i$ instances among the ones satisfying $Cond$ under the null hypothesis of statistical independence of $Class_i$ and $Cond$.[2]
– *Unusualness* of a rule is computed by the *weighted relative accuracy* of a rule (Lavrač, Flach, & Zupan, 1999), defined as follows:

$$WRAcc(Class \leftarrow Cond) = p(Cond) \cdot [p(Class \mid Cond) - p(Class)]$$

  Weighted relative accuracy can be understood as trading off rule *coverage* $p(Cond)$ and *accuracy gain* $p(Class \mid Cond) - p(Class)$.

As shown in Section 3.3, *WRAcc* is appropriate for measuring the unusualness of separate subgroups, because it is proportional to the vertical distance of the subgroup to the ascending diagonal in ROC space. As such, *WRAcc* also reflects rule significance—the larger *WRAcc* is, the more significant the rule is, and vice versa. As both *WRAcc* and rule significance measure the distributional unusualness of a subgroup, they are the most important quality measures for subgroup discovery.

Significance and unusualness can be used also as search heuristics in rule construction. While significance only measures distributional unusualness, computed in terms of correctly classified covered examples of all classes, *WRAcc* takes explicitly the rule coverage into account, therefore we consider *unusualness* to be the most appropriate measure for subgroup quality evaluation. As a result, we have replaced the rule significance heuristic in CN2 by the *WRAcc* heuristic in the implementation of the CN2-SD subgroup discovery algorithm, used in the marketing application in Section 5.

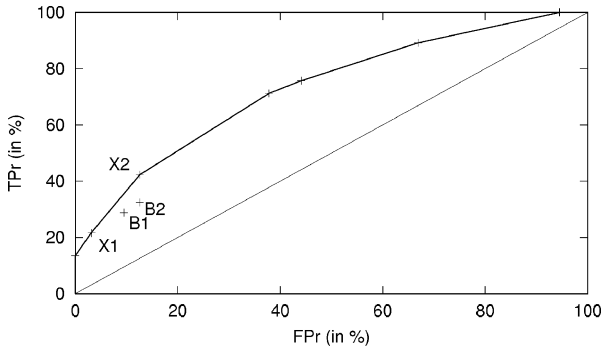

Figure 1. Labels $B1$ and $B2$ denote suboptimal subgroups, and $X1$ and $X2$ denote two of the seven subgroups (marked by $+$) forming the ROC convex hull. The ascending diagonal connecting points (0,0) and (100,100) represents rule positions with zero significance.

3.3. Subgroup evaluation in ROC space

Receiver Operating Characteristic (ROC) analysis (Provostand & Fawcett, 1998) enables us to plot the performance of classifiers in ROC space, where the performance of a classifier is characterized by its false positive rate (*FPr*), plotted on the X-axis, and its true positive rate *TPr*), plotted on the Y-axis (see figure 1). Points on the ascending diagonal connecting (0,0) and (100,100) correspond to classifiers that perform prediction by random guessing. Since ROC analysis is applicable to subgroup discovery, we briefly describe it here.

In the perspective of predictive induction, each point in ROC space represents a classifier. The ROC convex hull is a simple method to select the best ones among a set of classifiers. The ROC curve is a piecewise linear curve connecting a selection of 'best' points (classifiers with the best TPr/FPr tradeoff) in ROC space, such that all other classifiers are below it. This convex hull supports the choice of a single best classifier, provided that for a particular problem domain one knows the operating characteristics determined by the class and cost distribution (see the ROC analysis for a given *operational context* in Section 5). Alternatively, the ROC methodology allows, through the construction of a convex hull, to identify classifiers that are optimal for various TPr/FPr tradeoffs; as such, it identifies an integrated set of solutions, together with their optimality conditions in terms of *TPr* and *FPr*, which provides decision support for classifier selection as well as model combination.

In the context of descriptive induction, and subgroup discovery in particular, each point in ROC space represents a pattern (e.g., a rule). Thus, individual subgroups can be plotted in ROC space. The ascending diagonal represents subgroups with the same class distribution as the overall population, i.e., subgroups without distributional unusualness, while the interesting subgroups are those sufficiently distant from the diagonal.

There are essentially two ways to define the notion of 'distance from the diagonal'. The first is to select the subgroups on the convex hull, as described before. This method selects subgroups that are optimal under varying TPr/FPr tradeoffs. Figure 1 shows seven

rules on the convex hull (marked by +), including $X1$ and $X2$, while two rules $B1$ and $B2$ below the convex hull are of lower quality in terms of their TPr/FPr tradeoff.

In the second approach, distance from the diagonal is measured in geometric terms. This requires that we apply a fixed TPr/FPr tradeoff. In particular, the *WRAcc* measure gives equal weight to increasing the true positive rate and decreasing the false positive rate. Specifically, in true/false positive rate notation we have $WRAcc = p(Class) \cdot p(\overline{Class}) \cdot [TPr - FPr]$.[3] It follows that *WRAcc* iso-performance lines are parallel to the diagonal (Flach, 2003; Fürnkranz & Flach, 2003). Consequently, a point on the line $TPr = FPr + a$, where a is the vertical distance of the line to the diagonal, has $WRAcc = a \cdot p(Class) \cdot p(\overline{Class})$. Thus, given a fixed class distribution, *WRAcc* is proportional to the vertical distance a to the diagonal.

4. First case study: Coronary heart disease risk group detection

Having covered the essentials of the subgroup discovery approaches, this section presents one of the three case studies in which the subgroup discovery approaches were used.

4.1. Problem definition

Early detection of artherosclerotic coronary heart disease (CHD) is an important and difficult medical problem. CHD risk factors include artherosclerotic attributes, living habits, hemostatic factors, blood pressure, and metabolic factors. 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 thrombogenic factors.
C: Collecting ECG at rest test results, including measurements of heart rate, left ventricular hypertrophy, ST segment depression, cardiac arrhythmias and conduction disturbances.

In this application, the goal was to construct at least one relevant and interesting CHD risk group for each of the stages A, B, and C, respectively.

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 (Gamberger, Lavrač, & Krstačić, 2003). The 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, the database is very valuable since it 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 considered by general practitioners as potential CHD patients, and hence sent for further investigations to the Institute. This biased dataset is

appropriate for CHD risk group discovery, but it is inappropriate for measuring the success of CHD risk detection and for subgroup performance estimation in general medical practice.

4.2. Subgroup discovery with the SD algorithm

The subgroup discovery algorithm, Algorithm SD (Gamberger & Lavrač, 2002), is publicly available as part of the on-line Data Mining Server (DMS) at `http://dms.irb.hr`. The algorithm assumes that the user selects one class as a *target class*, and learns subgroup descriptions of the form *Class* ← *Cond* [*TP*, *FP*]. The result is a set of best rules, induced by heuristic beam search for rules with a maximal q value, where the quality function $q = \frac{TP}{FP+g}$ is defined by setting a value of the user-defined generalization parameter g (selection of larger g will result in a larger number of examples covered by the induced rule, the default being $g = 1$). Function q defines a tradeoff between true positives *TP* and false positives *FP* covered by the rule. By searching for rules with high quality q, the algorithm tries to find rules that cover many target class examples and a low number of non-target examples.

Algorithm SD is similar to association rule learning in the sense that in order for a rule to be included in the induced ruleset, the rule must satisfy the minimal support requirement. Typically, Algorithm SD generates many rules of high quality q satisfying the requested condition of a minimal number of covered target class examples, defined by the algorithm's *min_support* parameter. Accepting all these rules is generally not desirable 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 features and have similar prediction properties (define similar subgroups). A solution to this problem is to reduce generated rulesets to include only a relatively small number of rules which are as diverse as possible. The DMS approach to rule subset selection (Gamberger & Lavrač, 2002) accepts as diverse those rules that cover diverse sets of target class examples. The approach can not guarantee statistical independence of the selected rules, but ensures their diversity. The rule subset selection algorithm is similar to the weighted covering algorithm implemented in CN2-SD (briefly described in Section 5.3).

4.3. Results of patient risk group detection

The process of expert-guided subgroup discovery was performed as follows. For every data stage A, B and C, DMS was run for values g in the range 0.5 to 100 (values 0.5, 1, 2, 4, 6, ...), 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, following the suggestions of the medical expert to limit the number of features in the rule body and avoid the generation of rules whose features would involve expensive and/or unreliable laboratory tests.

In the iterative process of rule generation and selection, the expert has selected five most interesting CHD risk groups. Table 1 shows the induced subgroup descriptions in a simplified *Class* ← *Cond* form, without the information on their *TP* and *FP* values. As mentioned in Section 3.1, the features appearing in the conditions of rules describing the subgroups are called the *principal factors*. The described iterative process was successful

Table 1. Induced subgroups in the form of rules. 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 (A, B and C, respectively) with different g parameter values (14, 8, 10, 12 and 10, respectively).

Expert selected subgroups			
A1	CHD	←	Positive family history & age over 46 year
A2	CHD	←	Body mass index over 25 kgm^{-2} & age over 63 years
B1	CHD	←	Total cholesterol over 6.1 mmolL^{-1} & age over 53 years & body mass index below 30 kgm^{-2}
B2	CHD	←	Total cholesterol over 5.6 mmolL^{-1} & fibrinogen over 3.7 gL^{-1} & body mass index below 30 kgm^{-2}
C1	CHD	←	Left ventricular hypertrophy

for data at stages B and C, but it turned out that medical history data on its own (stage A data) is not informative enough for inducing subgroups, i.e., it failed to fulfil the expert's subjective criteria of interestingness. Only after engineering the domain, by separating male and female patients, interesting subgroups $A1$ and $A2$ have actually been discovered.

Separately for each data stage A, B and C, we have investigated which of the induced rules are the best in terms of the TPr/FPr tradeoff in ROC space,[4] i.e., which of them are used to define the ROC convex hull. At stage B, for instance, seven rules (marked by +) are on the convex hull of the ROC space shown in figure 1. Two of these rules, $X1$ and $X2$, are listed in Table 2. Notice that the expert-selected subgroups B1 and B2 are significant, but are not among those lying on the convex hull in figure 1. The reason for selecting exactly those two rules at stage B are their simplicity (consisting of three features only), their generality (covering relatively many positive cases) and the fact that the used features are, from the medical point of view, inexpensive laboratory tests. Moreover, the two rules B1 and B2 were deemed interesting by the expert.

Additionally, rules B1 and B2 are interesting because of the feature *body mass index below* 30 kgm^{-2}, which is intuitively in contradiction with the expert knowledge that both increased body weight as well as increased total cholesterol values are CHD risk factors. It is known that increased body weight typically results in increased total cholesterol values

Table 2. Two of the best induced subgroups induced for stage B, $X1$ and $X2$, induced using the g values 4 and 6, respectively. The position of subgroups in the ROC space are marked in figure 1.

Best induced subgroups			
X1	CHD	←	Age over 61 years & tryglicerides below 1.85 mmolL^{-1} & high density lipoprotein below 1.25 mmolL^{-1}
X2	CHD	←	Body mass index over 25 & high density lipoprotein below 1.25 mmolL^{-1} & uric acid below 360 mmolL^{-1} & glucose below 7 mmolL^{-1} & fibrinogen over 3.7 gL^{-1}

while subgroups B1 and B2 actually point out the importance of increased total cholesterol when it is not caused by obesity as a relevant disease risk factor.

4.4. Statistical characterization of subgroups

The next step in the proposed subgroup discovery process starts from the discovered subgroups. In this step, statistical differences in distributions are computed for two populations, the target and the reference population. The target population consists of true positive cases (CHD patients included into the analyzed subgroup), whereas the reference population are all available non-target class examples (all the healthy subjects). Statistical differences in distributions for all the descriptors (attributes) between these two populations are tested using the χ^2 test with 95% confidence level ($p = 0.05$).

To enable testing of statistical significance, numerical attributes have been partitioned in up to 30 intervals so that in every interval there are at least 5 instances. Among the attributes with significantly different value distributions there are always those that form the features describing the subgroups (the principal factors), but usually there are also other attributes with statistically 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 factors*.

Supporting factors are very important for subgroup descriptions to become more complete and acceptable for medical practice. Medical experts dislike long conjunctive rules which are difficult to interpret. On the other hand, they also dislike short rules providing insufficient supportive evidence. In this work, we found an appropriate tradeoff between rule simplicity and the amount of supportive evidence by enabling the expert to inspect all the statistically significant supporting factors, whereas the decision whether they indeed increase the user's confidence in the subgroup description is left to the expert. In the CHD application the expert has decided whether the proposed supporting factors are meaningful, interesting and actionable, how reliable they are and how easily they can be measured in practice. Table 3 lists the expert selected supporting factors.

5. Second case study: Decision support in a direct mailing campaign

One of the most important tasks of a marketing expert is to efficiently target a subset of the population for advertising a specific product or service (Berry & Linoff, 2000). Usually, there is a trade-off between (a) the cost of communicating a message to the entire population, and (b) lower revenues caused by selecting too narrow a population segment, which may result in missing some of the potential customers. Since the nature of the decision-making problem is quite complex, common solutions usually rely on statistical analysis of data gathered from surveys.

In this section we present a method resulting in the maximization of the expected profit of a marketing campaign. This method is based on ROC analysis and is illustrated by the case study of direct mailing to consumers that do not yet recognize a particular yoghurt brand.

Table 3. Statistical characterizations of induced subgroup descriptions (supporting factors).

	Supporting factors
A1	• psychosocial stress
	• cigarette smoking
	• hypertension
	• overweight
A2	• positive family history
	• hypertension
	• slightly increased LDL cholesterol
	• normal but decreased HDL cholesterol
B1	• increased triglycerides value
B2	• positive family history
C1	• positive family history
	• hypertension
	• diabetes mellitus

5.1. The decision making context

In direct mailing, a mailing is sent out to all people in a subgroup of the general population, for which the expected profit from sending a mailing is larger than the actual mailing cost.

The direct mailing problem can be viewed as a generic decision-making problem in which profit and cost of an individual marketing action (sending out a mailing, in this case) are taken into account, and the goal is to distinguish between the negatives and the positives defined below. The *negatives* are people who will not spend additional money on the product, even if they receive the mailing (for instance because they are not interested in the product, are using a product from a competitor, or are already using the product. The *positives* are people who might spend money on the product if they knew about it. A direct mailing problem is guided by two parameters: a marginal cost c per mailing, and an average profit p per true positive, i.e., potential customer reached by the mailing. Average profit includes the cost of the mailing, and a certain response percentage among the true positives. So, for instance, if 1% of the positives reached by the mailing become customers and spend \$1000 each, while the marginal cost per mailing is \$1, then average profit p is \$9.

The default decision a marketing analyst can take is to send a mailing to everybody in the population, resulting in a default profit: $profit = p \cdot Pos - c \cdot Neg$, where Pos (Neg) is the number of people that responded positively (negatively) to the marketing campaign, and $N = Pos + Neg$ is the size of the entire population. More generally, the expected profit is

$$profit = p \cdot TP - c \cdot FP = p \cdot TPr \cdot Pos - c \cdot FPr \cdot Neg \tag{3}$$

Who is (and who is not) a potential customer in a given application is a matter of choice of the marketing analyst, given the decision making *context*, which is in this problem defined

by the four-tuple (c, p, Pos, Neg). Note that the default profit may be negative, if positives are rare and/or mailings are expensive. Hence, in general one should not use the default strategy but try to detect population subgroups for which sending a mailing improves upon the default profit (or yield positive profit if the default profit is negative).

5.2. Problem definition

The dataset investigated in this work is a relational database obtained by interviewing potential customers. It consists of customers' answers about how they recognize, use and appreciate tested brands. Questions that are interesting for the marketing analyst, whose task is to design a marketing campaign, are: Which brands have the potential to improve their recognition or usage rate? What are the characteristics of the people appreciating or using a specific brand? What is the nature of the relationship between brand recognition and brand usage?

The problem addressed in this section is the selection of potential customer subgroups from the general population that can be successfully targeted by advertising campaigns. In this task, the customers can be classified into two groups according to whether or not they recognize the brand. In our approach, the group of people who do not know the brand is selected as the target ('positive') class for the data mining process. The task is (a) to find their significant characteristics, relative to the characteristics of the population that recognizes and regularly uses the brand ('negative' class) and (b) to determine if and how an advertising campaign could increase the expected brand recognition, and consequently, the profit of the company. Moreover, the goal is also to optimize the cost of the campaign.

We have designed a special questionnaire to gather data for the given marketing problem. In the design phase we first made sure that every input variable, required for the analysis, could be obtained from the completed questionnaires. However, we found that there is often a tradeoff between the desired quantity of answers from each respondent and the ability of the respondent to provide high quality answers. Therefore, we invested large efforts to optimize the demands of the questionnaire on customers, which resulted in increased quality of the gathered data. Specifically, we dealt with three issues: avoiding questionnaire fatigue, measuring recognition of brand names with words and pictures, and validating the brand name recognition.

The aim of the questionnaire was to evaluate 300 given brand names. The obvious idea of asking every respondent to evaluate each brand name would result in loss of concentration and would eventually decrease the quality of obtained results. So, we decided to leverage the burden by allowing each respondent to evaluate only 15 randomly selected brands. Further optimization was obtained by the use of genetic algorithms to improve the odds of brand names from similar categories to appear on the same questionnaire. As a result, the accuracy of mutually comparing competitors improved, which was of high interest for the end-user.

Every brand under study had a name and a logo. When asking respondents to recognize brands, we first assessed their recognition of written brand name. On the second screen, they were presented with the corresponding graphical logos and again asked to state their recognition. So, we were actually able to measure the correlation between the brand name and its logo recognition, which turned out to be very valuable in subsequent

studies. However, for the purpose of subgroup discovery we relied solely on the logo recognition.

The input data consists of two relational tables: (1) the general customer responses and demographic facts, and (2) the responses about specific brand names. The first table contains customer responses to general questions and demographic facts. A unique key Q identifies each customer. There are 2013 records (customers) in the table. The customers are described by their age, level of education, occupation, address, consumer preferences and habits (for example the TV programs they watch and the newspapers they read regularly). In total, the table consists of 55 attributes.

The second table contains questionnaire responses about specific brand names. There are 300 different brand names analyzed in the survey; and to avoid response fatigue, each customer was given only a subset of 15 brand names to evaluate with respect to their recognition, reputation and usage. Therefore, the second table contains Q as a foreign key and D as a key for a specific brand. In addition, there are three attributes that represent how the respondent recognized, used and valued the product representing the questioned brand. There are in total $2013 \cdot 15 = 30195$ answers (database records).

In order to obtain a single table from the first two tables, we extended the first table with 300 attributes derived from the second table representing the frequency of consumption of each particular brand. The frequency of consumption of a particular brand B has values from 1 to 5, 1 meaning that the customer does not know the brand (therefore does not use it), and 5 meaning that he or she regularly uses it.

In summary, the final data table that is used in subgroup discovery consists of 2013 records (customers) described by 55 attributes from the first table, followed by 300 binary attributes describing the usage of the corresponding brands. In the direct mailing application described in this section, the target class is formed as a single column representing one selected brand (the class 'does not use the yoghurt brand' being labeled class *positive*, and 'uses the yoghurt brand' being labeled class *negative*, due to the desire of targeting a marketing campaign to predict potential customers who do not yet use the brand). Note that the marketing application of Section 6 the target class was a combination of several columns (brands) included in the description of the target concept.

In this application, the need for data preprocessing arose from the fact that the concept 'user of brand B' can only be determined for a limited number of respondents: i.e., only those that were originally asked this question. As every respondent was only asked to evaluate 15 brands out of 300, only one customer in every 20 was asked about the recognition and reputation of a specific brand. So, each record contains only 15 actual frequency consumption values; the rest of 285 attribute values were unknown. To address this problem we used the probabilistic classification rule-learning algorithm CN2 (Clark & Niblett, 1989; Clark & Boswell, 1991) to construct a classifier that was used for missing value imputation: to fill in the class label data about using/non-using a specific brand. The input variables for each classifier were the attributes that originated from the first table and were therefore known for every record. As CN2 allows for probabilistic predictions, probability of class assignment was used as a class value as follows: value 1 for the probabilities in the range [0, 0.2], value 2 for (0.2, 0.4], ..., and value 5 for the probabilities in the range (0.8, 1]. Finally, the original frequency of brand consumption answer was split into two values (binarization)

and the probabilistic answers were discretized. Binarization was performed as follows. We replaced answers 4 and 5 by class label 'uses the yoghurt brand', and answers 1, 2 and 3 by class label 'does not use/know the yoghurt brand'.

The outcome of the data preprocessing step, which was performed by applying a rule learner to predict the missing values in the dataset, are uncertain predictions. Low statistical significance of predictions is also due to the inherent nature of the domain (targeting a population in marketing) which is inexact and probabilistic. So, it usually suffices to induce a rule stating that, for example, the readers of a specific newspaper will buy a certain product with probability higher than average.

5.3. Subgroup discovery with CN2-SD and profit maximization

Algorithm CN2-SD (Lavrač et al., 2002) adapts classical classification rule learning algorithm CN2 (Clark & Niblett, 1989) to subgroup discovery. The CN2 algorithm uses the *covering algorithm* for ruleset construction. However, 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, i.e., 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 the *weighted covering algorithm* used in CN2-SD, positive examples covered by the induced rule are not deleted from the current training set. Instead, their weights are modified so that the probability that an example with a modified weight will be covered by subsequent rules is decreased. Initial example weights of all target class examples are set to 1, while in the following iterations weights of positive examples covered by the constructed rule decrease according to the formula $\frac{1}{i+1}$, where i is the number of rules covering the example. Example weights are also taken into account when evaluating the weighted relative accuracy heuristic used in CN2-SD rule construction.

The ROC convex hull method, which is used in CN2-SD to evaluate and select best subgroups is in this application enhanced with the aim of profit maximization in a given context, defined by the four-tuple (c, p, Pos, Neg) (see Section 5.1).

In the rest of this section we give an outline of the approach to profit maximization (Flach & Gamberger, 2001). Maximum profit can be obtained when all potential customers are targeted without any expense lost on non-potential customers. This situation corresponds to the ideal subgroup with $TPr = 1$ and $FPr = 0$, where TPr and FPr stand for true positive rate and false positive rate, respectively. In this case, the associated profit is $p \cdot Pos$, where p is the average gain per true positive. By manipulating Eq. (3) in Section 5.1 we see that lines with equal profit in ROC space are defined by the equation:

$$TPr = \frac{c \cdot Neg}{p \cdot Pos} FPr + \frac{profit}{p \cdot Pos} \tag{4}$$

This means that in a given context (c, p, Pos, Neg), rules with different values (TPr, FPr) will have the same profit values if lying on the same line with the intended slope in ROC space, where the intended slope is given by $\frac{c \cdot Neg}{p \cdot Pos}$.

All subgroups, evaluated in ROC space, which lie on the same equal profit line, will obviously result in an equal total amount of profit. Subgroups above (below) this line will result in a higher (lower) profit. As suggested by Eq. (4), it is sufficient to work with *normalized profit* rather than absolute profit in order to select the optimal subgroup, where normalized profit is defined as profit divided by $p \cdot Pos$. For instance, a normalized profit of 40% may mean that we reached 50% of positives, but 1/5 of the profit was spent on the negatives addressed; or it may mean that we in fact reached 40% of the positives and no negatives. From the perspective of profit maximization, both situations are equivalent.

Notice that the slope of equal profit lines is completely defined by parameters determining the context (c, p, Pos, Neg). A rule, which is optimal for the given context, is determined by the point on the ROC convex hull that has an equal profit line as its tangent (if a segment of the convex hull has the same slope as the equal profit lines, either point on the end of the segment can be selected).

In the direct mailing application, the candidate subgroups were induced by the CN2-SD subgroup discovery algorithm. To decide which subgroup to target in order to maximize the expected profit for the given yoghurt brand, the marketing analyst computed the intended slope, which determined the optimal subgroup to be targeted in the direct mailing campaign.

6. Third case study: Decision support in targeting an advertising campaign

Instead of trying to describe a subgroup by a rule, as done in Sections 4 and 5 where SD and CN2-SD were used for subgroup discovery, an alternative approach is to present a population subgroup by listing only the supporting factors and the opposing factors. Listing of supporting factors in line with the approach presented in Section 4, where supporting factors are used to reinforce the medical expert's confidence in a subgroup, once constructed by a subgroup discovery algorithm.

6.1. Uncovering supporting factors for a single target concept

As in Section 5, the dataset investigated in this section is the relational database consisting of customers' answers about how they recognize, use and appreciate tested brands.

In this study, a population subgroup is presented by listing its supporting factors as well as its *opposing factors*, which follows the basic principles of Bayesian analysis (Berger, 1985). These factors are found in such a way that they respectively maximize or minimize the conditional probability of concept X under consideration; the opposing factors for X are the supporting factors for $\bar{X}$. Only the factors with statistical significance higher than 99% are selected as influential and are included in the listings.

This novel approach to subgroup discovery, appropriate for data analysis from uncertain data, was used in the decision support application designed to target an advertising campaign for a given natural non-alcoholic sparkling beverage brand (Cestnik et al., 2002). For confidentiality reasons it will be called brand X. More specifically, the task is to identify the characteristics of those consumers that do not yet recognize and/or use brand X.

Table 4. Induced subgroup descriptions through statistical characterizations (supporting factors) for the concept 'drinker of brand X' and the negation of this concept.

	Supporting factors
X	• The customers are from a central Slovenian region
	• Label 'monitored food' is neither important nor unimportant
	• They regularly read the *Daily News* newspaper and/or the *Youth* magazine
	• Their education degree is higher or equal to the university degree
$\bar{X}$	• Availability of a product in different quantities is not important
	• Product price is not so important

In our case, we first have the concept of 'user of brand X'. This group of customers can be characterized by the supporting factors listed in Table 4 for the target class denoted as X, while the non-users of brand X are listed for the class denoted by $\bar{X}$.

The marketing expert found such simple disjunctive descriptions very intuitive and easy to apply in practice, especially in the cases where the corresponding subgroup can be named with a suitable metaphor (see Section 7.1). It seems that such a disjunctive approach is particularly suitable in marketing (and possibly related domains), where the task is to increase the probability of a certain event (order, buy, reply) in a target population and not to accurately describe a portion of the target population.

6.2. Uncovering supporting factors for a combined target concept

In the above approach to extracting the supporting factors for a simple target group, the population was segmented into the consumers of brand X and those who do not yet recognize or use brand X. Based on the discussion with the marketing expert, it was decided that the target population should be further segmented to drinkers of other non-alcoholic sparkling beverages, and others that do not drink any beverages of this kind. The latter can be excluded from our target subgroup since it is fair to assume that they are not inclined to use the product generically. In other words, the aim of the campaign is to contact the users of competitive brands and present them with the qualities of product X. However, the marketing expert in our team emphasized the existence of one particular brand (named brand Y) that was so firmly positioned in the market that it was wise to exclude the users of this brand Y from the target population. In fact, it was reasonable to expect that the regular drinkers of brand Y were very unlikely to change their prevalent behavior no matter what the arguments in favor of brand X were presented to them.

Let us restate our target population. The combined target concept consists of people who do not yet know or use brand X, but drink other non-alcoholic sparkling beverages, with the exception of those who regularly drink brand Y. One approach to describe this target population is to use the combination of the supporting factors describing the above three concepts individually. Thus, in addition to learning the concept 'non-user of brand X', we applied the same approach to supporting factors discovery for the concepts 'user of non-alcoholic sparkling beverage brands', and 'non-user of brand Y'. For each of these

concepts, the supporting factors were computed and used to describe both the target concept and its converse.

Alternatively, we can also directly find the supporting factors for the combined concept: 'non-user of brand *X* & user of non-alcoholic sparkling beverage brands & non-user of brand *Y*'. This is only possible due to the data preprocessing phase in which we have learned labels for the missing concepts. Note that in the original dataset the concepts 'non-user of brand *X*' could only be determined for one out of 20 respondents who were asked this question. The same holds for two other concepts: 'non-user of brand *Y*' and 'user of non-alcoholic sparkling beverage brands'. Combining two sparse concepts with the logical operator & would result in only one customer out of 400 to be used for the analysis. Therefore, in order to combine several different concepts it has been necessary to augment the concepts to all respondents, which was done by filling in the missing values, using the probabilistic classification rule learning algorithm CN2, as mentioned in Section 5.2.

After data preprocessing we were able to directly find supporting factors for the combined target concept *non-user of brand X & user of non-alcoholic sparkling beverage brands & non-user of brand Y* (concept *Com*). The factors describing the customers that belong to the combined concept are listed in Table 5, which also lists the supporting factors that speak against the combined concept (concept $\overline{Com}$).

One important observation to be made is that the supporting factors of the combined target concept are not necessarily part of the three basic concept descriptions. For example, the factor of reading more than 4 newspapers did not appear in any of the basic concept descriptions. However, there are also some factors that can be traced from the combined concept to the basic ones. For instance, the consumers from the central Slovenian region tend to be excluded from the combined concept, because they tend to be more than average consumers of brand *X*.

Table 5. Induced subgroup descriptions through statistical characterizations (supporting factors) for the combined concept and the negation of the combined concept.

	Supporting factors
Com	• Availability of a product in different quantities is not important
	• Good commercials are not important
	• Different tastes of a product are not important
	• The good name of a product is not important
	• Popularity of a product is not important
	• They regularly read *Evening News*
$\overline{Com}$	• Good commercials are important
	• They read *Daily News*, *Sunday News*, *Youth* and/or *Our Home*
	• The good name of a product is important
	• They regularly read more than 4 newspapers
	• They are from the central Slovenian region
	• Their education level is higher or equal to a university degree

7. Lessons learned from subgroup discovery applications

Lessons learned from the applications indicate that subjective measures of interestingness such as the ability to trigger metaphoric descriptions of subgroups and subgroup actionability are very important when deciding which subgroups to select. We have also learned many lessons concerning questionnaire design and data preprocessing needed for appropriately defining the target concept.

7.1. Supporting factors and metaphoric descriptions of subgroups

In addition to confirming the appropriateness of if-then rules as a subgroup representation formalism, the following lessons were learned.

Lesson 1: Provide sufficient supporting evidence. One of the lessons learned in the medical application is that the final subgroup description is primarily based on the features defining the subgroup: the so-called *principal factors*. However, equally important are the *supporting factors*, uncovered by statistical analysis as having significantly different values for instances in the subgroup in comparison with the set of negative instances. The inclusion of supporting factors into the induced description is important for its actionability, enabling easier recognition of target cases, and providing for some redundant information that supports the classification. This lesson has triggered the development of the subgroup discovery approach appropriate for domains with high uncertainty (outlined in Section 6) where subgroups are described only by the supporting and opposing factors.

Lesson 2: Metaphoric descriptions provide crucial leverage. When describing subgroups with a set of supporting factors it is important to be able to substitute a set of factors with a proper metaphor. For example, the first five factors in Table 5 describing the target population (the combined concept *Com*) can be, according to the marketing expert, formulated as *store-brand consumers*. Store-brand consumers do not buy established popular brands. They settle for no-brand products that are usually sold under the store brand name, sold in simple packaging and offer good quality for a reasonable price. Such consumers can be addressed by low-profile advertising. According to the marketing expert the discovery of this piece of knowledge is substantial for the marketing analyst when planning and directing a marketing campaign.

The above use of metaphoric knowledge demonstrates how, with the use of background expert knowledge, non-actionable descriptors can turn into an actionable or even operational 'chunk' of knowledge that can be used in decision support. The description of the combined concept, which at first glance seemed useless, gained considerable value when represented by a single metaphor.

7.2. Subjective measures of interestingness and actionability of induced subgroup descriptions

Lessons learned in the subgroup discovery applications concern also increased awareness of the importance of *subjective* measures of interestingness. For automated rule induction only the objective quality criteria apply. However, for evaluating the usefulness of induced

subgroup descriptions for decision support, the subjective criteria are more important, but also harder to evaluate.

Lesson 3: Subgroup descriptions should be actionable. In the medical problem of detection and description of coronary heart disease risk groups we have learned that in this type of problem, there are no predefined specificity or sensitivity rule quality levels to be satisfied. The actionability of induced subgroup descriptions depends mainly on (a) whether the attributes used in the induced rules can be easily and reliably measured, and (b) how interesting/unexpected the subgroup descriptions in the given population are. Evaluation of such properties is completely based on expert knowledge and the success of the search depends on the expert's involvement, while the aim of machine learning based subgroup detection is to enable the domain expert to effectively search the hypothesis space, ranging from very specific to very general rules.

Lesson 4: Operational knowledge can operate on the target population. In addition to the subjective measures of interestingness introduced by other authors (usefulness, actionability, unexpectedness and redundancy), we have proposed another measure called *operationality*. If an operational rule is effectively executed, the performed operation can change the rule coverage.

We have discussed the notions of actionability and operationality on several examples and pointed out the importance of distinguishing between the two. For instance, the supporting factors induced in the application of targeting potential clients of a natural non-alcoholic sparkling drink differ in how actionable or operational they really are. If the description includes readers of a specific newspaper, it is actionable and not operational, since the information can be used by the decision maker for targeting whereas there is not much that she can do about the target audience of the newspaper. Also, if one of the characteristics of the target population is that people in the target group do not value good commercials, one cannot reach them by making bad commercials. On the other hand, if they think that healthy food is important or that attractive product packaging is important, one can address their need by stating the healthy ingredients of the product or by improving its packaging, which may lead to increased coverage of the target group.

In general, operationality of induced descriptions is harder to achieve than actionability. If, for example, the learned concept includes customers of a certain age and living in a certain area, those are the attributes that cannot be manipulated. The only thing one can do is to take them into account when targeting the commercial message. The induced concept description is, therefore, actionable but not operational. Alternatively, if the learned concept includes customers that were sent promotional material, then the induced description is operational because we can actively increase the coverage of the subgroup by sending out some additional catalogs.

7.3. Questionnaire design and data preprocessing

In questionnaire design for the two marketing applications we dealt with three issues: avoiding questionnaire fatigue, measuring recognition of brand names with words and pictures, and validating the brand name recognition.

Lesson 5: Avoid questionnaire fatigue. The most important lesson learned in questionnaire design was how to avoid questionnaire fatigue which would have occurred if a respondent had to evaluate each of the 300 brand names. A remedy was to randomly select just 15 brand names for evaluation, but this resulted in sparse data. In further data processing, this sparseness was overcome by missing value imputation through probabilistic value prediction by a classification rule learning algorithm.

Lesson 6: Target class definitions can be complex. Another lesson deals with the choice of the target class. While in the medical application the choice was intuitive (the positive class being the target), the target class definition was less obvious in the two marketing applications. In the direct mailing application, the label 'does not use the selected yoghurt brand' has the role of the target class labeled positive, and 'uses/knows the yoghurt brand' is the negative class.

In the campaign of targeting potential clients of a natural non-alcoholic sparkling drink, the target class is more complex, consisting of people who do not use or know of the brand, but who do drink other non-alcoholic sparkling drinks, except those people who are regular drinkers of a world-famous brand. Why should consumers of world-famous brands be excluded from the target? According to the marketing expert, these consumers are very unlikely to change their habits; therefore it makes no sense to direct a campaign at these consumers. Moreover, in the discussion with the marketing expert it became clear that the negative class should not be formed of all the other consumers. Limiting the population to non-alcohol drinkers makes more sense for uncovering specific properties of the target population. If, for example, alcohol drinkers were included in the class negative, the subtle differences between people who do not use brand X, but do drink other non-alcoholic drinks would be lost in much stronger regularities discriminating non-alcohol drinkers to those drinking alcohol drinks. Note that even subtler properties could be uncovered if the entire population were limited to consumers of non-alcoholic dark-colored sparkling drinks, since the color of the analyzed brand is dark.

8. Lessons learned from the development of subgroup discovery algorithms

We argue that learners that induce discriminant descriptions are inappropriate for subgroup discovery. Next, despite the fact that rule learners induce characteristic descriptions, we show that their use for subgroup discovery is hindered if rule learning is performed within the covering algorithm for ruleset construction. Moreover, we show the appropriateness of the rule unusualness heuristic (*WRAcc*) and the ROC methodology for subgroup discovery.

Lesson 7: Distinguish between discriminant and characteristic descriptions. In symbolic predictive induction, the two most common approaches are rule learning and decision tree learning. Let us first show that classification rules serve two different purposes: characterization and discrimination.

The usual goal of classification rule learning is to generate separate models, one for each class, inducing class characteristics in terms of properties occurring in the descriptions of training examples. Therefore, classification rule learning results in *characteristic*

descriptions, generated separately for each class by repeatedly applying the covering algorithm.

In decision tree learning, on the other hand, the rules which can be formed from paths leading from the root node to class labels in the leaves represent *discriminant descriptions*, formed from properties that best discriminate between the classes. As rules formed from decision tree paths form discriminant descriptions, they are inappropriate for solving subgroup discovery tasks which aim at describing subgroups by their characteristic properties.

8.1. Inappropriateness of the standard covering algorithm

The reasons for the inappropriateness of classification rules for subgroup discovery, which are due to the induction methods used, are listed below. These shortcomings of classification rule learning are illustrated on a 'classical' and well-known rule learner CN2, which we have analysed and upgraded to a subgroup discovery algorithm CN2-SD, outlined in Section 5.3.

Lesson 8: Standard covering algorithm is biased. Classification rules forming characteristic descriptions are expected to be appropriate for subgroup discovery. However, the fact that they have been generated by a covering algorithm (used in AQ (Michalski et al., 1986), CN2 (Clark & Niblett, 1989; Clark & Boswell, 1991), and most other rule learners) hinders their usefulness for subgroup discovery. Only the first few rules induced by a covering algorithm may be of interest as subgroup descriptions with sufficient coverage. Subsequent rules are induced from smaller and strongly biased example subsets, excluding the positive examples covered by previously induced rules. This bias prevents the covering algorithm from inducing descriptions uncovering significant subgroup properties of the entire population.

In CN2, the induced rules can be ordered or unordered. Ordered rules are interpreted as a decision list (Rivest, 1987) in a straight-forward manner: when classifying a new example, the rules are sequentially tried and the first rule that covers the example is used for prediction. Ordered rules may be very appropriate for classification, but an individual rule does not represent a separate 'chunk' of knowledge about the problem, which makes decision lists inappropriate for discovering interesting subgroup properties of the entire population. The ordering problem is solved in CN2 by its ability of inducing unordered rulesets, where rules can be interpreted individually, however, when used for classification even unordered rules can not be used separately from each other without losing information. Other problems that are due to the fact that rules were induced by the covering algorithm also remain, including a large number of rules, low coverage and low significance (see the discussion on the significance of CN2 rules in Section 8.2).

Lesson 9: Weighted covering algorithm is advantageous. The above-mentioned shortcomings of classification rule learning algorithms for subgroup discovery are overcome by the subgroup discovery algorithms SD and CN2-SD outlined in this paper. One of their features is the use of a weighted covering algorithm, where subsequently induced rules with high coverage allow for discovering interesting subgroup properties of the entire population. In addition to improved coverage, comprehensibility, compactness and significance of

rules, the advantage of subgroup discovery is also its ability of handling skewed distributions (empirical evidence for this claim is provided in Lavrač et al. (2004)).

8.2. Objective quality measures and ROC analysis

While the subjective interesting measures are crucial for the expert evaluation of induced subgroup descriptions, objective quality measures are crucial for automated rule induction and the comparative evaluation of different subgroup discovery algorithms. Classification rule learners consider rule accuracy/precision to be one of the most important evaluation measures; on the other hand, we consider rule *significance* and *unusualness* to be the most important quality measures for subgroup discovery.

Lesson 10: Low coverage, support and significance of CN2 rules. In one variant of the CN2 rule learner, the search procedure used in single rule learning performs beam search employing rule accuracy/precision $p(Class \mid Cond)$ as a heuristic function. Since the main goal of this heuristic is accuracy optimization, the heuristic leads to the induction of specific rules, consisting of conjunctions of many features. Consequently, induced rules have low coverage. Accurate rules may be very useful for classification, but an individual rule does not represent a separate 'chunk' of knowledge about the problem, due to low coverage and low support.

A variant of CN2 guarantees that induced rules are significant by employing the significance stopping criterion in single rule construction. Another variant of CN2 uses significance as a search heuristic. Empirical evaluation in Clark and Boswell (1991) shows that applying a significance test reduces the number of induced rules, while slightly decreasing the predictive accuracy. If the significance test is used within the covering algorithm, only the first few rules induced by a covering algorithm will be subgroup descriptions with sufficient significance w.r.t. the entire population. Since subsequent rules are induced from smaller and strongly biased example subsets, these are significant for subpopulations, but tend to be less significant or even insignificant w.r.t. the entire population.

Lesson 11: Appropriateness of WRAcc as a rule quality measure. It is generally accepted that objective quality measures can be successfully used for rule selection and evaluation in the context of the ROC methodology. Rule unusualness, measured by weighted relative accuracy *WRAcc*, and rule significance both measure the distributional unusualness of a subgroup. It was shown in Section 3.3 that *WRAcc* iso-performance lines are parallel to the ROC diagonal, and that, given a fixed class distribution, *WRAcc* is proportional to the vertical distance to the diagonal. As such, *WRAcc* is appropriate for measuring the unusualness of separate subgroups. Moreover, *WRAcc* also reflects rule significance: the larger *WRAcc* is, the more significant the rule is, and vice versa. However, while significance only measures distributional unusualness, computed in terms of correctly classified covered examples of all classes, *WRAcc* takes explicitly the rule coverage into account, therefore we consider *unusualness* to be the most appropriate measure for subgroup quality evaluation. As such, the *WRAcc* heuristic can be used in the search for optimal subgroups, and as a measure for evaluating the quality of induced subgroup descriptions.

Lesson 12: The final decision should be made by the expert. Although in objective terms the best subgroups are those on the ROC convex hull, the expert may decide to choose suboptimal subgroups, as they are more interesting according to some subjective measure of interestingness. ROC analysis and search for optimal subgroups in the ROC space is of ultimate importance for automated subgroup discovery. However, in expert-guided subgroup discovery for subgroups with high CHD risk in Section 4, ROC analysis was used for expert-guided search of the space of interesting subgroups, and for the evaluation of the expert-selected subgroups in comparison with the best induced subgroups forming the ROC convex hull.

9. Summary and further work

This paper presents three approaches to subgroup discovery and their application to a medical problem and two marketing problems.

An important aspect, to which we have devoted ample attention, was the development of evaluation criteria for measuring the success of subgroup discovery. To this end, much attention was devoted to the *objective* measures of quality of induced patterns, used both in the process of induction as heuristics for guiding the search for patterns, as well as in the evaluation of induced patterns in ROC space. One of the heuristics appropriate for subgroup discovery is the weighted relative accuracy heuristic used in this work which trades off the generality of a rule ($p(Cond)$, i.e., rule coverage) and its relative accuracy ($p(Class \mid Cond) - p(Class)$). Various similar rule evaluation measures and heuristics have been studied for subgroup discovery by Kloesgen (1996) and Wrobel (1997), 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 (the so-called 'p-g-space') have been extensively studied (Kloesgen, 1996).

Besides the *objective* measures of subgroup quality, this paper discusses also numerous *subjective* measures of interestingness. We have specifically addressed pattern actionability and operationality, while other interestingness measures evaluated by the experts and used in the process of expert-guided subgroup discovery were discussed in less detail.

The paper conveys the lessons learned from the applications of subgroup discovery for decision support in solving real-life problems, and the lessons learned in the development of recent subgroup discovery algorithms. Despite the fact that some of the described shortcomings of classification rule learners are specific to CN2, similar shortcomings could be given for other classification rule learners, even if we chose to analyse RIPPER (Cohen, 1995) or some other more sophisticated classification rule learner whose goal is to maximize the classification accuracy. In future work, we plan to compare our subgroup discovery approaches with SLIPPER (Cohen & Singer, 1999), a successor of RIPPER, which may turn out to be of interest for subgroup discovery due to its similarity with the CN2-SD algorithm (the most important difference being the way examples are reweighted).

We also plan to develop other subgroup discovery approaches, including the approaches that adapt standard association rule learning to subgroup discovery. In association rule learning (Agrawal et al., 1996), each rule is an individual pattern, describing an individual 'chunk' of knowledge, equipped with two standard quality measures:

support and confidence. As shown in Kavšek, Lavrač, and Jovanoski (2003), association rule learning can naturally be adapted to subgroup discovery. Results of subgroup discovery algorithm APRIORI-SD are similar to those of subgroup discovery algorithm CN2-SD (Lavrač et al., 2004), while experimental comparisons with CN2, RIPPER and APRIORI-C demonstrate that subgroup discovery algorithms CN2-SD and APRIORI-SD produce substantially smaller rulesets, where individual rules have higher coverage and significance.

Another line of development will be the incorporation of subgroup discovery approaches into the inductive database and constraint-based data mining framework. In this approach, heuristic search for subgroup descriptions will be replaced by complete search within the given language and quality constraints. Special attention will be devoted also to the development of novel applications of subgroup discovery, with the emphasis on the experiments demonstrating the success of subgroup discovery in terms of various subjective measures of interestingness such as operationality, actionability, unexpectedness, redundancy, novelty and usefulness.

Acknowledgments

This work was supported by the Slovenian Ministry of Education, Science and Sport, the Croatian Ministry of Science, Education and Sport, and the European project Data Mining and Decision Support for Business Competitiveness: A European Virtual Enterprise (IST-1999-11495). We are grateful to Mihael Kline for his collaboration in marketing domains, and to Goran Krstačić for his collaboration in the coronary heart disease risk group detection experiments. Thanks are also due to three anonymous reviewers for their detailed and insightful comments, which helped us to improve the paper.

Notes

1. For a binary classification problem, ruleset accuracy is computed as $\frac{TP+TN}{N}$.
2. Note that although for each generated subgroup description one class is selected as the target class, the significance criterion measures the distributional unusualness unbiased to any particular class; as such, it measures the significance of the rule condition only. In two-class problems this statistic is distributed approximately as χ^2 with one degree of freedom.
3. This can be derived as follows:

$$\begin{aligned} WRAcc &= p(Cond) \cdot [p(Class \mid Cond) - p(Class)] = p(Class \cdot Cond) - p(Class) \cdot p(Cond) \\ &= p(Class \cdot Cond) - p(Class) \cdot [p(Class \cdot Cond) + p(\overline{Class} \cdot Cond)] \\ &= (1 - p(Class)) \cdot p(Class \cdot Cond) - p(Class) \cdot p(\overline{Class} \cdot Cond) \\ &= p(\overline{Class}) \cdot p(Class) \cdot p(Cond \mid Class) - p(Class) \cdot p(\overline{Class}) \cdot p(Cond \mid \overline{Class}) \end{aligned}$$

4. Actually, the SD algorithm looks for the best subgroups in the TP/FP space, equivalent to the TPr/FPr space, as it performs heuristic search using the q heuristic which takes into account TP and FP, instead of TPr and FPr, respectively.

References

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., & Verkamo, A. I. (1996). Fast discovery of association rules. In *Advances in knowledge discovery and data mining*. Menlo Park, CA: AAAI Press.

Berger, J. (1985). *Statistical decision theory and bayesian analysis*. Springer-Verlag.

Berry, M., & Linoff, G. (2000). *Mastering data mining, the art and science of customer relationship managemen*. John Wiley.

Cestnik, B., Lavrač, N., Železny, F., Gamberger, D., Todorovski, L., & Kline, M. (2002). Data mining for decision support in marketing: A case study in targeting a marketing campaign. In *Proceedings of the ECML/PKDD-2002 Workshop on Integration and Collaboration Aspects of Data Mining, Decision Support and Meta-Learning* (pp. 25–34).

Clark, P., & Boswell, R. (1991). Rule induction with CN2: Some recent improvements. In *Proc. Fifth European Working Session on Learning* (pp. 151–163). Springer.

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

Cohen, W. W. (1995). Fast effective rule induction. In A. Prieditis & S. Russell (Eds.), *Proc. of the 12th International Conference on Machine Learning* (pp. 115–123). Morgan Kaufmann.

Cohen, W. W., & Singer, Y. (1999). A simple, fast, and effective rule learner. In *Proceedings of the 17th National Conference on Artificial Intelligence*. American Association for Artificial Intelligence.

De Raedt, L., Blockeel, H., Dehaspe, L., & Laer, W. V. (2001). Three companions for data mining in first order logic. In S. Džeroski & N. Lavrač (Eds.), *Relational Data Mining*. Springer-Verlag.

De Raedt, L., & Dehaspe, L. (1997). Clausal discovery. *Machine Learning, 26*, 99–146.

Flach, P. (2003). The geometry of ROC space: Understanding machine learning metrics through ROC isometrics. In *Proc. 20th International Conference on Machine Learning (ICML03)* (pp. 194–201). AAAI Press.

Flach, P., & Gamberger, D. (2001). Subgroup evaluation and decision support for direct mailing problem. In *Proceedings of the ECML/PKDD-2001 Workshop on Integration Aspects of Data Mining, Decision Support and Meta-Learning* (pp. 45–56).

Fürnkranz, J., & Flach, P. (2003). An analysis of rule evaluation metrics. In *Proc. 20th International Conference on Machine Learning (ICML03)* (pp. 202–209). AAAI Press.

Gamberger, D., & Lavrač, N. (2002). Expert guided subgroup discovery: Methodology and application. *Journal of Artificiel Intelligence Research, 17*, 501–527.

Gamberger, D., Lavrač, N., & Krstačić, G. (2003). Active subgroup mining: A case study in coronary heart disease risk group detection. *Artificial Intelligence in Medicine, 28*, 27–57.

Holte, R. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine Learning, 11*, 63–91.

Kavšek, B., Lavrač, N., & Jovanoski, V. (2003). APROPRI-SD: Adapting association rule learning to subgroup discovery. In M. Berthold, H. J. Lenz, E. Bradley, R. Kruse, & C. Borgelt (Eds.), *Advances in intelligent data analysis* (pp. 230–241). Springer-Verlag.

Kloesgen, W. (1996). EXPLORA: A multipattern and multistrategy discovery assistant. In M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy (Eds.), *Advances in knowledge discovery and data mining*. Menlo Park, CA: AAAI Press.

Lavrač, N., Flach, P., Kavšek, B., & Todorovski, L. (2002). Adapting classification rule induction to subgroup discovery. In V. Kumar, S. Tsumoto, N. Zhong, P. Yu, & X. Wu (Eds.), *Proceedings of the 2002 IEEE International Conference on Data Mining* (pp. 266–273). IEEE Computer Society.

Lavrač, N., Flach, P., & Zupan, B. (1999). Rule evaluation measures: A unifying view. In S. Džeroski & P. Flach (Eds.), *Proceedings of the 9th International Workshop on Inductive Logic Programming* (pp. 174–185). Springer-Verlag.

Lavrač, N., Kavšek, B., Flach, P., & Todorovski, L. (2004). Subgroup discovery with CN2-SD. *Journal of Machine Learning Research, 5*, 153–188.

Michalski, R., Mozetič, I., Hong, J., & Lavrač, N. (1986). The multi-purpose incremental learning system AQ15 and its testing application on three medical domains. In *Proc. 5th National Conference on Artificial Intelligence* (pp. 1041–1045). Morgan Kaufmann.

Myers, J. (1996). *Segmentation and positioning for strategic marketing decisions*. American Marketing Association.

Piatetsky-Shapiro, G., & Matheus, C. (1994). The interestingness of deviation. In *Proceedings of the AAAI-94 Workshop on Knowledge Discovery in Databases* (pp. 25–36).

Provost, F. J., & Fawcett, T. (1998). Robust classification systems for imprecise environments. In *Proceedings of the 19th National Conference on Artificial Intelligence* (pp. 706–713).

Rivest, R. L. (1987). Learning decision lists. *Machine Learning, 2:3*, 229–246.

Silberschatz, A., & Tuzhilin, A. (1995). On subjective measures of interestingness in knowledge discovery. In *Knowledge Discovery and data mining* (pp. 275–281).

Wrobel, S. (1997). An algorithm for multi-relational discovery of subgroups. In J. Komorowski & J. Zytkow (Eds.), *Proc. First European Symposion on Principles of Data Mining and Knowledge Discovery (PKDD-97)* (pp. 78–87). Springer Verlag.

Wrobel, S. (2001). Inductive logic programming for knowledge discovery in databases. In S. Džeroski & N. Lavrač (Eds.), *Relational data mining*. Springer-Verlag.

Wrobel, S., & Džeroski, S. (1995). The ILP description learning problem: Towards a general model-level definition of data mining in ILP. In K. Morik & J. Herrmann (Eds.), *Proc. Fachgruppentreffen Maschinelles Lernen (FGML-95)*. 44221 Dortmund, Univ. Dortmund.

Received April 26, 2003
Accepted April 8, 2004
Final manuscript April 8, 2004