Appl Intell (2007) 27: 205–217
DOI 10.1007/s10489-007-0068-9

Clinical data analysis based on iterative subgroup discovery: experiments in brain ischaemia data analysis

Dragan Gamberger · Nada Lavrač · Antonija Krstačić · Goran Krstačić

Published online: 16 June 2007

Abstract This paper presents a case study of the process of insightful analysis of clinical data collected in regular hospital practice. The approach is applied to a database describing patients suffering from brain ischaemia, either permanent as brain stroke with positive computer tomography (CT) or reversible ischaemia with normal brain CT test. The goal of the analysis is the extraction of useful knowledge that can help in diagnosis, prevention and better understanding of the vascular brain disease. This paper demonstrates the applicability of subgroup discovery for insightful data analysis and describes the expert's process of converting the induced rules into useful medical knowledge. Detection of coexisting risk factors, selection of relevant discriminative points for numerical descriptors, as well as the detection and description of characteristic patient subpopulations are important results of the analysis. Graphical representation is extensively used to illustrate the detected dependencies in the available clinical data.

This work was supported by Croatian Ministry of Science, Education and Sport project "Machine Learning Algorithms and Applications", Slovenian Ministry of Higher Education, Science and Technology project "Knowledge Technologies", and EU FP6 project "Heartfaid: A knowledge based platform of services for supporting medical–clinical management of the heart failure within the elderly population".

D. Gamberger (✉)
Rudjer Bošković Institute, Bijenička 54, 10000 Zagreb, Croatia
e-mail: dragan.gamberger@irb.hr

N. Lavrač
Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia

N. Lavrač
University of Nova Gorica, Vipavska 13, 5000 Nova Gorica, Slovenia

A. Krstačić
Department of Neurology, University Hospital of Traumatology, Draškovićeva 19, 10000 Zagreb, Croatia

G. Krstačić
Institute for Cardiovascular Diseases and Rehabilitation, Draškovićeva 13, 10000 Zagreb, Croatia

1 Introduction

Data analysis in medical applications is characterized by ambitious goals of extracting relevant, general and potentially new relations from data collected in regular medical practice, surveys or epidemiological studies. These are not simple tasks because medical manifestations are results of exceptionally complex processes in the human body and collected data sets in the best cases contain information about a restricted part of a population, typically using very indirect descriptors to present the state of human health. An even more serious problem is the patients' follow-up in time and the detection of time dependent characteristics. An important reason for the lack of representatively general data is the time complexity and high cost of their collection. Medical ethical reasons and the fact that healthy people are not willing to take part in medical data collection actions also hinder the applicability of data analysis approaches in medicine. On the other hand, a positive aspect is the availability of a very large corpus of medical expert knowledge. Active involvement of medical experts is extremely important for the extraction of useful knowledge from medical data, as experts can explicitly add valuable information missing in the data and give meaningful interpretations of the results of analyses. To make experts' cooperation possible, results obtained by data analysis methodology should be at every level, in-

cluding rules and graphical representations, in a form that is easily interpretable by humans.

Insightful data analysis is a process aimed at the detection and presentation of relevant relations that exist in available data sets [1]. Applications of quantitative statistical methods seldom lead to insightful results, leaving a large workload on human experts who have to provide appropriate interpretations of results, with no guarantees that—due to a huge search space of possible hypotheses—the most relevant combinations will be tested at all [2]. The goal of intelligent data analysis is to effectively detect the most relevant dependencies in an explicit qualitative form and to enable that quantitative analysis and human expert interpretation can concentrate on a relatively small set of potentially relevant hypotheses. This approach is specially well suited for medical data analysis, as large amounts of available medical expert knowledge allow for appropriate interpretation of detected relations.

This work demonstrates that rules induced by the methodology of supervised subgroup discovery [3] can serve as an appropriate basis for data analysis, if supplemented by a sufficiently large intellectual effort of medical experts, willing to convert machine-induced rules into adequate medical interpretations. The novelty of the work is in the approach which systematically generates subgroup descriptions of different generality and then demonstrates how very general rules can be used to describe global properties of the target population, how very specific rules can be used to detect relevant co-existing risk factors, how rules at a medium level of generality and specificity can be useful to determine discriminative points for numerical descriptors, and how they can be used to detect and describe relevant subpopulations of the target concept. In contrast to the complete automation of the first part of the data analysis process which generates rules describing subgroups of different generality, the second part of the data analysis process strongly depends on the expected goals of the data analysis process, existing domain knowledge, and the rules induced by the first part of the process. The aim of this paper is to provide guidelines for this process and to illustrate different options existing in this part of the analysis. The analysis process, the experts' reasoning process, and the achieved results are presented for a typical clinical database collected in regular hospital practice.

This paper is organized as follows. Section 2 presents the available brain ischaemia dataset. The constraint-based subgroup discovery methodology used for the induction of subgroup descriptions is presented in Sect. 3. The central part of the paper is Sect. 4 which presents the results of the iterative application of the subgroup discovery methodology, followed by medical specialists' interpretation of generated rules, the statistical evaluation of detected relevant sub-populations, the visualization of obtained results, and a discussion concerning the reasoning process.

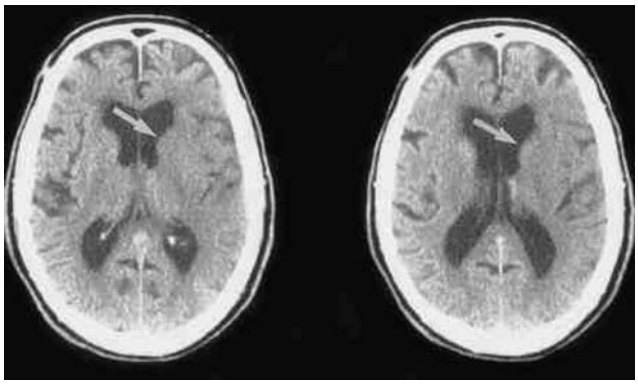

Fig. 1 CT brain scans of a patient with brain stroke caused by the occlusion of the left middle cerebral artery (MCA). The figure shows changes in the brain in two cross sections one hour after brain stroke, characterized by the loss of the bounds between the white and the grey matter of the brain

Table 1 The first part of descriptors in the brain ischaemia domain with abbreviations used in the induced rules. The last column provides reference values representing the ranges typically accepted as normal in the medical practice

Descriptor	Abbreviation	Characteristics
sex	sex	m, f
age	age	continuous (years)
family anamnesis	fhis	positive, negative
present smoking	smok	yes, no
stress	str	yes, no
alcohol consumption	alcoh	yes, no
body mass index	bmi	continuous (kg m^{-2})
		ref. value 18.5–25
systolic blood pressure	sys	continuous (mmHg)
		normal value <139 mmHg
diastolic blood pressure	dya	continuous (mmHg)
		normal value <89 mmHg
uric acid	ua	continuous (μmol L^{-1})
		ref. value for men <412
		ref. value for women <380
fibrinogen	fibr	continuous (g L^{-1})
		ref. value 2.0–3.7
glucose	gluc	continuous (mmol L^{-1})
		ref. value 3.6–5.8
total cholesterol	chol	continuous (mmol L^{-1})
		ref. value 3.6–5.0
trygliceride	tryg	continuous (mmol L^{-1})
		ref. value 0.9–1.7

2 Brain ischaemia data

The brain ischaemia database consists of records of patients who were treated at the Intensive Care Unit of the Department of Neurology, University Hospital Center "Zagreb", in

Table 2 The second part of descriptors in the brain ischaemia domain

Descriptor	Abbreviation	Characteristics
heart rate	ecgfr	continuous
		ref. value 60–100 beats/min
platelets	plat	continuous
		ref. value 150000–400000
protrombin time	pt	continuous
		ref. value without th. 0.7–1.2
		with anticoagulant th. 0.25–0.4
atrial fibrillation	af	yes, no
left ventricular hypertrophy	ecghlv	yes, no
fundus ocular	fo	discrete value 0–4
aspirin therapy	asp	yes, no
anticoagulant therapy	acoag	yes, no
antihypertensive therapy	ahyp	yes, no
antiarrhytmic therapy	aarrh	yes, no
statins (antihyperlipoproteinaemic t.)	stat	yes, no
hypoglycemic therapy	hypo	none, yesO (oral), yesI (insulin)

Zagreb, Croatia, in year 2003. In total, 300 patients are included in the database: 209 with the computed tomography (CT) confirmed diagnosis of brain attack (stroke), and 91 patients who entered the same hospital department with adequate neurological symptoms and disorders, but were diagnosed (based on the outcomes of neurological tests and CT) as patients with transition ischaemic brain attack (TIA, 33 patients), reversible ischaemic neurological deficit (RIND, 12 patients), and serious headache or cervical spine syndrome (46 patients). In this paper, the goal of data analysis experiments is to discover regularities that characterize brain stroke patients.

Patients are described with 26 different descriptors representing anamnestic data, physical examination data, laboratory test data, ECG data, and information about previous medical therapies [4, 5]. Descriptors used in the rules induced in the experiments presented in this work, including abbreviations used in these rules, are listed in Tables 1 and 2. All the patients in the control group have normal brain CT in contrast with the positive CT test result for patients with a confirmed brain attack. Figure 1 presents CT scans of a patient with stroke caused by the occlusion of the left middle cerebral artery (MCA). Scans show loss of the bounds between the white and the grey matter of the brain. Additionally, hypodensity of the brain tissue can be noticed due to the occluded supply to basal ganglia.

It should be noted that the target class are the patients with brain stroke and the control group does not consist of healthy persons but of patients with suspected serious neurological symptoms and disorders. In this sense, the available database is particularly appropriate for studying the specific characteristics and subtle differences that distinguish patients with stroke. The detected relationships can be accepted as the actual characteristics for these patients. However, the computed evaluation measures—including probability, specificity and sensitivity of induced rules—only reflect characteristics specific to the available data, not necessarily holding for the general population or other medical institutions [6].

3 The methodology of insightful data analysis

The presented approach to insightful data analysis is based on the application of supervised learning, intended at finding descriptions of the target (positive) class cases (in this domain brain stroke cases) in contrast to cases in the non-target (negative or control) class (in this domain TIA, RIND and problems with cervical spine or headache cases). This means that examples of two classes have to be available for the analysis. Sometimes the decision about what is the target class is not simple and the complete data analysis process can have a few task definitions with different choices of target and non-target classes. For example, in the same brain ischaemia domain the target class could be also patients with stroke taking some therapy, and the non-target class being stroke patients not taking the therapy. In this setting, the process of data analysis is far from completely automatic. Moreover, the process should be sometimes repeated for different subpopulations with specific properties, like sex or age range, or with different subsets of descriptors. In this work we demonstrate only the process performed for the complete database with patients who experienced stroke selected as the target class. The same approach may be repeated for differently defined problems, potentially leading

to other relevant results. Selection and definition of subproblems that have to be analysed completely depends on medical expert suggestions.

Subgroup mining is a form of supervised inductive learning, resulting in induced patterns that describe subgroups of the target class. As in all inductive rule learning tasks, the language bias is determined by the syntactic restrictions of the pattern language and the vocabulary of terms in the language. In this paper the hypothesis language is restricted to simple if-then rules of the form *Class* ← *Cond* (interpreted as if *Cond* then *Class*), where *Class* is the target class and *Cond* is a conjunction of features. Features are logical conditions that have values *true* or *false*, depending on the values of descriptors which describe the examples in the problem domain.

The goal of rule induction is to extract rules with optimal covering properties on the available example set. A rule with ideal covering properties should be *true* for all target class (positive) examples and *false* for all non-target class (negative) examples. Target class examples covered by rule R are called *true positives* $TP(R)$ and their number denoted by $p = |TP(R)|$, non-target class examples covered by the rule are called *false positives* $FP(R)$ and their number denoted by $n = |FP(R)|$, while all remaining non-target class examples not covered by the rule are called *true negatives* $TN(R)$ and their number denoted by $\overline{n} = N - n$. An ideal rule would be characterized by $p = |P|$ and $n = 0$, where P is the set of positive examples, N the set of negative examples, and $E = P \cup N$ is the entire set of training examples. The quality of induced rules is measured by two values: *sensitivity* which represents the proportion of positive cases correctly classified by R as true positives, $sens(R) = \frac{p}{P}$, and *specificity*, which represents the proportion of negative cases correctly classified by R as true negatives, $spec(R) = \frac{\overline{n}}{N}$.

3.1 Subgroup mining

In this work, subgroup mining is performed by the SD algorithm [7], an iterative beam search rule learner. The SD algorithm heuristically searches for rules R maximizing the $q_g(R)$ heuristic, defined as $q_g(R) = \frac{p}{n+g}$, where g is a *generalization parameter*. It was shown in [7] that if SD is used in the expert-guided mode, varying of the value of generalization parameter g will enable the expert to guide subgroup discovery towards the most interesting rules which have a significantly different distribution of covered positives and negatives, compared to the prior class distribution in the training set.

The SD algorithm can be best described in a constraint-based data mining framework, in which a formal definition of subgroup discovery involves a set of constraints that the induced subgroup descriptions have to satisfy. In the SD subgroup discovery algorithm the following constraints are used to formalize the SD constraint-based subgroup discovery task: language constraints (described in Sect. 3.2) and evaluation/optimization constraints (described in Sect. 3.3). A brief sketch of the SD rule learning algorithm is given in Sect. 3.4, while the DMS mechanism which ensures rule set diversity is presented in Sect. 3.5.

3.2 Language constraints

Features For discrete (categorical) attributes, features have the form *Attribute* = *value* or *Attribute* ≠ *value*, for continuous (numerical) attributes they have the form *Attribute* > *value* or *Attribute* ≤ *value*. Note that features can have values which are *true* and *false* only, that every feature has its logical complement (for feature f_1 being $A_1 = v_1$ its logical complement $\overline{f_1}$ is $A_1 \neq v_1$, for $A_2 > v_2$ its logical complement is $A_2 \leq v_2$), and that features are different from binary valued attributes because for every attribute at least two different features are constructed.

Let values v_{ix} $(x = 1, \ldots, k_{ip})$ denote the k_{ip} different values of attribute A_i that appear in the positive examples and w_{iy} $(y = 1, \ldots, k_{in})$ the k_{in} different values of A_i appearing in the negative examples. A set of features F, formed in the SD data preprocessing phase, is constructed as follows:

- For discrete attributes A_i, features of the form $A_i = v_{ix}$ and $A_i \neq w_{iy}$ are generated.
- For continuous attributes A_i, similar to [8], features of the form $A_i \leq (v_{ix} + w_{iy})/2$ are generated for all neighboring value pairs (v_{ix}, w_{iy}), and features $A_i > (v_{ix} + w_{iy})/2$ for all neighboring pairs (w_{iy}, v_{ix}).
- For integer valued attributes A_i, features are generated as if A_i were both discrete and continuous, resulting in features of four different forms: $A_i \leq (v_{ix} + w_{iy})/2$, $A_i > (v_{ix} + w_{iy})/2$, $A_i = v_{ix}$, and $A_i \neq w_{iy}$.

Rules Individual subgroup descriptions have the form of rules *Class* ← *Cond*, where *Class* is the property of interest (the target class), and *Cond* is a conjunction of features.

Rule length To simplify rule interpretation and increase rule actionability, subgroup discovery is aimed at finding short rules. This is formalized by a language constraint that every induced rule R has to satisfy: rule length (i.e., the number of features in *Cond*) has to be below a user-defined threshold: $length(R) \leq MaxRuleLength$. This is achieved by restricting the main repeat loop of the SD algorithm to *MaxRuleLength* iterations.

3.3 Evaluation/optimization constraints

Support To ensure that induced subgroups are sufficiently large, each induced rule R must have high support, i.e.,

$sup(R) \geq MinSup$, where $MinSup$ is a user-defined threshold, and $sup(R)$ is the relative frequency of correctly covered examples of the target class in example set E:

$$sup(R) = \frac{p}{|E|}.$$

Rule quality This constraints aims to ensure that the induced subgroups are highly significant (ensuring that the distribution of target class examples covered by the subgroup description will be statistically significantly different from their distribution in the original training set). This could be achieved in a straight-forward way by imposing a significance constraint on rules, e.g., by requiring that rule significance is above a user-defined threshold. Instead, in the SD subgroup discovery algorithm the following rule quality measure assuring rule significance, implemented as a heuristic in the rule construction process, is used:

$$q_g(R) = \frac{p}{n+g}. \quad (1)$$

High quality rules will cover relatively many target class examples and few non-target class examples. The number of tolerated non-target class cases, relative to the number of covered target class cases, is determined by parameter g. For low g ($g \leq 1$), induced rules will have high specificity (i.e., high proportion of correctly classified negative cases $\frac{n}{N}$, or equivalently, low false alarm rate) since covering of every single non-target class example is made relatively very 'expensive'. Typically, such rules will cover also a relatively small number of positive examples. On the other hand, by selecting a high g value, more general rules will be induced, covering many examples among which there can be also relatively many non-target class examples.

Rule relevancy Besides absolute relevancy, which is ensured by previously described constraint and which means that each rule should be true for many positive examples and false for many negative examples, each rule should be also relatively relevant. This is tested in the central part of the SD algorithm. This constraint means that rule R is acceptable only if there is no other rule R' in the beam such that $TP(R) \subseteq TP(R')$ and at the same time $TN(R) \subseteq TN(R')$. By applying this constraint, copies of the same rule that differ only in the order of features can be detected and eliminated from the beam. But even more importantly, this constraint ensures that only best features and their best combinations will enter the rule construction process, which is important for overfitting prevention [9].

3.4 The SD algorithm

The input to SD consists of a training set of examples $E = P \cup N$ and a set of features F constructed from the given example set (as outlined in Sect. 3.2). Parameters of the SD algorithm are g—generalization parameter, $MinSup$—minimal support for rule acceptance, $BeamWidth$—maximal number of rules in beam search of best rules, and $MaxRuleLength$—maximal rule complexity.[1] The output of the SD algorithm is a set of rules with good covering properties on the given example set, which is ensured by using rule quality heuristic $q_g(R)$ outlined in Sect. 3.3.

To construct subgroup describing rules for given $Class$, SD starts with rules with empty antecedents (if part) and the selected target class $Class$ as a consequent (then part). Consider the algorithm outlined in Table 3. Note that the empty antecedent of such a rule is satisfied by all examples in the training set, and not only those of selected $Class$. SD then progressively refines the antecedent by conjunctively adding features to the current rule condition. Consider a partially built rule. Rule conclusion $Class$ is fixed and there are some (possibly none) conditions in the rule antecedent. The SD algorithm now considers which feature to add to the rule condition. For example, if values 4.0, 1.0, and 2.0 for attribute A appear in the training set, according to the language constraints used for feature construction, outlined in Sect. 3.2, conditions $A \leq 1.5$, $A > 1.5$, $A \leq 3.0$, and $A > 3.0$ are among the candidates to be added to the rule antecedent.

While the set of features (attribute values for discrete attributes and intervals for continuous attributes) of the rule have been predetermined, the actual condition to be included in a partially built rule depends on the number of true positive and false positive examples covered by the refined rule and the heuristic estimate of rule quality $q_g(R) = \frac{p}{n+g}$.

3.5 Diversity of generated subgroups

If the SD algorithm is used in the expert-guided mode, varying of the value of generalization parameter g enables the expert to guide subgroup discovery towards the most interesting rules which have a significantly different distribution of covered positives and negatives, compared to the prior class distribution in the training set. However, in the experiments described in Sect. 4—in order to get a good insight into the available data—subgroups with best $q_g(R)$ values were systematically generated for different values of generalization parameter g, $g = \{5, 10, 20, 50, 100\}$. For each g, three ($MaxRules = 3$) different subgroup descriptions were induced. These rules have not been selected simply according to three best $q_g(R)$ values because such an approach would typically lead to similar—and because of that not very informative—subgroup descriptions. With the intention to

[1] In the experiments described in Sect. 4, values of these parameters were set as follows: $g = 5$–100, $MinSup = \frac{\sqrt{|P|}}{|E|}$, $BeamWidth = 20$, and $MaxRuleLength = 4$.

Table 3 Heuristic beam search rule construction algorithm SD

SD_algorithm(E, F, g, *MinSup*, *BeamWidth*, *MaxRuleLength*)

Input: set of examples $E = P \cup N$, set of features F
Parameters: g, *MinSup*, *BeamWidth*, *MaxRuleLength*
for each *Rule* in *Beam* and *NewBeam* **do**
    initialize *Rule* condition $cond(Rule) \leftarrow \{\}$
    initialize *Rule* quality $q_g(Rule) \leftarrow \frac{|P|}{|N|+g}$
end for
while there are improvements in *Beam* and $length(Rule) \leq MaxRuleLength$ **do**
    for each $Rule \in Beam$ **do**
        for each $Feature \in F$ **do**
            $NewRule \leftarrow cond(Rule) \wedge Feature$
            $q_g(Rule) \leftarrow \frac{p}{n+g}$
            if $sup(Rule) \geq MinSup$ **and** $q_g(Rule)$ is larger than the quality of any
                rule in *NewBeam* **and** *Rule* is *relevant* **do**
                replace the worst rule in *NewBeam* with *Rule*
        end for
    end for
    $Beam \leftarrow NewBeam$
end while
Output: *SD_Rules* (all rules from *Beam*)

Table 4 Iterative subgroup construction algorithm based on the weighted covering approach. Counts $c(e)$ for positive examples are increased for the examples covered in previous iterations

DMS_algorithm(E, F, g, *MinSup*, *BeamWidth*, *MaxRuleLength*, *MaxRules*)

Input: set of examples $E = P \cup N$, set of features F
Parameters: g, *MinSup*, *BeamWidth*, *MaxRuleLength*, *MaxRules*
initialize *DMS_Rules* $\leftarrow \{\}$ (empty set)
for each $e \in P$ **do** initialize $c(e) \leftarrow 1$
repeat *MaxRules* times
    SD_Rules $\leftarrow$ *SD_algorithm*(E, F, g, *MinSup*, *BeamWidth*, *MaxRuleLength*)
        using $q'_g(R) = \frac{1}{n+g} \cdot \sum_{TP(R)} \frac{1}{c(e)}$
    select one R from *SD_Rules* with best $q'_g(R)$ value
    for each $e \in TP(R)$ covered by rule R **do**
        $c(e) \leftarrow c(e) + 1$
    end for
    add R into *DMS_Rules*
end repeat
Output: *DMS_Rules*

ensure diversity of induced rules, the DMS algorithm with the implemented weighted covering approach was used [7, 10].[2]

The core of the DMS algorithm is presented in Table 4. The DMS algorithm calls iteratively the SD algorithm. In each iteration it selects a single best rule from the beam, and includes it into the output rule set (the number of rules in output rule set *RuleSet* is determined by the *MaxRules* parameter).

The classical covering algorithm, most frequently used in rule learning, works as follows: the learner constructs a rule that correctly classifies some examples, removes the positive examples covered by the rule from the training set and repeats the process until no more examples remain. In contrast to this approach, DMS uses the weights of positive examples with the intention to ensure the diversity of rules induced in different iterations. After selecting a rule, the weights of positive examples covered by the rule are decreased. To do

[2]The name DMS comes from Data Mining Server, a public service available at `http://dms.irb.hr`, in which a public version of this algorithm is made available.

so, the number of rules covering each positive example are counted. All counts $c(e)$ are initially set to 1. The weights are computed as $w(e) = \frac{1}{c(e)}$, and in each iteration of the algorithm the example counts are recomputed, leading to decreased example weights. For that purpose, the DMS algorithm uses—instead of the unweighted $q_g(R)$ measure defined in (1)—the weighted rule quality measure defined as

$$q'_g(R) = \frac{\sum_{TP(R)} w(e)}{n+g}. \quad (2)$$

Although this approach can not guarantee the statistical independence of generated rules, it practically ensures good diversity of induced rules. This can be verified also from the results presented in the following section.

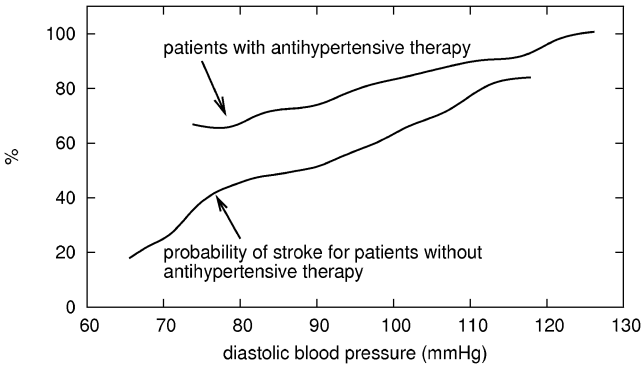

Fig. 2 Probability of brain stroke, estimated by the proportion of patients with brain attack (stroke) relative to the total number of patients in the hospital department, shown in dependence of diastolic blood pressure values, presented separately for patients with and without antihypertensive therapy

4 The process and results of insightful data analysis

Having determined the analysis task, defined by the selection of target and non-target class examples, the analysis process continues by systematically inducing rules at different generalization levels. By applying the methodology presented in Sect. 3 to brain ischaemia patient data, 15 rules presented in Table 5 were induced. For each of the five selected generalization levels there are three rules. By selecting a low generalization parameter value (g-value), the subgroup discovery algorithm tends to construct very specific rules with relatively low sensitivity. With the increase of the g-value the sensitivity of rules typically improves at the cost of decreased specificity. The sensitivity and the specificity values for each rule are given in columns 3 and 4, respectively. The last column indicates the overlap between the current rule and one/two rules induced previously for the same g-value. The overlap value is defined as the number of positive cases that are covered both by the current rule and the previously generated rule(s) divided by the number of positive cases covered by either the current rule or the previously generated rule(s), whichever is smaller. Low overlap values mean relative independence of rules.

As inductions with different generalization parameters are independent, there is a possibility that the same rule (e.g., $ahyp = yes$) is induced with different generalization parameter values. The order of rules in each group is the order selected by the DMS algorithm and is determined by the q'_g rule quality value that takes into the account the covering relations between the current rule and other rules previously selected for the same g-value.

The interpretation of induced rules starts by independent interpretations of each individual rule. There is no prior preference of either more specific or more sensitive rules. But typically more sensitive rules covering many cases tend to be shorter. This fact can be verified also from Table 5: very sensitive subgroup descriptions are described by a small number of conditions and because of that they are easier to be analysed by the domain experts.

4.1 Analysis of sensitive rules

Highly sensitive rules, like those induced with parameter $g = 100$ describe general characteristics of the target class. In the given domain we can observe that brain stroke is characteristic for middle aged or elderly population ($age > 52.00$), that people with stroke typically have normal or increased diastolic blood pressure ($dya > 75.00$), and that they have already detected hypertension problems and take some therapy (anti-hypertension therapy $ahyp = yes$). We also see that the selected boundary values are relatively low (52 years for the age and 75 mmHg for the diastolic pressure) which is due to the fact that the rules should satisfy a large number of cases. This is the reason why the rules are not applicable as decision rules but they provide useful descriptive information about the target class.

Expert interpretation of each individual rule is essential for the generation of useful knowledge. For example, the interpretation of rules like ($age > 52.00$) or ($dya > 75.00$) is straightforward. In contrast, the interpretation of the rule ($ahyp = yes$) could lead to the conclusion that antihypertensive therapy itself is dangerous for the incidence of stroke. A much more appropriate interpretation is that hypertension is dangerous, therefore people with detected hypertension problems, characterized by the fact that they already take antihypertensive therapy, have a greater probability of having a stroke. Indirectly, this rule also means that we have little chance to recognize the danger of high blood pressure, as suggested by rule g100b from Table 5, directly from their measured values. The reason is that many seriously ill patients have these values artificially low due to a previously prescribed therapy.

Table 5 Rules induced for g-values 5, 10, 20, 50, and 100, together with their sensitivity and specificity values measured on the available data set as well as their overlap with previously induced rule(s) in the same g-value group

Ref.	Rule	Sens.	Spec.	Overlap
generalization parameter $g = 5$				
g5a	(*fibr* > 4.55) and (*str* = *no*)	25%	100%	–
g5b	(*fibr* > 4.45) and (*age* > 64.00)	41%	100%	94%
g5c	(*af* = *yes*) and (*ahyp* = *yes*)	28%	95%	36%
generalization parameter $g = 10$				
g10a	(*fibr* > 4.45) and (*age* > 64.00)	41%	100%	–
g10b	(*af* = *yes*) and (*ahyp* = *yes*)	28%	95%	34%
g10c	(*str* = *no*) and (*alcoh* = *yes*)	28%	95%	67%
generalization parameter $g = 20$				
g20a	(*fibr* > 4.55)	46%	97%	–
g20b	(*ahyp* = *yes*) and (*fibr* > 3.35)	65%	73%	71%
g20c	(*sys* > 153.00) and (*age* > 57.00) and (*asp* = *no*)	45%	88%	80%
generalization parameter $g = 50$				
g50a	(*ahyp* = *yes*)	74%	54%	–
g50b	(*fibr* > 3.35) and (*age* > 58.00)	79%	63%	76%
g50c	(*age* > 52.00) and (*asp* = *no*)	64%	63%	96%
generalization parameter $g = 100$				
g100a	(*age* > 52.00)	96%	20%	–
g100b	(*dya* > 75.00)	98%	8%	98%
g100c	(*ahyp* = *yes*)	74%	54%	100%

This is a good example of expert reasoning stimulated by an induced rule. In this situation we may try to answer the question how the probability of stroke with respect to the transitory ischaemia cases changes with the increasing diastolic blood pressure. From the induced rule we have learned that we should compare only patients without antihypertension therapy. The result is presented in Fig. 2. It can be noticed that the probability of stroke grows significantly with the increase of diastolic blood pressure. The same dependency can be drawn also for the patients with the therapy. The differences between the two curves are significant and from them a few potentially relevant conclusions can be made. The first is that antihypertensive therapy helps in reducing the risk of stroke: this can be concluded from the fact that the probability of stroke is decreasing with the decrease of diastolic blood pressure also for the patients with the therapy. But it is also true that for diastolic blood pressure up to 100 mmHg the probability of stroke is significantly higher for patients with recognized hypertension problems than for other patients. The interpretation is that also in cases when successful treatment of hypertension is possible, the risk of stroke still remains relatively high and it is higher than for patients without hypertension problems.

As a conclusion of this reasoning it can be said that medical interpretations follow from a relatively simple visualization of existing statistical properties of the collected data. In this respect the methodology of data visualization and medical reasoning following from it, can not be characterized as novel. The significance of the approach is in the fact that the selection of properties that will be analysed (diastolic blood pressure) and conditions of the analysis (antihypertensive therapy present or absent) are suggested by medical reasoning based on induced subgroups. Obviously the same set of rules may stimulate other types of analyses, like the probability of stroke with respect to age with different levels of blood pressure as a parameter. What will be actually analysed and how the analysis will be performed depends on medical experts' preferences.

4.2 Analysis of specific rules

As noticed earlier, very sensitive rules are appropriate for extracting general properties of the target class. In contrast, very specific rules induced by generalization parameter values 5 or 10 are good as reliable classification rules for the target class. For example, rule g5c (*af* = *yes*) and (*ahyp* = *yes*) well reflects the existing expert knowledge that hypertension and atrial fibrillation are important risk factors for brain stroke. The rule is actually significant as it emphasizes the importance of the combination of these two risk factors, which is not a generally known fact. The relevance of the detected association is illustrated in Fig. 3. It shows that the probability of stroke is at least 85% in the age range 55–80 years for persons with both risk factors measured on the available hospital population. We can not estimate this probability on the general population but we can assume

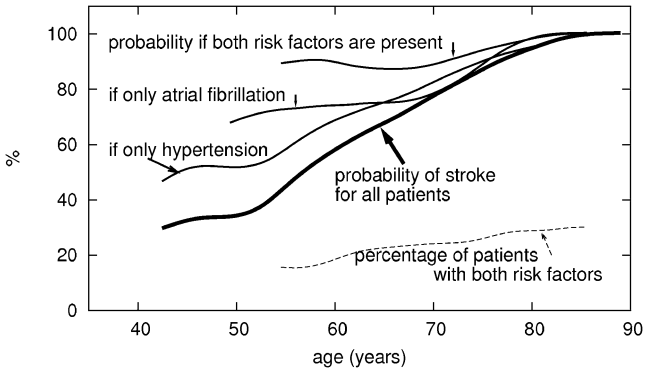

Fig. 3 Probability of brain stroke, estimated by the proportion of stroke patients, shown in dependence of patients' age presented for all patients in the available hospital population (*thick line*), probability of stroke for persons with hypertension problems, with atrial fibrillation problems, and with both hypertension and atrial fibrillation problems (*thin solid lines*). The percentage of patients with both risk factors is about 20–25% of the given hospital population (*dashed line*). The curves are drawn only for the range with sufficiently large numbers of patients in the database

that it is even larger. The observation might be important for prevention purposes in general medical practice, especially because both factors can be easily detected.

Other two rules induced for the g-value equal 5 contain conditions based on the fibrinogen values of about 4.5 or more (reference values for negative fibrinogen finding are in the range 2.0–3.7 gL^{-1}). The rules without doubt demonstrate the importance of high fibrinogen values for brain stroke patients. In the first rule the second necessary condition is the absence of stress, while in the second rule the second condition is age over 64 years. The interpretation of the second rule is relatively easy, leading to the conclusion that fibrinogen above 4.5 is itself very dangerous, which is confirmed also by rule g20a, being especially dangerous for elderly people. The interpretation of rule ($fibr > 4.55$) and ($stres = no$) is not so easy because it includes contradictory elements 'high fibrinogen value' and 'no stress', knowing the fact that stress increases fibrinogen values and increases the risk of stroke. The first part of the interpretation is that 'no stress' is characteristic of elderly people and this conclusion is confirmed by the high overlap value of rules g5a and g5b (see the last column for the g5b rule). The second part of the interpretation is that high fibrinogen values can be the result of stress and such fibrinogen is not as dangerous for stroke as fibrinogen resulting from other changes in the organism such as coagulation problems.

It can be concluded that the analysis of coexisting factors in subgroup descriptions may lead to very interesting insights. Two different situations are possible. Rules like g5b and g5c belong to the first situation in which detected conditions present known risk factors (like high fibrinogen value and diagnosed hypertension). In such situations the rules indicate the relevance of coexisting factors. In other cases when there is a surprising condition, like no stress in the g5a subgroup description of brain ischaemia, the interpretation should necessarily be based on the existing expert knowledge. The rule does not suggest the conclusion that 'no stress' is dangerous; instead, the conclusion is that increased fibrinogen is dangerous, *specially* when it is detected for patients that are not under stress. If there is a patient with increased fibrinogen value and the patient is under stress, it is possible to understand the reason for the increased fibrinogen value. In these circumstances the doctor will suggest the patient to avoid stressful situations. On the other hand, the situation when the patient has increased fibrinogen value without being exposed to stress is very different. In this case the fibrinogen value is increased without a known reason and, according to rule g5a, this may be a dangerous condition.

Very similar situation has been reported also in [3] when—in a coronary heart disease domain—two rules connected increased total cholesterol values with body mass index *below* 30. Again we had the situation that high body mass index and increased total cholesterol are known risk factor for the disease. The appropriate interpretation is that increased total cholesterol values are dangerous, *specially* if detected for patients without significant overweight problems.

Rules with 'surprising' conditions are interesting because they may open different hypotheses, sometimes stimulating further research that is out of the scope of this paper. In the case of brain stroke we can speculate that actually various subtypes of fibrinogen exist: one as a result of stress which is not very dangerous for stroke and the other subtype which is more dangerous but with unknown causes. The other possible speculation may be that increased fibrinogen is not dangerous by itself, but is dangerous because of some other unknown, typically co-occurring phenomenon that is difficult to detect or measure. And stress is the exception in the sense that it results in increased fibrinogen values but without so dangerous co-occurring phenomena.

4.3 Analysis of moderately sensitive and specific rules

From the rules induced with generalization parameter values 10–50 it can be noticed that conditions on age and fibrinogen values repeat often, confirming already made conclusions about their importance. Generally, rules obtained in the middle range of parameter g may be analysed in the same way as very sensitive or very specific rules. Potentially interesting subgroup descriptions are ($ahyp = yes$) and ($fibr > 3.35$), or another rule ($age > 52.00$) and ($asp = no$). The later rule stimulated the analysis presented in Fig. 4 which provides an excellent motivation for patients to accept prevention based on aspirin therapy. From the figure it can be easily noticed that the inductive learning approach correctly recognized the importance of the therapy for persons older than 52 years.

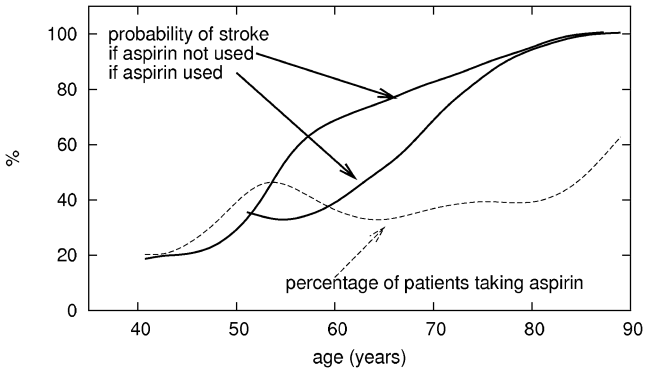

Fig. 4 The probability of brain stroke, estimated by the proportion of stroke patients, shown in dependence of patient age presented for patients taking aspirin as the prevention therapy, and the probability of stroke for patients without this therapy. The percentage of patients with the aspirin therapy is presented by a dashed line

In addition, the moderately sensitive and specific rules are relevant also for the selection of appropriate boundary values for numeric descriptors included into rule conditions. Examples are age over 57 or 58 years, fibrinogen over 3.3, and systolic blood pressure over 153. These values, if significantly different from generally accepted reference values, can initialize research in the direction of possibly accepting them as new decision points in medical decision making practice. Even more importantly, notice that different boundary points can be suggested in combinations with different conditions. This is in contradiction with existing medical practice which tends to define unique reference values irrespective of the disease that has to be described and irrespective of other patient characteristics. In the case of fibrinogen, reference values above 3.7 are treated as positive while rules induced for brain stroke domain suggest 4.55 as a stand alone decision point, 4.45 in combination with age over 64 years, and 3.35 in combination with hypertension or age over 58 years for very sensitive detection of stroke.

Boundary values for numerical descriptors suggested by decision points in conditions included into subgroup descriptions present an important part of insightful data analysis. It must be noted that suggested decision points are the result of unbiased search for optimal decision functions, incorporated in the used machine learning process. In this respect these values represent a result that is difficult to achieve by classical statistical approaches. Their importance is in the fact that the detected values nicely integrate the properties of several collected cases into values that can be easily compared to generally acceptable reference values or results obtained on other databases. In this way, major discrepancies may indicate diagnostic, methodological, or organizational problems in the organizations where data has been collected.

4.4 Analysis of rule groups

Besides the possibility to analyse each rule separately, combinations of co-occurring rules can trigger other interpretations. In this respect it is useful to look at the overlap values of rules. A good example is a group of three rules induced for g-value 10. These rules have low overlap values, meaning that they describe relatively diverse subpopulations of the target class. Their analysis enables global understanding of the hospital population in the Intensive Care Unit of the Neurology Department. Results of the analysis are presented in Fig. 5.

The figure graphically and numerically illustrates the importance of each population subgroup and its overlap with other subgroups. The textual description is also important, reflecting the results of basic statistical analysis (mean values of age and fibrinogen, as well as sex distribution) for the subpopulation described by the rule, followed by the so-called *supporting factors* [7]. The supporting factors are those descriptor values that are characteristic for the subgroup population in contrast to the cases in the negative class. These factors are important as they can help to confirm that a patient is a member of a subpopulation, also providing a better description of a typical member of a subgroup. The results show that the induced subgroups describe three relatively different types of stroke among elderly people (mean age between 70 and 75 years).

The largest subgroup can be called *elderly patients*; it is characterized by extremely high fibrinogen values (mean value 5.5) and increased glucose values (mean value 8.4). In most cases these are women (about 70%) that do not smoke, do not suffer from stress, and do not have problems with lipoproteins. Very different is the subpopulation that can be called *patients with serious cardiovascular problems* characterized with diagnosed hypertension and atrial fibrillation. It is a mixed male-female population. Its main characteristic is that they typically receive many different therapies but still they have increased—but inside reference—heart rate frequency (about 90) and uric acid (about 360). In between these two populations—in terms of age—is a subpopulation that can be called *do-not-care patients* characterized by alcohol consumption and no stress. It is a mixed male-female population characterized by the increased glucose values of laboratory tests, which one would not expect to find among the stroke patients because of their negative family history. Their do-not-care attitude is visible also from not taking aspirin as the prevention therapy.

Rule group analysis is an important part of insightful data analysis. Induction of subgroups is followed by statistical analysis resulting in supporting factors for detected subgroups [7]. Both descriptions of subgroups and their supporting factors are inputs for non-trivial medical expert reasoning. The problem is that this process can not be formalized because it strongly depends on human experience. In

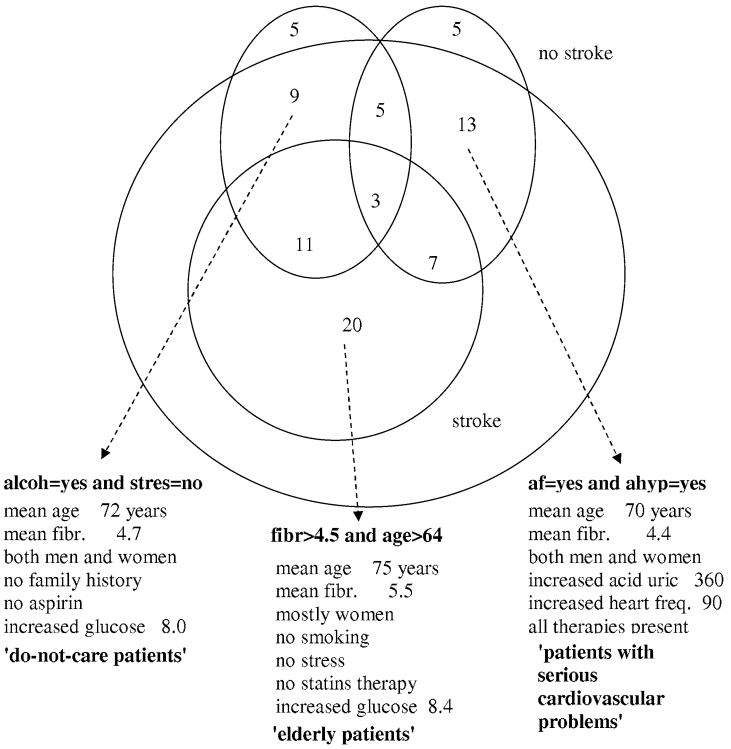

Fig. 5 Comparative study of three important subgroups of stroke patients detected by rules induced with generalization parameter $g = 10$. The *large circle* presents the stroke patients, negative cases are *outside the large circle*. *Small circles* present three detected subgroups. One of them includes only positive cases while the other two include also a small portion of negative cases. The *numbers* present the percentages of patients that satisfy the conditions of one, two, or all three rules. In total, 68% of positive cases are included in at least one subgroup. The definitions of patient groups (in *bold-face letters*) are followed by a list of most relevant properties that characterize the patient group (the supporting factors). The list ends with the concept name given to the group by the expert (in *bold-face letters*)

successful cases the process results in expert understanding of basic properties of detected subpopulations and the recognition of their agreement with previous medical experience. The final point of this reasoning process is when experts are able to give names to the detected subpopulations and when these names start to be used as concepts that describe existing medical practice or as novel knowledge that can be communicated to other people.

5 Related work

This section outlines related subgroup discovery approaches and applications.

5.1 Subgroup discovery approaches

Related subgroup discovery systems include EXPLORA [11], CN2-SD [12], MIDOS [13, 14], SubgroupMiner [15] and RSD [16]. EXPLORA and CN2-SD treat the learning task as a single relation problem, i.e., all the data are assumed to be available in one table (relation), while MIDOS, SubgroupMiner and RSD perform subgroup discovery from multiple relational tables. Like the SD algorithm, the CN2-SD and RSD algorithm also use a weighted covering algorithm and modify the computation of the search heuristic by example weights. The subgroup discovery component of RSD shares common basic principles with CN2-SD: the fundamental search strategy and the heuristic function employed therein (the weighted relative accuracy heuristic function is just slightly modified). The distinguishing feature of RSD compared to MIDOS and SubgroupMiner is that the latter two systems assume as input the tabular representation of training data and background relations. On the other hand, RSD input data has the form of ground Prolog facts and background knowledge either in the form of facts or intentional rules, including functions and recursive predicate definitions. Exception rule learning [17] also deals with finding interesting population subgroups.

A variety of rule evaluation measures and heuristics have been studied for subgroup discovery [11, 13, 14], with the aim of balancing the size of a group (referred to as factor g) with its distributional unusualness (referred to as factor p). The properties of functions that combine these two factors have been extensively studied in the so-called 'p-g-space' [11]. An alternative measure $q_g = \frac{p}{n+g}$, used by the SD algorithm in the experiments described in this paper, aims at minimizing the number of false positives n, and maximizing true positives p, balanced by generalization parameter g.

In the design of the SD algorithm, special attention was devoted to the problem of overfitting prevention. The topic, including the experiments with randomly generated data, has been extensively discussed in [9].

5.2 Subgroup discovery applications

An overview of exploratory data mining techniques and their applications in medical and health datasets is presented in

[18]. In [19] Pazzani and co-authors concentrate on rule learning as the starting point for medical expert reasoning, pointing out that monotonicity constraints in machine learning may help to induce rules that can be accepted as both accurate and meaningful by medical experts. The possibility of discovering functional interactions and co-influences among variables from topological properties of Bayesian networks has been presented in [20]. Decision tree learning algorithms [21] and rule learning algorithms [22] have also been extensively applied in very different intelligent data analysis applications in medical domains. It can be noticed that all these approaches enable the induction of classifiers that can be the source of insights concerning the relevant dependencies in the available data. In contrast with standard classification rule and decision tree learning, the major advantage of the subgroup discovery approach presented in this work is that the variations of the generalization parameter value enable an effective construction of rules with very different covering properties. Additionally, the application of the weighted covering algorithm, especially in combination with the possibility to use example weights different from 1, enables the control of overlap of the induced subgroups. The major advantage of the subgroup discovery approach is that the rules present the relevant relations in an explicit form. Although we stress the relevance of the existing expert knowledge for the proposed process of insightful analysis, it must be recognized that implicit or explicit incorporation of expert knowledge is required prior to any multifactorial data analysis, as well as for experiment organization and result interpretation when machine learning algorithms are used [23].

The subgroup discovery methodology has already been successfully applied to different medical and non-medical domains. In [3] its application to risk group modeling for the coronary heart disease has been presented. The work suggested an active mining framework, in which medical experts decide on the parameters used in the subgroup induction process, in contrast to this work which is based on a systematic evaluation of subgroup descriptions of different generality. The application of the subgroup discovery approach for the induction of comprehensible models from gene expression datasets [24] is also interesting. Typical gene expression domains namely include between 10 and 20 thousands of variables and the successful construction of short and simple rules with satisfactory prediction quality demonstrated the robustness and scalability of the subgroup discovery algorithm.

6 Discussion and conclusions

This work demonstrates that rules induced by the subgroup discovery methodology can be an appropriate starting point for data analysis leading to insightful descriptions generalizing the available data. The extensive presentation of the medical expert reasoning which is based on induced subgroup descriptions for the brain ischaemia domain is the central part of the work. The presentation intends to illustrate the intellectual effort necessary to convert the induced rules into reasonable medical knowledge and has the intention to set up the guidelines for this creative process. The effort to systemize the reasoning process according to the sensitivity level of induced rules is the main novelty of this work.

It must be noted that the presented approach does not cover all possible aspects of insightful data analysis. Application of other supervised and non-supervised machine learning approaches may enable the detection of other types of useful information about the available data, like explicit noise detection, relative distance among cases, construction of prototype cases for specified classes, detection of relative importance of descriptors, and so on. The identification of an appropriate set of tools that should be used in the data analysis process and the suggestions how their results should be interpreted, is one of the long term goals in the field of intelligent data analysis. It can be expected that the approach based on supervised learning of subgroup descriptions will be a part of the final solution.

According to the experience collected on a few, mainly medical domains, it can be said that the principal advantage of the subgroup discovery methodology is in the possibility to search over a range of different descriptions, in terms of their specificity, diversity of rules, and diversity of used descriptors. Another, equally important advantage of the induced rules is their form, which is easily understandable and interpretable by domain experts. This property is achieved by systematic use of constraints that are applied in the rule construction process. The constraints deserve special attention not only because they ensure the interpretability of the results but also because they ensure that the induced subgroup descriptions reflect the most relevant actual dependencies between descriptor values and classes of examples. The final result of insightful analysis strongly depends on this quality of subgroup descriptions and a lot of effort has been invested in the selection of the most appropriate set of constraints.

References

1. Pazzani MJ (2000) Knowledge discovery from data? IEEE Intell Syst 15(2):10–13
2. Fayyad UM, Piatetsky-Shapiro G, Smyth P (1996) From data mining to knowledge discovery in databases. AI Mag 17(3):37–54
3. Gamberger D, Lavrač N, Krstačić G (2003) Active subgroup mining: a case study in a coronary heart disease risk group detection. Artif Intell Med 28:27–57
4. Amarenco P et al (1994) Atherosclerotic disease of the aortic arch and the risk of ischemic stroke. New Engl J Med 331:1474–1479
5. Barnett HJM et al (eds) (1998) Stroke. Pathophysiology, diagnosis, and management, 3rd edn. Elsevier Science, Churchill

6. Victor M, Ropper AH (2001) Cerebrovascular disease. In: Adams & Victor's principles of neurology. McGraw–Hill, New York, pp 821–924
7. Gamberger D, Lavrač N (2002) Expert-guided subgroup discovery: methodology and application. J Artif Intell Res 17:501–527
8. Fayyad UM, Irani KB (1992) On the handling of continuous-valued attributes in decision tree generation. Mach Learn 8:87–102
9. Gamberger D, Lavrač N (2004) Avoiding data overfitting in scientific discovery: experiments in functional genomics. In: Proceedings of the 16th European conference on artificial intelligence (ECAI 2004), pp 470–474
10. Fürnkranz J (2005) From local to global patterns: evaluation issues in rule learning algorithms. In: Morik K, Boulicaut J-F, Siebes A (eds) Local pattern detection. Springer, Berlin, pp 20–38
11. Klösgen W (1996) Explora: a multipattern and multistrategy discovery assistant. In: Fayyad UM, Piatetski-Shapiro G, Smyth P, Uthurusamy R (eds) Advances in knowledge discovery and data mining. MIT, Cambridge, pp 249–271
12. Lavrač N, Kavšek B, Flach P, Todorovski L (2004) Subgroup discovery with CN2-SD. J Mach Learn Res 5:153–188
13. Wrobel S (1997) An algorithm for multi-relational discovery of subgroups. In: Proceedings of the 1st European conference on principles of data mining and knowledge discovery, pp 78–87
14. Wrobel S (2001) Inductive logic programming for knowledge discovery in databases. In: Džeroski S, Lavrač N (eds) Relational data mining. Springer, Berlin, pp 74–101
15. Klösgen W, May M (2002) Census data mining—an application. In: Proceedings of the 6th European conference on principles and practice of knowledge discovery in databases, pp 65–79
16. Lavrač N, Železný F, Flach P (2003) RSD: relational subgroup discovery through first-order feature construction. In: Proceedings of the 12th international conference on inductive logic programming, pp 149–165
17. Suzuki E (2004) Discovering interesting exception rules with rule pair. In: Proceedings of the ECML/PKDD workshop on advances in inductive rule learning, pp 163–178
18. Roddick JF, Fule P, Graco WJ (2003) Exploratory medical knowledge discovery: experiences and issues. ACM SIGKDD Explor Newslett 5(1):94–99
19. Pazzani MJ, Mani S, Shankle R (2001) Acceptance by medical experts of rules generated by machine learning. Methods Inf Med 40(5):380–385
20. Lucas PJF, van der Gaag LC, Abu-Hanna A (2004) Editorial: Bayesian models in biomedicine and health-care. Artif Intell Med 30(3):201–214
21. Quinlan JR (1993) C4.5: programs for machine learning. Kaufmann, San Mateo
22. Clark P, Niblett T (1989) The CN2 induction algorithm. Mach Learn 3(4):261–283
23. Okada T (2001) Medical knowledge discovery on the meningoencephalitis diagnosis studied by the cascade model. In: Proceedings of the new frontiers in artificial intelligence, Joint JSAI workshop, pp 533–540
24. Gamberger D, Lavrač N, Železný F, Tolar J (2004) Induction of comprehensible models for gene expression datasets by subgroup discovery methodology. J Biomed Inf 37(4):269–284