Data analysis based on subgroup discovery: Experiments in brain ischaemia domain

Dragan Gamberger
Rudjer Bošković Institute
Zagreb, Croatia
dragan.gamberger@irb.hr

Antonija Krstačić
Univ. Hospital of Traumatology
Dept. of Neurology,
Zagreb, Croatia

Goran Krstačić
Institute for Cardiovascular Prevention and Rehabilitation,
Zagreb, Croatia

Nada Lavrač
Jožef Stefan Institute
Ljubljana, Slovenia

Michèle Sebag
Université Paris-Sud
Orsay, France

Abstract

This paper presents insightful analysis of medical data collected in regular hospital practice. The domain consists of patients suffering from brain ischaemia, either permanent as brain attack (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 vascular brain disease. The work demonstrates the applicability of subgroup rule induction as the basis for insightful data analysis and describes intellectual process of converting rules into reasonable medical concepts. Detection of coexisting risk factors, selection of relevant discriminative points for numerical descriptors, as well detection and description of characteristic patient subpopulations are important results of the analysis. Graphical representation is extensively used to illustrate the detected regularities.

1 Introduction

Data analysis in medical applications is characterized by the ambitious goal of extracting potentially new relationships from the data, and providing insightful representations of detected relationships. Applications of quantitative statistical methods seldom lead to insightful results, leaving a large workload on the human experts who have to provide appropriate interpretations of results, with no guarantees that—due to a huge search space of possible solutions—the most relevant combinations have been tested at all [Fayyad *et al.*, 1996]. The goal of intelligent data analysis is to effectively detect 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 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 existing methodology of supervised subgroup discovery [Gamberger *et al.*, 2003] can serve as an appropriate basis for data analysis, if supplemented by the sufficient intellectual effort of medical experts, willing to convert machine-induced rules into adequate medical interpretations. The proposed approach, applied to a typical database collected in regular hospital practice describing brain ischaemia patients, is used to illustrate this expert-guided approach to knowledge discovery. The next section presents the problem domain. Section 3 presents the proposed data analysis approach leading to insightful knowledge, interpreted by medical specialists in Section 4.

2 Brain ischaemia data

The database consists of records of patients who have been treated in the Intensive Care Unit of the Department of Neurology, University Hospital Center "Zagreb", in Zagreb, Croatia during the year 2003. In total, 300 patients are included in the database: 209 with the confirmed diagnosis of brain attack (stroke), and 91 patients who entered the same department with adequate neurological symptoms and disorders, but were diagnosed (based on the outcomes of neurological tests) as patients with transition ischaemic brain attack (TIA, 33 patients), reversible ischaemic neurological deficit (RIND, 12 patients), and serious headache or cervical spine syndrom (46 patients). In this paper, the goal of data analysis experiments is to discover regularities that characterize brain stroke patients.

Patients are described with 27 different descriptors representing anamnestic data, physical examination data, laboratory test data, ECG data, CT test result and information about previous hospital therapies. Descriptors used in the analyses are listed in Table 1.

It must be noted that the control group does not consist of healthy persons but patients with serious neurological symptoms and disorders. In this sense, the available database is particularly appropriate for studying specific characteristics and subtle differences that distinguish patients with stroke. The detected relationships can be accepted as true 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 [Victor and Ropper, 2001].

3 Data analysis process

This section presents the data analysis process, using rules induced by the SD subgroup discovery algorithm [Gam-

Descriptor	Abbreviation
sex (f,m)	sex
age (years)	age
family anamnesis (n,p)	fhis
present smoking (y,n)	smok
stress (y,n)	str
alcohol consumption (y,n)	alcoh
systolic blood pressure cont. (mmHg) normal value < 139 mmHg	sys
diastolic blood pressure continuous (mmHg) normal value < 89 mmHg	dya
uric acid continuous ($\mu mol\ L^{-1}$) ref. value for men < 412 ref. value for women < 380	ua
fibrinogen continuous ($g\ L^{-1}$) ref. value 2.0-3.7	fibr
glucose continuous ($mmol\ L^{-1}$) ref. value 3.6-5.8	gluc
heart rate continuous ref. value $60 - 100$ /min	ecgfr
atrial fibrillation (y,n)	af
left ventricular hypertrophy (y,n)	ecghlv
aspirin therapy (y,n)	asp
anticoagulant therapy (y,n)	acoag
antihypertensive therapy (y,n)	ahyp
antiarrhytmic therapy (y,n)	aarrh
statins (antihyperlipoproteinaemic t.) yes, no	stat
hypoglycemic therapy none, yesO (oral), yesI (insulin)	hypo

Table 1: List of most relevant descriptors in the brain ischaemia domain with abbreviations used in induced rules. Included are also reference values representing the range typically accepted as normal in the medical practice.

berger *et al.*, 2003]. The process begins with a set of rules that are obtained by repetitively applying the SD algorithm with different generalization parameter values. In the experimental setting determined for the ischemia domain, the process of expert-guided subgroup discovery was performed as follows. The SD algorithm was run for values g in the range 5 to 100, and a fixed number of selected output rules equal to 3. The rules induced in this iterative process were shown to the expert for selection and interpretation. The intention of this paper is to illustrate what type of insights are possible by the analysis based on individual rules and what can be additionally obtained if rules are analysed in groups. The SD algorithm,[1] described in detail in [Gamberger *et al.*, 2003] is—due to paper length restrictions—out of scope of this paper.

The basic characteristic of the presented approach is supervised learning of subgroup defining rules that characterize the target (positive) class cases (in this domain stroke cases) in contrast to cases in the non-target (negative or control) class (in this domain transitory ischaemia cases). This means that examples of two classes have to be available. 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 section we demonstrate only the process performed for the complete database with patients who experienced stroke selected as the target class. We have performed a series of experiments also with patients separated in different age and sex groups, some of them also with reduced descriptor sets. Although the results are very interesting, specially due to the possibility of the comparative analysis of rules, they are not included in this paper due to space restrictions.

[1]The algorithm is available as part of the publicly available Data Mining Server at http://dms.irb.hr, and can be used to induce rules for domains with up to 250 cases.

4 Results of rule analysis

Table 2 presents rules generated for the class stroke. There are in total 15 rules, three for each of the five selected g-values in the range $5 \leq g \leq 100$. By selecting a low g-value, the subgroup discovery algorithm tends to construct very specific rules with relative low sensitivity. With the increase of the g parameter the sensitivity 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 previosly generated rule(s), whichever is the smaller. Low overlap values mean relative independence between the rules.

Because 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 algorithm and it is determined by the rule quality value and the rule covering properties.

4.1 Analysis of individual rules

The interpretation of induced rules starts by independent interpretation of each individual rule. There is no apriori preference of either more specific or more sensitive rules. Highly sensitive rules, like those induced with parameter $g = 100$ describe general characteristics of the target class. In the given domain we see that stroke is characteristic for middle aged or elderly population $(age > 52.00)$, that people with the stroke typically have normal or increased dyastolic blood pressure $(dya > 75.00)$, and that they have already detected hypertension problems and take some therapy (anti-hypertension therapy yes). We also see that the

Ref.	Rule		
	Sens.	**Spec.**	**Overlap**
generalization parameter value 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 value 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 value 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 value 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 value 100			
g100a	$(age > 52.00)$		
	96%	20%	-
g100b	$(dya > 75.00)$		
	98%	8%	98%
g100c	$(ahyp = yes)$		
	74%	54%	100%

Table 2: Rules induced for generalization parameter g values in the range[5,100]. Presented are their sensitivity and specificity values measured on the available data set as well as their overlap with previouly induced rule(s) in the same g-value group.

selected boundary values are relative low (52 years for the age and 75 mmHg for the dyastolic 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 give 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 better interpretation is that hypertension is dangerous and because of that people with detected hypertension problems, characterized but the fact that they already take antihypertensive therapy, have larger probability of having a stroke. Indirectly, this rule also means that we have little chance to recognize the danger of high

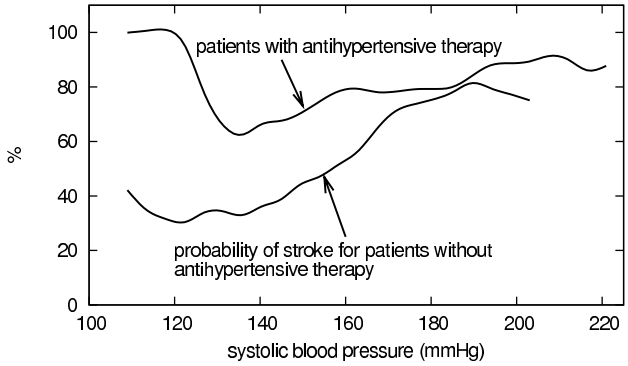

Figure 1: The proportion of patients with brain attack (stroke) in dependence of the total number of patients in the hospital department presented separately for patients with and without antihypertensive therapy for different systolic blood pressure values.

blood pressure directly from their measured values because many serious patients have these values artificially low due to a previously prescribed therapy. This is a good example of expert reasoning stimulated by an induced rule. In this situation we try to answer the question how the probability of stroke with respect to the transitory ischemia cases changes with the increasing systolic blood pressure. From the rule we have learned that we should compare only patients without anti-hypertension therapy. The result is presented in Figure 1. It can be noticed that the probability of stroke grows significantly with the increase of systolic blood pressure. The same dependency can be drawn also for the patients with the therapy. The differences between the two curves are significant and 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 systolic blood pressure also for the patients with the therapy (as long as the systolic blood pressure is not lower than 130 mmHg). But it is also true that for systolic blood pressure between 130 and 170 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 treatement of hypertension is possible, the risk of stroke still remains relatively high and it is higher than for patients without hypertension problems.

As noticed earlier, very general rules are good for extracting general properties of the target class. In contrast to that, very specific rules induced by generalization parameter values 5 or 10 are good as 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 the stroke. The rule is significant as it emphasizes the importance of the combination of these two risk factors, what is not a generally known fact. The relevancy of detected correlation is illustrated in Figure 2. 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

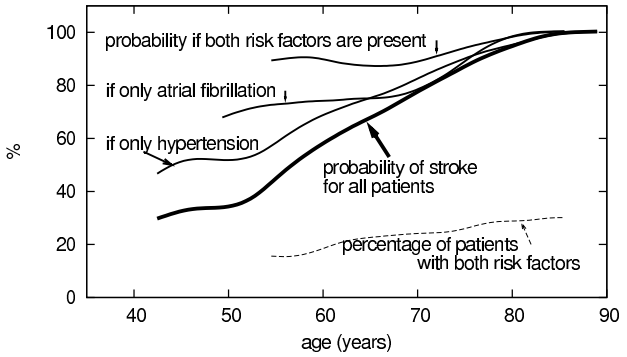

Figure 2: Probability of stroke in dependence of patient age presented for all patients in the available hospital population (thick line), probability of stroke for persons with hypertension problems, with atrial fibrillation problems, and with both hypertension and atrial fibrillation problems (thin solid lines). The percentage of patients with both risk factors is about 20-25% for the available hospital population (dashed line). The curves are drawn only for the range with a sufficiently large numbers of patients in the database.

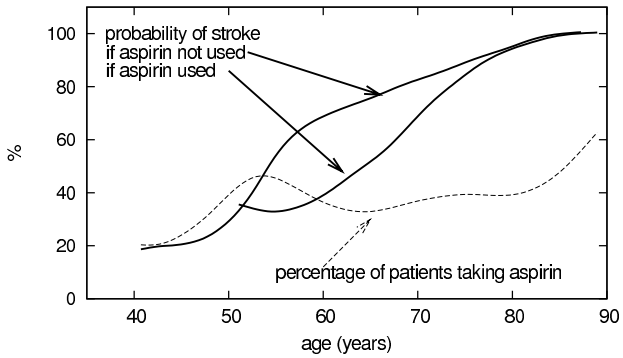

Figure 3: The probability of stroke in dependence of patient age presented for patients taking aspirin as the prevention therapy, and the probability of stroke for patients without this therapy. The percentage of patients with the aspirin therapy is presented by a dashed line.

probability on the general population but we can assume 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 g-value equal 5 contain conditions based on the fibrinogen values about 4.5 or more (reference values for negative fibrinogen finding are in the range 2.0 - 3.7 $g \cdot L^{-1}$). The rules without doubt demonstrate the importance of high fibrinogen values for the 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.

From the rules induced with generalization parameter values 10–50 we notice that conditions on age and fibrinogen values repeat often, confirming already made conclusions about their importance. Also they suggest much more reasonable boundary values for the numerical descriptors (age over 57 or 58 years, fibrinogen over 3.3, systolic blood pressure over 153) which, if different from generally accepted reference values, can initialize research in the direction of accepting them as new decision points in medical decision making practice.

Also rules in this middle range of parameter g stress relevant relations among different descriptors like $(ahyp = yes) and (fibr > 3.35)$ or $(age > 52.00) and (asp = no)$. The later rule stimulated the analysis presented in Figure 3 which seems as excellent motivation for patients to accept prevention based on aspirin therapy. It can be easily noticed that the inductive learning approach correctly recognized the importance of the therapy for persons older than 52 years.

4.2 Analysis of rule groups

Besides the possibility to analyse each rule separately, combinations of co-occurring rules can give some additional information. 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 relative 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 Figure 4.

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. The supporting factors are those descriptor values that are characteristic for the subpopulation in contrast to the cases in the negative class. The importance of these factors lies in the fact that they can help to confirm that a patient is a member of a subpopulation, also giving 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 subpopula-

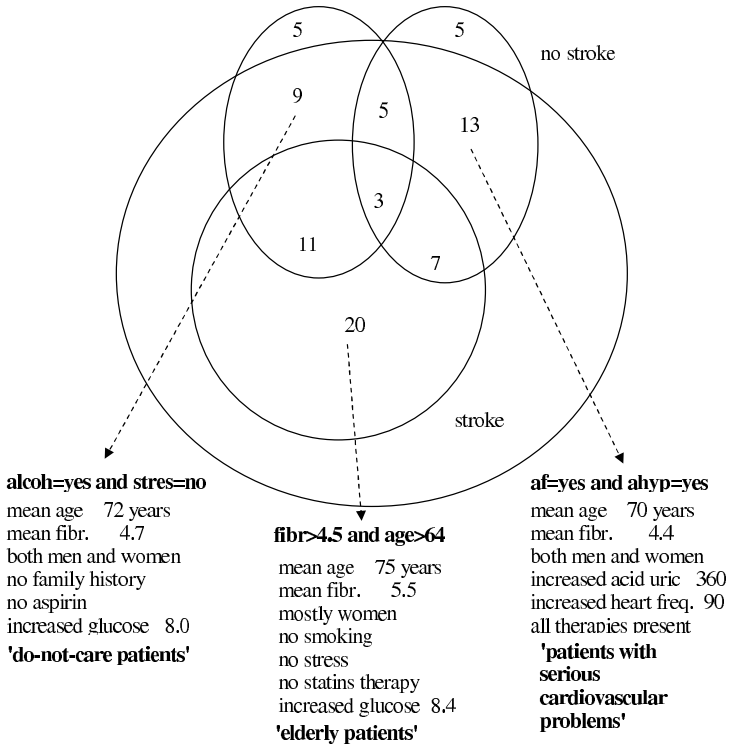

Figure 4: Comparative study of three important subgroups of stroke patients detected by rules induced with generalization parameter value 10. The large circle presents the stroke patients, negative cases are outside the large circle. Small circles present three detected subgroups. One of them includes only positive cases while the other two include also a small part 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 list ends with the expert's name given to the group (in bold-face letters).

tion 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 acid uric (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 also a mixed male-female population characterized only by the increased glucose values of laboratory tests. It seems as these people would have the largest chance not to be among patients with stroke because their relevant property is negative family history. Their do-not-care attitude is visible also from not taking aspirin as the prevention therapy.

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 of the available data. The extensive presentation of the analysis process intends to illustrate the intellectual effort necessary to convert the induced rules into reasonable medical knowledge. Special attention was devoted to the selection of appropriate visualization, enabling effective and convincing presentation of obtained results. The paper demonstrates that this type of data analysis, besides expert knowledge, requires also a lot of human imagination. Further work is expected in developing the methodology which could be used for semi-automated insightful data analysis.

References

[Fayyad *et al.*, 1996] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37–54, 1996.

[Gamberger *et al.*, 2003] Dragan Gamberger, Nada Lavrač, and Goran Krstačić. Active subgroup mining: A case study in a coronary heart disease risk group detection. *Artificial Intelligence in Medicine*, 28:27–57, 2003.

[Victor and Ropper, 2001] Maurice Victor and Allan H. Ropper. *Principles of Neurology*, chapter Cerebrovascular Disease, 821–924. McGraw-Hill New York, 2001.