Supporting Factors to Improve the Explanatory Potential of Contrast Set Mining: Analyzing Brain Ischaemia Data

N. Lavrač[1,2], P. Kralj[1], D. Gamberger[3] and A. Krstačić[4]

[1] Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
[2] University of Nova Gorica, Vipavska 13, 5000 Nova Gorica, Slovenia
[3] Rudjer Bošković Institute, Bijenička 54, 10000 Zagreb, Croatia
[4] University Hospital of Traumatology, Draškovičeva 19, 10000 Zagreb, Croatia

Abstract— **The goal of exploratory pattern mining is to find patterns that exhibit yet unknown relationships in data and to provide insightful representations of detected relationships. This paper explores contrast set mining and an approach to improving its explanatory potential by using the so called supporting factors that provide additional descriptions of the detected patterns. The proposed methodology is described in a medical data analysis problem of distinguishing between similar diseases in the analysis of patients suffering from brain ischaemia.**

Keywords— **Exploratory data analysis, contrast set mining, subgroup discovery, supporting factors, brain ischaemia**

I. INTRODUCTION

Data analysis in medical applications is characterized by the ambitious goal of extracting potentially new relationships from data, and providing insightful representations of detected relationships. Methods for symbolic data analysis are preferred since highly accurate but non-interpretable classifiers are frequently considered useless for medical practice.

The task of descriptive induction is to construct patterns or models describing data properties in a symbolic, human understandable form. Descriptive induction methods subgroup discovery [1], contrast set mining [2] and emerging patterns [3] are specifically designed to extract patterns (in the form of rules) from class labeled data. Unlike methods for inducing classification models (such as decision tree induction [4] and classification rule learning [5]), the patterns discovered by descriptive induction methods represent individual chunks of knowledge and are appropriate for being interpreted one-by-one.

The descriptive induction task is not concluded when individual rules are discovered. A property of the discovered rules is that they contain only the minimal set of principal characteristics of the target class that distinguish the target class examples (positive examples) from the control set (negative examples). For interpretation and understanding purposes other properties that support the detected rules are also relevant. In subgroup discovery these properties are called supporting factors. They are used for better human understanding of the principal factors and for the support in the decision making process [6].

A special data mining task dedicated to finding differences between contrasting groups is contrast set mining [2]. In our recent work [7] we have shown the similarity of contrast set mining and subgroup discovery and proposed a method for contrast set mining through subgroup discovery. The focus of this paper is to extend the concept of supporting factors from subgroup discovery to contrast set mining. We present our approach on the problem of discriminating between two groups of ischaematic brain stroke patients: patients with thrombolic stroke and those with embolic stroke.

This paper is organized as follows: Section II introduces the brain ischaemia data analysis problem. Section III presents the subgroup discovery approach to contrast set mining, including the results on the brain ischemia data. Section IV presents the statistical approach to discovering supporting factors in subgroup discovery and its adaptations to contrast set mining, as well as the results and the medical interpretation of the discovered contrast sets from brain ischaemia data.

II. THE BRAIN ISCHAEMIA DATA ANALYSIS PROBLEM

A stroke occurs when blood supply to a part of the brain is interrupted, resulting in tissue death and loss of brain function. Thrombi or emboli due to atherosclerosis commonly cause ischemic arterial obstruction. Atheromas, which underlie most thrombi, may affect any major cerebral artery. Atherothrombotic infarction occurs with atherosclerotic involving selected sites in the extracranial and major intracranial arteries. Cerebral emboly may lodge temporarily or permanently any where in the cerebral arterial tree. They usually come from atheromas (ulcerated atheroscleritic plaques) in extracranial vessels or from thrombi in a damaged heart (from mural thrombi in atrial Fibrillation). Atherosclerotic or hypertensive stenosis can also cause a stroke. Embolic strokes, thrombolic strokes and stokes caused by stenosis of blood vessels are categorized as ischaemic strokes. 80% of all strokes are ischaemic while the remaining 20% are caused by bleeding [8].

The brain ischaemia database, that is the focus of our analysis, consists of records of patients who were treated at the Intensive Care Unit of the Department of Neurology, University Hospital Center "Zagreb", Zagreb, Croatia, in year 2003. In total, 300 patients are included in the database:

- 209 patients with the computed tomography (CT) confirmed diagnosis of brain stroke: 125 with embolic stroke, 80 with thrombolic stroke, and 4 undefined.
- 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 severe headache or cervical spine syndrome (46 patients).

Patients are described with 26 descriptors representing anamnestic, physical examination, laboratory test and ECG data, and their diagnosis.

III. CONTRAST SET MINING THROUGH SUBGROUP DISCOVERY

A data mining task devoted to finding differences between groups is contrast set mining (CSM). It was defined by Bay and Pazzani [2] as "finding conjunctions of attributes and values that differ meaningfully across groups". If was later shown that contrast set mining is a special case of a more general rule discovery task [5]. Finding all the patterns that discriminate one group of individuals from all other contrasting groups is not appropriate for human interpretation. Therefore, as is the case in other descriptive induction tasks, the goal of contrast set mining is to find only the descriptions that are "unexpected" and "most interesting" to the end-user [2].

On the other hand, a subgroup discovery (SD) task is defined as follows: Given a population of individuals and a property of those individuals that we are interested in, find population subgroups that are statistically "most interesting", i.e., are as large as possible and and have the most unusual statistical (distributional) characteristics with respect to the property of interest [1].

Putting these two tasks in a broader rule learning context, note that there are two main ways of inducing rules in multi-class learning problems: learners either induce the rules that characterize one class compared to the rest of the data (the standard one-versus-all setting, used in most clasification rule learners), or alternatively, they search for rules that discriminate between all pairs of classes (known as the round robin approach used in classification rule learning, proposed by [9]). Subgroup discovery is typically performed in a one-versus-all rule induction setting, while contrast set mining implements a round robin approach (of course, with different heuristics and goals compared to classification rule learning).

Even though the definitions of subgroup discovery and contrast set mining seem different, the tasks are compatible [7]. From a dataset of class labeled instances (the class label being the property of interest) by means of subgroup discovery [1] we can find contrast sets in a form of short interpretable rules. Note, however, that in subgroup discovery we have only one property of interest (class) for which we are building subgroup descriptions, while in contrast set mining each contrasting group can be seen as a property of interest. It is easy to show that a two-group contrast set mining task CSM(G1;G2) can be directly translated into the following two subgroup discovery tasks: SD(Class = G1 vs. Class = G2) and SD(Class = G2 vs. Class = G1). And since this translation is possible for a two-group contrast set mining task, it is - by induction - also possible for a general contrast set mining task.

Our experiments show that the round robin approach is not appropriate when looking for characteristic differences between two similar diseases if data about normal (healthy) people is also available. The reason is that the algorithm could – by coincidence – find features that distinguish between two diseases but are at the same time characteristic for normal people. Therefore we use a one-versus-all approach which is standard in subgroup discovery. To find characteristics of the embolic patients we perform subgroup discovery on the embolic group compared to the rest of the patients (thrombolic and those with a normal CT). Similarly, when searching for characteristics of thrombolic patients, we compare them to the rest of the patients (embolic and those with a normal CT). In this setting, we ran the contrast set mining experiment with the Orange [10] implementation of the Apriori-SD subgroup discovery algorithm [11] with the following parameters: minimal support = 15%, minimal confidence = 30%, k = 5. The results are displayed in Figures 1 and 2.

Strokes caused by embolism are most commonly caused by heart disorders. The first rule displayed in Figure 1 has only one condition confirming this medical knowledge as atrial fibrillation (af = yes) as an indicator for brain stroke. The combination of features from the second rule also shows that patients with antihypertensive therapy (ahyp = yes) and antiarrhytmic therapy (aarrh = yes), therefore patients with heart disorders are prone to embolic stroke.

Thrombolic stroke is most common with older people, and often there is underlying atherosclerosis or diabetes. In the rules displayed in Figure 2 the features presenting diabetes do not appear. The rules rather describe patients without hart (or other) disorders but with elevated dyastolic blood pressure and fibrinogen. High cholesterol, age and fibrinogen values appear characteristic for all ischeamatic strokes.

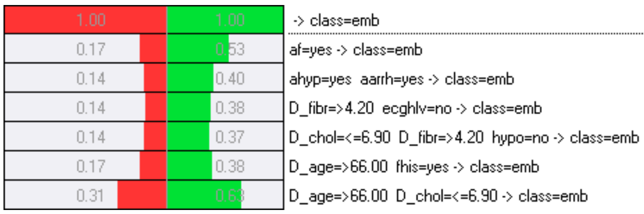

Fig. 1 Characteristic descriptions of embolic patients displayed in the bar chart subgroup visualization: on the right side the positive cases, in our case embolic patients, and on the left hand side the others - thombolic and normal CT.

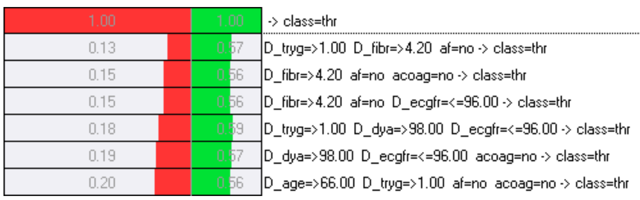

Fig. 2 Characteristic descriptions of thrombolic patients displayed in the bar chat subgroup visualization

IV. Supporting factors

Exploratory pattern discovery is not concluded when individual rules are discovered. The interpretation and insightful knowledge discovery is the goal that needs to be further perused. As shown in the previous section, some rules can be interpreted directly. But the discovered rules contain only a minimal set of principal differences between the detected subset of target (positive) and the control (negative) class examples – in our case up to four features per rule. For a domain expert, in our case a medical doctor, the information about other characteristics that support and enforce the discovered patterns is very relevant.

A. Supporting factors in subgroup discovery

In subgroup discovery the factors that appear in subgroup descriptions are called the principal factors, while the additional properties that are also characteristic for the detected subgroup are called supporting factors. They are used for better human understanding of the principal factors and for the support in the decision making process [12].

The supporting factors detection process is for every detected subgroup repeated for every attribute separately. For numerical attributes their mean values are computed while for categorical attributes the relative frequency of the most frequent or medically most relevant category is computed. The mean and relative frequency values are computed for three example sets: for the subset of positive examples that are included into the pattern, the set of all positive examples, and finally for the set of all negative examples (the control set).

The necessary condition for an attribute to be potentially used to form a supporting factor is that its mean value or the relative frequency of the given attribute value must be significantly different between the target pattern and the control example set. Additionally, the values for the pattern must be significantly different from those in the complete positive population. The reason is that if there is no such difference then such a factor is supporting for the whole positive class and not specific for the pattern.

The statistical significance between example sets can be determined using the Mann-Whitney test for numerical attributes and using the chi-square test of association for categorical attributes. A practical tutorial on using these tests can be found in [13] (Ch. 11a and 8, respectively). The decision which statistical significance is sufficiently large can depend on the medical context. We set the cut-off values at $P < .01$ for the significance of the difference with respect to the control set and $P < .05$ for the significance with respect to the positive set.

B. Supporting factors for contrast sets

Even though contrast set mining and subgroup discovery are very similar, there is a crucial difference between these two data mining techniques: in subgroup discovery there is only one property of interest and the goal is to find characteristics of the individuals that have this property of interest. In contrast set mining there are several groups of individuals and the goal is to find differences between the individuals belonging to these groups. Therefore the notion of supporting factor from subgroup discovery can not be directly adopted in the contrast set mining situation.

We propose and show in our experiments a way of extending the supporting factors from subgroup discovery to contrast set mining. Instead of presenting to the domain expert only the supporting factors for the positive class, we also show the distribution (for discrete) or the average (for numeric) attributes appearing in the supporting factor for the negative set and for the entire positive set. This is similar to the work presented in [14], but the methodology proposed here is tailored for helping explaining contrast sets. Since the interpretation of all the patterns discovered and presented in Section III is out of the scope of this paper, we focus only on two contrast sets:

Contrast set 1: (TPr=0.4, FPr=0.14)
ahyp=yes & aarrh=yes → class=emb
Contrast set 2: (TPr=0.56, FPr=0.2)
age>66 & trig>1 & af=no & acoag=no → class=thr

The first of the selected contrast sets is intuitive to interpret since both principal factors are treatments for cardiovascular disorders. The supporting factors for this set are shown in Table 1. We can see that the supporting factors (including two principal factors) for this contrast set are all about cardiovascular disorders and therefore they substantiate the original interpretation. It is therefore legitimate to say that embolic stroke patients are patients with cardiovascular disorders while cardiovascular disorders are not characteristic for thrombolic stroke patients.

The second selected contrast set is vague and is not directly connected with medical knowledge. High age and triglyceride values are characteristic for thrombolic stroke, but the boundary values in the contrast set are not high. The rest of the features in this contrast set say no atrial fibrillation and no anticoagulant therapy: again nothing specific. The supporting factors for this set are shown in Table 2. The supporting factors include high cholesterol and fibrinogen, low fundus ocular and non smoker. These patients are old and they do not have cardiovascular disorders. These examples indicate how supporting factors enforce the principal factors and help the interpretation to move from speculation toward legitimate conclusions.

Table 1 Supporting factors for contrast set 1

	CS1	thrombolic	embolic
fo high	0.82	0.73	0.76
af = yes	80%	13%	53%
ahyp = yes	100%	81%	70%
aarrh = yes	100%	19%	45%

V. Conclusions

We have generalized the notion of supporting factor form subgroup discovery to contrast set mining. We have applied the proposed methodology of supporting factors for contrast set mining in the analysis of the brain ischemia domain and have achieved interpretable and useful contrast set. The experiments show how much benefit can be gained from such in depth analysis. The presented approach to the detection of supporting factors enables in depth analysis. This approach nicely supplements contrast set mining and can be also easily implemented in domains with a very large number of attributes (e.g. gene expression domains).

Table 2 Supporting factors for contrast set 2

	CS2	embolic	thrombolic
age high	74.2	69.85	69.29
chol high	6.30	5.69	6.59
fibr high	5.25	4.51	4.85
fo low	0.64	0.76	0.73
af = no	100%	47%	88%
smoke = no	73%	46%	55%

References

1. S. Wrobel (1997) An algorithm for multi-relational discovery of subgroups. In Proc. of the First European Conference on Principles of Data Mining and Knowledge Discovery, 1997, pp. 78–87, Springer
2. Bay S D, Pazzani M J (2001) Detecting group differences: Mining contrast sets. Data Min. Knowl. Discov., 5(3):213–246, 2001.
3. Dong G, Li J (1999) Efficient Mining of Emerging Patterns: Discovering Trends and Differences. In Proc. of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, 1999, pp 43-52
4. decision trees
5. Clark P, Niblett T (1989) The CN2 induction algorithm. Machine Learning, 3(4):261–283, 1989.
6. Gamberger D, Lavrač Nada, Krstačić G (2003) Active subgroup mining: a case study in coronary heart disease risk group detection. Artif. intell. med.. [Print ed.], 2003, vol. 28, pp. 27-57.
7. Kralj P, Lavrač N, Gramberger D, Krstačić A (2007) Contrast Set Mining through Subgroup Discovery: Applied to Brain Ischaemina Data. In proc. of the11th Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2007, in press.
8. Victor M, Ropper A H (2001) Cerebrovascular disease. In Adams and Victor's Principles of Neurology, 2001, pp. 821-924
9. Fürnkranz J (2001) Round robin rule learning. In Proc. of the 18th International Conference on Machine Learning, 2001, pp 146-153
10. Demšar J, Zupan B, Leban G (2004) Orange: From Experimental Machine Learning to Interactive Data Mining, White Paper (www.ailab.si/orange), Faculty of Computer and Information Science, University of Ljubljana.
11. Kavšek B, Lavrač N (2006) APRIORI-SD: Adapting association rule learning to subgroup discovery. Appl. artif. intell., 2006, pp.543-583
12. Gamberger D, Lavrač N, Krstačić G (2003) Active subgroup mining: a case study in coronary heart disease risk group detection. Artif. intell. med., 28:27-57
13. Lowry R (2007) Concepts and applications of inferential statistics. http://faculty.vassar.edu/lowry/webtext.html
14. Lavrač N, Cestnik B, Gamberger D, Flach P (2004) Decision support through subgroup discovery: three case studies and the lessons learned. Mach. learn. [Print ed.], 2004, vol. 57, pp. 115-143.

Author: Petra Kralj
Institute: Jožef Stefan Institute
Street: Jamova 39
City: 1000 Ljubljana
Country: Slovenia
Email: Petra.Kralj@ijs.si