Contrast Set Mining for Distinguishing between Similar Diseases

Petra Kralj[1], Nada Lavrač[1,2], Dragan Gamberger[3], Antonija 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, Draškovićeva 19, 10000 Zagreb, Croatia

Abstract. The task addressed and the method proposed in this paper aim at improved understanding of differences between similar diseases. In particular we address the problem of distinguishing between thrombolic brain stroke and embolic brain stroke as an application of our approach of contrast set mining through subgroup discovery. We describe methodological lessons learned in the analysis of brain ischaemia data and a practical implementation of the approach within an open source data mining toolbox.

1 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.

A special data mining task dedicated to finding differences between contrasting groups is contrast set mining [1]. The goal of our research is to find discriminative differences between two groups of ischaematic brain stroke patients: patients with thrombolic stroke and those with embolic stroke. The problem is introduced in Section 2.

Contrast set mining can be performed by a specialized algorithm STUCCO [1], through decision tree induction and rule learning [13], and—as shown in our recent work—through subgroup discovery [7]. Section 3 presents the results of decision tree induction on our contrast set mining task and discuss advantages and disadvantages of this approach. In Section 4 we show an approach to contrast set mining through subgroup discovery by providing a mathematically correct translation from contrast set mining to subgroup discovery [7] and an implementation of the approach in the Orange [2] open source data mining toolbox. Next

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

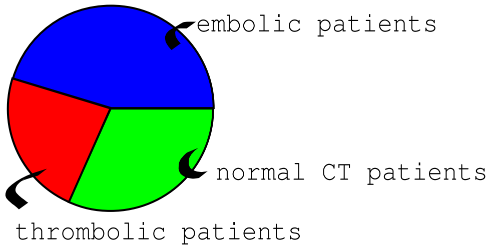


Fig. 1. Distribution of diagnosis of patients in our dataset

we show that the direct "round robin" contrast set mining approach to solve our descriptive induction task leads to rather disappointing results. We discuss the reasons for this undesired performance. This lesson learned resulted in a different, more appropriate "one-versus-all" transformation of contrast set mining to subgroup discovery, justified by the improved results of our experiments, confirmed by the medical expert.

2 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 anywhere 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.

We analyze the brain ischaemia database, which 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 (Figure 1):

- 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. Anamnestic data: aspirin therapy *(asp: yes, no)*, anticoagulant therapy *(acoag: yes, no)*, antihypertensive therapy *(ahyp: yes, no)*, antiarrhytmic therapy *(aarrh: yes, no)*, antihyperlipoproteinaemic therapy – statin *(stat: yes, no)*, hypoglycemic therapy *(hypo: none, yesO – oral, yesI – insulin)*, sex *(m or f)*, age *(in years)*, present smoking *(smok: yes, no)*, stress *(str: yes, no)*, alcohol consumption *(alcoh: yes, no)* and family anamnesis *(fhis: yes, no)*. Physical examination data are: body mass index *(bmi: ref. value 18.5–25)*, systolic blood pressure *(sys: normal value* < 139 *mmHg)*, diastolic blood pressure *(dya: normal value* < 89 *mmHg)* and fundus ocular *(fo: discrete value 0-4)*. Laboratory test data: uric acid *(ua: ref. value for men* < 412 $\mu mol\ L^{-1}$*, for women* < 380 $\mu mol\ L^{-1}$*)*, fibrinogen *(fibr: ref. value 2.0–3.7 g* L^{-1} *)*, glucose *(gluc: ref. value 3.6–5.8 mmol* L^{-1}*)*, total cholesterol *(chol: ref. value 3.6–5.0 mmol* L^{-1}*)*, triglyceride *(trig: ref. value 0.9–1.7 mmol* $L^{-}1$*)*, platelets *(plat: ref. value 150000–400000)* and prothrombin time *(pt: ref. value without th. 0.7–1.2, with anticoagulant th. 0.25–0.4)*. ECG data: heart rate *(ecgfr: ref. value 60–100 beats/min)*, atrial fibrillation *(af: yes, no)* and left ventricular hypertrophy *(ecghlv: yes, no)*.

In this paper, the goal of data analysis is to discover regularities that discriminate between thrombolic and embolic stroke patients. Despite the fact that the immediate treatment for all kinds of ischeamic strokes is the same, the distinction between thrombolic and embolic stroke patients is important in later phases of patient recovery and to better determine the risk factors of the specific diseases.

It must be noted that this dataset does not consist of healthy individuals but of patients with 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 different neurological disorders. The detected relationships can be accepted as generally true characteristics for these patients. However, the computed evaluation measures only reflect characteristics specific to the available data, not necessarily holding for the general population or other medical institutions [12].

3 Searching for Contrast Sets by Decision Tree Induction

Decision trees [11] are a classical machine learning technique. By selecting the attribute that best distinguishes between the classes and putting it as a root node, they partition the examples into subsets of examples where the same method is recursively applied. We have induced a decision tree in Figure 2 for a problem of distinguishing between two types of patients with brain stroke (marked "emb" and "thr") and patients with normal (marked "norm") brain CT test results.[1]

[1] In the experiments we used rigorous pruning parameters to induce small and comprehensible decision trees, using a decision tree learner implemented in the Orange data mining toolbox [2]. Due to noisy data and harsh pruning the decision tree has low classification accuracy (58% accuracy estimated by 10 fold crossvalidation).

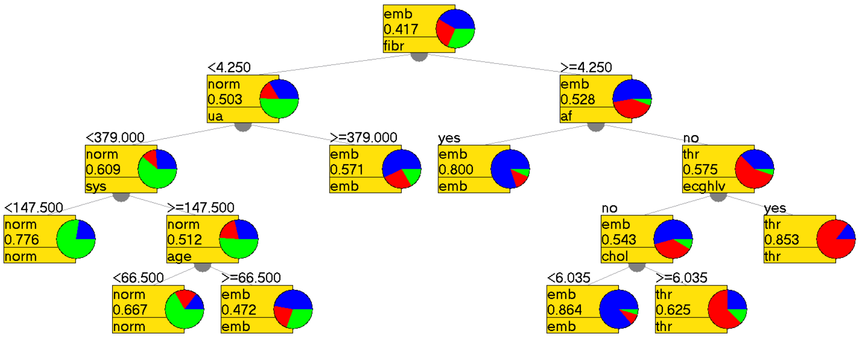


Fig. 2. A decision tree distinguishing between patients with embolic brain stroke, thrombolic brain stroke and patients with normal brain CT test results.

The interpretation of the decision tree by the medical expert is that fibrinogen ("fibr") is the "most informative" attribute distinguishing between patients with and without brain ischaemia, and that atrial fibrillation ("af") is the attribute that best distinguishes between groups of embolic and thrombolic patients. While the induced decision tree well represents the medical knowledge applied in patient diagnosis, the intention of this experiment was not to produce a classifier, but to generate a descriptive model and to investigate the advantages and disadvantages of decision tree induction for contrast set mining.

In the contrast set mining setting, the main advantage of decision trees is the simplicity of their interpretation. On the other hand, there are many disadvantages. A decision tree partitions the space of covered examples, disallowing the overlapping of the discovered patterns. All the contrasting patterns (rules formed of decision tree paths) include the same root attribute (*fibrinogen*), which is disadvantageous compared to contrast set rule representations. Moreover, due to attribute repetitions and thus a limited set of attributes appearing in decision tree paths, the variety of contrasting patterns is too limited.

4 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 [1] 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 [13]. 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 [1].

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

Contrast Set Mining (CSM)	Subgroup Discovery (SD)	Rule Learning (RL)
contrast set	subgroup description	rule condition
group	class (property of interest)	class
attribute value pair	feature	condition
examples in groups $G_1, \ldots G_n$	examples of $Class$ and $\overline{Class}$	examples of $C_1 \ldots C_n$
examples for which contrast set is true	subgroup	covered examples

Table 1. Table of synonyms from different communities.

interested in, find population subgroups that are statistically "most interesting", i.e., are as large as possible and have the most unusual statistical (distributional) characteristics with respect to the given property of interest [14].

Putting these two tasks in a broader rule learning context, note that there are two main ways of inducing rules in multiclass 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 classification 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 [3]). Subgroup discovery is typically performed in a one vs. 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).

Section 4.1 shows that, using a round robin setting, a CSM task can be directly translated into a SD task. The experiments in brain ischaemia data analysis were performed using a novel implementation of our subgroup discovery algorithms in the Orange data mining toolbox, characterized by excellent data and model visualization facilities (see Section 4.2).

The direct transformation of a CSM task into a SD task in the round robin setting showed some problems when used for contrast set mining for distinguishing between thrombotic and embolic patient groups (see Section 5). This lead to a modified task transformation, following the more "natural" one-versus-all subgroup discovery setting (see Section 6).

4.1 Round Robin Transformation: Unifying CSM and SD

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 [4] 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.

Moreover, using the dictionary of Table 1, it is now easy to show that a two-group contrast set mining task $CSM(G_1, G_2)$ can be directly translated into the following two subgroup discovery tasks: $SD(Class = G_1 \text{ vs. } \overline{Class} = G_2)$ and $SD(Class = G_2 \text{ vs. } \overline{Class} = G_1)$. 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. This induction step is as follows:

```
$CSM(G_1, \ldots, G_n)$
   **for** i=1 to n **do**
      **for** j=1, j$\neq$ i to n **do**
            $SD(Class = G_i \ vs. \ \overline{Class} = G_j)$
```

4.2 Implementations of Subgroup Discovery Algorithms and Subgroup Visualization in Orange

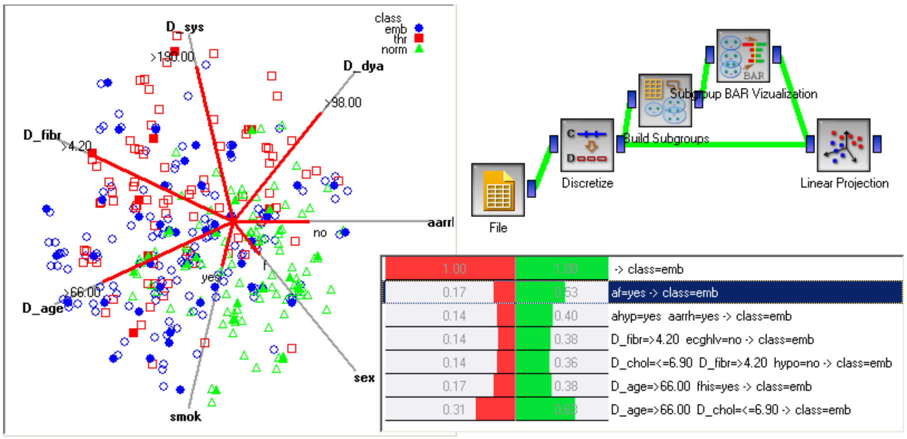
There are several algorithms that are adaptations of rule learners to perform the subgroup discovery task: SD [4], CN2-SD [10] and Apriori-SD [6]. We have reimplemented these algorithms [9] in Orange [2] with some minor adaptations compared to the descriptions in the original papers. The difference arises from the internal representation of the data in Orange, based on attributes and not on features (attribute values). Data need to be discretized in the preprocessing phase, as the implementations construct attribute-value pairs from discretized data on the fly while constructing rules. Despite this data representation limitation, the algorithm reimplementation in Orange is worthy, as it offers various data and model visualization tools and has excellent facilities for building new visualizations.

We here briefly describe just the APRIORI-SD algorithm [6], an adaptation of the algorithm for mining classification rules with association rule learning techniques APRIORI-C [5], which was used in our experiments. The main modifications of APRIORI-C, making it appropriate for subgroup discovery, involve the implementation of an example weighting scheme in rule post-processing, a modified rule quality function incorporating example weights and a probabilistic classification scheme.

Orange goes beyond static visualization, by allowing the interaction of the user and combination of different visualization techniques. In Figure 3 an example of a visual program in the Orange visual programming tool Orange Canvas is shown.[2] The first widget from the left (*File*) loads the dataset (in this example we load the Brain Ischemia dataset with three classes). The following widget (*Discretize*) takes care of data discretization in the preprocessing phase. It is followed by the new widget *Build Subgroups* which is in charge of building subgroups. In this widget the user chooses the algorithm for subgroup discovery and sets the algorithm parameters.

We have implemented a new subgroup visualization technique called the *visualization by bar charts* [8], described in the next paragraph. The widget *Subgroup Bar Visualization* provides the visualization of the subgroups. It can

[2] This visual program is just one example of what can be done by using the Subgroup discovery tool implemented in Orange. Subgroup evaluation and different method for visualizing the contents of subgroups are also available.



1.00	1.00	-> class=emb
0.17	0.53	af=yes -> class=emb
0.14	0.40	ahyp=yes aarrh=yes -> class=emb
0.14	0.38	D_fibr=>4.20 ecghlv=no -> class=emb
0.14	0.36	D_chol=<=6.90 D_fibr=>4.20 hypo=no -> class=emb
0.17	0.38	D_age=>66.00 fhis=yes -> class=emb
0.31	0.63	D_age=>66.00 D_chol=<=6.90 -> class=emb

Fig. 3. An example of a visual program in the interactive interface for subgroup discovery implemented in Orange.

be connected to several other widgets for data visualization. In our case we connected it to existing *Linear Projection* visualization (see the left-hand side of Figure 3) which visualizes the entries of the entire dataset as empty shapes and the entries belonging to the group selected in the *Subgroup Bar Visualization* widget as full shapes. By moving the mouse over a certain shape in the *Linear Projection* widget the detailed description of the entry is displayed.

In the bar chart visualization (shown below the Orange Canvas in Figure 3) the first line's purpose is to visualize the distribution of the entire example set. The area on the right represents the positive examples and the area on the left represents the negative examples. Each following line represents one subgroup. The positive and the negative examples of each subgroup are drawn below the positive and the negative examples of the entire example set. Subgroups are sorted by the relative share of positive examples. Examples of this visualization are shown in Figures 4 and 5.

This visualization method allows simple comparison between subgroups and is therefore useful. It is very intuitive and attractive to end-users. All the displayed data is correct and not misleading. It is very simple and does not display the contents of data, but it can be connected to other data visualizations in Orange (Figure 3) in order to allow in depth investigations.

5 Experimental Evaluation of the Round Robin CSM

The goal of our experiments was to find characteristic differences between patients with thrombolic and embolic ischeamic stroke. We approached this task by applying the round robin transformation from contrast set mining to subgroup discovery, described in Section 4.1. We ran this experiment and asked the experts for interpretation.

The resulting rules mainly include the feature $AF = no$ for thrombolic patients and $AF = yes$ for embolic patients, which are very typical for the corresponding diseases. However, the rules turned out to be non-intuitive to the medical expert. For example, the rule

$$af = yes \ \& \ sys < 185 \ \& \ fo = 1 \rightarrow embolic$$

covering many embolic and just one thrombolic patient (TP =33, FP = 1) was interpreted as "people with suspected thromb in the heart ($af = yes$) and visible consequences of hypertension in the eyes ($FO = 1$)". The feature $sys < 185$ says: patients with not extremely high systolic blood pressure, though high blood pressure is characteristic for both the diseases and the boundary 185 is very high, since everything above 139 is considered high in medical practice.[3]

We investigated further the reasons why the rules were difficult to interpret for domain experts. The reason comes from the task itself: Medical physicians are not used to distinguish between two types of disease given the condition that a patient has a disease, but are rather used to find characteristics for a specific type of a disease compared to the entire population. Another motivation is to avoid rules as

$$fhis = yes \ \& \ smok = yes \ \& \ asp = no \ \& \ dya < 112.5 \rightarrow embolic$$

This rule has good covering properties (TP=28, FP=4), but practically describes healthy people with family history of brain stroke. It is undoubtedly true that this pattern is present in the dataset, but it is not the reason why these patients have a certain type of disease. The algorithm just could not know that the combination of these features is not characteristic for group differentiation simply because it did not have normal people as a reference.

6 Experimental Evaluation of the One-Versus-All CSM

As the medical expert was not satisfied with the results of the comparison of thrombolic and embolic patients, we investigated the reasons and learned a lesson in medical contrast set mining. To overcome the problems related to the original definition of contrast set mining we need to modify the task: instead of using the round robin approach where we compare classes pairwise, we use a one vs. all approach which is standard in subgroup discovery. In this way we give the algorithm also the information about healthy patients.

Our dataset is composed of three groups of patients, as described in Section 2 and shown on Figure 1. An approach we claim is applicable in many similar domains where the differences between two varieties of one disease are as follows: 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).

[3] In our dataset there are 56 patients with $sys > 185$.

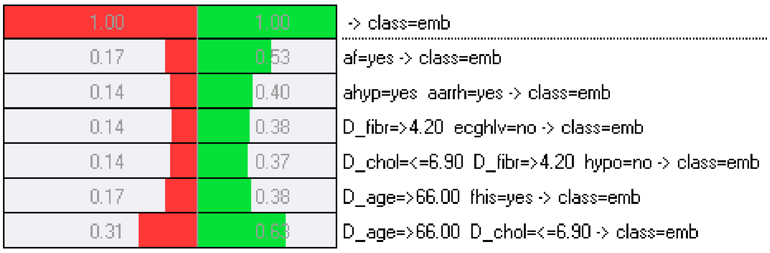


Fig. 4. Characteristic descriptions of embolic patients displayed in the bar chart subgroup visualization: on the right side the positive cases, in our case embolic patients, and on the left hand side the others - thombolic and normal CT.

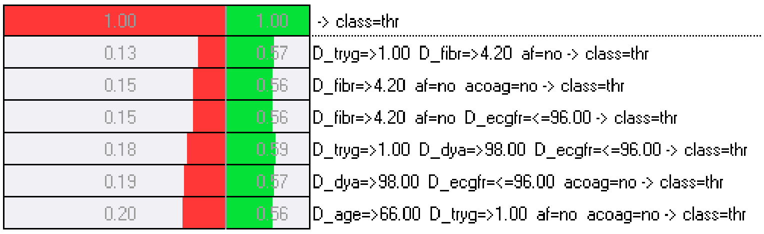


Fig. 5. Characteristic descriptions of thrombolic patients.

In this setting, we ran the experiment with Orange implementation of Apriori-SD. We used the following parameter values: minimal support = 15%, minimal confidence = 30%, the parameter for tuning the covering properties k = 5. The results are displayed in Figures 4 and 5.

Strokes caused by embolism are most commonly caused by heart disorders. The first rule displayed on Figure 4 has only one condition confirming atrial fibrillation ($af = yes$) as an indicator for embolic 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 5 the features presenting diabetes do not appear. The rules rather describe patients without heart or other disorders but with elevated diastolic blood pressure and fibrinogen. High cholesterol, age and fibrinogen values appear characteristic for all ischeamatic strokes.

7 Conclusions

This paper has shown that contrast set mining and subgroup discovery are very similar data mining tasks, and has presented approaches to solving a contrast set mining task by decision tree learning and by transforming the contrast set mining problem to a subgroup discovery problem. As shown in [7], the subgroup discovery approach to contrast set mining has several advantages. Its application in brain ischemia data analysis has shown that sometimes the right task to address is one-vs-all contrast set mining rather then the classical round robin formulation of contrast set mining.

Acknowledgment. This work was partially supervised by Blaž Zupan from Faculty of Computer and Information Science, Ljubljana, Slovenia.

References

1. S. D. Bay and M. J. Pazzani. Detecting group differences: Mining contrast sets. *Data Mining and Knowledge Discovery*, 5(3):213–246, 2001.
2. J. Demšar, B. Zupan, and G. Leban. Orange: From experimental machine learning to interactive data mining, white paper (www.ailab.si/orange). Faculty of Computer and Information Science, University of Ljubljana, 2004.
3. J. Fürnkranz. Round robin rule learning. In *ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning*, pages 146–153, 2001.
4. D. Gramberger and N. Lavrač. Expert-guided subgroup discovery: Methodology and application. *Journal of Artificial Intelligence Research*, (17):501–527, 2002.
5. V. Jovanovski and N. Lavrač. Classification rule learning with APRIORI-C. In *Proceedings of the 10th Portuguese Conference on Artificial Intelligence*, pages 44–51, 2001.
6. B. Kavšek and N. Lavrač. APRIORI-SD: Adapting association rule learning to subgroup discovery. *Applied Artificial Intelligence*, pages 543–583, 2006.
7. P. Kralj, N. Lavrač, D. Gramberger, and A. Krstačić. Contrast set mining through subgroup discovery applied to brain ischaemia data. In *PAKDD '07: Proceedings of the 11th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2007.
8. P. Kralj, N. Lavrač, and B. Zupan. Subgroup visualization. In *IS '05: Proceedings of the 8th International Multiconference Information Society*, pages 228–231, 2005.
9. P. Kralj, Nada Lavrač, Blaž Zupan, and Dragan Gramberger. Experimental comparison of three subgroup discovery algorithms: Analysing brain ischemia data. In *IS '05: Proceedings of the 8th International Multiconference Information Society*, pages 220–223, 2005.
10. N. Lavrač, B. Kavšek, P. Flach, and L. Todorovski. Subgroup discovery with CN2-SD. *Journal of Machine Learning Research*, 5:153–188, 2004.
11. J. R. Quinlan. *C4.5: Programs for Machine Learning.* Morgan Kaufman Publishers Inc, 1993.
12. M. Victor and A. H. Ropper. Cerebrovascular disease. In *Adams and Victor's Principles of Neurology*, pages 821–924, 2001.
13. G. I. Webb, S. Butler, and D. Newlands. On detecting differences between groups. In *KDD '03: Proceedings of the 9th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 256–265, 2003.
14. S. Wrobel. An algorithm for multi-relational discovery of subgroups. In *Proceedings of the 1st European Conference on Principles of Data Mining and Knowledge Discovery*, pages 78–87, Springer, 1997.