Contrast Set Mining through Subgroup Discovery Applied to Brain Ischaemina Data

Petra Kralj[1], Nada Lavrač[1,2], Dragan Gamberger[3], Antonija Krstačić[4] ⋆

[1] Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
[2] Nova Gorica Polytechnic, 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. Contrast set mining aims at finding differences between different groups. This paper shows that a contrast set mining task can be transformed to a subgroup discovery task whose goal is to find descriptions of groups of individuals with unusual distributional characteristics with respect to the given property of interest. The proposed approach to contrast set mining through subgroup discovery was successfully applied to the analysis of records of patients with brain stroke (confirmed by a positive CT test), in contrast with patients with other neurological symptoms and disorders (having normal CT test results). Detection of coexisting risk factors, as well as description of characteristic patient subpopulations are important outcomes of the analysis.

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. Medical data analysis is frequently performed by applying rule learning, as the induced rules are easy to be interpreted by human experts.

The goal of standard classification rule learners [5] is to induce classification/prediction models from labeled examples. Opposed to these *predictive induction* algorithms which induce a model in the form of a set of rules, *descriptive induction* algorithms aim to discover individual patterns in the data, described in the form of individual rules. Descriptive induction algorithms include association rule learners [1], and subgroup discovery systems [2, 6, 8, 11].

This paper addresses a data analysis task where groups of labeled examples are given and the goal is to find differences between the groups. This data analysis task, named *contrast set mining*, was first presented in [3]. In this paper we

⋆ 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".

propose to solve this task by transforming the contrast set mining task to a subgroup discovery task and to apply the subgroup discovery methodology to solve the task. This approach solves some open issues of existing contrast set mining approaches, like dealing with continuous valued attributes, choosing an appropriate search heuristic, selecting the level of generality of induced rules, avoiding of overlapping rules, and presenting the results to the end-users.

Although the goals of contrast set mining, which aims at finding differences between contrasting groups, and subgroup discovery, which aims at finding descriptions of population subgroups, seem different, this paper proves that the goals are the same and the results can be interpreted in both ways. The proposed approach of contrast set mining through subgroup discovery (presented in Section 4) was applied to a real-life problem of analyzing patients with brain ischaemia (presented in Section 2), where we provide insightful data analysis helping to answer questions about the severity of the brain damage based on risk factors obtained from physical examination data, laboratory test data, ECG data and anamnestic data. The usefulness of the approach is shown by the achieved results (Section 5) interpreted by medical specialists.

2 Brain Ischaemia Data

The brain ischaemia dataset 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 dataset: 209 with the computed tomography (CT) confirmed diagnosis of brain stroke, and 91 patients who entered the same hospital department with adequate neurological symptoms and disorders, but were diagnosed 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). In this paper, the goal of the experiments is to characterize brain stroke patients confirmed by a positive CT test in contrast with the patients with a normal CT test.

Patients are described with 26 descriptors representing anamnestic, physical examination, laboratory test and ECG data, and their diagnosis. Anamnestic data: aspirin therapy *(asp)*, anticoagulant therapy *(acoag)*, antihypertensive therapy *(ahyp)*, antiarrhytmic therapy *(aarrh)*, antihyperlipoproteinaemic therapy - statin *(stat)*, hypoglycemic therapy *(hypo)*, sex *(sex)*, age *(age)*, present smoking *(smok)*, stress *(str)*, alcohol consumption *(alcoh)*, family anamnesis *(fhis)*. Physical examination data are: body mass index *(bmi)*, systolic blood pressure *(sys)*, diastolic blood pressure *(dya)*, fundus ocular *(fo)*. Laboratory test data: uric acid *(ua)*, fibrinogen *(fibr)*, glucose *(gluc)*, total cholesterol *(chol)*, triglyceride *(trig)*, platelets *(plat)*, prothrombin time *(pt)*. ECG data: heart rate *(ecgfr)*, atrial fibrillation *(af)*, left ventricular hypertrophy *(ecghlv)*.[1]

The diagnosis of patients is based on the physical examination confirmed by the CT test. All the patients in the control group have a normal brain CT

[1] Details can be found on http://lis.irb.hr/PAKDD2007paper/.

test in contrast with the positive CT test of patients with a confirmed brain attack. It should be noted that the group of patients with brain stroke and the control group do not consist of healthy persons but of patients with suspected severe neurological symptoms and disorders. In this sense, the available dataset is particularly appropriate for studying the specific characteristics and subtle differences that distinguish the two groups. While the detected relationships can be accepted as the actual characteristics for these patients, the computed evaluation measures—including probability, specificity and sensitivity of induced rules—only reflect characteristics specific to the available data set, not necessarily holding for the general population or other medical institutions.

3 Methodological Background

A common question of exploratory data analysis is "What is the difference between the given groups?" where the groups are defined by a selected property of individuals that distinguishes one group from the others. For example, the distinguishing property that we want to investigate could be the gender of patients and a question to be explored can be "What is the difference between males and females affected by a certain disease?" or, if the property of interest was the response to a treatment, the question can be "What is the difference between patients reacting well to a selected drug and those that are not?" Searching for differences is not limited to any special type of individuals: we can search for differences between molecules, patients, organizations, etc.

Data analysis tasks that try to find differences between contrasting groups are very common and the approach presented here can be applied in many of these tasks. When the end-users ask for differences characterizing different groups, they are usually not interested in all the differences; they may prefer a small set of representative and interpretable patterns. Finding all the patterns that discriminate one group of individuals from the other contrasting groups is not appropriate for human interpretation. Therefore, as is the case in other descriptive induction tasks, the goal is to find descriptions that are unexpected and interesting to the end-user.

The approach presented in this paper offers this kind of analysis. From a dataset of class labeled instances (the class label being the property of interest) by means of subgroup discovery [7] we can find interpretable rules that offer a good starting point for human analysis of contrasting groups.

Contrast set mining. The problem of mining contrast sets was first defined in [3] as finding "conjunctions of attributes and values that differ meaningfully in their distributions across groups." They proposed the STUCCO algorithm [3], which is based on Bayardo's Max-Miner [4] rule discovery algorithm. In the level-wise search for contrast sets, formed of conjunctions of attribute-value pairs of length i, the interestingness of the conjunct is estimated by its statistical significance, assessed using a χ^2 test with a Bonferroni correction. Domain specific parameters need to be set, like the minimum support difference between groups. The algorithm works only on domains with nominal attributes.

It was shown in [10] that contrast set mining is a special case of a more general rule learning task, and that a contrast set can be interpreted as an antecedent of a rule and $Group_i$, for which it is characteristic, as the rule consequent: $ContrastSet \rightarrow Group_i$.

When using rule learners (OPUS-AR and C4.5 rules) for contrast set mining [10], the user needs to select a quality measure (choosing between support, confidence, lift, coverage and leverage). In this setting the number of generated rules largely exceeds the number of rules generated by STUCCO, unless pruned by the user-defined maximum number of rules parameter. Expert interpretation of rules is difficult due to a large amount of rules and sometimes also their specificity.

Subgroup discovery. A subgroup discovery 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', e.g. are as large as possible and have the most unusual statistical (distributional) characteristics with respect to the property of interest" [11]. The result of subgroup discovery is a relatively small set of *subgroup descriptions* formed of conjunctions of features. Members of a subgroup are examples from the dataset that correspond to the subgroup description. Good subgroups are large (descriptions covering many examples with the given property of interest), and have a significantly different distribution of examples with the given property compared to its distribution in the entire population.

Subgroup discovery algorithms include adaptations of rule learning algorithms to perform subgroup discovery [7, 8] algorithms for relational subgroup discovery [9, 11] and algorithms for exploiting background knowledge for discovering non-trivial subgroups [2], among others.

Since subgroup descriptions are conjunctions of features that are characteristic for a selected class of individuals (property of interest), a subgroup description can be seen as a condition of a rule $SubgroupDescription \rightarrow Class$ and therefore subgroup discovery can be seen as a special case of a more general rule learning task.

4 Contrast Set Mining through Subgroup Discovery

We present an approach to contrast set mining by means of subgroup discovery. Even though the definitions of subgroup discovery and contrast set mining seem different, we here provide a proof of the compatibility of the tasks. Furthermore, by subgroup discovery means, we solve the following open issues in contrast set mining [10]: proposing appropriate heuristics for identifying interesting contrast sets, appropriate measures of quality of contrast sets, and appropriate methods for presenting contrast sets to the end-users. The issue of dealing with continuous attributes is also solved by subgroup discovery algorithm SD [7].

Translating contrast set mining tasks to subgroup discovery tasks.

Contrast set mining and subgroup discovery were developed in different communities, each developing its own terminology that needs to be clarified before

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, G_2 $(G_3 \ldots G_n)$	examples of $Class$ and $\overline{Class}$	examples of C_1, C_2 $(C_3 \ldots, C_n)$
examples for which contrast set is true	subgroup	covered examples
support of contrast set on G_1	true positive rate	true positive rate
support of contrast set on G_2	false positive rate	false positive rate

Table 1. Table of synonyms from different communities.

proceeding. In order to show the compatibility of contrast set mining and subgroup discovery tasks, we first define the *compatibility* of terms used in different communities as follows: terms are compatible if they can be translated into equivalent logical expressions and if they bare the same meaning, i.e., if terms from one community can replace terms used in another community.

To show that terms used in contrast set mining (CSM) can be translated to terms used in subgroup discovery (SD), Table 1 provides a term dictionary through which we translate the terms used in CSM and SD into a unifying terminology of classification rule learning.

We now wish to show that every contrast set mining task (CSM) can be translated into a subgroup discovery task (SD). The definitions of contrast set mining and subgroup discovery appear different: contrast set mining searches for discriminating characteristics of groups called contrast sets, while subgroup discovery searches for subgroup descriptions.

A contrast set is formally defined as follows: Let $A_1, A_2, ..., A_k$ be a set of k variables called attributes. Each A_i can take values from the set $\{v_{i1}, v_{i2}, ..., v_{im}\}$. A contrast set is a conjunction of attribute value pairs defined on user defined groups $G_1, G_2, ..., G_n$ of data instances, whose characteristics we wish to uncover through contrast set mining [3]. A special case of contrast set mining considers only two contrasting groups G_1 and G_2. In such cases, we wish to find characteristics of one group discriminating it from the other and vice versa.

In subgroup discovery, subgroups are described as conjunctions of features of the form $A_i = v_{ij}$ for nominal attributes, and $A_l > value$ or $A_l \leq value$ for continuous attributes. The subgroup discovery task aims at finding population subgroups that are as large as possible and have the most unusual statistical (distributional) characteristics with respect to the property of interest [11].

Using the dictionary of Table 1 it is trivial 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 $SD(Class = G_2 \text{ vs. } \overline{Class} = G_1)$. Since this translation is possible for two-group contrast set mining, it is—by induction—also possible for a general contrast set mining task.

g	rule	stroke 209	normal 91
g=10	(fibr > 4.45) and (age > 64) → stroke	41 %	0 %
	(af = yes) and (ahyp = yes) → stroke	28 %	5 %
	(str = no) and (alcoh = yes) → stroke	28 %	5 %
	(ahyp = no) and (fibr ≤ 4.55) and (dya ≤ 95.5) → normal	6 %	36 %
	(fibr ≤ 4.55) and (af = no) and (stat=no) and (dya ≤ 95.5) and (age ≤ 70) → normal	8 %	42 %
	(age ≤ 61) and (Fhis = no) and (asp = yes) → normal	0 %	12 %
g=50	(ahyp = yes) → stroke	74 %	46 %
	(fibr > 3.35) and (age > 58) → stroke	79 %	37 %
	(age > 52) and (asp = no) → stroke	64 %	37 %
	(fibr ≤ 4.55) and (af = no) and (fo ≤ 1) and (RRsys ≤ 190) → normal	25 %	71 %
	(fibr ≤ 4.55) and (fo ≤ 1) and (acoag = no) and (age ≤ 75) → normal	37 %	82 %
	(age ≤ 70) and (str = yes) and (stat = no) and (RRsys ≤ 190) → normal	19 %	54 %

Fig. 1. Contrast sets for groups (classes) brain stroke and normal, induced for g-values 10 and 50, visualized with the bar visualization.

Solving open issues of CSM with SD.

In this paper, contrast set mining is performed by subgroup discovery algorithm SD [7], an iterative heuristic beam search rule learner.

Handling continuous attributes: SD uses a feature-based data representation, where attribute values needed for the construction of features are generated automatically from the data. In this way, the SD algorithm overcomes a deficiency of CSM: handling of continuous attributes.

Rule quality heuristic: At each run, the SD algorithm finds subgroups for a selected property of interest and a selected generalization parameter g. The output of the SD algorithm is a set of rules with good covering properties on the given example set, which is obtained by using rule quality heuristic $q_g(R) = \frac{TP}{FP+g}$, where TP (true positives) denotes the number of covered examples from the positive class, FP (false positives) the number covered negative examples, and generalization parameter g offers the user the opportunity to influence the degree of specificity of rules, since with large g general rules are preferred by the q_g heuristic, while with small g each covered negative example is severely punished thus generating specific rules.[2]

Rule diversity: To obtain diverse rules in different iterations, the algorithm implements weighting of covered positive examples after selecting a rule. Instead of the unweighted $q_g(R)$ measure, the weighted rule quality measure replaces TP with the sum of weights of covered positive examples. Although this approach can not guarantee the statistical independence of generated rules, it aims at ensuring good diversity of induced rules. This can be verified also from the results presented in the following section.

Presenting the results to end-users: In the next section we present some visualization methods with the results of our experiments. The visualizations proved to be intuitive and useful to the domain experts, and can help estimating the quality of the results.

[2] Generalization parameter values are usually selected in the range between 1 and 100; in our experiments values 10 and 50 were used.

5 Results of Brain Ischaemia Data Analysis

In this section we illustrate the usage of the presented approach of contrast set mining through subgroup discovery including the visualizations of the results.

There are several questions that medical doctors find interesting and that can be investigated by using the presented method and dataset. Due to space restrictions of this paper, we concentrate only on the question *"What is the difference between patients with confirmed stroke and patients with other severe neurological disorders?"* Other questions that could be addressed in a similar manner are: "What is the difference between patients with TIA and RIND and the confirmed stroke patients?", "What is the difference between patients with thrombolic ischaemia and embolic ischaemia", and others.

For each of the two classes, Figure 1 shows three best rules induced by selecting $g = 10$ and $g = 50$, visualized with the bar visualization along with their TP and FP values. The order of rules is selected by the iterative SD algorithm and is determined by the q_g rule quality value that takes into account the covering relations between the current rule and other rules previously selected for the same g-value.

An interesting subgroup description is rule (age>52.00) *and* (asp=no), which stimulated the analysis presented in Figure 2. This analysis provides an excellent motivation for patients to accept prevention based on aspirin therapy, as the rule explicitly recognizes the importance of the aspirin therapy for persons older than 52 years.

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 58 years, and fibrinogen over 3.35. In the case of fibrinogen, reference values above 3.7 are treated as positive while rules induced for brain stroke domain suggest 4.45 in combination with age over 64 years, and 3.35 in combination with age over 58 years for more sensitive detection of stroke. 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, the fact that various boundary points can be suggested in combinations with different conditions is better than the 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.

6 Conclusions

This work demonstrates that subgroup discovery methodology is appropriate for solving contrast set mining tasks. It shows the results of contrast set mining through subgroup discovery applied to the problem of distinguishing between patients with and without brain stroke. Attention was devoted also to the selection of appropriate visualizations, enabling effective presentations of obtained results. The presented theory and experimental results show that using subgroup discovery for contrast set mining solves many open issues of contrast set mining.

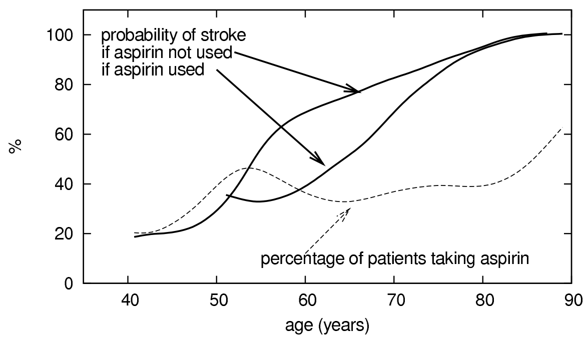

Fig. 2. The probability of brain stroke, estimated by the proportion of stroke patients, shown in dependence of patient age presented for patients taking aspirin as the prevention therapy, and the probability of stroke for patients without this therapy. The percentage of patients with the aspirin therapy is presented by a dashed line.

References

1. R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo. Fast discovery of association rules. *Advances in Knowledge Discovery and Data Mining*, AAAI Press:307–328, 1996.
2. M. Atzmueller, F. Puppe, and H.P. Buscher. Exploiting background knowledge for knowledge-intensive subgroup discovery. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI-05)*, pages 647–652, 2005.
3. S. D. Bay and M. J. Pazzani. Detecting group differences: Mining contrast sets. *Data Min. Knowl. Discov.*, 5(3):213–246, 2001.
4. R. J. Bayardo. Efficiently mining long patterns from databases. In *SIGMOD '98: Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data*, pages 85–93. ACM Press, 1998.
5. P. Clark and T. Niblett. The CN2 induction algorithm. *Machine Learning*, 3(4):261–283, 1989.
6. D. Gamberger and N. Lavrač. Descriptive induction through subgroup discovery: A case study in a medical domain. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pages 163–170, Morgan Kaufmann, 2002.
7. D. Gramberger and N. Lavrač. Expert-guided subgroup discovery: Methodology and application. *Journal of Artificial Intelligence Research*, (17):501–527, 2002.
8. N. Lavrač, B. Kavšek, P. Flach, and L. Todorovski. Subgroup discovery with CN2-SD. *Journal of Machine Learning Research*, 5:153–188, 2004.
9. F. Železný and N. Lavrač. Propositionalization-based relational subgroup discovery with RSD. *Machine Learning*, 62:33–63, 2006.
10. G. I. Webb, S. Butler, and D. Newlands. On detecting differences between groups. In *KDD '03: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 256–265, New York, NY, USA, 2003. ACM Press.
11. S. Wrobel. An algorithm for multi-relational discovery of subgroups. In *Proceedings of the First European Conference on Principles of Data Mining and Knowledge Discovery*, pages 78–87, Springer, 1997.