Journal of Biomedical Informatics 42 (2009) 113–122

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

CSM-SD: Methodology for contrast set mining through subgroup discovery ☆

Petra Kralj Novak [a,*], Nada Lavrač [a,b], Dragan Gamberger [c], Antonija Krstačić [d]

[a] Department of Knowledge Technologies, Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
[b] University of Nova Gorica, Vipavska 13, 5000 Nova Gorica, Slovenia
[c] Rudjer Bošković Institute, Bijenička 54, 10000 Zagreb, Croatia
[d] University Hospital of Traumatology, Draškovićeva 19, 10000 Zagreb, Croatia

ARTICLE INFO

Article history:
Received 11 February 2008
Available online 22 August 2008

Keywords:
Contrast set mining
Subgroup discovery
Supporting factors
Descriptive rules
Brain ischemia

ABSTRACT

This paper addresses a data analysis task, known as contrast set mining, whose goal is to find differences between contrasting groups. As a methodological novelty, it is shown that this task can be effectively solved by transforming it to a more common and well-understood subgroup discovery task. The transformation is studied in two learning settings, a one-versus-all and a pairwise contrast set mining setting, uncovering the conditions for each of the two choices. Moreover, the paper shows that the explanatory potential of discovered contrast sets can be improved by offering additional contrast set descriptors, called the supporting factors. The proposed methodology has been applied to uncover distinguishing characteristics of two groups of brain stroke patients, both with rapidly developing loss of brain function due to ischemia:those with ischemia caused by thrombosis and by embolism, respectively.

1. Introduction

The goal of automated data analysis is to construct models or discover interesting patterns in the data. In many domains, including medical data analysis, model construction and pattern discovery are frequently performed by rule learning, as the induced rules are easy to be interpreted by human experts. The standard classification rule learning task is to induce classification/prediction models from labeled examples [4]. Opposed to *predictive rule induction*, which goal is to induce a model in the form of a set of rules, the goal of *descriptive rule induction* is to discover individual patterns in the data, described in the form of individual rules. Descriptive induction algorithms include association rule learners [1], clausal discovery algorithms [20,19], as well as contrast set mining [3,24] and subgroup discovery algorithms [25,8,17,2].

This paper addresses a data analysis task where groups of examples are given and the goal is to find differences between these contrasting groups. This data analysis task, named *contrast set mining*, was first presented in Ref. [3]. We transform the contrast set mining task to a subgroup discovery task [25,8,17,2], whose goal is to find descriptions of groups of individuals with unusual distributional characteristics with respect to the given property of interest. By doing so, this paper shows that even though the contrast set mining and subgroup discovery tasks are different, subgroup discovery techniques can de used to achieve the goal of contrast set mining. It also shows that the subgroup discovery approach to contrast set mining—as implemented in the Orange [6] open source data mining toolbox—can solve some open issues of existing contrast set mining approaches, like choosing an appropriate search heuristic, selecting the level of generality of induced rules, avoiding of overlapping rules, and presenting the results to the end-user.

The formally justified pairwise transformation of contrast set mining to subgroup discovery—called the *round robin* subgroup discovery approach to contrast set mining—is performed pairwise, for every pair of contrasting groups (i.e., for every pair of classes in a multi-class problem setting). This setting can, however, in some circumstances lead to poor results. The analysis of the reasons for this undesired performance has triggered the development of an alternative method, called the *one-versus-all* transformation of contrast set mining to subgroup discovery, justified by improved results in our experiments, as confirmed by the medical expert.

We argue that a descriptive induction task should not be concluded when individual rules are discovered, as the discovered rules typically uncover only the principal characteristics of the analyzed groups. To enable a better interpretation and improve the understanding of the uncovered characteristics, other properties that support the extracted rules are also important. In subgroup discovery these additional properties are called the *supporting factors* [10]. In this paper we adapt the concept of

☆ 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".

* Corresponding author. Fax: +386 1 477 3315.
E-mail address: Petra.Kralj@ijs.si (P. Kralj Novak).
URL: http://kt.ijs.si/PetraKralj (P. Kralj Novak).

1532-0464/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2008.08.007

supporting factors from subgroup discovery to contrast set mining, to fit the definition and the goals of contrast set mining.

The proposed approach to contrast set mining through subgroup discovery is in this paper applied to a real-life problem of analyzing a dataset of patients with brain ischemia, where the goal of data analysis is to determine the type of brain ischemia from risk factors obtained from anamnesis, physical examination, laboratory tests and ECG data. The achieved results are interpreted by a medical specialist.

This paper is organized as follows. Section 2 presents the background technologies: contrast set mining and subgroup discovery. Section 3 provides the motivation for the new approach of contrast set mining through subgroup discovery, by presenting the brain ischemia data analysis problem, and the motivation for developing specific techniques for contrast set mining (illustrated by the shortcomings of the standard machine learning techniques). This section also presents the implementation and a novel method for contrast set visualization. Section 4 provides a unifying view on contrast set mining and subgroup discovery by unifying the terminology, the tasks and the rule quality measures. In Section 5 we present the experiments performed on the brain ischemia data and a refinement of the contrast set mining setting that is appropriate for distinguishing between similar diseases. Section 6 is dedicated to supporting factors as a mechanism to improve the explanatory potential of contrast set mining.

2. Background technologies: contrast set mining and subgroup discovery

Data analysis tasks that try to find differences between contrasting groups are very common. When end-users are interested in analyzing different groups, they are usually not interested in analyzing all the patterns that discriminate one group of individuals from the other contrasting groups, as the interpretation of large amounts of patterns is too difficult. They typically prefer a small set of representative and interpretable patterns that are novel, potentially interesting and preferably unexpected.

This paper investigates two approaches to finding interesting group descriptors: contrast set mining and subgroup discovery. Contrast set mining is a data mining technique specifically developed for finding differences between contrasting groups (described in Section 2.1). Subgroup discovery is aimed at finding descriptions of interesting subgroups in the data (described in Section 2.2). In Section 4.1 we show how to unify the terminology used in these two—until now separate—areas of research.

2.1. Contrast set mining

The problem of mining contrast sets was first defined in [3] as finding contrast sets as "conjunctions of attributes and values that differ meaningfully in their distributions across groups". Our definitions are epitomized from [24], which are based on the definitions from [3] with some notational differences for better enabling the comparison with subgroup discovery. Let $A_1, A_2, \ldots, A_k$, be a set of k variables called attributes. Each A_i can take on values from the set $\{v_{i1}, v_{i2}, \ldots, v_{im}\}$. Given a set of mutually exclusive user defined groups $G_1, G_2, \ldots, G_n$ of data instances, a *contrast set* is a conjunction of attribute–value pairs (with no A_i occurring more than once). A contrast set is equivalent to an itemset in association-rule discovery when applied to attribute–value data. Similar to an itemset, we measure the support of a contrast set. However, support is defined with respect to each group. The *support* of a contrast set X with respect to a group G_i is the percentage of examples in G_i for which contrast set X is true (denoted as $support(X, G_i)$).

It was shown in [24] that contrast set mining can be viewed as 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 G_i—for which it is characteristic—as the rule consequent: $X \rightarrow G_i$.

Contrast set discovery seeks to find all contrast sets whose support differs meaningfully across groups. Once all significant (Eq. 1) and large (Eq. 2) contrast sets are found, a subset which is 'interesting' should be presented to the end user [3]. Formally,

$$(X|G_i) \neq p(X|G_j) \tag{1}$$

$$SuppDiff(X, G_i, G_j) = |support(X, G_i) - support(X, G_j)| > \delta \tag{2}$$

where X is the contrast set and δ is a user-defined threshold called the *minimum support-difference*. Contrast sets for which Eq. (1) is statistically supported are called significant and those for which Eq. (2) is satisfied are called large. Note that these are different expressions of the same core principle, that the frequency of the contrast set must differ meaningfully across groups. Eq. (1) provides the basis of a statistical test of 'meaningful', while Eq. (2) provides a quantitative test thereof.

The STUCCO algorithm (Search and Testing for Understandable Consistent Contrasts), proposed in the original contrast set mining paper [3], is based on the Max-Miner rule discovery algorithm [13]. STUCCO discovers a set of contrast sets along with their supports on groups. STUCCO employs a number of pruning mechanisms. A potential contrast set X is discarded if it fails a statistical test for independence with respect to group variable G_i. It is also subjected to what is in [23] called a test for *productivity* which is based on the notion of confidence.[1] A rule $X \rightarrow G_i$ is productive iff

$$\forall Z \subset X : confidence(Z \rightarrow G_i) < confidence(X \rightarrow G_i)$$

that is, a more specific contrast set must have higher confidence than any of its generalizations. Further tests for minimum counts and effect sizes may also be imposed. STUCCO introduced a novel variant of the Bonferroni correction for multiple tests which applies ever more stringent critical values to the statistical tests employed as the number of conditions in a contrast set is increased. When using rule learners (e.g., OPUS-AR and C4.5 rules) for contrast set mining [24], 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 can be difficult due to a large amount of rules and sometimes also due to their specificity.

2.2. Subgroup discovery

The task of subgroup discovery 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 distributional characteristics with respect to the property of interest [25]. The result of subgroup discovery is a set of *subgroup descriptions*, where a subgroup description is a conjunction of *features* defined as follows.

Let $A_1, A_2, \ldots, A_k$, be a set of k variables called attributes. An attribute A_i is categorical if it has a predefined and limited set of possible values $\{v_{i1}, v_{i2}, \ldots, v_{im}\}$ and is continuous if it can take any value within a certain range $[\min, \max]$. Features are of the form $A_i = v_{ij}$ for categorical attributes, and $A_i > value$ or $A_i \leqslant value$ for continuous attributes.

[1] *Confidence* is the proportion of positive examples in all examples covered by the rule. This metric is known under many different names, e.g., confidence in association rule mining, or precision in information retrieval.

Members of a subgroup are instances 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.

Since subgroup descriptions are conjunctions of features that are characteristic for a selected class of individuals (class C, representing the investigated property of interest), a subgroup description can be seen as a condition part of a rule $X \rightarrow C$, therefore subgroup discovery can be seen as a special case of a more general rule learning task.[2]

Subgroup discovery algorithms include adaptations of rule learning algorithms to perform subgroup discovery [9,14,17], algorithms for relational subgroup discovery [22,25] and algorithms for exploiting background knowledge for discovering non-trivial subgroups [2], among others. Presenting subgroup discovery results to end-users has also been explored [15,11].

3. Motivation and methodology overview

This section provides a motivation for the development of a new methodology for contrast set mining. First, it presents the brain ischemia data analysis problem which is used to illustrate the potential of the proposed methodology. Next, it presents results of standard machine learning approaches to distinguishing between patients with stroke due to ischemia caused by thrombosis, patients with stroke due to ischemia caused by embolism, and patients with normal CT test results, and discusses the disadvantages of these approaches when used for distinguishing between these contrasting groups of patients. Finally, it provides the methodology overview by explaining the individual steps of the methodology, and discusses some implementation issues.

3.1. Brain ischemia data analysis problem

Stroke or cerebrovascular accident (CVA) is the clinical designation for a rapidly developing loss of brain function due to a disturbance in the blood vessels supplying blood to the brain. This phenomenon can be due to ischemia caused by thrombosis or embolism, or due to a hemorrhage (bleeding). About 80% of all strokes are ischemic while the remaining 20% are caused by bleeding.

A stroke occurs when blood supply to a part of the brain is interrupted, resulting in tissue death and loss of brain function [21]. 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 atherosclerosis involving selected sites in the extracranial and major intracranial arteries. Cerebral emboli 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.

For simplicity, in this paper we refer to brain stroke due to ischemia caused by embolism as *embolic stroke*, and brain stroke due to ischemia caused by thrombosis as *thrombotic stroke*.

The brain ischemia dataset available for the 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:

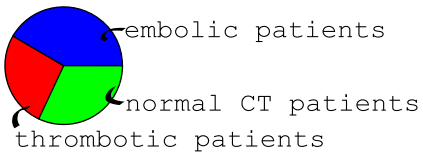


Fig. 1. Distribution of diagnosis of patients in the brain ischemia dataset.

- Two hundred and nine patients with the computed tomography (CT) confirmed diagnosis of stroke: 125 with embolic stroke, 80 with thrombotic stroke and 4 undefined.
- 91 patients who entered the same hospital department with brain stroke neurological symptoms and disorders, but were diagnosed (based on outcomes of neurological tests and CT) as patients with transient ischemic attack (TIA, 33 patients), reversible ischemic neurological deficit (RIND, 12 patients), and severe headache or cervical spine syndrome (46 patients). For simplicity, these patients are referred to as patients with *normal CT*.

The distribution of patients is shown in Fig. 1. Patients are described with their diagnosis and 26 descriptors representing anamnesis, physical examination, laboratory tests data and ECG data. Anamnesis data: aspirin therapy (*asp*), anticoagulant therapy (*acoag*), antihypertensive therapy (*ahyp*), antiarrhytmic therapy (*aarrh*), lipid-lowering therapy—statin (*stat*), hypoglycemic therapy (*hypo*), sex (*sex*), age (*age*), present smoking (*smok*), stress (*str*), alcohol consumption (*alcoh*), family anamnesis (*fhis*). Physical examination data: body mass index (*bmi*), systolic blood pressure (*sys*), diastolic blood pressure (*dya*), and examination of the fundus oculi (*fo*). Laboratory tests data: uric acid (*ua*), fibrinogen (*fibr*), glucose (*gluc*), total cholesterol (*chol*), triglyceride (*trig*), platelets (*plat*), and prothrombin time (*pt*). ECG data: heart rate (*ecgfr*), presence of atrial fibrillation (*af*), and signs of left ventricular hypertrophy (*ecghlv*).

It must be noted that this dataset does not include any healthy individuals but consists 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 between patients with different neurological disorders. The detected relationships can be accepted as generally true characteristics for these patients.[3]

In this paper, the goal of data analysis is to discover regularities that discriminate between thrombotic stroke and embolic stroke patients. Despite the fact that the immediate treatment for both types of ischemic strokes is the same, the distinction between thrombotic stroke and embolic stroke patients is important in later phases of patient recovery and to better determine the risk factors of the specific diseases. An example rule, induced by our methodology of contrast set mining through subgroup discovery, is

$$ahyp = yes \ \texttt{AND} \ aarrh = yes \ \rightarrow \ class = emb$$

This rule, interpreted as "ischemic stroke patients with antihypertensive therapy and antiarrhytmic therapy tend to have emboli as main cause of stroke", represents a contrast set for embolic stroke patients in contrast with thrombotic stroke patients. It should be further interpreted as "since both antihypertensive therapy and antiarrhytmic therapy are therapies for cardiovascular dis-

[2] Notice that in concept learning the task is to find rules that describe concept C. Examples of concept C are considered the positive examples while the others, belonging to $\overline{C}$, are considered the negative examples of concept C.

[3] Not that the computed evaluation measures only reflect characteristics specific to the available database, not necessarily holding for the general population or other medical institutions.

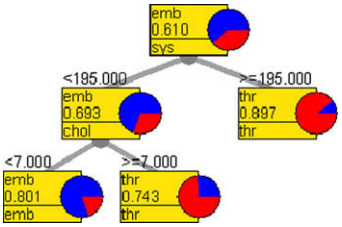


Fig. 2. A strongly pruned decision tree aimed at distinguishing between patients with embolic stroke and thrombotic stroke. Every node of the decision tree is represented by a circle–rectangle pair. In the circle, the distribution of the classes of the examples belonging to the node is visualized. The rectangle contains the information on the majority class of the node (first line), the percentage of the majority class (second line) and, depending on whether the node is an inner node of the tree or a leaf, the attribute to test or the prediction of the leaf.

orders, ischemic stroke patients with cardiovascular disorders tend to have emboli as main cause of stroke". Therapies themselves are, in line with medical knowledge, not causing strokes.

3.2. Motivation for contrast set mining

A common question of exploratory data analysis is "What are the differences between the given groups?" where the groups are defined by a 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 are the differences 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 are the differences 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. In this paper we address the problem of exploring the differences between two groups of ischemic stroke patients: patients with thrombotic stroke and those with embolic stroke.

Despite the availability of specific contrast set mining techniques, some of which adapt classification rule learners to contrast set mining [24], we provide further motivat for the development of our methodology by showing the inadequacy of standard machine learning techniques for contrast set mining. To do so, we use a standard decision tree learner and a standard classification rule learner, and show their shortcomings for contrast set mining.

3.2.1. Inadequacy of decision tree learners for CSM

We used a decision tree learner [18], implemented in the Orange data mining toolbox [6], to induce decision trees shown in Figs. 2 and 3, contrasting between patient groups with embolic stroke (*emb*) and thrombotic stroke (*thr*) with and without the presence of the third group of patients with normal (*normCT*) brain CT test results, respectively. To explore the capability of decision tree learning for contrast set mining we have applied harsh pruning parameters to induce small and comprehensible decision trees from the available data.[4]

Let us evaluate decision tree learning as a potential method 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 several disadvantages. All the contrasting patterns (rules formed of decision tree paths) include the same root attribute, which is disadvantageous compared to contrast set rule representations. Due to attribute repetition and thus a limited set of attributes appearing in decision tree paths, the variety of contrasting patterns is very limited. Another well-known problem of decision trees is their sensitivity to changes in the data: a small change in the training set may completely change the set of attributes appearing in the nodes of the tree.

3.2.2. Inadequacy of Classification Rule Learners for CSM

Classification rules overcome some disadvantages of decision trees. We experimented with JRip; the Java implementation of the Ripper algorithm [5]. From the results in Table 1 we can see that classification rules do not all share the same key feature, but there are other disadvantages of classification rules making them inappropriate for contrast set mining. First, the rules are generated consequently by a covering algorithm, which implies that they also need to be read and interpreted consequently—they are not independent 'chunks of knowledge'. The second disadvantage is the low coverage of classification rules which is undesired in contrast set mining. Last, in the concrete example in Table 1, only the last 'generic' rule has as a consequent embolic stroke patients—in the entire ruleset there is no description of embolic stroke patients at all.

3.3. Overview of the proposed methodology and its implementation

The novel contrast set mining methodology, proposed in this paper, is performed in the following steps:

- preprocess the data (to comply with the data format of the selected data mining toolbox),
- for each target class, transform the contrast set mining problem into adequate subgroup discovery problems (see Section 4),
- induce a set of subgroup descriptions for every subgroup discovery problem,
- list and visualize the induced subgroup descriptions (see Section 5),
- provide additional explanations by inducing the supporting factors (see Section 6), and
- evaluate the results in collaboration with the domain expert.

We here briefly describe the APRIORI-SD subgroup discovery algorithm [14] which was used in our experiments. APRIORI-SD is an adaptation of the APRIORI-C algorithm [12] for mining classification rules with association rule learning techniques. The main modifications of the APRIORI-C classification rule learner, making it appropriate for subgroup discovery, involve the implementation of an example weighting scheme in rule post-processing, a modified weighted relative accuracy heuristic incorporating example weights (see Eqs. 4 and 5 for the original *WRAcc* heuristic and its modification with example weights), and a probabilistic classification scheme. In brief, in APRIORI-SD, the set of potential rules (subgroup descriptions) is generated by executing the APRIORI-C algorithm. When selecting individual rules, APRIORI-SD repeatedly finds a subgroup with the highest weighted relative accuracy (by taking into account example weights) among subgroup description candidates (APRIORI-C rules) and decreases example weights of covered examples. This is repeated until *WRAcc* is greater than zero.

We have chosen to implement the proposed methodology in the Orange data mining toolbox [6]. We implemented three algorithms that are adaptations of rule learners to perform the subgroup discovery task: SD [9], CN2-SD [17] and APRIORI-SD [14] with some minor adaptations compared to the descriptions in the original papers. The implementation differences arise from the internal representation of the data in Orange, based on attributes and not on features (attribute–values). Data need to be dis-

[4] Note that the data is very noisy, hence the induced decision trees have a low classification accuracy: 75.61% accuracy for a two-class problem, and 58% accuracy for a three-class problem, estimated by 10 fold cross-validation, respectively.

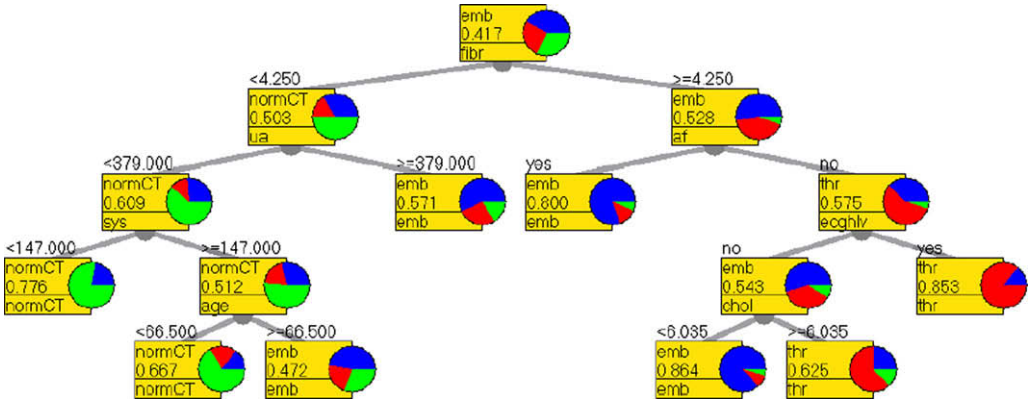


Fig. 3. A pruned decision tree aimed at distinguishing between patients with embolic stroke, thrombotic stroke and patients with normal brain CT test results.

Table 1
Classification rules generated by JRip aimed at distinguishing between patients with embolic stroke, thrombotic stroke and patients with normal brain CT test results

sys ⩾ 200 AND *chol* ⩾ 5.1 → *class* = *thr*
chol ⩾ 7.1 AND *plat* ⩾ 198 → *class* = *thr*
fibr ⩾ 5 AND *af* = *no* AND *ecghlv* = *yes* → *class* = *thr*)
fibr ⩽ 3.8 AND *au* ⩽ 305 → *class* = *normal*
fibr ⩽ 4.2 AND *chol* ⩾ 6.3 → *class* = *normal*
age ⩽ 66 AND *ecgfr* ⩾ 75 AND*pt* ⩾ 0.8 AND *gluc* ⩽ 6.8 → *class* = *normal*
→ *class* = *emb*

Tenfold cross-validated classification accuracy is 65%.

cretized in the preprocessing phase, as the implementations construct attribute–value pairs from discretized data on the fly while constructing the subgroup describing rules. Despite this data representation limitation, the algorithm reimplementation in Orange is valuable, as it offers various data and model visualization tools and has excellent facilities for building new visualizations.

Orange goes beyond static visualization, by allowing the interaction of the user and combination of different visualization techniques. In Fig. 4 an example of a visual program in the Orange visual programming tool Orange Canvas is shown.[5] 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 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.

The widget *Subgroup Bar Visualization* provides the visualization of the subgroups. It can be connected to several other widgets for data visualization. In our case we connected it to the existing *Linear Projection* visualization (see the left-hand side of Fig. 4) 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 a detailed description of the entry is displayed.

4. Proposed methodology: contrast set mining by transformation to subgroup discovery

Even though the definitions of subgroup discovery and contrast set mining appear to be substantially different, this section provides a proof of the compatibility of the two tasks and of the used rule quality measures. It is also shown that by transforming a contrast set mining task to a subgroup discovery task, one can solve the following currently open issues of contrast set mining [24]: selecting the most appropriate heuristics for identifying interesting contrast sets, avoiding of overlapping rules, and presenting contrast sets to the end-user.

4.1. Unifying the terminology of subgroup discovery and contrast set mining

As contrast set mining and subgroup discovery were developed in different research communities, each has developed its own terminology, therefore a common terminology needs to be established before 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 2 provides a term dictionary through which we translate the terms used in CSM and SD into a unifying terminology of rule learning, or more specifically, concept learning. In concept learning, class C is considered as the property of interest and examples with this property as positive examples of C. The negative examples are formed of examples of all other classes.

Note at this point the main terminological and conceptual mismatch between contrast set mining and subgroup discovery. First, in contrast set mining, the *contrasting groups* are the input to the algorithm, while in subgroup discovery, the *subgroups* are the output of the algorithm. Furthermore, in contrast set mining all the contrasting groups have the same importance while in subgroup discovery there is only one *property of interest* and all the terminology is centralized around this property (the true positives, true positive rate, etc.).

4.2. Task transformation

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. Despite these apparent differences this section shows that every contrast set mining task can be translated into a sequence of subgroup discovery tasks.

[5] 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.

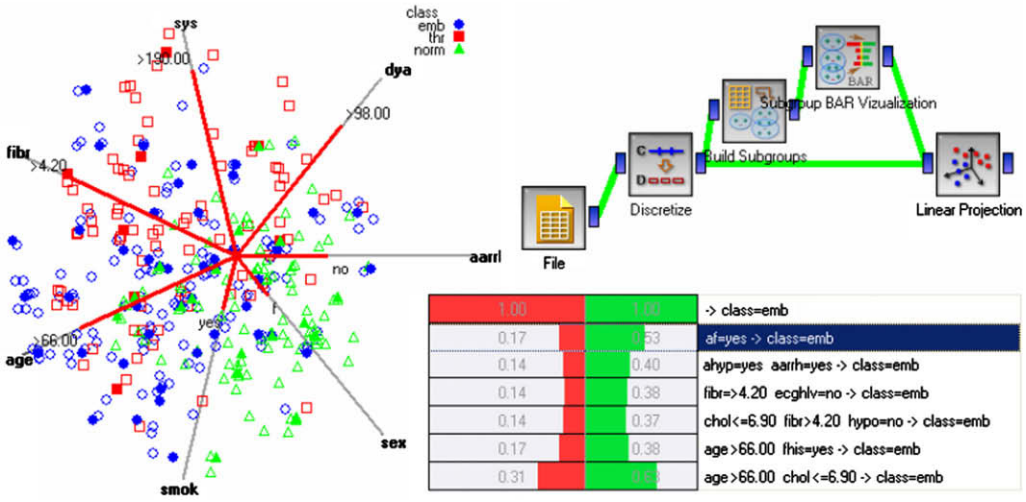


Fig. 4. An example of a visual program in the interactive interface for subgroup discovery implemented in Orange.

Table 2
Synonyms for terms used in contrast set mining and subgroup discovery

Contrast set mining (CSM)	Subgroup discovery (SD)	Rule learning (RL)
Contrast set	Subgroup description	Rule conditions
Groups	Class/property	Classes/concepts
$G_1, \ldots, G_n$	C	$C_1, \ldots, C_n$
Attribute–value pair	Feature	Condition
Examples in groups	Examples of	Examples of
$G_1, \ldots, G_n$	C and $\overline{C}$	$C_1, \ldots, C_n$
Examples for which the contrast set is true	Subgroup of examples	covered examples

A special case of contrast set mining considers only two contrasting groups G_i and G_j. In this situation, the task of contrast set mining is to find characteristics of one group discriminating it from the other and vice versa. Using the dictionary of Table 2 it is trivial to show that a two-group contrast set mining task $CSM(G_i, G_j)$ can be directly translated into the following two subgroup discovery tasks: $SD(C = G_i$ vs. $\overline{C} = G_j)$ and $SD(C = G_j$ vs. $\overline{C} = G_i)$. 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 involving n contrasting groups. The induction step is as follows:

$CSM(G_1, \ldots, G_n)$
  for $i = 2$ to n **do**
    for $j = 1, j \neq i$ to $n - 1$ **do**
      $SD(C = G_i$ vs. $\overline{C} = G_j)$

Putting contrast set mining and subgroup discovery 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 classification rule learners), or alternatively, they search for rules that discriminate between all pairs of classes (known as the *round robin* approach to classification rule learning, proposed in [7]). Subgroup discovery is typically performed in a one-versus-all rule learning setting, typically focusing on generating subgroup descriptions of a single target class. On the other hand, contrast set mining implements a round robin approach (of course, with different heuristics and goals compared to classification rule learning). Note that we have shown above that using a round robin setting, a general n group contrast set mining task can be translated into a sequence of subgroup discovery tasks.

4.3. Compatibility of rule quality measures

Rule quality measures are usually based on the covering property of rules, given the positive (target) class in the rule head. For instance, the true positive rate $TPr(X \rightarrow Y)$ is defined as the percentage of positive examples correctly classified as positive by rule $X \rightarrow Y$, and the false positive rate $FPr(X \rightarrow Y)$ is defined as percentage of negative examples incorrectly classified as positive by rule $X \rightarrow Y$. We illustrate these measures in Table 3 and in Fig. 5.

In this section we show that the rule quality measures *support difference* (*SuppDiff*) used in contrast set mining and *weighted relative accuracy* (*WRAcc*) used in subgroup discovery are compatible, using the following definition of compatibility: rule quality measures h_1 and h_2 are compatible if

$$\forall \text{ pairs of rules } R_i \text{ and } R_j: \; h_1(R_i) > h_1(R_j) \Longleftrightarrow h_2(R_i) > h_2(R_j).$$

A measure of contrast set quality defined in [3] is the support difference (see Eq. 2). We here show that the support difference heuristic can be rewritten, using the dictionary in Table 3 and equations from Fig. 5, as follows:

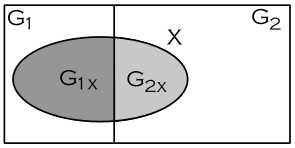
$$\begin{aligned} SuppDiff(X, G_1, G_2) &= support(X, G_1) - support(X, G_2) \\ &= TPr(X \rightarrow G_1) - TPr(X \rightarrow G_2) \\ &= TPr(X \rightarrow G_1) - FPr(X \rightarrow G_1) \end{aligned}$$

where *TPr* and *FPr* denote the true positive rate and the false positive rate, respectively.

Several heuristics have been developed and used in the subgroup discovery community. We will consider here only the *weighted relative accuracy* which is used in subgroup discovery algorithms CN2-SD [17] and APRIORI-SD [14]. The weighted relative accuracy heuristic optimizes two contrasting factors: rule cov-

Table 3
Rule quality measures used in two-group contrast set mining and subgroup discovery, where group G_1 from contrast set mining is considered as property of interest C in subgroup discovery

Contrast set mining (CSM)	Subgroup discovery (SD)	Rule learning (RL)
Groups G_1 and G_2	Classes C and $\overline{C}$	Classes C and $\overline{C}$
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



$$TPr(X \rightarrow G_1) = \frac{|G_{1x}|}{|G_1|} = support(X, G_1)$$
$$FPr(X \rightarrow G_1) = \frac{|G_{2x}|}{|G_2|} = support(X, G2)$$
$$TPr(X \rightarrow G_2) = \frac{|G_{2x}|}{|G_2|} = support(X, G2)$$

Fig. 5. On the left: the large rectangle represents the whole dataset divided into two groups: G_1 and G_2. The ellipse represents the subgroup of examples defined by conditions X. On the right: the formulas for the true and false positive rate, showing that $FPr(X \rightarrow G_1) = TPr(X \rightarrow G_2)$.

erage $p(X)$ (the size of the subgroup), and distributional unusualness $p(Y|X) - p(Y)$ (the difference between the proportion of positive examples in the subgroup describing rule and the proportion of positives in the entire example set). The weighted relative accuracy heuristic is here written in terms of probabilities as follows:

$$WRAcc(X \rightarrow Y) = p(X) \cdot (p(Y|X) - p(Y)) \quad (3)$$

Below we demonstrate that the weighted relative accuracy known from subgroup discovery and the support difference between groups used in contrast set mining are compatible, which is derived as follows.[6]:

$$\begin{aligned} WRAcc(X \rightarrow Y) &= p(X) \cdot [p(Y|X) - p(Y)] = p(Y \cdot X) - p(Y) \cdot p(X) \\ &= p(Y \cdot X) - p(Y) \cdot [p(Y \cdot X) + p(\overline{Y} \cdot X)] \\ &= (1 - p(Y)) \cdot p(Y \cdot X) - p(Y) \cdot p(\overline{Y} \cdot X) \\ &= p(\overline{Y}) \cdot p(Y) \cdot p(X|Y) - p(Y) \cdot p(\overline{Y}) \cdot p(X|\overline{Y}) \\ &= p(\overline{Y}) \cdot p(Y) \cdot [p(X|Y) - p(X|\overline{Y})] \\ &= p(\overline{Y}) \cdot p(Y) \cdot [TPr(X \rightarrow Y) - FPr(X \rightarrow Y)] \end{aligned}$$

Since the distribution of examples among classes is constant for any dataset, the first two factors $p(Y)$ *and* $p(\overline{Y})$ are constant within a dataset. Therefore, when maximizing the weighted relative accuracy, one is maximizing the second factor $[TPr(X \rightarrow Y) - FPr(X \rightarrow Y)]$, which actually is the support difference in a two group contrast set mining problem:

$$\begin{aligned} WRAcc(X \rightarrow Y) &= WRAcc(X \rightarrow G_1) \\ &= p(G_1) \cdot p(G_2) \cdot [support(X, G_1) - support(X, G_2)] \end{aligned}$$

4.4. Solving other contrast set mining open issues through subgroup discovery

Open issues of contrast set mining, identified by [24] are: choosing an appropriate search heuristic (see the solution to this open issue in Section 4.3 above), avoiding of too many overlapping rules, and presenting the results to the end-user. We have also identified dealing with continuous attribute values as an open issue.

4.4.1. Avoiding of too many overlapping rules

Webb et al. [24] show that contrast set mining is a special case of the more general rule discovery task, but the comparison of STUCCO, OPUS_AR and C4.5 shows that rules obtained from standard rule learners are a superset of rules obtained by STUCCO. Moreover, the number of rules generated by OPUS_AR largely exceeds the number of rules generated by STUCCO, unless pruned by the user-defined maximum number of rules parameter.

Complicated pruning mechanisms are used in STUCCO in order to overcome this problem. Pruning of generated contrast sets removes contrast sets that, while significant and large, derive these properties only due to being specializations of more general contrast sets: any specialization is pruned that has similar support to its parent or that fails a χ^2 test of independence with respect to its parent. Details of the relatively complex pruning mechanisms are elaborated in [3].

In subgroup discovery algorithms like CN2-SD [17] this problem is elegantly solved by using the weighted covering approach with the intention to ensure the diversity of rules induced in different iterations. The weighted covering algorithm starts by constructing and selecting the first rule, i.e., the 'best' rule with the highest value of the *WRAcc* heuristic, defined in Eq. 3 and computed as follows:

$$WRAcc(X, Y) = \frac{p+n}{P+N} \cdot \left(\frac{p}{p+n} - \frac{P}{P+N} \right) \quad (4)$$

where p and n are the numbers of covered positive and negative examples (i.e., $p = |TP|$ and $n = |FP|$, the numbers of true positives and false positives, respectively), and P and N are the numbers of all positive and negative examples in the dataset. Having selected the first rule, the weights of positive examples covered by the rule are decreased. To do so, the rules covering each positive example are counted. All example counts $c(e)$ are initially set to 1. The example weights are computed as $w(e) = \frac{1}{c(e)}$, and in each iteration of the algorithm the example counts are recomputed, leading to decreased example weights. For that purpose, the CN2-SD and the APRIORI-SD algorithm use the weighted relative accuracy heuristic, modified with example weights, as defined in Eq. (5) below:

$$WRAcc'(X, Y) = \frac{p'+n}{P'+N} \cdot \left(\frac{p'}{p'+n} - \frac{P}{P+N} \right) \quad (5)$$

where $p\prime = \sum_{TP(R)} w(e)$ is the sum of the weights of all covered positive examples, and $P\prime$ is the sum of the weights of all positive examples.

Although the weighted covering approach cannot guarantee the statistical independence of generated rules, it aims at ensuring good diversity of a relatively small set of rules.

4.4.2. Handling continuous attribute values

Subgroup discovery algorithms SD [9], CN2-SD [17] and APRIORI-SD [14] use a feature-based data representation, where attribute values needed for the construction of features are generated automatically from the data. In this way, subgroup discovery algorithms overcome this deficiency of contrast set mining.

4.4.3. Presenting the results to the end-user

Presenting subgroup discovery results to the end-user is an interesting research problem. Several methods for subgroup visualization have been proposed (see an overview in [11]). When visualizing contrast set mining results on two groups, these methods can be easily adopted without much adaptation. For example, the pie chart visualization can easily be adapted for multi-class visualization, while more advanced visualizations, like the distribution of a subgroup by a continuous attribute, require more inventiveness for being used for multi-class results visualizations.

In this work we propose a new subgroup visualization technique called *visualization by bar charts*, shown in Figs. 6 and 7. In this visualization, the first row is used to visualize the distribution of positive and negative examples in the entire example set. The

[6] These equations were derived by Peter Flach in another context, see [16]

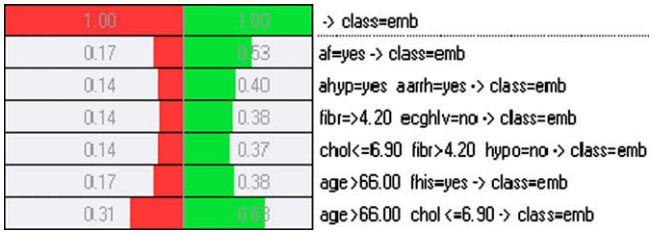


Fig. 6. Characteristic descriptions of embolic stroke patients displayed in the bar chart subgroup visualization: on the right side the positive cases, in our case embolic stroke patients, and on the left hand side the others—thombotic stroke patients and those with normal CT.

area at the right hand side represents the positive examples (one group, in the contrast set mining terminology), and the area at the left hand side represents the negative examples (the other group). The following rows present the induced subgroup descriptions, together with the fractions of positive and negative examples covered. Subgroups are sorted by the relative share of the positive examples in the subgroup.

This visualization method can help estimating the quality of the results by allowing for simple comparisons between subgroups. It is intuitive and simple, and therefore easy to be interpreted by the end-user. However, as this visualization does not display the contents of the data, it should best be used in hand with other visualization methods, e.g., together with those available in the Orange data mining toolbox (see Fig. 4) in order to allow for more detailed exploration.

5. Application of contrast set mining to the problem of distinguishing between similar diseases

The goal of our experiments was to find characteristic differences between patients with embolic and thrombotic stroke. We have approached this problem in three ways: first by standard machine learning algorithms (see Section 3.2), second by the round robin transformation of contrast set mining to subgroup discovery (Section 5.1), and finally by a one-versus-all transformation of contrast set mining to subgroup discovery (Section 5.2). The latter two are outlined below.

5.1. Experimental evaluation of the round robin CSM

To find characteristic differences between patients with embolic and thrombotic stroke we applied the mathematically correct *round robin* transformation from contrast set mining to subgroup discovery, described in Section 4. We ran this experiment and asked the expert for interpretation.

The resulting rules mainly include the feature $af = no$ for thrombotic stroke patients and $af = yes$ for embolic stroke 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

$$af = yes \;\; AND \; sys < 185 \; AND \; fo = 1 \rightarrow class = emb$$

covering many embolic and just one thrombotic stroke patient ($p = |TP| = 33,\ n = |FP| = 1$) was interpreted as *patients with suspected thromb in the heart in atrial fibrillation* ($af = yes$), *visible consequences of hypertension in the eyes* ($fo = 1$), *and with normal or high—but not extremely high (not over 185)—systolic blood pressure.*[7]

[7] High blood pressure is characteristic for both diseases and the boundary 185 is very high, since blood pressure above 139 is already considered high in medical practice. In our dataset there are 56 patients with $sys > 185$.

Fig. 7. Characteristic descriptions of thrombotic stroke patients.

We have further investigated the reasons why the rules were relatively difficult to be interpreted by the medical expert. One reason is the difficulty of the contrast set mining task itself: physicians are not used to distinguish between two types of the disease given the condition that a patient has a disease, but are rather used to find characteristics for a specific disease compared to the entire population. Another reason are rules like the rule listed below:

$$fhis = yes \;\texttt{AND}\; smok = yes \;\texttt{AND}\; asp = no \;\texttt{AND}\; dya < 112.5 \rightarrow class = emb$$

This contrast set describing rule has good covering characteristics ($|TP| = 28,\ |FP| = 4$), but practically describes healthy people with family history of brain stroke. It is undoubtedly true that this pattern is present in the dataset, but the discovered pattern does not describe the reason why these patients are embolic stroke patients; the round robin CSM algorithm could not detect that the combination of these features is not useful for group differentiation from the medical point of view as it simply did not have the normal CT people as a reference. This lesson learned has lead us to the development of a different approach to contrast set mining: the one-versus-all CSM algorithm whose experimental evaluation is described below.

5.2. Experimental evaluation of the one-versus-all CSM

As the medical expert was not satisfied with the results of the comparison of thrombotic and embolic stroke patients induced by the round robin CSM algorithm, we further investigated the reasons for the expert's dissatisfaction 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 definition of the contrast set mining task as addressed in this paper as follows. Instead of using the round robin approach where we compare classes pairwise, we may better use the one-versus-all approach which is standard in classification rule learning and subgroup discovery. In this way we give the algorithm also the information about the normal CT patients.

In particular, in our dataset composed of three groups of patients (as described in Section 3.1 and shown in Fig. 1), to find the characteristics of embolic stroke patients we should perform subgroup discovery on the embolic stroke group compared to the rest of the patients (thrombotic stroke patients and those with a normal CT). Similarly, when searching for characteristics of thrombotic stroke patients, we should compare them to the rest of the patients (those with embolic stroke and those with a normal CT).

In this setting, we ran the experiment with the Orange implementation of APRIORI-SD,[8] and got the results shown in Figs. 6 and 7.

Note that stroke caused by embolism is most commonly caused by heart disorders. The first rule shown in Fig. 6 has only one condition confirming the presence of atrial fibrillation ($af = yes$) as an

[8] We used the following parameter values: minimal support = 15%, minimal confidence = 30%, the parameter for tuning the covering properties $k = 5$.

indicator for embolic 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.

Thrombotic stroke is most common with older people, and often there is underlying atherosclerosis or diabetes. In the rules displayed in Fig. 7 the features presenting diabetes do not appear. The rules describe patients with elevated diastolic blood pressure and fibrinogen, but without heart or other disorders. High cholesterol, age and fibrinogen values appear characteristic for all ischemic strokes.

6. Supporting factors for contrast set mining

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 for distinguishing between the classes. 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 improved human understanding of the principal factors and for the support in decision making processes. This section explores an approach to improving contrast set mining explanatory potential by using supporting factors.

6.1. Supporting factors in subgroup discovery

In subgroup discovery the features that appear in subgroup descriptions are called the *principal factors*, while the additional features that are also characteristic for the detected subgroup are called the *supporting factors* [10]. For every detected subgroup the supporting factors detection process is 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, for the set of all positive examples, and finally for the set of all negative examples (the control set).

The necessary condition for a feature to be determined as 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 χ^2 test of association for categorical attributes. The decision which statistical significance is sufficiently large can depend on the medical context. Typically the cut-off values are set at $p < 0.01$ for the significance with respect to the control set and $p < 0.05$ for the significance with respect to the positive set.

6.2. 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 tasks: in subgroup discovery there is only one property of interest and the goal is to find characteristics common to subgroups of individuals that have this property. On the other hand, in contrast set mining there are several groups of individuals and the goal is to find differences between these groups. Therefore the notion of supporting factor from subgroup discovery cannot be directly adopted for contrast set mining.

Table 4
Supporting factors for contrast set CS1

	CS1	Thrombotic	Embolic
fo high	0.82	0.73	0.76
af = yes	80%	13%	53%
ahyp = yes	100%	81%	70%
aarrh = yes	100%	19%	45%
chol low	5.8	6.59	5.69
rrsys low	159	178	159
rrdya low	92	100	92
ecgfr high	87	77	94
acoag = yes	24%	5%	16%

Table 5
Supporting factors for contrast set CS2

	CS2	Embolic	Thrombotic
age high	74.2	69.85	69.29
chol high	6.3	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%
rrsys high	180	159	178
ecghlv = yes	60%	37%	61%
acoag = no	100%	84%	95%
aarh = no	93%	55%	81%

We propose and show in our experiments a way of generalizing the supporting factors from subgroup discovery to contrast set mining. Since the goal of contrast set mining is to find differences between contrasting groups, there is no need for the values of supporting factors being significantly different from those in the entire positive population. Another difference from subgroup discovery supporting factors is that instead of presenting to the domain expert only the values of supporting factors for the positive class, we also show the distribution (for categorical) or the average (for numeric) attributes for the negative set and for the entire positive set.

Since the interpretation of all the patterns discovered and presented in Section 5.2 is out of the scope of this paper, we focus only on two contrast sets: Contrast set $CS1 : (TPr = 0.4, FPr = 0.14)$

$$ahyp = yes \ \texttt{AND} \ aarrh = yes \rightarrow class = emb$$

Contrast set $CS2 : (TPr = 0.56, FPr = 0.2)$

$$age > 66 \ \texttt{AND} \ trig > 1 \ \texttt{AND} \ af = no \ \texttt{AND} \ acoag = no \rightarrow class = thr$$

The first of the selected contrast sets is intuitive to interpret since both primary factors are treatments for cardiovascular disorders. The supporting factors for this set are shown in Table 4. We can see that the first four supporting factors (as well as the two primary 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 thrombotic stroke patients.[9]

[9] Note that the computation of supporting factors differs if *CS*1 is interpreted as a subgroup or as a contrast set. In Table 4 the top four supporting factors are characteristic for group *CS*1, regardless if it is considered as a subgroup or as a contrast set, while the next five supporting factors are characteristic for *CS*1 only if considered as a contrast set to thrombotic.

The second selected contrast set is vague and is not directly connected with medical knowledge. High age and triglyceride values are characteristic for thrombotic stroke, but the boundary values in the contrast set are not very high. The rest of the features in this contrast set indicate no presence of atrial fibrillation and no anticoagulant therapy: again nothing specific. The supporting factors for this set are shown in Table 5. They include high cholesterol and fibrinogen, low fundus oculi and non-smoker. These patients are old and they do not have cardiovascular disorders.

The experiments show the advanced interpretability of the discovered contrast sets achieved by adding the supporting factors. The presented approach to the detection of supporting factors nicely supplements contrast set mining and enables in depth analysis. These examples indicate that the supporting factors appropriately complement the primary factors and can help the expert interpretation to move from speculation towards better justified medical conclusions.

7. Conclusions

This paper has shown that contrast set mining and subgroup discovery are very similar data mining tasks, and has presented approaches to contrast set mining by transforming the contrast set mining task to a subgroup discovery task. We have also shown that the subgroup discovery approach to contrast set mining solves several open issues of contrast set mining. Moreover, in the brain ischemia data analysis application, we have demonstrated that, in the problem of distinguishing between similar classes, the right task to address is the one-versus-all contrast set mining task rather then the classical pairwise (round robin) formulation of the task. Finally, we have improved the explanatory potential of discovered contrast sets by offering additional contrast set descriptors, called the supporting factors. A remaining open issue of contrast set mining is the evaluation and the visualization of contrast set mining results on several contrasting groups, which is the topic of further work.

References

[1] Agrawal R, Mannila H, Srikant R, Toivonen H, Verkamo AI. Fast discovery of association rules. Advances in Knowledge Discovery and Data Mining 1996:307–28.
[2] Atzmueller M, Puppe F, Buscher HP, Exploiting background knowledge for knowledge-intensive subgroup discovery. In: Proceedings of the 19th international joint conference on artificial intelligence (IJCAI-05), 2005, pp. 647–652.
[3] Bay SD, Pazzani MJ. Detecting group differences: mining contrast sets. Data Mining and Knowledge Discovery 2001;5(3):213–46.
[4] Clark P, Niblett T. The CN2 induction algorithm. Machine Learning 1989;3(4):261–83.
[5] Cohen WW. Fast effective rule induction. In: Proceedings of the 12th international conference on machine learning, 1995, pp. 115–123.
[6] Demšar J, Zupan B, Leban G. 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.
[7] Fürnkranz J. Round robin rule learning. In: Proceedings of the 18th international conference on machine learning, 2001, pp. 146–153.
[8] Gamberger D, Lavrač N. Descriptive induction through subgroup discovery: a case study in a medical domain. In: Proceedings of the 19th international conference on machine learning, 2002, pp. 163–170.
[9] Gamberger D, Lavrač N. Expert-guided subgroup discovery: methodology and application, Journal of Artificial Intelligence Research 2002; 17: 501–527.
[10] Gamberger D, Lavrač N, Krstačić G. Active subgroup mining: a case study in coronary heart disease risk group detection. Artificial Intelligence in Medicine 28;2003:27–57.
[11] Gamberger D, Lavrač N, Wettschereck D. Subgroup visualization: a method and application in population screening. In: Proceedings of the 7th international workshop on intelligent data analysis in medicine and pharmacology, 2002, pp. 31–35.
[12] Jovanovski V, Lavrač N. Classification rule learning with APRIORI-C. In: Proceedings of the 10th portuguese conference on artificial intelligence, 2001, pp. 44–51.
[13] Bayardo Jr RJ. Efficiently mining long patterns from databases. In: Proceedings of the 1998 ACM SIGMOD international conference on management of data, ACM Press, New York, NY, USA, 1998, pp. 85–93.
[14] Kavšek B, Lavrač N. APRIORI-SD: adapting association rule learning to subgroup discovery. Applied Artificial Intelligence 2006;20(7):543–583.
[15] Kralj P, Lavrač N, Zupan B. Subgroup visualization. In: Proceedings of the 8th international multiconference information society, 2005, pp. 228–231.
[16] Lavrač N, Cestnik B, Gamberger D, Flach P. Decision support through subgroup discovery: three case studies and the lessons learned. Machine Learning Journal Special Issue on Data Mining Lessons Learned 2003.
[17] Lavrač N, Kavšek B, Flach P, Todorovski L. Subgroup discovery with CN2-SD. Journal of Machine Learning Research 2004;5:153–88.
[18] Quinlan JR. C4.5: programs for machine learning. Morgan Kaufman Publishers Inc; 1993.
[19] De Raedt L, Blockeel H, Dehaspe L, Van Laer W. Three companions for data mining in first order logic. Relational data mining. Springer; 2001.
[20] De Raedt L, Dehaspe L. Clausal discovery. Machine Learning 1997;26:99–146.
[21] Victor M, Ropper AH. Cerebrovascular disease. In: Adams and Victor's principles of neurology, 2001, pp. 821–924.
[22] Železný F, Lavrač N. Propositionalization-based relational subgroup discovery with RSD. Machine Learning 2006;62:33–63.
[23] Webb GI. Discovering significant patterns. Machine Learning 2007;68(1):1–33.
[24] Webb GI, Butler S, Newlands D. On detecting differences between groups. In: Proceedings of the 9th ACM SIGKDD international conference on knowledge discovery and data mining, 2003, pp. 256–265.
[25] Wrobel S. An algorithm for multi-relational discovery of subgroups. In: Proceedings of the 1st european conference on principles of data mining and knowledge discovery, 1997, pp. 78–87.