Journal of Machine Learning Research 10 (2009) 377-403 Submitted 1/08; Revised 8/08; Published 2/09

Supervised Descriptive Rule Discovery: A Unifying Survey of Contrast Set, Emerging Pattern and Subgroup Mining

Petra Kralj Novak PETRA.KRALJ.NOVAK@IJS.SI
Nada Lavrač* NADA.LAVRAC@IJS.SI
Department of Knowledge Technologies
Jožef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia

Geoffrey I. Webb GEOFF.WEBB@INFOTECH.MONASH.EDU.AU
Faculty of Information Technology
Monash University
Building 63, Clayton Campus, Wellington Road, Clayton
VIC 3800, Australia

Editor: Stefan Wrobel

Abstract

This paper gives a survey of contrast set mining (CSM), emerging pattern mining (EPM), and subgroup discovery (SD) in a unifying framework named *supervised descriptive rule discovery*. While all these research areas aim at discovering patterns in the form of rules induced from labeled data, they use different terminology and task definitions, claim to have different goals, claim to use different rule learning heuristics, and use different means for selecting subsets of induced patterns. This paper contributes a novel understanding of these subareas of data mining by presenting a unified terminology, by explaining the apparent differences between the learning tasks as variants of a unique supervised descriptive rule discovery task and by exploring the apparent differences between the approaches. It also shows that various rule learning heuristics used in CSM, EPM and SD algorithms all aim at optimizing a trade off between rule coverage and precision. The commonalities (and differences) between the approaches are showcased on a selection of best known variants of CSM, EPM and SD algorithms. The paper also provides a critical survey of existing supervised descriptive rule discovery visualization methods.

Keywords: descriptive rules, rule learning, contrast set mining, emerging patterns, subgroup discovery

1. Introduction

Symbolic data analysis techniques aim at discovering comprehensible patterns or models in data. They can be divided into techniques for *predictive induction*, where models, typically induced from class labeled data, are used to predict the class value of previously unseen examples, and *descriptive induction*, where the aim is to find comprehensible patterns, typically induced from unlabeled data. Until recently, these techniques have been investigated by two different research communities: predictive induction mainly by the machine learning community, and descriptive induction mainly by the data mining community.

---

*. Also at University of Nova Gorica, Vipavska 13, 5000 Nova Gorica, Slovenia.

Data mining tasks where the goal is to find humanly interpretable differences between groups have been addressed by both communities independently. The groups can be interpreted as class labels, so the data mining community, using the association rule learning perspective, adapted association rule learners like Apriori by Agrawal et al. (1996) to perform a task named *contrast set mining* (Bay and Pazzani, 2001) and *emerging pattern mining* (Dong and Li, 1999). On the other hand, the machine learning community, which usually deals with class labeled data, was challenged by, instead of building sets of classification/prediction rules (e.g., Clark and Niblett, 1989; Cohen, 1995), to build individual rules for exploratory data analysis and interpretation, which is the goal of the task named *subgroup discovery* (Wrobel, 1997).

This paper gives a survey of contrast set mining (CSM), emerging pattern mining (EPM), and subgroup discovery (SD) in a unifying framework, named *supervised descriptive rule discovery*. Typical applications of supervised descriptive rule discovery include patient risk group detection in medicine, bioinformatics applications like finding sets of overexpressed genes for specific treatments in microarray data analysis, and identifying distinguishing features of different customer segments in customer relationship management. The main aim of these applications is to understand the underlying phenomena and not to classify new instances. Take another illustrative example, where a manufacturer wants to know in what circumstances his machines may break down; his intention is not to predict breakdowns, but to understand the factors that lead to them and how to avoid them.

The main contributions of this paper are as follows. It provides a survey of supervised descriptive rule discovery approaches addressed in different communities, and proposes a unifying supervised descriptive rule discovery framework, including a critical survey of visualization methods. The paper is organized as follows: Section 2 gives a survey of past research done in the main supervised descriptive rule discovery areas: contrast set mining, emerging pattern mining, subgroup discovery and other related approaches. Section 3 is dedicated to unifying the terminology, definitions and the heuristics. Section 4 addresses visualization as an important open issue in supervised descriptive rule discovery. Section 5 provides a short summary.

2. A Survey of Supervised Descriptive Rule Discovery Approaches

Research on finding interesting rules from class labeled data evolved independently in three distinct areas—contrast set mining, mining of emerging patterns and subgroup discovery—each area using different frameworks and terminology. In this section, we provide a survey of these three research areas. We also discuss other related approaches.

2.1 An Illustrative Example

Let us illustrate contrast set mining, emerging pattern mining and subgroup discovery using data from Table 1, a very small, artificial sample data set,[1] adapted from Quinlan (1986). The data set contains the results of a survey on 14 individuals, concerning the approval or disapproval of an issue analyzed in the survey. Each individual is characterized by four attributes—`Education` (with values `primary` school, `secondary` school, or `university`), `MaritalStatus` (`single`, `married`, or `divorced`), `Sex` (`male` or `female`), and `HasChildren` (`yes` or `no`)—that encode rudimentary information about the sociodemographic background. The last column `Approved` is the designated

1. Thanks to Johannes Fürnkranz for providing this data set.

Education	Marital Status	Sex	Has Children	Approved
primary	single	male	no	no
primary	single	male	yes	no
primary	married	male	no	yes
university	divorced	female	no	yes
university	married	female	yes	yes
secondary	single	male	no	no
university	single	female	no	yes
secondary	divorced	female	no	yes
secondary	single	female	yes	yes
secondary	married	male	yes	yes
primary	married	female	no	yes
secondary	divorced	male	yes	no
university	divorced	female	yes	no
secondary	divorced	male	no	yes

Table 1: A sample database.

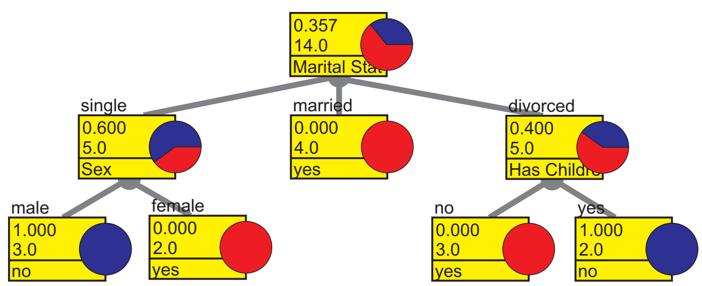

Figure 1: A decision tree, modeling the data set shown in Table 1.

class attribute, encoding whether the individual approved or disapproved the issue. Since there is no need for expert knowledge to interpret the results, this data set is appropriate for illustrating the results of supervised descriptive rule discovery algorithms, whose task is to find interesting patterns describing individuals that are likely to approve or disapprove the issue, based on the four demographic characteristics.

The task of *predictive induction* is to induce, from a given set of *training examples*, a domain model aimed at predictive or classification purposes, such as the *decision tree* shown in Figure 1, or a *rule set* shown in Figure 2, as learned by C4.5 and C4.5rules (Quinlan, 1993), respectively, from the sample data in Table 1.

```
Sex = female  →  Approved = yes
MaritalStatus = single  AND  Sex = male  →  Approved = no
MaritalStatus = married  →  Approved = yes
MaritalStatus = divorced  AND  HasChildren = yes  →  Approved = no
MaritalStatus = divorced  AND  HasChildren = no  →  Approved = yes
```

Figure 2: A set of predictive rules, modeling the data set shown in Table 1.

```
MaritalStatus = single AND Sex = male  →  Approved = no
Sex = male  →  Approved = no
Sex = female  →  Approved = yes
MaritalStatus = married  →  Approved = yes
MaritalStatus = divorced AND HasChildren = yes  →  Approved = no
MaritalStatus = single  →  Approved = no
```

Figure 3: Selected descriptive rules, describing individual patterns in the data of Table 1.

In contrast to predictive induction algorithms, *descriptive induction* algorithms typically result in rules induced from unlabeled examples. E.g., given the examples listed in Table 1, these algorithms would typically treat the class `Approved` no differently from any other attribute. Note, however, that in the learning framework discussed in this paper, that is, in the framework of *supervised descriptive rule discovery*, the discovered rules of the form $X \rightarrow Y$ are induced from class labeled data: the class labels are taken into account in learning of patterns of interest, constraining Y at the right hand side of the rule to assign a value to the class attribute.

Figure 3 shows six descriptive rules, found for the sample data using the Magnum Opus (Webb, 1995) software. Note that these rules were found using the default settings except that the critical value for the statistical test was relaxed to 0.25. These descriptive rules differ from the predictive rules in several ways. The first rule is redundant with respect to the second. The first is included as a strong pattern (*all* 3 single males do not approve) whereas the second is weaker but more general (4 out of 7 males do not approve, which is not highly predictive, but accounts for 4 out of all 5 respondents who do not approve). Most predictive systems will include only one of these rules, but either may be of interest to someone trying to understand the data, depending upon the specific application. This particular approach to descriptive pattern discovery does not attempt to second guess which of the more specific or more general patterns will be the more useful.

Another difference between the predictive and the descriptive rule sets is that the descriptive rule set does not include the pattern that divorcees without children approve. This is because, while the pattern is highly predictive in the sample data, there are insufficient examples to pass the statistical test which assesses the probability that, given the frequency of respondents approving, the apparent correlation occurs by chance. The predictive approach often includes such rules for the sake of completeness, while some descriptive approaches make no attempt at such completeness, assessing each pattern on its individual merits.

Exactly which rules will be induced by a supervised descriptive rule discovery algorithm depends on the task definition, the selected algorithm, as well as the user-defined constraints concerning minimal rule support, precision, etc. In the following section, the example set of Table 1 is used to illustrate the outputs of emerging pattern and subgroup discovery algorithms (see Figures 4 and 5, respectively), while a sample output for contrast set mining is shown in Figure 3 above.

2.2 Contrast Set Mining

The problem of mining contrast sets was first defined by Bay and Pazzani (2001) as finding contrast sets as "conjunctions of attributes and values that differ meaningfully in their distributions across groups." The example rules in Figure 3 illustrate this approach, including all conjunctions of attributes and values that pass a statistical test for productivity (explained below) with respect to attribute `Approved` that defines the 'groups.'

2.2.1 CONTRAST SET MINING ALGORITHMS

The STUCCO algorithm (Search and Testing for Understandable Consistent Contrasts) by Bay and Pazzani (2001) is based on the Max-Miner rule discovery algorithm (Bayardo, 1998). STUCCO discovers a set of contrast sets along with their supports[2] 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 the group variable Y. It is also subjected to what Webb (2007) calls a test for *productivity*. Rule $X \rightarrow Y$ is productive iff

$$\forall Z \subset X : \mathit{confidence}(Z \rightarrow Y) < \mathit{confidence}(X \rightarrow Y)$$

where $\mathit{confidence}(X \rightarrow Y)$ is a maximum likelihood estimate of conditional probability $P(Y|X)$, estimated by the ratio $\frac{\mathit{count}(X,Y)}{\mathit{count}(X)}$, where $\mathit{count}(X,Y)$ represents the number of examples for which both X and Y are true, and $\mathit{count}(X)$ represents the number of examples for which X is true. Therefore 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. In comparison, the other techniques discussed below do not, by default, employ any form of correction for multiple comparisons, as result of which they have high risk of making *false discoveries* (Webb, 2007).

It was shown by Webb et al. (2003) that contrast set mining is a special case of the more general rule learning task. A contrast set can be interpreted as the antecedent of rule $X \rightarrow Y$, and group G_i for which it is characteristic—in contrast with group G_j—as the rule consequent, leading to rules of the form $\mathit{ContrastSet} \rightarrow G_i$. A standard descriptive rule discovery algorithm, such as an association-rule discovery system (Agrawal et al., 1996), can be used for the task if the consequent is restricted to a variable whose values denote group membership.

In particular, Webb et al. (2003) showed that when STUCCO and the general-purpose descriptive rule learning system Magnum Opus were each run with their default settings, but the consequent restricted to the contrast variable in the case of Magnum Opus, the contrasts found differed mainly as a consequence only of differences in the statistical tests employed to screen the rules.

Hilderman and Peckham (2005) proposed a different approach to contrast set mining called CIGAR (ContrastIng Grouped Association Rules). CIGAR uses different statistical tests to STUCCO or Magnum Opus for both independence and productivity and introduces a test for *minimum support*.

Wong and Tseng (2005) have developed techniques for discovering contrasts that can include negations of terms in the contrast set.

In general, contrast set mining approaches require discrete data, which is in real world applications frequently not the case. A data discretization method developed specifically for set mining purposes is described by Bay (2000). This approach does not appear to have been further used by the contrast set mining community, except for Lin and Keogh (2006), who extended contrast set mining to time series and multimedia data analysis. They introduced a formal notion of a time series contrast set along with a fast algorithm to find time series contrast sets. An approach to quantitative contrast set mining without discretization in the preprocessing phase is proposed by Simeon

2. The support of a contrast set *ContrastSet* with respect to a group G_i, $\mathit{support}(\mathit{ContrastSet}, G_i)$, is the percentage of examples in G_i for which the contrast set is true.

and Hilderman (2007) with the algorithm Gen_QCSets. In this approach, a slightly modified equal width binning interval method is used.

Common to most contrast set mining approaches is that they generate all candidate contrast sets from discrete (or discretized) data and later use statistical tests to identify the interesting ones. Open questions identified by Webb et al. (2003) are yet unsolved: selection of appropriate heuristics for identifying interesting contrast sets, appropriate measures of quality for sets of contrast sets, and appropriate methods for presenting contrast sets to the end users.

2.2.2 SELECTED APPLICATIONS OF CONTRAST SET MINING

The contrast mining paradigm does not appear to have been pursued in many published applications. Webb et al. (2003) investigated its use with retail sales data. Wong and Tseng (2005) applied contrast set mining for designing customized insurance programs. Siu et al. (2005) have used contrast set mining to identify patterns in synchrotron x-ray data that distinguish tissue samples of different forms of cancerous tumor. Kralj et al. (2007b) have addressed a contrast set mining problem of distinguishing between two groups of brain ischaemia patients by transforming the contrast set mining task to a subgroup discovery task.

2.3 Emerging Pattern Mining

Emerging patterns were defined by Dong and Li (1999) as itemsets whose support increases significantly from one data set to another. Emerging patterns are said to capture emerging trends in time-stamped databases, or to capture differentiating characteristics between classes of data.

2.3.1 EMERGING PATTERN MINING ALGORITHMS

Efficient algorithms for mining emerging patterns were proposed by Dong and Li (1999) and Fan and Ramamohanarao (2003). When first defined by Dong and Li (1999), the purpose of emerging patterns was "to capture emerging trends in time-stamped data, or useful contrasts between data classes". Subsequent emerging pattern research has largely focused on the use of the discovered patterns for classification purposes, for example, classification by emerging patterns (Dong et al., 1999; Li et al., 2000) and classification by jumping emerging patterns[3] (Li et al., 2001). An advanced Bayesian approach (Fan and Ramamohanara, 2003) and bagging (Fan et al., 2006) were also proposed.

From a semantic point of view, emerging patterns are association rules with an itemset in rule antecedent, and a fixed consequent: $ItemSet \rightarrow D_1$, for given data set D_1 being compared to another data set D_2.

The measure of quality of emerging patterns is the *growth rate* (the ratio of the two supports). It determines, for example, that a pattern with a 10% support in one data set and 1% in the other is better than a pattern with support 70% in one data set and 10% in the other (as $\frac{10}{1} > \frac{70}{10}$). From the association rule perspective, $GrowthRate(ItemSet, D_1, D_2) = \frac{confidence(ItemSet \rightarrow D_1)}{1 - confidence(ItemSet \rightarrow D_1)}$. Thus it can be seen that growth rate provides an identical ordering to confidence, except that growth rate is undefined when confidence = 1.0.

3. Jumping emerging patterns are emerging patterns with support zero in one data set and greater then zero in the other data set.

```
MaritalStatus = single AND Sex = male  →  Approved = no
MaritalStatus = married  →  Approved = yes
MaritalStatus = divorced AND HasChildren = yes  →  Approved = no
```

Figure 4: Jumping emerging patterns in the data of Table 1.

Some researchers have argued that finding all the emerging patterns above a minimum growth rate constraint generates too many patterns to be analyzed by a domain expert. Fan and Ramamohanarao (2003) have worked on selecting the interesting emerging patterns, while Soulet et al. (2004) have proposed condensed representations of emerging patterns.

Boulesteix et al. (2003) introduced a CART-based approach to discover emerging patterns in microarray data. The method is based on growing decision trees from which the emerging patterns are extracted. It combines pattern search with a statistical procedure based on Fisher's exact test to assess the significance of each emerging pattern. Subsequently, sample classification based on the inferred emerging patterns is performed using maximum-likelihood linear discriminant analysis.

Figure 4 shows all jumping emerging patterns found for the data in Table 1 when using a minimum support of 15%. These were discovered using the Magnum Opus software, limiting the consequent to the variable *approved*, setting minimum confidence to 1.0 and setting minimum support to 2.

2.3.2 SELECTED APPLICATIONS OF EMERGING PATTERNS

Emerging patterns have been mainly applied to the field of bioinformatics, more specifically to microarray data analysis. Li et al. (2003) present an interpretable classifier based on simple rules that is competitive to the state of the art black-box classifiers on the acute lymphoblastic leukemia (ALL) microarray data set. Li and Wong (2002) have focused on finding groups of genes by emerging patterns and applied it to the ALL/AML data set and the colon tumor data set. Song et al. (2001) used emerging patterns together with unexpected change and the added/perished rule to mine customer behavior.

2.4 Subgroup Discovery

The task of subgroup discovery was defined by Klösgen (1996) and Wrobel (1997) 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', for example, are as large as possible and have the most unusual statistical (distributional) characteristics with respect to the property of interest".

2.4.1 SUBGROUP DISCOVERY ALGORITHMS

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 the condition part of a rule $SubgroupDescription \rightarrow Class$. Therefore, subgroup discovery can be seen as a special case of a more general rule learning task.

Subgroup discovery research has evolved in several directions. On the one hand, exhaustive approaches guarantee the optimal solution given the optimization criterion. One system that can use both exhaustive and heuristic discovery algorithms is Explora by Klösgen (1996). Other algo-

```
Sex = female  →  Approved = yes
MaritalStatus = married  →  Approved = yes
MaritalStatus = divorced AND HasChildren = no  →  Approved = yes
Education = university  →  Approved = yes
MaritalStatus = single AND Sex = male  →  Approved = no
```

Figure 5: Subgroup descriptions induced by Apriori-SD from the data of Table 1.

rithms for exhaustive subgroup discovery are the SD-Map method by Atzmüller and Puppe (2006) and Apriori-SD by Kavšek and Lavrač (2006). On the other hand, adaptations of classification rule learners to perform subgroup discovery, including algorithm SD by Gamberger and Lavrač (2002) and algorithm CN2-SD by Lavrač et al. (2004b), use heuristic search techniques drawn from classification rule learning coupled with constraints appropriate for descriptive rules.

Relational subgroup discovery approaches have been proposed by Wrobel (1997, 2001) with algorithm Midos, by Klösgen and May (2002) with algorithm SubgroupMiner, which is designed for spatial data mining in relational space databases, and by Železný and Lavrač (2006) with the algorithm RSD (Relational Subgroup Discovery). RSD uses a propositionalization approach to relational subgroup discovery, achieved through appropriately adapting rule learning and first-order feature construction. Other non-relational subgroup discovery algorithms were developed, including an algorithm for exploiting background knowledge in subgroup discovery (Atzmüller et al., 2005a), and an iterative genetic algorithm SDIGA by del Jesus et al. (2007) implementing a fuzzy system for solving subgroup discovery tasks.

Different heuristics have been used for subgroup discovery. By definition, the interestingness of a subgroup depends on its unusualness and size, therefore the rule quality evaluation heuristics needs to combine both factors. Weighted relative accuracy (*WRAcc*, see Equation 2 in Section 3.3) is used by algorithms CN2-SD, Apriori-SD and RSD and, in a different formulation and in different variants, also by MIDOS and EXPLORA. Generalization quotient (q_g, see Equation 3 in Section 3.3) is used by the SD algorithm. SubgroupMiner uses the classical binominal test to verify if the target share is significantly different in a subgroup.

Different approaches have been used for eliminating redundant subgroups. Algorithms CN2-SD, Apriori-SD, SD and RSD use weighted covering (Lavrač et al., 2004b) to achieve rule diversity. Algorithms Explora and SubgroupMiner use an approach called subgroup suppression (Klösgen, 1996). A sample set of subgroup describing rules, induced by Apriori-SD with parameters *support* set to 15% (requiring at least 2 covered training examples per rule) and *confidence* set to 65%, is shown in Figure 5.

2.4.2 SELECTED APPLICATIONS OF SUBGROUP DISCOVERY

Subgroup discovery was used in numerous real-life applications. The applications in medical domains include the analysis of coronary heart disease (Gamberger and Lavrač, 2002) and brain ischaemia data analysis (Kralj et al., 2007b,a; Lavrač et al., 2007), as well as profiling examiners for sonographic examinations (Atzmüller et al., 2005b). Spatial subgroup mining applications include mining of census data (Klösgen et al., 2003) and mining of vegetation data (May and Ragia, 2002). There are also applications in other areas like marketing (del Jesus et al., 2007; Lavrač et al., 2004a) and analysis of manufacturing shop floor data (Jenkole et al., 2007).

2.5 Related Approaches

Research in some closely related areas of rule learning, performed independently from the above described approaches, is outlined below.

2.5.1 CHANGE MINING

The paper by Liu et al. (2001) on *fundamental rule changes* proposes a technique to identify the set of fundamental changes in two given data sets collected from two time periods. The proposed approach first generates rules and in the second phase it identifies changes (rules) that can not be explained by the presence of other changes (rules). This is achieved by applying statistical χ^2 test for homogeneity of support and confidence. This differs from contrast set discovery through its consideration of rules for each group, rather than itemsets. A change in the frequency of just one itemset between groups may affect many association rules, potentially all rules that have the itemset as either an antecedent or consequent.

Liu et al. (2000) and Wang et al. (2003) present techniques that identify differences in the decision trees and classification rules, respectively, found on two different data sets.

2.5.2 MINING CLOSED SETS FROM LABELED DATA

Closed sets have been proven successful in the context of compacted data representation for association rule learning. However, their use is mainly descriptive, dealing only with unlabeled data. It was recently shown that when considering labeled data, closed sets can be adapted for classification and discrimination purposes by conveniently contrasting covering properties on positive and negative examples (Garriga et al., 2006). The approach was successfully applied in potato microarray data analysis to a real-life problem of distinguishing between virus sensitive and resistant transgenic potato lines (Kralj et al., 2006).

2.5.3 EXCEPTION RULE MINING

Exception rule mining considers a problem of finding a set of rule pairs, each of which consists of an exception rule (which describes a regularity for fewer objects) associated with a strong rule (description of a regularity for numerous objects with few counterexamples). An example of such a rule pair is “using a seat belt is safe” (strong rule) and “using a seat belt is risky for a child” (exception rule). While the goal of exception rule mining is also to find descriptive rules from labeled data, in contrast with other rule discovery approaches described in this paper, the goal of exception rule mining is to find “weak” rules—surprising rules that are an exception to the general belief of background knowledge.

Suzuki (2006) and Daly and Taniar (2005), summarizing the research in exception rule mining, reveal that the key concerns addressed by this body of research include interestingness measures, reliability evaluation, practical application, parameter reduction and knowledge representation, as well as providing fast algorithms for solving the problem.

2.5.4 IMPACT RULES, BUMP HUNTING, QUANTITATIVE ASSOCIATION RULES

Supervised descriptive rule discovery seeks to discover sets of conditions that are related to deviations in the class distribution, where the class is a qualitative variable. A related body of research seeks to discover sets of conditions that are related to deviations in a target quantitative variable.

Contrast Set Mining	Emerging Pattern Mining	Subgroup Discovery	Rule Learning
contrast set	itemset	subgroup description	rule condition
groups $G_1, \dots G_n$	data sets D_1 and D_2	class/property C	class/concept C_i
attribute-value pair	item	logical (binary) feature	condition
examples in groups $G_1, \dots G_n$	transactions in data sets D_1 and D_2	examples of C and $\overline{C}$	examples of $C_1 \dots C_n$
examples for which the contrast set is true	transactions containing the itemset	subgroup of instances	covered examples
support of contrast set on G_i support of contrast set on G_j	support of EP in data set D_1 support of EP in data set D_2	true positive rate false positive rate	true positive rate false positive rate

Table 2: Table of synonyms from different communities, showing the compatibility of terms.

Such techniques include Bump Hunting (Friedman and Fisher, 1999), Quantitative Association Rules (Aumann and Lindell, 1999) and Impact Rules (Webb, 2001).

3. A Unifying Framework for Supervised Descriptive Rule Induction

This section presents a unifying framework for contrast set mining, emerging pattern mining and subgroup discovery, as the main representatives of supervised descriptive rule discovery approaches. This is achieved by unifying the terminology, the task definitions and the rule learning heuristics.

3.1 Unifying the Terminology

Contrast set mining (CSM), emerging pattern mining (EPM) and subgroup discovery (SD) were developed in different communities, each developing their own terminology that needs to be clarified before proceeding. Below we show that terms used in different communities are compatible, according to the following definition of compatibility.

Definition 1: Compatibility of terms. *Terms used in different communities are compatible if they can be translated into equivalent logical expressions and if they bare the same meaning, that is, if terms from one community can replace terms used in another community.*

Lemma 1: *Terms used in CSM, EPM and SD are compatible.*
Proof The compatibility of terms is proven through a term dictionary, whose aim is to translate all the terms used in CSM, EPM and SD into the terms used in the rule learning community. The term dictionary is proposed in Table 2. More specifically, this table provides a dictionary of equivalent terms from contrast set mining, emerging pattern mining and subgroup discovery, in a unifying terminology of classification rule learning, and in particular of concept learning (considering class C_i as the concept to be learned from the positive examples of this concept, and the negative examples formed of examples of all other classes). ■

3.2 Unifying the Task Definitions

Having established a unifying view on the terminology, the next step is to provide a unifying view on the different task definitions.

CSM A contrast set mining task is defined as follows (Bay and Pazzani, 2001). Let A_1, A_2, ..., A_k be a set of k variables called attributes. Each A_i can take values from the set $\{v_{i1}, v_{i2}, \ldots, v_{im}\}$. Given a set of user defined groups G_1, G_2, ..., G_n of data instances, a contrast set is a conjunction of attribute-value pairs, defining a pattern that best discriminates the instances of different user-defined groups. 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.

EPM An emerging patterns mining task is defined as follows (Dong and Li, 1999). Let $I = \{i_1, i_2, \ldots, i_N\}$ be a set of items (note that an item is equivalent to a binary feature in SD, and an individual attribute-value pair in CSM). A transaction is a subset T of I. A *dataset* is a set D of transactions. A subset X of I is called an *itemset*. Transaction T contains an itemset X in a data set D, if $X \subseteq T$. For two data sets D_1 and D_2, emerging pattern mining aims at discovering itemsets whose support increases significantly from one data set to another.

SD In subgroup discovery, subgroups are described as conjunctions of features, where features are of the form $A_i = v_{ij}$ for nominal attributes, and $A_i > value$ or $A_i \leq value$ for continuous attributes. Given the property of interest C, and the population of examples of C and $\overline{C}$, 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 C (Wrobel, 1997).

The definitions of contrast set mining, emerging pattern mining and subgroup discovery appear different: contrast set mining searches for discriminating characteristics of groups called contrast sets, emerging pattern mining aims at discovering itemsets whose support increases significantly from one data set to another, while subgroup discovery searches for subgroup descriptions. By using the dictionary from Table 2 we can see that the goals of these three mining tasks are very similar, it is primarily the terminology that differs.

Definition 2: Compatibility of task definitions. *Definitions of different learning tasks are compatible if one learning task can be translated into another learning task without substantially changing the learning goal.*

Lemma 2: *Definitions of CSM, EPM and SD tasks are compatible.*

Proof To show the compatibility of task definitions, we propose a unifying table (Table 3) of task definitions, allowing us to see that emerging pattern mining task $EPM(D_1, D_2)$ is equivalent to $CSM(G_i, G_j)$. It is also easy 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: $SD(G_i)$ for $C = G_i$ and $\overline{C} = G_j$, and $SD(G_j)$ for $C = G_j$ and $\overline{C} = G_i$.

Contrast Set Mining	Emerging Pattern Mining	Subgroup Discovery	Rule Learning
Given	**Given**	**Given**	**Given**
examples in G_1 vs. G_j	transactions in D_1 and D_2	in examples C	examples in C_i
from $G_1, \ldots G_i$	from D_1 and D_2	from C and $\overline{C}$	from $C_1 \ldots C_n$
Find	**Find**	**Find**	**Find**
$ContrastSet_{i_k} \rightarrow G_i$	$ItemSet_{1_k} \rightarrow D_1$	$SubgrDescr_k \rightarrow C$	$\{RuleCond_{i_k} \rightarrow C_i\}$
$ContrastSet_{j_l} \rightarrow G_j$	$ItemSet_{2_l} \rightarrow D_2$		

Table 3: Table of task definitions from different communities, showing the compatibility of task definitions in terms of output rules.

Having proved that the subgroup discovery task is compatible with a two-group contrast set mining task, it is by induction compatible with a general contrast set mining task, as shown below.

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

Note that in Table 3 of task definitions column 'Rule Learning' again corresponds to a concept learning task instead of the general classification rule learning task. In the concept learning setting, which is better suited for the comparisons with supervised descriptive rule discovery approaches, a distinguished class C_i is learned from examples of this class, and examples of all other classes $C_1, \ldots, C_{i-1}, C_{i+1}, C_N$ are merged to form the set of examples of class $\overline{C_i}$. In this case, induced rule set $\{RuleCond_{i_k} \rightarrow C_i\}$ consists only of rules for distinguished class C_i. On the other hand, in a general classification rule learning setting, from examples of N different classes a set of rules would be learned $\{\ldots, RuleCond_{i_k} \rightarrow C_i, RuleCond_{i_{k+1}} \rightarrow C_i, \ldots, RuleCond_{j_l} \rightarrow C_j, \ldots, Default\}$, consisting of sets of rules of the form $RuleCond_{i_k} \rightarrow C_i$ for each individual class C_i, supplemented by the default rule. ■

While the primary tasks are very closely related, each of the three communities has concentrated on different sets of issues around this task. The contrast set discovery community has paid greatest attention to the statistical issues of multiple comparisons that, if not addressed, can result in high risks of false discoveries. The emerging patterns community has investigated how supervised descriptive rules can be used for classification. The contrast set and emerging pattern communities have primarily addressed only categorical data whereas the subgroup discovery community has also considered numeric and relational data. The subgroup discovery community has also explored techniques for discovering small numbers of supervised descriptive rules with high coverage of the data.

3.3 Unifying the Rule Learning Heuristics

The aim of this section is to provide a unifying view on rule learning heuristics used in different communities. To this end, we first investigate the rule quality measures.

Most rule quality measures are derived by analyzing the covering properties of the rule and the class in the rule consequent considered as positive. This relationship can be depicted by a confusion

actual	predicted # of positives	# of negatives	
# of positives	$p = \|TP(X,Y)\|$	$\overline{p} = \|FN(X,Y)\|$	P
# of negatives	$n = \|FP(X,Y)\|$	$\overline{n} = \|TN(X,Y)\|$	N
	$p+n$	$\overline{p}+\overline{n}$	$P+N$

Table 4: Confusion matrix: $TP(X,Y)$ stands for true positives, $FP(X,Y)$ for false positives, $FN(X,Y)$ for false negatives and $TN(X,Y)$ for true negatives, as predicted by rule $X \rightarrow Y$.

matrix (Table 4, see, e.g., Kohavi and Provost, 1998), which considers that rule $R = X \rightarrow Y$ is represented as (X,Y), and defines p as the number of true positives (positive examples correctly classified as positive by rule (X,Y)), n as the number of false positives, etc., from which other covering characteristics of a rule can be derived: true positive rate $TPr(X,Y) = \frac{p}{P}$ and false positive rate $FPr(X,Y) = \frac{n}{N}$.

CSM Contrast set mining aims at discovering contrast sets that best discriminate the instances of different user-defined groups. The support of contrast set X with respect to group G_i, *support*(X,G_i), is the percentage of examples in G_i for which the contrast set is true. Note that *support of a contrast set with respect to group G* is the same as *true positive rate* in the classification rule and subgroup discovery terminology, that is, $support(X,G_i) = \frac{count(X,G_i)}{|G_i|} = TPr(X,G_i)$. A derived goal of contrast set mining, proposed by Bay and Pazzani (2001), is to find contrast sets whose support differs meaningfully across groups, for δ being a user-defined parameter.

$$SuppDiff(X,G_i,G_j) = |support(X,G_i) - support(X,G_j)| \geq \delta.$$

EPM Emerging pattern mining aims at discovering itemsets whose support increases significantly from one data set to another Dong and Li (1999), where *support* of itemset X in data set D is computed as $support(X,D) = \frac{count(X,D)}{|D|}$, for $count(X,D)$ being the number of transactions in D containing X. Suppose we are given an ordered pair of data sets D_1 and D_2. The *GrowthRate* of an itemset X from D_1 to D_2, denoted as $GrowthRate(X,D_1,D_2)$, is defined as follows:

$$GrowthRate(X,D_1,D_2) = \frac{support(X,D_1)}{support(X,D_2)}. \qquad (1)$$

Definitions of special cases of $GrowthRate(X,D_1,D_2)$ are as follows, if $support(X,D_1) = 0$ then $GrowthRate(X,D_1,D_2) = 0$, if $support(X,D_2) = 0$ then $GrowthRate(X,D_1,D_2) = \infty$.

SD Subgroup discovery 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 (Wrobel, 1997). There were several heuristics developed and used in the subgroup discovery community. Since they follow from the task definition, they try to maximize subgroup size and the distribution difference at the same time. Examples of such heuristics are the *weighted relative accuracy* (Equation 2, see Lavrač et al., 2004b) and the *generalization*

Contrast Set Mining	Emerging Pattern Mining	Subgroup Discovery	Rule Learning
$SuppDiff(X, G_i, G_j)$		$WRAcc(X, C)$	Piatetski-Shapiro heuristic leverage
	$GrowthRate(X, D_1, D_2)$	$q_g(X, C)$	odds ratio for $g = 0$ accuracy/precision, for $g = p$

Table 5: Table of relationships between the pairs of heuristics, and their equivalents in classification rule learning.

quotient (Equation 3, see Gamberger and Lavrač, 2002) , for g being a user-defined parameter.

$$WRAcc(X,C) = \frac{p+n}{P+N} \cdot \left(\frac{p}{p+n} - \frac{P}{P+N}\right), \tag{2}$$

$$q_g(X,C) = \frac{p}{n+g}. \tag{3}$$

Let us now investigate whether the heuristics used in CSM, EPM and SD are compatible, using the following definition of compatibility.

Definition 3: Compatibility of heuristics.
Heuristic function h_1 is compatible *with h_2 if h_2 can be derived from h_1 and if for any two rules R and R', $h_1(R) > h_1(R') \Leftrightarrow h_2(R) > h_2(R')$.*

Lemma 3: *Definitions of CSM, EPM and SD heuristics are pairwise compatible.*
Proof The proof of Lemma 3 is established by proving two sub-lemmas, Lemma 3a and Lemma 3b, which prove the compatibility of two pairs of heuristics, whereas the relationships between these pairs is established through Table 5, and illustrated in Figures 6 and 7. ■

Lemma 3a: *The support difference heuristic used in CSM and the weighted relative accuracy heuristic used in SD are compatible.*
Proof Note that, as shown below, weighted relative accuracy (Equation 2) can be interpreted in terms of probabilities of rule antecedent X and consequent Y (class C representing the property of interest), and the conditional probability of class Y given X, estimated by relative frequencies.

$$WRAcc(X,Y) = P(X) \cdot (P(Y|X) - P(Y)).$$

From this equation we see that, indeed, when optimizing weighted relative accuracy of rule $X \rightarrow Y$, we optimize two contrasting factors: rule coverage $P(X)$ (proportional to the size of the subgroup), and distributional unusualness $P(Y|X) - P(Y)$ (proportional to the difference of the number of positive examples correctly covered by the rule and the number of positives in the original training set). It is straightforward to show that this measure is equivalent to the Piatetski-Shapiro measure, which evaluates the conditional (in)dependence of rule consequent and rule antecedent as follows:

$$PS(X,Y) = P(X \cdot Y) - P(X) \cdot P(Y).$$

Weighted relative accuracy, known from subgroup discovery, and support difference between groups, used in contrast set mining, are related as follows:[4]

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

Since the distribution of examples among classes is constant for any data set, the first two factors $P(Y)$ *and* $P(\overline{Y})$ are constant within a data set. Therefore, when maximizing the weighted relative accuracy, one is maximizing the second factor $TPr(X,Y) - FPr(X,Y)$, which actually is support difference when we have a two group contrast set mining problem. Consequently, for $C = G_1$, and $\overline{C} = G_2$ the following holds:

$$WRAcc(X,C) = WRAcc(X,G_1) = P(G_1) \cdot P(G_2) \cdot [support(X,G_1) - support(X,G_2)].$$

■

Lemma 3b: *The growth rate heuristic used in EPM and the generalization quotient heuristic used in SD are compatible.*
Proof Equation 1 can be rewritten as follows:

$$GrowthRate(X,D_1,D_2) = \frac{support(X,D_1)}{support(C,D_2)} =$$

$$= \frac{count(X,D_1)}{count(X,D_2)} \cdot \frac{|D_2|}{|D_1|} = \frac{p}{n} \cdot \frac{N}{P}.$$

Since the distribution of examples among classes is constant for any data set, the quotient $\frac{N}{P}$ is constant. Consequently, the growth rate is the generalization quotient with $g = 0$, multiplied by a constant. Therefore, the growth rate is compatible with the generalization quotient.

$$GrowthRate(X,C,\overline{C}) = q_0(X,C) \cdot \frac{N}{P}.$$

■

The lemmas prove that heuristics used in CSM and EPM can be translated into heuristics used in SD and vice versa. In this way, we have shown the compatibility of CSM and SD heuristics, as well as the compatibility of EPM and SD heuristics. While the lemmas do not prove direct compatibility of CSM and EPM heuristics, they prove that heuristics used in CSM and EPM can be translated into two heuristics used in SD, both aiming at trading-off between coverage and distributional difference.

4. Peter A. Flach is acknowledged for having derived these equations.

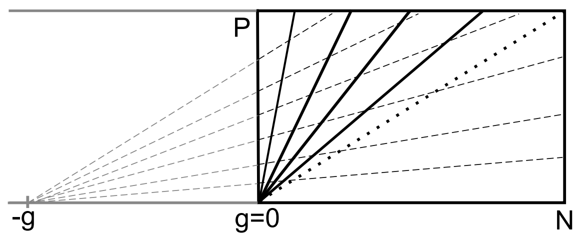

Figure 6: Isometrics for q_g. The dotted lines show the isometrics for a selected $g > 0$, while the full lines show the special case when $g = 0$, compatible to the EPM *growth rate* heuristic.

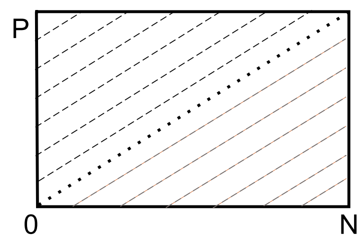

Figure 7: Isometrics for *WRAcc*, compatible to the CSM *support difference* heuristic.

Table 5 provides also the equivalents of these heuristics in terms of heuristics known from the classification rule learning community, details of which are beyond the scope of this paper (an interested reader can find more details on selected heuristics and their ROC representations in Fürnkranz and Flach, 2003).

Note that the growth rate heuristic from EPM, as a special case of the generalization quotient heuristic with $g = 0$, does not consider rule coverage. On the other hand, its compatible counterpart, the generalization quotient q_g heuristic used in SD, can be tailored to favor more general rules by setting the g parameter value, as for a general g value, the q_g heuristic provides a trade-off between rule accuracy and coverage. Figure 6[5] illustrates the q_g isometrics, for a general g value, as well as for value $g = 0$.

Note also that standard rule learners (such as CN2 by Clark and Niblett, 1989) tend to generate very specific rules, due to using accuracy heuristic $Acc(X,Y) = \frac{p+\bar{n}}{P+N}$ or its variants: the Laplace and the *m*-estimate. On the other hand, the CSM support difference heuristic and its SD counterpart *WRAcc* both optimize a trade-off between rule accuracy and coverage. The *WRAcc* isometrics are plotted in Figure 7.[6]

3.4 Comparison of Rule Selection Mechanisms

Having established a unifying view on the terminology, definitions and rule learning heuristics, the last step is to analyze rule selection mechanisms used by different algorithms. The motivation for rule selection can be either to find only significant rules or to avoid overlapping rules (too many too similar rules), or to avoid showing redundant rules to the end users. Note that rule selection is not always necessary and that depending on the goal, redundant rules can be valuable (e.g., clas-

5. This figure is due to Gamberger and Lavrač (2002).
6. This figure is due to Fürnkranz and Flach (2003).

sification by aggregating emerging patterns by Dong et al., 1999). Two approaches are commonly used: statistic tests and the (weighted) covering approach. In this section, we compare these two approaches.

Webb et al. (2003) show that contrast set mining is a special case of the more general rule discovery task. However, an experimental comparison of STUCCO, OPUS_AR and C4.5 has shown that standard rule learners return a larger set of rules compared to STUCCO, and that some of them are also not interesting to end users. STUCCO (see Bay and Pazzani 2001 for more details) uses several mechanisms for rule pruning. Statistical significance pruning 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 a similar support to its parent or that fails a χ^2 test of independence with respect to its parent.

In the context of OPUS_AR, the emphasis has been on developing statistical tests that are robust in the context of the large search spaces explored in many rule discovery applications Webb (2007). These include tests for independence between the antecedent and consequent, and tests to assess whether specializations have significantly higher confidence than their generalizations.

In subgroup discovery, the *weighted covering approach* (Lavrač et al., 2004b) is used with the aim of ensuring the diversity of rules induced in different iterations of the algorithm. In each iteration, after selecting the best rule, the weights of positive examples are decreased according to the number of rules covering each positive example $rule\_count(e)$; they are set to $w(e) = \frac{1}{rule\_count(e)}$. For selecting the best rule in consequent iterations, the SD algorithm (Gamberger and Lavrač, 2002) uses—instead of the unweighted q_g measure (Equation 3)—the weighted variant of q_g defined in Equation 4, while the CN2-SD (Lavrač et al., 2004b) and APRIORI-SD (Kavšek and Lavrač, 2006) algorithms use the weighted relative accuracy (Equation 2) modified with example weights, as defined in Equation 5, where $p' = \sum_{TP(X,Y)} w(e)$ is the sum of the weights of all covered positive examples, and P' is the sum of the weights of all positive examples.

$$q'_g(X,Y) = \frac{p'}{n+g}, \tag{4}$$

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

Unlike in the sections on the terminology, task definitions and rule learning heuristics, the comparison of rule pruning mechanisms described in this section does not result in a unified view; although the goals of rule pruning may be the same, the pruning mechanisms used in different subareas of supervised descriptive rule discovery are—as shown above—very different.

4. Visualization

Webb et al. (2003) identify a need to develop appropriate methods for presenting contrast sets to end users, possibly through contrast set visualization. This open issue, concerning the visualization of contrast sets and emerging patterns, can be resolved by importing some of the solutions proposed in the subgroup discovery community. Several methods for subgroup visualization were developed by Wettschereck (2002), Wrobel (2001), Gamberger et al. (2002), Kralj et al. (2005) and Atzmüller and Puppe (2005). They are here illustrated using the coronary heart disease data set, originally analyzed by Gamberger and Lavrač (2002). The visualizations are evaluated by considering their

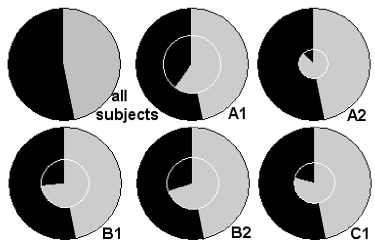

Figure 8: Subgroup visualization by pie charts.

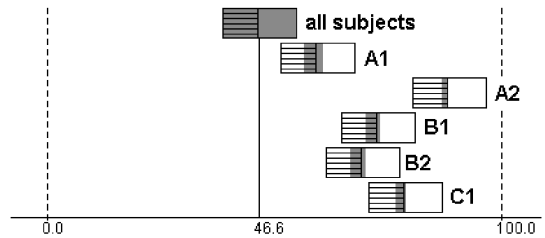

Figure 9: Subgroup visualization by box plots.

intuitiveness, correctness of displayed data, usefulness, ability to display contents besides the numerical properties of subgroups, (e.g., plot subgroup probability densities against the values of an attribute), and their extensibility to multi-class problems.

4.1 Visualization by Pie Charts

Slices of pie charts are the most common way of visualizing parts of a whole. They are widely used and understood. Subgroup visualization by pie chart, proposed by Wettschereck (2002), consists of a two-level pie for each subgroup. The base pie represents the distribution of individuals in terms of the property of interest of the entire example set. The inner pie represents the size and the distribution of individuals in terms of the property of interest in a specific subgroup. An example of five subgroups (subgroups A1, A2, B1, B2, C1), as well as the base pie "all subjects" are visualized by pie charts in Figure 8.

The main weakness of this visualization is the misleading representation of the relative size of subgroups. The size of a subgroup is represented by the radius of the circle. The faultiness arises from the surface of the circle which increases with the square of its radius. For example, a subgroup that covers 20% of examples is represented by a circle that covers only 4% of the whole surface, while a subgroup that covers 50% of examples is represented by a circle that covers 25% of the whole surface. In terms of usefulness, this visualization is not very handy since—in order to compare subgroups—one would need to compare sizes of circles, which is difficult. The comparison of distributions in subgroups is also not straightforward. This visualization also does not show the contents of subgroups. It would be possible to extend this visualization to multi-class problems.

4.2 Visualization by Box Plots

In subgroup visualization by box plots, introduced by Wrobel (2001), each subgroup is represented by one box plot (all examples are also considered as one subgroup and are displayed in the top box). Each box shows the entire population; the horizontally stripped area on the left represents the positive examples and the white area on the right-hand side of the box represents the negative examples. The grey area within each box indicates the respective subgroup. The overlap of the grey area with the hatched area shows the overlap of the group with the positive examples. Hence, the more to the left the grey area extends the better. The less the grey area extends to the right of the hatched area, the more specific a subgroup is (less overlap with the subjects of the negative class). Finally, the location of the box along the X-axis indicates the relative share of the target class within each subgroup: the more to the right a box is placed, the higher is the share of the target value within this subgroup. The vertical line (in Figure 9 at value 46.6%) indicates the default accuracy, that is, the number of positive examples in the entire population. An example box plot visualization of five subgroups is presented in Figure 9.

On the negative side, the intuitiveness of this visualization is relatively poor since an extensive explanation is necessary for understanding it. It is also somewhat illogical since the boxes that are placed more to the right and have more grey color on the left-hand side represent the best subgroups. This visualization is not very attractive since most of the image is white; the grey area (the part of the image that really represents the subgroups) is a relatively tiny part of the entire image. On the positive side, all the visualized data are correct and the visualization is useful since the subgroups are arranged by their confidence. It is also easier to contrast the sizes of subgroups compared to their pie chart visualization. However, this visualization does not display the contents of the data. It would also be difficult to extend this visualization to multi-class problems.

4.3 Visualizing Subgroup Distribution w.r.t. a Continuous Attribute

The distribution of examples w.r.t. a continuous attribute, introduced by Gamberger and Lavrač (2002) and Gamberger et al. (2002), was used in the analysis of several medical domains. It is the only subgroup visualization method that offers an insight of the visualized subgroups. The approach assumes the existence of at least one numeric (or ordered discrete) attribute of expert's interest for subgroup analysis. The selected attribute is plotted on the X-axis of the diagram. The Y-axis represents the target variable, or more precisely, the number of instances belonging to target property C (shown on the $Y+$ axis) or not belonging to C (shown on the $Y-$ axis) for the values of the attribute on the X-axis. It must be noted that both directions of the Y-axis are used to indicate the number of instances. The entire data set and two subgroups A1 and B2 are visualized by their distribution over a continuous attribute in Figure 10.

This visualization method is not completely automatic, since the automatic approach does not provide consistent results. The automatic approach calculates the number of examples for each value of the attribute on the X-axis by moving a sliding window and counting the number of examples in that window. The outcome is a smooth line. The difficulty arises when the attribute from the X-axis appears in the subgroup description. In such a case, a manual correction is needed for this method to be realistic.

This visualization method is very intuitive since it practically does not need much explanation. It is attractive and very useful to the end user since it offers an insight in the contents of displayed

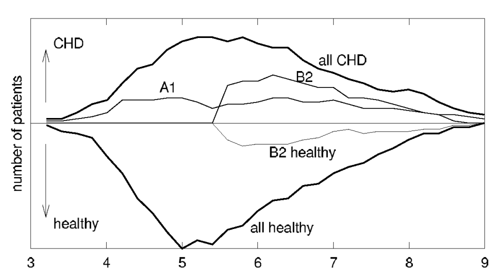

Figure 10: Subgroup visualization w.r.t. a continuous attribute. For clarity of the picture, only the positive (Y+) side of subgroup A1 is depicted.

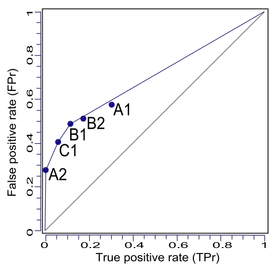

Figure 11: Representation of subgroups in the ROC space.

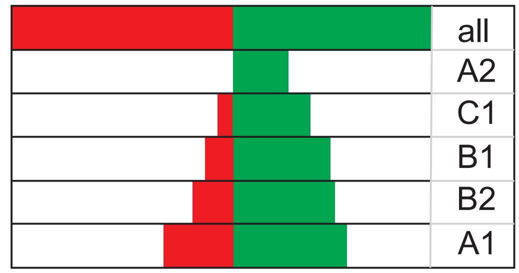

Figure 12: Subgroup visualization by bar charts.

examples. However, the correctness of displayed data is questionable. It is impossible to generalize this visualization to multi-class problems.

4.4 Representation in the ROC Space

The ROC (Receiver Operating Characteristics) (Provost and Fawcett, 2001) space is a 2-dimensional space that shows classifier (rule/rule set) performance in terms of its false positive rate (FPr) plotted on the X-axis, and true positive rate (TPr) plotted on the Y-axis. The ROC space is appropriate for measuring the success of subgroup discovery, since subgroups whose $\frac{TPr}{FPr}$ tradeoffs are close to the main diagonal (line connecting the points (0, 0) and (1, 1) in the ROC space) can be discarded as insignificant (Kavšek and Lavrač, 2006); the reason is that the rules with the $\frac{TPr}{FPr}$ ration on the main diagonal have the same distribution of covered positives and negatives ($TPr = FPr$) as the distribution in the entire data set. An example of five subgroups represented in the ROC space is shown in Figure 11.

Even though the ROC space is an appropriate rule visualization, it is usually used just for the evaluation of discovered rules. The ROC convex hull is the line connecting the potentially optimal subgroups. The area under the ROC convex hull (AUC, area under curve) is a measure of quality of the resulting ruleset.[7]

This visualization method is not intuitive to the end user, but is absolutely clear to every machine learning expert. The displayed data is correct, but there is no content displayed. An advantage of this method compared to the other visualization methods is that it allows the comparison of outcomes of different algorithms at the same time. The ROC space is designed for two-class problems and is therefore inappropriate for multi-class problems.

4.5 Bar Charts Visualization

The visualization by bar charts was introduced by Kralj et al. (2005). In this visualization, the purpose of the first line 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 of the target class. 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 (precision).

7. Note that in terms of $\frac{TPr}{FPr}$ ratio optimality, two subgroups (A1 and B2) are suboptimal, lying below the ROC convex hull.

An example of five subgroups visualized by bar charts is shown in Figure 12. It is simple, understandable and shows all the data correctly. This visualization method allows simple comparison between subgroups and is therefore useful. It is relatively straight-forward to understand and can be extended to multi-class problems. It does not display the contents of data, though.

4.6 Summary of Subgroup Visualization Methods

In this section, we (subjectively) compare the five different subgroup visualization methods by considering their intuitiveness, correctness of displayed data, usefulness, ability to ability to display contents besides the numerical properties of subgroups, (e.g., plot subgroup probability densities against the values of an attribute), and their extensibility to multi-class problems. The summary of the evaluation is presented in Table 6.

	Pie chart	Box plot	Continuous attribute	ROC	Bar chart
Intuitiveness	+	-	+	+/-	+
Correctness	-	+	-	+	+
Usefulness	-	+	+	+	+
Contents	-	-	+	-	-
Multi-class	+	-	-	-	+

Table 6: Our evaluation of subgroup visualization methods.

Two visualizations score best in Table 6 of our evaluation of subgroup visualization methods: the visualization of subgroups w.r.t. a continuous attribute and the bar chart visualization. The visualization of subgroups w.r.t. a continuous attribute is the only visualization that directly shows the contents of the data; its main shortcomings are the doubtful correctness of the displayed data and its difficulty to be extended to multi-class problems. It also requires a continuous or ordered discrete attribute in the data. The bar chart visualization combines the good properties of the pie chart and the box plot visualization. In Table 6, it only fails in displaying the contents of the data. By using the two best visualizations, one gets a very good understanding of the mining results.

To show the applicability of subgroup discovery visualizations for supervised descriptive rule discovery, the bar visualizations of results of contrast set mining, jumping emerging patterns and subgroup discovery on the survey data analysis problem of Section 2 are shown in Figures 13, 14 and 15, respectively.

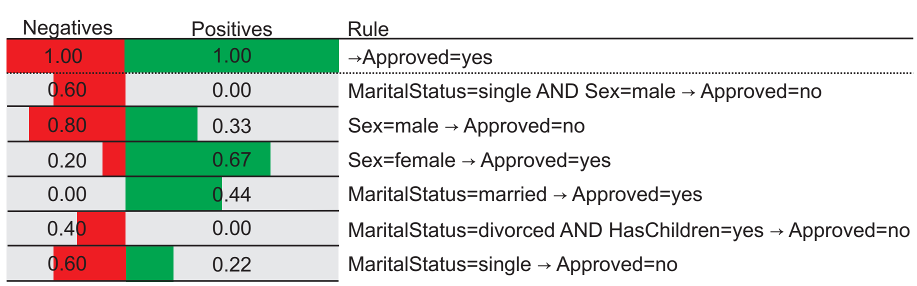

Figure 13: Bar visualization of contrast sets of Figure 3.

Negatives	Positives	Rule
1.00	1.00	→Approved=yes
0.60	0.00	MaritalStatus=single AND Sex=male → Approved=no
0.00	0.44	MaritalStatus=married → Approved=yes
0.40	0.00	MaritalStatus=divorced AND HasChildren=yes → Approved=no

Figure 14: Bar visualization of jumping emerging patterns of Figure 4.

Negatives	Positives	Rule
1.00	1.00	→Approved=yes
0.00	0.44	MaritalStatus=married → Approved=yes
0.00	0.33	MaritalStatus=divorced AND HasChildren=no → Approved=yes
0.20	0.67	Sex=female → Approved=yes
0.20	0.33	Education=university → Approved=yes

Figure 15: Bar visualization of subgroups of Figure 5 of individuals who have approved the issue.

5. Conclusions

Patterns in the form of rules are intuitive, simple and easy for end users to understand. Therefore, it is not surprising that members of different communities have independently addressed supervised descriptive rule induction, each of them solving similar problems in similar ways and developing vocabularies according to the conventions of their respective research communities.

This paper sheds a new light on previous work in this area by providing a systematic comparison of the terminology, definitions, goals, algorithms and heuristics of contrast set mining (CSM), emerging pattern mining (EPM) and subgroup discovery (SD) in a unifying framework called supervised descriptive rule discovery. We have also shown that the heuristics used in CSM and EPM can be translated into two well-known heuristics used in SD, both aiming at trading-off between coverage and distributional difference. In addition, the paper presents a critical survey of existing visualization methods, and shows that some methods used in subgroup discovery can be easily adapted for use in CSM and EPM.

Acknowledgments

The work of Petra Kralj and Nada Lavrač was funded by the project Knowledge Technologies (grant no. P2-0103) funded by the Slovene National Research Agency, and co-funded by the European 6FP project IQ - *Inductive Queries for Mining Patterns and Models* (IST-FP6-516169). Geoff Webb's contribution to this work has been supported by Australian Research Council Grant DP0772238.

References

Rakesh Agrawal, Heikki Mannila, Ramakrishnan Srikant, Hannu Toivonen, and A. Inkeri Verkamo. Fast discovery of association rules. *Advances in Knowledge Discovery and Data Mining*, pages 307–328, 1996.

Martin Atzmüller and Frank Puppe. Semi-automatic visual subgroup mining using VIKAMINE. *Journal of Universal Computer Science (JUCS), Special Issue on Visual Data Mining*, 11(11):

1752–1765, 2005.

Martin Atzmüller and Frank Puppe. SD-Map - a fast algorithm for exhaustive subgroup discovery. In *Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD-06)*, pages 6–17, 2006.

Martin Atzmüller, Frank Puppe, and Hans-Peter 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, 2005a.

Martin Atzmüller, Frank Puppe, and Hans-Peter Buscher. Profiling examiners using intelligent subgroup mining. In *Proceedings of the 10th Workshop on Intelligent Data Analysis in Medicine and Pharmacology (IDAMAP-05)*, pages 46–51, 2005b.

Yonatan Aumann and Yehuda Lindell. A statistical theory for quantitative association rules. In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-99)*, pages 261–270, 1999.

Stephen D. Bay. Multivariate discretization of continuous variables for set mining. In *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2000)*, pages 315–319, 2000.

Stephen D. Bay and Michael J. Pazzani. Detecting group differences: Mining contrast sets. *Data Mining and Knowledge Discovery*, 5(3):213–246, 2001.

Roberto J. Bayardo. Efficiently mining long patterns from databases. In *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data (SIGMOD-98)*, pages 85–93, 1998.

Anne-Laure Boulesteix, Gerhard Tutz, and Korbinian Strimmer. A CART-based approach to discover emerging patterns in microarray data. *Bioinformatics*, 19(18):2465–2472, 2003.

Peter Clark and Tim Niblett. The CN2 induction algorithm. *Machine Learning*, 3(4):261–283, 1989.

William W. Cohen. Fast effective rule induction. In *Proceedings of the 12th International Conference on Machine Learning (ICML-95)*, pages 115–123, 1995.

Olena Daly and David Taniar. Exception rules in data mining. In *Encyclopedia of Information Science and Technology (II)*, pages 1144–1148. 2005.

María José del Jesus, Pedro González, Francisco Herrera, and Mikel Mesonero. Evolutionary fuzzy rule induction process for subgroup discovery: A case study in marketing. *IEEE Transactions on Fuzzy Systems*, 15(4):578–592, 2007.

Guozhu Dong and Jinyan Li. Efficient mining of emerging patterns: Discovering trends and differences. In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-99)*, pages 43–52, 1999.

Guozhu Dong, Xiuzhen Zhang, Limsoon Wong, and Jinyan Li. CAEP: Classification by aggregating emerging patterns. In *Proceedings of the 2nd International Conference on Discovery Science (DS-99)*, pages 30–42, 1999.

Hongjian Fan and Kotagiri Ramamohanara. A bayesian approach to use emerging patterns for classification. In *Proceedings of the 14th Australasian Database Conference (ADC-03)*, pages 39–48, 2003.

Hongjian Fan and Kotagiri Ramamohanarao. Efficiently mining interesting emerging patterns. In *Proceeding of the 4th International Conference on Web-Age Information Management (WAIM-03)*, pages 189–201, 2003.

Hongjian Fan, Ming Fan, Kotagiri Ramamohanarao, and Mengxu Liu. Further improving emerging pattern based classifiers via bagging. In *Proceedings of the 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD-06)*, pages 91–96, 2006.

Jerome H. Friedman and Nicholas I. Fisher. Bump hunting in high-dimensional data. *Statistics and Computing*, 9(2):123–143, 1999.

Johannes Fürnkranz and Peter A. Flach. An analysis of rule evaluation metrics. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 202–209, 2003.

Dragan Gamberger and Nada Lavrač. Expert-guided subgroup discovery: Methodology and application. *Journal of Artificial Intelligence Research*, 17:501–527, 2002.

Dragan Gamberger, Nada Lavrač, and Dietrich Wettschereck. Subgroup visualization: A method and application in population screening. In *Proceedings of the 7th International Workshop on Intelligent Data Analysis in Medicine and Pharmacology (IDAMAP-02)*, pages 31–35, 2002.

Gemma C. Garriga, Petra Kralj, and Nada Lavrač. Closed sets for labeled data. In *Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD-06)*, pages 163 – 174, 2006.

Robert J. Hilderman and Terry Peckham. A statistically sound alternative approach to mining contrast sets. In *Proceedings of the 4th Australia Data Mining Conference (AusDM-05)*, pages 157–172, 2005.

Jože Jenkole, Petra Kralj, Nada Lavrač, and Alojzij Sluga. A data mining experiment on manufacturing shop floor data. In *Proceedings of the 40th International Seminar on Manufacturing Systems (CIRP-07)*, 2007. 6 pages.

Branko Kavšek and Nada Lavrač. APRIORI-SD: Adapting association rule learning to subgroup discovery. *Applied Artificial Intelligence*, 20(7):543–583, 2006.

Willi Klösgen. Explora: A multipattern and multistrategy discovery assistant. *Advances in Knowledge Discovery and Data Mining*, pages 249–271, 1996.

Willi Klösgen and Michael May. Spatial subgroup mining integrated in an object-relational spatial database. In *Proceedings of the 6th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD-02)*, pages 275–286, 2002.

Willi Klösgen, Michael May, and Jim Petch. Mining census data for spatial effects on mortality. *Intelligent Data Analysis*, 7(6):521–540, 2003.

Ron Kohavi and Foster Provost, editors. *Editorial for the Special Issue on Applications of Machine Learning and the Knowledge Discovery Process, Glossary of Terms*, 1998.

Petra Kralj, Nada Lavrač, and Blaž Zupan. Subgroup visualization. In *8th International Multiconference Information Society (IS-05)*, pages 228–231, 2005.

Petra Kralj, Ana Rotter, Nataša Toplak, Kristina Gruden, Nada Lavrač, and Gemma C. Garriga. Application of closed itemset mining for class labeled data in functional genomics. *Informatica Medica Slovenica*, (1):40–45, 2006.

Petra Kralj, Nada Lavrač, Dragan Gamberger, and Antonija Krstačić. Contrast set mining for distinguishing between similar diseases. In *Proceedings of the 11th Conference on Artificial Intelligence in Medicine (AIME-07)*, pages 109–118, 2007a.

Petra Kralj, Nada Lavrač, Dragan Gamberger, and Antonija Krstačić. Contrast set mining through subgroup discovery applied to brain ischaemia data. In *Proceedings of the 11th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining : (PAKDD-07)*, pages 579–586, 2007b.

Nada Lavrač, Bojan Cestnik, Dragan Gamberger, and Peter A. Flach. Decision support through subgroup discovery: Three case studies and the lessons learned. *Machine Learning Special issue on Data Mining Lessons Learned*, 57(1-2):115–143, 2004a.

Nada Lavrač, Branko Kavšek, Peter A. Flach, and Ljupčo Todorovski. Subgroup discovery with CN2-SD. *Journal of Machine Learning Research*, 5:153–188, 2004b.

Nada Lavrač, Petra Kralj, Dragan Gamberger, and Antonija Krstačić. Supporting factors to improve the explanatory potential of contrast set mining: Analyzing brain ischaemia data. In *Proceedings of the 11th Mediterranean Conference on Medical and Biological Engineering and Computing (MEDICON-07)*, pages 157–161, 2007.

Jinyan Li and Limsoon Wong. Identifying good diagnostic gene groups from gene expression profiles using the concept of emerging patterns. *Bioinformatics*, 18(10):1406–1407, 2002.

Jinyan Li, Guozhu Dong, and Kotagiri Ramamohanarao. Instance-based classification by emerging patterns. In *Proceedings of the 14th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD-2000)*, pages 191–200, 2000.

Jinyan Li, Guozhu Dong, and Kotagiri Ramamohanarao. Making use of the most expressive jumping emerging patterns for classification. *Knowledge and Information Systems*, 3(2):1–29, 2001.

Jinyan Li, Huiqing Liu, James R. Downing, Allen Eng-Juh Yeoh, and Limsoon Wong. Simple rules underlying gene expression profiles of more than six subtypes of acute lymphoblastic leukemia (ALL) patients. *Bioinformatics*, 19(1):71–78, 2003.

Jessica Lin and Eamonn Keogh. Group SAX: Extending the notion of contrast sets to time series and multimedia data. In *Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD-06)*, pages 284–296, 2006.

Bing Liu, Wynne Hsu, Heng-Siew Han, and Yiyuan Xia. Mining changes for real-life applications. In *Proceedings of the 2nd International Conference on Data Warehousing and Knowledge Discovery (DaWaK-2000)*, pages 337–346, 2000.

Bing Liu, Wynne Hsu, and Yiming Ma. Discovering the set of fundamental rule changes. In *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-01)*, pages 335–340, 2001.

Michael May and Lemonia Ragia. Spatial subgroup discovery applied to the analysis of vegetation data. In *Proceedings of the 4th International Conference on Practical Aspects of Knowledge Management (PAKM-2002)*, pages 49–61, 2002.

Foster J. Provost and Tom Fawcett. Robust classification for imprecise environments. *Machine Learning*, 42(3):203–231, 2001.

J. Ross Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

J. Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

Mondelle Simeon and Robert J. Hilderman. Exploratory quantitative contrast set mining: A discretization approach. In *Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence - Vol.2 (ICTAI-07)*, pages 124–131, 2007.

K.K.W. Siu, S.M. Butler, T. Beveridge, J.E. Gillam, C.J. Hall, A.H. Kaye, R.A. Lewis, K. Mannan, G. McLoughlin, S. Pearson, A.R. Round, E. Schultke, G.I. Webb, and S.J. Wilkinson. Identifying markers of pathology in SAXS data of malignant tissues of the brain. *Nuclear Instruments and Methods in Physics Research A*, 548:140–146, 2005.

Hee S. Song, Jae K. Kimb, and Soung H. Kima. Mining the change of customer behavior in an internet shopping mall. *Expert Systems with Applications*, 21(3):157–168, 2001.

Arnaud Soulet, Bruno Crmilleux, and Franois Rioult. Condensed representation of emerging patterns. In *Proceedings of the 8th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD-04)*, pages 127–132, 2004.

Einoshin Suzuki. Data mining methods for discovering interesting exceptions from an unsupervised table. *Journal of Universal Computer Science*, 12(6):627–653, 2006.

Ke Wang, Senqiang Zhou, Ada W.-C. Fu, and Jeffrey X. Yu. Mining changes of classification by correspondence tracing. In *Proceedings of the 3rd SIAM International Conference on Data Mining (SDM-03)*, pages 95–106, 2003.

Geoffrey I. Webb. OPUS: An efficient admissible algorithm for unordered search. *Journal of Artificial Intelligence Research*, 3:431–465, 1995.

Geoffrey I. Webb. Discovering associations with numeric variables. In *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-01)*, pages 383–388, 2001.

Geoffrey I. Webb. Discovering significant patterns. *Machine Learning*, 68(1):1–33, 2007.

Geoffrey I. Webb, Shane M. Butler, and Douglas Newlands. On detecting differences between groups. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-03)*, pages 256–265, 2003.

Dietrich Wettschereck. A KDDSE-independent PMML visualizer. In *Proceedings of 2nd Workshop on Integration Aspects of Data Mining, Decision Support and Meta-Learning (IDDM-02)*, pages 150–155, 2002.

Tzu-Tsung Wong and Kuo-Lung Tseng. Mining negative contrast sets from data with discrete attributes. *Expert Systems with Applications*, 29(2):401–407, 2005.

Stefan Wrobel. An algorithm for multi-relational discovery of subgroups. In *Proceedings of the 1st European Conference on Principles of Data Mining and Knowledge Discovery (PKDD-97)*, pages 78–87, 1997.

Stefan Wrobel. Inductive logic programming for knowledge discovery in databases. In Sašo Džeroski and Nada Lavrač, editors, *Relational Data Mining*, chapter 4, pages 74–101. 2001.

Filip Železný and Nada Lavrač. Propositionalization-based relational subgroup discovery with RSD. *Machine Learning*, 62:33–63, 2006.