SUPERVISED DESCRIPTIVE RULE INDUCTION

Doctoral Dissertation
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia, March 2009

Supervisor: Prof. Dr. Nada Lavrač, Jožef Stefan Institute, Ljubljana, Slovenia

Evaluation Board:
Prof. Dr. Sašo Džeroski, Chairman, Jožef Stefan Institute, Ljubljana, Slovenia
Assoc. Prof. Dr. Ljupčo Todorovski, Member, Faculty of Administration, University of Ljubljana, Slovenia
Dr. Igor Mozetič, Member, Jožef Stefan Institute, Ljubljana, Slovenia

Petra Kralj Novak

SUPERVISED DESCRIPTIVE RULE INDUCTION

Doctoral Dissertation

NADZOROVANO UČENJE OPISNIH PRAVIL

Doktorska disertacija

Supervisor: Prof. Dr. Nada Lavrač

March 2009

MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA
JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL
Ljubljana, Slovenia

To my husband

Contents

Abstract

The goal of knowledge discovery in databases is to construct models or discover interesting patterns in data. 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.

In contrast to predictive rule induction where the 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.

This thesis introduces the term *supervised descriptive rule induction* (SDRI), as a unification of several areas of machine learning that deal with finding comprehensible rules from class labeled data. We developed a unifying framework for contrast set mining, emerging pattern mining and subgroup discovery, as representatives of supervised descriptive rule induction approaches, which includes the unification of the terminology, definitions and heuristics. By using our SDRI framework, we overcame some open issues and limitations of SDRI sub-areas, like presenting the results to end users by visualization and supporting factors. Applications of SDRI methods to real-life datasets are also demonstrated. In collaboration with domain experts, this led to new insights in the analyzed domains and to methodology developments. A new method called mining of closed sets for labeled data (with the algorithm RelSets) and its application in microarray data analysis is also presented. The main algorithms used in the experiments are available on-line.

Povzetek

Na vseh področjih človekovega delovanja smo priča razkoraku med količino podatkov, ki se shranjujejo na elektronskih medijih, ter človeško sposobnostjo interpretiranja teh podatkov. Kot odgovor na izzive analiziranja vse večjih količin podatkov se uveljavlja raziskovalno področje, ki se imenuje *odkrivanje zakonitosti v podatkih*. V disertaciji obravnavamo posebno podpodročje odkrivanja zakonitosti v podatkih, ki se ukvarja z avtomatskim učenjem pravil, ki so namenjena človeški interpretaciji.

Uvod

Kot odgovor na izzive analiziranja vse večjih količin podatkov se v zadnjih letih uveljavlja raziskovalno področje, ki se imenuje odkrivanje zakonitosti v podatkih (*ang. knowledge discovery in data(bases)*, s kratico KDD) (Cios *et al.*, 2007; Frawley *et al.*, 1991). Namen področja je razvoj metodologij, tehnologij in standardov za odkrivanje novih, veljavnih, netrivialnih, zanimivih in potencialno uporabnih zakonitosti iz (baz) podatkov. KDD združuje metode, razvite na področjih podatkovnih baz, strojnega učenja, razpoznavanja vzorcev, statistike, umetne inteligence in vizualizacije podatkov. Posebno vlogo v KDD imajo metode strojnega učenja (Michalski *et al.*, 1986; Mitchell, 1997), ki iz podatkov izluščijo modele in vzorce.

Metode strojnega učenja lahko v grobem razdelimo v dve kategoriji: nadzorovano učenje (Kotsiantis *et al.*, 2006) in nenadzorovano učenje (Ghahramani, 2004). Pri nadzorovanem učenju so podatki označeni s ciljno spremenljivko in cilj učenja je zgraditi model, ki bo preslikal ostale podatke v vrednosti ciljne spremenljivke. Pogosto se nadzorovano učenje enači z *napovedno indukcijo* (*ang. predictive induction*) (Weiss in Indurkhya, 1998), kjer je namen uporabiti zgrajen model za napovedovanje vrednosti ciljne spremenljivke pri novih primerih. Pri nenadzorovanem učenju ciljna spremenljivka ni dana, cilj učenja pa je odkrivanje splošno veljavnih vzorcev, ki so v podatkih. Nenadzorovano učenje se enači z *opisno indukcijo* (*ang. descriptive induction*), kjer uporabnika zanima razumevanje podatkov.

V disertaciji smo osredotočeni na situacijo, ko imamo podatke v obliki primerni za nadzorovano učenje, torej imamo dano ciljno spremenljivko, iz njih pa želimo izluščiti opisna pravila, ki naj služijo razumevanju podatkov. Pravila so oblike "*ČE pogoji POTEM ciljna spremenljivka = vrednost*", kar je običajna oblika za klasifikacijska pravila (Clark in Niblett, 1989), a neobičajna za opisna pravila (Agrawal *et al.*, 1996).

Cilji disertacije

Glavni cilj disertacije je združitev področij, ki se ukvarjajo z iskanjem opisnih pravil iz označenih podatkov, s katero ustvarimo nov poenoten teoretski okvir z imenom *nadzorovano učenje opisnih pravil* (*ang. supervised descriptive rule induction*, s kratico SDRI). S tem pridobijo posamezna področja (podpodročja nadzorovanega učenja opisnih pravil), ki so bila do sedaj ločena in so se neodvisno razvijala, saj so določeni problemi še nerešeni na enem področju, medtem ko so rešitve za isti problem na drugem področju že dobro razvite. Prednosti razvitega teoretskega okvira demonstriramo z uspešnimi aplikacijami v medicini in biologiji.

Nadzorovano učenje opisnih pravil

V disertaciji definiramo termin *nadzorovano učenje opisnih pravil* (Kralj Novak *et al.*, 2009b), v okvire katerega združimo odkrivanje podskupin (*ang. subgroup discovery*) (Lavrač *et al.*, 2004; Wrobel, 1997), odkrivanje kontrastnih množic (*ang. contrast set mining*) (Bay in Pazzani, 2001; Webb *et al.*, 2003), odkrivanje porajajočih se vzorcev (*ang. emerging pattern mining*) (Dong in Li, 1999) in druga sorodna področja. Našteta področja se ukvarjajo z učenjem vzorcev v obliki pravil iz označenih podatkov, uporabljajo pa različne terminologije, definicije ciljev in različne formulacije hevristik. V skupnem teoretskem okviru poenotimo terminologijo, kar omogoči tudi poenotenje definicij. Kompatibilnost hevristik dokažemo s prevedbami formul in s prikazom izometrik v ROC (*ang. reciever operator characteristic*) prostoru.

Uporaba poenotenega teoretskega okvira ima več prednosti. Prvi rezultat uporabe je posplošitev vizualizacijskih metod, ki so bile razvite za odkrivanje podskupin (Atzmüller in Puppe, 2005; Gamberger *et al.*, 2002; Kralj *et al.*, 2005; Wettschereck, 2002; Wrobel, 2001), na nadzorovano učenje opisnih pravil (Kralj Novak *et al.*, 2009b). S tem rešimo odprto vprašanje predstavitve rezultatov uporabnikom s področja odkrivanja kontrastnih množic, ki so ga izpostavili Webb *et al.* (2003).

Analiza podatkov o pacientih z možgansko kapjo, ki smo jo opravili v tesnem sodelovanju z ekspertom s tega področja, je privedla do razvoja ustreznejše metodologije za odkrivanje kontrastnih množic in novih spoznanj na področju problemske domene. Cilj analize je bil odkriti dejavnike tveganja za možgansko kap. Podatki so zajemali tako vsebino kartotek pacientov z možgansko kapjo (razlikovali smo med dvema vrstama možganske kapi) kot tudi kartoteke pacientov z drugimi nevrološkimi motnjami. Problem smo definirali v obliki odkrivanja kontrastnih množic med dvema vrstama možganske kapi in z uporabo SDRI teoretičnega okvira razvili teoretično korektno transformacijo odkrivanja kontrastnih množic z uporabo odkrivanja podskupin (Kralj Novak *et al.*, 2009a). Diskusija o rezultatih je privedla do drugačne, a v danih okoliščinah primernejše transformacije problema, ki nas je pripeljala tudi do boljših rezultatov. Raziskave so potekale v več iteracijah in v vsaki iteraciji so bile uporabljene tudi vizualizacijske metode. Za podkrepitev končnih rezultatov smo uporabili podporne dejavnike (*ang. supporting factors*) (Gamberger *et al.*, 2003), ki smo jih s pomočjo SDRI teoretskega okvira posplošili iz odkrivanja podskupin na odkrivanje kontrastnih množic (Kralj Novak *et al.*, 2009a).

Razvili smo novo metodo nadzorovanega učenja opisnih pravil, ki omogoča prilagoditev metode odkrivanja *zaprtih množic* (*ang. closed sets*) za delo z označenimi podatki. Nova metoda se imenuje *odkrivanje zaprtih množic za označene podatke* (*ang. mining of closed sets for labeled data*) (Garriga *et al.*, 2008) in ima lepe teoretske lastnosti, kot so zagotavljanje optimalnosti v ROC prostoru in zagotavljanje neredundantnosti odkritih pravil. To metodo smo uporabili za analizo podatkov mikromrež krompirja, kjer smo iskali razlike med na virus občutljivimi in neobčutljivimi transgenimi linijami krompirja (Garriga *et al.*, 2008; Kralj *et al.*, 2006). Za razumevanje problema, pripravo podatkov in interpretacijo rezultatov smo sodelovali z eksperti, ki so iz rezultatov lahko razbrali, kateri geni vplivajo na občutljivost in časovni odziv teh genov (Baebler *et al.*, 2009). Metoda odkrivanje zaprtih množic za označene podatke je primerna za analizo podatkov mikromrež, ker nima težav z nesorazmerjem med velikim številom atributov in malim številom primerov, tipičnim za podatke mikromrež, kar predstavlja težavo pri uporabi drugih metod nadzorovanega učenja opisnih pravil.

V okolju Orange (Demšar *et al.*, 2004) smo razvili orodje za odkrivanje podskupin *Subgroup Discovery Toolkit for Orange*, dostopno pod GPL licenco na spletni strani `http://kt.ijs.si/petra_kralj/SubgroupDiscovery/`. Orodje vključuje tri algoritme za odkrivanje podskupin: SD (Gamberger in Lavrač, 2002), CN2-SD (Lavrač *et al.*, 2004) in Apriori-SD (Kavšek in Lavrač, 2006), dve metodi vizualizacije (palična vizualizacija in vizualizacija v ROC prostoru (Kralj *et al.*, 2005)) in postopek za evalvacijo podskupin (Kavšek in Lavrač, 2006). Algoritem za iskanje zaprtih množic za označene podatke *RelSets* je dostopen na spletni strani `http://kt.ijs.si/petra_kralj/RelSets/` kot spletni servis. Ostali algoritmi, uporabljeni v disertaciji (npr. Magnum Opus), so dostopni pri njihovih avtorjih.

Prispevki k znanosti

V disertaciji so opisani naslednji prispevki k znanosti.

- Pregled področij odkrivanja podskupin, odkrivanja kontrastnih množic, porajajočih se vzorcev in ostalih sorodnih pristopov (2. poglavje).
- Razvoj poenotenega teoretskega okvira za odkrivanje podskupin, odkrivanje kontrastnih množic in odkrivanje porajajočih se vzorcev z imenom *nadzorovano učenje opisnih pravil*. Teoretski okvir poenoti terminologijo, definicije in hevristike navedenih področij (2. poglavje).
- Kritični pregled metod za vizualizacijo nadzorovanih opisnih pravil in njihova posplošitev v sklopu razvitega teoretskega okvira (2. poglavje).
- Metodologija za *odkrivanje kontrastnih množic* z *odkrivanjem podskupin* (3. poglavje).
- Prilagoditev koncepta *podpornih dejavnikov* (supporting factors) iz odkrivanja podskupin na odkrivanje kontrastnih množic (3. poglavje).
- Praktična aplikacija algoritmov nadzorovanega učenja opisnih pravil na domeni možganske kapi. (3. poglavje).
- Prilagoditev metode za odkrivanje *zaprtih množic* (closed sets) za delo z označenimi podatki. Nova metoda se imenuje *odkrivanje zaprtih množic za označene podatke* (*ang. mining of closed sets for labeled data*) (4. poglavje).

- Aplikacija metode odkrivanja zaprtih množic za označene podatke za analizo podatkov mikromrež. (4. poglavje).
- Implementacija algoritmov in vizualizacjskih metod za odkrivanje podskupin, odkrivanje kontrastnih množic in odkrivanje porajajočih se vzorcev (5. poglavje).

Znanstveni prispevki disertacije so bili objavljeni v treh uglednih mednarodnih revijah s področja strojnega učenja (Garriga *et al.*, 2008; Kralj Novak *et al.*, 2009a,b) in na številnih mednarodnih znanstvenih konferencah. Seznam publikacij je podan na koncu disertacije.

Zaključki in nadaljnje delo

V disertaciji smo z uvedbo enotnega teoretskega okvira za nadzorovano učenje opisnih pravil dosegli zastavljene cilje, kar smo dokazali z uspešnimi aplikacijami tega pristopa. Orodja za analizo podatkov, ki smo jih uporabili v disertaciji, so prosto dostopna na svetovnem spletu v obliki paketa za programsko okolje Orange (Demšar *et al.*, 2004).

Ena od smernic za nadaljnje delo je dekompozicija SDRI pristopov na predprocesiranje, algoritme same in evalvacijo, in njihova implementacija v obliki povezljivih spletnih servisov. Z definicijo primernega vmesnika med servisi bi lahko omogočili povezovanje in uporabo kombinacij vseh pristopov, ki so na voljo.

Področji, ki sta zaenkrat še razmeroma neraziskani in jih v disertaciji nismo obravnavali, sta tudi odkrivanje zakonitosti iz kompleksnih podatkovnih struktur in semantično odkrivanje zakonitosti v podatkih. Na področju odkrivanja podskupin je Wrobel (1997, 2001) razvil algoritem Midos, Klösgen in May (2002) pa sta razvila algoritem SubgroupMiner, ki je prilagojen odkrivanju zakonitosti v prostorskih podatkih. Na to področje spada tudi algoritem RSD (*ang. Relational subgroup discovery*) avtorjev Železný in Lavrač (2006). Algoritem SEGS (*ang. Search for enriched gene sets*) avtorjev Trajkovski *et al.* (2008) predstavlja uspešen način semantičnega odkrivanja zakonitosti v podatkih v obliki opisnih pravil, saj uporablja specializirane biološke ontologije kot predznanje za gradnjo pravil, ki razlagajo podatke mikromrež.

Kljub temu, da je disertacija osredotočena na metode strojnega učenja za gradnjo razumljivih pravil, so uporaba in širjenje SDRI pristopa še kako pomembni.

S tem, ko smo naredili algoritme dostopne na svetovnem spletu, smo naredili prvi korak k približevanju le teh končnim uporabnikom. Objave v poljudnoznanstvenih medijih in predavanja različnim publikam bi gotovo veliko prispevali k uporabi metod v praksi.

Abbreviations

CSM	=	contrast set mining
DBMS	=	data base management system
DM	=	data mining
EPM	=	emerging pattern mining
KDD	=	knowledge discovery in databases
ROC	=	receiver operator characteristics
SD	=	subgroup discovery
SDRI	=	supervised descriptive rule induction

1 Introduction

This chapter introduces the terminology used in the dissertation, presents the motivation, the hypothesis and goals of this work, and provides a list of specific scientific contributions of this thesis.

1.1 Background

Nowadays, computer-based systems are applied in almost every aspect of everyday life. In various domains, like business, science and medicine, huge amounts of data are being collected on a daily basis. The data can then be analyzed and the analysis results can be used to discover market trends, to improve customer service, to gain insights in the domain and for decision support in general. Data analysis is frequently performed in the knowledge discovery process (Chapman *et al.*, 1999; Cios *et al.*, 2007; Frawley *et al.*, 1991), which is defined as "the non-trivial extraction of implicit, unknown, and potentially useful information from data" (Frawley *et al.*, 1991).

The concept *knowledge discovery from databases* (KDD) emerged in 1989 to refer to the process of finding interesting patterns and models in data. According to Fayyad *et al.* (1996), the KDD process is interactive and iterative (with many decisions made by the user), involving numerous steps, summarized as:

1. Learning the application domain: includes relevant prior knowledge and the goals of the application;

2. Creating a target dataset: includes selecting a dataset or focusing on a subset of variables or data samples on which discovery is to be performed;

3. Data cleaning and preprocessing: includes basic operations, such as removing noise or outliers if appropriate, collecting the necessary information to model or account for noise, deciding on strategies for handling missing data fields, and accounting for time sequence information and known changes, as well as deciding DBMS issues, such as data types, schema, and mapping of missing and unknown values;

4. Data reduction and projection: includes finding useful features to represent the data, depending on the goal of the task, and using dimensionality reduction or transformation methods to reduce the effective number of variables under consideration or to find invariant representations for the data;

5. Choosing the function of data mining: includes deciding the purpose of the model derived by the data mining algorithm (e.g., summarization, classification, regression or clustering);

6. Choosing the data mining algorithm(s): includes selecting method(s) to be used for searching for patterns in the data, such as deciding which models and parameters may be appropriate (e.g., models for categorical data are different from models on vectors over reals) and matching a particular data mining method with the overall criteria of the KDD process (e.g., the user may be more interested in understanding the model than in its predictive capabilities);

7. Data mining: includes searching for patterns of interest in a particular representational form or a set of such representations, including classification rules or trees, regression, clustering, sequence modeling, dependency, and others;

8. Interpretation: includes interpreting the discovered patterns and possibly returning to any of the previous steps, as well as possible visualization of the extracted patterns, removing redundant or irrelevant patterns, and translating the useful ones into terms understandable by the user;

9. Using discovered knowledge: includes incorporating this knowledge into the performance system, taking actions based on the knowledge, or simply documenting it and reporting it to interested parties, as well as checking for and resolving potential conflicts with previously believed (or extracted) knowledge.

In summary, the knowledge discovery process is an iterative process of searching for valuable information in large volumes of data. It is a cooperative effort of humans and computers: humans design databases, describe problems, set goals and interpret results, while computers search through the data looking for patterns that meet the human-defined goals. In steps 5, 6 and 7 of the knowledge discovery process, data mining algorithms are introduced. In this context, data mining can be viewed as an application of particular algorithms for extracting patterns or models from data.

Data mining is being influenced by many other disciplines like statistics, machine learning and artificial intelligence, pattern recognition, and data visualization. In the rest of this section, the relation between machine learning and data mining will be clarified (Section 1.1.1), and rule learning will be introduced (Section 1.1.2).

1.1.1 Machine learning and data mining

Machine learning builds on concepts from the fields of artificial intelligence, statistics, information theory, and many others. It studies computer programs that automatically improve with experience (Mitchell, 1997). The main types of machine learning approaches are supervised learning (Kotsiantis *et al.*, 2006) and unsupervised learning (Ghahramani, 2004), supplemented by semi-supervised learning (Chapelle *et al.*, 2006), reinforcement learning (Sutton and Barto, 1998), transduction (Vapnik, 1998), meta learning (Vilalta and Drissi, 2002) and others. Most automatic data mining methods have their origin in machine learning. However, machine learning can not be seen as a true subset of data mining as it also encompasses fields not utilized for data mining (e.g., theory of learning, computational learning theory, and reinforcement learning) (Kononenko and Kukar, 2007).

As a sub-field of artificial intelligence (Michalski *et al.*, 1986), machine learning is inspired by one of the main properties of intelligent systems: learning. Consequently, machine learning terminology is based on human learning terminology. In human supervised learning, there is a teacher, tutor or supervisor who knows both the questions and the correct answers and, by giving the answers to the learner, helps him to learn a new skill. *Supervised machine learning* emulates the supervisor by providing the learner with labeled data for training. The training data consist of pairs of input examples - questions (usually in the form of vectors of attribute values), and desired outputs - answers (called target values or labels). The target values are most commonly either categories (in such a case we talk about classification where each possible category is one class) or numbers. The supervised learning task is to learn the skill of correctly predicting the value of outputs for valid inputs after having seen a number of labeled training examples (i.e., pairs of input and output). To achieve this, the learner has to generalize from the presented data to unseen situations. In summary, supervised learning is about learning a function that maps the inputs to the outputs based on given labeled data (Russell and Norvig, 2003). A review of supervised machine learning techniques is available in Kotsiantis *et al.* (2006).

A special case of supervised learning is *concept learning*, defined by Bruner *et al.* (1956) as "the search for and listing of attributes that can be used to distinguish exemplars from non exemplars of various categories." The learning system aims at determining a description of a given concept from a set of concept examples provided by the teacher. Concept examples can be either positive or negative. Compared to the general classification setting where the number of classes between which one aims to distinguish can be greater than two, in concept learning, one class is taken as the target and the goal to find characteristics that distinguish this class from the others. Some machine learning techniques take inspiration from concept learning, for example subgroup discovery (Wrobel, 1997).

Compared to supervised learning, *unsupervised learning* is about learning without a supervisor. To translate this to the machine learning terminology, unsupervised machine learning is about learning from data where no target values are supplied (Russell and Norvig, 2003). The learner's goal is to build representations of the input. In a sense, unsupervised learning can be thought of as finding patterns in the data above and beyond what would be considered pure unstructured noise (Ghahramani, 2004).

While machine learning focuses on the development of data modeling techniques, data mining is more application oriented (Kononenko and Kukar, 2007). Each data mining application has a motivating story that should lead to a problem specification. Supervised and unsupervised machine learning methods are the most frequently used methods in data mining.

Predictive data mining is usually associated with supervised machine learning, since it uses the data to infer predictions. For predictive data mining, representative examples with known target values, summarizing past experiences, must be available. The technical mission of predictive data mining is to induce a model for assigning labels to new unlabeled cases (Weiss and Indurkhya, 1998).

Descriptive data mining aims at finding interesting patterns in the data. Descriptive data mining describes the data in a concise way and presents interesting characteristics of the data, usually without having any predefined target. The result of descriptive data mining is a description of a set of data in a concise and summarized manner, the presentation of the general properties of the data as well as the description of local patterns in the data.

Supervised machine learning is used in predictive data mining and unsupervised machine learning is used in descriptive data mining. Besides using supervised machine learning, predictive data mining uses also other methods like sequential prediction and interpolation. Similarly, besides using unsupervised machine learning, descriptive data mining uses also other methods like correlation, associations and dependencies. Summarized and simplified, predictive data mining is supervised machine learning used in practice and descriptive data mining is unsupervised machine learning used in practice.

1.1.2 Rule induction

Rule induction is an area of machine learning in which formal rules are extracted from a set of observations. The extracted rules may represent a full model of the data (in the form of a ruleset) like in Clark and Niblett (1989), or represent local patterns in the data (in the form of individual rules) like in Agrawal *et al.* (1996). The general form of each rule is an if-then rule:

$$IF\ Conditions\ THEN\ Conclusion.$$

Conditions contain one or more (conjunction of) attribute tests, i.e., features of the form $A_i = v_{ij}$ for categorical attributes (where v_{ij} is one of the possible values of attribute A_i), and $A_i < v$ or $A_i \geq v$ for numeric attributes (where v is a threshold value that does not need to correspond to a value of the attribute observed in the examples). The form of the *Conclusion* part of the rule depends on the type of the rule.

In the supervised rule learning setting, rules are induced from labeled data. Since rules learned in a supervised manner are usually used for classification, supervised rule learning is usually associated to classification rule learning. The classification rule learning task can be defined as follows: Given a set of training examples (instances for which the classification is known), find a set of classification rules that can be used for predicting or classifying new instances, i.e., cases that have not been presented to the learner before. Classification rules are of the form:

$$IF\ Conditions\ THEN\ Class = c_i.$$

In this setting, the $Conclusion$ consists of the target variable associated with one of the examples' labels (classes). Classification rules are not individual rules, but rather parts of models (rulesets) that work together to classify new instances. Classification rulesets can either be ordered (in the if-then-else form) or unordered where each rule votes for a class label. Several classification rule learning algorithms have been developed so far, for example CN2 (Clark and Boswell, 1991; Clark and Niblett, 1989) and RIPPER (Cohen, 1995). Most classification rule learning algorithms use heuristic approaches for rule induction.

In the unsupervised rule learning setting, rules are induced from unlabeled data. The goal of unsupervised rule learning is to discover interesting relations between variables (attributes). The most famous and well researched unsupervised rule learning method is association rule learning, which was introduced by Agrawal *et al.* (1996). Association rules are of the form

$$IF\ Conditions\ THEN\ Conclusions,$$

where both the *Conditions* and the *Conclusions* are conjunctions of attribute values (or items, depending on the data format). Each rule is an individual local pattern in the data, not related to other rules. Compared to classification rule learning where heuristic approaches are commonly used, the approaches used in association rule mining are usually exhaustive and therefore guarantee optimality of results in terms of support and confidence.

Traditional association rule mining produces an abundance of redundant rules, which is due to their individuality. To overcome this redundancy, association rules based on closed frequent itemsets were introduced by Pasquier *et al.* (1999). The number of non-redundant rules produced by the new approach is substantially smaller than the rule set from the traditional approach, since closed sets provide compacted data representations.

1.2 Supervised descriptive rule induction

In Section 1.1.1, we have introduced, on the one hand, supervised and unsupervised machine learning, and, on the other hand, predictive and descriptive data mining. If simplified, it can be summarized as follows:

supervised machine learning	$\sim$	predictive data mining
unsupervised machine learning	$\sim$	descriptive data mining

However, even if the learning setting is supervised (labeled data is given) the goal can be to build symbolic descriptions intended for human interpretation. This thesis addresses this less common situation, which is a merge of supervised learning and descriptive data mining in the context of rule induction, which we have named supervised descriptive rule induction.

1.2.1 Motivation

A common question is "What is the difference between groups of individuals?", where groups are defined by a selected property of the individuals. For example, one could specify as the property of interest the gender of patients and ask the question "What is the difference between males and females affected by a certain disease?", or, if the property of interest was the response to a treatment, the question could be "What is the difference between patients reacting well to the treatment and those who are not?". Searching for differences is not limited to any special type of individuals: one can search for differences between molecules, patients, projects, organizations, etc.

Data mining tasks where the goal is to find comprehensible differences between groups have been addressed by many researchers from both the descriptive and predictive data mining side. On the one hand, in descriptive data mining - using the association rule learning perspective - association rule learners like Apriori by Agrawal *et al.* (1996) were adapted to perform the tasks named *contrast set mining* (CSM) (Bay and Pazzani, 2001) and *emerging pattern mining* (EPM) (Dong and Li, 1999). On the other hand, in predictive data mining, algorithms for building accurate classifiers (Clark and Niblett, 1989; Cohen, 1995) have been adapted to build individual rules for exploratory data analysis and interpretation, i.e., to solve the task named *subgroup discovery* (SD) (Lavrač *et al.*, 2004; Wrobel, 1997).

Past research in three distinct areas of data mining - contrast set mining, emerging pattern mining and subgroup discovery - all dealing with finding comprehensible rules for distinguishing between groups, has been performed independently of each other, using different frameworks and terminology. Since researchers in these three areas were not aware of each others' work, they could not make sufficient use of each others' discoveries.

The purpose of this dissertation is to unify data mining tasks that deal with finding differences between groups in a novel unifying framework, named *supervised descriptive rule induction* (SDRI). By doing so, our aim is to improve individual supervised descriptive rule induction methods by cross-fertilizing the approaches developed in individual subareas of supervised descriptive rule induction. Furthermore, we aim at developing novel SDRI methods through component exchange (e.g., enabling the use of subgroup discovery components like visualization and evaluation in contrast set mining). Finally, we aim at showing the advantages of this approach in applications in important real life problems in medicine and biology.

1.2.2 Supervised descriptive rule induction definition

In this dissertation, *supervised descriptive rule induction* is defined as follows. We assume a set of data exists, described by attributes and their values and a selected nominal attribute that is of interest (called the target attribute). The goal of supervised descriptive rule induction (SDRI) is to induce rules in the classification rule form to explain the relation between the target attribute and the other attributes in the data. In other words, supervised descriptive rule induction is a process of inducing a set of comprehensible rules in the classification rule form from class-labeled data. The classification rule format constrains the conditions part of the rules to be in conjunctive form, and to have the target attribute and one of its values in the conclusions part of the rules.

The purpose of supervised descriptive rule induction is to allow the user, interested in the data, to browse through the rules and, by doing so, to gain insight in the data domain. The final goal of supervised descriptive rule induction is to allow the user to understand the phenomena underlying the data.

1.2.3 Supervised descriptive rule induction areas

We have identified three main supervised descriptive rule induction areas: contrast set mining, emerging pattern mining and subgroup discovery. Other related approaches include change mining (Liu *et al.*, 2001), mining of closed sets for labeled data (Garriga *et al.*, 2008), exception rule mining (Daly and Taniar, 2005; Suzuki, 2006), bump hunting (Friedman and Fisher, 1999), quantitative association rules (Aumann and Lindell, 1999) and impact rules (Webb, 2001). The first three approaches are outlined in this section.

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 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[1] 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 the 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 a result of which they have a 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

[1] 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.

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. Other work in the contrast set mining area includes Bay (2000); Hilderman and Peckham (2005); Lin and Keogh (2006); Wong and Tseng (2005) and Simeon and Hilderman (2007).

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.

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[1] (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: $\mathit{ItemSet} \rightarrow D_1$, 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, growth rate provides an identical ordering to confidence.

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

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

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.

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', e.g., are as large as possible and have the most unusual statistical (distributional) characteristics with respect to the property of interest".

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 $\mathit{SubgroupDescription} \rightarrow \mathit{Class}$. 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 algorithms 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.* (2004), 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 spatial 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 have been developed, including an algorithm for exploiting background knowledge in subgroup discovery (Atzmüller *et al.*, 2005), 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*) is used by algorithms CN2-SD, Apriori-SD and RSD and, in a different formulation and in different variants, also by MIDOS and EXPLORA. The generalization quotient (q_g) is used by the SD algorithm, and SubgroupMiner uses the classical binominal test to verify if the target share is significantly different in a subgroup as compared to the whole population.

Different approaches have been used for eliminating redundant subgroups. Algorithms CN2-SD, Apriori-SD, SD and RSD use weighted covering (Lavrač *et al.*, 2004) to achieve rule diversity. Algorithms Explora and SubgroupMiner use an approach called subgroup suppression (Klösgen, 1996).

1.3 Hypothesis and goals

The hypothesis of this thesis is that contrast set mining, emerging pattern mining, subgroup discovery and other similar data mining approaches can be re-interpreted as special cases of a general task of supervised descriptive rule induction (SDRI), and that this novel perspective can help solving current open issues of SDRI sub-areas. Moreover, SDRI approaches can be interpreted as workflows of individual SDRI algorithmic components and therefore interchanged, leading to the development of novel SDRI algorithms, possibly improved due to the cross-fertilization of ideas from the different SDRI sub-areas. Finally, the applications of SDRI algorithms to problems in selected biomedical domains can lead to improved problem solutions and novel domain insights.

The individual research goals of this dissertation are as follows:

1. Identification of supervised descriptive rule induction (SDRI) tasks (contrast set mining, emerging pattern mining, subgroup discovery and other related approaches) and a survey of past SDRI research;

2. Unification of the SDRI terminology, definitions and heuristics;

3. Selection of SDRI evaluation measures;

4. Experimental comparison of SDRI methods (existing and novel);

5. Avoiding current limitations of SDRI approaches by improving the presentation of SDRI results to end users, which includes visualization and explanation in terms of supporting factors;

6. Applications of SDRI algorithms to practical problem domains from medicine and biology.

1.4 Scientific contributions

The contributions of this dissertation to data mining are the following.

- A state-of-the-art survey of subgroup discovery, contrast set mining, emerging pattern mining and other similar data mining approaches (Chapter 2).
- The development of a unifying framework for subgroup discovery, contrast set mining and emerging pattern mining, named supervised descriptive rule induction (SDRI). The unifying framework includes the unification of the terminology, definitions and heuristics (Chapter 2).
- A critical survey of existing supervised descriptive rule visualization methods. The visualization methods were developed in the subgroup discovery context while visualization and the presentation of results to the end user was considered an open issue in contrast set mining. By using the SDRI framework, we have analyzed the visualization methods for general SDRI purposes (Chapter 2).
- A methodology for contrast set mining though subgroup discovery (Chapter 3).
- The adaptation of the *supporting factor* concept from subgroup discovery to contrast set mining, also achieved by using the SDRI unifying framework (Chapter 3).
- An application of SDRI algorithms in a practical problem domain from medicine: the real-life dataset of patients with brain ischemia. The analysis of this dataset and interaction with a domain expert have lead to improved problem definition and solutions, and novel domain insights (Chapter 3).
- The adaptation of closed sets mining to classification and discrimination purposes. The new SDRI method is called *mining of closed sets for labeled data* (Chapter 4).
- An application of the closed sets for labeled data method in microarray data analysis (Chapter 4).
- Implementation of supervised descriptive rule induction approaches and their availability on the web (Chapter 5).

The main scientific contributions of this work were published in the following journal and conference papers:

Journal papers

- [Kralj Novak *et al.*(2009a)] Kralj Novak, P., Lavrač, N., Gamberger, D., and Krstačić, A. (2009a). CSM-SD: Methodology for contrast set mining through subgroup discovery. *Journal of Biomedical Informatics*, **42**(1), 113–122.

- [Kralj Novak *et al.*(2009b)] Kralj Novak, P., Lavrač, N., and Webb, G. I. (2009b). Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. *Journal of Machine Learning Research*, **10**, 377–403. http://www.jmlr.org/papers/volume10/kralj-novak09a/kralj-novak09a.pdf.

- [Garriga *et al.*(2008)] Garriga, G. C., Kralj, P., and Lavrač, N. (2008). Closed sets for labeled data. *Journal of Machine Learning Research*, **9**, 559–580. http://www.jmlr.org/papers/volume9/garriga08a/garriga08a.pdf.

- [Kralj *et al.*(2006)] Kralj, P., Rotter, A., Toplak, N., Gruden, K., Lavrač, N., and Garriga, G. C. (2006). Application of closed itemset mining for class labeled data in functional genomics. *Informatica Medica Slovenica*, (1), 40–45.

Conference papers

- [Kralj *et al.*(2007a)] Kralj, P., Lavrač, N., Gamberger, D., and Krstačić, A. (2007a). Contrast set mining for distinguishing between similar diseases. In *Proceedings of the 11th Conference on Artificial Intelligence in Medicine (AIME 2007)*, pages 109–118.

- [Kralj *et al.*(2007b)] Kralj, P., Lavrač, N., Gamberger, D., and Krstačić, A. (2007b). Contrast set mining through subgroup discovery applied to brain ischaemia data. In *Proceedings of the 11th Pacific-Asia conference on Knowledge Discovery and Data Mining (PAKDD 2007)*, pages 579–586.

- [Lavrač *et al.*(2007)] Lavrač, N., Kralj, P., Gamberger, D., and Krstačić, A. (2007). 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 2007)*, pages 157–161.

- [Garriga *et al.*(2006)] Garriga, G. C., Kralj, P., and Lavrač, N. (2006). Closed sets for labeled data. In *Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2006)*, pages 163–174.

- [Kralj *et al.*(2005a)] Kralj, P., Lavrač, N., Zupan, B., and Gamberger, D. (2005a). Experimental comparison of three subgroup discovery algorithms: Analysing brain ischemia data. In *Proceedings of the 8th International Multiconference Information Society (IS 2005)*, pages 220 – 223.
- [Kralj *et al.*(2005b)] Kralj, P., Lavrač, N., and Zupan, B. (2005b). Subgroup visualization. In *Proceedings of the 8th International Multiconference Information Society (IS 2005)*, pages 228–231.

Papers by Petra Kralj (Novak) that are not related to the thesis:

- [Hren *et al.*(2007)] Hren, M., Boben, J., Rotter, A., Kralj, P., Gruden, K., and Ravnikar, M. (2007). Real-time PCR detection systems for flavescence dorée and bois noir phytoplasmas in grapevine: comparison with conventional PCR detection and application in diagnostics. *Plant Pathology*, **56**, 785–796.
- [Jenkole *et al.*(2007)] Jenkole, J., Kralj, P., Lavrač, N., and Sluga, A. (2007). A data mining experiment on manufacturing shop floor data. In *Proceedings of the 40th International Seminar on Manufacturing Systems (CIRP 2007)*. 6 pages.
- [Kralj *et al.*(2007)] Kralj, P., Lavrač, N., Gruden, K., Rotter, A., Štebih, D., Morisset, D., and Žel, J. (2007). A prototype decision support system for gmo traceability. In *Proceedings of the 10th International Multiconference Information Society (IS 2007)*, pages 214–217.

1.5 Thesis structure

This thesis is structured as follows. The background, motivation and definition of supervised descriptive rule induction (SDRI) are provided in Chapter 1, where the main terminology is also introduced. This chapter includes also a brief survey of three main approaches to SDRI: contrast set mining, emerging pattern mining and subgroup discovery. The main part of this thesis are three original research papers on supervised descriptive rule induction co-authored by the author of this dissertation, which were published in internationally recognized journals of the machine learning field: Kralj Novak *et al.* (2009b), Kralj Novak *et al.* (2009a) and Garriga *et al.* (2008), presented in Chapters 2, 3 and 4, respectively. Chapter 5 is dedicated to the methodology used to prove the dissertation's thesis, the summary and discussion of results, and software availability. Chapter 6 is devoted to conclusions and further work.

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

In this chapter, the paper (Kralj Novak *et al.*, 2009b) titled "Supervised Descriptive Rule Discovery: A Unifying Survey of Contrast Set, Emerging Pattern and Subgroup Mining" by Petra Kralj Novak, Nada Lavrač and Geoffrey I. Webb is presented. The paper was published in the Journal of Machine Learning Research in February 2009.

This paper represents the core of the dissertation, since it provides a survey of supervised descriptive rule induction approaches, a unifying framework for supervised descriptive rule induction, which includes unifications of the terminology, definitions and heuristics, and a survey of visualization methods. Three representative supervised descriptive rule induction approaches are discussed in detail (subgroup discovery, emerging pattern and contrast set mining), while other related approaches are discussed in relation to the previously mentioned three approaches. The approaches are presented on a very small, artificial sample dataset, adapted from Quinlan (1986), and the visualization methods are presented on a real-life coronary heart disease dataset, both aimed at improving the clarity and understandability of the paper.

All three authors contributed significantly to this paper. Petra Kralj Novak and Nada Lavrač conceived the unification of the three supervised descriptive rule induction approaches and framed the paper, while the brainstorming with Geoffrey I. Webb lead to the final idea to make the paper a survey. After considering the comments from reviewers, we have decided to add a section on visualization (Section 4), which was adapted from the conference paper Kralj *et al.* (2005).

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.

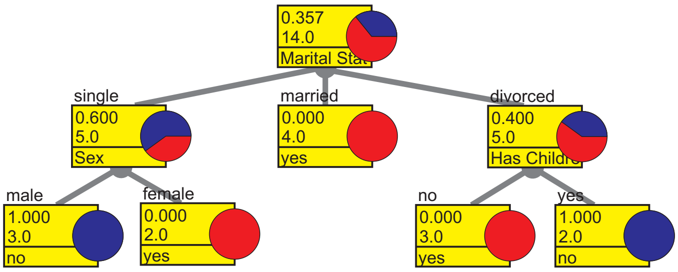


Figure 1: A decision tree, modeling the data set shown in Table 1.

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 $G_1, \ldots G_n$	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 $G_1, \ldots G_n$	transactions in data sets D_1 and D_2	examples of C and $\overline{C}$	examples of $C_1 \ldots C_n$
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

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

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

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
$\mathit{SuppDiff}(X,G_i,G_j)$		$\mathit{WRAcc}(X,C)$	Piatetski-Shapiro heuristic leverage
	$\mathit{GrowthRate}(X,D_1,D_2)$	$q_g(X,C)$	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.

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

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

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.

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

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.

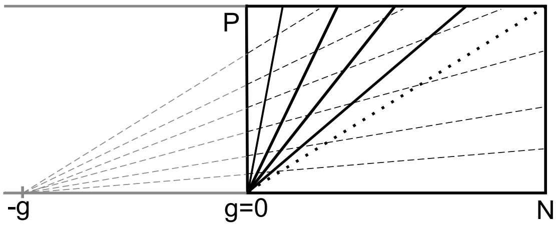


Figure 6: Isometrics for q_g. The dotted lines show the isometrics for a selected $g > 0$, while the full lines show the special case when $g = 0$, compatible to the EPM *growth rate* heuristic.

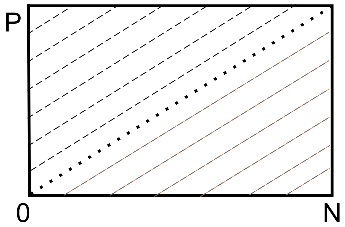


Figure 7: Isometrics for *WRAcc*, compatible to the CSM *support difference* heuristic.

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

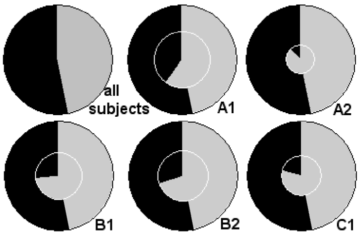


Figure 8: Subgroup visualization by pie charts.

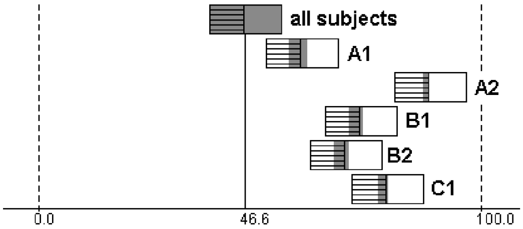


Figure 9: Subgroup visualization by box plots.

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

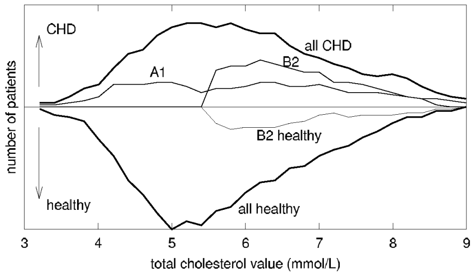


Figure 10: Subgroup visualization w.r.t. a continuous attribute. For clarity of the picture, only the positive (Y+) side of subgroup A1 is depicted.

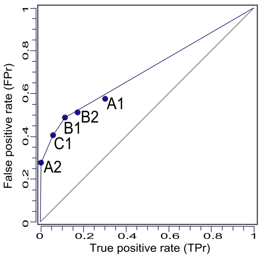


Figure 11: Representation of subgroups in the ROC space.

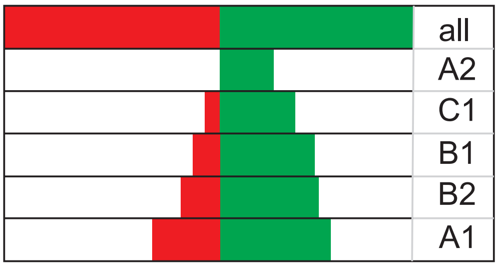


Figure 12: Subgroup visualization by bar charts.

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

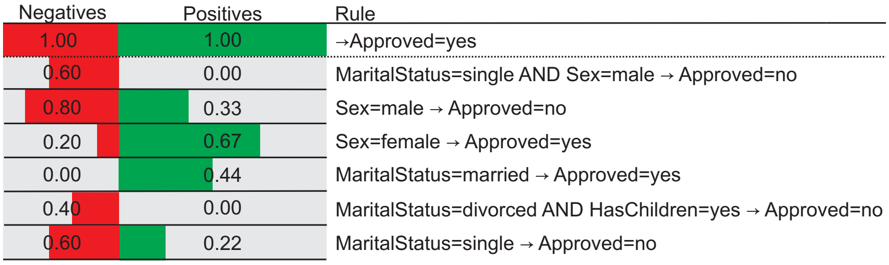
In this section, we (subjectively) compare the five different subgroup visualization methods by considering their intuitiveness, correctness of displayed data, usefulness, ability to ability to display contents besides the numerical properties of subgroups, (e.g., plot subgroup probability densities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

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

Table 6: Our evaluation of subgroup visualization methods.

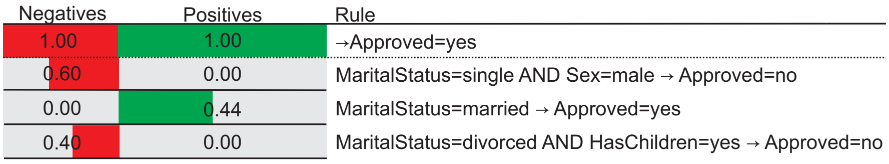
Two visualizations score best in Table 6 of our evaluation of subgroup visualization methods: the visualization of subgroups w.r.t. a continuous attribute and the bar chart visualization. The visualization of subgroups w.r.t. a continuous attribute is the only visualization that directly shows the contents 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.



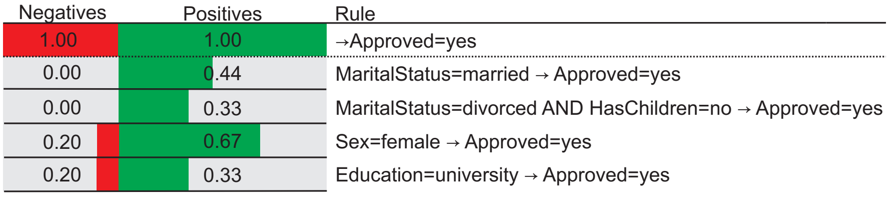
Negatives	Positives	Rule
1.00	1.00	→Approved=yes
0.60	0.00	MaritalStatus=single AND Sex=male → Approved=no
0.80	0.33	Sex=male → Approved=no
0.20	0.67	Sex=female → Approved=yes
0.00	0.44	MaritalStatus=married → Approved=yes
0.40	0.00	MaritalStatus=divorced AND HasChildren=yes → Approved=no
0.60	0.22	MaritalStatus=single → Approved=no

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.

3 CSM-SD: Methodology for Contrast Set Mining through Subgroup Discovery

In this chapter, the paper (Kralj Novak *et al.*, 2009a) titled "CSM-SD: Methodology for Contrast Set Mining through Subgroup Discovery" by Petra Kralj Novak, Nada Lavrač, Dragan Gamberger and Antonija Krstačić is presented. The paper was published on-line on the Journal of Biomedical Informatics web site in August 2008 and was published in Journal of Biomedical Informatics (Elsevier) in February 2009.

Compared to the theory-focused paper presented in the previous chapter, this paper is application driven, since all the discoveries and conclusions were achieved while applying supervised descriptive rule induction approaches to a real-life data analysis problem of distinguishing between groups of patients with similar diseases. The interaction and discussion with the medical practitioner Antonija Krstačić, who collected the data and also co-authored the paper, was crucial in several steps of our research. Petra Kralj Novak and Nada Lavrač in collaboration with Dragan Gamberger developed the theory, adapted the algorithms, ran experiments and wrote the majority of the paper.

Parts of the research presented in this paper were published at scientific conferences. First, the paper Kralj *et al.* (2007b), which compares contrast set mining and subgroup discovery, was presented at the 11th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2007). The data mining task was later refined and the paper Kralj *et al.* (2007a) discussing differences and advantages of pairwise and one-versus-all contrast set mining was presented at the 11th Conference on Artificial Intelligence in Medicine (AIME 2007). Finally, the paper Lavrač *et al.* (2007), which generalizes supporting factors from subgroup discovery to contrast set mining, was presented at the 11th Mediterranean Conference on Medical and Biological Engineering and Computing (Medicon 2007). All the listed papers were co-authored by the same authors. Only the journal paper Kralj Novak *et al.* (2009a) is enclosed in this dissertation.

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:

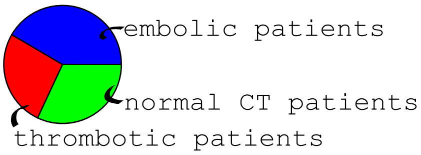


Fig. 1. Distribution of diagnosis of patients in the brain ischemia dataset.

- 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.

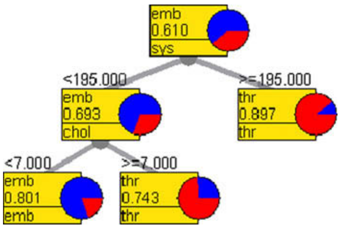


Fig. 2. A strongly pruned decision tree aimed at distinguishing between patients with embolic stroke and thrombotic stroke. Every node of the decision tree is represented by a circle–rectangle pair. In the circle, the distribution of the classes of the examples belonging to the node is visualized. The 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.

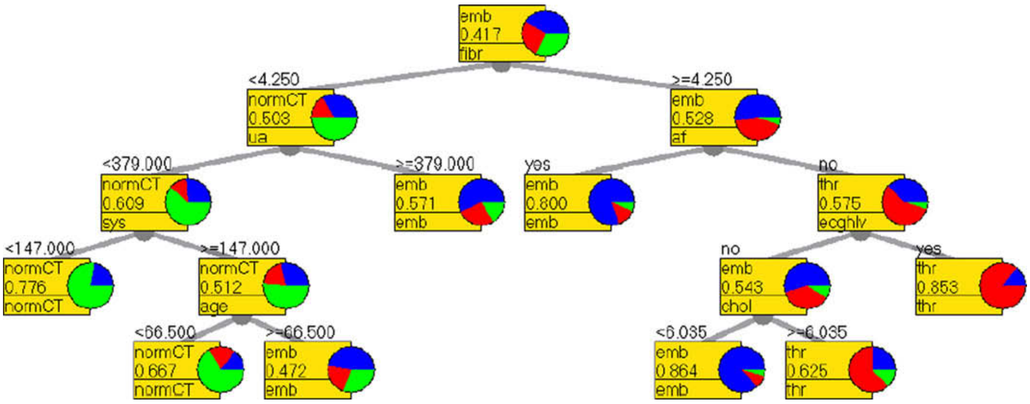


Fig. 3. A pruned decision tree aimed at distinguishing between patients with embolic stroke, thrombotic stroke and patients with normal brain CT test results.

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

$sys \geqslant 200 \text{ AND } chol \geqslant 5.1 \rightarrow class = thr$
$chol \geqslant 7.1 \text{ AND } plat \geqslant 198 \rightarrow class = thr$
$fibr \geqslant 5 \text{ AND } af = no \text{ AND } ecghlv = yes \rightarrow class = thr)$
$fibr \leqslant 3.8 \text{ AND } au \leqslant 305 \rightarrow class = normal$
$fibr \leqslant 4.2 \text{ AND } chol \geqslant 6.3 \rightarrow class = normal$
$age \leqslant 66 \text{ AND } ecgfr \geqslant 75 \text{ AND}pt \geqslant 0.8 \text{ AND } gluc \leqslant 6.8 \rightarrow class = normal$
$\rightarrow class = emb$

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.

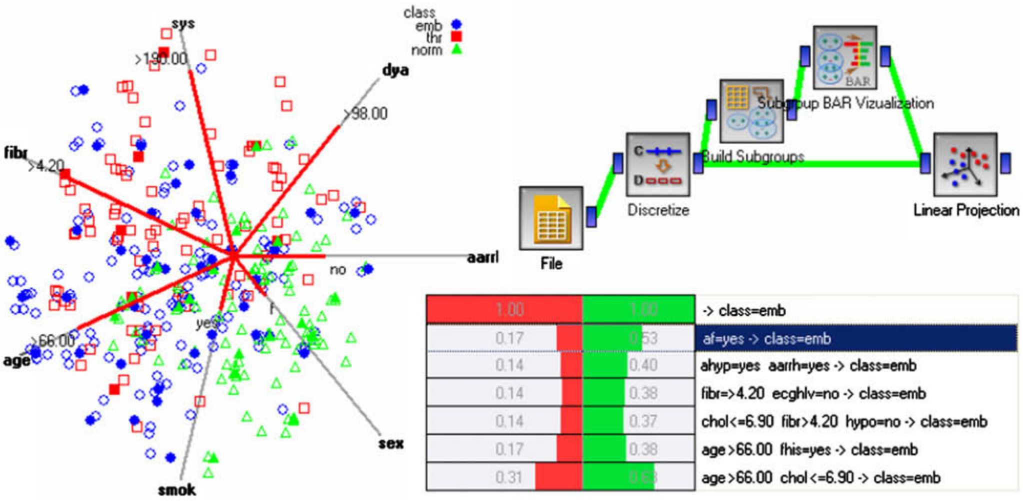
To show that terms used in contrast set mining (CSM) can be translated to terms used in subgroup discovery (SD), Table 2 provides a term dictionary through which we translate the terms used in CSM and SD into a unifying terminology of rule learning, or more specifically, concept learning. In concept 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.



1.00	1.00	-> class=emb
0.17	0.53	af=yes -> class=emb
0.14	0.40	ahyp=yes aarrh=yes -> class=emb
0.14	0.38	fibr=>4.20 ecghlv=no -> class=emb
0.14	0.37	chol<=6.90 fibr>4.20 hypo=no -> class=emb
0.17	0.38	age>66.00 fhis=yes -> class=emb
0.31	0.63	age>66.00 chol<=6.90 -> class=emb

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 $G_1, \ldots, G_n$	Examples of C and $\overline{C}$	Examples of $C_1, \ldots, C_n$
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: $SD(C = G_i \text{ vs. } \overline{C} = G_j)$ and $SD(C = G_j \text{ vs. } \overline{C} = G_i)$. 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
            $SD(C = G_i \text{ vs. } \overline{C} = G_j)$
```

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

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

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:

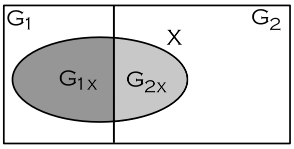
$$\begin{aligned} SuppDiff(X, G_1, G_2) &= support(X, G_1) - support(X, G_2) \\ &= TPr(X \rightarrow G_1) - TPr(X \rightarrow G_2) \\ &= TPr(X \rightarrow G_1) - FPr(X \rightarrow G_1) \end{aligned}$$

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



$$TPr(X \to G_1) = \frac{|G_{1x}|}{|G_1|} = support(X, G_1)$$

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

$$TPr(X \to G_2) = \frac{|G_{2x}|}{|G_2|} = support(X, G2)$$

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 $FPr(X \to G_1) = TPr(X \to G_2)$.

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:

$$WRAcc(X \to Y) = p(X) \cdot (p(Y|X) - p(Y)) \tag{3}$$

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]:

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

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 $[TPr(X \to Y) - FPr(X \to Y)]$, which actually is the support difference in a two group contrast set mining problem:

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

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:

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

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:

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

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

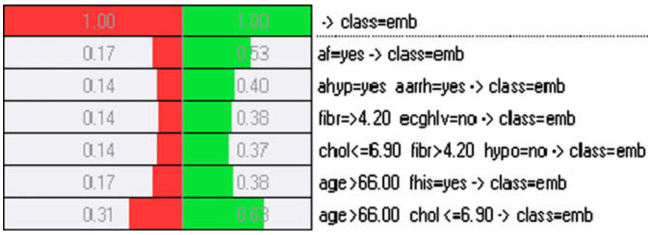
Subgroup discovery algorithms SD [9], CN2-SD [17] and APRIORI-SD [14] use a feature-based data representation, where attribute values needed for the construction of features are generated automatically from the data. In this way, subgroup discovery algorithms overcome this deficiency of contrast set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



1.00	1.00	-> class=emb
0.17	0.53	af=yes -> class=emb
0.14	0.40	ahyp=yes aarrh=yes -> class=emb
0.14	0.38	fibr=>4.20 ecghlv=no -> class=emb
0.14	0.37	chol<=6.90 fibr>4.20 hypo=no -> class=emb
0.17	0.38	age>66.00 fhis=yes -> class=emb
0.31	0.63	age>66.00 chol<=6.90 -> class=emb

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.

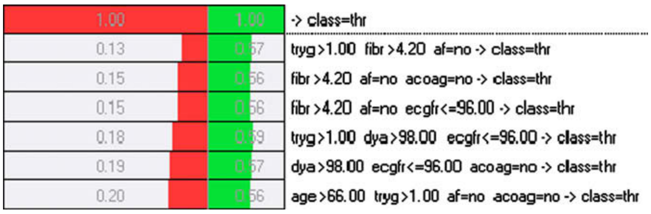
5.1. Experimental evaluation of the round robin CSM

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]



1.00	1.00	-> class=thr
0.13	0.57	tryg>1.00 fibr>4.20 af=no -> class=thr
0.15	0.56	fibr>4.20 af=no acoag=no -> class=thr
0.15	0.56	fibr>4.20 af=no ecgfr<=96.00 -> class=thr
0.18	0.59	tryg>1.00 dya>98.00 ecgfr<=96.00 -> class=thr
0.19	0.57	dya>98.00 ecgfr<=96.00 acoag=no -> class=thr
0.20	0.56	age>66.00 tryg>1.00 af=no acoag=no -> class=thr

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

[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 $sys > 185$.

[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.

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%

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.

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 $CS1$ is interpreted as a subgroup or as a contrast set. In Table 4 the top four supporting factors are characteristic for group $CS1$, regardless if it is considered as a subgroup or as a contrast set, while the next five supporting factors are characteristic for $CS1$ 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.

4 Closed Sets for Labeled Data

In this chapter, the paper (Garriga *et al.*, 2008) titled "Closed Sets for Labeled Data" by Gemma C. Garriga, Petra Kralj and Nada Lavrač is presented. The paper was published in the Journal of Machine Learning Research in April 2008.

The main part of the research presented in this paper was performed during a visit of the first author of the paper, Gemma C. Garriga, to the Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia. Gemma C. Garriga is the main author of the paper, since she has contributed most of the theory which connects closed sets with the *relevance theory* by Lavrač and Gamberger (2005). Petra Kralj contributed the experimental section, which includes the implementation of the algorithm named *RelSets*, and the evaluation in the ROC space. Nada Lavrač identified the potential of merging closed sets and subgroup discovery, and supervised the whole work by providing support and valuable advice.

Part of the research presented in this paper (Garriga *et al.*, 2006) was first published at the 10^{th} European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2006). The paper was later extended and improved to be published as a journal paper.

Closed sets for labeled data were used on a real life problem of analyzing microarray data from a potato experiment. Part of the results are published in the presented paper (Section 6.3: Subgroup Discovery in Microarray Data Analysis), while an extended version was published in Kralj *et al.* (2006). The biological aspect is described in Baebler *et al.* (2009).

The research presented in this paper was performed before the supervised descriptive rule induction framework was developed, therefore it does not use the unifying terminology and definitions. However, Section 5 of the paper already evaluates closed sets for labeled data as rules/ruleset in the ROC space and Section 6 compares closed sets for labeled data with subgroup discovery and emerging patterns — as discussed in Chapter 2. The paper fits well in the context of supervised descriptive rule induction, since closed sets for labeled data are rules that are used for descriptive data mining.

Journal of Machine Learning Research 9 (2008) 559-580 Submitted 12/06; Revised 11/07; Published 4/08

Closed Sets for Labeled Data

Gemma C. Garriga GEMMA.GARRIGA@HUT.FI
Helsinki Institute for Information Technology
Helsinki University of Technology
02015 Helsinki, Finland

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

Editor: Stefan Wrobel

Abstract

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. This paper shows 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. We formally prove that these sets characterize the space of relevant combinations of features for discriminating the target class. In practice, identifying relevant/irrelevant combinations of features through closed sets is useful in many applications: to compact emerging patterns of typical descriptive mining applications, to reduce the number of essential rules in classification, and to efficiently learn subgroup descriptions, as demonstrated in real-life subgroup discovery experiments on a high dimensional microarray data set.

Keywords: rule relevancy, closed sets, ROC space, emerging patterns, essential rules, subgroup discovery

1. Introduction

Rule discovery in data mining mainly explores unlabeled data and the focus resides on finding itemsets that satisfy a minimum support constraint (namely frequent itemsets), and from them, constructing rules over a certain confidence. This is the case of the well-known Apriori algorithm of Agrawal et al. (1996), and its successors, for example, Brin et al. (1997), Han and Pei (2000) and Zaki (2000b) among others. From a different perspective, machine learning is mainly concerned with the analysis of class labeled data, mainly resulting in the induction of classification and prediction rules, and—more recently—also descriptive rules that aim at discovering insightful knowledge from the data (subgroup discovery, contrast set mining). Traditional rule learning algorithms for classification include CN2 (Clark and Niblett, 1989) and Ripper (Cohen, 1995). Other approaches have been proposed that are based on the association rule technology but applied to class labeled data, for example, a pioneer work towards this integration is Liu et al. (1998), and later followed by others, for example, the Apriori-C classifier by Jovanoski and Lavrač (2001), and the Essence algorithm for inducing "essential" classification rules based on the covering properties of frequent itemsets, by Baralis and Chiusano (2004).

Subgroup discovery is a learning task directed at finding subgroup descriptions that are characteristic for examples with a certain property (class) of interest. Special rule learning algorithms for subgroup discovery include Apriori-SD (Kavšek and Lavrač, 2006), CN2-SD (Lavrač et al., 2004) or SD (Gamberger and Lavrač, 2002). The goal of these descriptive mining algorithms is to find characteristic rules as combinations of features with high coverage. If there are several rules with the same coverage, most specific rules (with more features) are appropriate for description and explanation purposes. On the other hand, the closely related task of contrast set mining aims at capturing discriminating features that contrast instances between classes. Algorithms for contrast set mining are STUCCO (Bay and Pazzani, 2001), and also an innovative approach presented in the form of mining emerging patterns (Dong and Li, 1999). Basically, Emerging Patterns (EP) are sets of features in the data whose supports increase significantly from one class to another. Interestingly, also good classifiers can be constructed by using the discriminating power of the mined EPs, for example, see Li et al. (2000). A condensed representation of EPs, defined in terms of a support growth rate measure, has been studied in Soulet et al. (2004).

Indeed, we can see all these tasks on labeled data (learning classification rules, subgroup discovery, or contrast set mining) as a rule induction problem, that is, a process of searching a space of concept descriptions (hypotheses in the form of rule antecedents). Some descriptions in this hypothesis space may turn out to be more relevant than others for characterizing and/or discriminating the target class. The question of relevance has attracted much attention in the context of feature selection for propositional learning (Koller and Sahami, 1996; Liu and Motoda, 1998). This is an important problem since non-relevant features can be excluded from the learning process, thus facilitating the search for the final solution and increasing the quality of the final rules. Feature filtering can be applied during the learning process, or also, by pre-processing the set of training examples (Lavrač et al., 1999; Lavrač and Gamberger, 2005).

Searching for relevant descriptions for rule construction has been extensively addressed in descriptive data mining as well. A useful insight was provided by closure systems (Carpineto and Romano, 2004; Ganter and Wille, 1998), aimed at compacting the whole space of descriptions into a reduced system of relevant sets that formally conveys the same information as the complete space. The approach has successfully evolved towards mining closed itemsets (see, for example, Pasquier et al., 2001; Zaki, 2004). Intuitively, closed itemsets can be seen as maximal sets of items/features covering a maximal set of examples. Despite its success in the data mining community, the use of closed sets is mainly descriptive. For example, they can be used to limit the number of association rules produced without information loss (see, for example, how to characterize rules with respect to their antecedent in Crémilleux and Boulicaut, 2002).

To the best of our knowledge, the notion of closed sets has not yet been exported to labeled data, nor used in the learning tasks for labeled data described above. In this paper we show that raw closed sets can be adapted for discriminative purposes by conveniently contrasting covering properties on positive and negative examples. Moreover, by exploiting the structural properties and the feature relevancy theory of Lavrač et al. (1999) and Lavrač and Gamberger (2005), we formally justify that the obtained closed sets characterize the space of relevant combinations of features for discriminating the target class.

In practice, our notion of closed sets in the labeled context (described in Sections 3 and 4) can be naturally interpreted as non-redundant descriptive rules (discriminating the target class) in the ROC space (Section 5). We also show that finding closed sets in labeled data turns out to be very useful in many applications. We have applied our proposal to reduce the number of emerging patterns (Section 6.1), to compress the number of essential rules (Section 6.2), and finally, to learn descriptions for subgroup discovery on potato microarray data (Section 6.3).[1]

2. Background

Features, used for describing the training examples, are logical variables representing attribute-value pairs (called items in the association rule learning framework of Agrawal et al., 1996). If $F = \{f_1, \ldots, f_n\}$ is a fixed set of features, we can represent a training example as a tuple of features $f \in F$ with an associated class label. For instance, Table 1 contains examples for the simplified problem of contact lens prescriptions (Witten and Frank, 2005). Patients are described by four attributes: Age, Spectacle prescription, Astigmatism and Tear production rate; and each tuple is labeled with a class label: none, soft or hard. Then, F is the set of all attribute-value pairs in the data, that is, $F = \{$Age=young, …, Tear=normal$\}$ (the class label is not included in F), and each example (a patient) corresponds to a subset of features in F with an associated class label. This small data set will be used throughout the paper to ease the understanding of our proposals.

We consider two-class learning problems where the set of examples E is divided into positives (P, target-class examples identified by label $+$) and negatives (N, labeled by $-$), and $E = P \cup N$. Multi-class problems can be translated to a series of two-class learning problems: each class is once selected as the target class (positive examples), while examples of all the other classes are treated as non-target class examples (thus, negative examples). For instance, when class soft of Table 1 is the target class, all examples with label soft are considered as positive, as shown in Table 2, and all examples labeled none and hard are considered as negative.

Given a rule $X \rightarrow +$ formed from a set of features $X \subseteq F$, *true positives* (TP) are those positive examples covered by the rule, that is, $p \in P$ such that $X \subseteq p$; and *false positives* (FP) are those negative examples covered by the rule, that is, $n \in N$ such that $X \subseteq n$; reciprocally, *true negatives* (TN) are those negative examples not covered by X. Later, we will see that some combinations of features $X \subseteq F$ produce more relevant antecedents than others for the rules $X \rightarrow +$. Our study will focus specifically on the combinations of features from the universe F which best define the space of non-redundant rules for the target class. We will do it by integrating the notion of closed itemsets and the concept of feature relevancy proposed in previous works.

2.1 Closed Itemsets

From the practical point of view of data mining algorithms, closed itemsets are the largest sets (w.r.t. set-theoretic inclusion) among those other itemsets occurring in the same examples (Bastide et al., 2000a; Crémilleux and Boulicaut, 2002; Pasquier et al., 2001; Taouil et al., 2000; Zaki, 2000a, 2004; Zaki and Ogihara, 1998). Formally, let *support of itemset* $X \subseteq F$, denoted by $\text{supp}(X)$, be the number of examples in the data where X is contained. Then: a set $X \subseteq F$ is said to be *closed* when there is no other set $Y \subseteq F$ such that $X \subset Y$ and $\text{supp}(X) = \text{supp}(Y)$.

In the example of Table 2, the itemset corresponding to $\{$Age=young$\}$ is not closed because it can be extended to the maximal set $\{$Age=young, Astigmatism=no, Tear=normal$\}$ that has the same support in this data. Notice that by treating positive examples separately, the positive label will be already implicit in the closed itemsets mined on the target class data. So, here we will work by

[1] A preliminary version of this work appeared in Garriga et al. (2006). This paper is improved based on the valuable reviewers' comments, incorporates proofs, detailed explanations, extended comparisons with related work and more experiments.

Id	Age	Spectacle prescription	Astig.	Tear prod.	Lens
1	young	myope	no	normal	soft
2	young	hypermetrope	no	normal	soft
3	pre-presbyopic	myope	no	normal	soft
4	pre-presbyopic	hypermetrope	no	normal	soft
5	presbyopic	hypermetrope	no	normal	soft
6	young	myope	no	reduced	none
7	young	myope	yes	reduced	none
8	young	hypermetrope	no	reduced	none
9	young	hypermetrope	yes	reduced	none
10	pre-presbyopic	myope	no	reduced	none
11	pre-presbyopic	myope	yes	reduced	none
12	pre-presbyopic	hypermetrope	no	reduced	none
13	pre-presbyopic	hypermetrope	yes	reduced	none
14	pre-presbyopic	hypermetrope	yes	normal	none
15	presbyopic	myope	no	reduced	none
16	presbyopic	myope	no	normal	none
17	presbyopic	myope	yes	reduced	none
18	presbyopic	hypermetrope	no	reduced	none
19	presbyopic	hypermetrope	yes	reduced	none
20	presbyopic	hypermetrope	yes	normal	none
21	young	myope	yes	normal	hard
22	young	hypermetrope	yes	normal	hard
23	pre-presbyopic	myope	yes	normal	hard
24	presbyopic	myope	yes	normal	hard

Table 1: The contact lens data set, proposed by Witten and Frank (2005).

Id	Age	Spectacle prescription	Astig.	Tear prod.	Class
1	young	myope	no	normal	$+$
2	young	hypermetrope	no	normal	$+$
3	pre-presbyopic	myope	no	normal	$+$
4	pre-presbyopic	hypermetrope	no	normal	$+$
5	presbyopic	hypermetrope	no	normal	$+$

Table 2: The set of positive examples when class soft of the contact lens data of Table 1 is selected as the target class. These examples form the set P of positive examples, while instances of classes none and hard are considered non-target, thus treated together as negative examples N. Note that examples are represented here in a simplified tabular form instead of the feature set representation.

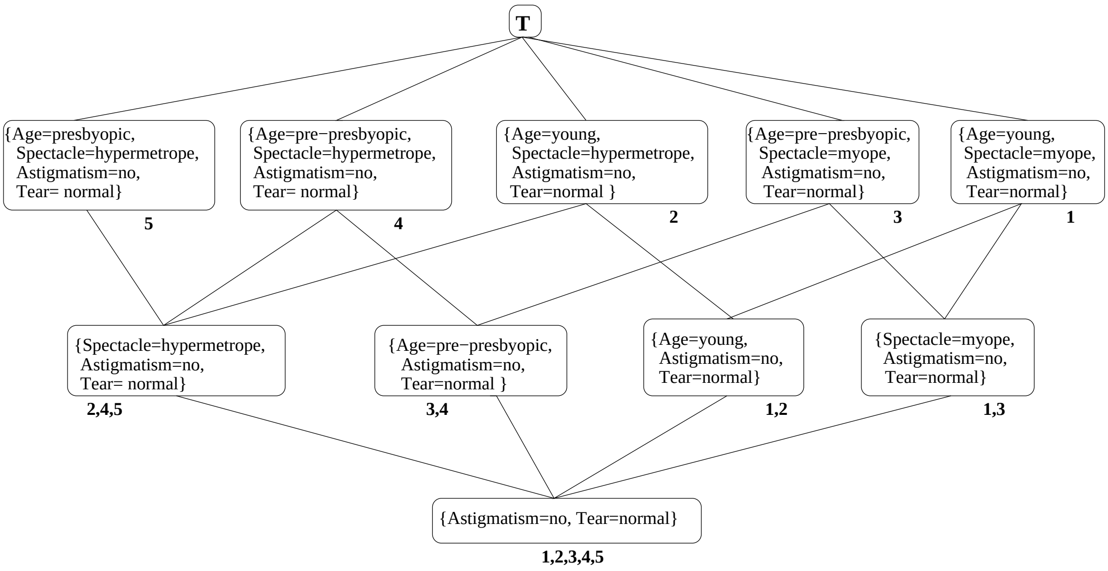

T
{Age=presbyopic, Spectacle=hypermetrope, Astigmatism=no, Tear= normal}
5
{Age=pre−presbyopic, Spectacle=hypermetrope, Astigmatism=no, Tear= normal}
4
{Age=young, Spectacle=hypermetrope, Astigmatism=no, Tear=normal }
2
{Age=pre−presbyopic, Spectacle=myope, Astigmatism=no, Tear=normal}
3
{Age=young, Spectacle=myope, Astigmatism=no, Tear=normal}
1
{Spectacle=hypermetrope, Astigmatism=no, Tear= normal}
2,4,5
{Age=pre−presbyopic, Astigmatism=no, Tear=normal }
3,4
{Age=young, Astigmatism=no, Tear=normal}
1,2
{Spectacle=myope, Astigmatism=no, Tear=normal}
1,3
{Astigmatism=no, Tear=normal}
1,2,3,4,5

Figure 1: The lattice of closed itemsets for data in Table 2.

constructing the closure system of items on our positive examples and use this system to study the structural properties of the closed sets to discriminate the implicit label. Many efficient algorithms have been proposed for discovering closed itemsets over a certain minimum support threshold; see a compendium of them in Goethals and Zaki (2004).

The foundations of closed itemsets are based on the definition of a closure operator on a lattice of items (Carpineto and Romano, 2004; Ganter and Wille, 1998). The standard closure operator Γ for items acts as follows: the closure $\Gamma(X)$ of a set of items $X \subseteq F$ includes all items that are present in all examples having all items in X. According to the classical theory, operator Γ satisfies the following properties: Monotonicity: $X \subseteq X' \Rightarrow \Gamma(X) \subseteq \Gamma(X')$; Extensivity: $X \subseteq \Gamma(X)$; and Idempotency: $\Gamma(\Gamma(X)) = \Gamma(X)$.

From the formal point of view of Γ, closed sets are those coinciding with their closure, that is, for $X \subseteq F$, X is *closed* iff $\Gamma(X) = X$. Also, when $\Gamma(Y) = X$ for a set $Y \neq X$, it is said that Y is a *generator* of X. By extensivity of Γ we always have $Y \subseteq X$ for Y generator of X. Intensive work has focused on identifying which collection of generators is good to ensure that all closed sets can be produced. The named δ-free sets in Boulicaut et al. (2003) are minimal generators when $\delta = 0$, and these are equivalent to key patterns in Bastide et al. (2000b). Different properties of these δ-free sets generators in Boulicaut et al. (2003) have been studied for different values of δ.

Considering Table 2, we have the following $\Gamma(\{\text{Age=young}\}) = \{\text{Age=young, Astigmatism=no, Tear=normal}\}$. Then, $\{\text{Age=young}\}$ is a generator of this closed set. Note that for $\Gamma(Y) = X$, both Y and X are sets with exactly the same support in the data, but X being a largest set of items, that is, $Y \subset X$ for all Y such that $\Gamma(Y) = X$. This property is ensured by the extensivity of this operator. Moreover, closed sets formalized with operator Γ are exactly those sets obtained in closed set mining process and defined above, which present many advantages (see, for example, Balcázar and Baixeries, 2003; Crémilleux and Boulicaut, 2002).

Closed itemsets are lossless in the sense that they uniquely determine the set of all frequent itemsets and their exact support (cf. Pfaltz, 1996; Zaki and Ogihara, 1998, for more theoretical details). Closed sets of items can be graphically organized in a Hasse diagram, where each node corresponds to a closed itemset, and there is an edge between two nodes if and only if they are comparable (w.r.t. set-theoretic inclusion) and there is no other intermediate closed itemset in the lattice. In this partial order organization, ascending/descending paths represent the subset/superset relation. Typically, the top of this lattice is represented by a constant T corresponding to a set of items not included in any example.

Figure 1 shows the lattice of closed itemsets obtained from data from Table 2. Each node is depicted along with the set of example identifiers where the closed set occurs. Notice that all closed itemsets with the same support cover a different subset of transactions of the original data. In practice, such exponential lattices are not completely constructed, as only a list of closed itemsets over a certain minimum support suffices for practical purposes. Therefore, instead of closed sets one needs to talk about *frequent closed sets*, that is, those closed sets over the minimum support constraint given by the user. Also notice the difference of frequent closed sets from the popular concept of maximal frequent sets (see, for example, Tan et al., 2005), which refers to those sets for which none of their supersets are frequent.

Obviously, imposing a minimum support constraint will eliminate the largest closed sets whose support is typically very low. The impact of such constraint depends on the application. In general, there exists a trade-off between quality and speed up of the process. In the following we consider a theoretical framework with all closed sets; in practice though, we will need a minimum support constraint to consider only the frequent ones.

2.2 Relevant Features for Discrimination

The main aim of the theory of relevancy, described in Lavrač et al. (1999) and Lavrač and Gamberger (2005), is to reduce the hypothesis space by eliminating irrelevant features from F in the pre-processing phase. Other related work, such as Koller and Sahami (1996) and Liu and Motoda (1998), eliminate features in the model construction phase. However, here we concentrate on the elimination of irrelevant features in the preprocessing phase, as proposed by Lavrač and Gamberger (2005):

Definition 1 (Coverage of features) *Feature $f \in F$ covers another feature $f' \in F$ if and only if true positives of f' are a subset of true positives of f, and true negatives of f' are a subset of true negatives of f. In other words,* $\mathrm{TP}(f') \subseteq \mathrm{TP}(f)$ *and* $\mathrm{TN}(f') \subseteq \mathrm{TN}(f)$ *(or equivalently,* $\mathrm{TP}(f') \subseteq \mathrm{TP}(f)$ *and* $\mathrm{FP}(f) \subseteq \mathrm{FP}(f')$*).*

Using the definition of feature coverage, we further define that $f' \in F$ is *relatively irrelevant* if there exists another feature $f \in F$ such that f covers f'. To illustrate this notion we take the data of Table 1: if examples of class none form our positives and the rest of examples are considered negative, then the feature Tear=reduced covers Age=young, hence making this last feature irrelevant for the discrimination of the class none.

Other notions of irrelevancy described in Lavrač and Gamberger (2005) consider a minimum coverage constraint in the true positives or accordingly, on the true negatives.

3. Closed Sets on Target-class Data

Given a set of examples $E = P \cup N$ it is trivial to realize that for any rule $X \rightarrow +$ with a set of features $X \subseteq F$, the support of itemset X in P (target class examples) exactly corresponds to the number of true positives (TP) of the rule; reciprocally, the support of X in N (non-target class examples) is the number of false positives (FP) of the rule. Also, because of the anti-monotonicity property of support (i.e., $Y \subseteq X$ implies $\text{supp}(X) \leq \text{supp}(Y)$) the following useful property can be easily stated.

Proposition 2 *Let $X, Y \subseteq F$ such that $Y \subseteq X$, then* $\text{TP}(X) \subseteq \text{TP}(Y)$ *and* $\text{FP}(X) \subseteq \text{FP}(Y)$.

Proof The anti-monotonicity property of support on the set of positive examples ensures that $|\text{TP}(X)| \leq |\text{TP}(Y)|$. Since $Y \subseteq X$, we necessarily have $\text{TP}(X) \subseteq \text{TP}(Y)$. The same reasoning applies to the set of negative examples. ■

For convenience, let $\text{supp}^+(X)$ denote the support of the set X in the positive set of examples P, and $\text{supp}^-(X)$ the support in the negative set of examples N. Notice that for a rule $X \rightarrow +$ we indeed have that $\text{supp}^+(X) = |\text{TP}(X)|$ and $\text{supp}^-(X) = |\text{FP}(X)|$. In the following we will use one notation or the other according to the convenience of the context.

Following from the last proposition, the next property can be readily seen.

Lemma 3 *Feature $f \in F$ covers another feature $f' \in F$ (as in Definition 1), iff* $\text{supp}^+(\{f'\}) = \text{supp}^+(\{f, f'\})$ *and* $\text{supp}^-(\{f\}) = \text{supp}^-(\{f, f'\})$.

Proof That f covers f' can be formulated as $\text{TP}(f') \subseteq \text{TP}(f)$ and $\text{FP}(f) \subseteq \text{FP}(f')$. Because all the true positives of f' are also covered by f, it is true that $\text{TP}(f') = \text{TP}(f, f')$; similarly, because all the false positives of f are also covered by f' we have $\text{FP}(f) = \text{FP}(f, f')$. These two facts directly imply that $\text{supp}^+(\{f'\}) = \text{supp}^+(\{f, f'\})$ and $\text{supp}^-(\{f\}) = \text{supp}^-(\{f, f'\})$.

The other direction is proved as follows. The anti-monotonicity property of Proposition 2 applied over $\{f'\} \subseteq \{f, f'\}$ leads to $\text{TP}(f, f') \subseteq \text{TP}(f')$. Indeed, from $\text{supp}^+(\{f'\}) = \text{supp}^+(\{f, f'\})$ we have $|\text{TP}(f')| = |\text{TP}(f, f')|$, which along with $\text{TP}(f, f') \subseteq \text{TP}(f')$ implies an equivalence of true positives between these two sets: that is, $\text{TP}(f, f') = \text{TP}(f')$. From here we deduce $\text{TP}(f') \subseteq \text{TP}(f)$. Exactly the same reasoning applies to the negatives. Proposition 2 ensures that $\text{FP}(f, f') \subseteq \text{FP}(f)$ because $\{f\} \subseteq \{f, f'\}$. But from $\text{supp}^-(\{f\}) = \text{supp}^-(\{f, f'\})$ we have $|\text{FP}(f)| = |\text{FP}(f, f')|$, which together with $\text{FP}(f, f') \subseteq \text{FP}(f)$ leads to the equivalence of the false positives between these two sets: that is, $\text{FP}(f) = \text{FP}(f, f')$. Then, we deduce $\text{FP}(f) \subseteq \text{FP}(f')$. That is f covers f' as in Definition 1. ■

Indeed, this last result allows us to rewrite, within the data mining language, the definition of relevancy proposed by Lavrač et al. (1999) and Lavrač and Gamberger (2005): a feature f is *more relevant* than f' when $\text{supp}^+(\{f'\}) = \text{supp}^+(\{f, f'\})$ and $\text{supp}^-(\{f\}) = \text{supp}^-(\{f, f'\})$. For instance, the support of {Age=young} over the class none of data from Table 1 is equal to the support of {Age=young, Tear=reduced} in this same class none ; at the same time, the support of {Tear=reduced} is zero in the negatives (formed here by the classes soft and hard together), thus equal to the support in the negatives of {Age=young, Tear=reduced}. So, the feature Age=young is irrelevant with respect to Tear=reduced, as we identified in Section 2.1. In other words, f' is irrelevant with respect to f if the occurrence of f' always implies the presence of f in the positives, and at the same time, f always implies the presence of f' in the negatives.

To the effect of our later arguments it will be useful to cast the result of Lemma 3 in terms of the formal closure operator Γ. This will provide the desired mapping from relevant sets of features to the lattice of closed itemsets constructed on target class examples. Again, because we need to formalize our arguments against positive and negative examples separately, we will use Γ^+ or Γ^- for the closure of itemsets on P or N respectively.

Lemma 4 *A feature f is more relevant than f' iff $\Gamma^+(\{f'\}) = \Gamma^+(\{f,f'\})$ and $\Gamma^-(\{f\}) = \Gamma^-(\{f,f'\})$.*

Proof It follows immediately from Lemma 3 and the formalization of operator Γ. A feature f is more relevant than f' when f covers f' according to Definition 1. Then, by Lemma 3 we have that $\text{supp}^+(\{f'\}) = \text{supp}^+(\{f,f'\})$ and $\text{supp}^-(\{f\}) = \text{supp}^-(\{f,f'\})$. By construction of Γ, this means that the sets $\{f'\}$ and $\{f,f'\}$ have the same closure on the positives, and the sets $\{f\}$ and $\{f,f'\}$ have the same closure on the negatives. That is: because Γ is an extensive operator, we can rewrite it as $\Gamma^+(\{f'\}) = \Gamma^+(\{f,f'\})$ and $\Gamma^-(\{f\}) = \Gamma^-(\{f,f'\})$. ■

Interestingly, operator Γ is formally defined for the universe of sets of items, so that these relevancy results on single features can be directly extended to sets of features. This provides a proper generalization, which we express in the following definition.

Definition 5 (Relevancy of feature sets) *Set of features $X \subseteq F$ is more relevant than set $Y \subseteq F$ iff $\Gamma^+(Y) = \Gamma^+(X \cup Y)$ and $\Gamma^-(X) = \Gamma^-(X \cup Y)$.*

To illustrate Definition 5 take the positive examples from Table 2, with negative data formed by classes none and hard together. Feature Spectacle=myope alone cannot be compared to feature Astigmatism=no alone with Definition 1 (because Astigmatism=no does not always imply Spectacle=myope in the negatives). For the same reason, Spectacle=myope cannot be compared to feature Tear=normal alone. However, when considering these two features together, then Spectacle=myope turns out to be irrelevant w.r.t. the set {Astigmatism=no, Tear=normal}. So, the new semantic notion of Definition 5 allows us to decide if a set of features is structurally more important than another for discriminating the target class. In the language of rules: rule $Y \to +$ is *irrelevant* if there exists another rule $X \to +$ satisfying two conditions: first, $\Gamma^+(Y) = \Gamma^+(X \cup Y)$; and second, $\Gamma^-(X) = \Gamma^-(X \cup Y)$. E.g., when soft is the target class: the rule Spectacle=myope $\to +$ is not relevant because at least the rule {Astigmatism=no, Tear=normal} $\to +$ will be more relevant.

Finally, from the structural properties of operator Γ and from Proposition 2, we can deduce that the semantics of relevant sets in Definition 5 is consistent.

Lemma 6 *A set of features $X \subseteq F$ is more relevant than set $Y \subseteq F$ (Definition 5) iff $\text{TP}(Y) \subseteq \text{TP}(X)$ and $\text{FP}(X) \subseteq \text{FP}(Y)$.*

Proof That X is more relevant than Y means $\Gamma^+(Y) = \Gamma^+(X \cup Y)$ and $\Gamma^-(X) = \Gamma^-(X \cup Y)$. Proposition 2 ensures that $\text{TP}(X \cup Y) \subseteq \text{TP}(Y)$ because $Y \subseteq X \cup Y$. Then, from $\Gamma^+(Y) = \Gamma^+(X \cup Y)$ we naturally have that $|\text{TP}(Y)| = |\text{TP}(X \cup Y)|$ (by formalization of Γ), which together with $\text{TP}(X \cup Y) \subseteq \text{TP}(Y)$ leads to the equality of the true positives between the following sets: $\text{TP}(X \cup Y) = \text{TP}(Y)$.

From here, $\mathrm{TP}(Y) \subseteq \mathrm{TP}(X)$. On the other hand, it is implied by the definition of relevancy that $Y \subseteq X$, thus directly from Proposition 2 we have that $\mathrm{FP}(X) \subseteq \mathrm{FP}(Y)$.

The other direction is proved as follows. Let X and Y be two sets such that $\mathrm{TP}(Y) \subseteq \mathrm{TP}(X)$ and $\mathrm{FP}(X) \subseteq \mathrm{FP}(Y)$. As all the true positives of Y are also covered by X, it is true that $\mathrm{TP}(Y) = \mathrm{TP}(X \cup Y)$; similarly, as all the false positives of X are also covered by Y we have that $\mathrm{FP}(X) = \mathrm{FP}(X \cup Y)$. This directly implies that $\mathrm{supp}^+(Y) = \mathrm{supp}^+(X \cup Y)$ and $\mathrm{supp}^-(X) = \mathrm{supp}^-(X \cup Y)$. By construction of Γ, this means we can directly rewrite this as $\Gamma^+(Y) = \Gamma^+(X \cup Y)$ and $\Gamma^-(X) = \Gamma^-(X \cup Y)$. That is: set X is more relevant than Y by Definition 5. ■

In the language of rules, Lemma 6 implies that when a set of features $X \subseteq F$ is more relevant than $Y \subseteq F$, then rule $Y \rightarrow +$ is less relevant than rule $X \rightarrow +$ for discriminating the target class. Moreover, Lemma 6 proves the consistency of Definition 5. If we consider $X = \{f\}$ and $Y = \{f'\}$, then the definition is simply reduced to the coverage of Definition 1. Yet, the interestingness of Definition 5 is that we can use this new concept to study the relevancy of itemsets (discovered in the mining process) for discrimination problems. Also, it can be immediately seen that if X is more relevant than Y in the positives, then Y will be more relevant than X in the negatives (by just reversing Definition 5).

Next subsection characterizes the role of closed itemsets to find relevant sets of features for discrimination. Notice that the first condition to consider a set X more relevant than Y in the discrimination of target class examples is that $\Gamma^+(Y) = \Gamma^+(X \cup Y)$. So, the closure system constructed on the positive examples will be proved to be structurally important for inducing target class rules.

3.1 Closed Sets for Discrimination

Together with the result of Lemma 6, it can be shown that only closed itemsets mined in the set of positive examples suffice for discrimination.

Theorem 7 *Let $Y \subseteq F$ be a set of features such that $\Gamma^+(Y) = X$ and $Y \neq X$. Then, set Y is less relevant than X (as in Definition 5).*[2]

Proof By the extensivity property of Γ we know $Y \subseteq X$. Then, Proposition 2 ensures that $\mathrm{TP}(X) \subseteq \mathrm{TP}(Y)$ and $\mathrm{FP}(X) \subseteq \mathrm{FP}(Y)$. However, by hypothesis we have $\Gamma^+(Y) = X$, which by construction ensures that $|\mathrm{TP}(Y)| = |\mathrm{TP}(X)|$; but because $Y \subseteq X$, it must be true that $\mathrm{TP}(Y) = \mathrm{TP}(X)$. In all, we obtained that $\mathrm{TP}(Y) = \mathrm{TP}(X)$ and $\mathrm{FP}(X) \subseteq \mathrm{FP}(Y)$, and from Lemma 6 we have that X is more relevant than Y. ■

Typically, in approaches such as Apriori-C (Jovanoski and Lavrač, 2001), Apriori-SD (Kavšek and Lavrač, 2006) or RLSD (Zhang et al., 2004), frequent itemsets with very small minimal support constraint are initially mined and subsequently post-processed in order to find the most suitable rules

[2] We are aware that some generators Y of a closed set X might be exactly equivalent to X in terms of TP and FP, thus forming equivalence classes of rules (i.e., $Y \rightarrow +$ might be equivalent to $X \rightarrow +$). The result of this theorem characterizes closed sets in the positives as those representatives of relevant rules; so, any set which is not closed can be discarded, and thus, efficient closed mining algorithms can be employed for discrimination purposes. The next section will approach the notion of the shortest representation of a relevant rule, which will be conveyed by these mentioned equivalent generators.

for discrimination. The new result presented here states that not all frequent itemsets are necessary: as shown in Theorem 7 only the closed sets have the potential to be relevant.

To illustrate this result we use again data in Table 2, where $\Gamma^{+}(\{\text{Astigmatism=no}\}) = \{\text{Astigmatism=no, Tear=normal}\}$. Thus, rule Astigmatism=no $\rightarrow +$ can be discarded: it covers exactly the same positives as {Astigmatism=no, Tear=normal}, but more negatives. Thus, a rule whose antecedent is {Astigmatism=no, Tear=normal} would be preferred for discriminating the class soft.

However, Theorem 7 simply states that those itemsets which are not closed in the set of positive examples cannot form a relevant rule to discriminate the target class, thus they do not correspond to a relevant combination of features. In other words, closed itemsets suffice but some of them might not be necessary to discriminate the target class. It might well be that a closed itemset is irrelevant with respect to another closed itemset in the system.

As illustrated above, when considering class soft as the target class (identified by $+$), we had that feature Spectacle=myope is irrelevant with respect to set {Astigmatism=no, Tear=normal}; yet, set {Spectacle=myope, Astigmatism=no, Tear=normal} is closed in the system (see the lattice of Figure 1). Indeed, this latter closed set is still irrelevant in the system according to our Definition 5 and can be pruned away. The next section is dedicated to the task of reducing the closure system of itemsets to characterize the final space of relevant sets of features.

4. Characterizing the Space of Relevant Sets of Features

This section studies how the dual closure system on the negative examples is used to reduce the lattice of closed sets on the positives. This reduction will characterize a complete space of relevant sets of features for discriminating the target class. First of all, we raise the following two important remarks following from Proposition 2.

Remark 8 *Given two different closed sets on the positives X and X' such that $X \not\subseteq X'$ and $X' \not\subseteq X$ (i.e., there is no ascending/descending path between them in the lattice), then they cannot be compared in terms of relevancy, since they cover different positive examples.*

We exemplify Remark 8 with the lattice in Figure 1. The two closed sets: {Age=young, Astigmatism=no, Tear=normal} and {Spectacle=myope, Astigmatism=no, Tear=normal}, are not comparable with subset relation: they cover different positive examples and they cannot be compared in terms of relevance.

Remark 9 *Given two closed sets on the positives X and X' with $X \subset X'$, we have by construction that $\mathrm{TP}(X') \subset \mathrm{TP}(X)$ and $\mathrm{FP}(X') \subseteq \mathrm{FP}(X)$ (from Proposition 2). Notice that because X and X' are different closed sets in the positives, $\mathrm{TP}(X')$ is necessarily a proper subset of $\mathrm{TP}(X)$; however, regarding the coverage of false positives, this inclusion is not necessarily proper.*

To illustrate Remark 9 we use the lattice of closed itemsets in Figure 1. By construction the closed set {Spectacle=myope, Astigmatism=no, Tear=normal} from Figure 1 covers fewer positives than the proper predecessor {Astigmatism=no, Tear=normal}. However, both closed sets cover exactly one negative example. In this case {Astigmatism=no, Tear= normal} is more relevant than {Spectacle=myope, Astigmatism=no, Tear=normal}.

Remark 9 points out that two different closed sets in the positives, yet being one included in the other, may end up covering exactly the same set of false positives. In this case, we would like

Transaction occurrence list	Closed Set
$1,2,3,4,5$	{Astigmatism=no, Tear=normal }
$2,4,5$	{Spectacle=hypermetrope, Astigmatism=no, Tear=normal }
$3,4$	{Age=pre-presbyopic, Astigmatism=no, Tear=normal }
$1,2$	{Age=young, Astigmatism=no, Tear=normal }

Table 3: The four closed sets corresponding to the space of relevant sets of features for data in Table 2.

to discard the closed set covering less true positives. Because of the anti-monotonicity property of support, the smaller one will be the most relevant.

From these two remarks we obtain the following result.

Theorem 10 *Let $X \subseteq F$ and $X' \subseteq F$ be two different closed sets in the positives such that $X \subset X'$. Then, we have that X' is less relevant than X (as in Definition 5) iff $\Gamma^-(X) = \Gamma^-(X')$.*

Proof That X' is less relevant than X is defined as: $\Gamma^+(X') = \Gamma^+(X' \cup X)$ and $\Gamma^-(X) = \Gamma^-(X' \cup X)$. Since $X \subset X'$ by hypothesis, we always have that $X' = X' \cup X$, so that the above two conditions can be rewritten as $\Gamma^+(X') = \Gamma^+(X')$ (always true) and $\Gamma^-(X) = \Gamma^-(X')$, as we wanted to prove.

In the backward direction we start from $\Gamma^-(X) = \Gamma^-(X')$, where $X \subset X'$ as stated by hypothesis of the theorem. Because $X \subset X'$ it is true that $X' = X' \cup X$. Then, we can rewrite $\Gamma^-(X) = \Gamma^-(X')$ as $\Gamma^-(X) = \Gamma^-(X' \cup X)$, thus satisfying already the first condition of Definition 5. Also, $\Gamma^+(X')$ is simply the same as $\Gamma^+(X') = \Gamma^+(X' \cup X)$, thus satisfying the second condition of Definition 5. ■

Thus, by Theorem 10 we can reduce the closure system constructed on the positives by discarding irrelevant nodes: if two closed itemsets are connected by an ascending/descending path on the lattice of positives (i.e., they are comparable by set inclusion $\subset$), yet they have the same closure on the negatives (i.e., they cover the same false positives, or equivalently, their support on the negatives is exactly the same), then just the shortest set is relevant.

Finally, after Theorem 7 and Theorem 10, we can characterize the space of relevant sets of features for discriminating the selected target class as follows.

Definition 11 (Space of relevant sets of features) *The space of relevant combinations of features for discriminating the target class is defined as those sets X for which it holds that: $\Gamma^+(X) = X$ and there is no other closed set $\Gamma^+(X') = X'$ such that $\Gamma^-(X') = \Gamma^-(X)$.*

It is trivial to see after Remarks 8 and 9, that by construction, any two sets in this space always cover a different set of positives and a different set of negatives. These final sets can be directly interpreted as antecedents of rules for classifying the target class (i.e., for each relevant $X \subseteq F$ in the space, we have a relevant rule $X \rightarrow +$ for classifying the positives).

The four closed sets forming the space of relevant sets of features for the class soft are shown in Table 3. It can be checked that the CN2 algorithm (Clark and Niblett, 1989) would output a single rule whose antecedent corresponds to the closed set in the first row of Table 3. On the other hand, Ripper (Cohen, 1995) would obtain the most specific relevant rules, that is, those corresponding to the three last rows from Table 3. Finally, other algorithms such as Apriori-C would also output rules whose antecedents are not relevant as such, for example, Astigmatism=no → Lenses= soft.

To complete the example of the contact lenses database: the lattice of closed itemsets on the class hard contains a total of 7 nodes, which is reduced to only 3 relevant sets; on the other hand, the lattice of closed itemsets on the class none contains a total of 61 nodes, which is reduced to 19 relevant sets.

The space of relevant combinations defines exhaustively all the relevant antecedents for discriminating the target class. Not to generate this space completely, in large sets of data a minimum support threshold will be usually imposed (see more details in the experimental section). As expected, too large relevant sets will be naturally pruned by the minimum support constraint, which might have an undesired effect depending on the application. Still, it is known that very long closed sets, that is, too specific sets of features in our contribution, tend to overestimate when constructing a classifier or learning a discriminative model. In general, it will be up to the user to find a proper trade off between quality of the results and speed up of the process.

4.1 Shortest Representation of a Relevant Set

Based on Theorem 7 we know that generators Y of a closed set X are characterized to cover exactly the same positive examples, and at least the same negative examples. Because of this property, any generator will be redundant w.r.t. its closure. That is:

Remark 12 *Let Y be a generator of X in the closure system on the positives; then, $\Gamma^{+}(Y) = X$ always implies $\mathrm{TP}(Y) = \mathrm{TP}(X)$ and $\mathrm{FP}(X) \subseteq \mathrm{FP}(Y)$ (from Lemma 6 and Theorem 7). However, note that the inclusion between the set of false positives is not necessarily proper.*

However, we have $\mathrm{FP}(X) \subseteq \mathrm{FP}(Y)$ for Y generator of X; so, it might happen that some generators Y are equivalent to their closed set X in that they cover exactly the same true positives and also the same false positives.

Definition 13 (Equivalent generators) *Let $\Gamma^{+}(Y) = X$ and $Y \neq X$. We say that a generator Y is equivalent to its closure X iff $\mathrm{FP}(X) = \mathrm{FP}(Y)$.*

The equivalence between true positives of Y and X is guaranteed because $\Gamma^{+}(Y) = X$. Therefore, it would be only necessary to check if generators cover the same false positives than its closure to check equivalence. Generators will provide a more general representation of the relevant set (because $Y \subset X$ by construction). So, $Y \to +$ is shorter than the rule $X \to +$ and it is up to the user to choose the more meaningful to her or to the application. For example, this may depend on a minimum-length criterion of the final classification rules: a generator Y equivalent to a closed set X satisfies by construction that $Y \subset X$, so $Y \to +$ is shorter than the rule $X \to +$. Then, the minimal equivalent generators of a closed itemset X naturally correspond to the minimal representation of the relevant rule $X \to +$.

In terms of the closure operator of negatives, we have the following way of characterizing these equivalent generators.

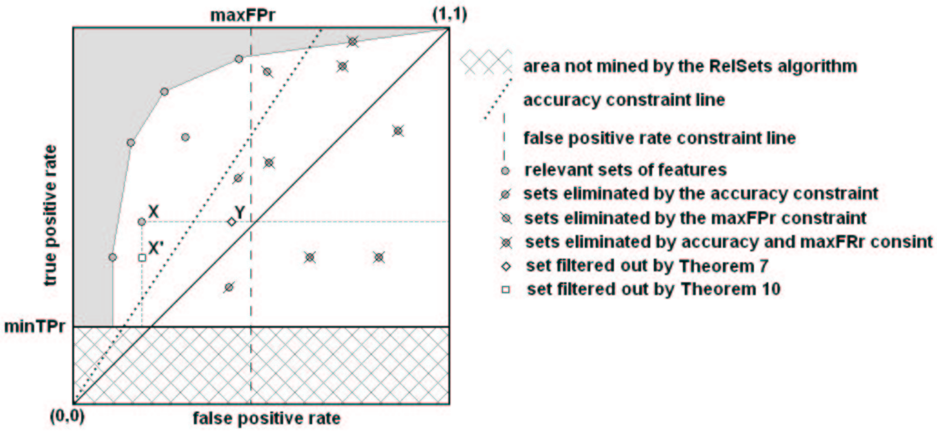


Figure 2: The evaluation of relevant combinations of features in the ROC space.

Proposition 14 *Let* $\Gamma^+(Y) = X$ *and* $Y \neq X$*. Then* Y *is an equivalent generator of* X *iff* $\Gamma^-(X) = \Gamma^-(Y)$.

Proof It is defined that the generator Y is equivalent to its closure X when $\mathrm{FP}(X) = \mathrm{FP}(Y)$, which directly implies $\Gamma^-(X) = \Gamma^-(Y)$ by construction of Γ. On the other direction: $\Gamma^-(X) = \Gamma^-(Y)$ implies $|\mathrm{FP}(Y)| = |\mathrm{FP}(X)|$, but because $Y \subseteq X$ by the extensivity of Γ, we necessarily have that $\mathrm{FP}(Y) = \mathrm{FP}(X)$. ■

It is well-known that minimal generators of a closed set X can be computed by traversing the hypergraph of differences between X and their proper predecessors in the system (see, for example, Pfaltz and Taylor, 2002). In practice, efficient algorithms have been designed for computing free sets and their generalizations (see, for example, Calders and Goethals, 2003).

5. Evaluation of Relevant Sets in the ROC Space

The ROC space (Provost and Fawcett, 2001) is a 2-dimensional space that shows a classifier (rule/ ruleset) performance in terms of its *false positive rate* (also called 'false alarm'), $\mathrm{FPr} = \frac{|\mathrm{FP}|}{|\mathrm{TN}|+|\mathrm{FP}|} = \frac{|\mathrm{FP}|}{|N|}$ plotted on the X-axis, and *true positive rate* (also called 'sensitivity') $\mathrm{TPr} = \frac{|\mathrm{TP}|}{|\mathrm{TP}|+|\mathrm{FN}|} = \frac{|\mathrm{TP}|}{|P|}$ plotted on the Y-axis. The ROC space is appropriate for measuring the quality of rules since rules with the best covering properties are placed in the top left corner, while rules that have similar distribution of covered positives and negatives as the distribution in the entire data set are close to the main diagonal.

A set of features from Definition 5 can be interpreted as a condition part of a rule or also as a subgroup description. A set of relevant sets of features from Definition 11 can therefore be visualized and evaluated in the ROC space as a ruleset.

Relevant sets are induced with a minimum support constraint on the positives (as discussed in Section 4). This means that in the ROC space they all lie above the minimum true positive rate constraint line (in Figure 2 denoted as minTPr). Relevant sets are depicted in Figure 2 as circles.

Sometimes, depending on the application, additional filtering criteria are applied. In such cases a maximum false positive rate constraint can be imposed (in Figure 2 this constraint is represented by a dashed line, rules eliminated by this constraint are shown as circles with backslash), or we can apply a minimum confidence constraint (represented by a dotted line, rules eliminated by this constraint are shown as slashed circles in Figure 2). Alternatively we may simply select just the rules on the convex hull.

Let us interpret and visualize Theorems 7 and 10 in the ROC space. According to Theorem 7, sets of features Y, s.t. $Y \subset X$, that cover the same positives as X (i.e., $TP(Y) = TP(X)$), are filtered out. Since Y and X have the same true positive rate (i.e., $TPr(Y) = TPr(X)$), both lie on the same horizontal line in the ROC space. Since Y is a subset of X, which in rule learning terminology translates into "rule X is a specialization of rule Y", $FPr(X) \leq FPr(Y)$ so Y is located at the right hand side of X. In Figure 2, a sample feature set filtered out according to Theorem 7 is depicted as a diamond. Note that this captures exactly the notion of relevancy defined by Lavrač and Gamberger (2005) and Lavrač et al. (1999).

According to Theorem 10, sets of features X', s.t. $X \subset X'$, that cover the same negatives as X (i.e., $FP(X') = FP(X)$), are filtered out. Since X' and X have the same false positive rate (i.e., $FPr(X') = FPr(X)$), both lie on the same vertical line in the ROC space. Since X is a subset of X', which in rule learning terminology translates into "rule X' is a specialization of rule X ", $TPr(X) \geq TPr(X')$, therefore X is located above X' in the ROC space. In Figure 2, a sample feature set filtered out according to Theorem 10 is depicted as a square.

Note that the feature sets filtered out by the relevancy filter are never those on the ROC convex hull. Furthermore, it can be proved that there are no sets of features outside the convex hull (grey area on Figure 2 denotes an area without sets/rules).

6. Experimental Evaluation

The results presented above lead to the concept of closed sets in the context of labeled data. In practice, closed sets can be discovered from labeled data as follows.

1. First, mining the set $S = \{X_1, \ldots, X_n\}$ of frequent closed itemsets from the target class (Theorem 7). This requires a minimum support constraint on positives. For our experiments we will use the efficient LCM algorithm by Uno et al. (2004).

2. Second, reducing S to the space of relevant set of features by checking the coverage in the negatives (Theorem 10). Schematically, for any closed set $X_i \in S$, if there exists another closed set $X_j \in S$ such that both have the same support in the negatives and $X_j \subset X_i$, then X_i is removed.

The first step of this process usually requires a minimum support constraint on true positives, while the second step can be computed automatically without any constraints. However, depending on the purpose of the application we can apply an extra filtering criterion (such as forcing a maximum false positive constraint on the negatives, or a minimum accuracy constraint), or compute minimal equivalent generators of the relevant sets as described above. For short, we will name this computing process as *RelSets* (i.e., the process of discovering the Relevant Sets of features of Definition 5).

			Emerging Patterns					
			Growth rate > 1.5			Growth rate ∞		
Data set	Class	Distrib. %	EPs	RelSets	CF%	EPs	RelSets	CF%
Lenses	soft	20.8	31	4	87.10	8	3	62.5
	hard	16.9	34	3	91.18	6	2	66.67
	none	62.5	50	12	76.00	42	4	90.48
Iris	setosa	33.3	83	16	80.72	71	7	90.14
	versicolor	33.3	134	40	70.15	63	10	84.13
	virginica	33.3	92	16	82.61	68	6	91.18
Breast-w	benign	65.5	6224	316	94.92	5764	141	97.55
	malignant	34.5	3326	628	81.12	2813	356	87.34
SAheart	0	34.3	4557	1897	58.37	2282	556	75.64
	1	65.7	9289	2824	69.60	3352	455	86.43
Balance-scale	B	7.8	271	75	72.32	49	49	0.00
	R	46	300	84	72.00	90	90	0.00
Yeast	MIT	16.4	3185	675	78.81	250	40	84.00
	CYT	31.2	3243	808	75.08	68	16	76.47
	ERL	0.3	1036	5	99.52	438	4	99.09
Monk-1	0	64.3	1131	828	26.79	321	18	94.39
	1	35.7	686	9	98.69	681	4	99.41
Lymphography	metastases	54.72	36435	666	98.17	10970	90	99.18
10% min supp.	malign	41.21	61130	740	98.79	19497	55	99.72
Crx	$+$	44.5	3366	782	76.76	304	26	91.44
10% min supp.	$-$	55.5	3168	721	77.24	12	5	58.33

Table 4: Compression factor (CF% = $(1 - \frac{|RelSets|}{|EPs|}) \times 100$) of EPs in several UCI data sets. Note that we did not impose any minimum true positive threshold on any data set, except for Lymphography and Crx, where all EPs and RelSets were discovered with a 10% threshold on true positives.

As discussed above, the minimum support constraint on the first phase will tend to prune too long closed sets and this might have an impact in the application. In practice however, it is known that the longest sets of features are sometimes too specific, thus leading to overfitting problems. It is up to the user to trade off between the specificity of the closed sets and the speed up of the process. Also notice that the lowest the minimum support constraint, the largest the number of closed sets, and thus, the most expensive it becomes to compute the second phase of the approach. Our goal is not to present efficient algorithms but to illustrate the concept of relevancy.

Still we find important to point out that the notion of relevancy explored in the paper prefers typically the shortest closed sets. This is obvious by the second reduction phase shown in Theorem 10, where the shortest sets are always more relevant than the longest ones if they cover the same negative examples. Thus, finding a proper threshold level for the minimum support is not critical in our experiments as different minimum support thresholds lead to very similar results.

6.1 Emerging Patterns on UCI data

Emerging Patterns (EP) (Dong and Li, 1999; Li et al., 2000; Dong et al., 1999) are sets of features in the data whose supports change significantly from one class to another. More specifically, EPs are itemsets whose growth rates (the ratio of support from one class to the other, that is, $\frac{\mathrm{TPr}}{\mathrm{FPr}}$ of the pattern) are larger than a user-specified threshold. In this experimental setting we want to show that some of the EPs mined by these approaches are redundant, and that our relevant sets correspond to the notion of compacted data representation for labeled data. Indeed, EPs are a superset of the result returned by RelSets.

In our comparisons we calculate relevant sets over a certain growth rate threshold (1.5 and infinite), and we compare this with the number of EPs by using the same growth rate constraint. Numerical attributes in the data sets are discretized when necessary by using four equal frequency intervals. Although being a very simple discretization scheme, we want to point out that our goal in this experiment is to compare the number of EPs with our relevant sets, and thus, any preprocessing decision on the original data will affect in the same way the two methods we wish to compare.

Results are shown in Table 4. We observe that compression factor may vary according to the data set. When data is structurally redundant, compression factors are higher since many frequent sets are redundant with respect to the closed sets. However, in data sets where this structural redundancy does not exist (such as the Balance-scale data), the compression factor is zero, or close to zero.

A set of relevant properties of EPs have been studied in Soulet et al. (2004). This latter work also identifies condensed representations of EPs from closed sets mined in the whole database. Our approach is different in that we deal with pieces of the data for each class separately, and this allows for a reduction phase given by Theorem 10. Indeed, the amount of compression that this second phase provides in our approach depends on the distribution of the negative examples in the data, but at least, the number of relevant sets obtained by RelSets will be always smaller than the number of condensed EPs from Soulet et al. (2004).

6.2 Essential Rules on UCI Data

Essential rules were proposed by Baralis and Chiusano (2004) to reduce the number of association rules to those with nonredundant properties for classification purposes. Technically, they correspond to mining all frequent itemsets and removing those sets X such that there exists another frequent Y with $Y \subset X$ and having both the same support in positives and negatives. This differs from our proposal in the way of treating the positive class with closed sets. The compression factor achieved for these rules is shown in Table 5. Note that essential rules are not pruned by growth rate threshold, and this is why their number is usually higher than the number of emerging patterns shown in previous subsection.

6.3 Subgroup Discovery in Microarray Data Analysis

Microarray gene expression technology offers researchers the ability to simultaneously examine expression levels of hundreds or thousands of genes in a single experiment. Knowledge about gene regulation and expression can be gained by dividing samples into control samples (in our case mock infected plants), and treatment samples (in our case virus infected plants). Studying the differences between gene expression of the two groups (control and treatment) can provide useful insights into complex patterns of host relationships between plants and pathogens (Taiz and Zeiger, 1998).

Data set	Class	Distrib. %	Essential rules	RelSets	CF%
Lenses	soft	20.8	43	4	90.69
	hard	16.9	39	3	92.30
	none	62.5	89	19	78.65
Iris	setosa	33.3	76	20	73.68
	versicolor	33.3	111	41	63.06
	virginica	33.3	96	27	71.87
Breast-w	benign	65.5	3118	377	87.90
	malignant	34.5	2733	731	73.25
SAheart	0	34.3	6358	4074	35.92
	1	65.7	9622	4042	58
Balance-scale	B	7.8	415	147	88.67
	R	46	384	364	5.20
Yeast	MIT	16.4	2258	1125	50.17
	CYT	31.2	2399	1461	80.78
	ERL	0.3	417	5	98.80
Monk-1	0	64.3	1438	1135	21.07
	1	35.7	1477	363	75.42
Lymphography 10% min supp.	metastases	54.72	1718	369	78.52
	malign	41.21	2407	476	80.22
Crx 10% min supp.	$+$	44.5	2345	1091	53.47
	$-$	55.5	2336	1031	55.86

Table 5: Compression factor (CF% = $(1 - \frac{|RelSets|}{|EPs|}) \times 100$) of essential rules in UCI data sets. Note that essential rules and RelSets are not pruned by any growth rate threshold.

Microarray data analysis problems are usually addressed by statistical and data mining/machine learning approaches (Speed, 2003; Causton et al., 2003; Parmigiani et al., 2003). State-of-the-art machine learning approaches to microarray data analysis include both supervised learning (learning from data with class labels) and unsupervised learning (such as conceptual clustering). A review of these various approaches can be found in Molla et al. (2004). It was shown by Gamberger et al. (2004) that microarray data analysis problems can be approached also through subgroup discovery, where the goal is to find a set of subgroup descriptions (a rule set) for the target class, that preferably has a low number of rules while each rule has high coverage and accuracy (Lavrač et al., 2004; Gamberger and Lavrač, 2002).

The goal of the real-life experiment addressed in this paper is to investigate the differences between virus sensitive and resistant transgenic potato lines. For this purpose, 48 potato samples were used, leading to 24 microarrays. The laboratory experiment was carried out at the National Institute of Biology, Ljubljana, Slovenia.

Our data set contains 12 examples. Each example is a pair of microarrays (8 and 12 hours after infection) from the same transgenic line. All the data was discretized by using expert background knowledge. Features of the form $|gene\ expression\ value| > 0.3$ were generated and enumerated. Three groups of features were generated: first group corresponding to gene expression levels 8 hours after infection (feature numbers $\in [1, 12493]$); second group corresponding to gene expression levels 12 hours after infection (feature numbers $\in [12494, 24965]$); finally, a third group corresponding

Data set	Class	Num. of rules			AUC		Time	
		RelSets	RelSets-ROC	SD	RelSets	SD	RelSets	SD
potatoes	sensitive	1	1	20	100%	100%	<1s	>1h
	resistant	1	1	20	100%	91%	<1s	>1h

Table 6: Comparison of algorithms RelSets and SD on the potato microarray data. Column RelSets-ROC shows the number of RelSets rules on the ROC convex hull.

to the difference between gene expression levels 12 and 8 hours after infection (feature numbers $\in [24966, 37559]$).

We used the RelSets algorithm to analyze the differences between gene expression levels characteristic for virus sensitive potato transgenic lines, discriminating them from virus resistant potato transgenic lines and vice versa. We ran it twice: once the sensitive examples were considered positive and once the resistant ones were considered positive. In both cases the constraint of minimal true positive count was set to 4, and in the first phase the algorithm returned 22 closed sets on positives. Rule relevancy filtering according to Definition 5, filtered the rules to just one relevant rule with a 100% true positive rate and a 0% false positive rate for each class. The results gained are shown below, where features are represented by numbers.

Twelve features determine the virus sensitive class for the potato samples used:

{*13031, 13066, 19130, 23462, 24794, 25509, 29938, 33795, 33829, 35003, 35190, 36266*} → *sensitive*

Sixteen features determine the virus resistant class for the potato samples used:

{*16441, 20474, 20671, 24030, 25141, 29777, 30111, 32459, 33225, 33248, 33870, 34108, 34114, 34388, 37252, 37484*} → *resistant*

When comparing our results with the SD algorithm for subgroup discovery (Gamberger and Lavrač, 2002), we observe that the running time of SD degrades considerably due to the high dimensionality of this data set. Moreover, SD obtains a larger set of rules which are less interpretable and do not have the same quality as the rules obtained with RelSets. Table 6 shows the numbers of discovered rules, area under ROC curve and the running time of both algorithms.

The results obtained with RelSets were validated by the experts from the National Institute of Biology, Ljubljana, Slovenia, and evaluated as insightful. Based on the tested samples, the experts have observed that the response to the infection after 8 hours is not strong enough to distinguish between resistant transgenic lines and sensitive ones. None of the gene expression changes after 8 hours appeared significant for the RelSets algorithm. However, selected gene expression levels after 12 hours and the comparison of gene expression difference (12-8) characterize the resistance to the infection with potato virus for the transgenic lines tested.[3]

[3] Details of this analysis are beyond the scope of this paper: first qualitative analysis results have appeared in Kralj et al. (2006), while a more thorough analysis is to appear in a biological journal.

7. Conclusions

We have presented a theoretical framework that, based on the covering properties of closed itemsets, characterizes those sets of features that are relevant for discrimination. We call them closed sets for labeled data, since they keep similar structural properties of classical closed sets, yet taking into account the positive and negative labels of examples. We show that these sets define a nonredundant set of rules in the ROC space.

This study extends previous results where the notion of relevancy was analyzed for single features (Lavrač and Gamberger, 2005; Lavrač et al., 1999), and it provides a new formal perspective for relevant rule induction. In practice the approach shows major advantages for compacting emerging patterns and essential rules and solving hard subgroup discovery problems. Thresholds on positives make the method tractable even for large databases with many features. The application to potato microarray data, where the goal was to find differences between virus resistant and virus sensitive potato transgenic lines, shows that our approach is not only fast, but also returns a small set of rules that are meaningful and easy to interpret by domain experts.

Future work will be devoted to adapting efficient algorithms of emerging patterns by Dong and Li (1999) for the discovery of the presented relevant sets.

Acknowledgments

This work was partially funded by the Pascal Network of Excellence through a visit of the first author to the Jožef Stefan Institute, Ljubljana, Slovenia, and the Slovenian Research Agency grant Knowledge Technologies (2004–2008). We wish to thank Ana Rotter, Nataša Toplak and Kristina Gruden from the National Institute of Biology, Ljubljana, Slovenia, for the biological data and the joint research on the application of closed sets in functional genomics.

References

R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A.I. Verkamo. Fast discovery of association rules. *Advances in Knowledge Discovery and Data Mining*, pages 307–328, 1996.

J.L. Balcázar and J. Baixeries. Discrete deterministic datamining as knowledge compilation. In *SIAM Int. Workshop on Discrete Mathematics and Data Mining*, 2003.

E. Baralis and S. Chiusano. Essential classification rule sets. *ACM Trans. Database Syst.*, 29(4): 635–674, 2004.

Y. Bastide, N. Pasquier, R. Taouil, G. Stumme, and L. Lakhal. Mining minimal non-redundant association rules using frequent closed itemsets. *Lecture Notes in Computer Science*, 1861:972–986, 2000a.

Y. Bastide, R. Taouil, N. Pasquier, G. Stumme, and L. Lakhal. Mining frequent patterns with counting inference. *SIGKDD Explor. Newsl.*, 2(2):66–75, 2000b.

S.D. Bay and M.J. Pazzani. Detecting group differences: Mining contrast sets. *Data Min. Knowl. Discov.*, 5(3):213–246, 2001. ISSN 1384-5810.

J.F. Boulicaut, A. Bykowski, and C. Rigotti. Free-sets: A condensed representation of boolean data for the approximation of frequency queries. *Data Min. Knowl. Discov.*, 7(1):5–22, 2003. ISSN 1384-5810.

S. Brin, R. Motwani, J.D. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. In *Proceedings ACM SIGMOD Int. Conference on Management of Data*, pages 255–264, 1997.

T. Calders and B. Goethals. Minimal *k*-free representations of frequent sets. In *Proceedings of the 7th European Conference on Principles and Knowledge Discovery in Data mining*, pages 71–82, 2003.

C. Carpineto and G. Romano. *Concept Data Analysis. Theory and Applications.* Wiley, 2004.

H.C. Causton, J. Quackenbush, and A. Brazma. *Microarray Gene Expression Data Analysis: A Beginner's Guide.* Blackwell Publishing, Oxford, United Kingdom, 2003.

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

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

B. Crémilleux and J. F. Boulicaut. Simplest rules characterizing classes generated by delta-free sets. In *Proceedings of the 22nd Annual International Conference Knowledge Based Systems and Applied Artificial Intelligence*, pages 33–46, 2002.

G. Dong and J. Li. Efficient mining of emerging patterns: discovering trends and differences. In *Proceedings of the 5th Int. Conference on Knowledge discovery and data mining*, pages 43–52, 1999.

G. Dong, X. Zhang, L. Wong, and J. Li. CAEP: classification by aggregating emerging patterns. In *Proceedings of the 2nd In. Conference on Discovery Science*, pages 30–42, 1999.

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

D. Gamberger, N. Lavrač, F. Železný, and J. Tolar. Induction of comprehensible models for gene expression datasets by subgroup discovery methodology. *Journal of Biomedical Informatics*, 37 (4):269–284, 2004.

B. Ganter and R. Wille. *Formal Concept Analysis. Mathematical Foundations*. Springer, 1998.

G.C. Garriga, P. Kralj, and N.Lavrač. Closed sets for labeled data. In *Proceedings of the 10th Int. Conference on Principles and Knowledge Discovery on Databases*, pages 163–174, 2006.

B. Goethals and M. Zaki. Advances in frequent itemset mining implementations: report on FIMI'03. *SIGKDD Explor. Newsl.*, 6(1):109–117, 2004.

J. Han and J. Pei. Mining frequent patterns by pattern-growth: methodology and implications. *SIGKDD Explor. Newsl.*, 2(2):14–20, 2000.

V. Jovanoski and N. Lavrač. Classification rule learning with APRIORI-C. In *Proceedings of the10th Portuguese Conference on Artificial Intelligence on Progress in Artificial Intelligence, Knowledge Extraction, Multi-agent Systems, Logic Programming and Constraint Solving (EPIA '01)*, pages 44–51. Springer-Verlag, 2001.

B. Kavšek and N. Lavrač. APRIORI-SD: Adapting association rule learning to subgroup discovery. *Applied Artificial Intelligence*, To appear, 2006.

D. Koller and M. Sahami. Toward optimal feature selection. In *Proceedings of the 13th Int. Conference on Machine Learning*, pages 284–292, 1996.

P. Kralj, A. Grubešič, K. Gruden N. Toplak, N. Lavrač, and G.C. Garriga. Application of closed itemset mining for class labeled data in functional genomics. *Informatica Medica Slovenica*, 2006.

N. Lavrač and D. Gamberger. Relevancy in constraint-based subgroup discovery. *Constraint-Based Mining and Inductive databases*, 3848:243–266, 2005.

N. Lavrač, D. Gamberger, and V. Jovanoski. A study of relevance for learning in deductive databases. *Journal of Logic Programming*, 40(2/3):215–249, 1999.

N. Lavrač, B. Kavšek, P. Flach, and L. Todorovski. Subgroup discovery with CN2-SD. *Journal of Machine Learning Research*, 5:153–188, 2004.

J. Li, G. Dong, and K. Ramamohanarao. Instance-based classification by emerging patterns. In *Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery*, pages 191–200, 2000.

B. Liu, W. Hsu, and Y. Ma. Integrating classification and association rule mining. In *Proceedings of the 4th Int. Conference on Knowledge Discovery and Data Mining*, pages 571–574, 1998.

H. Liu and H. Motoda. *Feature Selection for Knowledge Discovery and Data Mining*. Kluwer Academic Publishers, 1998.

M. Molla, M. Waddell, D. Page, and J. Shavlik. Using machine learning to design and interpret gene-expression microarrays. *AI Magazine*, 25(1):23–44, 2004.

G. Parmigiani, E.S. Garrett, R.A. Irizarry, and S.L. Zeger, editors. *The Analysis of Gene Expression Data: Methods and Software*. Springer-Verlag, New York, 2003.

N. Pasquier, Y. Bastide, R. Taouil L., and Lakhal. Closed set based discovery of small covers for association rules. *Networking and Information Systems*, 3(2):349–377, 2001.

J.L. Pfaltz. Closure lattices. *Discrete Mathematics*, 154:217–236, 1996.

J.L. Pfaltz and C.M. Taylor. Scientific knowledge discovery through iterative transformations of concept lattices. In *SIAM Int. Workshop on Discrete Mathematics and Data Mining*, pages 65–74, 2002.

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

A. Soulet, B. Crémilleux, and F. Rioult. Condensed representation of eps and patterns quantified by frequency-based measures. In *Proceedings of Knowledge Discovery in Inductive Databases Workshop*, pages 173–190, 2004.

T.P. Speed, editor. *Statistical Analysis of Gene Expression Microarray Data*. Chapman & Hall/CRC, Boca Raton, 2003.

L. Taiz and E. Zeiger. *Plant Physiology*. Sinauer Associates, second edition (372:374) edition, 1998.

P-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2005.

R. Taouil, Y. Bastide, N. Pasquier, and L. Lakhal. Mining bases for association rules using closed sets. In *Proceedings of the 16th Int. Conference on Data Engineering*, page 307. IEEE Computer Society, 2000.

T. Uno, T. Asai, Y. Uchida, and H. Arimura. An efficient algorithm for enumerating closed patterns in transaction databases. In *Discovery Science*, pages 16–31, 2004.

I.H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 2005.

M. Zaki. Generating non-redundant association rules. In *Proceedings of the 6th Int. Conference on Knowledge Discovery and Data Mining*, pages 34–43, 2000a.

M. Zaki. Mining non-redundant association rules. *Data Mining and Knowledge Discovery: An International Journal*, 4(3):223–248, 2004.

M. Zaki. Scalable algorithms for association mining. *IEEE Transactions on Knowledge and Data Engineering*, 12(3):372–390, 2000b.

M. Zaki and M. Ogihara. Theoretical foundations of association rules. In *SIGMOD-DMKD Int. Workshop on Research Issues in Data Mining and Knowledge Discovery*, 1998.

J. Zhang, E. Bloedorn, L. Rosen, and D. Venese. Learning rules from highly unbalanced data sets. In *Proceedings of the 4th. IEEE Int. Conference on Data Mining (ICDM'04)*, pages 571–574, 2004.

5 Results Summary

The purpose of this dissertation is to unify data mining tasks that deal with finding differences between groups in a novel unifying framework, named supervised descriptive rule induction (SDRI). By doing so, our aim is to improve individual supervised descriptive rule induction methods by cross-fertilizing the approaches developed by individual subareas of supervised descriptive rule induction. Furthermore, we aim at developing novel SDRI methods through component exchange (e.g., enabling the use of subgroup discovery components in contrast set mining). Finally, we aim at showing the advantages of this approach in applications in important real life problems in medicine and biology.

In this chapter, we present and discuss the achieved results. The discoveries are not treated in their chronological order, since the research process was iterative and different goals are entangled. We rather describe the results and discoveries in the following logical order: first the contributions to data mining and then the contributions to application areas. Besides the scientific contributions, the methodology used and the availability of the developed software are also discussed.

5.1 Methodology

We used the following methodology to prove the research hypothesis (stated in Section 1.3) and achieve the goals of the thesis. First, the existing literature on different data mining tasks focused on distinguishing between groups was elaborated in an overview. Second, we proposed a theoretical framework that unifies the terminology, definitions and heuristics of supervised descriptive rule induction. We continued our research with experiments in real-life domains. We next proposed a way of adapting subgroup discovery visualizations to contrast set mining and emerging patterns. All in all, the methodology used to prove our hypothesis was to implement the SDRI framework and show that it really brings the desired benefits. The hypothesis has indeed been proven by successful applications in important real-life problems in medicine (analysis of brain ischemia) and biology (analysis of virus infected potato plants).

5.2 Contributions to data mining

The basis of this thesis is the identification of supervised descriptive rule induction (SDRI) tasks (contrast set mining, emerging pattern mining, subgroup discovery and other related approaches) and a survey of past SDRI research (Chapter 2). The survey is a fundamental step that enabled us to generalize from specific tasks that deal with finding descriptive rules on labeled data to a general SDRI framework.

The proposed SDRI framework (Chapter 2) unifies the supervised descriptive rule induction terminologies, definitions and heuristics. Once the general framework is available, the cross-fertilization of approaches developed by individual sub-areas of supervised descriptive rule induction can begin. We next describe a methodology for contrast set mining though subgroup discovery, supporting factors, and visualization as examples of the cross-fertilization process enabled by the SDRI unifying framework (Chapter 3). They are followed by the RelSets methodology of closed sets for labeled data (Chapter 4).

Chronologically, one of our first steps towards the unifying supervised descriptive rule induction framework was the identification of contrast set mining and subgroup discovery as very similar data mining tasks. A methodology for contrast set mining though subgroup discovery, named CSM-SD (Chapter 3), presents two possible transformations of a contrast set mining problem to a subgroup discovery problem: the formally justified *pairwise* transformation and the application driven *one-versus-all* transformation. Besides presenting the contrast set mining results to the end user, we have identified other contrast set mining open issues, that can be solved through subgroup discovery: avoiding of overlapping rules, handling attributes with continuous values and choosing the appropriate search heuristics. The concrete application of our SDRI framework to contrast set mining resulted in the novel CSM-SD methodology, developed in tight collaboration with the domain expert while analyzing data about brain ischemia patients. This research was application driven.

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 discovered subgroup are called the *supporting factors* (Gamberger *et al.*, 2003). Supporting factors are useful for presenting the discovered rules to the end user, since they complement the principal factors when distinguishing between the classes. Since presenting the contrast set mining results to the end user was one of the contrast set mining open issues, we adapted this subgroup discovery method to contrast set mining. Supporting factors from subgroup discovery can not be directly used for contrast set mining, but, by using the unifying SDRI framework, they can be effectively adapted to contrast set mining and therefore improve the contrast set mining explanatory potential (Chapter 3).

A relatively trivial, but nevertheless significant cross-fertilization achievement is the generalization of subgroup discovery visualization methods to supervised descriptive rule induction. Presenting the results of contrast set mining results to the end user (which includes visualization) was identified as an open issue of contrast set mining by Webb *et al.* (2003). On the other hand, the visualization problem has been addressed by many authors in another SDRI sub-area: subgroup discovery (Atzmüller and Puppe, 2005; Gamberger *et al.*, 2002; Kralj *et al.*, 2005; Wettschereck, 2002; Wrobel, 2001). The SDRI framework allows for the detection of similarities and differences between contrast set mining and subgroup discovery and, by doing so, it proposes subgroup discovery visualization methods that fit directly the desired contrast set mining problem and also shows the way of adaptation of other methods that are not directly applicable. As mentioned above, we have not only generalized the subgroup discovery visualization methods to contrast set mining, but to SDRI tasks in general (Chapter 2).

To summarize, the unifying framework is not only a scientific achievement on its own, but it allows for the cross-fertilization of different supervised descriptive rule induction algorithms and approaches.

Besides contributing the unifying SDRI framework and its successful applications, we have also developed a new method for SDRI named mining of closed sets for labeled data and the algorithm *RelSets* (Chapter 4). We have presented a theoretical framework that, based on the covering properties of closed itemsets, characterizes those sets of features that are relevant for discrimination while keeping similar structural properties to classical closed sets. We show that these sets define a non-redundant set of rules in the ROC space. The RelSets algorithm returns all the relatively relevant rules which fulfill the minimum true positive count constraint. The algorithm is complete in the sense that it finds all the most specific rules satisfying the constraints. This is, for example, appropriate for microarray data analysis since not many examples are available and exhaustive search of the space can be desired (Kralj *et al.*, 2006).

5.3 Applications of supervised descriptive rule induction methods

We have applied SDRI methods to practical problem domains in medicine and biology.

In the medical application (Chapter 3), we developed and used our CSM-SD methodology for contrast set mining through subgroup discovery on a real-life problem of analyzing a dataset of patients with brain ischemia, where the goal of data analysis was to determine the type of brain ischemia from risk factors obtained from anamnesis, physical examination, laboratory tests and ECG data. The data analysis process was iterative and included interaction with the domain expert in each iteration. First, standard data mining methods were used, like decision tree (Quinlan, 1986) and classification rule (Cohen, 1995) learning, but both lead to results that were not satisfactory to the domain expert. Second, the data mining task was formulated as a contrast set mining task and a formally justified transformation of contrast set mining to subgroup discovery was introduced, named the *pairwise* transformation. In this iteration, the domain expert was not fully satisfied with the result. Brainstorming on the methods used and the discussion of the expectations and way-of-thinking in the analyzed domain led to important new insights that triggered a new approach. In the next phase, an application driven transformation from contrast set mining to subgroup discovery was developed, which incorporated the experiences from the previous iterations and was well fitted to the problem and the expert's expectations. We named it the *one-versus-all* transformation. Last, we improved the explanatory potential of the discovered patterns by providing the supporting factors for each discovered pattern. The supporting factors, which have been defined and used in subgroup discovery, were adapted to the contrast set mining task and successfully used in our experiments. Visualization was also used in every iteration.

The analysis results were interpreted by the medical domain expert. The main aim of our research was to discover the differences between two types of stroke: embolic stroke and thrombotic stroke, to be able to define the risk factors for the diseases. Both types of stroke are ischemic (a clot blocks the blood flow to the brain), but the origin of the clot is different. The results confirmed some already known risk factors for stroke and new insights were also gained. For example, high systolic blood pressure (above 139) is in medical practice considered characteristic for both diseases. Our results confirm this finding and also indicate that extremely high systolic blood pressure (above 185) is not typical for embolic stroke patients. To summarize, our application proved successful: known risk factors were confirmed and new insights were gained.

Several lessons have been learned from this experiment. First, iterative knowledge discovery is a necessity. Second, the descriptive data analysis task is not concluded when individual patterns are discovered; presenting the results to the end user with proper visualization methods and additional information (in our case the supporting factors) makes the discovered patterns more tangible and therefore more acceptable to the end user. Third, the involvement of the end user is beneficial for achieving better analysis results. Last, the involvement of the end user is beneficial also for the development of the theory and methodology. To summarize, the interaction with the end user is of vital importance, not only for the application itself, but also for the development of the theory and methodology.

In the biological application (Chapter 4, Section 6.3), we applied the mining of closed sets for labeled data (RelSets for short) to potato microarray data, where the goal was to find differences in response to viral infection of virus resistant and virus sensitive potato transgenic lines (see Kralj *et al.*, 2006, for more details). The RelSets method was developed independently of any application and the motivation for its development was mainly theoretical. The reason for applying it to microarray data was that data mining tasks on microarray data differ from traditional data mining tasks because microarray domains are characterized by very large numbers of attributes (genes) relative to the number of examples (observations, samples). Standard SDRI algorithms do not perform well on microarray data because of this high dimensionality problem. In contrast, RelSets does not have any difficulty when faced with the high dimensionality problem.

The involvement of the biological expert was crucial in the non-trivial data preparation phase. Besides the data cleaning and normalization, which are standard preprocessing steps in microarray data analysis, expert-driven data discretization was also performed for semi-automatic feature generation, which was the most complex step due to the complicated biological experimental setup. Such data preprocessing can be used in other similar settings. Once the adequate features were generated, the algorithm was run and the results were visualized with heatmaps, which are a standard method in micorarray data visualization. RelSets is not only fast—much faster than other SDRI algorithms on microarray data—but also returns a small set of rules that are meaningful and easy to be interpreted by domain experts (see Table 6 in Chapter 4).

The analysis results were interpreted by a biology expert. The expert was, for example, able to determine the categories of genes that influence the sensitivity of potato plants to the tested virus. The analysis results also helped to elucidate the time response of the plants to the virus: all the plants responded similarly in the first eight hours after infection, while the response twelve hours after the infection was different for resistant and sensitive transgenic potato lines (Baebler *et al.*, 2009).

In summary, we have shown the adequacy of supervised descriptive rule induction methods in real-life data analysis problems where the goal is to gain new insights into the domain. In our applications, the domain experts were intensively involved in the knowledge discovery process, which was beneficial for both, the expert and the methodology development.

5.4 Software availability

The software that was used in most of the experiments in this dissertation is grouped in the *Subgroup Discovery Toolkit for Orange*, implemented in the Orange data mining toolbox (Demšar *et al.*, 2004). The algorithm implementation in Orange is valuable, as it offers various data structures, data and model visualization tools and has excellent facilities for building new visualizations.

The *Subgroup Discovery Toolkit for Orange* includes the implementation of:

- three subgroup discovery algorithms: SD (Gamberger and Lavrač, 2002), CN2-SD (Lavrač *et al.*, 2004) and Apriori-SD (Kavšek and Lavrač, 2006),
- two visualization methods: the visualization by bar charts and the representation of rules in the ROC space (Kralj *et al.*, 2005; Kralj Novak *et al.*, 2009b),
- six evaluation measures for subgroup discovery (Kavšek and Lavrač, 2006).

The algorithms are implemented in such a way that they can be used for both predictive and descriptive data mining. It follows that also all the Orange's facilities for classifiers can be used for subgroup discovery.

The *Subgroup Discovery Toolkit for Orange* is available under the GPL (General Public Licence) terms on the web page `http://kt.ijs.si/petra_kralj/SubgroupDiscovery/`. The description of the toolkit, system requirements, instructions for installation, screenshots and contact information are also available on the web page.

The *Subgroup Discovery Toolkit for Orange* does not include the implementation of the algorithms Magnum Opus (Webb, 1995), used in Chapter 2, and the algorithm for finding supporting factors (Gamberger *et al.*, 2003), used in Chapter 3. For these implementations, one should contact the original authors of the algorithms, who also co-authored the papers where the algorithms were used (for Magnum Opus contact Geoffrey I. Webb, for supporting factors contact Dragan Gamberger).

The implementation of the RelSets algorithm, described in Chapter 4, is available as a web service on the web site `http://kt.ijs.si/petra_kralj/RelSets/`.

An effort was made to make the software available for use by the general public. Even if such efforts do not bring scientific merits directly, they enable the continuity and spreading of the methods and contribute to the popularization of machine learning in general. These are important issues, even if they are not entirely scientific.

6 Conclusions and Further Work

In this dissertation, we have introduced the term supervised descriptive rule induction (SDRI), as a unification of several areas of machine learning that deal with finding comprehensible rules from class labeled data. We have developed a unifying framework for subgroup discovery, contrast set mining and emerging pattern mining, as representatives of supervised descriptive rule induction approaches, which includes the unification of the terminology, definitions and heuristics. By using our SDRI framework, we were able to overcome some open issues and limitations of SDRI sub-areas, like presenting the results to the end user by visualization and explanation through supporting factors. We have also developed a new method called mining of closed sets for labeled data - RelSets. It adapts closed sets to classification and discrimination purposes. We have successfully applied this method to the analysis of microarray data.

Applications of SDRI methods to real-life datasets and interaction with interested domain experts lead to new insights in the analyzed domains and to new methodology developments. For example, we have developed a methodology for contrast set mining though subgroup discovery, which proposes two transformations from contrast set mining to subgroup discovery. Depending on the problem at hand, a proper transformation should be used.

The main algorithms used in the experiments in this dissertation were made available on-line, mostly as downloadable tools. One direction for further research is to decompose SDRI algorithms, preprocessing and evaluation methods into basic components and their re-implementation as connectable web services, which includes the definition of interfaces between SDRI services. For instance, this can include the adaptation and implementation of subgroup discovery techniques to solving open problems in the area of contrast set mining and emerging patterns. This would allow for the improvement of algorithms due to the cross-fertilization of ideas from the different SDRI sub-areas.

Another issue that has not been addressed in this dissertation are complex data types and background knowledge. The SDRI attempts in this direction include relational subgroup discovery approaches 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). An attempt to building rules when using specialized biological background knowledge in ontological form is the method SEGS by Trajkovski *et al.* (2008). It is a step towards semantically enabled creative knowledge discovery in the form of descriptive rules, which may become the next knowledge discovery paradigm.

Even if the focus of this dissertation is on new theory and methods, the purpose of the methods is to be used in practical data analysis. Dissemination of the methods to additional applications is vital. A dissemination attempt is making the algorithm implementations available on-line and providing user friendly interfaces to their use. Other ideas for dissemination include publications in popular science media and lectures to diverse audiences. By promoting the methods in this way, they can achieve their purpose of being used in practical data analysis.

The core problem that motivated this thesis and the subsequent core contribution of this dissertation arises from the granularity of the scientific community. For instance, at least three "cliques" of data mining researchers in the field of supervised descriptive rule induction formed in time—contrast set mining, emerging pattern mining and subgroup discovery—each using their own terminology and background knowledge. In this thesis, we have unified the terminology and, by doing so, we were able to unify the supervised descriptive rule induction field and aggregate its achievements. We believe that the supervised descriptive rule induction field is not the only one using non-standardized terminology and having its researchers fractioned into closed communities. All researchers should grow awareness of the importance of standardized terminology and familiarity with related work.

7 Acknowledgments

There are many people I wish to thank for a huge variety of reasons.

Firstly, I wish to thank my supervisor Prof. Dr. Nada Lavrač, who—besides being my supervisor—was also my mentor, adviser, motivator, colleague and friend. We have bounded on several levels and helped each other in many situations, but certainly I am the one who benefited the most from our relationship. Due to her experienced guidance in the scientific world, I was able to perform high quality research, promote it at important scientific conferences and publish it in internationally recognized scientific journals.

Special thanks goes to the co-authors of our research papers Geoffrey I. Webb, Dragan Gamberger, Antonia Krstačić and Gemma C. Garriga. They contributed long hours of work, fruitful discussions, corrections, patience and much more to jointly succeed in our research and its publishing in internationally recognized journals and conferences.

Thanks goes also to Igor Trajkovski, who saved me a lot of time and effort by providing me with the template for the thesis format.

I should thank my examiners Prof. Dr. Sašo Džeroski, Assoc. Prof. Dr. Ljupčo Todorovski and Dr. Igor Mozetič for their valuable comments and remarks.

Let me express my respect to my colleagues—who in the meanwhile became also friends—and every co-worker at the Department of Knowledge Technologies at Jožef Stefan Institute, Ljubljana, Slovenia. Special thanks to Ivica Slavkov and Panče Panov, who, besides reading the manuscript and providing useful comments, provided moral support in the crucial moments.

Last but not least, I wish to thank my husband, who is always supportive and understanding.

8 References

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., and Verkamo, A. I. (1996). Fast discovery of association rules. *Advances in Knowledge Discovery and Data Mining*, pages 307–328.

Atzmüller, M. and Puppe, F. (2005). Semi-automatic visual subgroup mining using VIKAMINE. *Journal of Universal Computer Science (JUCS), Special Issue on Visual Data Mining*, **11**(11), 1752–1765.

Atzmüller, M. and Puppe, F. (2006). 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.

Atzmüller, M., Puppe, F., and Buscher, H.-P. (2005). Exploiting background knowledge for knowledge-intensive subgroup discovery. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI'05)*, pages 647–652.

Aumann, Y. and Lindell, Y. (1999). 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.

Baebler, Š., Krečič Stres, H., Rotter, A., Kogovšek, P., Cankar, K., Kok, E., Gruden, K., Kovač, M., Žel, J., Pompe Novak, M., and Ravnikar, M. (2009). PVY^{NTN} elicits a diverse gene expression response in different potato genotypes in the first 12h after inoculation. *Molecular Plant Pathology*, **10**(2), 263–275. doi:10.1111/j.1364-3703.2008.00530.x.

Bay, S. D. (2000). Multivariate discretization of continuous variables for set mining. In *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'00)*, pages 315–319.

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

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

Boulesteix, A.-L., Tutz, G., and Strimmer, K. (2003). A CART-based approach to discover emerging patterns in microarray data. *Bioinformatics*, **19**(18), 2465–2472.

Bruner, J. R., Goodnow, J. J., and Austin, G. A. (1956). *A study of thinking*. Wiley, New York.

Chapelle, O., Schölkopf, B., and Zien, A., editors (2006). *Semi-Supervised Learning*. MIT Press, Cambridge, MA.

Chapman, P., Kerber, R., Clinton, J., Khabaza, T., Reinartz, T., and Wirth, R. (1999). The CRISP-DM process model. Discussion Paper. http://www.crisp-dm.org.

Cios, K. J., Swiniarski, R. W., Pedrycz, W., and Kurgan, L. A. (2007). *Data Mining: A Knowledge Discovery Approach*, chapter 2: The Knowledge Discovery Process, pages 9–24. Springer US.

Clark, P. and Boswell, R. (1991). Rule induction with CN2: Some recent improvements. In *Proceedings of the 5th European Working Session on Learning (EWSL'91)*, pages 151–163.

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

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

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

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

Demšar, J., Zupan, B., and Leban, G. (2004). Orange: From experimental machine learning to interactive data mining, white paper (www.ailab.si/orange). Faculty of Computer and Information Science, University of Ljubljana.

Dong, G. and Li, J. (1999). 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.

Dong, G., Zhang, X., Wong, L., and Li, J. (1999). CAEP: Classification by aggregating emerging patterns. In *Proceedings of the 2nd International Conference on Discovery Science (DS'99)*, pages 30–42.

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

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

Fan, H., Fan, M., Ramamohanarao, K., and Liu, M. (2006). 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.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. In *Communication of the ACM*, volume 29, pages 27–34.

Frawley, W. J., Piatetsky-Shapiro, G., and Matheus, C. J. (1991). Knowledge discovery in databases: An overview. In *Knowledge Discovery in Databases*, pages 1–30. AAAI/MIT Press.

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

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

Gamberger, D., Lavrač, N., and Wettschereck., D. (2002). 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.

Gamberger, D., Lavrač, N., and Krstačić, G. (2003). Active subgroup mining: a case study in coronary heart disease risk group detection. *Artificial Intelligence in Medicine*, **28**, 27–57.

Garriga, G. C., Kralj, P., and Lavrač, N. (2006). 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.

Garriga, G. C., Kralj, P., and Lavrač, N. (2008). Closed sets for labeled data. *Journal of Machine Learning Research*, **9**, 559–580. http://www.jmlr.org/papers/volume9/garriga08a/garriga08a.pdf.

Ghahramani, Z. (2004). Unsupervised learning. In O. Bousquet, G. Raetsch, and U. von Luxburg, editors, *Advanced Lectures on Machine Learning*, pages 72–112. Springer-Verlag.

Hilderman, R. J. and Peckham, T. (2005). A statistically sound alternative approach to mining contrast sets. In *Proceedings of the 4th Australia Data Mining Conference (AusDM'05)*, pages 157–172.

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

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

Klösgen, W. and May, M. (2002). 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.

Kononenko, I. and Kukar, M. (2007). *Machine Learning and Data Mining: Introduction to Principles and Algorithms*. Horwood Publishing Limited, West Sussex.

Kotsiantis, S., Zaharakis, I., and Pintelas, P. (2006). Supervised machine learning: a review of classification techniques. *Artificial Intelligence Review*, **26**, 159–190.

Kralj, P., Lavrač, N., and Zupan, B. (2005). Subgroup visualization. In *Proceedings of the 8th International Multiconference Information Society (IS'05)*, pages 228–231.

Kralj, P., Rotter, A., Toplak, N., Gruden, K., Lavrač, N., and Garriga, G. C. (2006). Application of closed itemset mining for class labeled data in functional genomics. *Informatica Medica Slovenica*, (1), 40–45.

Kralj, P., Lavrač, N., Gamberger, D., and Krstačić, A. (2007a). Contrast set mining for distinguishing between similar diseases. In *Proceedings of the 11th Conference on Artificial Intelligence in Medicine (AIME'07)*, pages 109–118.

Kralj, P., Lavrač, N., Gamberger, D., and Krstačić, A. (2007b). Contrast set mining through subgroup discovery applied to brain ischaemia data. In *Proceedings of the 11th Pacific-Asia conference on Knowledge Discovery and Data Mining (PAKDD'07)*, pages 579–586.

Kralj Novak, P., Lavrač, N., Gamberger, D., and Krstačić, A. (2009a). CSM-SD: Methodology for contrast set mining through subgroup discovery. *Journal of Biomedical Informatics*, **42**(1), 113–122. doi:10.1016/j.jbi.2008.08.007.

Kralj Novak, P., Lavrač, N., and Webb, G. I. (2009b). Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. *Journal of Machine Learning Research*, **10**, 377–403. http://www.jmlr.org/papers/volume10/kralj-novak09a/kralj-novak09a.pdf.

Lavrač, N. and Gamberger, D. (2005). Relevancy in constraint-based subgroup discovery. *Constraint-Based Mining and Inductive databases*, **3848**, 243–266.

Lavrač, N., Kavšek, B., Flach, P. A., and Todorovski, L. (2004). Subgroup discovery with CN2-SD. *Journal of Machine Learning Research*, **5**, 153–188.

Lavrač, N., Kralj, P., Gamberger, D., and Krstačić, A. (2007). 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.

Li, J., Dong, G., and Ramamohanarao, K. (2000). Instance-based classification by emerging patterns. In *Proceedings of the 14th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'00)*, pages 191–200.

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

Lin, J. and Keogh, E. (2006). 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.

Liu, B., Hsu, W., and Ma, Y. (2001). 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.

Michalski, S. R., Carbonell, G. J., and Mitchell, M. T., editors (1986). *Machine learning an artificial intelligence approach volume II*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Mitchell, T. (1997). *Machine Learning*. McGraw Hill.

Pasquier, N., Bastide, Y., Taouil, R., and Lakhal, L. (1999). Discovering frequent closed itemsets for association rules. In *Proceedings of the 7th International Conference on Database Theory (ICDT'99)*, pages 398–416.

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

Russell, S. J. and Norvig, P. (2003). *Artificial Intelligence: A Modern Approach*. Pearson Education.

Simeon, M. and Hilderman, R. J. (2007). 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.

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

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning: An Introduction (Adaptive Computation and Machine Learning)*. The MIT Press.

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

Trajkovski, I., Lavrač, N., and Tolar, J. (2008). SEGS: Search for enriched gene sets in microarray data. *Journal of Biomedical Informatics*, **41**(4), 588–601.

Vapnik, V. N. (1998). *Statistical Learning Theory*. Wiley, New York.

Vilalta, R. and Drissi, Y. (2002). A perspective view and survey of meta-learning. *Artificial Intelligence Review*, **18**, 77–95.

Železný, F. and Lavrač, N. (2006). Propositionalization-based relational subgroup discovery with RSD. *Machine Learning*, **62**, 33–63.

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

Webb, G. I. (2001). 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.

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

Webb, G. I., Butler, S. M., and Newlands, D. (2003). 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.

Weiss, S. M. and Indurkhya, N. (1998). *Predictive data mining: a practical guide*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Wettschereck, D. (2002). 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.

Wong, T.-T. and Tseng, K.-L. (2005). Mining negative contrast sets from data with discrete attributes. *Expert Systems with Applications*, **29**(2), 401–407.

Wrobel, S. (1997). 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.

Wrobel, S. (2001). Inductive logic programming for knowledge discovery in databases. In S. Džeroski and N. Lavrač, editors, *Relational Data Mining*, chapter 4, pages 74–101.

Appendix 1: Publications Related to this Thesis

The main scientific contributions of this work were published in the following papers:

Journal papers:

- [Kralj Novak *et al.*(2009a)] Kralj Novak, P., Lavrač, N., Gamberger, D., and Krstačić, A. (2009a). CSM-SD: Methodology for contrast set mining through subgroup discovery. *Journal of Biomedical Informatics*, **42**(1), 113–122.

- [Kralj Novak *et al.*(2009b)] Kralj Novak, P., Lavrač, N., and Webb, G. I. (2009b). Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. *Journal of Machine Learning Research*, **10**, 377–403. http://www.jmlr.org/papers/volume10/kralj-novak09a/kralj-novak09a.pdf.

- [Garriga *et al.*(2008)] Garriga, G. C., Kralj, P., and Lavrač, N. (2008). Closed sets for labeled data. *Journal of Machine Learning Research*, **9**, 559–580. http://www.jmlr.org/papers/volume9/garriga08a/garriga08a.pdf.

- [Kralj *et al.*(2006)] Kralj, P., Rotter, A., Toplak, N., Gruden, K., Lavrač, N., and Garriga, G. C. (2006). Application of closed itemset mining for class labeled data in functional genomics. *Informatica Medica Slovenica*, (1), 40–45.

Conference papers:

- [Kralj *et al.*(2007a)] Kralj, P., Lavrač, N., Gamberger, D., and Krstačić, A. (2007a). Contrast set mining for distinguishing between similar diseases. In *Proceedings of the 11th Conference on Artificial Intelligence in Medicine (AIME 2007)*, pages 109–118.

- [Kralj *et al.*(2007b)] Kralj, P., Lavrač, N., Gamberger, D., and Krstačić, A. (2007b). Contrast set mining through subgroup discovery applied to brain ischaemia data. In

Proceedings of the 11th Pacific-Asia conference on Knowledge Discovery and Data Mining (PAKDD 2007), pages 579–586.

- [Lavrač *et al.*(2007)] Lavrač, N., Kralj, P., Gamberger, D., and Krstačić, A. (2007). 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 2007)*, pages 157–161.

- [Garriga *et al.*(2006)] Garriga, G. C., Kralj, P., and Lavrač, N. (2006). Closed sets for labeled data. In *Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2006)*, pages 163–174.

- [Kralj *et al.*(2005a)] Kralj, P., Lavrač, N., Zupan, B., and Gamberger, D. (2005a). Experimental comparison of three subgroup discovery algorithms: Analysing brain ischemia data. In *Proceedings of the 8th International Multiconference Information Society (IS 2005)*, pages 220 – 223.

- [Kralj *et al.*(2005b)] Kralj, P., Lavrač, N., and Zupan, B. (2005b). Subgroup visualization. In *Proceedings of the 8th International Multiconference Information Society (IS 2005)*, pages 228–231.

Appendix 2: Biography

Petra Kralj Novak was born in Šempeter pri Gorici, Slovenia, on November 4, 1980.

She completed the Bachelor of Science degree in computer science at the Faculty of Computer and Information Science, University of Ljubljana in 2005. Afterwards, she enrolled at the Ph.D. programme New Media and E-science at the Jožef Stefan International Postgraduate School.

During her Ph.D studies, she was research assistant at the Department of Knowledge Technologies at Jožef Stefan Institute, Ljubljana, Slovenia, and teaching assistant for knowledge discovery related subjects at the Jožef Stefan International Postgraduate School, Ljubljana, Slovenia, and at the University of Nova Gorica, Slovenia. From 2006 to 2009, she was the secretary and treasurer of SLAIS - Slovenian Artificial Intelligence Society. She was also a program committee member of the following conferences: ICDM 2008 - IEEE International Conference on Data Mining, PAKDD 2009 - The 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining and LeGo 2008 - From Local Patterns to Global Models - ECML/PKDD-08 Workshop.