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 $X \to +$ 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 $X \to +$. 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.

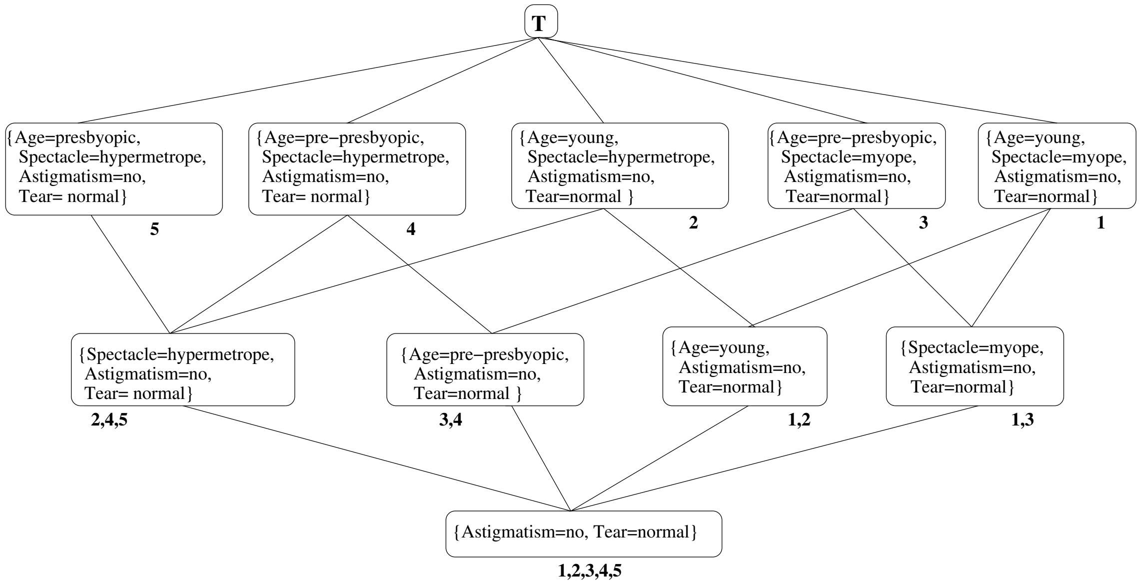

Figure 1: The lattice of closed itemsets for data in Table 2.

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 $\Gamma(\{$Age=young$\}) = \{$Age=young, Astigmatism=no, Tear=normal$\}$. Then, $\{$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 $X \to +$ 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 $X \to +$ 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 $Y \rightarrow +$ is *irrelevant* if there exists another rule $X \rightarrow +$ 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 $\rightarrow +$ is not relevant because at least the rule {Astigmatism=no, Tear=normal} $\rightarrow +$ 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 $Y \to +$ is less relevant than rule $X \to +$ 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., $Y \to +$ might be equivalent to $X \to +$). 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 $X \nsubseteq X'$ and $X' \nsubseteq X$ (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 $X \to +$ 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 $\rightarrow$ 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, $Y \rightarrow +$ is shorter than the rule $X \rightarrow +$ 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 $Y \rightarrow +$ is shorter than the rule $X \rightarrow +$. Then, the minimal equivalent generators of a closed itemset X naturally correspond to the minimal representation of the relevant rule $X \rightarrow +$.

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

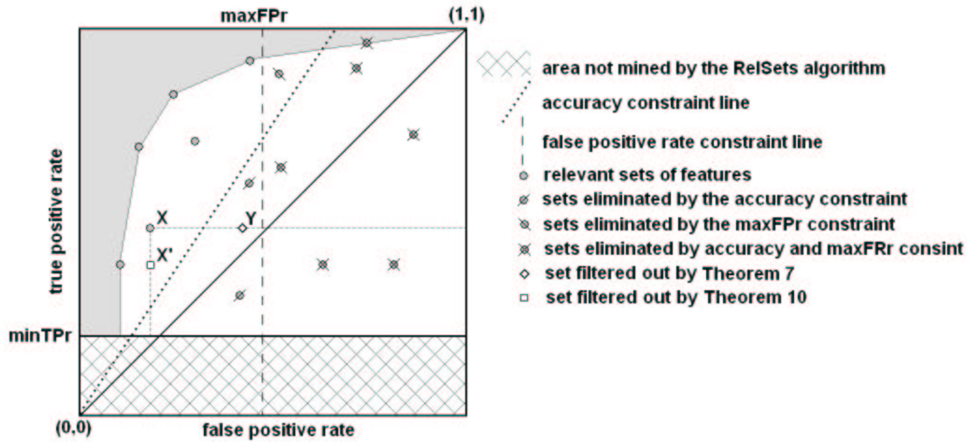

Figure 2: The evaluation of relevant combinations of features in the ROC space.

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, $\frac{\text{TPr}}{\text{FPr}}$ 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	metastases	54.72	1718	369	78.52
10% min supp.	malign	41.21	2407	476	80.22
Crx	$+$	44.5	2345	1091	53.47
10% min supp.	$-$	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*$\} \rightarrow$ *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*$\} \rightarrow$ *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.