Closed Sets for Labeled Data*

Gemma C. Garriga[1], Petra Kralj[2], and Nada Lavrač[2,3]

[1] Universitat Politècnica de Catalunya, Jordi Girona 1-3, 08034 Barcelona, Spain
garriga@lsi.upc.edu
[2] Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
[3] University of Nova Gorica, Vipavska 13, 5000 Nova Gorica, Slovenia
petra.kralj@gmail.com
nada.lavrac@ijs.si

Abstract. Closed sets are being successfully applied in the context of compacted data representation for association rule learning. However, their use is mainly descriptive. This paper shows that, when considering labeled data, closed sets can be adapted for prediction and discrimination purposes by conveniently contrasting covering properties on positive and negative examples. We formally justify 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. Here we apply it to compacting emerging patterns and essential rules and to learn descriptions for subgroup discovery.

1 Introduction

Rule discovery has been addressed from two different perspectives: data mining and machine learning. Data mining mainly explores unlabeled data, and the focus resides on finding all rules over a certain confidence that summarize the original data. On the other hand, machine learning is mainly concerned with the analysis of class labeled data, resulting in the induction of classification and prediction rules, and—more recently—also descriptive rules that aim at providing insightful knowledge from the data (subgroup discovery, contrast set mining). Traditional rule learning algorithms for classification include CN2 [3] and Ripper [4]. Other approaches have been proposed that are based on the association rule technology but are applied to class labeled data (e.g., the Apriori-C classifier [8] and the Essence algorithm for inducing "essential" classification rules based on the covering properties of frequent itemsets [1]).

In subgroup discovery the aim is to find subgroup descriptions that are characteristic for examples with a certain property of interest, and the closely related contrast set mining aims at capturing discriminating features that contrast instances between classes. Special rule learning algorithms for subgroup discovery include Apriori-SD [9], CN2-SD [11] or SD [7]. These descriptive mining algorithms aim at finding characteristic rules as combinations of features with high

* This work has been partly supported by the PASCAL Network of Excellence.

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, algorithms for contrast set mining are STUCCO [2], and recently, an innovative approach was presented in the form of mining Emerging Patterns [5]. Basically, Emerging Patterns (EP) are sets of features in the data whose supports change significantly from one class to another.

Indeed, we can see all these described tasks on labeled data (learning classification rules, subgroup discovery, or contrast set mining) as 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. Searching for relevant descriptions for rule construction has been extensively addressed in descriptive data mining. A useful insight was provided by closure systems, 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 e.g. [12, 14]). 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.

To the best of our knowledge, the notion of closed sets has not yet been exported to labeled data, neither 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, thanks to the final structural properties and the feature filtering theory of [10], we formally justify that our obtained sets characterize the space of relevant combinations of features for discriminating the target class.

In practice, our approach to discovering closed sets from labeled data, (described in Sections 3 and 4) turns out to be very useful in many applications: from constructing rule based classifiers of increased accuracy, to finding most interpretative descriptions for subgroup discovery, among others. In particular, we have applied our proposal to reduce the number of EPs and to compress the number of essential rules (Section 6.1), and finally, to learn descriptions for subgroup discovery (Section 6.2).

2 Background

Features, used for describing the training examples, are logical variables representing attribute-value pairs (called items in association rule learning). 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 [13]. Patients are described by four attributes and each tuple is labeled with a class label: none, soft or hard. Here F is the set of all attribute-value pairs in the data, i.e. $F = \{\text{Age=young}, \ldots, \text{Tear=normal}\}$ (the class label is not included in F). This dataset is known to be complete and we will use it 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, labeled by $+$) and negatives (N, labeled by $-$), and $E = P \cup N$. Multi-class problems can be translated to a series of two-class learning problems. For instance, when the class soft of Table 1 is the target class (in Table 2), all examples labeled with none and hard are considered negative.

Table 1. The contact lens data set.

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 2. In this table we show the positive examples when the class soft is selected as the target class (thus, forming the set of examples in P). Instances of the classes none and hard will be considered non-target, thus treated together as negative data N.

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	+

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, i.e. $p \in P$ such that $X \subseteq p$, and false positives (FP) are those negative examples covered by the rule, i.e. $n \in N$ such that $X \subseteq n$; reciprocally, true negatives (TN) are those negative examples not covered by X.

2.1 Relevant Features for Discrimination

The main aim of the theory of relevancy, described in [10] is to reduce the hypothesis space by eliminating irrelevant features from F in the pre-processing phase. As proposed by the authors:

Definition 1 (Coverage of features). *Feature $f \in F$ covers another feature $f' \in F$ if and only if* $\mathrm{TP}(f') \subseteq \mathrm{TP}(f)$ *and* $\mathrm{TN}(f') \subseteq \mathrm{TN}(f)$ *(or equivalently,* $\mathrm{TP}(f') \subseteq \mathrm{TP}(f)$ *and* $\mathrm{FP}(f) \subseteq \mathrm{FP}(f')$*).*

Then, it is stated that $f' \in F$ is *relatively irrelevant* if there exists another feature 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 none class.

2.2 Closed Itemsets

From the practical point of view of data mining algorithms, closed itemsets are maximal sets among those other itemsets occurring in the same examples. Formally, let $\text{supp}(X)$ denote the number of examples where the itemset $X \subseteq F$ 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 from 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. Efficient algorithms for discovering closed itemsets over a certain minimum support threshold can be found in [6].

The foundations of closed itemsets are based on the definition of a closure operator on a lattice of items. The standard closure operator Γ for items acts as follows: given a binary relation, 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: (1) Monotonicity: $X \subseteq X' \Rightarrow \Gamma(X) \subseteq \Gamma(X')$; (2) Extensivity: $X \subseteq \Gamma(X)$; and (3) 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 extensitivity of Γ we always have $Y \subseteq X$ for Y generator of X. Closed sets of items can be graphically organized in a Hasse diagram, such as the one depicted in Figure 1 for the closed itemsets mined from data in Table 2.

3 Closed Sets on Target-class Data

Given an example set $E = P \cup N$ 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 of the rule; reciprocally, the support of X in N (non-target class examples) is the number of false positives 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. For the sake of simplicity and due to a lack of space, proofs are omitted, although a proof sketch will be provided to justify important results.

Proposition 1. *Let* $X, Y \subseteq F$ *such that* $Y \subseteq X$*, then* $\text{TP}(X) \subseteq \text{TP}(Y)$ *and* $\text{FP}(X) \subseteq \text{FP}(Y)$*.*

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

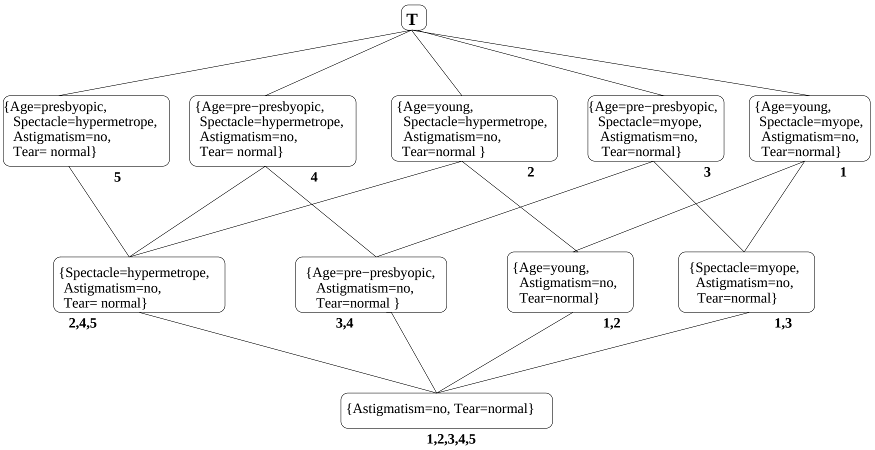
T
{Age=presbyopic, Spectacle=hypermetrope, Astigmatism=no, Tear= normal}
5
{Age=pre−presbyopic, Spectacle=hypermetrope, Astigmatism=no, Tear= normal}
4
{Age=young, Spectacle=hypermetrope, Astigmatism=no, Tear=normal }
2
{Age=pre−presbyopic, Spectacle=myope, Astigmatism=no, Tear=normal}
3
{Age=young,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

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

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

Indeed, this last result allows us to rewrite, within the data mining language, the definition of relevancy proposed in [10]: a feature f is *more relevant* than f' when $\text{supp}^+(\{f'\}) = \text{supp}^+(\{f, f'\})$ and $\text{supp}^-(\{f\}) = \text{supp}^-(\{f, f'\})$. 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 1 in terms of the formal closure operator Γ. 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 2. *A feature f is more relevant than f' iff $\Gamma^+(\{f'\}) = \Gamma^+(\{f, f'\})$ 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:

Definition 2 (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 2 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 2 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 $X \rightarrow +$ with $\Gamma^+(Y) = \Gamma^+(X \cup Y)$ and $\Gamma^-(X) = \Gamma^-(X \cup Y)$.

Moreover, from the structural properties of operator Γ and from Proposition 1, we can deduce that the semantics of relevant sets in Definition 2 is consistent:

Lemma 3. *A set of features $X \subseteq F$ is more relevant than set $Y \subseteq F$ (Definition 2) iff* $\mathrm{TP}(Y) \subseteq \mathrm{TP}(X)$ *and* $\mathrm{FP}(X) \subseteq \mathrm{FP}(Y)$.

The forward proof of this result is based on Proposition 1, which ensures that $\mathrm{TP}(X \cup Y) \subseteq \mathrm{TP}(Y)$ when X is more relevant than Y. Moreover, since we have that $\Gamma^+(Y) = \Gamma^+(X \cup Y)$ by hypothesis, we can derive after a couple of steps that $\mathrm{TP}(Y) \subseteq \mathrm{TP}(X)$. Similarly, we can conclude $\mathrm{FP}(X) \subseteq \mathrm{FP}(Y)$ for the negative part. The backward direction of Lemma 3 is also simple: if X and Y are two sets with $\mathrm{TP}(Y) \subseteq \mathrm{TP}(X)$ and $\mathrm{FP}(X) \subseteq \mathrm{FP}(Y)$, we can imply after some deduction steps that $\mathrm{supp}^+(Y) = \mathrm{supp}^+(X \cup Y)$ and $\mathrm{supp}^-(X) = \mathrm{supp}^-(X \cup Y)$. By construction of Γ this means $\Gamma^+(Y) = \Gamma^+(X \cup Y)$ and $\Gamma^-(X) = \Gamma^-(X \cup Y)$.

3.1 Closed Sets for Discrimination

Together with the result of Lemma 3, it can be shown that only closed itemsets mined in the set of positive examples suffice for discrimination.

Theorem 1. *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 2).* [4]

The proof of this theorem is mainly based on the construction of Γ: $\Gamma^+(Y) = X$ ensures that $|\mathrm{TP}(Y)| = |\mathrm{TP}(X)|$; but because $Y \subseteq X$ it must be true that $\mathrm{TP}(Y) = \mathrm{TP}(X)$. This, together with Proposition 1 leads to X being more relevant than Y according to Definition 2.

Typically, in approaches such as Apriori-C [8], Apriori-SD [9] and RLSD [15] frequent itemsets with very small minimal support constraint are initially mined and subsequently post-processed in order to find the most suitable rules for discrimination. The new result presented here states that not all frequent itemsets are necessary: as shown in Theorem 1 only the closed sets have the potential to be relevant.

[4] 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. 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.

To illustrate this result we use again the data in Table 2. There, we have $\Gamma^+(\{\text{Astigmatism=no}\} = \{\text{Astigmatism=no,Tear=normal}\}$. Thus, rule Astigmatism=no → + 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 1 simply states that 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. 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 of the lattice 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 1.

Remark 1. 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 1 with the lattice of Figure 1. The following 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 2. 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 1). 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.

Remark 2 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 to discard the closed set covering less true positives. Because of the monotonicity property of support, the smaller one will be the most relevant. From these two remarks we have:

Theorem 2. *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 2) iff $\Gamma^-(X) = \Gamma^-(X')$.*

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

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 }

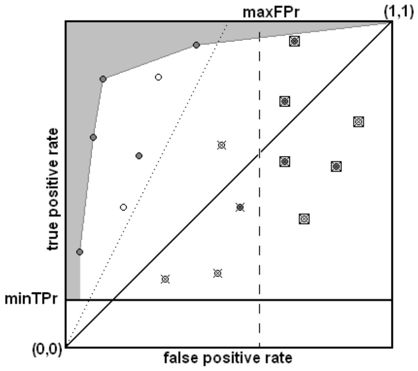

Fig. 2. The evaluation of relevant combinations of features in the ROC space.

The forward direction of this proof is simple: when $X \subset X'$ then $X' = X' \cup X$, so that $\Gamma^+(X') = \Gamma^+(X' \cup X)$ and $\Gamma^-(X) = \Gamma^-(X' \cup X)$ gets trivial, thus satisfying Definition 2. The backward direction of the proof is also based on rewriting $\Gamma^-(X) = \Gamma^-(X')$ as $\Gamma^-(X) = \Gamma^-(X' \cup X)$, and $\Gamma^+(X')$ as $\Gamma^+(X') = \Gamma^+(X' \cup X)$; therefore, we also satisfy conditions of the definition.

Thus, by Theorem 2 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 survives as a relevant set.

Finally, after Theorem 1 and Theorem 2, we can characterize the space of relevant sets of features for discriminating the selected target class as follows. These final sets can be directly interpreted as antecedents of discriminating rules.

Definition 3 (Space of relevant sets of features). *The space of relevant combinations of features for discriminating the target class is defined as those sets such 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 1 and 2, that by construction, any two sets in this space cover always a different set of positives and a different set of negatives.

The three closed sets forming the space of relevant sets of features for the class soft are shown in Table 4. It can be checked that CN2 algorithm [3] would output the rule whose antecedent corresponds to the closed set in the first entry of Table 4; Ripper [4], would obtain the most specific relevant rules, i.e. those corresponding to the three last entries from Table 4. Finally, other algorithms such as Apriori-C would also output rules whose antecedents are not relevant such e.g. Astigmatism=no → Lenses=soft.

4.1 Shortest Representation of a Relevant Set

Based on Theorem 1 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, any generator will be redundant w.r.t. its closure. 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 4. *Let $\Gamma^+(Y) = X$ and $Y \neq X$. We say that a generator Y is equivalent to its closure X if* $\mathrm{FP}(X) = \mathrm{FP}(Y)$.

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.

In terms of the closure operator of negatives, we have that Y is an equivalent generator of X iff $\Gamma^-(X) = \Gamma^-(Y)$.

5 Evaluation of Relevant Sets in the ROC space

The ROC space is a 2-dimensional space that shows classifier (rule/ruleset) performance in terms of its *false positive rate* (also called 'false alarm'), $FPr = \frac{FP}{TN+FP} = \frac{FP}{|N|}$ plotted on the X-axis, and *true positive rate* (also called 'sensitivity') $TPr = \frac{TP}{TP+FN} = \frac{TP}{|P|}$ 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.

The combinations of features of Definition 2 can be interpreted as condition parts of rules. Since they are induced with a minimal support constraint on the positives, they all lie above the minimum true positive rate constraint line (in Figure 2 denoted as minTPr). The rules removed by the relevancy filter are never those on the ROC convex hull (the empty circles are removed while the other remain). Furthermore, it can be trivially proved that we discover all the rules in the dataset on the ROC convex hull above the minimum true positives constraint (the full circles connected with a line). Therefore there are no rules outside the convex hull (grey area on the Figure 2 denotes an area without rules).

Sometimes an extra filtering criterion is required. In such cases we can imply a maximum FPr constraint covered by our relevant sets (in Figure 2 this constraint is represented by the dashed line, the rules eliminated by this constraint are shown in squares), or we can imply a minimum confidence constraint (represented by the dotted line, the rules eliminated by this constraint are crossed in Figure 2), or simply output rules on the convex hull, among others.

6 Experimental Evaluation

The presented theoretical study can be briefly summarized in the following steps:

- First, mining the set $S = \{X_1, \ldots, X_n\}$ of frequent closed itemsets from the target class (Theorem 1). This requires a minimum support constraint on true positives. Here we will use the efficient LCM algorithm [6].
- Second, reducing S to the space of relevant set of features by checking the coverage in the negatives (Theorem 2). Schematically, for any closed set $X_i \in S$, if there exists another closed set $X_j \in S$ s.t. both have same support in the negatives and $X_j \subset X_i$, then X_i is removed.

Finally, depending on the purpose of the application we can apply an extra filtering criterion, or compute minimal equivalent generators of the relevant sets as described above. For short, we will name this computing process as *RelSets* (for the Relevant Sets of features of Definition 2 we are discovering).

6.1 Emerging Patterns and Essential Rules on UCI data

Emerging Patterns (EP) [5] 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, i.e. $\frac{TPr}{FPr}$ 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 rate growth threshold (1.5 and infinite), and we compare this with the number of EPs by using the same rate growth constraint. Numerical attributes in the datasets are discretized when necessary by using four equal frequency intervals. Results are shown in the first part of Table 4.

Essential rules were proposed in [1] 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 s.t. 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 we do for these rules is shown in the second part of Table 4. Note that essential rules are not pruned by rate growth threshold, and this is why their number is usually higher than the number of Emerging Patterns.

6.2 Subgroup Discovery in New Application Domains

Subgroup discovery [11, 7] is a supervised descriptive induction task. The result of subgroup discovery is a set of subgroup descriptions (a rule set) that preferably has a low number of rules while each rule has high coverage and accuracy.

Table 4. Compression factor (CF% $= (1 - \frac{|Relsets|}{|EPs|}) \times 100$) of EPs and essential rules in UCI datasets. Note that we did not impose any minimum true positive threshold on any dataset, except for Lymphography and Crx, where all EPs, Relsets and essential rules were discovered with a 10% threshold on true positives. Also, note that in the second part of the table, essential rules and RelSets are not pruned by any rate growth threshold.

			EMERGING PATTERNS						ESSENTIAL RULES		
			Rate growth > 1.5			Rate growth ∞					
Dataset	**Class**	**Distrib. %**	**EPs**	**RelSets**	**CF%**	**EPs**	**RelSets**	**CF%**	**Essence**	**RelSets**	**CF%**
Lenses	soft	20.8	31	4	87.10	8	3	62.5	43	4	90.69
	hard	16.9	34	3	91.18	6	2	66.67	39	3	92.30
	none	62.5	50	12	76.00	42	4	9.52	89	19	78.65
Iris	setosa	33.3	83	16	80.72	71	7	90.14	76	20	73.68
	versicolor	33.3	134	40	70.15	63	10	84.13	111	41	63.06
	virginica	33.3	92	16	82.61	68	6	91.18	96	27	71.87
Breast-w	benign	65.5	6224	316	94.92	5764	141	97.55	3118	377	87.90
	malignant	34.5	3326	628	81.12	2813	356	87.34	2733	731	73.25
SAheart	0	34.3	4557	1897	58.37	2282	556	75.64	6358	4074	35.92
	1	65.7	9289	2824	69.60	3352	455	86.43	9622	4042	58
Balance-scale	B	7.8	271	75	72.32	49	49	0.00	415	147	88.67
	R	46	300	84	72.00	90	90	0.00	384	364	5.20
Yeast	MIT	16.4	3185	675	78.81	250	40	84.00	2258	1125	50.17
	CYT	31.2	3243	808	75.08	68	16	76.47	2399	1461	80.78
	ERL	0.3	1036	5	99.52	438	4	99.09	417	5	98.80
Monk-1	0	64.3	1131	828	26.79	321	18	94.39	1438	1135	21.07
	1	35.7	686	9	98.69	681	4	99.41	1477	363	75.42
Lymphography	metastases	54.72	36435	666	98.17	10970	90	99.18	1718	369	78.52
10% min supp.	malign	41.21	61130	740	98.79	19497	55	99.72	2407	476	80.22
Crx	+	44.5	3366	782	76.76	304	26	91.44	2345	1091	53.47
10% min supp.	−	55.5	3168	721	77.24	12	5	58.33	2336	1031	55.86

Table 5. Comparison of algorithms RelSets and SD on new subgroup discovery problems. Column RelSets-ROC shows the number of RelSets rules on the ROC convex hull.

Dataset	Class	Num. of rules			AUC		Time	
		RelSets	RelSets-ROC	SD	RelSets	SD	RelSets	SD
potato	sensitive	1	1	20	100%	100%	<1s	>1h
microarray	resistant	1	1	20	100%	91%	<1s	>1h
dribble	kick	110	7	20	89%	61%	<1s	3min
pass	pass	8	4	0	88%	0%	<1s	3min
	shoot	1	1	20	100%	100%	<1s	3min

The first experiment is performed on a real life potato microarray dataset with high dimensionality on the number of attributes. The goal is to distinguish between two different classes of resistance of four transgenic potato lines. After data preprocessing, we have only 12 examples (6 virus resistant and 6 virus sensitive examples) and 19,131 attributes. In Table 5 it can be seen how slowly the subgroup discovery algorithm SD performs and that RelSets performs better in terms of the area under the ROC curve (AUC) in the ROC space. Moreover, standard subgroup discovery algorithms present subgroups that are not as satisfactory to end users as subgroups found by RelSets.

The second experiment was performed on a real world strategy learning problem of robots playing soccer. In this dataset we have four classes: three classes represent successful moves made by the robots and the majority class (92%) when nothing interesting happens. We ran RelSets with a minimum true positive rate constraint of 20%. In Table 5 we show that we do not only outperform the algorithm SD in time, but also in quality (area under ROC convex hull).

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 dimension of examples. 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. Future work may adapt efficient algorithms of EPs in [5] for discovering relevant sets.

References

1. E. Baralis and S. Chiusano. Essential classification rule sets. *ACM Trans. Database Syst.*, 29(4):635–674, 2004.
2. Stephen D. Bay and Michael J. Pazzani. Detecting group differences: Mining contrast sets. *Data Min. Knowl. Discov.*, 5(3):213–246, 2001.
3. P. Clark and T. Niblett. The CN2 induction algorithm. *Mach. Learn.*, 3(4):261–283, 1989.
4. W.W. Cohen. Fast effective rule induction. In *Proc. 12th Int. Conf. on Machine Learning*, pages 115–123, 1995.
5. G. Dong and J. Li. Efficient mining of emerging patterns: discovering trends and differences. In *KDD '99: Proc. of the fifth ACM SIGKDD Int. Conf. on Knowledge discovery and data mining*, pages 43–52, 1999.
6. B. Goethals and M. Zaki. Advances in frequent itemset mining implementations: report on fimi'03. *SIGKDD Explor. Newsl.*, 6(1):109–117, 2004.
7. D. Gramberger and N. Lavrač. Expert-guided subgroup discovery: Methodology and application. *Journal of Artificial Intelligence Research*, 17:501–527, 2002.
8. V. Jovanoski and N. Lavrač. Classification rule learning with APRIORI-C. In *EPIA '01*, pages 44–51. Springer-Verlag, 2001.
9. B. Kavšek and N. Lavrač. APRIORI-SD: Adapting association rule learning to subgroup discovery. *Applied Artificial Intelligence*, To appear, 2006.
10. 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.
11. N. Lavrač, B. Kavšek, P. Flach, and L. Todorovski. Subgroup discovery with CN2-SD. *Journal of Machine Learning Research*, 5:153–188, 2004.
12. N. Pasquier, Y. Bastide, R. Taouil L., and Lakhal. Closed set based discovery of small covers for association rules. In *Proc. ICAD*, pages 361–381, 1999.
13. I.H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java implementations.* Morgan Kaufmann, 2005.
14. M. Zaki. Mining non-redundant association rules. *Data Mining and Knowledge Discovery: An Int. Journal*, 4(3):223–248, 2004.
15. J. Zhang, E. Bloedorn, L. Rosen, and D. Venese. Learning rules from highly unbalanced data sets. In *ICDM'04*, pages 571–574, 2004.