# Feature Construction Based on Closedness Properties Is Not That Simple

Dominique Gay[1], Nazha Selmaoui[1], and Jean-François Boulicaut[2]

[1] ERIM EA 3791, University of New Caledonia,
BP R4, F-98851 Nouméa, New Caledonia
{dominique.gay, nazha.selmaoui}@univ-nc.nc
[2] INSA-Lyon, LIRIS CNRS UMR5205
F-69621 Villeurbanne Cedex, France
jean-francois.boulicaut@insa-lyon.fr

**Abstract.** Feature construction has been studied extensively, including for 0/1 data samples. Given the recent breakthrough in closedness-related constraint-based mining, we are considering its impact on feature construction for classification tasks. We investigate the use of condensed representations of frequent itemsets (closure equivalence classes) as new features. These itemset types have been proposed to avoid set counting in difficult association rule mining tasks. However, our guess is that their intrinsic properties (say the maximality for the closed itemsets and the minimality for the $\delta$-free itemsets) might influence feature quality. Understanding this remains fairly open and we discuss these issues thanks to itemset properties on the one hand and an experimental validation on various data sets on the other hand.

## 1 Introduction

Feature construction is one of the major research topics for supporting classification tasks. Based on a set of original features, the idea is to compute new features that may better describe labeled samples such that the predictive accuracy of classifiers can be improved. When considering the case of 0/1 data (i.e., in most of the cases, collections of attribute-value pairs that are true or not within a sample), several authors have proposed to look at feature construction based on patterns that satisfy closedness-related constraints [1,2,3,4,5,6]. Using patterns that hold in 0/1 data as features (e.g., itemsets or association rules) is not new. Indeed, pioneering work on classification based on association rules [7] or emerging pattern discovery [8,9] have given rise to many proposals. Descriptive pattern discovery from unlabeled 0/1 data has been studied extensively during the last decade: many algorithms have been designed to compute every set pattern that satisfies a given constraint (e.g., a conjunction of constraints whose one conjunct is a minimal frequency constraint). One breakthrough into the computational complexity of such mining tasks has been obtained thanks to condensed

T. Washio et al. (Eds.): PAKDD 2008, LNAI 5012, pp. 112–123, 2008.

representations for frequent itemsets, i.e., rather small collections of patterns from which one can infer the frequency of many sets instead of counting for it (see [10] for a survey). In this paper, we consider closure equivalence classes, i.e., frequent closed sets and their generators [11]. Furthermore, when considering the $\delta$-free itemsets with $\delta > 0$ [12,13], we can consider a "near equivalence" perspective and thus, roughly speaking, the concept of almost-closed itemsets. We want to contribute to difficult classification tasks by using a method based on: (1) the efficient extraction of set patterns that satisfy given constraints, (2) the encoding of the original data into a new data set by using extracted patterns as new features. Clearly, one of the technical difficulties is to discuss the impact of the intrinsic properties of these patterns (i.e., closedness-related properties) on a classification process.

Our work is related to pattern-based classification. Since [7], various authors have considered the use of association rules. These proposals are based on a pruned set of extracted rules built w.r.t. support and confidence ranking. Differences between these methods mainly come from the way they use the selected set of rules when an unseen example $x$ is coming. For example, `CBA` [7] ranks the rules and it uses the best one to label $x$. Other algorithms choose the class that maximizes a defined score (`CMAR` [14] uses *combined effect* of subsets of rules when `CPAR` [15] uses *average expected accuracy* of the best $k$ rules). Also, starting from ideas for class characterization [16], [17] is an in-depth formalization of all these approaches. Another related research stream concerns emerging patterns [18]. These patterns are frequent in samples of a given class and infrequent for samples from the other classes. Several algorithms have exploited this for feature construction. Some of them select essential ones (`CAEP` classifier [8]) or the most expressive ones (`JEPs` classifier [9]). Then, an incoming example is labeled with the class $c$ which maximizes scores based on these sets. Moreover, a few researchers have considered condensed representations of frequent sets for feature construction. Garriga et al. [3] have proposed to characterize a target class with a collection of relevant closed itemsets. Li et al. [1] invoke `MDL` principle and suggest that free itemsets might be better than closed ones. However, classification experimental results to support such a claim are still lacking. It turns out that the rules studied in [17] are based on 0-free sets such that a minimal body property holds. The relevancy of such a minimality property is also discussed in terms of "near equivalence" in [19]. In [2], we have considered preliminary results on feature construction based on $\delta$-freeness [12,13]. Feature construction approaches based on closedness properties differ in two main aspects: *(i)* mining can be performed on the whole database or per class, and *(ii)* we can mine with or without the class labels. The pros and cons of these alternatives are discussed in this paper.

In Section 2, we provide more details on state-of-the-art approaches before introducing our feature construction method. Section 3 reports on our experimental results for UCI data sets [20] and a real-world medical database. Section 4 concludes.

## 2 Feature Construction Using Closure Equivalence Classes

A *binary database* $r$ is defined as a binary relation $(\mathcal{T}, \mathcal{I}, R)$ where $\mathcal{T}$ is a set of objects (or transactions), $\mathcal{I}$ is a set of attributes (or items) and $R \subseteq \mathcal{T} \times \mathcal{I}$. The *frequency* of an itemset $I \subseteq \mathcal{I}$ in $r$ is $freq(I, r) = |Objects(I, r)|$ where $Objects(I, r) = \{t \in \mathcal{T} \mid \forall i \in I, (t, i) \in R\}$. Let $\gamma$ be an integer, an itemset $I$ is said to be $\gamma$-*frequent* if $freq(I, r) \geq \gamma$.

Considering that *"what is frequent may be interesting"* is intuitive, Cheng et al. [4] brought some evidence to support such a claim and they have linked frequency with other interestingness measures such as Information Gain and Fischer score. Since the number of frequent itemsets can be huge in dense databases, it is now common to use condensed representations (e.g., free itemsets, closed ones, non derivable itemsets [10]) to save space and time during the frequent itemset mining task and to avoid some redundancy.

**Definition 1 (Closed itemset).** *An itemset $I$ is a* closed itemset *in $r$ iff there is no superset of $I$ with the same frequency than $I$ in $r$, i.e., $\nexists I' \supset I$ s.t. $freq(I', r) = freq(I, r)$. Another definition exploits the closure operation $cl : \mathcal{P}(\mathcal{I}) \rightarrow \mathcal{P}(\mathcal{I})$. Assume that Items is the dual operator for Objects: given $T \subseteq \mathcal{T}$, $Items(T, r) = \{i \in \mathcal{I} \mid \forall t \in T, (t, i) \in R\}$, and assume $cl(I, r) \equiv Items(Objects(I, r), r)$: the itemset $I$ is a closed itemset in $r$ iff $I = cl(I, r)$.*

Since [11], it is common to formalize the fact that many itemsets have the same closure by means of *closure equivalence relation*.

**Definition 2 (Closure equivalence).** *Two itemsets $I$ and $J$ are said to be* equivalent *in $r$ (denoted $I \sim_{cl} J$) iff $cl(I, r) = cl(J, r)$. Thus, a* closure equivalence class *(CEC) is made of itemsets that have the same closure, i.e., they are all supported by the same set of objects ($Objects(I, r) = Objects(J, r)$).*

Each CEC contains exactly one maximal itemset (w.r.t. set inclusion) which is a closed itemset. It may contain several minimal itemsets which are 0-free itemsets according to the terminology in [12] (also called key patterns in [11]).

*Example 1.* Considering Tab. 1, we have $r = (\mathcal{T}, \mathcal{I}, R)$, $\mathcal{T} = \{t_1, \ldots, t_6\}$, and $\mathcal{I} = \{A, B, C, D, c_1, c_2\}$, $c_1$ and $c_2$ being the class labels. For a frequency threshold $\gamma = 2$, itemsets $AB$ and $AC$ are $\gamma$-frequent. $ABCc_1$ is a $\gamma$-frequent closed itemset. Considering the equivalence class $\mathcal{C} = \{AB, AC, ABC, ABc_1, ACc_1, ABCc_1\}$, $AB$ and $AC$ are its minimal elements (i.e., they are 0-free itemsets) and $ABCc_1$ is the maximal element, i.e., one of the closed itemsets in this toy database.

### 2.1 Freeness or Closedness?

Two different approaches for feature construction based on condensed representations have been considered so far. In, e.g., [1,5], the authors mine free itemsets and closed itemsets (i.e., CECs) once the class attribute has been removed from

**Table 1.** A toy example of a binary labeled database

| $r$ | $A$ | $B$ | $C$ | $D$ | $c_1$ | $c_2$ |
|---|---|---|---|---|---|---|
| $t_1$ | 1 | 1 | 1 | 1 | 1 | 0 |
| $t_2$ | 1 | 1 | 1 | 0 | 1 | 0 |
| $t_3$ | 0 | 1 | 1 | 0 | 1 | 0 |
| $t_4$ | 1 | 0 | 0 | 1 | 1 | 0 |
| $t_5$ | 0 | 1 | 1 | 0 | 0 | 1 |
| $t_6$ | 0 | 1 | 0 | 1 | 0 | 1 |

the entire database. Other proposals, e.g., [3,4], consider (closed) itemset mining from samples of each class separately.

Looking at the first direction of research, we may consider that closed sets, because of their maximality, are good candidates for characterizing labeled data, but not necessarily suitable to predict classes for unseen samples. Moreover, thanks to their minimality, free itemsets might be better for predictive tasks. Due to closedness properties, every itemset of a given closure equivalence class $\mathcal{C}$ in $r$ covers exactly the same set of objects. Thus, free itemsets and their associated closed are equivalent w.r.t. interestingness measures based on frequencies. As a result, it is unclear whether choosing a free itemset or its closure to characterize a class is important or not. Let us now consider an incoming sample $x$ (test phase) that is exactly described by the itemset $Y$ (i.e., all its properties that are true are in $Y$). Furthermore, assume that we have $F \subseteq Y \subseteq cl(F, r)$ where $F$ is a free itemset from the closure equivalence class $\mathcal{C}_F$. Using free itemsets to label $x$ will not lead to the same decision than using closed itemsets. Indeed, $x \supseteq F$ and it satisfies rule $F \Rightarrow c$ while $x \not\supseteq cl(Y, r)$ and it does not satisfy rule $cl(F, r) \Rightarrow c$. Following that direction of work, Baralis et al. have proposed classification rules based on free itemsets [17].

On the other hand, for the "per-class" approach, let us consider w.l.o.g a two-class classification problem. In such a context, the equivalence between free itemsets and their associated closed ones is lost. The intuition is that, for a given free itemset $Y$ in $r_{c_1}$ –database restricted to samples of class $c_1$– and its closure $X = cl(Y, r_{c_1})$, $X$ is more relevant than $Y$ since $Objects(X, r_{c_1}) = Objects(Y, r_{c_1})$ and $Objects(X, r_{c_2}) \subseteq Objects(Y, r_{c_2})$. The closed itemsets (say $X = cl(X, r_{c_1})$) such that there is no other closed itemset (say $X' = cl(X', r)$) for which $cl(X, r_{c_2}) = cl(X', r_{c_2})$ are chosen as relevant itemsets to characterize $c_1$. In some cases, a free itemset $Y$ could be equivalent to its closure $X = cl(Y, r_{c_1})$, i.e., $Objects(X, r_{c_2}) = Objects(Y, r_{c_2})$. Here, for the same reason as above, a free itemset may be chosen instead of its closed counterpart. Note that relevancy of closed itemsets does not avoid conflicting rules, i.e., we can have two closed itemsets $X$ relevant for $c_1$ and $Y$ relevant for $c_2$ with $X \subseteq Y$.

Moreover, these approaches need for a post-processing of the extracted patterns. Indeed, we not only look for closedness-related properties but we have also to exploit interesting measures to keep only the ones that are discriminating. To avoid such a post-processing, we propose to use syntactic constraint (i.e., keeping

the class attribute during the mining phase) to mine class-discriminant closure equivalence classes.

### 2.2 What Is Interesting in Closure Equivalence Classes?

In Fig. 1, we report the different kinds of CECs that can be obtained when considering class attributes during the mining phase.

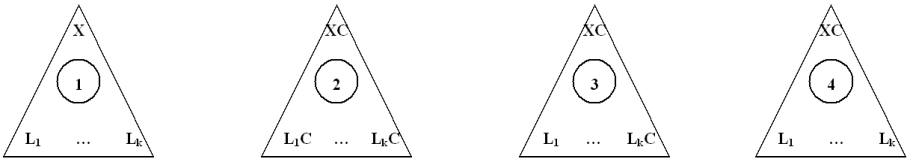


**Fig. 1.** Different types of CECs

These CECs have nice properties that are useful to our purpose: since association rules with a maximal confidence (no exception, also called hereafter exact rules) stand between a free itemset and its closure, we are interested in CECs whose closure contains a class attribute to characterize classes. Thus, we may neglect Case 1 in Fig. 1.

**Definition 3 (Association rule).** *Given* $r = \{\mathcal{T}, \mathcal{I}, R\}$*, an* association rule $\pi$ *on* $r$ *is an expression* $I \Rightarrow J$*, where* $I \subseteq \mathcal{I}$ *and* $J \subseteq \mathcal{I} \backslash I$*. The* frequency *of the rule* $\pi$ *is* $freq(I \cup J, r)$ *and its* confidence *is* $conf(\pi, r) = freq(I \cup J, r)/freq(I, r)$*. It provides a ratio about the numbers of exceptions for* $\pi$ *in* $r$*. When* $J$ *turns to be a single class attribute,* $\pi$ *is called a* classification rule*.*

From Case 3 (resp. Case 4), we can extract the exact classification rule $\pi_3 : L_1 \Rightarrow C$ (resp. the exact rules $\pi_{4_1} : L_1 \Rightarrow C \cdots \pi_{4_k} : L_k \Rightarrow C$). Note that if we are interested in exact rules only, we also neglect Case 2: $L_1C$ is a free itemset and it implies there is no exact rule $I \Rightarrow J$ such that $I \cup J \subseteq L_1C$. Thus, we are interested in CECs whose closed itemset contains a class attribute and whose free itemsets (at least one) do not contain a class attribute. This also leads to a closedness-related condensed representation of Jumping Emerging Patterns [21]. Unfortunately, in pattern-based classification (a fortiori in associative classification), for a given frequency threshold $\gamma$, mining exact rules is restrictive since they can be rare and the training database may not be covered by the rule set. In a relaxed setting, we consider association rules that enable exceptions.

**Definition 4 ($\delta$-strong rule, $\delta$-free itemset).** *Let* $\delta$ *be an integer. A* $\delta$-strong rule *is an association rule of the form* $I \Rightarrow^{\delta} J$ *which is violated in at most* $\delta$ *objects, and where* $I \subseteq \mathcal{I}$ *and* $J \subseteq \mathcal{I} \setminus I$*. An itemset* $I \subseteq \mathcal{I}$ *is a* $\delta$-free itemset *iff there is no* $\delta$*-strong rule which holds between its proper subsets. When* $\delta = 0$*,* $\delta$ *is omitted, and we talk about* strong rules*, and* free itemsets*.*

When the right-hand side is a single item $i$, saying that $I \Rightarrow^{\delta} i$ is a $\delta$-strong rule in $r$ means that $freq(I,r) - freq(I \cup \{i\}) \leq \delta$. When this item is a class attribute, a $\delta$-strong rule is called a $\delta$-strong classification rule [16].

The set of $\delta$-strong rules can be built from $\delta$-free itemsets and their $\delta$-closures.

**Definition 5 ($\delta$-closure).** *Let $\delta$ be an integer. The* $\delta$-closure *of an itemset $I$ on $r$ is $cl_\delta : \mathcal{P}(\mathcal{I}) \rightarrow \mathcal{P}(\mathcal{I})$ s.t. $cl_\delta(I,r) = \{i \in \mathcal{I} \mid freq(I,r) - freq(I \cup \{i\}) \leq \delta\}$. Once again, when $\delta = 0$, $cl_0(I,r) = \{i \in \mathcal{I} \mid freq(I,r) = freq(I \cup \{i\})\}$ and it corresponds to the closure operator that we already defined. We can also group itemsets by* $\delta$-closure equivalence classes*: two $\delta$-free itemsets $I$ and $J$ are $\delta$-equivalent ($I \sim_{cl_\delta} J$) if $cl_\delta(I,r) = cl_\delta(J,r)$.*

The intuition is that the $\delta$-closure of a set $I$ is the superset $X$ of $I$ such that every added attribute is almost always true for the objects which satisfy the properties from $I$: at most $\delta$ false values (or exceptions) are enabled. The computation of every frequent $\delta$-free set (i.e., sets which are both frequent and $\delta$-free) can be performed efficiently [13]. Given threshold values for $\gamma$ (frequency) and $\delta$ (freeness), the used `AC_like`[1] implementation outputs each $\delta$-free frequent itemset and its associated $\delta$-closure. Considering Table 1, a frequency threshold $\gamma = 3$ and a number of exceptions $\delta = 1$, itemset $C$ is a 3-frequent 1-free itemset ; items $B$ and $c_1$ belong to its $\delta$-closure and $\pi : C \Rightarrow^{\delta} c_1$ is a 1-strong classification rule.

### 2.3 Information and Equivalence Classes

We get more information from $\delta$-closure equivalence classes than with other approaches. Indeed, when considering contingency tables (See Tab. 2), for all the studied approaches, $f_{*1}$ and $f_{*0}$ are known (class distribution). However, if we consider the proposals from [3,4] based on frequent closed itemsets mined per class, we get directly the value $f_{11}$ (i.e., $freq(X \cup c, r)$) and the value for $f_{01}$ can be inferred. Closure equivalence classes in [5] only inform us on $f_{1*}$ (i.e., $freq(X,r)$) and $f_{0*}$. In our approach, when mining $\gamma$-frequent $\delta$-free itemsets whose closure contains a class attribute, $f_{1*} \geq \gamma$ and we have a lower bound $f_{11} \geq \gamma - \delta$ and an upper bound $f_{10} \leq \delta$ for frequencies on $X$. We can also infer other bounds for $f_{01}$ and $f_{00}$[2].

**Table 2.** Contingency table for a $\delta$-strong classification rule $X \Rightarrow^{\delta} c$

| $X \Rightarrow c$ | $c$ | $\bar{c}$ | $\Sigma$ |
|---|---|---|---|
| $X$ | $f_{11}$ | $f_{10}$ | $f_{1*}$ |
| $X$ | $f_{01}$ | $f_{00}$ | $f_{0*}$ |
| $\Sigma$ | $f_{*1}$ | $f_{*0}$ | $f_{**}$ |

Moreover, $\gamma$-frequent $\delta$-free itemsets, bodies of $\delta$-strong classification rules are known to have a minimal body property. Some constraints on $\gamma$ and $\delta$ can help

[1] `AC_like` implementation is available at http://liris.cnrs.fr/jeremy.besson/

[2] Note the confidence of a $\delta$-strong classification rule $\pi$ is $f_{11}/f_{1*} \geq 1 - (\delta/\gamma)$.

to avoid some of the classification conflicts announced at the end of Section 2.1. Indeed, [16] has shown that setting $\delta \in [0; \lfloor \gamma/2 \rfloor[$ ensures that we can not have two classification rules $\pi_1 : I \Rightarrow^{\delta} c_i$ and $\pi_2 : I \Rightarrow^{\delta} c_j$ with $i \neq j$ s.t. $I \subseteq J$. This constraint also enforces confidence to be greater than $\frac{1}{2}$. Furthermore, we know that we can produce $\delta$-strong classification rules that exhibit the discriminant power of emerging patterns if $\delta \in [0; \gamma \cdot (1 - \frac{|r_{c_i}|}{|r|})[$, $r_{c_i}$ being the database restricted to objects of the majority class $c_i$ [6]. One may say that the concept of $\gamma$-frequent $\delta$-free itemsets ($\delta \neq 0$) can be considered as an interestingness measures (function of $\gamma$ and $\delta$) for feature selection.

### 2.4 Towards a New Space of Descriptors

Once $\gamma$-frequent ($\delta$)-free itemsets have been mined, we can build a new representation of the original database using these new features. Each selected itemset $I$ will generate a new attribute $NewAtt_I$ in the new database. One may encode $NewAtt_I$ to a binary attribute, i.e., for a given object $t$, $NewAtt_I$ equals 1 if $I \subseteq Items(t, r)$ else 0. In a relaxed setting and noise-tolerant way, we propose to compute $Newatt_I$ as follows:

$$NewAtt_I(t) = \frac{|I \cap Items(t, r)|}{|I|}$$

This way, $I$ is a multivalued ordinal attribute. It is obvious that for an object $t$, $NewAtt_I(t) \in \{0, 1, \ldots, \frac{p-1}{p}, 1\}$ where $p = |I|$. Then, the value $NewAtt_I(t)$ is the proportion of items $i \in I$ that describe $t$. We think that multivalued encoding –followed by an entropy-based supervised discretization step[3]– should hold more information than binary encoding. Indeed, in the worst case, the split will take place between $\frac{p-1}{p}$ and 1, that is equivalent to binary case; in other better cases, split may take place between $\frac{j-1}{p}$ and $\frac{j}{p}$, $1 \leq j \leq p-1$ and this split leads to a better separation of data.

## 3 Experimental Validation

The frequency threshold $\gamma$ and the accepted number of exceptions $\delta$ are important parameters for our Feature Construction (FC) proposal. Let us discuss how to set up sensible values for them. Extreme values for $\gamma$ bring either (for lowest values) a huge amount of features –some of which are obviously irrelevant– or (for highest values) not enough features to correctly cover the training set. Furthermore, in both cases, these solutions are of limited interest in terms of Information Gain (see [4]). Then, $\delta$ varies from 0 to $\gamma \cdot (1 - \frac{|r_{c_i}|}{r})$ to capture discriminating power of emerging patterns. Once again, lowest values of $\delta$ lead to *strong* emerging patterns but a potentially low coverage proportion of data and features with high values of $\delta$ lacks of discriminating power.

[3] The best split between 2 values is recursively chosen until no more information is gained.

Intuitively, a high coverage proportion implies a relatively good representation of data. In Fig. 2, we plotted proportion of the database coverage w.r.t. $\delta$ for a given frequency threshold. Results for `breast`, `cleve`, `heart` and `hepatic` data (from UCI repository) are reported. We easily observe that coverage proportion grows as $\delta$ grows. Then, it reaches a saturation point for $\delta_0$ which is interesting: higher values of $\delta > \delta_0$ are less discriminant and lower values $\delta < \delta_0$ cover less objects. In our following experiments, we report (1) maximal accuracies over all $\gamma$ and $\delta$ values (denoted `Max`), and (2) average accuracies of all $\gamma$ values with $\delta = \delta_0$ (denoted `Av`).

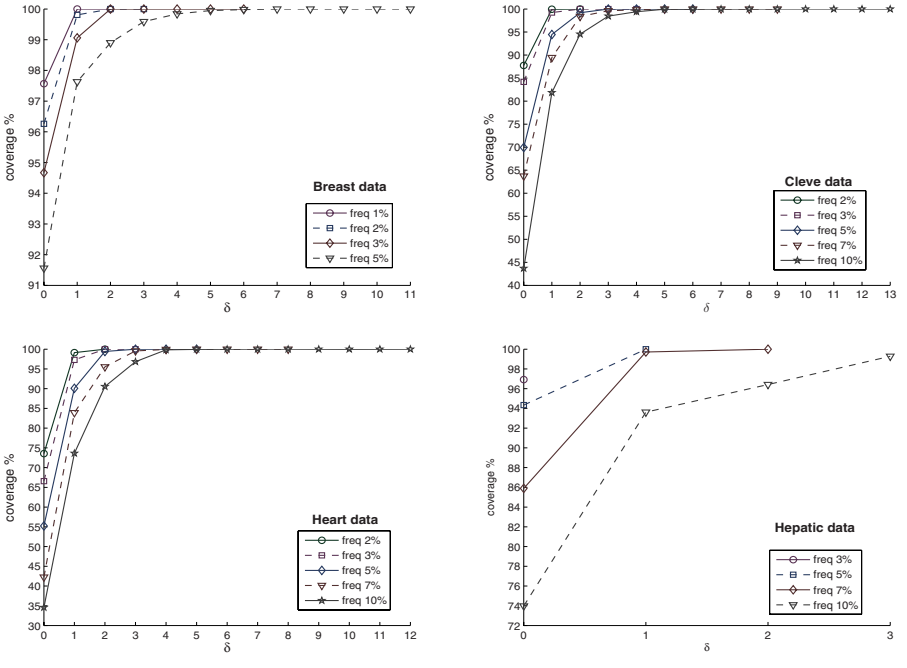


**Fig. 2.** Evolution of training database coverage proportion w.r.t. $\gamma$ and $\delta$

To validate our feature construction (`FC`) process, we used it on several data sets from UCI repository [20] and a real-world data set `meningitis`[4]. We have been using popular classification algorithms such as `NB` and `C4.5` on both the original data and the new representation based on extracted features. As a result, our main objective criterion is the accuracy of the obtained classifiers.

Notice that before performing feature construction, we translated all attributes into binary ones. While the translation of nominal attributes is straightforward, we decided to discretize continuous attributes with the entropy-based method by Fayyad et al. [22]. Discretizations and classifier constructions have been performed with `WEKA` [23] (10-folds stratified cross validation).

[4] `meningitis` concerns children hospitalized for acute bacterial or viral meningitis.

**Table 3.** Accuracy results improvement thanks to FC

| databases | NB | FC & NB (Av/Max) | C4.5 | FC & C4.5 (Av/Max) |
|---|---|---|---|---|
| breast | 95.99 | **97.32/97.54** | 94.56 | **96.12/96.43** |
| car | **85.53** | 81.95/84.64 | 92.36 | **98.49/99.13** |
| cleve | 83.5 | 83.35/**84.33** | 76.24 | **81.39/83.18** |
| crx | 77.68 | **85.91/86.46** | 86.09 | 83.95/**86.33** |
| diabetes | 75.91 | 75.56/**76.59** | 72.26 | **76.03/77.75** |
| heart | 84.07 | 83.62/**84.81** | 80 | **84.56/85.55** |
| hepatic | 83.22 | **84.09/84.67** | 81.93 | **85.29/86.83** |
| horse | 78.8 | **81.09/83.74** | 85.33 | 83.35/**85.40** |
| iris | **96** | 94.26/**96** | 96 | 94.26/**96.67** |
| labor | 94.74 | 93.5/**95.17** | 78.95 | **83.07/87.17** |
| lymph | **85.81** | 83.35/85.46 | 76.35 | **81.08/83.46** |
| meningitis | **95.74** | 93.24/93.64 | 94.83 | 92.54/**95.13** |
| sonar | 69.71 | **85.17/86.28** | 78.85 | **79.88/83.86** |
| vehicle | 45.03 | **59.72/62.88** | 71.04 | **70.70/71.28** |
| wine | 96.63 | 96.42/**97.83** | 94.38 | **95.57/96.29** |

**Table 4.** Our FC Feature Construction proposal vs. state-of-the-art approaches

| databases | BCEP | LB | FC&NB(Av/Max) | SJEP | CBA | CMAR | CPAR | FC&C4.5(Av/Max) |
|---|---|---|---|---|---|---|---|---|
| breast | – | 96.86 | 97.32/97.54 | **96.96** | 96.3 | 96.4 | 96.0 | 96.12/96.43 |
| car | – | – | 81.95/84.64 | – | 88.90 | – | 92.65 | **98.49/99.13** |
| cleve | 82.41 | 82.19 | **83.35/84.33** | 82.41 | 82.8 | 82.2 | 81.5 | 81.39/**83.18** |
| crx | – | – | 85.91/86.46 | **87.65** | 84.7 | 84.9 | 85.7 | 83.95/86.33 |
| diabetes | **76.8** | 76.69 | 75.56/76.59 | 76.18 | 74.5 | 75.8 | 75.1 | 76.03/**77.75** |
| heart | 81.85 | 82.22 | **83.62/84.81** | 82.96 | 81.9 | 82.2 | 82.6 | **84.56/85.55** |
| hepatic | – | 84.5 | 84.09/**84.67** | 83.33 | 81.8 | 80.5 | 79.4 | **85.29/86.83** |
| horse | – | – | 81.09/83.74 | 84.17 | 82.1 | 82.6 | 84.2 | 83.35/**85.40** |
| iris | – | – | 94.26/96 | – | 94.7 | 94.0 | 94.7 | 94.26/**96.67** |
| labor | – | – | 93.5/95.17 | 82 | 86.3 | **89.7** | 84.7 | 83.07/87.17 |
| lymph | 83.13 | 84.57 | 83.35/**85.46** | – | 77.8 | 83.1 | 82.3 | 81.08/**83.46** |
| meningitis | – | – | 93.24/93.64 | – | 91.79 | – | 91.52 | **92.54/95.13** |
| sonar | 78.4 | – | **85.17/86.28** | **85.10** | 77.5 | 79.4 | 79.3 | 79.88/83.86 |
| vehicle | 68.05 | **68.8** | 59.72/62.88 | **71.36** | 68.7 | 68.8 | 69.5 | 70.70/71.28 |
| wine | – | – | 96.42/97.83 | 95.63 | 95.0 | 95.0 | 95.5 | 95.57/**96.29** |

We report in Tab. 3 the accuracy results obtained on both the original data and its new representation. NB, C4.5 classifiers built on the new representation often perform better (i.e., it lead to higher accuracies) than respective NB and C4.5 classifiers built from the original data. One can see that we have often (12 times among 15) a combination of $\gamma$ and $\delta$ for which NB accuracies are improved by feature construction (column Max). And this is experimentally always the case for $C4.5$. Now considering average accuracies (column Av), improvement is still there w.r.t. C4.5 but it appears less obvious when using NB.

Then, we also compared our results with state-of-the-art classification techniques: FC & NB is compared with other bayesian approaches, LB [24] and BCEP [25]. When accessible, accuracies were reported from original papers within Tab. 4. Then, we have compared FC & C4.5 with other associative classification approaches, namely CBA [7], CMAR [14], CPAR [15], and an EPs-based classifier SJEP-classifier [26]. Accuracy results for associative classifiers are taken from [14]. Others results are taken from the published papers. FC allows to often achieve better accuracies than the state-of-the-art classifiers, e.g., FC & C4.5 wins 9 times over 15 against CPAR, 8 times over 13 against CMAR, 10 times over 15 against CBA when considering average accuracies (column Av). Considering optimal $\gamma$ and $\delta$ values (column Max), it wins 10 times over 15 (see bold faced results).

## 4 Conclusion

We study the use of closedness-related condensed representations for feature construction. We pointed out that differences about "freeness or closedness" within existing approaches come from the way that condensed representations are mined : with or without class label, per class or in the whole database. We proposed a systematic framework to construct features. Our new features are built from mined ($\delta$)-closure equivalence classes – more precisely from $\gamma$-frequent $\delta$-free itemsets whose $\delta$-closures involve a class attribute. Mining these types of itemsets differs from other approaches since (1) mined itemsets hold more information (such as emergence) and (2) there is no need for post-processing the set of features to select interesting features. We also proposed a new numeric encoding that is more suitable than binary encoding. Our FC process has been validated by means of an empirical evaluation. Using C4.5 and NB on new representations of various datasets, we demonstrated improvement compared with original data features. We have also shown comparable accuracy results w.r.t. efficient state-of-the-art classification techniques. We have now a better understanding of critical issues w.r.t. feature construction when considering closedness related properties. One perspective of this work is to consider our FC process in terms of constraints over sets of patterns and its recent formalization in [27].

**Acknowledgments.** The authors wish to thank B. Crémilleux for exciting discussions and the data set meningitis. They also thank J. Besson for technical support during this study. Finally, this work is partly funded by EU contract IST-FET IQ FP6-516169.

## References

1. Li, J., Li, H., Wong, L., Pei, J., Dong, G.: Minimum description length principle: generators are preferable to closed patterns. In: Proceedings AAAI 2006, pp. 409–415. AAAI Press, Menlo Park (2006)
2. Selmaoui, N., Leschi, C., Gay, D., Boulicaut, J.F.: Feature construction and delta-free sets in 0/1 samples. In: Todorovski, L., Lavrač, N., Jantke, K.P. (eds.) DS 2006. LNCS (LNAI), vol. 4265, pp. 363–367. Springer, Heidelberg (2006)

3. Garriga, G.C., Kralj, P., Lavrac, N.: Closed sets for labeled data. In: Fürnkranz, J., Scheffer, T., Spiliopoulou, M. (eds.) PKDD 2006. LNCS (LNAI), vol. 4213, pp. 163–174. Springer, Heidelberg (2006)
4. Cheng, H., Yan, X., Han, J., Hsu, C.W.: Discriminative frequent pattern analysis for effective classification. In: Proceedings IEEE ICDE 2007, pp. 716–725 (2007)
5. Li, J., Liu, G., Wong, L.: Mining statistically important equivalence classes and delta-discriminative emerging patterns. In: Proceedings ACM SIGKDD 2007, pp. 430–439 (2007)
6. Gay, D., Selmaoui, N., Boulicaut, J.F.: Pattern-based decision tree construction. In: Proceedings ICDIM 2007, pp. 291–296. IEEE Computer Society Press, Los Alamitos (2007)
7. Liu, B., Hsu, W., Ma, Y.: Integrating classification and association rule mining. In: Proceedings KDD 1998, pp. 80–86. AAAI Press, Menlo Park (1998)
8. Dong, G., Zhang, X., Wong, L., Li, J.: CAEP: Classification by aggregating emerging patterns. In: Arikawa, S., Furukawa, K. (eds.) DS 1999. LNCS (LNAI), vol. 1721, pp. 30–42. Springer, Heidelberg (1999)
9. Li, J., Dong, G., Ramamohanarao, K.: Making use of the most expressive jumping emerging patterns for classification. Knowledge and Information Systems 3, 131–145 (2001)
10. Calders, T., Rigotti, C., Boulicaut, J.F.: A survey on condensed representations for frequent sets. In: Boulicaut, J.-F., De Raedt, L., Mannila, H. (eds.) Constraint-Based Mining and Inductive Databases. LNCS (LNAI), vol. 3848, pp. 64–80. Springer, Heidelberg (2006)
11. Bastide, Y., Taouil, R., Pasquier, N., Stumme, G., Lakhal, L.: Mining frequent patterns with counting inference. SIGKDD Explorations 2, 66–75 (2000)
12. Boulicaut, J.F., Bykowski, A., Rigotti, C.: Approximation of frequency queries by means of free-sets. In: Zighed, A.D.A., Komorowski, J., Żytkow, J.M. (eds.) PKDD 2000. LNCS (LNAI), vol. 1910, pp. 75–85. Springer, Heidelberg (2000)
13. Boulicaut, J.F., Bykowski, A., Rigotti, C.: Free-sets: A condensed representation of boolean data for the approximation of frequency queries. Data Mining and Knowledge Discovery 7, 5–22 (2003)
14. Li, W., Han, J., Pei, J.: CMAR: Accurate and efficient classification based on multiple class-association rules. In: Proceedings IEEE ICDM 2001, pp. 369–376 (2001)
15. Yin, X., Han, J.: CPAR: Classification based on predictive association rules. In: Proceedings SIAM SDM 2003 (2003)
16. Boulicaut, J.F., Crémilleux, B.: Simplest rules characterizing classes generated by delta-free sets. In: Proceedings ES 2002, pp. 33–46. Springer, Heidelberg (2002)
17. Baralis, E., Chiusano, S.: Essential classification rule sets. ACM Trans. on Database Systems 29, 635–674 (2004)
18. Dong, G., Li, J.: Efficient mining of emerging patterns: discovering trends and differences. In: Proceedings ACM SIGKDD 1999, pp. 43–52 (1999)
19. Bayardo, R.: The hows, whys and whens of constraints in itemset and rule discovery. In: Boulicaut, J.-F., De Raedt, L., Mannila, H. (eds.) Constraint-Based Mining and Inductive Databases. LNCS (LNAI), vol. 3848, pp. 1–13. Springer, Heidelberg (2006)
20. Newman, D., Hettich, S., Blake, C., Merz, C.: UCI repository of machine learning databases (1998)
21. Soulet, A., Crémilleux, B., Rioult, F.: Condensed representation of emerging patterns. In: Dai, H., Srikant, R., Zhang, C. (eds.) PAKDD 2004. LNCS (LNAI), vol. 3056, pp. 127–132. Springer, Heidelberg (2004)

22. Fayyad, U.M., Irani, K.B.: Multi-interval discretization of continous-valued attributes for classification learning. In: Proceedings IJCAI 1993, pp. 1022–1027 (1993)
23. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)
24. Meretakis, D., Wuthrich, B.: Extending naïve bayes classifiers using long itemsets. In: Proceedings ACM SIGKDD 1999, pp. 165–174 (1999)
25. Fan, H., Ramamohanarao, K.: A bayesian approach to use emerging patterns for classification. In: Proceedings ADC 2003, pp. 39–48. Australian Computer Society, Inc. (2003)
26. Fan, H., Ramamohanarao, K.: Fast discovery and the generalization of strong jumping emerging patterns for building compact and accurate classifiers. IEEE Trans. on Knowledge and Data Engineering 18, 721–737 (2006)
27. De Raedt, L., Zimmermann, A.: Constraint-based pattern set mining. In: Proceedings SIAM SDM 2007 (2007)