# $\delta$-strong classification rules for characterizing chemical carcinogens

Jean-François Boulicaut[1], Bruno Crémilleux[1,2]
[1] Laboratoire d'Ingénierie des Systèmes d'Information
INSA Lyon, Bâtiment Blaise Pascal
F-69621 Villeurbanne Cedex, France
Jean-Francois.Boulicaut@insa-lyon.fr
[2] Université de Caen - GREYC - CNRS UMR 6072
Campus Côte de Nacre
F-14032 Caen Cedex, France
Bruno.Cremilleux@info.unicaen.fr

**Abstract**

This is a contribution to the PKDD'01 Predictive Toxicology Discovery Challenge with a classification point of view. The goals of this challenge are to obtain models that predict carcinogenicity of chemicals using information related to chemical structure only. Our aim is to show the potential impact of $\delta$-strong classification rules in such a domain. This technique relies on recent results in the association rule mining area in order to mine a condensed representation of potentially interesting rules for the characterization of classes.

## 1 Introduction

One popular data mining technique concerns knowledge discovery from frequent association rules. This kind of process has been studied a lot of times since the definition of the mining task in [1].

Association rules can tell something like "When properties $A_1$ and $A_2$ are true within the data, then property $A_3$ is often true". We provide a simple formalization of this task in Section 2.1.

Classification is another popular data mining technique. Starting from a collection of examples associated with a known class value, classification concerns the design of models that enable to predict accurate class values for unseen examples. The set of examples for which the class value is given is the so-called learning set. Various knowledge representation formalisms have been used for building the so-called classifiers. Classification rules are quite popular for that purpose and the literature is abundant (see for example [10, 13]). In that context, a classification rule is a rule that concludes on one class value.

The key point of this research is to use efficient association rule mining techniques in order to identify classification rules and, as a first step, provide a symbolic characterization of the classes.

The starting idea is quite simple. Roughly speaking, we first mine association rules from the learning set and then we select in such a collection the rules that conclude on a class value. In [8], Freitas has shown the limitations of such a naive approach, emphasizing the differences between classification rules and association rules. Other researchers have studied the selection of classification rules from a collection of association rules, e.g., [3, 4, 11]. Interestingly, in these proposals, the identification of the classification rules is performed mainly as a post-processing step on standard association rules.

We use recent results in the association rule mining area in order to mine efficiently a condensed representation of potentially interesting rules, the so-called $\delta$-strong rules, for the characterization of classes. It is possible to work on difficult contexts such as large, dense and highly-correlated learning sets. Next, we show that it is rather easy to process the discovered $\delta$-strong rules in order to get a cover of the classification rules that characterize the classes. We see in Section 2.2 that our aim is to produce the simplest rules with respect to (w.r.t.) their left-hand sides and the technique provides *every* simplest classification rule: given a classification rule, one wants that any own and proper subset of its left-hand side does not enable to conclude on the same class value.

Our aim is to show the potential impact of $\delta$-strong classification rules in domains like predictive toxicology. This is a preliminary work and more investigations have to be done to validate the interest of this approach on such domains. Section 2.2 introduces association rule mining and the concept of $\delta$-strong classification rule. Section 3 presents the data preparation stage and Section 4 gives the results of this application.

# 2 $\delta$-strong rules to characterize classes

## 2.1 Association rule mining

Let us provide a simple formalization of $\delta$-strong rule mining task.

**Definition 1 (item, itemset, example)** *Assume* $\mathbf{R} = \{A_1, \ldots, A_n\}$*, is a schema of boolean attributes. One attribute from* $\mathbf{R}$ *is called an* item *and a subset of* $\mathbf{R}$ *is called an* itemset*.* $\mathbf{r}$*, an instance of* $\mathbf{R}$*, is a multi-set of* examples*. Thus,* $\mathbf{r}$ *can be considered as a boolean matrix.*

In the context of this challenge, $\mathbf{R}$ is made up by the chemical descriptors. For instance, with the molecular substructures represented with fingerprints, an attribute is a fragment identifying an atom pair with a distance between atoms of any length desired, an atom sequence of any length desired, etc. We will see in Section 3 that here, the experimental data concern 6,150 fragments. This is obviously a difficult mining context.

**Definition 2 (association rule)** *Given* $\mathbf{r}$*, an instance of* $\mathbf{R}$*, an* association rule *on* $\mathbf{r}$ *is an expression* $X \Rightarrow B$*, where the itemset* $X \subseteq \mathbf{R}$ *and* $B \in \mathbf{R} \setminus X$*.*

The intuitive meaning of a potentially interesting association rule $X \Rightarrow B$ is that when an example contains true, i.e., 1, for each item of $X$, then this example tends to contain true for item $B$ too. This semantics is captured by the classical measures of *frequency* and *confidence* [1].

**Definition 3 (frequency, confidence, support)** *Given* $W \subseteq \mathbf{R}$*,* $\mathcal{F}(W, \mathbf{r})$ *(or frequency of* $W$*) is the number of examples in* $\mathbf{r}$ *that contain* 1 *for each item in* $W$*. The frequency of* $X \Rightarrow B$ *in* $\mathbf{r}$ *is defined as* $\mathcal{F}(X \cup \{B\}, \mathbf{r})$ *and its confidence is* $\mathcal{F}(X \cup \{B\}, \mathbf{r})/\mathcal{F}(X, \mathbf{r})$*. Notice that in the theoretical background of this paper, we use an absolute frequency (a number of examples* $\leq |\mathbf{r}|$*) instead of the relative frequency* $\mathcal{F}(X \cup \{B\}, \mathbf{r})/|\mathbf{r}|$ *in* $[0, 1]$*. Frequency is also called* support*.*

The standard association rule mining task concerns the discovery of *every* rule having its frequency and its confidence higher than user-specified thresholds. In other words, one wants rules that are frequent "enough" and valid. The main algorithmic issue concerns the computation of every frequent set.

**Definition 4 (frequent itemset)** *Given* $\gamma$ *a frequency threshold* $\leq |\mathbf{r}|$*. An itemset* $X$ *is said* frequent *or* $\gamma$*-frequent if* $\mathcal{F}(X, \mathbf{r}) \geq \gamma$*.*

Algorithmic complexity of frequent itemset discovery is exponential with the number of attributes. Many research works (e.g. [2, 15, 12, 6]) concern the practical contexts for which such a discovery remains tractable, even though a trade-off is needed with the exact knowledge of the frequencies (a fundamental issue for classification, see Section 2.2) and/or the completeness of the extractions.

## 2.2 $\delta$-strong rules

For class characterization, interesting association rules must conclude on class values with a rather high confidence. $\delta$-strong rules introduced in [6] satisfy such a constraint.

**Definition 5 ($\delta$-strong rules)** *Given* $\mathbf{R}$*, a matrix* $\mathbf{r}$*, a frequency threshold* $\gamma$*, and an integer* $\delta$*, a* $\delta$-strong rule *on* $\mathbf{r}$ *is an association rule* $X \Rightarrow B$*, where* $\mathcal{F}(X \cup \{B\}, \mathbf{r}) \geq \gamma$*,* $\mathcal{F}(X, \mathbf{r}) - \mathcal{F}(X \cup \{B\}, \mathbf{r}) \leq \delta$*,* $X \subseteq \mathbf{R}$*, and* $B \in \mathbf{R} \setminus X$*.*

A $\delta$-strong rule is violated by at most $\delta$ examples. In other words, its confidence is at least equal to $1 - (\delta/\gamma)$. When using $\delta$-strong rules, we assume that $\delta$ is rather small.

From a technical perspective, $\delta$-strong rules can be built from $\delta$-free itemsets that will constitute their left-hand sides [6]. Due to the space limitation, we do not explain in details what is a $\delta$-free itemset and how they are extracted (see [6, 7] for details). We just provide an intuition for the concept of $\delta$-free itemset. It is clearly related to the concepts of closed itemsets in [12] and almost-closures in [5]. An itemset $X$ is called $\delta$-free if there is no a $\delta$-strong rule that holds between two of its own and proper subsets. The case $\delta = 0$ is important: no rule with confidence equal to 1 holds between proper subsets of $X$. In fact, frequent closed itemsets are the closures of 0-free sets: the closure of an itemset $X$ is the largest superset of $X$ (w.r.t. set inclusion) that has the same frequency as $X$. When $\delta > 0$, we are interested in the almost-closures of a frequent $\delta$-free set $X$: $B$ belongs to the almost-closure of $X$ if $\mathcal{F}(X, \mathbf{r}) - \mathcal{F}(X \cup \{B\}, \mathbf{r}) \leq \delta$. It is easy to provide $\delta$-strong rules from a $\gamma$-frequent $\delta$-free set and its almost-closure. We use the research prototype `ac-miner-12` [6] for that purpose[1]. Given thresholds $\gamma$ and $\delta$, it provides the collection of frequent $\delta$-free itemsets, their frequencies and the attributes in their almost-closures. It has been shown efficient even in the case of dense and higly-correlated data, i.e., in practical applications where `apriori`-like algorithms clearly fail.

Let us now indicate the property of minimal body which allows us to build $\delta$-strong rules with a minimal left-hand side.

**Property 1 (minimal body)** *If $X$ is a $\delta$-free itemset and $X \Rightarrow B$ $[\delta]$ is a $\delta$-strong rule with exactly $\delta$ exceptions, then $X$ is the minimal set of items from which we can conclude on $B$ with at most $\delta$ exceptions.*

It means that if $X \Rightarrow B$ $[\delta_1]$ is a $\delta$-strong rule with $\delta_1$ exceptions, there is no itemset $Y$, $Y \subset X$, such that $Y \Rightarrow B$ $[\delta_2]$ is a $\delta$-strong rule with $\delta_2 \leq \delta_1$. In other words, it is possible to get the simplest rules, i.e., a cover of $\delta$-strong rules. We argue that it is a fundamental issue for classification. Not only it prevents from over-fitting [14] but also it makes the classification of an example easier to explain. Experts are generally interested in an explicit characterization of the concepts that support classification. It provides a feedback on the application domain expertise that can be reused for further analysis.

Let us consider a classification task where the class can take $k$ values. Assuming $C_1, \ldots, C_k$ are the $k$ items that denote class values.

**Definition 6 ($\delta$-strong classification rule)** *A* $\delta$-strong classification rule *is a $\delta$-strong rule that concludes on one class value (i.e., $C_i$).*

It is shown in [7] that if $\delta < \gamma$, then some rule conflicts are avoided. For instance, if there is the $\delta$-strong classification rule $R_1 : X \Rightarrow C_i$, then a $\delta$-strong classification rule $R_2 : X \Rightarrow C_j$ with $i \neq j$ cannot appear. Furthermore, if $\delta < \gamma$, there is no $\delta$-strong classification rule $R_3 : X \cup Y \Rightarrow C_j$ with $i \neq j$. As this sufficient condition on $\gamma$ and $\delta$ is quite reasonable in practice, our experiments have been done under this assumption.

# 3 Data preparation

Chemicals are available with seven sets of descriptors (see `http://www.informatik.uni-freiburg.de/~ml/ptc/`) according to several paradigms: propositional

[1] `ac-miner-12` has been implemented by A. Bykowski at INSA Lyon

logic, multi-relational representation where atoms and bonds are encoded as t-uples with their own attributes to denote their properties, and predicate logic. Chemical characteristics are functional groups, atomic and bond properties, molecular substructures represented with fingerprints. Several sets of descriptors require a chemical knowledge to be transformed into a suitable format to extract $\delta$-strong classification rules. So, we considered Barnard Chemical Information (BCI) fingerprints because data can be used without a sound chemical knowledge (among other things, there are no quantitative attributes which would have to be discretized). Each molecule is represented by a fingerprint made of 6150 fragments (each fragment is encoded as 0 if absent, 1 otherwise). There are in average almost 277 fragments present for each chemical. A fingerprint captures the information from the raw data, i.e., the initial 57240 raw features. The data concern 417 molecules. It has been identified as a difficult classification task (the correct classification score for experts in the domain ranges from 28% to 78% [9]). Let us notice also that it is a quite hard context for association rule mining since we have few examples and a huge number of boolean attributes.

As required by the purpose of this challenge, we split data into four files according to the populations (male rats (`MR`), female rats (`FR`), male mice (`MM`), female mice (`FM`)). We joined the class contained in the file `corrected_results.txt` (see `http://www.informatik.uni-freiburg.de/~ml/ptc/`). Class values are a mixture between the US National Toxicology Program (NTP) classification, i.e., 5 values about the carcinogenic activity[2] and earlier designations[3]. Table 1 gives class value frequencies w.r.t. the populations.

| File | P | CE | SE | EE | E | NE | N | IS | Total |
|---|---|---|---|---|---|---|---|---|---|
| MR | 70 | 48 | 34 | 23 | 21 | 66 | 126 | 12 | 400 |
| FR | 63 | 41 | 17 | 24 | 15 | 89 | 141 | 10 | 400 |
| MM | 69 | 43 | 17 | 19 | 22 | 84 | 123 | 15 | 392 |
| FM | 80 | 46 | 17 | 10 | 12 | 82 | 124 | 8 | 379 |

Table 1: Class value frequencies w.r.t. the populations

Given that the challenge requires a predicting outcome coded as `POS` or `NEG` and that there is no official rule to move from the previous classifications to this binary one, we decided to recode `CE` and `SE` into `POS` and `NE` in `NEG`. Furthermore, as we know that all equivocal and inadequate studies were removed from the test set used in this challenge, we decided not to recode instances with class values `EE`, `E` and `IS`. Finally, we got the four following sets (see Table 2).

[2] `CE`: Clear Evidence ; `SE`: Some Evidence ; `EE`: Equivocal Evidence ; `NE`: No Evidence ; `IS`: Inadequate Study

[3] `P`: Positive ; `E`: Equivocal ; `N`: Negative

| File | POS | NEG | Total |
|---|---|---|---|
| MR | 152 | 192 | 344 |
| FR | 121 | 230 | 351 |
| MM | 129 | 207 | 336 |
| FM | 143 | 206 | 349 |

Table 2: Class value frequencies w.r.t. the populations

# 4 Results and discussion

Data have to be translated into a binary format to be processed by `ac-miner-12`. This is automated by producing a binary item for each pair of attribute/value. From a technical point of view, that means 12,302 items: two items (`present` or `true`, `false`) are derived from each fragment and two items are used to denote the class (`NEG`, `POS`). In order to minimize the computational cost, we restricted the search to $\delta$-strong classification rules linking fragments that are present. Let us remark that there is no missing value in these data.

The prototype `ac-miner-12` is implemented in C++. We used a PC with 768 MB of memory and a 500 MHz Pentium III processor under Linux operating system.

We focus now on `MR` data. This file gathers 344 chemicals, 152 (44%) are classified as `POS` and 192 (56%) as `NEG`. For different values of $\delta$ and $\gamma$, Table 3 gives the extraction time, the number of $\delta$-free itemsets and almost-closures that contain a class value. This last number can be seen as the number of potential $\delta$-strong classification rules (i.e., with any support and confidence values). In this first experiment, the training was done with 9/10 of data (i.e., 310 examples), and we have $\delta < \gamma$. Class has the same frequency distribution in each file and in the whole data.

| $\gamma/\lvert\mathbf{r}\rvert$ | $\delta$ | Time (sec.) | No. of $\delta$-free sets | No. of almost-closures with a class value |
|---|---|---|---|---|
| 0.15 | 15 | intractable | - | - |
| 0.15 | 17 | 3814 | 24671 | 2835 |
| 0.15 | 20 | 1563 | 17173 | 4529 |
| 0.20 | 10 | 3300 | 26377 | 0 |
| 0.20 | 15 | 850 | 12071 | 8 |
| 0.20 | 20 | 323 | 7109 | 305 |
| 0.30 | 10 | 69 | 3473 | 0 |
| 0.40 | 0 | intractable | - | - |
| 0.40 | 10 | 36 | 922 | 0 |
| 0.50 | 0 | 201 | 56775 | 0 |

Table 3: Time, $\delta$-free itemsets and almost-closures w.r.t. $\delta$ and $\gamma$

When the extraction turns to be intractable, it comes from an excessive memory requirement because of the management of huge collections of candidates (itemsets that are candidates for frequent $\delta$-freeness).

On these data where there is no strong association between the class value and the items (i.e., the fragments), given that the extracted almost-closures are the most general, the frequency of the classification rules tends towards $\gamma - \delta$. The number of $\delta$-strong classification rules depends on the values for the thresholds $\gamma$ and $\delta$. Also, we check experimentally that the more we increase the value of $\delta$, the more we can have tractable extractions for lower frequency thresholds. Let us note that with $\delta = 0$, there is no classification rule for the frequency threshold we can use. It illustrates the added-value of the relaxed constraint on $\delta$.

Almost every classification rules we got conclude on `NEG`. With $\gamma/|\mathbf{r}| = 0.20$ and $\delta = 15$ and with $\gamma/|\mathbf{r}| = 0.20$ and $\delta = 20$, all classification rules conclude on `NEG`. With $\gamma/|\mathbf{r}| = 0.15$ and $\delta = 17$, there are 2828 rules concluding on `NEG` and only 7 on `POS` and with $\gamma/|\mathbf{r}| = 0.15$ and $\delta = 20$, we get 4443 rules concluding on `NEG` and 86 on `POS`. Examples of $\delta$-strong classification rules are given at the end of this section.

Collections of discovered $\delta$-strong classification rules can be used to predict chemical carcinogens. In this experiment, we give classification results (see Table 4) achieved on the files according to the populations of the four rodents (cf. Section 3). We used every rule of the cover. When there was a conflict (several rules with different conclusions were triggered from a same chemical), a score incorporating the support and the confidence of each rule was computed and the class value having the best score was predicted. To better evaluate results, for each experiment, files have been split into a training file (4/5 of data) and a test file (1/5 of data). Class has the same frequency distribution in each file and in the whole data. On each file, we ran `ac-miner-12` with the lowest value of $\gamma$ and a sensible value for $\delta$ (we experimented also with $\gamma/|\mathbf{r}| = 0.11$ but it led to choose $\delta$ around 30 to ensure the extraction tractability).

| File | $\gamma/\|\mathbf{r}\|$ | $\delta$ | No. of rules on `POS` | No. of rules on `NEG` | Well-classified (%) |
|---|---|---|---|---|---|
| `MR` | 0.15 | 17 | 17 | 4914 | 55.88 |
| `FR` | 0.15 | 17 | 0 | 9470 | 68.66 |
| `MM` | 0.15 | 15 | 1 | 5723 | 63.49 |
| `FM` | 0.15 | 15 | 0 | 11369 | 60.61 |

Table 4: Classification results with all rules

As Table 4 shows, almost all extracted rules conclude on `NEG`. Even with `MR` and `MM` (where there are very few rules on `POS`) no chemical is classified as `POS`. 4 chemicals (1 in `MR`, 1 in `MM` and 2 in `FM`) are not classified (i.e., no rule is triggered). In fact, almost all `POS` chemicals are classified as `NEG` with this strategy (otherwise, they are not classified) and almost all `NEG` chemicals are classified as `NEG` (so, for each file, the number of well-classified chemicals is similar to the number of NEG chemicals). Let us recall that the prediction is here based on a cover of the classification rules that characterizes the classes. This cover includes rules with low support and/or confidence. Such rules with a poor quality may introduce errors. The design of a classifier stemming from the classification rule cover still needs some research. We are giving now a preliminary approach.

For each cover of rules, we computed for each rule a score (denoted $\Delta$) which is the difference between the well-classified and the miss-classified examples of the test file. For each cover, we sorted its rules w.r.t. $\Delta$ and we defined a family of nested sets of rules: for a value $\Delta_1$ of $\Delta$, we kept rules having $\Delta$ higher or equal to $\Delta_1$ (all rules belonging to the set defined by $\Delta_1$ belong to sets defined by $\Delta_2$ with $\Delta_2 \leq \Delta_1$). Table 5 shows results with the higher $\Delta$ values. Then, we used these sets of rules to classify again examples of test files. All rules belonging to the selected subsets of the cover of Table 5 conclude on NEG. To cope with this lack on rules on POS, we decided to use the following process (that we call *default rule*): when a chemical triggers no rule, it is classified as POS. Using this default rule, classification results are much better (see Table 5).

| File | $\Delta$ | No. of rules | Well-classified (%) | Well-classified (%) (using default rule) |
|---|---|---|---|---|
| MR | *12* | *14* | *30.88* | *66.18* |
| | 11 | 35 | 35.29 | 66.18 |
| | 10 | 95 | 36.76 | 64.71 |
| | 9 | 197 | 48.53 | 72.06 |
| | 8 | 326 | 51.47 | 67.65 |
| FR | 14 | 21 | 44.78 | 65.67 |
| | *13* | *76* | *50.75* | *71.64* |
| | 12 | 226 | 58.21 | 73.13 |
| | 11 | 552 | 62.69 | 73.13 |
| | 10 | 1103 | 64.18 | 71.64 |
| MM | *15* | *9* | *46.03* | *74.60* |
| | 14 | 16 | 49.21 | 76.19 |
| | 13 | 112 | 53.97 | 71.43 |
| | 12 | 267 | 55.56 | 65.08 |
| FM | 15 | 18 | 24.24 | 60.61 |
| | *14* | *42* | *45.45* | *69.70* |
| | 13 | 48 | 45.45 | 68.18 |
| | 12 | 99 | 51.52 | 69.70 |
| | 11 | 203 | 54.55 | 71.21 |

Table 5: Classification results with subsets of the cover

The comparison between the number of well-classified chemicals with all rules (Table 4) and selected subsets of the cover (Table 5) used with the default rule shows that the selection of rules improves the rate of well-classified chemicals by more than 10% (except on FR).

For predicting carcinogenic activity of the test file chemicals for the challenge, it was necessary to choose a subset of the extracted $\delta$-strong classification rules. For each experiment, we selected the subset emphasized in italic in Table 5 (we did empirically a trade-off between the well-classified rate and the number of rules). We used the default rule to classify test file chemicals. Table 6 gives the predicted class value frequencies w.r.t. the populations on this test file (185 examples).

| File | POS | POS (%) | NEG |
|---|---|---|---|
| MR | 48 | 25.95 | 137 |
| FR | 13 | 7.03 | 172 |
| MM | 37 | 20.00 | 148 |
| FM | 28 | 15.14 | 157 |

Table 6: Predicted class value frequencies w.r.t. the populations

More investigations are required to confirm such results. Among other things, a more sophisticated classification test strategy should be defined. Indeed, it seems unfair to use the same test file to select the subset of $\delta$-strong classification rules and classification results. Performances on the unseen chemicals of this challenge might provide a feedback about a potential bias.

Let us have a look on the rules we use for prediction. The 14 selected rules on MR data include the presence of fragment number 122. All these rules have a confidence value between 76.5% and 79.7%. Except the rule given below, all other rules have a support value between 18.5% and 19.2% (conf. is the abbreviation of confidence, sup. of support and frag. of fragment):

frag. 122 and frag. 158 $\Rightarrow$ NEG   $conf. = 77.9\%$   $sup. = 21.7\%$

On FR data, rules have a confidence between 62.2% and 81.5% and support ranges between 10.5% and 26.6%. Fragment number 1017 belongs to 22 rules (among 76). Here are some rules:

frag. 48 and frag. 1017 $\Rightarrow$ NEG   $conf. = 80.7\%$   $sup. = 26.6\%$
frag. 112 and frag. 1017 $\Rightarrow$ NEG   $conf. = 81.5\%$   $sup. = 24.7\%$
frag. 48 and frag. 818 $\Rightarrow$ NEG   $conf. = 80.0\%$   $sup. = 25.5\%$

As we selected only 9 rules on MM data, we provide the whole set of rules.

| | | |
|---|---|---|
| frag. 15 and frag. 818 $\Rightarrow$ NEG | $conf. = 75.0\%$ | $sup. = 17.6\%$ |
| frag. 15 and frag. 178 and 255 $\Rightarrow$ NEG | $conf. = 68.1\%$ | $sup. = 12.5\%$ |
| frag. 15 and frag. 178 and 257 $\Rightarrow$ NEG | $conf. = 71.2\%$ | $sup. = 14.5\%$ |
| frag. 15 and frag. 178 and 256 $\Rightarrow$ NEG | $conf. = 70.0\%$ | $sup. = 13.7\%$ |
| frag. 15 and frag. 178 and 872 $\Rightarrow$ NEG | $conf. = 68.1\%$ | $sup. = 12.5\%$ |
| frag. 266 and frag. 1017 $\Rightarrow$ NEG | $conf. = 76.7\%$ | $sup. = 18.0\%$ |
| frag. 257 and frag. 1017 $\Rightarrow$ NEG | $conf. = 76.3\%$ | $sup. = 17.6\%$ |
| frag. 80 and frag. 1017 $\Rightarrow$ NEG | $conf. = 76.0\%$ | $sup. = 16.0\%$ |
| frag. 15 and frag. 1017 $\Rightarrow$ NEG | $conf. = 76.6\%$ | $sup. = 19.1\%$ |

On `FM` data, rules have a confidence between 68.8% and 84.1% and support (except for the four rules given below) ranges between 12.0% and 16.2%. Fragment number 80 belongs to 36 rules (among 42). Here are the four best rules according to support and confidence (let us remark that these rules include just four fragments):

| | | |
|---|---|---|
| frag. 15 and frag. 1017 $\Rightarrow$ NEG | $conf. = 84.1\%$ | $sup. = 21.8\%$ |
| frag. 26 and frag. 818 $\Rightarrow$ NEG | $conf. = 82.3\%$ | $sup. = 19.9\%$ |
| frag. 15 and frag. 818 $\Rightarrow$ NEG | $conf. = 82.3\%$ | $sup. = 19.9\%$ |
| frag. 26 and frag. 1017 $\Rightarrow$ NEG | $conf. = 82.1\%$ | $sup. = 20.7\%$ |

These four experiments show that fragments 818 and 1017 are present in rules coming from `FR`, `MM` and `FM` (but not from `MR`). Fragment number 122 belongs only to rules on `MR`. Fragments numbers 15 and 80 are included only in rules coming from male and female mice. Identities of BCI fragments can be looked up in `http://www.informatik.uni-freiburg.de/~ml/ptc/train.bci.dictionary`. For instance, fragment 818 is an extremely general fragment that includes a wide variety of commonly occurring functional groups (e.g., alcohols, phenols, ethers including oxiranes, carbonyl compounds such as aldehydes, ketones, acids, and esters) and fragment 122 is a wide variety of aliphatic compounds or unsaturated compounds that have 4-atom substituents.

To finish, we give below some of the few rules concluding on `POS` and belonging to the cover extracted from `MR` file (see Table 4):

| | | |
|---|---|---|
| frag. 249 and frag. 1312 $\Rightarrow$ POS | $conf. = 64.3\%$ | $sup. = 9.8\%$ |
| frag. 256 and frag. 1312 $\Rightarrow$ POS | $conf. = 63.6\%$ | $sup. = 10.1\%$ |
| frag. 256 and frag. 1565 $\Rightarrow$ POS | $conf. = 62.8\%$ | $sup. = 9.8\%$ |
| frag. 257 and frag. 1312 $\Rightarrow$ POS | $conf. = 63.6\%$ | $sup. = 10.1\%$ |
| frag. 257 and frag. 1565 $\Rightarrow$ POS | $conf. = 62.8\%$ | $sup. = 9.8\%$ |

All the 17 rules concluding on `POS` have two fragments in their left-hand side and 14 fragments in total are used: numbers 233, 249, 255, 256, 257, 267, 420, 756, 872, 879, 1042, 1312, 1565 and 3599.

Let us note that Table 3 highlights the role of $\delta$. With $\delta = 0$, no classification rule is found. In real-world domains, it is likely that data intrinsically embed uncertainty. In other words, the same cause does not always produce the same effect and/or a number of parameter values which could explain the phenomenon are unknown. That is why such tasks are generally hard. In such situations, it is hopeless to look for sound and general rules with a confidence of 100% (i.e., $\delta = 0$). Furthermore, rules without exceptions (or too few exceptions) w.r.t. $\gamma$ may be over-specified and do not reflect a sound knowledge about the domain. Mining with $\delta > 0$ is a way to avoid over-fitting and to improve predictive performances.

## 5 Conclusion and further work

We showed the potential impact of $\delta$-strong classification rules in the discovery challenge for predicting chemical carcinogens. The method relies on recent results in the association rule mining area and provides a set of rules characterizing classes. With a positive value of $\delta$, we have shown that classification rules can be extracted from data sets for which no rule is discovered when $\delta = 0$. Furthermore, even if it were possible to mine 0-strong rules for lower frequency thresholds, those rules would have a very low frequency and could be spoilt by over-fitting for real-world domains like chemistry.

About the data of this discovery challenge, we found that only very few rules conclude on `POS` while a lot of rules conclude on `NEG`. It may be useful to study the fragment distributions w.r.t. the classes (are `NEG` chemicals correlated with more fragments?). Interestingly, such a method seems to highlight few rules (between 9 and 76, according to the populations) to predict the carcinogenic activity of `NEG` chemicals. Performances on the unseen chemicals of this challenge will provide a feedback about this approach. Further work is needed to improve the classification strategy.

**Acknowledgments**. The authors thank Christophe Rigotti for stimulating discussions and referees for their very helpful comments.

## References

[1] Agrawal, R. and Imielinski, T. and Swami, A. Mining association rules between sets of items in large databases, In *Proceedings SIGMOD'93*, ACM Press, pages 207–216, 1993.

[2] Agrawal, R. and Mannila, H. and Srikant, R. and Toivonen, H. and Verkano, I. Fast discovery of association rules, In *Advances in Knowledge Discovery and Data Mining*, AAAI Press, pages 307–328, 1996.

[3] Ali, K. and Manganaris, S. and Srikant, R. Partial classification using association rules, In *Proceedings KDD'97*, AAAI Press, pages 115-118, 1997.

[4] Bayardo, R. J. Brute-force mining of high-confidence classification rules, In *Proceedings KDD'97*, AAAI Press, pages 123-126, 1997.

[5] Boulicaut, J. F. and Bykowski, A. Frequent closures as a concise representation for binary data mining, In *Proceedings PAKDD'00*, Springer-Verlag LNAI 1805, pages 62-73, 2000.

[6] Boulicaut, J. F. and Bykowski, A. and Rigotti, C. Approximation of frequency queries by means of free-sets, In *Proceedings PKDD'00*, Springer-Verlag LNAI 1910, pages 75-85, 2000.

[7] Boulicaut, J. F. and Crémilleux B. $\delta$-strong rules to characterize classes, Research Report INSA Lyon-LISI, July 2001, 12 p. Submitted.

[8] Freitas, A. A. Understanding the crucial differences between classification and discovery of association rules - a position paper, In *SIGKDD Explorations*, Vol. 2(1), pages 65-69, 2000.

[9] Helma, C. and Gottmann, E. and Kramer, S. Knowledge Discovery and data mining in toxicology Technical Report, University of Freiburg, 2000.

[10] King, R.D. and Feng, C. and Sutherland, A. Statlog : Comparison of classification algorithms on large real-world problems, In *Applied Artificial Intelligence*, 1995.

[11] Liu, B. and Hsu, W. and Ma, Y. Integrating classification and association rules mining, In *Proceedings KDD'98*, AAAI Press, pages 80-86, 1998.

[12] Pasquier, N. and Bastide, Y. and Taouil, R and Lakhal, L. Efficient mining of association rules using closed itemset lattices. In *Information Systems* 24(1), pages 25-46. 1999.

[13] Salzberg, S. On comparing classifiers: pitfalls to avoid and a recommended approach, In *Data Mining and Knowledge Discovery*, Vol. 3(1), pages 317-327, 1997.

[14] Schaffer, C. Overfitting avoidance as bias, In *Machine Learning*, Vol. 10, pages 153-178, 1993.

[15] Toivonen, H. Sampling large databases for association rules, In *Proceedings VLDB'96*, Morgan Kaufmann, pages 134-145, 1996.