δ-strong classification rules for characterizing chemical carcinogens

Bruno Crémilleux[1,2], Jean-François Boulicaut[1]
[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 Predictive Toxicology Challenge for 2000-2001. 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 impact of δ-strong classification rule mining in such a domain. This new approach relies on recent results for association rule mining. It is based on δ-free set computation and provide the simplest classification rules. These rules characterize the classes and are used to build classifiers. This technique is operational for large data sets and can be used even in the difficult context of highly-correlated data where other algorithms fail.

1 Introduction

One popular data mining technique concerns knowledge discovery from frequent association rules. This kind of process has been studied a lot since its definition in [1]. Association rules can tell something like "It is frequent that when properties A_1 and A_2 are true within an example, then property A_3 tends to be true". We provide a simple formalization of this task in Section 2.1.

Finding classification rules is an important research focus as well. Starting from a collection of examples associated with a known class value, it 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 [7, 12]). In that context, a classification rule is a rule that concludes on one class value. We provide a formal definition at the end of Section 2.2.

Mining classification rules can be viewed as a special form of association rule mining where conclusions of rules are pre-specified. Unfortunately, naive approaches are not able to tackle huge data and, as the identification of the classification rules is performed mainly as a post-processing step [8, 2], the large number of produced rules leads to rule conflicts and over-fitting. Indeed, how to identify the most relevant rule to classify a new case might be quite difficult. To cope with these drawbacks, we propose to extract efficiently the collection of the simplest classification rules.

Recent works revisit these questions and propose algorithms for mining relevant sets of classification rules. CMAR [9] uses statistical techniques to avoid bias. The minimal subset of classification rules having the same prediction power (defined by the pessimistic estimation statistical measure or the confidence) as the complete class rule set is computed in [10].

The main contribution of this paper is to study the impact of δ-strong classification rule mining in real-world domains like predictive toxicology. This is a preliminary work and more investigations have to be done to validate the interest of this approach when classifiers have to be built on such domains.

We use a recent efficient association rule mining technique, the so-called δ-strong rule mining technique [4], to extract a condensed representation of classification rules [5]. We are then able to exhibit the simplest classification rules w.r.t. their left-hand sides (which is a key point in classification) and we get a cover of classification rules. The intuition is that, 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. The technique provides *every* simple classification rule. Furthermore, in [5], it has been shown that a simple property enables to avoid important classification conflicts for unseen examples, allowing a straightforward use of such rules for class characterization. As a result, it becomes possible to work on difficult domains such as large and dense learning sets for which other techniques based on association rule mining and then postprocessing might fail.

Section 2.2 introduces association rule mining and the concept of δ-strong classification rule. Section 3 presents the data preparation stage for the Predictive Toxicology Challenge 2000-2001 and Section 4 gives our results of this application.

2 δ-strong rules to characterize classes

2.1 Association rule mining

Let us provide a simple formalization of the δ-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 the Predictive Toxicology Challenge (PTC), $\mathbf{R}$ 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 all the items in $X \cup \{B\}$ are true (value 1) for enough examples and that when an example contains true 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) *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* $\mathcal{F}(X \cup \{B\}, \mathbf{r})/\mathcal{F}(X, \mathbf{r})$. *We defined an absolute frequency (a number of examples* $\leq |\mathbf{r}|$*). We also use the relative frequency* $\mathcal{F}(X \cup \{B\}, \mathbf{r})/|\mathbf{r}|$*, i.e., a value in* $[0, 1]$.

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

The complexity of frequent itemset mining is exponential with the number of attributes. Many research works (e.g. [14, 4]) concern the 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 δ-strong rules

A classification rule must conclude on class values with a rather high confidence. δ-strong rules introduced in [4] satisfy such a constraint.

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

A δ-strong rule is violated by at most δ examples, i.e., its confidence is at least equal to $1 - (\delta/\gamma)$. Notice also that its frequency relies on the frequency of its left-hand side. From a technical perspective, δ-strong rules can be built from δ-free itemsets that will constitute their left-hand sides. It is out of the scope of this paper to provide details about the concept of δ-free itemset that has been recently designed in our group (see [4]). It is related to the concepts of closed itemset [11] and almost-closure [3]. An itemset X is called δ-free if there is no δ-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 [1]. When $\delta > 0$, we are interested in the almost-closures of a frequent δ-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 δ-strong rules from the γ-frequent δ-free sets and their almost-closures.

We use the prototype `ac-miner-12`[2]. Given thresholds γ and δ, it provides the collection of frequent δ-free itemsets, their frequencies and the attributes in their almost-closures. The δ-strong rules formalism offers a property of minimal body which is a key point for a classification purpose.

Property 1 (minimal body) *If* X *is a* δ*-free itemset and* $X \Rightarrow B$ $\{\delta\}$ *is a* δ*-strong rule with exactly* δ *exceptions (figure between braces indicated the exact number of exceptions), then* X *is the minimal set of items from which we can conclude on* B *with at most* δ *exceptions.*

It means that if $X \Rightarrow B$ $\{\delta_1\}$ is a δ-strong rule with δ_1 exceptions, there is no itemset Y, $Y \subset X$, such that $Y \Rightarrow B$ $\{\delta_2\}$ is a δ-strong rule with $\delta_2 \leq \delta_1$. In other terms, it is possible to get the simplest rules, i.e., a cover of δ-strong rules. We argue that this property of minimal body is a fundamental issue for classification. Not only it prevents from over-fitting [13] but also it makes the classification of an example easier to explain. Furthermore, 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 now 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 (δ-strong classification rule) *A* δ-strong classification rule *is a* δ*-strong rule that concludes on one class value (i.e.,* C_i*).*

It is shown in [5] that if $\delta < \gamma$, then some rule conflicts are avoided. For instance, if there is the δ-strong classification rule $R_1 : X \Rightarrow C_i$, then a δ-strong classification rule $R_2 : X \Rightarrow C_j$ with $i \neq j$ cannot appear. Furthermore, if $\delta < \gamma$, there is no δ-strong classification rule $R_3 : X \cup Y \Rightarrow C_j$ with $i \neq j$. As this sufficient condition on γ and δ is quite reasonable in practice, our experiments have been done under this assumption.

3 Data preparation

Chemicals in PTC are available with seven sets of descriptors. Chemical characteristics are functional groups, atomic and bond properties, molecular substructures represented with fingerprints. Most of the sets of descriptors require a chemical knowledge to be used and we were lacking from such an expertise. 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 6,150 fragments (each fragment

[1] 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.

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

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 57,240 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% [6]). 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 the PTC, 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[3] and earlier designations[4]. 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: Initial class value frequencies

Given that the PTC 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 by the PTC, we decided not to recode instances with class values EE, E and IS. Finally, we got the four following sets (see Table 2). Let us remark that there is no missing value in these data.

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

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

Table 2: Computed class value frequencies

4 Results and discussion

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 δ and γ, Table 3 gives the extraction time, the number of δ-free itemsets (noted "δ-FIS") and almost-closures ("AC") that contain a class value. This last number can be seen as the number of potential δ-strong classification rules (i.e., with any frequency and confidence values). All these rules compose a cover of all classification rules (see Section 2.2). Let us remark that with most values of γ indicated in Table 3, usual apriori-like algorithms fail due to an excessive memory requirement). 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.

$\gamma/\lvert\mathbf{r}\rvert$	δ	Time (sec.)	No. of δ-FIS	No. of AC
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, δ-free itemsets and almost-closures with a class value w.r.t. δ and γ

When the extraction turns to be intractable, it comes from an excessive memory requirement because of the management of huge collections of candidates for frequent δ-freeness.

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

[4] P: Positive ; E: Equivocal ; N: Negative

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 δ-strong classification rules depends on the values for the thresholds γ and δ. Also, we check experimentally that the more we increase the value of δ, the more we can have tractable extractions for lower frequency thresholds. 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 δ.

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 2,828 rules concluding on NEG and only 7 on POS. With $\gamma/|\mathbf{r}| = 0.15$ and $\delta = 20$, we get 4,443 rules concluding on NEG and 86 on POS. Examples of δ-strong classification rules are given at the end of this section.

The cover (or a collection) of discovered δ-strong classification rules can be used to predict chemical carcinogens. Nevertheless, such a cover includes rules with low support and/or confidence. These rules with a poor quality may introduce errors. To evaluate this issue, we used the cover as a classifier. 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). 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 has been computed and the class value having the best score has been 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 used ac-miner-12 with the lowest value of γ and a sensible value for δ (we experimented also with $\gamma/|\mathbf{r}| = 0.11$ but it led to choose δ around 30 to ensure the extraction tractability).

File	$\gamma/\lvert\mathbf{r}\rvert$	δ	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 which includes rules with low support and/or confidence. The design of a classifier stemming from the classification rule cover to remove rules with a poor quality still needs some research. Let us give a preliminary approach.

It concerns one experiment and it is similar for the others. For each rule, we compute a score (denoted Δ) which is the difference between the well-classified and the miss-classified examples of the test file. Then rules are sorted out w.r.t. Δ and, by varying Δ, we define a family of nested sets of rules by the following way: for a value Δ_1 of Δ, we keep the rules having $\Delta \geq \Delta_1$. All rules belonging to the set defined by Δ_1 belong to the sets defined by Δ_2 with $\Delta_2 < \Delta_1$.

All rules belonging to the selected subsets of the cover of Table 5 conclude on NEG. We used these sets of rules to classify again the examples of the test files. Table 5 shows classification results with the higher Δ values. To cope with the lack of rules on POS, we decided to use the following *default rule*: when a chemical triggers no rule, it is classified as POS. Using this rule, classification results are much better (see Table 5, "WC" noted well-classified and "default" that the default rule is used).

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

File	Δ	No. of rules	WC (%)	WC (%) default
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

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 PTC, it was necessary to choose a subset of the extracted δ-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 for which true class values were known (this one contains 185 chemicals). The numbers of positive chemicals are: 52 on MR, 36 on FR, 29 on MM and 35 on FM.

It is then possible to evaluate performances on the test file. Table 6 gives classification results. True predictions (noted “True P”) are made of true positive (“TP”) and true negative (“TN”). False predictions (“False P”) are composed of false negative (“FN”, chemicals predicted as negative whereas they are positive) and false positive (“FP”, chemicals predicted as positive whereas they are negative).

File	TP	TN	True P (%)	FN	FP	False P (%)
MR	14	99	61.1	38	34	38.9
FR	3	139	76.8	33	10	23.2
MM	6	125	70.8	23	31	29.2
FM	7	129	73.5	28	21	26.5

Table 6: Classification results

The global percentage of true predictions is 70.55%. Nevertheless, especially on female populations, models built from δ-strong classification rules tend to favour to predict no carcinogenic activity. This is confirmed by the ROC analysis of models of all participants (see http://www.informatik.uni-freiburg.de/~ml/ptc/#ROC).

Let us have a look on the rules used for prediction. The numbers within the rules denote fragments. The first figure after a rule is its confidence and the second one is its relative frequency. To enable rule understanding, Table 7 indicates the identities of fragments as given in http://www.informatik.uni-freiburg.de/~ml/ptc/train.bci.dictionary. Some fragments have several identities (indicated on different lines or separated by “-” in Table 7).

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 : “122 and 158 ⇒ NEG 77.9% 21.7%”, all other rules have an absolute frequency value between 18.5% and 19.2%.

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

	Conf.	Freq.
48 and 1017 ⇒ NEG	80.7%	26.6%
112 and 1017 ⇒ NEG	81.5%	24.7%
48 and 818 ⇒ NEG	80.0%	25.5%

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

	Conf.	Freq.
15 and 818 ⇒ NEG	75.0%	17.6%
15 and 178 and 255 ⇒ NEG	68.1%	12.5%
15 and 178 and 257 ⇒ NEG	71.2%	14.5%
15 and 178 and 256 ⇒ NEG	70.0%	13.7%
15 and 178 and 872 ⇒ NEG	68.1%	12.5%
266 and 1017 ⇒ NEG	76.7%	18.0%
257 and 1017 ⇒ NEG	76.3%	17.6%
80 and 1017 ⇒ NEG	76.0%	16.0%
15 and 1017 ⇒ NEG	76.6%	19.1%

Num.	Fragment identities
15	APAA2 3 3 AA0 1
26	APAA2 3 2 AA0 1
48	AA4Aar4Aar4A - AAC arC arC
80	APAA2 2 5 AA0
112	RC4Aaa4Aaa4Aaa4Aaa4Aaa4Aaa RC4Aar4Aar4Aar4Aar4Aar4Aar RCC aaC aaC aaC aaC aaC aa RCC arC arC arC arC arC ar
122	ASAAcsAAcsAAcsAA
158	AS4Aaa4Aaa4Aaa4A ASC aaC aaC aaC
178	APAA2 3 5 AA0 1
249	AA4Aaa4Aaa4Aaa5A AAC aaC aaC aaN
255	AS5Aaa4Aaa4Aaa4Aaa4Aaa4Aaa4A ASN aaC aaC aaC aaC aaC aaC
256	AS5Aaa4Aaa4Aaa4Aaa4A ASN aaC aaC aaC aaC
257	AS5Aaa4Aaa4Aaa4A - ASN aaC aaC aaC
266	ASAAarAAarAAarAAarAAarAAacAAacAA
818	AAO aaC - AAO acC
872	AS5Aaa4Aaa4Aaa4Aaa4Aaa4A ASN aaC aaC aaC aaC aaC
1017	AA6Aaa4A - AA6Aac4A
1312	APAA2 2 5 AA2 2
1565	APAA2 2 6 AA2 2

Table 7: Identities of fragments

On FM data, rules have a confidence between 68.8% and 84.1% and frequency (except for the four rules given below) ranges between 12.0% and 16.2%. Fragment number 80 belongs to 36 rules (among 42).

Let us provide the four best rules according to frequency and confidence (let us remark that these rules include just four fragments).

	Conf.	Freq.
15 and 1017 ⇒ NEG	84.1%	21.8%
26 and 818 ⇒ NEG	82.3%	19.9%
15 and 818 ⇒ NEG	82.3%	19.9%
26 and 1017 ⇒ NEG	82.1%	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. For instance, fragment number 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 number 122 is involved in 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):

	Conf.	Freq.
249 and 1312 ⇒ POS	64.3%	9.8%
256 and 1312 ⇒ POS	63.6%	10.1%
256 and 1565 ⇒ POS	62.8%	9.8%
257 and 1312 ⇒ POS	63.6%	10.1%
257 and 1565 ⇒ POS	62.8%	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.

5 Conclusion

We discussed the potential impact of δ-strong classification rules in the Predictive Toxicology Challenge for 2000-2001. The method relies on recent results in the association rule mining area and provides a set of rules that characterize the classes.

With a positive value of δ, we have shown that classification rules can be extracted from data sets for which no rule is discovered when $\delta = 0$. It means that the most effective algorithms based on closed set discovery might not help. In real-world like chemistry, 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. γ may be over-specified and do not reflect a sound knowledge about the domain. Mining with $\delta > 0$ is a way to avoid overfitting and to improve predictive performances. This approach is effective even in the case of huge, dense and/or highly correlated learning data sets.

About the data of this PTC 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. Further work is needed to improve the classification strategy.

Acknowledgments. The authors thank Christophe Rigotti for stimulating discussions. This work has been done while the first author was an invited researcher at INSA Lyon.

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] Bayardo, R. J. and Agrawal, R. and Gunopulos, D. Constraint-based rule mining in large, dense database, In *Proceedings ICDE'99*, pp. 188-197, 1999.

[3] 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.

[4] 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.

[5] Crémilleux B. and Boulicaut, J. F. Simplest classification rules generated by δ-free sets, Research Report INSA Lyon-LISI, 2001, 12 pages. Submitted. *A French version of this paper is to appear in the proceedings of the French-speaking biennal conference on Artificial Intelligence RFIA'02.*

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

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

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

[9] Li, W. and Han, J. and Pei, J. CMAR: Accurate and Efficient Classification Based on Multiple Class-Association Rules, In *Proceedings of ICDM 01*, San Jose, California, 2001.

[10] Li, J. and Shen, H. and Topor, R. Mining the Smallest Association Rule Set for Predictions, In *proceedings of ICDM'01*, San Jose, California, 2001.

[11] 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.

[12] 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.

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

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