Mining δ-strong Characterization Rules in Large SAGE Data

Céline Hébert[1], Sylvain Blachon[2], and Bruno Crémilleux[1]

[1] GREYC - CNRS UMR 6072, Université de Caen
Campus Côte de Nacre
F-14032 Caen cedex, France
{Forename.Surname}@info.unicaen.fr
[2] CGMC - CNRS UMR 5534, Université Lyon 1
Bat. Grégoire Mendel
F-69622 Villeurbanne cedex, France
blachon@cgmc.univ-lyon1.fr

Abstract. SAGE data provide the level-expression of a large amount of genes and few biological situations. Geometrical dimensions of such context make difficult the running of data mining methods. In this paper, we propose a new method to extract all the δ-strong characterization rules in large data sets. Such rules enable to characterize classes. We use this method to mine the SAGE data and highlight a set of characterization rules and few potential relevant genes which may be associated to cancer.

1 Introduction

A critical issue in genomic research is to derive relevant knowledge from huge gene expression datasets generated at high throughput. Biologists are interested in looking for associations (i.e., patterns) under several kinds of constraints like for instance groups of co-regulated genes, also known as synexpression groups. The ECML/PKDD 2005 discovery challenge deals with publicly available human serial analysis of gene expression (SAGE) data. SAGE is an experimental technique designed to quantify gene expression. SAGE data provide expression values for given biological situations (i.e., the lines) and given genes (i.e., the columns). These datasets are characterized by a *large* number of columns (e.g., ten of thousands of gene expressions) and few biological situations. For instance, one of the dataset available in this discovery challenge gathers 27,679 gene expressions for 90 biological situations. As the algorithms extracting patterns run through a search space growing exponentially according to the number of columns, the extraction of the complete collection of patterns under various kinds of constraints remains a challenge. Nevertheless, we claim that it is important to provide complete methods to capture all the information embedded in the data. Columns are often called *attributes* and lines *rows* or *objects*.

Assume that the biologist is interested in all patterns of overexpressed genes occurring in at least a given number of biological situations. To be able to answer this query, it is necessary to tackle the common constraint of frequency (a pattern is said *frequent* if it is supported by at least γ objects in the dataset, γ

being a given threshold). Even for this basic constraint, which furthermore satisfies relevant properties (anti-monotonous with respect to (w.r.t.) the specialization relation defined by the attributes [10]), classical algorithms based on a Apriori-like approach fail [13].

However, in the particular case of the closed patterns, thanks to the join use of the transposition of data and the properties of the Galois connections, it is possible to mine all closed patterns and infer all frequent patterns in such large data [13] (briefly speaking, a closed pattern is a maximal set of attributes (w.r.t. the set inclusion) shared by a set of objects). This approach has been successfully applied on gene expression data including SAGE data [14]. By using this technique, let us note that the extracted patterns are object patterns and no longer attribute patterns and it is necessary to define the transposed form of the constraint. Unfortunately, there is no straightforward generalization of this approach for a lot of useful constraints. For instance, this approach cannot be used to extract free (or key) patterns [3, 12] and δ-free patterns in large data. Furthermore, even if the constraint of δ-freeness is efficiently pushed by the algorithms extracting the so-called condensed representations of frequent patterns [4], δ-free patterns cannot be extracted by these algorithms [13] in such data. In a recent work [8], by using the associated objects (i.e., *extension*) to a pattern, we have proposed a method to mine frequent and δ-free patterns from large data. The success of this approach relies on the fact that the extension of a pattern has few objects in large datasets.

The δ-free patterns are of a great interest because their uses are highly interesting. For instance, they enable to build rules with a bounded number of exceptions [3], non redundant rules [15] and it is known that their capacity to indicate the minimal part of attributes highlighting a phenomenon is precious in classes characterization and classification [1, 5]. There is an intense need of classes characterization and classification techniques. For instance, in this discovery challenge, the collected biological situations gather 59 cancerous samples and 31 normal and the biologists would like to better understand the relationships between the genes and the class (cancer versus normal) of the biological situations. In this paper, we tackle the search of δ-strong characterization rules in large datasets. From a technical point of view, we will see in Section 2 that δ-strong characterization rules are inferred from δ-free patterns and their almost-closures (the almost-closure is defined in Section 2.1).

The contribution of this paper is twofold. First, we propose a method to mine all the δ-strong characterization rules in large datasets. Indeed, even if a method to extract δ-free patterns is provided in [8], the almost-closures are not given. Due to their huge size in large data, it seems to be a pitfall to try to build them. We show how we get round this difficulty by focussing on the attributes describing the class values. These attributes are the only attributes belonging to the almost-closures which are required for mining δ-strong characterization rules. Second, we provide several δ-strong characterization rules w.r.t. the class attribute (cancer versus normal) of the datasets. Experiments tackle the large matrix (27,679 gene expressions for 90 biological situations) available at this discovery challenge.

Section 2 presents the background (δ-strong characterization rules and mining δ-free patterns in large datasets) which is required to understand the rest of this paper. Section 3 describes our method to mine δ-strong characterization rules in large datasets. Section 4 gives an overview of the used datasets. Finally, experiments and results are describing in Section 5.

2 Characterization rules and mining δ-free patterns in large datasets

2.1 Notations and δ-freeness

Let us specify the definitions and notations. Let r be the dataset. An *attribute pattern* (resp. *object pattern*) is a set of attributes (resp. objects). The class attribute is denoted by a. We say that an attribute pattern X is supported by an object if this object contains X. X is *γ-frequent* if it is supported by at least γ objects in r, γ being a given threshold. The frequency of X is denoted by $\mathcal{F}(X)$. Each object is labelled by a class value in $\{c_1, \ldots, c_n\}$. Table 1 provides an example of dataset where 5 biological situations (i.e., objects) $o_1, \ldots, o_5$ are described by 8 gene expressions (i.e., attributes) $a_1, \ldots, a_8$. The class values are `cancer` and `no cancer` (for instance, `cancer` $= c_1$ and `no cancer` $= c_2$). For example, the attribute pattern $a_1a_3a_5$ is supported by the objects o_1 and o_3 and thus $a_1a_3a_5$ is 2-frequent.

	gene expressions								class: a	
situations	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	`cancer`	`no cancer`
o_1	1	0	1	0	1	0	1	0	1	0
o_2	0	1	1	0	1	0	1	1	1	0
o_3	1	0	1	0	1	0	0	1	1	0
o_4	0	1	1	0	1	1	0	0	1	0
o_5	1	0	0	1	0	1	0	1	0	1
o_6	0	1	1	0	1	1	0	1	0	1
o_7	1	0	1	0	0	1	0	1	0	1

Table 1. An example of gene expression dataset

Let us provide a formal definition of the δ-freeness. X is a *δ-free* pattern if there is no association rule between two of its proper subsets with less than δ exceptions (i.e., $\forall Y \subset X, \forall Z \subset X$, and $Y \cup Z = X$ and $Y \cap Z = \emptyset$, there is no rule $Y \rightarrow Z$ with $\mathcal{F}(YZ) + \delta \geq \mathcal{F}(Y)$). An equivalent definition [4] is: X is a *δ-free* pattern if for each $Y \subset X$, $\mathcal{F}(X) + \delta < \mathcal{F}(Y)$. In Table 1, with $\delta = 1$, a_5a_8 is 1-free since $\mathcal{F}(a_5a_8) = 3$ and one have $\mathcal{F}(a_5a_8) + \delta = 4 < \mathcal{F}(a_5) = \mathcal{F}(a_8) = 5$. Let us recall that the *confidence* of the rule $Y \rightarrow Z$ is $\mathcal{F}(YZ)/\mathcal{F}(Y)$. We specify the notion of *almost-closure* which is needed for the rest of this paper: let δ be a positive integer, the almost-closure of X is the set of attributes a_i satisfying $\mathcal{F}(X) - \mathcal{F}(Xa_i) \leq \delta$.

2.2 δ-strong characterization rules

Let us have a look on the δ-strong characterization rules introduced in [5]. The key idea is to mine all rules with a *minimal body* (i.e., the shortest premises). We argue that this property of minimal body is a fundamental issue for characterization. Not only it prevents from over-fitting (i.e. over-specified rules) but also it makes the characterization easier to explain. To be used in practice, a reliable characterization rule must have a rather high confidence (i.e., we say that the rule is *strong* [3]). The following definition of δ-strong characterization rules ensures that such a rule has a minimal body and a number of exceptions bounded by an integer δ (in other words, the confidence of such a rule is greater than or equal to $1 - (\delta/\gamma)$).

Definition 1 (δ-strong characterization rules). *Given a frequency threshold γ and an integer δ, a rule $X \rightarrow a_i$ is a δ-strong characterization rule if there is no rule $Y \rightarrow a_i$ with $Y \subset X$ and a confidence greater than or equal to $1 - (\delta/\gamma)$.*

Given a frequency threshold γ, this definition means that we consider only the minimum sets of attributes (i.e., the minimal bodies) to end up on a_i, the uncertainty being controlled by δ. The value of δ is fundamental to obtain relevant rules. When $\delta = 0$, every rule has a confidence value of 1 (i.e., *exact* rule). In practical applications, especially in biology or medicine, we generally got few exact rules due to the non-determinism of the phenomena and we have to relax the condition on the value of δ to accept exceptions (the more δ raises, the more the confidence decreases). For instance, in our running example (Table 1), with $\delta = 1$ and $\gamma = 2$, $a_5 \rightarrow \texttt{cancer}$ is a δ-strong characterization rule, but $a_5a_8 \rightarrow \texttt{cancer}$ is not a δ-strong characterization rule.

Let us suppose now that the conclusion of a δ-strong characterization rule is a class value (i.e., c_i). We consider here the typical case where each object is associated to a unique class value. Then, the δ-strong characterization rules satisfy Property 1 which indicates that, under the assumption $\delta < \gamma$ (which is very sensible in practice!), some classification conflicts are avoided. This property is proved in [5] and it is precious to design classifiers [5]. This property will be also useful in the following of this paper (Section 3) to prove Property 5 which leads to a safe pruning criterion.

Property 1 (included bodies conflict) *Given a frequency threshold γ, an integer δ, X and Y two attribute patterns, $Y \subset X$, then the pair of δ-strong characterization rules $X \rightarrow c_i$ and $Y \rightarrow c_j$ cannot exist.*

Let us say few words on the mining of δ-strong characterization rules. From a technical point of view, these rules come from the collection of δ-free patterns as highlighted by Property 2.

Property 2 *Given a frequency threshold γ, an integer δ and a δ-strong characterization rule $X \rightarrow a_i$, then X is a δ-free pattern.*

The proof of Property 2 is in [5] (its converse is false [5]). The usual method to mine the δ-strong characterization rules requires the computation of all frequent δ-free patterns and their almost-closures [5] to check if a class attribute belongs to the almost-closure of X and does not belong to the almost-closures of the subsets of X. For datasets with usual geometrical dimensions, this can be efficiently performed thanks to a level-wise search [5]. Unfortunately, in large datasets, the computation of the almost-closures is unfeasible and this motivated our work to mine δ-strong characterization rules in large datasets. The next section addresses the mining of δ-free patterns in large data.

2.3 Mining δ-free patterns in large datasets

As indicated in introduction, usual techniques fail to extract δ-free patterns in large data. This is mainly due to the computation of the closures which contain a lot of attributes or the large number of the candidates patterns and their storage requires a large amount of memory. But, in large data, there are only few objects which contain a set of attributes. Starting from this observation, we have recently proposed [8] a method to mine frequent δ-free patterns in large datasets. The key idea is to use the extension of a pattern to check these constraints, because the extension has few objects which are easily handled in large databases. More formally, the *extension* of an attribute pattern X, denoted by $g(X)$, is the maximal set of objects containing X. Our approach relies on the following property which underlines the relation between the frequency of an attribute pattern and its extension:

Property 3 *The frequency of an attribute pattern X is equal to the cardinal of its extension $|g(X)|$.*

It is clear that this property is well known but its use is interesting because it enables to establish the frequency and the δ-freeness of any pattern only with its extension.

Furthermore, we have shown an efficient property to compute the extension of a pattern (i.e., let X and Y be two patterns, the extension of $X \cup Y$ is equal to $g(X) \cap g(Y)$) and a new safe pruning criterion based on the common use of the minimal frequency and the δ-freeness properties. The simultaneous use of this property and the pruning criterion is on the core of the FTminer (FT for Free faT[3]) algorithm [8]. FTminer succeeds in mining large datasets like the ones of this discovery challenge. Nevertheless, due to their huge size in large data, FTminer does not compute the almost-closures of the δ-free patterns. We have seen in Section 2.2 that the almost-closures are required to mine the δ-strong characterization rules. The next section proposes a new approach to get round this difficulty by focussing on the attributes describing the class values.

3 Mining characterization rules in large datasets

This section presents our approach to mine δ-strong characterization rules in large data. As for mining δ-free patterns in large datasets, the key idea is to use the extensions to check if an attribute belongs to an almost-closure. As the almost-closures in large data are huge, we only compute this test for the attributes of the class values. Furthermore, we give a new pruning criterion which speeds up the

[3] The word "fat" is also used to refer to a large dataset as indicated by D. Hand during his invited talk at PKDD'04.

extraction of δ-patterns which are on the basis of the δ-strong characterization rules.

Property 4 shows that we can check if an attribute a_i belongs to the almost-closure of a pattern X thanks to the extensions of X and $X \cup \{a_i\}$. The number of exceptions (i.e., number of objects containing X and not a_i) is also known, which will be required to compute the confidence of the δ-strong characterization rules.

Property 4 *Let X be an attribute pattern and δ a positive integer. An item a_i belongs to the almost-closure of X if and only if*

$$|g(X)| - |g(X) \cap g(a_i)| \leq \delta$$

Moreover, the number of exceptions (number of objects containing X and not a_i) is equal to $|g(X)| - |g(X) \cap g(a_i)|$.

Proof. Property 3 indicates that the frequency of a pattern X is $|g(X)|$. As $\mathcal{F}(X \cup \{a_i\})$ is equal to $|g(X) \cap g(a_i)|$ [8], the definition of the almost-closure (see Section 2.1) is equivalent to $|g(X)| - |g(X) \cap g(a_i)| \leq \delta$. Furthermore, by the definition of the almost-closure, the number of exceptions is $|g(X)| - |g(X) \cap g(a_i)|$.

Let us provide an example. With $\gamma = 2$ and $\delta = 1$, in our running example (Table 1), a_5a_8 is 1-free (see Section 2.1). $|g(a_5a_8)| - |g(\{a_5a_8\} \cup \{c_1\})| = 3 - 2 = 1$, thus c_1 belongs to the almost-closure of a_5a_8. As c_1 does not belong to the almost-closure of a_5 nor a_3, $a_5a_8 \rightarrow c_1$ is a 1-strong rule characterization of frequency 2 with one exception in r (i.e., its confidence is 2/3).

We give now the following important property:

Property 5 (pruning criterion) *Let c_i and c_j be two class attributes, X and Y two attribute patterns. If $X \rightarrow c_i$ is a δ-strong characterization rule, then $\forall\ Y \supset X,\ \ Y \rightarrow c_j$ is not a δ-strong characterization rule.*

Proof. Let $X \rightarrow c_i$ be a δ-strong characterization rule, c_i denotes a class attribute. Assume that $Y \rightarrow c_j$ with $Y \supset X$ is a δ-strong characterization rule. This is contradictory with Property 1.

Property 5 means that a level-wise algorithm can prune the search space from X and it leads to a pruning criterion to mine δ-free patterns from which δ-strong characterization rules are inferred.

We designed FTCminer (FTC for Free faT Characterization) algorithm. FTCminer follows the outline of a level-wise algorithm. Its originality is that there is no generation phase of all the candidates at a given level since the candidates are generating one at a time.

Given γ and δ, FTCminer mines the sound and complete collection of frequent δ-strong characterization rules.

Proof (Correctness). By construction, the body of a rule `r` produced by FTCminer is a frequent δ-free pattern X and the conclusion of `r` is a class attribute c_i belonging to the almost-closure of X w.r.t. δ (i.e., $\mathcal{F}(X \cup \{c_i\}) - \mathcal{F}(X) \leq \delta$), thus the confidence of `r` is greater than or equal to $1 - (\delta/\gamma)$. Let Z be an attribute pattern. The use of the pruning criterion (Property 5) ensures that there is no rule $Z \rightarrow c_i$ with $Z \subset X$, thus `r` satisfies Definition 1.

Proof (Completeness). FTCMINER mines all the frequent δ-free patterns with an almost-closure containing a class attribute. As the pruning criterion (Property 5) is a safe pruning criterion (i.e. the pruned rules do not satisfy Definition 1), FTCMINER is complete.

4 SAGE dataset and data preparation

From `http://lisp.vse.cz/challenge/CURRENT/`, we downloaded the large gene expression dataset which provides the level of expression of 27,679 genes in 90 biological situations. To be more precise, each attribute is a *tag*. The identification of genes is closely related to the tags and biologists are able to associate the genes to the corresponding tags. We did not use the small (74×822) dataset also available at this discovery challenge because biologists are more interested in knowledge which may be extracted from a large set of genes. In the following of this paper, we speak only of the large dataset.

The values of tags vary from 0 to 26021. The percentage of values of tags different from 0 is 19.86% and the arithmetic mean is around 4. As already said, the biological situations are divided into two classes: `cancer` and `no cancer`. 59 situations are labelled by `cancer` and 31 by `no cancer` (i.e., `normal`).

Gene expressions are quantitative values. Starting from such values, the property of overexpression has to be encoded for each gene. For that, three discretized datasets were kindly provided by the CGMC Laboratory. We briefly summarize the methods of discretization (see [2] for more details). For each discretization, the value `1` encodes overexpressed genes.

- `Xmax` method. For each tag, this consists of identifying the biological situations in which its value is in the 5% of the highest values. These values are encoded `1`, `0` otherwise.
- `max-Xmax` method. The threshold is fixed w.r.t. the maximal value (*max*) observed for each tag. All the values which are greater than $(100 - X\%)$ of *max* are assigned to `1`, `0` for the others. Here, $X = 25$.
- `mid-range` method. The cut-off is fixed according to the highest (*max*) and the lowest (*min*) values of each tag and their arithmetic mean is computed. Then, for each tag, all the values greater than the arithmetic mean are set to `1`, `0` otherwise.

Obviously, the discretization process preserves the dimensions of the dataset and the label of each situation. Table 2 gives the main characteristics of the three resulting datasets which are called `Xmax`, `max-Xmax` and `mid-range`. The density is the number of `1` divided by the total size of the dataset (i.e. the number of attributes times the number of objects) given as a percentage. For instance, the density of Table 1 is 53.57% (thirty 1 divided by $56 = 8 \times 7$).

data	Xmax	max-Xmax	mid-range
density (%)	4.49	2.01	3.64
overexpressed genes/row	1242.03	554.98	1008.12

Table 2. Main characteristics of the three discretized datasets

5 Results and discussion

In these experiments, we give the main features on the whole sets of δ- characterization rules and on the specific tags and rules which might be relevant in order to characterize the biological situations according to `cancer` and `no cancer`. More precisely, we discuss the interest of a particular rule. Obviously, we use FTCminer.

Overview of the whole sets of δ-characterization rules. Table 3 gives the number of δ-characterization rules from `Xmax`, `max-Xmax` and `mid-range` according to the parameters γ and δ. As expected, the number of rules increases when γ decreases (and δ has a constant value). On the contrary, the number of rules decreases when δ decreases (see for γ fixed to 5 or 10).

Note that when $\gamma = 15$ and $\delta = 3$, we do not find rules in `Xmax` and `max-Xmax`. Moreover, in this case, all extracted rules on `mid-range` conclude on `cancer`. More generally, we got more rules on `cancer` than on `no cancer`. The imbalance between the classes may explain this phenomenon (as there are only 31 situations labelled by `no cancer`, $\gamma = 15$ means that a rule must appear in nearly 50% of the situations).

γ	15		10		10		9	
δ	3		3		2		2	
rules	cancer	no cancer	cancer	no cancer	cancer	no cancer	cancer	no cancer
Xmax	0	0	8	0	2	0	12	0
max-Xmax	0	0	10	1	1	0	9	0
mid-range	45	0	638	8	369	1	777	3

γ	7		5		5		4	
δ	2		2		1		1	
rules	cancer	no cancer	cancer	no cancer	cancer	no cancer	cancer	no cancer
Xmax	278	29	4837	1322	2838	341	12602	2952
max-Xmax	89	4	761	135	489	31	1367	186
mid-range	3543	104	23872	4548	20622	996	80965	11676

Table 3. Number of rules according to the class value in `Xmax`, `max-Xmax` and `mid-range`

The number of rules varies according to the datasets (thus, the used discretization method) and the largest number of rules is achieved with `mid-range`. With $\gamma = 7$ and $\delta = 2$, there are 13 (resp. 40) times more rules extracted from `mid-range` than from `Xmax` (resp. `max-Xmax`). Surprisingly, `mid-range` has not the highest density (see Table 2), which ought to be an explanation. Currently, we have no sound explanation of this record. Note that an identical phenomenon is reported in [2]. In the following, we focus on the rules extracted with $(\gamma, \delta) \in \{(5, 2), (4, 1)\}$. This leads to a relevant number of rules in each discretized dataset and for each class value.

Comparison of sets of rules. In a certain sense, the rules belonging to several sets of rules computed on the different datasets are independent of the encoding of the overexpression. Let us have a look on such rules. For instance, with $\gamma = 4$ and $\delta = 1$, there are 1496 common rules between Xmax and max-Xmax, 768 between max-Xmax and mid-range, 4398 rules between Xmax and mid-range and only a single rule is shared by the three datasets. Table 4 is an excerpt of these rules. As a common rule may have different frequency and confidence values with regard to the dataset, this is expressed in Table 4 by the use of several lines for a unique rule. This table gives the rule shared by the three datasets, which could be high of interest, $\{8091 \quad 19351\} \rightarrow$ no cancer. This rule appears in situations 12, 16, 38 and 84.

common to max-Xmax and mid-range					
Body	Conclusion	Exceptions	Frequency	Confidence	Data
7259 14143	cancer	0	4	1	max-Xmax
		1	6	0.83	mid-range
11695 17436	cancer	1	5	0.8	both
12719 19258	cancer	1	4	0.75	max-Xmax
		1	7	0.86	mid-range
22218 26894	cancer	0	4	1	both
6756 26019	no cancer	1	4	0.75	max-Xmax
		1	6	0.83	mid-range
13954 27489	no cancer	0	4	1	both

common to Xmax and mid-range					
Body	Conclusion	Exceptions	Frequency	Confidence	Data
566 11119	cancer	1	4	0.75	Xmax
		1	5	0.8	mid-range
1525 9002	cancer	0	4	1	both
9739 27441	cancer	1	4	0.75	Xmax
		1	5	0.8	mid-range
11119 21930	cancer	0	4	1	both
2467 20091	no cancer	1	4	0.75	both
20091 27139	no cancer	1	4	0.75	both

common to Xmax, max-Xmax and mid-range					
Body	Conclusion	Exceptions	Frequency	Confidence	Data
8091 19351	no cancer	1	4	0.75	all

Table 4. Characteristics of the common rules when $\gamma = 4$ and $\delta = 1$.

Potential relevant rules. As said in introduction, biologists are interested in associations on genes (e.g., synexpression groups), so we do not examine trivial rules such as rules with only one tag in their body. Due to space limitation, we present only a selection of rules with at least two tags in their body and a rather high confidence and frequency (the latter point is to get associations conveying sound relationships). Following these pragmatic selection criteria, with $\gamma = 4$ and $\delta = 1$, the rules coming from mid-range seem the more interesting and some of them are presented in Table 5. Due to the lack of space, we provide the description of tags (identification number, sequence and description) only for the tags which appear the most frequently in our results (Table 6). Some tags are identified by several genes: their identifications are separated by ";".

Table 7 gives examples of selected rules with $\gamma = 5$ and $\delta = 2$. The rules have a higher frequency in mid-range than in Xmax and max-Xmax. In max-Xmax, the confidence of the extracted rules is low.

`mid-range`				
Body	Conclusion	Exceptions	Frequency	Confidence
11115 19811	`cancer`	1	13	0.92
5961 11115	`cancer`	0	12	1
8279 23600	`cancer`	1	12	0.92
10960 11115	`cancer`	1	12	0.92
11115 20766	`cancer`	1	12	0.92
4602 7259 18882	`cancer`	1	10	0.9
4602 7259 24686	`cancer`	1	10	0.9
8255 11115 19811	`cancer`	1	10	0.9
4602 7259 20461	`cancer`	1	9	0.89
4602 7259 25202	`cancer`	1	9	0.89
4602 18882 24686	`cancer`	1	9	0.89
4287 4602 7818	`cancer`	1	8	0.88
4287 4602 19811	`cancer`	**1**	**8**	**0.88**
4602 7259 19734	`cancer`	1	8	0.88
4602 24686 25202	`cancer`	1	8	0.88
4602 25128 25202	`cancer`	1	8	0.88
7259 12667 16807	`cancer`	1	8	0.88
8255 11115 13642	`cancer`	0	8	1
8255 11115 26846	`cancer`	1	8	0.88
8255 19811 26846	`cancer`	1	8	0.88
22619 25202 26846 27358	`cancer`	1	5	0.8
16786 26715	`no cancer`	1	7	0.86
22129 25356	`no cancer`	1	7	0.86
22129 27414	`no cancer`	1	7	0.86
22647 25356	`no cancer`	1	7	0.86
1722 25202 26715	`no cancer`	1	6	0.83

Table 5. Examples of potential relevant rules in `mid-range` with $\gamma = 4$ and $\delta = 1$

Potential relevant tags. Few tags (e.g., 4602, 8255, 11115, 22129) clearly arise in many rules concluding on `cancer`. They may have an influence on the development of this disease. It is interesting to note that the frequencies of these tags strongly varies from one class to another. For example, the tag 11115 appears 28.7 times more in rules characterizing `cancer` than `no cancer` (with $\gamma = 4$ and $\delta = 1$). The tag 11115 is identified as GPX1. The expression of GPX1 has been found in various studies to be correlated with cancerous situations [9, 11]. On the contrary, the tag 22129 appears 22 times more in rules concluding on `no cancer` than concluding on `cancer`. It might mean that this tag is related to normal development.

A biologist's point of view . The rule $\{4287 \quad 4602 \quad 19811\} \rightarrow$ `cancer` (in bold font in Table 5) extracted from `mid-range` with $\gamma = 4$ and $\delta = 1$ seems particularly interesting. In 8 situations (20, 37, 45, 46, 48, 51, 53, 76), the association of the three tags 4287, 4602 and 19811 leads to cancer. The confidence of the rule is 0.88. These tags are described in Table 6. On these three tags, two of them are identified as ribosomal proteins and one is identified as a transmembrane protein. The interest of such a rule lies in its originality: NIFIE14 is a recently discovered protein. The role of transmembrane proteins in cancer developpement is well studied: a perturbation in the cellular communication mediated via transmembrane proteins is often invoked as one major cause of cancers [7]. Moreover, ribosomal proteins are more and more found linked with tumourous. For example, one recent study has shown that the inhibition of the phosphorylation of the ribosomal protein S6 combined with other inhibitions was a possible way for treatment of cancerous

Number	Sequence	Description
4287	AGCTCTCCCT	RPL17 CDNA sequence BC022357; PIGK Phosphatidylinositol glycan, class K
4602	AGGCTACGGA	Similar to ribosomal protein L13a, 60S ribosomal protein L13a, 23 kD highly basic protein
8255	CATCCAAAAC	HNRPH1 Heterogeneous nuclear ribonucleoprotein H1 (H)
11115	CTCTTCGAGA	GPX1 Glutathione peroxidase 1
19811	GTTGCTGCCC	NIFIE14 Seven transmembrane domain protein
22129	TCAGAGAATA	SLC25A22 Solute carrier family 25 (mitochondrial carrier: glutamate), member 22; IRS2 Insulin receptor substrate 2
25202	TGTGCTAAAT	RPL34 Ribosomal protein L34; USP36 Ubiquitin specific protease 36

Table 6. Characteristics of potential relevant tags

Xmax				
Body	Conclusion	Exceptions	Frequency	Confidence
431 9002	cancer	1	5	0.8
6497 6544	cancer	1	5	0.8
18271 21701	no cancer	1	5	0.8

max-Xmax				
Body	Conclusion	Exceptions	Frequency	Confidence
3401 27230	cancer	2	5	0.6
5371 19950	cancer	2	5	0.6

mid-range				
Body	Conclusion	Exceptions	Frequency	Confidence
4602 24686	cancer	2	17	0.88
8255 11115	cancer	2	15	0.87
4602 7259	cancer	2	14	0.86
8255 19811	cancer	2	14	0.86
16306 16690 24686	cancer	2	9	0.78
7259 24686 25202	cancer	2	8	0.75
8083 8925 19811	no cancer	2	6	0.67

Table 7. Examples of potential relevant rules with $\gamma = 5$ and $\delta = 2$

cells [6]. We also notice that the 8 situations supporting this rule are quite homogeneous: five of them concern prostate libraries. The three remaining situations are divided in two situations dealing with pancreas library and one concerning cerebral tumor.

6 Conclusion

In this paper, we have proposed a new method to extract all the δ-strong characterization rules in large datasets. These geometrical dimensions are encountered in a lot of domains such as gene expression data. We have performed this approach in the large SAGE data set.

We have shown the potential impact of these rules to characterize cancer versus no cancer biological situations. Several tags (e.g., 4602, 8255, 11115, 22129) might be associated to cancer. The association of tags {4287 4602 19811} which concludes on cancer seems promising. More investigations have to be done to validate the interest of such rules.

Acknowledgements. This work has been partially funded by the ACI "masse de donnes" (MD 46, CNRS 2004-2007) Bingo (Bases de donnes INductives pour la Gn0mique).

References

[1] R. J. Bayardo. The hows, whys, and whens of constraints in itemset and rule discovery. In *proceedings of the workshop on Inductive Databases and Constraint Based Mining*, 2005.

[2] C. Becquet, S. Blachon, B. Jeudy, J.-F. Boulicaut, and O. Gandrillon. Strong-association-rule mining for large-scale gene-expression data analysis: a case study on human sage data. *Genome Biology 2002*, 3(12):research0067.1–0067.16, 2002.

[3] J.-F. Boulicaut, A. Bykowski, and C. Rigotti. Approximation of frequency queries by means of free-sets. In *The Fourth European Conference on Principles and Practice of Knowledge Discovery in Databases*, Lyon, 2000.

[4] J. F. Boulicaut, A. Bykowski, and C. Rigotti. Free-sets: a condensed representation of boolean data for the approximation of frequency queries. *Data Mining and Knowledge Discovery journal*, 7(1):5–22, 2003.

[5] B. Crémilleux and J.-F. Boulicaut. Simplest rules characterizing classes generated by delta-free sets. In *22nd Int. Conf. on Knowledge Based Systems and Applied Artificial Intelligence (ES'02)*, pages 33–46, Cambridge, UK, December 2002.

[6] J. Graff et al. The protein kinase cbeta-selective inhibitor, enzastaurin (ly317615.hcl), suppresses signaling through the akt pathway, induces apoptosis, and suppresses growth of human colon cancer and glioblastoma xenografts. *Cancer Research*, 65(6):7462–7469, 2005.

[7] K. Hajra and E. Fearon. Cadherin and catenin alterations in human cancer. *Genes Chromosomes Cancer*, 34(3):255–268, 2002.

[8] C. Hébert and B. Crémilleux. Mining frequent delta-free patterns in large databases. In *proceedings of the 8th International Conference on Discovery Science (DS'05)*, Lecture notes in artificial intelligence, Singapore, 2005. Springer.

[9] R. Korotkina et al. Activity of glutathione-metabolizing and antioxidant enzymes in malignant and benign tumors of human lungs. *Bulletin of Experimental Biology and Medicine*, 133(6):606–608, 2002.

[10] H. Mannila and H. Toivonen. Levelwise search and borders of theories in knowledge discovery. *Data Mining and Knowledge Discovery*, 1(3):241–258, 1997.

[11] M. Nasr, M. Fedele, K. Esser, and A. Diamond. Gpx-1 modulates akt and p70s6k phosphorylation and gadd45 levels in mcf-7 cells. *Free Radical Biology and Medicine*, 37(2):187–195, 2004.

[12] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Efficient mining of association rules using closed itemset lattices. *Data Mining and Knowledge Discovery journal*, 24(1):25–46, 1999.

[13] F. Rioult, J.-F. Boulicaut, B. Crémilleux, and J. Besson. Using transposition for pattern discovery from microarray data. In M. J. Zaki and C. C. Aggarwal, editors, *proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD'03)*, pages 73–79, San Diego, CA, 2003.

[14] F. Rioult, C. Robardet, S. Blachon, B. Crémilleux, O. Gandrillon, and J.-F. Boulicaut. Mining concepts from large sage gene expression matrices. In *proceedings of the second International Workshop on Knowledge Discovery in Inductive Databases (KDID'03) co-located with the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases PKDD'03*, pages 107–118, Dubvronik, Croatia, 2003.

[15] M. J. Zaki. Generating non-redundant association rules. In *proceedings of the 6th International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD'00)*, pages 34–43, Boston, MA, 2000.