Statistical inference and data mining: false discoveries control

Stéphane Lallich[1] and Olivier Teytaud[2] and Elie Prudhomme[1]

[1] Université Lyon 2, Equipe de Recherche en Ingénierie des Connaissances
5 Avenue Pierre Mendès-France, 69676 BRON Cedex - France
stephane.lallich@univ-lyon2.fr, eprudhomme@eric.univ-lyon2.fr
[2] TAO-Inria, LRI, CNRS-Université Paris-Sud, bat. 490
91405 Orsay Cedex France
teytaud@lri.fr

Summary. Data Mining is characterised by its ability at processing large amounts of data. Among those are the data "features"- variables or association rules that can be derived from them. Selecting the most interesting features is a classical data mining problem. That selection requires a large number of tests from which arise a number of false discoveries. An original non parametric control method is proposed in this paper. A new criterion, UAFWER, defined as the risk of exceeding a pre-set number of false discoveries, is controlled by BS_FD, a bootstrap based algorithm that can be used on one- or two-sided problems. The usefulness of the procedure is illustrated by the selection of differentially interesting association rules on genetic data.

Key words: feature selection, multiple testing, false discoveries, bootstrap.

Introduction

The emergence of Data Mining is linked to the increase in storage capacity and computing power of computers. It is also linked to the increased number of information systems and to the automation of data collection. This emergence follows from the development of Tukey's Exploratory Data Analysis [17] and of Benzecri's *Analyse des Données* [4], while integrating lessons from databases and artificial intelligence. Whereas statistics organises data collection and analysis for an objective set a priori, data mining extracts relevant information a posteriori from the collected data. This creates some difficulties for statistical inference when working in a data mining context. More specifically, the statistical control of false discoveries when performing a large number of tests is of interest here.

The paper is organized as follows. First, we analyze the specificities of Data Mining which impede the application of statistical inference techniques

(Sect. 1). The problem of controlling the false discoveries with multiple tests will then be reviewed (Sect. 2), and *BS_FD*, a non parametric method to control the number of false discoveries will be introduced (Sect. 3). In the last section, we show how *BS_FD* allows the selection of the most differentially interesting association rules from a gene expression micro-array data (Sect. 4).

1 Data Mining Specificities and Statistical Inference

Data Mining is typically used on corporate databases, yielding large volumes of data, individuals or variables. Those databases ore often populated by automated systems (e.g. transactional databases). Moreover, the complexity of the data (sound, image, numbers, and text) contributes to the multiplication of the number of variables (e.g. medical record databases).

The large number of individuals (n) makes algorithms with complexity linear in n appealing and introduces the problem of selecting individuals. Selection can be done by mere sampling [5] or by reducing the learning set [20]. From a theoretical point of view, high-dimensional data present several weird specificities under the *i.i.d* hypothesis, as quoted by Verleysen [19]: curse of dimensionality, concentration of measures, empty space phenomenon. In most cases, the real data are located near a manifold of smaller dimension than the number of variables. Variable selection is then an important task of data mining.

The tools of statistical inference can be used at every step of data mining: (1) to detect outliers and/or to select the relevant variables during data preparation; (2) to control the learning process, especially step-wise algorithms; (3) to validate results under supervised learning, or to assess the stability of the results under unsupervised learning. These tools, and the p-values, can be used for statistical testing, or as selection criteria. Given the specificities of data mining, new problems arise:

- **overfitting:** When the model fits the learning data too well, in part due to its complexity, it incorporates some amount of sampling variability, which reduces its performance when generalising to new data. The reason for the underperformance is often that standard statistical inference formulae are used on an optimized empirical result. Cross-validation or using distinct learning and testing sets often solves that problem.
- **status of records:** The status of the individuals is not always clear and this impedes the validation of the inferred results. Are the data a sample? If so, what sampling plan was used? It is important that the validation techniques accounts for the sampling plan (see [5] for cross-validation of results obtained from a cluster sample). Are they rather an exhaustive population? If so, can it be considered a sample of a super-population? Rather than working on a population, wouldn't it be better to work on a sample (see [5] for the case of decision trees)?

- **large number of records:** All usual tests become significant when the sample is large enough. The null hypothesis is rejected by slightest of differences. It is just like everything was happening under a microscope. P-values, particularly, become extremely small, which makes comparisons difficult. Facing this problem, Lebart et al. [13] suggested to use the test values instead of p-values, which provide an equivalent information, but easier to use and interpret. Briefly, the test value associated with a probability level p is a standardized normal variable u corresponding to this level: for example, a test value of $u = 1.96$ will correspond to bilateral probability level $p = 0.05$. In a more drastic way, Morineau and Rakotomalala [14] propose an alternative criterion, *TV100*, a modification to the test value. The *TV100* test value is calculated as if the empirical results had been obtained from a sample of size 100.
- **multiple testing:** The multiplicity of tests inflates the number of Type I errors (false discoveries). This problem is often encountered when selecting relevant attributes (e.g. selection of most differentially expressed genes from micro-array data [6]) or when comparing the efficiency of several algorithms [8].

The remainder of this paper addresses this latter problem.

2 Validation of Interesting Features

2.1 Searching Interesting Features

The problem of selecting interesting features is encountered in supervised learning (e.g. selection of discriminant variables) and unsupervised learning (e.g. selection of interesting association rules). From a sample, m features (variables or rules) are examined in turn, and a decision must be made with respect to their selection.

Discriminant variables are those whose behaviour, in the real world, changes with classification variables. The discriminating power of a variable is commonly assessed by comparing its mean value conditional to a class (Student's t test for two classes, ANOVA, otherwise), by comparing its average ranks (Mann-Whitney rank test for two classes, Friedman's, otherwise), or using a permutation test. The null hypothesis, noted H_0, assumes that the means are equal, stating the lack of discriminating power of the variable of interest. For example, the two-class situation is written as $H_0 : \mu_1 - \mu_2 = 0$, where μ_i is the theoretical mean of the variable of interest in class $i, i = 1, 2$.

Association rules were originally developed in the context of transactional databases. Each record is a transaction, where the fields are the possible items of the transaction. Considering two sets of items (itemset) A and B having no common item, an association rule $A \rightarrow B$ means that if somebody buys the items of A, then they probably will buy the items of B. The rule $A \rightarrow B$ has

support s if $s\%$ of transactions contain both A and B. The rule $A \rightarrow B$ holds with confidence c if $c\%$ of transactions that contain A also contain B. Let n be the number of transactions, n_x the number of transactions containing a given itemset X, $p_x = n_x/n$ the proportion of transactions containing X and π_x the corresponding real world proportion. Then, $s = p_{ab}$ and $c = p_{b/a}$. More generally, A and B can be conjunctions of binary attributes having no common attributes. Following *Apriori* [1], the founding algorithm, support-confidence extraction algorithms exhaustively seek the association rules whose support and confidence exceed some user-defined thresholds. A set $\mathcal{R}$ of admissible rules, of cardinality $m = \#\mathcal{R}$ is then obtained. An admissible rule is interesting if the consequent occurs more often when the antecedent has in effect occurred. The null hypothesis of independence between A and B, noted $H_0 : \pi_{b/a} = \pi_b$, must be tested against the alternative of positive dependence $H_1 : \pi_{b/a} > \pi_b$.

In both situations, selecting variables or selecting rules, the selection is the result of m replications of the test of H_0 at the predetermined level α_0. This multiplicity of tests inflates the number of false discoveries (features wrongly selected). In effect, if m tests are developed, each with a probability of Type I error set at α_0, even if no feature is truly interesting, the procedure automatically creates $m\alpha_0$ false discoveries.

2.2 Constructing Multiple Tests

Significance Test

The "interestingness" of a feature f is assessed by a measure $M(f)$. For example, the measure may be the difference of means in the case of two-class discriminant variables selection, or confidence in case of selection of association rules. The feature is said to be significant under M with respect to μ_0 if $M_{\text{obs}} = M(f)$ is significantly far from some preset value μ_0. The alternative hypothesis may be bilateral ($H_1 : \mu \neq \mu_0$), or unilateral (the more often right-sided hypothesis, $H_1 : \mu > \mu_0$). H_0 is rejected whenever M_{obs} is too far from H_0 in the direction of H_1, with a Type I error risk set at $\alpha = \alpha_0$. The p-value for M_{obs} is computed as the probability of obtaining a value as exceptional as M_{obs} in direction of H_1, assuming H_0 is true. The feature is selected if the p-value for M_{obs} is less than α_0. Obviously, this requires the knowledge of the distribution of $M(f)$ under H_0 or the estimation of p-values by resampling.

Risk and Type I Error

The identification of the significant features under M among the m features extracted from a database requires m tests. This raises the problem of false discoveries, a recurrent problem in data mining. If m uninteresting features are tested at the level α_0, then $m\alpha_0$ features will mechanically be erroneously selected. For example, with $\alpha_0 = 0.05$, and a base of extracted features comprising $m = 10,000$ features, even if all were non-significant, 500 features would mechanically be selected!

Table 1. Synthesis of the results of m tests

Reality \ Decision	Acceptation	Reject	Total
H_0 true	U	V	m_0
H_1 true	T	S	m_1
Total	W	R	m

The fundamental idea of Benjamini and Hochberg [2] is to consider the number of errors over m iterations of the test, rather than the risk of being wrong on one test. From Table 1 (where an upper case represents observable random variables and lower case are fixed yet unknown quantities m_0 and m_1), one can derive several indicators. The two most common are described next, *FWER* (Family Wise Error Rate) and *FDR* (False Discovery Rate).

FWER is the probability of erroneously rejecting H_0 at least once,

$$FWER = P(V > 0) \ .$$

The well-known Bonferroni correction, that is setting the risk at $\frac{\alpha_0}{m}$ so that the *FWER* be α_0, is not a good solution for two reasons:

- *FWER* is in fact not controlled, $\frac{\alpha_0}{m} \leq FWER \leq \alpha_0$, and equal to α_0 only when the features are mutually independent;
- *FWER* is conservative, thus increasing the risk of a Type II error, that is, not finding an interesting feature.

One needs a criterion less stringent than *FWER* for a large number of tests and to exert some form of control, especially when the tests are not independent. The authors proposed the *User Adjusted Family Wise Error Rate*, an original and more flexible variant [12] which allows V_0 false discoveries,

$$UAFWER = P(V > V_0) \ .$$

It can be controlled using a bootstrap based algorithm (Sect. 3.3).

Other quantities using the expectation of V, the number of false discoveries, possibly standardised, have been proposed to remedy the difficulties inherent to *FWER*. The best known is *FDR* [2], the expected proportion of erroneous selections among the selected features. When $R = 0$, define $\frac{V}{R} = 0$, that is, $FDR = E(Q)$, where $Q = \frac{V}{R}$ if $R > 0$, 0 otherwise. Then

$$FDR = E(\frac{V}{R} \mid R > 0)P(R > 0) \ .$$

Benjamini and Liu [3] proposed a sequential method for the control of *FDR* under the assumption of independence. The p-values are examined in increasing order and the null hypothesis is rejected if the p-value at hand $p_{(i)}$ is less than $\frac{i\alpha_0}{m}$. This procedure ensures that $FDR = \frac{m_0}{m}\alpha_0$ under independence. It

is compatible with positively dependent data. Storey [15] proposed the $pFDR$, a variation on FDR, using the knowledge that H_0 has been rejected at least once, $pFDR = E(\frac{V}{R} \mid R > 0)$. At the cost of a fixed proportion of erroneous selections, these quantities are less severe, thus augmenting the probability of selecting an interesting feature (increased power). One has $FDR \leq FWER$ and $FDR \leq pFDR$, hence $FDR \leq pFDR \leq FWER$ when m is large, because $P(R > 0)$ goes to 1 as m increases. The problem of controlling these criteria is resolved, in the literature, by the use of p-values. A remarkable summary of can be found in Ge et al. [8].

3 Controlling UAFWER Using the *BS_FD* Algorithm

3.1 Notations

- $\mathcal{C}$: set of cases; $n = \#\mathcal{C}$; p: number of attributes;
- $\mathcal{F}$: base of admissible features with respect to some predefined measures; $m = \#\mathcal{F}$;
- M: measure; $\mu(f)$: theoretical value of M for feature f; $M(f)$: empirical value of M for f on $\mathcal{F}$;
- V: number of false discoveries; δ: risk level of the control procedure, with V_0 the number of false discoveries not to be exceeded given δ; $\mathcal{F}^*$ a subset of $\mathcal{F}$ comprising the significant features as determined by M and μ_0.

3.2 Objective

The objective is to select the features f of $\mathcal{F}$ that are statistically significant as measured by M, meaning that $M(f)$ is significantly larger than $\mu_0(f)$, the expected value of $M(f)$ assuming H_0 true. The authors have suggested various algorithms that use the VC-dimension and other tools of statistical learning so that 100% of the identified features be significant for a given α [16]. A bootstrap-based algorithm *BS* was also proposed for the same purpose [11]. Experience has shown that this approach might be too prudent, therefore not powerful enough. Allowing a small number of false discoveries, after Benjamini's work (Sect. 2.2), the authors propose *BS_FD*, an adaptation of *BS* that controls the number of false discoveries. *BS_FD* selects features so that $UAFWER = P(V > V_0)$, which ensures that the number of false discoveries does not exceed V_0 at the level δ. The algorithm guarantees that $P(V > V_0)$ converges to δ when the size of the sample of cases increases.

3.3 Unilateral *BS_FD*

Given $\mathcal{C}$, $\mathcal{F}$, and M, $\mu(f) > \mu_0(f)$ is guaranteed by setting $\mu(f) > 0$, without loss of generality simply by shifting $\mu(f)$ to $\mu(f) - \mu_0(f)$. V_0 false discoveries are allowed at risk δ.

1. *Empirical assessment.* All features of $\mathcal{F}$ are measured using M on the set of cases $\mathcal{C}$, creating the $M(f), f \in \mathcal{F}$.
2. *Bootstrap.* The following operations are repeated l times:
   a. Sample with replacement and equal probability m cases from $\mathcal{C}$, thus creating $\mathcal{C}'$, $\#\mathcal{C}' = \#\mathcal{C}$. Some cases of $\mathcal{C}$ will not be in $\mathcal{C}'$ while some others will be there many times. All features are measured on $\mathcal{C}'$ using M, creating the $M'(f)$, $f \in \mathcal{F}$.
   b. Compute the differences $M'(f) - M(f)$, then compute $\varepsilon(V_0, i)$, the smallest value such that $\#\{M'(f) > M(f) + \varepsilon(V_0, i)\} \leq V_0$. Hence, $\varepsilon(V_0, i)$ is the $(V_0 + 1)^{\text{st}}$ largest element of the $M'(f) - M(f)$, during the i^{th} iteration, $i = 1, 2...l$.
3. *Summary of bootstrap samples.* There are l values $\varepsilon(V_0, i)$. Compute $\varepsilon(\delta)$, $(1-\delta)^{\text{th}}$ quantile of the $\varepsilon(V_0, i)$: that is, $\varepsilon(V_0, i)$ was larger than $\varepsilon(\delta)$ only $l\delta$ times in l.
4. *Decision.* Keep in $\mathcal{F}^*$ all features f such that $M(f) > \varepsilon(\delta)$.

3.4 Bilateral *BS_FD*

The procedure *BS_FD* can easily be extended to bilateral tests. Let $V_{0\text{l}}$ and $V_{0\text{r}}$, the number of false discoveries tolerated at the left and right, respectively, be such that $V_{0\text{l}} + V_{0\text{r}} = V_0$ for a risk δ. The idea behind *BS_FD* is to estimate using a bootstrap by how much M can move to the left or to the right still maintaining $V_{0\text{l}}$ false discoveries to the left and $V_{0\text{r}}$ false discoveries to the right, at the global level δ. It is then sufficient to modify steps 2.b., 3. and 4. of *BS_FD* like so :

2.*b* Set $V_{0\text{l}}$ and $V_{0\text{r}}$ such that $V_{0\text{l}} + V_{0\text{r}} = V_0$. At the i^{th} iteration of the bootstrap, compute $\varepsilon(V_{0\text{r}}, i)$, the smallest number such that $\#\{M'(f) - M(f) > \varepsilon(V_{0\text{r}}, i)\} \leq V_{0\text{r}}$. Thus, $\varepsilon(V_{0\text{r}}, i)$ is the $(V_{0\text{r}} + 1)^{\text{st}}$ largest element of the $M'(f) - M(f)$. Then, compute $\varepsilon(V_{0\text{l}}, i)$, the smallest number such that $\#\{M'(f) - M(f) < -\varepsilon(V_{0\text{l}}, i)\} \leq V_{0\text{l}}$. Then, $\varepsilon(V_{0\text{l}}, i)$ is the $(V_{0\text{l}} + 1)^{\text{st}}$ largest element of the $M(f) - M'(f)$.

3. *Summary of bootstrap samples.* At the completion of the l bootstrap iterations, l pairs have been created, $(\varepsilon(V_{0\text{l}}, i), \varepsilon(V_{0\text{r}}, i)), i = 1, 2...l$. Compute $\varepsilon(\delta) = (\varepsilon(V_{0\text{l}}), \varepsilon(V_{0\text{r}}))$, the $(1-\delta)$ quantile of the $(\varepsilon(V_{0\text{l}}, i), \varepsilon(V_{0\text{r}}, i))$, where $(a, b) > (c, d) \Leftrightarrow (a > c)$ et $(b > d)$. Only $l\delta$ times in l was $(\varepsilon(V_{0\text{l}}, i), \varepsilon(V_{0\text{r}}, i))$ larger than $(\varepsilon(V_{0\text{l}}), \varepsilon(V_{0\text{r}}))$.
4. *Decision.* Keep in $\mathcal{F}^*$ all features f of F such that $M(f) < -\varepsilon(V_{0\text{l}})$ or $M(f) > \varepsilon(V_{0\text{r}})$.

At step 3., there are many possible maxima as the order is not total. Among the many maxima, it is suggested to choose that which maximises the number of discoveries; stated in a different manner, this choice maximises the power of the test. This solution is both efficient and flexible, but rather hard to implement.

A different solution, easier to implement but less powerful, is to execute step 2. with $V_{0\mathrm{l}} = V_{0\mathrm{r}} = \frac{V_0}{2}$ and to compute the corresponding $\varepsilon(V_{0\mathrm{l}}, i)$ and $\varepsilon(V_{0\mathrm{r}}, i)$. The quantities $\varepsilon(V_{0\mathrm{l}})$, the $1 - \frac{\delta}{2}$ quantile of the $\varepsilon(V_{0\mathrm{l}}, i)$, and $\varepsilon(V_{0\mathrm{r}})$, the $1 - \frac{\delta}{2}$ quantile of the $\varepsilon(V_{0\mathrm{r}}, i)$ are then obtained by bootstrap. There is yet another possibility for step 3. Define $\varepsilon(\delta)$ as the $(1 - \delta)$ quantile of the set of $2l$ values $\varepsilon(V_{0\mathrm{l}}, i)$ and $\varepsilon(V_{0\mathrm{l}}, i)$, where $i = 1, 2, ..., l$. $\mathcal{F}^*$ retains all features f of F such that $M(f) < -\varepsilon(\delta)$ or $M(f) > \varepsilon(\delta)$. The difficulty is that the same iteration i can give both $\varepsilon(V_{0\mathrm{l}}, i)$ and $\varepsilon(V_{0\mathrm{r}}, i)$ exceeding $\varepsilon(\delta)$. Applying *BS_FD* on the $|M'(f) - M(f)|$ (or on the bilateral p-values when they are known) can not be considered a solution, as this procedure masks certain variations between the original sample and the bootstrap replicates.

3.5 Rationale

Bootstrap methods [7] approximate the distance between the empirical and true distributions by the distance between the bootstrap and empirical distributions. At the i^{th} bootstrap iteration, there are V_0 features whose evaluation augments by more than $\varepsilon(V_0, i)$. Given the definition of $\varepsilon(\delta)$, the number of features whose evaluation augments by more than $\varepsilon(\delta)$ is larger than V_0 in a proportion δ of the l iterations. Consequently, selecting features for which $M(f)$ exceeds $\varepsilon(\delta)$, one is guaranteed to have at most V_0 false discoveries at the risk level δ.

Moreover, bootstrap-based methods have solid mathematical foundations [9] which require a clearly posed question. Formally, the objective is that the distribution function of the number of features such that $\mu(f) < 0$ while $M(f) > \epsilon$, be at least $1 - \delta$ for V_0. One gets $\#\{\mu(f) \leq 0 \text{ et } M(f) > \epsilon\} \leq \#\{M(f) \geq \mu(f) + \epsilon\}$. Theorems on bootstrap applied to a family of functions verifying the minimal conditions [18] yield the approximation of this quantity by $\#\{M'(f) \geq M(f) + \epsilon\}$, which serves as a basis for $\varepsilon(V_0, i)$ and $\varepsilon(\delta)$ described in this section.

3.6 Extension to Multiple Measures

In practice, more that one measure will be of interest to evaluate feature interestingness. The extension of *BS_FD*, noted *BS_FD_mm*, is achieved by using as a summary measure the minimum of the various measures. Hence, for 3 measures M_1, M_2 and M_3, one considers $M(f) = min\{M_1(f), M_2(f), M_3(f)\}$. Using *BS_FD_mm* on M at the level δ will select features which comply with M_1, M_2 and M_3, at level δ. Risk of Type II errors can be optimised by working with Hadamard differentiable transformations of the M_i that will make the measures homogenous, for example, p-values or reductions, through standardisation.

3.7 Complexity of *BS_FD*

The complexity of *BS_FD* is proportional to $l \times m \times n$, assuming that the random number generator operates in constant time. In effect, the complexity of the search for the k^{th} largest element of a table is proportional to the size of the table. The value of l must be large enough so that the finiteness of l impedes not the global reliability, and be independent of both m and n. The algorithm is globally linear in $m \times n$, to a constant l linked to the bootstrap.

4 Experimentation

4.1 Introduction

The data used here are from Golub et al. [10]. They represent the level of expression of $7,129$ genes in 62 tissue samples of Acute Myeloid Leukemia (AML, 34 tissue samples) or Acute Lymphoblastic Leukemia (ALL, 28 tissue samples), two types of human cancer. They are available as standardised Affymetrix data. Thus, for a gene and a tissue sample, the level of expression and its presence or absence in the tissue sample are known. Moreover, the class of each tissue sample is known. Rules of the type "if gene A is present, then it is very likely that gene B be present as well" are of interest. Here, we seek rules that are differentially interesting, that is, relevant to a class and not for another.

4.2 Notation

- n: the number of tissue samples; n_i: the number of tissue samples in class i; p: the number of genes.
- D^i: a $n_i \times p$ Boolean matrix. $D^i_{jk} = 1$ if tissue j expresses gene k, 0 otherwise.
- p_{a_i}: prevalence of the antecedent on D^i and p_{b_i} the prevalence of the consequent on D^i.
- $Sup_i(r)$ and $Conf_i(r)$: the support and confidence for rule r on D^i; m: the number of rules examined.

4.3 Process

Differentially interesting rules are identified by a two-step process:

a. Selection of differentially expressed genes.

These are genes more frequent on average in one class than in another. By limiting the search for rules to those genes, the number of rules is reduced, and the search is focused on the best explanatory variables of the class variable.

These genes are determined by computing for each the p-value of a Student's t test. The FDR control procedure [3] (Sect. 2.2) is used to identify the genes for which the frequency in classes ALL and AML are significantly different, at the 0.05 level. Then, let p be the number of such genes and D^i the matrix reduced to those only genes.

b. Selection of differentially interesting rules

- *Rule extraction.* Build $\mathcal{R}_i$, the set of association rules of D^i such that $\forall r \in \mathcal{R}_i,\ Sup_i(r) > 0.25$ and $Conf_i(r) > 0.5$, using the Apriori algorithm.
- *Filtering admissible rules.* Build $\mathcal{R} = \{r \in \mathcal{R}_1 \cup \mathcal{R}_2 \mid\ p_{a_1} > 0.5 \text{ and } p_{a_2} > 0.5\}$. This ensures that each rule is meaningful in both D^1 and D^2 by assuring that the antecedent is present in at least 50% of the cases for both classes.
- *Measure of interestingness.* $\forall r \in \mathcal{R},\ M(r) = Conf_1(r) - Conf_2(r)$. A rule is interesting if it has high confidence for the tissue samples of one class and low confidence on the other class. The rules that characterise a class are thus privileged.
- *Selecting differentially interesting rules.* The set $\mathcal{R}^*$ of truly interesting rules under M is determined by the application of the bilateral *BS_FD* ($V_0 = 10, \delta = 0.05$) on $\mathcal{R}$.

4.4 Results and Interpretation

By applying FDR, the number of genes is reduced from $7,129$ to 980. From these 980 discriminating genes, $174,412$ admissible rules (as defined above) are obtained. The bilateral BS_FD procedure (equal shares of V_0 and δ between the left and right tails) identifies 799 differentially interesting rules. In our case, the bilateral *BS_FD* selects the rules for which $M(r) < -0.76$ or $M(r) > 0.74$. Among those, 235 are characteristic of the AML class and 564 of the ALL class.

Inspection of these rules shows that they contain only 26 different consequents (table 2). For each one, the table lists the name, the number of rules where it appears (#), the class for which the rules are valid (Class), its probability of occurrence in each class (p_{b_1} and p_{b_2}) and its rank when the genes are sorted by increasing p-values (Rank).

These consequents correspond to differentially expressed genes. For on the two tissue types, they are strongly co-expressed with genes (the antecedents) themselves differentially expressed (for example, gene M31211-s is co-expressed with 228 other genes on tissues of the ALL class). Conversely, for the other tissue type, the co-expressions do not occur, though the antecedents are still present (rules are such that $p_{a_1} > 0.5$ and $p_{a_2} > 0.5$). Their interest is not only the difference in expression levels between tissue types, but also the disappearance of the context (the set of co-expressed genes) that allows their presence. These genes are thus robust and specific indicators of the difference between ALL and AML cancers. Still, sorting in increasing order of p-values,

Table 2. Consequent of interesting rules

Name	#	**Class**	p_{b_1}	p_{b_2}	**Rank**	**Name**	#	**Class**	p_{b_1}	p_{b_2}	**Rank**
X95735	30	AML	0.09	0.82	1	M89957	2	ALL	0.68	0.11	136
M23197	4	AML	0.18	0.86	3	M12959-s	3	ALL	0.93	0.32	183
M84526	14	AML	0	0.71	4	D88270	8	ALL	0.73	0.11	197
U46499	2	AML	0.20	0.89	8	L08895	2	ALL	0.77	0.21	224
S50223	3	ALL	0.66	0.07	14	L41870	2	ALL	0.84	0.25	258
M31211-s	228	ALL	0.95	0.21	22	Z14982-rna1	1	ALL	0.59	0.07	261
U05259-rna1	16	ALL	0.82	0.18	28	U79285	9	ALL	0.73	0.14	277
M96326-rna1	185	AML	0.05	0.82	30	AB000449	10	ALL	0.77	0.14	298
M92287	30	ALL	0.91	0.25	41	D21262	11	ALL	0.82	0.18	343
L47738	93	ALL	0.86	0.14	71	D86983	2	ALL	0.73	0.11	472
U53468	1	ALL	0.59	0.04	73	X62535	15	ALL	0.86	0.21	581
M83233	19	ALL	0.73	0.07	75	U37352	1	ALL	0.84	0.32	617
M11722	103	ALL	0.91	0.18	83	X79865	5	ALL	0.84	0.21	629

these genes do not correspond to the 26 first (table 2). Their discovery adds some qualitative information (these genes are robust indicators of ALL and AML cancers and the context in which they are expressed) to quantitative information (the probability of having different levels of expression between two tissue types). Proper use of these preliminary results requires further study in partnership with biologists.

Conclusion and Perspectives

The control of the *UAFWER* criterion by *BS_FD* is a doubly original solution to the increased number of false discoveries. Accepting a pre-determined number of false discoveries, at a pre-determined risk level, this procedure allows the selection of interesting features without the computation or estimation of p-values. Bilateral and unilateral tests can be handled by the procedure. The identification of differentially interesting rules opens up a new field of research in the domain of rules of association. *BS_FD* offers the further advantage of using measures of interest more sophisticated than confidence, eventually more than one measure at once.

References

1. R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In J. B. Bocca, M. Jarke, and Zaniolo C., editors, *Proceedings of the 20th Very Large Data Bases Conference*, pages 487–499. Morgan Kaufmann, 1994.
2. Y. Benjamini and Y. Hochberg. Controlling the false discovery rate : a practical and powerful approach to multiple testing. *J. R. Stat. Soc., B*, 57:289–300, 1995.

3. Y. Benjamini and W. Liu. A step-down multiple-hypothesis procedure that controls the false discovery rate under independance. *J. Stat. Planng Inf.*, 82: 163–170, 1999.
4. J.P. Benzécri. *Analyse des Données*. Dunod, Paris, 1973.
5. J.H. Chauchat. *Echantillonnage, validation et généralisation en extraction des connaissances à partir des données*. Habilitation à diriger des recherches, Université Lyon 2, 2002.
6. J. Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, Jan 2006.
7. B. Efron. Bootstrap methods: Another look at the jacknkife. *Annals of statistics*, 7:1–26, 1979.
8. Y. Ge, S. Dudoit, and T.P. Speed. Resampling-based multiple testing for microarray data analysis. Tech. rep. 663, Univ. of California, Berkeley, 2003.
9. E. Giné and J. Zinn. Bootstrapping general empirical measures. *Annals of probability*, 18:851–869, 1984.
10. T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, 1999.
11. S. Lallich and O. Teytaud. Évaluation et validation de l'intérêt des règles d'association. *RNTI-E-1*, pages 193–217, 2004.
12. S. Lallich, E. Prudhomme, and O. Teytaud. Contrôle du risque multiple en sélection de règles d'association significatives. *RNTI-E-2 (EGC 2004)*, 2:305–316, 2004.
13. L. Lebart, A. Morineau, and M. Piron. *Statistique exploratoire multidimensionnelle*. Dunod, Paris, 1995.
14. A. Morineau and R. Rakotomalala. Critère VT100 de sélection des règles d'association. In *Conférence EGC 06*, 2006.
15. J. D. Storey. A direct approach to false discovery rates. *J. R. Statisc. Soc., Series B*, 64:479–498, 2002.
16. O. Teytaud and S. Lallich. Bornes uniformes en extraction de règles d´association. In *Conférence Apprentissage CAp´2001*, pages 133–148, 2001.
17. J.W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, 1977.
18. A. Van Der Vaart and J.A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer-Verlag Publishers, 1996.
19. M. Verleysen. *Limitations and future trends in neural computation*, chapter Learning high-dimensional data, pages 141–162. IOS Press, 2003.
20. D. R. Wilson and T. R. Martinez. Reduction techniques for instance-based learning algorithms. *Machine Learning*, 38(3):257–286, 2000.