# The 0-1 Outcomes Feature Selection Problem : a Chi-2 Approach

Cyril Duron, Jean-Marie Proth

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

# *The 0-1 Outcomes Feature Selection Problem : a Chi-2 Approach*

Cyril Duron — Jean-Marie Proth

**N° 4709**

Janvier 2003

THÈME 4


ISSN 0249-6399 ISRN INRIA/RR--4709--FR+ENG

# The 0-1 Outcomes Feature Selection Problem : a Chi-2 Approach

Cyril Duron [*†] , Jean-Marie Proth [‡ †]

Thème 4 — Simulation et optimisation
de systèmes complexes
Projet Sagep

Rapport de recherche n° 4709 — Janvier 2003 — 10 pages

**Abstract:** This paper addresses the 0-1 outcome feature selection problem. In such a problem, a set of features leads to an outcome that is 0 or 1, depending upon the values of the features. The goal is to extract subsets of features that characterize at best outcome 1. This kind of problem arises in medical analysis, quality control and, generally, in any domain that requires series of expensive tests to evaluate the state of a system.

**Key-words:** Feature selection, Data mining, Chi-2 analysis

[*] Corresponding author. E-mail : duron_ca@yahoo.fr, Phone : 00 33 (0)3 87 31 54 85 Fax : 00 33 (0)3 87 31 54 78
[†] Location : Inria / Sagep, UFR Scientifique, Université de Metz, Ile du Saulcy, 57000 Metz
[‡] Corresponding author. E-mail : proth@loria.fr, Phone : 00 33 (0)3 87 31 54 58 Fax : 00 33 (0)3 87 31 54 78

Unité de recherche INRIA Lorraine
LORIA, Technopôle de Nancy-Brabois, Campus scientifique,
615, rue du Jardin Botanique, BP 101, 54602 Villers-Lès-Nancy (France)
Téléphone : +33 3 83 59 30 00 — Télécopie : +33 3 83 27 83 19

# Le Problème de Sélection de Caractéristiques de Type 0-1 : Une approche utilisant le Chi-2.

**Résumé :** Ce papier concerne le problème de la sélection de caractéristiques de type 0-1. Dans le problème que nous considérons, un ensemble de caractéristiques conduit à une décision motivée par les valeurs de cet ensemble, et caractérisée par un 0 ou un 1. Le but est d'extraire des sous-ensembles de caractéristiques qui caractérisent au mieux la décision 1. Ce type de problème est courant dans le domaine de l'analyse de données médicales, du contrôle de qualité, ou plus généralement, de tout domaine nécessitant des séries de tests onéreux pour évaleur l'état d'un système.

**Mots-clés :** Sélection de caractéristiques, Forage de données, Analyse Chi-2

# 1 Introduction

Consider a set of tests that are performed to diagnose cancer, for instance. Each test leads either to a positive result (denoted by 1) or a negative result (denoted by 0). Thus, the results of the tests appear as a sequence of 0 or 1, and the diagnosis, based on the set of results, concludes that the person suffers from cancer (denoted by 1) or not (denoted by 0). But tests are expensive. Furthermore, the result of a test can be wrong or imprecise : it is the reason why several tests are performed. An important problem is to reduce as much as possible the number of tests to perform the diagnosis while keeping the probability of wrong diagnosis as low as possible.

The same problem arises in quality control, where the goal is to select the smallest (and cheapest) sequence of tests to evaluate efficiently the quality of the products.

The outcome feature selection problem is also of importance in :

- manufacturing, to determine the production stages that are of utmost importance for the quality of the products,
- marketing, to define the media that are the best support to promote a product or service,
- oil prospecting, to select the probes that are the most significant,
- finance, to select the ratios of importance.

Thus, the problem addressed in this paper, called feature selection problem, consists in selecting subsets of feature values that characterize at best outcome 1.

Note that :

- Several subsets of features may be used to characterize outcome 1, and the goal is to select the best one in terms of quality of the results, size of the subset and cost related to the use of this subset.
- A subset of $k$ features being selected, it may take $2^k$ sets of values. Some of these sets characterize outcome 1, other characterize outcome 0. An algorithm should be introduced to decode if a set of values characterize a 1-outcome or a 0-outcome.
- Two constraints should be considered :
  1. The probability of finding an outcome 1 when the outcome is 0 must remain under a given threshold denoted by $\eta$ in the remainder of this paper.
  2. The proportion of feature values that characterize outcome 1 covered by the subset of features under consideration is greater than a given threshold denoted by $\alpha$ in the remainder of this paper.

In other words, the goal is to extract a subset of tractable features that characterizes at best a given state of the system while keeping the probability of error under a given threshold.

Numerous research works are available in the literature to provide some solutions to the feature selection problem. In their paper, Bradley et al. (1998) propose a mathematical programming approach with a parametric objective function and linear constraints. The objective is to discriminate between the sets of features using a separating plane defined by as few features as possible. In other words, authors are trying to discriminate between the sets of feature values having an outcome equal to 1 and the ones having an outcome equal to 0 using a subset of features that is as small as possible. As they claim, "having a minimal number of features often leads to better generalization and simpler models that can be more easily interpreted". Bredensteiner and Bennett (1998) also propose a solution that is based on a linear program with equilibrium constraints. Quinlan (1990) focuses on techniques that represent classification problems in the form of decision trees. His universe is made with objects (that are sets of feature values in our terminology), these objects belonging to one of the disjoint classes. In our approach, we consider only two classes: the class characterized by outcome 1 and the class characterized by outcome 0. Boros et al. (1997) propose an approach derived from the "*Logical Analysis of Data*" (LAD) that aims at discovering "hidden structural information in data sets". The problem is presented as a set of observations, an observation being a set of feature values in the vocabulary used in this paper. A feature value is an attribute in Boros's vocabulary. It is assumed that an observation belongs either to the class of positive observations (i.e. outcome 1) or negative observations (i.e. outcome 0). Authors are interested in studying the binarization process that consists of replacing each feature value either by 0 or 1, depending on the sign of the difference between this value and a threshold that has to be defined.

The approach presented hereafter is, as far as we know, totally different from the ones mentionned above. It is based on $\chi^2$ tests that are successively used for :

- selecting the features that discriminate the 0 and 1 outcomes,

- creating the subset of features whose values can be used to characterize at best outcome 1 while avoiding the selection of outcome 1 when the real outcome is 0.

The problem is formalized in section 2. Section 3 presents the $\chi^2$ test applied to select the most significant features. Section 4 is devoted to the building of the optimal set of feature values that characterize outcome 1. In section 4.1, we use a $\chi^2$ test to select the candidate subsets of features. In section 4.2, we show how to use a set of feature values to decide if the corresponding outcome is 0 or 1. Section 4.3 shows how to compute the confidence intervals of the probabilities of successful and wrong decisions concerning the outcomes. In section 4.4, we propose a way to use the previous results to define the optimal subset of feature values and we provide a numerical example. In section 4.5, we summarize the algorithm.

Section 5 is the conclusion.

# 2 Problem formulation

The data are gathered in two matrices :

- A matrix $A(m_1 * n)$ of 0 and 1 values that corresponds to outcome 1. In this matrix, $m_1$ is the number of sets of features values and $n$ the number of features under consideration.
- A matrix $B(m_2 * n)$ of 0 and 1 values that corresponds to outcome 0. In this matrix, $m_2$ is the number of sets of features values and $n$ is, as in matrix $A$, the number of features under consideration.

We do not address the case of missing values. The objective is to find one subset of features and an algorithm. This algorithm concludes that the outcome is 1 or 0 based on the values of the selected features that :

- leads to the conclusion that the outcome is 1 when it is really 1 with a probability greater than $\alpha$. This parameter could be 0.99 by instance.
- leads to the conclusion that the outcome is 1 when it is really 0 with a probability less than $\eta$. For instance, we may choose $\eta = 0.01$.

In practice, we compute the interval of confidence of these probabilities at a given threshold (0.95 for instance) and compare the lower bound of the first interval of confidence to $\alpha$ and the upper bound of the second to $\eta$.

The strategy applied hereafter is twofold :

- selecting the features that discriminate at best the 0 and 1 outcomes,
- gathering the features corresponding to $A$-columns that can be considered as extracted from the same population.

# 3 Selection of features having a discriminatory power

Let us consider the $j^{th}$ columns of $A$ and $B$, $j \in \{1,2,\ldots,n\}$. The goal is to decide whether these two columns are issued from the same population or not.

We denote by $n_1$ ($n_1 \leq m_1$) the number of 1 values in the $j^{th}$ column of $A$ and by $n_2$ ($n_2 \leq m_2$) the number of 1 values in the $j^{th}$ column of $B$. An estimation of the probability of the 1-value associated to the union of these two columns is :

$$p = \frac{n_1 + n_2}{m_1 + m_2}$$

Thus the $\chi^2$ associated to the $j^{th}$ column of matrix $A$ is :

$$\chi^2_A = \frac{(n_1 - m_1 p)^2}{m_1 p} + \frac{(m_1 - n_1 - m_1(1-p))^2}{m_1(1-p)} = \frac{(n_1 - m_1 p)^2}{m_1 p(1-p)}$$

Similarly, the $\chi^2$ associated to the $j^{th}$ column of matrix $B$ is :

$$\chi^2_B = \frac{(n_2 - m_2 p)^2}{m_2 p(1-p)}$$

Finally, the $\chi^2$ associated to the two columns is : $\chi^2_j = \chi^2_A + \chi^2_B$ and the degree of freedom of $\chi^2_j$ is 1, taking into account the fact that $p$ is an estimated value.

Let $\chi_0^2$ be the $\chi^2$ such that :

$$Pr\{\chi_j^2 > \chi_0^2\} = (1-\alpha)$$

For instance, if $\alpha = 0.99$, then $\chi_0^2 = 6.635$ : this value is given by the $\chi^2$ table for a degree of freedom equal to one.

Thus, if $\chi_j^2$ is greater than $\chi_0^2$, we reject the fact that the $j^{th}$ column of $A$ and the $j^{th}$ column of $B$ are extracted from the same population. More precisely, the probability for this two columns to be extracted from the same population is less than $1-\alpha$, and thus, this hypothesis can be rejected.

In other words, we consider that the feature corresponding to the $j^{th}$ column of $A$ and $B$ discriminates the outcomes 0 and 1 if $\chi_j^2 > \chi_0^2$.

Indeed, $\alpha$ is chosen big enough to reduce the probability of error.

Making the same computation for $j = 1, 2, \ldots, n$ leads to a subset of features that discriminate the 0 and 1 outcomes.

# 4 Optimal subsets of features

## 4.1 Selecting the candidate subsets

At this point of the computation, we have a subset $S \in \{1, 2, \ldots, n\}$ of features whose elements have a discriminatory power. For a given $j \in S$, the bigger $\chi_j^2$, the stronger the discriminatory power. But if we consider, for instance, two features $j_1$ and $j_2$ belonging to $S$, it may happen that the columns $j_1$ and $j_2$ of matrix $A$ cannot be considered as extracted from the same population. In other words, several subsets $S_1, S_2, \ldots, S_k$, with $\bigcup_{k=1}^{K} S_k = S$, may characterize outcome 1. The goal of this section is to extract these subsets from $S$ and select the ones that leads to the lowest error ratio.

Considering features $j_1$ and $j_2$ mentioned above, the objective is to decide whether they are extracted from the same population or not. In terms of columns of matrix $A$, we have to decide whether columns $j_1$ and $j_2$ are independent from each other or not.

Let $a_{ij}$ be the element of row $i$ and column $j$ of $A$. Indeed, $a_{ij}$ is equal to 0 or 1. A pair $(a_{i,j_1}, a_{i,j_2})$ is either (0,0), (0,1), (1,0) or (1,1).

We denote by $p_{0,0}, p_{0,1}, p_{1,0}$ and $p_{1,1}$ the probabilities of these occurrences, and by :

$p_{u,.}, u \in \{0,1\}$, the probability of $u$ in column $j_{1,.}$

$p_{.,v}, v \in \{0,1\}$, the probability of $v$ in column $j_{2,.}$

Indeed, since the feature values are independant from each other : $p_{u,v} = p_{u,.} * p_{.,v}$

To apply the $\chi^2$ test, we have to estimate $p_{0,.}, p_{1,.}, p_{.,0}$ and $p_{.,1}$. The best estimation of these probabilities is : $p_{0,.} = \frac{k_{0,0}+k_{0,1}}{m_1}$, where $k_{u,v}$ is the number of pairs $(u,v)$ in columns $(j_1, j_2)$ of matrix $A$.

Similarly : $p_{1,.} = \frac{k_{1,0}+k_{1,1}}{m_1}$, $p_{.,0} = \frac{k_{0,0}+k_{1,0}}{m_1}$, $p_{.,1} = \frac{k_{0,1}+k_{1,1}}{m_1}$

The $\chi^2$ of a pair of columns $j_1$ and $j_2$ is :

$$\chi_{j_1,j_2}^2 = \frac{(k_{0,0} - m_1 p_{0,.} p_{.,0})^2}{m_1 p_{0,.} p_{.,0}} + \frac{(k_{1,1} - m_1 p_{1,.} p_{.,1})^2}{m_1 p_{1,.} p_{.,1}} + \frac{(k_{0,1} - m_1 p_{0,.} p_{.,1})^2}{m_1 p_{0,.} p_{.,1}} + \frac{(k_{1,0} - m_1 p_{1,.} p_{.,0})^2}{m_1 p_{1,.} p_{.,0}}$$

Replacing the probabilities by their estimation, we obtain :

$$\chi_{j_1,j_2}^2 = m_1 \left(\frac{k_{0,0}^2}{k_{0,.} k_{.,0}} + \frac{k_{1,1}^2}{k_{1,.} k_{.,1}} + \frac{k_{0,1}^2}{k_{0,.} k_{.,1}} + \frac{k_{1,0}^2}{k_{1,.} k_{.,0}} - 1\right) \quad (1)$$

We estimated four parameters (the probabilities) that are linked by two relations : $p_{1,.} + p_{0,.} = 1$ and $p_{.,1} + p_{.,0} = 1$

As a consequence, the number of parameters to take into account is 4-2=2. Since the probability rule concerns binary variables, we have to substract 1 from the number of parameters to take into account in order to obtain the number of degrees of freedom. Finally, the degree of freedom is 1.

The feature $j_1$ and $j_2$ are dependant if $\chi_{j_1,j_2}^2$ does not exceed the value $\chi_0^2$ such that : $Pr\{\chi_{j_1,j_2}^2 > \chi_0^2\} = 1-\alpha$ where $\alpha$ is large (0.99 for instance).

This test allows us to match features that collaborate in characterizing the 1-outcome. Applying the test to each pair $(j_1, j_2)$, $j_1, j_2 \in S$, we obtain $\chi_{j_1,j_2}^2$. We select the pairs of features that depend on each other, or in other words, that lead to $\chi_{j_1,j_2}^2 \leq \chi_0^2$.

We then consider the subsets made with three features. They are built based on this rule : $(j_1, j_2, j_3)$ is selected if and only if $(j_1, j_2)$, $(j_1, j_3)$ and $(j_2, j_3)$ have been selected in the previous step.

We can continue the extension of the set in the same way : $(j_1, j_2, j_3, j_4)$ is select if and only if $(j_1, j_2)$, $(j_2, j_3)$, $(j_3, j_4)$ $(j_1, j_3)$, $(j_1, j_4)$ and $(j_2, j_4)$ have been selected at step 2, and so on.

Assume, for instance, that 10 features have been selected for their discriminatry power.

Starting from those features, we can create $C_{10}^2 = 45$ pairs of features. After computing the $\chi^2_{j_1,j_2}$ of each pair $(j_1, j_2)$, we select the pairs whose $\chi^2_{j_1,j_2} < \chi^2_0$. From this set of selected pairs, we can derive the subsets of 3,4,...,10 elements, if any.

At this stage of the computation, we have a collection of subsets made with 2,3,...,$k$ elements. The goal is to select the *best* subset, taking into account the fact that the selection should be based on three criteria :

- the selected subset should be able to find out outcome 1 with a probability as close as possible to 1,
- the selected subset should be able to avoid selecting outcome 1 when the outcome is 0 : the probability to select outcome 1 in this case should be as close as possible to 0,
- the selected subset should be as small as possible or, alternatively, the use of this subset should be as cheap as possible.

## 4.2 The algorithm to extract the outcome

At this point of the explanation, we have several subsets of features and we associate, to each of these features, either 0 or 1, depending on the binary value that is in a majority in the corresponding column of $A$.

We also assume that, for each feature $j$, we know the probability $p_j$ that the value taken by the feature is wrong for outcome 1. If $j$ is a pathology analysis, $p_j$ is the probability that the result of analysis is wrong or meaningless for output 1.

Consider a set $E$ of $R$ features and assume that $r_1$ features take the wrong value. The probability that this set corresponds to a 1 output is :

$$Q_E = \prod_{j \in E_1} p_j . \prod_{j \in E_2} (1 - p_j) \tag{2}$$

where $E_1$ is the set of $r_1$ feature values that do not correspond to the values required by outcome 1 and $E_2$ is the set of $R - r_1$ feature values that correspond to the value required by outcome 1.

Indeed, $E_1 \cup E_2 = E$.

The same approach is used to evaluate a set of features with regard to outcome 0. Let $q_j$ be the probability that the value taken by feature $j$ is wrong, when outcome is 0. Consider a set $F$ of $R$ features and assume that $r_2$ features take the wrong value with regard to output 0, that is a value that does not correspond to output 0.

Let $F_1$ the set of $r_2$ features that are wrong with regard to output 0 and $F_2$ the set of $R - r_2$ features that fit with output 0. The probability that set $F$ corrresponds to output 0 is :

$$H_F = \prod_{j \in F_1} q_j . \prod_{j \in F_2} (1 - q_j) \tag{3}$$

and $F_1 \cup F_2 = F$.

Thus, when a set $L$ of feature values is given, we compute $Q_L$ and $H_L$. If $Q_L > H_L$, we decide that $L$ corresponds to output 1, otherwise we decide that $L$ corresponds to output 0.

The values of the probabilities $p_j$ and $q_j$ are evaluated are as follows. Assume that the column $j$ of a matrix $A$ contains $n_1$ $(n_1 \leq m_1)$ values that are different from the value required for output 1. Then we set :

$$p_j = \frac{n_1}{m_1}$$

Similarly, if the column $j$ of matrix $B$ contains $n_2$ $(n_2 \leq m_2)$ values that are different from the value required for output 0, then we set :

$$q_j = \frac{n_2}{m_2}$$

Using this approach, we are now able to decide whether a set of feature values defines an output 0 or 1.

### 4.3 Evaluation of the probabilities

For each subset, we thus have to evaluate two probabilities. Assume, for instance, that $k_1 \geq m_1$ is the number of 1-outcomes that have been found in $A$ based on the diagnosis rule presented in section 4.2. The value of $m_1$ is supposed to be *large enough*, say greater than or equal to 20. In this case, the random variable:

$$\frac{\sqrt{m_1}}{\sigma_1}(\frac{k_1}{m_1} - M_1) \tag{4}$$

follows a Gaussian distribution of mean value 0 and standard deviation 1. In this formulation, $M_1$ is the mean value of the number of 1-outcomes that have been found in $A$, and $\sigma_1$ is the related standard deviation. In our example,

$$\sigma_1 \sim \sqrt{\frac{k_1(m_1 - k_1)}{m_1(m_1 - 1)}} \tag{5}$$

Replacing $\sigma_1$ in (4) by its evaluation (5), we can say that :

$$\sqrt{\frac{m_1^2(m_1 - k_1)}{k_1(m_1 - k_1)}}(\frac{k_1}{m_1} - M_1)$$

follows a Gaussian distribution of mean value 0 and standard deviation 1. In other words :

$$P_2\{-a \leq \sqrt{\frac{m_1^2(m_1 - 1)}{k_1(m_1 - k_1)}}(\frac{k_1}{m_1} - M_1) \leq a\} = \frac{1}{\sqrt{2\Pi}} \int_{-a}^{+a} e^{-\frac{x^2}{2}} dx \tag{6}$$

Assume that we want to evaluate $M_1$ with a confidence probability of 0.95. The statistic table related to the Gaussian distribution provides $a = 1.96$.

We derive, from (6) :

$$\frac{k_1}{m_1} - 1.96\sqrt{\frac{k_1(m_1 - k_1)}{m_1^2(m_1 - 1)}} \leq M_1 \leq \frac{k_1}{m_1} + 1.96\sqrt{\frac{k_1(m_1 - k_1)}{m_1^2(m_1 - 1)}}$$

with a probability equal to 0.95. The lower and upper bounds of $M_1$ define the so-called *confidence interval* of $M_1$ at the threshold 0.95.

Similarly, if $M_2$ is the mean value of the 1-outcome wrongly selected instead of a 0-outcome, and $k_2$ the number of wrongly selected 1-outcomes in matrix $B$, then :

$$\frac{k_2}{m_2} - 1.96\sqrt{\frac{k_2(m_2 - k_2)}{m_2^2(m_2 - 1)}} \leq M_2 \leq \frac{k_2}{m_2} + 1.96\sqrt{\frac{k_2(m_2 - k_2)}{m_2^2(m_2 - 1)}}$$

### 4.4 Selecting the optimal subset : an example

We want to select a subset such that $M_1$ is as close as possible to 1 and $M_2$ to 0.

Using the previous limits, we will choose a subset having a lower limit of the confidence interval of $M_1$ as big as possible and an upper limit of the confidence interval of $M_2$ as low as possible.

The example presented here has been obtained as follows :

We generated two matrices $A(60 * 14)$ and $B(60 * 14)$ as follows :

1. We defined a key, that is a sequence of 14 elements that are either 0 or 1 or *.
2. To generate a feature value of $A$, we procceed as follows :
   - if this value corresponds to 0 in the key, it will be 0 with the probability 0.9 and 1 with the probability 0.1
   - if this value corresponds to 1 in the key, it will be 1 with the probability 0.9 and 0 with the probability 0.1.
   - if this value corresponds to * in the key, it will be 1 with the probability 0.5 and 0 with the same probability.
3. $B$ is generated similarly, starting from the key obtained by changing 0 in 1 and 1 in 0 in the previous key.

Table 1: Selection of the pairs of features

| | | *A* : Test efficiency | | | *B* : Errors | | |
|---|---|---|---|---|---|---|---|
| Pairs | $\chi^2$ | Min Prob. | Probability | Max Prob. | Min Prob. | Probability | Max Prob. |
| (1,3) | 0.1478 | 0.9283 | 0.9667 | 1 | 0.05 | 0.0035 | 0.0965 |
| (1,4) | 0.2256 | 0.8147 | 0.8833 | 0.9518 | 0 | 0 | 0 |
| (1,8) | 0.4762 | 1 | 1 | 1 | 0.0359 | 0.1 | 0.1640 |
| (1,10) | 0.07264 | 0.8576 | 0.9167 | 0.9757 | 0 | 0.0167 | 0.044 |
| (1,12) | 0.7563 | 0.8801 | 0.9333 | 0.9866 | 0.0035 | 0.05 | 0.09653 |
| (1,14) | 1.0714 | 0.956 | 0.9833 | 1 | 0.1 | 0.1640 | 0.0359 |
| (3,4) | 0.1089 | 0.9283 | 0.9666 | 1 | 0.0034 | 0.05 | 0.0965 |
| (3,8) | 0.2299 | 0.9283 | 0.9667 | 1 | 0.0034 | 0.05 | 0.09653 |
| (3,10) | 0.0350 | 0.9034 | 0.95 | 0.9965 | 0 | 0.0333 | 0.0716 |
| (3,12) | 1.9878 | 0.9283 | 0.9667 | 1 | 0.0035 | 0.05 | 0.0965 |
| (3,14) | 0.2299 | 0.9283 | 0.9667 | 1 | 0.0035 | 0.05 | 0.0965 |
| (4,8) | 0.3509 | 0.9035 | 0.95 | 0.9965 | 0.0134 | 0.0667 | 0.1199 |
| (4,10) | 0.0535 | 0.8801 | 0.9334 | 0.9867 | 0 | .0167 | 0.044 |
| (4,12) | 6.6116 | 0.9035 | 0.95 | 0.9965 | 0.0134 | 0.0667 | 0.1199 |
| (4,14) | 0.3509 | 0.9035 | 0.95 | 0.9965 | 0.0134 | 0.0667 | 0.1199 |
| (8,10) | 0.1129 | 0.8147 | 0.8833 | 0.9519 | 0 | 0.0167 | 0.044 |
| (8,12) | 0.0145 | 0.956 | 0.9833 | 1 | 0.0607 | 0.1333 | 0.2059 |
| (8,14) | 0.3292 | 0.956 | 0.9833 | 1 | 0.0359 | 0.1 | 0.1641 |
| (10,12) | 0.1795 | 0.956 | 0.9833 | 1 | 0.0607 | 0.1333 | 0.2059 |
| (10,14) | 0.1129 | 0.8148 | 0.8833 | 0.9519 | 0 | 0 | 0 |
| (12,14) | 0.0145 | 0.956 | 0.9833 | 1 | 0.0675 | 0.1333 | 0.2059 |

In our example, the key is :

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | * | 1 | 1 | * | * | * | 1 | * | 0 | * | 0 | * | 0 |

We select features (1,3,4,8,10,12,14) by applying the test introdced in section 3. We observe that the selected features are those corresponding to the elements of the key that are different from *. From this set of features, we can extract $C_7^2$=21 pairs of features.

All these pairs of feature are selected using the second test (see section 4.1). Table 1 provides the results of this selection.

Column 1 provides the pairs that have been selected. The value of $\chi^2$ is given in the second column. This value should be less than 6.635 that corresponds to a probability of 0.99.

The three next columns provide the lower bound, the evaluation and the upper bound of the probability of success (threshold 0.95) when using the pair to evaluate a 1-output. The last three columns provide similar informations concerning the probability of concluding that the output is 1 when it is 0.

As table 1 shows, none of the pairs is perfect. Note that pair (1,8) is perfect to extract a 1-outcome, but may conclude that the outcome is 1 when it is 0. Pair (10,14) never concludes that the output is 1 when it is 0, but its performance is quite bad for recognizing a 1 output.

Considering the sets composed of three features, we observe that sets (1,3,8) and (1,4,8) are with a confidence interval [1,1] for the probability of extracting a 1-output and the confidence interval [0,0] for extracting a 1-output instead of a 0-output : we will select one of these sets.

## 4.5 The algorithm

To summarize, the algorthm is as follows :

1. Selecting the features having a discriminatory power.

   Let $j$ be a feature. We compute $\chi_j^2 = \chi_A^2 + \chi_B^2$ as explained in section 3, and we accept or reject a feature depending on $\chi_0^2$ defined according to the probability chosen by the user. If $\chi_j^2 > \chi_0^2$, we decide that feature $j$ has a discriminatory power.

2. For each pair $(j_1, j_2)$ of features, $j_1$ and $j_2$ being the features selected in the previous step, compute $\chi^2_{j_1,j_2}$ (see (1)). If $\chi^2_{j_1,j_2} < \chi^2_0$, then $(j_1, j_2)$ is selected since this means that features $j_1$ and $j_2$ are complementary.
3. Compute the subsets of features having more than two elements starting from the pairs obtained in step 2, as explained in section 4.1.
4. For each subset of two elements and more, evaluate their efficiency with regard to :
   - their ability to detect a 1-output,
   - their ability not to detect a 1-output when the output is 0.

   This leads to a confidence interval as explained in section 4.3.
5. Select the subset. The decision to be made is of a multi-criteria type, since we have to choose :
   - a maximum value for the lower bound of the confidence interval of the probability to detect the 1-outcomes,
   - a minimum value for the upper bound of the confidence interval of the probability to detect a 1-outcome instead of a 0-outcome,
   - the cost associated to the sets. In the previous example, the cost was the number of components in the subset.

# 5 Conclusion

In this article, we used the $\chi^2$-test to extract discriminatory features. We then build sets containing discriminatory features whose values lead to the diagnosis.

Usually, one feature is not enough to conclude if the output is 0 or 1, due to the fact that tests may lead to wrong anr unprecise results. Several features may correct each other and provide a safer result. In the numerical examples that have been developped, we observed that the efficiency first increases, and then decreases with the number of features in the set.

The next step will be to consider the case of missing data, as well as the case when the number of values taken by the features is greater than two.

# References

[1] M. R. Anderberg. *Cluster analysis for applications*. Academic Press, New York, 1973.

[2] G. K. Bhattacharyya and R. A. Johnson. *Statistical concepts and methods*. Wiley, New York, 1977.

[3] E. Boros, P. L. Hammer, T. Ibaraki, and A. Kogan. Logical analysis of numerical data. *Mathematical Programming*, vol. 79, pp. 163-190, 1997.

[4] P. S. Bradley, O. L. Mangasarian, and W. N. Street. Feature selection via mathematical programming. *INFORMS Journal on Computing*, vol. 10, n° 2, pp. 209-217, 1998.

[5] E. J. Bredensteiner and K. P. Bennett. Feature minimization within decision trees. *Computational Optimization and Applications*, vol. 10, pp. 111-126, 1998.

[6] J. C. Deville and G. Saporta. Correspondence analysis with an extension towards nominal time series. *Journal of econometrics*, vol. 22, pp. 169-189, 1983.

[7] R. Gittins. *Canonical Analysis*. Springer Verlag, New York, 1984.

[8] B. Green. *Analyzing multivariate data*. Winston, New York, 1978.

[9] M. J. Greenacre. *Theory and application of correspondence analysis*. Academic Press, New York, 1984.

[10] J. R. Quinlan. Decision trees and decisionmaking. *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 20, n°2, pp. 339-346, 1990.

Unité de recherche INRIA Lorraine
LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)

Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38330 Montbonnot-St-Martin (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399