Fast Feature Selection using Partial Correlation for Multi-valued Attributes

S. Lallich and R. Rakotomalala

ERIC Laboratory - University of Lyon 2
5, av Pierre Mendes-France
F-69676 Bron
FRANCE

Abstract. We propose a fast feature selection method in supervised learning for multi-valued attributes. The main idea is to rewrite the multi-valued problem in the space of examples into a boolean problem in the space of pairwise examples. On basis of this approach, we can use point correlation coefficient which is null in the case of conditional independence, and verifies a formula connecting partial coefficients with marginal coefficients. This property allows to reduce considerably the computing times because a single pass over the database is necessary to compute all coefficients. We test our algorithm on benchmark databases.
Keywords: feature selection, partial association, marginal association

1 Introduction

Feature selection is a key step in any machine learning process. In supervised learning, only the relevant variables are selected, reducing the volume of computation and making the classifier more efficient in generalization, as shown in a variety of papers on sensitivity of noisy attribute classifiers (nearest neighbour [1], naive bayes classifier [11]).

With the development of searches in very large databases [7], preselection -even reduction of- the variables becomes more crucial and more resource consuming [4]. Fast feature selection methods that are general enough to deal with multi-valued categorical variables are thus needed. Typically, there are two kinds of variable selection methods [17]: stepwise filtering [16] and wrapper strategies [12].

Wrapper strategies explicitly use the classifier to select the subset of predictive attributes which minimizes the generalization error rate obtained using cross validation. The main difficulty is choosing between exploring all solutions and the greedy elementary strategy while maintaining the generalization error rate and a reasonable computation time. For this very reason, studies have often relied on so-called rapid learning methods such as decision trees and naive bayes models [11]. Large volumes of data make this strategy harder to apply.

A filtering-type method is suggested here; this quicker method takes advantage of a measure of association that allows the derivation of partial association directly from the marginal pairwise associations. With Boolean variables, we

D.A. Zighed, J. Komorowski, and J. Zytkow (Eds.): PKDD 2000, LNAI 1910, pp. 221−231, 2000.

have shown that such a filtering method can be built on point correlation coefficients [21]. For categorical attributes, we suggest in this paper to reduce them into boolean attributes, by rewriting the problem in the space spanned by the co-labels of the original attributes.

In section 2, the principle of our pairwise correlation measure and the linking formula from marginal coefficient to partial coefficient are laid out. The corresponding feature selection algorithm is given in section 3. Results of experiments on real and artificial datasets using a naive bayes classifier are given in section 4. Related works are described in Section 5, while a conclusion is given in Section 6.

2 Principles

Let's first review stepwise selection using learning set. At the first step, the variable showing the strongest predictive association with the class attribute, say Y, is selected. At each following step, the variable adding the most to the quality of the prediction is selected and added to the set of variables already selected. A stopping rule is needed, and a measure of predictive association in marginal form (first step) or partial form (following steps) that can show the marginal gain brought about by each of the added remaining variables. Thus, we seek a measure of (partial) association between X and Y, given Z, with two important features:

P_1 : the partial measure can be written as a function of the marginal measures between all of the variables taken two at a time (linking formula);

P_2 : if X and Y are independent given Z, than the partial measure is null (conditional independence).

2.1 A linking formula

With a linking formula, the partial coefficients can be computed gradually from the mere marginal coefficients, which represents an important reduction in the need for computational resources [24].

Pearson's correlation coefficient, defined for continuous attributes, verifies such a linking formula:

$$r(Y;X/Z) = \frac{r(Y;X) - r(Y;Z)r(X;Z)}{\sqrt{(1-r^2(Y;Z))\,(1-r^2(X;Z))}} \tag{1}$$

Kendall's rank correlation for ordinal variables verifies a similar equation. Conversely, Saporta [24] has proposed a partial coefficient related to Tschuprow's coefficient[1] that is formally derived from the linking formula.

[1] a Chi-2 coefficient standardized to account for the sample size and the table dimensions

2.2 Conditional independence

Conditional independence means that when X and Y are independent, given Z, then the partial measure is null. This is an important feature for stepwise procedures, as adding a predictor X, independent of Y given Z, to a predictor Z will add nothing to the quality of the model.

Using combinatorics, Lerman [15] studied measures of association for categorical variables, giving, for each type of cross-classification, a formulation for the null hypothesis and an expression for the partial correlation coefficient that guarantees nullity under the null hypothesis. Unfortunately, to our knowledge, there is no linking formula.

As Lerman [15] points out, partial Tschuprow's coefficient proposed by Saporta [24] is not null under conditional independence, except, noted Daudin [5], if the conditioning variable has only 2 values, which is indeed the case for boolean attributes !

Coefficients similar to Proportional Reduction in Error [9] could also be used; these have been generalized as the Proportional Reduction in Entropy, be it Daroczy's-type entropy or rank entropy [13],[22]. The partial association coefficient is then defined as the weighted mean of the marginal associations given Z; weights can be the probabilities of Z [9], or better yet, those probabilities times the conditional entropy of Y given Z [20]. All these coefficients are null under conditional independence as long as the marginal coefficients are null under independence, but they do not yield to a linking formula !

2.3 Boolean attributes

Boolean attributes, whether binary discretized continuous variables or categorical variables rewritten as set of Boolean variables, can be treated using a point correlation coefficient, as we have proposed [21]; this coefficient satisfies P1 and P2 above. Saporta's [24] partial coefficient after Tschuprow's, while null for conditionally independent Boolean variables, can be negative for certain 2 × 2 × 2 tables.

Cross tables of Boolean variables show some interesting properties linked to the fact that those variables can be regarded, without loss of generality, as 0-1 variables since any other coding can be deduced by a linear transformation. Let Y and X be two Boolean variables; let the joint proportion of 1 be p_{11}, the marginal proportions be p_{1+} and p_{+1}; then the expected value and variance of X are p_{+1} and $p_{+1}(1-p_{+1})$ respectively; those of Y are p_{1+} and $p_{1+}(1-p_{1+})$ respectively; and the covariance is $p_{11}-p_{+1}p_{1+}$. The linear correlation coefficient (in this very case called the point correlation coefficient), invariant under linear transformation, is obtained as [2]:

$$r(Y;X) = \varphi = \frac{p_{00}p_{11} - p_{01}p_{10}}{\sqrt{p_{0+}p_{1+}p_{+0}p_{+1}}} \tag{2}$$

[2] In a 2×2 table, all standardizations of the Chi-2 are equivalent : r^2, ϕ^2, $\frac{\chi^2}{n}$, Cramer's V, Kendall's tau, Tschuprow, Goodman and Kruskal's tau.

Examples	**Y**	**X**
1	1	1
2	1	2
3	2	2
4	3	3

Pairwise	I_Y	I_X
1,1	1	1
1,2	1	0
1,3	0	0
1,4	0	0
2,1	1	0
2,2	1	1
2,3	0	1
2,4	0	0
3,1	0	0
3,2	0	1
3,3	1	1
3,4	0	0
4,1	0	0
4,2	0	0
4,3	0	0
4,4	1	1

Table 1. From original dataset to pairwise co-labeled dataset

There are many advantages to reason with the point correlation coefficient: it is naturally signed, it verifies a linking formula being a special case of Pearson's correlation, and it is null under conditional independence. This last point can be shown easily as any regression on a Boolean regressor is linear. As with Tschuprow's coefficient, $r(Y;X/Z)$ can be null even if X and Y are not independent. This is true for all coefficients verifying the linking equation. Consider a 2×2×2 table for which a new predictor X is uncorrelated with a former predictor Z and uncorrelated with the class attribute.

2.4 Multi-valued Attributes

For multi-valued attributes, there is no method that satisfies P_1 and P_2, as stated above. Categorical attributes can be reduced to Boolean attributes by rewriting the problem in the space spanned by the co-labels[3] of the original attributes, and thus the selection method described earlier can be applied. Hence, step by step, co-labels associated with the regressors that explain the best the class attribute will be selected using the point correlation coefficient. The $n \times (p+1)$ matrix of data (n is the number of examples in the learning set, $(p+1)$ the number of multi-valued attributes including the class) changes to a $n^2 \times (p+1)$ Boolean matrix (e.g. Table 1).

Formally, we will prefer to define individuals pairs with replacement taking order into account. Indeed, in this case, one can demonstrate that independence of categorical variables Y and X implies independence between co-labels associated with each variable [18] and then the point correlation coefficient $r(I_Y, I_X)$

[3] A co-label I_X associated with an attribute X is set to 1 if both individuals have the same value for the attribute of interest, 0 otherwise.

has zero value. If the pairs are without replacement and non ordered, we have demonstrated that under Y and X independence hypothesis, $r(I_Y, I_X)$ has a negative value.

In order to compute the point correlation coefficient between the co-labels, it is simpler to use a contingency-type formula based on the original Y and X, rather than the customary "r" based on I_Y and I_X flags. Consequently, it is not necessary to explicitly form the I_Y and I_X co-labels attributes. The computational cost of the point correlation $r(I_Y, I_X)$ remains in $O(n)$, which is an essential condition of the quickness of the algorithm.

Frequencies of the cross-classified flags, say $g_{ij}, i = 0, 1; j = 0, 1$, are written as functions of the frequencies of the cross-classified attributes Y and X, say $n_{kl}, k = 1, 2, ..., K; l = 1, 2..., L$.

Y\X	$\neq$ (different label on X)	$=$ (same label on X)
$\neq$	$g_{00} = \sum_{k=1}^{K} \sum_{l=1}^{L} n_{kl} \left[n - n_{k+} - n_{+l} + n_{kl}\right]$	$g_{01} = \sum_{k=1}^{K} \sum_{l=1}^{L} n_{kl} \left[n_{+l} - n_{kl}\right]$
$=$	$g_{10} = \sum_{k=1}^{K} \sum_{l=1}^{L} n_{kl} \left[n_{k+} - n_{kl}\right]$	$g_{11} = \sum_{k=1}^{K} \sum_{l=1}^{L} n_{kl}^2$

$$r(I_Y, I_X) = \frac{g_{11}g_{00} - g_{10}g_{01}}{\sqrt{g_{1+}g_{0+}g_{+1}g_{+0}}}$$

Once the table of the point correlation coefficients is completed, partial correlations are derived, thus selecting attributes step by step. The transition from the space of examples to the space of pairs changes the predictions: now, whether two examples have the same Y-labels given their X-labels are identical or not is predicted, rather than the Y-label given the X-labels. This Boolean set-up is particular as the roles of the labels are not interchangeable. Here, "r", not "r^2", must be maximized, since a strong negative correlation is a sign that X is ill-adapted at predicting Y. Similarly, "r_{part}" and not "r_{part}^2" will be maximized in the following steps. Let's consider the second step of the procedure, when the best X is sought, given the Z selected at the first step. For a given Z, either all pairs are concordant in Z or all pairs are discordant in Z; in either case, concordance in X should correspond to the concordance in Y, that is maximizing "r_{part}".

A rigorous stopping rule is still needed. Value zero is not convenient because we work on a sample. For the time being, we use an empirical stopping rule which operates when "r_{part}" is less than $\frac{2.5}{n}$. This would correspond approximately to the upper 0.5% point of the distribution of "r" under independence, namely, $r \sim N(0, \frac{1}{n})$. A theoretically sound critical value for "r_{part}" that accounts for the number of tests is needed.

3 A Greedy Algorithm for Feature Selection

Building the partial correlations from the marginal correlation without any additional passes though the data set is a key feature of our algorithm. Only the marginal correlations require passing through the data set, a $(p + 1) \times (p + 1)$

```
$S = \emptyset$, $k = 0$
Compute marginal coefficients $T_{r_0(Y,X)}$
Repeat
   Find $X^* = \arg\max_{X} r_k(Y, X \,/\, S)$
   If $r_k(Y, X^*/\, S) > \frac{2.5}{n}$
     Then $S = S \cup \{X^*\}$
            $k = k + 1$
            Compute $T_{r_k(Y,X)}$ from $T_{r_{k-1}(Y,X)}$
   End if
Until ”Last add refused”
Return $S$
```

Table 2. G3 greedy algorithm for feature selection

table containing the marginal correlations $r_{Y,X_j} (j = 1, \ldots , p)$ and $r_{X_i,X_j} (i, j = 1, \ldots , p)$, say $T_{r(Y,X)}$, being then constructed (p is the number of descriptors on the data set). As correlations are symmetrical, only the upper triangle of $T_{r(Y,X)}$ need be computed.

First, S, the set of selected attributes, is empty. The attribute X^* the most correlated with Y, as indicated in $T_{r(Y,X)}$, is sought. If this search yields a significant solution, then X^* is selected and inserted in S, and the table T is refreshed with the partial correlations $r(Y, X/X^*)$ using the linking formula. In the next step, X^{**}, the attribute showing the strongest correlation with Y given X^*, is selected. This goes on until the stopping rule is activated. Thus, with a very simple greedy algorithm, T is updated each time an attribute is selected. Pseudo-code for the corresponding algorithm $G3$ is shown in Table 2.

The one and only pass through the data set is of magnitude $O(p^2 \times n)$, where n is the number of individuals on the dataset. All other computations can be derived from the coefficients computed first. The maximum complexity, if all attributes were selected, is of magnitude $O(p^2)$.

4 Experiments

Whether our algorithm indeed selects the ”right” attributes, and the impact of the selection on a classifier have to be assessed. The naive bayes classifier will be used for two reasons: its complexity is easy to compute $[O(p \times n)]$ and hence the reduction of computing time due to the reduction of attributes is easily seen; it is sensitive to noisy attributes [11], hence eliminating irrelevant attributes should improve its performances.

Databases extracted from the UCI Irvine [3] server were used. We used a very diversified set of databases so that the performance of the algorithm could be assessed in a variety of situations. Some are real-life cases (adult, auto, dermatology, heart, iris, lung cancer), some are artificial (monks1, monks3, mushroom);

Base	Examples	Err. init	Naive bayes	G3	MIFS
Adult	48842	0.24	14 (0.161)	7 (0.144)	8 (0.146)
Autos	205	0.55	25 (0.248)	12 (0.249)	23 (0.243)
Breast noisy	699	0.345	18 (0.037)	8 (0.036)	9 (0.040)
Dermatology	366	0.70	34 (0.102)	23 (0.085)	34 (0.101)
Heart	270	0.44	13 (0.169)	10 (0.175)	12 (0.174)
Iris	150	0.66	4 (0.060)	2 (0.028)	2 (0.027)
Led noisy	10000	0.90	24 (0.264)	7 (0.264)	24 (0.264)
Lung-cancer	32	0.60	56 (0.742)	3 (0.308)	38 (0.700)
Monks-1	556	0.50	6 (0.254)	1 (0.254)	4 (0.254)
Monks-3	554	0.48	6 (0.036)	2 (0.036)	3 (0.036)
Mushroom	8416	0.47	22 (0.004)	13 (0.007)	7 (0.023)
Segmentation	2310	0.86	11 (0.163)	11 (0.163)	4 (0.080)
Wave noisy	5000	0.67	40 (0.224)	14 (0.219)	30 (0.221)

Table 3. Databases characteristics and results - Number of selected attributes (Error rate)

the last set are sets to which random noise was added (wave noisy, breast cancer noisy, led noisy).

Some continuous attributes were discretized using FUSINTER [25]. This is a supervised discretization method which has the advantage of suggesting partitions maximizing the link of each attribute with the class attribute. Compared to less sophisticated methods (mainly unsupervised strategies) [6], we could think that this approach gives the advantage to discretized continuous attributes compared to other categorical one. Experiments show that this undesirable effect does not appear in practice. If we refer to the analogy with decision trees which we will explain in more details below, doing a global discretization before learning process is a practicable strategy [8].

Data bases characteristics and results are shown in Table 3; 10 cross-validations were used to measure the error rate. Results for G3 were compared to results for MIFS [2], an alternative greedy algorithm that will be reviewed in the next section. Table 3 shows the number of cases, the default classifier error rate (the default classifier always predicts the class with the highest frequency), the number of selected attributes and the error rate for the naive bayes algorithm, G3, and finally MIFS. Some observations follow:

- where noise was deliberately added (wave noisy, breast cancer noisy, led noisy), all random attributes were eliminated; thus the method is effective when a large number of attributes are present, not all relevant;
- more interesting yet, for the bases where the "right" attributes were known, these were among the first selected (iris, wave, breast, led); the test error rate should not be severely affected by the reduction of attributes that would follow the introduction of a more stringent stopping rule;

- on the other bases, even UCI Irvine's that have been "worked at", with few irrelevant attributes [10], the selection almost always drastically reduces the number of attributes;
- for the monks datasets (*monks-1, monks-3*), where the concepts to learn are disjunctions of conjunctions, the greedy algorithm failed to select the right attributes; in both cases, only one set of the disjunction was identified. This also happens when a strong interaction is present among the variables (mushroom);.
- error rates are insensitive to reductions in the number of attributes, with two exceptions (dermatology and especially lung cancer); in the lung cancer case, curse of dimensionality disturbs, and a reduction of the number of dimensions improved learning;
- how well MIFS operates depends on how well β, the parameter that rules how many attributes are to be selected, is set. In practice, it appears that β varies with the problem at hand; a constant β=1 was used, following [14], to obtain a stable comparison. In some instances (auto, dermatology, led, lung cancer, wave) this value seems insufficient; in other cases (mushroom) it appears too restrictive; it appears appropriate for iris and segmentation. From this point of view, G3 is more stable. Tuning G3 is done by adjusting the stopping rule threshold, and even when it was set to 0, results were comparable to those displayed in Table 3;
- as regards calculation times, $G3$ and $MIFS$ are theoretically equivalent since they have the same computational complexity. However, in our experiments, we can note that the $MIFS$ method is a little slower, simply because it selects a greater number of attributes.

5 Related works

In this paper, we are mainly concerned with filtering feature selection methods. The basic idea is to identify the most discriminating subset of attributes before learning starts [17]. Contrary to the wrapper method, filtering methods do not use the classifier's characteristics to build the classifier. There is no guarantee that the best set of attributes will be selected for a given learning algorithm. For example, in the case of the Boolean XOR, a filtering method could select the right attributes, but the decision tree used later on to learn the concept, short-sightedness, fails to see the correct model [19]. This apparent flaw can become an advantage. As it does not depend on any learning algorithm, the filtering aims at outlining the right representation space. Then, a suitable learning algorithm can be found for the definitive set of attributes. Using a multilayer perceptron in the example above, the solution is trivial. The second appealing feature of filtering algorithms is their computational speed. In data mining situations [7], the huge size of the database is such that direct processing of all the data is hardly ever possible. The data base must be reduced in length (by sampling cases) and in width (by selecting variables) to make computations possible and affordable [4]. Of course, this requires a fast selection algorithm.

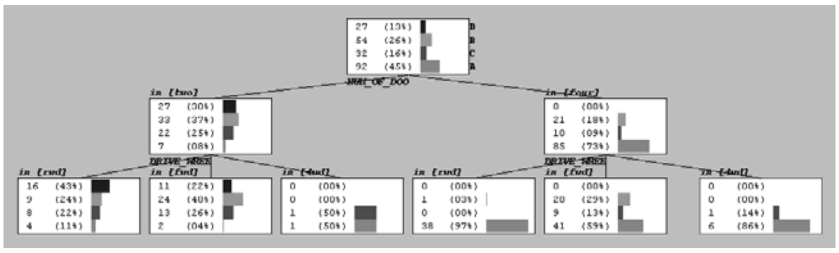

Fig. 1. Constrained Decision Tree with two level on AUTOS dataset

The best representation space is determined by a stepwise selection of the predictors that exhibit the strongest correlation with the class attribute. Hence, exploring the solution space ought to be greedy. In this sense, FGMIFS [14] is similar to our algorithm. FGMIFS constructs a decision tree under the constraint that, for each level, the leaves be split using the same attribute (Figure 1). Here, the measure for the attribute selection is akin to information gain [26]. Typically, it is the description of a stepwise regression using discrete explicative variables. G3 implicitly does the same, but differs from FGMIFS on two essential points: (1) at each node, the partial correlations are derived from previously computed parameters and do not require an additional pass over the databases; (2) the tree not being explicitly constructed, it is not penalized by data fragmentation. Indeed, with FGMIFS, for a large tree (assume 10 Boolean attributes, hence 2^{10} leaves), the number of individuals on each leave is too small for reliable estimation of probabilities.

In our opinion, MIFS [2] is closer to G3 than FGMIFS: the additional information about the class attribute Y brought by an additional X, given the already selected attributes, is computed. Here, the main difference with G3 is that the author uses mutual information, and while partial association is derived from marginal, it is defined empirically. Let S be the set of the already selected attributes, the additional information about Y brought by X is given by:

$$I(Y,X/S) = I(Y,X) - \beta \times \sum_{Z \in S} \frac{I(X,Z)}{card(S)}$$

This is in fact a valid procedure whatever the number of categories. But the success of the procedure relies heavily on the choice of β. In practice, many tests are required before a suitable β is found, and the expected gain in time is lost.

Lastly, work in [23] is older, yet close to ours. The main idea is to derive the partial coefficients from the marginals. While we seek coefficients satisfying some properties, namely the iterative derivation of the partial coefficients, coefficients in [23] are defined after the linking equation. Such coefficients need not be null under conditional independence, and may lie outside the domain of the marginal coefficient (Tschuprow's is bounded by 0 and 1; the partial Tschuprow, as defined by its author, may be negative) which makes interpretation difficult.

6 Conclusion and Future Work

In this paper, a fast multi-valued attribute selection method based on the recurrent computation of partial correlations is developed. It is quite fast as a single pass through the data is necessary, and can thus be used as a starting point far an induction process.

Tests on real and artificial data showed that this approach is pragmatic, and have identified situations where it is quite advantageous. Where the method appears to fail, namely in disjunction problems or with data showing large interaction, the greediness of the method seems at fault. Similarities with decision trees could open new venues: borrowing sophisticated algorithms such as lookahead search, post pruning, or synthetic attributes [26].

Acknowledgements

We are very grateful to the two anonymous referees who provided helpful comments, and to M. Jean Dumais, Section Chief, Methodology Branch, Statistics Canada and guest lecturer at Université Lumière for his careful proof reading of the manuscript and his suggestions.

References

1. D. Aha. Tolerating noisy, irrelelvant and novel attributes in instance-based algorithms. *International Journal of Man-Machine Studies*, 36:267–287, 1992.
2. R. Battiti. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, 5(4):537–550, 1994.
3. S.D. Bay. The uci kdd archive [http://kdd.ics.uci.edu]. Irvine, CA: University of California, Department of Computer Science, 1999.
4. A. Blum and P. Langley. Selection of relevant feature and examples in machine learning. *Artificial Intelligence*, pages 245–271, 1997.
5. J.J. Daudin. Analyse factorielle des dependances partielles. *Revue de Statistique Appliquee*, 29(2):15–29, 1981.
6. J. Dougherty, R. Kohavi, and M. Sahami. Supervised and unsupervised discretization of continuous attributes. In Morgan Kaufmann, editor, *Machine Learning : Proceedings of the* 12^{th} *International Conference (ICML-95)*, pages 194–202, 1995.
7. U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. Knowledge discovey and data mining : Towards an unifying framework. In *Proceedings of the* 2^{nd} *International Conference on Knowledge Discovery and Data Mining*, 1996.
8. Eibe Frank and Ian H. Witten. Making better use of global discretization. In *Proc. 16th International Conf. on Machine Learning*, pages 115–123. Morgan Kaufmann, San Francisco, CA, 1999.
9. L.A. Goodman and W.H. Kruskall. Measures of association for cross classifications. *Journal of American Statistical Association*, 49:732–764, 1954.
10. R.C. Holte. Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, 11:63–91, 1993.
11. G. John and P. Langley. Static versus dynamic sampling for data mining. In *Proceedings of the* 2^{nd} *International Conference on Knowledge Discovery in Databases and Data Mining.* AAAI/MIT Press, 1996.

12. R. Kohavi and G. John. Wrappers for feature subset selection. *Journal of Artificial Intelligence, Special issue on Relevance*, 1997.
13. S. Lallich and R. Rakotomalala. Les entropies de rangs généralisés en induction par arbres. In *Proceedings of* 7^{mes} *Journées de la Société Francophone de Classification - SFC'99*, pages 101–107, September 1999.
14. K-C. Lee. A technique of dynamic feature selection using the feature group mutual information. In *Proceedings of the Third PAKDD-99*, pages 138–142, 1999.
15. I.C. Lerman. Correlation partielle dans le cas qualitatif. Technical Report 111, INRIA, 1982.
16. H. Liu and H. Motoda. *Feature Extraction, Construction and Selection : A Data Mining Perspective.* Kluwer Academic Publishers, 1998.
17. H. Liu and H. Motoda. *Feature selection for knowledge discovery and data mining*, volume 454 of *The kluwer international series in engineering and computer science.* Kluwer, 1998.
18. F. Marcotorchino. Utilisation des comparaisons par paires en statistique des contingences - partie iii. Technical Report F 081, Centre Scientifique IBM-France, 1985.
19. D. Michaud. *Filtrage et Selection D'attributs En Apprentissage.* PhD thesis, Universite de Franche-Comte, 1999.
20. M. Olszak and G. Ritschard. The behaviour of nominal and ordinal partial association measures. *The statistician*, 44(2):195–212, 1995.
21. R. Rakotomalala and S. Lallich. Sélection rapide de variables booléennes en apprentissage supervisé. In *Proceedings of 2nd Conférence Apprentissage - CAP'2000*, pages 225–234, 2000.
22. R. Rakotomalala, S. Lallich, and S. Di Palma. Studying the behavior of generalized entropy in induction trees using a m-of-n concept. In *Proceedings of the Third European Conference PKDD'99*, pages 510–517, 1999.
23. G. Saporta. *Liaisons entre plusieurs ensembles de variables et codage de données qualitatives.* PhD thesis, 1975.
24. G. Saporta. Quelques applications des operateurs d'Escouffier au traitement des variables qualitatives. *Statistique et Analyse de Donnees*, (1):38–46, 1976.
25. D.A. Zighed, S. Rabaseda, and R. Rakotomalala. Fusinter : a method for discretization of continuous attributes for supervised learning. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(33):307–326, 1998.
26. D.A. Zighed and R. Rakotomalala. *Graphes d'Induction - Apprentissage et Data Mining.* Hermes, 2000.