Multivariate discretization by recursive supervised bipartition of graph

Sylvain Ferrandiz[1][2] and Marc Boullé[1]

[1] France Télécom R&D
2, avenue Pierre Marzin,
22307 LANNION Cedex, France
[2] Université de Caen, GREYC,
Campus Côte de Nacre, boulevard du Maréchal Juin
BP 5186
14032 Caen Cedex, France
sylvain.ferrandiz@francetelecom.com
marc.boulle@francetelecom.com

Abstract. In supervised learning, discretization of the continuous explanatory attributes enhances the accuracy of decision tree induction algorithms and naive Bayes classifier. Many discretization methods have been developped, leading to precise and comprehensible evaluations of the amount of information contained in one single attribute with respect to the target one.
In this paper, we discuss the multivariate notion of neighborhood, extending the univariate notion of interval. We propose an evaluation criterion of bipartitions, which is based on the Minimum Description Length (MDL) principle [1], and apply it recursively. The resulting discretization method is thus able to exploit correlations between continuous attributes. Its accuracy and robustness are evaluated on real and synthetic data sets.

1 Supervised partitioning problems

In supervised learning, many inductive algorithms are known to produce better models by discretizing continuous attributes. For example, the naive Bayes classifier requires the estimation of probabilities and the continuous explanatory attributes are not so easy to handle, as they often take too many different values for a direct estimation of frequencies. To circumvent this, a normal distribution of the continuous values can be assumed, but this hypothesis is not always realistic [2]. The same phenomenon leads rules extraction techniques to build poorer sets of rules. Decision tree algorithms carry out a selection process of nominal attributes and cannot handle continuous ones directly. Discretization of a continuous attribute, which consists in building intervals by merging the values of the attribute, appears to be a good solution to these problems.

Thus, as the results are easily interpretable and lead to more robust estimations of the class conditional probabilities, supervised discretization is widely use. In [2], a taxonomy of discretization methods is proposed, with three dimensions : supervised vs. unsupervised (considering a class attribute or not), global vs. local (evaluating the partition as a whole or locally to two adjacent intervals) and static vs. dynamic (performing the discretizations in a preprocessing step or imbedding them in the inductive algorithm). This paper is placed in the supervised context.

The aim of the discretization of a single continuous explanatory attribute is to find a partition of its values which best discriminates the class distributions between groups. These groups are intervals and the evaluation of a partition is based on a compromise : fewer intervals and stronger class discrimination are better. Discrimination can be performed in many different ways. For example,

- Chimerge [3] applies chi square measure to test the independance of the distributions between groups,
- C4.5 [4] uses Shannon's entropy based information measures to find the most informative partition,
- MDLPC [5] defines a description length measure, following the MDL principle,
- MODL [6] states a prior probability distribution, leading to a bayesian evaluation of the partitions.

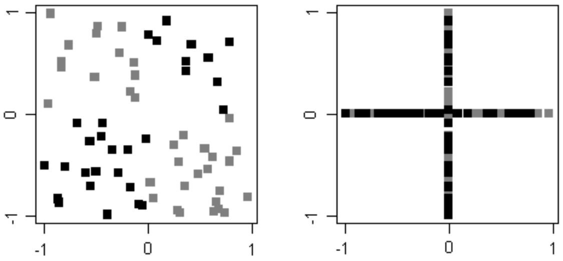

Fig. 1. The XOR problem : projection on the axes leads to an information loss.

The univariate case does not take into account any correlation between the explanatory attributes and fails to discover conjointly defined patterns. This fact is usually illustrated by the XOR problem (cf. Figure 1) : the contributions of the axes have to be considered conjointly. Many authors have thus introduced a fourth category in the preceding taxonomy : multivariate vs. univariate (searching for cut points simultaneously or not), and proposed multivariate methods (see for examples [7] and [8]). These aim at improving rules extraction algorithms and build conjonctions of intervals. It means that considered patterns are parallelepipeds. This can be a limiting condition as underlying structures of the data are not necessarily so squared (cf. Figure 2). We then distinguish these *strongly biased* multivariate techniques from *weakly biased* multivariate ones, that consider more generic patterns. This opposition is slightly discussed in [2], where the authors talk about *feature space* and *instance space* discretizations respectively.

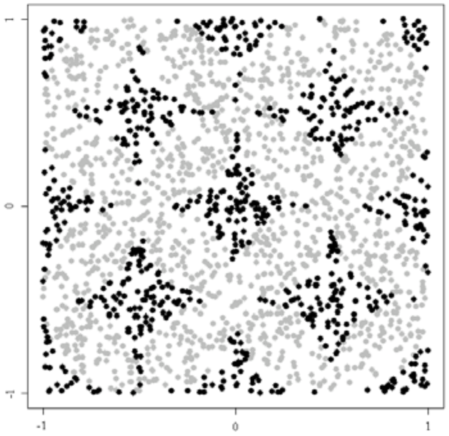

Fig. 2. A challenging synthetic dataset for strongly biased multivariate discretization methods.

We present in this paper a new discretization method, which is supervised, local, static, multivariate and weakly biased. As for the MDLPC method, an evaluation criterion of a bipartition is settled following the MDL principle and applied recursively.

The remainder of the paper is organized as follow. We first set the notations (section 2). Then, we describe the MDLPC technique (section 3) and our framework (section 4). We propose a new evaluation criterion for bipartitions (section 5) and test its validity on real and synthetic datasets (section 6). Finally, we conclude and point out future works (section 7).

2 Notations

Let us set the notations we will use throughout this paper. Let $O = \{o_1, \ldots, o_N\}$ be a finite set of objects. A target class l_n lying in an alphabet of size J is associated to every object o_n. For a subset A of O, $N(A)$ stands for the size of A, $J(A)$ for the number of class labels represented in A and $N_j(A)$ for the number of elements in this groups with label j ($1 \leq j \leq J$). The Shannon entropy of A, which measures the amount of information in bits needed to specify the class labels in A, is then

$$Ent(A) = -\sum_{j=1}^{J} \frac{N_j(A)}{N(A)} \log_2 \frac{N_j(A)}{N(A)}.$$

The problem consists in setting an evaluation criterion of the hypothesis $\mathcal{H}(A, A_1, A_2)$: split the subset A so that $A = A_1 \bigsqcup A_2$. We distinguish the *null* hypothesis $\mathcal{H}(A, A, \emptyset) (= \mathcal{H}(A, \emptyset, A))$ from the family of *split* hypotheses $(\mathcal{H}(A, A_1, A_2))_{A_1 \subsetneq A}$.

Following the MDL principle, a description length $l(A, A_1, A_2)$ must be assigned to each hypothesis and the best hypothesis is the one with the shortest description. Two steps are considered for the description : description of the hypothesis (leading to a description length $l_h(A, A_1, A_2)$) and description of the data given the hypothesis (leading to a description length $l_{d/h}(A, A_1, A_2)$), so that $l(A, A_1, A_2) = l_h(A, A_1, A_2) + l_{d/h}(A, A_1, A_2)$.

3 MDLPC discretization method

In the univariate case, O is a set of ordered real values (i.e. $o_{n_1} \leq o_{n_2}$ if $n_1 \leq n_2$) and the considered groups are intervals. The MDLPC method [5] seeks for the best split of an interval I into a couple of sub-intervals (I_1, I_2), applying the MDL principle.

We begin by considering a split hypothesis. This is determined by the position of the boundary point, and the numbers $J(I_1), J(I_2)$ of class labels represented in I_1 and I_2 respectively. Description lengths are no other than negative log of probabilities, and assuming a uniform prior leads to write :

$$l_h(I, I_1, I_2) = \log_2(N(I) - 1) + \log_2(3^{J(I)} - 2),$$

as there is $N - 1$ possibilities for the choice of the boundary point and the number of admissible values for the couple $(J(I_1), J(I_2))$ has been evaluated to $3^{J(I)} - 2$.

The description of the data given the hypothesis consists in first specifying the frequencies of the class labels in each interval and second the exact sequences of class labels. The evaluation of the lengths is based on the entropy of the intervals I_1 and I_2 :

$$l_{d/h}(I, I_1, I_2) = J(I_1)Ent(I_1) + N(I_1)Ent(I_1) + J(I_2)Ent(I_2) + N(I_2)Ent(I_2).$$

The evaluation of $\mathcal{H}(I, I_1, I_2)$ finally relies on the following formula :

$$\begin{aligned} l(I, I_1, I_2) = {} & \log_2(N(I) - 1) + \log_2(3^{J(I)} - 2) \\ & + J(I_1)Ent(I_1) + N(I_1)Ent(I_1) + J(I_2)Ent(I_2) + N(I_2)Ent(I_2). \end{aligned}$$

For the null hypothesis, the class labels in I have to be described only (i.e $l_h(I, I, \emptyset) = 0$) :

$$l(I, I, \emptyset) = J(I)Ent(I) + N(I)Ent(I).$$

The MDL principle states that the best hypothesis is the one with minimal description length. As partitioning always decreases the value of the entropy function, considering the description lengths of the hypotheses allows to balance the entropy gain and eventually accept the null hypothesis. Performing recursive bipartitions with this criterion leads to a discretization of the continuous explanatory attribute at hand.

4 Multivariate framework

Extending the univariate case mainly requires the definition of a multivariate notion of neighborhood corresponding to the notion of interval. The univariate case does not actually consider the whole set of intervals but those whose bounds are midpoints between two consecutive values. The resulting set of "patterns" is thus discrete, data dependent and induced by a simple underlying structure : the Hasse diagram, which links two consecutive elements of O.

We thus begin by supposing that a non-oriented graph structure G on O conveying a well-suited notion of proximity is provided. Some cases of natural underlying structure arise, like road networks, web graphs, etc ... If the objects in O are tuples of an euclidean space $\mathbb{R}^d$ and a natural structure does not exist, proximity graphs [9] provide definitions of neighborhood.

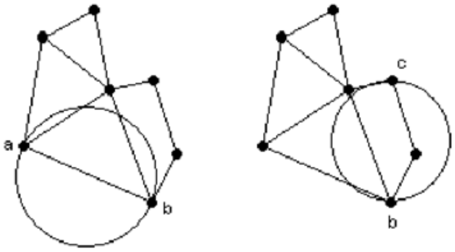

Fig. 3. Example of a Gabriel graph. The ball of diameter $[ab]$ contains no other point : a and b are Gabriel-adjacent. The ball of diameter $[bc]$ contains another point : b and c are not Gabriel-adjacent.

For example, as we work with vectorial data in practice, the Gabriel discrete structure can be chosen. Two multivariate instances o_1 and o_2 are *adjacent in the Gabriel sense* (cf Figure 3) if and only if

$$L(o_1, o_2)^2 \leq \min_{o \in O} L(o_1, o)^2 + L(o_2, o)^2,$$

where L is any distance measure defined on O.

The related discrete metric will be called the *Gabriel metric* on O and will be used throughout the experiments. Any prior knowledge of the user would eventually lead him to select another discrete metric, and it's noteworthy that the use of the Gabriel one is a general choice, made without any further knowledge.

Once a discrete structure G is chosen, we define partitions on the basis of elementary "patterns" related to G. We consider the balls induced by the discrete metric δ related to G : $\delta(o_1, o_2)$ is the minimum number of edges needed to link o_1 and o_2 ($o_1, o_2 \in O$). The resulting set of balls is denoted $\mathcal{B}$ (cf. Figure 4).

We can now express a multivariate analog of the univariate family of split hypotheses, considering balls as basic patterns. In the multivariate case, a local

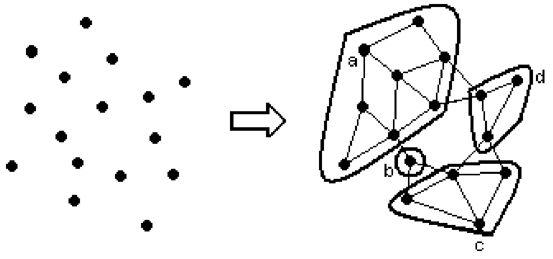

Fig. 4. Multivariate analog of intervals : examples of discrete balls. For example, the ball centered in a with radius 2 contains 8 objects of the dataset.

bipartitioning hypothesis consists in spliting a subset S of O into a ball $B \in \mathcal{B}$, included in S, and its complement. $\mathcal{H}(S, B)$ denotes such a hypothesis. As we utilize a connected discrete structure (the Gabriel graph), eventually obtaining partitions with non-connected groups can be somewhat counterintuitive. We do not try to alleviate this conceptual fact in the present paper.

5 Evaluation of a bipartition

The proposed framework leads to the study of the hypothesis $\mathcal{H}(S, B)$, where S is a subset of O, B a ball included in S. We now introduce an evaluation criterion $l(S, B)$ for such a hypothesis. Following the MDL principle, we have to define a description length $l_h(S, B)$ of the bipartition and a description length $l_{d/h}(S, B)$ of the class labels given the bipartition.

We first consider a split hypothesis : $B \neq S$. In the univariate case, the bipartition results from the choice of a cut point. In the general case, the bipartition is determined by the ball B and the description of B relies on two parameters : its size $N(B)$ and its index in the set of balls of size $N(B)$ included in S.

Description lengths are negative log of probabilities and, if $\beta(S, B)$ stands for the number of balls of size $N(B)$ included in S, we obtain

$$l_h(S, B) = \log_2 N(S) + \log_2 \beta(S, B)$$

assuming a uniform prior.

Let us now specify the distribution of the class labels in a subset A of O (A will be S, B or $S \setminus B$). This is the same as putting the elements of A in J boxes. We begin specifying the numbers of elements to put in the j^{th} box, that is, the frequencies $(N_1(A), \ldots, N_J(A))$. It then remains to give the index of the actual partition in the set of partitions of A in J groups of sizes $N_1(A), \ldots, N_J(A)$.

Each possible J-uple of frequencies satisfies the property that the sum of its components equals $N(A)$. The set of possible frequencies is then of size $\binom{N(A)+J-1}{J-1}$. Counting the set of partitions of A in J groups of fixed sizes

$N_1(A), \ldots, N_J(A)$ is a multinomial problem and the size of this set is the multinomial coefficient $\frac{N(A)!}{N_1(A)! \ldots N_J(A)!}$.

Still assuming a uniform prior, the description length of the distribution of the labels in A is then :

$$l_d(A) = \log_2 \binom{N(A)+J-1}{J-1} + \log_2 \frac{N(A)!}{N_1(A)! \ldots N_J(A)!}.$$

For fixing $l_{d/h}(S,B)$, we suppose the distributions of the labels in B and its complement independant. This results in setting :

$$l_{d/h}(S,B) = l_d(B) + l_d(S \setminus B).$$

Finally, the description length of a split hypothesis is given by the formula :

$$\begin{aligned} l(S,B) = \log_2 N(S) + \log_2 \beta(S,B) \\ + \log_2 \binom{N(B)+J-1}{J-1} + \log_2 \frac{N(B)!}{N_1(B)! \ldots N_J(B)!} \\ + \log_2 \binom{N(S \setminus B)+J-1}{J-1} + \log_2 \frac{N(S \setminus B)!}{N_1(S \setminus B)! \ldots N_J(S \setminus B)!}. \end{aligned}$$

The null hypothesis relies on the description of the size of the considered subset (S) and the distribution of the labels in S. Indicating that the size is that of S amounts to pointing the null hypothesis. Thus, $l_h(S,S,\emptyset) = \log_2 N(S)$ and $l_{d/h}(S,S,\emptyset) = l_d(S)$, giving

$$l(S,S) = \log_2 N(S) + \log_2 \binom{N(S)+J-1}{J-1} + \log_2 \frac{N(S)!}{N_1(S)! \ldots N_J(S)!}.$$

Still, the decision results from an optimal compromise between an entropy gain and a structural cost of the considered split hypotheses, taking into account the null hyptohesis as well. But the latter does not employ the Shannon entropy (as MDLPC does), replacing it by a binomial evaluation of the frequencies of the distributions. The former exploits a multinomial definition of the notion of entropy, overcoming the asymptotic validity of the Shannon entropy.

6 Experiments

The multivariate discretization algorithm consists in applying recursively the following decision rule :

1. S a subset of O (initialy, $S = O$)
2. select the ball B_0 which minimizes $l(S,B)$ over the balls $B \in \mathcal{B}$ contained in S,
3. if $l(S,B_0) < l(S,S)$, performs step 1 on $S = B$ and $S = S \setminus B$, else stop.

		Continuous	Class	Majority
Dataset	Size	Attributes	Values	Class
Iris	150	4	3	0.33
Wine	178	13	3	0.40
Heart	270	10	2	0.56
Bupa	345	6	2	0.58
Ionosphere	351	34	2	0.64
Crx	690	6	2	0.56
Australian	690	6	2	0.56
Breast	699	9	2	0.66
Pima	768	8	2	0.65
Vehicle	846	18	4	0.26
German	1000	24	2	0.7

Table 1. Tested datasets.

Constructing the Gabriel graph requires $O(N^3)$ operations. If D is the diameter of the graph, the overall number of balls is in $O(DN)$ and each decision thus results from evaluating $O(DN)$ hypotheses. Each evaluation can be performed with $O(J)$ operations, storing the $O(DN)$ sizes of the balls. At most N splits can be triggered, giving a time complexity in $O(JDN^2)$ and a space complexity in $O(DN)$ for the optimisation algorithm. In practice, the method performs few splits and the number of available balls quickly decreases, giving an $O(N^3)$ algorithm.

We perform three experiments, one on real datasets and two on synthetic datasets. The metric is chosen to be the euclidean one. We do not consider any other metric or weighting scheme, as the experiments aim at comparing our methods with others, in a single framework.

The main advantage of partitioning methods lies in their intrinsic capacity for providing the user with an underlying structure of the analysed data. However, this structural gain may be balanced by an information loss. The first experiment aims at evaluating how our method is affected by such a flaw. We consider the resulting partition as a basic predictive model : a new instance is classified according to a majority vote in the nearest group. We thus compare the accuracy of the discretization method to the accuracy of the Nearest Neighbor rule (NN), which gives the class label of its nearest neighbor to an unseen instance [10].

The tests are performed on 11 datasets (cf Table 1) from the UCI machine learning database repository [11]. As we focus on continuous attributes, we discard the nominal attributes of the Heart, Crx and Australian database. The evaluation consists in a stratified five-fold cross-validation. The predictive accuracy of the classifiers are reported in the Table 2, as well as the robustness (i.e the ratio of the test accuracy by the train accuracy) of our classifier.

The overall predictive accuracy does not significantly suffers from the partitioning of the data (72% against 73%). But with some datasets (Iris, Wine, Vehicle), the disadvantage of making local decision is evidenced. Indeed, as illus-

	Test accuracy		Robustness	
Dataset	Partition	NN	Partition	NN
Iris	0.92 ± 0.05	0.96 ± 0.02	0.98 ± 0.06	0.96 ± 0.02
Wine	0.69 ± 0.09	0.76 ± 0.07	0.90 ± 0.14	0.76 ± 0.07
Heart	0.62 ± 0.04	0.55 ± 0.03	0.88 ± 0.04	0.55 ± 0.03
Bupa	0.61 ± 0.06	0.61 ± 0.05	0.85 ± 0.07	0.61 ± 0.05
Ionosphere	0.85 ± 0.04	0.87 ± 0.02	0.99 ± 0.04	0.87 ± 0.02
Crx	0.66 ± 0.03	0.64 ± 0.05	0.90 ± 0.06	0.64 ± 0.05
Australian	0.69 ± 0.02	0.68 ± 0.02	0.95 ± 0.05	0.68 ± 0.02
Breast	0.97 ± 0.01	0.96 ± 0.01	1.00 ± 0.01	0.96 ± 0.01
Pima	0.68 ± 0.01	0.68 ± 0.02	0.94 ± 0.04	0.68 ± 0.02
Vehicle	0.54 ± 0.04	0.65 ± 0.02	0.90 ± 0.07	0.65 ± 0.02
German	0.70 ± 0.01	0.67 ± 0.02	0.96 ± 0.01	0.67 ± 0.02
Mean	0.72	0.73	0.93	0.73

Table 2. Predictive accuracy and robustness of our method and predictive accuracy of the NN rule for the tested datasets.

trated by the Figure 5, a succession of local decisions can lead to the constitution of some border groups, which is especially harmful in the context of separable distributions, producing a decrease of the accuracy. While our method takes on a safe approach, handling the boundary data with cautions, the NN rule builds more hazardous decision boundaries without being penalized in term of test accuracy.

The robustness of the NN rule is equal to its test accuracy, and we observe that building a well-suited partition of the data sharply increases the robustness of the prediction (0.93 against 0.73).

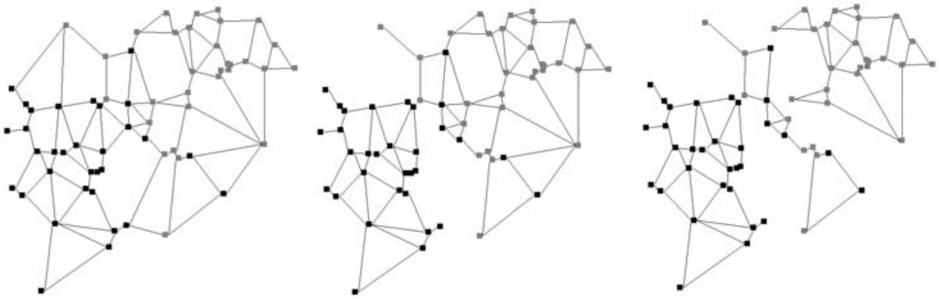

Fig. 5. Partitioning of a 2 separable classes problem : creation of a buffer zone, containing a mixture of the two classes.

In a second experiment, we compare our method and the well-known decision tree algorithm C4.5 when faced with the challenging pattern presented in

Method	Test accuracy	Robustness	Group number
Partition	0.83 ± 0.01	0.95 ± 0.01	29.5 ± 0.35
C4.5	0.71 ± 0.04	0.94 ± 0.01	17 ± 1.41
NN	0.90 ± 0.00	0.90 ± 0.00	-

Table 3. Predictive accuracy, robustness and number of groups of our method, C4.5 and the NN rule on the "challenging" dataset.

Figure 2. The dataset contains 2000 instances and we carry out a stratified two-fold cross-validation. We report the predictive accuracy, the robustness and the number of groups in the Table 3.

From this experiment, we notice quite a strong difference between the predictive performances : our method perfoms a better detection than C4.5 (0.83 against 0.71). This is not surprising and illustrates the distinction between weakly and strongly biased multivariate partitioning. C4.5, which is a strongly biased method, is forced to detect parallelepipeds, limiting its detection ability as evidenced on this example. This experiment shows the robustness of the partitioning methods once again.

On the negative side, we notice a loss of predictive accuracy of our method compared with the NN rule. Examining the two produced partitions, we find that after the detection of a few clean balls (i.e objects in a ball sharing the same class label), a group containing about 600 instances marked Grey and 100 marked Black remains uncut. As the set of balls is updated by deleting balls only, the descriptive capacity of our method becomes poorer after each triggered cut. This results from the fact that we consider balls defined in the whole set O and not a locally defined set of balls. As the method makes local optimizations, performing better updates would enhance its predictive accuracy.

The third experiment consists in evaluating the tolerance of our method to the presence of mislabelled data. The method is applied to 11 Datasets, each containing 1000 instances uniformly generated in $[-1, 1] \times [-1, 1]$, representing the XOR problem with increasing mislabelled data rate, from 0 (XOR problem) to 0.5 (pure noise). The evolution of the predictive accuracy and the robustness (evaluated by a stratified 5-fold cross-validation) is shown in Figure 6, and compared with NN rule results again.

The expected optimal accuracy curve is the line passing through $(0, 1)$ and $(0.5, 0.5)$. The partitioning algorithm is up to 10% more accurate than the NN rule and far more robust. This is its main advantage : still building accurate and robust partitions in presence of noise.

7 Conclusion and further works

In this paper, we have discussed the usefulness of supervised partitioning methods for data preparation in the univariate case. We have proposed an extension to the multivariate case, relying on the multivariate definition of discrete neigh-

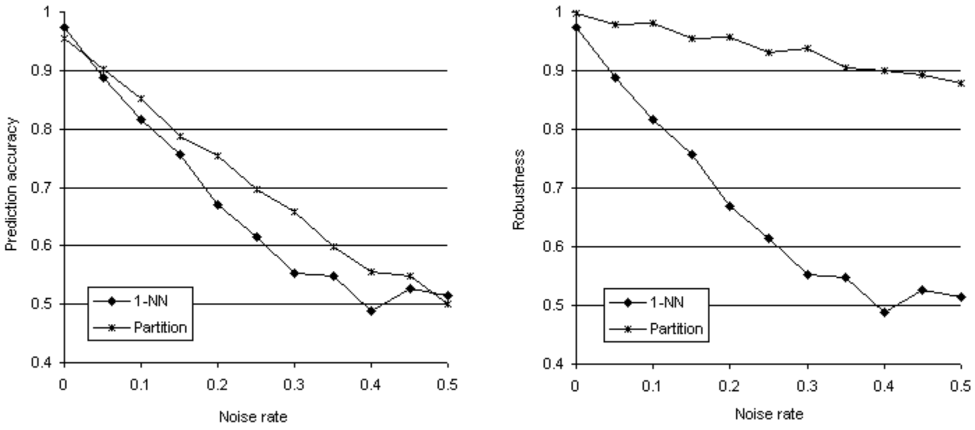

Fig. 6. Evolution of the predictive accuracy and the robustness with the mislabelled data rate of the partitioning technique and the NN rule on the XOR pattern.

borhood by means of a non-oriented graph structure. A framework for supervised bipartitioning has been proposed, which applied recursively leads to a new multivariate discretization algorithm. Finally, this algorithm has been tested on real and synthetic datasets.

The proposed method builds an underlying structure of the data, producing understandable results without fitting parameters and without loss of predictive information (as shown by the experiments on real datasets). Defining basic patterns (the balls) from the data allows the technique to better partition the dataset, compared with classical strongly biased multivariate algorithm like C4.5. Furthermore, its demonstrated robustness is a main advantage, particularly since it's very tolerant to the presence of noise.

Still, more experiments have to be carried out. In the reported experiments, our method is evaluated as a classifier not as a data preparation technique. We plan to evaluate the impact of our method when considered as a preprocessing step of a naive bayes classifier, for example. Furthermore, the presented criterion can be improved, by considering local sets of balls rather than updating the global set.

8 Acknowledgement

We are grateful to the anonymous referees who provided helpful suggestions, and to Fabrice Clérot, Senior Expert at France Télécom R&D, for his careful proofreading of the manuscript and his comments.

References

[1] Rissanen, J.: Modeling by shortest data description. Automatica **14** (1978) 465–471

[2] Dougherty, J., Kohavi, R., Sahami, M.: Supervised and unsupervised discretization of continuous features. In Proc. of the 12^{th} ICML (1995) 194–202

[3] Kerber, R.: Chimerge discretization of numeric attributes. Tenth International Conference on Artificial Intelligence (1991) 123–128

[4] Quinlan, J.R.: C4.5: Programs for machine learning. Morgan Kaufmann (1993)

[5] Fayyad, U., Irani, K.: On the handling of continuous-valued attributes in decision tree generation. Machine Learning **8** (1992) 87–102

[6] Boullé, M.: A bayesian approach for supervised discretization. Data Mining V, Zanasi and Ebecken and Brebbia, WIT Press (2004) 199–208

[7] Bay, S.: Multivariate discretization of continuous variables for set mining. In Proc. of the 6^{th} ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (2000) 315–319

[8] Kwedlo, W., Kretowski, M.: An evolutionary algorithm using multivariate discretization for decision rule induction. In Proc. of the European Conference on Principles of Data Mining and Knowledge Discovery (1999) 392–397

[9] Jaromczyk, J.W., Toussaint, G.T.: Relative neighborhood graphs and their relatives. P-IEEE **80** (1992) 1502–1517

[10] Cover, T.M., Hart, P.E.: Nearest neighbor pattern classification. Institute of Electrical and Electronics Engineers Transactions on Information Theory **13** (1967) 21–27

[11] Blake, C.L., Merz, C.J.: UCI repository of machine learning databases. http://www.ics.uci.edu/m̃learn/MLRepository.html