Supervised evaluation of dataset partitions : advantages and practice

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 the context of large databases, data preparation takes a greater importance : instances and explanatory attributes have to be carefully selected. In supervised learning, instances partitioning techniques have been developped for univariate representations, leading to precise and comprehensible evaluations of the amount of information contained in an attribute, with respect to the target attribute. Still, the multivariate case remains unstated.
In this paper, we describe the partitioning intrinsic convenience for data preparation and we settle a framework for supervised partitioning. A new evaluation criterion of labelled objects partitions, which is based on Minimum Description Length principle, is then set and tested on real and synthetic data sets.

1 Supervised partitioning problems in data preparation

In a data mining project, the data preparation phase is a key one. Its main goal is to provide a clean and representative database for the consecutive modelling phase [3]. Typically, topics like instances representation, instances selection and/or aggregation, missing values handling, attributes selection, are to be carefully dealt with. Among the many designed methods, partition-based one are often used, for their ability to comprehensibly summarize the information.

The first examples that come in mind are clustering techniques, like the most popular one : K-means [11], which aim at partitioning instances. Building partitions hierarchy or mixture models is another way of doing unsupervised classification [5]. Combining clustering and attributes selection has led to the description of self-organizing feature maps [10].

In the supervised context, induction tree models are plainly partition-based [2],[12],[8]. These models build a hierarchy of instances groups relying on the discriminating power of the explanatory attributes with respect to the categorical

target attribute. As the naive Bayes classifier, they need to discretise the continuous explanatory attributes to make probability estimations more accurate. As discretisation is the typical univariate supervised partitioning problem, we now take a closer look at it.

The objective of the discretisation of a single continuous explanatory attribute is to find a partition of the values of this attribute which best discriminates the target distributions between groups. These groups are intervals and the partition evaluation is based on a compromise : fewer intervals and stronger target discrimination are better. There are mainly two families of search algorithms : bottom-up greedy agglomerative heuristics and top-down greedy divisive ones.

Discrimination can be evaluated in four ways using statistical test, entropy, description length or bayesian prior :

- Chimerge [9] applies chi square measure to test the independance of the distributions between groups,
- C4.5 [12] uses Shannon's entropy based information measures to find the most informative partition,
- MDLPC [6] defines a description length measure, following the Minimum Description Length principle [13],
- MODL [1] states a prior probability distribution, leading to a bayesian evaluation of the partitions.

The discretisation problem is illustrative of the convenience of supervised partitioning methods for data preparation since it addresses simultaneously the three following problems :

- Data representation : a suitable representation of the objects at hand have to be selected. Partitioning is an efficient mean to evaluate representations quality (in the supervised context, statistical test for class separability is another one, cf. [14]).
- Interpretability : labelled groups result from an understandable compromise between partition simplicity and target discrimination.
- Comparison capacity : explanatory attributes effects on the target can be quickly compared.

These themes are intertwined and play a crucial role in the data preparation phase (cf. Table 1 for an intuitive illustration in the multivariate case). The goal of this paper is to set a framework for supervised partitioning and to specify an evaluation criterion, preserving the interpretability bias and allowing not to consider single continuous attributes only.

In the remainder of the paper, we first set our framework and a description method of partitions (section 2). Then, we propose a new evaluation criterion (section 3) and we test its validity on real and synthetic datasets (section 4). Finally, we conclude and point out future works (section 5).

	Labels distributions in groups								
	Group 1			Group 2			Group 3		
Explanatory attributes	Set.	Ver.	Vir.	Set.	Ver.	Vir.	Set.	Ver.	Vir.
Sepal width, sepal length, petal width, petal length	50	0	0	0	50	0	0	0	50
Petal width, petal length	50	0	0	0	50	1	0	0	49
Sepal width, sepal length	50	2	1	0	48	49			
Petal width	50	0	0	0	48	0	0	2	50

Table 1. Examples of resulting partitions of Fisher's Iris database for different representation spaces. Partitioning techniques allow, among other things, to carry out the selection of an attribute subset in an intelligible way, as the results are quickly interpretable and easily comparable. Here, we see that the three iris categories (Setosa, Versicolor and Virginica) are completely discrimated by the four attributes. However, one can consider petal width only. Furthermore, one can state that setosas distinguish themselves by their sepal width and length.

2 Graph constrained supervised partitioning

Let $O = \{o_1, \ldots, o_N\}$ be a finite set of objects. A target l_n lying in an alphabet of size J is associated to each object o_n and a graph structure G is set on O. This structure can be natural (road networks, web graphs, ...) or imposed (proximity graphs, partial orders, ...). In the remainder, we will suppose G non-oriented. Our problem consists in finding an optimal partition of G, considering partitions composed of connected groups with respect to the discrete structure (i.e *connected partitions*). As explained above, optimality of a partition relies on the correct balance between the structure of its groups and its discriminating power (cf Figure 1). The setting of the balance requires the definition of description parameters both for the structure and the target distribution.

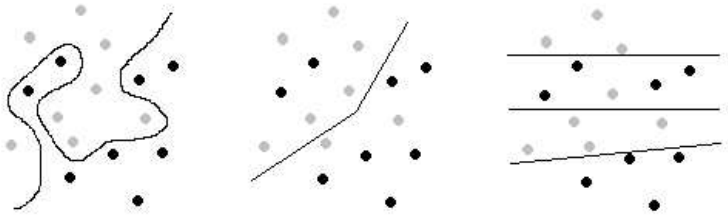

Fig. 1. 2 classes problem : which is the "best" partition?

Let π be a connected partition of G. We now introduce an effective and interpretable bias. We consider the balls induced by the discrete metric $\delta : \delta(o_1, o_2)$

is the minimum number of edges needed to link o_1 and o_2. As illustrated by Figure 2, each group of π is then covered with δ-balls.

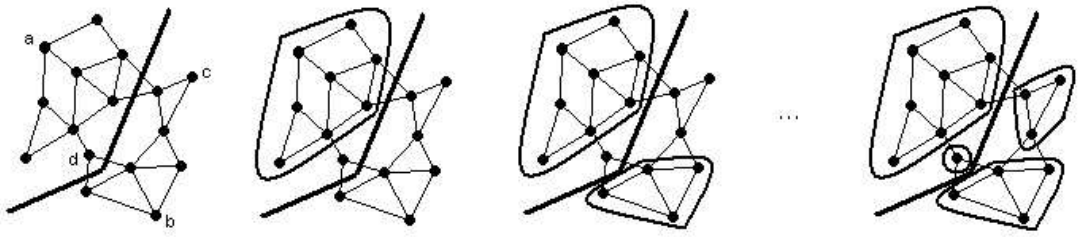

Fig. 2. Applying algorithm 1 : description of a partition with non-intersecting balls $(B(a,2), B(b,1), B(c,1), B(d,0))$ defined by the graph distance.

The method consists in selecting non-intersecting balls that are included in a group of π. At each step, the biggest one is picked up :

Algorithm 1 :

- $A \leftarrow O$
- $B \leftarrow \emptyset$
- **While** $A \neq \emptyset$ **Do**
  - $S \leftarrow$ the ball with maximal size included in A and in a group of π
  - $B \leftarrow B \bigcup \{S\}$
  - $A \leftarrow A \setminus S$

However, the set B does not characterise π : different partitions can give the same set B (cf Figure 3). But if the number of groups K is considered as a description parameter, obtaining π from B is the same as putting these balls in K different boxes. Finally, π is fully described by the set of balls B, the number K and a partition of B in K groups. This is not a compact description as we do not take into account the graph structure in the second step. Indeed, some partitions of B in K groups do not lead to connected partitions and should not be taken into account.

The description parameters of the target distribution are more easily catched. If π_k is one of the K groups in π, describing its inner labels distribution is the same as putting the objects contained in π_k in J boxes. This is done by firstly assigning the numbers N_{kj} of objects in π_k to put in the j^{th} box, and secondly specifying the partition of the group π_k in J groups of sizes $N_{k1}, \ldots, N_{kJ}$.

The description bias allows to define the *structural complexity* of π relying on its ball decomposition in an interpretable way : fewer and bigger balls means simpler structure. The description of the target disribution in terms of frequency

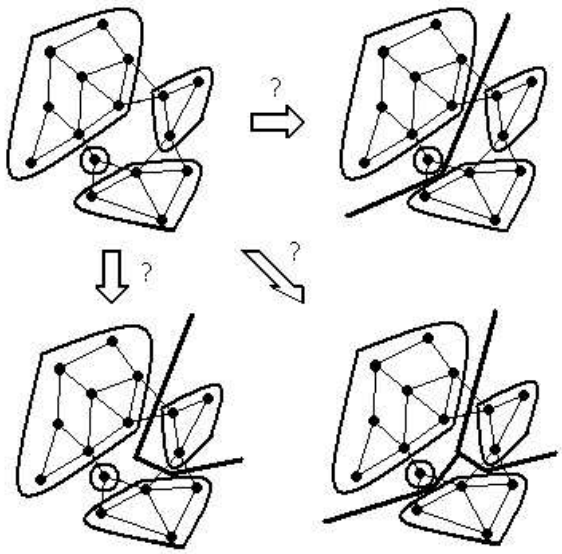

Fig. 3. Examples of possible partitions obtained with different groupings of the balls.

parameters leads to an informational definition of the *target discrimination* : strong discrimination is related to low entropy. The evaluation of a partition must result from a compromise as strong discrimination goes with high structural complexity.

3 Evaluation criterion

Let π be a connected partition of O. To set an evaluation criterion $l(\pi)$, the Minimum Description Length principle is applied [13], for its intrinsic ability to handle compromises. The problem turns into a two-step description problem : description of the parameters defining groups and description of the labelling parameters. This leads to write

$$l(\pi) := l_{structure}(\pi) + l_{labels/structure}(\pi),$$

with l standing for description lengths function. A description protocol must be designed from which description lengths can be specified.

As the structure is characterised by a set of balls B and a partition of B, its description length is split into the sum of the descritpion length of B and that of the partition of B :

$$l_{structure}(\pi) := l_{ballsset}(\pi) + l_{ballsgrouping}(\pi).$$

As the distributions are characterised in each group by the frequencies of the labels and a partition, the related description length is split into the following

way :

$$l_{labels/structure}(\pi) := \sum_{k=1}^{K} l_{frequencies}(\pi_k) + \sum_{k=1}^{K} l_{partitioning/frequencies}(\pi_k),$$

where $\pi = (\pi_1, \ldots, \pi_K)$.

In the first place, let's form a description protocol of the balls set B. The balls in B are ordered by decreasing sizes $d_1 > \cdots > d_p$ and if t_i is the number of balls of size d_i, B_j^i refers to the j^{th} ball ($1 \leq j \leq t_i$) of size d_i ($1 \leq i \leq p$). The protocol consists in specifying by decreasing size which is the next ball of the description and, when needed, the next size to be considered. Precisely :

- $p \leftarrow 1$
- describe d_p
- **if** $d_p < N$ **then**
  - **While** $d_p > 1$
    - describe successively $B_1^p, \ldots, B_{t_p}^p$
    - $p \leftarrow p + 1$
    - describe d_p

As description lengths can be interpreted as negative log of probabilities, we just have to assign probabilities to obtain $l_{ballsset}(\pi)$. We choose a uniform prior for each parameter description step and description lengths are computed using counting. For example, description length of d_1 is $\log_2(N)$ as possible values of d_1 are $1, \ldots, N$. Description length of d_2 is $\log_2(d_1 - 1)$ as, at this step, the possible values of d_2 are $1, \ldots, d_1$ and so on. Besides, the description length of B_1^1 is $\log_2 \beta_1^1$, where β_1^1 is the total number of balls of size d_1 induced by the discrete structure G. That of B_2^1 is $\log_2 \beta_2^1$, where β_2^1 stands for the total number of balls of size d_1 induced by G that do not intersect B_1^1, etc... The overall sum of these lengths defines $l_{ballsset}(\pi)$.

In the second place, to set $l_{ballsgrouping}(\pi)$, we describe the group number K of π and the partition of B in K groups. Once again, a uniform prior is applied. As K lies between 1 and the size K_B of B and as the number of partitions of B in less than K groups is $B(K_B, K)$ (the sum of the K first Stirling numbers), we have

$$l_{ballsgrouping}(\pi) = \log_2 K_B + \log_2 B(K_B, K).$$

In the third and final place, applying a uniform prior to obtain the description lengths of the target leads to set

$$l_{labels/structure}(\pi) = \sum_{k=1}^{K} \log_2 \binom{N_k + J - 1}{J - 1} + \sum_{k=1}^{K} \log_2 \frac{N_k!}{N_{k1}! \ldots N_{kJ}}.$$

The first sum results from the description of the labels frequencies $(N_{k1}, \ldots, N_{kJ})$ in each group k. These J-tuples satisfy the property $\sum N_{kj} = N_k$ (with N_k the size of group k), and $\binom{N_k+J-1}{J-1}$ is the number of such tuples. The number of partitions of a set of size N_k in J groups of sizes $N_{k1}, \ldots, N_{kJ}$ is the multinomial coefficient $\frac{N_k!}{N_{k1}! \ldots N_{kJ}!}$. That gives the second sum.

4 Experiments

The experiments are performed using the standard hierarchical greedy bottom-up heuristic, the initial partition being that with one object per group :

Algorithm 2

- $\pi \leftarrow InitialPartition$
- **For** $k = 2$ *to* N **Do**
  - $\pi \leftarrow$ the best partition resulting from the merging of two groups of π
- **Return** the overall best partition encountered

Thus, $O(N^2)$ partitions are evaluated. The greedy character of this heuristic does not allow to evaluate a significant part of the partitions set and such a method easily falls into local optima. To alleviate these facts, we select a more appropriate initial partition : initial groups are the biggest clean balls (i.e objects in a ball share the same label).

A graph structure has to be selected. As the objects are always imbedded in an euclidean space, the experiments are carried out with the Gabriel graph, which is a proximity graph [7]. The distance between two objects o_1 and o_2 is taken to be the imbedding euclidean one L and these objects are adjacent in the Gabriel sense (cf Figure 4) if and only if

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

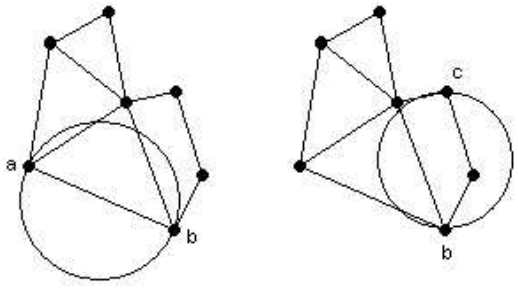

Fig. 4. Example of a Gabriel graph. The ball of diameter $[ab]$ contains no other point : a and b are Gabriel-adjacent. The ball of diameter $[bc]$ contains another point : b and c are not Gabriel-adjacent.

We perform two experiments on synthetic datasets and one on real datasets. In a first one, we check the criterion ability to detect the independence between the descriptive attributes and the target one, on synthetic datasets. These are

two-classes problems, with points uniformly generated inside the Hamming hypercube and each point label uniformly assigned. The varying parameters are the number N (from 1 to 100) of points and the space dimension d (taking values 1, 2, 3, 5 and 10). For each couple of values, 25 datasets are generated. For every dataset, our method builds a partition composed of one single group. This is exactly the expected behavior : no discrimination has to be done since the target is independent of the explanatory attributes.

In a second experiment, we test the criterion discrimination ability for gaussian mixture models in the plane. We settle a four gaussians problem, centered in $(1,1),(-1,1),(-1,-1)$ and $(1,-1)$, with diagonal covariance matrix $\begin{pmatrix} 1/4 & 0 \\ 0 & 1/4 \end{pmatrix}$. The varying parameter is the number N of points and for each value, 25 datasets were generated. Figure 5 shows that, with sufficiently many points, the four groups are detected. The detection threshold could however be better. Indeed, as we do not take into account the graph structure for the description of the balls set partition, the description length $l_{ballsgrouping}$ is over-estimated. To obtain a decrease of the total description length, the (too big) increasing of the structural length induced by the decision of creating a new group must be balanced by a (very) strong resulting discrimination.

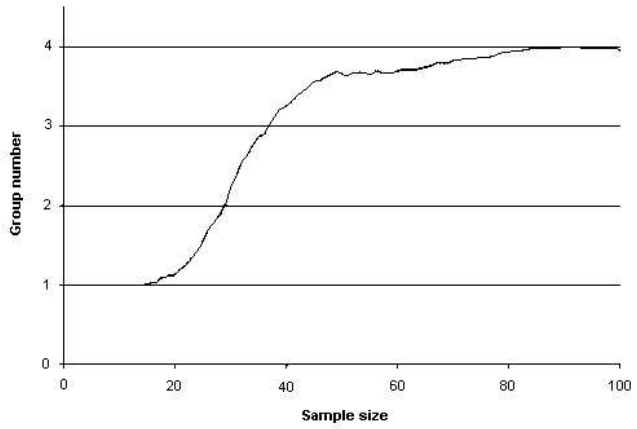

Fig. 5. Resulting number of groups on the four gaussians problem.

In a third experiment, we consider the resulting partition as a predictive model : a new instance is classified according to a majority vote in the nearest group. The evaluation consists in a stratified five-fold cross-validation and results of the Nearest Neighbor (NN) rule [4] are given for comparison. The tests were carried through 3 datasets from the UCI machine learning database repository.

The well-known Fisher's Iris dataset contains 150 instances of such flowers described by 4 continuous explanatory attributes and belonging to one of the

three classes Setosa, Versicolor and Virginica (as previously seen). The Wine database results from the analysis of 13 components found in each of 3 types of wines, and is composed of 178 instances. Finally, the Breast dataset aims at studying the malignant character of a breast cancer for 699 subjects through 9 descriptive attributes.

In order to limit scale effects on the distance measure, each explanatory attribute is linearly transformed to lie in $[0, 1]$. Table 2 summarizes the results of the evaluation. The main advantage of partitioning methods lies in the fact that they detect or supply an underlying structure of the analysed data. As this structural gain may be balanced by an information loss, it's noteworhty that, on the three datasets, our technique does not suffer from such a curse.

Dataset	NN	Partition	Group number
Iris	0.95 ± 0.03	0.94 ± 0.04	3 ± 0.0
Wine	0.95 ± 0.03	0.94 ± 0.04	3 ± 0.0
Breast	0.96 ± 0.01	0.96 ± 0.01	2 ± 0.0

Table 2. Prediction accuracy of both NN and partition-based rules and group number of the partition. Our method gives additional information : each class lies in a single cluster, for every datasets.

5 Conclusion and further works

In this paper, we have discussed the usefulness of supervised partitioning methods for data preparation, set a framework for supervised partitioning, proposed and tested an evaluation criterion of labelled partition. The representation quality of the objects and their inner amount of information about the target attribute can be subtly and simply evaluated, whatever may be the kind of the objects. Specifically, multivariate representations can be considered.

The proposed method builds an underlying structure of the data : a partition. This is done in an understandable way (with the use of balls) and without loss of predictive information (as shown by the experiments on real datasets). The settled criterion is able to detect independance too. If the explanatory attributes contain no information with respect to the target attribute, the "best" partition should be that with one group and that's the way the criterion behaves.

Still, this is preliminary work. The presented criterion can be improved. The "balls grouping" description part could take into account the graph structure, leading to a more accurate evaluation criterion.

As well, the greedy agglomerative approach is not effective and easily falls into a local optimum. Furthermore, the heuristic lacks of computational efficiency : the complexity's polynomial order is too high for real applications. In future works, we plan to design a heuristic founded on the description bias (the use of balls).

References

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

[2] Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and regression trees. California: Wadsworth International (1984)

[3] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R.: CRISP-DM 1.0 : step-by-step data mining guide. Applied Statistics Algorithms (2000)

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

[5] Duda, R.O., Hart, P.E., Stork, D.G.: Pattern classification. Wiley-Interscience (2000)

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

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

[8] Kass, G.: An exploratory technic for investigating large quantities of categorical data. Journal of Applied Statistics **29** (1980) 119–127

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

[10] Kohonen, T.: Self-organized formation of topologically correct feature maps. Biological Cybernetics **43** (1982) 59–69

[11] McQueen, J.: Some methods for classification and analysis of multivariate observations. Fifth Berkeley Symposium on Mathematical Statistics and Probability, Le Cam and Neyman **1** (1967) 281–297

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

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

[14] Zighed, D.A., Lallich, S., Muhlenbach, F.: Séparabilité des classes dans $\mathbb{R}^p$. VIIIème Congrès de la Société Francophone de Classification (2001) 356–363