A Non-parametric Semi-supervised Discretization Method

Bondu, A. and Boulle, M. and Lemaire, V.
Orange Labs
2 av. Pierre Marzin
22300 Lannion - France
alexis.bondu@orange-ftgroup.com

Loiseau, S. and Duval, B.
LERIA, Université d'Angers
2 Boulevard Lavoisier,
49045 Angers Cedex 01, France

Abstract

Semi-supervised classification methods aim to exploit labelled and unlabelled examples to train a predictive model. Most of these approaches make assumptions on the distribution of classes. This article first proposes a new semi-supervised discretization method which adopts very low informative prior on data. This method discretizes the numerical domain of a continuous input variable, while keeping the information relative to the prediction of classes. Then, an in-depth comparison of this semi-supervised method with the original supervised MODL approach is presented. We demonstrate that the semi-supervised approach is asymptotically equivalent to the supervised approach, improved with a post-optimization of the intervals bounds location.

1 Introduction

Data mining can be defined as the non trivial process of identifying valid, novel, potentially useful, ultimately understandable patterns in data [10]. Even though the modeling phase is the core of the process, the quality of the results rely heavily on data preparation which usually takes around 80% of the total time [17]. An interesting method for data preparation is to discretize the input variables.

Discretization methods aim to induce a list of intervals which splits the numerical domain of a continuous input variable, while keeping the information relative to the output variable [6] [12] [8] [14] [15]. A naïve Bayes classifier can exploit a discretization of its input space [3] as the intervals set which is used to estimate conditional probabilities of classes given the data. Discretization methods are useful for data mining, to explore, prepare and model data.

The objective of semi-supervised learning is to exploit unlabelled data to improve a predictive models. This article focuses on semi-supervised classification, a well known problem in the literature. Most of semi-supervised approaches deal with particular cases where information about unlabelled data is available. Semi-supervised learning without strong assumption on data distribution is a great challenge.

This article proposes a new semi-supervised discretization method which adopts very low informative priors on data. Our semi-supervised discretization method is based on the MODL framework [3] *("Minimal Optimized Description Length")*. This approach turns the discretization problem into a model selection one. A Bayesian approach is applied and leads to an analytical evaluation criterion. Then, the best discretization model is selected by optimizing this criterion.

The organization of this paper is as follows: Section 2 presents the motivation for non-parametric semi-supervised learning; Section 3 formalizes our semi-supervised approach; our discretization method is compared with the supervised approach in Section 4; in Section 5, empirical and theoretical results are exploited to demonstrate that the semi-supervised approach is asymptotically equivalent to the supervised approach, improved with a post-optimization of the intervals bounds location. Future work is discussed in Section 7.

2 Related works

This section introduces the semi-supervised learning owing to a short state of the art. Previous works on supervised discretization are then summarized.

2.1 Semi-supervised algorithms

Semi-supervised classification methods [7] exploit labelled and unlabelled examples to train a predictive model. The main existing approaches are the following:

- The **Self-training** approach is a heuristic which iteratively uses the predictions of a model to label new examples. The new labelled examples in turns are used to

train the model. The uncertainty of predictions is evaluated in order to label only the most confident examples [18].

- The **Co-training** approach involves two predictive models which are independently trained on disjoint subfeature sets. This heuristic uses the predictions of both models to label two examples at every iteration. Each model labels one example and "teaches" the other classifier with its prediction [2] [16].
- The **Covariate shift** approach estimates the distributions of labelled and unlabelled examples [21]. The covariate shift formulation [20] weights labelled examples according to the disagreement between these distributions. This approach incorporates this disagreement into the training algorithm of a supervised model.
- **Generative model** based approaches estimate the distribution of classes, under hypothesis on data. These methods make the assumption that the distributions of classes belong to a known parametric family. Then training data is exploited in order to fit parameters values [11].

Semi-supervised learning without making hypothesis on data distribution is a great challenge. Therefore, most of semi-supervised approaches make assumptions on the distribution of classes.

For instance, generative model based approaches aim to estimate $P(x,y) = P(y)P(x|y)$ the joint distribution of data and classes (with data denoted by $x \in \mathbb{X}$ and classes denoted by $y \in \mathbb{Y}$). The distribution $P(x,y)$ is assumed to belong to a parametric family $\{P(x,y)_\theta\}$. The vector θ of finite size corresponds to the modeling parameters of $P(x,y)$. The joint distribution can be rewritten as $P(x,y)_\theta = P(y)_\theta P(x|y)_\theta$. The term $P(y)_\theta$ is defined by a prior knowledge on the distribution of classes. $P(x|y)_\theta$ is identified in a given family of distributions, thanks to the vector θ.

Let U be the set of unlabelled examples and L the set of labelled examples. The set L contains couples (x,y), with x a scalar value and $y \in [1,J]$ a discrete class value. The set U contains scalar values without labels. Semi-supervised generative model based approaches aim to find the parameters θ which maximize $P(x,y)_\theta$ on the data set $D = U \cup L$. The quantity to be maximized is $p(L,U|\theta)$, the probability of data given the parameters θ. The Maximum Likelihood Estimation (MLE) is widely employed to maximize $p(L,U|\theta)$ (with $(x_i,y_i) \in L$ and $x_{i'} \in U$):

$$\max_{\theta \in \Theta} \left[\sum_{i=1}^{|L|} \log \left[p(y_i)_\theta p(x_i|y_i)_\theta \right] + \sum_{i'=1}^{|U|} \log \left[\sum_{j'=1}^{|\mathbb{Y}|} p(y_{j'})_\theta p(x_{i'}|y_{j'})_\theta \right] \right] \tag{1}$$

These approaches are usable only if information about the distribution of classes is available. The hypothesis that $P(x,y)$ belongs to a known family of distributions is a strong assumption which could be invalid in practice.

The objective of a non-parametric semi-supervised method is to estimate the distribution of classes without making strong hypothesis on these distributions. Therefore, our approach can be put in opposition with the generative approaches. This article exploits the MODL framework [3] and proposes a new semi-supervised discretization method. This "objective" Bayesian approach makes very low assumptions on the data distribution.

2.2 Summary of the supervised MODL discretization method

The discretization of a descriptive variable aims at estimating the conditional distribution of class labels, owing to a piece wise constant density estimator. In the MODL approach [3], the discretization is turned into a model selection problem. First, a space of discretization models is defined. The parameters of a specific discretization are the number of intervals, the bounds of the intervals and the output frequencies in each interval. A Bayesian approach is applied to select the best discretization model, which is found by maximizing the probability $P(M|D)$ of the model M given the data D. Using the Bayes rule and since the probability $P(D)$ is constant under varying the model, this is equivalent to maximizing $P(M)P(D|M)$.

Let N^l be the number of labelled examples, J the number of classes, I the number of intervals for the input domain. N_i^l denotes the number of labelled examples in the interval i, and N_{ij}^l the number of labelled examples of output value j in the interval i. A discretization model is then defined by the parameter set $\left\{ I, \left\{ N_i^l \right\}_{1 \leq i \leq I}, \left\{ N_{ij}^l \right\}_{1 \leq i \leq I, 1 \leq j \leq J} \right\}$.

Owing to the definition of the model space and its prior distribution, the prior $P(M)$ and the conditional likelihood $P(D|M)$ can be calculated analytically. Taking the negative log of $P(M)P(D|M)$, we obtain the following criterion to minimize:

$$\mathcal{C}_{sup} = \underbrace{\log N^l + \log \binom{N^l + I - 1}{I - 1} + \sum_{i=1}^{I} \log \binom{N_i^l + J - 1}{J - 1}}_{-\log P(M)} + \underbrace{\sum_{i=1}^{I} \log \frac{N_{i.}^l !}{N_{i1}^l ! N_{i2}^l ! \ldots N_{iJ}^l !}}_{-\log P(D|M)} \tag{2}$$

The first term of the criterion $\mathcal{C}_{sup}$ stands for the choice of the number of intervals and the second term for the

choice of the bounds of the intervals. The third term corresponds to the choice of the output distribution in each interval and the last term represents the conditional likelihood of the data given the model. Therefore "complex" models with large numbers of intervals are penalized. This discretization method for classification provides the most probable discretization given the data sample. Extensive comparative experiments showed high performances [3].

3 A new semi-supervised discretization method

This section presents a new semi-supervised discretization method which is based on previous work described above. The same modeling hypothesis as [3] are adopted. A prior distribution $P(M)$ which exploits the hierarchy of the model parameters is first proposed. This prior distribution is uniform at each stage of this hierarchy. Then, we define $P(D|M)$ the conditional likelihood of data given the model. This leads to an exact analytical criterion for the posterior probability $P(M|D)$.

Discretization models:
Let $\mathbb{M}$ be a family of semi-supervised discretization models denoted $M(I, \{N_i\}, \{N_{ij}\})$. These models consider unlabelled and labelled examples together, and N is the total number of examples in the data set. The models parameters are defined as follows: I is the number of intervals; $\{N_i\}$ the number of examples in each interval; $\{N_{ij}\}$ the number of examples of each class in each interval.

3.1 Prior distribution

A prior distribution $P(M)$ is defined on the parameters of the models. This prior exploits the hierarchy of the parameters. The number of intervals is first chosen, then the bounds of the intervals and finally the output frequencies are chosen. The joint distribution $P(I, \{N_i\}, \{N_{ij}\})$ can be written as follows:

$$P(M) = P(I, \{N_i\}, \{N_{ij}\})$$

$$P(M) = P(I) \times P(\{N_i\}|I) \times P(\{N_{ij}\}|\{N_i\}, I)$$

The number of intervals is assumed to be uniformly distributed between 1 and N. Thus we get:

$$P(I) = \frac{1}{N}$$

We now assume that all data partitions into I intervals are equiprobable for a given number of intervals. Computing the probability of one set of intervals turns into the combinatorial evaluation of the number of possible intervals sets, which is equal to $\binom{N+I-1}{I-1}$. The second term is defined as:

$$P(\{N_i\}|I) = \frac{1}{\binom{N+I-1}{I-1}}$$

The last term $P(\{N_{ij}\}|\{N_i\}, I)$ can be rewritten as a product, assuming the independence of the distribution of classes between the intervals. For a given interval i containing N_i examples, all the distributions of the class values are considered equiprobable. The probabilities of distributions are computed as follows:

$$P(\{N_{ij}\}|\{N_i\}, I) = \prod_{i=1}^{I} \frac{1}{\binom{J-1}{N_i+J-1}}$$

Finally, the prior distribution of the model is similar to the supervised approach [3]. The only one difference is that the semi-supervised prior takes into account all examples, including unlabelled ones:

$$P(M) = \frac{1}{N} \times \frac{1}{\binom{N+I-1}{I-1}} \times \prod_{i=1}^{I} \frac{1}{\binom{J-1}{N_i+J-1}} \qquad (3)$$

3.2 Likelihood

This section focuses on the conditional likelihood $P(D|M)$ of the data given the model. First, the family Λ of labelling models has to be defined. Semi-supervised discretization handles labelled and unlabelled pieces of data, Λ represents all possible labellings. Each model $\lambda(N^l, \{N_i^l\}, \{N_{ij}^l\}) \in \Lambda$ is characterized by the following parameters: N^l is the total number of labelled examples; $\{N_i^l\}$ the number of labelled examples in the interval i; $\{N_{ij}^l\}$ the number of labelled examples of the class j in the interval i.

Owing to the formula of the total probability, the likelihood can be written as follows:

$$P(D|M) = \sum_{\lambda \in \Lambda} P(\lambda|M) \times P(D|M, \lambda)$$

$P(D|M)$ can be drastically simplified considering that $P(D|M, \lambda)$ is equal to 0 for all labelling models which are incompatible with the observed data D and the discretization model M. The only one compatible labelling model that is considered is denoted λ^*. The previous expression can be rewritten as follows:

$$P(D|M) = P(\lambda^*|M) \times P(D|M, \lambda^*) \qquad (4)$$

The first term $P(\lambda^*|M)$ can be written as a product using the hypothesis of independence of the likelihood between the intervals. In a given interval i which contains N_{ij} examples of each class, the computation of $P(\lambda^*|M)$ consists in finding the probability of observing $\{N_{ij}^l\}$ examples of each class, drawing N_i^l examples. Once again, this problem is turned into a combinatorial evaluation. The number of draws which induce $\{N_{ij}^l\}$ can be calculated, assuming the N_i^l labelled examples are uniformly drawn:

$$P(\lambda^*|M) = \prod_{i=1}^{I} \frac{\prod_{j=1}^{J} \binom{N_{ij}}{N_{ij}^l}}{\binom{N_i}{N_i^l}} \quad (5)$$

Let us consider a very simple and intuitive problem to explain Equation 5. An interval i can be compared with a "bag" containing N_{i1} "black balls" and N_{i2} "white balls". Given the parameters $N_{i1} = 6$ and $N_{i2} = 20$, what is the probability to simultaneously draw $N_{i1}^l = 2$ black balls and $N_{i2}^l = 3$ white balls? Let $\binom{26}{5}$ be the number of possible draws, and $\binom{6}{2} \times \binom{20}{3}$ the number of draws which are composed of 2 black balls and 3 white balls. Assuming that all draws are equiprobable, the probability to simultaneously draw 2 black balls and 3 white balls is given by: $\frac{\binom{6}{2}\times\binom{20}{3}}{\binom{26}{5}}$.

The second term $P(D|M, \lambda^*)$ of Equation 4 is estimated considering a uniform prior over all possible permutations of $\{N_{ij}^l\}$ examples of each class among N_i^l. The independence assumption between the intervals gives:

$$P(D|M, L^*) = \prod_{i=1}^{I} \frac{1}{\frac{N_i^l!}{N_{i1}^l! N_{i2}^l! \ldots N_{iJ}^l!}} = \prod_{i=1}^{I} \frac{\prod_{j=1}^{J} N_{ij}^l!}{N_i^l!}$$

Finally, the likelihood of the model is:

$$P(D|M) = \prod_{i=1}^{I} \frac{\prod_{j=1}^{J} \binom{N_{ij}}{N_{ij}^l} \times N_{ij}^l!}{\binom{N_i}{N_i^l} \times N_i^l!}$$

In every interval, the number of unlabelled examples is denoted by $N_{ij}^u = N_{ij} - N_{ij}^l$ and $N_i^u = N_i - N_i^l$. The previous expression can be rewritten:

$$P(D|M) = \prod_{i=1}^{I} \frac{\prod_{j=1}^{J} \frac{N_{ij}!}{N_{ij}^u!}}{\frac{N_i!}{N_i^u!}}$$

$$P(D|M) = \prod_{i=1}^{I} \left[\frac{\prod_{j=1}^{J} N_{ij}!}{N_i!} \times \frac{N_i^u!}{\prod_{j=1}^{J} N_{ij}^u!} \right] \quad (6)$$

3.3 Evaluation criterion

The best semi-supervised discretization model is found by maximizing the probability $P(M|D)$. A Bayesian evaluation criterion is obtained exploiting Equations 3 and 6. The maximum a posteriori model, denoted "M_{map}", is defined by:

$$M_{map} = \max_{M \in \mathbb{M}} \left[\frac{1}{N} \times \frac{1}{\binom{N+I-1}{I-1}} \times \prod_{i=1}^{I} \frac{1}{\binom{N_i+J-1}{J-1}} \times \prod_{i=1}^{I} \left[\frac{\prod_{j=1}^{J} N_{ij}!}{N_i!} \times \frac{N_i^u!}{\prod_{j=1}^{J} N_{ij}^u!} \right] \right] \quad (7)$$

Taking the negative log of the probabilities, the maximization problem turns into the minimization of the criterion $\mathcal{C}_{semi\ sup}$:

$$\begin{aligned} M_{map} &= \min_{M \in \mathbb{M}} \mathcal{C}_{semi\ sup}(M) \\ &= \min_{M \in \mathbb{M}} \left[\log(N) + \log \binom{N+I-1}{I-1} \right. \\ &\quad + \sum_{i=1}^{I} \log \binom{N_i+J-1}{J-1} + \sum_{i=1}^{I} \log \left(\frac{N_i!}{\sum_{j=1}^{J} N_{ij}!} \right) \\ &\quad \left. - \sum_{i=1}^{I} \log \left(\frac{N_i^u!}{\sum_{j=1}^{J} N_{ij}^u!} \right) \right] \end{aligned} \quad (8)$$

4 Comparison: semi-supervised *vs.* supervised criteria

In this section, the semi-supervised criterion $\mathcal{C}_{semi\ sup}$ of Equation 8 is compared with the supervised criterion $\mathcal{C}_{sup}$ of Equation 2:

- both criteria are analytically equivalent when $U = \emptyset$;
- the semi-supervised criterion corresponds to the prior distribution when $L = \emptyset$, in this case, semi-supervised and supervised approaches give the same discretization;
- the semi-supervised approach is penalized by a high modeling cost when the data set includes labelled and unlabelled examples, in this case, the optimization of the criterion $\mathcal{C}_{semi\ sup}$ gives a model with less intervals than the supervised approach.

4.1 Labelled examples only

In this case, all training examples are supposed to be labelled: $D = L$ and $U = \emptyset$. We have $N_i^u = 0$ for each interval and $N_{ij}^u = 0$ for each class. Therefore, the last term of Equation 8 is equal to zero. The criterion $\mathcal{C}_{semi\ sup}$ can be rewritten as follows:

$$\begin{aligned} &\log(N) + \log \binom{N+I-1}{I-1} + \sum_{i=1}^{I} \log \binom{N_i+J-1}{J-1} \\ &+ \sum_{i=1}^{I} \log \left(\frac{N_i!}{\sum_{j=1}^{J} N_{ij}!} \right) \end{aligned} \quad (9)$$

When all the training examples are labelled, $N = N^l$, $N_i = N_i^l$ and $N_{ij} = N_{ij}^l$. The semi-supervised criterion $\mathcal{C}_{semi\ sup}$ and the supervised criterion $\mathcal{C}_{sup}$ are equivalent.

4.2 Unlabelled examples only

In the case where no example is labelled we have $D = U$ and $L = \emptyset$. For each interval $N_i^u = N_i$ and for each class $N_{ij}^u = N_{ij}$. Therefore, the term $P(D|M)$ is equal to 1 for any model. The conditional likelihood (Equation 6) can be rearranged as follows :

$$P(D|M) = \prod_{i=1}^{I} \left[\frac{\prod_{j=1}^{J} N_{ij}!}{N_i!} \times \frac{N_i!}{\prod_{j=1}^{J} N_{ij}!} \right]$$

$$P(D|M) = 1$$

The posterior distribution is only composed by the prior distribution $P(M|D) = P(M)$, in which case the model M_{map} includes a single interval. Both criteria give the same discretization, as long as supervised approach is not able to cut the numerical domain of the input variable in this case. $\mathcal{C}_{semi\ sup}$ can be rewritten as:

$$\log(N) + \log \binom{N+I-1}{I-1} + \sum_{i=1}^{I} \log \binom{N_i+J-1}{J-1}$$

4.3 Mixture of labelled and unlabelled examples

The main difference between the semi-supervised and the supervised approaches consists in the prior distribution $P(M)$. In semi-supervised approach, the space of discretization models is bigger than in the supervised approach. Unlabelled examples represent additional possible locations for the intervals bounds. Therefore, the modeling cost of the prior distribution is more important for the semi-supervised criterion. When the number of unlabelled examples increases, the criterion $\mathcal{C}_{semi\ sup}$ prefers models with less intervals.

This behaviour is illustrated with a very simple experiment. Let us consider a binary classification problem. All examples belonging to the class "0" [*respectively* "1"] are located at $x = 0$ [*respectively* $x = 1$]. During the experiment, N the number of examples increases. The number of labelled examples is always the same in both classes. For every value of N, we evaluate N_{min}^l the minimal number of labelled examples which induces a M_{map} with two intervals (and not a single interval).

Figure 1 plots N_{min}^l against $N = N^l + N^u$ for both criteria. For the criterion $\mathcal{C}_{sup}$, the minimal number of labelled examples necessary to split data does not depend on N. In this case, $N_{min}^l = 6$ for every value of N. A different behaviour is observed for $\mathcal{C}_{semi\ sup}$. Figure 1 quantifies the influence of N on the selection of the model M_{map}. When

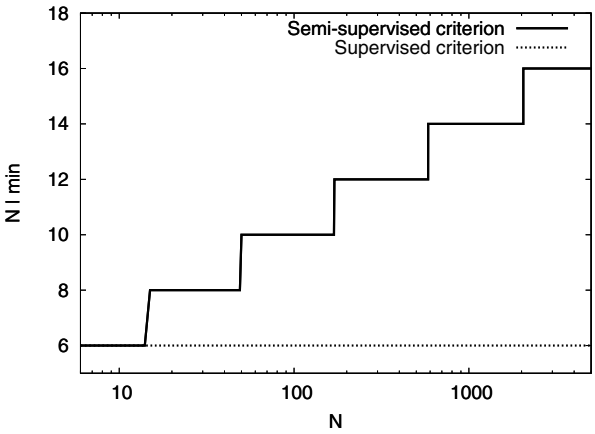

Figure 1. Mixture of labelled and unlabelled examples. The vertical axis represents the minimal number of labelled examples necessary to obtain a model with two intervals, rather than a model with a single interval. The horizontal axis represents the total number of examples using a logarithmic scale.

the number of examples N grows, N_{min}^l increases approximately as $\log(N)$. Therefore, the criterion $\mathcal{C}_{sup}$ gives a model M_{map} with less intervals than the supervised approach, due to its high modeling cost.

5 Theoretical and empirical results

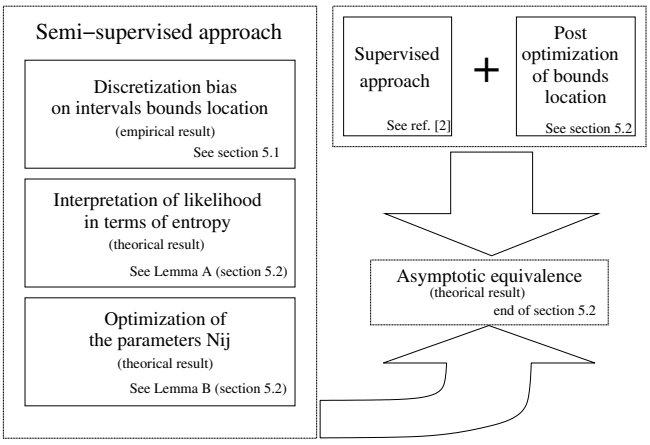

Figure 2. Structure of Section 5

Figure 2 illustrates the structure of the results presented in this section, and their relations. An additional discretization bias is first empirically established for our semi-supervised discretization method. Then, two theoretical re-

sults are demonstrated: an interpretation of the likelihood in terms of entropy; and an analytical expression of the optimal N_{ij}. Taking into account these empirical and theoretical results, we demonstrate that the semi-supervised approach is asymptotically equivalent to the supervised approach, associated with a post-optimization of the bounds location.

5.1 Discretization bias

The semi-supervised and the supervised discretization approaches are based on the ranks statistics. Therefore, the location of the bounds between intervals of the optimal model are defined in a discrete space, thanks to the number of examples in every interval. The discretization bias aims to define the bounds location in the numerical domain of the continuous input variable.

5.1.1 How to position a bound between two training examples?

The parameters $\{N_i\}$ [*respectively* $\{N_i^l\}$] given by the optimization of $\mathcal{C}_{semi\ sup}$ [*respectively* $\mathcal{C}_{sup}$] are not sufficient to define continuous bounds location. Indeed, there is an infinity of possible locations between two training examples. A prior is adopted in [3] which considers the best bound location as the median of the distribution of the true bound locations, denoted P_{tb}. This median minimizes the generalization Means Square Error, for any P_{tb}. The objective is to place a bound between two examples without information about the distribution P_{tb}. In this case P_{tb} is assumed to be uniform, and a bound is placed midway between the two concerned examples.

5.1.2 How to position a bound in an unlabelled area?

The optimization of the semi-supervised criterion $\mathcal{C}_{semi\ sup}$ does not indicate the best bounds location, when the parameters $\{N_i^l\}$ are constant. This phenomenon is observed on a toy example below. Considering an area of the input space $\mathbb{X}$ where no example is labelled, all possible bounds locations have the same cost according to the criterion $\mathcal{C}_{semi\ sup}$. Therefore, the semi-supervised approach is not able to determine bounds location in such an unlabelled area. The same prior as [3] which aims minimizing the generalization Means Square Error is adopted, in order to define continuous bounds location. The unlabelled examples are supposed to be drawn from the distribution P_{tb}. In this case, the median of P_{tb} is estimated exploiting the unlabelled examples. The intervals bounds are placed in the middle of unlabelled areas.

Finally, the supervised and the semi-supervised approaches are not able to position a continuous bound between two labelled examples. In both cases, the same prior on the best bounds location is adopted. The only one interest of the unlabelled examples is to bring information about P_{tb}, and to refine the median of this distribution.

5.1.3 Empirical evidence :

Let us consider an univariate binary classification problem. Training examples are uniformly distributed in the interval $[0, 1]$. This data set contains three separate areas denoted "A", "B", "C". The part "A" [*respectively* "C"] includes 40 labelled examples of class "0" [*respectively* "1"] and corresponds to the interval $[0, 0.4]$ [*respectively* $[0.6, 1]$]. The part "B" corresponds to the interval $[0.4, 0.6]$ and contains 20 unlabelled examples.

As part of this experiment, the family of discretization models $\mathbb{M}$ is restricted to the models which contain two intervals. This toy problem consists in finding the best bound $b \in [0, 1]$ between the two intervals of the model. Every bound is related to the number of examples in each intervals, $\{N_1, N_2\}$.

There are a lot of possible models for a given bound (due to the N_{ij} parameters). We estimate the probability of a bound by a Bayesian averaging over all possible models which are compatible with the bound. This evaluation is not biased by the choice of a particular model among all possible models. For a given bound b, the parameters $\{N_{ij}\}$ are not defined, we have:

$$P(b|D) = \sum_{\{N_{ij}\}} P(\underbrace{b, \{N_{ij}\}}_{M \in \mathbb{M}} | D)$$

Using the Bayes rule, we get:

$$P(b|D) \times P(D) = \sum_{\{N_{ij}\}} P(D|b, \{N_{ij}\}) \times P(b, \{N_{ij}\})$$

Figure 3 plots $-\log P(b|D)$ against the bound's location b. Minimal values of this curve give the best bound's locations. This figure indicates that it is neither wise to cut the data set in part "A" nor in part "C". All bound's locations in part "B" are equivalent and optimal according to the criterion $\mathcal{C}_{semi\ sup}$.

This experiment empirically shows that the criterion $\mathcal{C}_{semi\ sup}$ can not distinguish between bounds' location in an unlabelled area of the input space $\mathbb{X}$. This result is unexpected and difficult to demonstrate formally (due to the Bayesian averaging over models). Intuitively, this phenomenon can be explained by the fact that the criterion $\mathcal{C}_{semi\ sup}$ has no expressed preferences on bounds' location. This is consistent with an "objective" Bayesian approach [1].

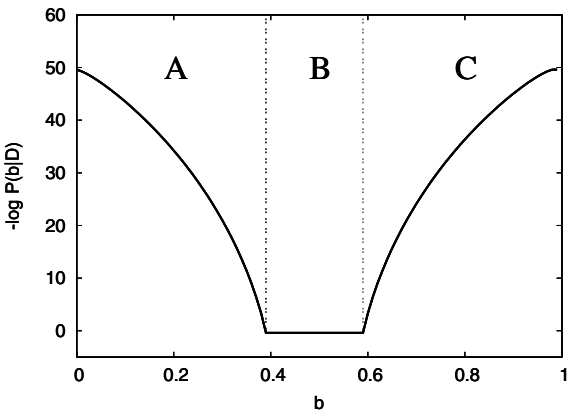

Figure 3. Bound's quantity of information *vs.* **bound's location**

5.2 A post-optimisation of the supervised approach

This section demonstrates that the semi-supervised approach is asymptotically equivalent to the supervised approach improved with a post-optimization on the bounds location. This post-optimization consists in exploiting unlabelled examples in order to position the intervals bounds in the middle of unlabelled areas.

5.2.1 Equivalent prior distribution

The discretization bias established in Section 5.1 modifies our a priori knowledge about the distribution $P(M)$. From now, the bounds are forced to be placed in the middle of unlabelled areas. The number of possible locations for each bound is substantially reduced. The criterion $\mathcal{C}_{semi\ sup}$ considers $N-1$ possible locations for each bound. Exploiting the discretization bias of Section 5.1, only N^l-1 possible locations are considered. In these conditions, the prior distribution $P(M)$ (see Equation 3) can be easily rewritten as in the supervised approach (see Equation 2).

5.2.2 Asymptotically equivalent likelihood

Lemma A: *The conditional likelihood of the data given the model can be expressed using the entropy (denoted H_M) of the sets U, L and D, given the model M:*

- *Supervised case*
  $-\log P(D|M)^* = N^l H_M(L) + \mathcal{O}(\log N)$
- *Semi-supervised case*
  $-\log P(D|M) = NH_M(D) - N^u H_M(U) + \mathcal{O}(\log N)$

Proof :

• Let us denote $H_M(D)$ the Shannon's entropy [19] of the data, given a discretization model M. We assume that $H_M(D)$ is equals to its empirical evaluation:

$$H_M(D) = N \times \sum_{i=1}^{I} \left[\frac{N_i}{N} - \sum_{j=1}^{J} \log \frac{N_{ij}}{N_i} \right]$$

• In the semi-supervised case $P(D|M) = \prod_{i=1}^{I} \frac{\prod_{j=1}^{J} \frac{N_{ij}!}{N_{ij}^u!}}{\frac{N_i!}{N_i^u!}}$.

Consequently:

$$-\log P(D|M) = \sum_{i=1}^{I} \left[\log(N_i!) - \log(N_i^u!) - \sum_{j=1}^{J} \log(N_{ij}!) + \sum_{j=1}^{J} \log(N_{ij}^u!) \right]$$

The Stirling's approximation gives $\log(n!) = n\log(n) - n + \mathcal{O}(\log n)$:

$$\begin{aligned} -\log P(D|M) = \sum_{i=1}^{I} \Big[& N_i \log(N_i) - N_i - N_i^u \log(N_i^u) + N_i^u \\ & - \sum_{j=1}^{J} \left[N_{ij} \log(N_{ij}) - N_{ij} \right] \\ & + \sum_{j=1}^{J} \left[N_{ij}^u \log(N_{ij}^u) - N_{ij}^u \right] \\ & + \mathcal{O}(\log N_i) - \mathcal{O}(\log N_i^u) \\ & - \sum_{j=1}^{J} \mathcal{O}(\log N_{ij}) + \sum_{j=1}^{J} \mathcal{O}(\log N_{ij}^u) \Big] \end{aligned}$$

Exploiting the fact that :

$$\sum_{j=1}^{J} N_{ij} = N_i$$

and

$$\sum_{j=1}^{J} N_{ij}^u = N_i^u,$$

we obtain:

$$\begin{aligned} -\log P(D|M) = \sum_{i=1}^{I} \Big[& \sum_{j=1}^{J} N_{ij}^u \left(\log N_{ij}^u - \log N_i^u \right) \\ & - N_{ij} \left(\log N_{ij} - \log N_i \right) + \mathcal{O}(\log N_i) \Big] \end{aligned}$$

$$\begin{aligned} -\log P(D|M) = \sum_{i=1}^{I} \Big[& - N_i \sum_{j=1}^{J} \frac{N_{ij}}{N_i} \log \left(\frac{N_{ij}}{N_i} \right) \\ & + N_i^u \sum_{j=1}^{J} \frac{N_{ij}^u}{N_i^u} \log \left(\frac{N_{ij}^u}{N_i^u} \right) + \mathcal{O}(\log N_i) \Big] \end{aligned}$$

The entropy is additive on disjoint sets. We get:

$$-\log P(D|M) = NH_M(D) - N^u H_M(U) + \mathcal{O}(\log N)$$

Lemma B: *The values of parameters $\{N_{ij}\}$ which minimize the criterion $\mathcal{C}_{semi\ sup}$ (denoted $\{N_{ij}^{\diamond}\}$) correspond to the proportion of labels observed in each interval* :*

$$N_{ij}^{\diamond} = \left\lceil (N_i + 1) \times \frac{N_{ij}^l}{N_i^l} - 1 \right\rceil$$

* If $\sum_{j=1}^{J} N_{ij}^{\diamond} = N_i - 1$, simply choose one of the $N_{ij}^{\diamond}$ and add 1. All possibilities are equivalent and optimal for $\mathcal{C}_{semi\ sup}$

Proof :
This proof handles the case of a single interval model. Since data distribution is assumed to be independent between the intervals, this proof can be independently repeated on I intervals. We consider a binary classification problem. Let the function $f(N_{i1}, N_{i2})$ denote the criterion $\mathcal{C}_{semi\ sup}$, with all parameters fixed except N_{i1} and N_{i2}. We aim to find an analytical expression of the minimum of the function $f(N_{i1}, N_{i2})$:

$$f(N_{i1}, N_{i2}) = \log\left(\frac{(N_{i1} - N_{i1}^l)!}{N_{i1}!}\right) + \log\left(\frac{(N_{i2} - N_{i2}^l)!}{N_{i2}!}\right)$$

The terms N_{i1}^l and N_{i2}^l are constant, and $N_{i2} = N_i - N_{i1}$. f can be rewritten as a single parameter function:

$$\begin{aligned} f(N_{i1}) &= \log\left(\frac{(N_{i1} - N_{i1}^l)!}{N_{i1}!}\right) + \log\left(\frac{(N_i - N_{i1} - N_{i2}^l)!}{(N_i - N_{i1})!}\right) \\ &= \sum_{k=1}^{N_{i1}-N_{i1}^l} \log k - \sum_{k=1}^{N_{i1}} \log k + \sum_{k=1}^{N_i-N_{i1}-N_{i2}^l} \log k - \sum_{k=1}^{N_i-N_{i1}} \log k \\ &= -\sum_{k=N_{i1}-N_{i1}^l+1}^{N_{i1}} \log k - \sum_{k=N_i-N_{i1}-N_{i2}^l+1}^{N_i-N_{i1}} \log k \end{aligned}$$

And:

$$f(N_{i1}+1) = -\sum_{k=N_{i1}-N_{i1}^l+2}^{N_{i1}+1} \log k - \sum_{k=N_i-N_{i1}-N_{i2}^l}^{N_i-N_{i1}-1} \log k$$

Consequently:

$$\begin{aligned} f(N_{i1}) - f(N_{i1}+1) &= \log(N_{i1}+1) - \log(N_{i1}+1-N_{i1}^l) \\ &\quad - \log(N_i - N_{i1}) + \log(N_i - N_{i2}^l - N_{i1}) \\ &= \log\left(\frac{(N_{i1}+1)(N_i - N_{i2}^l - N_{i1})}{(N_{i1}+1-N_{i1}^l)(N_i - N_{i1})}\right) \end{aligned}$$

$f(N_{i1})$ decreases if:

$$\begin{aligned} & f(N_{i1}) - f(N_{i1}+1) > 0 \\ \Leftrightarrow \quad & \frac{(N_{i1}+1)(N_i - N_{i2}^l - N_{i1})}{(N_{i1}+1-N_{i1}^l)(N_i - N_{i1})} > 1 \\ \Leftrightarrow \quad & -N_{i2}^l \times N_{i1} - N_{i2}^l > -N_{i1}^l \times N_i + N_{i1}^l \times N_{i1} \\ \Leftrightarrow \quad & N_{i1} < \frac{-N_{i2}^l + N_{i1}^l \times N_i}{N_{i1}^l + N_{i2}^l} \end{aligned}$$

In the same way, $f(N_{i1})$ increases if:

$$f(N_{i1}) - f(N_{i1}+1) < 0 \Leftrightarrow N_{i1} > \frac{-N_{i2}^l + N_{i1}^l \times N_i}{N_{i1}^l + N_{i2}^l}$$

As $f(N_{i1})$ is a discrete function, its maximum is reached for $N_{i1} = \lceil \frac{-N_{i2}^l + N_{i1}^l \times N_i}{N_{i1}^l + N_{i2}^l} \rceil$. This expression can be generalized to the case of J classes[1]:

$$N_{ij}^{\diamond} = \left\lceil (N_i + 1) \times \frac{N_{ij}^l}{N_i^l} - 1 \right\rceil$$

Theorem 5.1 *Given the best model M_{map},* **Lemma B** *states that the proportion of the labels are the same in the sets L and D. Thus, L and D have the same entropy. The set U also has the same entropy because $U = D \setminus L$. Exploiting* ***lemma A****, we have for the semi-supervised case:*

$$\begin{aligned} -\log P(D|M_{map}) &= N H_{M_{map}}(D) - N^u H_{M_{map}}(U) \\ &\quad + \mathcal{O}(\log N) \\ -\log P(D|M_{map}) &= (N - N^u) H_{M_{map}}(L) + \mathcal{O}(\log N) \\ -\log P(D|M_{map}) &= N^l H_{M_{map}}(L) + \mathcal{O}(\log N) \end{aligned}$$

We have:

$$-\log P(D|M_{map}) + \log P(D|M_{map})^* = \mathcal{O}(\log N)$$

$$\lim_{N \to +\infty} \frac{-\log P(D|M_{map}) + \log P(D|M_{map})^*}{-\log P(D|M_{map})} = 0$$

With $P(D|M_{map})$ [respectively $P(D|M_{map})^*$] corresponding to the semi-supervised [respectively supervised] approach.

The conditional likelihood $P(D|M_{map})$ is asymptotically the same in the supervised and the semi-supervised cases. Both approaches aim to solve the same optimization problem. Owing to this result, the semi-supervised approach can be reformulated a posteriori. Our approach is equivalent to [3] improved with a post-optimization on the bounds location.

[1] The generalized expression of $N_{ij}^{\diamond}$ has been empirically verified on multi-class data sets.

6 Experiments

A toy problem is exploited to evaluate the behavior of the post-optimized method (see Section 5.2). This problem consists in estimating a step function from data. The artificial dataset is constituted by examples which belong to the class "1" on left-hand part of Figure 4 and to the class "2" on the right-hand part. The objective of this experiment is to find the step location with less labelled examples as possible.

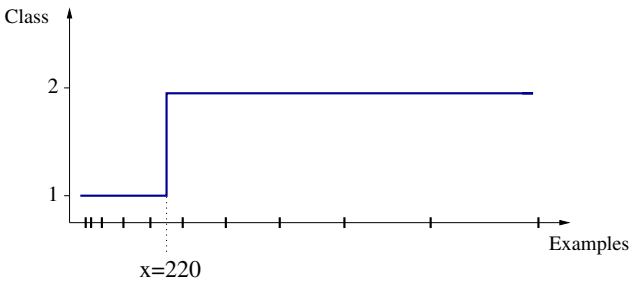

Figure 4. Step dataset

The values of the variable (horizontal axis) which characterizes the 100 examples x_i are drawn according to the expression $x_i = e^{\alpha}$, α varying between 0 and 10 with a step of 0.1. The step location is placed at $x = 220$: 47 examples belong to the class 1 and 53 to the class 2. The train set and the test set are both constituted by 100 examples.

We compare two discretization methods: the supervised method (see Section 2) and the supervised method with a post-optimisation of bounds location (see Section 5.2). For both methods, the $\mathcal{M}_{map}$ is exploited to discretize the input variable. Then this variable is placed on the input of a naive Bayes classifier. The predictive model is evaluated using the area under the ROC curve (AUC) [9]. The number of labelled examples is the only free parameter and allows comparisons between both methods, examples to be labelled are drawn randomly. The experiment of this section is realized considering discretization models with one or two intervals, that is consistent with theoretical proofs demonstrated above.

Figure 5 plots the average AUC versus the number of labelled examples. For each value of the number of labelled examples the experiment has been realized 10 times. Points reprensent the mean AUC and natches the variance of the results. Considering less than 6 labelled examples, both discretisation methods give a M_{map} with a single interval. In this cas, the AUC is equal to 0.5. From 6 labelled examples, the M_{map} includes two intervals for both discretization methods. The bound between these two intervals becomes better and better when examples are labelled. The Figure 5 shows that the post-optimization of the bound location improve the supervised discretization method. This improvement is weak but always present : (i) the post-optimization always improves the discretization; ii) this improvement is more important as the number of labelled examples is low.

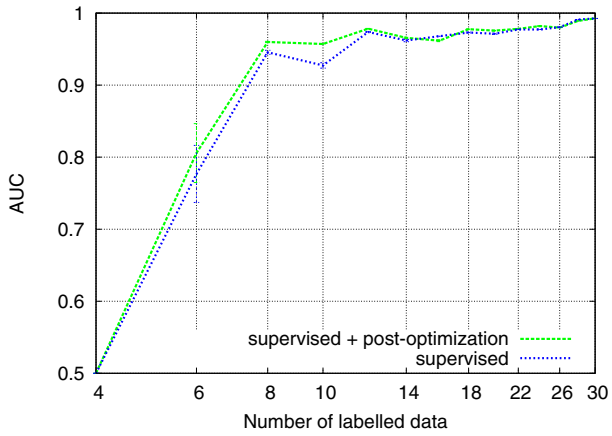

Figure 5. Comparison between the post-optimised method and the supervised method.

Results on this artificial dataset, which is often used to test discretization method [4], are very promising. This section shows that our post-optimization of bounds location improve the optimal discretization model, when the distribution of examples is not uniform. The method described in this paper will be tested on step function [5] with or whithout noise as in [4] [13]. We will show its robustness compare to other methods which discretize a single dimension into two intervals.

7 Conclusion

This article presents a new semi-supervised discretization method based on very few assumptions on the data distribution. It provides an in-depth analysis of the problem which consists in dealing with a set of labelled and unlabelled examples.

This paper significantly extends the previous research of Boulle in [3] on supervised discretization method MODL - i.e. it presents a semi-supervised generalization of it where additional unlabelled learning examples are taken into account. The results have been proved in an intuitive manner, and mathematical proofs have also been given.

Our approach gives an important result: the intervals bounds must be placed in the middle of unlabelled areas to minimize the mean square error. The main contribution of this article is to demonstrate that unlabelled examples provide useful information, even with a minimum of assumptions on the data distribution. We also proposed a post-optimization which allows the supervised MODL approach to be equivalent to our semi-supervised discretization method. This post-optimization makes an intuitive bridge between both approaches, and can be exploited to efficiently implement the semi-supervised discretization method.

In practice, the use of [3] to carry out a semi-supervised discretization offers advantages. First, the supervised approach is faster than the semi-supervised one, due to the less important number of possible bounds' locations which are considered. Second, the supervised approach gives best M_{map} with most intervals, due to the less important modeling cost of the prior distribution. We plan to incorporate this semi-supervised preprocessing step in datamining algorithms, such as decision trees or naive Bayes classifier.

An efficient search algorithm which optimizes the evaluation criterion to find the optimal discretization model is necessary to exploit our semi-supervised approach on real data set. The optimization algorithm used in [3] performs a post-optimization on the result of a standard greedy bottom-up heuristic which is based on hill-climbing search in the neighborhood of a discretization. The time complexity of this algorithm is $O(JN \log N)$. Our "semi-supervised" and "post-optimized supervised" approaches will be implemented using the same efficient algorithm. Empirical results support the conclusions though both approaches have to be compared more in depth on large number of real world data sets in future work.

Acknowledgement : Thank to Oliver Bernier for his wise advices on this article.

References

[1] J. Berger. The case of objective bayesian analysis. *Bayesian Analysis*, 1(3):385–402, 2006.

[2] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *COLT' 98: Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100, New York, NY, USA, 1998. ACM.

[3] M. Boulle. MODL: A bayes optimal discretization method for continuous attributes. *Machine Learning*, 65(1):131–165, 2006.

[4] M. V. Burnashev and K. S. Zigangirov. An interval estimation problem for controlled observations. In *Problems in Information Transmission*, volume 10, pages 223–231, 1974.

[5] R. Castro and R. Nowak. *Foundations and Application of Sensor Management*, chapter Active Learning and Sampling. Springer-Verlag, 2008.

[6] J. Catlett. On changing continuous attributes into ordered discrete attributes. In *EWSL-91: Proceedings of the European working session on learning on Machine learning*, pages 164–178. Springer-Verlag New York, Inc., 1991.

[7] O. Chapelle, B. Schölkopf, and A. Zien. *Semi-Supervised Learning*. The MIT Press, 2007.

[8] J. Dougherty, R. Kohavi, and M. Sahami. Supervised and Unsupervised Discretization of Continuous Features. In *International Conference on Machine Learning*, pages 194–202, 1995.

[9] T. Fawcett. Roc graphs: Notes and practical considerations for data mining researchers. Technical Report HPL-2003-4, HP Labs, 2003., 2003.

[10] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery: An overview. In *Advances in Knowledge Discovery and Data Mining*, pages 1–34. 1996.

[11] A. Fujino, N. Ueda, and K. Saito. A hybrid generative/discriminative approach to text classification with additional information. *Inf. Process. Manage.*, 43:379–392, 2007.

[12] R. Holte. Very Simple Classification Rules Perform Well on Most Commonly Used Datasets. *Machine Learning*, 11:63–91, 1993.

[13] M. Horstein. Sequential decoding using noiseless feedback. In *IEEE Transmition Information Theory*, volume 9, pages 136–143, 1963.

[14] R. Kohavi and M. Sahami. Error-Based and Entropy-Based Discretization of Continuous Features. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pages 114–119, 1996.

[15] H. Liu, F. Hussain, C. Tan, and M. Dash. Discretization: An Enabling Technique. *Data Mining Knowledge Discovery*, 6(4):393–423, 2002.

[16] B. Maeireizo, D. Litman, and R. Hwa. Analyzing the effectiveness and applicability of co-training. In *ACL '04: The Companion Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, 2004.

[17] D. Pyle. *Data preparation for data mining*. Morgan Kaufmann Publishers, Inc. San Francisco, USA, 19, 1999.

[18] C. Rosenberg, M. Hebert, and H. Schneiderman. Semi-Supervised Self-Training of Object Detection Models. In *Seventh IEEE Workshop on Applications of Computer Vision*, January 2005.

[19] C. Shannon. A Mathematical Theory of Communication. *Key Papers in the Development of Information Theory* , 1948.

[20] M. Sugiyama, M. Krauledat, and K. Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 2007.

[21] M. Sugiyama and K. Müller. Model selection under covariate shift. In *ICANN, International Conference on Computational on Artificial Neural Networks: Formal Models and Their Applications*, 2005.