Knowl Inf Syst
DOI 10.1007/s10115-009-0230-2

REGULAR PAPER

A non-parametric semi-supervised discretization method

Alexis Bondu · Marc Boullé · Vincent Lemaire

Received: 18 March 2009 / Revised: 3 June 2009 / Accepted: 20 June 2009

Abstract Semi-supervised classification methods aim to exploit labeled and unlabeled 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.

Keywords Bayesian · Semi-supervised · Discretization

1 Introduction

Data mining can be defined as the non-trivial process of identifying valid, novel, potentially useful, ultimately understandable patterns in data [10,26]. 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 [19]. An interesting method for data preparation is to discretize the input variables [13].

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 [5,7,12,14,16]. A naïve Bayes classifier [15] can exploit a discretization of its input space

A. Bondu (✉)
EDF R&D (ICAME/SOAD), 1 av. Général de Gaulle, 92140 Clamart, France
e-mail: alexis.bondu@edf.fr

M. Boulle · V. Lemaire
ORANGE LABS (TECH/EASY/TSI), 2 av. Pierre Marzin, 22300 Lannion, France
e-mail: vincent.lemaire@orange-ftgroup.com

Published online: 01 August 2009

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 unlabeled data to improve a predictive model [27]. This article focuses on semi-supervised classification, a well-known problem in the literature. Most of the semi-supervised approaches deal with particular cases where information about unlabeled data are 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 [4] *("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 Sect. 4; in Sect. 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 location of the interval boundaries. Section 6 presents an experimental evaluation comparing the supervised method and the supervised method improved with a post-optimization of the location of the interval boundaries. Finally, future work is discussed in Sect. 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 [6] exploit labeled and unlabeled 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 labeled examples in turns are used to train the model. The uncertainty of predictions is evaluated in order to label only the most confident examples [21].
- The **Co-training** approach involves two predictive models, which are independently trained on disjoint sub-feature 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 [2,17].
- The **Covariate shift** approach estimates the distributions of labeled and unlabeled examples [25]. The covariate shift formulation [24] weights labeled 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 are exploited in order to fit parameters values [11].

Semi-supervised learning without making hypothesis on data distribution is a great challenge. Therefore, most of the 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 belongs 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 unlabeled examples and L the set of labeled 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 [4] 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 [4], 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 labeled examples, J the number of classes, I the number of intervals for the input domain. N_i^l denotes the number of labeled examples in the interval i, and N_{ij}^l the number of labeled examples of output value j in the interval i. A discretization model is then defined by the parameter set $\left\{ I, \left\{ N_i^l \right\}_{1 \le i \le I}, \left\{ N_{ij}^l \right\}_{1 \le i \le I, 1 \le j \le J} \right\}$.

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 [4].

The MODL approach was extensively compared with other supervised discretization methods in [4], including entropy-based approaches. This previous comparative study showed that the entropy-based approaches tend to overfit and require to be regularized owing to additional parameters, such as the maximum number of intervals, or the minimum number of instances in each interval. The MDLPC [9] method exploits a variation of the entropy which is regularized by a *"Minimum Description Length"* [20] approach. The MDLPC is relatively close to MODL, however, significant differences remain between both approaches. On the one hand, the MDLPC method recursively applies a dichotomic partitioning in intervals. On the other hand, the MODL approach directly optimizes a K-partitioning criterion which leads to a better discretization model.

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 [4] is 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 unlabeled and labeled 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, and $\{N_{ij}\}$ 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:

$$P(M) = P(I, \{N_i\}, \{N_{ij}\}) \tag{3}$$

$$P(M) = P(I) \times P(\{N_i\}|I) \times P(\{N_{ij}\}|\{N_i\}, I) \tag{4}$$

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

$$P(I) = \frac{1}{N} \tag{5}$$

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:

$$P(\{N_i\}|I) = \frac{1}{\binom{N+I-1}{I-1}} \tag{6}$$

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:

$$P(\{N_{ij}\}|\{N_i\}, I) = \prod_{i=1}^{I} \frac{1}{\binom{J-1}{N_i+J-1}} \tag{7}$$

Finally, the prior distribution of the model is similar to the supervised approach [4]. The only one difference is that the semi-supervised prior takes into account all examples, including unlabeled ones:

$$P(M) = \frac{1}{N} \times \frac{1}{\binom{N+I-1}{I-1}} \times \prod_{i=1}^{I} \frac{1}{\binom{J-1}{N_i+J-1}} \tag{8}$$

3.2 Likelihood

This section focuses on the conditional likelihood $P(D|M)$ of the data given the model. First, the family Λ of labeling models has to be defined. Semi-supervised discretization handles labeled and unlabeled pieces of data, Λ represents all possible labelings. 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 labeled examples, $\{N_i^l\}$ the number of labeled examples in the interval i, and $\{N_{ij}^l\}$ the number of labeled examples of the class j in the interval i.

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

$$P(D|M) = \sum_{\lambda \in \Lambda} P(\lambda|M) \times P(D|M, \lambda) \tag{9}$$

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

$$P(D|M) = P(\lambda^*|M) \times P(D|M, \lambda^*) \tag{10}$$

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 labeled examples are uniformly drawn:

$$P(\lambda^*|M) = \prod_{i=1}^{I} \frac{\prod_{j=1}^{J} \binom{N_{ij}}{N_{ij}^l}}{\binom{N_i}{N_i^l}} \tag{11}$$

Let us consider a very simple and intuitive problem to explain Eq. 11. 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 Eq. (10) 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:

$$P(D|M, L^*) = \prod_{i=1}^{I} \frac{1}{\frac{N_i^l!}{N_{i1}^l! N_{i2}^l! \ldots N_{iJ}^l!}} = \prod_{i=1}^{I} \frac{\prod_{j=1}^{J} N_{ij}^l!}{N_i^l!} \tag{12}$$

Finally, the likelihood of the model is:

$$P(D|M) = \prod_{i=1}^{I} \frac{\prod_{j=1}^{J} \binom{N_{ij}}{N_{ij}^l} \times N_{ij}^l!}{\binom{N_i}{N_i^l} \times N_i^l!} \tag{13}$$

In every interval, the number of unlabeled 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:

$$P(D|M) = \prod_{i=1}^{I} \frac{\prod_{j=1}^{J} \frac{N_{ij}!}{N_{ij}^u!}}{\frac{N_i!}{N_i^u!}} \tag{14}$$

$$P(D|M) = \prod_{i=1}^{I} \left[\frac{\prod_{j=1}^{J} N_{ij}!}{N_i!} \times \frac{N_i^u!}{\prod_{j=1}^{J} N_{ij}^u!} \right] \tag{15}$$

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 Eqs. (8) and (15). The maximum a posteriori model, denoted as "M_{map}" is defined by:

$$M_{map} = \max_{M \in \mathbb{M}} \left[\frac{1}{N} \times \frac{1}{\binom{N+I-1}{I-1}} \times \prod_{i=1}^{I} \frac{1}{\binom{N_i+J-1}{J-1}} \times \prod_{i=1}^{I} \left[\frac{\prod_{j=1}^{J} N_{ij}!}{N_i!} \times \frac{N_i^u!}{\prod_{j=1}^{J} N_{ij}^u!} \right] \right] \tag{16}$$

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

$$\begin{aligned} M_{map} &= \min_{M\in\mathbb{M}} \mathcal{C}_{semi\ sup}(M) \\ &= \min_{M\in\mathbb{M}} \Bigg[\log(N) + \log\binom{N+I-1}{I-1} \\ &\quad + \sum_{i=1}^{I} \log\binom{N_i+J-1}{J-1} + \sum_{i=1}^{I} \log\left(\frac{N_i!}{\sum_{j=1}^{J} N_{ij}!}\right) \\ &\quad - \sum_{i=1}^{I} \log\left(\frac{N_i^u!}{\sum_{j=1}^{J} N_{ij}^u!}\right)\Bigg] \end{aligned} \tag{17}$$

4 Comparison: semi-supervised versus supervised criteria

In this section, the semi-supervised criterion $\mathcal{C}_{semi\ sup}$ of Eq. (17) is compared with the supervised criterion $\mathcal{C}_{sup}$ of Eq. (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 labeled and unlabeled 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 Labeled examples only

In this case, all training examples are supposed to be labeled: $D = L$ 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 Eq. (17) is equal to zero. The criterion $\mathcal{C}_{semi\ sup}$ can be rewritten as follows:

$$\log(N) + \log\binom{N+I-1}{I-1} + \sum_{i=1}^{I} \log\binom{N_i+J-1}{J-1} + \sum_{i=1}^{I} \log\left(\frac{N_i!}{\sum_{j=1}^{J} N_{ij}!}\right) \tag{18}$$

When all the training examples are labeled, $N = N^l$, $N_i = N_i^l$ 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 Unlabeled examples only

In the case, where no example is labeled 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 (Eq. 15) can be rearranged as follows:

$$P(D|M) = \prod_{i=1}^{I} \left[\frac{\prod_{j=1}^{J} N_{ij}!}{N_i!} \times \frac{N_i!}{\prod_{j=1}^{J} N_{ij}!}\right] \tag{19}$$

$$P(D|M) = 1 \tag{20}$$

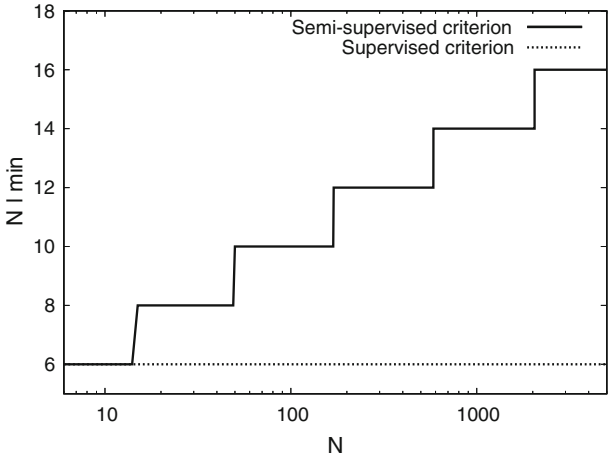

Fig. 1 Mixture of labeled and unlabeled examples. The *vertical axis* represents the minimal number of labeled examples necessary to obtain a model with two intervals, rather than a model with a single interval. The *horizontal axis* represents the total number of examples using a logarithmic scale

The 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:

$$\log(N) + \log\binom{N+I-1}{I-1} + \sum_{i=1}^{I} \log\binom{N_i+J-1}{J-1} \tag{21}$$

4.3 Mixture of labeled and unlabeled 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. Unlabeled 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 unlabeled examples increases, the criterion $\mathcal{C}_{semi\ sup}$ prefers models with less intervals.

This behavior 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 labeled examples is always the same in both classes. For every value of N, we evaluate N^l_{min} the minimal number of labeled examples which induces a M_{map} with two intervals (and not a single interval).

Figure 1 plots N^l_{min} against $N = N^l + N^u$ for both criteria. For the criterion $\mathcal{C}_{sup}$, the minimal number of labeled examples necessary to split data does not depend on N. In this case, $N^l_{min} = 6$ for every value of N. A different behavior is observed for $\mathcal{C}_{semi\ sup}$. Figure 1 quantifies the influence of N on the selection of the model M_{map}. When the number of examples N grows, N^l_{min} 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 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 results are demonstrated: an interpretation of the likelihood in terms of entropy, and an analytical expression of the optimal N_{ij}. Taking into account of 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 boundary 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 the continuous boundary location. Indeed, there is an infinity of possible locations between two training examples with input values x_1 and x_2. In [4], the location of the boundary b is chosen as $b = (x_1 + x_2)/2$.

Let us analyze what is the theoretical foundation for this choice. Let us consider two adjacent intervals, with x_1, the last train input value of the first interval and x_2, the first train input value of the second interval. Let us assume that the true conditional distributions are $\{p_{1,j}\}_{1\leq j\leq J}$ in the the first interval and $\{p_{2,j}\}_{1\leq j\leq J}$ in the second interval. Let $B, x_1 \leq B \leq x_2$ be the true boundary location between the two intervals, and $b, x_1 \leq b \leq x_2$, the choice our boundary location, which is the only unknown parameter given our hypotheses. We then predict the conditional probabilities $\{p_{1,j}\}$ below b and $\{p_{2,j}\}$ above b. Let L be a loss function between the predicted and the true conditional probabilities. For example, L is

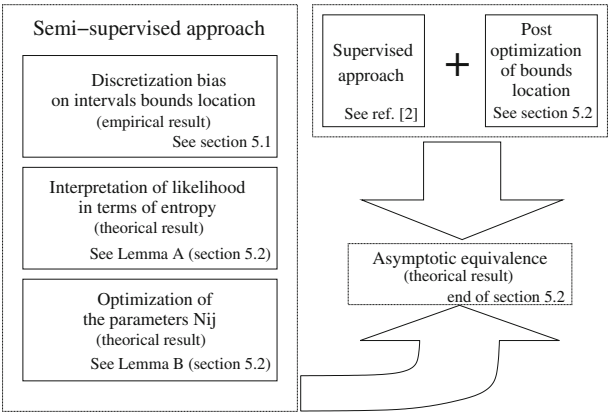

Fig. 2 Structure of Sect. 5

the quadratic loss function. According to our assumptions, $L = 0$ for $x \in [x_1, \min(b, B)] \cup [\max(b, B), x_2]$ and $L = L_{Max} = L(\{p_{1,j}\}, \{p_{2,j}\})$ for $x \in [\min(b, B), \max(b, B)]$. The expectation $E_Y(L)$ of the loss function w.r.t Y is then constant and non null only on the sub-interval $[\min(b, B), \max(b, B)]$. The expected loss w.r.t X and Y is equal to

$$E_{XY}(L) = E_Y(L) \int_{x=\min(b,B)}^{\max(b,B)} p(X = x)\mathrm{d}x.$$

If we assume that X is uniformly distributed on $[x_1, x_2]$, we get

$$E_{XY}(L) \sim E_Y(L)|B - b|.$$

Assuming that the true boundary B is uniformly distributed on $[x_1, x_2]$, the expected loss is

$$\int_{x=x_1}^{x_2} E_{XY}(L)p(B = x)\mathrm{d}x \sim \frac{(b - x_1)^2 + (x_2 - b)^2}{2} E_Y(L).$$

Therefore, the expected loss is minimized for $b = (x_1 + x_2)/2$.

To summarize, the choice of the mean value between the two input values for the location of the interval boundary is optimal if we assume that the input data and the location of the true interval boundary are uniformly distributed in the range of the possible interval boundaries.

5.1.2 How to position a boundary in an unlabeled area?

The optimization of the semi-supervised criterion $\mathcal{C}_{semi\ sup}$ does not indicate the best boundary 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 labeled, all possible boundary locations have the same cost according to the criterion $\mathcal{C}_{semi\ sup}$. Therefore, the semi-supervised approach is not able to determine boundary location in such an unlabeled area. We adopt the same approach as [4] to define the boundary location, and use the unlabeled instances to better exploit the assumption of uniform distribution of the input values on the possible interval boundaries. Indeed, if we replace the input values by their empirical ranks, we are more likely to follow the uniform assumption (ranks are always uniformly distributed). We thus have to choose the median value (instead of the mean value) for the best boundary location, and we estimate this median value using the empirical distribution of the input ranks provided by the unlabeled instances.

Finally, the supervised and the semi-supervised approaches are not able to position a continuous boundary between the two labeled examples. In both cases, the same prior on the best boundary location is adopted. The only one interest of the unlabeled examples is to bring information about the input values, in order 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 labeled 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 unlabeled examples.

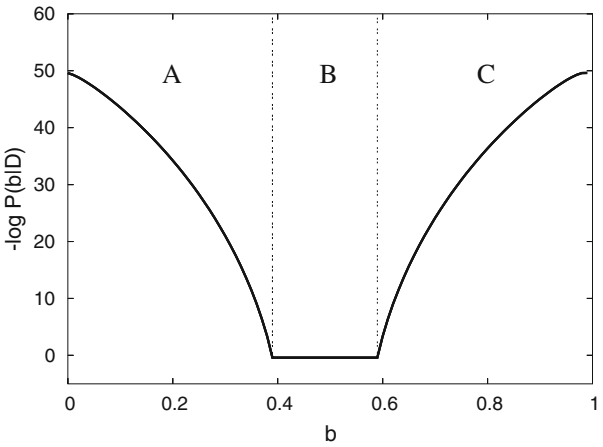

Fig. 3 Bound's quantity of information versus bound's location

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:

$$P(b|D) = \sum_{\{N_{ij}\}} P(\underbrace{b, \{N_{ij}\}}_{M \in \mathbb{M}} |D) \tag{22}$$

Using the Bayes rule, we get:

$$P(b|D) \times P(D) = \sum_{\{N_{ij}\}} P(D|b, \{N_{ij}\}) \times P(b, \{N_{ij}\}) \tag{23}$$

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 $\mathcal{C}_{semi\ sup}$ cannot distinguish between bounds' location in an unlabeled 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].

5.2 A post-optimization 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 unlabeled examples in order to position the intervals bounds in the middle of unlabeled areas.

5.2.1 Equivalent prior distribution

The discretization bias established in Sect. 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 unlabeled 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 Sect. 5.1, only N^l-1 possible locations are considered. In these conditions, the prior distribution $P(M)$ (see Eq. 8) can be easily rewritten as in the supervised approach (see Eq. 2).

5.2.2 Asymptotically equivalent likelihood

Lemma 1 *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* $-\log P(D|M) = N H_M(D) - N^u H_M(U) + \mathcal{O}(\log N)$

Proof • Let us denote $H_M(D)$ the Shannon's entropy [23] 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:

$$-\log P(D|M) = \sum_{i=1}^{I} \left[\log(N_i!) - \log(N_i^u!) - \sum_{j=1}^{J} \log(N_{ij}!) + \sum_{j=1}^{J} \log(N_{ij}^u!) \right] \tag{24}$$

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

$$\begin{aligned} -\log P(D|M) = \sum_{i=1}^{I} \Bigg[& N_i \log(N_i) - N_i - N_i^u \log(N_i^u) + N_i^u \\ & - \sum_{j=1}^{J} \left[N_{ij} \log(N_{ij}) - N_{ij} \right] + \sum_{j=1}^{J} \left[N_{ij}^u \log(N_{ij}^u) - N_{ij}^u \right] \\ & + \mathcal{O}(\log N_i) - \mathcal{O}(\log N_i^u) - \sum_{j=1}^{J} \mathcal{O}(\log N_{ij}) + \sum_{j=1}^{J} \mathcal{O}(\log N_{ij}^u) \Bigg] \end{aligned} \tag{25}$$

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:

$$\begin{aligned} -\log P(D|M) &= \sum_{i=1}^{I} \left[\sum_{j=1}^{J} N_{ij}^u \left(\log N_{ij}^u - \log N_i^u \right) - N_{ij} \left(\log N_{ij} - \log N_i \right) + \mathcal{O}(\log N_i) \right] \\ &= \sum_{i=1}^{I} \left[-N_i \sum_{j=1}^{J} \frac{N_{ij}}{N_i} \log \left(\frac{N_{ij}}{N_i} \right) + N_i^u \sum_{j=1}^{J} \frac{N_{ij}^u}{N_i^u} \log \left(\frac{N_{ij}^u}{N_i^u} \right) + \mathcal{O}(\log N_i) \right] \end{aligned} \tag{26}$$

The entropy is additive on disjoint sets. We get:

$$-\log P(D|M) = N H_M(D) - N^u H_M(U) + \mathcal{O}(\log N) \tag{27}$$

□

Lemma 2 *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*:*

$$N_{ij}^{\diamond} = \left\lceil (N_i + 1) \times \frac{N_{ij}^l}{N_i^l} - 1 \right\rceil \tag{28}$$

* *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 $f(N_{i1}, N_{i2})$:

$$f(N_{i1}, N_{i2}) = \log \left(\frac{(N_{i1} - N_{i1}^l)!}{N_{i1}!} \right) + \log \left(\frac{(N_{i2} - N_{i2}^l)!}{N_{i2}!} \right) \tag{29}$$

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:

$$\begin{aligned} f(N_{i1}) &= \log \left(\frac{(N_{i1} - N_{i1}^l)!}{N_{i1}!} \right) + \log \left(\frac{(N_i - N_{i1} - N_{i2}^l)!}{(N_i - N_{i1})!} \right) \\ &= \sum_{k=1}^{N_{i1} - N_{i1}^l} \log k - \sum_{k=1}^{N_{i1}} \log k + \sum_{k=1}^{N_i - N_{i1} - N_{i2}^l} \log k - \sum_{k=1}^{N_i - N_{i1}} \log k \\ &= - \sum_{k=N_{i1} - N_{i1}^l + 1}^{N_{i1}} \log k - \sum_{k=N_i - N_{i1} - N_{i2}^l + 1}^{N_i - N_{i1}} \log k \end{aligned} \tag{30}$$

And

$$f(N_{i1} + 1) = - \sum_{k=N_{i1} - N_{i1}^l + 2}^{N_{i1} + 1} \log k - \sum_{k=N_i - N_{i1} - N_{i2}^l}^{N_i - N_{i1} - 1} \log k \tag{31}$$

Consequently,

$$\begin{aligned} f(N_{i1}) - f(N_{i1}+1) &= \log(N_{i1}+1) - \log(N_{i1}+1-N^l_{i1}) - \log(N_i - N_{i1}) \\ &\quad + \log(N_i - N^l_{i2} - N_{i1}) \\ &= \log\left(\frac{(N_{i1}+1)(N_i - N^l_{i2} - N_{i1})}{(N_{i1}+1-N^l_{i1})(N_i - N_{i1})}\right) \end{aligned} \tag{32}$$

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

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

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

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

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

$$N^{\diamond}_{ij} = \left\lceil (N_i + 1) \times \frac{N^l_{ij}}{N^l_i} - 1 \right\rceil \tag{33}$$

□

Theorem 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:

$$-\log P(D|M_{map}) = N H_{M_{map}}(D) - N^u H_{M_{map}}(U) \tag{34}$$

$$+ \mathcal{O}(\log N) \tag{35}$$

$$-\log P(D|M_{map}) = (N - N^u) H_{M_{map}}(L) + \mathcal{O}(\log N) \tag{36}$$

$$-\log P(D|M_{map}) = N^l H_{M_{map}}(L) + \mathcal{O}(\log N) \tag{37}$$

We have:

$$-\log P(D|M_{map}) + \log P(D|M_{map})^* = \mathcal{O}(\log N) \tag{38}$$

$$\lim_{N \to +\infty} \frac{-\log P(D|M_{map}) + \log P(D|M_{map})^*}{-\log P(D|M_{map})} = 0 \tag{39}$$

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

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

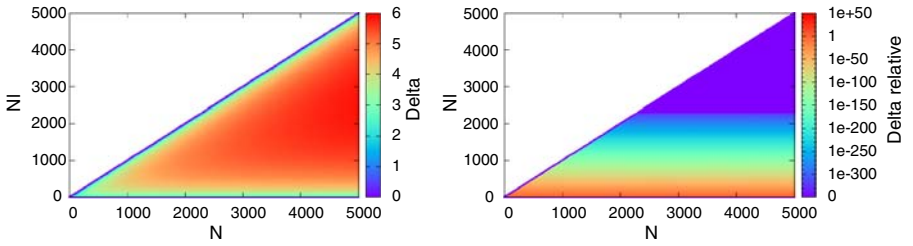

Fig. 4 Δ and $\Delta_{\mathcal{R}elative}$ versus N and N^l

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 [4] improved with a post-optimization on the bounds location.

Illustration: Supervised and semi-supervised evaluation criteria are respectively, defined in Eqs. (2) and (17). The objective of this section is to characterize the gap between the likelihood terms of both criteria, under varying the size of the data set. Let Δ be the difference of the "$-\log$" of likelihood terms :

$$\Delta = -\log P(D|M)_{semi\ super} + \log P(D|M)_{super} \tag{40}$$

Let $\Delta_{\mathcal{R}elative}$ be the relative difference:

$$\Delta_{\mathcal{R}elative} = \frac{-\log P(D|M)_{semi\ super} + \log P(D|M)_{super}}{-\log P(D|M)_{semi\ super}} \tag{41}$$

Considering a fixed discretization model M, Δ and $\Delta_{\mathcal{R}elative}$ depend only on the data set. This experiment considers the same data set D as the Sect. 4.3: a binary classification problem in which all examples belonging to the class "0" [*respectively* "1"] are located at $x = 0$ [*respectively* $x = 1$].

A range of values of the couple (N, N^l) is considered during the experiment, with $N \in [0, 5{,}000]$ and $N^l \in [0, N]$. N denotes the total number of examples and N^l denotes the number of labeled examples. The fixed discretization model M includes two intervals ($I = 2$), and the single bound, which is defined by $\{N_1, N_2\}$ is placed at $x = 0.5$. Δ and $\Delta_{\mathcal{R}elative}$ are evaluated for each value of the couple (N, N^l), given the above described D and M.

The left chart of Fig. 4 plots Δ under varying (N, N^l) using a color code. This chart shows the $\mathcal{O}(\log N)$ variation of Δ when N and N^l increase. For instance, Δ is approximately equals to 5 for $N = 5{,}000$ and $N^l = 2{,}500$. This observation is consistent with the entropy-based interpretation of **Lemma A** which gives $\Delta = \mathcal{O}(\log N)$.

When N tends to infinity, Δ is insubstantial compared to the likelihood $\Delta \ll -\log P(D|M)$. The right chart of Fig. 4 plots the relative difference between the "$-\log$" of likelihood terms, under varying N and N^l. This chart shows that $\Delta_{\mathcal{R}elative}$ tends to "0" when N and N^l has a constant ratio and jointly tend to infinity.

To conclude, the supervised approach improved with a post-optimization on the bounds location and our semi-supervised approach tend to resolve the same optimization problem, given the dicretization bias defined in Sect. 5.1.

5.3 Algorithm for criterion optimization

The supervised and semi-supervised discretization approaches involve the optimization of the associated evaluation criterion. First, this section presents a baseline algorithm able to find a sub-optimal solution in low time complexity. Second, an improvement of this algorithm based on the notion of neighborhood is presented.

Greedy heuristic: The bottom-up greedy heuristic is presented in Algorithm 1. This generic algorithm is used to optimize a univariate partition [28]. In our case, the purpose is to find the discretization model $M \in \mathbb{M}$ which minimizes $\mathcal{C}ost(M)$, i.e the value of the criterion $\mathcal{C}_{semi\ super}$.

The initial model $M_{initial}$ handles as many intervals as training examples. This heuristic consists in the evaluation of all the possible merges m between two adjacent intervals. The best merge is performed if the cost of the current model is reduced. This iterative algorithm is repeated as long as the model is improved.

Algorithm 1 Bottom-up greedy heuristic

Notations:

* The cost function $\mathcal{C}ost : \mathbb{M} \rightarrow \mathbb{R}$, corresponding to the value of the criterion
* $M_{initial}$, the initial disretization model, such as $I = N$ and $N_i = 1, \quad \forall i \in [1, I]$
* M', the optimal discretization model

/* Variables Initialization*/
$M' \leftarrow M_{initial}$
$improvement = true$

Repeat
  /* Look for the best improvement */
  $M_{actual} \leftarrow M'$
  For for all the merges m of two adjacent intervals **do**
    /* Evaluation of the merge m */
    $M_{merge} \leftarrow M' + m$
    If $\mathcal{C}ost(M_{merge}) < \mathcal{C}ost(M_{actual})$ **then**
      $M_{actual} \leftarrow M_{merge}$
    end If
  end For
  /* Test of the improvement */
  If $\mathcal{C}ost(M_{actual}) < \mathcal{C}ost(M')$ **then**
    $M' \leftarrow M_{actual}$ $improvement = true$
  else
    $improvement = false$
  end If
until $improvement = true$

The discretization model returned by this algorithm is a sub-optimal solution since the space of the models $\mathbb{M}$ is only partially scanned.

A naive implementation of this algorithm has a $\mathcal{O}(N^3)$ time complexity, where N is the size of the data set. But the additivity of the criterion (see Eq. 17) can be exploited to memorize intermediate results and reduce the impact of each merge to the two considered intervals. At the end, the greedy heuristic is implemented in $\mathcal{O}(N \log N)$.

Improvement of the bottom-up greedy heuristic: The greedy heuristic is followed by two post-optimization steps improving the quality of the model returned M'.

The first post-optimization step has an effect on the number of discretization intervals (I'). The Algorithm 1 is repeated until the model M' contains a single one interval : at each iteration the best merge is done even if this merge does not improve immediately the current model. The best seen discretization model is ultimately retained.

The second post-optimization step focuses on the intervals bounds $\{N_i''\}$ and considers adding or deleting an interval or moving interval boundaries. This second step exploits a local neighborhood of M', which is based on elementary operations between two adjacent intervals:

- deletion of an interval merging three adjacent intervals and splitting the merged interval;
- moving the boundary between two intervals : merge of two adjacent intervals followed by a split of the merged interval;
- addition of a new interval splitting an existing interval.

The additivity of the criterion $\mathcal{C}_{semi\ super}$ allows an exhaustive scan of the neighborhood of the model M' in $\mathcal{O}(N)$ time complexity. The systematic exploration of this neighborhood turns down many local optimum and improves the quality of the discretization model [4].

6 Evaluation on UCI data sets

This section presents an experimental evaluation comparing the supervised method, which exploits the labeled instances only (see Sect. 2) and the supervised method improved with a post-optimization of bounds location (see Sect. 5.2), which is a solution of the semi supervised problem and exploits the labeled and unlabeled instances. Let M_{super} [*respectively* M^*_{super}] be the best discretization model resulting from the supervised method [*respectively* from the supervised method improved with a post-optimization]. For both methods, the $\mathcal{M}_{map}$ is exploited to discretize the input variable. Then this variable is placed as input of a naive Bayes classifier (NBC) [15]. The predictive model is evaluated using the area under the ROC curve, denoted AUC [8] of the obtained Bayes classifier. The classification results using [M_{super}+NB] and [M^*_{super}+NB] are called below respectively M_{S+NB} and M^*_{S+NB}.

6.1 Experimental setup

Our benchmark involves 15 data sets[2], which come from the repository of University of California at Irvine [18]. Some properties of these data sets are given in Table 1.

Each data set is split into two subsets, respectively dedicated to the training of the naive Bayes classifier and its evaluation. A stratified two-fold cross validation is repeated 50 times to generate these two subsets. At each iteration, the twofolds permute and play both roles: train set and evaluation set. Overall, the experiments are repeated 100 times.

At the beginning of the experiments, all the training examples are considered as unlabeled. Then, labels are progressively unmasked. The new labeled examples are randomly chosen. The classifier is evaluated for several values of N^l : 4, 6, 8, 12, 16, 4, 32, 48, 64, 96, 128, 192, 256. For a given value of N^l, the discretization models M_{super} and M^*_{super} result from the same labeled examples: the difference is that the M^*_{super} model exploits the unlabeled training instances to post-optimize the location of the boundaries.

6.2 Results

This section evaluates the effects of our post-optimization of bounds location on the performance of supervised discretization method. First, theoretical expected results are established. Then, observed results are discussed.

[2] In our benchmark, categorical variables are handled owing to the grouping method MODL [3].

Table 1 UCI data sets

Name	Instances	Numerical variables	Categorical variables	Classes	Majority accuracy
Adult	48,842	7	8	2	76.1
Australian	690	6	8	2	55.5
Breast	699	10	0	2	65.5
Crx	690	6	9	2	55.5
German	1,000	24	0	2	70.0
Heart	270	10	3	2	55.6
Hepatitis	155	6	13	2	79.4
Hypothyroid	3,163	7	18	2	95.2
Ionosphere	351	34	0	2	64.1
Iris	150	4	0	3	33.3
Pima	768	8	0	2	65.1
SickEuthyroid	3,163	7	18	2	90.7
Vehicle	846	18	0	4	25.8
Waveform	5,000	21	0	3	33.9
Wine	178	13	0	3	39.9

Expected results: As shown in Sect. 5.2, the post-optimization of bounds location shifts each bound between two labeled examples. On the one hand, the expected improvement of the best discretization model is about $\frac{I_{map}}{N^l}$, where I_{map} denotes the number of intervals of the M_{map} and N^l denotes the number of labeled examples of the data set. On the other hand, the best model varies depending on the labeled examples in the same way as a binomial function. The statistical variance of M_{map} is about $1/\sqrt{(}N^l)$, that is superior to its expected improvement. Therefore, the improvement of the best discretization model will be difficult to highlight in practice.

Observed results: Figures 5 and 6 plot the average AUC of the naive Bayes classifier, which intervals are given either by M_{super} or by M^*_{super}. Each chart correspond to a data set, curves represent the performance of the classifier (on test and train set) under varying the number of labeled examples.

These figures show that the post-optimization of the bounds location has no significant effect on the performance of considered classifiers. Examining the detailed results, the differences between both discretization methods are in the range [0.01–0.1%], and the variances of results represent several percents. This behavior is consistent with the above described expected results.

Although the post-optimization of bounds location is not significant in practice, Figs. 5 and 6 [3] exhibit other interesting points:

- For each data set, performance on the train and test sets is relatively close to each other. The performance monotonically increases in both cases with the number of labeled examples. This point underlines the robustness of our discretization method.
- The results using M_{super} are very competitive and therefore they are very difficult to beat using M^*_{super}.

[3] In this case, the limit of the vertical axis is not the same for all the database, this limit is set to see the optimal test AUC.

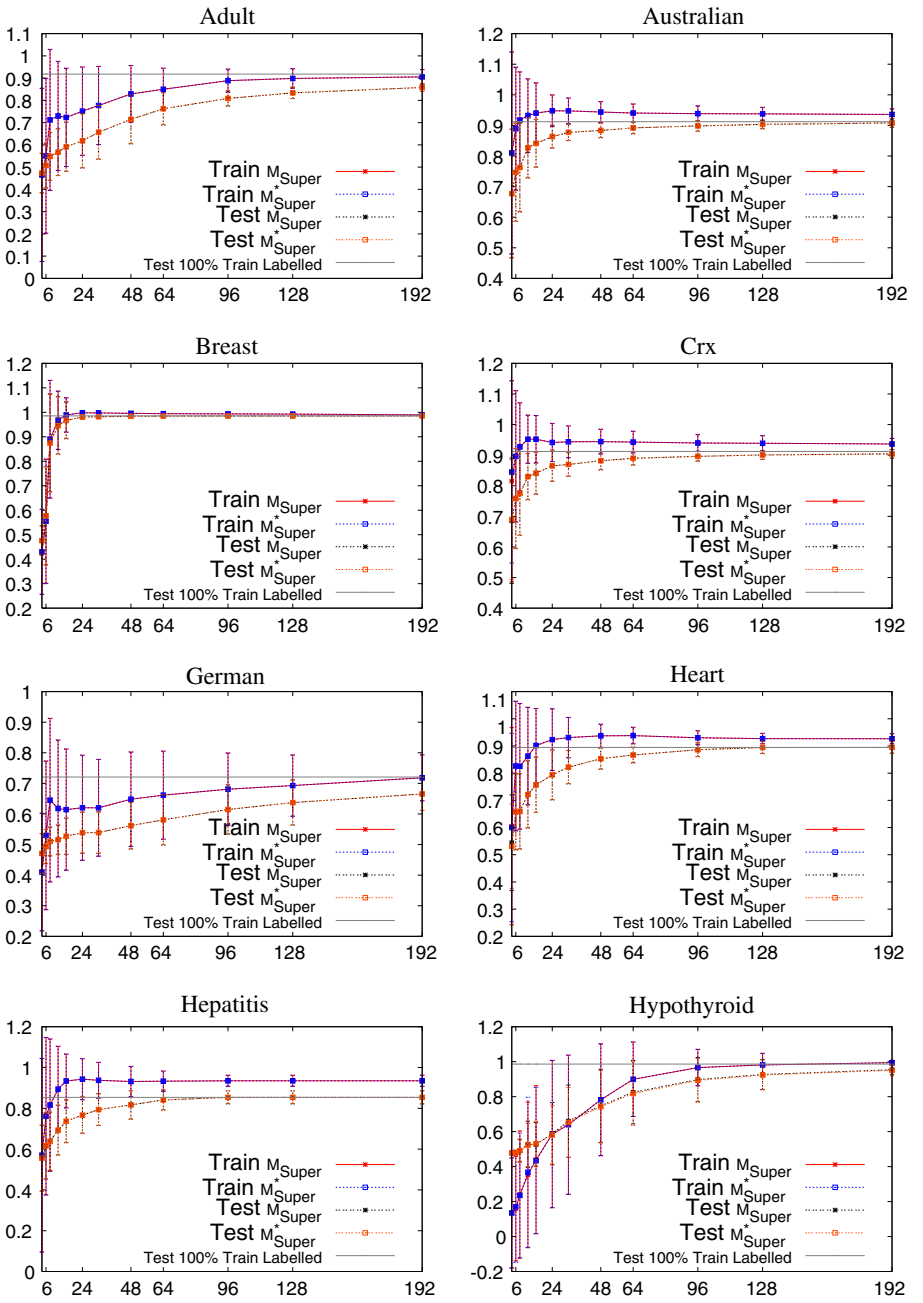

Fig. 5 Evaluation of the naive Bayes classifier where the discretizations are given either by M_{super} or by M^*_{super}. On each chart, the *vertical axis* corresponds to the average AUC and the *horizontal axis* corresponds to the number of labeled examples. AUC on train and test sets are plotted for both discretization methods. The optimal AUC, which is observed when all examples are labeled is also represented. On each curves, natches represent the variance of the AUC($\pm\sigma$)

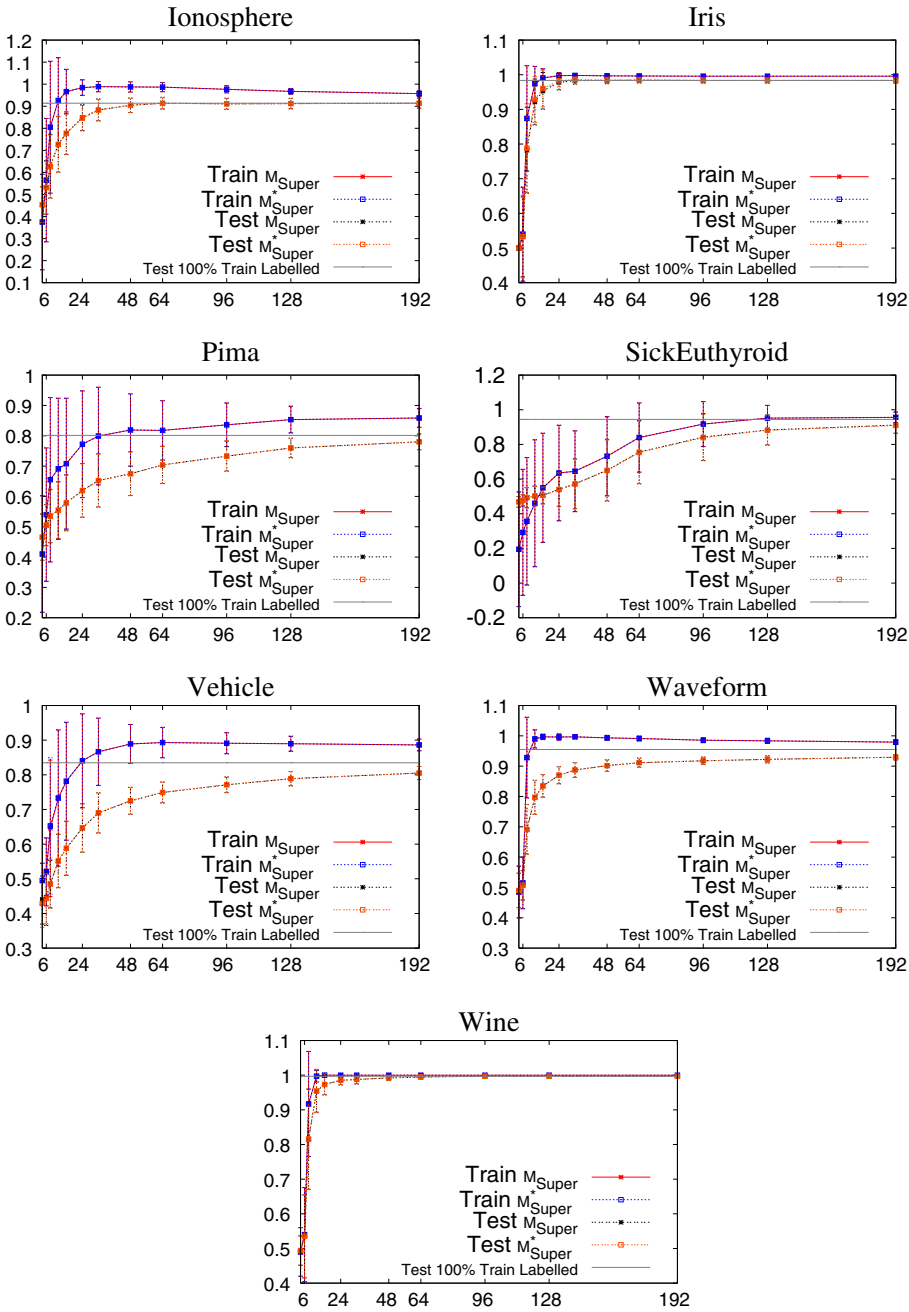

Fig. 6 Evaluation of the naive Bayes classifier where the discretizations are given either by M_{super} or by M^*_{super}. On each chart, the *vertical axis* corresponds to the average AUC and the *horizontal axis* corresponds to the number of labeled examples. AUC on train and test sets are plotted for both discretization methods. The optimal AUC, which is observed when all examples are labeled, is also represented. On each curves, natches represent the variance of the AUC($\pm\sigma$)

– The quality of the classifiers quickly increase, namely the best performance is reached labeling only few examples. The high convergence speed is an interesting behavior, which will be studied in an active learning context [22] in future work.

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 labeled and unlabeled examples.

This paper significantly extends the previous research of Boullé in [4] on supervised discretization method MODL, i.e., it presents a semi-supervised generalization of it where additional unlabeled 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 interval bounds must be placed in the middle of unlabeled areas to minimize the mean square error. The main contribution of this article is to demonstrate that the unlabeled 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 [4] 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.

According to our experimental results, the semi supervised discretization is not better than the supervised discretization. Our interpretation is that relying on few assumptions on the data distribution do not allow to take benefit from the unlabeled instances. This raises the question of whether the semi supervised framework is valuable in the non-parametric modeling, where less prior knowledge is available.

References

1. Berger J (2006) The case of objective Bayesian analysis. Bayesian Anal 1(3):385–402
2. Blum A, Mitchell T (1998) Combining labeled and unlabeled data with co-training. In: COLT '98: Proceedings of the eleventh annual conference on Computational learning theory. ACM Press, New York, pp 92–100
3. Boullé M (2005) A Bayes optimal approach for partitioning the values of categorical attributes. J Mach Learn Res 6:1431–1452
4. Boullé M (2006) MODL: a Bayes optimal discretization method for continuous attributes. Mach Learn 65(1):131–165
5. Catlett J (1991) On changing continuous attributes into ordered discrete attributes. In: EWSL-91: Proceedings of the European working session on learning on machine learning. Springer, New York, pp 164–178
6. Chapelle O, Schölkopf B, Zien A (2007) Semi-supervised learning. MIT Press, Cambridge
7. Dougherty J, Kohavi R, Sahami M (1995) Supervised and unsupervised discretization of continuous features. In: International conference on machine learning, pp 194–202
8. Fawcett T (2003) Roc graphs: notes and practical considerations for data mining researchers. Technical Report HPL-2003-4, HP Labs. http://citeseer.ist.psu.edu/fawcett03roc.html

9. Fayyad U, Irani K (1992) On the handling of continuous-valued attributes in decision tree generation. Mach Learn 8:87–102
10. Fayyad U, Piatetsky-Shapiro G, Smyth P (1996) From data mining to knowledge discovery: an overview. Adv Knowl Discov Data Min 1–34
11. Fujino A, Ueda N, Saito K (2007) A hybrid generative/discriminative approach to text classification with additional information. Inf Process Manage 43:379–392
12. Holte R (1993) Very simple classification rules perform well on most commonly used datasets. Mach Learn 11:63–91
13. Jin R, Breitbart Y, Muoh C (2009) Data discretization unification. Knowl Inf Syst 19(1):1–29
14. Kohavi R, Sahami M (1996) Error-based and entropy-based discretization of continuous features. In: Proceedings of the second international conference on knowledge discovery and data mining, pp 114–119
15. Langley P, Iba W, Thomas K (1992) An analysis of Bayesian classifiers. In: Press A (ed) Tenth national conference on artificial intelligence, pp 223–228
16. Liu H, Hussain F, Tan C, Dash M (2002) Discretization: an enabling technique. Data Min Knowl Discov 6(4):393–423
17. Maeireizo B, Litman D, Hwa R (2004) 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
18. Newman DJ, Hettich S, Blake CL, Merz CJ (1998) UCI repository of machine learning databases. Department of Information and Computer Sciences, University of California, Irvine. http://www.ics.uci.edu/~mlearn/MLRepository.html
19. Pyle D (1999) Data preparation for data mining. Morgan Kaufmann, San Francisco, p 19
20. Rissanen J (1978) Modeling by shortest data description. Automatica 14:465–471
21. Rosenberg C, Hebert M, Schneiderman H (2005) Semi-supervised self-training of object detection models. In: Seventh IEEE workshop on applications of computer vision
22. Settles B (2009) Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison
23. Shannon C (1948) A mathematical theory of communication. Key papers in the development of information theory
24. Sugiyama M, Krauledat M, Müller K (2007) Covariate shift adaptation by importance weighted cross validation. J Mach Learn Res 8:985–1005
25. Sugiyama M, Müller K (2005) Model selection under covariate shift. In: ICANN, International conference on computational on artificial neural networks: formal models and their applications
26. Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng A, Liu B, Yu PY, Zhou Z, Steinbach M, Hand DJ, Steinberg D (2008) Top 10 algorithms in data mining. Knowl Inf Syst 14(1)
27. Zhou ZH, Li M (2009) Semi-supervised learning by disagreement. Knowl Inf Syst doi:10.1007/s10115-009-0209-z
28. Zighed D, Rakotomalala R (2000) Graphes d'induction. Hermes, France

Author Biographies

Alexis Bondu was born in 1982 and he recently obtained a Ph.D. in Computer Science from the University of Angers (France). His Ph.D. focused on active learning using local models and was conducted in partnership with the "Statistical Processing of Information" research group of France Telecom R&D. Currently, he is a researcher in the "Commercial Innovations and markets analysis" Department of EDF R&D. His main research interests include stream mining, data mining and supervised classification.

Marc Boullé was born in 1965 and graduated from Ecole Polytechnique (France) in 1987 and Sup Telecom Paris in 1989. Currently, he is a senior researcher in the "Statistical Processing of Information" research group of France Telecom R&D. His main research interests include statistical data analysis, data mining, especially data preparation and modelling for large databases. He developed regularized methods for feature preprocessing, feature selection and construction, model averaging of selective naive Bayes classifiers and regressors.

Vincent Lemaire was born in 1968 and he obtained his undergraduate degree from the University of Paris 12 in signal processing and was in the same period an Electronic Teacher. He obtained a Ph.D. in Computer Science from the University of Paris 6 in 1999. He thereafter joined the R&D Division of France Télécom where he became a senior expert in data-mining. His research interests are the application of machine learning in various areas for telecommunication companies with an actual main application in data mining for business intelligence. He developed exploratory data analysis and classification interpretation tools. He obtained a HDR thesis ("Habilitation à diriger des recherches") in Computer Science from the University of Paris-Sud 11 (Orsay) in 2008.