Exploration *vs.* exploitation in active learning : a Bayesian approach

A. Bondu, V. Lemaire, M. Boullé

***Abstract*— The labeling of training examples could be a costly task in numerous cases of supervised learning. Active learning strategies address this problem and select unlabeled examples which are considered as the most useful for the training of a predictive model. The choice of examples to be labeled can be considered as a dilemma between the exploration and the exploitation of the input data space. In this article, a new active learning strategy that manages this compromise is proposed. This strategy is based on a Bayesian formalism that minimizes assumptions on data. An experimental validation is conducted on a unidimensional dataset, the objective is to assess the position of a step function from noisy examples. Our approach is favorably compared to a ad hoc strategy : the probabilistic dichotomy.**

I. Introduction

Machine learning refers to a wide range of methods and algorithms that allow a predictive model to learn behavior by using observations. In practice, the collection and the labeling of training examples could be a costly task for numerous supervised classification problems. This cost can be due to the requirement of a human expert, the use of measuring equipment, or a prohibitive computation time. Active learning strategies [1] select the unlabeled examples which are considered as the most useful to build a training set iteratively. An expert labels the selected examples. All active learning strategies have the same objective : to label the fewest examples as possible given a fixed performance the predictive model has to reach. This article considers the framework of selective sampling [2] where the predictive model observes a finite subset of examples : the creation of new examples is prohibited.

In this article, a new active learning strategy based on a semi-supervised Bayesian discretization method [3] is proposed. Section II briefly presents the semi-supervised discretization method on which our approach is based. The section III-A introduces the dilemma between exploitation vs. exploration. Our active learning strategy is formalized in Section III-B.

An experimental validation is conducted in Section IV. An applicative framework is defined : i) the handled discretization models include one or two intervals ; ii) the learning problem to be solved is a binary classification ; iii) the examples of the dataset are characterized by a single explicative variable. Although restrictive, this framework constitutes a first application of the proposed method. Comparative experiments which aim at detecting a step function from noisy examples are also presented in Section IV. These experiments show the behavior of our strategy and characterize the influence of the level of labeling noise on the quality of the predictive model. In this section, our approach is favorably compared to a ad hoc strategy : the probabilistic dichotomy [4].

Alexis Bondu is with EDF R&D, 1 avenue du Général de Gaulle 92140 Clamart France

Vincent Lemaire is with Orange Labs, 2 avenue Pierre Marzin, 22300 Lannion France. http ://perso.rd.francetelecom.fr/lemaire/

Marc Boullé is with Orange Labs, 2 avenue Pierre Marzin, 22300 Lannion France. http ://perso.rd.francetelecom.fr/boulle/

II. Semi-supervised Bayesian discretization

Notations : The data D includes two subsets L and U that respectively correspond to labeled and unlabeled training examples, with $D = L \cup U$. The set L is composed by couples (x, y), where $x \in \mathbb{R}$ and $y \in \mathbb{Y}$ is a discrete target value. The set U includes unlabeled examples denoted by $(x, ?)$. The following notations are adopted : N, the number of observable examples ($N = |D|$) ; N^l, the number of labeled examples ($N^l = |T|$) ; J, the number of possible classes ($J = |\mathbb{Y}|$).

The active learning strategy proposed in this article is based on a semi-supervised Bayesian discretization method which comes from the MODL framework [3]. In the case of semi-supervised learning, this method discretizes explicative variables in order to estimate conditional densities of classes. These estimated densities are supposed to be piecewise constant functions. The MODL approach turns the discretization problem into a model selection problem.

A family of feasible discretization models based on the order statistic is defined. A discretization model $M(I, \{N_i\}, \{N_{ij}\})$ is defined by the following parameters :

- I is the number of intervals ;
- $\{N_i\}$ is the number of examples in each interval
- $\{N_{ij}\}$ is the number of examples belonging to each class, in each interval.

The parameters $\{N_i\}$ specify the bounds of intervals through the rank of explicative values, and the parameters $\{N_{ij}\}$ characterize conditional densities by counting of each target value in the interval i.

A Bayesian approach is applied to select the best discretization model, denoted by $\mathcal{M}_{map}$ *(Maximum a posteriori)*. The best discretization model maximizes $P(M|D)$, the probability of the model M given the data D. Exploiting the Bayes formula and considering that $P(D)$ is constant over all possible models, this approach aims at maximizing $P(M)P(D|M)$. The prior distribution $P(M)$ and the likelihood of data $P(D|M)$ are analytically developed exploiting

the discrete family of models. The employed Bayesian approach adopts very low informative hypothesis on data, this is an objective Bayesian approach [5]. Eventually, the $\mathcal{M}_{map}$ minimizes Equation 1.

The criterion $\mathcal{C}_{semi\ super}$ corresponds to the negative logarithm of a probability, that represents a quantity of information [6]. The first term corresponds to the choice of the number of intervals, the second term describes the choice of bounds locations. The third term represents the choice of estimated distributions in each interval. The penultimate term corresponds to the probability to observe values of the predictive variable given the discretization model. The last term can be interpreted as a penalty due to unlabeled examples in each interval of the model (denoted by N_i^u and N_{ij}^u).

$$\overbrace{\mathcal{C}_{semi\ super}(M) = \log(N) + \log C_{N+I-1}^{I-1} \sum_{i=1}^{I} \log C_{N_i+J-1}^{J-1} + \underbrace{\sum_{i=1}^{I} \log\left(\frac{N_i!}{\sum_{j=1}^{J} N_{ij}!}\right) - \sum_{i=1}^{I} \log\left(\frac{N_i^u!}{\sum_{j=1}^{J} N_{ij}^u!}\right)}_{-logP(D|M)}}^{-logP(M)} \quad (1)$$

III. Active learning

This section introduces active learning and underlines the importance of the dilemma between the exploitation and the exploration of the input data space. The active learning strategy that we propose is formalized and presented in detail.

A. Dilemma between exploration and exploitation

During an active leaning process, the choice of examples to be labeled can be viewed as a dilemma between the exploration and the exploitation of the input data space (denoted by $\mathbb{X}$). The selection of an unlabeled example in a non-sampled area of $\mathbb{X}$ contributes to **explore** the data. In this case the space $\mathbb{X}$ tends to be uniformly sampled, that potentially limits the areas on which the predictive model is mistaken. The larger the dimension of $\mathbb{X}$, the more exploration of this space requires a large number of labeled examples. The selection of an unlabeled example in a sampled area of $\mathbb{X}$ contributes to **exploit** data. In this case, the active leaning strategy focuses on an area already populated with labeled examples, and locally refines the predictive model. The dilemma between exploration and exploitation can be illustrated by two extreme behaviors. On the one hand, an active learning strategy that only exploits data may ignore a large part of the input data space. The predictive model will be specialized on some areas of $\mathbb{X}$ but it will be definitely incorrect on the whole data space. On the other hand, an active learning strategy that only explores data does not focus on interesting area of $\mathbb{X}$ where the labeling of new examples could improve the predictive model. In such conditions, an active learning strategy represents little interest compared to a random sampling. These two extreme behaviors show that active learning strategies need to find a **compromise** between exploration and exploitation. In the literature, several approaches try to resolve this dilemma. Three of the most common approaches are presented below.

1) Simultaneous use of multiple strategies: Several active learning strategies can be jointly exploited in order to find a compromise between exploration and exploitation. If we consider two strategies which are respectively dedicated to the exploitation and the exploration of data, such as the uncertainty sampling [7] and the random sampling. At each iteration, one of these strategies is used. The choice of the strategy is probabilistic, p represents the probability to explore the data and $1-p$ is the probability to exploiting ; the probabilities are updated during the active learning process and settle the compromise exploration *vs.* exploitation. T. Osugi [8] draws a parallel with reinforcement learning. The predictive model is considered as an *"agent"* able to carry out two actions : explore or exploit data. At each iteration, the agent executes one of these actions and receives an award [*respectively* a retribution] if this action is appropriat [*respectively* inappropriate]. A metric is used to assess changes of the hypothesis that is learned by the predictive model : an action is awarded proportionally to the observed changes.

Some heuristics comming from the combinatorial optimization can also be used to respond to the exploration *vs.* exploitation dilemma. T. Zoller [9] exploits the simulated annealing algorithm to adjust p over time. This heuristic is inspired by the thermodynamics and decreases the probability p according to a predefined function (also called cooling schema). This heuristic condenses the exploration of the data space at the beginning of the active learning process, and exploits this space again using representative examples of $\mathbb{X}$.

2) pre-clustering: H. Nguyen improves the uncertainty sampling strategy [7] applying a pre-clustering on data, this technique allows diversifying the labeled examples. Only centroids of clusters are candidates to be labeled. Examples which are assigned to a given cluster are supposed to belong to the same class and the centroid is supposed to be representative of these examples. H. Nguyen defines a criterion that selects the centroid which most contributes to the current error. The user changes the size of cluster over time. The decrease of the sizes of clusters responds to the compromise exploration *vs.* exploitation. When clusters include many examples, the centroids are distant from each other. In this case, this strategy mainly explores the input data space. By contrast, when clusters include few examples the centroids are potentially close to each other. In this case, this pre-clustering based approach mainly exploits data. In the same way as the *simulated annealing* [9], this strategy explores the data more at the beginning of the active learning process than at the end.

3) Use of similarity measure: Some active learning strategies manage the compromise exploration / exploitation by measuring the dissimilarity between selected examples. Xu [10] proposes a multi-criteria active learning strategy that is applied to documentary research. This approach maximizes

the distance between the new example and the closest labeled one : this strategy labels the examples the most distant from each other. Brinker [11] presents a kernel based strategy that uses dissimilarity measure. At each iteration, this strategy labels the set of examples that most reduces the versions space of the predictive model. Each unlabeled example corresponds to a hyperplane in the prehilbertian space that is induced by the kernel. This strategy selects the unlabeled examples for which the corresponding hyperplanes are the most distant from each other. An angle measure is defined by exploiting the kernel trick.

Finaly, the 3 approaches presented above and the state of the art [12, 13] underline that the compromise between exploration and exploitation is a focal question in the active learning field. The strategy that we propose in this article find a compromise based on a Bayesian formalism that minimizes assumptions on data. Our strategy does not require adjusting user parameters.

B. A new strategy

This section presents an original active learning strategy based on the semi-supervised discretization method that is described on Section II. The quality of a discretization model is given by the probability of this model given the data. The criterion $\mathcal{C}_{semi\ super}$ is an analytical expression of $P(M|D)$, in the meaning of the modeling hypothesis of the MODL approach [14]. Our strategy aims at labeling the example that will maximize the quality of the future predictive model, without knowing the label of the new example and without knowing the best model of the next iteration. Our approach takes into account these uncertainties conducting an expectancy calculation over all possible cases. A optimization criterion designates the example $x_{t+1} \in U$ that maximizes the expectation of $P(M|D, x_{t+1})$ is defined.

Let $P_{(.|D)}(M) = P(M|D)$ be the posterior distribution of discretization models given the data. Our strategy selects the example $x_{t+1} \in U$ that maximizes the expectation of $P(M|D, x_{t+1})$ over the family of models $\mathbb{M}$:

$$\underset{x_{t+1} \in U}{ArgMax} \ \mathrm{E}_{\substack{P_{(.|D)} \\ M \in \mathbb{M}}} \ [P(M|D, x_{t+1})]$$

$$= \underset{x_{t+1} \in U}{ArgMax} \sum_{M \in \mathbb{M}} P(M|D) \times P(M|D, x_{t+1})$$

The label y_{t+1} is not known, but the probability $P(y|M, D, x_{t+1})$ of the class $y \in \mathbb{Y}$ given the model and the data can be estimated. Owing to the formula of total probability, we write :

1) $\sum_{y \in \mathbb{Y}} P(y|M, D, x_{t+1}) = 1$
2) $P(M|D, x_{t+1}) = \sum_{y \in \mathbb{Y}} P(y|M, D, x_{t+1}) P(M|D, x_{t+1}, y)$

At last ,

$$\underset{x_{t+1} \in U}{ArgMax} \sum_{M \in \mathbb{M}} P(M|D) \times \left[\sum_{y \in \mathbb{Y}} P(y|M, D, x_{t+1}) \times P(M|D, x_{t+1}, y) \right]$$

This expression is developed exploiting the Bayes formula :

$$\underset{x_{t+1} \in U}{ArgMax} \sum_{M \in \mathbb{M}} \left[\frac{P(M) \times P(D|M)}{P(D)} \times \right.$$

$$\left. \sum_{y \in \mathbb{Y}} \left[P\left(y|M, D, x_{t+1}\right) \times \frac{P(M) \times P(D, x_{t+1}, y|M)}{P(D, x_{t+1}, y)} \right] \right] \quad (2)$$

The joined probability $P(D, x_{t+1}, y)$ is developed :

$$P(D, x_{t+1}, y) = P(D) \times P(x_{t+1}|D) \times P(y|D, x_{t+1}) \quad (3)$$

The example x_{t+1} is considered as uniformly drawn from the set U, a priori, all unlabeled examples have the same probability to be selected. Consequently, the terms $P(D)$ and $P(x_{t+1}|D)$ of Equation 2 are constant under varying the model M. This equation can be written as follows :

$$\underset{x_{t+1} \in U}{ArgMax} \sum_{M \in \mathbb{M}} \left[\overbrace{P(M) \times P(D|M)}^{A} \times \sum_{y \in \mathbb{Y}} \right.$$

$$\left. \left[\underbrace{P(y|M, D, x_{t+1})}_{C} \times \frac{\overbrace{P(M) \times P(D, x_{t+1}, y|M)}^{B}}{\underbrace{P(y|D, x_{t+1})}_{D}} \right] \right] \times Cste \quad (4)$$

Where :

– The term "A" of Equation 4 is deduced by the criterion $\mathcal{C}_{semi\ super}$:

$$P(M)P(D|M) = \frac{1}{N} \times \frac{1}{C_{N+I-1}^{I-1}} \times \prod_{i=1}^{I} \frac{1}{C_{N_i+J-1}^{J-1}} \times \prod_{i=1}^{I} \left[\frac{\prod_{j=1}^{J} N_{ij}!}{N_i!} \times \frac{N_i^u!}{\prod_{j=1}^{J} N_{ij}^u!} \right] \quad (5)$$

– The term "B" is calculated by the same way, adding the couple (x_{t+1}, y) to the training set L.
– The term "C" is evaluated by the prediction of the model M : the current model estimates the probability to observed the class y given the example x_{t+1}. This prediction is based on the proportion of examples labeled with the value y in the interval which includes the example x_{t+1}.
– The term "D" represents the probability to observe the class y, given the example x_{t+1} and the data. This term is difficult to assess because any particular discretization model is involved. In order to estimate this term, we choose to integrate the calculation over the family of models $\mathbb{M}$. Exploiting the total probability formula $[\sum_{M' \in \mathbb{M}} P(M'|D) = 1]$, we can write :

$$P(y|D, x_{t+1}) = \sum_{M' \in \mathbb{M}} P(M'|D) \times P(y|D, M', x_{t+1})$$

$$= \sum_{M' \in \mathbb{M}} \frac{P(D|M')P(M')}{P(D)} \times P(y|D, M', x_{t+1})$$

The probability of data $P(D)$ is constant under varying the model M.

Finally, the expectation of $P(M|D, x_{t+1})$ is evaluated by the criterion $\mathcal{C}_{active}(x_{t+1})$:

$$\mathcal{C}_{active}(x_{t+1}) = \sum_{M \in \mathbb{M}} \Bigg[P(M)P(D|M) \times \sum_{y \in \mathbb{Y}} \Bigg[\frac{P(y|M, D, x_{t+1}) \times P(M) \times P(D, x_{t+1}, y|M)}{\sum_{M' \in \mathbb{M}} P(M') \times P(D|M') \times P(y|D, M', x_{t+1})} \Bigg] \Bigg] \quad (6)$$

IV. Experimental validation

In this section our active strategy is evaluated in a simple case : the estimation of the location of a step function from noisy examples [15]. This learning problem constitutes a preliminary experimentation and highlights the behaviors of our stategy compared to an ad-hoc method. In this particular case, where examples are defined by a single explicative variable and where the discretization model includes one or two intervals, the criterion $\mathcal{C}_{active}$ can be optimized with a 'reasonable' time complexity.

First, this section presents the considered dataset. Two competitor strategies are considered : the probabilistic dichotomy that is an ad-hoc method toward the detection of noisy step function, and the random sampling strategy that gives baseline results. Our experiments lead to several types of results : (i) illustrative results that exhibit the behavior of our strategy in term of selection of examples ; (ii) comparative results that evaluate the performance of strategies depending on the number of labeled examples.

A. Data

The exploited dataset contains 100 examples which are uniformly distributed in the interval $[0, 1]$. The variation domain of the explicative variable x is split into two parts $[0, \theta[$ and $[\theta, 1]$ where θ is the step location. The objective is to "detect" θ from noisy examples that are potentially mislabeled. The majority of training examples located in the interval $[0, \theta]$ [*respectively* $]\theta, 1]$] belong to the class '1' [*respectively* of the class "2"]. The probability that an example is mislabeled is denoted by $p \in [0, 0.5]$. The Figure 1 draws the probability to observe the class '1' given the value of x.

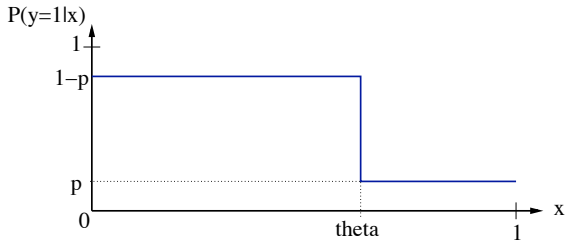

Fig. 1. Step function : Probability to observe the class '1' knowing the value of x.

B. Rival strategies

1) Random strategy: The random sampling strategy uniformly selects the examples to be labeled. This baseline strategy gives minimum performances and allows to estimate the contribution of the two others strategies.

2) probabilistic Dichotomy: In [4] the authors generalize of the dichotomy to the use of noisy examples. This ad-hoc strategy is specialized in finding the location of a step function, and assumes the level of noise is known. Algorithm 1 describes this strategy, the density of the step location is denoted by $P_\theta(x)$. Initially no label is available. The density $P_\theta(x)$ is initialized on step (I) adopting an uniform prior. At each iteration, the unlabeled example $x^* \in U$ which 'splits' the distribution $P_\theta(x)$ in two equal parts is selected (Step A). x^* is the example which is the closest to the median of $P_\theta(x)$. The label $f(x^*)$ is sought from an expert (Step B) and the couple $(x^*, f(x^*))$ is added to the training set L. The density $P_\theta(x)$ is updated at each iteration exploiting the new target value$f(x^*)$ (Step C). This step takes into account the level of noise p.

Notations :

- $P_\theta(x)$ the a priori probability distribution of θ.
- p the probability that a label is false.
- A noisy step function, such as the interval $[0, \theta]$ [*respectively* $[\theta, 1]$] contains a majority of examples which belong to the class "1" [*respectively* to the class "2"].
- n the number of (desired) training examples.
- U and L the unlabeled and labeled set, with $L \cup U = \Phi$
- T the labelling budget, with $|T| < n$

/*initialization of L et U*/
$L = \emptyset$ et $U = \Phi$

/*initialization of $P_\theta(x)$*/
(I) $P_\theta(x) \leftarrow \frac{1}{N} \quad \forall x \in \Phi$

Repeat
(A) Find the unlabeled example $x^* \in U$ such as : $\sum_{x \in [0, x^*]} P_\theta(x) = \frac{1}{2}$
(B) Request of the label $f(x^*)$, add $(x^*, f(x^*))$? to L, withdraw x^* of U.
(C) Update $P_\theta(x)$ such as :
If $f(x^*) = 1$ **then**
(i) $P_\theta(x) \leftarrow 2.p.P_\theta(x) \quad \forall x \in [0, x^*]$
(ii) $P_\theta(x) \leftarrow 2.(1-p).P_\theta(x) \quad \forall x \in]x^*, 1]$
end If
If $f(x^*) = 2$ **then**
(iii) $P_\theta(x) \leftarrow 2.(1-p).P_\theta(x) \quad \forall x \in [0, x^*]$
(iv) $P_\theta(x) \leftarrow 2.p.P_\theta(x) \quad \forall x \in]x^*, 1]$
end If
until $|T| < n$

Algorithm 1: probabilistic Dichotomy

C. Illustrative Results

This section presents illustrative results of our stategy and shows the changes of the criterion $\mathcal{C}_{active}(x_{t+1})$ (Equation 6) over time. The selected examples are indicated as well as their corresponding labels. The location of the step is fixed at $\theta = 0.5$.

1) Step without noise: Figure 2 shows the selection of the examples during the first iterations of the active learning process. On each sub-figure the vertical axis represents the

expectanty [1] of $\mathcal{C}_{active}(x_{t+1})$ versus the location of the candidate examples to be labeled. The maximum value of each curve is symbolized by a "▼" and corresponds to the location of the selected example at each iteration. Labeled examples belonging to the class "1" [*respectively* "2"] are symbolized by a "•" [*respectively* "•"].

Initially, there is no labeled examples. During the first iteration (chart "A", Figure 2), the criterion $\mathcal{C}_{active}(x_{t+1})$ reaches its highest values for $x_{t+1} = 0$ and $x_{t+1} = 1$. The semi-supervised version of the MODL criterion (see Section II) penalizes the unlabeled examples and thus generates the selection of an example that is located on one end of the interval $[0, 1]$. In case of equality, one of these two possible examples is randomly chosen. Here, the example at $x_{t+1} = 1$ is labeled by the class "2". At the second iteration (chart "B", Figure 2), the example $x_{t+1} = 0$ is labeled by the class "1". At the third iteration the curve seems to be flat (chart "C", Figure 2) but the criterion $\mathcal{C}_{active}(x_{t+1})$ reaches its maximum value at two symmetrical locations. The example at $x_{t+1} = 0.28$ is selected and labeled. From these 2 labels per class, our strategy adopts a behavior which seems to be similar to the dichotomy. The charts D to F of the Figure 2 show next iterations. Our strategy converges and finds the real location of the step by labeling 9 examples.

a) Noisy Step: The same experiments as Section IV-C.1 has been realized introducing a mislabeled example. In this case our strategy adopts a behavior which has two principal characteristics :

1) In a first period, our strategy tries to find the location of the step 'around' the noisy example. This behavior is consistent since nothing presumes that a label is false.
2) Then the strategy "detects" the noisy example (the two examples on each sides of the noisy example have been labeled) and from here adopts the same behavior as Section IV-C.1 (Figure 2).

Finally, our strategy adopts a correct behavior face to noisy examples. The real location of the step function is determined labeling 12 examples.

D. Comparative results

This section presents comparative experiments realized on a step function, the step location is set at $\theta = 0.675$. The objective of these experiments is to evaluate the influence of the level of labeling noise (denoted by p) on the quality of the predictive model. The performances of the three active learning strategies are evaluated under varying p in the interval $[0.0 - 0.20]$. All the active strategies are evaluated using the same predictive model, therefore only the examples selection influences one the results. The used predictive model is the MAP model ($\mathcal{M}_{map}$) defined at Section II, this model includes one or two intervals and has the choice to split (or not) the variation domain of the variable x. The AUC is used to compare the results of the three strategies depending on the number of labeled examples.

All the experiments have been done 100 times to obtain a mean and a variance of the performance. Initially, there is no labeled example $L = \emptyset$. At each iteration, one example is selected and labeled. The experiments are stopped when the budget of 20 labels is reached. By contrast with to the others strategies, the probabilistic dichotomy has to be informed by the level of noise p. Three cases have been considered below.

1) Case where the probabilistic dichotomy is correctly informed by the level of noise: Figure 3 plots the AUC [2] (vertical axis) depending on the number of labeled examples (horizontal axis). The 4 charts correspond to different levels of labeling noise $p = 0.0, 0.05, 0.10, 0.15$. In each chart the random strategy is plot by the red curve, the probabilistic dichotomy is plot by the blue curve and our strategy is plot by the green curve.

The first chart, where $p = 0$, can ben interpreted as follows :

- The random strategy (red curve) has a constant AUC that equals to 0.5 when $|L|$<12. In this case the MAP model has a single interval and estimates the conditional distribution of classes as uniform. Then, when $|L| \geq 12$, a sufficient number of labeled examples is present to produce a MAP model which has 2 intervals. The random strategy progresses and reach an AUC of 0.95 for $L = 20$.
- The strategy based on the probabilistic dichotomy has the same behavior as the random strategy when $|L|$<12. Then, when $|L| \geq 12$, this strategy progresses very quickly to reach the optimal performance since $\mathcal{M}_{map}$ has 2 intervals ($|L| = 13$). In the case where the noise is null this strategy is the best one.
- Our strategy, based on the maximization of the expectancy of $P(M|D, x_{t+1})$, presents 2 interesting characteristics : (i) it is better than the random strategy, (ii) it needs less labeled examples to produce an optimal model with 2 intervals (AUC>0.5 for L>10). This strategy reaches the optimal performance for L=15 and is (for $p = 0$) slightly less good than the probabilistic dichotomy.

When the level of noise is known and null (a very favorable case for the dichotomy) the 3 strategies are ranked as follows : 1) the probabilistic dichotomy, 2) our strategy based on the maximization of the expectancy of $P(M|D, x_{t+1})$ and 3) the random sampling strategy.

This analysis can be done identically for the others charts ($p = 0.05, 0.10, 0.15$). When the level of noise increases the performances of the three strategies decreases. Our strategy stays better than the random sampling for $p = 0.05$ and $p = 0.10$. The probabilistic dichotomy declines clearly when p increases. When the noise increases the number of labeled examples required to obtain a $\mathcal{M}_{map}$ with two intervals is greater.

2) Case where the dichotomy is informed that $p = 0$: Several other experiments have been realized in the case

1. The expectancy of $\mathcal{C}_{active}(x_{t+1})$ is normalized such as its maximum equals to be 1.

2. Mean and variance of AUC are plot on Figure 3 over 100 repeated experiments, with whiskers = $\pm 2\sigma$

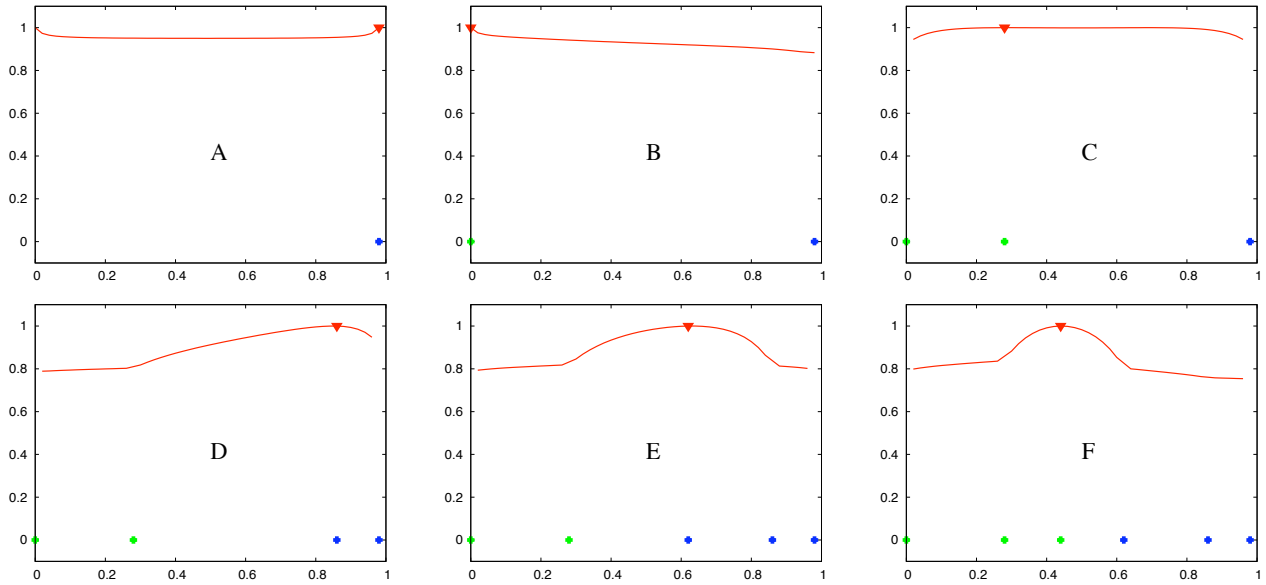

Fig. 2. Visualization of the positions of the labeled examples for the un-noisy step function.

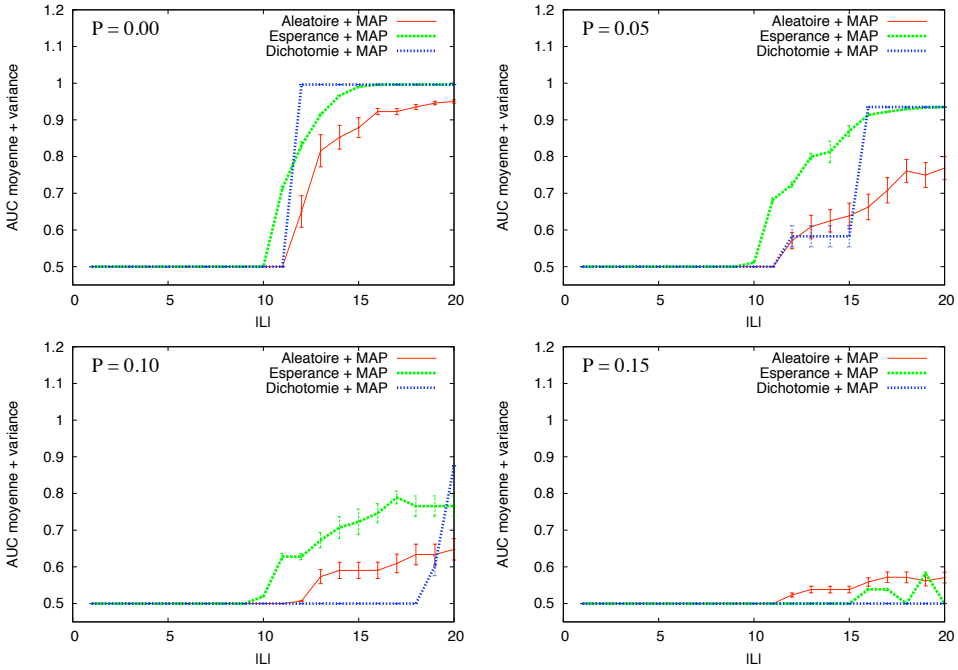

Fig. 3. Performances of the three active strategies *vs.* the number of labeled examples, in the case where the probabilistic dichotomy is correctly informed of the level of noise.

where the probabilistic dichotomy is informed that $p = 0$ while p is varying. In this case, our strategy gives better results and the gap increases when the level of noise in the data increases.

3) Case where the dichotomy is informed that $p_{false} \in [0, 0.20]$, even though $p = 0$: In the last experiments the dichotomy is misinformed by p_{false} even though $p = 0$. In this case, the performance of the probabilistic dichotomy are very low when p_{false} reaches 0.1 : the MAP model ($\mathcal{M}_{map}$) includes a single interval irrespective of the number of labeled examples.

To conclude, these experiments show that our active learning strategy outperforms the probabilistic dichotomy, in particular when the dichotomy is misinformed of the level of labeling noise in data.

V. Conclusion and Perspectives

In this article a new active learning strategy that is based on a semi-supervised discretization method from the MODL

family [14] has been presented. This strategy selects the unlabeled example which maximizes the expectancy of the probability of the discretization models, given the data and an additional example x_{t+1}. Our approach leads to an optimization criterion, $\mathcal{C}_{asset}(active)(x_{t+1})$. Our framework was restricted here to unidimensional dataset and to models that include one or two intervals. In the case of the detection of a step location from noisy examples, our approach has been favorably compared to a ad-hoc strategy : the probabilistic dichotomy. The comparative experiments realized in Section IV-D exhibits interesting results mainly when the level of noise in not known. This result is promising for future research.

Our active learning strategy could be exploited by other learning methods which need a dichotomy on noisy data. To elaborate decision tree, incremental methods use progressively the information in the training set [16]. These methods do not allow the active selection of training examples to be labeled. A binary tree would be defined using the criterion $\mathcal{C}_{active}$ presented in this paper. Further works on active trees using our active strategy could be investigated.

RÉFÉRENCES

[1] R. Castro, R. andWillett and R. Nowak. Faster rate in regression via active learning. In *NIPS (Neural Information Processing Systems)*, Vancouver, 2005.

[2] Nicholas Roy and Andrew McCallum. Toward optimal active learning through sampling estimation of error reduction. In *Proc. 18th International Conf. on Machine Learning*, pages 441–448. Morgan Kaufmann, San Francisco, CA, 2001.

[3] A. Bondu, M. Boullé, and V. Lemaire. A Non-parametric Semi-supervised Discretization Method. In *ICDM (International Conference on DataMining)*, Pise, december 2008.

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

[5] C.P. Robert. *Le choix bayésien Principes et pratique*. Springer, 2006.

[6] C.E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3) :379–423, 1948.

[7] Sebastian B. Thrun and Knut Möller. Active exploration in dynamic environments. In John E. Moody, Steve J. Hanson, and Richard P. Lippmann, editors, *Advances in Neural Information Processing Systems*, volume 4, pages 531–538. Morgan Kaufmann Publishers, Inc., 1992.

[8] T. Osugi, D. Kun, and S. Scott. Balancing exploration and exploitation : A new algorithm for active machine learning. In *ICDM '05 : Proceedings of the Fifth IEEE International Conference on Data Mining*, pages 330–337, Washington, DC, USA, 2005. IEEE Computer Society.

[9] T. Zoller and M. Buhmann. Active learning for hierarchical pairwise data clustering. In *ICPR '00 : Proceedings of the 15th International Conference on Pattern Recognition*, pages 186–189, 2000.

[10] Z. Xu, R. Akella, and Y. Zhang. Incorporating diversity and density in active learning for relevance feedback. In *ECIR (European Conference on Information Retrieval)*, volume 4425, pages 246–257, 2007.

[11] K. Brinker. Incorporating Diversity in Active Learning with Support Vector Machines. In *ICML : International Conference on Machine Learning*, pages 59–66, 2003.

[12] Anonyme. *Manuscrit de these du premier auteur*. Phd thesis, XXX, 2008.

[13] Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

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

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

[16] P.E. Utgoff. Incremental Induction of Decision Trees. *Machine Learning*, 4 :161–186, 1989.