

Adaptive Curiosity for Emotions Detection in Speech

Alexis Bondu, Vincent Lemaire

Abstract—Exploratory activities seem to be crucial for our cognitive development. According to psychologists, exploration is an intrinsically rewarding behaviour. The developmental robotics aims to design computational systems that are endowed with such an intrinsic motivation mechanism. There are possible links between developmental robotics and machine learning. Affective computing takes into account emotions in human machine interactions for intelligent system design. The main difficulty to implement automatic detection of emotions in speech is the prohibitive labelling cost of data. Active learning tries to select the most informative examples to build a training set for a predictive model. In this article, the adaptive curiosity framework is used in terms of active learning terminology, and directly compared with existing algorithms on an emotion detection problem.

I. INTRODUCTION AND NOTATION

Human beings develop in an autonomous way, carrying out exploratory activities. This phenomenon is an intrinsically motivated behavior. Psychologists [1] have propose theory which explain exploratory behavior as a source of self rewarding. Building a robot with such behavior is a great challenge of developmental robotics. The ambition of this field is to build a computational system that tries to capture curious situations. Adaptive curiosity [2] is one possibility to reach this objective, it pushes a robot towards situations in which it maximizes its learning progress. The robot first spends time in situations that are easy to learn, then shifts progressively its attention to more difficult situations, avoiding situations in which nothing can be learnt.

On the one hand, active learning brings into play a predictive model that explores the space of unlabelled examples, in order to find the most informative ones. On the other hand, adaptive curiosity allows a robot to explore its environment in an intelligent way, and tries to deal with the exploration / exploitation dilemma. This article uses the bridge elaborated in [3] between developmental robotic and classical machine learning to explore the data.

The organization of this paper is as follow: in section II adaptive curiosity is presented in a generic way, and original choices of implementation are described. The next section shows a possible implementation of adaptive curiosity for classification problems, a new criterion of zones selection is proposed. Section III compares the new adaptive curiosity strategy with two other active learning strategies, on a detection of emotions problem. Finally, possible improvements of this adaptive curiosity are discussed.

Notations:

$\mathcal{M} \in \mathbb{M}$ is the predictive model that is trained with an

algorithm \mathcal{L} . $\mathbb{X} \subseteq \mathbb{R}^n$ represents all possible input examples of the model and $x \in \mathbb{X}$ is a particular example. \mathbb{Y} is the set of possible outputs of the model; $y \in \mathbb{Y}$ refers to a class label which is associated to $x \in \mathbb{X}$.

The point of view of selective sampling is adopted [4] in this paper. The model observes only one restricted part of the universe $\Phi \subseteq \mathbb{X}$ which is materialized by training examples without label. The image of a “bag” containing examples for which the model can ask for associated labels is usually used to describe this approach. The set of examples for which the labels are known (at one step of the training algorithm) is called L and the set of examples for which the labels are unknown is called U with $\Phi = U \cup L$ and $U \cap L = \emptyset$.

The concept which is learnt can be seen as a function, $f : \mathbb{X} \rightarrow \mathbb{Y}$, with $\hat{f}(x_1)$ the desired answer of the model for the example x_1 . $\hat{f} : \mathbb{X} \rightarrow \mathbb{Y}$ is the answer of the model; an estimate of the concept. The elements of L and the associated labels constitute a training set T . The training examples are pairs of input vectors and desired labels such as $(x, f(x))$.

II. ADAPTIVE CURIOSITY - ORIGINAL CHOICES

A. Generic Algorithm

Adaptive curiosity [2] involves a double strategy. The first strategy makes a recursive partitioning of \mathbb{X} , the input space of the model. The second strategy selects zones to be fed with labelled examples (and to be split by recursive partitioning). It is an active learning as long as the selection of a zone defines the subset of examples which can be labelled (those which belong to the zone). Adaptive curiosity is described below in a generic way and illustrated by an algorithm.

The input space \mathbb{X} is recursively partitioned in zones (some of them are included in others). Each zone corresponds to a type of situations the robot must learn. Adaptive curiosity uses a criterion to select zones and preferentially splits areas of input space \mathbb{X} in which learning improves. The main idea is to schedule situations to be learnt in order to accelerate the robot’s training.

Each zone is associated with a sub-model which is trained with examples belonging only to the zone. Sub-models are trained at the same time, on disjointed examples sets. The partitioning of the input space is progressively realized while new examples are labelled. Just before the partitioning of a zone, the sub-model of the “parent” zone is duplicated in “children” zones. Duplicated sub-models continue independently its learning thanks to the examples that appear in their own zones.

Algorithm (1) shows the general steps of adaptive curiosity. It is an iterative process during which examples are selected and labelled by an expert. A first criterion chooses a

Orange Labs, 2 avenue Pierre Marzin, 22307 Lannion Cedex - France, (email: {alexis.bondu,vincent.lemaire}@orange-ftgroup.com)

zone to be fed with examples (stage A). The following stage consists in drawing an example from the selected zone (stage B). The expert gives the associated label (stage C) and the sub-model is trained with an additional example (stage D). A second criterion determines if the current zone must be partitioned. In this case, one seeks adequate separations in the “parent” zone to create “children” zones (stage i). Lastly, the sub-model is duplicated into the “children” zones (stage ii).

Given:

- a learning algorithm \mathcal{L}
- a set $\mathcal{M} = \{m_1, m_2, \dots, m_n\}$ of n predictive sub-models
- $U = \{u_1, u_2, \dots, u_n\}$, n subsets of unlabelled examples
- $L = \{l_1, l_2, \dots, l_n\}$, n subsets of labelled examples
- $T = \{t_1, t_2, \dots, t_n\}$ the training subsets corresponding to sub-models, with $t_i = \{(x, f(x))\} \forall x \in l_i$

$n \leftarrow 1$

Repeat

(A) Choose a sub-model m_i to be fed with examples

(B) Draw a new example x^* from u_i

(C) Label the instance x^* , $t_i \leftarrow t_i \cup (x^*, f(x^*))$

(D) Train the sub-model m_i thanks to \mathcal{L} , U and t_i

If the split criterion is satisfied **then**

(i) Separate l_i in two sub-sets l_j and l_k the most homogeneous as possible

(ii) Duplicate m_i into two sub-models m_j and m_k

(iii) $n \leftarrow n + 1$

end If

until $U \neq \emptyset$

Algorithm 1: Adaptive Curiosity

B. Parameters - Original Choices

The main purpose of this algorithm is to seek interesting zones in the input space while the machine discovers data to learn. The algorithm chooses, as soon as possible, the examples belonging to the zones where there is possible progress. Five questions appear: (i) How to decide if a zone must be partitioned? (ii) How to carry out the partitioning? (iii) How many “children” zones? (iv) How to choose zones to be fed with labelled examples? and (v) What kind of sub-models must be used?

The following paragraphs describe the original answers of P. Y. Oudeyer to these questions [2].

Partitioning:

A zone must be partitioned when the number of labelled examples exceeds a certain threshold. Partitioned zones are those which were preferentially chosen during previous iterations. These zones are interesting to be partitioned when

more populated. Associated sub-models have done important progress.

To cut a “parent” zone into two “children” zones, all dimensions of the input space \mathbb{X} are considered. For each dimension, all possible cut values are tested using the sub-model to calculate the variance of example’s predictions on both sides of the separation. During this stage, observable data Φ is used. This criterion¹ consists in finding a dimension to cut and a cut value minimizing the variance. This criterion elaborates preferentially pure zones to facilitate the learning of associated sub-models. Another constraint is added by the authors, the cut has to separate labelled examples into two subsets whose cardinalities are about balanced.

Zones selection:

At every iteration, the sub-model that most improves results is considered as having the strongest potential of improvement. Consequently, adaptive curiosity needs an estimation of sub-model’s progress. Firstly, performances of sub-models are measured on labelled data. The choice of a measure of performance is required. Secondly, sub-models’ performances are evaluated on a temporal window. The sub-model that realizes the most important progress is chosen to be fed with new examples that are uniformly drawn.

III. ADAPTIVE CURIOSITY FOR CLASSIFICATION

A. Introduction

The original criterion of zones selection is not adapted to classification problems and it is difficult to implement [3]. Indeed, this criterion requires a measure of performance which variations are examined on a temporal window to estimate robot’s progresses.

Adaptive curiosity tries to deal with the dilemma exploration / exploitation drawing new examples in zones where progress is possible. To take in consideration the dilemma exploration / exploitation in a better way, a new criterion of zones selection is proposed in this section. The rest of the adaptive curiosity method is not modified. The new criterion is composed by two terms which respectively correspond to the exploitation and the exploration. A compromise between both terms is provided by the new criterion.

Others implementation elements are exposed in section VI such parameters of the partitioning strategy (see VI-C), or as the experimental protocol (see VI-E).

¹This recursive partitioning uses a discretization method. For a state of the art on discretization methods, interested readers can refer to [5].

B. Exploitation: Mixture rate

Among existing splitting criteria [6], we use the entropy as a mixture rate. The function $MixRate(l)$ (equation 1) uses labels of examples $l \subseteq L$, that belong to the zone, to calculate the entropy over classes.

Part “A” of equation 1 corresponds to the entropy of classes that appear in a zone. Probabilities of classes $P(y_i)$ are empirically estimated by a counting of examples which are labelled with the considered class.

The entropy belongs to the interval $[0, \log |\mathbb{Y}|]$ with $|\mathbb{Y}|$ the number of classes. Part “B” of equation 1 normalizes mixture rate in the interval $[0, 1]$.

$$MixRate(l) = - \underbrace{\sum_{y_i \in \mathbb{Y}} P(y_i) \log P(y_i)}_A \times \underbrace{\frac{1}{\log |\mathbb{Y}|}}_B \quad (1)$$

$$\text{with } P(y_i) = \frac{|x \in l, f(x) = y_i|}{|l|}$$

Mixture rate is the “exploitation” term of the proposed zones selection criterion. By choosing zones that have the strongest entropy, the hidden pattern is locally clarified thanks to new labelled examples that are drawn in these zones. The model (see VI-B) becomes very precise, on some area of the space. Figure 1 shows an experiment that is realized on a toy example (see VI-A), using only entropy to select interesting zones. Selected examples are grouped around the boundary, but there is a large part of the space that is not explored.

C. Exploration: Relative density

Relative density is the proportion of labelled examples among available examples in the considered zone. Equation 2 expresses relative density, with $\phi \subseteq \Phi$ the subset of observable examples that belong to a zone. As mixture rate, relative density varies in the interval $[0, 1]$.

$$RelativeDensity(l, \phi) = \frac{|l|}{|\phi|} \quad (2)$$

Relative density is the “exploration” term of the criterion. The homogeneity of drawn examples over the input space is ensured by choosing zones that have the lowest relative density. This strategy is different from a random sampling because homogeneity of drawn examples is forced. Figure 1 shows an experiment that is realized on the toy example, using relative density to select interesting zones. Input space partitioning and examples drawing are homogeneous.

D. Exploitation vs. Exploration Compromise

The criterion evaluates the interest of zones, taking into account both terms; mixture rate and relative density. Equation 3 shows how each term is used. The parameter $\alpha \in [0, 1]$ corresponds to a compromise between exploitation of already known mixture zones and exploration of new zones.

$$Interest(l, \phi) = (1 - \alpha) MixRate(l) + \alpha (1 - RelativeDensity(l, \phi)) \quad (3)$$

The notion of progress is included in the criterion: the relative density (that increases at the same time new examples are labelled) forces the algorithm to leave zones in which mixture rate does not increase quickly. If there is nothing else to discover in a zone, the criterion naturally avoids it. In some cases, the criterion prefers none mixed zones which are insufficiently explored. This criterion does not need a temporal window to evaluate the progress of sub-models (see section II-B). So its implementation is easier than original adaptive curiosity approach. Figure 3 shows an experiment that is realized on the toy example, using the criterion with $\alpha = \frac{1}{2}$. Input space partitioning and examples drawing are organized around the boundary considering every region of space.

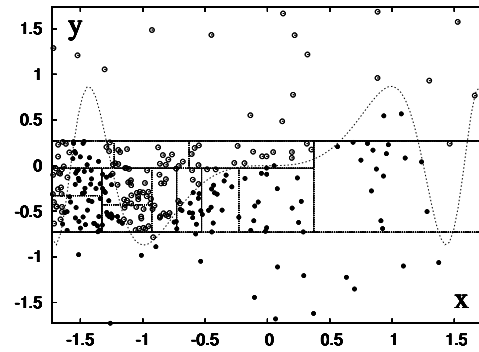


Fig. 1. Selected examples using Mixture Rate only in \mathbb{X} , with “o” points of first class, and “•” points of second class

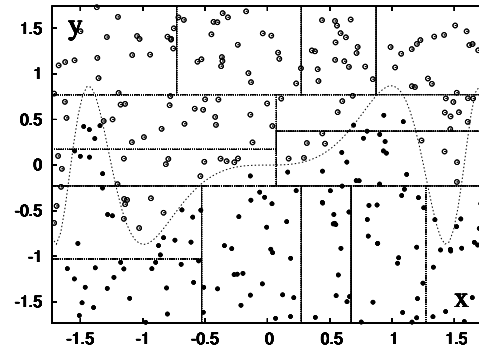


Fig. 2. Selected examples using Relative Density only in \mathbb{X} , with “o” points of first class, and “•” points of second class

Figure 4 shows performances (see VI-D) of the proposed strategy for various values of α . When $\alpha = 0$ only mixture rate is considered by the criterion. In this case, the observed performances are significantly lower than the

“stochastic” strategy considering less than 100 examples. This phenomenon can be intuitively interpreted by a strong exploitation of detected mixture zones, to the detriment of the remaining space. When $\alpha = 1$ only relative density is considered. In this case, adaptive curiosity gives lower performances than the “stochastic” strategy considering less than 70 examples. The best performances are observed for $\alpha = 0.25$. In this case, the maximum AUC is reached very early (with 60 labelled examples). Observed performances are superior to stochastic strategy for all considered number of learnt examples. This value obviously offers a good compromise between exploration and the exploitation.

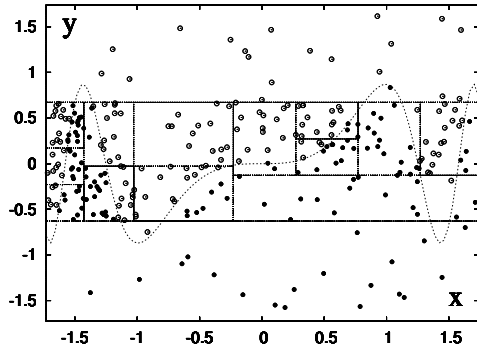


Fig. 3. Selected examples with $\alpha = 0.5$ in \mathbb{X} , with “o” points of first class, and “•” points of second class

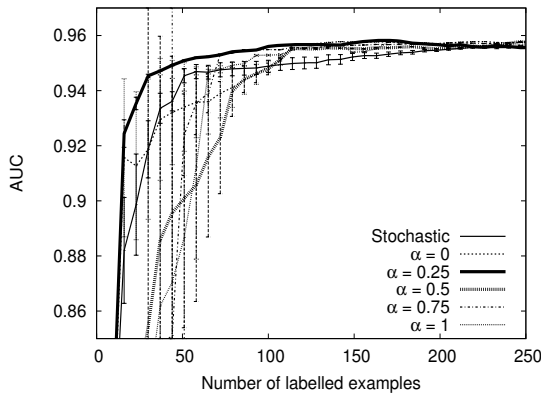


Fig. 4. AUC vs number of examples

These results show that adaptive curiosity can be beneficially used in active learning framework, with the proviso of using an adapted zones selection strategy. Moreover, the new strategy of zones selection is only based on data typology. Sub-models are only used to carry out the partitioning and not to choose interesting zones.

IV. APPLICATION TO EMOTION DETECTION

A. Introduction

Owing to recent techniques of speech processing, many automatic phone call centers appear. These vocal servers are used by customers to carry out various tasks conversing with a machine. Companies aim to improve their customer’s satisfaction by redirecting them towards a human operator, in the event of difficulty. The shunting of unsatisfied users is carried out detecting the negative emotions in their dialogues with the machine, under the assumption that a problem of dialogue generates a particular emotional state in the subject.

The detection of expressed emotions in speech is generally considered as a supervised learning problem. The detection of emotions is limited to a binary classification since taking into account more classes raises the problem of the objectivity of labelling task [7]. In this application, the acquisition and the labelling of data are costly. Active learning can reduce this cost by labelling only the examples considered to be informative for the predictive model.

B. Characterization of data

This study is based on a previous work [8] which characterizes vocal exchanges, in optimal way, for the classification of expressed emotions in speech. The objective is to control the dialogue between users and a vocal server. More precisely, this study deals with relevance of variables describing data, according to the detection of emotions.

The used data results from an experiment involving 32 users who test a stock exchange service implemented on a vocal server. According to the users point of view, the test consists in managing a virtual portfolio of stock options, the goal is to realize the strongest profit. The obtained vocal traces constitute the corpus of this study: 5496 “turns of speech” exchanged with the machine. Turns of speech are characterized by 200 acoustic variables, describing variations of the sound intensity, variations of voice height, frequency of elocution... Data is also characterized by 8 dialogical variables describing the rank of a turn of speech in a dialogue, the duration of the dialogue... Each turn of speech is manually labelled as containing positive (or neutral) or negative emotions.

The subset of the most informative variables with respect to the detection of expressed emotions in speech is given thanks to a naive Bayesian selector [5]. At the beginning of the selection of the most informative variables, the set of attributes is empty. At each iteration, the attribute that most improves the quality of the predictive model is added. The algorithm stops when the addition of attributes does not improve any more the quality of the model. Finally, 20 variables were selected to characterize vocal exchanges. In this article, used data comes from the same corpus from this previous study [8]. So, every turn of speech is characterized by 20 variables (see VI-G).

C. The choice of the model

Parameters that must be adjusted to use a model may represent a bias for measuring the contribution of a learning strategy. A Parzen window, with a Gaussian kernel [9], is used in experiments below since this predictive model uses a single parameter and is able to work with few examples. This model has been chosen to compare obtained results using adaptive curiosity and previous results [10] using classical active learning strategy. The “output” of this model is an estimate of the probability to observe the label y_j conditionally to the instance u :

$$\hat{P}(y_j|u) = \frac{\sum_{n=1}^N \mathbb{1}_{\{f(l_n)=y_j\}} K(u, l_n)}{\sum_{n=1}^N K(u, l_n)} \quad (4)$$

with

$$l_n, \in L_x \text{ et } u \in U_x \cup L_x$$

and

$$K(u, l_n) = e^{-\frac{\|u-l_n\|^2}{2\sigma^2}}$$

The optimal value ($\sigma^2=0.24$) of the kernel parameter was found thanks to a cross-validation using the whole of available training data [11]. Thereafter, this value is used to fix the Parzen window parameter. The obtained results (using all pieces of training data) are similar with the previous study that was realised with a naive Bayesian classifier [8]. Thus, Parzen windows are considered satisfying and valid for the following active learning procedures. Moreover, kernel methods and closer neighbour methods are usually employed in classification of expressed emotions in speech [12].

The model must be able to assign a label $\hat{f}(u)$ to an input data u , so a decision threshold noted $Th(L_x)$ is calculated at each iteration. This threshold maximizes the AUC of the model on the available training set. The predicted label is $\hat{f}(u_n) = 1$ if $\{\hat{P}(y_1|u_n) > Th(L_x)\}$, else $\hat{f}(u_n) = 0$. Since the single parameter of the Parzen window is fixed, the training stage is reduced to count instances “inside” the Gaussian kernel. In such conditions, strategies of examples selection are comparable without influence of the training of the model.

D. Used Active Learning strategies

The objective of this section is to compare adaptive curiosity with active learning strategies already described in the literature. Two alternating strategies are considered in this paper: uncertainty sampling and sampling by risk reduction.

Uncertainty sampling [13] is based on the confidence that the model has on its predictions. The used model must be able to produce an output and to estimate the relevance of its answers. In the case of the Parzen window, the confidence of a prediction is based on the estimated probability to observe the predicted class. More precisely, a prediction is considered as uncertain when the probability to observe the predicted class is weak. This strategy selects unlabelled examples that

maximize the uncertainty of the model. The uncertainty can be expressed as follows:

$$Uncertain(x) = \frac{1}{\argmax_{y_j \in \mathbb{Y}} \hat{P}(y_j|x)} \quad x \in \mathbb{X}$$

Sampling by risk reduction aims to reduce the generalization error, $E(\mathcal{M})$, of the model [14]. This strategy chooses examples that minimize this generalization error. In this paper, the generalization error ($E(\mathcal{M})$) is estimated using the empirical risk [15]:

$$\hat{E}(\mathcal{M}) = R(\mathcal{M}) = \sum_{i=1}^{|L|} \sum_{y_j \in \mathbb{Y}} \mathbb{1}_{\{f(x_i) \neq y_j\}} P(y_j|x_i) P(x_i)$$

where $f(x_i)$ is the predicted class of the instance x_i , $\mathbb{1}$ the indicating function equal to 1 if $f(x_i) \neq y_i$ and equal to 0 else, and $P(y_j|x_i)$ is the probability to observe the class y_j for the example $x_i \in L$. Therefore $R(\mathcal{M})$ is the sum of the probabilities that the model makes a bad decision on the training set (L). Using a uniform prior to estimate $P(x_i)$, one can write:

$$\hat{R}(\mathcal{M}) = \frac{1}{|L|} \sum_{i=1}^{|L|} \sum_{y_j \in \mathbb{Y}} \mathbb{1}_{\{f(x_i) \neq y_j\}} \hat{P}(y_j|x_i)$$

In order to select examples, the model is re-trained several times considering one more “potential” example. Each instance $x \in U$ and each label $y_j \in \mathbb{Y}$ can be associated to constitute the additional example. The expected risk of an example $x \in U$ that is added to the training set is then:

$$\hat{R}(\mathcal{M}^{+x}) = \sum_{y_j \in \mathbb{Y}} \hat{P}(y_j|x) \hat{R}(\mathcal{M}^{+(x,y_j)}) \quad \text{with } x \in U$$

E. Results

Several experiments were realised. Each experiment has been done five times² in order to obtain average performances provided with a variance. At the beginning of each experiment, the training set contains only two randomly chosen examples (one positive and one negative). At each iteration, ten examples are selected to be labelled and added to the training set. The considered classification problem is unbalanced: there are 92% of positive (or neutral) emotions and 8% of “negative” emotions. To observe correctly the classification profits when examples are labelled, the model is evaluated using the AUC (see VI-D) on the test examples set³.

For this real world problem no information to adjust parameters of adaptive curiosity is available, so we use $\alpha = 0.5$ as a default value. Because of the important size of Φ (1200 examples), the partitioning step is very long to be computed. So, the partitioning threshold increases to 100 examples in a zone. In such conditions, adaptive curiosity

²the natches on the curves of the figure 5 correspond to 4 times the variance of the results ($\pm 2\sigma$).

³The test set includes 1613 examples and the training set 3783 examples

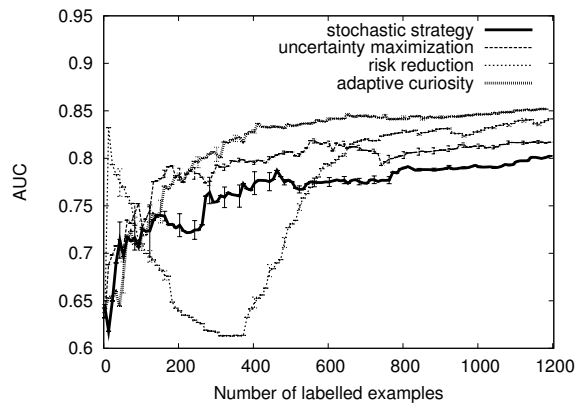


Fig. 5. Focus of the results on the test set using [0:1200] training examples

is the strategy that maximizes the quality of the predictive model. Adaptive curiosity is significantly better than the other strategies for a number of labelled examples in the range [80:1200]. Moreover the observed variance of the results is very low.

The two other active strategies are more difficult to differentiate. Between 100 and 700 labelled examples the uncertainty sampling wins, and beyond 700 labelled examples the sampling by risk reduction is better than the uncertainty sampling.

On this real problem, active strategies allow to obtain the optimal performance using fewer examples than the stochastic strategy. Adaptive curiosity reaches the optimal AUC (0.84) with only 500 examples. These results show adaptive curiosity is a competitive active learning strategy for detection of emotions in speech.

V. CONCLUSION

This paper shows adaptive curiosity can be used as an active learning strategy in machine learning framework. More precisely, adaptive curiosity seems to be very efficient for detection of emotions in speech.

Adaptive curiosity is a strategy that is not dependent of the predictive model. In this article, two different predictive models are used (a logistic regression in part III, a Parzen window in part IV). Consequently, this strategy can be applied on numerous real problems and is easy to use with existing systems.

We have defined a new zones' selection criterion that gives good results on the considered toy example and on emotions detection. However, this criterion balances exploitation and exploration using a parameter. Future works will be done to make the algorithm autonomous to adjust this parameter [16].

Adaptive curiosity was initially developed to deal with high dimensionality input spaces, where large parts are unlearnable or quasi-random. Future works will be realized to estimate the interest of our new criterion in such conditions.

The influence of the complexity of the problem to be learnt (that is to say, the number of examples necessary to solve it) will be also studied.

The partitioning step of adaptive curiosity has a $O(n^3)$ complexity and is prohibitive to treat high dimensionality datasets. Moreover, the cut criterion involves two parameters: the maximum number of labelled examples belonging to a zone, and the maximum balance rate of labelled examples subsets of a zone split. The use of non parametric discretization method [5] could be an efficient way to decide "when" and "where" a zone has to be split. This aspect will be considered in future works.

REFERENCES

- [1] R. White, "Motivation reconsidered: The concept of competence," *Psychological Review*, vol. 66, pp. 297–333.
- [2] P.-Y. Oudeyer and F. Kaplan, "Intelligent adaptive curiosity: a source of self-development," in *Proceedings of the 4th International Workshop on Epigenetic Robotics*, L. Berthouze, H. Kozima, C. G. Prince, G. Sandini, G. Stojanov, G. Metta, and C. Balkenius, Eds., vol. 117. Lund University Cognitive Studies, 2004, pp. 127–130.
- [3] A. Bondu and V. Lemaire, "Active learning using adaptive curiosity," in *International Conference on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*, 2007.
- [4] R. Castro, R. Willett, and R. Nowak, "Faster rate in regression via active learning," in *NIPS (Neural Information Processing Systems)*, Vancouver, 2005.
- [5] M. Boullé, "An enhanced selective naive bayes method with optimal discretization," in *Feature extraction, foundations and Application*, I. Guyon, S. Gunn, M. Nikravesh, and L. Zadeh, Eds. Springer, August 2006, pp. 499–507.
- [6] L. Breiman, "Technical note: Some properties of splitting criteria," *Machine Learning*, vol. 24, no. 1, pp. 41–47, July 1996.
- [7] J. Liscombe, G. Riccardi, and D. Hakkani-Tr, "Using context to improve emotion detection in spoken dialog systems," in *InterSpeech*, Lisbon, 2005.
- [8] B. Poulain, "Sélection de variables et modélisation d'expressions d'émotions dans des dialogues hommes-machine," in *EGC (Extraction et Gestion de Connaissance)*, Lille. + Technical Report available here: <http://perso.rd.francetelecom.fr/lemaire> (in french), 2006.
- [9] E. Parzen, "On estimation of a probability density function and mode," *Annals of Mathematical Statistics*, vol. 33, pp. 1065–1076, 1962.
- [10] A. Bondu, V. Lemaire, and B. Poulain, "Active learning strategies: a case study for detection of emotions in speech," in *ICDM' (Industrial Conference of Data Mining)*, Leipzig, july 2007.
- [11] O. Chappelle, "Active learning for parzen windows classifier," in *AI & Statistics*, Barbados, 2005, pp. 49–56.
- [12] V. Guide, Rakotomamonjy, and S. Canu, "Méthode noyaux pour l'identification d'émotion," in *RFIA (Reconnaissance des Formes et Intelligence Artificielle)*, 2003.
- [13] S. B. Thrun and K. Möller, "Active exploration in dynamic environments," in *Advances in Neural Information Processing Systems*, J. E. Moody, S. J. Hanson, and R. P. Lippmann, Eds., vol. 4. Morgan Kaufmann Publishers, Inc., 1992, pp. 531–538.
- [14] N. Roy and A. McCallum, "Toward optimal active learning through sampling estimation of error reduction," in *Proc. 18th International Conf. on Machine Learning*. Morgan Kaufmann, San Francisco, CA, 2001, pp. 441–448.
- [15] X. Zhu, J. Lafferty, and Z. Ghahramani, "Combining active learning and semi-supervised learning using gaussian fields and harmonic functions," in *ICML (International Conference on Machine Learning)*, Washington, 2003.
- [16] T. Osugi, D. Kun, and S. Scott, "Balancing exploration and exploitation: A new algorithm for active machine learning," in *Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM'05)*, 2005.
- [17] W. S. Sarle, "Neural networks and statistical models," in *Proceedings of the Nineteenth Annual SAS Users Group International Conference, April, 1994*. Cary, NC: SAS Institute, 1994, pp. 1538–1550.

A. Toy example

The toy example is a binary classification problem in a two dimensional space $\mathbb{X} = x \times y$. We consider two classes that are separated by the boundary $y = \sin(x^3)$, on intervals $x \in [-2, 2]$ and $y \in [-2, 2]$. 2000 training examples were used (Φ) and 30000 test examples both uniformly generated over the space \mathbb{X} .

B. Used model for the toy example

A logistic regression implemented by a neural network is used [17]. The outputs of this model are normalized by a soft max function in the interval $[0, 1]$. Outputs correspond to probabilities of observing classes, conditionally to the instance that is placed as input of the model. Neural network's training is stopped when the training error does not decrease more than 10^{-8} , and the training step is fixed to 10^{-2} . Logistic regression is used as a global model that is trained independently of the input space partitioning, using examples that are selected by sub-models. Sub-models play only a role in the selection of interesting zones and in the selection of instances to be labelled. A global model is trained using these examples. The global model allows making a coherent comparison between adaptive curiosity and others strategies that handle a single model. Performances of the global model report only the quality of selected examples.

C. Partitioning

Zones containing at least 30 labelled examples are split. A cut separates labelled examples into two $\pm 25\%$ balanced subsets (according to the criterion of section II-B). These arbitrary choices are preserved for all experiments in this paper.

D. Measure of performances

ROC curves plot the rate of good predictions against the rate of bad predictions on a two dimensional space. These curves are built sorting instances of test set according to the output of the model. ROC curves are usually built considering a single class. Consequently, $|\mathbb{Y}|$ ROC curves are considered. AUC is computed for each ROC curve, and the global performance of the model is estimated by the mathematical expected value of AUC, over all classes: $AUC_{global} = \sum_{i=1}^{|\mathbb{Y}|} P(y_i) \cdot AUC(y_i)$

E. Protocol

Beforehand, data is normalized using mean and variance. At the beginning of experiments, the training set contains only two labelled examples which are randomly chosen among available data. At every iteration, a single example is drawn in the current zone to be labelled and added to the training set. Active learning stops when 250 examples are labelled.

F. Stochastic strategy

The "stochastic" strategy handles a global model and uniformly selects examples according to their probability distribution. This strategy plays a role of reference and is used to measure the contribution of adaptive curiosity.

G. data of emotion detection

This part enumerates the 20 variables which characterize vocal exchanges in emotion detection problem.

- 1) System shut down (the user closes the dialog)
- 2) Number of words of the current turn of speech
- 3) The user comments the dialog
- 4) Number of errors on the current task
- 5) Total number of errors on nested tasks
- 6) Increase of the signal intensity
- 7) Decrease of the signal intensity
- 8) Maximum coefficient of the first harmonic of the signal (Fourier transform)
- 9) Average of the distribution of voice's timbre variation
- 10) Maximum value of standard variance of voice's timbre variation
- 11) Standard variance of voice's timbre variation
- 12) Average of the distribution of power of high-frequency / low frequency ratio.
- 13) Standard variance of signal energy
- 14) Sum of standard variance of signal energy
- 15) Maximum value of standard variance of signal energy
- 16) Derivative of signal energy
- 17) Jitter of signal energy
- 18) Complete reformulation of the previous turn of speech
- 19) Complete repetition of the previous turn of speech
- 20) Partial repetition of the previous turn of speech