Selective Sampling Based on the Variation in Label Assignments

Piotr Juszczak, Robert P.W. Duin
Information and Communication Theory Group,
Faculty of Electrical Engineering, Mathematics and Computer Science,
Delft University of Technology, The Netherlands
p.juszczak@ewi.tudelft.nl, r.p.w.duin@ewi.tudelft.nl

Abstract

In this paper, a new selective sampling method for the active learning framework is presented. Initially, a small training set T and a large unlabeled set Ω are given. The goal is to select, one by one, the most informative objects from Ω such that, after labeling by an expert, they will guarantee the best improvement in the classifier performance. Our sampling strategy relies on measuring the variation in label assignments (of the unlabeled set) between the classifier trained on T and the classifiers trained on T with a single unlabeled object added with all possible labels.

We compare the performance of our algorithm with two traditional procedures random sampling and uncertainty sampling. We show empirically across a range of datasets that the proposed selective sampling method decreases the number of labeled instances needed to achieve the desired error for the fixed size of T. Experimental results on toy problems and the UCI datasets are presented.

1. Assume that a small number of objects with the true labels is given, constituting the initial training set T.
2. Train a specified classifier ω on the training set T.
3. Select a number of objects from the unlabeled data Ω according to the chosen criterion E.
4. Ask an expert for the labels of these objects and add them to the training set T.
5. Repeat the steps 2-4 until the stopping criterion, e.g. specified by the final size of T.

Figure 1. Active learning.

1. Introduction

In many problems, a small labeled training set T and a large unlabeled dataset Ω are available. For instance, the classification of webpages, mp3-s or images. Semi-supervised learning methods use both labeled and unlabeled data e.g. to maximize the class posterior probabilities [3]. On the contrary, active learning offers a possibility to select which data points are added to the training set. This is useful if labeling of data points is expensive; see Fig. 1 for a general framework. The desired behavior of the performance of an active learner during the sampling process is shown in Fig. 2. The abscissas represents the information needed for learning, i.e. the number of labeled examples used for training, and the ordinate represents the error rate of a current classifier on an independent test set. Ideally, in every phase of the selective sampling process only the most informative patterns from the unlabeled set Ω should be chosen. These are the objects, which after adding them with the true labels, to the training set T and retraining the classifier, they will lower the true error the most.

In the field of machine learning several selective sampling techniques were introduced, such as sampling from regions in the instance space, where no data are present [12] or which yield low confidence [11]: *uncertainty sampling* [6] and *query-by-committee* [10], sampling from regions where the classifier performs poorly [7], or where the previously found data was used in learning [9].

Most of these methods sample in the vicinity of the current decision boundary. This is just appropriate for a classifier which is well established in the instance space. For a poor classifier the mentioned methods will allow only for small improvement steps in the classifier performance for growing number of samples added to T. Consequently, to reach a desired test error, longer queries are required than in case when the most informative patterns are added. To solve this problem, Cohn [2] proposed an active learning approach to the statistical learning models by focusing on these examples in Ω that change the most the second order statistics e.g. means and variances. This was developed for classifiers like the Mixture of Gaussians. Based on the ideas of Cohn, Roy and McCallum [8] introduced another selective sampling criterion for classifiers which are not based on the second order statistics. They suggested to use an additional test set in the selection process. In their strategy, the examples from Ω with all possible labels are considered and the

0-7695-2128-2/04 $20.00 (C) 2004 IEEE

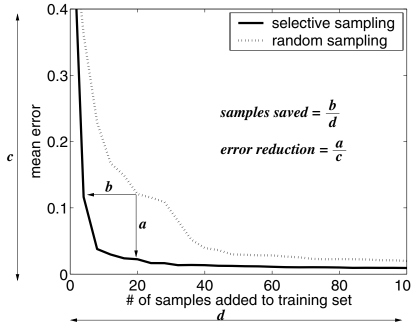

Figure 2. Learning curves for the selective sampling and random sampling.

ones are chosen which yield the lowest error on the test set. In this paper, we argue that the use of a separate labeled test set is in fact in conflict with an assumption of the selective sampling problem, namely that the labeling is expensive and the training should be based on a minimal set of objects. Based on the methods of Cohn and Roy, we propose a selection criterion which makes use of the variation in label assignments by a single classifier ω. On the contrary, the variation is measured as a difference of classification labels of the unlabeled dataset $[\Omega/\{x_i\}]$ between the classifier ω trained on the labeled data T and ω_i^j trained on $[T \cup \{x_i\}]$ where for all x_i all possible labels j are considered that exist in the learning problem.

In comparison to the methods of Roy and McCallum, our technique has the advantage of not using a test set. Instead of making the selection criterion being based on the error measured on the test set, we rather measure the expected change in the architecture of the classifier itself as measured on Ω. Our selection criterion is designed to be advantageous for a weak classifier.

In the following sections we address the problem of the selection of learning queries knowing that the initial classifier is weak. We compare our method to the traditional ones: *uncertainty sampling* and *random sampling*. *Uncertainty sampling* relies on the distance to a classifier estimated from the posterior probabilities. It selects samples that yield the minimum distance. *Random sampling* samples unlabeled dataset according to the data distribution.

2. Selective sampling based on label variation

Assume that for a C-class problem, an initial training set T and an unlabeled set Ω are given. Our selective sampling technique relies on the variation in label assignments for the unlabeled dataset. The true error of the classifier (hence on a test set) is expected to be reduced by selecting those objects which cause the largest change of the classifier. For a particular object $x_i \in \Omega$, this is measured by the statistic E:

```
1. Train the classifier $\omega$ on the labeled set $T$.
2. Train the classifier $\omega_i^j$ on $T_i = [T \cup \{x_i\}]$ with the label $j$ for $x_i$.
3. Let $\Omega_{-i} = [\Omega/\{x_i\}]$. Label $\Omega_{-i}$ by $\omega$; $L_{-i}^{\omega} \leftarrow \omega(\Omega_{-i})$.
4. Label $\Omega_{-i}$ by $\omega_i^j$; $L_{-i}^{\omega_i^j} \leftarrow \omega_i^j(\Omega_{-i})$.
5. Repeat the steps 2-4 for all labels $j \in C$. Compute $E(x_i)$ or $E'(x_i)$.
6. Repeat 2-5 for all $x_i \in \Omega$.
7. Select $x_k$ to be labeled by the expert such that $x_k = arg\max[E(x_i)]$.
```

Figure 3. The variation algorithm

$$E(x_i) = \sum_{j=1}^{C} P_j(x_i)\Phi_j, \qquad (1)$$

where $P_j(x_i)$ is the posterior probability that object x_i belongs to the j-th class ($j = 1 \ldots C$), estimated from the classifier ω, trained on the labeled set T. ω_i^j is a classifier trained on $T_i = [T \cup \{x_i\}]$, where x_i has been assigned to the j-th class. Classifiers ω and the set of ω_i^j are applied to $\Omega_{-i} = [\Omega/\{x_i\}]$ yielding the label sets L_{-i}^{ω} and $L_{-i}^{\omega_i^j}$, respectively. Φ_j in the equation (1) measures the number of objects in Ω_{-i} which have different class memberships for the two classifiers. Hence,

$\Phi_j = \sum_{x \in \Omega_{-i}} \mathcal{I}(L_{-i}^{\omega}(x) \neq L_{-i}^{\omega_i^j}(x))$ where $\mathcal{I}(a)$ takes the value 1 if the condition a is true and 0, otherwise. $L_{-i}^{\omega}(x)$ and $L_{-i}^{\omega_i^j}(x)$ stand for the labels of a single object x.

Instead of using 'hard' $0/1$ for Φ_j in the equation (1), we can replaced it by probabilities. If the object x changes its class membership to the class c, the posterior probability $P_c(x)$ that x belongs to that class is included. These probabilities are estimated by the classifier ω. Then, we have:

$$\Phi'_j = \sum_{x \in \Omega_{-i}} P_c(x)\, \mathcal{I}(L_{-i}^{\omega}(x) \neq L_{-i}^{\omega_i^j}(x)) \qquad (2)$$

$$E'(x_i) = \sum_{j=1}^{C} P_j(x_i)\Phi'_j \qquad (3)$$

Equations (1) and (3) are used inside the active learning framework, phase 3 in Fig. 1, to calculate the expected change in the classifier ω over the example space Ω. The algorithm is presented in Fig. 3.

3. Experiments

In this section we compare the performance of the proposed selective sampling method (variation in labels) to *uncertainty sampling* and *random sampling*, using equation (3) as the selective sampling criterion. In all the experiments, the Parzen classifier was used with the smoothing parameter optimized by the leave-one-out approach as introduced by Duin [4]. For simplicity, we

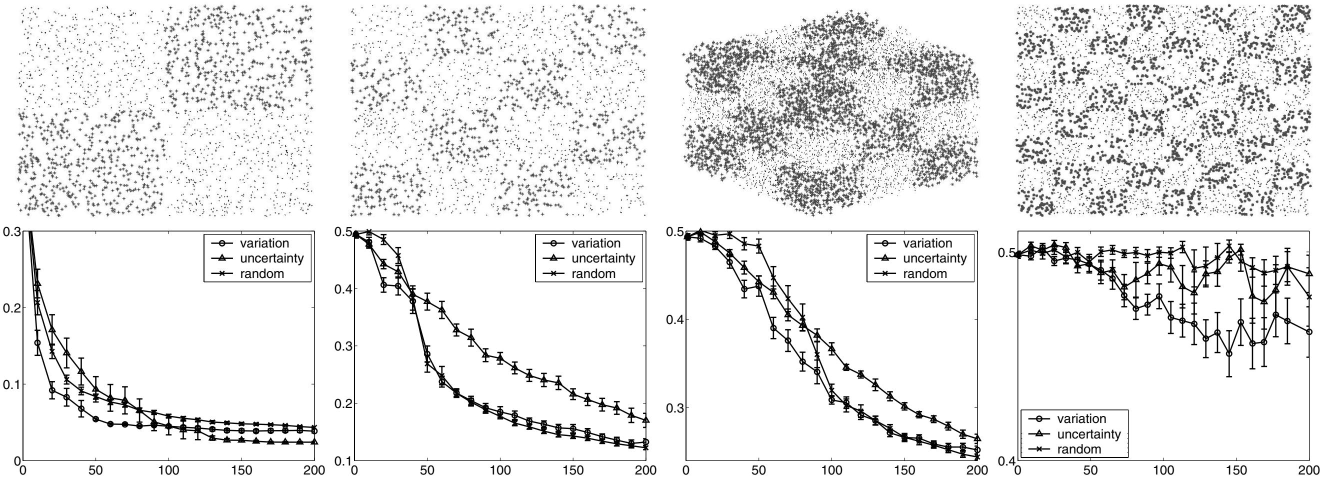

Figure 4. Learning curves for the chess board datasets with different levels of multi-modality.

consider two-class problems only. Two-class toy problems of different multi-modality (see Fig. 4) and some datasets from the UCI repository [1] were used to verify the performance of the selective sampling methods. In Tab. 1 the first three datasets are simple two-class problems, while the following five datasets are converted to the two-class problems by assigning the first half of the classes to one class and the rest to the other class. This introduces multi-modality in data. Due to space limits, for the real-world data (see Tab. 1), instead of presenting the learning curves we compute three statistics (see Fig. 2) on them :

$$\text{below random} = \tfrac{1}{N}\textstyle\sum_{i=1}^{N}\mathcal{I}[f_s(x_i) < f_r(x_i)]$$

$$\text{samples saved} = \tfrac{1}{|\Omega_{10\%}||A|}\textstyle\sum_{err\in A}[f_r^{-1}(err) - f_s^{-1}(err)]$$

$$\text{error reduction} = \tfrac{1}{err_0\cdot N}\textstyle\sum_{i=1}^{N}[f_r(x_i) - f_s(x_i)]$$

Where f_r and f_s are the values of the learning curves for random and selective samplings respectively. N is the size of the complete query $\{x_1, x_2 \ldots x_N\}$ and $A = \{0 : 0.01 : 0.5\}$. The results are based on the size of the query, set to 10% of the original cardinality of Ω. Additionally, the percentage of the points on the learning curves for the selective sampling methods below the relevant points on the learning curve for the random method is computed (below random). The experiments were designed as follows: the datasets were split into three subsets: a small initial training set (two objects per class), an unlabeled set Ω and a test set, equally split from the remaining objects. In each sampling phase, a single object from the unlabeled set Ω was selected according to the specified selection criteria and included with the correct label in the training set. The performance of the Parzen classifier ω was measured on an independent test set. The results were averaged over 10 randomly chosen initial training sets, unlabeled sets and test sets.

From the results shown in Fig. 4 it can be seen that the number of queries up to which our method outperforms the uncertainty sampling and depends on the complexity of a problem. For the first problem in Fig. 4, it is about 75 queries, for more complex problems, it is larger than 200. The same pattern can be seen in Tab. 1. For single-mode datasets uncertainty sampling outperforms our method for the first 10% queries. However, if the problem is complex e.g. by introducing multi-modality, the sampling method based on variation in labels tends to outperform the uncertainty sampling. The decision which selective sampling should be used can be based e.g. on the number of queries that was correctly labeled by the current classifier so far [5]. If the unlabeled dataset is very large, many researchers use a subset of randomly sampled examples as a substitute for the complete set Ω. This will reduce the computational burden, but it will not ensure that the chosen subset consists of potentially informative patterns, even more, the most informative patterns might be disregarded from the analysis. In such a case, we propose to cluster the unlabeled data before computing any active learning statistic. The compactness hypothesis states that similar objects are close in their representation space. Therefore, the computation of the active learning statistic for queries consisting of more than one object will cause the selection of objects containing similar information. Therefore, the selection of the cluster centers as potential queries can be beneficial. It will reduce the unlabeled set to a subset of potentially the most informative patterns and, moreover, it will penalize the selection of objects that contain similar information.

4. Conclusions

Many selective sampling methods are designed to improve classifiers that are well established in the instance space. We showed that in such cases the methods like *uncertainty sam-*

	dataset	size/dim	sampling method	below random [%]	samples saved [%]	error reduction [%]
single-mode	HEART	297/13	variation	7	-40(10)	-15(28)
			uncertainty	30	-14(32)	-2(25)
	DIABETES	768/8	variation	49	-1(58)	0.9(26)
			uncertainty	76	14(58)	7(25)
	BREAST	699/10	variation	72	9(53)	2(10)
			uncertainty	67	16(71)	2(10)
multi-modal problems	MFEAT-FOU	2000/76	variation	83	17(47)	2(5)
			uncertainty	31	-6(20)	-3(-4)
	MFEAT-ZER	2000/53	variation	90	12(34)	2(5)
			uncertainty	86	10(26)	-3(4)
	CBANDS	2000/30	variation	96	20(40)	6(9)
			uncertainty	97	14(36)	4(6)
	IONOSPHERE	2000/5	variation	100	29(64)	6(44)
			uncertainty	87	18(47)	1(8)
	SATELLITE	2000/36	variation	100	37(74)	6(48)
			uncertainty	100	35(78)	5(41)

Table 1. Results for the UCI datasets averaged over 10 trials. The maximum (in brackets) and the average value on 10% of Ω of some of statistics (see text and Fig. 2) on learning curves are presented.

pling which samples in the neighborhood of the the current decision boundary performs the best. However, when the dataset is multi-modal or when only a small initial training set is provided, our proposed method, based on the variations in label assignments, (measured by the number of objects that change their labels in the unlabeled dataset), outperforms the uncertainty and random samplings.

To chose a selective sampling method for a given problem without using an independent test set, the following procedure has been proposed. First, it is observed how many labels of the subsequent queries are guessed by the current classifier ω. If the initial classifier is far from the optimal one and it makes many errors on the incoming data, the selective sampling technique to be used, should 'swap' the instance space (e.g. by using our selective sampling method) instead of trying to sample in the vicinity of a classifier. Next, after subsequent correct guesses of the labels of the selected queries, the selective sampling method can be replaced e.g. by the *uncertainty sampling* for the final tuning of the classifier.

To reduce the computational cost, it has been proposed to cluster the unlabeled dataset Ω. The random reduction of Ω, usually applied for this purpose, can remove the most informative patterns from the considered set. On the contrary, clustering focuses on the natural structure being present in data. In addition, by taking the cluster centers as potential queries the selection of queries containing the same information are disregard when a single query consists more than a single object.

Acknowledgments

This work was partly supported by the Dutch Organization for Scientific Research (NWO).

References

[1] C. Blake and C. J. Merz. UCI repository of machine learning databases, 1998.

[2] D. Cohn, Z. Ghahramani, and M. Jordan. Active learning with statistical models. In *ANIS*, 1995.

[3] F. Cozman, I. Cohen, and M. Cirelo. Semi-supervised learning of mixture models. In *ICML*, 2003.

[4] R. P. W. Duin. On the choice of the smoothing parameters for parzen estimators of probability density functions. *IEEE Trans. on Comp.*, 1976.

[5] P. Juszczak and R. P. W. Duin. Selective sampling methods in one-class classification problems. In *ICANN*, 2003.

[6] D. Lewis and W. Gale. A sequential algorithm for training text classifiers. In *SIGIR*, 1994.

[7] A. Linden and F. Weber. Implementing inner drive by competence reflection. In *ICSAB*, 1993.

[8] N. Roy and A. McCallum. Toward optimal active learning through sampling estimation of error reduction. In *ICML*, 2001.

[9] J. Schmidhuber and J. Storck. Reinforcement driven information acquisition in non-deterministic environment. In *ICANN*, 1995.

[10] H. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *COLT*, 1992.

[11] S. Thrun and K. Möller. Active exploration in dynamic environments. In *ANIS*, 1992.

[12] S. Whitehead. A study of cooperative mechanisms for faster reinforcement learning. In *TR Uni. of Rochester*, 1991.