Uncertainty sampling methods for one-class classifiers

Piotr Juszczak PIOTR@PH.TN.TUDELFT.NL
Pattern Recognition Group, Delft University of Technology, Lorenzweg 1, 2628 CJ Delft, The Netherlands

Robert P.W. Duin BOB@PH.TN.TUDELFT.NL
Pattern Recognition Group, Delft University of Technology, Lorenzweg 1, 2628 CJ Delft, The Netherlands

Abstract

Selective sampling, a part of the active learning method, reduces the cost of labeling supplementary training data by asking for the labels only of the most informative, unlabeled examples. This additional information added to an initial, randomly chosen training set is expected to improve the generalization performance of a learning machine. We investigate some methods for a selection of the most informative examples in the context of one-class classification problems (OCC) i.e. problems where only (or nearly only) the examples of the so-called target class are available. We applied selective sampling algorithms to a variety of domains, including real-world problems: mine detection and texture segmentation. The goal of this paper is to show why the best or most often used selective sampling methods for two- or multi-class problems are not necessarily the best ones for the one-class classification problem. By modifying the sampling methods, we present a way of selecting a small subset from the unlabeled data to be presented to an expert for labeling such that the performance of the retrained one-class classifier is significantly improved.

1. Introduction

In many classification problems, a large number of unlabeled examples may be available in addition to a small training set. To benefit from such examples, one usually exploits either implicitly or explicitly the link between the marginal density $P(x)$ over the examples of a class x and the conditional density $P(y|x)$ representing the decision boundary for the label y. For example, high density regions or clusters in the data can be expected to fall solely in one or another class. One technique to exploit the marginal density $P(x)$ between classes is selective sampling, which is a part of the active learning method (Cohn 1996). In this technique the performance of classifiers is improved by adding supplementary information to a training set. In general, there is a small set of labeled data and a large set of unlabeled data. In addition, there exists a possibility of asking an expert (oracle) for labeling additional data. However, this may not be used excessively e.g. for economic reasons. The question is: how to select an additional subset of unlabeled data such that after labeling and including it in the training set the performance of a particular classifier improves the most. These examples are called the most informative patterns. Many methods of selective sampling have already been considered in two- or multi-class problems. They select objects:

- which are close to the decision boundary (Cohn 1992) e.g. close to a margin or inside a margin for the support vector classifier (Cambell),
- or which have the most evenly split labels over a variation of classifiers:
  - trained on multiple permutations of the labeled data (Warmuth),
  - differing by the settings,
  - trained on independent sets of features (Muslea).
- or that reduce the size of the version space (Mitchell, Tong and Koller)

These sampling methods are looking for the most informative patterns in the vicinity of a current classifier. It means they select patterns, to be labeled by an oracle, which have a high probability of incorrect classifi-

Workshop on Learning from Imbalanced Datasets II, ICML, Washington DC, 2003.

cation if they are not included in the training set. The classification performance is improved in small steps. In this paper, we will test a number of selective sampling methods for several one-class classification problems (De Ridder 1998, Japkowicz 1999, Tax 2001).

In the problem of one-class classification, the goal is to accurately describe one class of objects, called the target class, as opposed to a wide range of other objects which are not of interest, called outliers in this paper. Many standard pattern recognition methods are not well equipped to handle this type of problem; they require complete descriptions for both classes. Especially when one class is very diverse and ill-sampled, normal (two-class) classifiers obtain very bad generalization for this class.

The problem of one-class classification is harder than the standard two-class classification problem. In two-class classification, when examples of outliers and targets are both available, a decision boundary is supported from both sides by examples of each of the classes. Because in case of one-class classification only the target class is available, just one side of the boundary is supported. Based on the examples of one class only, it is hard to decide how tight the boundary should fit around the target class.

The absence of outlier examples makes it also very hard to estimate the classification error. The error of the first kind $\mathcal{E}_{\mathcal{I}}$, referring to the target objects that are classified as outlier objects, can be estimated on the available data. However, the error of the second kind $\mathcal{E}_{\mathcal{II}}$ referring to the outlier objects that are classified as target objects, cannot be estimated without assumptions on the distribution of the outliers. If no information on the outlier class is given we assume a uniform distribution of the outliers.

Figure 1 illustrates for a multi-class problem the difference between discrimination by multi-class classification and description by a one-class approach. The first solution to the problem divides the entire data space and assigns each of its parts to the particular class. The second one assigns a new data point only to the particular class if it is in one of the described regions:

- in the discriminant approach a new object x_i **has to** be assigned to one of the classes being present in the training set.
- in the description approach if a new object x_i is not inside a region described by the target class it is be assigned to a not-recognized class, called the outlier class.

In this paper, we will show that the standard selective sampling methods for multi-class problems, which look in the vicinity of the classifier, do not perform well in a one-class classification problem. To justify this, a distance measure to the description boundary defined by the classification confidence (called also uncertainty sampling (Lewis and Gale)), will be used.

The layout of this paper is as follows: in the section 2, the selective sampling techniques will be introduced. In the next section we will show some results of uncertainty sampling for several one-class classifiers on an artificially created problem. We will then go on to show some results on a real-world mine detection problem and discuss the relative merits and disadvantages of the uncertainty sampling methods.

2. A formal framework

In selective sampling algorithms the challenge is to determine which unlabeled examples will be the most informative (e.g. improve the classification performance the most) if they were labeled and added into an existing training set. These are the examples which are presented as a query to an oracle - an expert who can label any new data without error. We begin with a preliminary, weak classifier that has to be first determined by a small set of labeled samples. In particular, in selective sampling algorithms, mentioned in section 1, the distributions of query patterns will be dense near the final decision boundaries (where examples are informative) rather than at the region of the highest prior probabilities (where patterns are typically less informative). At the beginning, the training set consists of a few randomly selected samples. To reach the desired classification error, we would like to add as few as possible new examples (labeled by the expert) from the unlabeled data using a selective sampling method 1. If the sampling method selects patterns close to the boundary given by the current classifier, then the probability of an incorrect classification is higher for such examples than for examples being far from the description boundary. This approach was proved to work for several multi-class problems (Blum 1998, Cambell, Cohn 1992, Freund 1997).

Because it is usually not possible to compute the distance between a pattern and a nonlinear classifier, we propose to base this distance measure on the raw output of a classifier $y(x)$, where $y(x) \in (-\infty, +\infty)$ and:

$$y(x) < 0 \text{ for objects classified as outliers}$$

$$y(x) \geq 0 \text{ for objects classified as targets}$$

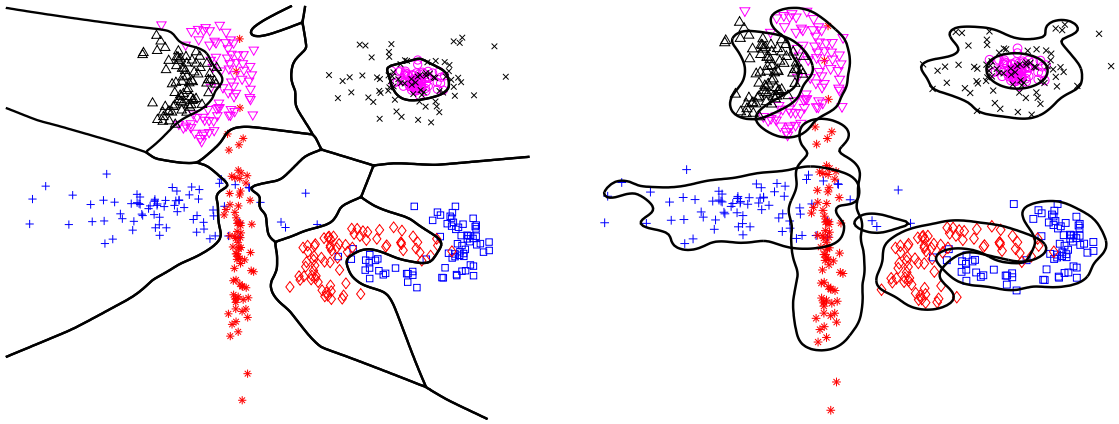

Figure 1. The multi-class problem solved by discriminant, multi-class support vector classifier (left) and by description, one-class support vector classifier (right)

Table 1. Active learning with selective sampling - The algorithm

---

```
1. assume that a small number of the target
objects with true labels is given
constituting an initial training set

2. train a specified classifier on the
training set

3. select a number of objects classified
as targets and outliers according to the
chosen selective sampling method

4. ask an oracle for labels of these
objects and include them in the training
set

5. repeat the steps 2-4 or STOP if e.g. the
training set is larger than a specified
size
```

$y(x)$ is converted to a relative confidence Γ_y^c indicating that object x belongs to class c assigned by the classifier to one of the classes (target or outlier):

The confidence Γ_y^c is computed as follows:

$$\Gamma_y^c = \frac{f^c(y)}{\sum_{x \in c} f^c(y)}$$

Where c indicates either a target (t) or an outlier (o) class assigned by the classifier, $f^t(y) = \frac{1}{1+e^{(-y)}}$ for objects classified as targets and $f^o(y) = \frac{1}{1+e^y}$ for objects classified as outliers.

$$\sum_{x \in c}(\Gamma_y^c) = 1;^1 \qquad 0 \leq \Gamma_y^c \leq 1$$

For objects classified as targets only the confidences Γ_y^t are computed, for objects classified as outliers only the confidences Γ_y^o are computed.

There are two interesting types of regions considering the classification confidences:

1. high confidence regions, defined by the objects far from the decision boundary for which Γ_y^c has a high value
2. low confidence regions, defined by the objects close to the decision boundary for which Γ_y^c has a low value

Based on the confidence regions of a classifier, we can describe four selective sampling methods that choose

[1] If $f^c(y) > 0$, then x is assigned to the class c. So, the confidences of all objects, within a class (as classified by the actual classifier) sum to one. We realize that this is a nonstandard way of using the 'confidence' concept.

an additional set of examples (e.g. 5 from each target/outlier class) for an oracle to be labeled:

ll - a low confidence for both the target and the outlier classes. This method is an approximation of the standard selective sampling methods used in multi-class problems because it samples from the vicinity of the current classifier.

lh - a low confidence for the target and a high confidence for the outlier class.

hl - a high confidence for the target and a low confidence for the outlier class.

hh - a high confidence for both the target and the outlier class.

We compare these sampling techniques with the two methods that are not dependent on the classification confidence:

hr - a half-random method, which first classifies the unlabeled set of examples and then selects randomly an equal number of examples from each of the two classification sets $rand(x \in t)$ and $rand(x \in o)$. This method selects objects based just on the classification labels; the classification confidences Γ_x^c are not considered during the selection process.

ra - a random selective sampling method, $rand(x \in t \vee o)$. In this method the classification labels as well as the confidences are not considered during the selection process.

To avoid the selection of patterns being 'really far' from the current description boundary we will assume that the class examples: targets and outliers, in the one-class classification problem are bounded by a box. In our experiments with the artificial data, the lengths of the bounding box edges are set up to 10 times the feature ranges of the initial training set.

The artificial data used in experiments: the target class contains merged, normally distributed clouds; see figure 2:

$N([1\ 1\ 1],\ [4\ 0\ 0;\ 0\ 0.5\ 0;\ 0\ 0\ 0.5])$
$N([2\ 0\ 1],\ [0.01\ 0\ 0;\ 0\ 8\ 0;\ 0\ 0\ 0.01])$
$N([10\ 2\ 1],\ [0\ 0.5\ 0;\ 4\ 0\ 0;\ 0\ 0\ 0.5])$
$N([10\ 1\ 1],\ [0\ 8\ 0;\ 0.01\ 0\ 0;\ 0\ 0\ 0.01])$
$N([20\ 10\ 5],\ [0\ 0\ 0.5;\ 4\ 0\ 0;\ 0\ 0.5\ 0])$
$N([20\ 10\ 5],\ [0\ 0\ 0.01;\ 0.01\ 0\ 0;\ 08\ 0])$

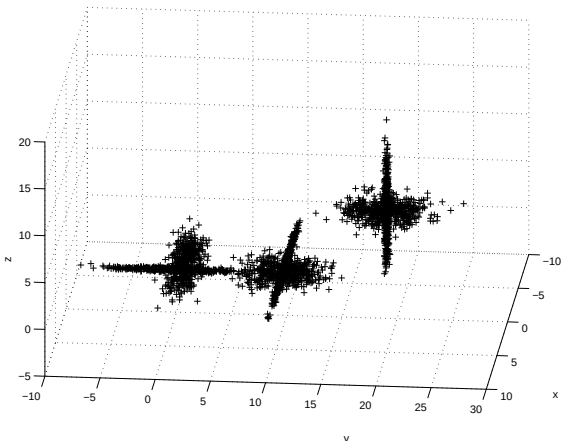

Figure 2. The artificial created target set used in experiments

As the outlier class, we considered objects uniformly distributed in the bounding box with 5% overlap with the target class.

To see how well a classifier fits the data both errors $\mathcal{E}_{\mathcal{I}}$ and $\mathcal{E}_{\mathcal{II}}$ should be considered. Because the initial training set contains a small set of the target objects at the beginning $\mathcal{E}_{\mathcal{I}}$ is high and $\mathcal{E}_{\mathcal{II}}$ is relatively low. The correct selective sampling methods chosen for the particular classifier should reduce the error $\mathcal{E}_{\mathcal{I}}$ and should not increase $\mathcal{E}_{\mathcal{II}}$ at the same time. In section 4 with the results on real-world data, for clarity we present just the result for target class. The error for outlier class was just slightly increasing like in examples with an artificial data.

3. Experiments with the artificial data

Now we will present the results of experiments performed on the $3D$ artificially created classes, using the uncertainty selective sampling methods described in section 2. A number of different one-class classifiers is taken into account (Tax 2001): Support Vector Data Description(SVDD), Autoencoder Neural Network(ANN) and the Parzen classifier. The dataset contains 3000 target objects and 7000 outlier objects chosen in the bounding box. At the beginning, we randomly select 6 patterns from the target class and train a classifier. First, in every sampling step, 5 objects currently classified as targets and 5 objects currently classified as outliers are chosen according to the selective sampling method. Next, the true object labels are retrieved and the classifier is retrained. The error of the first kind $\mathcal{E}_{\mathcal{I}}$ for all the classifiers is set to 0.1 on the training set. The size of the bounding box equals 10. In Table 3 the averaged results over 10 runs are

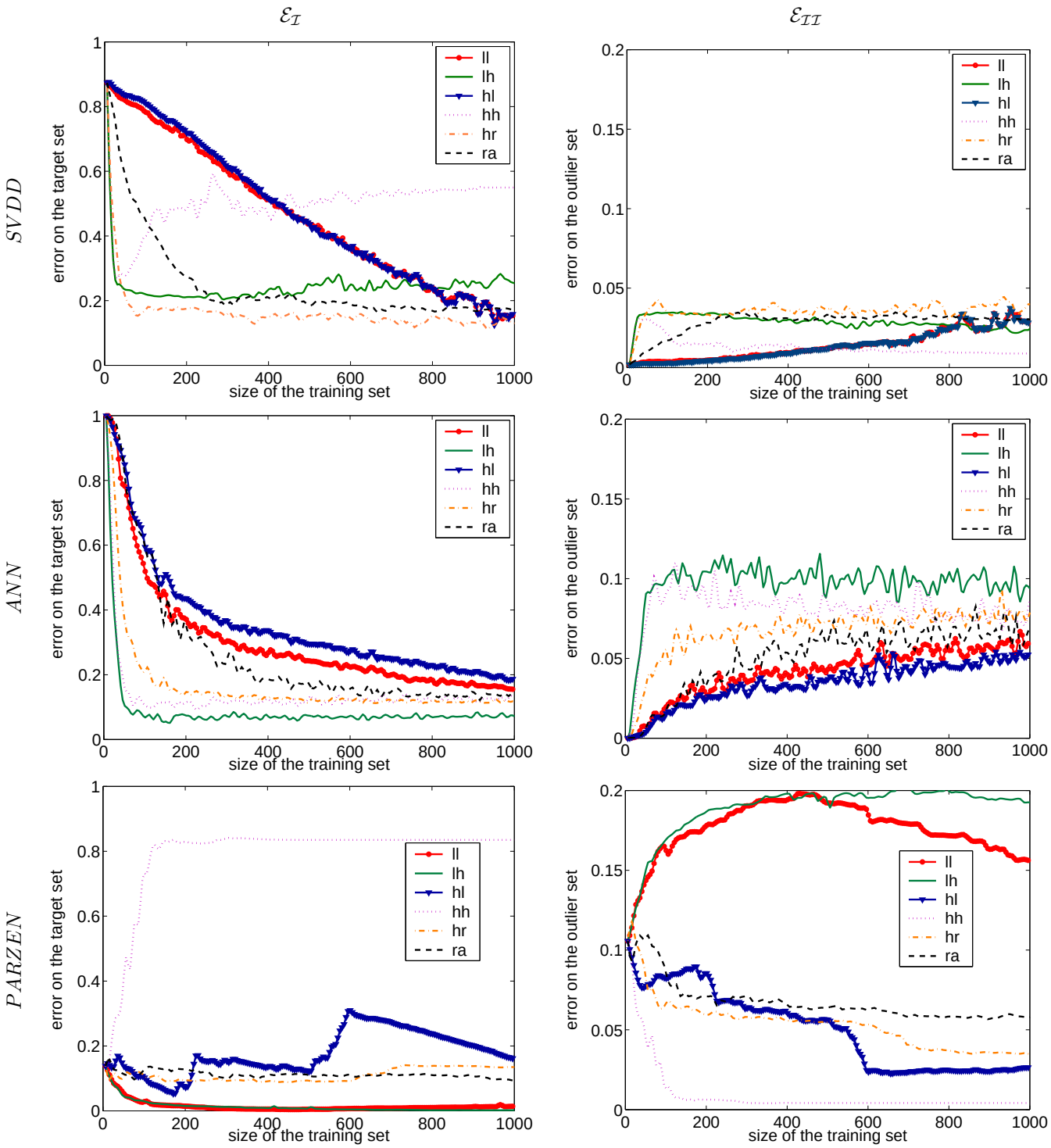

Table 2. The classification error $\mathcal{E}_{\mathcal{I}}$ and $\mathcal{E}_{\mathcal{II}}$ for SVDD, autoencoder (ANN) and Parzen classifier, on the artificial dataset for different selective sampling methods. The results are averaged over 10 runs.

presented.

3.1. Support vector data description (SVDD)

In this experiment, the SVDD with kernel whitening (Tax and Juszczak) is used. From Table 3, it can be seen that:

- the **ll** and **hl** methods are the slowest ones; they require to label more samples than the other methods to reach the same classification error. The low performance of the **ll** method in the combination with SVC is surprising because SVC considers just objects close to a decision boundary during computation, in multi-class approach methods that samples close to the decision boundary perform the best. The difference between one- and multi-class approach is that in occ it is more important to expand the target class regions rather then refine the boundary.
- the **lh** method is the fastest one; it requires to label less samples than the other methods. This method allows to evolve the classifier fast by asking for the true labels of highly confident patterns, classified as outliers and supports the description boundary by patterns of a low confidence classified as targets.
- the **hh** method also allows to evolve the classifier fast by asking for the true labels of highly confident patterns classified as outliers, but the description boundary is not supported by patterns classified as targets close to the boundary. In consequence, the boundary is collapsing around the training size of 50; see Table 3.

3.2. Autoencoder neural network (ANN)

We train two autoencoder neural networks with 5 hidden units: one for the target class and one for the outlier class. For this classifier, both the **lh** and **hh** methods perform almost equally well, since they allow for fast classification improvement by finding the true labels of the patterns classified as outliers with high confidences. Also here the **ll** and **hl** methods are the worst ones and **hh** cause a rise of the classification error.

3.3. Density based classifiers

For density estimation classifiers based on: Parzen, gaussian distributions, mixture of gaussians or on other density types, all selective sampling methods based on distances to a description boundary do not perform well, especially the **hh** method; see Table 3. They spoil the density estimation. For this type of classifiers the best sampling algorithm is the random method **ra**, because it uniformly samples the classes over entire distributions.

3.4. Different size of the bounding box

The size of the bounding box has an influence on the performance of the selective sampling methods introduced in section 2. This influence is stronger for methods that do not use the information about classification during selection or the distance to the currently trained classifier. In figure 3, the classification error for different sizes of the bounding box is presented(8 (top) and 20 (bottom) of the maximum distance, within the target class, in the respective feature direction). For

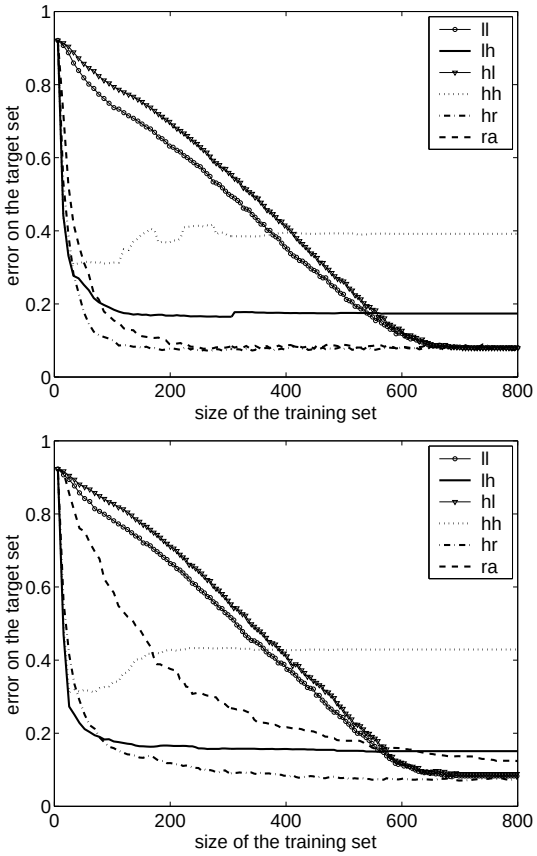

Figure 3. The classification error $\mathcal{E}_{\mathcal{I}}$ for the SVDD trained on merged Higleyman classes for different size of the bounding box 8 (top) and 20 (bottom). The results are averaged over 10 runs.

selective sampling methods not based on the distance to the classifier - (**hr**) and classification knowledge - (**ra**), the probability that the most informative patterns will be selected and presented to an expert is lower when the size of the bounding box is larger; see

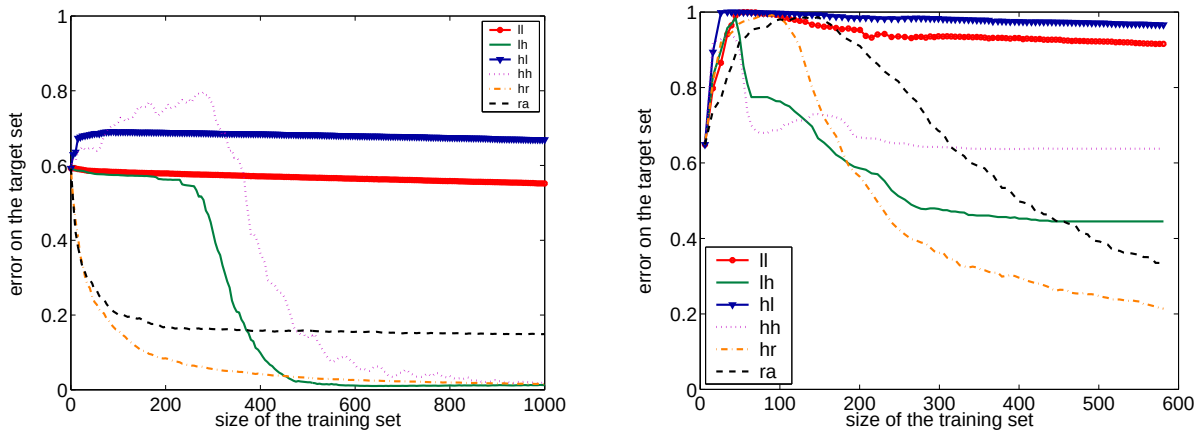

Figure 4. The classification error $\mathcal{E}_{\mathcal{I}}$ for the target class (left) for the SVDD with kernel whitening, trained on the mine data with the sand type of soil, (right) for the Parzen classifier trained on texture data, for different selective sampling methods. The results are averaged over 10 runs.

figure 3. For the (**hh**) and (**lh**) methods only the selection of objects classified as outliers depends on the size of the bounding box, so they are less dependent on it. These methods select patterns, closer to edges of the bounding box than to the classifier. For the very large size of the bounding box the best performance has the (**ll**) method, it samples from the regions that are in the vicinity of the description boundary.

4. Experiments with the real-world data

4.1. Texture data

This image data contains five different type of textures, where one of them was chosen as the target class and all others became the outlier class. The 7-dimensional data set contains the following features: the outputs of Gabor and Gauss filters and the second derivative estimates. It contains 13231 target examples and 52305 outlier examples.

4.2. Mine data

Land mines are hidden in a test bench of different soils: sand, clay, peat and ferruginous. Features are infrared images taken at different day time (12-dimensional feature space). Only the approximated positions of the mines are known (consequently some mine pixel labels are incorrect). Because of this and because the collection of soil samples is easier and safer than the collection of mine samples and some of the mine pixel labels are incorrect, soil was taken as the target class and mines as the outlier class. The data contains 3456 examples of the target class and 23424 examples of outlier class. We built a classifier for each type of soil separately. We did not consider mixtures of soils.

In this experiment the Parzen and the SVDD with kernel whitening was used. For each dataset, the initial training sets contain 40 randomly chosen target objects. In each iteration step, 5 objects currently classified as targets and 5 objects currently classified as outliers are added to the training set with their true labels. The classification errors for the target class for the selective sampling methods, described in section 2, are shown in figure 4.

Similar as for artificial data, the results for the **hl** and **ll** methods are very bad, because the initial training set might have been too small. The **hl** and **ll** selective sampling methods select mainly those target objects that are close to the actual description boundary. As a result, the classifier can only grow slowly.

5. Conclusions

We have described several methods in which unlabeled data can be used to augment labeled data based on the confidence of classifiers. Many selective sampling methods try to improve the performance of a classifier by adding supplementary patterns from the vicinity of the classifier. These patterns have a high probability to be wrongly classified. Because they are close to the current classifier including them in the training set, with their true labels, will improve the classification performance slightly. One-class classification differs from the standard, half-spaces, two-class problem because of the assumption that the domain of one of the classes, the target class, is limited to a certain area.

If in this problem only a small, labeled, target set is available, with the size e.g. twice the data dimensionality and we would like to improve the performance of a classifier by asking an expert for labels of the supplementary data, then the selection of patterns close to the description boundary (**ll**, **lh** methods) will build a more dense distribution of the target class.

The choice of a selective sampling method depends on the classifier considered. For some classifiers, like the SVDD or the ANN, selective sampling methods based on the distance to the decision boundary will perform well. Patterns close to the decision boundary influence them the most. For classifiers based on density estimation, like the Parzen classifier, selective sampling methods based on the distance to the decision boundary could spoil the estimation of the density. It could happen that adding more samples to the training set will, in fact, increase the classification error.

In problems where only a small target set is available and the task is to select a small unlabeled set to be labeled by an expert, to reach the desired classification error, it is worth to base the selection procedure on the confidence of the classifier. Our experiments showed that by selecting objects far from the description boundary it is possible to lower the number of necessary objects to be labeled by the expert. If the classes are not overlapping it is possible to improve further the classifier by changing the selective sampling method to one that chooses the most informative patterns close to the decision boundary (**ll**, **lh**).

The performance of the methods, based on the confidence of the classifier, presented in this paper depends on the size of the bounding box. The size of the box has the strongest influence on the random method **ra**. For very large size of the bounding box the best performance will be given by the **ll** selective method.

Acknowledgments

We would like to thank to dr. K. Schutte from FEL-TNO and W.A.C.M. Messelink from TU Delft for providing us with the mine data. This work was partly supported by the Dutch Organization for Scientific Research (NWO).

References

Blum A., Mitchell T.M.*Combining labeled and unlabeled data with co-training.* Proceedings of the 1998 Conference on Computational Learning Theory

Cambell C., Cristianini N., Smola A., *Query learning with large margin classifiers*

Cohn D., Atlas L., Ladner R., *Improving generalization with active learning*, 1992

Cohn D., Ghahramani Z., Jordan M.I. *Active learning with statistical models*, Journal of artificial intelligence research 4, 1996 129-145

Freund Y., Seung H.S., Shamir E., Tishby N., *Selective sampling using the query by committee algorithm*, Machine Learning, 28, 133-168 (1997)

Japkowicz N., *Concept-learning in the absence of counter-examples: an autoassociation-based approach to classification*, PhD thesis 1999

Lewis D.D. and Gale W.A. *A sequential algorithm for training text classifiers* Proceedings of 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval 1994

Michell T.M.*Generalization as search.* Artificial Intelligence, 18, 1982

Muslea I., Minton S., Knoblock C., *Selective sampling with redundant views*, Proceedings of the 15th National Conference on Artificial Intelligence, 621-626, AAAI-2000.

Tax D.M.J. and Juszczak P., *Kernel whitening for data description*, International Workshop on Pattern Recognition with Support Vector Machines 2002

De Ridder Dick, Tax D.M.J. and Duin Robert P.W. *An experimental comparison of one-class classification methods* Proceedings ASCI'98, 4th Annual Conference of the Advanced School for Computing and Imaging

Tax David M.J., Duin Robert P.W., *Support Vector Data Description*, Pattern Recognition Letters, December 1999, vol. 20(11-13), pg. 1191-1199

Tax D.M.J., 'One-class classification' , PhD Thesis, Delft University of Technology, ISBN: 90-75691-05-x, 2001

Tong S. and Koller D. *Support vector machine active learning with applications to text classification.* Proceedings of the 17th International Conference on Machine Learning

Warmuth M.K., Rätsch G., Mathieson M., Liao J., Lemmen C., *Active learning in the drug discovery process*