How to ask an expert for new samples in the one-class classification problem?

Piotr Juszczak, Robert P.W. Duin

Pattern Recognition Group, Faculty of Applied Sciences,
Delft University of Technology, Lorentzweg 1,
2628 CJ Delft, The Netherlands
email:{piotr,bob}@ph.tn.tudelft.nl

Keywords: the one-class classification problem, selective sampling methods, active learning

Abstract

Selective sampling, a part of the active learning method, reduces the cost of labeling supplementary training data by asking only for the labels 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, there can be a large number of unlabeled examples available. 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 labels 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 [4]. 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 should not be used excessively. The question is: how to select an additional subset of unlabeled data such that by including it in the training set would improve the performance of a particular classifier 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 description boundary [3] e.g. close to a margin or inside a margin for the support vector classifier [2],
- which have the most evenly split labels over a variation of classifiers:
  - trained on multiple permutations of the labeled data [12],
  - differing by the settings,
  - trained on independent sets of features [8].

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 classification. The classification performance is improved in small steps. In this paper, we will test a number of selective sampling methods for one-class classification problems [10, 6].

In the problem of one-class classification, one class of objects, called the target class, has to be distinguished from all the other possible objects, called

outliers. The description should be constructed such that the acceptance of objects not originating from the target class should be minimized. The problem of the one-class classification is harder than the standard two-class classification problem. In a 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; see Figure 1. Because in case of a one-class classification only the target class is available, only 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 error that the classifier would make. 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.

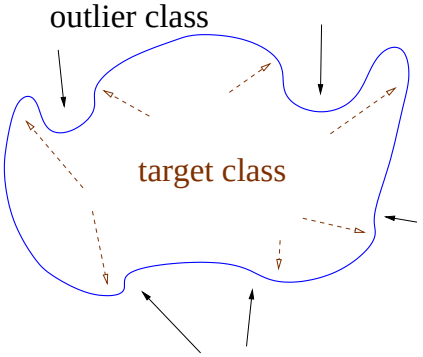

Figure 1: Influence of the target (generalization) and the outlier (specialization) classes on the description boundary.

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 [7], will be used.

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, presented 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 Table 2.

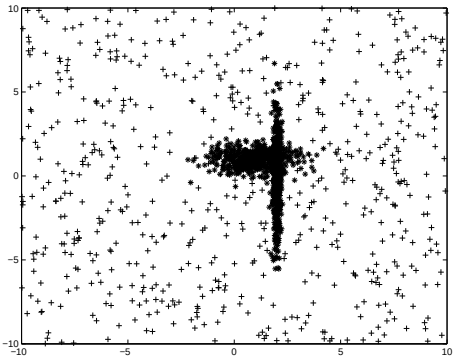

Figure 2: The merged Higleyman classes {N([1 1],[1 0; 0 0.25]; N([2 0],[0.01 0; 0 4])} with a uniformly distributed outlier class.

If a 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 [1, 2, 3, 5].

Because it is usually not possible to compute the distance between a pattern and a classifier, we propose to base this distance measure on the preprocessed output of the classifier $f^c(x)$, where c indicates either a target (t) or an outlier (o) class. The raw output of the classifier $f^c(x)$ is converted to the confidence Γ_x^c of the classifier that the object x belongs to one of the classes (target or outlier). Where:

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

The confidence Γ_x^c is computed as follows:

$$\Gamma_x^c = \frac{f^c(x)}{\sum_{x \in c}[f^c(x)]} \tag{1}$$

where $c = target \vee outlier$ class.

[1] If $f^c(x) > 0$, 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.

	ll	**lh**	**hl**	**hh**
target class	$\Gamma_x^t \ll 0.5$	$\Gamma_x^t \ll 0.5$	$\Gamma_x^t \gg 0.5$	$\Gamma_x^t \gg 0.5$
outlier class	$\Gamma_x^o \ll 0.5$	$\Gamma_x^o \gg 0.5$	$\Gamma_x^o \ll 0.5$	$\Gamma_x^o \gg 0.5$

Table 1: The description of selective sampling methods

```
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
```

Table 2: Active learning with selective sampling - The algorithm

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

There are two interesting regions of the classification confidences:

1. a high confidence, $\Gamma_x^c \gg 0.5$; the objects are far from the decision boundary,

2. a low confidence, $\Gamma_x^c \ll 0.5$; the objects are close to the decision boundary.

Based on the confidence regions of a classifier, we can describe four selective sampling methods that choose a set of examples (e.g. 5 from each target/outlier class) for an oracle to label them:

ll - a low confidence for both the target and the outlier classes

lh - a low/high confidence for the target/outlier class

hl - a high/low confidence for the target/outlier class

hh - a high confidence for both the target and the outlier classes

We compare these sampling techniques with the two methods that are not dependent on the distance to the description boundary:

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 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.

In experiments with the artificial data we used the following datasets: banana [10], multidimensional Gaussian and the merged Higleyman classes {N([1 1],[1 0; 0 0.25]; N([2 0],[0.01 0; 0 4])} ; see Figure 2. As the outlier class, we considered objects uniformly distributed in the bounding box. The results for all these datasets were similar. For clarity, in section 3, we present only the outcomes on the merged Higleyman classes.

3 Experiments with the artificial data

Now we will present the results of experiments performed on the $2D$ Higleyman classes, using the selective sampling methods described in section 2. A number of different classifiers is taken into account: Support Vector Data Description(SVDD) [11], Autoencoder Neural Network(ANN) and the Parzen classifier. The dataset contains 1000 target objects and 5000 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 objects' labels are retrieved and the classifier

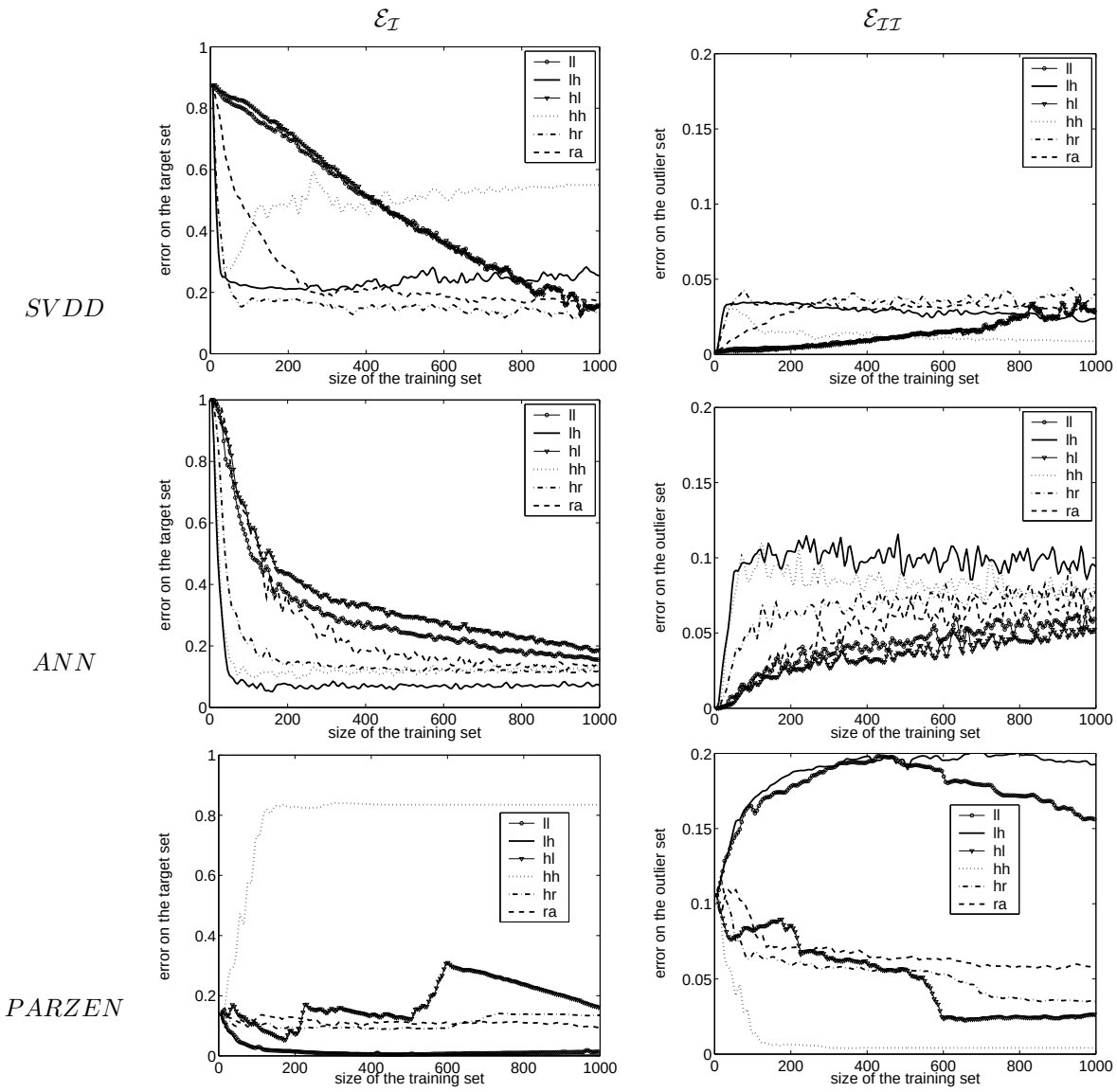

Table 3: The classification error $\mathcal{E}_{\mathcal{I}}$ and $\mathcal{E}_{\mathcal{II}}$ for SVDD, autoencoder (ANN) and Parzen classifier, on merged Higleyman classes for different selective sampling methods. The results were averaged over 20 runs.

is retrained. The error of the first kind $\mathcal{E}_{\mathcal{I}}$ [10] 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 20 runs are presented. To see how well a classifier fits the data both errors $\mathcal{E}_{\mathcal{I}}$ and $\mathcal{E}_{\mathcal{II}}$ should be considered.

3.1 Support vector data description (SVDD)

In this experiment, the SVDD [11] with kernel whitening [9] 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 **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; see Figure 1.
- 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 in 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 the fast classification improvement by finding the true labels of the patterns classified as outliers with high confidences. Because the low confidence region $\Gamma_x^t \ll 0.5$, and the high confidence region $\Gamma_x^t \gg 0.5$ for the target class are relatively close to each other compared to the low confidence region $\Gamma_x^o \ll 0.5$ and the high confidence region $\Gamma_x^o \gg 0.5$ for the outlier class, almost no difference between performance of the **lh** and **hh** methods can be observed.

3.3 Density based classifiers

For density estimation classifiers based on: Parzen, gaussian distribution, mixture of gaussians or for other types like the nearest neighbor classifier, all selective sampling methods based on distances to a description boundary do not perform well, especially **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 the entire distribution.

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 during a selection the information about classification or the distance to the currently trained classifier. In Figure 3, the classification error for different size (8 (upper) and 20 (lower) of the maximum distance, within the target class, in the respective feature direction) of the bounding box is presented. For selective sampling methods do not based

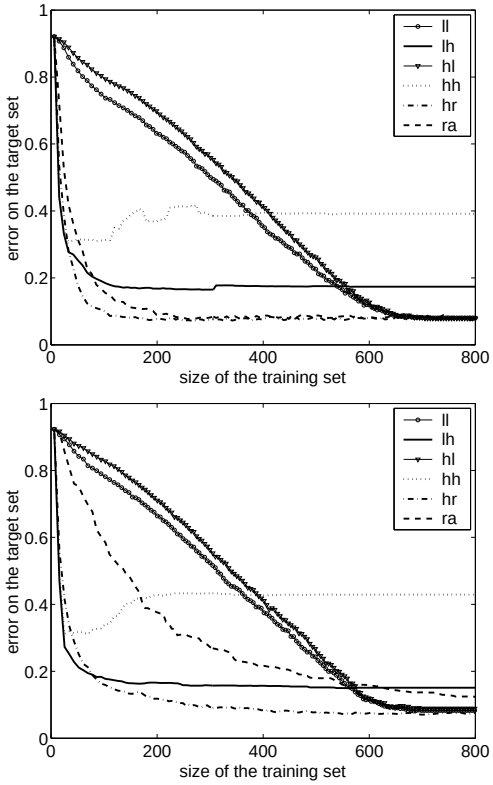

Figure 3: The classification error $\mathcal{E}_{\mathcal{I}}$ for the SVDD train on merged Higleyman classes for different size of the bounding box 8 (top) and 20 (bottom). The results were averaged over 20 runs.

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 bigger; see Figure 3. For **hh** and **lh** methods only the selection of objects classified as outliers depends on the size of the bounding box, so there are less dependent on it. These methods select patterns, rather close

to edges of the bounding box than to the classifier. For the very large size of the bounding box the best performance has **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 become 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.

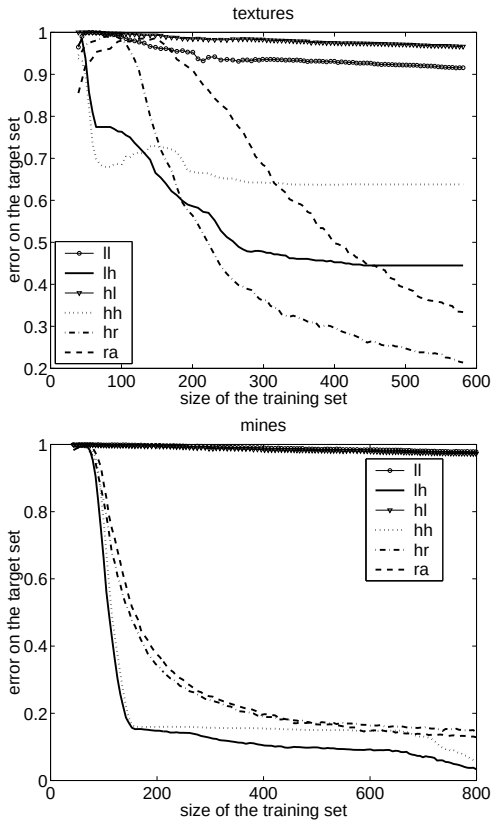

Figure 4: The classification error $\mathcal{E}_{\mathcal{I}}$ for the SVDD with kernel whitening, trained on the texture data (top) and the mine data with the sand type of soil (bottom), for different selective sampling methods. The results were averaged over 10 runs.

4.2 Mine data

Mines are hidden in a test bench of different soils: sand, clay, peat and ferruginous. Features are infra-red images taken at different day time (12-dimensional feature space). Only the approximated positions of the mines are known (some mine pixel labels are incorrect). 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 SVDD [11] with kernel whitening [9] 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 selective sampling methods, described in section 2, are shown in Figure 4.

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, it can only grow slowly.

5 Conclusions and open questions

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 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, for reaching 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.

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.

6 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

[1] A. Blum, T. Mitchell 'Combining labeled and unlabeled data with co-training' Proceedings of the 1998 Conference on Computational Learning Theory

[2] C. Cambell, N. Cristianini, A. Smola, 'Query learning with large margin classifiers'

[3] D. Cohn, L. Atlas, R. Ladner, 'Improving generalization with active learning', 1992

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

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

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

[7] M.J. Kearns, U.V. Vazirani, 'An introduction to computational learning theory', The MIT Press 1994, ISBN 0-262-11193-4

[8] Ion Muslea, Steve Minton, Craig Knoblock, 'Selective sampling with redundant views', Proceedings of the 15th National Conference on Artificial Intelligence, 621-626, AAAI-2000.

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

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

[11] D.M.J. Tax, R.P.W. Duin, 'Support Vector Data Description', Pattern Recognition Letters, December 1999, vol. 20(11-13), pg. 1191-1199

[12] M.K. Warmuth, G. Rätsch, M. Mathieson, J. Liao, C. Lemmen, 'Active learning in the drug discovery process'