

An Evaluation of Dimension Reduction Techniques for One-Class Classification

Santiago D. Villalba

Pádraig Cunningham

University College Dublin
Technical Report UCD-CSI-2007-9
August 13th, 2007

Abstract

Dimension reduction (DR) is important in the processing of data in domains such as multimedia or bioinformatics because such data can be of very high dimension. Dimension reduction in a supervised learning context is a well posed problem in that there is a clear objective of discovering a reduced representation of the data where the classes are well separated. By contrast DR in an unsupervised context is ill posed in that the overall objective is less clear. Nevertheless successful unsupervised DR techniques such as Principal Component Analysis (PCA) exist – PCA has the pragmatic objective of transforming the data into a reduced number of dimensions that still captures most of the variation in the data. While one-class classification falls somewhere between the supervised and unsupervised learning categories, supervised DR techniques appear not to be applicable at all for one-class classification because of the absence of a second class label in the training data. In this paper we evaluate the use of a number of up-to-date unsupervised DR techniques for one-class classification and we show that techniques based on *cluster coherence* and *locality preservation* are effective.

1 Introduction

In recent years, the traditional distinction in machine learning between supervised and unsupervised techniques has been blurred due to the emergence of real-world problems that sit somewhere between these two extremes. In supervised classification problems, *discriminating* classifiers are trained using positive and negative examples. However, for a number of practical problems, counter-examples are either rare, entirely unavailable or statistically unrepresentative. Such problems include industrial process control, text classification and analysis of chemical spectra.

One-class classifiers (OCCs) have emerged as a set of techniques for situations where labelled data exists for only one of the classes in a classification problem. For instance, in industrial inspection tasks, abundant data may only exist describing the process operating correctly. It is difficult to gather training data describing the myriad of ways the system might operate incorrectly. A related problem is where negative examples

exist, but their distribution cannot be characterised. For example, it is reasonable to provide characteristic examples of family pictures but impossible to provide examples of pictures that are “typical” of non-family pictures. One-class classifiers are emerging as a solution, which *characterises* the target class, to distinguish it from all other classes.

In practice, one-class problems are typically of high dimension so DR is an important pre-processing step. Indeed the evaluation presented by Manevitz and Yousef [1] shows that one-class Support Vector Machine (SVM) performance is quite sensitive to the number of features used. This contrasts with two-class SVMs which are generally considered to be robust to high data dimensionality. This provides additional justification for DR in one-class classifier construction. However, the absence of counter-examples means it is difficult to identify a feature subset that encodes a *discriminating* description of the concept.

In this paper we review a range of unsupervised DR techniques and evaluate their performance on a number of OCCs. We find that DR based on *locality preservation* and *cluster coherence* principles seem particularly promising for OCC. However locality preservation is only effective when there are no irrelevant features in the full feature set; i.e. locality in the original space must be meaningful.

In the next section we provide an overview of OCC and describe the OCC techniques included in the evaluation. In section 3 we describe the DR techniques considered in the evaluation – the evaluation is presented in section 4. The paper concludes with a summary and some proposals for further research.

2 One-Class Classifiers

Traditionally machine learning tasks are divided into supervised and unsupervised categories. Roughly speaking, in unsupervised learning we are provided with a dataset (set of examples describing a real world concept) and the objective is to uncover some structure in the data. In supervised learning we are provided with a dataset where the information to be modeled is explicitly stated in the form of a label (a “class” label in the case of so called “classification problems”) and the task is to predict the label for new (as yet unseen) examples.

One-class classification, also referred to as novelty or outlier detection, is sometimes thought of as a weaker form of supervised classification, where the only information we are given about the training examples is that they belong to the same class, arbitrarily called “positive” or “target”. The task here is to accept or reject unseen examples depending on their similarity to the known positive examples. OCC approaches consequently can operate with very few, or no, negative training examples. In other words, one-class learning handles the “no-counter-example” and “imbalanced-data” problems by considering only positive examples. When unlabeled examples and/or small (or large) amounts of negative examples are available for training, several OCC techniques can also use them to fine-tune their performance.

In the current study we choose four different one class classifiers (Support Vector Data Description (SVDD), a k -Nearest Neighbours approach, a k -Means Clustering approach

and a Gaussian Model), all of them available in the Data Description toolbox [2], an open source Matlab software library of one-class classification tools.

Support Vector Data Description [3, 4]: The SVDD learns the hypersphere, defined by a center a and a radius R , that encloses (almost) all the training set while covering as little volume as possible. It can employ the kernel trick for learning more flexible boundaries, and the solution is found by solving a convex quadratic optimization problem analogous to the one found in Support Vector Machines.

Clustering (k -Means): Another approach to one-class classification is that of learning clusters, modeling the target class as a reduced set of cluster prototypes or centers onto which new examples are projected. Examples of clustering methods that can be used are the Self Organizing Map, Learning Vector Quantization or k -Means, the one we choose here [3, 5]. When a new example is to be classified, its distance to the nearest prototype is used to score the extent to which it is an outlier.

Lazy learning (k -Nearest-Neighbours): The Nearest Neighbour approach can be used for constructing one-class classifiers. The training data is stored and an *outlierness* criterion is calculated for new examples based on their nearest neighbours, i.e. their position relative to the seen examples. Several criteria have been proposed to measure the outlierness of an example [6, 7]. Here we use γ [6] which is the average of the distances to the k nearest neighbours.

Density estimation (Gaussian model) [3, 5]: The Gaussian model is a simple parametric one-class classifier which models the training data under the assumption that it comes from a unimodal multivariate normal distribution. These assumptions fit a lot of natural processes, but when they are violated this model introduces a large bias. The mean and the covariance matrix are estimated using an Expectation-Maximization approach and the Mahalanobis distance is used as the resemblance criterion.

We selected these OCC strategies and not others because of their conceptual simplicity, their well established properties and because we wish to explore the hypothesis that local learners are particularly well suited to one class problems. All but SVDD are local learners.

3 Dimension Reduction Techniques

Research on dimension reduction has itself two dimensions. The first design decision is whether to select a subset of the existing features or to transform to a new reduced set of features. The other dimension in which DR strategies differ is the question of whether the learning process is supervised or unsupervised (see Figure 1). For OCC problems it seems that both feature selection and feature transformation strategies are relevant. However, given that labelled data is only available for one class, it seems that supervised DR techniques cannot be directly applied to OCC problems.

In supervised learning the objective of DR is to optimize the performance of the final system, that is, minimize the classification error. However, in one-class classification performance estimation is difficult because the absence of counterexamples makes the estimation of the false positive rate hard. This makes it difficult also to tune the bias of

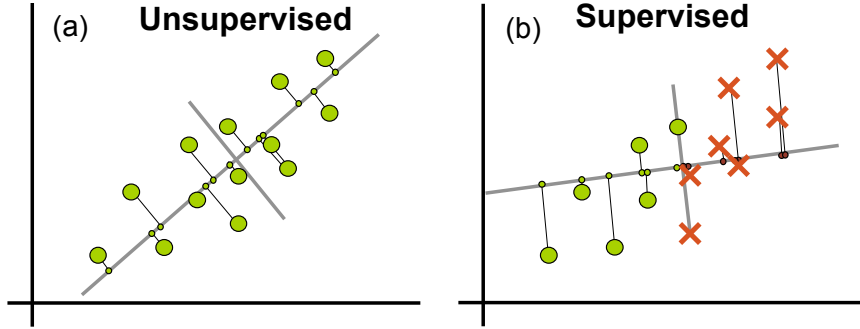


Figure 1: In unsupervised DR (a) the “best” that can be done is to find a representation that maximises the variance in the data. When the data is labelled (b) a representation can be sought that improves the separation in the data.

the classifier and the best strategy to address this problem depends on the specifics of the data available.

A sensible approach is to try to synchronize the assumptions of both DR and classification. In our evaluation we consider four DR techniques; the first two are Principal Component Analysis and the $Q-\alpha$ algorithm presented by Wolf and Shashua [8]. The final two are based on the principle of *locality preservation* and these are described in section 3.1. We believe that locality preservation is of particular relevance to DR in the OCC domain because, usually one class classifiers rely on local neighbourhood relationships (see section 2).

Principal Component Analysis (PCA): PCA is the most commonly used technique for unsupervised dimensionality reduction [9]. It aims at finding the linear projections that best capture the variability of the data. In this study we use the common approach of keeping those directions that explains most of the variance. In [10] it is shown that retaining the high variance dimensions is not always optimal for one-class classification, so a minor components analysis (use the smallest variance directions) can be better under some circumstances.

The $Q-\alpha$ Algorithm: A well motivated criterion of cluster quality is cluster coherence, in graph theoretic terms this is expressed by the notion of objects within clusters being well connected and individual clusters being weakly linked. The whole area of spectral clustering captures these ideas in a well founded family of clustering algorithms based on the idea of minimising the *graph-cut* between clusters [11].

The principles of spectral clustering have been extended by Wolf and Shashua [8] to produce the $Q-\alpha$ algorithm that simultaneously performs feature subset selection and discovers a good partition of the data. As with spectral clustering, the fundamental data structure is the affinity matrix \mathbf{A} where each entry \mathbf{A}_{ij} captures the similarity (typically as a dot-product) between data points i and j . In order to facilitate feature selection the affinity matrix for $Q-\alpha$ is expressed as $\mathbf{A}_\alpha = \sum_{i=1}^p \alpha_i \mathbf{m}_i \mathbf{m}_i^\top$ where \mathbf{m}_i is the i^{th} feature vector in the data matrix that has been normalised so to be centered on 0 and be of unit L_2 norm (this is the set of values in the data set for feature i). $\mathbf{m}_i \mathbf{m}_i^\top$ is the *outer-product* of \mathbf{m}_i with itself. α is the weight vector for the p features – ultimately

the objective is for some of these weight terms to be set to 0.

In spectral clustering \mathbf{Q} is an $n \times k$ matrix composed of the k eigenvectors of \mathbf{A} corresponding to the largest k eigenvalues. Wolf and Shashua show that the relevance of a feature subset as defined by the weight vector α can be quantified by:

$$Rel(\alpha) = trace(\mathbf{Q}^T \mathbf{A}_\alpha^T \mathbf{A}_\alpha \mathbf{Q}) \quad (1)$$

They show that feature selection and clustering can be performed as a single process by optimising:

$$\max_{\mathbf{Q}_\alpha} trace(\mathbf{Q}^T \mathbf{A}_\alpha^T \mathbf{A}_\alpha \mathbf{Q}) \quad (2)$$

subject to $\alpha^T \alpha = 1$ and $\mathbf{Q}^T \mathbf{Q} = \mathbf{I}$.

Wolf and Shashua show that this can be solved by solving two inter-linked eigenvalue problems that produce solutions for α and \mathbf{Q} . They show that a process of iteratively solving for α then fixing α and solving for \mathbf{Q} will converge. They also show that the process has the convenient property that the α_i weights are biased to be positive and sparse, i.e. many of them will be zero.

So the $Q - \alpha$ algorithm performs feature selection in the spirit of spectral clustering, i.e. it discovers a feature subset that will support a partitioning of the data where clusters are well separated according to a graph-cut criterion.

3.1 Locality Preservation

Locality preservation in dimensionality reduction techniques refers to the aim of keeping neighbourhood properties, e.g. objects that are close in the input space should also be close in the reduced space. Several linear and nonlinear techniques exploiting this criterion have recently been proposed. For the OCC problem it is rational to think that a locality preserving dimension reduction technique would be more practical in some cases than a global based one. Locality and density are frequently used in the OCC literature and both are present in the locality preservation bias.

Locality Preserving Projections (LPP): The idea behind LPP is that of finding subspaces which preserve the *local structure* in the data [12, 13]. Given a matrix \mathbf{A} (symmetric, positive, invertible and, usually, sparse) which captures information about the relationships between the data points, for example the similarity in a neighbourhood, LPP finds the optimal linear embedding that respects the structure present in that matrix. LPP preserves cluster structures when clustering is based on locality, such as in the k -means algorithm, which renders an attractive quality for being used together with cluster analysis based OCCs. The details of LPP are described in Algorithm 1.

The embedding is defined by the bottom eigenvectors in the solution of equation 3. The construction of the weighted graph in the first and second steps of Algorithm 1 can be accomplished using any criterion. This permits more utilitarian approaches to be used by including *a priori* information about the problem.

Laplacian Score for Feature Selection (LS): The same criterion of locality preservation found in LPP can be applied in the feature selection context, where the merit of each feature is measured according to its locality preservation power [14].

As with $Q - \alpha$, there is no explicit enumeration of the feature subsets. Rather a nearest neighbour based graph is constructed from the training set and its spectrum is analysed to rank each variable. The first two steps of the algorithm are identical to those of LPP (Algorithm 1). For ranking each feature, its *Laplacian score* is computed. For the i -th feature we define:

$$\tilde{\mathbf{m}}_i = \mathbf{m}_i - \frac{\mathbf{m}_i^\top \mathbf{D} \mathbf{1}}{\mathbf{1}^\top \mathbf{D} \mathbf{1}} \mathbf{1} \quad (4)$$

where $\mathbf{1} = [1, \dots, 1]^\top$

The Laplacian Score (LS_i) for the i -th feature is:

$$LS_i = \frac{\tilde{\mathbf{m}}_i^\top \mathbf{L} \tilde{\mathbf{m}}_i}{\tilde{\mathbf{m}}_i^\top \mathbf{D} \tilde{\mathbf{m}}_i} \quad (5)$$

4 Evaluation

In this study we use the three biomedical datasets summarized in Table 1. One difficulty in assessing the performance of the combination of OCC and DR techniques on these datasets is the parameter optimisation (model selection) required for the different techniques. We followed a simple approach of fixing the values of the parameters to sensible values: $k = 6$, the number of clusters for k -means and neighbours for k -NN and the rest of the are left to the `dd_tools` default values. For computing the threshold of the resemblance function we set a 90% of training objects to be accepted (i.e. we consider that a 10% of the training examples are outliers).

As parameters for the dimension reduction techniques, when applicable, we follow the principle of using the same parameters used in the counterpart classifiers. For example, the same value of k in the k -nearest neighbour is used when constructing the adjacency graph for LPP and LS, and the value of k in the k -means algorithm is set as the target number of clusters for $Q - \alpha$. In both LPP and LS the “simple minded” weighting approach is followed. No further model selection is done and we also fix the rest of the parameters to *a priori* defined default values. The choice of target dimensionality is also an important issue. In this case we just explored all possibilities, from dimensionality 1 to the original dimensionality in feature selection or to the maximum defined by the feature transformation embedding.

In addition to the techniques described in section 3, we add two dimension reduction methods; a random feature selection process to provide a baseline and a supervised ranking of the features using information gain over the original multiclass datasets (this is “cheating”). In this way we try to establish if using a supervised feature selection criterion provides an upper bound for the accuracy of the dimension reduction system for one-class tasks.

Algorithm 1: LPP computation [12]

Construct the adjacency graph: let \mathcal{S} be the training set and G denote a graph with $|\mathcal{S}|$ nodes. We put an edge between nodes i and j if \mathbf{x}_i and \mathbf{x}_j are "close". There are two variations:

- ε -neighbourhoods (parameter $\varepsilon \in \mathbb{R}$). Nodes i and j are connected if $\|\mathbf{x}_i - \mathbf{x}_j\|^2 < \varepsilon$ where the norm is the usual Euclidean norm in \mathbb{R}^{d_x} .
- k nearest neighbours (parameter $k \in \mathbb{N}$). Nodes i and j are connected if i is among the k -nearest neighbours of j or viceversa.

Choose the weights for the graph edges: Here, as well, we have two variations for weighting the edges. \mathbf{A} is a sparse symmetric $|\mathcal{S}| \times |\mathcal{S}|$ matrix with \mathbf{A}_{ij} having the weights of the edge joining vertices i and j , and 0 if there is no such edge.

- Heat kernel (parameter $t \in \mathbb{R}$). When nodes i and j are connected put $\mathbf{A}_{ij} = e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{t}}$
- Simple minded (no parameter). When nodes i and j are connected, put $\mathbf{A}_{ij} = 1$.

Eigenmaps: Compute the eigenvectors and eigenvalues for the generalized eigenvector problem:

$$\mathbf{S}\mathbf{L}\mathbf{S}^\top \mathbf{a} = \lambda \mathbf{S}\mathbf{D}\mathbf{S}^\top \mathbf{a} \quad (3)$$

where \mathbf{D} is a diagonal matrix whose entries are column sums of \mathbf{A} , $\mathbf{D}_{ii} = \sum_j \mathbf{A}_{ji}$. $\mathbf{L} = \mathbf{D} - \mathbf{A}$ is the Laplacian matrix.

Table 1: Summary of the three datasets used in the evaluation.

Dataset	n# Examples	n# Features	Target Class (n#)	Source
<i>Bronchiolitis</i>	118	22	1-Day (37)	[15]
<i>Arrhythmia</i>	452	279	Normal (245)	[16]
<i>TIS-5%</i>	668	927	TIS (178)	[17]

Table 2: Evaluation of the DR-OCC combinations. Balanced Accuracy Rate estimations for the winning dimensionality (in brackets) for each classifier / dimension reduction technique pair. In italics the *cheating* supervised feature selection. In **boldface** the unsupervised winning dimension reduction techniques for each classifier. It is clear that they are beneficial over the *No DR* case.

(a) Bronchiolitis

	No DR	IG	Random	Q- α	LS	LPP	PCA
Gauss	0.64(22)	<i>0.72(12)</i>	0.67(16)	0.73(7)	0.68(13)	0.72(13)	0.73(17)
KMeans	0.73(22)	<i>0.71(20)</i>	0.67(18)	0.75(15)	0.70(19)	0.70(15)	0.72(19)
KNN	0.66(22)	<i>0.70(16)</i>	0.66(21)	0.66(21)	0.68(21)	0.72(12)	0.65(20)
SVDD	0.65(22)	<i>0.72(1)</i>	0.68(19)	0.67(16)	0.70(20)	0.64(20)	0.67(16)

(b) Arrhythmia

	No DR	IG	Random	Q- α	LS	LPP	PCA
Gauss	0.55(279)	<i>0.78(57)</i>	0.71(83)	0.71(167)	0.70(167)	0.68(30)	0.76(24)
KMeans	0.68(279)	<i>0.77(28)</i>	0.70(195)	0.74(167)	0.68(279)	0.68(139)	0.72(20)
KNN	0.67(279)	<i>0.77(35)</i>	0.68(223)	0.71(167)	0.67(279)	0.68(139)	0.70(20)
SVDD	0.68(279)	<i>0.75(83)</i>	0.68(167)	0.70(167)	0.66(279)	0.65(167)	0.69(29)

(c) TIS

	No DR	IG	Random	Q- α	LS	LPP	PCA
Gauss	0.54(927)	<i>0.83(3)</i>	0.54(463)	0.82(27)	0.70(18)	0.54(185)	0.77(4)
KMeans	0.45(927)	<i>0.79(3)</i>	0.49(1)	0.67(18)	0.56(6)	0.52(3)	0.74(4)
KNN	0.45(927)	<i>0.81(5)</i>	0.49(4)	0.70(19)	0.55(19)	0.54(3)	0.77(2)
SVDD	0.38(927)	<i>0.82(2)</i>	0.49(32)	0.69(3)	0.59(1)	0.49(3)	0.72(2)

The results are shown in Table 2. The Balanced Accuracy Rate, defined as the average of the true positive rate (sensitivity) and true negative rate (specificity), is estimated by 10-fold cross validation. The figures shown are those obtained by the winning target dimensionality in each case (in parenthesis).

In Table 2 dimension reduction is beneficial in all cases. This is not surprising since we chose datasets that require dimension reduction to achieve good results in a supervised setting. In most cases the supervised DR technique provide an upper bound for the performance. Q- α is very promising, it gives high scores to relevant features and yields consistent improvements.

In the case of the Arrhythmia and especially the TIS dataset the locality preserving principle of LPP is not competitive with the rest of the unsupervised criteria. This is due to the fact that the presence of a lot of irrelevant features in the full feature set renders locality in that space inappropriate. Further, when used with a non-local learner such as SVDD, LPP compares adversely even with the random feature selection criterion.

In Figure 2 we show the evolution of the sensitivity / specificity tradeoff for two dataset / classifier combinations: k -means in TIS and k -NN in Bronchiolitis. For locality-

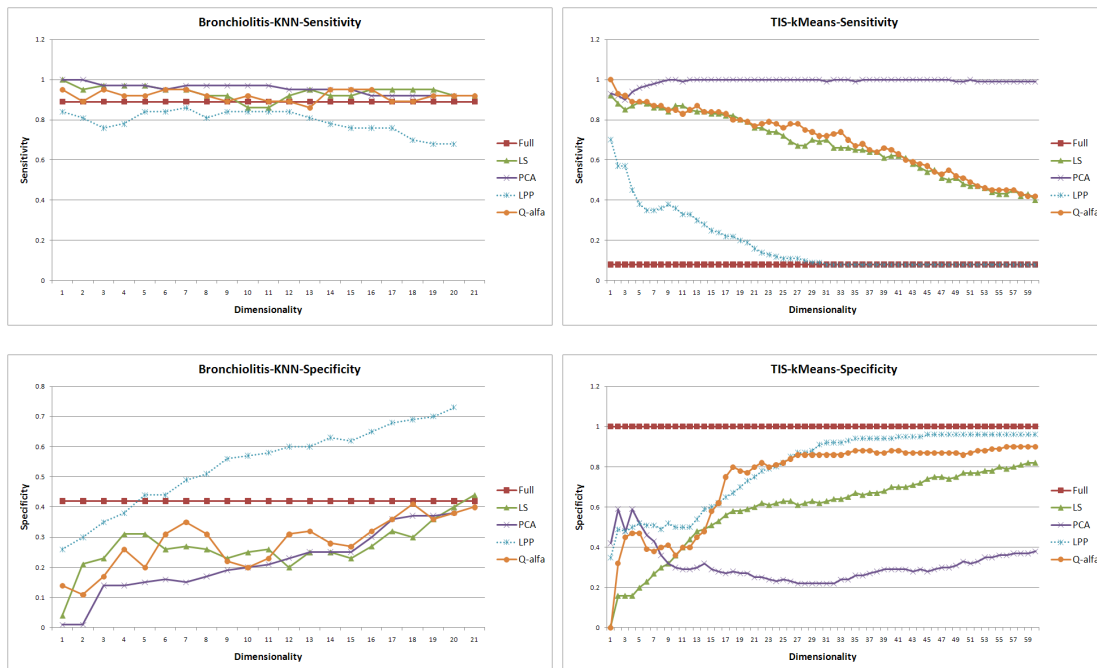


Figure 2: Evolution of Sensitivity (up) vs Specificity (bottom) tradeoff with increasing target dimensionality. On the left applying k -NN to the bronchiolitis dataset. On the right, applying k -means to the TIS dataset with dimensionality varying from 1 to 59. (This figure is best viewed in colour).

based classifiers it usually holds that the more the dimensionality is reduced, the better the sensitivity and the worse the specificity. This is due to the fact that computing description of objects in low dimensional spaces is easier while discriminative power is lost, so descriptions also capture outliers. When the dimensionality is high the descriptions become inaccurate and so the classifiers become accept-all or, more often, reject-all machines. This phenomenon is more significant in the case of LPP.

5 Conclusions

This paper reports progress in research on the applicability of DR techniques (specifically techniques from unsupervised learning) for OCC problems. At this stage we feel there are two key findings:

- We have demonstrated the potential improvements to be had by applying carefully selected DR techniques prior to one-class classification.
- In some circumstances techniques based on *cluster coherence* and *locality preservation* are particularly effective. We believe that LP-based techniques are appropriate

when there are few irrelevant features in the data set, i.e. in order for locality to be meaningful in the original feature space it is necessary for most if not all of the features to be relevant.

As already stated, *locality preservation* seems an appropriate criterion for OCC tasks but it contains the implication that none of the input features are irrelevant; they may be just redundant. The key issue here is how meaningful the distance functions are. We encounter the paradoxical situation that for reducing the dimensionality of the data one needs to rely on distance measures which, most probably, are not meaningful in the original high dimensional space. How to learn a proper metric from the data itself instead of imposing a pre-specified one is an active research field in several areas of classification. For one-class classification, once again, the current techniques are not directly applicable because they use information from both sides of the classification boundary. However, related techniques could lead to useful one-class metric learning techniques.

References

- [1] Manevitz, L.M., Yousef, M.: One-class SVMs for document classification. *Journal of Machine Learning Research* **2** (2001) 139–154
- [2] Tax, D.M.J.: DDtools, the Data Description Toolbox for Matlab (April 2007)
- [3] Tax, D.M.J.: One-class classification. Concept learning in the absence of counterexamples. PhD thesis, Delft University of Technology (2001)
- [4] Tax, D.M.J., Duin, R.P.W.: Support vector data description. *Mach. Learn.* **54**(1) (January 2004) 45–66
- [5] Juszczak, P.: Learning to recognise, a study on one-class classification and active learning. PhD thesis, Delft University of Technology (2006)
- [6] Harmeling, S., Dornhege, G., Tax, D., Meinecke, F., Muller, K.R.: From outliers to prototypes: Ordering data. *Neurocomputing* **69**(13-15) (August 2006) 1608–1618
- [7] Rieck, K., Laskov, P.: Language models for detection of unknown attacks in network traffic. *Journal in Computer Virology* **2**(4) (February 2007) 243–256
- [8] Wolf, L., Shashua, A.: Feature selection for unsupervised and supervised inference: The emergence of sparsity in a weight-based approach. *Journal of Machine Learning Research* **6** (2005) 1855–1887
- [9] Jolliffe, I.T.: *Principal Component Analysis*. Springer (October 2002)
- [10] Tax, D., Muller, K.R.: Feature extraction for one-class classification. In: *ICANN/ICONIP 2003*, Springer Berlin / Heidelberg (2003) 342–349

- [11] Ng, A.Y., Jordan, M.I., Weiss, Y.: On spectral clustering: Analysis and an algorithm. In Dietterich, T.G., Becker, S., Ghahramani, Z., eds.: *Advances in Neural Information Processing Systems*. Volume 14. (2001)
- [12] He, X., Niyogi, P.: Locality preserving projections. In: *NIPS: Advances in Neural Information Processing Systems*. (2003)
- [13] He, X.: *Locality Preserving Projections*. PhD thesis, University of Chicago (2005)
- [14] He, X., Cai, D., Niyogi, P.: Laplacian score for feature selection. In: *NIPS: Advances in Neural Information Processing Systems*. (2005)
- [15] Doyle, D.: *A Knowledge-Light Mechanism for Explanation in Case-Based Reasoning*. PhD thesis, University of Dublin, Trinity College (2005)
- [16] Bay, S.D., Hettich, S.: The UCI KDD archive [<http://kdd.ics.uci.edu>] (1999)
- [17] Liu, H., Wong, L.: Data mining tools for biological sequences. *J Bioinform Comput Biol* **1**(1) (April 2003) 139–167