A Mixture Approach to Novelty Detection Using Training Data with Outliers

Martin Lauer

Institut für Logik, Komplexität und Deduktionssysteme
Universität Karlsruhe
D-76128 Karlsruhe, Germany
lauer@ira.uka.de

Abstract. This paper describes an approach to handle multivariate training data which contain outliers. The aim is to analyze the training patterns and to detect anomalous patterns. Therefore we explicitly model the existence of outliers in the training data using a widespread outlier distribution. Indicator variables assign each pattern to either the outlier distribution or the distribution of normal patterns. Thus we can estimate the data distribution using the EM-algorithm or Data Augmentation. We present the general approach as well as a concrete realization where we use Gaussian mixture models to describe the patterns' distribution. Experimental results show the applicability of this approach for practical studies.

1 Introduction

Novelty detection is concerned with the identification of anomalous patterns in data sets. These patterns are often faulty values generated by flaws in data ascertainment. Possible reasons are faulty measurement instruments or mistakes when feeding the computer with data, among others.

Furthermore outlying patterns may influence the analysis of data a lot. Standard approaches for regression analysis like the least sum of squares approach or for density estimation like the method of moments suffer a lot from their sensitivity to outliers. Even an outlier rate fewer than one per cent may corrupt the result of such a statistical analysis. Some examples illustrating this problem are given in [11]. Therefore it is necessary to detect outlying patterns and to eliminate them from the training data.

Several approaches have been proposed to tackle the task of novelty detection. Almost all approaches are based on the idea to learn a model of the data distribution and afterwards classify the patterns according to a density level. Two types of approaches can be distinguished: a) approaches working only with outlier-free training sets which are designed to find outliers in test data and b) approaches working on training data with a known number of outliers. The models which are used for data description are Gaussian mixture models (GMMs) [10,9], auto-associating neural networks [6,15], self organizing maps [16,8] and a class of sets based on support vector representation [12,2], among others.

L. De Raedt and P. Flach (Eds.): ECML 2001, LNAI 2167, pp. 300–311, 2001.
© Springer-Verlag Berlin Heidelberg 2001

In this paper we want to extend the approach based on GMMs for the use of training data which contain outliers themselves. Up to now GMMs were only used to estimate the density of a data distribution from outlier-free training data. Afterwards these estimates can be used to find anomalous patterns in other data sets which are taken from the same distribution but which contain supplementary outliers. Our approach is more robust against outliers so that it tolerates even a small amount of outliers in the training data. To calibrate the algorithm we need to have either classified validation patterns or a rough knowledge of the outlier proportion in the training data. But in contrast to the above mentioned approaches we do not need to know the exact number of outliers.

2 Key Idea

We start from the presumption that the training data contain a small amount of outliers. Thus we can model the pattern distribution as the composition of a) a big percentage of normal patterns[1] and b) a small proportion of corrupted patterns. If we denote the proportion of anomalous patterns with λ, the distribution of normal patterns with P_N and the distribution of outliers with P_O we can describe the distribution of the whole training set by:

$$P(x) = (1 - \lambda) \cdot P_N(x) + \lambda \cdot P_O(x) \tag{1}$$

λ can be interpreted as the prior probability for outlying patterns. The modeling described in (1) is called the "mixture alternative" in [1].

The learning task can now be split up into three steps:

1. estimate $P(x)$ from the given training set
2. decompose $P(x)$ into $P_N(x)$ and $P_O(x)$
3. decide whether x is an outlier given the probabilities $P_N(x)$, $P_O(x)$ and the prior probability λ

The second step is a delicate task since there are many possibilities to decompose $P(x)$ into two parts. Additionally we neither know which training patterns are outlying nor the exact number of anomalous patterns. If we assume an outlier percentage of $\leq 1\%$ and a number of training patterns ≤ 1000 the number of outliers is anyway too small to estimate the distribution $P_O(x)$ reliably. Therefore we cannot perform the second step directly.

Instead we assume that we know the outliers' distribution P_O and the outlier proportion λ or at least have a rough idea which we can use as an approximation for P_O and λ. We focus on the special problem of determining appropriate P_O and λ in sect. 5. Then we can derive the distribution of normal patterns from (1):

$$P_N(x) = \frac{1}{1 - \lambda} P(x) - \frac{\lambda}{1 - \lambda} \cdot P_O(x) \tag{2}$$

[1] We use the term "normal pattern" as complement to "outlier" or "anomalous pattern", not in the sense of a pattern derived from a Gaussian distribution.

The second step is now to estimate $P_N(x)$ from the training patterns. The rough knowledge of $P_O(x)$ and λ will help to restrain the outliers' influence on the estimation of $P_N(x)$. Thus we can work with training data which contain anomalous patterns.

If we would omit to consider the outlier distribution $P_O(x)$ and try to estimate $P_N(x)$ directly from the training patterns we would actually learn $P(x)$ which can be substantially different from $P_N(x)$. In particular $P(x)$ will be much more widespread than $P_N(x)$. If we assume the patterns to be distributed according to a parameterized model, e.g. a Gaussian distribution, the estimated parameters will be corrupted by the outliers, e. g. the variances of the Gaussian will be too large. Thus the explicit modeling of an outlier distribution P_O preserves the estimate from being corrupted by outliers.

The third step of classifying patterns according to $P_N(x)$, $P_O(x)$ and λ is an application of the Bayesian classification approach. λ is the prior probability for outliers, $P_O(x)$ and $P_N(x)$ are the distributions for anomalous and normal patterns, respectively. Thus a new pattern x is classified as outlier if the probability of belonging to the set of outliers is larger than the probability of being a normal pattern, i.e. equation (3) holds:

$$x \text{ is classified as outlier} \quad \text{if and only if} \quad \lambda \cdot P_O(x) > (1 - \lambda) \cdot P_N(x) \tag{3}$$

3 Implementation of the Outlier Detection Approach

So far we have described the general approach. Now we want to show how to estimate the probability distribution $P_N(x)$. In this paper we want to concentrate on the use of parameterized models. Thus the estimation of $P_N(x)$ becomes the determination of a distribution's parameters. Firstly we want to show how to use an arbitrary parameterized distribution for $P_N(x)$ and afterwards we will describe the case of Gaussian mixture models as a special choice.

Now we assume that the distribution P_N is parameterized by a parameter vector ϑ which we want to estimate from the given training patterns $x_1, \ldots, x_n$. Unfortunately we do not know which patterns are normal and which are anomalous. Therefore we introduce an indicator variable z_i for every training pattern x_i which is either one if x_i is anomalous or zero if the respective pattern is normal. If we knew the indicator variables z_i we could use a standard estimation procedure like maximum likelihood or another appropriate approach to estimate the parameter ϑ from the normal training patterns indicated by $z_i = 0$.

Since we do not know the z_i we have to estimate them in parallel with the parameter ϑ, i.e. we have to estimate the vector $(\vartheta, z_1, \ldots, z_n)$. Thereto we can use the EM-algorithm [4] or Data Augmentation [14].

Both algorithms split up the vector $(\vartheta, z_1, \ldots, z_n)$ into ϑ and $(z_1, \ldots, z_n)$. Alternately the EM-algorithm estimates the indicator variables' distribution by its expectation value given a current estimate of ϑ and the parameter ϑ given a current estimate of the indicator variables. A convergence theorem guarantees that the EM-algorithm converges into a local maximum of the likelihood function for a large number of iterations. It is therefore a maximum likelihood estimator.

In contrast the Data Augmentation algorithm is a Bayesian approach. It alternately samples parameters ϑ given current indicator variables $(z_1, \ldots, z_n)$ and the indicator variables given the current parameter ϑ. Although the current parameters are random it can be shown that they are anyhow samples from the distribution of parameters given the training data $P(\vartheta|x_1, \ldots, x_n)$. Taking the expectation value of this distribution yields an estimate for ϑ.

The approach based on indicator variables can be interpreted as the estimation of the parameters of a mixture distribution composed of two components, namely the distribution of normal patterns $P_N(x|\vartheta)$ and the outlier distribution $P_O(x)$ weighted by $1 - \lambda$ and λ, respectively. The indicator variables assign each pattern to one of the mixture components, either to the outlier component or to the component of normal patterns. The only difference from the standard applications of mixture distributions is the fact that the outlier component has no parameters to estimate.

4 Gaussian Mixture Models

In this section we want to describe a special choice for the distribution $P_N(x)$, namely Gaussian mixture models (GMMs). They are a very flexible class of probability distributions and thus they are well suited for the modeling of unknown data. Subsequently we will give a brief review of GMMs and explain the peculiarity of GMM fitting in our framework.

Given d-variate patterns the density of a GMM is defined as:

$$p(x) = \sum_{j=1}^{k} \frac{w_j}{\sqrt{(2\pi)^d \cdot det(\Sigma_j)}} \exp\left(-\frac{1}{2}(x - \mu_j)^T \Sigma_j^{-1}(x - \mu_j)\right) \tag{4}$$

A GMM can be understood as the combination of k Gaussian distributions. Each component j is described by its mean vector μ_j, its covariance matrix Σ_j and its contribution to the mixture $w_j \geq 0$ which we name the mixing weight. w_j can be understood as the prior probability of a pattern belonging to the j-th component. Thereto the mixing weights have to fulfill the condition $\sum_{j=1}^{k} w_j = 1$. The complete parameter vector of a GMM contains the mixing weights, means and covariances of each component: $\vartheta = (w_1, \ldots, w_k, \mu_1, \ldots, \mu_k, \Sigma_1, \ldots, \Sigma_k)$. The number of components k controls the amount of distributions which can be approximated by a GMM. A survey of the topic of mixture models is given in [7].

Up to now the modeling is a two-level approach: on the top level the complete data distribution $P(x)$ is a mixture with two components, the distribution of normal patterns and the outlier distribution. On the second level the distribution of normal patterns is again modeled as a mixture, i.e. a GMM. These two mixtures can be unified such that the overall distribution $P(x)$ is a single mixture distribution with a) a single fix component modeling the outliers and b) several Gaussians which constitute the distribution of normal patterns. The mixing weight of the outlier component is λ and the mixing weights of the components of normal patterns sum to $1 - \lambda$.

The estimation of the parameters of the complete mixture can again be done using either the EM-algorithm or Data Augmentation. The necessary extension from the 2-component case described above and the case with k components is that the indicator variables can now take on values in the range of $1, \ldots, k$.

The EM-algorithm has two disadvantages: a) the dependency of the result from the initialization and b) the danger of overfitting. Both problems are not really serious if the parameters of all components are adapted and the number of components is chosen adequately. But the first difficulty becomes a snare when at least one component is not adapted to the data. This phenomenon is explained by the fact that the EM-algorithm locally maximizes the likelihood-function and that the existence of the outlier component bounds the log-likelihood below. Think of a training pattern x lying in an area of input space where the density of the outlier component is much greater than the density of the other components according to a current parameter vector ϑ_t. Then there are two aspects to consider: 1.) the influence of the pattern x onto the non-outlier components is very small and 2.) changing the parameter vector does not reduce the contribution of pattern x to the log-likelihood function since its contribution is bounded below by the logarithm of the the outlier component's density at x. Thus the next parameter vector ϑ_{t+1} is computed without considering x. If x is an outlier this behavior is desired but if x is a normal pattern at the edge of the distribution this behavior leads to a GMM fitting that treats x by mistake as an outlier.

Due to the described problem we used Data Augmentation for all experiments. It has the advantage to search the parameter space globally. Therefore it does not get stuck in local optima and the result does not depend so much from the initial choice of parameters.

As priors for Data Augmentation we used non-informative priors which do not bias the resulting parameters. Additionally we used Data Augmentation to control the number of components k of the mixture distribution: we started with a large number of components and examined in every iteration whether there is a component j without assignments, i.e. all indicator variables z_i are $\neq j$. Then this component was deleted. A similar proceeding is described in [5].

5 Determining the Outlier Distribution

An essential of our approach is the use of an explicit outlier distribution P_O and an outlier rate λ. They cannot be estimated from the training data but have to be derived from a model of outlier generation.

If there is no explicit model of outlier generation we have to determine P_O and λ as plausible as possible. E. g. in an application working on measured data a typical outlier is generated by an erroneous shift of a decimal point when copying the measured value. Thus a plausible outlier distribution is normal with ten times the standard deviation of the normal data.

But, however, the determination of the outlier distribution and the outlier rate is delicate. A change in these parameters may influence the result a lot.

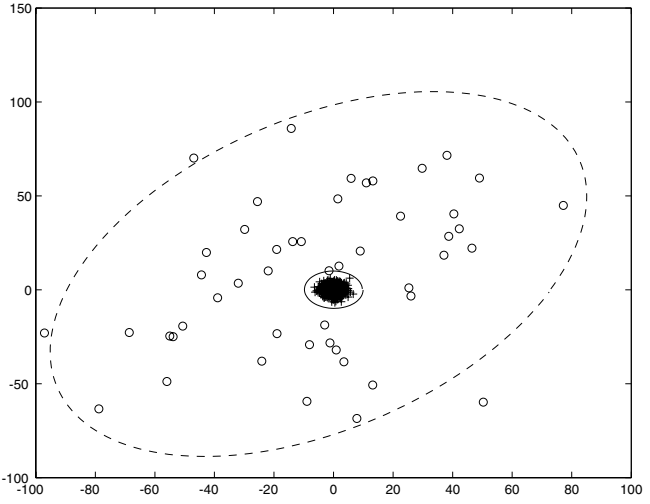

Fig. 1. Influence of λ onto the result of the outlier detection. The pattern set contains 1000 normal patterns indicated by "+" and 50 outliers indicated by "o". The dashed line shows the computed contour discriminating outliers from normal patterns for $\lambda = 0.01$ while the solid line shows the discriminating contour for $\lambda = 0.05$. The assumed outlier distribution P_O was set to be Gaussian with covariance equal to 100 times the estimated covariance of the pattern set. Further increasing λ would lead the algorithm to classify normal patterns by mistake as outliers.

In any case, the outlier distribution should be widespread since we expect the outliers to lie widespread in the data space.

A possibility to determine the outlier distribution and λ is to choose P_O to be an arbitrary widespread distribution, e.g. a Gaussian with very large variances or an improper distribution with constant density. Running our approach with different values for λ and fix P_O results in different outlier quota. Thus we can determine λ so that the resulting outlier rate resembles the expected proportion. Although this way of determining λ does not match the theoretical analysis exactly it can successfully be used in practice. Certainly, λ cannot be interpreted as the outlier rate anymore but rather as a parameter of the algorithm to adjust the sensitivity against anomalous patterns.

Furthermore, if a validation set with tagged outliers is available or if we know the outlier rate in a validation set we can use cross-validation to calibrate the unknown λ. Figure 1 shows the outcome of the presented algorithm for different values of λ on a 2-dimensional pattern set. Note that a modification of λ does not only influence the classification of patterns given a fix distribution of normal patterns but also influences the estimation of the distribution of normal patterns itself: a smaller λ increases the number of training patterns which contribute to the distribution of normal data.

6 Experimental Study

Firstly we want to illustrate the algorithm's behavior in the case where P_N is a Gaussian distribution. Therefore we used an artificial training set composed of 300 bivariate normal patterns distributed from a Gaussian distribution with zero mean vector and diagonal covariance matrix with both entries 4. 15 outliers were added derived from a Gaussian with also zero mean and a diagonal covariance matrix with both entries 400.

We set the distribution P_O to be Gaussian with the mean vector equal to the estimated mean of the complete training data and the covariance matrix equal to 100 times the estimated covariance of the complete training data. The parameter λ varied between 0.01 and 0.999.

We trained two Gaussian distributions on the training data. The first one was trained with the new approach described above and the second one was fitted using a maximum likelihood approach. Of course the maximum likelihood approach did not consider the existence of outliers and thus the second estimate was corrupted by the outliers. In contrast the first estimate was very similar to the original distribution of normal data for all $\lambda \in [0.01, 0.9]$. Larger values of λ produced too many false outliers on a test set of normal patterns while very small values did not avoid the estimate to be corrupted by the outliers. The resulting estimates for the maximum likelihood estimate and the new approach are illustrated by their covariance ellipses in Fig. 2. Although the choice of λ was not critical in this example it is in general not easy to calibrate.

In a second experimental study we investigated the Biomed dataset [3] from the StatLib archive [13]. This benchmark has already been used in [2] to analyze the performance of a novelty detection algorithm.

The Biomed data are taken from a study of medical diagnosis. The aim is to detect the carriers of a rare disease. The patterns consist of four measurements on blood samples. 127 patterns of healthy patients and 67 of carriers are available. We used 27 patterns of healthy patients and 57 patterns of carriers as test set and the remaining patterns for training. The training sets were composed of

1. 100 patterns from healthy patients, no carriers
2. 100 patterns from healthy patients and 5 patterns from carriers
3. 100 patterns from healthy patients and 10 patterns from carriers

As model for P_N we used a GMM with a variable number of components. The training was performed by Data Augmentation. The outlier distribution was set to be Gaussian with the mean equal to the mean of the training patterns and the covariance equal to 100 times the covariance of the training patterns.

Figure 3 shows the error rates on the test sets for the model trained on the second training set with various λ between 0.001 and 0.999. Comparing this figure with Fig. 4 in [2] shows two important similarities:

- the rate of undetected outliers is less than 60% even if the parameters of the algorithms are chosen inappropriate (small λ)

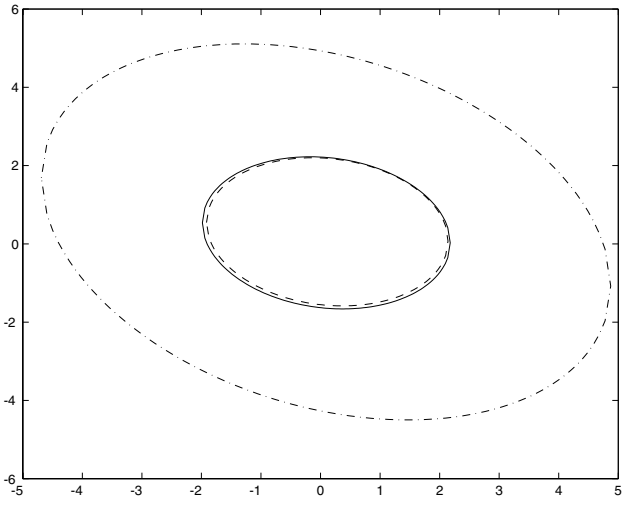

Fig. 2. Estimated covariance ellipses for the training patterns derived from a Gaussian distribution. The solid line shows the covariance ellipse of the true distribution while the dashed line shows the estimated distribution computed with the above presented new approach with $\lambda = 0.5$. The dash-dot line gives the maximum likelihood estimate.

– the point of intersection between the rate of misclassified normal patterns and undetected carriers is about 15% in both approaches

These circumstances suggest that both algorithms lead to a comparable result. But in contrast to the Linear Programming approach of [2] our algorithm was trained on data which contained 5% anomalous patterns while the Linear Programming approach was trained on outlier-free data. Figure 4 shows that even an increase in the rate of anomalous patterns in the training set does not worsen the results critically.

In a third experiment we want to show that our approach is also able to model more complex data distributions. Thereto we used an artificial bivariate data set. It consisted of 300 (1000) normal patterns which are distributed in the shape of a horseshoe. Additional 15 (50) outliers were randomly generated according to a uniform distribution in a square area of the two-dimensional plane.

The outlier distribution P_O was again set to be Gaussian with 100 times the covariance of the training patterns. As a model for normal patterns we used a GMM with variable number of components. The Fig. 5 and 6 show the 315 (1050) training patterns and the computed contour discriminating outliers from normal patterns. In both cases the algorithm recognized that the distribution of normal patterns is not convex and it classified the anomalous patterns in the inner of the horseshoe correctly as outliers.

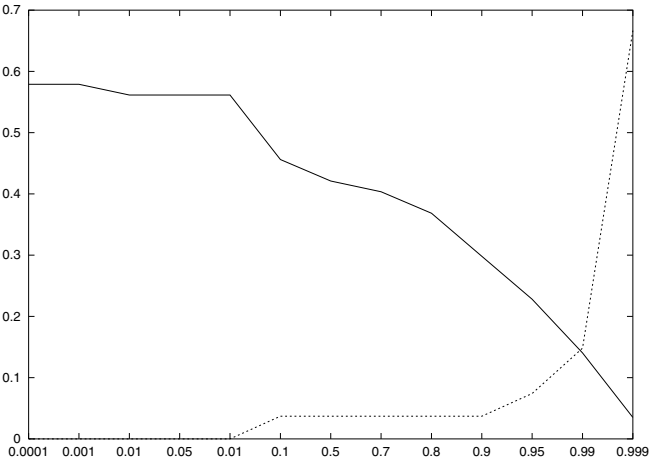

Fig. 3. Error rates (on the y-axis) for the Biomed data versus λ. The solid line shows the rate of undetected outliers, the dotted line shows the rate of misclassified normal patterns. Note that the labels on the x-axis are not equidistant.

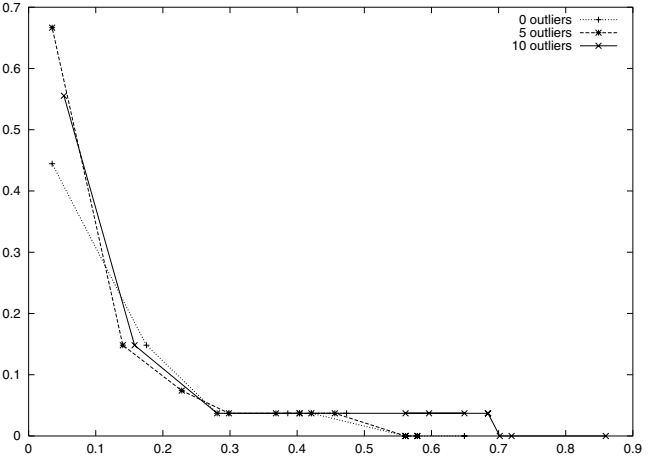

Fig. 4. Rate of misclassified outliers (x-axis) versus rate of misclassified normal patterns on a test set in the Biomed data domain. The three models are trained on an outlier-free training set (dotted line) of 100 patterns, a training set with additional 5 outliers (dashed line) and a training set with additional 10 outliers (solid line).

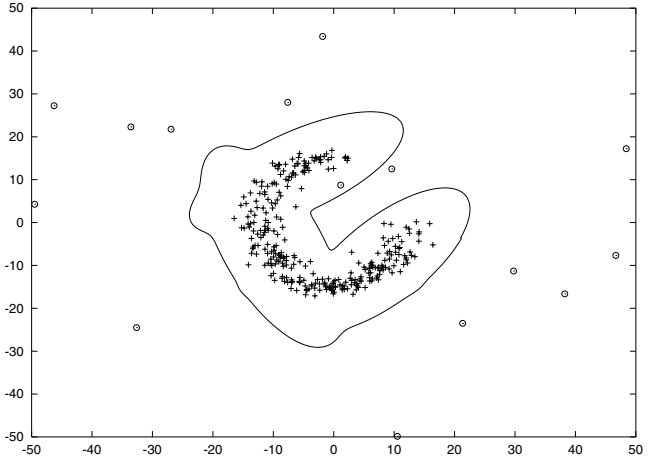

Fig. 5. Training patterns of the horseshoe data. Outliers are indicated by "o" and normal patterns by "+". The solid line gives the contour which was computed to discriminate outliers from normal patterns. λ was set to 0.05. The number of training patterns was 300 normal and 15 anomalous patterns.

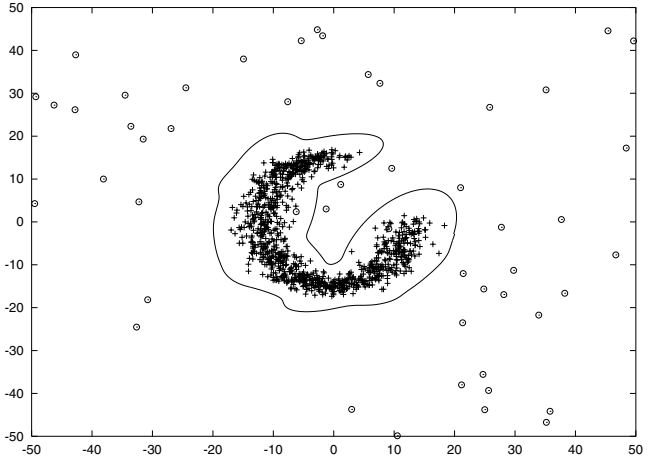

Fig. 6. Training patterns of the horseshoe data. Outliers are indicated by "o" and normal patterns by "+". The solid line gives the contour which was computed to discriminate outliers from normal patterns. λ was set to 0.2. The number of training patterns was 1000 normal and 50 anomalous patterns.

7 Discussion and Future Work

In this paper we developed an approach to deal with outliers in training data. The key idea is to explicitly model the occurrence of outliers with an outlier distribution. Therefore we use indicator variables which assign every pattern to either the set of normal patterns or the set of outliers. Since we do not know the outliers we have to estimate the indicator variables. This modeling is closely related to the estimation of a mixture distribution with two components: the distribution of normal patterns and the distribution of outliers.

This allows us to estimate the distribution of the normal patterns from training data that contain a small amount of outliers: the outlying patterns are assigned to the outlier distribution and thus do not disturb the estimation of the normal data distribution. In a second step we can use this distribution to detect the anomalous patterns in the training set or in test data.

A difficult task is the determination of the outlier distribution and the outlier rate. Mostly there are not enough outliers in the training data so that a reliable estimation is not possible. Thus these parameters have to be set explicitly. We propose the use of an arbitrary widespread distribution and to vary the outlier rate. Cross validation helps to find a suitable parameter. In our future work we hope to find a rule of thumb how to set these parameters depending on the dimensionality of the data and the spread of the training patterns.

Another problem sometimes occurs when the normal pattern distribution is modeled by a flexible class of distributions like GMMs. Then it may happen that a component of the GMM specializes on the outlying patterns or on a subset of the outliers. Especially if the GMM parameters are estimated with the EM-algorithm such a component overfits the outlying data. Therefore these outliers cannot be detected. If instead Data Augmentation with non-informative priors is used this problem is not so serious because Data Augmentation does not overfit the data but estimates a very broad component with a very small density which is often smaller than the density of the outlier distribution. Thus the outliers are found anyhow. In our future work we will focus on this phenomenon and examine in which way the outlier distribution influences the occurrence of such a component.

If we use a model for the distribution of normal patterns which is not so flexible like a single Gaussian distribution the above described problem does not occur since fitting the distribution of normal patterns to outliers would seriously worsen the description of normal patterns.

The experimental results presented in this paper show that our approach can successfully be applied in practical studies. It bounds the influence of the outliers on the estimation of the distribution of normal patterns and it is able to model even complex data distributions.

Acknowledgments. Thanks to Martin Riedmiller for helpful comments that improved this paper.

References

[1] Vic Barnett and Toby Lewis. *Outliers in Statistical Data*. John Wiley & Sons, 1978.

[2] Colin Campbell and Kristin P. Bennett. A linear programming approach to novelty detection. In *Advances in Neural Information Processing Systems 13 (to appear)*, 2001.

[3] L. H. Cox, M. M. Johnson, and K. Kafadar. Exposition of statistical graphics technology. In *ASA Proceedings of the Statistical Computation Section*, pages 55–56, 1982.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B*, 39:1–38, 1977.

[5] Jean Diebolt and Christian P. Robert. Estimation of finite mixtures through bayesian sampling. *Journal of the Royal Statistical Society Series B*, 56(2):363–375, 1994.

[6] Nathalie Japkowicz, Catherine Myers, and Mark Gluck. A novelty detection approach to classification. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 518–623, 1995.

[7] Geoffrey McLachlan and David Peel. *Finite Mixture Models*. John Wiley & Sons, 2000.

[8] Alberto Munõz and Jorge Muruzabál. Self-organizing maps for outlier detection. *Neurocomputing*, 18(1-3):33–60, 1998.

[9] Alexandre Nairac, Timothy A. Corbett-Clark, Ruth Ripley, Neil W. Townsend, and Lionel Tarassenko. Choosing an appropriate model for novelty detection. In *Proceedings of the Fifth International Conference on Artificial Neural Networks*, pages 117–122, 1997.

[10] Stephen Roberts and Lionel Tarassenko. A probabilistic ressource allocation network for novelty detection. *Neural Computation*, 6(2):270–284, 1994.

[11] Peter J. Rousseeuw and Annick M. Leroy. *Robust Regression and Outlier Detection*. John Wiley & Sons, 1987.

[12] Bernhard Schölkopf, Robert C. Williamson, Alex Smola, and John Shawe-Taylor. SV estimation of a distribution's support. In *Advances in Neural Information Processing Systems 12*, pages 582–588, 2000.

[13] Statlib–datasets archive. cf. `http://lib.stat.cmu.edu/datasets`.

[14] Martin A. Tanner and Wing Hung Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Society*, 82(398):528–550, 1987.

[15] Geoffrey G. Towell. Local expert autoassociators for anomaly detection. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 255–262, 2000.

[16] Alexander Ypma and Robert P. W. Duin. Novelty detection using self-organizing maps. In *Progress in Connectionist-Based Information Systems*, volume 2, pages 1322–1325, 1997.