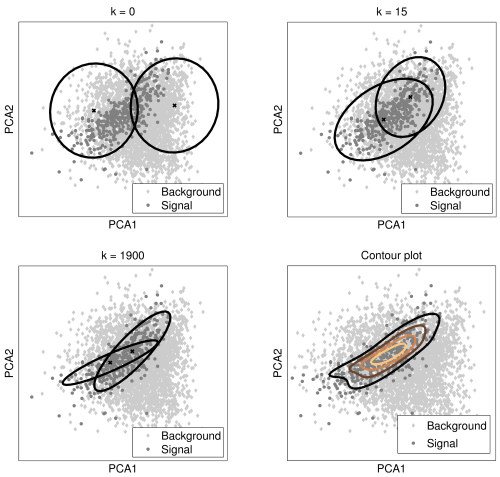
Fixed-Background EM Algorithm for Semi-Supervised Anomaly Detection

Tommi Vatanen, Mikael Kuusela, Eric Malmi, Tapani Raiko, Timo Aaltonen and Yoshikazu Nagai



Aalto University

SCIENCE + TECHNOLOGY

TECHNICAL REPORT

Aalto University publication series
SCIENCE + **TECHNOLOGY** 22/2011

Fixed-Background EM Algorithm for Semi-Supervised Anomaly Detection

Tommi Vatanen, Mikael Kuusela, Eric Malmi, Tapani Raiko, Timo Aaltonen and Yoshikazu Nagai

Aalto University
School of Science
Department of Information and Computer Science
Adaptive Informatics Research Centre

Aalto University publication series
SCIENCE + **TECHNOLOGY** 22/2011

ISBN 978-952-60-4319-7 (pdf)
ISSN-L 1799-4896
ISSN 1799-490X (pdf)

Aalto Print
Helsinki 2011

Finland

Author
Tommi Vatanen, Mikael Kuusela, Eric Malmi, Tapani Raiko, Timo Aaltonen and Yoshikazu Nagai

Name of the publication
Fixed-Background EM Algorithm for Semi-Supervised Anomaly Detection

Publisher School of Science

Unit Department of Information and Computer Science

Series Aalto University publication series SCIENCE + TECHNOLOGY 22/2011

Field of research Computer science

Abstract

We study a semi-supervised anomaly detection problem where anomalies lie among the normal data. Instead of analyzing individual observations, anomalies are identified collectively based on deviations from the distribution of the normal data. We first model the normal data using a mixture of Gaussians and then use a variant of the EM algorithm to fit a mixture of the normal model and a number of additional Gaussians to an unlabeled data set. The statistical significance of the model is verified using a likelihood ratio test based on nonparametric bootstrapping. Using artificial data, we show that the proposed methodology provides accurate models for the anomalous data and good estimates for the proportion of anomalies in the sample. We apply the method to the search of the Higgs boson in particle physics and show that it is applicable to this type of tasks with little a priori knowledge of the new phenomenon.

Keywords anomaly detection, semi-supervised learning, EM algorithm, mixture of Gaussians, nonparametric bootstrap, high energy physics

ISBN (printed) **ISBN (pdf)** 978-952-60-4319-7

ISSN-L 1799-4896 **ISSN (printed)** 1799-4896 **ISSN (pdf)** 1799-490X

Location of publisher Espoo **Location of printing** Helsinki **Year** 2011

Pages 31

Contents

Contents **3**

1 Introduction **5**

1.1 Related Work 6

2 Fixed-Background Model for Anomaly Detection **9**

3 Methods **11**

3.1 Mixture of Multivariate Gaussian Distributions 11

3.2 EM Algorithm for the Normal Model 11

3.3 The Fixed-Background EM Algorithm 12

3.4 Additional Remarks 13

3.5 Statistical Significance of the Anomaly Model 15

4 Experiments with Artificial Data **17**

5 Demonstration: Search for the Higgs Boson **21**

5.1 Description of the Data Set 21

5.2 Modeling the Higgs Data 22

5.3 Anomaly Detection Results 23

6 Discussion **25**

7 Conclusions **27**

Bibliography **29**

1 Introduction

Anomaly detection refers to the process of locating observations (instances, events, data points) in a collection of data which deviate from what is standard, normal or expected. The term *anomaly* is often used interchangeably with *outlier* or *novelty* and refers to "an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism" [9]. Thus, to detect anomalies, a reference model on what is normal is needed. Typically, a training data set of normal observations is used either directly or indirectly by, e.g., training a statistical model to estimate the probability density of normal data.

Anomaly detection has applications, e.g., in credit card fraud detection, network intrusion detection, aircraft engine damage detection, video and electronic surveillance and health-care informatics. In each problem, the nature of the data and the notion of an anomaly varies greatly which makes anomaly detection a very complex and diverse problem domain.

Semi-supervised anomaly detection is a particular anomaly detection problem with two data sets. The first one is called the normal data, and it is labeled as not containing anomalies. The second data set is unlabeled and can contain both normal and anomalous observations. Traditional anomaly detection methods would compare observations in the second data set one-by-one to the model created from the normal data, and label unexpected instances as anomalies. Instead of this, we propose to study the unlabeled data set as a whole: We form a probabilistic model for the second data set which is a mixture of the normal model and an additional anomaly model. This way we can detect anomalous observations even if they lie among the normal data, as long as the distribution changes strongly enough. Such previously unobserved patterns are sometimes called *collective anomalies* which according to Chandola et al. [2] are "a subset of instances that occur together as a collection and whose occurrence is not normal with respect to a normal behavior".

The proposed model has many desirable features. First, it is fully proba-

bilistic, thus providing models and outputs that can be easily interpreted. Second, there is a single model parameter that directly gives an estimate for the amount of anomalies in the unlabeled data. Third, the approach has a wide application potential with diverse data sets as the Gaussian distributions used in this study can be easily replaced with any other parametric distribution.

To achieve this, we make a number of assumptions on the problem setting. First, we assume that the normal data has a fixed distribution. In particular, the method is not applicable to situations where there are temporal changes in the normal data. Second, anomalies are assumed to occur collectively, i.e., a single isolated anomalous observation might not be detected. Also, too small of a proportion of anomalous events, say less than a few percent, handicaps the method. Third, anomalies are assumed to occur as an excess in the distribution of the normal data.

This paper is organized as follows: We start with a brief overview of recent work on anomaly detection below in Sect. 1.1. We then introduce our probabilistic anomaly detection model in Sect. 2 and describe the fixed-background expectation-maximization (EM) algorithm for locating the anomalous patterns in Sect. 3. This section also provides implementational details of the algorithm and the means for testing the statistical significance of the anomaly model. In Sections 4 and 5, we report the performance of the method with artificial data and a data set from high energy physics related to the search of the Higgs boson. We discuss other potential applications of the method and directions for future work in Sect. 6 and summarize our findings of Sect. 7.

1.1 Related Work

There are many comprehensive reviews available about the domain of anomaly detection. A recent survey by Chandola et al. [2] covers practically the whole field. Markou and Singh [14, 15] cover statistical and neural network based approaches. They point out that the most of the statistical techniques are based on modeling the normal data and classifying observations that fall in the regions of low density as anomalous.

Previously, Argwal [1], Eskin [7] and Lauer [13], among others, have used parametric mixture models in anomaly detection. Eskin [7] used the EM algorithm to train a mixture model to represent the normal and

anomaly classes. His technique requires some prior knowledge about the classes of the observations and, moreover, he uses a uniform distribution to model the anomalous data. A common assumption in most of the literature is that anomalies have a more widespread distribution compared to the normal data.

In background subtraction (see, e.g., [21]), one uses separate models for the normal and the unlabeled data sets. Anomalies can then be detected by calculating the difference between the two models. However, this approach does not allow for probabilistic interpretation of the results.

In semi-supervised anomaly detection, there has recently been at least neural network, support vector machine and Markov model based approaches. Hawkins [10] uses a replicator multi-layer feed-forward neural network to form a compressed model of the normal data. Anomalies can then be detected by their large reconstruction errors. In the field of support vector machines, this kind of a problem is usually referred to as one-class classification and is studied, for instance, in [8, 22, 24].

Statisticians have approached similar problem settings as ours using empirical distribution function based goodness-of-fit tests, such as the Anderson-Darling test [4]. Such tests are however only able to indicate if the observed data follows the hypothesized distribution while our framework also performs pattern recognition for the anomalies. Wang et al. [26] propose another hypothesis testing based anomaly detection method in the context of detecting nuclear explosions. While extending this work, Sain et al. [23] also show how the method fails if the nuclear blasts are too similar compared to other seismic activity.

2 Fixed-Background Model for Anomaly Detection

When the anomalies are among the normal data, an event-by-event classification is usually difficult. Nevertheless, one can detect changes in the distribution of the data—there are more observations in the regions containing anomalies than one would expect according to the model of the normal data. To detect such changes, we proceed in two steps. First, we utilize parametric density estimation to learn a *normal model*, $p_{\mathrm{N}}(\boldsymbol{x})$ using the labeled normal data. The next step is to model the unlabeled data with a *fixed-background model*, $p_{\mathrm{FB}}(\boldsymbol{x})$, which is a mixture of the normal model and a new *anomaly model* $p_{\mathrm{A}}(\boldsymbol{x})$:

$$p_{\mathrm{FB}}(\boldsymbol{x}) = (1-\lambda)p_{\mathrm{N}}(\boldsymbol{x}) + \lambda p_{\mathrm{A}}(\boldsymbol{x}). \tag{2.1}$$

Above, the anomaly model, $p_{\mathrm{A}}(\boldsymbol{x})$, represents the unexpected data and λ is the proportion of anomalous observations in the model.

Figure 2.1a illustrates a univariate data set of normal data generated from a Gaussian distribution and a maximum likelihood Gaussian density $p_{\mathrm{N}}(\boldsymbol{x})$ estimated using the data set. Figure 2.1b shows a very simple anomalous pattern that can be modeled with a single additional univariate Gaussian. Given a sample contaminated with these anomalies, our goal is to find an optimal combination of the parameters of the anomaly model (μ_{A}, σ_{A}) and the mixing coefficient λ in (2.1). The resulting model $p_{\mathrm{FB}}(\boldsymbol{x})$ is shown with a black line and the anomaly model $p_{\mathrm{A}}(\boldsymbol{x})$ with a gray line in Fig. 2.1b.

For an event-by-event anomaly detection a discriminant function $\mathcal{D}(\boldsymbol{x})$ is needed. A natural choice is to use the posterior probability

$$p(\text{anomaly}|\boldsymbol{x}) = \frac{\lambda p_{\mathrm{A}}(\boldsymbol{x})}{(1-\lambda)p_{\mathrm{N}}(\boldsymbol{x}) + \lambda p_{\mathrm{A}}(\boldsymbol{x})} \equiv \mathcal{D}(\boldsymbol{x}). \tag{2.2}$$

The decision rule for selecting events is as follows

$$\mathcal{D}(\boldsymbol{x}) = \begin{cases} \geq T \;\Rightarrow\; \boldsymbol{x} \text{ is an anomaly}, \\ < T \;\Rightarrow\; \boldsymbol{x} \text{ is normal}, \end{cases} \tag{2.3}$$

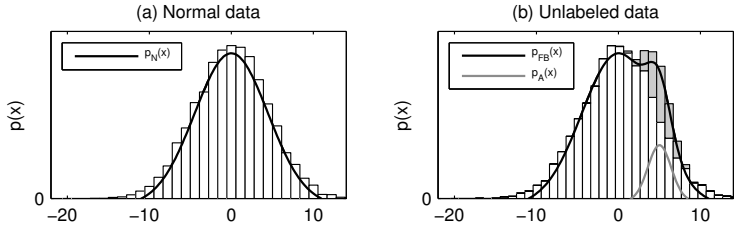


Figure 2.1. (a) A histogram of a one dimensional data set of normal data from a Gaussian distribution and an estimated normal model $p_{\mathrm{N}}(x)$. (b) An illustration of the fixed-background model in a univariate case. The histogram shows the unlabeled data (the light gray excess in the histogram denotes the anomalous observations) and the plot shows the fixed-background model estimated using the data. The fixed-background model $p_{\mathrm{FB}}(x)$ is shown with a black line and the anomaly model $p_{\mathrm{A}}(x)$ with a gray line.

where the constant $T \in [0, 1]$ is a threshold which can be used to control the sensitivity of the classifier. As extreme cases, if $T = 0$, all events are classified as anomalies, and if $T = 1$, all events are classified as normal.

3 Methods

In this section, we describe the methods used in our experiments. We first review the EM algorithm for multivariate mixtures of Gaussians (MoG) and then describe in detail a specific variant of the algorithm for learning the fixed-background model (2.1). We conclude this section by showing how the statistical significance of the model can be verified using non-parametric bootstrapping.

3.1 Mixture of Multivariate Gaussian Distributions

Finite mixtures of distributions are a flexible method for modeling complex data sets [16]. In this work, we use mixtures of multivariate Gaussian distributions or shortly mixtures of Gaussians (MoG) to represent the distribution of the data. Even though the data might not in reality be a sample from a MoG, it can often be modeled with a sufficient accuracy using a mixture of Gaussian components. The mixture of J multivariate Gaussian distributions is defined as

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_{j=1}^{J} \pi_j \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j), \tag{3.1}$$

where $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ denotes the probability density of a multivariate Gaussian with mean $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{\Sigma}_j$ at $\boldsymbol{x}$. The π_j are mixture proportions (or mixing coefficients) which satisfy $\pi_j \geq 0$ and $\sum_{j=1}^{J} \pi_j = 1$, and $\boldsymbol{\theta} = \{\pi_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\}_{j=1}^{J}$ represents the parameters of the mixture model with J components.

3.2 EM Algorithm for the Normal Model

Let us first consider the case of fitting a MoG model with J components to the normal data with N observations $\boldsymbol{x}_i, i = 1, \ldots, N$. The log-likelihood

of the parameters $\boldsymbol{\theta}$ is

$$l(\boldsymbol{\theta}) = \log(\mathcal{L}(\boldsymbol{\theta})) = \sum_{i=1}^{N} \log \left(\sum_{j=1}^{J} \pi_j \mathcal{N}(\boldsymbol{x}_i | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \right). \tag{3.2}$$

Here we have assumed that the observations are independent and identically distributed (i.i.d.).

The maximum likelihood (ML) estimate of the parameters can be obtained by maximizing (3.2) which is carried out by using the EM algorithm [5, 18]. The algorithm proceeds in two steps. In the *expectation step* (E-step), the posterior probabilities for each data point $\boldsymbol{x}_i$ being generated by the jth component

$$p(z_{ij} = 1 | \boldsymbol{x}_i, \boldsymbol{\theta}^k) = \frac{\pi_j^k \mathcal{N}(\boldsymbol{x}_i | \boldsymbol{\mu}_j^k, \boldsymbol{\Sigma}_j^k)}{\sum_{j'=1}^{J} \pi_{j'}^k \mathcal{N}(\boldsymbol{x}_i | \boldsymbol{\mu}_{j'}^k \boldsymbol{\Sigma}_{j'}^k)} \equiv \gamma_{ij}^k \tag{3.3}$$

are computed. Here, $\boldsymbol{\theta}^k$ contains the parameter estimates at the kth iteration and $\boldsymbol{z}_i$ indicates the component which generated the ith observation.

In the subsequent *maximization step* (M-step), the parameter values are updated according to the following equations

$$\pi_j^{k+1} = \frac{1}{N} \sum_{i=1}^{N} \gamma_{ij}^k, \tag{3.4}$$

$$\boldsymbol{\mu}_j^{k+1} = \frac{\sum_{i=1}^{N} \gamma_{ij}^k \boldsymbol{x}_i}{\sum_{i=1}^{N} \gamma_{ij}^k}, \tag{3.5}$$

$$\boldsymbol{\Sigma}_j^{k+1} = \frac{\sum_{i=1}^{N} \gamma_{ij}^k (\boldsymbol{x}_i - \boldsymbol{\mu}_j^{k+1})(\boldsymbol{x}_i - \boldsymbol{\mu}_j^{k+1})^{\mathrm{T}}}{\sum_{i=1}^{N} \gamma_{ij}^k}. \tag{3.6}$$

A detailed derivation of the EM algorithm for mixtures of Gaussians can be found in [18] where it is also shown that each iteration of the EM algorithm increases the log-likelihood until a local maximum is found.

3.3 The Fixed-Background EM Algorithm

In this section, we elaborate how to use the EM algorithm to estimate models of the form (2.1). We call this variant of the algorithm the *fixed-background EM algorithm*.

The goal is to search for unmodeled anomalies in the unlabeled data set. Now, the normal model $p_{\mathrm{N}}(\boldsymbol{x})$ in equation (2.1) is fixed and both λ and the parameters of $p_{\mathrm{A}}(\boldsymbol{x})$ need to be optimized to maximize the log-likelihood. Here, $p_{\mathrm{A}}(\boldsymbol{x})$ can be either a single Gaussian or more generally a MoG with

Q components. We can now write (2.1) as follows

$$\begin{aligned} p_{\mathrm{FB}}(\boldsymbol{x}) &= (1-\lambda)p_{\mathrm{N}}(\boldsymbol{x}) + \lambda \sum_{q=J+1}^{J+Q} \tilde{\pi}_q \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q) \\ &= \pi_{\mathrm{N}} p_{\mathrm{N}}(\boldsymbol{x}) + \sum_{q=J+1}^{J+Q} \pi_q \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q), \end{aligned} \tag{3.7}$$

where we have defined $\pi_{\mathrm{N}} = 1-\lambda$ and $\pi_q = \lambda\tilde{\pi}_q$, $q = J+1, \ldots, J+Q$. The mixture proportions satisfy $\pi_{\mathrm{N}} + \sum_{q=J+1}^{J+Q} \pi_q = 1$ and $\sum_{q=J+1}^{J+Q} \pi_q = \sum_{q=J+1}^{J+Q} \lambda\tilde{\pi}_q = \lambda$. This anomaly detection model and its components are illustrated in Fig. 3.1.

The EM update equations for model (3.7) are easily found by straightforward analogy to the standard EM algorithm. In the E-step, the posterior probabilities of the normal model and the components of the anomaly MoG are updated as follows

$$p(z_{i\mathrm{N}} = 1|\boldsymbol{x}_i, \boldsymbol{\theta}^k) = \frac{\pi_{\mathrm{N}}^k p_{\mathrm{N}}(\boldsymbol{x}_i)}{\pi_{\mathrm{N}}^k p_{\mathrm{N}}(\boldsymbol{x}_i) + \sum_{q'=J+1}^{J+Q} \pi_{q'}^k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_{q'}^k, \boldsymbol{\Sigma}_{q'}^k)} \equiv \gamma_{i\mathrm{N}}^k, \tag{3.8}$$

$$p(z_{iq} = 1|\boldsymbol{x}_i, \boldsymbol{\theta}^k) = \frac{\pi_q^k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_q^k, \boldsymbol{\Sigma}_q^k)}{\pi_{\mathrm{N}}^k p_{\mathrm{N}}(\boldsymbol{x}_i) + \sum_{q'=J+1}^{J+Q} \pi_{q'}^k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_{q'}^k, \boldsymbol{\Sigma}_{q'}^k)} \equiv \gamma_{iq}^k. \tag{3.9}$$

In the first equation, $z_{i\mathrm{N}} = 1$ denotes that the ith observation was generated by the normal model $p_{\mathrm{N}}(\boldsymbol{x})$. In the second equation $q = J+1, \ldots, J+Q$. In the consequent M-step, means and covariances are updated using (3.5) and (3.6) for indices $j = J+1, \ldots, J+Q$. The mixture proportions for these indices are also updated with (3.4), while the mixture proportion of the normal model follows from the normalization constraint

$$\pi_{\mathrm{N}}^{k+1} = 1 - \sum_{q=J+1}^{J+Q} \pi_q^{k+1} \left(= \frac{1}{N} \sum_{i=1}^{N} \gamma_{i\mathrm{N}}^k \right). \tag{3.10}$$

3.4 Additional Remarks

Assessing the number of components in mixture models is a hard problem which has not been completely resolved [16]. We use the cross-validation-based information criterion (CVIC) for model selection in the Higgs experiments of Sect. 5, but take the correct number of components as given in our artificial data experiments. Naturally, any known information criterion can be used to perform model selection for the normal model. Further discussion about model selection can be found in, e.g, [16, 25].

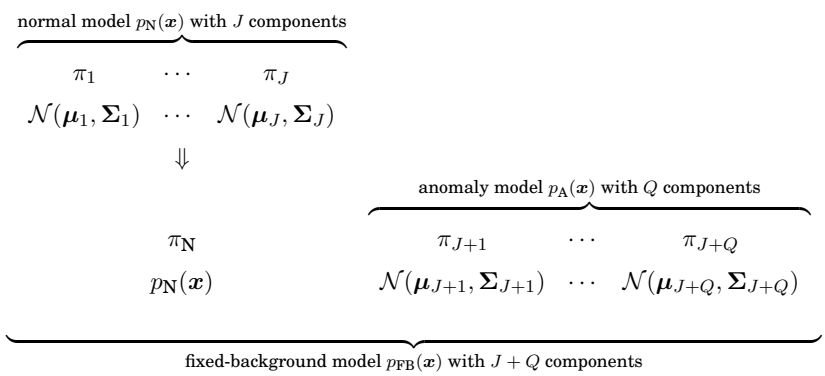


Figure 3.1. Illustration of the proposed anomaly detection model. The normal model $p_{\mathrm{N}}(\boldsymbol{x})$ and anomaly model $p_{\mathrm{A}}(\boldsymbol{x})$ are mixtures of Gaussians with J and Q components, respectively. The normal model is combined with the anomaly model with an additional mixture proportion π_{N} to give the fixed-background model p_{FB}.

The maximization of the log-likelihood function of a Gaussian mixture model is not a well-posed problem due to the singularities corresponding to one of the Gaussian components "collapsing" onto a single data point, i.e., $\sigma_j \to 0$ in the one-dimensional case. With multivariate data, this corresponds to the case where the smallest eigenvalue of the covariance matrix $\boldsymbol{\Sigma}$ tends to zero. In this work, we avoid this problem by resetting the mean of a collapsing component to a randomly chosen data point while also resetting its variance or covariance matrix to some large value. We also reset the components with a very small mixture proportion to avoid unnecessary nuisance components.

We assess the "goodness" of the components in the anomaly model using a simple likelihood comparison. Using the likelihood of the normal model as a reference, we take components of the anomaly model one at a time and combine them with the normal model. Components that have learned some anomalous patterns in the unlabeled data should increase the likelihood compared to the normal model. On the other hand, if the component under investigation decreases the likelihood, it is most probably useless. Again, components that do not appear to capture any anomalies in the data are reset to a random data point.

We also exploit the resetting heuristics above in order to remove excess components from the anomaly model. We assume that a component can be removed if it has been reset too many times and, consequently, hinders the convergence of the fixed-background EM algorithm. Finally, while es-

timating the fixed-background model, the convergence of the algorithm is denied if the fixed-background model decreases the log-likelihood compared to the normal model. Instead, poor components are reset and the EM iteration continues until a model that increases the log-likelihood is found or all anomalous components have been removed.

3.5 Statistical Significance of the Anomaly Model

Once we have fitted the fixed-background model $p_{\mathrm{FB}}(\boldsymbol{x})$ to the unlabeled data that potentially contains anomalies, we should be able to say if the anomaly model represents statistical fluctuations in the normal data or a real anomalous contribution. To this end, we perform a likelihood ratio test for the significance of $p_{\mathrm{FB}}(\boldsymbol{x})$ (see e.g. [11, 3]). We test the background-only null hypothesis H_0, i.e., $\pi_{\mathrm{N}} = 1$, against the anomaly hypothesis $\pi_{\mathrm{N}} < 1$. The test is based on the statistic

$$\Lambda = \frac{\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} \mathcal{L}(\boldsymbol{\theta})}{\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\theta})}, \tag{3.11}$$

where $\boldsymbol{\Theta}_0$ refers to the set of parameters allowed by the null hypothesis and $\mathcal{L}(\boldsymbol{\theta})$ is the likelihood function. In our case, the nominator is simply the likelihood of the normal model and the denominator the likelihood of the fixed-background model.

Small values of Λ give evidence against the null hypothesis. We may also equivalently reject the null for large values of the test statistic

$$D = -2 \log \Lambda. \tag{3.12}$$

A result known as Wilks' theorem states that under certain regularity conditions D is asymptotically χ^2 distributed under the null H_0. Unfortunately, these conditions are not satisfied for mixture models [17] and hence one needs to consider alternative methods for recovering the distribution of D.

We follow the approach taken by Wang et al. [26] and use nonparametric bootstrap simulation [6] to estimate the distribution of D. The algorithm is as follows:

1. Sample with replacement N observations from the normal data set used to learn the normal model $p_{\mathrm{N}}(\boldsymbol{x})$. Here N equals to the number of data points in the unlabeled data set.

2. Use the fixed-background EM algorithm to learn $p_{\mathrm{FB}}(\boldsymbol{x})$.

3. Compute D.

4. Repeat from 1. until R observations of D have been obtained.

Note that also parametric bootstrapping where one generates samples from $p_{\mathrm{N}}(\boldsymbol{x})$ could have been used. However, the nonparametric version makes the test more robust against misspecification of $p_{\mathrm{N}}(\boldsymbol{x})$.

The obtained bootstrap sample allows us to estimate the $100(1-\alpha)$th percentile of the distribution of D. Let us denote this by D_α. We then reject the null hypothesis H_0 at a significance level α if $D_{\mathrm{obs}} \geq D_\alpha$, where D_{obs} denotes the value of D for the unlabeled data set. Additionally, the simulated distribution of D can be used to obtain the p-value of D_{obs}.

4 Experiments with Artificial Data

We test the fixed-background EM algorithm with artificial data generated from mixtures of Gaussians. Two data sets are generated for each model: a collection of normal data for training the normal model and an unlabeled test data set consisting of some small amount of anomalies among a new sample of normal data. The data are generated using five components for the normal data and three additional anomalous components for the test data. The means and variances of the components are randomly generated in such a way that the anomalies appear as clusters among the normal data. Figure 4.1 shows examples of anomalous instances denoted by the gray proportions on the histogram bars on top of the normal data. We use 10 different generative models and generate for each 10 different pairs of data sets consisting of 100 000 data points. Furthermore, we test each data set with different proportions of anomalies ranging from 1 % to 20 %.

For each data set we train a fixed-background model as described in Sections 2 and 3.3. The model is then used to classify the data in the test set as normal or anomalous with different thresholds according to (2.2) and (2.3). This allows us to construct the receiver operating characteristic (ROC) curves for each experiment, and use the area under the ROC curve (AUC) as a single measure for the classifier performance, $0 < \text{AUC} \leq 1$. We use the original generative model as an optimal model to obtain gold standard AUC for each test data set. We also compare the results with a traditional outlier detection model where data points at low density areas of the normal model are considered anomalies.

Figure 4.2a shows the median of the AUC values obtained using the fixed-background model. The dashed line denotes the median AUC obtained using the generative model itself as a classifier for the test data. Given that the test data contains a sufficient amount of anomalies, the resulting AUC values are practically identical. However, the robustness of the fixed-background EM algorithm starts to suffer when the test data

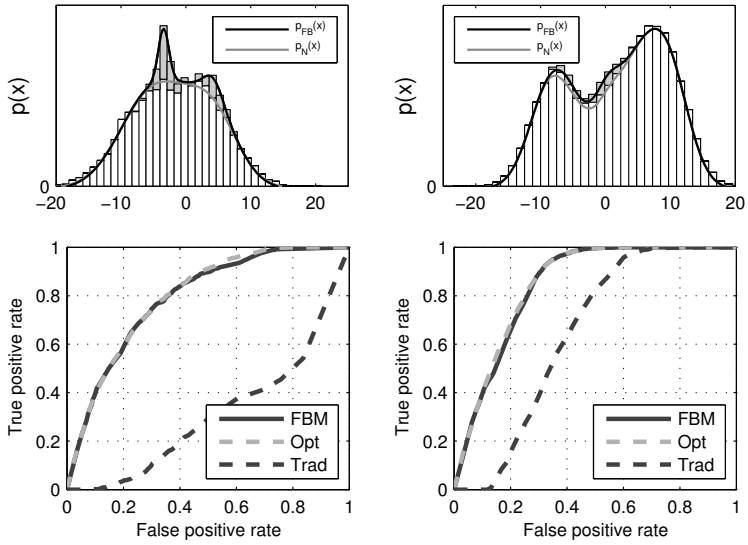


Figure 4.1. Two examples of artificial test data sets with 100 000 observations containing 10 000 (10 %, left column) and 3000 (3 %, right column) anomalous observations (light gray area on the histograms). The top row shows the estimated models such that the gray line denotes the normal model and the black line denotes the fixed-background model. The estimated anomaly proportions λ are 0.082 and 0.033, respectively. The bottom row shows the ROC curves for the models. The area under curve (AUC) is practically the same for the fixed-background model (FBM) and the optimal model (Opt). The traditional method (Trad), which treats instances that fall in regions of low normal model density as anomalies, performs significantly worse with this kind of data.

contains less than 3 % of anomalies. Figure 4.2b shows a box plot of the estimated anomaly proportions λ. The small boxes on the diagonal show the interquartile range of the estimated λs which are in good agreement with the correct results. The whiskers show the full range of the estimates. The wide downward range results from the algorithm occasionally being able to find only a portion of the anomalous data.

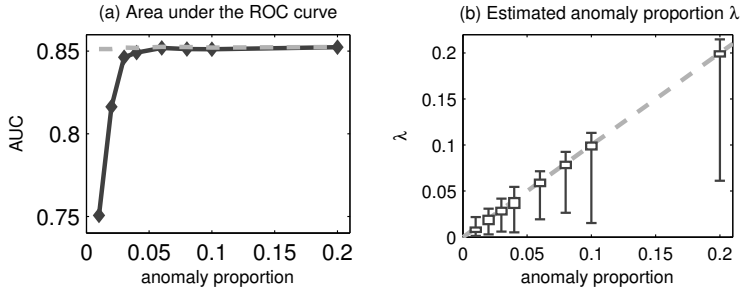


Figure 4.2. The results of our artificial data experiments with unlabeled test data sets containing 100 000 data points. (a) Comparison between AUC of the fixed-background EM (solid line) and the generative model used to generate the data (gray dashed line) with different amount of anomalies in the test data. (b) Estimation of the anomaly proportion (λ) using fixed-background EM. The small boxes show the interquartile range and the whiskers show the full range of the estimates. Gray dashed line shows the correct anomaly proportion.

5 Demonstration: Search for the Higgs Boson

We demonstrate the applicability of the fixed-background EM algorithm to real world problems by considering searches for new particles in high energy particle physics. Throughout this section, we use the terms background and signal data instead of normal and anomalous data to conform with the physics terminology. The new physics signals usually manifest themselves as an excess of certain types of collision events in particle detectors. These events can be simulated with a Monte Carlo generator, which is fairly accurate for the background data of known physics, but for the unknown new physics, the simulation might contain inaccuracies or free parameters which results in uncertainty in the exact nature of the signal. To avert the risk of missing the signal or some part of it, one would like to search for new signals without relying on any particular Monte Carlo model. Such approaches are called model-independent, one example of which is our fixed-background EM algorithm. For more information on the physics motivations of the algorithm as well as a comparison to more traditional model-dependent data analysis methods, see [12].

5.1 Description of the Data Set

We apply our method to a data set containing a simulated signal produced by the Higgs boson. This is a particle predicted by the Standard Model of particle physics to explain the mass of the other particles in the model. More precisely, we consider a data set produced by the CDF collaboration [19, 20] containing background events and Monte Carlo simulated Higgs events where the Higgs is produced in association with the W boson and decays into two bottom quarks, $q\bar{q} \to WH \to l\nu b\bar{b}$. This signal looks different for different Higgs masses m_{H} which is an unknown free parameter in the Standard Model. The advantage of the semi-supervised anomaly detection approach is that one is potentially able to detect the signal without

knowledge of m_{H}.

Each observation in the data set corresponds to a single collision event in the CDF detector at the Tevatron proton-antiproton collider. The data vectors consist of 8 variables corresponding to different characteristics of the topology of a collision event. To facilitate density estimation, the dimensionality of the logarithmically normalized data was reduced to 2 using PCA on the background data.

We used 3406 data points to train the normal model which was then used to detect signals of 400 data points of masses $m_{\mathrm{H}} = 100$, 115, 135, 150 GeV among another sample of 3406 observations of background data. Hence, the unlabeled sample contained 10.5 % of signal events. In reality, the expected signal is roughly 5 to 50 times weaker than this, but due to the limited number of background events available, the signal had to be amplified for this demonstration. As shown by the experiments with artificial data, we expect to be able to find also weaker signals should more background observations be available.

5.2 Modeling the Higgs Data

We used cross-validation-based information criterion (CVIC) [25] in order to select a suitable number of components J for the normal model. When a 5-fold cross-validation was performed, the evaluation log-likelihood was maximized with $J = 5$. Figure 5.1a shows contours of the resulting normal model in the two-dimensional principal subspace.

We then ran the fixed-background EM algorithm for the signals with different masses starting with $Q = 3$ and allowed for heuristic removal of unnecessary components as described in Sect. 3.4. The algorithm converged with one anomalous component for $m_{\mathrm{H}} = 100$ GeV and two components for the rest of the masses. The resulting anomaly model for $m_{\mathrm{H}} = 150$ GeV is shown in Fig. 5.1b.

The statistical significance of these models was then evaluated using the bootstrap technique described in Sec. 3.5 based on $R = 50000$ resamplings. It was found out that at 5 % significance level the background-only null hypothesis was rejected for all the considered mass points. Figure 5.2a shows the distribution of the test statistic and the p-values of the models. It turns out that the higher the mass, the more significant the model becomes. The peak of the test statistic distribution at the origin results from

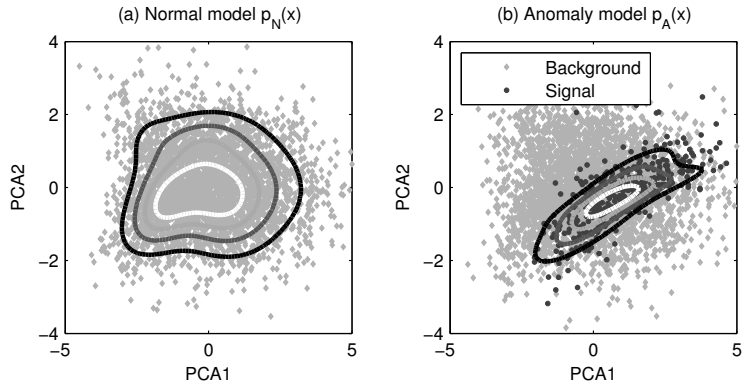


Figure 5.1. (a) A projection of the Higgs background data into its two-dimensional principal subspace. The solid lines show contours of the estimated 5-component MoG for the background. (b) A projection of the $m_{\mathrm{H}} = 150$ GeV test data set into the two-dimensional principal subspace. The solid lines show contours of the estimated 2-component MoG for the signal.

situations where all the components of the anomaly model are correctly removed by the removal heuristics.

5.3 Anomaly Detection Results

The fixed-background models can be used for event classification using (2.2) and (2.3). Figure 5.2b shows ROC curves for the classifiers with different Higgs masses. One can see that regardless of the mass of the Higgs, the fixed-background EM is able to identify the signal with a good accuracy. The classification results are slightly better with higher masses because the high-mass signal lies on a region of the data space with slightly lower background density than the low-mass signal.

From the mixture proportion λ, we get an estimate for the amount of anomalous events in the test data set. In the case of particle physics, this is proportional to the cross section of the process, the measurement of which is the typical goal of many physics analyses. Starting from the lowest mass, we get the estimates $\lambda = 0.100,\ 0.121,\ 0.118,\ 0.122$ which are all in agreement with the real proportion of 0.105.

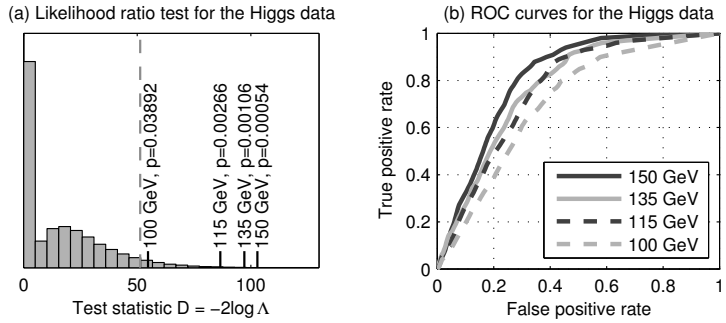


Figure 5.2. (a) Test for the significance of the anomaly model for various Higgs masses. The histogram shows the probability distribution of the likelihood ratio test statistic under the background-only null hypothesis. The vertical dashed line shows the critical value of the test at 5 % significance level and the black markers denote the test statistics for the fixed-background models with respective p-values. All observed test statistics fall on the critical region of the test leading to the rejection of the null hypothesis. (b) ROC curves for the Higgs signal with various Higgs masses m_{H} with the fixed-background model. The method is able to identify the signal without a priori knowledge of the mass.

6 Discussion

The proposed semi-supervised anomaly detection method is applicable to problems where anomalies lie among the normal data, or put in other words, to problems where we want to find an unexpected, unknown or uncertain signal that does not appear in the known background data. We showed that the method can be applied to searches of new particles in high energy physics, but other potential application domains can appear in many fields of life. One example is epidemiology: when a new type of a flu appears, one could take measurements of patients with flu symptoms and compare the distribution to the previous years. The proposed framework could give a hint on how to find the patients that most probably have the new disease which could prove out to be crucial in the first stages of a disease outbreak.

The method could also be useful in defense applications and in particular electronic surveillance. In this case, the normal model would be trained using day-to-day surveillance data. New measurements could then be compared to this distribution using the fixed-background EM algorithm in order to detect any increase of certain types of signals or communication patterns and sort out the suspicious observations for further scrutiny. The same applies to detection of network intrusions.

The general idea of semi-supervised anomaly detection with the fixed-background model can be implemented in a number of different ways. First, instead of using mixtures of Gaussians, one could use some other density estimation method, parametric or nonparametric, especially for the normal model. Second, even when mixtures of Gaussians are used, one could use some other statistical learning method instead of the EM algorithm. For instance, Bayesian approaches based on Markov chain Monte Carlo sampling or variational approximations might be more robust and allow taking into account more flexibly prior information about the anomalies.

One obvious shortcoming of the proposed algorithm is that it is only able

to detect anomalies that manifest themselves as an excess on top of the expected normal data. In its current form, the method is not applicable to situations where there is a deficit in the data. For example, in high energy physics there could be defects in the detectors which cause the observed number of collision events to be lower than expected. In such situations, it might possible to extend the methodology to cover cases where some π_q are negative.

Another practical limitation of the algorithm is the *curse of dimensionality* which refers to the fact that the higher the dimensionality of the data, the larger the number of observations required to achieve density estimates of certain precision. Thus, a suitable dimensionality reduction method combined with application-specific preprocessing steps are needed, as shown by our Higgs demonstration. When implementing these steps, one should take into account that anomalies should remain as well separated from the normal data as possible while maintaining the compatibility of the normal data samples of the two data sets used in the method.

7 Conclusions

We have presented a semi-supervised anomaly detection framework based on the so called fixed-background model. The proposed model assumes that the normal data follows a fixed distribution, thus providing the means to detect anomalous patterns that lie among the normal data and manifest themselves as collective deviations from this distribution. The most important features of the framework are its ability to perform pattern recognition of anomalies within the normal data and its fully probabilistic construction.

Learning of the models is carried out using a variant of the EM algorithm called the fixed-background EM. We showed that after some heuristic adjustments, the algorithm is robust enough to consistently find anomalous patterns that make up only a few percent of an unlabeled data set. In these situations, the method is able to accurately model the distribution of the anomalies and their percentage among the data.

We demonstrated one possible application of the method within the field of high energy physics where it could serve as a means of detecting unexpected new particles without exact a priori knowledge of their properties. However, given the generality of the framework, it should be straightforward to find future applications also on other fields of science and technology.

Acknowledgments.

The authors are grateful to the CDF collaboration for providing access to the Higgs signal and background Monte Carlo samples, to the Academy of Finland for financial support and to Matti Pöllä, Timo Honkela and Risto Orava for valuable discussions and feedback on the manuscript.

Bibliography

[1] D. Agarwal. An empirical Bayes approach to detect anomalies in dynamic multidimensional arrays. In *Proceedings of the Fifth IEEE International Conference on Data Mining*, ICDM '05, pages 26–33, Washington, DC, USA, 2005. IEEE Computer Society.

[2] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Compututing Surveys*, 41:15:1–15:58, 2009.

[3] D. R. Cox and D. V. Hinkley. *Theoretical Statistics*. Chapman and Hall, 1974.

[4] R. B. D'Agostino and M. A. Stephens. *Goodness-of-Fit Techniques*. Marcel Dekker, New York, N.Y., 1986.

[5] 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 (Methodological)*, 39(1):1–38, 1977.

[6] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, 1993.

[7] E. Eskin. Anomaly detection over noisy data using learned probability distributions. In *Proceedings of the International Conference on Machine Learning*, pages 255–262. Morgan Kaufmann, 2000.

[8] E. Eskin, A. Arnold, M. Prerau, L. Portnoy, and S. Stolfo. A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data. In *Applications of Data Mining in Computer Security*. Kluwer, 2002.

[9] D. M. Hawkins. *Identification of Outliers*. Monographs on Applied Probability and Statistics. Springer, 1980.

[10] S. Hawkins, H. He, G. J. Williams, and R. A. Baxter. Outlier detection using replicator neural networks. In *Proceedings of the 4th International Conference on Data Warehousing and Knowledge Discovery*, DaWaK 2000, pages 170–180, London, UK, 2002. Springer-Verlag.

[11] K. Knight. *Mathematical Statistics*. Chapman and Hall, 2000.

[12] M. Kuusela, E. Malmi, T. Vatanen, R. Orava, T. Aaltonen, and Y. Nagai. Detection of new physics using density estimation based anomaly search. CDF/DOC/EXOTIC/CDFR/10227 (Internal note), 2010.

[13] M. Lauer. A mixture approach to novelty detection using training data with outliers. In *Lecture Notes in Computer Science*, pages 300–311. Springer, 2001.

[14] M. Markou and S. Singh. Novelty detection: A review - part 1: Statistical approaches. *Signal Processing*, 83:2481–2497, 2003.

[15] M. Markou and S. Singh. Novelty detection: A review - part 2: Neural network based approaches. *Signal Processing*, 83:2499–2521, 2003.

[16] G. McLachlan and D. Peel. *Finite Mixture Models*. Wiley Series in Probability and Statistics. Wiley-Interscience, 2000.

[17] G. J. McLachlan and K. E. Basford. *Mixture models: Inference and applications to clustering*. Marcel Dekker, New York, N.Y., 1988.

[18] G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions (Wiley Series in Probability and Statistics)*. Wiley-Interscience, 2 edition, 2008.

[19] Y. Nagai. *Search for the Standard Model Higgs Boson in the $WH \to \ell\nu b\bar{b}$ Channel in 1.96-TeV Proton-Antiproton Collisions*. PhD thesis, University of Tsukuba, 2010. FERMILAB-THESIS-2010-21.

[20] Y. Nagai et al. Search for the Standard Model Higgs boson production in association with a W boson using 4.3/fb. CDF/PUB/EXOTIC/PUBLIC/9997, 2009.

[21] M. Piccardi. Background Subtraction Techniques: A Review. In *IEEE International Conference on Systems, Man and Cybernetics*, volume 4, pages 3099–3104, 2004.

[22] G. Rätsch, S. Mika, B. Schölkopf, and K.-R. Müller. Constructing boosting algorithms from SVMs: An application to one-class classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24:1184–1199, 2002.

[23] S. R. Sain, H. L. Gray, W. A. Woodward, and M. D. Fisk. Outlier detection from a mixture distribution when training data are unlabeled. *Bulletin of the Seismological Society of America*, 89(1):294–304, 1999.

[24] B. Schölkopf, J. C. Platt, J. C. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Comput.*, 13:1443–1471, 2001.

[25] P. Smyth. Model selection for probabilistic clustering using cross-validated likelihood. *Statistics and Computing*, 10:63–72, 2000.

[26] S. Wang, W. A. Woodward, H. L. Gray, S. Wiechecki, and S. R. Sain. A new test for outlier detection from a multivariate mixture distribution. *Journal of Computational and Graphical Statistics*, 6(3):285–299, 1997.

Aalto-ST 22/2011

ISBN 978-952-60-4319-7 (pdf)
ISSN-L 1799-4896
ISSN 1799-490X (pdf)

Aalto University
School of Science
Department of Information and Computer Science
www.aalto.fi

BUSINESS +
ECONOMY

ART +
DESIGN +
ARCHITECTURE

**SCIENCE +
TECHNOLOGY**

CROSSOVER

DOCTORAL
DISSERTATIONS