Pattern Recognition 58 (2016) 121–134

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

# High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning

Sarah M. Erfani *, Sutharshan Rajasegarar [1], Shanika Karunasekera, Christopher Leckie

*NICTA Victoria Research Laboratory, Department of Computing and Information Systems, Room 7.14, Dough MacDonell Building, The University of Melbourne, VIC 3010, Australia*

ARTICLE INFO

*Article history:*
Received 8 April 2015
Received in revised form
29 February 2016
Accepted 28 March 2016
Available online 14 April 2016

*Keywords:*
Anomaly detection
Outlier detection
High-dimensional data
Deep belief net
Deep learning
One-class SVM
Feature extraction

ABSTRACT

High-dimensional problem domains pose significant challenges for anomaly detection. The presence of irrelevant features can conceal the presence of anomalies. This problem, known as the '*curse of dimensionality*', is an obstacle for many anomaly detection techniques. Building a robust anomaly detection model for use in high-dimensional spaces requires the combination of an unsupervised feature extractor and an anomaly detector. While one-class support vector machines are effective at producing decision surfaces from well-behaved feature vectors, they can be inefficient at modelling the variation in large, high-dimensional datasets. Architectures such as deep belief networks (DBNs) are a promising technique for learning robust features. We present a hybrid model where an unsupervised DBN is trained to extract generic underlying features, and a one-class SVM is trained from the features learned by the DBN. Since a linear kernel can be substituted for nonlinear ones in our hybrid model without loss of accuracy, our model is scalable and computationally efficient. The experimental results show that our proposed model yields comparable anomaly detection performance with a deep autoencoder, while reducing its training and testing time by a factor of 3 and 1000, respectively.



## 1. Introduction

The growth in pervasive network infrastructure, such as the Internet of Things (IoT), is enabling a wide range of physical objects and environments to be monitored in fine spatial and temporal detail [1,2]. A key challenge in the development of IoT applications is how to model and interpret the large volumes of high-dimensional data that are generated in such domains [2]. Further, the lack of ground truth (labels) in the data that are collected from large-scale networks in the IoT require unsupervised algorithms to process the data. Anomaly detection aims to detect unusual behaviours caused by either faulty devices or events of interest in the monitoring environment, and thus is of great importance in IoT applications. However, a major challenge for anomaly detection in such domains is how to cope with noisy, large-scale datasets [3-6]. In this work we address this challenge by proposing an unsupervised hybrid architecture for anomaly detection in large-scale high-dimensional problem domains.

A core challenge for anomaly detection that distinguishes it from other classification problems is that in many cases anomaly detection algorithms should be trained with unlabelled records, i.e., trained in an unsupervised manner. Obtaining a large training set of clean and labelled data is often a labour and time intensive task. Moreover, anomaly detection becomes more challenging when applied to high-dimensional datasets that contain a large number of records. Many of the available methods for identifying anomalies assume small datasets with low numbers of features.

High-dimensional datasets pose a challenge for anomaly detection due to the following factors [7]: (i) *Exponential search space* — The number of potential feature subspaces grows exponentially with increasing input dimensionality, resulting in an exponential search space. (ii) *Data-snooping bias* — Every point in a high-dimensional space appears as an anomaly. Given enough alternative subspaces, at least one feature subspace can be found for each point such that it appears as an anomaly. (iii) *Irrelevant features* — A high proportion of irrelevant features effectively creates noise in the input data, which masks the true anomalies. The challenge is to choose a subspace of the data that highlights the relevant attributes.

Our objective is to find a large-scale, high-dimensional anomaly detection algorithm that is *robust*, i.e., generates an accurate model for data drawn from a wide range of probability distributions, and is not unduly affected by small departures from the trained model. In addition, it is desirable that the algorithm be efficient in terms

* Corresponding author. Tel. +61 3 83440366.
*E-mail address:* sarah.erfani@unimelb.edu.au (S.M. Erfani).
[1] This author is now with: School of Information Technology, Deakin University, Australia

http://dx.doi.org/10.1016/j.patcog.2016.03.028
0031-3203/

of *time complexity*, *memory complexity* and the *required number of labelled records*.

One-class Support Vector Machines (1SVMs) [8–10] are a popular technique for unsupervised anomaly detection. Generally, they aim to model the underlying distribution of normal data while being insensitive to noise or anomalies in the training records. A kernel function implicitly maps the input space to a higher dimensional feature space to make a clearer separation between normal and anomalous data. When properly applied, in principle a kernel-based method is able to model any non-linear pattern of normal behaviour. For clarity in the rest of the paper, the notation of 1SVM is used to denote (an unsupervised) one-class SVM; *l*SVMs — short for labeled SVM — to denote (supervised) binary and multi-class SVM classifiers; and SVMs when both 1SVMs and *l*SVMs are considered.

SVMs are theoretically appealing for the following reasons [11,12]: they provide good generalisation when the parameters are appropriately configured, even if the training set has some bias; they deliver a unique solution, since the loss function is convex; and in principal they can model any training set, when an appropriate kernel is chosen.

In practice, however, training SVMs is memory and time intensive. SVMs are non-parametric learning models, whose complexity grows quadratically with the number of records [13]. They are best suited to small datasets with many features, and so far large-scale training on high-dimensional records (e.g., $10^6 \times 10^4$) has been limited with SVMs [14]. Large numbers of input features result in the curse of dimensionality phenomenon, which causes the generalisation error of shallow architectures (discussed in Section 2.1), such as SVMs, to increase with the number of irrelevant and redundant features. The curse of dimensionality implies that to obtain good generalisation, the number of training samples must grow exponentially with the number of features [14,4,15]. Furthermore, shallow architectures have practical limitations for efficient representation of certain types of function families [16]. To avoid these major issues, it is essential to generate a model that can capture the large degree of variation that occurs in the underlying data pattern, without having to enumerate all of them. Therefore, a compact representation of the data that captures most of the variation can alleviate the curse of dimensionality as well as reducing the computational complexity of the algorithm [16,17].

An alternative class of classification algorithms that have emerged in recent years are Deep Belief Nets (DBNs), which have been proposed as a multi-class classifier and dimensionality reduction tool [18–20]. DBNs are multi-layer generative models that learn one layer of features at a time from unlabelled data. The extracted features are then treated as the input for training the next layer. This efficient, greedy learning can be followed by fine-tuning the weights to improve the generative or discriminative performance of the whole network.

DBNs have a deep architecture, composed of multiple layers of parameterised non-linear modules. There are a range of advantageous properties that have been identified for DBNs [16]: they can learn higher-level features that yield good classification accuracy; they are parametric models, whose training time scales linearly with the number of records; they can use unlabelled data to learn from complex and high-dimensional datasets.

A major limitation of DBNs is that their loss function is non-convex, therefore the model often converges on local minima and there is no guarantee that the global minimum will be found. In addition, DBN *classifiers* are semi-supervised algorithms, and require some labelled examples for discriminative fine-tuning, hence an unsupervised generative model of DBNs, known as autoencoders, are used for anomaly detection.

The open research problem we address is how to overcome the limitations of one-class SVM architectures on complex, high-dimensional datasets. We propose the use of DBNs as a feature reduction stage for one-class SVMs, to give a hybrid anomaly detection architecture. While a variety of feature reduction methods — i.e., feature selection and feature extraction methods — have been considered for SVMs (e.g., [21–25] — see [26] for a survey) none have studied the use of DBNs as a method for deep feature construction in the context of anomaly detection, i.e., with a one-class SVM. In this paper, we design and evaluate a new architecture for anomaly detection in high-dimensional domains. To the best of our knowledge, this is the first method proposed for combining DBNs with one-class SVMs to improve their performance for anomaly detection.

The contributions of this paper are two-fold. The performance of DBNs against one-class SVMs is evaluated for detecting anomalies in complex high-dimensional data. In contrast, the reported results in the literature from DBN classification performance only cover *multi-class* classification, e.g., [14,27–29]. A novel unsupervised anomaly detection model is also proposed, which combines the advantages of deep belief nets with one-class SVMs. In our proposed model an unsupervised DBN is trained to extract features that are reasonably insensitive to irrelevant variations in the input, and a 1SVM is trained on the feature vectors produced by the DBN. More specifically, for anomaly detection we show that computationally expensive non-linear kernel machines can be replaced by linear ones, when aggregated with a DBN. To the best of our knowledge, this is the first time these frameworks have been combined this way. The result of experiments conducted on several benchmark datasets demonstrate that our hybrid model yields significant performance improvements over the stand-alone systems. The combination of the hybrid DBN-1SVM avoids the complexity of non-linear kernel machines, and reaches the accuracy of a state-of-the-art autoencoder while considerably lowering its training and testing time.

The remainder of the paper is organised as follows. Section 2 begins with an introduction to deep architectures and their strengths and weaknesses compared to their shallow counterparts. Then it reviews some of the leading 1SVM methods, and motivates the requirements for the hybrid model by considering the shortcomings of SVMs for processing large datasets. Section 3 presents our proposed unsupervised anomaly detection approach DBN-1SVM. Section 4 describes the empirical analysis and provides a detailed statistical comparison of the performance of autoencoder, 1SVM and DBN-1SVM models on various real-world and synthetic datasets. It demonstrates the advantages of the DBN-1SVM architecture in terms of both accuracy and computational efficiency. Section 5 summarises the paper and outlines future research.

## 2. Background

### 2.1. Shallow and deep architectures

Classification techniques with shallow architectures typically comprise an input layer together with a single layer of processing. Kernel machines such as SVMs, for example, are a layer of kernel functions that are applied to the input, followed by a linear combination of the kernel outputs. In contrast, deep architectures are composed of several layers of nonlinear processing nodes. The widely used form of the latter type of architectures are multi-layer neural networks with multiple hidden layers.

While shallow architectures offer important advantages when optimising the parameters of the model, such as using convex loss functions, they suffer from limitations in terms of providing an efficient representation for certain types of function families. In