Expert Systems With Applications 91 (2018) 374–385

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Manifold learning techniques for unsupervised anomaly detection

C.C. Olson*, K.P. Judd, J.M. Nichols

Naval Research Laboratory, 4555 Overlook Ave. SW, Washington, D.C. 20375, USA

ARTICLE INFO

Article history:
Received 25 July 2016
Revised 31 July 2017
Accepted 1 August 2017
Available online 25 August 2017

Keywords:
Manifolds
Manifold learning
Image processing
Anomaly detection
Target detection

ABSTRACT

Appropriately identifying outlier data is a critical requirement in the decision-making process of many expert and intelligent systems deployed in a variety of fields including finance, medicine, and defense. Classical outlier detection schemes typically rely on the assumption that normal/background data of interest are distributed according to an assumed statistical model and search for data that deviate from that assumption. However, it is frequently the case that performance is reduced because the underlying distribution does not follow the assumed model. Manifold learning techniques offer improved performance by learning better models of the background but can be too computationally expensive due to the need to calculate a distance measure between all data points. Here, we study a general framework that allows manifold learning techniques to be used for unsupervised anomaly detection by reducing computational expense via a uniform random sampling of a small fraction of the data. A background manifold is learned from the sample and then an out-of-sample extension is used to project unsampled data into the learned manifold space and construct an anomaly detection statistic based on the prediction error of the learned manifold. The method works well for unsupervised anomaly detection because, by definition, the ratio of anomalous to non-anomalous data points is small and the sampling will be dominated by background points. However, a variety of parameters that affect detection performance are introduced so we use here a low-dimensional toy problem to investigate their effect on the performance of four learning algorithms (kernel PCA, two versions of diffusion map, and the Parzen density estimator). We then apply the methods to the detection of watercraft in an ensemble of 22 infrared maritime scenes where we find kernel PCA to be superior and show that it outperforms a commonly employed baseline algorithm. The framework is not limited to the tested image processing example and can be used for any unsupervised anomaly detection task.

Published by Elsevier Ltd.

1. Introduction

We consider the problem of detecting points that are rare within a data set dominated by the presence of ordinary background points. The goal is to assign unknown data to either a background or anomaly class and the numerous algorithms that have been devised to handle this problem can be categorized as supervised, semi-supervised, or unsupervised depending on how much information is available to the training algorithm.

Supervised approaches require labeled training data for both classes. Models that maximize the difference between classes are then constructed; some common algorithms include neural networks (Markou & Singh, 2003b), Gaussian mixture models (Tarassenko, Nairac, Townsend, & Cowley, 1999), principal components analysis (for linearly separable data), and kernel support vector machines (SVMs) (Schölkopf & Smola, 2002). Semi-supervised approaches, some examples of which can be found in Fujimaki, Yairi, and Machida (2005) and Bouchachia (2007), only require labels for the background class. Unsupervised algorithms are the most generically applicable and are often based on measures of similarity between data vectors. Examples include thresholding distances between neighboring data vectors (Knorr, Ng, & Tucakov, 2000), the local outlier factor (Breunig, Kriegel, Ng, & Sander, 2000), one-class SVMs (Schölkopf, Williamson, Smola, Shawe-Taylor, & Platt, 2000), and fuzzy c-means clustering (Bezdek, Ehrlich, & Full, 1984). Some reviews of the anomaly detection problem are provided in Markou and Singh (2003a) and Chandola, Banerjee, and Kumar (2009).

Supervised techniques are preferred over the unsupervised case whenever possible as we would expect the presence of training data to improve classification. We are not, however, always afforded this luxury. This is the case in many image-based detection scenarios where neither the background pixels nor the anomalous (often man-made) pixels are expected to be consistent be-

* Corresponding author.
E-mail addresses: colin.olson@nrl.navy.mil, enceladus.co@gmail.com (C.C. Olson), kyle.judd@nrl.navy.mil (K.P. Judd), jonathan.nichols@nrl.navy.mil (J.M. Nichols).

http://dx.doi.org/10.1016/j.eswa.2017.08.005
0957-4174/Published by Elsevier Ltd.

tween scenes. Variations in the type of background pixels as well as changes in lighting and scene viewing angle can invalidate *a priori* assumptions about scene composition. Thus, we are motivated to develop unsupervised detection techniques.

Kernel and spectral methods comprise a family of algorithms commonly used for clustering and classification. Of these, spectral methods in particular are also used for dimensionality reduction. Algorithms such as Laplacian eigenmaps (Belkin & Niyogi, 2003), locally-linear embedding (Roweis & Saul, 2000), Isomap (Tenenbaum, de Silva, & Langford, 2000), and diffusion map (Coifman & Lafon, 2006) (which we discuss below) assume the observed high-dimensional data were actually generated by a lower-dimensional process and that the associations between the two can be learned. The goal of a manifold learning algorithm is therefore to map the original data onto a new coordinate system in which the classification problem is made simpler. These methods are similar in that they organize the data into clusters based on the eigenvalues and eigenvectors of a distance (adjacency) matrix calculated from the data. The data are viewed as nodes in a graph and the edges connecting the nodes are weighted by the similarity between the data as determined by a distance-measuring kernel.

In recent years such methods have been used for the analysis of hyperspectral images. For example, they have been used for classification (Bachmann, Ainsworth, & Fusina, 2005; Chen, Crawford, & Ghosh, 2005), target detection (Ziemann & Messinger, 2015; Ziemann, Theiler, & Messinger, 2015), and change detection (Albano, Messinger, Schlamm, & Basener, 2011). Within the context of anomaly detection, Kwon and Nasrabadi (2005) introduced a kernelized version of the standard RX algorithm (Reed & Yu, 1990) under the assumption that background and target would be described by Gaussian distributions in the high-dimensional feature space describing kernelized spectra. The TAD approach was introduced by Basener *et al* and found to perform well against a variety of benchmark algorithms (Basener, Ientilucci, & Messinger, 2007). Messinger and Albano also considered the anomaly detection problem by measuring the connectivity of individual pixels within a locally-constructed graph (Messinger & Albano, 2011).

The motivation for using such kernel-based or manifold-learning algorithms is that a background model that is more appropriate to the specifics of a given scene can be learned using data-driven techniques rather than assuming a statistical model *a priori* as is done with, for example, RX. Estimating the parameters governing an assumed statistical distribution and constructing decision surfaces as a function of the learned parameters would be preferred, but real-world data frequently fail to follow assumed distribution models and it has been shown (see, e.g., Theiler, Foy, & Fraser, 2007) that sensitivity to outliers may be reduced if the assumptions underlying the model are not met by the data.

Adoption of such data-driven techniques is hampered, however, by the expense of calculating an adjacency matrix. In Olson, Nichols, Michalowicz, and Bucholtz (2010) we proposed a statistically uniform "skeleton" subsampling of a hyperspectral scene to reduce the computational cost of building an adjacency matrix and performed a preliminary study of out-of-sample extension (Bengio et al., 2004; Lafon, Keller, & Coifman, 2006) as a means of developing a detection statistic for the remaining unsampled points. We performed an additional study of the subsampling method in Olson and Doster (2016). Bachmann et al. (2005) have previously considered the use of subsampled pixel sets as a means of building a global manifold backbone against which local manifolds built from sub-segments of a scene could be aligned, but they found the method to be too computationally expensive for classification and did not consider the anomaly detection problem. Graph-based methods have been used previously in a semi-supervised manner for classification tasks (Blum & Chawla, 2001; Szummer & Jaakkola, 2002) and, more recently, Belkin and Niyogi (2002) and Belkin, Niyogi, and Sindhwani (2006) demonstrated that semi-supervised techniques can be used to learn a data manifold for classification. Although similar to our method, we are not aware of any work besides our own that extends these techniques to unsupervised anomaly detection in imagery.

Building an adjacency matrix from a subset of the data is conceptually simple but enables application of the wide variety of data-driven learning techniques to the anomaly detection problem and offers the prospect of improved detection performance over classical techniques. The tradeoff is the introduction of a set of unique considerations relative to previous approaches. The following are a few of the most fundamental considerations: (1) What data-driven learning algorithm should be applied to the sampled skeleton subset?; (2) What fraction of the data set must be sampled in order to guarantee with some probability that all background classes will be sufficiently sampled without over-sampling the anomalous class?; (3) What should the parameter settings be for a given learning algorithm and how are they affected by the size of the subsample?; (4) How stable is detection performance as a function of parameter settings and subsample size?; (5) How best to extend the learned model space to the out-of-sample points?

In this work we primarily focus on considerations (1), (3), and (4). In particular we address consideration (1) by using kernel PCA (Schölkopf, Smola, & Müller, 1998), two versions of diffusion map (Coifman & Lafon, 2006), and the Parzen density estimator (Parzen, 1962) to learn background models for panchromatic (not hyperspectral) images that have been tiled to form superpixels. With all three techniques the basic idea is the same: learn a model based on previously acquired background data, project in new pixel data, and compute a measure of error between data and model as our detection statistic. The performance of the algorithms on a toy problem and real-world data set are quantified using receiver operating characteristic (ROC) curves (Kay, 1998) over a wide range of algorithm parameter settings (consideration 3) and over multiple skeleton samples (consideration 4). In all cases, good detection performance is obtained on the toy problem; however, kernel PCA outperforms the other learning algorithms on the real-world target detection task. We provide a more complete description of each technique in Section 2, describe the experiments and compare to an established algorithm in Section 3, and discuss results in Section 4 before concluding in Section 5.

2. Methods and motivation

The idea behind any anomaly detection approach is to model the background distribution using either assumed physical principles or by learning its description from the data. It is the latter route that we consider here. We begin with the set, Ω, of N pixel intensities $\mathbf{x}_i \in \mathbb{R}^M,\ i = 1 \cdots N$ that comprise an image. Most of the pixels are assumed to contain background information while only a very few ($< 1\%$) are assumed to contain a "target" point of interest.

In general, we seek to find a function, $f(\cdot)$, that maps the $\mathbf{x}_i$ into a new coordinate system where we can draw decision surfaces that more accurately separate anomaly from background. We don't, however, know $f(\cdot)$ *a priori* and must form an estimate, $\hat{f}(\cdot)$, from our data. In this work we compare a number of methods, both linear and nonlinear, for learning $\hat{f}(\cdot)$ and compare their resulting detection performance (although we drop the $\hat{f}(\cdot)$ from here on out and work with $f(\cdot)$ for notational parsimony).

Given $f(\cdot)$, each datum can be represented in the new coordinate system by performing an analysis step $\theta_i = f(\mathbf{x}_i)$ where $\theta_i \in \mathbb{R}^M$. Conversely, we may model (synthesize) each datum as $\hat{\mathbf{x}}_i = f^{-1}(\theta_i)$ where we allow $f^{-1} : \mathbb{R}^M \rightarrow \mathbb{R}^m$ with $m \leq M$. Of course, a unique inverse and $\hat{\mathbf{x}}_i = \mathbf{x}_i$ can only be guaranteed when $m = M$. In