Neurocomputing 130 (2014) 63–72

Contents lists available at ScienceDirect

Neurocomputing

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

A novelty detection machine and its application to bank failure prediction

Shukai Li [a,*], Whye Loon Tung [b,c], Wee Keong Ng [b]

[a] Institute for Infocomm Research, A*Star, Singapore
[b] School of Computer Engineering, Nanyang Technological University, Singapore
[c] Service Platform Lab, HP Labs, Singapore

ARTICLE INFO

Article history:
Received 15 January 2012
Received in revised form
8 February 2013
Accepted 27 February 2013
Available online 27 July 2013

Keywords:
Novelty detection
Cluster assumption
Bank failure prediction

ABSTRACT

Novelty detection has been well-studied for many years and has found a wide range of applications, but correctly identifying the outliers is still a hard problem because of the diverse variation and the small quantity of such outliers. We address the problem using several distinct characteristics of the outliers and the normal patterns. First, normal patterns are usually grouped together, forming clusters in the high density regions of the data space. Second, outliers are characteristically very different from the normal patterns, and hence tend to be located far away from the normal patterns in the data space. Third, the number of outliers is generally very small in a given dataset. Based on these observations, we can envisage that the appropriate decision boundary segregating the outliers and the normal patterns usually lies in some low density regions of the data space. This is referred to as cluster assumption. The resultant optimization problem to learn the decision function can be solved using the mixed integer programming approach. Following that, we present a cutting plane algorithm together with a multiple kernel learning technique to solve the convex relaxation of the optimization problem. Specifically, we make use of the scarcity of the outliers to find a violating solution to the cutting plane algorithm. Experimental results with several benchmark datasets show that our proposed novelty detection method outperforms existing hyperplane and density estimation-based novelty detection techniques. We subsequently apply our method to the prediction of banking failures to identify potential bank failures or high risk banks through the traits of financial distress.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

Novelty detection, also known as outlier detection, anomaly detection, or one-class classification, is an important problem in data mining and machine learning. The primary task of novelty detection is to differentiate the known objects (normal patterns) from the deviant samples (outliers) [1–3]. The typical assumption of novelty detection is that the normal patterns are in abundance and frequently observed; while the outliers are very rare or may even be previously unseen samples that are characteristically very different from the normal patterns. Novelty detection has found many real-world applications; for instance, in mechanical diagnosis to isolate the faulty jet engines [4], to perform intrusion detection in network systems [5], and to detect fraudulent credit card transactions [3], etc. In finance, bank failure is an important issue in credit risk management, and the number of failing banks is small within the entire banking system. However, the collapse and failure of a bank could have devastating consequences to the entire banking system and an adverse repercussion effect on other banks and financial institutions. Some of the negative impacts are the massive bail out cost for a failing bank and the negative sentiments and loss of confidence developed by investors and depositors. Hence, bank failure prediction is an important issue for regulators of the banking industries. In this paper, we will apply our proposed technique of novelty detection to predict banking failures.

Traditionally, outliers are often detected by estimating the density functions of some pre-defined models (e.g. multivariate Gaussian (MVG), Gaussian mixtures, kernel density estimation (KDE) [6], etc.) to fit the normal patterns. However, such methods may fail miserably when the pre-defined models cannot capture the true distribution of the normal patterns. Alternatively, several k nearest neighbors (kNN) based outlier detection methods such as local outlier factor (LOF) [7] and prototype-based domain description (PDD) [8] have been proposed, and they showed encouraging results for detecting outliers. However, these kNN based methods cannot handle previously unseen data and may scale poorly on large and high dimensional datasets.

Instead of estimating the density distribution of the normal patterns or using the computationally expensive k nearest neighbor approaches, a simpler and more intuitive methodology is to directly model the support of the normal pattern distribution.

* Corresponding author. Tel.: +65 8178 6016; fax: +65 6792 6559.
E-mail addresses: lisk@i2r.a-star.edu.sg (S. Li), wltung@pmail.ntu.edu.sg (W.L. Tung), AWKNG@ntu.edu.sg (W.K. Ng).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.02.043

Tax and Duin proposed the support vector data description (SVDD) technique [9], which uses a small hypersphere (enclosing ball) to enclose most of the normal patterns. Computationally, this leads to a convex quadratic programming (QP) problem, which has the important feature that the solution obtained is always globally optimal. Moreover, as with other kernel methods, SVDD works well with high-dimensional datasets and can be easily extended to nonlinear generalization by replacing the dot products between the data patterns with the corresponding *kernel* evaluations. Besides the enclosing ball methodology, Schölkopf et al. [10] proposed a one-class support vector machine (SVM) that uses a hyperplane to separate the normal patterns from the outliers with a large margin. Again, this leads to a QP problem. Moreover, when a Gaussian kernel is used, the solution of the one-class SVM is equivalent to that of SVDD. For further details, the interested reader can refer to [11,2,3] for several comprehensive surveys of existing novelty detection algorithms.

In this paper, we employ the hyperplane approach for novel detection as it can handle high dimensional datasets as well as achieve good generalization performance on previously unseen data. Moreover, it is easy to incorporate constraints to encode prior knowledge of the outliers into the training process of the hyperplane approach. Several empirical observations provide the motivation for our proposed novelty detection method. First, normal patterns are usually grouped together, forming clusters in the high density regions of the data space. Second, the outliers are characteristically very different from the normal patterns, and hence far away from the normal patterns. Third, the number of outliers is generally small for a given dataset. Based on these observations, we envisage that the decision boundary between the outliers and the normal patterns lies in some low density regions of the data space. To the best of our knowledge, such empirical knowledge has not been explicitly exploited for novelty detection. Moreover, the scarcity of the outliers may serve as a constraint to help identify such outliers from the normal patterns.

The major contributions of this paper are outlined as follows:

(1) We observe that the decision function to segregate the normal patterns and the outliers lies in some low density regions and can be employed for novelty detection. This is known as cluster assumption. Hence, we explicitly define a constraint based on this empirical observation in the corresponding optimization problem to facilitate novelty detection. The resultant optimization process can learn the decision function (decision boundary) and the labels of the data samples simultaneously.
(2) Due to the combinatorial nature of the problem, the resultant optimization process is NP hard. Hence, we introduce a convex relaxation to this non-convex optimization problem, and subsequently proposed an efficient cutting plane algorithm to solve this convex relaxation.
(3) By exploiting the scarcity of the outliers, we have devised an efficient method to find an approximation to the most violating labeling for our proposed cutting plane algorithm.
(4) Comprehensive experimental results on several benchmark datasets demonstrated that our proposed novelty detection technique outperforms the existing hyperplane-based novelty detection approaches.

The rest of this paper is organized as follows. Section 2 gives a brief review on the one-class SVM and the notion of cluster assumption. Section 3 describes our proposed methodology for novelty detection, and the corresponding experimental results are presented in Section 4. Section 5 concludes the paper.

For simplicity, the transpose of a vector/matrix will be denoted by the superscript $'$, and $\mathbf{0}, \mathbf{1} \in \mathbb{R}^n$ denote the zero vector and the vector of all ones, respectively. In addition, the inequality $\mathbf{v} = [v_1, \ldots, v_k]' \geq \mathbf{0}$ means that $v_i \geq 0$ for $i = 1, \ldots, k$. tr($\mathbf{X}$) is short for tr($\mathbf{X}$) which means the sum of the diagonal elements of the matrix $\mathbf{X}$.

2. Review of related works

2.1. One-class support vector machine (SVM)

We first review the one-class SVM. Given a set of unlabeled patterns $\{\mathbf{x}_i\}_{i=1}^n$ where $\mathbf{x}_i \in \mathcal{X}$ is the input, these patterns are first mapped to the feature space $\mathcal{H}$ via a nonlinear mapping function ϕ induced by a kernel k. Next, we assume that the outliers generally lie in the low density regions of the data space. In the one-class SVM, the outliers are presumed to be close to the origin [10], and a decision function $f(\mathbf{x}) = \mathbf{w}'\phi(\mathbf{x})$ is found to separate the majority of the data (the normal patterns) from the origin (the reference for the outliers) with a large margin $\rho/\|\mathbf{w}\|$ by solving the following structural risk functional:

$$\min_{\mathbf{w},\rho,\xi_i} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i^p - \rho : \mathbf{w}'\phi(\mathbf{x}_i) \geq \rho - \xi_i, \quad \xi_i \geq 0 \ \ \forall i = 1, \ldots, n, \tag{1}$$

where C is a parameter that trades off the empirical risk $\sum_{i=1}^n \xi_i^p$ (ξ_i is the slack variable) and the model complexity $\|\mathbf{w}\|^2$, and $p = 1$ or 2 corresponds to the hinge loss and the squared hinge loss, respectively.

This constrained optimization problem (for $p = 1$ or 2) is usually solved using its dual form. For simplicity, we just present the case of $p = 2$ in this paper, i.e.

$$\max_{\boldsymbol{\alpha}} -\frac{1}{2}\sum_{i,j=1}^{n} \alpha_i\alpha_j\left(k(\mathbf{x}_i, \mathbf{x}_j) + \frac{1}{C}\delta_{ij}\right) : \quad \alpha_i \geq 0, \quad \sum_{i=1}^{n} \alpha_i = 1, \tag{2}$$

where α_i is a dual variable for each inequality constraint in (1), and δ_{ij} is an indicator function (i.e., $\delta_{ij} = 1$ if $i = j$; and 0 otherwise). Let $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_n]'$ be the vector of dual variables, and $\mathbf{K} = [k(\mathbf{x}_i, \mathbf{x}_j)] \in \mathbb{R}^{n\times n}$ be the kernel matrix, and $\mathcal{A} = \{\boldsymbol{\alpha} | \boldsymbol{\alpha} \geq \mathbf{0}, \boldsymbol{\alpha}'\mathbf{1} = 1\}$. Then the QP in (2) can be re-expressed as

$$\max_{\boldsymbol{\alpha}\in\mathcal{A}} -\frac{1}{2}\boldsymbol{\alpha}'\left(\mathbf{K} + \frac{1}{C}\mathbf{I}\right)\boldsymbol{\alpha}. \tag{3}$$

When the Gaussian kernel is used, this QP problem is equivalent to the dual of SVDD, and hence they share the same solution. Moreover, the decision function to identify the outliers can be expressed as

$$f(\mathbf{x}) = \sum_{i:\alpha_i > 0} \alpha_i k(\mathbf{x}_i, \mathbf{x}), \tag{4}$$

which is an expansion of the kernel evaluations on the support vectors only. Thus, identifying the outliers can be very efficient when the number of support vectors is small.

2.2. Cluster assumption

It may be too restrictive to simply assume that the outliers are located close to the origin. As discussed in [11,3,7,12], outliers generally are located far away from the majority of the data (i.e. the normal patterns). These normal patterns often lie in the high density regions of the data space forming clusters. Intuitively, the corresponding decision boundary to separate the normal patterns from the outliers should lie in some low density regions of the data space. This is referred to as cluster assumption [13]. Such a notion has been widely used in semi-supervised learning (SSL) algorithms [13,14] such as transductive SVM (TSVM) to perform text categorization [15]. Besides SSL, cluster assumption has also been employed in the development of unsupervised learning models such as maximum margin clustering [16].

In practice, cluster assumption is generally realized using transductive learning. Transductive learning is employed to learn

Download English Version:

https://daneshyari.com/en/article/410092

Download Persian Version:

https://daneshyari.com/article/410092

Daneshyari.com