Pattern Recognition Letters 49 (2014) 77–84

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Entropy-based outlier detection using semi-supervised approach with few positive examples ☆

Armin Daneshpazhouh [1], Ashkan Sami *

CSE and IT Department, School of Electrical Engineering and Computer, Shiraz University, Shiraz, Iran

ARTICLE INFO

Article history:
Received 13 July 2013
Available online 6 July 2014

Keywords:
Data mining
Fraud detection
Outlier detection
Semi-supervised learning

ABSTRACT

Outlier detection is an important problem in data mining that aims to discover useful exceptional and unusual patterns hidden in large data sets. Fraud detection, time series monitoring, intrusion detection and medical condition monitoring are some of the most common applications of outlier detection. Most existing outlier detection methods are based on supervised or unsupervised learning while some others use semi-supervised approaches. However, in many real world applications, there are not enough labeled data for training and only a few positive labeled samples are available. This paper presents an entropy-based solution. The proposed method consists of two phases. First, reliable negative examples are extracted from positive and unlabeled data and then, as the second phase, the entropy-based outlier detection algorithm is employed to detect top N outliers. Many experiments on real and synthetic data sets are performed. The experimental results on synthetic and real data demonstrated superiority of the proposed outlier detection method in comparison of unsupervised state-of-the-art outlier detection strategies.

1. Introduction

Most of the data mining researches focus on finding frequent patterns in the data, such as clustering, association rule mining and frequent item set mining. Whereas, fewer studies are conducted to tackle other aspects of data mining such as identifying uncommon behavior in the data called outliers. Outlier detection approaches focus on discovering patterns that occur infrequently in the data as opposed to other data mining techniques.

Outliers are generally viewed as observations that are far away from, or inconsistent with the main body of the data set. No formal or widely accepted definition of an outlier exists. Since the exact definition of an outlier often depends on the context of the application domain, Hawkins [8] definition captures the spirit: "An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism".

Outlier detection is used extensively in many applications. Current application areas of outlier detection include detection of credit card frauds, detecting fraudulent applications or potentially problematic customers in loan application processing, intrusion detection in computer networks, medical condition monitoring such as heart-rate monitoring, identifying abnormal health conditions, detecting abnormal changes in stock prices and fault diagnosis.

The importance of outlier detection is in the view of the fact that outliers can provide interesting insight about data set. For instance, the network activity that is surprisingly high with respect to its network, may indicate an error or a network attack in the system. In finance, detecting outliers such as credit card frauds can trigger actions to prevent more monetary loss of customers and the banks. In criminal investigations, outlier detection techniques can be used to identify the networks of terrorists or criminals by analyzing and detecting their unusual connections.

Outlier detection methods can be categorized based on the availability of training data. These methods further fall into three categories: supervised, semi-supervised, or unsupervised. Generally, supervised methods have a high precision rate. However, they require pre-labeled data tagged as normal or outlier. In addition, supervised methods are suitable for data whose characteristics do not change through time. Semi-supervised methods depend on the availability of training data set of normal (non-outlier) observations. Semi-supervised methods may incorrectly identify a normal observation that falls outside the trained boundary, as an outlier. The limitation of supervised and semi-supervised approaches is that the training data set must represent all possible

☆ This paper has been recommended for acceptance by G. Moser.

* Corresponding author.
E-mail addresses: daneshpajooh@cse.shirazu.ac.ir (A. Daneshpazhouh), ashkan.sami@shirazu.ac.ir (A. Sami).
[1] Tel.: +98 341 3239064.

http://dx.doi.org/10.1016/j.patrec.2014.06.012
0167-8655/

classes. Unsupervised methods process data without any prior knowledge. The advantage of unsupervised methods is that no labeled data is required. However, unsupervised outlier detection methods usually have much higher false alarm rates than the false alarm rates of supervised and semi-supervised methods.

In few real world applications, training data are available only for outliers. For example, in fraud detection scenario, the information of few fraudulent samples (positive instances), which previously have been detected, are available. No prior knowledge about the normal observations (negative instances) is available since obtaining this information is time-consuming and takes much effort. The previous studies in semi-supervised outlier detection use both negative and positive instances as training data. The current methods are not applicable for particular cases that there is lack of examples for negative class.

As opposed to former works, this study proposes an innovative approach to address the problem of detecting outliers with only few positive examples as training data. In this paper, an entropy-based method is utilized to obtain adequate information about unlabeled data.

Specifically, an innovative two-phase strategy (**EODSP**: **E**ntropy-based **O**utlier **D**etection based on **S**emi-supervised learning from **P**ositive data) is presented to solve the problem of detecting outliers when there are only few outlier samples available. In the first phase, a revised entropy-based technique is represented to find some reliable negative instances. In the second phase, a modified outlier detection algorithm is suggested, which efficiently utilizes both negative and positive examples.

The rest of the paper is organized as follows. Section 2 reviews related works. In Section 3, the proposed outlier detection approach is presented. Section 4 presents the experimental results. The last section concludes the paper and provides some future directions.

2. Related work

There are three kinds of approaches to the problem of outlier detection. Supervised, semi-supervised and unsupervised. The majority of outlier detection techniques are unsupervised, yet a few of them are based on supervised learning [23,25]. Recently semi-supervised approaches are applied to outlier detection [31,6,7,19,34] in order to overcome the time-consuming process of labeling training data. In Xue et al. [31], fuzzy rough semi-supervised outlier detection (FRSSOD) is proposed, which combines the fuzzy set theory, rough set theory, and semi-supervised learning to detect outliers. However, all these semi-supervised techniques employ both positive and negative instances as training data.

The current unsupervised approaches to outlier detection can be broadly classified into the following categories:

(1) *Distribution-based approaches* these are some classical methods used in statistics [28,3]. The user uses a statistical distribution to model the data points. Then the points, which deviate from the model are identified as outliers. However, in many practical applications the distribution of data is unknown.
(2) *Depth-based approaches* these methods compute the different layers of k dimensional convex hulls and flags objects in the outer layer as outliers [14]. They do not require to fit a distribution to the data, but still it is a well known fact that the employed algorithms suffer from the curse of dimensionality and cannot cope with large k.
(3) *Distance-based approaches* distance-based outlier was originally proposed by Knorr et al. [15] and improved by Ramaswamy et al. [27] and Angiulli and Pizzuti [2]. An object o in a data set T is a distance-based outlier if at least a fraction p of the objects in T are further than distance D from o. This outlier definition is based on a single global criterion determined by the parameters p and D. The problem of these methods is that they cannot cope with data sets having both dense and sparse regions. This is referred to as the multi-density problem.
(4) *Density-based approaches* to avoid the multi-density problem, Breunig et al. [4] assigned a local outlier factor (LOF) to each object, indicating its degree of outlierness. LOF depends on the local density of its neighborhood, where the neighborhood is defined as the distance to the *MinPtsth* nearest neighbor. Failing to deal with the multi-granularity problem is the disadvantage of this solution. When there are clusters of various numbers of points, the method is unexpectedly sensitive to parameters defining the neighborhood i.e. *MinPts* value [26].
(5) *Clustering-based approaches* these methods identify outliers as clusters of small sizes [12]. Hierarchical based approach is also proposed to recognize the outliers by using the resultant clusters as an indicator to identify the outliers.

In summary, most of the current methods are based on unsupervised approaches and the semi-supervised outlier detection algorithms need labeled examples from both normal data and abnormal classes. In Liu et al. [21], the assumption is that most of the unlabeled samples are negative (outlier) samples, which is different from the goal of this paper. In Li et al. [20], the idea of *learning from positive and unlabeled data* is used, yet the method is only suitable for text data. Contrary to these methods, this paper proposes an innovative strategy to tackle the problem of semi-supervised outlier detection with only few positive data.

3. The proposed method: EODSP

In this section the proposed EODSP method is introduced. This technique is an innovative two-phase strategy to solve the problem of outlier detection when only few positive instances are available for training data. Before discussing the method, the following background on entropy is given.

3.1. Entropy

Entropy is the measure of information and uncertainty of a random variable [29]. If x is a random variable, the entropy $E(x)$ of the probability distribution $p(x)$ on $x = \{x_1, \ldots, x_n\}$ is defined as:

$$E(x) = -\sum_{i=1}^{n} p(x_i) \lg p(x_i) \tag{1}$$

Given a data set including n objects with m attributes, the entropy $E(x)$ of a multivariable vector $\vec{x} = \{X_1, \ldots, X_m\}$, where X_i is a random variable whose realizations belong to the set of $\{x_{i1}, \ldots, x_{in}\}$, can be calculated as Eq. (2).

$$E(\vec{x}) = -\sum_{x_1 \in \{x_{11}, \ldots, x_{1n}\}} \cdots \sum_{x_m \in \{x_{m1}, \ldots, x_{mn}\}} p(x_i, \ldots, x_m) \lg p(x_i, \ldots, x_m) \tag{2}$$

However, with assumption of the objects independency, the joint probability of combined attribute values becomes the product probabilities of each attribute. Therefore $E(\vec{x})$ can be computed as the sum of entropies of attributes, which is shown in Eq. (3).

$$E(\vec{x}) = E(X_1) + E(X_2) + \cdots + E(X_n) \tag{3}$$