Information Sciences 324 (2015) 88–107

Contents lists available at ScienceDirect

Information Sciences

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

A precise ranking method for outlier detection

Jihyun Ha, Seulgi Seok, Jong-Seok Lee*

Department of Industrial Engineering, Sungkyunkwan University, Suwon 440-746, Republic of Korea

ARTICLE INFO

Article history:
Received 29 January 2015
Revised 9 May 2015
Accepted 16 June 2015
Available online 26 June 2015

Keywords:
Outlier detection
Observability factor
Random sampling
Nearest neighbors
Information entropy

ABSTRACT

Recent research studies on outlier detection have focused on examining the nearest neighbor structure of a data object to measure its outlierness degree. This leads to two weaknesses: the size of nearest neighborhood, which should be predetermined, greatly affects the final detection results, and the outlierness scores produced by existing methods are not sufficiently diverse to allow precise ranking of outliers. To overcome these problems, in this research paper, a novel outlier detection method involving an iterative random sampling procedure is proposed. The proposed method is inspired by the simple notion that outlying objects are less easily selected than inlying objects in blind random sampling, and therefore, more inlierness scores are given to selected objects. We develop a new measure called the observability factor (OF) by utilizing this idea. In order to offer a heuristic guideline to determine the best size of nearest neighborhood, we additionally propose using the entropy of OF scores. An intensive numerical evaluation based on various synthetic and real-world datasets shows the superiority and effectiveness of the proposed method.

1. Introduction

Many data mining techniques used for classification, clustering, or association rules mining in general focus on finding the predominant patterns formed by a greater part of the objects in the data. However, outlier detection, also referred to as anomaly detection, is to determine rare but important patterns that may be caused by a very small part of the data. Outliers are also referred to as abnormalities, discordants, deviants, anomalies, etc. A well-known definition of an outlier is given by Hawkins [14]: "*An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism.*" The main idea of this definition is that unexpected observations indicate the user's lack of understanding of a particular process, or are produced by a different process [43]. For this reason, an outlier sometimes contains useful information about the abnormal characteristics of a system, which may reflect the process of data generation [1]. For example, in a computer-related system, an intrusion is very different from the normal behavior of the system, and hence, anomaly detection techniques can be applied to the intrusion detection domain [4]. For meteorological applications, Lu et al. proposed a systematic methodology to detect region outliers in a sequence of meteorological data frames [27]. In [10], a new method for finding outliers in transactional data was proposed to recognize unexpected transactions. Capozzoli et al. used statistical pattern recognition techniques and artificial neural networks in order to automatically detect outliers in building energy consumption [7]. In other applications, such as abnormal weather detection [37], credit card fraud detection [5], insider trading detection [11], medical and public health outlier detection [26], ECG signal analysis [19], image processing [30], abnormal crowd behavior detection [28], anomalous GPS traces

* Corresponding author. Tel.: +82 31 290 7608; fax: +82 31 290 7610.
E-mail addresses: haaforever@skku.edu (J. Ha), smile9188@skku.edu (S. Seok), jongseok@skku.edu (J.-S. Lee).

http://dx.doi.org/10.1016/j.ins.2015.06.030
0020-0255/

[9], and anomalous topic detection in text data [35], the detection of outliers provides valuable domain knowledge and insight. This wide applicability has led to the development of many outlier detection methods.

Outlier detection algorithms have been developed for automatically identifying valuable but irrelevant objects in data. Two scenarios, supervised and unsupervised, are used when developing and applying outlier detection methods. In this study, we developed a new outlier detection algorithm by focusing on the unsupervised case. A supervised scenario is a situation in which a dataset contains information about the class of objects that is normal or abnormal. One-class classifiers, such as one-class support vector machines [32] and support vector data descriptions [39], are available for describing either the normal or the abnormal class. The supervised outlier detection problem is often regarded as a difficult case in classification tasks, which is called unbalanced data classification [15]. In an unsupervised scenario, no information about the class distribution is available. The majority of existing outlier detection algorithms addresses this situation [16]. Most unsupervised outlier detection algorithms share a common model parameter, the size of the neighborhood around an object. Since it is difficult to set the model parameter appropriately, ordinary users sometimes experience poor detection results caused by inappropriate setting of the parameter. In our previous work [13], we pointed out that, since existing algorithms are very sensitive to the choice of model parameter, the development of robust and parameter-free algorithms is a critical issue. In addition, the outlierness scores produced by existing methods are not sufficiently diverse to allow precise ranking of outliers. A good outlier detection method should be able to discriminate a far outlier from a farther outlier. Recently, many outlier detection methods have been developed in various fields of study, but there is still a lack of awareness of these problems.

In this research, we propose a new method for finding outliers to overcome the weaknesses of the existing methods mentioned above. The motivation of the proposed method is the simple notion that outlying objects are not as easily selected as inlying objects in blind random sampling. Imagine a box containing a set of arbitrarily distributed objects. We try to grasp a part of the objects by putting one hand into the box. Objects located at a distance from the others have a lower probability to be selected. By implementing this idea, we develop a new measure that we named the observability factor (OF). The proposed measure indicates the degree of inlierness of each object in a dataset. The object corresponding to a low OF value is therefore a good outlier candidate. Since the value of OF ranges from 0 to 1, it can be interpreted as the probability of a candidate object being an inlier. It is important to note that $1 - \mathrm{OF}$ can be used as the degree of outlierness. A second advantage of the proposed algorithm is that it does not only attempt to detect outliers but also ranks all the objects in the data in their order of outlierness degree. This enables us to offer a guideline for determining the model parameter, the size of the nearest neighborhood, by quantifying the entropy of the computed OF scores of the objects in the data.

The rest of this paper is organized as follows. In Section 2, we review previous research work on outlier detection. Several well-known benchmarking algorithms are briefly introduced in this section. In Section 3, we propose a new outlier detection method and explain our new measure in detail. In Section 4, a numerical evaluation based on various synthetic and real-world datasets is presented that shows the effectiveness of the proposed method. Finally, we conclude the paper in Section 5.

2. Literature review

In the previous section, we mentioned that there are supervised and unsupervised scenarios in outlier detection and our method focuses on the latter case. We now review several well-known approaches to unsupervised outlier detection. Most unsupervised outlier detection methods compute the outlierness score for each object in a dataset. All the objects are then sorted according to the computed scores and a part of those with high outlierness scores are declared outliers. In this section, we briefly describe the main idea of each existing method to help readers establish a better understanding of their characteristics. They all have a common model parameter, the number of nearest neighbors k, to describe the outlierness of an object [34]. We found that most unsupervised outlier detection methods can be grouped into two categories: distance- and density-based approaches. All the existing methods described in this section are compared with the proposed method in Section 4.1. Before describing each of these methods, first we need to define the following. Let $d_k(p)$ denote the distance between an object p and its kth nearest neighbor. The measure of distance can be any appropriate one, such as Euclidean distance, Mahalanobis distance, or some other measure of dissimilarity. The choice of distance measure usually depends on the types of the variables. Let $N_k(p)$ denote a set of k nearest neighbors of an object p, namely,

$$N_k(p) = \{q | d(p,q) \le d_k(p),\ q \ne p\} \tag{1}$$

where $d(p,q)$ is the distance between objects p and q.

2.1. Distance-based approaches

In distance-based approaches, the distances between an object and its nearest neighbors are computed, and then used to measure the outlierness of an object. The basic assumption of the distance-based approaches is that outliers are far apart from their neighbor objects [8]. Several well-known methods based on this idea are briefly described in this subsection.

Ramaswamy et al. [31] proposed simply using the distance between an object and its kth nearest neighbor as an outlierness score. As stated previously, any metrics such as Mahalanobis distance and Euclidean distance can be used to find outliers.