Optik 154 (2018) 145–156

Contents lists available at ScienceDirect

Optik

journal homepage: www.elsevier.de/ijleo

Original research article

Weight-based method for inside outlier detection

Sulan Zhang [a,c,d], Jiaqiang Wan [b,*]

[a] College of Computer Engineering, Yangtze Normal University, Chongqing 408100, China
[b] Country Garden, Foshan City, Guangdong 528312, China
[c] Hyperspectral Collaborative Innovation Center for Green Development in Wuling Mountain Areas, Yangtze Normal University, Chongqing 408100, China
[d] Hyperspectral Remote Sensing Monitoring Center for Ecological Environment of the Three Gorges Reservoir Area, Yangtze Normal University, Chongqing 408100, China

ARTICLE INFO

Article history:
Received 16 June 2017
Accepted 29 September 2017

Keywords:
Outlier detection
Weight-based method
Inside outlier
LOF

ABSTRACT

Outlier detection becomes more and more important in our real life, such as network intrusion detection and credit card fraud detection, etc. In this paper, a weight-based method is proposed for inside outlier detection. According to the concepts of density and volume information, the weight is defined and introduced to construct a new measure of outlierness. Firstly, the total weight of a given object p and its neighbors is computed via their volume and average density. Then the estimated weight of the neighbors is obtained via the neighborhood's volume and p's density. If the total weight is not close to the estimated weight, p is an outlier. The weight-based method shows more superiority in inside outlier detection than LOF in low dimensions. Moreover, the proposed method performs as well as LOD in a high-dimensional space or when no inside outlier exists.

1. Introduction

Outliner detection or anomaly detection devotes to identify non-conforming patterns in data and find extensive use in a wide variety of applications such as network intrusion detection, credit card fraud identification, and valuable user mining [1]. Because of the implicit information contained in all sorts of rare events, outlier detection becomes a hot topic in data mining and machine learning areas [2,3].

Recent years have witnessed various outlier detection approaches which can be categorized as supervised, semi-supervised and unsupervised methods. Among those categorizes, unsupervised methods are more widely applied, because they do not require accurate and representative labels [4]. In unsupervised methods, a score for each point is calculated and the scores are ranked for finding top K outlier candidates. This kind study mainly relies on a measure of distance or similarity in order to detect outliers.

Many algorithms have been proposed to measure the similarity among instances including statistic-based methods [5], cluster-based methods [6,7], density-based methods [8,9] and angle-based methods [10,11]. Local outlier factor (LOF) [12] is one of the classic local outlier detection methods calculating the local density with average of the neighbors' density. Although local outlier detection methods perform better on accuracy than other methods, sometimes miss the inside outliers among local outliers as the density does not take the data shape into consideration.

* Corresponding author.
E-mail addresses: slzhang@cqu.edu.cn (S. Zhang), china_chongqing@sina.com (J. Wan).

https://doi.org/10.1016/j.ijleo.2017.09.116
0030-4026/

In this paper, we propose a weight-based measurement for outlier-ness which could accurately reflect the outlier-ness degree so that outliers can be easily identified. The key step in our outlier detection process is to create an outlier-ness metric which covers both volume and the density of an instance and its neighbors. For an object p and its neighbors, the total weight is computed based on their volume and average density. Then the estimated weight of p's neighbors is obtained via its neighborhood's volume and p's density. If the total weight is not close to the estimated weight, p is considered as an outlier.

Our contributions can be summarized as follows.

(1) Create a new measurement called weight to describe the outlier-ness degree. This measurement takes density and volume information into account. Based on the weight, the outlier-ness is derived and it performs better on measuring the degree of outlier-ness.
(2) Suppose a dataset obeys normal distribution, the degree of outlier-ness for p in m dimensions is defined with its density, mean and variance of the distance between p and its neighbors. Then an algorithm is designed on the degree of outlier-ness.
(3) In the experiments, the weight-based method is analyzed and compared with the LOF method on different datasets, and our method is more superiority in inside outlier detection.

The rest of paper is organized as follows. In Section 2, related work is addressed. In Section 3, we give the definitions for weight-based outlier-ness and degree of outlier-ness for normal distribution datasets and their proofs in detail, and then the algorithm is given in Section 4. Experiments and analysis are shown in Section 5. Last but not least we conduct the conclusion in Section 6.

2. Related work

In supervised and semi-supervised methods, outlier detection is formulated into classification problem which needs labels. Sometimes we have to resort to unsupervised methods sometimes because of the difficult in finding the label information. Therefore, we focus on the unsupervised methods for outlier detection in the following.

Most of previous studies on outlier detection focused on distribution-based approach. A key drawback of this kind of methods is that it is necessary to know or simulate data's distribution. However, most of the distributions used are univariate [13]. In depth-based approach, each object is presented as a point in a k–d space, and is assigned a depth. Chen et al. [14] used kernel spatial depth to develop a method independent of dimensionality. The objects with smaller depth are outliers. However, those algorithms suffer from dimensionality curse. About cluster-based approach, there are many methods such as spectral clustering-based methods and fuzzy c-means based methods. This kind of methods needs to get a clustering result [15,16], and detects outliers according to clusters. The main problem is that clustering increases time complexity but the terminal condition is hard to determine.

Now, density-based approach and distance-based approach attract more and more attentions and there are many advanced methods. The density-based approach identifies outliers as those lying in low-density regions. This type of methods is intuitive and makes good performance. A representative approach is LOF (local outlier factor) [12]. LOF is very outstanding in local outlier detection. Later, Jiang et al. [17] created a generalized local outlier factor (GLOF) on the base of LOF. Seung et al. did some researches to improve LOF's time complexity. However, LOF's performance is hard to be exceeded all the same in outlier detection.

With respect to distance-based approach, Knorr et al. proposed a distance-based approach firstly in 1998 [18], and then improved their method in 1999 [19] and 2000 [20]. Ramaswamy et al. put forward a partition-based method which used the kth nearest point to determine outliers [21]. In 2010, Chen et al. showed a method based on neighborhood [22]. It discriminates outliers by the sum of all attribute distances of neighbors. Szeto and Hung did some researches about improving time complexity [23]. To sum up, for distance-based approach, its rules are too inflexible to exactly detect outliers. Especially, when a distribution is very un-uniform, a sparse point would be mistaken as an outlier.

Yu et al. proposed k local outlier factor that use k-walk similarity to create a new measure based on LOF [24]. They attempted to identify local outliers in the center, where they are similar to some clusters of objects on one hand, and are unique on the other. What is worthy of speaking is Zhang et al.'s method-Local distance-based outlier factor (LDOF) [25]. The method is well-matched with LOF in performance. LDOF uses the relative location of an object to its neighbors to determine the degree to which the object deviates from its neighborhood. However, it is hard to deal with local outliers inside a dataset.

In this paper, we propose a weight-based approach which makes use of density and volume information to construct a new outlier-ness measure. Generally, this approach is easy to detect global outliers. Hence, our attention lies in local outlier detection. Especially, some outliers are inside data distribution. Most methods are used to find the outliers which are far from data's main body, instead of inside outliers. In fact, inside outliers are also very important in real applications. As for inside outlier detection, weight-based method is proposed. Not only does it make good performance in inside outlier detection, but also boasts of property of density-based detection method. Therefore, weight-based method can detect all sorts of outliers, and it is good at inside outlier detection.