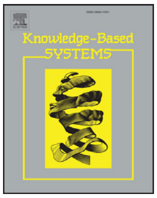
Knowledge-Based Systems 92 (2016) 71–77

Contents lists available at ScienceDirect

Knowledge-Based Systems

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



A non-parameter outlier detection algorithm based on Natural Neighbor

Jinlong Huang, Qingsheng Zhu*, Lijun Yang, Ji Feng

Chongqing Key Laboratory of Software Theory and Technology, College of Computer Science, Chongqing University, Chongqing 400044, China

ARTICLE INFO

Article history:
Received 2 April 2015
Revised 8 October 2015
Accepted 9 October 2015
Available online 30 October 2015

Keywords:
Outlier detection
Natural Neighbor
Natural Outlier Factor

ABSTRACT

Outlier detection is an important task in data mining with numerous applications, including credit card fraud detection, video surveillance, etc. Although many Outlier detection algorithm have been proposed. However, for most of these algorithms faced a serious problem that it is very difficult to select an appropriate parameter when they run on a dataset. In this paper we use the method of Natural Neighbor to adaptively obtain the parameter, named Natural Value. We also propose a novel notion that Natural Outlier Factor (NOF) to measure the outliers and provide the algorithm based on Natural Neighbor (NaN) that does not require any parameters to compute the NOF of the objects in the database. The formal analysis and experiments show that this method can achieve good performance in outlier detection.

1. Introduction

Outlier detection is an important data mining activity with numerous applications, including credit card fraud detection, discovery of criminal activities in electronic commerce, video surveillance, weather prediction, and pharmaceutical research [1–9].

An outlier is an observation that deviates so much from other observations so that it arouses that it is generated by a different mechanism [8]. At present, the studies on outlier detection is very active. Many outlier detection algorithms have been proposed. Outlier detection algorithm can be roughly divided into distribution-based, depth-based, distance-based, clustering-based and density-based act.

In distribution-based methods, the observations that deviate from a standard distribution are considered as outliers [7]. But distribution-based methods not applicable to dataset that multi-dimensional or the distribution unknown. The depth-based [10,11] methods can improve this problem. Depth-based methods relies on the computation of different layers of k–d convex hulls. In this way, outliers are objects in the outer layer of these hulls. However, the efficiency of depth-based algorithms is low on 4-dimensional or more than 4-dimensional dataset. In clustering-based methods, the outliers are by-products of clustering, such as DBSCAN [12], CLARANS [13], CHAMELEON [14], BIRCH [15], and CURE [16]. But the target of clustering-based methods is finding clusters, not detecting outliers, so the efficiency of detecting outliers is low too.

The distance-based algorithms was widely used for the effectiveness and simplification. In paper [4], a distance-based outlier is described as the object that with pct% of the objects in database having a distance of more than d_{min} away from it. However, since distance-based algorithms do not take into account the changes of local density, so distance-based algorithms can only detect the global outliers, fail to detect the local outliers.

The local outliers have received much attention recently. The density-based methods can solve this problem well. And many density-based outlier detection algorithms have been proposed. In paper [17], authors define the concept of a local outlier factor (LOF) that a measure of outlier degree in density between an object and its neighborhood objects. The article [18] made an improved on LOF and proposed an outlier detection algorithm, which defined the influenced outlierness (INFLO) computed by considering both neighbors and reverse neighbors as the outlier degree. This results in a meaningful outlier detection.

Given our motivation, through the above analysis, although the density-based methods can solve problem of local outliers well, density-based methods face the same problem that parameter selection as the first four methods. All of these algorithms almost cannot effectively detect the outliers without appropriate parameter. In other words, most of these algorithms have high dependency to the parameter. Once the parameter changed, the result of outlier detecting would have obvious difference. So the selection of parameter is very important for outlier detection algorithm. In fact, however, determination of parameter is dependent on the knowledge of researcher's experience and a lot of experiment. For example, it is difficult to select an appropriate parameter k that the number of neighbors when use LOF or INFLO to detect the outlier on database.

More detailed analysis of the problem with existing approaches can be available in paper [19]. Paper [19] also propose a new outliers detection algorithm (INS) using the instability factor. INS is

* Corresponding author. Tel.: +86 2365105660; fax: +86 2365104570.
E-mail address: qszhu@cqu.edu.cn (Q. Zhu).

http://dx.doi.org/10.1016/j.knosys.2015.10.014
0950-7051/© 2015 Elsevier B.V. All rights reserved.

insensitive to the parameter$k$ when the value of k is large as shown in Fig. 7(c). However, the cost is that the accuracy is low when the accuracy stabilized. Moreover INS hardly find a properly parameter to detect the local outliers and global outliers simultaneously. In other words, when the value of k is well to detect the global outliers, the effect on local outliers detection is bad, and vice versa.

In this paper, in order to solve the above problem, we first introduce a novel concept of neighbor named Natural Neighbor (NaN) and its search algorithm (NaN-Searching). Then we obtain the number of neighbors, the value of parameter k, use the NaN-Searching algorithm. We also define a new concept of Natural Influence Space (NIS) and Natural Neighbor Graph (NNG), and compute the Natural Outlier Factor (NOF). The bigger the value of NOF is, the greater the possibility of object is outlier.

The paper is organized as follows. In Section 2, we present the existing definition and our motivation. In Section 3, properties of Natural Neighbor are introduced. In Section 4, we propose a outlier detection algorithm based on Natural Neighbor. In Section 5, a performance evaluation is made and the results are analyzed. Section 6 concludes the paper.

2. Related work

In this section, we will briefly introduce concept of LOF and INS. LOF is a famous density-based outlier detection algorithm. And INS is a novel outlier detection algorithm proposed in 2014. Interested readers are referred to papers [17] and [19].

Let D be a database, p, q, and o be some objects in D, and k be a positive integer. We use $d(p,q)$ to denote the Euclidean distance between objects p and q.

Definition 1 (k-distance and nearest neighborhood of ***p***)**.** The k-distance of p, denoted as $k_{dist}(p)$, is the distance $d(p,o)$ between p and o in D, such that:

(1) For at least k objects $o' \in D/\{p\}$ is holds that $d(p, o') <= d(p, o)$, and

(2) For at most $(k-1)$ objects $o' \in D/\{p\}$ it holds that $d(o, o') < d(p, o)$

The $k_{dist}(p)$ can reflect the density of the object p. The smaller $k_{dist}(p)$ is, the much denser the area around p is.

Definition 2 (K-Nearest neighborhood)**.** The k-nearest neighborhood of p, $NN_k(p)$ is a set of objects X in D with $d(p, X) <= k_{dist}(p)$: $NN_k(p) = \{X \in D/\{p\} | d(p, X) <= k_{dist}(p)\}$.

Note that the number of objects in k-nearest neighborhood of p may be more than k. In other words, there may be more than k objects within $NN_k(p)$.

Definition 3 (Reachability distance of p w.r.t object o)**.** The reachability distance of object p with respect to object o is defined as follows:

$$reach-dist_k(p, o) = max\{k - distance(o), d(p, o)\} \tag{1}$$

Definition 4 (Local reachability density of p)**.** The local reachability density of an object p is the inverse of the average reachability distance from the k-nearest-neighbors of p. This is defined as follows:

$$lrd_k(p) = 1 \Bigg/ \frac{\sum_{o \in NN_k(p)} reach - dist_k(p, o)}{|NN_k(p)|} \tag{2}$$

Essentially, the local reachability density of an object p is the reciprocal of the average distance between p and the objects in its k-neighborhood. Based on local reachability density, paper [17] defines the local outlier factor as follows:

$$LOF_k(p) = \frac{\sum_{o \in NN_k(p)} \frac{lrd_k(o)}{lrd_k(p)}}{|NN_k(p)|} \tag{3}$$

Obviously, LOF is the average of the ratios of the local reachability density of p and p's k-nearest-neighbors. We can think about it in this way that LOF is the ratios of the local reachability density of p and the average local reachability density of p's k-nearest-neighbors. Intuitively, p's local outlier factor will be very high if its local reachability density is much lower than those of its neighbors. In this way, the bigger p's local outlier factor is, the more likely p is outlier.

Although LOF has been used widely, there are some problem existed in it. It is the main problem that LOF is sensitive to parameters. To solve this problem, paper [19] proposed a new algorithm (INS) using the instability factor. The follows are some briefly introduce to INS.

Definition 5 (The k center of gravity)**.** The k center of gravity of p is defined as a centroid of the objects in $NN_k(p)$, which is given by

$$m_k(p) = \frac{1}{k+1} \sum_{q \in NN_k(p)} X_q \tag{4}$$

where $X_q = (x_{q1}, x_{q2}, !‘, x_{qd})$ is the coordinates of the object q observed in a d-dimensional space (under the assumption that the space is Euclidean).

Let $_i(p)$ denote the distance between $m_i(p)$ and $m_{(i+1)}(p)$, which is defined by the following equation:

$$_i(p) = d(m_i(p), m_{(i+1)}(p)), i = 1, 2, \ldots, k-1 \tag{5}$$

Definition 6 (Absolute difference)**.** The absolute difference between $\theta_i(p)$ and $\theta_{(i+1)}(p)$, denoted as $\Delta\theta_i(p)$, which is defined as:

$$\Delta\theta_i(p) = |\theta_i(p) - \theta_{(i+1)}(p)|, i = 1, 2, \ldots, k-2 \tag{6}$$

Definition 7 (Instability factor)**.** The instability factor, and INS(p,k) are defined by the following equation:

$$INS(p, k) = \sum_{i=1}^{k-2} \Delta\theta_i(p) \tag{7}$$

INS improve the problem that are sensitive to parameter. The changes of accuracy, using INS, were minor when the parameter is changed. And INS can be flexibly used for both local and global detection of outliers by controlling its parameter. But the accuracy is not high when the accuracy stabilized. Moreover INS hardly find a properly parameter to detect the local outliers and global outliers simultaneously.

Though above analysis, we know that LOF and INS have their own advantages and disadvantages. However, no matter LOF or INS, there is a same problem that parameter selection. It is difficult to select an appropriate parameter k that the number of neighbors.

In order to solve the problem that parameter selection, we introduce the concept of Natural Neighbor. Natural Neighbor can adaptively obtain the appropriate value of k that the number of neighbors without any parameters. The outlier detection algorithm that we proposed in this paper use the k as parameter to detect the outliers. In the section that follows, a detailed introduction will be made to Natural Neighbor.

3. Natural Neighbor and NOF algorithm

3.1. NaN definition and algorithm

Natural Neighbor is a new concept of neighbor. The concept originates in the knowledge of the objective reality. The number of one's real friends should be the number of how many people are taken him or her as friends and he or she take them as friends at the same time. For data objects, object y is one of the Natural Neighbor of object x if object x considers y to be a neighbor and y considers x to be a neighbor at the same time. In particular, data points lying in sparse region should have small number of neighbors, whereas data points lying in