Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition Letters 29 (2008) 967–976
Pattern Recognition Letters
www.elsevier.com/locate/patrec

LDBOD: A novel local distribution based outlier detector

Yong Zhang [a], Su Yang [a,*], Yuanyuan Wang [b]

[a] Department of Computer Science and Engineering, Fudan University, Shanghai 200433, PR China
[b] Department of Electronic Engineering, Fudan University, Shanghai 200433, PR China

Received 30 March 2007; received in revised form 29 October 2007
Available online 6 February 2008

Communicated by R.P.W. Duin

Abstract

As an important research direction in KDD field, outlier detection has been drawing much attention from different communities. In this paper, two novel algorithms LDBOD and LDBOD+ for outlier detection are proposed. Similar to LOF, they also aim to find local outliers. However, LDBOD/LDBOD+ detects local outliers from the viewpoint of local distribution, which is characterized through three proposed measurements, local-average-distance, local-density, and local-asymmetry-degree. Several experiments were conducted to demonstrate the advantages of LDBOD/LDBOD+ compared with LOF.

Keywords: Outlier detection; Symmetry; Local distribution

1. Introduction

Different from clustering analysis, which aims at discovering the underlying structure of a data set, the goal of outlier detection is to explore abnormal data patterns corresponding with a minority of samples in a given data set. Outlier detection has many synonyms like novelty detection, anomaly detection, deviation detection, and exception mining (Hodge and Austin, 2004).

From the viewpoint of clustering analysis, outliers are a nuisance that must be quickly identified and eliminated so that they do not interfere with the analysis process. However, outliers contain useful information for many other important applications such as fraud detection, network intrusion detection, video surveillance, and weather prediction (Hodge and Austin, 2004). Hence, it is not a surprise that outlier detection has drawn much attention from different communities.

Although many algorithms (Lee and Cho, 2006; Jiang et al., 2006; Hu and Sung, 2003; Cao et al., 2003; He et al., 2003; Jiang et al., 2001; Aggarwal and Yu, 2001; Knorr and Ng, 1998; Breunig et al., 2000; Ramaswamy et al., 2000; Hodge and Austin, 2004; Jin et al., 2006) have been proposed in recent years for outlier detection, a generally accepted formal definition of outliers does not exist in the literature. Among all the different definitions, Hawkins' definition is used frequently, that is, "an outlier is an observation that deviates so much from other observations as to arouse suspicions that it is generated by a different mechanism" (Knorr and Ng, 1998). Obviously, this is still an abstract definition of an outlier, which can be quantified by different measurements in different algorithms.

Most of the previous studies on outlier detection are conducted in the field of statistics (Hodge and Austin, 2004). In this context, data points are modeled using an assumed stochastic distribution and points are determined to be outliers depending on how well they fit into this model. Just as pointed out by most researchers, a key drawback of such algorithms is that most of the distributions used are univariate and for many real applications, the underlying distribution is unknown at all. Another category of outlier detection algorithms in the sense of statistics is depth-based. Each data sample is represented as a point in a k-d space and assigned a

* Corresponding author. Tel.: +86 (21) 55664412.
E-mail address: suyang@fudan.edu.cn (S. Yang).

0167-8655/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2008.01.019

depth according to a certain definition. Then, the data points with smaller depths are more likely to be outliers. Theoretically, depth-based approaches could work with high-dimensional data. However, in practice, they become inefficient rapidly for large-scale and high-dimensional data sets due to the high time complexity in computing the k-d convex hull (Breunig et al., 2000).

In order to overcome these problems aforementioned, Knorr and Ng (Knorr and Ng, 1998) proposed a distance-based definition for outliers as well as a number of efficient algorithms for finding this type of outliers. Although the distance-based approaches could detect outliers efficiently without any prior knowledge about the data set of interest, the parameters like the distance threshold d and the portion p could be very difficult to determine, which usually requires several trials even for expert users. In (Ramaswamy et al., 2000), the authors extended the definition of the distance-based outliers using the distance to the kth-nearest neighbor, by which outliers can be ranked.

Different from the techniques discussed above, in (Breunig et al., 2000), the concept of local outliers was introduced. Unlike the distance-based outliers, local outliers are defined based on the densities of their neighborhoods. Specifically, the degree of outlierness of a data object is defined to be the ratio of its density to the average density of its neighboring objects.

The outlierness of an object usually appears to be more outstanding in contrast to its local neighborhood. For example, a network intrusion might cause a significant spike in the number of network events within a low traffic period, but this spike might be insignificant when a period of high network traffic is also included in the comparison (Jin et al., 2006). The LOF introduced in (Breunig et al., 2000) captures this very well. However, it only considers the local density factor that may not be enough to describe the characteristics of an outlier sufficiently.

Most of the existing outlier detection algorithms only take the number, distance, or density factor of the neighboring points of a point into account. The distribution of the neighboring points themselves is not considered. Compared with the previously proposed the density or distance-based measurements, the local distribution reveals more information so that it can describe the characteristics of outliers more appropriately. Hence, we try to utilize the distribution of the neighboring points to characterize an outlier in this paper. Two novel local outlier detectors LDBOD (Local Distribution Based Outlier Detector) and LDBOD+ are proposed in the following sections.

The main contributions of this paper are as follows:

- We propose using the homogeneousness degree of the local distribution to detect outliers. To the best of our knowledge, this is not considered in the previous studies of outlier detection.
- We propose two novel local outlier detection algorithms that are more efficient compared with LOF. We performed intrusion detection experiment. The proposed algorithms achieve competitive results compared with LOF, but the speed is faster.

For the two algorithms, the key issue lies in characterizing the local distribution of a point quantitatively. We discuss the details of local distribution measurement in Section 2. Section 3 presents the proposed algorithms LDBOD and LDBOD+. The experimental results are presented in Section 4. Finally, we conclude in Section 5.

2. Local distribution

2.1. Neighborhood selection

As mentioned earlier, we are interested in local outliers defined by the distributions of their neighboring points. Hence, at first, we must select a suitable neighborhood of interest for a data point. In other words, we must construct a neighborhood diagram among all the data points explicitly or implicitly. There exist many kinds of neighborhood diagrams, among which kNN diagram, ε-diagram, and Delaunay diagram are used frequently in related works.

As we know, the Voronoi diagram captures the proximity uniquely and represents the topology explicitly with its dual graph known as the Delaunay Triangulation. Each Delaunay edge is an explicit representation of a neighborhood relation between two points. However, Delaunay Triangulations are not unique when co-circularity occurs. Delaunay diagram, which is the result of removing all such edges in the Delaunay Triangulation that the three vertices of the corresponding triangles are co-circular with a fourth point, can be used instead. Even in the presence of co-circularity, Delaunay diagram guarantees a unique topology (Estivill-Castro and Lee, 2000). Fig. 1b shows the corresponding Delaunay diagram for a simple synthetic data set.

Compared with kNN diagram and ε-diagram, the key advantage of Delaunay diagram lies in that it is parame-

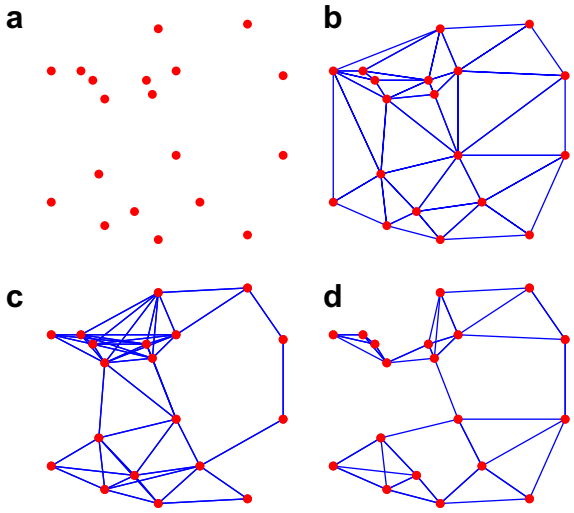

Fig. 1. An illustration of different neighborhood diagrams: (a) Original data points, (b) Delaunay diagram, (c) ε-diagram, and (d) kNN diagram (k=3).