Knowledge-Based Systems 63 (2014) 15–23

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Robust outlier detection using the instability factor

Jihyun Ha, Seulgi Seok, Jong-Seok Lee *

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

ARTICLE INFO

Article history:
Received 15 August 2013
Received in revised form 1 March 2014
Accepted 1 March 2014
Available online 24 March 2014

Keywords:
Outlier detection
Noise removal
Instability factor
Nearest neighbors
Data mining

ABSTRACT

Since outlier detection is applicable to various fields such as the financial, telecommunications, medical, and commercial industries, its importance is radically increasing. Receiving such great attention has led to the development of many detection methods, most of which pertain to either the distance-based approach or the density-based approach. However, each approach has intrinsic weaknesses. The former hardly detects local outliers, while the latter has the low density patterns problem. To overcome these weaknesses, we proposed a new detection method that introduces the instability factor of a data point by utilizing the concept of the center of gravity. The proposed method can be flexibly used for both local and global detection of outliers by controlling its parameter. In addition, it offers the instability plot containing useful information about the number and size of clusters in data. Numerical experiments based on artificial and real datasets show the effectiveness of the proposed method.

1. Introduction

Outlier detection (also known as anomaly detection) refers to detecting patterns in data that do not conform to an established normal behavior [3]. Either detecting or eliminating outliers is a necessary preprocessing step in data analysis. Filtering out noise information from data helps to improve the performance of predictive models. On the other hand, outlier detection sometimes helps us discover useful information from the detected outliers [15]. For example, we can recognize how irregular credit card transactions occur by examining the detected fraudulent transactions, which are considered as outliers in the credit card database. Nonetheless, this research considers outlier detection as noise removal. Two scenarios (*supervised* and *unsupervised*) are used when developing and applying outlier detection methods. The former is a situation in which a dataset contains information about the class of objects that is normal or abnormal, and the latter does not have such class information. In this study, we developed a new method for outlier detection by focusing on the unsupervised case. Since outlier detection can be applied to various fields such as intrusion detection, fraud detection, fault detection, health monitoring systems, and detection of ecosystem disturbances, many detection methods have been proposed. Earlier studies have used traditional statistical methods using reference distributions [1,8,22], depth-based approaches [11], and clustering approaches [9]. However, most of the recent detection methods are categorized as either distance-based or density-based approaches.

Distance-based approaches basically compute the distance between an object and its nearest neighbor, and then compare the distances for all of the objects. The underlying principle is that an object that is far from its neighbors is likely to be an outlier [13]. Examples of distance-based approaches include the $DB(\varepsilon, \pi)$-outliers method proposed by Knorr and Ng [14], the outlier scoring method based on k-nearest neighbor (kNN) distances [20], the in-degree-based detection method [7], and the resolution-based outlier factor developed by Fan et al. [5]. These approaches are simple and fast, and are therefore applied to large datasets. However, they do not work properly if various degrees of cluster density exist in data. Also, it is not easy to select an appropriate value for the model parameters that determine the size of the neighborhood around an object, which is critical to detection performance.

Density-based approaches are distinct from distance-based approaches in that they compare the density of an object's neighborhood with that of its neighbor's neighborhood. As the density difference becomes relatively larger than that of the other objects, the object under consideration is likely to be an outlier. An early study was conducted by Breunig et al. [2], and their method is called the local outlier factor (LOF). Rather than examining an individual object, the local correlation integral (LOCI) method proposed by Papadimitriou et al. [19] looks for groups of outliers. Jin et al. [10] developed a method of ranking outliers that considers both the neighbors and the reverse neighbors of an object in order to estimate its density distribution. The LOF method was extended

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

http://dx.doi.org/10.1016/j.knosys.2014.03.001
0950-7051/

in a study by Kriegel et al. [16]. They suggested local outlier probabilities (LoOP) ranging from 0 to 1, so that the outlierness scores can be interpreted directly. These methods perform well, although a given dataset has various degrees of cluster density. However, if low density patterns exist in data, the detection performance of these methods rapidly deteriorates [23]. Since density-based approaches are sensitive to the choice of parameter used to determine the size of the neighborhood to be examined, choosing the best parameter is a critical issue, as is the case with distance-based approaches.

In this study we propose a new method, based on the notion of the center of gravity in physics, which could overcome the weaknesses of the existing approaches. The center of gravity, which is the same as the center of mass, represents a geometric property of an object. It divides the weight of an object, so that it is not biased. An object with a low center of gravity from its surface is balanced, whereas an object with a high center of gravity is unstable. Our proposed method was inspired by the fact that the center of gravity is related to the stability of an object, and we applied this notion to develop a new outlier detection method. Each object in data is included in a group with its nearest neighbors, and the center of gravity of the group changes as the size of the group increases. If an outlier is included in a certain neighborhood set, then the size of the neighborhood increases. The center of gravity then moves from the current location to another location. The movement variation is greater for an outlier than for normal objects. Using this property, we introduced a new measure for outlierness that we named the *instability factor*. Similar to existing distance-based and density-based approaches, the algorithm of our method also begins with finding nearest neighbors of an object in order to find the center of gravity in the neighborhood set. Since the subsequent step is to measure distances between consecutive centers of gravity, the proposed method could be called a distance-based approach. However, our proposed method is distinct from other distance-based methods in that it does not directly utilize distances between objects. The proposed method has several properties that are better than those of existing detection methods.

The rest of this paper is organized as follows. In Section 2, we discuss the weaknesses of distance-based and density-based approaches. In Section 3, we propose a new method for outlier detection, and describe its strengths compared to those of existing methods. In Section 4, we discuss the results of numerical experiments based on synthetic and real datasets that were conducted to show the effectiveness of the proposed method. Section 5 features the conclusions of this study.

2. Problems with existing approaches

As we briefly mentioned in the previous section, there are several known weaknesses of the existing distance-based and density-based approaches. These include local density problem, low density patterns problem, and vulnerability to model parameters. In this section we review the problems in detail with illustrative examples.

2.1. Local density problem

The local density problem, which is related to different degrees of cluster density, generally occurs with distance-based approaches [2]. Fig. 1 shows a representative example of the local density problem. The 2-dimensional dataset in this figure consists of a total of 201 objects: 100 objects forming a dense cluster at the left corner, another 100 objects in a sparse cluster, and one object colored red that is regarded as an outlier. Since distance-based approaches rank outliers based on the distances between an object and its neighbors, they fail to detect the red outlier. Instead, such approaches declare the sparse cluster as outliers, which is definitely not our expectation. The local density problem can be addressed by introducing the concept of *local outliers*, which is what density-based approaches practice in their algorithms. However, there is another problem with density-based approaches, which is described in the next subsection.

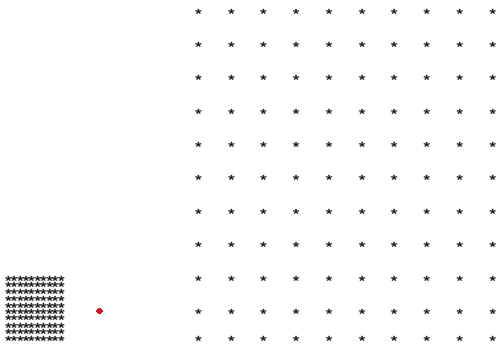

Fig. 1. Example illustrating a local density problem.

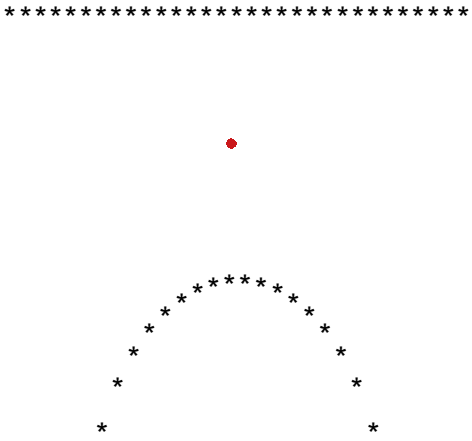

Fig. 2. Example illustrating a low density patterns problem.

2.2. Low density patterns problem

If an outlier is located near patterns or clusters with low densities, then density-based approaches barely detect the outlier [23]. This is called the local density patterns problem. This problem arises when there is not a large difference between the neighborhood density of an outlier and the neighborhood densities of its nearest objects. The example shown in Fig. 2 helps to illustrate the low density patterns problem, where the object that we would like to detect as an outlier is also colored red.[1]

Density-based approaches must first decide the size of the neighborhood of an object; in this example we chose 10. Imagine the smallest circle containing the red point and its 10 nearest neighbors, and another circle that encloses any point on the straight line and its 10 nearest neighbors. Since the densities of these circles are similar, comparing them becomes useless for detecting the outlier. This example demonstrates that density-based approaches would not work for cases of outliers with

[1] For interpretation of color in Figs. 1–3,5,6,8–10, the reader is referred to the web version of this article.