Enhancing Effectiveness of Outlier Detections for Low Density Patterns

Jian Tang[1⋆], Zhixiang Chen[2], Ada Wai-chee Fu[1], and David W. Cheung[3]

[1] Department of Computer Science and Engineering
Chinese University of Hong Kong
Shatin, Hong Kong
[2] Department of Computer Science
University of Texas at Pan-America
Texas, U.S.A
[3] Department of Computer Science and Information Systems
University of Hong Kong
Pokfulam, Hong Kong

Abstract. Outlier detection is concerned with discovering exceptional behaviors of objects in data sets. It is becoming a growingly useful tool in applications such as credit card fraud detection, discovering criminal behaviors in e-commerce, identifying computer intrusion, detecting health problems, etc. In this paper, we introduce a connectivity-based outlier factor (COF) scheme that improves the effectiveness of an existing local outlier factor (LOF) scheme when a pattern itself has similar neighbourhood density as an outlier. We give theoretical and empirical analysis to demonstrate the improvement in effectiveness and the capability of the COF scheme in comparison with the LOF scheme.

1 Introduction

Outlier detection is an important branch in the area of data mining. It is concerned with discovering the exceptional behaviors of certain objects. Revealing these behaviors is important since it signifies that something out of ordinary has happened and shall deserve people's attention. In many cases, such exceptional behaviors will cause damage to users and must be stopped. In other cases, there can be "good" outliers which can help users to make profits. Therefore, in some sense detecting outliers is at least as significant as discovering general patterns. Outlier detection is becoming a growingly useful tool in applications to which people have already paid attention, such as credit card fraud detection, calling card fraud detection, discovering criminal behaviors in e-commerce, discovering computer intrusion, and etc. [4, 6].

Hawkins [7] characterizes an outlier in a quite intuitive way as follows:

> *An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism.*

⋆ On leave from Memorial University of Newfoundland, Canada.

M.-S. Chen, P.S. Yu, and B. Liu (Eds.): PAKDD 2002, LNAI 2336, pp. 535–548, 2002.

Following the spirit of this definition, researchers have proposed various schemes for outlier detection. A large amount of the work was under the general topic of clustering [5, 10, 13, 14, 16]. These algorithms can also generate outliers as by-products. However, the outliers discovered this way are highly dependent on the clustering algorithms used and hence subject to the clusters generated. Most methods in the early work that detects outliers independently have been developed in the field of statistics [2]. These methods normally assume that the distribution of a data set is known in advance and try to detect outliers by examining the deviations of individual data objects based on such a distribution. In reality, however, a priori knowledge about the distribution of a data set is not always obtainable. Besides, these methods do not scale well for even modest number of dimensions as the size of a data set increases.

More recently, researchers proposed distance based schemes, which distinguish objects that are likely to be outliers from those that are not based on the number of objects in the neighborhood of an object [8, 9, 11]. These schemes do not make any assumptions about the distribution of a data set. Furthermore, since the counting process is restricted only to the neighborhood of an object, the scalability of these methods is better than that of their predecessors. As a result, distance based schemes are more appropriate for detecting outliers in large data sets without assuming a priori knowledge about their distributions.

Knorr and Ng [8] propose a distance based scheme, called $DB(n, q)$-outlier. In this scheme, if the neighborhood with the radius of q (called "q-neighborhood") of an object contains less than n objects, then it is called an outlier with respect to n and q, otherwise it is not. The advantage of this scheme is its simplicity while capturing the basic intuition given in Hawkins' definition. Its weakness is that it cannot deal with data sets that contain patterns with diverse characteristics. The scheme proposed by Ramaswamy, et al. [11], called (t, k)-nearest neighbor scheme, considers for each point its k-distance, i.e., the distance to its kth nearest neighbor(s). It ranks the top t objects with the maximum k-distances as the outliers. If there are multiple objects with the same k-distance ranked as the top k, they are all considered as outliers. Therefore, the number of outliers returned may be greater than t. This scheme is actually a special case of $DB(n, q)$-outlier. Thus it shares the same weakness as $DB(n, q)$-outlier has.

Recently, Breuning, et al, [3] proposed a density based formulation scheme as follows.

Let $p, o \in \mathcal{D}$ and k be a positive integer. Let $k\text{-}distance(o)$ be the distance from o to its k-th nearest neighbor, where if two neighbors are at same distance from o, the ordering of "nearest" for them is arbitrary. The k-distance neighbourhood of an object p is denoted by $N_{k\text{-}distance(p)}(p)$ and is the set of objects whose distance from p is not greater than k-distance.

The reachability distance of p with respect to o for k is defined as:

$$reach\text{-}disk_k(p, o) = max\{k\text{-}distance(o), dist(p, o)\}.$$