Ranking Outliers Using Symmetric Neighborhood Relationship

Wen Jin[1], Anthony K.H. Tung[2], Jiawei Han[3], and Wei Wang[4]

[1] School of Computing Science, Simon Fraser University
wjin@cs.sfu.ca
[2] Department of Computer Science, National University of Singapore
atung@comp.nus.edu.sg
[3] Department of Computer Science, Univ. of Illinois at Urbana-Champaign
hanj@cs.uiuc.edu
[4] Department of Computer Science, Fudan University
weiwang1@fudan.edu.cn

Abstract. Mining outliers in database is to find exceptional objects that deviate from the rest of the data set. Besides classical outlier analysis algorithms, recent studies have focused on mining **local outliers**, i.e., the outliers that have density distribution significantly different from their neighborhood. The estimation of density distribution at the location of an object has so far been based on the density distribution of its k-nearest neighbors [2, 11]. However, when outliers are in the location where the density distributions in the neighborhood are significantly different, for example, in the case of objects from a sparse cluster close to a denser cluster, this may result in wrong estimation. To avoid this problem, here we propose a simple but effective measure on local outliers based on a symmetric neighborhood relationship. The proposed measure considers both neighbors and reverse neighbors of an object when estimating its density distribution. As a result, outliers so discovered are more meaningful. To compute such local outliers efficiently, several mining algorithms are developed that detects top-n outliers based on our definition. A comprehensive performance evaluation and analysis shows that our methods are not only efficient in the computation but also more effective in ranking outliers.

1 Introduction

From a knowledge discovery standpoint, outliers are often more interesting than the common ones since they contain useful information underlying the abnormal behavior. Basically, an outlier is defined as an exceptional object that deviates much from the rest of the dataset by some measure. Outlier detection has many important applications in fraud detection, intrusion discovery, video surveillance, pharmaceutical test and weather prediction. Various data mining algorithms [1, 2, 3, 8, 10, 11, 12, 13, 15, 18, 17, 20, 21, 23, 24, 25] for outlier detection were proposed. The outlierness of an object typically appears to be more outstanding with respect to its local neighborhood. For example, a network intrusion might

W.K. Ng, M. Kitsuregawa, and J. Li (Eds.): PAKDD 2006, LNAI 3918, pp. 577–593, 2006.

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. In view of this, recent work on outlier detection has been focused on finding **local outliers**, which are essentially objects that have significantly lower density [1] than its local neighborhood [2]. As an objective measure, the degree of outlierness of an object p is defined to be **the ratio of its density and the average density of its neighboring objects** [2]. To quantify what are p's neighboring objects, users must specify a value k, and neighboring objects are defined as objects which are not further from p than p's k^{th} nearest objects [2]. As an example, let us look at Figure 1 in which k is given a value of 3. In this case, the three neighboring objects of p will have higher density than p and thus p will have a high degree of outlierness according to the definition in [2]. This is obviously correct based on our intuition.

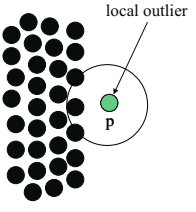

Fig. 1. A local outlier, p

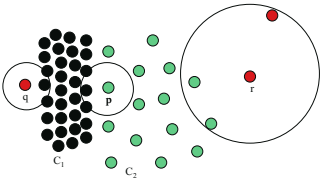

Fig. 2. Comparing the outlierness of p, q, r

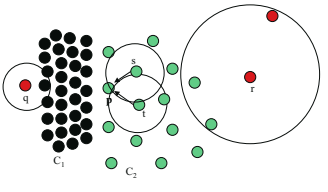

Fig. 3. Taking RNNs of p into account

Unfortunately, the same cannot hold in more complex situation. Let us look at the following example.

Example 1: We consider Figure 2 in which p is in fact part of a sparse cluster C_2 which is near the dense cluster C_1. Compared to objects q and r, p obviously displays less outlierness. However, if we use the measure proposed in [2], p could be mistakenly regarded to having stronger outlierness in the following two cases:
Case I: The densities of the nearest neighboring objects for both p and q are the same, but q is slightly closer to cluster C_1 than p. In this case, p will have a stronger outlierness measure than q, which is obviously wrong.

[1] The density of an object p is defined as $1/k_{dist}(p)$ where k is a user-supplied parameter and $k_{dist}(p)$ is the distance of the k^{th} nearest object to p.

[2] Note that p 's k^{th} nearest neighbor might not be unique and thus p could have more than k neighboring objects.