A New Local Distance-Based Outlier Detection Approach for Scattered Real-World Data

Ke Zhang[1], Marcus Hutter[1,2], and Huidong Jin[1,2,3]

[1] RSISE, Australian National University
[2] National ICT Australia (NICTA), Canberra Lab, ACT, Australia
[3] CSIRO Mathematical and Information Sciences, Acton ACT 2601, Australia
{ke.zhang, marcus.hutter}@rsise.anu.edu.au, Warren.Jin@csiro.au

Abstract. Detecting outliers which are grossly different from or inconsistent with the remaining dataset is a major challenge in real-world KDD applications. Existing outlier detection methods are ineffective on scattered real-world datasets due to implicit data patterns and parameter setting issues. We define a novel *Local Distance-based Outlier Factor* (LDOF) to measure the outlier-ness of objects in scattered datasets which addresses these issues. LDOF uses the relative location of an object to its neighbours to determine the degree to which the object deviates from its neighbourhood. We present theoretical bounds on LDOF's false-detection probability. Experimentally, LDOF compares favorably to classical KNN and LOF based outlier detection. In particular it is less sensitive to parameter values.

Keywords: local outlier; scattered data; k-distance; KNN; LOF; LDOF.

1 Introduction

Of all the data mining techniques that are in vogue, outlier detection comes closest to the metaphor of mining for nuggets of information in real-world data. It is concerned with discovering the exceptional behavior of certain objects [TCFC02]. Outlier detection techniques have widely been applied in medicine (e.g. adverse reactions analysis), finance (e.g. financial fraud detection), security (e.g. counter-terrorism), information security (e.g. intrusions detection) and so on. In the recent decades, many outlier detection approaches have been proposed, which can be broadly classified into several categories: distribution-based [Bar94], depth-based [Tuk77], distance-based (e.g. KNN) [KN98], cluster-based (e.g. DBSCAN) [EKSX96] and density-based (e.g. LOF) [BKNS00] methods.

However, these methods are often unsuitable in real-world applications due to a number of reasons. Firstly, real-world data usually has a scattered distribution, where objects are loosely distributed in the domain feature space, e.g. like stars in the universe, forming many mini-clusters rather than a few main clusters. Only the objects which do not belong to any mini-cluster are genuine outliers. Secondly, most outlier detection approaches require fine-tuning of their parameters through trial-and-error [FZFW06], which is impractical, because real-world data

T. Theeramunkong et al. (Eds.): PAKDD 2009, LNAI 5476, pp. 813–822, 2009.

usually do not contain labels for anomalous objects. Top-n style outlier detection methods alleviate the parameter setting problem somewhat. They provide a ranked list of objects that represent the degree of 'outlier-ness' of each object.

Top-n K^{th}-Nearest Neighbour distance [RRS00] is a typical top-n style outlier detection approach. In top-n KNN outlier, the distance from an object to its k^{th} nearest neighbour (denoted as k-distance for short) indicates outlier-ness of the object. Intuitively, the larger the k-distance is, the higher outlier-ness the object has. Top-n KNN outlier regards the n objects with the highest values of k-distance as outliers [RRS00].

A density-based outlier, Local Outlier Factor (LOF) [BKNS00], was proposed in the same year as top-n KNN. In LOF, an outlier factor is assigned for each object w.r.t its surrounding neighbourhood. The outlier factor depends on how the data object is closely packed in its locally reachable neighbourhood [FZFW06]. In recent real-world applications, researchers have found it more reliable to use LOF in a top-n manner [TCFC02], i.e. only objects with the highest LOF values will be considered outliers. Hereafter, we call it top-n LOF.

In this paper, we propose a new outlier detection definition which is sensitive to outliers in scattered datasets, named Local Distance-based Outlier Factor (LDOF). LDOF uses the relative distance from an object to its neighbours to measure how much objects deviate from their scattered neighbourhood. The higher the violation degree an object has, the more likely the object is an outlier.

In Section 2, we illustrate and discuss the problems of top-n KNN and top-n LOF on a real-world dataset. In Section 3, we formally introduce the outlier definition of our approach, and mathematically analyse properties of our outlier-ness factor in Section 4. In Section 5, the top-n LDOF outlier detection algorithm is described, together with an analysis of its complexity. Experiments are reported in Section 6, which show the superiority of our method to previous approaches, at least on the considered datasets. Finally, conclusions are presented in Section 7.

2 Problem Formulation

In real-world datasets, high dimensionality (e.g. 30 features) and sparse feature value range usually cause objects to be scattered in the feature space. The scattered data is similar to the distribution of stars in the universe. Locally, they seem to be randomly allocated in the night sky (i.e. stars observed from the Earth), whereas globally the stars constitute innumerable galaxies. Figure 1(a) illustrates a 2-D projection of a real-world dataset, *Wisconsin Diagnostic Breast Cancer* (WDBC)[1], which is typically 30-D. The green points are the benign diagnosis records (regarded as normal objects), and the red triangles are malignant diagnosis records (i.e. outliers we want to capture). Obviously, we cannot detect these outliers in 2-D space, whereas in high dimension (e.g. 30-D), these scattered normal objects constitute a certain number of loosely bounded mini-clusters, and we are able to isolate genuine outliers. Unlike galaxies, which always contain billions of stars, these mini-clusters in scattered datasets usually have

[1] WDBC dataset is from UCI ML Repository: http://archive.ics.uci.edu/ml