Pattern Recognition Letters 60–61 (2015) 24–31

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Outlier detection using neighborhood rank difference ☆

Gautam Bhattacharya [a], Koushik Ghosh [b], Ananda S. Chowdhury [c,*]

[a] Department of Physics, University Institute of Technology, University of Burdwan, Golapbag (North), Burdwan 713104, India
[b] Department of Mathematics, University Institute of Technology, University of Burdwan, Golapbag (North), Burdwan 713104, India
[c] Department of Electronics and Telecommunication Engineering, Jadavpur University, Kolkata 700032, India

ARTICLE INFO

Article history:
Received 28 August 2014
Available online 21 April 2015

Keywords:
Outlier
Rank-difference
kNN
RNN

ABSTRACT

Presence of outliers critically affects many pattern classification tasks. In this paper, we propose a novel dynamic outlier detection method based on neighborhood rank difference. In particular, reverse and the forward nearest neighbor rank difference is employed to capture the variations in densities of a test point with respect to various training points. In the first step of our method, we determine the influence space for a given dataset. A score for outlierness is proposed in the second step using the rank difference as well as the absolute density within this influence space. Experiments on synthetic and some UCI machine learning repository datasets clearly indicate the supremacy of our method over some recently published approaches.

1. Introduction

The problem of outlier detection is of great interest to the pattern recognition community. The major objective of an outlier detection method is to find the rare or exceptional objects with respect to the remaining (large amount of) data [1]. Outlier detection has several practical applications in diverse fields, e.g., in fault detection of machines [2,3], in anomaly detection in hyperspectral images [4], in novelty detection of image sequence analysis [5], in biomedical testing [6,7], in weather prediction [8], in geoscience and remote sensing [9], in medicine [10], in financial fraud detection [11,12], and in information security [13,14]. Different outlier detection methods have been proposed over the years based on the nature of application.

Outlier detection algorithms first create a normal pattern in the data, and then assign an outlier score to a given data point on the basis of its deviation with respect to the normal pattern [15]. Extreme value analysis models, probabilistic models, linear models, proximity-based models [16], information-theoretic models and high dimensional outlier detection models represent some prominent categories of outlier detection techniques. Proximity-based methods treat outliers as points which are isolated from the remaining data and can be further classified into three different sub-categories, namely, cluster-based [17], density-based and nearest neighbor-based [15]. The main difference between the clustering and the density-based methods is that the clustering methods segment the points, whereas the density-based methods segment the space [18]. Local outlier factor (LOF) [19], connectivity-based outlier factor (COF) [20] and influenced outlierness (INFLO) [21] are examples of some well-known density-based approaches for outlier detection. In contrast, rank based detection algorithm (RBDA) [22] and outlier detection using modified-ranks with Distance (ODMRD), [23] are two recently published approaches which use ranks of nearest-neighbors for the detection of the outliers.

In most of the density-based approaches, it is assumed that the density around a normal data object is similar to the density around its neighbors, whereas in case of an outlier the density is considerably low than that of its neighbors. In LOF [19], the densities of the points have been calculated within some local reachable distances and the degree of outlierness of a point has been assigned in terms of relative density of the test point with respect to its neighbors [19]. Tang et al. argued that lower density is not a necessary condition to be an outlier. Accordingly, in COF [20], a set based nearest path was used to select a set of nearest neighbors [20]. This nearest path was further employed to find the relative density of a test point within the average chaining distance. COF [20] is shown to be more effective when a cluster and a neighboring outlier have similar neighborhood densities. Both LOF [19] and COF [20], which use properties of kNN, are found to yield poor results when an outlier lies in between a sparse and a dense cluster. To handle such situations, Jin et al. proposed a new algorithm INFLO based on a symmetric neighborhood relationship. In this method both forward and reverse neighbors of a data point are considered while estimating its density distribution [21]. In case of density-based approaches all the neighbors of a test point are

☆ This paper has been recommended for acceptance by G. Moser.

* Corresponding author at: Jadavpur University, Electronics and Telecommunication Engineering, 188 Raja S.C. Mallik Road, Jadavpur, Kolkata 700032, India. Tel.: +91 33 2457 2405; fax: +91 33 2414 6217.
E-mail address: aschowdhury@etce.jdvu.ac.in, ananda.chowdhury@gmail.com (A.S. Chowdhury).

http://dx.doi.org/10.1016/j.patrec.2015.04.004
0167-8655/

assumed to have a similar density. So, if a neighbor is chosen from different clusters with uneven densities the above assumption may introduce some errors in outlier detection. In addition, the notion of density may not work properly for some special types of distributions. For example, if all data points lie on a single straight line, the normal density-based algorithm [22] assumes equal density around the test point and its neighbors. This occurs due to the equal closest neighbor distance for both the test-point and its neighbor points. In such situations, rank-based outlier detection schemes like RBDA [22] and ODMRD [23] yield better results as compared to the density-based algorithms. RBDA uses mutual closeness between a test point and its k-neighbors for rank assignment. In ODMRD [23] the ranks were given some weights and the distances between the test point and its neighbors were incorporated. Still, both RBDA and ODMRD are found to be adversely affected by the local irregularities of a dataset like the cluster deficiency effect and the border effect.

In order to address the shortcomings of density-based and rank-based methods, we propose a novel hybrid outlier detection approach using the concepts of density as well as neighborhood rank-difference. The first contribution of our method is: instead of local reachable distance [19], we employ a dataset-specific global limit in terms of k (number of forward neighbors) to estimate the spatial density. The second contribution of our method is: we can better capture the variations in density by using reverse as well as forward rank difference over rank-based methods [22,23] by minimizing both the cluster deficiency effect and the border effect. Our third contribution is: we can minimize information loss due to averaging [19] through an effective density sampling procedure. Experimental results clearly indicate that we can capture most of the m outliers within top-m instances.

The rest of the paper is organized in the following manner: In Section 2, we provide the theoretical foundations. In Section 3, we describe the proposed method and also analyze its time-complexity. In Section 4, we present the experimental results with detailed comparisons. Finally, the paper is concluded in Section 5 with an outline for directions of future research.

2. Theoretical foundations

Let D denotes the data set of all observations, k denotes the number of points in the set of k nearest neighbors $N_k(p)$ around some point $p \in D$, and $d(p, q)$ denotes the Euclidean distance between any two points $p, q \in D$. We consider Euclidean distance for its simplicity. Further, let R represents the reverse ranking of the point p with respect to the point $q \in N_k(p)$. In the proposed method, we employ the difference of the Reverse Nearest Neighbor (RNN) and forward Nearest Neighbor (kNN) ranks of a point. If q is the k^{th} neighbor of the point p at distance $d_k(p, q)$, then the forward density up to k^{th} neighbor is given by:

$$\Omega_k(p) = k/d_k(p, q) \tag{1}$$

Similarly, if p be the R^{th} neighbor of q for the same distance $d_k(p, q)$ then the reverse density around q at same distance $d_k(p, q)$ is given by:

$$\Omega_R(q) = R/d_k(p, q) \tag{2}$$

The positive value of the rank difference $(R-k)$ indicates that q has a denser surrounding than that of p. By denser surrounding, we mean presence of more number of points within the hypersphere with radius $d_k(p, q)$ and center q. Similarly, a negative value of the rank difference indicates that p has a denser surrounding than that of q. For same values of R and k, p and q have equally dense surroundings. An illustration is shown in Fig. 1, where $k = 4$ and $R = 6$. So, their rank difference according to our definition is 2. In this case, as $(R-k)$ is positive, the point q has denser surrounding than the point p.

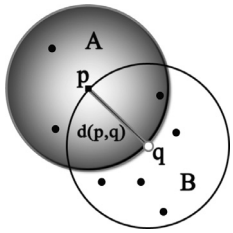

Fig. 1. Schematic diagram of the distribution of neighbors around a test point p and different regions of interest with respect to p and q.

3. Proposed method

Our proposed method consists of two steps. In the first step, we construct an influence space around a test point p. In the second step a rank difference based outlier score is assigned on the basis of this influence space.

3.1. Influence space construction

Influence space depicts a region with significantly high reverse density in the locality of a point under consideration. If the localities of the neighbors within the influence space [21,24] are more dense with respect to the locality of the concerned point, then a high value of outlierness score will be assigned to it. For an entire dataset, number of neighbors in the influence space is kept fixed.

In section 2, we have defined a reverse density Ω_R which captures the density surrounding the locality of the neighboring points of a particular point. As the distance is increased from the target point, more number of neighbors get included in its surroundings resulting in different values of Ω_R. With successive addition of neighboring points, a set of reverse densities is obtained for each point at varying depths (number of neighboring points). The average reverse density $\overline{\Omega}_R$ for each depth is determined next. Note that we have considered the depth and not the distance around the neighbors to handle situations where there is empty space (no neighboring point is present) surrounding a given point. To avoid random fluctuations, the variation in the average reverse density with respect to depth has been smoothed using a Gaussian kernel in the following manner:

$$\Omega_{\text{smoothed}} = \frac{1}{N h_{\text{optimal}}} \sum_{i=1}^{n} \frac{e^{-\frac{1}{2}\left(\frac{\overline{\Omega}_R - \overline{\Omega}_R^i}{h_{\text{optimal}}}\right)^2}}{\sqrt{2\pi}} \tag{3}$$

where,

$$h_{\text{optimal}} = \frac{0.9\sigma}{N^5} \tag{4}$$

and

$$\sigma = \frac{median(|\overline{\Omega}_R - median(\overline{\Omega}_R)|)}{0.6745} \tag{5}$$

where σ stands for an unbiased and consistent estimate of population standard deviation for large N [25,26].

In this smoothing process, an optimal width for the kernel h_{optimal} is determined using (4) [27] and (5) for better estimation of the significant density fluctuation around the neighbor points. We deem the first most significant peak [28,29] in this smoothed-kernel probability density function [25] as the limit of the influence space. The peak has been determined using the undecimated wavelet transform with Daubechies coefficients [30]. Such wavelet transforms can obtain peaks with maximum confidence by eliminating any surrounding noisy spurious peaks.