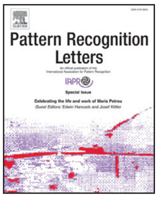
Contents lists available at ScienceDirect

Pattern Recognition Letters

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



Normalized residual-based constant false-alarm rate outlier detection[☆]

Xiaohu Ru[*], Zheng Liu, Zhitao Huang, Wenli Jiang

College of Electronic Science and Engineering, National University of Defense Technology, Changsha, Hunan 410073, P.R. China

ARTICLE INFO

Article history:
Received 10 June 2015
Available online 6 November 2015

Keywords:
Outlier detection
Supervised
Constant false-alarm rate (CFAR)
Normalized residual

ABSTRACT

Outlier detection is an important issue in machine learning and knowledge discovery. The aim is to find the patterns that deviate too much from others. In this paper, we consider constant false-alarm rate (CFAR) outlier detection, and propose a supervised detection method based on normalized residual (NR). For a query point, its NR value related to the training data is compared with a predefined threshold, indicating if it is an outlier. Heretofore, the choice of outlier threshold relied too much on experience, making CFAR detection impossible. We solve the problem by introducing a sufficiently training strategy applying to the given normal instances, gaining a large number of NR values of them, based on which the threshold can be located properly according to the desired false-alarm rate. Theoretical analysis proves that the proposed method can achieve CFAR detection and the most powerful test, regardless of pattern dimension and noise distribution, thus can be widely applied to outlier detection problems. Simulations and real-world data experiments also show that, the proposed method can effectively control the false-alarm rate even when a few training instances are available, and at the same time its operating characteristic is generally better than competing methods.

1. Introduction

Outlier is defined as an observation which deviates so much from others that it arouses suspicions that it was generated by a different mechanism [14]. Outliers may represent important events, or they may result in severe consequences, thus outlier detection problems arise in many areas such as machine learning, knowledge discovery and data mining. Typical applications include detecting abnormal regions in image processing, intrusion detection in computer networking, specific emitter verification (SEV) in signal processing [22], robust classification for mixed labeled/unlabeled data sets [24], anomaly detection for particle image velocimetry (PIV) data [8,30], and so on.

Constant false-alarm rate (CFAR) detection of signals is an important and classical problem in statistical signal processing [18,20]. In this paper, we focus on CFAR outlier detection, meaning that the probability that normal instances are wrongly classified as outliers should be controlled no higher than some specified significance level. If the false-alarm rate is too high, excessive normal instances will be seen as outliers. These false outliers may be immediately dropped, or they have to be further dealt with in case important events arise, which will increase the burden of the processing system. On the contrary, if the false-alarm rate is too low, some outliers may be unexpectedly identified as normal ones, resulting in false dismissal and influence on the processing of normal instances. Therefore, a desired value of the false-alarm rate should be set according to practical requirements. However, previous outlier detection methods either cannot directly control the false-alarm rate, or the effectiveness of CFAR detection is not as good as expected when a few training instances are given.

Some outlier detection techniques are unsupervised, but others rely on training data for the benefit of a high precision rate. The training data may come from normal or/and anomalous instances. In this study, we consider the case that only normal ones are provided for training. This is corresponding to wide applications in which the normal states can be more easily recorded in common situations, or the classes of interest have been collected in the dataset. However, the styles or distributions of incoming outliers are unknown, or anomalous instances have not been obtained yet. In this way, some normal instances can be chosen as the training data and then used for outlier detection.

In this paper, we choose normalized residual (NR) to indicate outliers and realize CFAR detection. The NR value of a query point $\boldsymbol{\eta}$ refers to the distance between $\boldsymbol{\eta}$ and its nearest neighbors, normalized by the median distance of the latter. The NR value of an outlier is usually large, and that of a normal instance is small. To identify outliers, a detection threshold h should be determined first, and then those instances whose NR values exceed h are claimed as outliers. To achieve CFAR detection, the key problem is how to choose the detection

[☆] This paper has been recommended for acceptance by Dr. D. Dembele

[*] Corresponding author. Tel.: +86 0731 84573490.
E-mail address: 870142339@qq.com, ruxiaohu88@163.com (X. Ru).

http://dx.doi.org/10.1016/j.patrec.2015.10.002
0167-8655/© 2015 Elsevier B.V. All rights reserved.

threshold according to the desired false-alarm rate. To solve this, we introduce a sufficiently training strategy to gain an effective understanding of the NR values of normal instances, based on which a proper threshold can then be chosen.

In detail, the training data are first randomly divided into two equal parts. Instances in one part are used to calculate the NR value of each point in another part, and vice versa. Then resplit the data and repeat the above procedure many times. The divisions of the training data should be different from each other in order to obtain diverse and numerous NR results, which can be used to represent the probability distribution of the NR values of normal instances. The detection threshold h can then be chosen accordingly considering the desired false-alarm rate. At testing stage, the NR value r of the query point $\boldsymbol{\eta}$ is calculated similarly, meaning that half of the training instances are randomly selected for the purpose of calculating r. $\boldsymbol{\eta}$ will be claimed to be an outlier if its NR value exceeds the predefined threshold h; otherwise it is claimed to be normal.

We prove through theoretical analysis that the proposed threshold definition and outlier detection method can achieve CFAR detection and simultaneously the most powerful test, regardless of pattern dimension and noise distribution, thus it has wide applications. We also investigate our method through simulations and real-world data experiments, which show that the proposed method can make the false-alarm rate be much closer to the desired value even when a few training instances are given, and meanwhile, the outlier detection probability is comparable with or even better than competing methods.

The remainder of this paper is organized as follows. Section 2 introduces related work. Section 3 gives the definition of NR for the supervised case. Section 4 describes in detail how to choose the detection threshold according to the desired false-alarm rate, and gives the NR-based outlier detection method. Section 5 discusses and proves the superior performance of the proposed method in aspects of false-alarm rate and outlier detection probability. The theoretical analysis results are validated through simulations and real-world data experiments in Section 6. At last, we conclude this work.

2. Related work

A great many techniques have been proposed for outlier detection. Considering the kinds of detection approaches, outlier detection techniques can be mainly divided into distribution-based [2], depth-based [12], distance-based [1,6,25–27,34], density-based [3] and clustering-based [9,15–17,28] techniques. Considering the nature of algorithms, there exist supervised, semi-supervised and unsupervised detection algorithms.

Recently proposed semi-supervised outlier detection methods [10,11,31,33] require both normal instances and outliers as training data. In [7], an entropy-based technique is proposed to solve the problem of detecting outliers when only a few outlier samples are available. In this paper, we consider the case that training data are all normal instances. Important methods concerning this case include one-class support vector machine (SVM) [9,17,28] and localized *p*-value estimation (LPE) method [6,26,34]. Just as existing semi-supervised and unsupervised methods, one-class SVM cannot directly control the false-alarm rate [5,26], but LPE method is believed to have the ability in this aspect. Moreover, LPE method can achieve the uniformly most powerful (UMP) test when outliers are drawn from a mixture of the normal density and uniform distribution. Although the mixed density hypothesis can be satisfied in general, LPE method may still fail to maintain the false-alarm rate when the training instances are not enough. NR has been used to detect outliers for unsupervised situations in [8,27,30]. However, in these works, the choice of h relies too much on experience or experiments, lacking theoretical guidance. And the problem of how to choose h to make the false-alarm rate controllable has still not been solved.

3. Definition of NR

Let $\boldsymbol{X}_{tr} = \{\boldsymbol{x}_i, i = 1, 2, \ldots, n\}$ be the training set which includes n normal instances. The given instances $\boldsymbol{x}_i$, which are N-dimensional vectors, are assumed to be independent and identically distributed (i.i.d.), and drawn from some underlying density f_0. For a query point $\boldsymbol{\eta}$, outlier detection problem is to determine whether $\boldsymbol{\eta}$ belongs to the same class with $\boldsymbol{x}_i$, that is to say if $\boldsymbol{\eta}$ is consistent with f_0.

For the query point $\boldsymbol{\eta}$, conventional definition of NR relies on the K nearest neighbors of $\boldsymbol{\eta}$ among all the instances. When outliers appear in the neighbors, the NR value of $\boldsymbol{\eta}$ will be affected and turns to be small, making it unfavorable for outlier detection. In this paper, the Kneighbors are selected from the training set, denoted by $\boldsymbol{x}_{tr}^k$, $k = 1, 2, \ldots, K$. Then the NR value of $\boldsymbol{\eta}$ can be defined as

$$r(\boldsymbol{\eta}) = \frac{\left\| \boldsymbol{\eta} - med\left(\alpha_{tr}^k \boldsymbol{x}_{tr}^k\right) \right\|}{med\left(||\boldsymbol{x}_{tr}^k - med(\boldsymbol{x}_{tr}^k)||\right) + \xi} \tag{1}$$

where $||\cdot||$ denotes 2-norm, $med(\cdot)$ denotes calculating the median value concerning the superscripts $k = 1, 2, \ldots, K$, $\alpha_{tr}^k = (sum(d_{tr}^k) \cdot d_{tr}^k)^{-1}$ is the weight of the training instance $\boldsymbol{x}_{tr}^k$, $d_{tr}^k = ||\boldsymbol{\eta} - \boldsymbol{x}_{tr}^k||$ is the distance between $\boldsymbol{\eta}$ and $\boldsymbol{x}_{tr}^k$, $sum(\cdot)$ means summation concerning k and ξ is the tolerance.

The weights α_{tr}^k aim at reducing the influence of those instances which are far away from the query point. In this way, local outliers are more likely to be detected. The tolerance ξ is used to ensure the effective calculation of NR when instance overlapping exists. In this paper, we set ξ to be $\gamma \cdot r_{num}$ if $r_{num} \neq 0$, and to be 1 otherwise, where r_{num} is the numerator of the right hand side of (1), and γ is a small number, e.g. 0.001. As a result, if $\boldsymbol{x}_{tr}^k$, $k = 1, 2, \ldots, K$ are entirely overlapped, but $\boldsymbol{\eta}$ deviates from them, a sufficiently large $r(\boldsymbol{\eta})$ can be obtained; if $\boldsymbol{x}_{tr}^k$ and $\boldsymbol{\eta}$ are all overlapped, the definition in (1) can ensure $r(\boldsymbol{\eta})$ to be 0; otherwise, ξ almost has no influence on the calculated result of $r(\boldsymbol{\eta})$.

If the query point $\boldsymbol{\eta}$ is an outlier, it will deviate from $\boldsymbol{x}_{tr}^k$, resulting in a big NR value; otherwise $r(\boldsymbol{\eta})$ will be relatively small. Then a detection threshold h can be set to determine whether $\boldsymbol{\eta}$ is an outlier or not. Heretofore, the choice of h relied heavily on experience or experiments, lacking theoretical basis, and at the same time making the false-alarm rate uncontrollable.

For clarity in the sequel, we use $r(\boldsymbol{\eta})$ to denote the NR value of the query point $\boldsymbol{\eta}$, r_i which has subscript i to represent the *ith* NR value calculating from a set of instances, and a single r to represent the general variable concerning NR.

4. NR-based outlier detection method

The proposed NR-based outlier detection method first trains the given normal instances to choose a suitable detection threshold, and then calculates the NR value of the query point to determine if it is an outlier. The procedure is shown as follows.

Given the training set $\boldsymbol{X}_{tr} = \{\boldsymbol{x}_i, i = 1, 2, \ldots, n\}$ and the query point $\boldsymbol{\eta}$, set the number of nearest neighbors to be K, the number of random divisions of $\boldsymbol{X}_{tr}$ to be B, and the desired false-alarm rate to be P_f.

4.1. Training stage: choosing the detection threshold

(1) Randomly divide $\boldsymbol{X}_{tr}$ into two parts: S_1 and S_2, where S_1 contains $\lfloor n/2 \rfloor$ training instances, $\lfloor\cdot\rfloor$ denotes the nearest integer no higher than $n/2$, and S_2 contains the other ones. Set K' to be the smaller value between K and $\lfloor n/2 \rfloor$.

(2) For each sample $\boldsymbol{x} \in S_1$, find its K' nearest neighbors in S_2 and then calculate its NR value according to (1). Similarly, for samples in S_2, calculate their NR values based on S_1.

(3) Repeat the above steps B times. Make sure that the divisions of $\boldsymbol{X}_{tr}$ are different from each other in order to obtain diverse