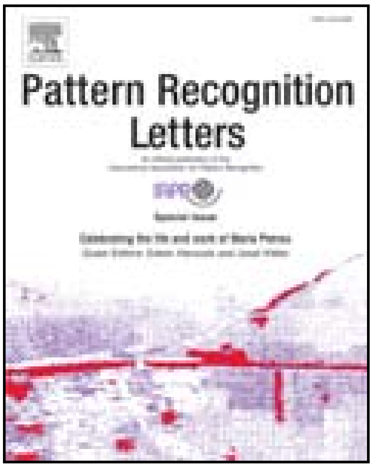
Journal Pre-proof




Robust Geodesic based Outlier Detection for Class Imbalance Problem

Canghong Shi, Xiaojie Li, Jiancheng Lv, Jing Yin, Imran Mumtaz

PII: S0167-8655(20)30044-1
DOI: https://doi.org/10.1016/j.patrec.2020.01.028
Reference: PATREC 7783

To appear in: *Pattern Recognition Letters*

Received date: 13 January 2020
Accepted date: 31 January 2020

*Pattern Recognition Letters*

# Authorship Confirmation

**Please save a copy of this file, complete and upload as the "Confirmation of Authorship" file.**

As corresponding author I, Xiaojie Li, hereby confirm on behalf of all authors that:

1. This manuscript, or a large part of it, has not been published, was not, and is not being submitted to any other journal.
2. If presented at or submitted to or published at a conference(s), the conference(s) is (are) identified and substantial justification for re-publication is presented below. A copy of conference paper(s) is(are) uploaded with the manuscript.
3. If the manuscript appears as a preprint anywhere on the web, e.g. arXiv, etc., it is identified below. The preprint should include a statement that the paper is under consideration at Pattern Recognition Letters.
4. All text and graphics, except for those marked with sources, are original works of the authors, and all necessary permissions for publication were secured prior to submission of the manuscript.
5. All authors each made a significant contribution to the research reported and have read and approved the submitted manuscript.

Signature Xiaojie Li Date Dec. 12th, 2019

---

**List any pre-prints:**

---

**Relevant Conference publication(s) (submitted, accepted, or published):**

**Justification for re-publication:**

**Research Highlights (Required)**

To create your highlights, please type the highlights against each `\item` command.

It should be short collection of bullet points that convey the core findings of the article. It should include 3 to 5 bullet points (maximum 85 characters, including spaces, per bullet point.)

- We present an interesting heuristic for unsupervised outliers detection facing an imbalanced class problem.
- We construct global disconnectivity score and local real degree to effectively consider the characteristics of points.
- We prove outlierness rises as the distance to cluster center, and reduces with higher local degree if cluster is more dense.

Pattern Recognition Letters
journal homepage: www.elsevier.com

# Robust Geodesic based Outlier Detection for Class Imbalance Problem

Canghong Shi[a], Xiaojie Li[b,**], Jiancheng Lv[c], Jing Yin[d], Imran Mumtaz[e]

[a]School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756, China
[b]College of Computer Science, Chengdu University of Information Technology, Chengdu 610103, China
[c]College of Computer Science, Sichuan University, Chengdu 610065, China
[d]Chongqing University of Technology, Chongqing, China
[e]University of Agriculture, Faisalabad 38000, Pakistan

ABSTRACT

Outlier detection is very useful in many applications, such as fraud detection and network intrusion detection. However, some existing methods often generate incorrect identification results due to the imbalanced distribution of data points. In this paper, we present a robust geodesic-based outlier detection algorithm which simultaneously considers both global disconnectivity score and local real degree as measures of outlierness. We first construct the global disconnectivity score to incorporate suitable global characteristics of data, then we provide the local real degree to effectively consider the local characteristics of points. Thus, we can identify local outliers with higher overall connectivity but in a smaller cluster with fewer points. Experimental results obtained for a number of synthetic and real-world data sets demonstrate the effectiveness and robustness of our method. In particular, we estimate an increase in average area under curve (AUC) on ten datasets of approximately 15%, with smaller RMSD than any of the competing methods.

*Keywords:* Outlier detection; structural stability; local structure




## 1. Introduction

Outlier detection is an important task in identifying abnormal points deviating from normal patterns, for example in credit card fraud detection. It can also ameliorate the deterioration of recognition performance due to outliers. Unlike the tasks of clustering, classification, and pattern analysis, which aim to find general patterns, outlier detection and boundary-point detection identify critical patterns that do not conform to the expected normal patterns Li et al. (2016); Zhai et al. (2016). For example, a liver disorder detection system might consider healthy patients as normal observations, patients with liver disorders as outlier observations, and at-risk patients as boundary points. Such a system would help in the study of the disease, and the set of people corresponding to outliers and boundary points would warrant special attention. This in an example for which detecting outliers becomes more critical than detecting the normal pattern.

Different outlier detection strategies have been proposed, but no consensus has been reached even on the definition of outlier Hawkins (1980); Aggarwal and Yu (2001). Distance-based technique is one popular approach, using the nearest-neighbor Euclidean distances between a given point and the other points. However, a single dissimilarity measure may not capture all possible anomalous patterns in many application domains Hsiao et al. (2016). The classical definition of an boundary point was proposed by Li and Maguire Li and Maguire (2011) and states that boundary points sit on the extremes of a class region, near free pattern space. The proposed border-edge pattern selection method plays an important role in identifying boundary points but requires a training data set. Although outliers and boundary points are different by definition, they are generally located around the margin of the high-density regions of the data set Li et al. (2016). Henceforth we can conflate the two.

Depending on whether point labels are available, detection methods can be classified as supervised, semi-supervised, or

**Corresponding author: Tel.: +86-28-85966901; fax: +86-28-85966901;
*e-mail:* lixj@cuit.edu.cn (Xiaojie Li)

unsupervised methods Xia et al. (2006); Ding et al. (2015); Campos et al. (2016). Due to the lack of label information, the unsupervised methods are extremely challenging. Moreover, an unbalanced distribution of data points (i.e., class imbalance problem) is predominant in detection scenarios, which can lead the detection model to generate largely incorrect identification results. A traditional approach to solve the problem is the local outlier model, whereby one calculates a local outlier factor to evaluate to what degree the data point is an outlier (e.g. LOF) Breunig et al. (2000). However, Zhai et al. (2016) states that the statistical power and accuracy of the detection methods depend on the model characterizing the data distribution. Unfortunately, conventional detection methods have limited capacity when only considering local similarity.

In this paper, we provide a robust geodesic-based unsupervised outlier detection algorithm by simultaneously calculating both global disconnectivity score and local real degree as measures of outlierness. Our method is based on the idea that boundary points and outliers are characterized by a lower local connectivity and large geodesic distance to their neighbors. To incorporate suitable global distributions of data, we first propose a global disconnectivity score, explicitly taking into account the global data structure, then the local real degree is provided to effectively consider the local characteristics of points. This can identify local outliers with higher connectivity but in a smaller class with fewer points. Two contributions are claimed in this paper. We propose an unsupervised geodesic-based detection method, which can effectively solve the imbalance distribution problem and obtain robust prediction. Our method better reflects the shape of the data. We show that even more points are identified as outliers but located near the margin of the clusters, which provide a flexible solution for identifying the number of outliers or boundary points. Experimental results obtained for a number of synthetic and real-world data sets demonstrate the effectiveness and robustness of the proposed method.

The remainder of this paper is organized as follows. Section 2 presents the preliminaries and motivation. Section 3 introduces our novel boundary detection method. In Section 4, experimental results on a number of synthetic and real data sets demonstrate the effectiveness of our method. Finally, conclusions are presented in Section 5.

## 2. Preliminaries & Motivation

### *2.1. Mathematical Formulation*

The data comprises a metric space $(\mathbf{R}^m, d)$ and data set $\mathbf{X} = \{\mathbf{x_1}, \mathbf{x_2}, \cdots, \mathbf{x_n}\}$ with $n$ samples and $m$ features. Although outliers and boundary points are different by definition, they are generally located around the margin of the data set with high density. More precisely, we state Assumption 1:

**Assumption 1.** *Suppose the set* $\mathbf{I}$ *includes inner points and* $\mathbf{B}$ *includes outliers and boundary points. It holds that* $\mathbf{I} \subseteq \mathbf{X}$, $\mathbf{B} \subseteq \mathbf{X}$ *and* $\mathbf{X} = \{\mathbf{I} \cup \mathbf{B}\}$. *In general, the local density of* $\mathbf{x}_i \in \mathbf{B}$ *is less than that of* $\mathbf{x}_j \in \mathbf{I}$.

This assumption is not as restrictive as it might first seem, because the local density of boundary points should be lower than that of inner points intuitively Li et al. (2016). Otherwise one cannot judge whether they are anomalous or not from the data distribution. Given data set $\mathbf{X}$, the detection schemes aim to find the set $\mathbf{B}$.

### *2.2. Distance-based methods*

Distance-based techniques rely implicitly or explicitly on the distance of each point from its neighbors, and several variants have been proposed Knorr and Ng (1998, 1999); Angiulli and Pizzuti (2002); Hautamaki et al. (2004). Intuitively, points with distances significantly larger than others are more likely to be outliers. Given a k-nearest neighbor (kNN) directed graph of the data, let $N_k(\mathbf{x}_i) = \{r_{i1}, \cdots, r_{ik}\}$ be the set of distances between $\mathbf{x}_i$ and its $k$ nearest neighbors. Suppose $r_{ij} \leq r_{ij+1}$, and denote $Ideg(\mathbf{x}_i)$ as the in-degree of $\mathbf{x}_i$, with $T$ a threshold. For $\mathbf{x}_i$, $Ideg(\mathbf{x}_i)$ calculates the number of head ends adjacent to $\mathbf{x}_i$. In Hautamaki et al. (2004), if $Ideg(\mathbf{x}_i)$ exits

$$Ideg(\mathbf{x}_i) \leq T,$$

then one marks $\mathbf{x}_i$ as an outlier. Meanwhile, two different variants, the mean kNN distance and the maximum kNN distance, have also been proposed by Hautamaki et al. (2004). Formally, denote $mn(x_i) = mean\{r_{i1}, \cdots, r_{ik}\}$ and $ma(x_i) = max\{r_{i1}, \cdots, r_{ik}\}$. For simplicity, we sort all values and suppose $mn(x_{\hat{1}}) \geq \cdots \geq mn(x_{\hat{n}})$ and $ma(x_{\hat{1}}) \geq \cdots \geq ma(x_{\hat{n}})$. If

$$mn(x_{\hat{i}}) - mn(x_{\hat{i+1}}) > T,$$

where $T = max\{mn(x_{\hat{i}}) - mn(x_{\hat{i+1}})\} * t, t \in [0, 1]$, then one marks $\mathbf{x}_i$ as outlier. Similarly, if

$$ma(x_{\hat{i}}) - ma(x_{\hat{i+1}}) > T,$$

then $\mathbf{x}_i$ can be identified as outlier Hautamaki et al. (2004). These methods are based on local structures and depend on the parameters $k$ and $T$.

In Knorr and Ng (1998), the classical seminal datasets-oriented paper studying the Distance-Based method, one determines local outliers DB$((\epsilon, R))$, where $\epsilon$ is the data fraction and $R$ is a radius. Point $\mathbf{x}_i$ in a dataset $\mathbf{X}$ is a DB$(\epsilon, R)$-outlier if at least a fraction $\epsilon$ of the points in $\mathbf{X}$ lie greater than distance $R$ from $\mathbf{x}_i$. Formally, this conditions holds if

$$\#\{\mathbf{x}_j \in \mathbf{X} | r_{ij} > R\} \geq \epsilon n.$$

For each point, Angiulli et al. Angiulli and Pizzuti (2002) considered the sum of distances from its k nearest neighbors, which can be found by linearizing the search space through the Hilbert space filling curve. However, the space filling curve can aggravate identification problems Kriegel et al. (2008), whereas a single distance-based measure may not capture all possible anomalous patterns in many application domains Hsiao et al. (2016). Moreover, the threshold problem in such cases remains a challenge.

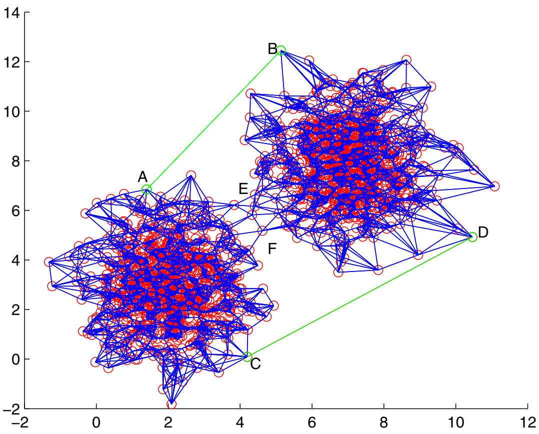



**Fig. 1. Geodesics and Euclidean distance. Green lines denote the Euclidean distance between points A and B (C and D). While blue lines implies geodesics.**

### 2.3. Data Structure and Geodesic

Over the past decades, numerous methods that learn a data representation and perform better than traditional methods in clustering and embedding have been established Elhamifar (2011); Li et al. (2013). Thus, structure preservation may be useful in detecting outliers and boundary points. The effect of global and local geometric structure needs to be studied in different detection modes.

Most representation-based methods rely implicitly or explicitly on global or local geometric data structure. Geodesic structure has the aforementioned benefits Tenenbaum et al. (2000). It focuses on globally-preserved pairwise sample similarity but does not ignore the local geometric structure of data. *Geodesics are locally shortest paths* Kimmel et al. (1995). More precisely, Figure 1 illustrates the difference between Euclidean distance and geodesics. It shows intuitively that boundary points (or outliers) have greater distances to their neighbors than other points.

### 2.4. Problem

Suppose the geodesic distance matrix is $\mathbf{D} \in R^{n\times n}$ Tenenbaum et al. (2000), and let $d_{ij}$ denote the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$. Let

$$\eta_i = \sum_{j=1}^{n} d_{ij}. \tag{1}$$

From Assumption 1, Remark 1 intuitively holds without the class imbalance problem.

**Remark 1.** *Assumption 1 indicates that the quantity $\eta_i$ increases as the distance of a point from the center increases. It holds that the values of $\eta_i$ of boundary points or outliers are much larger than that of points with higher density.*

Thus, outliers or boundary points are generally recognized as points for which the value of $\eta_i$ is anomalously large. It is difficult to conduct thorough quantitative research in theory alone. To illustrate this, consider the example shown in Figure 2. We bin the elements of $\{\eta_1, \eta_2, \cdots, \eta_n\}$ into 10 equally spaced containers $\{\theta_1, \theta_2, \cdots, \theta_{10}\}$, and plot the number of points in each

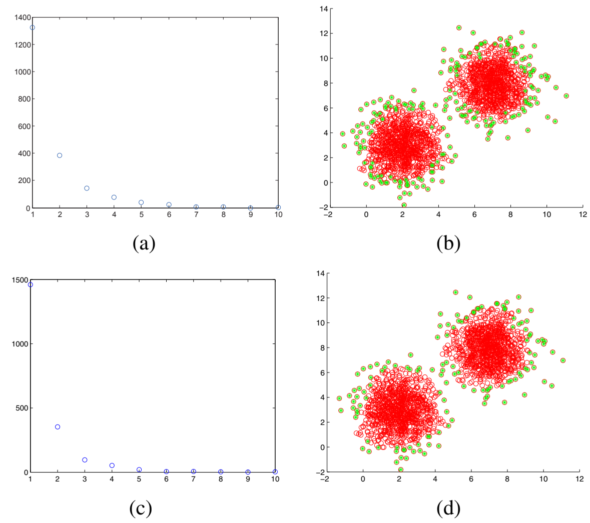



**Fig. 2. Synthetic data set consisting of two clusters. Figure 2(a) The elements $\eta_i$, divided into 10 equally spaced containers, showing the number of elements in each container. Figure 2(b) The data set, with green points lying within the $\{\theta_3, \theta_4, \cdots, \theta_{10}\}$ containers, with larger $\eta_i$ values. Figure 2(c) The elements of $\bar{\eta}_i$, divided into 10 equally spaced containers, showing the number of elements in each $\bar{\theta}_i$. Figure 2(d) The data set, with green points lying within the $\{\theta_3, \theta_4, \cdots, \theta_{10}\}$ containers, with larger $\bar{\eta}_i$ values.**

$\theta_i$, as in Figure 2(a). We find that the maximum number of elements are in the $\theta_1$ and $\theta_2$ containers, which we can identify as normal points or inner points. Following Assumption 1, only a few points are identified as boundary points and outliers. Figure 2(b) shows green stars corresponding to points within the $\{\theta_3, \theta_4, \cdots, \theta_{10}\}$ containers. We find that green stars with larger $\eta_i$ values are located near the margin of densely distributed data. This shows that Remark 1 is valid. However, it does not work well the class imbalance scenario.

In the following, we discuss how, and under what conditions, $\eta_i$ and the decision histogram can be applied to detecting boundary points and outliers, and how to deal with the class imbalance problem.

## 3. Detection method based on global disconnectivity and local real degree

Since geodesics capture sufficient information about the geometry of data, we present a geodesic-based method for detecting such points by calculating their global disconnectivity score and local real degree as measures of outlierness (GDLD). The method is based on the following definitions.

In graph theory, the degree of a vertex of a graph is the number of edges incident to the vertex Diestel (2005). Formally,

$$deg(\mathbf{x}_i) = \sum_{j=1}^{n} \chi(d_{ij}), j = \{1, 2, \cdots, n\} \tag{2}$$

where $\chi(x) = 1$ if $x \neq 0$ and $\chi(x) = 0$ otherwise. Clearly, $deg(\mathbf{x}_i)$ is equal to the number of points that directly connect to point $\mathbf{x}_i$. The following definition of outlier is proposed. Given geodesics graph $\mathbf{G}$ for dataset $\mathbf{X}$, an outlier or boundary point is a vertex, whose sum distance $\eta_i$ is bigger or whose degree $deg(\mathbf{x}_i)$ is less than that of all points in **I** generally.

*3.1. The global disconnectivity score*

As discussed in Section 2.4, $\eta_i$, which explicitly takes into account the global data structure, can identify boundary points and outliers. Its values are used to evaluate to what degree an observation is an outlier in some cases without the class imbalance problem. One geodesic may be equal to the sum of several geodesics of a point. Formally,

$$\eta_i = \eta_j, \qquad deg(\mathbf{x}_i) \neq deg(\mathbf{x}_j)$$

Therefore, $\eta_i$ may be inappropriate to evaluate what degree this observation is an outlier. A mean disconnectivity alternative method will be proposed. For simplicity, we describe the method in 2-D space. In the experimental section, we will illustrate its applicability to high-dimensional data. For each point $\mathbf{x}_i, i \in \{1, 2, \cdots, n\}$, the mean value, weighted by degree, is adopted. Formally,

$$\bar{\eta}_i = \frac{\eta_i}{deg(\mathbf{x}_i)} \tag{3}$$

Note that $\bar{\eta}_i$ first explicitly takes into account the global data structure and simultaneously considers its local structure. It is more interesting and obviously preferable. $\bar{\eta}_i$ can deal with the following cases without an extreme class imbalance problem:

$$\left.\begin{array}{l} \eta_i = \eta_j,\ deg(\mathbf{x}_i) \neq deg(\mathbf{x}_j) \\ \eta_i \neq \eta_j,\ deg(\mathbf{x}_i) = deg(\mathbf{x}_j) \end{array}\right\} \stackrel{by}{\Rightarrow} \bar{\eta}_i \neq \bar{\eta}_j. \tag{4}$$

$\bar{\eta}_i$ reveals the degree of disconnectivity between a data point and other points. The larger the value of $\bar{\eta}_i$, the greater the disconnectivity. More generally, $\bar{\eta}_i$ value of outliers is larger than that of boundary points. Similarly, we bin the elements of $\{\bar{\eta}_1, \bar{\eta}_2, \cdots, \bar{\eta}_n\}$ into 10 equally spaced containers $\{\bar{\theta}_1, \bar{\theta}_2, \cdots, \bar{\theta}_{10}\}$, and plot the number of points in each $\bar{\theta}_i$ (see Figure 2(c)). Figure 2(d) shows green points corresponding to the number of elements in the $\{\bar{\theta}_3, \bar{\theta}_4, \cdots, \bar{\theta}_{10}\}$ containers. Compared with Figure 2(b), Figure 2(d) better reflects the characteristics of data. We show that Figure 2(d) illustrated less points but located around the margin than point Figure 2(b). Therefore, Eq. (3) is obviously preferable.

**Comment 1.** *You'll need to elaborate on the differences in the figures, because they are not apparent just by looking.*

*3.2. Real degree for cluster imbalance*

From Fig.1, it is easy to see that point B is more of an outlier than point A, while $deg(A) = deg(B) = 7$ with $\eta_A \neq \eta_B$. Suppose A and B share the same conditions. Despite their same number of routes, nearby residents would prefer to go to A, rather than B. Eq.(3) explicitly takes into account the global data structure and reveals the degree of connectivity between a data point and other points.

However, Eq.(3) is lacking in some cases. We discover that extreme cluster imbalance is the central cause. This cluster imbalance causes two problems.

1. A group of points in a small cluster can be misclassified as outliers because of their small $\bar{\eta}_i$ values.
2. Some points with lower $\bar{\eta}_i$ but higher $\eta_i$, i.e. points between clusters, can be misclassified as interior points.

A common solution is to calculate local effective measures. We will show that our proposed local real degree naturally handles the cluster imbalance caused by the distance method and allows us to efficiently identify outliers and boundary points. Formally, define

$$\mathcal{R}deg(\mathbf{x}_i) = \#\{d_{i1} < \tau, d_{i2} < \tau, \cdots, d_{in} < \tau\}, \tag{5}$$

where $\tau$ is a cutoff constant determined according to Rodriguez and Laio (2014), and $\#\{\cdot\}$ computes the number of elements with true value $d_{ij} < \tau$. If $d_{ij} < \tau$ for each $j = 1, 2, \cdots, n$, then $\mathbf{x}_i$ can be called an remote point. More generally, the smaller the value of $\mathcal{R}deg(\mathbf{x}_i)$, the more likely the point is a outlier.

In contrast, there exist some boundary points that nonetheless have higher $\mathcal{R}deg(\mathbf{x}_i)$ values than their neighbors, such as points located between two classes (see Figure 1 points $\{E, F\}$). Using $\bar{\eta}_i$, the problems can be solved generally. The new solution is

$$\eta_i = \eta_j,\ deg(\mathbf{x}_i) = deg(\mathbf{x}_j) \stackrel{by}{\Rightarrow} \mathcal{R}deg(\cdot)$$

$$\eta_A \neq \eta_B,\ deg(A) \neq deg(B) \stackrel{by}{\Rightarrow} \bar{\eta} \cup \mathcal{R}deg(\cdot)$$

Denote $\mathfrak{R} = \{\mathcal{R}deg(\mathbf{x}_1), \mathcal{R}deg(\mathbf{x}_2) \cdots, \mathcal{R}deg(\mathbf{x}_n)\}$ and $\mathcal{H} = \{\bar{\eta}_1, \bar{\eta}_2, \cdots, \bar{\eta}_n\}$. Let the means of $\mathfrak{R}$ and $\mathcal{H}$ be $\bar{\mathfrak{R}}$ and $\bar{\mathcal{H}}$, respectively. Suppose $t$ is a constant and $t < 1$. The following definition of an outlier is then proposed. *Given geodesics graph* **G** *for dataset* **X**, *an outlier (boundary point) is a vertex with a small local real degree $\mathcal{R}deg(\mathbf{x}_i)$ or bigger $\bar{\eta}_i$. The smaller the $\mathcal{R}deg(\mathbf{x}_i)$ or the bigger $\bar{\eta}_i$, the more likely it is that the point is an outlier.*

To determine how many observations are detected as outliers, the following definition is proposed. More precisely, we state Definition 1.

**Definition 1.** *Given data set* **X**, *the set* $\mathbf{\Omega}$ *consists of points with local real degrees $\mathcal{R}deg(\mathbf{x}_i)$ less than $t\bar{\mathfrak{R}}$ and the set* $\mathbf{\Psi}$ *consists of points where $\bar{\eta}_i$ is larger than $\frac{\bar{\mathcal{H}}}{t}$. If $\mathbf{x}_i \in \mathbf{\Omega} \cup \mathbf{\Psi}$, the point $\mathbf{x}_i$ is an outlier.*

From Assumption 1 and Definition 1, we state Remark 2.

**Remark 2.** *The bigger the t, the more likely it is that more points are identified as outliers (boundary points) but located near the margins of the clusters.*

It is difficult to conduct thorough quantitative research in theory alone. To illustrate this, consider the example shown in Fig.2(d). Based on simple geometric intuition, Remark 2 has a natural interpretation. Outlierness increases with the distance from the cluster mean, while outlierness decreases if the cluster is more dense with higher $\mathcal{R}deg(\mathbf{x}_i)$ values.

### *3.3. Robustness*

To analyze a data structure, a proper choice of the neighborhood size $k$ is important. Due to Dijkstra's algorithm, employed to achieve a shortest-path graph search, our method is robust with respect to the choice of $k$. To this end we compute $\mathcal{R}deg(\mathbf{x}_i)$, $\bar{\eta}_i$ and area under curve (AUC) for different values of $k \in [5, 5log(n)]$ and then average the resulting AUCs to produce a final measure of outlierness. We show that the global shortest-path structure is considered, and that the value of k is not critical to our method.

### *3.4. The GDLD algorithm*

The GDLD method is summarized in Algorithm 1. The computational time complexity is $O(N^2)$ which is characterized by the following Remark 3. We now present practical experiments to illustrate the effectiveness and efficiency of our method.

**Algorithm 1** *GDLD* Algorithm

**Input:** $\mathbf{X} \in R^{m\times n}$, $\tau$, and $k$
**Output:** boundary points and outliers
  Construct geodesics distance **D** by Dijkstra algorithm or others
  **for** each $i \in [1, n]$ **do**
    Calculate the global mean distance $\bar{\eta}_i$ by Eq.(3)
    Calculate the local real degree $\mathcal{R}deg(\mathbf{x}_i)$ by Eq.(5)
  **end for**
  **for** each $i \in [1, n]$ **do**
    Identify outlier or boundary points by Definition 1
  **end for**

**Remark 3.** *Given* $\mathbf{X} \in R^{m\times n}$ *and positive integer k, Algorithm 1 calculates the* $\bar{\eta}_i$ *and* $\mathcal{R}deg(\mathbf{x}_i)$, $i = 1, 2, \cdots, n$ *in time* $O(n(mlog(k)log(n) + n(k + log(n)) + (k + 1)))$ *and* $O(n(m + k + 2))$ *space.*

## 4. Experiments

In this section, we evaluate the proposed algorithm on both synthetic and real-world data sets, and compare performance with that of four related outlier detection methods: One-Class SVM with RBF kernel Zhang et al. (2007); Schölkopf et al. (1999), Isolation Forest Liu et al. (2009), Local Outlier Factor Breunig et al. (2000), and Robust Covariance Zoubir et al. (2012). The algorithms are implemented in Python 2.7 using the NumPy and SciPy libraries. In this paper, we set $t = 2/3$ unless otherwise stated.

**Evaluation measure:** We consider a two-class prediction problem (binary classification) in this paper. Receiver operating characteristic (ROC) curves illustrate the diagnostic ability of a binary classifier system as its discrimination threshold is varied. Thus, we use the area under the receiver operating characteristic curve (AUC) for model comparison.

**Data sets:** To evaluate our approach, we use a variety of real-world data sets from the UCI Lichman (2013) as well as synthetic data sets. The relevant information about the data sets is summarized in Table 1. Note that all instances in the smallest class are outliers. For the synthetic data set, we generate normal instances from normal distribution $N(\mu, \sigma^2)$ and 50 outliers from a uniform distribution $uniform(low, high, 50)$ in the range from the minimum to the maximum values of the interior points. We set the variance $\sigma = 0.3$ and the mean vector $\mu = (-1, +1)$ for two dimensions, and let $low = -6$ and $high = 6$. We scale each feature by its maximum absolute value. This does not shift and center the data, and thus does not destroy any sparsity. From Table 1, the number of normal and anomalous instances, and the number of attributes, can be found.

**Table 1. Data set characteristics**

| Dataset | # Normal | # Outlier | # Attribute | # Instance |
|---|---|---|---|---|
| Hr | 45 | 2 | 2 | 47 |
| Dermatology | 112 | 7 | 34 | 119 |
| Hepatitis | 123 | 3 | 19 | 126 |
| Artificial | 150 | 50 | 2 | 200 |
| Ionosphere | 225 | 126 | 34 | 351 |
| Arrhythmia | 245 | 207 | 279 | 452 |
| Nhl2 | 730 | 1 | 2 | 731 |
| Digits (2,8) | 351 | 55 | 2 | 1797 |
| Spambase | 2788 | 1813 | 57 | 4601 |
| Optdigits(2,8) | 1111 | 135 | 64 | 5620 |

**Parameters:** For each data set, the value of $k$ ranges over the set $\{5, 6, \cdots, 5 * log_{10}(N)\}$ calculating each corresponding AUC. Then, we report the average AUC results on all the data sets. The AUC results for Isolation Forest and robust covariance are minuscule with different values of $k$, with a fluctuation of about 27%. In general, however, the different strategies return similar results, and the value of $k$ does not seem to be significant for the accuracy of our algorithm.

**Discussion:** The results of the experiments are shown in Table 2. The proposed method wins most cases (8 data sets over 10), while even when it does not, its AUC is close to the winner's. Moreover, it has the lowest average root mean squared deviation (RSMD) among all competing methods.

As an additional evaluation, we also perform statistical tests to show the effect of different values of parameter $k$. Our method can achieve more robust or stable performance, compared to the comparison methods. From Table 2, our method achieved the minimal standard deviation in experiments. Fig.3 illustrates AUC results with different values of $k$, ranging over $\{5, 6, \cdots, 5*log_{10}(N)\}$, using the optdigits data set. Our method achieves a more effective and stable performance than that of the robust covariance and Isolation Forest methods, which always have a certain degree of random variability. Moreover, Fig.4 illustrates stable performance of our method with different $k = \{5, 6, \cdots, 5 * log_{10}(N)\}$ values on different data sets.

There are no standards for determining boundary quality in a data set. The simple way to illustrate the performance of boundary detection methods is to show the relative locations of

**Table 2. Average AUC results with different values of $k$, in $\{5, 6, \cdots, 5 * log_{10}(N)\}$, on both real-world and synthetic data sets.**

| | One-Class SVM | Isolation Forest | our method | Local Outlier Factor | Robust covariance |
|---|---|---|---|---|---|
| Hr | 0.78 | 0.97 | 0.86 | 0.97 | 0.71 |
| nhl2 | 0.25 | 0.56 | **0.87** | 0.47 | 0.56 |
| ionosphere | 0.66 | 0.64 | **0.72** | 0.62 | 0.64 |
| arrhythmia | 0.67 | 0.60 | **0.69** | 0.58 | 0.60 |
| digits | 0.76 | 0.87 | **0.95** | 0.82 | 0.85 |
| hepatitis | 0.59 | 0.45 | **0.63** | 0.45 | 0.56 |
| spambase | 0.52 | 0.53 | **0.68** | 0.49 | 0.47 |
| optdigits | 0.75 | 0.78 | **0.88** | 0.74 | 0.79 |
| artificial | 0.82 | 0.70 | **0.89** | 0.70 | 0.70 |
| dermatology | 0.76 | 0.98 | 0.94 | 0.98 | 0.79 |
| Average | 0.66 | 0.71 | **0.81** | 0.68 | 0.67 |
| RMSD | 0.37 | 0.32 | **0.18** | 0.35 | 0.22 |
| Max AUC | 0.82 | **0.98** | 0.95 | **0.98** | 0.85 |
| Min AUC | 0.25 | 0.45 | **0.63** | 0.45 | 0.47 |

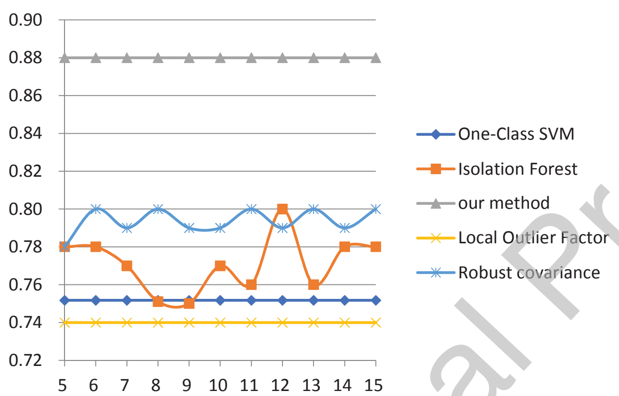



**Fig. 3. AUC results with different values of $k$, in $\{5, 6, \cdots, 5 * log_{10}(N)\}$, on the optdigits data set.**

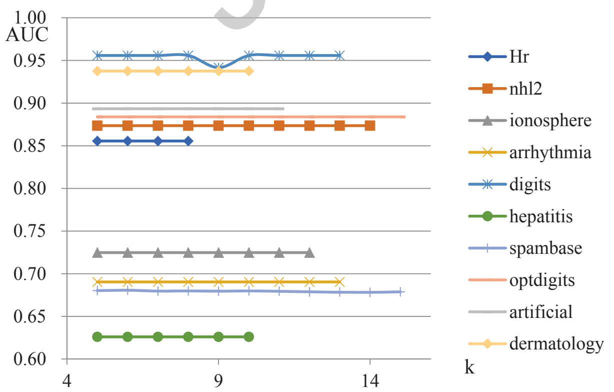



**Fig. 4. The AUC results of our method, with different values of $k$ in $\{5, 6, \cdots, 5 * log10(N)\}$, on different datasets.**

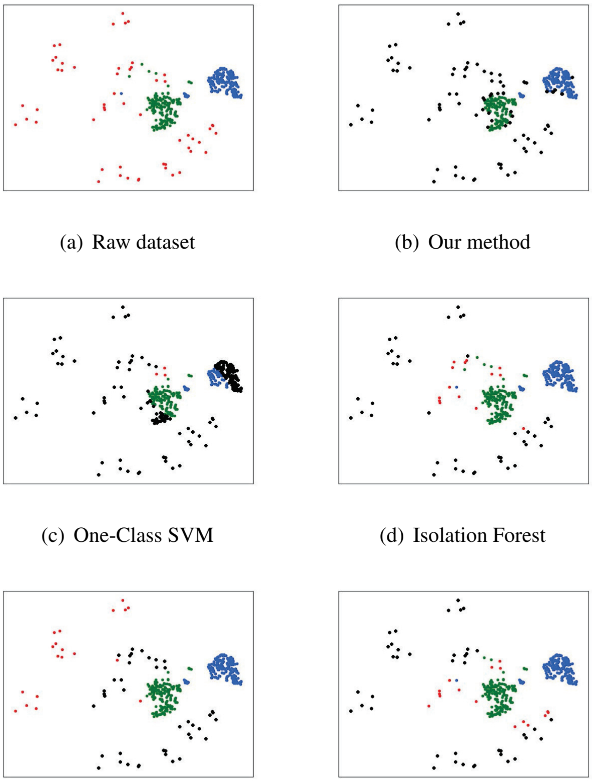
boundary points in visible space. Figure 5 illustrates boundary detection on the MNIST data set of hand-written digits. Note that MNIST is a special dataset that consists of handwritten digits (the size of each MNIST image is 8×8). We embedded these original data into 2D space using t-distributed stochastic neighbor embedding (SNE) van der Maaten (2014). Then classes two and eight were selected to evaluate our method. Inspired by Sugiyama and Borgwardt (2013), we deem the majority classes (the classes with the highest number of observations) as normal points. Then choose eight other small classes and randomly select 3% of points in these two classes as abnormal points. It is easy to see that our method gets better performance.

## 5. Conclusions

This paper proposed a robust geodesic-based method, calculating global disconnectivity score and local real degree as measures of outlierness, which provides a reliable solution for the class imbalance problem. Our method better reflects the characteristics of the data. We showed that additional points are identified as outliers, located near the margin of the clusters. Our method provides a flexible solution for identifying the number of outliers or boundary points. Experimental results obtained for a number of synthetic and real-world data sets demonstrated the effectiveness and robustness of our method.

## Declaration of Competing Interest

We would like to submit the enclosed manuscript entitled Robust Geodesic based Outlier Detection for Class Imbalance Problem, which we wish to be considered for publication in Pattern Recognition Letters. No conflict of interest exits in the submission of this manuscript, and manuscript is approved by all authors for publication. I would like to declare on behalf of my co-authors that the work described was original research that has not been published previously, and not under consideration for publication elsewhere, in whole or in part. All the authors listed have approved the manuscript that is enclosed.



(a) Raw dataset (b) Our method

(c) One-Class SVM (d) Isolation Forest

(e) Local Outlier Factor (f) Robust covariance

**Fig. 5. Outlier detection on MNIST data set embedded in 2D spaces. The different colors correspond to different digits. Black points denote outliers by different methods.**

## Acknowledgment

This work was supported by the National Science Foundation of China (Grant No. 61602066)and the National Key R&D Program of China under contract No. 2017YFB1002201 and supported by National Natural Science Fund for Distinguished Young Scholar (Grant No. 61625204), and the major Project of Education Department in Sichan (17ZA0063 and 2017JQ0030) and the Project Supported by the Scientific Research Foundation (KYTZ201608) of CUIT, and partially supported by the Sichuan international science and technology cooperation and exchange research program (2016HH0018).

## References

Aggarwal, C.C., Yu, P.S., 2001. Outlier detection for high dimensional data, in: ACM SIGMOD International Conference on Management of Data, pp. 37–46.

Angiulli, F., Pizzuti, C., 2002. Fast outlier detection in high dimensional spaces, in: European Conference on Principles of Data Mining and Knowledge Discovery, pp. 15–26.

Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J., 2000. Lof: Identifying density-based local outliers. SIGMOD Rec. 29, 93–104. doi:10.1145/335191.335388.

Campos, G.O., Zimek, A., Sander, J., Campello, R.J.G.B., Micenková, B., Schubert, E., Assent, I., Houle, M.E., 2016. On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. Data Mining and Knowledge Discovery 30, 891–927. URL: https://doi.org/10.1007/s10618-015-0444-8, doi:10.1007/s10618-015-0444-8.

Diestel, R., 2005. Graph Theory (Graduate Texts in Mathematics). Springer.

Ding, X., Li, Y., Belatreche, A., Maguire, L.P., 2015. Novelty detection using level set methods. IEEE Trans. Neural Netw. Learning Syst. 26, 576–588.

Elhamifar, E., 2011. Sparse manifold clustering and embedding, in: International Conference on Neural Information Processing Systems, pp. 55–63.

Hautamaki, V., Karkkainen, I., Franti, P., 2004. Outlier detection using k-nearest neighbour graph, in: International Conference on Pattern Recognition, pp. 430–433 Vol.3.

Hawkins, D.M., 1980. Identification of Outliers. Chapman and Hall,.

Hsiao, K.J., Xu, K.S., Calder, J., Hero, A.O., 2016. Multicriteria similarity-based anomaly detection using pareto depth analysis. IEEE Transactions on Neural Networks and Learning Systems 27, 1307–1321. doi:10.1109/TNNLS.2015.2466686.

Kimmel, R., Amir, A., Bruckstein, A.M., 1995. Finding shortest paths on surfaces using level sets propagation. IEEE Transactions on Pattern Analysis and Machine Intelligence 17, 635–640.

Knorr, E.M., Ng, R.T., 1998. Algorithms for mining distance-based outliers in large datasets, in: International Conference on Very Large Data Bases, pp. 392–403.

Knorr, E.M., Ng, R.T., 1999. Finding intensional knowledge of distance-based outliers. Vldb , 211–222.

Kriegel, H.P., Hubert, M.S., Zimek, A., 2008. Angle-based outlier detection in high-dimensional data, in: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 444–452.

Li, X., Lv, J., Yi, Z., 2016. An efficient representation-based method for boundary point and outlier detection. IEEE Transactions on Neural Networks and Learning Systems PP, 1–12;.

Li, X., Lv, J.C., Yi, Z., 2013. Manifold alignment based on sparse local structures of more corresponding pairs, in: International Joint Conference on Artificial Intelligence, pp. 2862–2868.

Li, Y., Maguire, L.P., 2011. Selecting critical patterns based on local geometrical and statistical information. IEEE Trans. Pattern Anal. Mach. Intell. 33, 1189–1201.

Lichman, M., 2013. UCI machine learning repository.

Liu, F.T., Kai, M.T., Zhou, Z.H., 2009. Isolation forest, in: Eighth IEEE International Conference on Data Mining, pp. 413–422.

van der Maaten, L., 2014. Accelerating t-sne using tree-based algorithms. Journal of Machine Learning Research 15, 3221–3245.

Rodriguez, A., Laio, A., 2014. Clustering by fast search and find of density peaks. Science 344, 1492–1496. URL: http://science.sciencemag.org/content/344/6191/1492, doi:10.1126/science.1242072.

Schölkopf, B., Platt, J., Shawe-Taylor, J., Smola, A., Williamson, R., 1999. Estimating the support of a high-dimensional distribution. Technical Report MSR-TR-99-87. Microsoft Research.

Sugiyama, M., Borgwardt, K., 2013. Rapid distance-based outlier detection via sampling, in: Advances in Neural Information Processing Systems, pp. 467–475.

Tenenbaum, J.B., de Silva, V., Langford, J.C., 2000. A global geometric framework for nonlinear dimensionality reduction. Science 290, 2319–2323.

Xia, C., Hsu, W., Lee, M.L., Ooi, B.C., 2006. Border: efficient computation of boundary points. Knowledge and Data Engineering, IEEE Transactions on 18, 289–303. doi:10.1109/TKDE.2006.38.

Zhai, S., Cheng, Y., Lu, W., Zhang, Z., 2016. Deep structured energy based models for anomaly detection. CoRR abs/1605.07717.

Zhang, R., Zhang, S., Muthuraman, S., Jiang, J., 2007. One class support vector machine for anomaly detection in the communication network performance data, in: Proceedings of the 5th Conference on Applied Electromagnetics, Wireless and Optical Communications, World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, Wisconsin, USA. pp. 31–37.

Zoubir, A.M., Koivunen, V., Chakhchoukh, Y., Muma, M., 2012. Robust estimation in signal processing: A tutorial-style treatment of fundamental concepts. Signal Processing Magazine IEEE 29, 61–80.