https://doi.org/10.2352/ISSN.2470-1173.2018.09.IRIACV-239

Accumulated Relative Density Outlier Detection For Large Scale Traffic Data

Sophia W.T.T. Liu[1], Henry Y.T. Ngan[1], Michael K. Ng[1], Steven J. Simske[2, 3]
[1]Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong
[2]Hewlett-Packard Labs, HP Inc. [3]Systems and Mechanical Engineering, Colorado State University, USA

Abstract

Outlier detection (OD) has been popularly developed in many fields such as medical diagnosis, network intrusion detection, fraud detection and military surveillance. This paper presents an accumulated relative density (ARD) OD method to identify outliers which possess relatively low or high local density. Previously, many density-based OD methods, such as local outlier factor (LOF) and Local Correlation Integral (LOCI), are applied to detect outliers which have low relative density in the data set. Relative local density (RLD) is measured and then compared with each other by statistics to label abnormities. In the proposed ARD method, a big circle centered at every data point is formed first. This big circle covers some data points with its radius. Then, for each encapsulated point inside this big circle, a small circle centered at itself is defined. Afterward, the ratio of number of covered data points inside the small circle of that particular point to the average number of data points in all small circles is defined as the RLD. After RLDs of all data points are calculated, a point whose RLD deviates greatly from the mean of all RLDs will be labeled as an outlier, otherwise as inliers. This ARD method was evaluated by a real world traffic data set which was originally represented as spatial-temporal (ST) traffic flow signals. The ST signals were processed by a principal component analysis (PCA) to reduce its dimension into two-dimensional 2D data points. An average 95% detection success rate (DSR) of OD can be achieved by this method.

Introduction

In recent years, many countries spend lot of efforts to develop smart cities, for which an intelligent transportation system plays a vital role in every modern city. In order to build the transportation being more intelligent like providing more useful information for traffic department, police officers, drivers and passengers, data accuracy and cleanliness is necessary and essential in dealing with massive traffic data. OD is a way to maintain the data quality and provide an alert for anomalies in patterned textures [1] or abnormal events, no matter in data transmission and storage, or real traffic situation [2-3].

Traffic data can be found in many domains [3] such as digital recording cameras, sensors, satellites, etc. As the traffic data is non-stop collected and keeps increasing in database, there are some challenges for the data management. One of the major challenges is the high similarity of data appearance in different spaces and times [3], sometimes it also includes a high dimension problem [3-4]. Another challenge is the lack of information about other factors which may contribute to or affect the traffic conditions, for instance, sudden bad weather, festival celebration and an undergoing pavement construction. Besides, assessing whether an OD is effective or not is not popular in traffic data because there are too few real-world large-scale traffic databases available now [5].

There are mainly 2 types of outliers [6][3]: one type is data with errors due to data set itself during data transmission and storage; another is abnormal event happening on the road, such as traffic jams, low velocity and accidents. Both types of outliers distort the data set, therefore detecting and discarding them before analyzing data are crucial [6]. In order to better manage the transportation system, OD in traffic data is necessary for mainly two reasons. First, an outlier itself can provide useful information. OD can locate an abnormal or undesirable event, for example an accident, so that traffic department or police can respond it on time. In an infrastructure point of view, it can even change any irrational design of transportation system. Second, the quality of data can be improved by removing the detected outliers. Having a city traffic database without being distorted by misleading outliers, normal traffic patterns can be utilized to help drivers make routes decisions.

In the past few decades, many OD methods were proposed and can be broadly classified as model-based [3, 7-12], proximity-based [5, 7, 10, 13], clustering-based [14] and classification-based [15-17] approaches. The model-based method is trying to find the best model to describe the inliers, then those points who do not match the model are considered as outliers. The clustering approach is to divide data into different groups according to some defined features. Similarly, the classification-based approach is to use some prior knowledge about data characteristics to classify outliers and inliers. In some situations, fitting data into a perfect model or finding obvious features are difficult in model-based and clustering-based approaches so that they may be impractical. In addition, the classification-based approach requires the knowledge of data within a data set which is not that straight forward. These limitations lead to our re-visit of the proximity–based approach, in which the main idea is to investigate and compare a data point to its nearby datum. The proposed ARD is a proximity-based method, in which the proximity is about density-how many data points are embedded in a local region (or denoted as big/small circle).

The proposed ARD OD method in this paper is a fusion of density-based and classic statistical OD methods. Generally speaking, a datum is considered as an outlier if there are too many or too little data points around it. Density of a datum is measured by the number of elements in a certain region near this datum, and the relative density is the ratio of the density of a datum to the average density of its nearby elements. We assume that a data set free of outliers can consist of several clusters whose densities may be varied one-by-one. Locally or within a cluster, the density should be at a

Table I. 19 Traffic Direction Distributions of The Traffic Dataset from The 4-arm Junction in Hong Kong.

Direction	z_1	z_2	z_3	z_4	z_5	z_6	z_7	z_8
Entry	E	S	W	N				
Exit					E	S	W	N

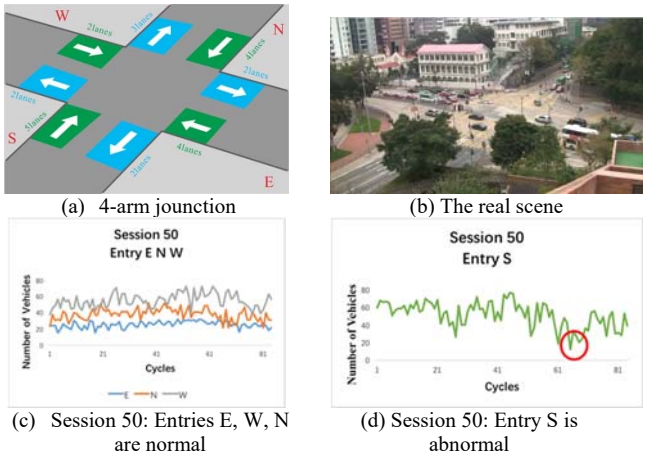
Direction	z_9	z_{10}	z_{11}	z_{12}	z_{13}	z_{14}	z_{15}	z_{16}	z_{17}	z_{18}	z_{19}
E.D.D.	E_l	E_r	E_s	S_l	S_r	S_s	W_l	W_s	N_l	N_r	N_s

Remark: S, N, W, E for South, North, West, East; E.D.D. for Entry Direction Distribution

certain range, or even the same for evenly-distributed cluster, so that the relative density should be around the value 1. For an outlier embedded in the cluster of inliers, its relative density is higher than other than the inliers; For an outlier far away from inliers, its relative density may be very small. Combining these two cases, the datum whose relative density deviates largely from the mean of relative density is suspected. According to the famous Chebyshev's theorem [18], for any number k greater than 1, at least $(1-\frac{1}{k^2})$ of the data values lie within k standard deviations of the mean. By applying Chebyshev's theorem, we can define a region for inliers which separate them from outliers. To determine how large the threshold is, optimizations of the best parameters are imperative.

The OD method proposed by this paper is tested in a real-world dataset collected by video camera of a crossroad in Hong Kong (Fig. 1) [3]. This is a 4-road crossroad with each road consisting of 2-3 paths. Totally, there are 19 directions as shown in Table I. The traffic flow of these directions were casted during rush hours in 31 days (23 workdays and 4 weekends). Each day consists of two rush hour sessions: 07:00 am -10:00 am, 17:00 pm - 20:00 pm and then be processed into signal data. Because of the high similarity lies in this spatial-temporal data, it is difficult to detect outliers, therefore a PCA was carried out to reduce the dimension of data from approximately 80 to 2. The proposed method has a good performance in this data set: average DSR is more than 95%.

The paper is organized as follows: Section II gives a literature review of OD methods. Section III describes how the ARD OD method works. Section IV presents the experimental results and Section V draws the conclusion.



(a) 4-arm jounction (b) The real scene

(c) Session 50: Entries E, W, N are normal (d) Session 50: Entry S is abnormal

Fig. 1. (a) Idea map of the 4-arm junction; (b) sample of the read scene; (c) normal ST signals; (d) abnormal ST signal.

Literature Review

A general definition of outlier is given in [19]: an observation (or subset of observations) which appears to be inconsistent with the remainders in one data set. OD aims at detecting the abnormal or outlying data embedded in a data set. The basic assumption for OD is that a data point with a large deviation from the normal model is likely to be anomalous [20]. In simplicity, we classify the existing OD methods into three main approaches: model-based, proximity-based and classification-based approaches.

2.1 Model-based Approach

The general idea of the model-based approach is to fit a model which is able to describe the data well. Intuitively, outliers are those largely deviate from the majority in the model. The model-based approach includes statistical-based, depth-based and deviation-based methods. Among all, the statistics one is more prevailing and popular. Model-based OD approach's main advantage is its simplicity in detecting if a suitable model is found or constructed which is usually difficult.

(a) Statistics-based Method

The statistical-based method [7] is to assume one data set following a specific statistical distribution such as normal distribution [12], Dirichlet process mixture model (DPMM) [3] and then detect the outliers by a discordancy test. This method is fast in computation with complexity $O(n)$. It is simple but difficult to determine a suitable distribution and parameters. Actually, in real world, many data sets are not fit to one particular mathematical distribution. [7] pinpoints another limitation that it can only deal with a data set in one dimension only. Also, sometimes a data set can be partitioned into several clusters so that implementing only one single method may not be sufficient [7]. Thus a multiple-distribution model can be applied to describe the data-set where some clusters belong to a particular distribution.

(b) Depth-based Method

A depth function defined in [8] is used to describe to what extent a data point belongs to a cluster or follows a distribution. Contours are used to encapsulate data points in different depths. The less degree the depth is, the more possible being an outlier this point is. Among different depth functions, a spatial depth is of high computational efficiency [9], but not very effective in detecting outliers to remote to a cluster of inliers due to ignorance of the importance of distance [5].

(c) Deviation-based Method

The main idea of a deviation–based OD method is to use a smoothing factor [10] to measure how much the deviation can be reduced if a data point is removed. By intuition, the larger a data point contributes to the deviation of whole data set, the more

TABLE II.
COMPARISON OF ARD, LOF AND LOCI OD METHODS.

	ARDS	LOF [8]	LOCI [11]
Basic assumption about outlier	Higher/ lower local density	Lower local density	
Outliers' feature (compared with inliers)	More or less nearby elements	More distant to its nearby elements	Less nearby elements
Definition of neighborhood	A circle whose radius auto-adjusts to size of data set	k^{th}-nearest-neighborhood	A circle of a fixed radius
Measurement of density	Number of elements in neighborhood	Sum distances between a point to its neighbors	Number of elements in neighborhood
Decision making	RLD exceeds an inlying region	RLD exceeds an inlying region	Density exceeds an inlying region
Outlier regions	2 parts	1 part	1 part
Data manipulation	Unsupervised	Supervised or semi-supervised	Unsupervised
How is inlier region decided	Statistics-based; Evaluating the mean, standard deviation of relative density for all points in a dataset	By training	Statistics-based; Evaluating the normalized mean, standard deviation of numbers of inside for points inside each neighborhood
Advantages	Free of training process; Sensitive to outliers embedded inside clusters of uniformly-distributed inliers	Insusceptible to extremely closed data points, thus density factor is less biased	Free of training process; parameter-free
Disadvantages	Difficulty in choosing parameters; Ineffective in detecting outlier far away from inlying outliers	Inefficient of detecting outliers which have a high local density	Difficulty in choosing parameters; Inefficient of detecting outliers with high density, or far away from inlying outliers; High false positive

outlying is this data point. Three depth-based algorithms are proposed in [11], two of them are based on dynamic programming and another one takes the advantage of deviation factor, all are very precise and efficient in computing.

2.2 Proximity-based Approach

The main idea of this approach [13-14] is to measure the proximity of a point to other points. Intuitively, an outlier is more dissonant with remaining inliers. This approach becomes the last resort when no suitable model can fit the data set. The proximity can be measured by comparing distance, density or angle, etc.

(a) Distance-based Method

Ramaswamy et al. [13] gave a new definition of the distance-based method based on the k^{th} smallest distance between a given point p and the remaining points, which is denoted by *DK (p)*. The top m points having the largest *DK (p)* are considered as outliers. By [7], this method is effective and can be used in multidimensional OD. The disadvantage is that it is highly dependent on the parameters k and m, for which it is difficult to find appropriate parameters in real-world situation. Another drawback is long computing time since it has to measure distance between every single point and any other point. According to [21], this method cannot detect outliers successfully when a dataset has non-dense neighborhoods. The distance-based method can mine global outliers effectively but it is difficult to detect outliers if the dataset has groups of different densities [10].

(b) Angle-based Method

A parameter-free angle based method is proposed in [5]. Its main idea by intuition is to use the variance of angles between any difference vectors to a data point, then the smaller the variance the more likely this is an outlier. This approach reaches high efficiency in high-dimensional data in comparison to LOF, but it is very time-consuming with complexity n^3. Another obvious drawback is this method cannot detect the outliers lying in the clusters of inlier. When 80-dimensional and PCA-processed 2-dimensional dataset mentioned in [22] tested on the angle-based OD, the result is not ideal because it cannot detect outliers lying inside the inliers.

2.3 Classification-based Approach

The basic assumption of the classification approach is that a classifier can learn from a given feature space in order to distinguish inliers and outliers [14-15]. This can be a learning supervised approach which consists of two main steps: training and testing, or just an unsupervised approach [2]. Every datum is assigned to a known class in comparison to clustering approaches applied in dataset where a definite classification is not given [16]. This approach can be very effective under the circumstance that outliers assemble to a cluster/class, yet this often is not the case. Besides, this approach is highly dependent on vast volume of prior information results in its unavailability in many situations.

(a) Density-based Method

Local Outlier Factor (LOF) is defined in [23] which describes what degree a single point belongs to the group of outliers. One variant is bounded LOF [17] and another variant is local correlation integral (LOCI) in [24], which uses an algorithm to calculate density of part of data points to save computing time. In LOCI, a multi-granularity deviation factor (MDEF) is used to compute the relative density of a point to its neighborhoods. Table II is used to summarize similarity and difference in main features of OD method based on ARD for LOF and LOCI.

Apart from detecting the outliers lying outside the clusters of inliers, the proposed ARD method aims to detect outliers appearing inside a group of inliers. Both LOF and LOCI methods neglect the fact that outliers may be embedded inside the regions of inliers which lead to a relatively high local density. However, this is one type of outliers that only our ARD method would like to tackle with. Besides, compared to the LOF method, the ARD method is free of

a learning process which can be carried out when no past information is available for training. In addition, in contrast to the LOCI method, the ARD method uses a radius the radius of which is auto-adjusted to the size of data set, to define neighbourhood(s). Finally, the ARD method is more reliable than the LOCI method. For example, the ARD method applies the relative density of all data points to form a distribution while the LOCI method employs only a few points to construct a distribution which may lead to a fluctuation of the performance of the method.

ARD OD Method

In this section, mathematics, algorithm and procedure of the ARD method will be offered based on a real traffic data set recorded in Hong Kong for evaluation.

3.1 Data description

The original traffic data [3] was video recorded and then transformed as ST signals, for which has too high a level of dimension. PAC process was carried out to reduce the dimension to 2 as well as maintain the major features of data. There are 2 data sets: the AM sessions (7:00 am -- 10:00 am) and PM sessions (5:00 pm -- 8:00pm). In each session, there are 19 directions (Table I) denoted by z_j where$j = 1,2 \ldots.,19$. 23 workdays of every direction was chosen to test, thus each direction data set z_j consists of 23 2D data points: $p_i\ (x_i\ ,y_i\ )$ for $i = 1,2 \ldots.,23$. For example, for north entry (z_4) of the AM session, $p_3\ (x_3\ ,y_3\ )$ refers to the $(x, y)-$coordinates of PCA-processed data on the 3rd day in the direction of z_4.

3.2 Mathematics

Definitions 1 (Diagonal distance). For a particular session in traffic direction j of the dataset, there are 23 $(x, y)-$datapoints. The maximum and minimum of x and y are obtained by the smallest rectangle which contains all datapoints with edges parallel to the x- or y-axis.

x_{min}=min $\{x_1\ , x_2\ \ldots\ldots\ x_{23}\}$, (1)
x_{max} =max$\{x_1\ , x_2\ \ldots\ldots\ x_{23}\}$, (2)
y_{min} =min $\{y_1\ , y_2\ \ldots\ldots\ y_{23}\}$, (3)
y_{max} =max $\{y_1\ , y_2\ \ldots\ldots\ y_{23}\}$ (4)

Then, a diagonal distance of the rectangle above is defined as

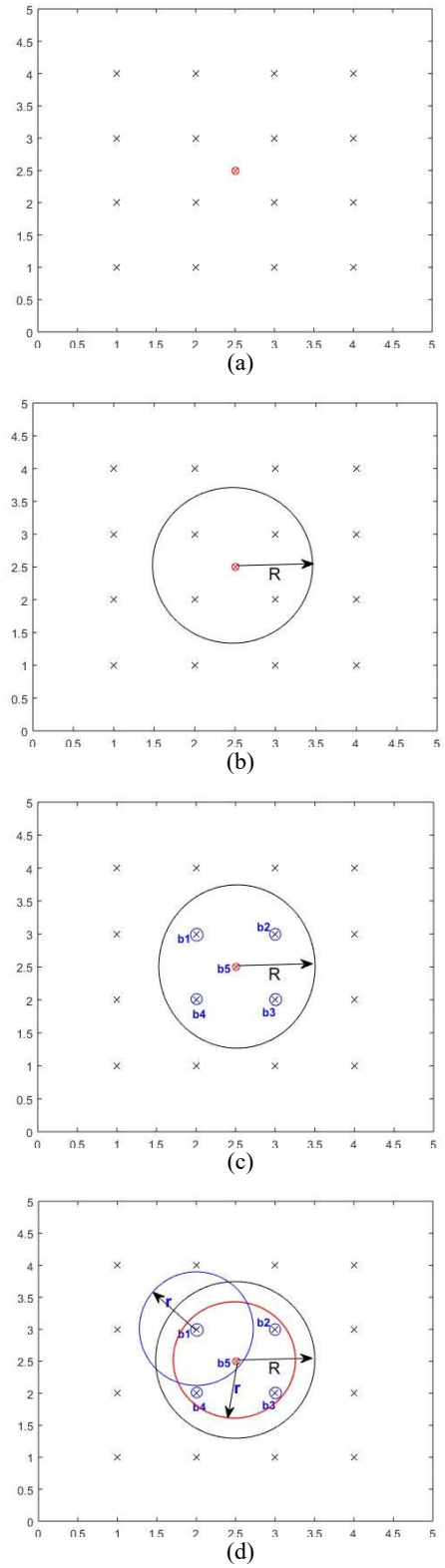
$$D\ = \sqrt{(x_{max} - x_{min})^2 + (y_{max} - y_{min})^2} \quad (5)$$

This procedure aims to approximate the size of the data set for further use.

Definition 2 (Radius). Two radii, namely big-circle and small-circle (defined in definition 3) radii, are defined as below,

Big-circle radius: $R^j = D^j \times \mathrm{d}\ ,$ (6)
Small-circle radius: $r^j = R^j \times \mathrm{K}\ ,$ (7)

where a diagonal-rate denoted by d, a small-to-big rate denoted by K.



(a) (b) (c) (d)

Fig. 2. Big and small circles (a) 17 points; (b) Big circle; (c) Elements in big circle; (d) Small circles.

Definition 3 (Big-circle and Small-circle neighborhoods). Let $O(p,q)$ denote a circle centered at p with a radius q. In a dataset z_j, for a point $p_i = (x_i, y_i)$, a big circle of p_i is defined as $O(p_i, R)$, then all points in z_j (including p_i itself) lying inside the region of $O(p_i, R)$ are big-circle's elements, denoted by $p_{i,k}$. A big circle- neighborhood(s) of p_i (denoted by b5 in Fig. 2(b)) is defined as

$$B(p_i) = \{p_{i,k} \in z_j \mid \text{norm}(p_i - p_{i,k}) \le R^j\}, \quad (8)$$

where $N_b(p_i) = |B(p_i)|$ which denotes the number of big-circle elements of p_i . $k = 1,2\ldots\ldots, N_b$ As shown in Fig. 2(c), the circle is the big circle of b5, and the elements within the circle: b1, b2, b3, b4, b5 are the big-circle elements of b5.

Similarly, each big circle element p_k^i has its own small circle's neighborhood(s) which is defined as the set of all points inside $O(p_{i,k}, r^j)$,

$$S(p_{i,k}) = \{p_{i,k,h} \in z_j \mid norm(p_{i,k,h} - p_{i,k}) \le r^j\}, \quad (9)$$

where $N_s(p_{i,k}) = |S(p_{i,k,h})|$. Particularly, the number of small neighbors of p_i is denoted by $N_s(p_i)$. As shown in Fig. 2(d), the blue circle is the small circle of b1 and the red circle is the small circle of b5. Elements within the red circle: b1, b2, b3, b4, b5 are small-circle elements of b5.

Definition 4 (RLD). A RLD is defined as

$$\rho(p_i, \text{K}, \text{d}) = \frac{N_s(p_i)}{L}, \quad (10)$$

where $L = \frac{\sum_{All\, p_{i,k}} N_s(p_{i,k})}{N_b(p_i)}$.

This definition is a variant from the MDEF defined in [25].

Definition 5 (Mean and standard deviation). The mean and standard deviation of the relative density of all $p_i \in z_j$ are denoted as μ_ρ^j and σ_ρ^j for a relative density ρ.

Definition 6 (Inlying region and outlying region) Inlying region of and outlying region of z_j are expressed as

$$(-\alpha \times \sigma_\rho^j + \mu_\rho^j, \alpha \times \sigma_\rho^j + \mu_\rho^j), \quad (11)$$

and

$$(-\infty, -\alpha \times \sigma_\rho^j + \mu_\rho^j] \cup [\alpha \times \sigma_\rho^j + \mu_\rho^j, \infty), \quad (12)$$

respectively.

3.3 Procedure

After possessing all the 2D data points. A rectangle is drawn to measure the size of the data set, then the radii of the big-circle and small-circle (both are scalable) are determined by the diagonal distance of this rectangle. For a data point, the number of all elements (aka data points) in its big-circle is counted. In the small-circle of each of these data points, the corresponding number of data points is also counted. Then, a RLD is calculated. After calculating the relative density of all points in the data set, the mean value and standard deviation can be computed and used to construct the inlying and outlying regions. Then, each data point falling outside of the inlying region is labeled as an outlier, otherwise it is an inlier. The algorithm and the flowchart of procedure (Fig. 3) are shown. There are three stages in Fig. 3: Stage 0 of the generation of 2D/3D data points, Stage 1 of the measurement of relative local density and Stage 2 of outlier decision. An algorithm of the pseudo-codes of the ARD method is presented as below too.

Algorithm: ARD statistics-based OD with auto-selected radius
Requirement: Input dataset consisting of 2D data points
Inputs: AM/PM sessions; testing set z_j ; parameters: d, α and K
Output: labels of 2D data points
1: Input session 'am' or 'pm';
3: x_{min}=min $\{x_1, x_2 \ldots\ldots x_{23}\}$, x_{max} =max $\{x_1, x_2 \ldots\ldots x_{23}\}$, y_{min} =min $\{y_1, y_2 \ldots\ldots y_{23}\}$, y_{max} =max $\{y_1, y_2 \ldots\ldots y_{23}\}$;
4: Calculate a diagonal distance: D^j
5: Calculate a big-circle radius: $R^j = D^j \times \text{d}$; a small-circle radius $r^j = R^j \times \text{K}$
6: **for** each point $p_i \in z_j$ **do**
7: Find big-circle neighborhood: $B(p_i)$, the number of big-circle elements: $N_b(p_i)$
8: **for** each big-circle element of $p_{i,k}$ **do**
9: Find small-circle neighborhood: $S(p_{i,k})$
10: Find the number of its small-circle element: $N_s(p_{i,k})$
11: **end**
12: Measure the RLD: $\rho(p_i, \text{K}, \text{d})$
13: Calculate mean (μ_ρ^j) and standard deviation (σ_ρ^j) of the ρ of all points $\in z_j$
14: Determine outlying region of z_j :
$(-\infty, -\alpha \times \sigma_\rho^j + \mu_\rho^j] \cup [\alpha \times \sigma_\rho^j + \mu_\rho^j, \infty)$
15: **for** each point p_i **do**
16: **If** $\rho(p_i, K, d) \in$ outlying region
17: Label p_i as an outlier
18: **else**
19: Label p_i as an inlier
20: **end**

The following example is used to illustrate the steps 1-5 of Stage 1 (Measurement of RLD) in Fig. 3. A simulated dataset consisting of 17 data points, with 16 inliers and 1 outlier (marked red) is plotted in Fig. 4. After the data is input, a diagonal distance as well as R and r can be calculated. An example of the procedure of the ARD OD method is as follows.

Step 1: For a point $\boldsymbol{p}$ (the red one in Fig. 4(a)), draw a circle centered at itself with radius R Fig. 4(b). All points in this circle are big-circle elements of $\boldsymbol{p}$: $b_1, b_2, b_3, b_4(p)$, $N_b = 4$.

Step 2: For each of these 4 big-circle elements, draw a circle centered at itself with radius r. Small circles for b_1 and b_4 are shown in Fig. 4(c),(d).

Step 3: All points in a small circle are small-circle elements of this big-circle element. As shown in Fig. 4(c), b_1 has 6 small-circle elements: $b_{1,1}, b_{1,2}, b_{1,3}, b_{1,4}, b_{1,5}(b_1), b_{1,6}(b_2)$, whereas b_4 has 2 small-circle elements: $b_{4,1}(b_3), b_{4,2}(b_4)$ (Fig. 4 (c)). In fact,

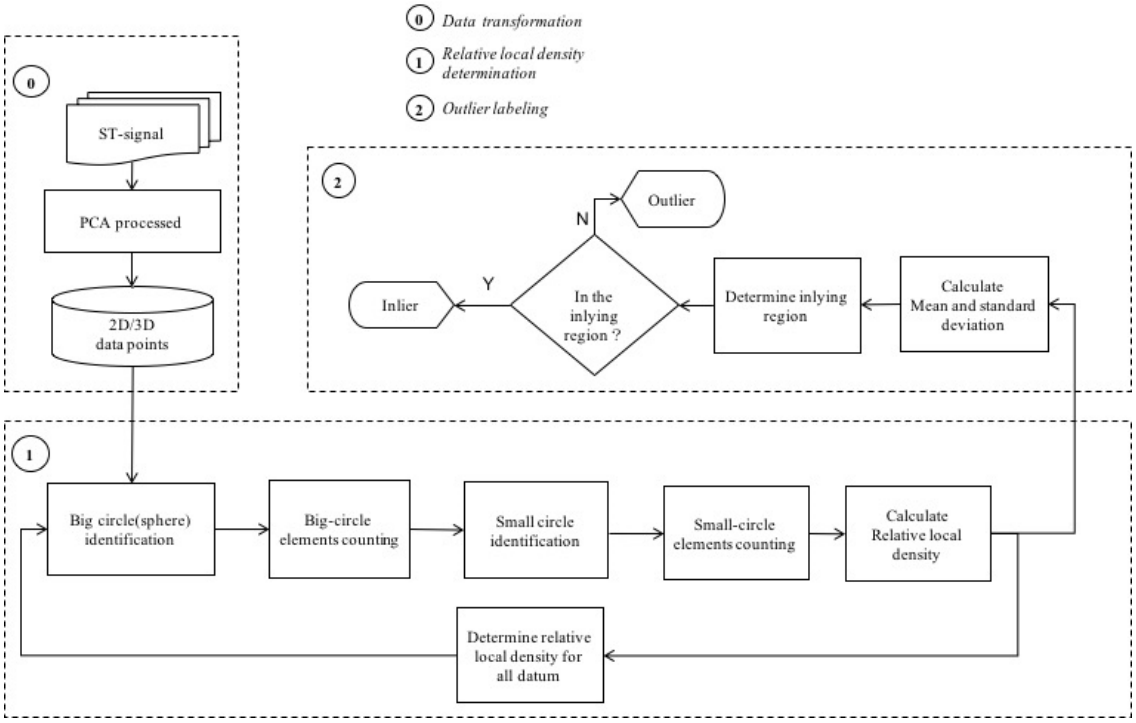


Fig. 3. Procedure of the proposed ARD method.

b_2, b_4 have 4 and 2 small-circle elements, respectively. Thus, $N_s(b_1) = 6,\ N_s(b_2) = 4,\ N_s(b_3) = N_s(\mathrm{p}) = 2$;

Step 4: The RLD of the red point:

$$\rho = \frac{N_s(\mathrm{p})}{\left(N_s(b_1) + N_s(b_2) + N_s(b_3) + N_s(\mathrm{p})\right)\Big/ N_b(\mathrm{p})} = \frac{2}{(6+4+2+2)/4} = 0.5714$$

Step 5: Repeat Step 1-4 to calculate the relative local density for other points.

Performance Evaluation

The experimental results were computed by DELL OptiPlex 9010, Core i7-3770 3.40GHz with MATLAB 2014A. The methods are evaluated by the following aspects: True positive (TP); True negative (TN); False positive (FP); False negative (FN); Detection success rate (DSR) $= \frac{TP+TN}{TP+FP+TN+FN}$; True positive rate (TPR) $= \frac{TP}{TP+FN}$; Positive predictive value (PPV) $= \frac{TP}{TP+FP}$; False positive rate (FPR) $= \frac{TP}{FR+TN}$; Negative predictive value (NPV) $= \frac{TN}{TN+FN}$. Meanwhile, DSR is one of key measurements in the evaluation of this paper.

4.1 Parameter selection

4.1.1 Parameter description

There are 3 parameters involved in the algorithm: *d, K and α* (denoted by alpha in the figures). *d* is used to decide the radius of a big neighborhood; *K*, along with d, is used to decide the radius of a small neighborhood; Given a data set, the bigger the parameters d and K are, the larger the radius of big and small circle. On one hand, if the radius is too big, it may cause a long computing time in a large data set because of the huge number of elements in the circle. Besides, too big a circle may cover some other clusters with different relative densities, which will affect the performance. On the other hand, if the radius is too small, there may be too few elements in the circle for the testing point, which is the center of the circle, to compare density with. *α* is used to decide the inlying and outlying regions, i.e. the threshold for outliers and inliers. The bigger *α* is, the less intendancy for the method to label a point as an outlier, therefore potentially higher FN and lower FP. In order to obtain the best parameters for ARD method, experiments were carried out to find the relationship among these 3 parameters independently. The data set described in Section 3.1 was used for the analysis below.

4.1.2 Relationship among different parameters

(a) Relationship between d and DSR

Fig. 5 illustrates how DSR changes along the selection of d. Given fixed K and *α,* DSR starts to be stabilized and maintained at a high level (above 90%) when d increases to and is larger than 0.2. When d is less than 0.2, the average DSR is low because the circle drawn is too small so that very few, or even none of the elements is in the big circle, except for the datum itself. As a result, we may fail to compare the density with enough reference. The bigger the circle has the more number of points for reference.

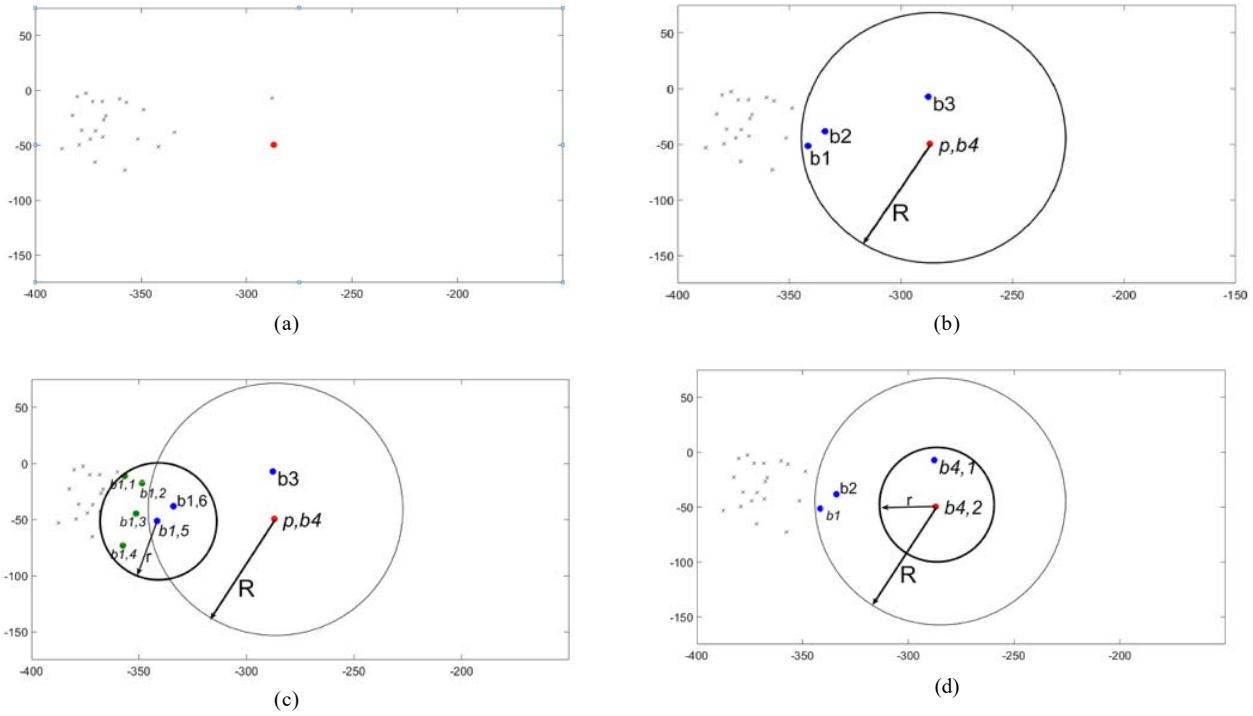


Fig. 4. Stage 1's procedure illustration of AM z5 PCA (x, y) −coordinates (a) plot of all points in z5; (b) Big circle; (c) Small circle of b1 (d) Small circle of b4.

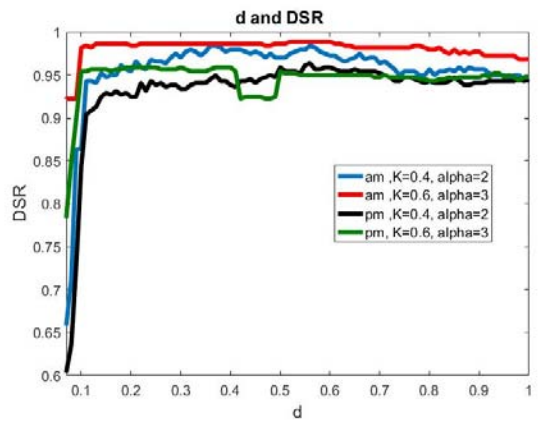


Fig. 5. DSR versus different d values

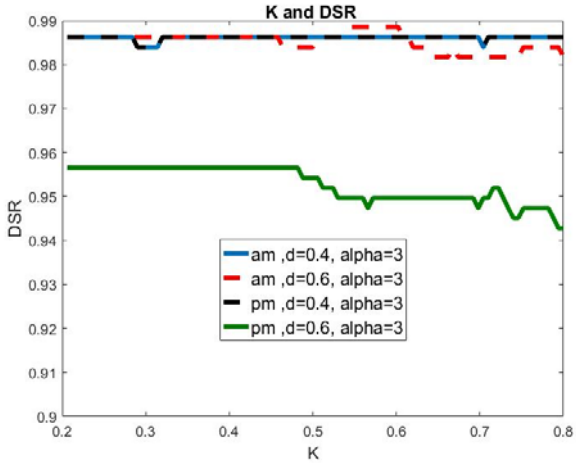


Fig. 6. DSR value for different K values

(b) Relationship between K and DSR

Fig. 6 shows how DSR changes along the selection of K. Given fixed d and α, there is no clear pattern for the DSR-K relationship, but DSR stays at a relative high level (above 90%) when d=0.6 or 0.4 and alpha α=3. Compared to the other two parameters, K does not affect DSR that much. However, it is not hard to see that the value of K has a direct impact on computing time needed.

(c) Relationship between α and DSR

Fig. 7 demonstrates how DSR changes along the selection of α. Given fixed K and d, DSR is found to be stable and staying at a high level when α is larger than 2 (sometimes decreases slightly for given K and d). So when α is too small (less than 2), which means a very narrow inlying region, many inliers will be wrongly classified as outliers and when α is too large (larger than 3), the DSR decreases because some outliers are missed, but the decline is not lethal as for in our case, the data set contains very few outliers.

The target of an overall DSR was set to be at 90% for three parameters that can result in higher DSR than 90% for following trials. From the results, we can determine 3 ranges for parameters that can lead to high DSR independently. However, high DSR is definitely not the only issue to be pursued. If an algorithm labels almost all the data points as inliers and rarely as outliers, the TN will

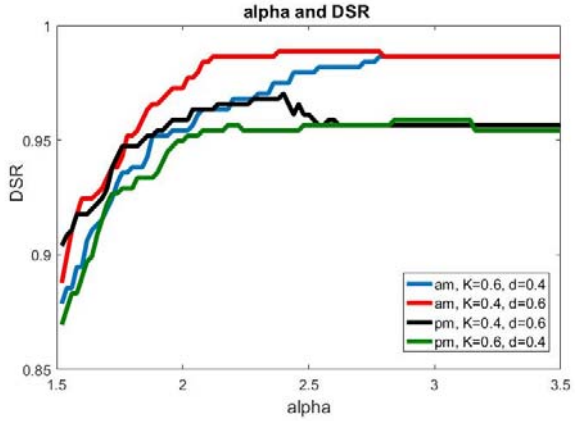


Fig. 7. DSR versus different alpha values

be very high with a relatively low TP. For this data set with a small percentage of outliers (6 out of 437 for the AM sessions and 20 out of 437 for the PM sessions), a high DSR can be achieved easily by enlarging the inlying region. The problem is obviously the relatively low TPR, which means the algorithm is not effective in detecting outliers. In order to find parameters that can lead to a relatively high TPR without sacrificing too much DSR, an experiment was carried out to compute average DSR and TP under the circumstances of different combinations of three parameters respectively from the ranges determined by aforementioned experiments. Tables III and IV deliver for each particular TP, the corresponding maximum and minimum of average DSRs.

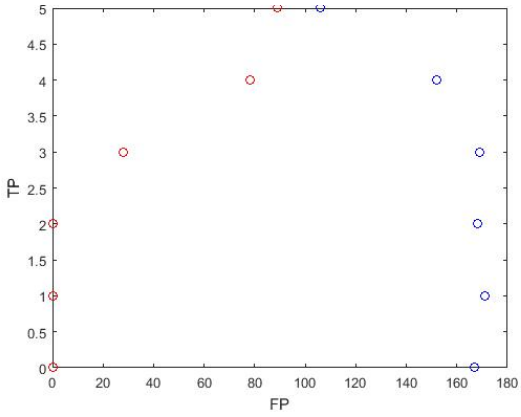
The ARD were carried out to test the TP and FP for each of these combination of parameters, which is shown in Fig. 8. The *x* and *y* coordinates of red circles are FP and TP of the results for given parameters which can lead to the maximum (blue circle for the minimum) average DSR at a specific TP. The parameter combinations corresponding to points lying on the left-upper corner are preferred because of the relatively high TP and low FP.

Concerning high DSR as well as an acceptable TPR, the following parameters are chosen: Parameters for the AM sessions: d=0.5, K=0.7, α=3.1 and parameters for the PM sessions: d=0.68, K=0.6, α=2.2.

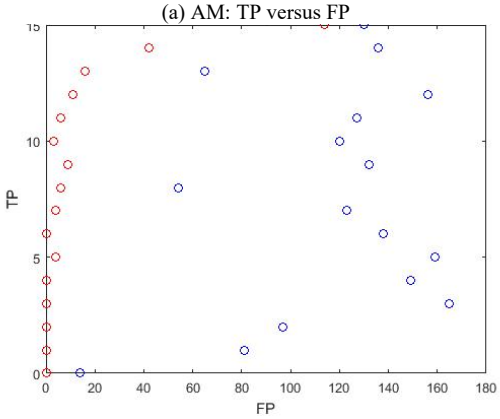
(d) Summary of evaluation

By the selected parameters with the optimal TP, DSR and other measurements of 19 different directions in the data set, the evaluation for the AM and PM sessions are tabulated in Tables V and VI.

Outliers lie outside the cluster of inliers, for example, z3 and z16 for the AM sessions and z8 for the PM sessions, the ARD method can detect them successfully because they have a relatively small relative local density compared to other points. But for outliers embedded inside clusters of inliers, for instance, z4 for the AM sessions, the method fails to detect the outlier. In the data set of z4 for the AM sessions, the cluster containing the outlier is unevenly distributed, to be specific, some points lie close to each other, while some are more distant to other points. This kind of distribution



(a) AM: TP versus FP



(b) PM: TP versus FP

Fig. 8. Plot of TP versus FP: (a) AM sessions; (b) PM sessions.

TABLE III. TRIALS OF ASSOCIATIVE PARAMETER FOR AM SESSIONS.

TP		DSR	d	K	alpha
	max	0.79	0.62	0.80	1.00
5	min	0.76	0.56	0.80	1.00
	max	0.82	0.74	0.70	1.00
4	min	0.65	0.62	0.40	1.00
	max	0.93	0.74	0.30	1.60
3	min	0.61	0.50	0.30	1.00
	max	0.99	0.50	0.70	3.10
2	min	0.61	0.44	0.30	1.00
	max	0.99	0.62	0.40	2.50
1	min	0.60	0.38	0.30	1.00
	max	0.99	0.44	0.40	2.20
0	min	0.60	0.38	0.40	1.00

enlarges the standard deviation of relative density of each point. Therefore, the relative density of the outlier is not big enough to excess the inlying region, which leads to failure of detection. In addition, an average computing time for one traffic direction is 0.021s.

(e) A study of LOCI-MDEF with a fixed R

The density-based OD method based on the LOCI-MDEF proposed in [25] is used to test the aforementioned data set. First, an experiment is carried out to select the best R. The DSR reaches a

high level when R is larger than 40. 40 is chosen for testing and corresponding results are shown in the Table VII. As shown in Table VIII, the ARD method outperforms the LOCI and supervised LOF OD methods by achieving relatively high DSR and TP rates.

TABLE IV. ASSOCIATIVE PARAMETER TRIAL FOR PM SESSIONS

TP		DSR	d	K	alpha
	max	0.73	0.68	0.30	1.0
15	min	0.69	0.62	0.30	1.0
	max	0.89	0.74	0.50	1.3
14	min	0.68	0.56	0.30	1.0
	max	0.95	0.68	0.60	2.2
13	min	0.84	0.62	0.30	1.3
	max	0.96	0.74	0.40	2.2
12	min	0.63	0.38	0.70	1.0
	max	0.97	0.56	0.40	2.2
11	min	0.69	0.44	0.60	1.0
	max	0.97	0.68	0.30	2.2
10	min	0.70	0.44	0.70	1.0
	max	0.96	0.50	0.60	2.2
9	min	0.68	0.32	0.80	1.0
	max	0.96	0.50	0.50	2.2
8	min	0.85	0.44	0.80	1.6
	max	0.96	0.68	0.40	2.5
7	min	0.69	0.44	0.50	1.0
	max	0.97	0.62	0.30	2.2
6	min	0.65	0.44	0.40	1.0
	max	0.96	0.50	0.70	2.8
5	min	0.60	0.38	0.30	1.0
	max	0.97	0.56	0.40	2.5
4	min	0.62	0.32	0.30	1.0
	max	0.96	0.50	0.30	2.5
3	min	0.59	0.20	0.30	1.0
	max	0.96	0.50	0.30	2.8
2	min	0.74	0.26	0.50	1.3
	max	0.96	0.38	0.30	2.5
1	min	0.77	0.32	0.50	1.3
	max	0.96	0.62	0.30	2.5
0	min	0.92	0.20	0.30	1.9

TABLE V. EVALUATION OF THE AM SESSIONS (d =0.5, K=0.7, α =3.1).

DIR.	TP	FP	TN	FN	DSR	PPV	NPV	TPR	FPR
Z1	0	0	23	0	1.000	NaN	1.000	NaN	0
Z2	0	0	22	1	0.957	NaN	0.957	0	0
Z3	1	0	22	0	1.000	1	1.000	1	0
Z4	0	0	22	1	0.957	NaN	0.957	0	0
Z5	0	0	23	0	1.000	NaN	1.000	NaN	0
Z6	0	0	23	0	1.000	NaN	1.000	NaN	0
Z7	0	0	22	1	0.957	NaN	0.957	0	0
Z8	0	0	23	0	1.000	NaN	1.000	NaN	0
Z9	0	0	23	0	1.000	NaN	1.000	NaN	0
Z10	0	0	23	0	1.000	NaN	1.000	NaN	0
Z11	0	0	23	0	1.000	NaN	1.000	NaN	0
Z12	0	0	23	0	1.000	NaN	1.000	NaN	0
Z13	0	0	23	0	1.000	NaN	1.000	NaN	0
Z14	0	0	23	0	1.000	NaN	1.000	NaN	0
Z15	0	0	23	0	1.000	NaN	1.000	NaN	0
Z16	1	0	22	0	1.000	1	1.000	1	0
Z17	0	0	23	0	1.000	NaN	1.000	NaN	0
Z18	0	0	22	1	0.957	NaN	0.957	0	0
Z19	0	0	23	0	1.000	NaN	1.000	NaN	0
Sum	2	0	431	4	0.991	NaN	0.991	NaN	0

TABLE VI. EVALUATION OF THE PM SESSION (PM: d=0.68, K=0.6, α =2.2).

DIR.	TP	FP	TN	FN	DSR	PPV	NPV	TPR	FPR
Z1	0	1	22	0	0.957	0	1.000	NaN	0.043
Z2	3	0	18	2	0.913	1	0.900	0.6	0.000
Z3	0	1	21	1	0.913	0	0.955	0	0.045
Z4	0	1	22	0	0.957	0	1.000	NaN	0.043
Z5	1	1	21	0	0.957	0.5	1.000	1	0.045
Z6	1	0	22	0	1.000	1	1.000	1	0.000
Z7	1	0	22	0	1.000	1	1.000	1	0.000
Z8	3	0	20	0	1.000	1	1.000	1	0.000
Z9	0	0	23	0	1.000	NaN	1.000	NaN	0.000
Z10	0	2	21	0	0.913	0	1.000	NaN	0.087
Z11	0	2	21	0	0.913	0	1.000	NaN	0.087
Z12	0	2	20	1	0.870	0	0.952	0	0.091
Z13	1	1	21	0	0.957	0.5	1.000	1	0.045
Z14	3	0	19	1	0.957	1	0.950	0.75	0.000
Z15	0	0	23	0	1.000	NaN	1.000	NaN	0.000
Z16	0	1	21	1	0.913	0	0.955	0	0.045
Z17	0	1	22	0	0.957	0	1.000	NaN	0.043
Z18	0	2	21	0	0.913	0	1.000	NaN	0.087
Z19	0	1	22	0	0.957	0	1.000	NaN	0.043
Sum	13	16	402	6	0.950	NaN	0.985	NaN	0.037

TABLE VII. RESULTS OF LOCI-MDEF [25] OF AM & PM SESSIONS DATA.

AM		PM	
TP	**DSR**	**TP**	**DSR**
2	0.9497	9	0.9314

TABLE VIII. COMPARISON OF 3 DENSITY-BASED OD METHODS.

Methods	**AM DSR**	**PM DSR**	**Average DSR**	**AM TP**	**PM TP**
ARD	0.99	0.95	0.97	2	13
LOCI [11]	0.95	0.93	0.94	2	9
LOF (supervised) [6]	0.95	0.91	0.93	2	12

Conclusion

This paper proposed an ARD OD method for large-scale traffic data, especially successfully in detecting outliers lying outside of clusters of inliers or embedded in an evenly distributed data set. In comparison to other density-based OD methods like LOCI [25] and LOF [4], this method achieves a higher DSR (an average 96%) in a real world data set. Previous DPMM [3], quanterion method [2], modulo-k [14], kNN [21] OD methods achieved 96.67%, 97.83%, 97.94% and 96.19% DSRs, respectively. Due to the fact that the data set for testing contains too small number of outliers, a simulated evenly distributed data set (as the example used to illustrate the procedure in Section III) is used to test whether this ARD method can detect outliers embedded in a cluster of inliers whose relative density are similar, and the result is ideal. The proposed ARD OD method can detect outliers far away from inliers or outliers embedded in evenly distributed inliers. There is still room for improvement: reducing the computing cost and detecting any outliers inside inliers which are not evenly distributed.

Acknowledgment

This research is supported by HKBU Science Elite Program 2016-2017 and the grants of Hong Kong RGC GRF: 12201814 and HKBU FRG/15-16/002.

References

[1] H.Y.T. Ngan, G.K.H. Pang & N.H.C. Yung, "Automated Fabric Defect Detection – A Review," *Image & Vision Computing*, 29(7), 442-458, 2011.

[2] L-L. Wang, H.Y.T. Ngan, W. Liu & N.H.C. Yung, "Anomaly Detection for Quaternion-valued Traffic Signals," *Proc. IEEE DICTA*, pp. 1-4, 2016.

[3] H.Y.T. Ngan, N.H.C. Yung & A.G.O. Yeh, "Detection of Outliers in Traffic Data based on Dirichlet Process Mixture Model," *IET Intelligent Transportation Systems*, vol. 9, no. 7, pp. 773-781, 2015.

[4] M.X. Ma, H.Y.T. Ngan & W. Liu. "Density-based Outlier Detection by Local Outlier Factor on Large-scale Traffic Data." *IS&T Int'l Sym. Electronic Imaging,* no. 4, pp. 1-4, 2016.

[5] H.-P. Kriegel, M.S. Hubert & A. Zimek, "Angle-based outlier detection in high-dimensional data," *ACM SIGKDD,* pp. 444-452*,* 2008.

[6] M. Last & A. Kandel, "Automated Detection of Outliers in Real-World Data," *Proc 2nd Int'l Conf. Intelligent Technologies* .pp.292--301.2002.

[7] S. Chen, W. Wang & H.V. Zuylen, "A comparison of outlier detection algorithms for its data," *Expert Systems with Applications,* 37(2), pp.1169-1178, 2010.

[8] R. Serfling, "Depth functions in nonparametric multivariate inference," *DIMACS Series in Discrete Mathematics & Theoretical Computer Science*, pp. 1-16, 2004.

[9] R. Serfling, "A Depth Function and a Scale Curve Based on Spatial Quantiles," *Statistics in Industry & Technology: Statistical Data Analysis,* pp. 25-38, 2002.

[10] J. Leng, "A novel subspace outlier detection approach in high dimensional data sets," *Proc. IEEE Int'l Conf. Computer Science & Electrical Engineering, VI-162-VI-165,* 2010.

[11] Z. Zhang & X. Feng, "New Methods for Deviation-Based Outlier Detection in Large Database," *Proc. IEEE Fuzzy Systems & Knowledge Discovery,* vol.1, pp.495-499, 2009.

[12] P. Lam, L. Wang, H.Y.T. Ngan, N.H.C. Yung & A.G.O. Yeh, "Outlier Detection in Large-scale Traffic Data by Naïve Bayes Method and Gaussian Mixture Model Method," *IS&T Int'l Sym. Electronic Imaging,* no. 6, pp. 73-78, 2017.

[13] S. Ramaswamy, R. Rastogi & K. Shim, "Efficient algorithms for mining outliers from large data set," *Proc. ACM SIDMOD*, pp.1–20, 2000.

[14] C.H.M. Wong, H.Y.T. Ngan & N.H.C. Yung, "Modulo-k Clustering based Outlier Detection for Large-scale Traffic Data," *ICITA*, 2016.

[15] S. Upadhyaya & K. Singh, "Classification based outlier detection techniques," *Int'l J. Computer Trends & Technology, 3*(2). pp. 294-298, 2012.

[16] Agrawal, S., & Agrawal, J. "Survey on anomaly detection using data mining techniques". *Procedia Computer Science*, 60(1), pp.708-713.2015.

[17] J. Tang & H.Y.T. Ngan, "Traffic Outlier Detection by Density-based Bounded Local Outlier Factors," *IT in Industry*, vol. 4, no. 1, pp. 6-18, 2016.

[18] P.S. Mann, *Introductory Statistics*, 8th Edition, Wiley, 2013.

[19] V. Barnett & T. Lewis, *Outliers in Statistical Data,* J. Wiley & Sons, 3rd Edition. 1995.

[20] X. Jin, Y. Zhang & J. Hu, "Robust PCA-based abnormal traffic flow pattern isolation and loop detector fault detection," *Tsinghua Science & Technology, 13*(6), pp. 829-835, 2008.

[21] T.T. Dang, H.Y.T. Ngan & W. Liu, "Distance-based k-nearest neighbors outlier detection method in large-scale traffic data," *Proc. IEEE DSP*, pp. 507-510. 2015.

[22] H.Y.T. Ngan, N.H.C. Yung & Y.G.O. Yeh, "A comparative study of outlier detection for large-scale traffic data by one-class SVM and kernel density estimation," *SPIE/IS&T Electronic Imaging*, vol. 9405, pp.94050I-94050I-10, 2015.

[23] M.M. Breunig, H.-P. Kriegel, R.T. Ng, J. Sander, "LOF: Identifying density-based local outliers." *Proc. ACM SIGMOD*, pp. 93–104, 2000.

[24] O. Maimon, A. Kandel & M. Last, "Information-Theoretic Fuzzy Approach to Data Reliability and Data Mining," *Fuzzy Sets and Systems,* vol. 117, no. 2, pp. 183-194, 2001.

[25] S. Papadimitriou, H. Kitagawa, P.B. Gibbons & C. Faloutsos, "LOCI: Fast Outlier Detection Using the Local Correlation Integral." *Proc. 19th IEEE ICDE,* pp. 315 - 326, 2003.

Author Biography

Sophia W.T.T. Liu is studying the B.Sc. (Hons) degree in Applied and Computational Mathematics at Hong Kong Baptist University, China and expected to complete the degree on May, 2018. Her research interests include data analytics, financial studies and machine learning.

Henry Y.T. Ngan received the B.Sc. degree in Mathematics (2001), the M. Phil. degree (2005) and the Ph.D. degree (2008) in Electrical & Electronic Engineering at The University of Hong Kong, China. He is currently an assistant professor of research in Mathematics, Hong Kong Baptist University. He was a conference chair of IS&T Electronic Imaging 2016, 2017. He is an editor of IS&T Journal of Imaging Science & Technology.

Michael K. Ng received his B.Sc. degree (1990) and M.Phil. degree (1992) in Mathematics at The University of Hong Kong and Ph.D. degree (1995) at Chinese University of Hong Kong, China. He is currently a Chair Professor at the Department of Mathematics, Hong Kong Baptist University. His research interests include Bioinformatics, Data Mining, Operations Research and Scientific Computing. He has published and edited 5 books, and published more than 200 journal papers. He is an SIAM fellow.

Steven J. Simske is a Professor in Systems and Mechanical Engineering at Colorado State University. He is an HP Fellow emeritus and a previous Director in HP Labs. As of January 2018, he is the author of more than 400 publications and more than 160 US patents (many more worldwide). He is an IS&T Fellow, and its current (2017-2019) President. Steve is the Steering Committee Chair for the ACM DocEng Symposium, which meets annually. Steve is an honorary professor at the University of Nottingham. Dr. Simske was a member of the World Economic Forum Global Agenda Councils from 2010-2016, including Illicit Trade, Illicit Economy and the Future of Electronics. At CSU, Steve is leading research on analytics, intelligent systems, sensing systems, imaging systems, and Situationally-Aware Collectors of Knowledge (SACKs)--the bridge between analytics and robotics. At HP, he directed teams in research on 3D printing, education, life sciences, sensing, authentication, packaging, analytics, imaging and manufacturing. His book "Meta-Algorithmics" addresses intelligent systems. He is currently co-authoring books on Industrial Inkjet Printing (Wiley), Fundamentals and Applications of Hardcopy Communication (Springer), and Meta-Analytics (Elsevier). He has degrees/Post-Docs in biomedical, electrical and aerospace engineering.