Traffic Outlier Detection by Density-Based Bounded Local Outlier Factors

Jialing Tang and Henry Y.T. Ngan
Department of Mathematics
Hong Kong Baptist University
Kowloon Tong, Hong Kong
14452847@life.hkbu.edu.hk, ytngan@hkbu.edu.hk

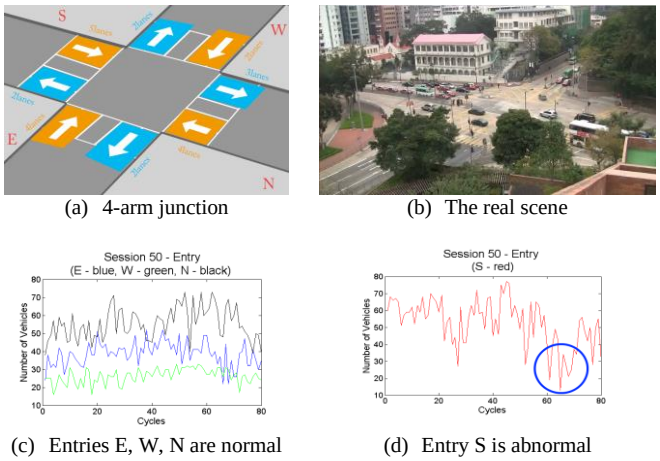
***Abstract*—Outlier detection (OD) is widely used in many fields, such as finance, information and medicine, in cleaning up datasets and keeping the useful information. In a traffic system, it alerts the transport department and drivers with abnormal traffic situations such as congestion and traffic accident. This paper presents a density-based bounded LOF (BLOF) method for large-scale traffic video data in Hong Kong. A dimension reduction by principal component analysis (PCA) was accomplished on the spatial-temporal traffic signals. Previously, a density-based local outlier factor (LOF) method on a two-dimensional (2D) PCA-proceeded spatial plane was performed. In this paper, a three-dimensional (3D) PCA-proceeded spatial space for the classical density-based OD is firstly compared with the results from the 2D counterpart. In our experiments, the classical density-based LOF OD has been applied to the 3D PCA-proceeded data domain, which is new in literature, and compared to the previous 2D domain. The average DSRs has increased about 2% in the PM sessions: 91% (2D) and 93% (3D). Also, comparing the classical density-based LOF and the new BLOF OD methods, the average DSRs in the supervised approach has increased from 94% (LOF) to 96% (BLOF) for the AM sessions and from 93% (LOF) to 95% (BLOF) for the PM sessions.**

Keywords—outlier; density-based; local outlier factor; supervised approach; traffic data

I. INTRODUCTION

Data is available everywhere growing in volume every day and minute in every domain. There are lots of data mining techniques of classification, clustering, association rule mining to deal with every data domain. Meanwhile, OD [1],[2] is the main way to detect and identify outliers (a.k.a. anomalies, errors) in order to clean a dataset, especially useful for massive or big dataset, and help keep useful information for the users. At first, an outlier is usually defined as an observation (or a datum) deviated from other observations. Hence, how much deviation from the others is regarded as an outlier? In most situations, an outlier is a data point generated accidentally, or in other words, by a different mechanism of the majority in data. This research aims to uncover such mechanism and perform an effective OD.

OD has been widely used in fraud detection [3], patient vital sign detection in medical care [4], intrusion detection [5], detecting measure errors [6], inspecting defects in automation [7],[8],[9], etc. For example, manual operators at the traffic control surveillance system (TCSS) are required to monitor the real-time traffic situation and discover any abnormal traffic situations (i.e. outliers). They would report any transportation problems and offer instant responses as soon as possible. A common series of outliers in traffic data are detector faults [10], transmission distortion [10], traffic accident [10], abnormal traffic behaviors [6], etc. There are two types of outliers. One is caused by measurement error. The other one is due to real traffic anomaly. A promising OD method should be able to detect outliers accurately and make less erroneous judgement on inlier data. In the literature, a number of OD methods have been reported, namely learning-based [3], statistical-based [6], proximity-based [3] and ranking-based [11] approaches.



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

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

Fig. 1. (a) The 4-arm junction; (b) the real scene; (c) normal ST signal; (d) abnormal ST signal.

In performance evaluation, the OD results can be affected by the settings. For example, the range of threshold boundaries can be adjusted to judge a testing point is an outlier or not. In this research, OD is performed on a large-scale traffic video data collected by a video camera at a 4-arm junction in Hong Kong (Fig. 1(a)(b)). The video data is further converted as traffic dynamic and traffic flow, as spatial-temporal (ST) signals (Fig. 1(c)(d)). There are 19 ST signals in total with respect to each traffic direction at the junction. Since the ST signals are in high dimensions, a PCA is applied to reduce the

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

TABLE I. DIFFERENT OD METHODS

Approaches	Learning-based	Statistical-based	Proximity-based	Ranking-based
Characteristics	Two-phased fashion in most learnt model	Model "Outlierness"	Measure nearness of objects	Produce scoring function in a unified fashion
Data process	One-class, binary-class, multi-class	Univariate, multivariate data	All types of data	Homogeneous vector and graph data on score or tree structures
Applications	Intrusion detection [5], patient vital sign detection [4]	Traffic problem about managing road network [6], outliers in wireless sensor network [16]	Credit card fraud [10], road accident [17], traffic anomaly detection [18],[19]	E-commerce marketplaces [3], fraud detection [3], network intrusion [3], inspecting errors in automation [7]
Advantages	Fast, apply to various normal classes	Easy to operate	Get satisfied result by many trials	Deal with large number of attributed data
Disadvantages	High dependency on the assumption for the system, high resource consuming	Based on fewer dimensions of data	Time-/space-consuming, not appropriate for large data set	Not fast as other approaches

dimension by projecting the ST signals as 2D/3D data points (i.e. each data point is extracted from first two/three leading coefficients in the PCA expression of each signal). In this paper, a bounded density-based LOF OD will be presented and applied in the PCA-proceeded 2D/3D domains of the traffic data.

The motivation of this research is three-folded: (1) In order to monitor the traffic system, it needs a clean data collection. However, the original collected traffic data often carries some noises and errors, which would be regarded as outliers. Hence, it needs to have a reliable OD initially for any further traffic incident detection; (2) OD has been widely used in many domains, such as credit fraud detection [12], medical care [4], etc., but less in traffic flow analysis; (3) Many OD methods are often applied on data points in the 1D or 2D domains. They are rarely performed in the 3D domain. This paper attempts to evaluate the density-based OD method on the 3D domain.

The objectives of this research are firstly to characterize different data models, such as original ST signals, after PCA-processed data in the 2D/3D domains in order to utilize their features to solve the OD problem. Secondly, it aims to apply a suitable OD method on different data dimensionalities and evaluate its performance. The contributions of this research are as follows.

1. The classical density-based OD method was evaluated in the 3D PCA-proceeded data domain and compared with the results of the previous 2D domain. Herein, its DSRs has been improved from 97.5% (2D) to 98.5% (3D) in the semi-supervised approach and the DSR has increased from 93% (2D) to 93.5% (3D) in the supervised approach.

2. The new BLOF method is proposed. The semi-supervised approach in the new BLOF method obtains similar results and the supervised-approach generates a better result. Specially, the DSR of the supervised-approach has increased 2.5% in the 2D domain, but the semi-supervised approach has decreased 1.5% by comparing original one.

3. In short, the BLOF OD on the 3D domain shows superior DSR and PPV in evaluation compared to the 2D domain. Its supervised approach has an obvious increase of 2% in DSR, but the OD results of the semi-supervised approach has a little decrease in DSR about 2.5%. When the BLOF has a change, the semi-supervised approach performs effectively while the supervised approach offers a poorer result than before (i.e. its DSR has decreased 2%).

Section II presents related work of major OD approaches. Section III describes the result of density based LOF OD of 3D data set and compare it with the result of 2D data set. Section IV proposes a new LOF bounds on OD and then the evaluation results. Section V discusses all the results of DSR and PPV. Conclusion with future improvement is drawn in Section VI.

II. RELATED WORK

Several main OD approaches: learning-based, statistical-based, proximity-based and ranking-based ones are reviewed in this Section. A comparison of these OD approaches are summarized in Table I.

A. Learning based Approach

In the learning approaches [4],[13], the OD is performed by learning a model from a set of labeled data instances (training), and each test instance is then classified into one of the classes. The technique often operates in a two-phase fashion. For the training phase, it would determine a classifier for learning, for which classifies a testing instance in the testing phase. And it has been assumed that the classifier could be learned in the given feature space. In this approach, a neural network (NN) [14] is commonly used for training data to learn the different normal classes. Then, a test instance would be input to the NN. If the instance is classified to the same class of the training data, then it is an inlier; otherwise it is an outlier.

In regard to computational complexity [6], the rule-based training decision tree is faster while SVMs can be more expensive. If a test phase has been carried out in the learnt model, then the rule-based technique could be very fast. Rule-based methods make use of powerful algorithms to distinguish various instances, especially for the multi-class techniques, because the test phase in learning based techniques could be fast [6]. Multi-class classification would be applied for a database with various normal classes. When a meaningful anomaly score is desired for the test instances, it sometimes could become a disadvantage for classification based techniques [6].

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

1) Bayesian Network

Bayesian network [15] is only used in the multi-class setting, which aggregates the per-attribute posterior probabilities for a test instance and uses the aggregated value to assign a class label to the test instance. For the one-class case, the naïve Bayesian network [6] is more suitable. It estimates the posterior probability of observing a class label. The class label with largest posterior is chosen as the predicted class for the given test instance.

2) Support Vector Machines

Support Vector Machines (SVMs) [3], including one class SVM [4],[13],[20] and multi-class SVM [5], are a standard classification technique popularly used in machine learning. SVMs is commonly employed for anomaly detection in audio signal data [21], novelty detection in power generation plants [22], intrusion detection [5], patient vital-sign detection [4], and temporal sequences detection [23]. The task of SVMs is to determine the optimal hyperplane which separates a set of training points into two categories. In order to solve an optimal problem, a Lagrangian technique and transformation to a dual problem would be usually constructed.

3) Rule-based Technique

Rule based anomaly detection techniques [11] learn rules that capture the normal behavior of a system, which applies to a one-class or multi-class classifier. For the one-class classifier, an association rule mining helps to detect anomalies. To ensure the rules correspond to strong patterns, a supporting threshold has been used to prune out rules with support. For the multi-class one, it initially would learn rules that have an associated confidence value, from the training data by using the rule learning algorithm. It is required to find the best rule for each test instance. The inverse of the confidence is as the anomaly score of the test instance.

B. Statistical-based Approach

A statistical method [3] exploits a statistical model to the given data and then applies a statistical inference test to determine if an unseen instance belongs to this model or not. They assume a normal instance (i.e. inlier) occurs in high probability regions, while an anomaly (i.e. outlier) lies in the low probability. It would have parametric and non-parametric tests in this region. For a parametric test, the normal data would fit a parametric distribution and probability density. The parameters are estimated from the given data. Usually a statistical hypothesis would help to judge whether the point is an outlier. The anomaly score would be the inverse of the probability density function. Some common models include Gaussian model [14], univariate Gaussian model [16] regression model [24] and mixture of parametric distributions [14]. For a non-parametric test, the model structure is not defined by a priori from the given data. It would have fewer assumptions compared to the parametric test. Histogram and kernel function methods [25] are two common seen examples in the non-parametric test.

Computational complexity of the statistical approach often depends on the nature of model chosen. Confidence interval could offer additional information in decision making. However, it often ties to known distributions and it is hard to deal with high dimensional real data sets.

1) Gaussian Model Based Method

The data is generated from Gaussian distribution. It could use maximum likelihood estimated to the parameters and the distance of a data instance to the estimated mean is the anomaly score. The statistical tests [6] have been used to detect the anomalies, such as box plot rule, Grubb's test, student's $t-test$ and χ^2 test. Box plot rule has been used in the medical domain data and turbine rotors data. In the box plot, the observations beyond the limits are regarded as an anomaly. The limits often would be distance of 1.5 times Inter Quartile Range, which is the difference between upper quartile and lower quartile. A Grubb's test could apply to the distance to detect the anomalies in a univariate or multivariate data set.

2) Regression Model Based Method

In this model [24], the residual could be used as the anomaly score. But it needs to find a suitable regression model to fit the data, and then get the residuals for each test instances. Akaike Information Content (AIC) would help to detect outliers when the model has been fitted. In order to improve the accuracy of the regression model, robust regression [23] is introduced to avoid the anomalies affecting the parameters of model.

C. Proximity-based Approach

Proximity-based approach [3] mainly measures the nearness of objects in terms of distance and density. It considers how to determine a proper distance or density so as to avoid high computational time and great complexity in spatial calculation. In fact, this approach requires many prerequisite parameters, which are desired for a huge amount of trial-and-errors to attain a desired result. When using either a distance or local density measure for the data, the proximity-based methods suffer from the curse of dimensionality. These methods are time- and space-consuming and consequently are not appropriate for large data sets.

1) Distance-based Method

The distance-based method [26] is to determine whether the number of the points, whose distance less than a specified distance from a point, is more than the number of data times the value of the fraction of objects minus one. The new definition [26] has been updated as the distance of a point p from its k^{th} nearest neighbor. The values of k and m are given. A point is an outlier when no more than $m-1$ other points in the data set have a higher value for D^k than p, which means that the top m points having the maximum D^k values could be considered as outliers. The weaknesses of this method are: (a) It needs to calculate the distance between all samples, hence lower efficiency. (b) It is hard to deal with high dimensional space.

The anomaly score of a data instance is defined as its distance to its k^{th} nearest neighbor in a given data set. A threshold on the anomaly score is applied to judge whether a test instance is anomalous. The largest anomaly scores also would be regarded as anomalies. It could count the number of nearest neighbors which are not more than d distance apart from the given data instances.

2) Density-based Method

In the density-based method [27], a parameter k is needed to be chosen by a heuristic method by a LOF [28] computation.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

Afterward, k is utilized to compute the density in the neighborhood of an object, which is a measure of the volume to determine the LOF of a data-point. The outlier score of an object is the reciprocal of the density in the object's neighborhood. A point of relative density is the outlier score for that point. It requires more complexity when the right k is not obvious. It could not do well when the number of data attributes increase. In general, the density-based method is an effective measure of outliers but the shortcomings are time- and space-consuming, and often not appropriate for large data sets, due to the curse of dimensionality. A fusion of distance-based and density-based method has been proposed in [7].

3) Deviation-based Method

This method [29] aims at removing outliers so that the variance of the data set could be minimized. A smoothing factor (SF) [29] helps judge which point to be removed that could make the variance minimum. If the SF values among two data sets are equal, the smaller set is often taken out. It is similar to the classical statistical approach, but it does not need to choose a specified distribution in the initial stage. Therefore, it could be applied to any data type as a global method.

D. Ranking-based Approach

Traditional outlier ranking or rating techniques focus on image [7], vectored/ graph structured data [11] and tree structure data [18],[30]. However, many data nowadays is required to find an approach to deal with different data types in a unified fashion. Previously, some OD techniques relied on a binary decision tree structure [18],[30]. Among them, a representative method is called GOutRank [11]. It produces scoring functions based on the selected subgraphs (as a relevant graph context of an outlier) and subspaces (as a relevant attribute set in which an outlier is deviating). These complex outliers must be detected by a combination of information on relations between products and a large set of attributes.

1) GOutRank

GOutRank is utilized in heterogeneous data of e-commerce marketplaces. People need to deal with the data focusing on the relations between co-purchased products and a lot of attributes. GOutRank detects outliers in attributed graphs [11]. It computes a subspace clustering at first by using different graph clustering algorithms. Each subspace contains connected subgraph with high attribute similarity. Each object is a graph vertex and the line connected between vertexes is an edge. At the same time, each object is represented as a vector in a multi-dimensional data space. An outlier ranking is sorted in an ascending order by a scoring function. An outlier would have low score and regular objects would have high scores. For objects in multiple subspace clusters, the score should depend on the occurrence of objects in different subspace clusters. For large numbers of given attributes, GOutRank gives a high OD accuracy. But, it is not fast by comparing with other approaches.

In short, it is found that little research was investigated in the traffic problems. Therefore, in this paper, it focuses OD on traffic data. However, the proximity-based approaches have many limitations and they were used widely in many domains by comparing with other approaches. While the density-based LOF approach has a high accuracy among all methodology of proximity-based approaches, then it was applied in this paper. Mathematically, 3D data-points implicitly carry more information than 2D data points, so 3D data points are applied in the developed OD method at this paper in order to get a higher accuracy.

III. Classical Density-based LOF OD Method

The traffic dataset evaluated in this paper was collected by video camera from a four-arm junction located the one of Hong Kong's busiest districts. The data [17] has two sessions: one from 07:00~10:00 (the AM session) and another one from 17:00~20:00 (the PM session). The data set covered a total of 31 days, but only 23 weekdays (from Monday to Friday per week) were selected for analysis due to different traffic situations between weekdays and weekend. There are 312,333 vehicles in the AM session, 251,694 vehicles in the PM sessions, so 764,027 vehicles in total. The video data was further converted as statistics of traffic flow dynamic as ST signals which represents the volume of Entry, Exit, and Entry direction distribution traffic flows in each session. Actually, an Entry or Exit per session would be interpreted as a signal. Due to four entry signals for this section, they are labeled as $\{z1, z2, z3, z4\}$. Four exit signals for the arm of E, S, W and N can be $\{z5, z6, z7, z8\}$. There are three entry direction distributions in this traffic junction for left (l), right (r) and straight ahead (s), so these signals are labeled as $\{z9, z10, z11, z12, z13, z14, z15, z16, z17, z18, z19\}$ which represent $E_l, E_r, E_s, S_l, S_r, S_s, W_l, W_s, N_l, N_r, N_s$.

Due to the large dimensions of ST signal, the original data has been preceded by a PCA to reduce its dimension. Ma et al. [28] performed a preliminary study of a density-based LOF kNN OD method by using these 2D PCA-proceeded data points. The semi-supervised approach achieved an average 97.5% DSR and the supervised approach obtained about 93% DSR. In this paper, a higher dimension of 3D PCA-proceeded data points will be firstly investigated for the classical density-based LOF OD method.

In general, there are three main data manipulation techniques for OD namely as unsupervised, semi-supervised and supervised techniques. In this research, semi-supervised and supervised techniques have been used. The semi-supervised approach would exploit some experienced outlier data for training while the supervised approach would use the normal inliers as the sole training data.

A. Review of Classical Density-based LOF OD

The underlying concept of the classical density-based OD method is the density around an outlier object is significantly different from the density around its neighbors. Therefore, using the relative density of an object against its neighbors could be the indicator of the degree of the object being outliers, LOF, which depends on how isolated the object is with respect to the surrounding neighborhoods. If the local reachability density of the object is lower, and the local reachability density of the kNN of the object becomes higher, the LOF will be higher.

A brief procedure of density-based LOF OD is as following:

1) Calculate the $k-$distance of a specified point;

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

2) Calculate local reachability distance;
3) Calculate local reachability density;
4) Get the local outlier factor (LOF);
5) Separate the outlier and inlier;
6) Evaluate the performance.

Mathematical expressions are given as follows. At first, it needs to get the k^{th} small distance [28] to a specified point. Define x is the target point, the $k-$distance neighborhood of x is $N_k(x)$ and q is the point of $D(x)$.

$$N_k(x) = \{q \in D(x) | d(x,q) \leq k - distance(x)\} \quad (1)$$

Then, it needs to get the local reachability distance [28], it should be calculated like this:

$$reach - disp_k(x,o) = \max\{k - distance(o), d(x,o)\} \quad (2)$$

Local reachability density is defined as the inverse of reachability distance –based on the k-nearest neighborhood [28],

$$lrd_k(x) = \frac{1}{\frac{\sum_{o \in N_k(x)} reach\_disp_k(o)}{|N_k(x)|}} \quad (3)$$

where $|N_k(x)|$ is the cardinality of $N_k(x)$.

The LOF [28] is computed as following:

$$lof_k(x) = \frac{\sum_{o \in N_k(x)} \frac{lrd_k(o)}{lrd_k(x)}}{|N_k(x)|} \quad (4)$$

Finally, a performance evaluation is carried out by the standard metrics: true positive (tp) as the outlier is regarded outlier; false positive (fp) as the normal is regarded as outlier; true negative (tn) as the normal (inlier) is regarded as normal; false negative (fn) as the outlier is regarded as normal; detection success rate (DSR) as the accuracy. Also, true positive rate (TPR), false positive rate (FPR), positive predictive rate (PPV), negative predictive rate (NPV). Their formula is listed as below.

$$\text{DSR} = \frac{tp+tn}{tp+fp+tn+fn} \quad (5)$$

$$\text{TPR} = \frac{tp}{tp+fp} \quad (6)$$

$$\text{FPR} = \frac{tn}{tp+fp} \quad (7)$$

$$\text{PPV} = \frac{tp}{tp+fp} \quad (8)$$

$$\text{NPV} = \frac{tn}{tn+fn} \quad (9)$$

B. Experimental Results on the 2D/3D data

The large-scale traffic data would be evaluated as the 2D/3D domains after the PCA process in the semi-supervised and supervised approaches. Then, the results between the 2D and 3D domains will be compared.

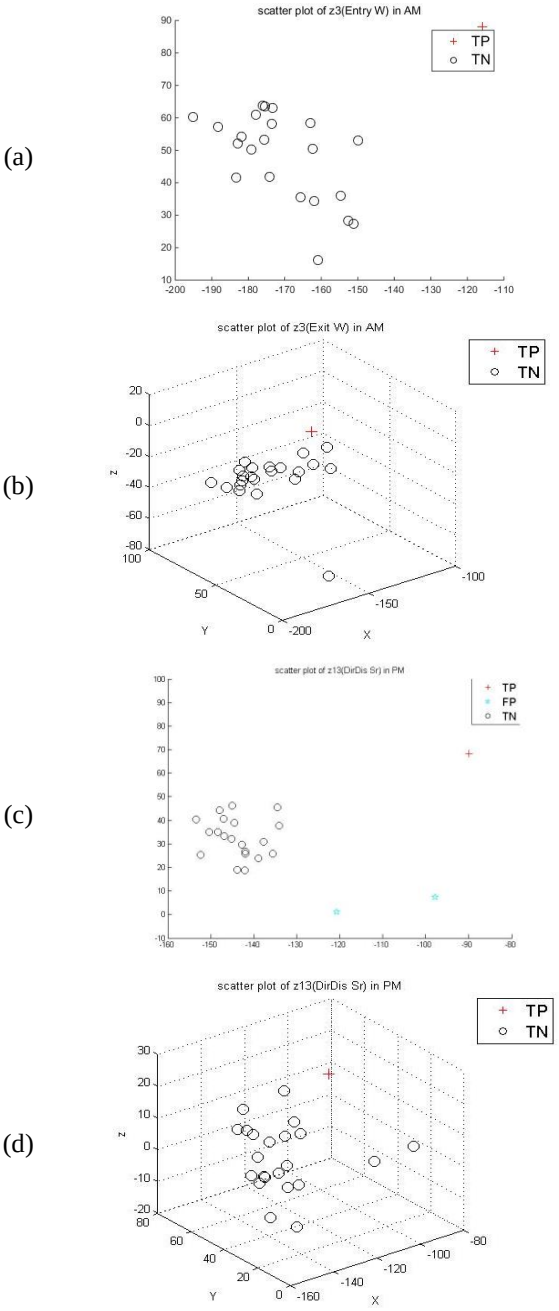


Fig. 2. Scatter plots of the OD results of (a) $z3$ in 2D (AM), (b) $z3$ in 3D (AM), (c) $z13$ in 2D (PM), (d) $z13$ in 3D (PM).

1) Semi-supervised Approach

In this approach, if the $lof_k(x) > \max\{lof_i(x)\}$ of a datapoint k, then the datapoint is k, otherwise it is an inlier. First, the $z3$ in the AM sessions and the $z13$ in the PM sessions are used for training and the remaining ones are employed for testing. Since both $z3$ and $z13$ have perfect filter, the classical LOF method can separate inliers and outliers effectively. Hence, there is no any errors and both DPRs are 100%. The scatter plots

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

of the OD results of $z3$ (AM) and $z13$ (PM) are shown in Fig. 2. The red "cross" represents an outlier. From the sub-figures, both $z3$ (AM), $z13$ (PM) have one outlier. For z13 (PM), it has two 2 FP cases in Fig. 2(c), but the OD results in the 3D domain perform better with one TP and no FP.

To apply the 3D data points in semi-supervised approach, the result is shown in the following table, and it was compared with the result of the 2D data points. By testing the criteria of LOF bounds, the range of [1.6, 5.0] was used in the AM sessions. Then, the range of [1.6, 3.0] is obtained from the ROC curves from the training set. Herein, the best bound for the AM sessions is 3.7 while the best bound for the PM sessions is 2.9 in the 2D domain. The optimal bound is 3.3 for the AM sessions and 2.3 for the PM sessions in the 3D data points.

The filter should be obtained from the training set by different values, and then the optimal filter is determined by the DSR and PPV results. In order to judge whether the filter of which approach is good, the training sets of $z3$ and $z13$ were plotted for a Receiver operating characteristic (ROC) [32] curve. Fig. 3 illustrates the ROC curve with the TPR (i.e. y-axis) against the FPR (i.e. x-axis) that both two samples are performed well. The blue lines in Fig. 3(a)(b) indicate that our training samples ($z3$ in AM, $z13$ in PM) have always 100% TPR along the change of FPR from 0 to 100 in the x-axis.

The experimental results of the semi-supervised classical LOF method in the AM sessions for the 2D and 3D data points are given in Tables 2 and 3, respectively. They coincidently show the same experimental results. Meanwhile, it is found that several cases only have 96% DSRs due to some false negative points (i.e. some outliers misclassified as normal points).

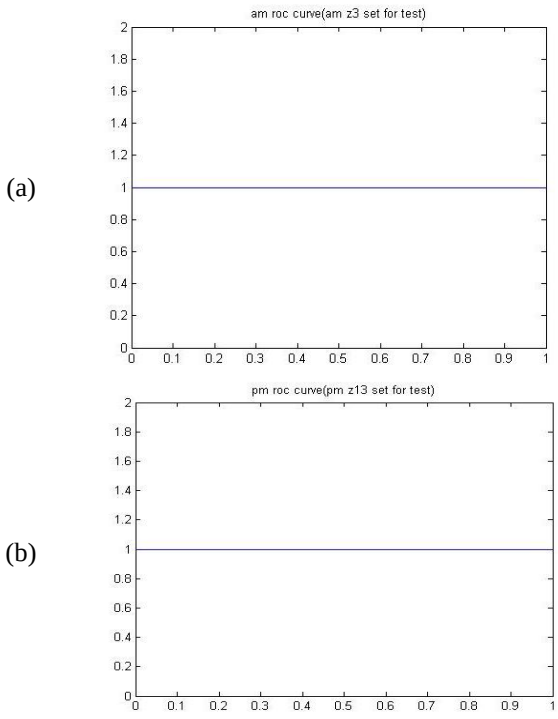


Fig. 3. ROC curves of (a) z3 (AM) and (b) z13 (PM).

TABLE II. CLASSICAL LOF ON 2D PCA-PROCEEDED DATA POINTS (AM, SEMI-SUPERVISED)

	TP	FP	TN	FN	DSR	DSR(%)	PPV	NPV
z1	0	0	23	0	23/23	100%	NaN	100%
z2	0	0	22	1	22/23	96%	NaN	96%
z4	0	0	22	1	22/23	96%	NaN	96%
z5	0	0	23	0	23/23	100%	NaN	100%
z6	0	0	23	0	23/23	100%	NaN	100%
z7	0	0	22	1	22/23	96%	NaN	96%
z8	0	0	23	0	23/23	100%	NaN	100%
z9	0	0	23	0	23/23	100%	NaN	100%
z10	0	0	23	0	23/23	100%	NaN	100%
z11	0	0	23	0	23/23	100%	NaN	100%
z12	0	1	22	0	22/23	96%	0%	100%
z13	0	0	23	0	23/23	100%	NaN	100%
z14	0	0	23	0	23/23	100%	NaN	100%
z15	0	0	23	0	23/23	100%	NaN	100%
z16	1	0	22	0	23/23	100%	100%	100%
z17	0	0	23	0	23/23	100%	NaN	100%
z18	0	0	22	1	22/23	96%	NaN	96%
z19	0	0	23	0	23/23	100%	NaN	100%
Total	1	1	408	4	409/414	99%	50%	99%

TABLE III. CLASSICAL LOF ON 3D PCA-PROCEEDED DATA POINTS (AM, SEMI-SUPERVISED)

	TP	FP	TN	FN	DSR	DSR(%)	PPV	NPV
z1	0	0	23	0	23/23	100%	NaN	100%
z2	0	0	22	1	22/23	96%	NaN	96%
z4	0	0	22	1	22/23	96%	NaN	96%
z5	0	0	23	0	23/23	100%	NaN	100%
z6	0	0	23	0	23/23	100%	NaN	100%
z7	0	0	22	1	22/23	96%	NaN	96%
z8	0	0	23	0	23/23	100%	NaN	100%
z9	0	0	23	0	23/23	100%	NaN	100%
z10	0	0	23	0	23/23	100%	NaN	100%
z11	0	0	23	0	23/23	100%	NaN	100%
z12	0	1	22	0	22/23	96%	0%	100%
z13	0	0	23	0	23/23	100%	NaN	100%
z14	0	0	23	0	23/23	100%	NaN	100%
z15	0	0	23	0	23/23	100%	NaN	100%
z16	1	0	22	0	23/23	100%	100%	100%
z17	0	0	23	0	23/23	100%	NaN	100%
z18	0	0	22	1	22/23	96%	NaN	96%
z19	0	0	23	0	23/23	100%	NaN	100%
Total	1	1	408	4	409/414	99%	50%	99%

Tables 4 and 5 list the experimental results of the semi-supervised classical LOF method in the PM sessions for the 2D and 3D data points, respectively. The DSRs on 2D and 3D data points are 96% and 98%, respectively. The PPVs in the 2D data points is only 30%, but that in the 3D data points can be up to 100%. Some directions perform better, the PPV percentage has increased to 100% (Table 4) from 33% (Table 3) in z13. In $z13$, its DSR has risen up to 100% from 91%. In these two tables, only the DSR in z2 has decreased from 2D to 3D data points. Two outliers were picked incorrectly as inliers at z2 in the 3D domain. These two outliers may be closer to other normal points in the spatial distributions no matter in the 2D/3D domain.

In the 2D domain, there are still some traffic directions such as z2, z6, z18 which offer poor OD results.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

TABLE IV. CLASSICAL LOF ON 2D PCA-PROCEEDED DATA POINTS (PM, SEMI-SUPERVISED)

	TP	FP	TN	FN	DSR	DSR(%)	PPV	NPV
z1	0	0	23	0	23/23	100%	NaN	100%
z2	4	0	18	1	22/23	96%	100%	95%
z3	0	1	21	1	21/23	91%	0%	95%
z4	0	1	22	0	22/23	96%	0%	100%
z5	0	1	21	1	21/23	91%	0%	95%
z6	1	0	22	0	23/23	100%	100%	100%
z7	0	0	22	1	22/23	96%	NaN	96%
z9	0	0	23	0	23/23	100%	NaN	100%
z10	0	1	22	0	22/23	96%	0%	100%
z11	0	0	23	0	23/23	100%	NaN	100%
z12	0	0	22	1	22/23	96%	NaN	96%
z13	1	2	20	0	21/23	91%	33%	100%
z14	3	0	19	1	22/23	96%	100%	95%
z15	0	0	23	0	23/23	100%	NaN	100%
z16	0	1	21	1	21/23	91%	0%	95%
z17	0	0	23	0	23/23	100%	NaN	100%
z18	0	2	21	0	21/23	91%	0%	100%
z19	0	1	22	0	22/23	96%	0%	100%
Total	9	10	388	7	397/414	96%	30%	98%

TABLE V. CLASSICAL LOF ON 2D PCA-PROCEEDED DATA POINTS (PM, SEMI-SUPERVISED)

	TP	FP	TN	FN	DSR	DSR(%)	PPV	NPV
z1	0	0	23	0	23/23	100%	NaN	100%
z2	4	0	18	1	22/23	96%	100%	95%
z3	0	1	21	1	21/23	91%	0%	95%
z4	0	1	22	0	22/23	96%	0%	100%
z5	0	1	21	1	21/23	91%	0%	95%
z6	1	0	22	0	23/23	100%	100%	100%
z7	0	0	22	1	22/23	96%	NaN	96%
z9	0	0	23	0	23/23	100%	NaN	100%
z10	0	1	22	0	22/23	96%	0%	100%
z11	0	0	23	0	23/23	100%	NaN	100%
z12	0	0	22	1	22/23	96%	NaN	96%
z13	1	2	20	0	21/23	91%	33%	100%
z14	3	0	19	1	22/23	96%	100%	95%
z15	0	0	23	0	23/23	100%	NaN	100%
z16	0	1	21	1	21/23	91%	0%	95%
z17	0	0	23	0	23/23	100%	NaN	100%
z18	0	2	21	0	21/23	91%	0%	100%
z19	0	1	22	0	22/23	96%	0%	100%
Total	9	10	388	7	397/414	96%	30%	98%

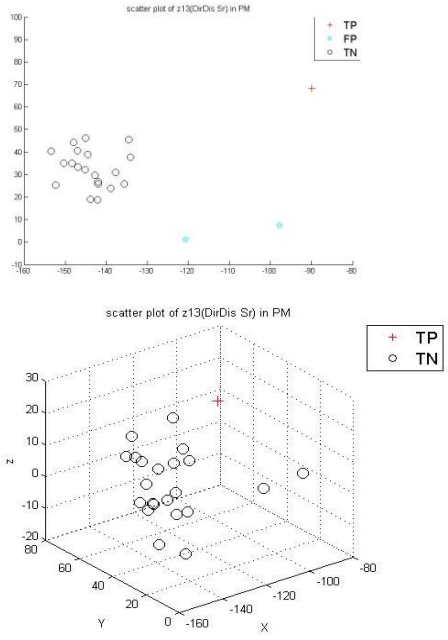
Fig. 4 demonstrates the scatter plot of the OD results in the 2D and 3D domains of $z13$ (Fig. 4(a)) and $z18$ (Fig. 4(b)) (PM) and their respective TP and FN points. It shows that the 3D domain is more effective for OD than the 2D domain. No more FPs in the 3D domain for $z13$ and $z18$.

2) Supervised-approach

In the supervised approach, 23 datapoints in one direction is divided into 12 ones as a training set, the other 11 ones are for testing.

The supervised classical LOF OD results of the 2D and 3D domains for the AM sessions are presented in Tables 6 and 7, respectively. The average DSR and PPV have decreased in the 3D dataset compared to the 2D dataset. Herein, the average DSR decreased 1% (i.e. 95% in 2D to 94% in 3D) and PPV decreased

(a)



3D	x	y	z
TP	-89.9	68.3	11.9

(b)

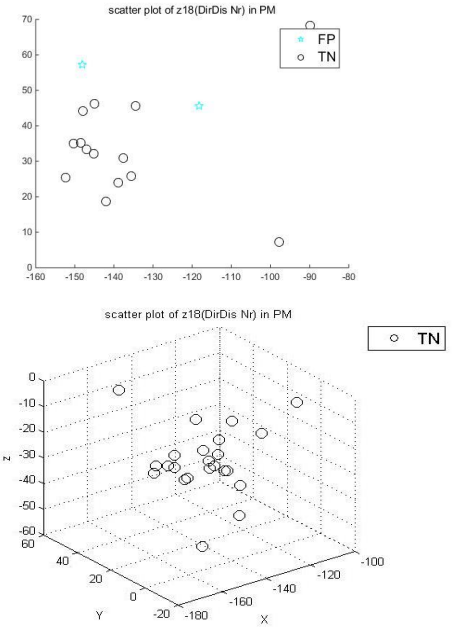


Fig. 4. Scatter plots of the OD results in 2D and 3D of (a) $z13$ and (b) $z18$ (all PM) and the corresponding (x,y,z) of the TP and FN cases.

16% (33% in 2D to 17% in 3D). The OD performances in some directions, such as $z3, z4$ and $z16$ in the 3D domain are worse than those directions in the 2D domain. Fig. 5 shows the scatter plot of the OD results of in the 2D and 3D domains for z11, z16 (the AM sessions) and their respective (x,y,z)-coordinates of TP and FN cases. The OD result of the 3D domain in z11 performs better than the 2D domain but that of 3D one in z16 shows poorer result than the 2D one.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

TABLE VI. CLASSICAL LOF ON 2D PCA-PROCEEDED DATA POINTS (AM, SUPERVISED)

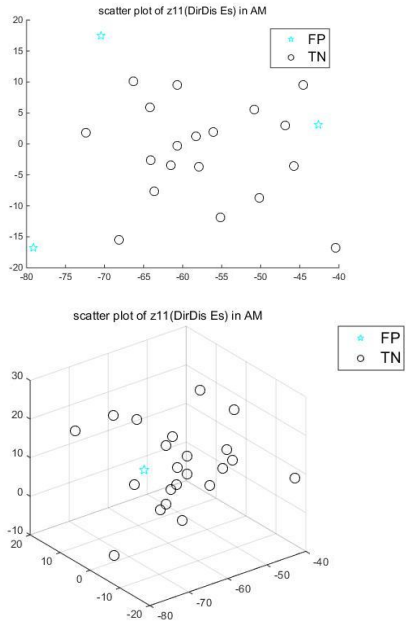
	TP	FP	TN	FN	DSR	DSR(%)	PPV	NPV
z1	0	0	11	0	11/11	100%	NaN	100%
z2	0	0	10	1	10/11	91%	NaN	91%
z3	1	0	10	0	11/11	100%	100%	100%
z4	0	0	10	1	10/11	91%	NaN	91%
z5	0	0	11	0	11/11	100%	NaN	100%
z6	0	0	11	0	11/11	100%	NaN	100%
z7	0	0	10	1	10/11	91%	NaN	91%
z8	0	0	11	0	11/11	100%	NaN	100%
z9	0	1	10	0	10/11	91%	0%	100%
z10	0	1	10	0	10/11	91%	0%	100%
z11	0	3	8	0	8/11	73%	0%	100%
z12	0	0	11	0	11/11	100%	NaN	100%
z13	0	0	11	0	11/11	100%	NaN	100%
z14	0	0	11	0	11/11	100%	NaN	100%
z15	0	2	9	0	9/11	82%	0%	100%
z16	1	0	10	0	11/11	100%	100%	100%
z17	0	0	11	0	11/11	100%	NaN	100%
z18	0	0	10	1	10/11	91%	NaN	91%
z19	0	0	11	0	11/11	100%	NaN	100%
Total	2	7	196	4	198/209	95%	33%	98%

TABLE VII. CLASSICAL LOF ON 3D PCA-PROCEEDED DATA POINTS (AM, SUPERVISED)

	TP	FP	TN	FN	DSR	DSR(%)	PPV	NPV
z1	0	0	11	0	11/11	100%	NaN	100%
z2	0	0	10	1	10/11	91%	NaN	91%
z3	1	1	9	0	10/11	91%	50%	100%
z4	0	1	9	1	9/11	82%	0%	90%
z5	0	0	11	0	11/11	100%	NaN	100%
z6	0	0	11	0	11/11	100%	NaN	100%
z7	0	0	10	1	10/11	91%	NaN	91%
z8	0	0	11	0	11/11	100%	NaN	100%
z9	0	0	11	0	11/11	100%	NaN	100%
z10	0	0	11	0	11/11	100%	NaN	100%
z11	0	1	10	0	10/11	91%	0%	100%
z12	0	0	11	0	11/11	100%	NaN	100%
z13	0	0	11	0	11/11	100%	NaN	100%
z14	0	0	11	0	11/11	100%	NaN	100%
z15	0	3	8	0	8/11	73%	0%	100%
z16	1	1	9	0	10/11	91%	50%	100%
z17	0	0	11	0	11/11	100%	NaN	100%
z18	0	0	10	1	10/11	91%	NaN	91%
z19	0	2	9	0	9/11	82%	0%	100%
Total	2	9	194	4	196/209	94%	17%	98%

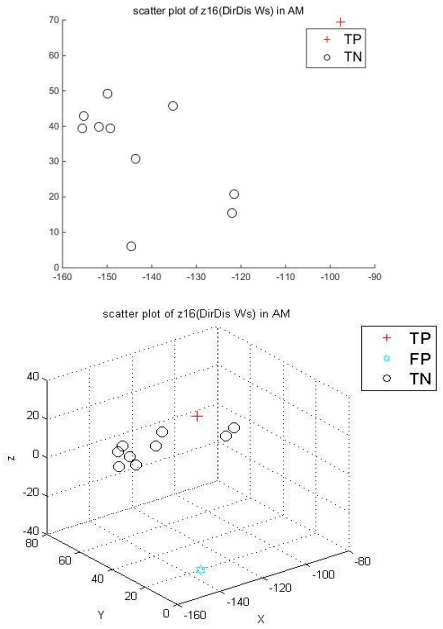
Table 8 and Table 9 deliver the supervised classical LOF OD results of the 2D and 3D domains for the PM sessions. The average DSR has improved from 91% in the 2D domain to 93% in the 3D domain. Also, the average PPV has increased from 44% (2D) to 50% (3D). Among all traffic directions, the OD results in z18 is not improved and fixed at 73% which have 3 false positive points (i.e. originally inliers are incorrectly determined as outliers).

(a)



3D	x	y	z
FN	-79.1	-16.8	23.3

(b)



	x	y	z
TP	-97.7	69.4	4.5
FN	-144.6	6.1	-29.9

Fig. 5. Scatter plots of the OD results in 2D and 3D of (a) $z11$ and (c) $z16$ (all AM) and corresponding (x,y,z) of the TP, FN and TN cases.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

TABLE VIII. CLASSICAL LOF ON 2D PCA-PROCEEDED DATA POINTS (PM, SUPERVISED).

	TP	FP	TN	FN	DSR	DSR(%)	PPV	NPV
z1	0	0	11	0	11/11	100%	NaN	100%
z2	3	0	6	2	9/11	82%	100%	75%
z3	0	0	10	1	10/11	91%	NaN	91%
z4	0	1	10	0	10/11	91%	0%	100%
z5	1	1	9	0	10/11	91%	50%	100%
z6	1	0	10	0	11/11	100%	100%	100%
z7	0	0	10	1	10/11	91%	NaN	91%
z8	3	0	8	0	11/11	100%	100%	100%
z9	0	2	9	0	9/11	82%	0%	100%
z10	0	0	11	0	11/11	100%	NaN	100%
z11	0	0	11	0	11/11	100%	NaN	100%
z12	0	0	10	1	10/11	91%	NaN	91%
z13	1	2	8	0	9/11	82%	33%	100%
z14	3	0	7	1	10/11	91%	100%	88%
z15	0	0	11	0	11/11	100%	NaN	100%
z16	0	0	10	1	10/11	91%	NaN	91%
z17	0	1	10	0	10/11	91%	0%	100%
z18	0	3	8	0	8/11	73%	0%	100%
z19	0	1	10	0	10/11	91%	0%	100%
Total	12	11	179	7	191/209	91%	44%	96%

TABLE IX. CLASSICAL LOF ON 3D PCA-PROCEEDED DATA POINTS (PM, SUPERVISED).

	TP	FP	TN	FN	DSR	DSR(%)	PPV	NPV
z1	0	0	11	0	11/11	100%	NaN	100%
z2	3	0	6	2	9/11	82%	100%	75%
z3	0	0	10	1	10/11	91%	NaN	91%
z4	0	1	10	0	10/11	91%	0%	100%
z5	1	1	9	0	10/11	91%	50%	100%
z6	1	0	10	0	11/11	100%	100%	100%
z7	0	0	10	1	10/11	91%	NaN	91%
z8	3	0	8	0	11/11	100%	100%	100%
z9	0	0	11	0	11/11	100%	NaN	100%
z10	0	0	11	0	11/11	100%	NaN	100%
z11	0	0	11	0	11/11	100%	NaN	100%
z12	0	0	10	1	10/11	91%	NaN	91%
z13	1	1	9	0	10/11	91%	50%	100%
z14	3	0	7	1	10/11	91%	100%	88%
z15	0	0	11	0	11/11	100%	NaN	100%
z16	0	0	10	1	10/11	91%	NaN	91%
z17	0	1	10	0	10/11	91%	0%	100%
z18	0	3	8	0	8/11	73%	0%	100%
z19	0	1	10	0	10/11	91%	0%	100%
Total	12	8	182	7	194/209	93%	50%	96%

IV. NEW BOUNDED LOF

A. Bounded LOF

In the original semi-supervised approach of the density-method from Ma et al. [28], the criteria of LOF are tested from the specific interval (i.e. a set of upper and lower bounds) in order to separate inliers and outliers perfectly. However, it was manually performed many times. If there is a wide set of bounds, then the computational work would become demanding. Therefore, there is a need to compute suitable bounds of LOF so that MATLAB could run it by itself. In the supervised approach, the criteria were defined based on this formula:

$$Range = lof_{max} + (lof_{max} - lof_{min}) \times 0.5 \qquad (10)$$

which is defined by Ma et al. that means LOF bounds can be changed in order to find more accurate bounds conveniently.

B. Review of LOF Bounds

A theorem [31] of the bounds of LOF is as follows. The p is an object from the database D, and $1 \leq K \leq |D|$.

Then it would have the case like that:

$$\frac{direct_{min}(p)}{indirect_{max}(p)} \leq LOF(p) \leq \frac{direct_{max}(p)}{indirect_{min}(p)} \qquad (11)$$

where the minimum and maximum reachability distances between p and a k-nearest neighbor of p are defined as

$$direct_{min}(p) = \min\{reach - dist(p,q) | \, q \in N_k(p)\} \qquad (12)$$

$$direct_{max}(p) = \max\{reach - dist(p,q) | \, q \in N_k(p)\} \qquad (13)$$

The minimum and maximum reachability distances between q and a k-nearest neighbor of q is defined as

$$indirect_{min}(p) = \min\{reach - dist(q,o) | \, q \in N_k(p) \; and \; o \in N_k(q)\} \qquad (14)$$

$$indirect_{max}(p) = \max\{reach - dist(q,o) | \, q \in N_k(p) \; and \; o \in N_k(q)\} \qquad (15)$$

where p is the target point; q is the k-nearest neighbor of p, and o is the k-nearest neighbor of q.

C. Experimental Results on the 2D/3D Data

After obtaining the new bounds (NBs), the BLOF OD method offered poor results in some specific traffic directions in both the semi-supervised and supervised approaches. In the semi-supervised approach, the upper bounds are attained from the specific interval from the ROC curves by testing for many times. When 3D data is used, the bound in the AM sessions has changed into 3.7 from 3.3 (2D) and the bound in the PM sessions has been into 2.9 from 2.3 (2D). In the supervised approach, the bounds are obtained by specific formula in Eq. (10). For the 2D data, the bounds are 1.1878 and 1.1272 for the AM and PM sessions, respectively. For the 3D data, the bounds are 1.1394 and 1.1954 for the AM and PM sessions, respectively. However, if the only upper bounds are chosen to be applied in these approaches, the result of the semi-supervised BLOF method is similar to the semi-supervised classical LOF method, but the supervised BLOF method has better results than the supervised LOF method. The details are as follows.

1) Semi-Supervised Approach

In the semi-supervised approach, the BLOF method gives an average of 97% DSR in Table 10 and the classical LOF method on the AM sessions offers an average 99% DSR in Table 2. Especially, the DSR of z13 and z17 are very low in the BLOF

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

TABLE X. BLOF ON 2D PCA-PROCEEDED DATA POINTS (AM, SEMI-SUPERVISED)

	TP	FP	TN	FN	DSR	DSR(%)	PPV	NPV
z1	0	0	23	0	1	100%	NaN	100%
z2	0	0	22	1	22/23	96%	NaN	96%
z4	0	0	22	1	22/23	96%	NaN	96%
z5	0	0	23	0	23/23	100%	NaN	100%
z6	0	0	23	0	23/23	100%	NaN	100%
z7	0	0	22	1	22/23	96%	NaN	96%
z8	0	0	23	0	23/23	100%	NaN	100%
z9	0	0	23	0	23/23	100%	NaN	100%
z10	0	0	23	0	23/23	100%	NaN	100%
z11	0	1	22	0	22/23	96%	0%	100%
z12	0	0	23	0	23/23	100%	NaN	100%
z13	0	4	19	0	19/23	83%	0%	100%
z14	0	0	23	0	23/23	100%	NaN	100%
z15	0	0	23	0	23/23	100%	NaN	100%
z16	1	0	22	0	23/23	100%	100%	100%
z17	0	2	21	0	21/23	91%	0%	100%
z18	0	0	22	1	22/23	96%	NaN	96%
z19	0	0	23	0	23/23	100%	NaN	100%
Total	1	7	402	4	403/414	97%	25%	99%

TABLE XI. BLOF ON 3D PCA-PROCEEDED DATA POINTS (AM, SEMI-SUPERVISED)

	TP	FP	TN	FN	DSR	DSR(%)	PPV	NPV
z1	0	1	22	0	22/23	96%	0%	100%
z2	0	0	22	1	22/23	96%	NaN	96%
z4	0	0	22	1	22/23	96%	NaN	96%
z5	0	0	23	0	23/23	100%	NaN	100%
z6	0	0	23	0	23/23	100%	NaN	100%
z7	0	0	22	1	22/23	96%	NaN	96%
z8	0	0	23	0	23/23	100%	NaN	100%
z9	0	0	23	0	23/23	100%	NaN	100%
z10	0	5	18	0	18/23	78%	0%	100%
z11	0	0	23	0	23/23	100%	NaN	100%
z12	0	0	23	0	23/23	100%	NaN	100%
z13	0	0	23	0	23/23	100%	NaN	100%
z14	0	0	23	0	23/23	100%	NaN	100%
z15	0	0	23	0	23/23	100%	NaN	100%
z16	1	1	21	0	22/23	96%	50%	100%
z17	0	0	23	0	23/23	100%	NaN	100%
z18	0	0	22	1	22/23	96%	NaN	96%
z19	0	0	23	0	23/23	100%	NaN	100%
Total	1	7	402	4	403/414	97%	13%	99%

TABLE XII. BLOF ON 2D PCA-PROCEEDED DATA POINTS (PM, SEMI-SUPERVISED)

	TP	FP	TN	FN	DSR	DSR(%)	PPV	NPV
z1	0	1	22	0	22/23	96%	0%	100%
z2	5	2	16	0	21/23	91%	71%	100%
z3	0	3	19	1	19/23	83%	0%	95%
z4	0	1	22	0	22/23	96%	0%	100%
z5	1	1	21	0	22/23	96%	50%	100%
z7	0	0	22	1	22/23	96%	NaN	96%
z8	3	0	20	0	23/23	100%	100%	100%
z9	0	0	23	0	23/23	100%	NaN	100%
z10	0	0	23	0	23/23	100%	NaN	100%
z11	0	0	23	0	23/23	100%	NaN	100%
z12	0	0	22	1	22/23	96%	NaN	96%
z13	1	2	20	0	21/23	91%	33%	100%
z14	3	0	19	1	22/23	96%	100%	95%
z15	0	0	23	0	23/23	100%	NaN	100%
z16	0	1	21	1	21/23	91%	0%	95%
z17	0	0	23	0	23/23	100%	NaN	100%
z18	0	2	21	0	21/23	91%	0%	100%
z19	0	1	22	0	22/23	96%	0%	100%
Total	13	14	382	5	401/414	95%	32%	99%

TABLE XIII. BLOF ON 3D PCA-PROCEEDED DATAPOINTS (PM, SEMI-SUPERVISED)

	TP	FP	TN	FN	DSR	DSR(%)	PPV	NPV
z1	0	0	23	0	23/23	100%	NaN	100%
z2	4	1	17	1	21/23	91%	80%	94%
z3	0	3	19	1	19/23	83%	0%	95%
z4	0	1	22	0	22/23	96%	0%	100%
z5	0	0	22	1	22/23	96%	NaN	96%
z7	0	0	22	1	22/23	96%	NaN	96%
z8	3	0	20	0	23/23	100%	100%	100%
z9	0	1	22	0	22/23	96%	0%	100%
z10	0	0	23	0	23/23	100%	NaN	100%
z11	0	0	23	0	23/23	100%	NaN	100%
z12	0	0	22	1	22/23	96%	NaN	96%
z13	1	3	19	0	20/23	87%	25%	100%
z14	3	0	19	1	22/23	96%	100%	95%
z15	0	0	23	0	23/23	100%	NaN	100%
z16	0	1	21	1	21/23	91%	0%	95%
z17	0	0	23	0	23/23	100%	NaN	100%
z18	0	3	20	0	20/23	87%	0%	100%
z19	0	1	22	0	22/23	96%	0%	100%
Total	11	14	382	7	393/414	95%	31%	98%

method (i.e. 83% and 91%) in Table 10 while their DSRs are 100% in the classical LOF method in Table 2. Both of them have several false positive points which mean several normal points are regarded as outliers. The upper bounds may be too large for z13 and z17.

The semi-supervised classical LOF method on 3D data points of the AM sessions in Table 3 achieves 99% DSR while the DSR for the new bounds in the BLOF method has been decreased to 97% in Table 11. Especially, the result of z10 is terrible that there are 5 false positive points. It is believed that the upper bound for z10 is too large so that several points were regarded as outliers.

From Table 12, the average DSR of the semi-supervised BLOF method is 95% and PPV is 32% for the PM sessions. The average DSR of the BLOF method decreased to 95% while the classical LOF bounds in Table 4 is 96%. The new BLOF OD result in z3 is not as well as before in the LOF one. Its DSR has only 83%. The number of false positive points is 3.

By comparing the semi-supervised approach in the PM sessions of Table 5 (LOF) and Table 13 (BLOF), the DSR for 3D data points also decreased from 98% to 95%. The OD results of z3 and z13 are 83% DSR and 87% DSR, respectively, that are not good enough. Both have 3 false positive cases that inliers were incorrectly regarded as outliers.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

In conclusion, the average DSR and PPV of the BLOF method are not as good as the classical LOF method by dropping 1% to 2%. By observations, most cases are due to too many false positive points. Spatially, some outliers are far away from the normal points which make the upper bounds be larger, then some normal points lie near the bounds' edge would be easily regarded as outliers.

2) Supervised Approach

In the supervised approach, the OD result for the AM sessions for the 2D data points has increased from 95% DSR (LOF in Tables 6) to 96% (BLOF in Table 14). The BLOF performance is superior in most traffic directions, yet the OD result of z5 is only 73% DSR. It is believed that the upper bound of BLOF is not effective enough.

For the 3D domain in the AM sessions, the average DSRs increased from 94% (LOF in Table 7) to 96% (BLOF in Table 15). Most traffic directions have improved OD results in DSR.

In regard to the AM sessions, the new LOF bounds in the BLOF method in the supervised approach has increased the DSRs and the PPVs, but its deficiency is the large number of false positive cases.

The OD results of the supervised LOF method in the PM sessions of DSR for 2D data points is 91% (Table 8) while that

TABLE XIV. BLOF ON 2D PCA-PROCEEDED DATA POINTS (AM, SUPERVISED)

	TP	FP	TN	FN	DSR	DSR(%)	PPV	NPV
z1	0	0	11	0	11/11	100%	NaN	100%
z2	0	0	10	1	10/11	91%	NaN	91%
z3	1	0	10	0	11/11	100%	100%	100%
z4	0	0	10	1	10/11	91%	NaN	91%
z5	0	3	8	0	8/11	73%	0%	100%
z6	0	1	10	0	10/11	91%	0%	100%
z7	0	0	10	1	10/11	91%	NaN	91%
z8	0	0	11	0	11/11	100%	NaN	100%
z9	0	0	11	0	11/11	100%	NaN	100%
z10	0	0	11	0	11/11	100%	NaN	100%
z11	0	0	11	0	11/11	100%	NaN	100%
z12	0	0	11	0	11/11	100%	NaN	100%
z13	0	0	11	0	11/11	100%	NaN	100%
z14	0	0	11	0	11/11	100%	NaN	100%
z15	0	0	11	0	11/11	100%	NaN	100%
z16	1	0	10	0	11/11	100%	100%	100%
z17	0	0	11	0	11/11	100%	NaN	100%
z18	0	0	10	1	10/11	91%	NaN	91%
z19	0	0	11	0	11/11	100%	NaN	100%
Total	2	4	199	4	201/209	96%	50%	98%

TABLE XV. BLOF ON 3D PCA-PROCEEDED DATA POINTS (AM, SUPERVISED)

	TP	FP	TN	FN	DSR	DSR(%)	PPV	NPV
z1	0	0	11	0	11/11	100%	NaN	100%
z2	0	0	10	1	10/11	91%	NaN	91%
z3	1	1	9	0	10/11	91%	50%	100%
z4	0	0	10	1	10/11	91%	NaN	91%
z5	0	0	11	0	11/11	100%	NaN	100%
z6	0	0	11	0	11/11	100%	NaN	100%
z7	0	0	10	1	10/11	91%	NaN	91%
z8	0	0	11	0	11/11	100%	NaN	100%
z9	0	0	11	0	11/11	100%	NaN	100%
z10	0	1	10	0	10/11	91%	0%	100%
z11	0	1	10	0	10/11	91%	0%	100%
z12	0	0	11	0	11/11	100%	NaN	100%
z13	0	0	11	0	11/11	100%	NaN	100%
z14	0	0	11	0	11/11	100%	NaN	100%
z15	0	0	11	0	11/11	100%	NaN	100%
z16	0	0	10	1	10/11	91%	NaN	91%
z17	0	0	11	0	11/11	100%	NaN	100%
z18	0	0	10	1	10/11	91%	NaN	91%
z19	0	0	11	0	11/11	100%	NaN	100%
Total	1	3	200	5	201/209	96%	17%	98%

TABLE XVI. BLOF ON 2D PCA-PROCEEDED DATA POINTS (PM, SUPERVISED)

	TP	FP	TN	FN	DSR	DSR(%)	PPV	NPV
z1	0	0	11	0	11/11	100%	NaN	100%
z2	3	0	6	2	9/11	82%	100%	75%
z3	0	1	9	1	9/11	82%	0%	90%
z4	0	0	11	0	11/11	100%	NaN	100%
z5	0	0	10	1	10/11	91%	NaN	91%
z6	0	0	10	1	10/11	91%	NaN	91%
z7	0	0	10	1	10/11	91%	NaN	91%
z8	3	0	8	0	11/11	100%	100%	100%
z9	0	0	11	0	11/11	100%	NaN	100%
z10	0	0	11	0	11/11	100%	NaN	100%
z11	0	0	11	0	11/11	100%	NaN	100%
z12	0	0	10	1	10/11	91%	NaN	91%
z13	1	1	9	0	10/11	91%	50%	100%
z14	3	0	7	1	10/11	91%	100%	88%
z15	0	0	11	0	11/11	100%	NaN	100%
z16	0	0	10	1	10/11	91%	NaN	91%
z17	0	0	11	0	11/11	100%	NaN	100%
z18	0	0	11	0	11/11	100%	NaN	100%
z19	0	0	11	0	11/11	100%	NaN	100%
Total	10	2	188	9	198/209	95%	70%	95%

TABLE XVII. BLOF ON 3D PCA-PROCEEDED DATA POINTS (PM, SUPERVISED)

	TP	FP	TN	FN	DSR	DSR(%)	PPV	NPV
z1	0	0	11	0	11/11	100%	NaN	100%
z2	3	0	6	2	9/11	82%	100%	75%
z3	0	0	10	1	10/11	91%	NaN	91%
z4	0	1	10	0	10/11	91%	0%	100%
z5	0	0	10	1	10/11	91%	NaN	91%
z6	0	0	10	1	10/11	91%	NaN	91%
z7	0	0	10	1	10/11	91%	NaN	91%
z8	3	0	8	0	11/11	100%	100%	100%
z9	0	0	11	0	11/11	100%	NaN	100%
z10	0	0	11	0	11/11	100%	NaN	100%
z11	0	0	11	0	11/11	100%	NaN	100%
z12	0	0	10	1	10/11	91%	NaN	91%
z13	1	0	10	0	11/11	100%	100%	100%
z14	3	0	7	1	10/11	91%	100%	88%
z15	0	0	11	0	11/11	100%	NaN	100%
z16	0	0	10	1	10/11	91%	NaN	91%
z17	0	0	11	0	11/11	100%	NaN	100%
z18	0	0	11	0	11/11	100%	NaN	100%
z19	0	1	10	0	10/11	91%	0%	100%
Total	10	2	188	9	18/19	95%	67%	95%

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

TABLE XVIII. COMPARISON OF RESULTS IN TERM OF DSR

DSR	AM				PM			
Dimension	2D LOF	2D BLOF	3D LOF	3D BLOF	2D LOF	2D BLOF	3D LOF	3D BLOF
Semi-supervised	99%	97%	99%	97%	96%	95%	98%	95%
Supervised	95%	96%	94%	96%	91%	95%	93%	95%

TABLE XIX. COMPARISON OF RESULTS IN TERM OF PPV

DSR	AM				PM			
Dimension	2D LOF	2D BLOF	3D LOF	3D BLOF	2D LOF	2D BLOF	3D LOF	3D BLOF
Semi-supervised	50%	25%	50%	13%	30%	32%	100%	31%
Supervised	33%	50%	17%	17%	44%	67%	50%	70%

of the BLOF method is 95% (Table 16). However, the BLOF OD results of z2 and z3 have only 82% DSRs. Herein, many false negative cases are found. We believe the upper bounds for OD are small at these two traffic directions.

In the 3D domain for the PM sessions, the supervised classical LOF method offers 93% (Table 9) and now the supervised BLOF method gives 95% (Table 17). In Table 17, the DSR for the OD in z2 is 82% which is poor indeed. It has two false negative points.

In short, the DSRs and PPVs of the BLOF methods in the PM sessions are better than that of the classical LOF method. Only some traffic directions in the BLOF method have high false negative cases.

V. COMPARISON AND DISCUSSION

Tables 18 and 19 tabulate the OD results for the classical LOF method and BLOF method in term of the semi-supervised or supervised approaches. Specially, in semi-supervised approach, there is not much difference between the traffic data from the 2D or 3D domains in the AM sessions because both obtain 99% DSR and 50% PPV. But using 3D data points of the PM sessions do help improve the average DSR and PPV compared to the 2D data points: 96% to 98% for LOF, 30% to 100% for PPV.

In contrast, in the supervised approach, applying 3D data points for OD have also helped the DSRs of the PM sessions to improve: 91% to 93% for LOF. For the AM sessions, the OD results of the 3D domain is similar with that of the 2D domain, which may be due to the fewer outliers in the AM sessions so some normal points are easily regarded as outliers.

However, for the supervised approach, it would perform better when taking direct reachability distances and indirect reachability distances to get the LOF bounds. In the 2D data points, the BLOF OD result is 50% PPV in the AM sessions while the classical LOF result is only 33% PPV. For supervised approach, it would fit the new LOF bounds better and the corresponding DSRs and PPVs for the 3D data points, therefore the OD performance is improved. In the PM sessions for the 3D domain, the supervised BLOF method have an outstanding results of 95% DSR and 70% PPV that is superior to the classical LOF method with 93% DSR and 50% PPV.

VI. CONCLUSION

This paper presents the new BLOF OD method for large-scale traffic data that demonstrates a superiority to the classical LOF method. There are two possible future works in this research. (1) For the semi-supervised approach, it is required to find a better methodology to get LOF bounds instead of testing by a human; (2) The performance of the supervised approach is not very good. It should have more improvements on the algorithmic design.

ACKNOWLEDGMENT

This research is supported by Hong Kong RGC GRF: 12201814, HKBU FRG1/15-16/002 and HKBU FRG2/14-15/054.

REFERENCES

[1] C. C. Aggarwal, *Outlier Analysis*, Springer, 2013.

[2] C. O'Reilly, A. Gluhak, M. A. Imran, S. Rajasegarar, "Anomaly detection in wireless sensor networks in a non-stationary environment," *IEEE Trans. Communications Surveys & Tutorals*, pp. 1413–1432, 2014.

[3] V. Chandola, A. Banerjee, and V. Kumar. "Anomaly detection: a survey," *ACM Computing Surveys (CSUR)*, vol. 41, issue 3, no. 15, 2009.

[4] L. Clifton, D. A. Clifton, Y. Zhang, P. Watkinson, L. Tarassenko, and H. Yin, "Probabilitic novelty detection with support vector machines*," IEEE Trans. Reliability*, vol. 63, no. 2, pp. 455–467, 2014.

[5] H. Lee, J. Song, and D. Park, "Intrusion detection system based on multi-class SVM," *RSFDGrC, LNAI 3642*, pp. 511–519, 2005.

[6] H.-P. Kriegel, P, Kröge, and A, Zimek, "Outlier detection techniques," *SIAM Int'l Conf. Data Mining,* 2010.

[7] C. S. C. Tsang, H. Y. T. Ngan, and G. K. H. Pang, "Fabric inspection based on the ELO rating method," *Pattern Recognition*, 51, pp. 378–394, 2016.

[8] H. Y. T. Ngan and G. K. H. Pang, "Robust defect detection in plain and twill fabric using directional Bollinger bands," *Optical Engineering*, vol. 54, no. 7, 073106, 2015.

[9] M. K. Ng, H. Y. T. Ngan, X. Yuan, and W. Zhang, "Patterned fabric inspection and visualization by the method of image decomposition." *IEEE Trans. Automation Science & Engineering*, vol. 11, no. 3, pp. 943–947, 2014.

[10] S. Chen, W. Wang, and H. van Zuylen, "A comparison of outlier detection algorithms for ITS data," *Expert Systems with Applications*, vol. 37, no. 2, pp. 1169–1178, 2010.

[11] E. Muller, P. I. Sanchez, Y. Mulle, and K. Bohm, "Ranking outlier nodes in subspaces of attributed graphs," *IEEE 29th Int'l Conf. Data Engineering Workshops (ICDEW)*, pp. 216–222, 2013.

[12] A. D. Pawar, P. N. Kalavadekar, and S. N. Tambe, "A survey on outlier detection techniques for credit card fraud detection," *IOSR Journal of Computer Engineering*, vol. 16, no .2, pp. 44–48, 2014.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

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

[14] D. M. J. Tax and R. P. W. Duin, "Outlier detection using classifier instability," *Proc. Joint IAPR Int'l Workshops on Advances in Pattern Recognition*, pp. 593–601, 1998.

[15] D.J. Hill, B.S. Minsker and E. Amir, "Real-time Bayesian Anomaly Detection for Environmental Senor Data," *Water Resources Reserrch*, vol. 45, no. 4, 2009.

[16] P. Gil, A. Santos, and A. Cardoso, "Dealing with outliers in wireless sensor networks: an oil refinery application," *IEEE Trans. Control Systems Technology*, vol. 22, no. 4, pp. 1589–1596, 2013.

[17] H. Y. T. Ngan, N. H. C. Yung, and A. G. O. Yeh, "Outlier detection in traffic data based on the Dirichlet process mixture model," *IET Intelligent Transport Systems*, vol. 9, no. 7, pp. 773–781, 2015.

[18] C. H. M. Wong, H. Y. T. Ngan and N. H. C. Yung, "Modulo-k clustering based outlier detection for large-scale traffic data," *Proc. Int'l Conf. IEEE Information Technology and Application (ICITA)*, 2016.

[19] T. T. Dang, H. Y. T. Ngan, and W.Liu, "Distance-based k-nearest neighbors outlier detection method in large-scale traffic data," *IEEE Int'l Conf. Digital Signal Processing (DSP)*, pp. 507–710, 2015.

[20] M. Amer, M. Goldstein and S. Abdennadher, "Enhancing one-class support vector machines unsupervised anomaly detection," *Proc. ACM ODD*, pp. 8–15, 2013.

[21] M. Davy and S. Godsill, "Detection of abrupt spectral changes using support vector machines: an application to audio signal segmentation," *Proc. IEEE ICASSP*, vol. 2, pp. II-1313–II-1316, 2002.

[22] N. A. Shrivastava, A. Khosravi, and B. K. Panigrahi, "Prediction interval estimation for wind farm power generation forecasts using support vector machines," *Proc. IEEE Int'l Joint Conf. Neural Networks*, pp.1–7, 2015.

[23] J. Ma and S. Perkins, "Online novelty detection on temporal sequences," *Proc. 9th ACM SIGKDD*, pp. 613–618, 2003.

[24] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*, John Wiley & Sons Inc., 2005.

[25] L. J. Latecki, A. Lazarevic, and D. Pokrajac, "Outlier detection with kernel density functions," *Proc. 5th Int'l Conf. Machine Learning & Data Mining in Pattern Recognition*, pp. 61–75, 2007.

[26] E. M. Knorr, R. T. Ng, and V. Tucakov, "Distance-based outliers: algorithms and applications," *The VLDB Journal*, vol. 8, no. 3–4, pp. 237–253, 2000.

[27] K. Cao, L. Shi, G. Wang, D. Han, and M. Bai, "Density-based local outlier detection on uncertain data," *WAIM, LNCS, v*ol. 8485, pp. 67–71, 2014.

[28] M. X. Ma, H. Y. T. Ngan, and W. Liu, "Density based outlier detection by local outlier factor on large-scale traffic data," *IS&T Int'l Sym. Electronic Imaging*, pp. 1–4, 2016.

[29] A. Arning, R. Agrawal, and P. Raghavan, "A linear method for deviation detection in large databases," *Proc. KDD*, pp. 164–169, 1996.

[30] F. T. Liu, K. M. Ting and Z. H. Zhou, "Isolation-based Anomaly Detection," *ACM Trans. Knowledge Discovery from Data*, vol. 6, no.1, 3:1–3:38, 2012.

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

[32] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731