Distance-Based k-Nearest Neighbors Outlier Detection Method in Large-Scale Traffic Data

Taurus T. Dang, Henry Y.T. Ngan
Department of Mathematics,
Hong Kong Baptist University,
Kowloon Tong, Hong Kong
12051160@life.hkbu.edu.hk, ytngan@hkbu.edu.hk

Wei Liu
Department of Electronic & Electrical Engineering,
University of Sheffield,
Sheffield S1 3JD, U.K.
w.liu@sheffield.ac.uk

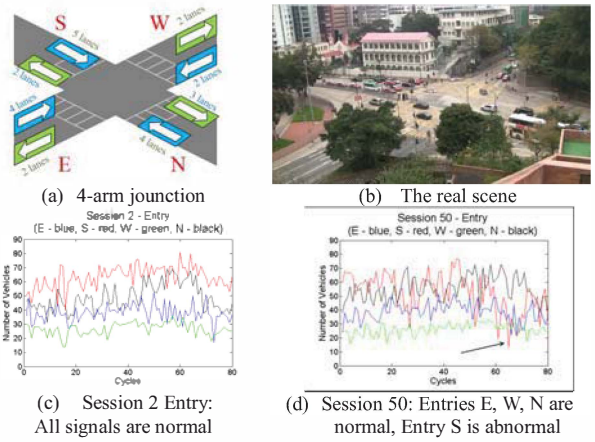
***Abstract*— This paper presents a k-nearest neighbors (kNN) method to detect outliers in large-scale traffic data collected daily in every modern city. Outliers include hardware and data errors as well as abnormal traffic behaviors. The proposed kNN method detects outliers by exploiting the relationship among neighborhoods in data points. The farther a data point is beyond its neighbors, the more possible the data is an outlier. Traffic data here was recorded in a video format, and converted to spatial-temporal (ST) traffic signals by statistics. The ST signals are then transformed to a two-dimensional (2D) (x, y) -coordinate plane by Principal Component Analysis (PCA) for dimension reduction. The distance-based kNN method is evaluated by unsupervised and semi-supervised approaches. The semi-supervised approach reaches 96.19% accuracy.**

Keywords— Outlier detection, large-scale, traffic data, distance-based, kNN.

I. INTRODUCTION

Nowadays, data analysis has a more important role than past decades. Modern cities around the world are creating an extremely huge amount of data [1] at each second in businesses, communications, and transportation, etc. Among such a sea of data, there exist, unavoidably, some outliers to exhibit errors, noise, and abnormal behaviors, etc., deviated from the majority of data. In particular, for vehicle traffic data, analysis on these outliers is vital for traffic forecasting and maintenance [2], and also beneficial to incident detection and management [2], and many other areas.

An outlier is a piece of data which deviates largely from the other observations, leading to suspicions that it emerges from a non-ordinary mechanism, as stated by Hawakins [3]. OD refers to finding any patterns in data which do not conform to an expected behavior [4]. The OD results can lead to actionable (and often critical) responses to the real situation, which is having more and more influence on TCSS in any modern city. For instance, abnormal traffic behaviors [5, 6] such as congestion, incident, low volume, and abrupt driving behavior from road network, detected by a suitable automated OD method could offer an instant and fast response to the transport department and transportation companies.



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

(c) Session 2 Entry: All signals are normal (d) Session 50: Entries E, W, N are normal, Entry S is abnormal

Fig. 1. (a) A generic diagram of the 4-arm junction; (b) sample of the real scene; (c) normal ST signals; (d) abnormal ST signals.

The traffic data used for this research is from a video database from a four-arm junction (Fig. 1(a), (b)) located in one of the most popular regions in Hong Kong [2]. This dataset was captured for 31 days, with each day divided into two sessions: AM (07:00-10:00) and PM (17:00-20:00). Hence, a total of 62 sessions were acquired (764,027 vehicles). This junction has 4 motion patterns (MPs) per cycle operated by traffic lights. Traffic data is measured as either an Entry or Exit per session which is called a ST signal (Fig. 1(c) shows normal signals, Fig. 1(d) shows abnormal one), which represents the volume of Entry or Exit traffic during a session.

Since the original traffic data is recorded in video format and vehicles are moving into and out of the scene with lot of occludes (i.e. trees blocking partial-view of the junction) in the junction, it is hard to analyze and distinguish manually. The video data used in this research has been firstly transformed to ST signals by statistics and then its dimension is reduced to pairs of (x, y)-coordinates by PCA. The reason of using PCA is to transform data into a simpler form but not losing the main information, and reduce the dimension of data to speed up and simplify the analysis.

In this research, the PCA-processed (x, y)-coordinates are fed into the proposed distance-based kNN OD method. The kNN algorithm, as one of the simplest machine learning algorithms, is commonly used in classification and regression [3]. It can have different internal metrics in addition to the

Authorized licensed use limited to: BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE. Downloaded on February 26,2021 at 04:30:20 UTC from IEEE Xplore. Restrictions apply.

distance-based one, such as the local outlier factor (LOF) [3]. Previously, some other statistical OD methods (such as DPMM) and learning approaches (such as SVM) for traffic data were suggested in [2]. However, the kNN method has not been applied to traffic data yet.

In this work, we apply the kNN method to traffic data by firstly calculating the distance between data points (PCA processed (x, y) -coordinates) in given data sets and then classifying them into corresponding groups by their distances. The basic assumption of this distance-based method is that abnormal data points are far from normal ones in location. Two approaches: unsupervised and semi-supervised, are investigated, and it is shown that the semi-supervised approach performs more effectively, with a 96.19% detection success rate.

The organization of this paper is as follows: Section II is a review of related work. Section III presents details of the proposed OD method. Performance evaluation results are provided in Section IV and conclusions are drawn in Section V.

II. Related Work

OD in a large set of data elements is an important data mining step [7, 8]. The importance of OD methods lies in the fact that the results can lead to actionable information. There exist various kinds of approaches for OD as below:

A. Statistical Approach

This is the simplest OD technique. Many OD applications for traffic data belongs to this approach. This approach **[6]** assumes the given data set as a statistical model such as normal (Gaussian) distribution and subsequently detects outliers corresponding to the model by a discordancy test. Its advantage is the simple design and its time complexity is $O(n)$. Its disadvantages include: (a) the result is highly dependent on the statistical model chosen since a data may be an outlier within a model but be a inlier in another [6]; (b) It needs a prior information to recognize the data distribution and its parameters, which is often very chanllenging [3, 6]**.**

B. Distance-based Approach

The distance-based approach [6] provides particular advantage when the data does not fit into any standard distribution model [6], and it can still discover outliers effectively in the multi-dimensional case. However, since the basic assumption of this approach is that outliers are far apart from their neighborhood, this approach may not be efficient for data sets which have non-dense neighborhoods [3]. There are three kinds of distance-based methods: $DB(p, d)$-outlier [6], kNN distance [6, 9, 10] and resolution-based outlier factor (ROF) [11].

C. Density-based Approach

Density-based OD approach [6] introduces a new outlier representation: a local outlier, to investigate the degree of an object to be an outlier corresponding to the local neighborhood's density. This degree is also identified as a local outlier factor (LOF) associating to each data object. Since distance-based methods may be difficult in dealing with data points in different density, the density-based method could be applied to handle such kind of data points. LOF has some weaknesses. The calculation of the standard deviation is very expensive [6]**.** As a result, the density-based method requires a very long running time**.** Three common density-based methods are LOF, Influenced Outlierness (INFLO) and Local outlier correlation integral (LOCI).

D. High-dimensional (HD) Approach

In data mining, when dealing with high-dimensional data, the performance of most existing approaches degrades due to the notorious "curse of dimensionality" [7]. Herein, some HD approaches for OD are listed for alleviating the effects of such curse compared to purely distance-based approaches. Two common HD methods are Angle-based (AB) Method [7], and Grid-based (GB) Method [3].

The state-of-the-art OD methods in traffic data include the density-based (90% accuracy in [6]) and threshold-based (100% accuracy in [5]) methods. As reviewed above, kNN is a classical OD method in the distance-based approach and many variations exist in literature. In this paper, a kNN-based OD method is employed on the large-scale traffic data.

III. kNN OD Method

This section provides basic definitions of the kNN OD method, the procedure and detection results of unsupervised, semi-supervised and supervised approaches.

A. Basic Definitions of the kNN OD Method

Definition 1 ($k-distance$). For a point $p(x, y)$, the distance between p and its k-th nearest neighbor $p_k(x_k, y_k)$ is the $k-distance$ (Shorthand as $k-dist$) of p and denotes as $D^k(p)$:

$$D^k(p) = \sqrt{(y - y_k)^2 + (x - x_k)^2} \quad (1)$$

Definition 2 (maximum D^k values) [6]. Given k and m, a data point p is considered as an outlier provided than less than $m - 1$ other data points possess a larger magnitude of D^k than p, i.e., the top m points with the largest D^k magnitudes are classified as outliers.

In real OD operations, m is unknown and cannot be explicitly determined. Thus, a threshold t acting as a cut line is involved to separate inliers and outliers instead of the parameter m.

B. Performance Evaluation Metrics

As mentioned, the traffic video data was recorded at a four-arm junction in Hong Kong. This junction has 19 traffic directions, in which 8 traffic directions' signals are used in this section. In the proposed distance-based kNN method, each traffic direction has the PCA-processed 23 (x, y)-coordinates as data points. Performance evaluation is carried out by the following metrics: Detection Success Rate (DSR), True Positive Rate (TPR), False Positive Rate (FPR), Positive Predictive Value (PPV), and Negative Predictive Value (NPV), True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). The details of the metrics can be referred to [12].

Authorized licensed use limited to: BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE. Downloaded on February 26,2021 at 04:30:20 UTC from IEEE Xplore. Restrictions apply.

C. Unsupervised kNN Approach

Fig. 2 illustrates the procedure of the unsupervised distance-based kNN approach for OD. First, all data are transformed into a pair of $(x, y$ -coordinate form through PCA operations. Secondly, the parameter k is determined manually. Then we apply **the unsupervised approach procedures** to the given data set. Finally, perform OD and obtain the outliers.

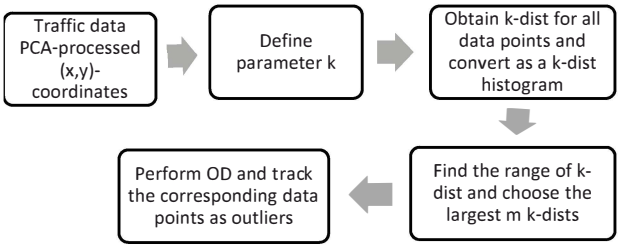


Fig. 2. Procedures of unsupervised distance-based kNN method

Definition 3 (outlier). Given k, for any data point p, p is considered as an outlier if $D^k(p) > t$, where t is a threshold value.

Definition 4 ($k-dist$ histogram). In **the unsupervised approach procedures**, there is a certain value of k-distance for every data point. The $k-dist$ histogram is a distribution of all k-distances in different intervals. Every pillar in histogram represents a $k-dist$ quantity at that k in a certain range.

In this case, the width of a pillar is set as 5. A simple test with two signals based on **the unsupervised approach procedures** is shown in Fig. 3. The two signals are both from the PM session. The signal of Entry E is a normal signal, which is chosen as a comparison group, while the signal of Entry S is

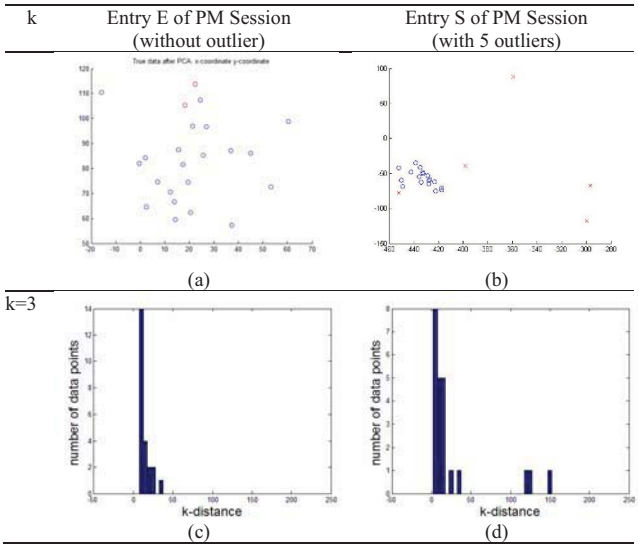


Fig. 3. The first row contains scatter plots of two signals. For every histogram in the other rows, x-axis represents the k-distance and y-axis represents the quantity of data points.

TABLE I.
OD RESULT FOR PM ENTRY E SIGNAL.

t	5	10	15	20	25	30	35	40
TP	0	0	0	0	0	0	0	0
FP	23	15	5	3	1	0	0	0
TN	0	0	18	20	22	23	23	23
FN	0	0	0	0	0	0	0	0
DSR(%)	0	0	78.26	86.96	95.65	100	100	100
TPR(%)	NA	NA	NA	NA	NA	NA	NA	NA
FPR(%)	100	65.22	21.74	13.04	4.35	0	0	0

TABLE II.
OD RESULT FOR PM ENTRY S SIGNAL

	5	10	15	20	25	30	35	40
TP	5	5	5	5	4	4	3	3
FP	17	7	3	0	0	0	0	0
TN	1	11	15	18	18	18	18	18
FN	0	0	0	0	1	1	2	2
DSR(%)	26.09	69.57	86.96	100	95.65	95.65	91.30	91.30
TPR(%)	100	100	100	100	80	80	60	60
FPR(%)	94.44	38.89	16.67	0	0	0	0	0

an abnormal signal with 5 outliers. In this test, there are only 23 data points in each group and k is tested from 1 to 3. Figs. 3(a),(b) show the PCA-processed (x, y)-coordinates of Entry E and Entry S, respectively and Fig. 3(c),(d) shows $k = 3$, the histogram of group Entry E (no outlier) does not change much. While in the histogram of group Entry S (with outlier), some data points have a k-distance value larger than the majority.

Then, a study of the effect of variation of the threshold t is shown in Table I, where we can see that FP decreases and TN increases when t increases, while TP and FN maintain the same level (all zero). The best DSR result is 100% when t is greater than 30.

In Table II, TP and FP decrease and TN and FN increase when t increases. The best DSR result is 100% when t is equal to 20. A ROC analysis is carried out for the PM Entry S signal, but there is no ROC plot for Table I because the denominator of TPR is zero for all t. According to the scatter plots and histograms in Fig. 1, the value of t is manually determined to be 30. Based on this t, through **the unsupervised approach procedures**, some results are obtained. The average DSR of the AM session is 96.11%, and the average DSR of the PM session is 94.97%.

Obviously, the OD result by **the unsupervised approach procedures** is not very promising. However, it provides a good reference baseline for semi-supervised kNN approaches.

D. Semi-supervised kNN Approach

In the semi-supervised approach, the distance-based kNN method determines the value of a *global* threshold t. It is divided into two phases: a training phase and a testing phase. Procedure of the semi-supervised approach is shown in Fig. 4. First, all data are transformed into the (x, y) coordinate form through PCA. Second, the parameter k is determined manually.

For the training phase, two abnormal signals are chosen as a training set, one Entry W of AM session and one Exit N of PM session, since both signals have significant outlier(s) that can be easily recognized from the scatter plots. The parameter

Authorized licensed use limited to: BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE. Downloaded on February 26,2021 at 04:30:20 UTC from IEEE Xplore. Restrictions apply.

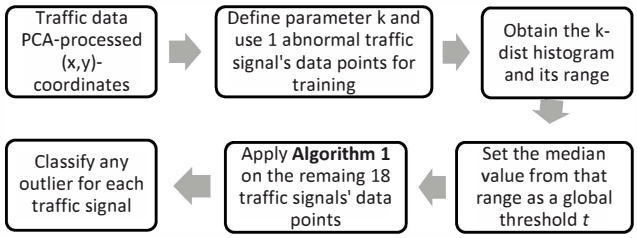


Fig. 4. Procedures of the semi-supervised distance-based kNN method.

k here is determined as 3 manually, same as in the step of the unsupervised method. Each signal generates a $k-dist$ histogram which looks similar to Fig. 3. With various t values, the optimized choices of a global threshold t are located in [19,59] and [22,62] for AM Entry W and PM Exit N signals, respectively. The value of t is chosen as the medium of the interval obtained in the training phase, i.e., $t = 39$ for AM session and $t = 42$ for PM session, because the OD results do not vary within those intervals.

For the testing phase, the average OD results for the remaining 8 traffic signals are 96.20% for the AM sessions (Table III) and 96.19% for the PM sessions (Table IV). The poorest result is found for the PM Entry S signal with 2 FN cases and 91.30% DSR. Herein, the semi-supervised approach is much more efficient than the unsupervised one.

In particular, for the result on Entry S of PM session where we see that the two FN points in the result are located very closely to the normal data points, even in the cluster of inliers, which is unusual to the whole data set. That is a reason for the failure of the semi-supervised distance-based OD method.

TABLE III.
OD RESULTS OF AM SESSIONS UNDER A GLOBAL THRESHOLD $t = 39$.

Signal	TP	FP	TN	FN	TPR	FPR	PPV(%)	NPV(%)	DSR	%
Entry E	0	0	23	0	NA	0/23	NA	100	23/23	100
Entry S	0	0	22	1	0/1	0/22	NA	95.65	22/23	95.65
Entry W	1	0	22	0	1/1	0/22	100	100	23/23	100
Entry N	0	2	20	1	0/1	2/22	0	95.24	20/23	86.96
Exit E	0	2	21	0	NA	2/23	0	100	21/23	91.30
Exit S	0	0	23	0	NA	0/23	NA	100	23/23	100
Exit W	0	0	22	1	0/1	0/22	NA	95.65	22/23	95.65
Exit N	0	0	23	0	NA	0/23	NA	100	23/23	100
Average										**96.20**

TABLE IV.
OD RESULTS OF PM SESSIONS UNDER A GLOBAL THRESHOLD $t = 42$.

Signal	TP	FP	TN	FN	TPR	FPR	PPV(%)	NPV(%)	DSR	DSR(%)
Entry E	0	0	23	0	NA	0/23	NA	100	23/23	100
Entry S	3	0	18	2	3/5	0/18	100	90.00	21/23	91.30
Entry W	0	1	21	1	0/1	1/22	0	95.45	21/23	91.30
Entry N	0	1	22	0	NA	1/23	0	100	22/23	95.65
Exit E	1	1	21	0	1/1	1/22	50	100	22/23	95.65
Exit S	1	0	22	0	1/1	0/22	100	100	23/23	100
Exit W	0	0	22	1	0/1	0/22	NA	95.65	22/23	95.65
Exit N	3	0	20	0	3/3	0/20	100	100	23/23	100
Average										**96.19**

IV. CONCLUSION

The distance-based kNN OD method has been studied for large-scale traffic data with unsupervised and semi-supervised approaches. Our results show that the unsupervised and semi-supervised methods give average DSRs of 95.54% and 96.19% in real-world traffic data. Previously, we have evaluated OD methods of kernel density estimation (95.20% DSR), S-estimator (76.20% DSR), one-class SVM (59.27% DSR), and Gaussian mixture model (80.86% DSR) in [2]. Therefore, the kNN method presented in this paper is robust. Since our distance-based kNN method used only one of the distance-based metrics for a certain point under a given k, more complicated variations such as multiple distances or different weights on different distances, can potentially improve the OD result further in future.

ACKNOWLEDGMENT

This research is supported by 2013-2014 summer research fellowship of Math. Dept., Hong Kong Baptist University (HKBU) for the 1st author and by Hong Kong RGC GRF: 12201814 and HKBU FRG/14-15/054 for the 2nd author.

REFERENCES

[1] S. Sscalera, X. Baro, O. Pujol, J. Vitria, and P. Radeva, *Traffic-Sign Recognition System*, Springer, 2011.
[2] H.Y.T. Ngan, H.C. Yung, and G.O. Yeh, "Detection of Outliers in Traffic Data based on Dirichlet Process Mixture Model," *IET ITS*, *(in press)*.
[3] H. Kriegel, P. Kröger, and A. Zimek, *Outlier Detection Techniques*. Ludwig-Maximilians-Universität München, 2010.
[4] M. Onderwater, "Detecting Unusual User Profiles with Outlier Detection Techniques," Master Thesis, 2010.
[5] E.S. Park, S. Turner, C.H. Spiegelman, "Empirical Approaches to Outlier Detection in Intelligent Transportation Systems Data," *Transportation Research Record*, 03-2990, pp. 21-30, 2003.
[6] S.Y. Chen, W. Wang, and H. Zuylen, "A Comparison of Outlier Detection Algorithms for ITS data," *Expert Systems with Applications,* vol. 37, pp.1169–1178, 2010.
[7] H.P. Kriegel, M. Schubert and A. Zimek, "Angle-based Outlier Detection in High-dimensional Data," *Proc. ACM SIGKDD,* pp. 444-452, 2008.
[8] V. Chandola, A. Banerjee, and V. Kumar, "Outlier Detection: A Survey," *ACM Computing Surveys*, vol. 41, no. 3, pp. 1-58, 2009.
[9] V. Hautamaki, I. Karkkainen, and P. Franti, "Outlier Detection Using k-Nearest Neighbour Graph," *Proc. IEEE ICPR, vol. 3, pp. 430-433,* 2004.
[10] D. Wang, Y. Zheng, and J. Cao, "Parallel Construction of Approximate kNN Graph," *Proc. IEEE DCABES,* pp. 22-26, 2012.
[11] M.R. Trad, A. Joly, and N. Boujemaa, "Large Scale KNN-Graph Approximation," *Proc. IEEE ICDMW, pp. 439-448,* 2012.
[12] M.K. Ng, H.Y.T. Ngan, X. Yuan and W. Zhang, "Patterned Fabric Inspection and Visualization by the Method of Image Decomposition," *IEEE TASE,* vol. 11, no. 3, pp. 943-947, 2014.

Authorized licensed use limited to: BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE. Downloaded on February 26,2021 at 04:30:20 UTC from IEEE Xplore. Restrictions apply.