DOI: 10.2352/ISSN.2470-1173.2016.14.IPMVA-385

Density-based Outlier Detection by Local Outlier Factor on Large-scale Traffic Data

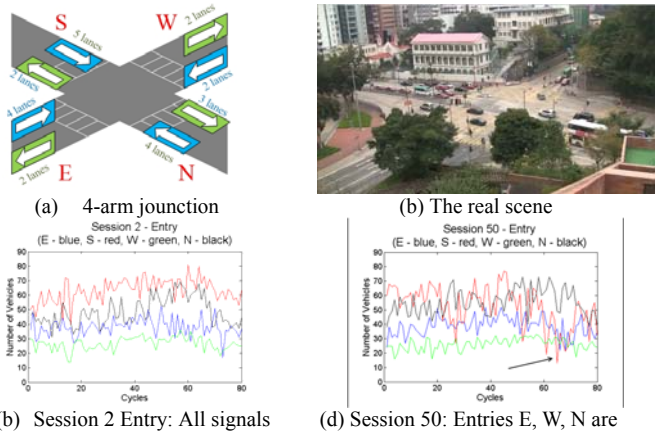
Mathew X. Ma[1], Henry Y.T. Ngan[1] and Wei Liu[2] ;Department of Mathematics, Hong Kong Baptist University, Hong Kong
[2]Communication Research Group, Department of Electronic & Electrical Engineering, University of Sheffield, Sheffield S1 3JD, U.K.

Abstract

A density-based outlier detection (OD) method is presented by measuring the local outlier factor (LOF) on a projected principal component analysis (PCA) domain from real world spatial-temporal (ST) traffic signals. Its aim is to detect traffic data outliers which are errors in data and traffic anomalies in real situations such as accidents, congestions and low volume. Since the ST traffic signals have a high degree of similarities, they are first projected to two-dimensional (2D) (x,y)-coordinates by the PCA to reduce its dimension as well as to remove noise, while keeping the anomaly information of the signals. Based on the designed LOF algorithm, a semi-supervised approach is employed to label any embedded outliers. It reaches an average detection success rate of 93.5%.

Introduction

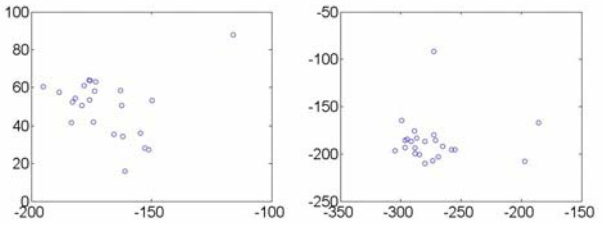
Traffic data reflects the dynamics of traffic through the recording of vehicle volume, speed, change of lanes, etc [1]. Traffic flows at the same location is roughly periodic and its spatial-temporal (ST) property in different time units has high degree of similarities [2]. However, unexpected traffic events including congestions and accidents will lead to deviation of the recorded data from the majority data and such anomalies may not be obvious in simple visualization of the ST traffic signals [3]. Therefore, an automated traffic OD method is needed for effective abnormal traffic event detection [4], localization and descriptions of such traffic flow feature. OD is commonly applied in areas of fraud and intrusion detection [5], data processing [6], trajectory monitoring [7], and classifying chromosome [8]. In traffic data area, commonly used approaches for OD include the statistical [9], the distance-based [6] and the density-based [6] approaches.



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

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

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



(a) AM Z3 data set after PCA (Entry West) (b) PM Z8 data set after PCA (Exit North)

Figure 2. Samples of the PCA-processed traffic data: (a) AM Z3, (b) PM Z8.

TABLE I.
19 Traffic Direction Distribution.

Entry \ Exit	E (Z5)	S (Z6)	W (Z7)	N (Z8)
E (Z1)		Z9	Z11	Z10
S (Z2)	Z13		Z12	Z14
W (Z3)	Z16	Not exist		Z15
N (Z4)	Z17	Z19	Z18	

Remark: E, S, W and N denote East, South, West, North direction.

The traffic data set employed in this paper was collected by video camera in a 4-arm Junction (Fig. 1(a) and (b)) in Hong Kong for 31 days [4]. The selected time span is 23 weekdays. The data set includes 19 subsets $Z1, Z2, \ldots, Z19$ (i.e. 19 traffic directions of vehicles for the entry, exit and direction distribution directions in the junction) as shown in Table 1. Each of the dataset from Z1 to Z19 denotes one case in direction distribution, e.g. row E column S denotes the vehicles that come from the East direction and go into the South direction. The traffic flows entry toward the junction are from $Z1$ to $Z4$ while the traffic flows exit from the junction are from $Z5$ to $Z8$.

In each day, the video data was collected from the A.M. peak hours (07:00-10:00) and the P.M. peak hours (17:00-20:00). The AM sessions have 312, 333 vehicles and the PM sessions have 451,694 vehicles. The dataset recorded 764, 027 vehicles in total. Fig. 1(c) and (d) demonstrate the high degree of similarities between the normal and abnormal ST traffic signals. Although the ST signals have a finite duration, the dimension of one ST signal is too large (e.g. usually over 80 traffic cycles) and contains much unnecessary information. Therefore, the actual OD is performed on the PCA-processed data points by reducing the high dimensional ST signals into first two coefficients as 2D (x,y)-datapoints.

©2016 Society for Imaging Science and Technology
DOI: 10.2352/ISSN.2470-1173.2016.14.IPMVA-385

In [4], we proved first two coefficients are effective to represent one ST signal. Now, each traffic signal is then modeled by 23 AM/23 PM 2D data points. The traffic anomaly are still maintained in these transformed datapoints.

The motivations of the research include that an investigation of a possibility for an OD to be performed on a new data domain by transforming the traffic ST signals to PCA-processed data points, and the spatial distribution property in the datapoints such as density can be utilized for the design of an OD method. Two main research objectives are to develop an effective OD for the PCA-processed datapoints, and to utilize the density property from the datapoints' distribution for measuring the LOF.

Fig. 2 shows two samples of PCA-processed (x,y)-coordinates from the AM Z3 (Fig. 2(a)) and PM Z8 (Fig. 2(b)) signals. Some data points, regarded as suspicious outliers, deviate from the major cluster. The density-based approach suggested in this paper is widely used in 2D and 3D OD [5]. The performance evaluation is carried out in a semi-supervised approach. In the AM/PM sessions, one traffic direction is input to the density-based LOF OD method as the corresponding training data. A receiver operating characteristics (ROC) analysis is performed to find the best threshold to separate outliers and inliers based on the LOF of each point. Finally, the threshold and LOF algorithm will be used to detect any outlier of the remaining 18 traffic signals. The above semi-supervised approach can achieve an average of 93.5% DSR.

This paper is organized as follows. Section 2 gives a review of the related work for OD. Details of our density-based LOF OD method and the corresponding results are provided in Section 3. Conclusions are drawn in Section 4.

Related work

In this research, our dataset is PCA-proceeded traffic data points. OD has generally three main approaches: statistical, distance and density. The statistical approach includes normal, Kai, F, student-t, Poisson, alpha, gamma distributions, etc [9]. Sometimes, more than one statistical distributions can fit the dataset. However, for most cases, this distribution is not known. Therefore, the statistical approach is not appropriate in our dataset. The main idea of the distance-based approach is to determine an outlier to its neighborhood by the Euclidean distance [6]. The distance-based approach, including the k-nearest neighbor (kNN) method, is usually employed when the data does not fit any distribution, and a model generating mechanism is not required in this approach. The distance-based methods are flexible and can be used for the data points derived from PCA processing of the traffic data in this research.

The density-based approach is a proximity-based method using various distance metrics [6], which is capable of solving the multi-cluster OD problem effectively. The LOF is an effective method to find an outlier and it is actually based on the concept from the distance-based approach. However, the LOF algorithm generates a relative density

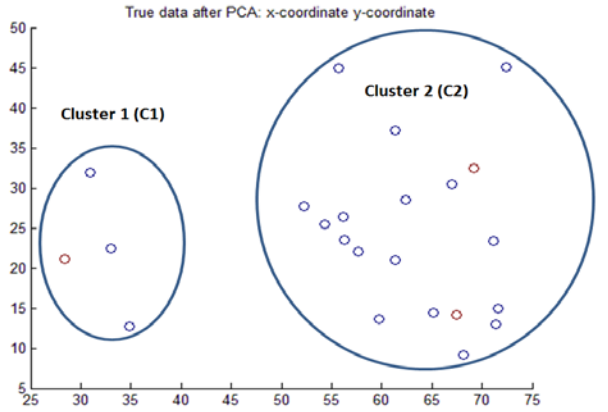


Figure 3. The PCA-processed data points after transformation (C1: cluster 1, C2: cluster 2)

value instead of a distance value. Therefore, the LOF algorithm can be applied to PCA-proceeded traffic data points with different density clusters. The key difference between the KNN and the LOF methods is that LOF computes the relative density of each data point while kNN only calculates the sum of distances to each neighbor.

Fig. 3 illustrates how the LOF works better than the distance-based approach in one kind of data distribution. In Fig. 3, it is obvious that cluster C2 has higher density than cluster C1, where the data points in C1 are regarded as outliers. Moreover, the neighbor distances of points in C1 are generally longer than points in cluster C2. In this case, the distance to its neighbor cannot directly show the outliers. Regarding this situation, an appropriate OD should depend on its cluster as well as relative density. Hence, Fig. 4 illustrates how the *k*-distance, the key proximity in LOF, is defined as the max distance of k nearest neighbor points.

In short, for the fitness of the above three approaches, the statistical approach is not suitable in our case for the traffic OD because the data distribution cannot be specified for most situations. Meanwhile, the distance-based and density-based ones are applicable to the problem of traffic OD.

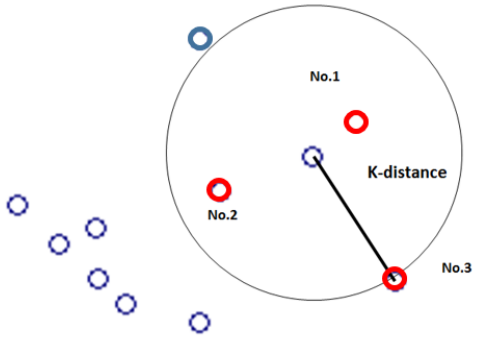


Figure 4. Illustration to show k-distance when $k = 3$. The k-distance is the distance to the k^{th} nearest points labeled as red circle points.

DOI: 10.2352/ISSN.2470-1173.2016.14.IPMVA-385

As regard to their complexities, the density-based and distance-based approaches are proportional to kn^2, where n is the sample size and k is an adjustable parameter, while the statistical approach is proportional to n. Because of the algorithmic difference among the three approaches, the average computer running time Cputime for each approach has the following relationship:

$$Cputime_{statistics} < Cputime_{distance} < Cputime_{density} \quad (1)$$

For the accuracy in most traffic OD studies [10], the density-based approach usually results in a higher accuracy than the distance-based approach.

Density-based LOF method

This section has five parts: mathematics and algorithm of the LOF OD method, evaluation metrics, parameter selection, experimental results of the semi-supervised approach.

Mathematics and Algorithmic Procedures

a. **Definition of k-neighbor distance**

A $kdist(m)$ is defined as the k^{th} smallest distance to a data point m (as shown in Fig. 4).

b. Definition of reachability distance

A reachability distance, as the intermediate parameter, is expressed as

$$rdist(m,p) = max\{dist(m,p), kdist(p)\} \quad (2)$$

where p is a target point and m is the current data point.

It is actually a replacement for Euclidean distance $dist(m,p)$. If the Euclidean distance of two points is very small, the following steps will give a bias ratio of distance. Therefore, the LOF algorithm uses reachability distance instead.

c. Definition of local reachability distance

A local reachability distance is defined as

$$lrd(m) = |R(m)|/(\sum_{p\in R(m)} rdist(m,p)) \quad (3),$$

where $R(m) = \{p|dist(m,p) < kdist(m)\}$.

It calculates the average reachability distance of k neighbors.

d. Definition of LOF

Lastly, a LOF is defined as

$$lof(m) = \frac{\sum_{p\in R(m)} lrd(p)}{|R(m)|}/lrd(m) \quad (4)$$

With (2)-(4), OD can be carried.

Evaluation Metrics

The key evaluation metric detection success rate (DSR) is defined as $DSR = (TP + TN)/(TP + FP + TN + FN)$. Other evaluation metrics such as true positive rate (TPR), false positive rate (FPR), positive predictive value (PPV), negative predictive value (NPV) are involved as well to measure the effectiveness of the proposed OD method. Details of the metrics' definitions can be found in [4].

Parameter Selection

Intuitively, the best value of k does not exsit in density-based OD because the deviation of an outlier cannot be exactly quantified. Therefore, we will choose k by mainly considering two factors: robustness and effectiveness. In our case, $k = 7$ is chosen after an analysis of LOF value versus k value.

Semi-supervised Approach

In this paper, the Z3 in AM (or Z8 in PM) signals are used as the training sets for the corresponding AM (or PM) sessions while the remaining ones are utilized as the testing sets. Since both of the training sets have perfect thresholds. The threshold means to separate the outlier group from the inlier group, and a perfect threshold can divide them without any errors. There exists at least one optimal x_b value to divide the PCA-processed data points into the inlier group and the outlier group. In experiments, a variety of x_b values is tested in both AM and PM sessions, i.e., $x_b = \{1.6,1.8, \dots,5.0\}$ for AM section; $x_b = \{1.6,1.8, \dots,3.0\}$ for PM section (differed by 0.2). After testing, we chose $x_b = 3.3$ for the global threshold of the AM sections, and $x_b = 2.3$ for the global threshold of the PM sessions.

The results are shown in TABLE II and TABLE III. A DSR of over 90% has been achieved in both cases. The AM sessions have a 20% TPR and 4% FPR while for the PM sessions we have obtained a 81% TPR and 8% FPR. However, the false positive cases are 15 out of 414 points (AM) and 32 out of 414 points (PM). Since only one signal was used as the training set for the AM/PM sessions, the decision boundary has not been optimized. A larger database is believed to be better for evaluation in the future.

Since the signal for each direction goes through the PCA separately, it is difficult to measure the LOF value altogether. Moreover, the outliers may overlap with other inliers across signals with different directions. Therefore, the semi-supervised approach does have potential problems in this context.

Result Analysis

The DSRs of the semi-supervised approach for the AM and PM sessions are 95% and 92%, respectively. Therefore,

DOI: 10.2352/ISSN.2470-1173.2016.14.IPMVA-385

TABLE II.
SEMI-SUPERVISED APPROACH: AM RESULT.

	TP	FP	TN	FN	DSR	DSR(%)	PPV	NPV
Z1	0	0	23	0	23/23	100%	NA	100%
Z2	0	0	22	1	22/23	96%	NA	96%
Z4	0	1	21	1	21/23	91%	0%	95%
Z5	0	3	20	0	20/23	87%	0%	100%
Z6	0	0	23	0	23/23	100%	NA	100%
Z7	0	0	22	1	22/23	96%	NA	96%
Z8	0	0	23	0	23/23	100%	NA	100%
Z9	0	1	22	0	22/23	96%	0%	100%
Z10	0	2	21	0	21/23	91%	0%	100%
Z11	0	1	22	0	22/23	96%	0%	100%
Z12	0	2	21	0	21/23	91%	0%	100%
Z13	0	2	21	0	21/23	91%	0%	100%
Z14	0	1	22	0	22/23	96%	0%	100%
Z15	0	0	23	0	23/23	100%	NA	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%	NA	96%
Z19	0	0	23	0	23/23	100%	NA	100%
Total	1	15	394	4	395/414	95%	6%	99%

TABLE III.
SEMI-SUPERVISED APPROACH: PM RESULT.

	TP	FP	TN	FN	DSR	DSR(%)	PPV	NPV
Z1	0	1	22	0	22/23	96%	0%	100%
Z2	5	0	18	0	23/23	100%	100%	100%
Z3	0	3	19	1	20/23	83%	0%	95%
Z4	0	1	22	0	22/23	96%	0%	100%
Z5	1	1	21	0	22/23	96%	50%	100%
Z6	1	0	22	0	23/23	100%	100%	100%
Z7	1	3	19	0	22/23	87%	25%	100%
Z9	0	0	23	0	23/23	100%	NA	100%
Z10	0	3	20	0	21/23	87%	0%	100%
Z11	0	4	19	0	21/23	83%	0%	100%
Z12	1	5	17	0	20/23	78%	17%	100%
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%	NA	100%
Z16	0	3	19	1	20/23	83%	0%	95%
Z17	0	2	21	0	23/23	91%	0%	100%
Z18	0	3	20	0	20/23	87%	0%	100%
Z19	0	1	22	0	22/23	96%	0%	100%
Total	13	32	366	3	379/414	92%	29%	99%

this approach offers high accuracy which is outstanding in traffic OD. However, the PPVs of the semi-supervised approach are 6% (AM) and 29% (PM) which can be further improved.

Conclusion

The semi-supervised density-based OD achieves an average 93.5% DSR. Therefore, it shows the novelty that the density-based LOF OD method is effective and efficient in PCA-proceeded traffic data. This performance is comparable to our previous evaluated OD methods of one-class SVM (59.27% DSR), S-estimator (76.20% DSR), Gaussian mixture model (80.86% DSR) and kernel density estimation (95.20% DSR) in [4]. In this paper, the measuring dynamic contains only traffic flow signals. With more dynamics like average speed, usage of different lanes in one direction, the proposed method may achieve a higher accuracy and can even reveal possible causes for the underlying traffic anomalies.

Acknowledgement

This research is supported by 2013-2014 summer research fellowship of Department of Mathematics, Hong Kong Baptist University for the first author and by Hong Kong RGC GRF: 12201814 and HKBU FRG1/15-16/002 for the second author.

References

[1] S. Wu, S. Wang, "Information-Theoretic Outlier Detection for Large-Scale Categorical Data", *IEEE Trans. KDE*, 25(3), pp.589, 2013

[2] M. Gupta, J. Gao, C.C. Aggarwal and J. Han, "Outlier Detection for Temporal Data: A Survey", *IEEE TKDE*, 26(9), pp. 2250-2267, 2014.

[3] F. Angiulli, S. Basta, S. Lodi, C. Sartori, "Distributed Strategies for Mining Outliers in Large Data Sets," *IEEE TKDE*, 25(7), pp. 1520-1532, 2013.

[4] 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 ITS*, vol. 9, no. 7, 773-781, 2015.

[5] C.C. Aggarwal, P.S. Yu, "Outlier Detection for High Dimensional Data"; *ACM SIGMOD*, pp. 37-46, 2001.

[6] V. Chandola, A. Banerjee, and V. Kumar, "Outlier Detection: A Survey," *ACM Computing Surveys*, vol. 41, no. 3, pp. 1-58, 2009.

[7] E. Masciari, "Trajectory Outlier Detection Using An Analytical Approach," *Proc. 23rd IEEE Conf. ICTAI*, pp. 377-384, 2011.

[8] G. Ritter, M.T. Gallegos, "Outliers in Statistical Pattern Recognition and An Application to Automatic Chromosome Classification," *Pat. Rec. Letters*, vol. 18, pp. 525-539, 1997.

[9] U. Balasooriya, "Detection of outliers in the exponential distribution based on prediction," *Communications in Statistics - Theory and Methods*, vol. 18, no. 2, pp. 711-720, 1989.

[10] S. Chen, W. Wang and H. van Zuylen, "A comparison of outlier detection algorithms for ITS data", *Expert Systems with Applications*, vol. 37, pp. 1169–1178, 2010.