Outlier Detection in Traffic Data Set

Teodora Mecheva[1, a)]

[1]*Technical University of Sofia – Plovdiv Branch, 25 Tsanko Dustabanov Street, Plovdiv, Bulgaria*

a)teodora.mecheva@std.tu-plovdiv.bg

Abstract. The article presents a comparison of three outlier detection methods - Local Outlier Factor, Isolation Forest, and One Class Support Vector Machine over traffic data set. The data are obtained from virtual detectors based on road cameras. Subsets corresponding to hours of the day for working days and holidays are formed. The algorithms are applied over each subset and the dispersion of purified and raw data is compared via coefficient of variation in percentage. Experimental results show that the algorithms handle the working day data set well, but have difficulty with the holidays datasets. When processing the holidays data sets Isolation Forest shows the best results, which confirms one of its main advantages – applicability over small data sets.

INTRODUCTION

With the emergence of the Internet of Things (IoT) one of the important issues facing data analysis is the quality of data and its pre-processing. Intelligent transport systems (ITS) are one of the sub-areas of IoT, which has been developing dynamically in recent years due to growing needs for mobility and the concentration of population in urban areas [1][2].

The main characteristic of the traffic that is used in describing, predicting and improving ITS is the estimation of the traffic flow, based on counting the number of vehicles that cross given location during some time interval. The flow varies over the day and between different days of the week. Finding generalized models and ignoring some characteristics can contribute to the detection of dependencies in ITS [3][4].

Traffic flow data can be collected from various sources - infrastructure, intelligent vehicles, consumer devices. Inaccuracy in measurements or data transmission errors can lead to deviation in the data. Single events like particular weather conditions, or reconstruction can also contaminate data [5][6].

Data preprocessing in the form of selection, purification and generalization plays considerable role in the process of traffic data analysis. Outlier detection, statistical analysis, classification, clustering, and sampling are common preprocessing approaches [7].

The detection of outliers in the traffic flow is one of the main stages of the analysis of urban traffic data. Outliers can skew statistical measures and data distributions, providing a misleading representation of the underlying data and relationships. Eliminating data deviations can lead to a more accurate and functional model. There are a variety of methods and tools for identifying outliers, providing alternate approaches [7][8][9].

The present work describes the application of three outlier detection methods, implemented in the python library *scikit-learn*, over traffic data from road cameras.

OUTLIER DETECTION METHODS (ODM)

The assessment of the applicability of ODM over a given data set is complex due to the volume, variation, and the multidimensionality of the data. This study examines three classical algorithms for outlier detection – Local Outlier Factor (LOF), Isolation Forest (IF), and One Class Support Vector Machine (OCSVM). The python libraries *scikit-*

learn (outlier detection) and *pandas* (data loading, manipulating and summarizing) are used. Table 1 presents the strengths, drawbacks and the main characteristic of the used ODMs [9][10][11][12][13][14][15].

TABLE 1. Outlier detection methods characteristics.

Method	**Local Outlier Factor**	**Isolation Forest**	**One class Support Vector Machine**
Description	based on an idea of nearest neighbours	builds multiple random isolation trees	capturing the density of the majority class to detect outliers
Scikit-learn class	LocalOutlierFactor	IsolationForest	OneClassSVM
Additional information	to each example is assigned a scoring of how isolated based on the size of its local neighborhood is	isolates anomalies that are both few in number and different in the feature space	a modification of the classification algorithm Support Vector Machine
Strengths	•can effectively identify the local outliers.	•simple and fast; •works very well with small sampling size; •performs well in data sets that contain only normal data.	•effective in high dimensional spaces; •effective in cases where the number of dimensions is greater than the number of samples; •relatively memory efficient.
Drawbacks	•with larger dimensionality data is less reliable; •the selection of a point as an outlier is user-dependent.	•has difficulty with multidimensional data; •when anomalies are close to normal samples, it is difficult to discriminate them (swamping); •when anomalies form a dense cluster, has difficulties in detecting them (masking).	•not suitable for large data sets; •does not perform very well when the data set contains large margin of noise.

INPUT DATA

The data in the research are obtained through virtual detectors built on the basis of road cameras. They are extracted from the database of the municipality of Plovdiv in the form of excel reports. Each report contains information about the number of vehicles passed through the line of a single intersection per hour. The available reports cover 3 main intersections in central Plovdiv for the periods from January 15th to January 27th 2021 and from February 19th to March 19th 2021.

Figure 1 presents graph of number of vehicles passed through one of the directions of the junction “Naicho Tsanov” Boulevard – “Hristo Botev” Boulevard for the period from January 15th to January 27th 2021.

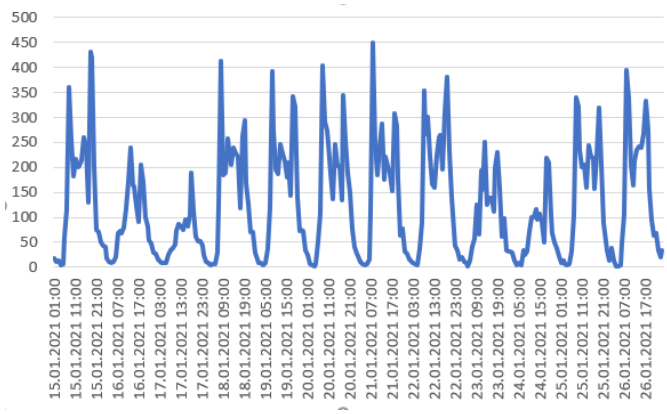

FIGURE 1. Traffic flow over "Hristo Botev" – "Naicho Tsanov" junction

The morning and evening peak hours are clearly visible on the graph, as well as the weekdays and holidays.

Via python script is generated general report of the flows for the analyzed area in csv format. Each row in the report corresponds to a time interval (one hour) and each column corresponds to a traffic direction. The values in the report are the number of vehicles passed through the particular direction per hour.

As the pattern of traffic on weekdays and holidays at certain hours of the day is repeated, the data set is divided into two parts - working days and holidays and is grouped by hour. In the formed subsets each of the streams consists of numerical code of junction, code of junction direction and hour of the day. The ODMs are applied over each subset.

APPLYING ODM OVER TRAFFIC DATA

LOF, IF and OCSVM are applied to each of traffic data subset. The coefficient of variation (CV) in percentage is used to compare the dispersion of purified and raw data:

$$CV=\sigma/\mu*100\% \quad (1)$$

Since the mean values in the subsets of data vary, the coefficient of variation in percentage is chosen as a measure independent of the mean. In this way, it is possible to compare the distribution of data in different subsets.

24 working days data subsets are formed, corresponding to each hour of the day. Each of the them consists of 10 columns. The number of rows varies from 27 to 29. Figures 2, 3 and 4 illustrate the comparison of CV of purified and raw data of working days subsets.

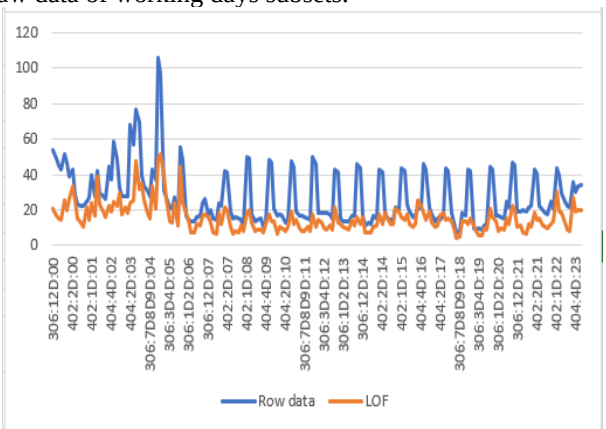

FIGURE 2. CV of LOF and raw data in working days subsets

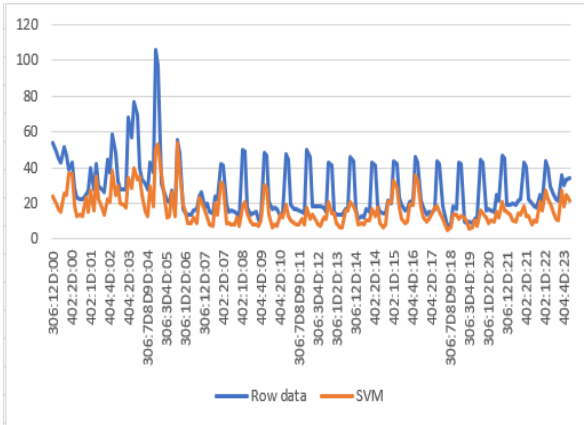

FIGURE 3. CV of OCSVM and raw data in working days subsets

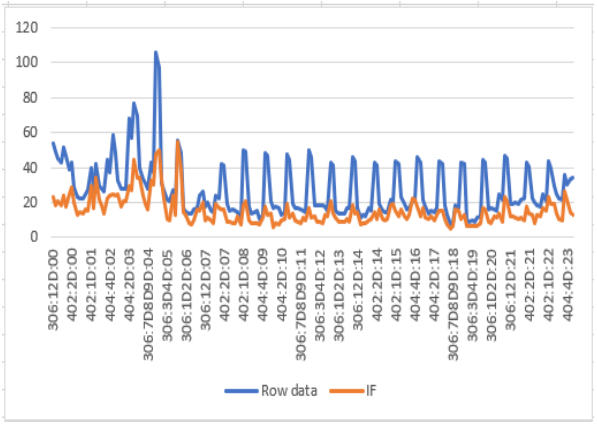

FIGURE 4. CV of IF and raw data in working days subsets

Figures 2, 3 and 4 show that the CV of the data processed with LOF, IF, and OCSVM looks similar. The CV of the raw data varies between 15 and 110%. The CV of the data purified with LOF, IF, and OCSVM varies between 5 and 60%.

24 holidays data subsets are formed, corresponding to each hour of the day. Each of them consists of 10 columns and 13 rows. Figures 5, 6 and 7 illustrate CV comparison of raw and purified data over holidays subsets.

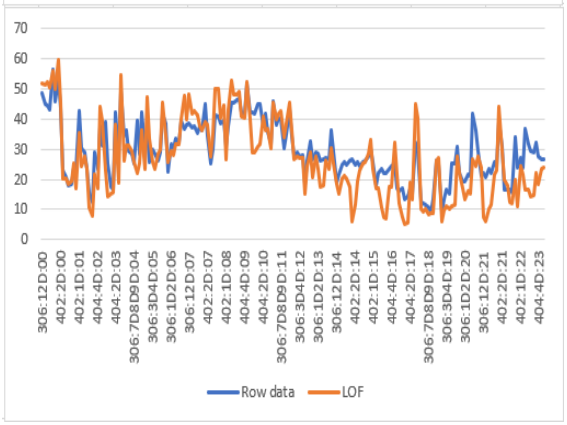

FIGURE 5. CV of LOF and raw data in holidays subsets

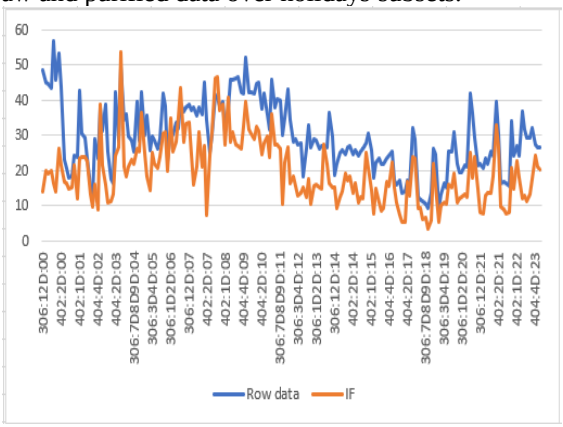

FIGURE 6. CV of IF and raw data in holidays subsets

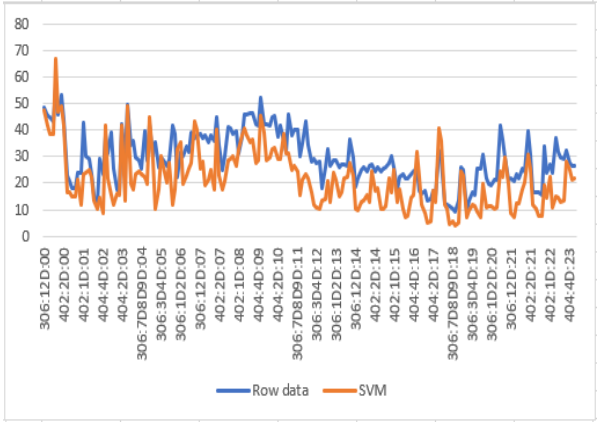

FIGURE 7. CV of IF and raw data in holidays subsets

Figures 5, 6 and 7 show that the CV of the data purified with each of the ODM increases for some samples in comparison with the raw data. However, these increases are observed at least in the data processed with IF. The CV of the raw data varies between 10 and 60%. The CV of the data purified with IF varies between 5 and 55%

CONCLUSION AND FUTURE WORK

The article presents the purification of traffic data with three popular outlier detection technics. The efficiency of the methods was evaluated by comparing the coefficient of variation in percentage of the raw and purified data.

The data set is divided into two parts - working days and holidays and is grouped by hour, in order to obtain groups of homogeneous data. 48 subsets are formed - for each hour of the day for working days and holidays. In the working days subsets the three algorithms achieved similar results. In the holidays subsets all outlier detection methods show increases in coefficient of variation for some samples of the purified data compared to the raw data. This result is probably due to the insufficient number of samples in the subsets. However, the Isolation Forest shows the best results, which confirms its stability with small set of data.

One direction for future developments is to use the averaged purified data as input of traffic simulation. The generalized model of traffic flows is very convenient due to the independence of extracting data for a specific time interval. It will be interesting to compare the averaged raw and purified data with data from another time period and to assess which of them is more representative for the traffic flow.

Another opportunity is to improve the methodology for the data purification by extending the size of dataset, or dividing the holidays dataset to subset.

ACKNOWLEDGMENTS

This research was partially supported by the European Regional Development Fund within the OP "Science and Education for Smart Growth 2014–2020", Project CoC "Smart Mechatronic, Eco- And Energy Saving Systems And Technologies", № BG05M2OP001-1.002-0023 23

I would like to express my special thanks to the Municipality of Plovdiv and in particular to Mr. Angel Velchev for the provided data and the time taken.

REFERENCES

1. M. Chowdhury, A. Apon, K. Dey, P. Bhavsar, N. Bouaynaya, R. R. Brooks, J. Deng, D. Dera, Y. Fu, V. N. Gudivada, O. Hambolu, N. Huynh, K. Kennedy, et al., *Data analytics for intelligent transportation systems 1st edition* (Elsevier, Amsterdam, 2017, pp. 44-81).
2. H. Y. The, A. W. K. Liehr and I. K. Wang, "Sensor Data Quality: a Systematic Review," in *Journal of Big Data* **7,** 11 (2020).
3. V. Astarita, V. P. Giofrè, G. Guido and A. Vitale, "The Use of Adaptive Traffic Signal Systems Based on Floating Car Data" *Wireless Communication Mobile Computing* **6** (2017)
4. C. Bachechi and L. Po, "Implementing an Urban Dynamic Traffic Model", *IEEE/WIC/ACM International Conference of Web Intelligence*, pp. 312-316 (2019)
5. S. Vergis, V. Komianos, G. Tsoumanis, A. Tsipis and K. Oikonomou, "A Low-cost Vehicular Traffic Monitoring System Using Fog Computing" *Smart Cities* **3**(1) pp.138–156 (2020)
6. Y. Boneva, T. Stoilov "Simulation of Tram Stops and Their Influence on Traffic – case study in Sofia, Bulgaria", *Information technologies and control* **3** (2019)
7. Y. Djenouri, A. Zimek, M. Chiarandini, "Outlier Detection in Urban Traffic Flow Distributions", *IEEE International Conference of Data Mining* pp. 935-940 (2018)
8. R. Bettinger and S. Spring, "Outlier Detection and Treatment", *DC SAS User Group* (2020)
9. J. Brownlee "*4 Automatic Outlier Detection Algorithms in Python*", available online: https://machinelearningmastery.com/model-based-outlier-detection-and-removal-in-python/
10. "*Applications for Python*", available online: https://www.python.org/about/apps/
11. *"Pandas Documentation",* available online: https://pandas.pydata.org/docs/
12. J. Brownlee "*A Gentle Introduction to Scikit-Learn: A Python Machine Learning Library*" available online: https://machinelearningmastery.com/a-gentle-introduction-to-scikit-learn-a-python-machine-learning-library/

13. "*Local Outlier Factor (LOF) — Algorithm for Outlier Identification"*, available online: https://towardsdatascience.com/local-outlier-factor-lof-algorithm-for-outlier-identification-8efb887d9843
14. K. Dhiraj "*Top 4 Advantages and Disadvantages of Support Vector Machine or SVM"* available online: https://dhirajkumarblog.medium.com/top-4-advantages-and-disadvantages-of-support-vector-machine-or-svm-a3c06a2b107
15. *"Isolation Forest Algorithm for Anomaly Detection"* available online: https://heartbeat.fritz.ai/isolation-forest-algorithm-for-anomaly-detection-2a4abd347a5