The Prediction of Traffic Flow with Regression Analysis

Ishteaque Alam, Dewan Md. Farid and Rosaldo J. F. Rossetti

Abstract Traffic data mining applying machine learning algorithms is necessary to analyse and understand the road traffic flow in busy cities. Also, it is very essential for making smart cities. Mining traffic data helps us to reduce travel delays and improve the city life. Currently, many cities in the developed countries use different sensors to collect the real-time traffic data and apply machine learning algorithms on the traffic data to improve the traffic condition. In this paper, we have collected the real-time traffic data from the city of Porto, Portugal, and applied five regression models: Linear Regression, Sequential Minimal Optimisation (SMO) Regression, Multilayer Perceptron, M5P model tree and Random Forest to predict/forecast the traffic flow of Porto city. Also, we have tested the performance of these regression models. The experimental results show that the M5P regression tree outperforms the other regression models.

Keywords Historical traffic data · Predictive model · Regression analysis Traffic flow forecast

1 Introduction

Mining traffic big data becomes an inevitable part in the way of building smart cities nowadays [3]. Every day we face a tremendous amount of trouble for the road traffic inconvenience. To improve the road traffic condition many cities are

I. Alam · D. Md. Farid (✉)
Department of Computer Science and Engineering, United International University, Dhaka, Bangladesh
e-mail: dewanfarid@cse.uiu.ac.bd
URL: http://cse.uiu.ac.bd

I. Alam
e-mail: ishteaque.ark@gmail.com

R. J. F. Rossetti
LIACC - Artificial Intelligence and Computer Science Laboratory, Faculty of Engineering, Department of Informatics Engineering, University of Porto, Porto, Portugal
e-mail: rossetti@fe.up.pt

A. Abraham et al. (eds.), *Emerging Technologies in Data Mining and Information Security*, Advances in Intelligent Systems and Computing 813,
https://doi.org/10.1007/978-981-13-1498-8_58

collecting road traffic information using modern technologies. Then, this information is analysed to understand the traffic flow and its characteristics to improve real-time Traffic Management System (TMS). Knowledge extraction from big traffic data applying machine learning and data mining algorithms has become an essential part for building Intelligent Transportation System (ITS) [13, 22]. We need real-time traffic information to take smart decision to minimise the traffic delay that also saves our time to move from one place to another [12]. Several types of sensors and wireless communication devices such as Wi-Fi, Bluetooth and GPS are used to collect real-time traffic data recently, and then the traffic data is analysed for traffic flow forecasting and decision-making for Traffic Management Systems by applying different machine learning algorithms [25].

Controlling the road traffic is always a challenging task. Still in many cities, the traditional traffic control devices like traffic lights and human traffic police are using to maintain the traffic flow [15]. However, the traditional traffic control approaches are not sufficient enough to improve road traffic condition in the busy cities like New York, Paris, London, etc. To improve the traffic management systems, intelligent computational researchers are applying several machine learning and data mining tools and algorithms for mining traffic data nowadays. Analysing historical traffic data helps us to minimise road accidents and travel delays. A traffic accident prediction model for Western Desert Road at Aswan city was developed using fuzzy modelling to examine and build an accident prediction model based on historical traffic data collected from the year 2011 to 2015 [14]. This model was also used for transportation management and planning on the desert roads to reduce highway accidents. Luis et al. [17] conducted a survey at Quito Metropolitan District and trained an artificial neural network (ANN) to evaluate the willingness to pay among people to minimise the road traffic noises. Oruc et al. [4] collected floating car data from Ankara City and extracted critical traffic patterns for urban traffic. Sasan et al. [5] came up with a real-time traffic control architecture stand on big data analytics. Although the study could not access the real traffic data, the architecture is capable to accommodate different data storage settings and analytical engines.

This work extends our previous work [3], where we developed a traffic data visualisation tool and applied regression models to understand the relationship among the features in traffic dataset. In this paper, we have taken a real-time traffic data from City of Porto, Portugal for 3 years from 2013 to 2015. Then, we have applied five regression models: Linear Regression, Sequential Minimal Optimisation (SMO) Regression, Multilayer Perceptron, M5P model tree and Random Forest on the data to forecast the real traffic flow of Porto city. The 23 sensors that are placed at the different streets of the Porto city collected the traffic data. Initially, we have used Python high-level programming language to visualise and pre-process the traffic data. Then, we have used Weka (Waikato Environment for Knowledge Analysis) libraries to develop regression models. Finally, we compared the predicted results of regression models in comparison with the actual real-time traffic flow.

The rest of the paper is organised as follows: Sect. 2 presents the related works regarding road traffic flow prediction. Section 3 presents the regression models that used for forecasting traffic flow. Section 4 presents the traffic data that collected from city of Porto, Portugal, and the experimental results. Finally, conclusion and future work is discussed in Sect. 5.

2 Related Work

Predicting short-term and long-term traffic flow is necessary for building smart cities. Xiaobo et al. [11] presented a short-term traffic flow prediction model applying hybrid genetic algorithm-based LSSVR (Least Squared Support Vector Regression) model using the dataset collected from 24 observation sites from freeways of Portland, United States. One advantage of the method is that it can interpret the relationship among the spatiotemporal variables.

Nicholas et al. [19] proposed an innovative architecture based on deep learning to predict traffic flow. The architecture was combined with a linear model with another sequence of tenth layers. The study identifies that deep learning performs better over linear models for prediction. Yalda et al. [21] used adaptive model to develop a real-time data flow prediction model. A two-step approach was employed to capture uncertainty. The generality of their method was tested through an open-access database called PeMS. PeMS is a performance measurement system that also provides a variety of support for historical data analysis. Their proposed method was executed for imputation of missing data and their results presented that the method is more efficient in comparison with PPCA and k-NN. Xianyao et al. [16] also proposed an algorithm (APSO-MSVM) based on Adaptive Particle Swarm Optimisation and Multi-kernel Support Vector Machine (MSVM) for short-term traffic flow prediction. The study collected real data from roadside units (RSU) and the algorithm had significantly lower error rate on both freeway and urban road.

In 2014, Marcin et al. [8] presented an approach to boost the performance of k-Nearest Neighbour (kNN) for predicting road traffic condition. The paper investigated the data segmentation method to figure out useful neighbours for better prediction. Their experimental result had a better accuracy by searching the nearest neighbour among the useful neighbours. In 2015, Afshin et al. [1] proposed a methodology to predict road traffic till 30 min ahead by linking their estimated demand with a limited real-time data. The study used Monte Carlo method for the evaluation of the algorithm in San Francisco and California road network. In 2016, Jinyoung et al. [2] presented a method to predict traffic flow in their paper using Bayesian Classifier and SVR. The performance of their method was later tested on the traffic data collected from Gyeongbu Expressway. Dick et al. [7] recommended a road traffic estimation model for less occupied roads in Wyoming. The study proposed a linear and another logistic regression models for estimating the traffic level.

In 2017, Hubert and Jerome [6] explained a methodology for an Ensemble Kalman Filter depending on the static detector measurements and vehicle data to estimate future traffic velocity and density. Their methodology was applied over a single road to simplify the task. Xi et al. [10] studied traffic demand by conducting an experimental case study on a school bus and people's psychological characteristics. Finally, they made a forecast model on traffic demand for different sunny, rainy and exam days based on the survey.

3 Regression Analysis

Regression analysis is often performed in data mining to study or estimate the relationship among several predictors or variables. Generally, a regression model gets created to perform analysis over a dataset. Then, it uses least squared method to estimate the parameter that fits the best. The model gets verified using one or more hypothesis tests. In this paper, we have applied Linear Regression, SMO Regression, Multilayer Perceptron, M5P model tree and Random Forest to forecast the future traffic flow for the fourth week of a month.

3.1 Linear Regression

Linear model predicts one variable based on another using statistical method [18]. It uses the equation of a line to draw an estimation line to predict result of a dependent variable from the result over an explanatory variable. The objective of a simple linear regression is to create best fit line that minimises the sum of the squared residuals. This is used to establish the relationship between two variables or to find a possible statistical relationship between the two variables.

3.2 SMO Regression

Sequential Minimal Optimisation (SMO) regression is an algorithm that deals solving the quadratic programming (QP) problem. The algorithm is widely used to find solutions to Support Vector Machine (SVM). The algorithm continues running the iteration loop to reach the most promising combination of pairs to optimise weights. The model also optimises time-series forecasting by reducing the operation runtime than general SVM algorithm as it avoids performing quadratic programming [24].

3.3 Multilayer Perceptron

The perceptron consists of weights and summation process along with activation function. Multilayer perceptron has the same functionality as a single perceptron, which is a class of artificial neural network (ANN) with one or more hidden layers.

One hidden layer gets connected to the input layer and using the inputs and weights, it forward passes the outputs from one layer to the next. The output gets computed utilising the sigmoid function as like nonlinear transfer functions.

3.4 M5Base Regression Tree

M5Base also known as M5P is basically a method for creating regression tree by using m5 algorithm. It implements a base routine for generating m5 model tree by using the separate-and-conquer technique to create decision lists. The tree is generated by employing a base routine, just as applying piecewise function over several linear models [20]. Later, an improved version of the algorithm was developed to handle missing feature values effectively by adding them artificially [23].

3.5 Random Forest

Random Forest is an often used algorithm in data science to make decision generated from random trees [9]. The algorithm combines a large number of decision trees to make a classification using bagging technique. The technique combines a number of learning models in order to increase the classification accuracy. This powerful supervised machine learning algorithm is capable to perform both classification and regression tasks. It chooses the classification, having the most votes among the generated decision trees. And, in the case of regression, it takes the average of the outputs calculated by different trees.

4 Experimental Analysis

4.1 Traffic Dataset

A real-time traffic data was collected from the city of Porto, which is the second largest city in Portugal. The city is very well known for its impressive bridges and wine production. However, the road traffic data was collected via 23 sensors placed on the city roads. Most of the sensors were positioned in Via de Cintura Interna (VCI). This road has a length of 21 Km portrayed as a ring-shaped motorway. The road is also referred as tambm chamada IC23. All the data collected through each sensor gets stored in PostgreSQL. This is an open-source SQL standard object-relational database management system, which supports a variety of data types. Students are allowed to use the dataset for research purpose without any cost. Table 1 shows the detail explanation of each column attributes of the collected dataset sensed through

Table 1 Feature description of traffic dataset

Name	Description
Equipment ID	Sensor ID
LANE_NR	Lane number
LANE_DIRECTION	Lane direction
TOTAL_VOLUME	Number of vehicles
AVG_SPEED_ARITHMETIC	Arithmetic average speed of vehicles
AVG_SPEED_HARMONIC	Harmonic average speed of vehicles
AVG_LENGTH	Average length of the vehicles
AVG_SPACING	Average spacing between vehicles
OCCUPANCY	Occupancy
INVERTED_CIRCULATION	No. of vehicles pass through with opposite direction
LIGHT_VEHICLE_RATE	The % is the ratio of light vehicles
VOLUME_CLASSE_A	No. of vehicles of class A
VOLUME_CLASSE_B	No. of vehicles of class B
VOLUME_CLASSE_C	No. of vehicles of class C
VOLUME_CLASSE_D	No. of vehicles of class D
VOLUME_CLASSE_0	No. of unidentified vehicles
AXLE_CLASS_VOLUMES	No. of vehicles per vehicle axis number
AGGREGATE_BY_LANEID	Is a unique ID fusing two IDs, the lane ID and the lane direction
AGG_PERIOD_LEN_MINS	Is the time between each measurement the sensors made. It is always 5 min
AGG_PERIOD_START	Time interval
NR_LANES	No. of lanes in the street of the event
AGGREGATE_BY_LANE_BUNDLEID	Similar to AGGREGATE_BY_LANEID
AGG_ID	Street ID

each sensor. Our initial dataset contained 2,83,97,311 instances starting from the year 2013 to 2015. In this paper, we present the forecast mechanism and experimental results performed focusing on the sub-dataset of the month April 2014. The study predicts future traffic flow of the month utilising previous historical road traffic data. The sub-dataset April 2014 has more than 10 lakh (10,39,077) instances.

4.2 Data Pre-processing

Our collected dataset had 23 attributes. We applied feature selection method to exclude some redundant attributes that are not useful features for forecasting the traffic flow, such as, AGG_PERIOD_LEN_MINS, INVERTED_CIRCULATION,

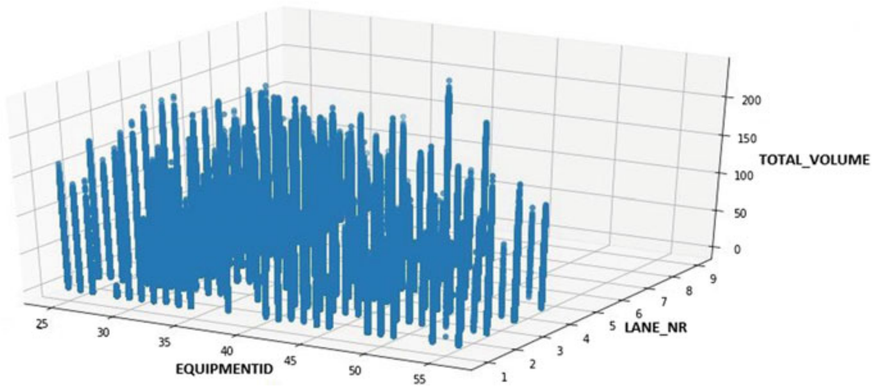

Fig. 1 Total volume of each EquipmentID and lane number for April 2014 of Porto City

etc. The main objective of this paper was to apply machine learning algorithms to forecast the Total Volume of the traffic flow. Therefore, other volume classes were not important to consider. The dataset had no missing values, but yet it was difficult to visualise for the randomness in between. To clear the randomness of the dataset, we serialised it utilising Pandas library from Python programming. In recent times, Pandas is a very dominant data analysis module provided by Python, which is also open source. Next, we plotted some figures to make sense of the dataset. Figure 1 explains that lane number varies with respect to each Equipment id. Equipment id represents the sensor's id that is placed on different roads of the city. So, we break the dataset with respect to each equipment id to forecast the traffic flow for each of the roads. Afterward, we merge together total volume of all the lanes of each road for a particular time. In this approach, we take the first 3 weeks data of the month and try to make a model that can predict the fourth week's traffic flow for the roads of Porto city. We have used five different regression models to perform the prediction.

4.3 Experimental Setup

This study has used WEKA which has a vast collection of machine learning algorithms, and it is widely popular in the field of data mining. A new package was installed in the workbench named 'timeseriesForecasting', which is specially made for time-series analysing. This new framework uses machine learning approaches to make models for time series. It transforms the data into a processable format for standard propositional learning algorithms.

Our train dataset contained the first three weeks preprocessed data for the month April 2014 of Porto city. This was use to train each of the regression models. Then, we set the number of units to forecast to 7 with a daily periodicity for predicting traffic flow for the upcoming week. We generated a linear model where m5 method was used for attribute selection with a ridge parameter set to default 1.0e−8. For SMO regression, we used normalised training data as filter type for better data integrity. Everything else was left with the default setup of the workbench. A lag variable was generated by the package itself to forecast based on the time series.

4.4 Experimental Results

The forecasting models generated by each of the machine learning algorithms were implemented for future traffic flow prediction. Figure 2 shows the patterns yielded by each of those algorithms. Figure 3 focuses specifically on the predicted pattern (From the date 22 to 28). Afterwards, the accuracy of the algorithms was measured by comparing the mean absolute errors. Table 2 describes the error rate for each of the regression models. The worst performance was found by the linear model with the highest error rate and the best performance was found by the M5P regression model with the lowest error rate of 2.88%. We have only presented the experimental result for a single road in Porto city in this paper. However, we applied the same technique over other roads of the city and find out that M5P performs the best over all the other algorithms.

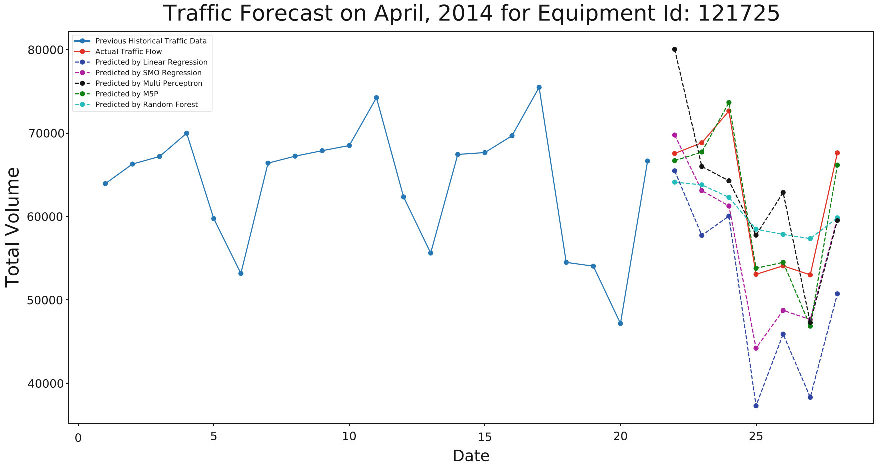

Fig. 2 Prediction of traffic flow by regression models for the fourth week of April 2014 based on previous 3 weeks of historical traffic data

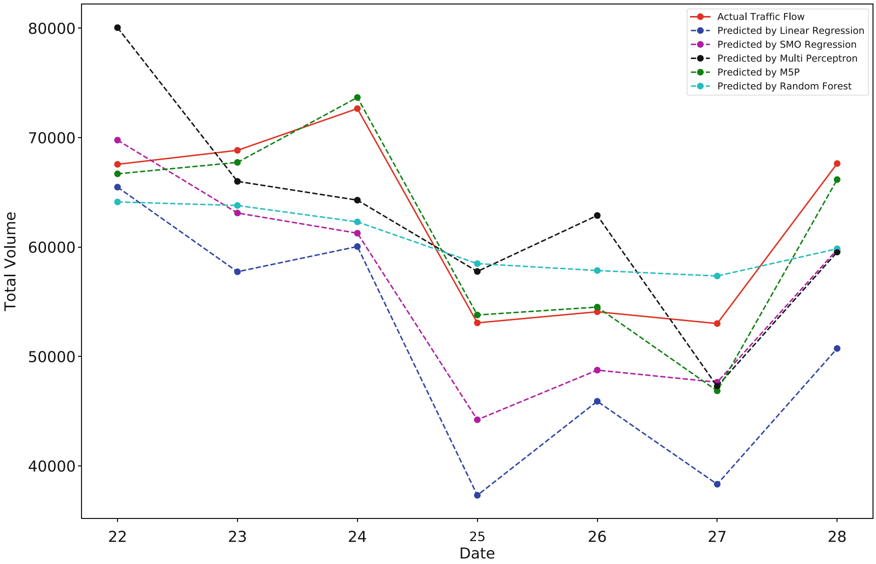

Fig. 3 Prediction comparison of the traffic flow of models with real traffic data for the fourth week of April 2014

Table 2 Result

Day	Linear regression (%)	SMO regression (%)	Multilayer perceptron (%)	M5P model tree (%)	Random forest (%)
22nd Apr	3.08	3.26	18.48	1.28	5.08
23nd Apr	16.12	8.32	4.12	1.60	7.32
24nd Apr	17.34	15.67	11.51	1.40	14.24
25nd Apr	29.69	16.68	8.84	1.35	10.19
26nd Apr	15.12	9.86	16.26	0.78	6.96
27nd Apr	27.69	10.10	10.81	11.60	8.19
28nd Apr	24.98	11.63	11.95	2.16	11.51
MAE	19.15	10.79	11.71	2.88	9.07

5 Conclusion

Predicting road traffic flow is one of the major challenges over the past decade in the field of data mining. This paper presented a new innovative approach to predict real traffic flow in a long term utilising popular machine learning algorithms. The study used Linear Regression, SMO Regression, Multilayer Perceptron, M5P model tree, and Random Forest. The traffic dataset from Porto city was collected and processed using Python programming. Later on, we applied the regression algorithms to generate patterns of traffic flow for the fourth week of each month using historical traffic data of the previous 3 weeks of that month. The results of the models were compared with the actual traffic flow. M5P model tree outperformed every other models with the lowest mean absolute error rate. For a particular road that we showed in this work, the error rate was 2.88%. This paper establishes high possibility of M5P algorithm in terms of forecasting real-world traffic flow. The same approach can be applied to forecast the traffic flow for a whole month also. In future, we would like to present a traffic flow visualising and forecasting system for the capital of Bangladesh, Dhaka City.

Acknowledgements We would like to thank "Armis" (http://www.armis.pt/) for sharing the traffic data.

References

1. Abadi, A., Rajabioun, T., Ioannou, P.A.: Traffic flow prediction for road transportation networks with limited traffic data. IEEE Trans. Intell. Transp. Syst. **16**(2), 653–662 (2015)
2. Ahn, J., Ko, E., Kim, E.Y.: Highway traffic flow prediction using support vector regression and Bayesian classifier. In: 2016 International Conference on Big Data and Smart Computing (BigComp), pp. 239–244. IEEE (2016)
3. Alam, I., Ahmed, M.F., Alam, M., Ulisses, J., Farid, D.M., Shatabda, S., Rossetti, R.J.F.: Pattern mining from historical traffic big data. In: IEEE Technologies for Smart Cities (TENSYMP 2017), pp. 1–5. IEEE (2017)
4. Altintasi, O., Yaman, H.T., Tuncay, K.: Detection of urban traffic patterns from floating car data (FCD). Transp. Res. Procedia **22**, 382–391 (2017)
5. Amini, S., Gerostathopoulos, I., Prehofer, C.: Big data analytics architecture for real-time traffic control. In: 2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS), pp. 710–715. IEEE (2017)
6. Andre, H., Ny, J.L.: A differentially private ensemble Kalman filter for road traffic estimation. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6409–6413. IEEE (2017)
7. Apronti, D., Ksaibati, K., Gerow, K., Hepner, J.J.: Estimating traffic volume on Wyoming low volume roads using linear and logistic regression methods. J. Traffic Transp. Eng. (Engl. Ed.) **3**(6), 493–506 (2016)
8. Bernaś, M., Płaczek, B., Porwik, P., Pamuła, T.: Segmentation of vehicle detector data for improved k-nearest neighbours-based traffic flow prediction. IET Intell. Transp. Syst. **9**(3), 264–274 (2014)
9. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
10. Chen, X., Peng, L., Zhang, M., Li, W.: A public traffic demand forecast method based on computational experiments. IEEE Trans. Intell. Transp. Syst. **18**(4), 984–995 (2017)
11. Chen, X., Wei, Z., Liu, X., Cai, Y., Li, Z., Zhao, F.: Spatiotemporal variable and parameter selection using sparse hybrid genetic algorithm for traffic flow forecasting. Int. J. Distrib. Sens. Netw. **13**(6), 1550147717713,376 (2017)
12. Csikós, A., Charalambous, T., Farhadi, H., Kulcsár, B., Wymeersch, H.: Network traffic flow optimization under performance constraints. Transp. Res. Part C: Emerg. Technol. **83**, 120–133 (2017)
13. Dell'Orco, M., Marinelli, M.: Modeling the dynamic effect of information on drivers' choice behavior in the context of an advanced traveler information system. Transp. Res. Part C: Emerg. Technol. **85**, 168–183 (2017)

14. Gaber, M., Wahaballa, A.M., Othman, A.M., Diab, A.: Traffic accidents prediction model using fuzzy logic: Aswan desert road case study. J. Eng. Sci. Assiut Univ. **45**, 2844 (2017)
15. Ghorghi, F.B., Zhou, H.: Traffic control devices for deterring wrong-way driving: historical evolution and current practice. J. Traffic Transp. Eng. **4**, 280–289 (2017)
16. Ling, X., Feng, X., Chen, Z., Xu, Y., Zheng, H.: Short-term traffic flow prediction with optimized multi-kernel support vector machine. In: 2017 IEEE Congress on Evolutionary Computation (CEC), pp. 294–300. IEEE (2017)
17. Moncayo, L.B., Naranjo, J.L., García, I.P., Mosquera, R.: Neural based contingent valuation of road traffic noise. Transp. Res. Part D: Transp. Environ. **50**, 26–39 (2017)
18. Montgomery, D.C., Peck, E.A., Vining, G.G.: Introduction to Linear Regression Analysis, vol. 821. Wiley (2012)
19. Polson, N.G., Sokolov, V.O.: Deep learning for short-term traffic flow prediction. Transp. Res. Part C: Emerg. Technol. **79**, 1–17 (2017)
20. Quinlan, J.R., et al.: Learning with continuous classes. In: 5th Australian Joint Conference on Artificial Intelligence, vol. 92, pp. 343–348. Singapore (1992)
21. Rajabzadeh, Y., Rezaie, A.H., Amindavar, H.: Short-term traffic flow prediction using time-varying Vasicek model. Transp. Res. Part C: Emerg. Technol. **74**, 168–181 (2017)
22. Talebpour, A., Mahmassani, H.S., Hamdar, S.H.: Effect of information availability on stability of traffic flow: percolation theory approach. Transp. Res. Procedia **23**, 81–100 (2017)
23. Wang, Y., Witten, I.H.: Inducing model trees for continuous classes. In: Proceedings of the Ninth European Conference on Machine Learning, pp. 128–137 (1997)
24. Yang, J.F., Zhai, Y.J., Xu, D.P., Han, P.: SMO algorithm applied in time series model building and forecast. In: 2007 International Conference on Machine Learning and Cybernetics, vol. 4, pp. 2395–2400. IEEE (2007)
25. Zhou, M., Qu, X., Li, X.: A recurrent neural network based microscopic car following model to predict traffic oscillation. Transp. Res. Part C: Emerg. Technol. **84**, 245–264 (2017)