ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

Survey on Classification and Prediction Approaches in Traffic Flow

Riya R. Gandewar[1], Anupama G. Phakatkar[2]

M.E Student, Department of Computer Engineering, Pune Institute of Technology, Pune, India[1]

Asst. Professor, Department of Computer Engineering, Pune Institute of Technology, Pune, India[2]

ABSTRACT: Obtaining accurate information about current and near-term future traffic flows of all links in a traffic network has a wide range of applications, including traffic forecasting, vehicle navigation devices, vehicle routing, and congestion management. In big data driven traffic flow prediction systems, accuracy and timeliness affects the robustness of prediction performance. However, the classification of large amounts of data is becoming a necessary task in a great number of real-world applications. This topic is known as big data classification, in which standard data mining techniques normally fail to tackle such volume of data. This model allows us to simultaneously classify large amounts of unseen cases against a big dataset. This system will use various classification approaches for Big-data-driven traffic flow prediction using Apache Spark. Apache Spark is used here for processing, which overcomes all the necessary drawback of previous system. So, automatically processing time gets decreases and memory utilized efficiently.

KEYWORDS: Traffic flow prediction, Apache Spark, Classification Approaches, Hadoop, etc.

I. INTRODUCTION

Traffic problems are crucial issues in the rapidly developing society. Traffic flow prediction can be an important problem in Intelligent Transportation System. Intelligent transport systems collect traffic data such as traffic volume and speed on every roads and provide statistical summary services, usually on traffic congestion. Traffic congestion affects on Vehicular queuing, travel time, cost, fuel consumption, pollution in the environment. In a big city, the change of traffic flow has a large impact on people's daily life, such as the route selection for drivers. The big data generated by the Intelligent Transportation Systems are worth further exploring to traffic management. The introduction and development of the intelligent transport system has resulted in more reliable traffic information gathering, analysing, and processing, thereby providing more time-relevant and precise traffic analytic and prediction to users.

The classification of big data is becoming an essential task in a wide variety of fields such as biomedicine, social media, marketing, etc. The recent advances in data gathering in many of these fields have resulted in an inexorable increment of the data that we have to manage. The volume, diversity and complexity that bring big data may hinder the analysis and knowledge extraction processes. Under this scenario, standard data mining models need to be re-designed or adapted to deal with this data.

A variety of techniques have been applied in the context of short-term traffic flow forecasting, including moving average methods, k-nearest-neighbor methods, auto regressive MA (ARIMA) or seasonal ARIMA (SARIMA) model, and neural networks (NNs). Forecasting tra□c flow heavily depends on historical and real-time tra□c data collected from various sensor sources, such as inductive loops, radars, cameras, mobile Global Positioning System, social media, etc.

Deep learning has drawn a lot of academic and industrial attention, and it has been applied with success in the tasks of classification, natural language processing, dimensionality reduction, object detection, motion modelling, etc. There are also early attempts to apply deep learning to tra□c flow forecasting.

As data is growing in nature, It is challenge to store and process this type of data. If the data is big in size then processing of different algorithms on data takes lots of time. For storing big data Apache Hadoop's HDFS is used. Data

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

is valuable, so storing all the data at one place is very risky. Hadoop distributed file system provides replication of data nodes on distributed connected nodes, which achieves fault tolerance.
Big data analysis takes longer time to draw necessary conclusions. Thus, require faster processing approach that will speed up the computation by distributive performs local computations on set of connected commodity hardware and return the result on master node. Apache Spark provides faster processing, which will significantly decrease overall execution time of our big data application. We discuss few technologies in Section 4.

Traffic flow prediction Process
Any data to be classified needs to go to certain classification process. But, as data is big in size and need to modify process to perform distributed operations compatible with processing framework, Hadoop.

A. Data Collection

The data is collected from different sources. The data contains its own identifier and many more attributes. These data is classified to our required categories. We also need to prepare our dataset train our classifier.

B. Pre-processing Data

Data pre-processing is done through various ways. In pre-processing missing data and noise is identified. Missing data is calculated through different ways. If missing data is in large in quantity then we have to calculate that data with more accurate algorithms.

C. Classification

For data to be classified into different categories, we considered using supervised machine learning algorithms. This will discuss more in detail in section 2. There are bunch of labeled points and uses them to learn how to label other points.

D. Prediction

Prediction is made from historical data. There are various prediction techniques, which comes up with different prediction results according to their algorithms. These algorithms will discuss more in detail in section 3. The accuracy of prediction is measured through various parameters.

E. Performance Measures

Following are the measures for performance evaluation:

1. Accuracy

Classifier accuracy: the ability of a classifier to predict class labels.
Predictor accuracy: how close is the predicted value from true one.

2. Speed

Time to construct the model (training time).
Time to use the model (classification/prediction time).

3. Robustness

Handling noise and missing values

4. Scalability

Efficiency in disk-resident databases.

5. Interpretability

II. RELATED WORK

Researchers have been trying to make traffic flow forecasting more accurate since last few years. Based on the length of time interval, there are long-term forecasting and short-term forecasting.
Long-term forecasting indicates making a prediction by month or by year. The volume of traffic flow is large and relatively stable, and it is slightly affected by daily accident. For example, Papagiannaki et al. [23] proposed a method

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

based on ARIMA to forecast traffic flow in a month after they revealed that their time series data had the long term trend and the fluctuations at the 12 hour time scale.

For short term forecasting, auto regressive integrated moving average (ARIMA) [17] models and artificial neural network (ANN) [19] models are widely exploited. Lv et al. [14] proposed a plan moving average algorithm for utilizing previous days historical data. In addition, some researchers have compared the seasonal ARIMA (SARIMA) model, which is a TSA model, and showed its excellent prediction performance.Smith et al.[21] compared SARIMA and nonparametric (data-driven regression) models. The authors in this research concluded that the SARIMA model has better performance than the nonparametric regression models.

Lippi et al.[20] additionally compared time-series analysis methods and SVR models and concluded that SARIMA showed the best performance. In addition, there are some integration models for this task. For example, Chang et al. [15] proposed a 3-stage model which integrates ARIMA and ANN, which uses the ARIMA forecasting data as a part of input of ANN. Moretti et al. [16] studied an ensemble model which consists of a statistical method and a neural network bagging model.

Yuqi Wang and Wengen Li proposed method for predicting traffic congestion correlation between road segments on GPS trajectories [22]. Method extract various features on each pair of road segments from road network. Result of this is input to the several classifiers to predict congestion correlation. It use classifiers like decision tree, Logistic regression, Random forest and Support vector machines. Besides, researchers also consider the use of integrated data.

As another example, Oh et al. proposed a Gaussian mixture model clustering (GMM) method to partition the data set for training ANN. Deep learning methods have also gained a lot of attention recently. Lv et al. [17] used a stacked autoencoder (SAE) to learn generic traffic flow features. Huang, Song and Huang et al.applied a deep belief networks model in traffic flow prediction, which adopts multitask learning to reduce the error.

To sum up, all the above mentioned methods have many desirable properties in different disciplines, and thus it is hard to conclude that which one of these is significantly superior to other methods in any situation. One of the best essential reasons is that the accuracy of prediction models which are developed with small-scale separate specific traffic data depends on the traffic flow features embedded in the collected traffic data. Furthermore, most of existing models are performed in stand-alone modes, and thus the computational effort is “expensive” [18] and the capability of data processing and storage is restricted. However, researchers also developed a general architecture of distributed modeling in a MapReduce framework for traffic flow forecasting, to efficiently process large-scale traffic data on a Hadoop platform.

As of now, there is diversified research in all the techniques and platform. This paper suggests, Traffic flow prediction of large volume traffic data using classification approach in Apache Spark.

III. CLASSIFICATION ALGORITHMS

Classification: This type of machine learning algorithm focuses on outcome variable and to be predicted from set of independent variables. Using our training dataset, it classifies data in different required categories. Training phase continues till our model gets higher accuracy.

Following are approaches of Machine learning:

1) Naive Bayes:

The Naive Bayes Classifier technique is based on Bayesian theorem. This technique is particularly used when the dimensionality of the inputs is too high. This Classifier is capable of calculating the most possible output based on the input. In this technique, It is also possible to add new raw data at runtime and have a better probabilistic classifier [5]. A naive Bayes classifier considers that the presence of a particular attribute of a class is unrelated to the presence of any other attribute when the class variable is given. Naive Bayes is has strong feature independence assumptions.

2) Support Vector Machines:

SVM are based on statistical learning theory and structural risk minimization principal. It has the aim of determining the location of decision boundaries also known as hyper plane that produce the optimal separation of classes. Viewing input data as two sets of vectors in an Viewing input data as two sets of vectors in an non-dimensional space, an SVM will construct a separating hyper-plane in that space, one which maximizes the margin between the two data sets

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

dimensional space, an SVM will construct a separating hyper-plane in that space, one which maximizes the margin between the two data sets. To calculate the margin, two parallel hyper-planes are constructed. A good separation is achieved by the hyper-plane that has the largest distance from the neighboring data points of both classes. In general the larger the margin the lower the generalization error of the classifier hyper-plane that has the largest distance to the neighboring data points of both classes [5]. In SVM, we have to deal with features that are truly relevant. The idea of SVMs is to find such linear separators as most text categorization problems are linearly separable. SVM is more suited for large and sparse instances.

3) K-Nearest Neighbor

K-nearest neighbor has nonparametric, small error ratio and good error distribution. The basic process of k-nearest neighbor prediction model is to build a representative historical database with large capacity. Then, set the model elements, including the state vector value of k and prediction algorithm [13]. The state vector and value of k is search mechanism. Finally, according to the observed values of the input and search mechanism, a close neighbor matching the current real-time observation data from the history database are picked up to predict the traffic flow at the next time.

The k-NN nonparametric regression method is a classic data-driven method for short-term traffic flow forecasting. This method argues that the traffic flow vector at the same clock time of the current traffic flow vector is viewed as a neighbor. According to the distances between the neighbors and the current traffic flow, the k-nearest neighbors are picked and the predicted traffic flow is obtained by using the corresponding outputs of these neighbors.

4) Artificial Neural Network

An artificial neural network operates by creating connections between many different processing elements, called as neuron. Each neuron takes many input signals and then based on an internal weighting produces a single output signal. That output is sent as input to another neuron [11]. These neurons are strongly interconnected and organized into different layers. The input layer receives the input and the output layer produces the final output. In general one or more hidden layers are there between the two input layer and output layer. This structure makes it impossible to forecast or know the exact flow of data.

Initially they must be trained to solve the particular problem for which they are proposed. A back-propagation ANN is trained by humans to perform specific tasks. During the training period, we can evaluate whether the ANN's output is correct by observing pattern. If it's correct the neural weightings that produced that output are reinforced; if the output is incorrect, those weightings responsible can be diminished. The parallel architecture allows ANNs to process very large amounts of data very efficiently in less time.

5) Decision Trees

Regarding decision trees, they are tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees. The most commonly used are the classification trees that are usually represented graphically as hierarchical structures, making them easier to interpret than other techniques. Classification trees are used to classify an object or an instance to a predefined set of classes. Decision tree generally consists of nodes that form a rooted tree. It is a directed tree with a node called a root that has no incoming edges. All other nodes have exactly one incoming edge. A node with outgoing edges is referred to as an internal node. All other nodes are called leaves.

In the decision tree, each internal node splits the instance space into two or more sub-spaces [9]. This split is done according to a certain discrete function of the input attribute values. In the simplest way each test considers a single attribute, such that the instance space is partitioned according to the attributes value. In the case of numeric attributes, each leaf is assigned to one class representing the most appropriate target value. Alternatively, the leaf may hold a probability vector affinity vector indicating the probability of the target attribute having a certain value.

6) Random Forest

It has developed an ensemble classification approach that displayed outstanding performance with regard prediction error on a suite of benchmark datasets [10]. Random Forest is a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error for forests assemble as to a limit as the number of trees in the forest becomes large. The

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them. Internal estimates monitor error, strength, and correlation and these are used to show the response to increasing the number of features used in the splitting. The common element in all of these procedures is that for the k-th tree, a random vector θ_k is generated, independent of the past random vectors (θ_1 to θ_{k-1}), but with the same distribution, and a tree is grown using the training set and θ_k , resulting in a classifier h(x,θ_k) where x is an input random vector.

Approaches	Pros	Cons
Decision Tree	1. Easy to interpret. 2. Classification without much calculation.	1. High Classification error rate. 2. Exponential calculation growth.
SVM	1. Useful for both Linearly separable and non-linearlyseparable data 2. Guaranteed optimality.	1. They can be painfully inefficient to train.
Naïve Bayes	1. Data set is small than high bias low variance classifier.	1. Data set is small than generative class will work well.
KNN	1. No training involved. 2. Simple and Powerful.	1. Expensive and slow.

Table-1: Analysis of different classification approaches

IV. PREDICTION ALGORITHMS

1) Linear Regression

We use linear regression to approximate the correlation between future traffic data and current traffic data. In this paper, multiple linear regressions will be used to decide the statistical correlation. It is a generalization of linear regression by considering more than one independent variable, and a specific case of general linear models formed by restricting the number of dependent variables to one [5]. It formalizes a simultaneous statistical relation between the single continuous outcome and the predictor variables.

Given a training data $D= \{(\ _t, y_t) \mid t = 1,2, \dots, n\}$, where each $x_t \subset R_n$ denotes the input dimension of the clique potentials and has a corresponding target value $y_t \subset R$ during n time indexes. The goal of the regression is to fit a function $g(x_t)$ which approximates the relation between the data set points and it can be used later to infer the output for a new input data point. Suppose the traffic condition yy in a road is influenced by the velocity of its neighbor's x_t on the defined cliques.

Thus the multiple linear regression model can be described as follows.

$y_t = (x_t) = \theta_0 + \theta_1 x_{t,1} + \theta_2 x_{t,2} + \cdots + \theta_k x_{t,k} + \varepsilon_t$

2) ARIMA model

It is the autoregressive integrated moving average model.1. The training data is preprocessed to get rid of the error data, and is classified into different data sets according to different time periods. 2. ARIMA model structure recognition: using the autocorrelation and partial autocorrelation to decide the step number of (p,q) for the ARIMA model. 3. Then, the ARIMA model parameters are estimated based on the data sets. 4. Model testing is to test whether the (p, q) is reasonable. If it can not pass the test, the model structure should be redefined. 5. Based on the model with the estimated parameters, the traffic flow is predicted [12].

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

3) Neural network

Multi-layers Convolutional Neural Network (CNN) contains convolution layer, pooling layer, loss-function layer, etc. Based on di□erent types of loss function layers, we can perform di□erent tasks such as classification and regression [2]. With the assist of convolution layer and pooling layer, CNN can help understanding the relations between the features. Properly handled, CNN can be a powerful tool for modeling the future tra□c flow. To use the 2-dimensional temporal-spatial tra□c flow features for forecasting, we adopt the Convolutional Neural Network (CNN) with its capacity to process the temporal-spatial features at the same time. CNN is built by convolution layers, sub-sampling layers, and full connected layers, etc. It stands with an array of alternating convolution and sub-sampling operations, and then continues with generic multi-layer network. The last few layers (closest to the outputs) are fully connected with 1-dimensional layers [2] for regression analysis.

V. CONCLUSION

In case of big data, KNN gives more accuracy and less execution time for classifying data. CNN gives more accurate prediction results because of their hidden and complex layer mechanism. It is important to processsuch big data in a distributive manner reducing overall execution time. There is communication overhead and latency involved for fetching data from various sources. Big data technologies are used to efficiently classify and predict data that is scalable and provides high performance computing.

REFERENCES

1. Su Yang, Shixiong Shi, Xiaobing Hu, “Discovering Spatial temporal weighted model on Map- Reduce for short term traffic low forecasting ”, IEEE international conference on transportation, pp. 364 – 367, 2015.
2. Donghai Yu, Yang Liu, and Xiaohui Yu, “A Data Grouping CNN Algorithm for Short-Term Tra□c flow Forecasting”,2016.
3. YuqiWang ,Jiannong Cao , Wengen Li and Tao Gu ,”Mining Traffic Congestion Correlation between Road Segments on GPS Trajectories” , IEEE, pp. 1 – 8, 2016.
4. Hao-Fan Yang, Tharam S. Dillon, Life Fellow, IEEE, and Yi-Ping, “Optimized Structure of the Traffic flow Forecasting Model With a Deep Learning Approach” IEEE transactions on neural network, pp. 1 – 11, 2016.
5. JinyoungAhn, EunjeongKo, Eun Yi Kim “Highway Traffic Flow Prediction using Support Vector Regression and Bayesian Classifier”, IEEE, pp. 239 – 244, 2016.
6. ZhongshengHou, Senior Member, IEEE, and XingyiLi,”Repeatability and Similarity of Freeway Traffic Flow and Long-Term Prediction Under Big Data”, IEEE transactions on intelligent transportation system pp.1786 – 1796, 2016.
7. JiwanLee ,Bonghee Hong , Kyungmin Lee and Yang-Ja Jang , “A Prediction Model of Traffic Congestion Using Weather Data” ,IEEE International Conference on Data Science and Data Intensive Systems, pp. 81 – 88, 2015.
8. Zhiyuan Ma and GuangchunLuo, “Short Term Traffic Flow Prediction Based on On-line Sequential Extreme Learning Machine”, 8th International Conference on Advanced Computational Intelligence, pp. 143 – 149, 2016.
9. KalliSrinivasaNageswara Prasad SeelamRamakrishn, “An Efficient Traffic Forecasting System Based on Spatial Data and Decision Trees”, Department of Computer Science, Sri Venkateswara University, 2012.
10. Guy Leshem, and Ya'acovRitov , “Traffic Flow Prediction using Adaboost Algorithm with Random Forests as a Weak Learner”, International Journal of Mathematical, Computational, Physical, Electrical and Computer Engineering, Volume 1, 2007.
11. Megha Gupta, Naveen Aggarwal, “CLASSIFICATION TECHNIQUES ANALYSIS”, National Conference on Computational Instrumentation, 2010.
12. Honghui Dong, LiminJia, Xiaoliang Sun, Chenxi Li, Yong Qin, “Road Traffic Flow Prediction with a Time-Oriented ARIMA Model”, Fifth International Joint Conference on INC, IMS and IDC, 2009.
13. FuyingYu ,Zhijie Song, “A MapReduce-Based Nearest Neighbor Approach for Big-Data-Drive Traffic Flow Prediction”, Sixth International Conference on Intelligent Systems Design and Engineering Applications, pp. 890 – 893, 2015.
14. Lv, Lei, Chen, Meng, Liu, Yang, Yu, Xiaohui, “A plane moving average algorithm for short-term traffic flow prediction”,Springer, vol. 9078, pp. 357−369,2015.
15. Chang, S.C., Kim, R.S., Kim, S.J., Ahn, M.H.,”Traffic-flow forecasting using a 3-stage model”,IEEE Intelligent Vehicles Symposium, pp. 451−456 ,2000.
16. Moretti, F., Pizzuti, S., Panzieri, S., Annunziato, M.,”Urban traffic flow forecasting through statistical and neural network bagging ensemble hybrid modeling”, Neurocomputing 167, pp. 3−7 ,2015.
17. Lv, Y., Duan, Y., Kang, W., Li, Z., Wang, F.Y.” Traffic flow prediction with big data: a deep learning approach”, IEEE Transportation Intelligent Transp. System ,pp. 865−873, 2015.
18. M.-L. Huang,”Intersection traffic flow forecasting based on v-GSVR with a new hybrid evolutionary algorithm”, Neurocomputing 147, pp. 343–349, 2015.
19. JiwanLee ,Bonghee Hong , Kyungmin Lee and Yang-Ja Jang , “A Prediction Model of Traffic Congestion Using Weather Data” ,IEEE International Conference on Data Science and Data Intensive Systems, pp. 81–88, 2015.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

20. M. Lippi, M. Bertini, and P. Frasconi, "Short-term traffic flow forecasting: An experimental comparison of time-series analysis and supervised learning," IEEE Trans. Intell. Transp. Syst., pp. 871−882, Jun. 2013
21. B. L. Smith, B. M. Williams, and R. K. Oswald, "Comparison of parametric and nonparametric models for traffic flow forecasting," Transp. Res. C,Emerg. Technol.,pp. 303−321, Aug. 2002.
22. YuqiWang ,Jiannong Cao , Wengen Li and Tao Gu ,"Mining Traffic Congestion Correlation between Road Segments on GPS Trajectories ", IEEE, pp. 1−8, 2016.
23. K Papagiannaki," Long-Term Forecasting of Internet Backbone Traffic", IEEE Trans Neural Network.pp. 1110−1124, 2005.

BIOGRAPHY

RiyaGandewar is student in the computer engineering department, Pune Institute of Technology, Pune.
Her research interest in Data mining and Machine Learning.

A.G.phakatkar is a Assistant professor in the computer engineering department, Pune Institute of Technology, Pune.