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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2016

A Survey on Implementation of k-NN on Traffic Flow Prediction Using Hadoop

Vaibhav Gholap[1], Sujit Nale[2], Pradeep Patil[3], Renuka Patil[4], Prof. Harshwardhan Bhosle[5]

B. E Students, Dept. of Computer Engineering, JSPM'S ICOER, Wagholi, Pune, Maharashtra, India[1,2,3,4]

Asst. Professor, Dept. of Computer Engineering, JSPM'S ICOER, Wagholi, Pune, Maharashtra, India[5]

ABSTRACT: In big-data-driven traffic flow prediction systems, the robustness of prediction performancedepends on accuracy and timeliness. This paper presents a new MapReduce-based nearest neighbor (NN)approach for traffic flow prediction using correlation analysis (TFPC) on a Hadoop platform. In particular,we develop a real-time prediction system including two key modules, i.e., offline distributed training(ODT) and online parallel prediction (OPP). Moreover, we build a parallel k-nearest neighbor optimization classifier, which incorporates correlation information among traffic flows into the classification process.Finally, we propose a novel prediction calculation method, combining the current data observed in OPPand the classification results obtained from large-scale historical data in ODT, to generate traffic flowprediction in real time. The empirical study on real-world traffic flow big data using the leave-oneoutcross validation method shows that TFPC significantly outperforms four state-of-the-art predictionapproaches, i.e., autoregressive integrated moving average, Naïve Bayes, multilayer perception neuralnetworks, and NN regression, in terms of accuracy, which can be improved 90.07% in the best case,with an average mean absolute percent error of 5.53%.

KEYWORDS: Big data analytics, traffic flow prediction, correlation analysis, parallel classifier, HadoopMapReduce.

I. INTRODUCTION

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. The k-Nearest Neighbor algorithm (k-NN) is consider done of the ten most influential data mining algorithms. It belongs to the lazy learning family of methods that do not need of an explicit training phase. This method requires that all of the data instances are stored and unseen cases classified by finding the class labels of the k closest instances to them. To determine how close two instances are, several distances or similarity measures can be computed. This operation has to be performed for all the input examples against the whole training dataset. Thus, the response time may become compromised when applying it in the big data context. The traffic data in transportation have beenexploding rapidly with the characteristics of heterogeneity, autonomous sources, and complex and evolving associations (HACE). The big data generated by the IntelligentTransportation Systems (ITS) are worth further exploring tobring all their full potential for more proactive traffic management. The ability to accurately predict the evolutionof traffic in an online and real-time manner that plays acrucial role in traffic management and control applicationsis particularly important. Data-driven intelligent transportation has drawn significant attention in recent years: provided a special session on big data servicesand computational intelligence for industrial systems,including ITS applications. Furthermore, a special issuehighlighted the most recent research progress in big data The traffic is the main problem that occurs in the major cities. It will affect many of the Peoples day to day life and may cause major problem to heavy vehicles to reach the destination on time. Many Problems that will made the traffic flow difficult such as Weather, Road works, Accidents, vehicles that was repaired and make the traffic jam on the road. So, the future the world is going to face traffic problem which is the major problem for the normal peoples. Many of the Projects are also available to control these kind of the traffic problems and still the research is going for these control traffic on Intelligent Traffic System (ITS). And our aim is to predict the traffic for frequent interval of time and also to give the suggestion to the people to reach the destination quickly. In other countries they have already some uses these type of prediction technique to help the peoples to reach the destination and also to reduce the traffic that occur in the cities. Some project the traffic signals are recorder with the cameras and sensors to know the Number of vehicles.

II. Related work

In the last several decades, considerable research studieswere reported on the applications of different empiricaland theoretical techniques to traffic flow prediction.These works can be roughly categorized as parametricmethods, nonparametric methods, and hybridintegration methods. Recently, there have been varioustraffic flow prediction systems, models and algorithmsusing statistics-based approachesand computational. Intelligence(CI)-based approaches. Lv et al. proposed a novel deep-learning-based traffic flow predictionmethod, using auto encoders as building blocks to representtraffic flow features for forecasting. Li et al.presented a method which picks the most relevant data fromthe ``Big Data" to build concise yet accurate traffic flowprediction model. Tchrakian Lv et al.developed an algorithmfor the implementation of short-term prediction oftraffic with real-time updating based on spectral analysis.Min and Wynter put forward an approach to providingpredictions of speed and volume over 5-min interval for upto 1 h in advance. However, most of them employed thestand-alone learning models with the sequential algorithmson a single machine and were still somewhat unsatisfactoryin processing the increasingly explosive traffic big data in realtime, such as massive taxi trajectory data used in this work.

A. *PROBLEM STATEMENT:*

The traffic data in transportation have been exploding rapidly with the characteristicsof heterogeneity, autonomous sources, and complex and evolving associations (HACE). The big data generated by the Intelligent Transportation Systems(ITS) are worth further exploring to bring all their full potential for more proactivetraffic management. The ability to accurately predict the evolution of traffic in anonline and real-time manner that plays a crucial role in traffic management andcontrol applications is particularly important.

B. *GOALS AND OBJECTIVES:*

To predict real time traffic flow prediction using current big traffic data.
To save memory consumption.
To reduce the computational costs of big calculations.
To display analysis of traffic flow data in form of speedup, scale-up andsize-up.

C. *PROPOSED SYSTEM:*

This paper aims to develop a robust approach for the accurate and real-time Traffic Flow Prediction using CorrelationAnalysis (TFPC). The major contributions of our work aresummarized as follows:
A new prediction system (RPS) on a Hadoop platformis built to improve the capacity of big traffic data processingfor forecasting traffic flow in real time, whichcontains two key modules of offline distributed training (ODT) and online parallel prediction (OPP).
A robust nearest neighbor classifier (ParKNNO) ona MapReduce framework is presented to enhance the accuracy, efficiency and scalability of traffic flow prediction,by discovering correlation information amongtraffic flows and incorporating it into the classificationprocess.
A novel prediction calculation method is put forward to generate the real-time traffic flow prediction, withthe current data observed in OPP and the classificationresults obtained from large-scale historical data in ODT.
The prediction performance of TFPC is investigated with real-world traffic flow data collected from 12,000 taxis of Beijing during 15 days. The empiricalstudy indicates that the proposed approach is superior toother comparable prediction methods in terms of accuracy,speedup, scaleup, and sizeup.

We aim to develop a new framework (RPS) to handlereal-time prediction applications. Also, a robust nonparametric classifier (ParKNNO) is put forward to improve the classification performance by effectively incorporating correlationof traffic flows, and a novel prediction calculation method is proposed to accurately generate the prediction inreal time.

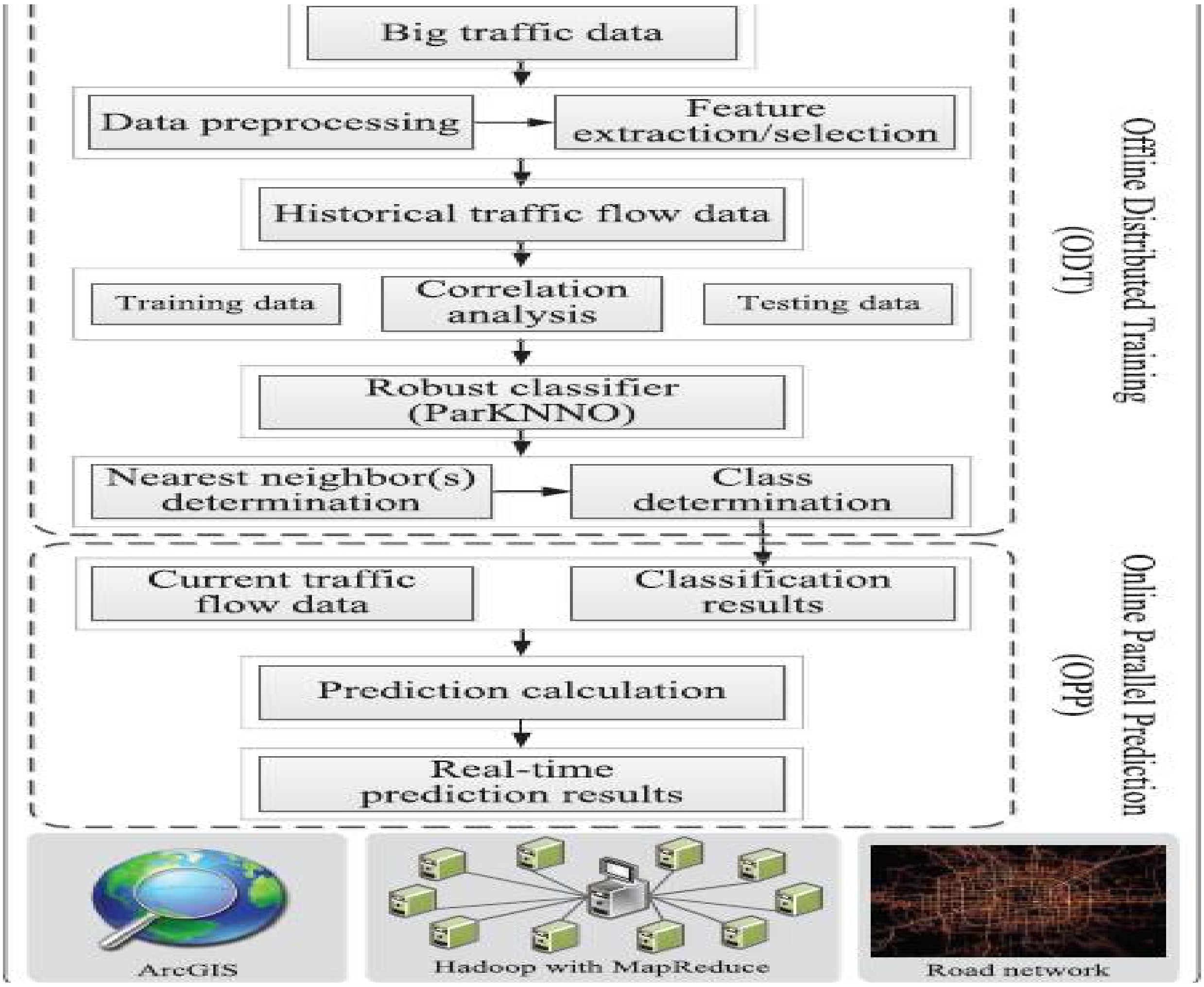
Big traffic data
Data preprocessing
Feature extraction/selection
Historical traffic flow data
Training data
Correlation analysis
Testing data
Robust classifier (ParKNNO)
Nearest neighbor(s) determination
Class determination
Offline Distributed Training (ODT)
Current traffic flow data
Classification results
Prediction calculation
Real-time prediction results
Online Parallel Prediction (OPP)
ArcGIS
Hadoop with MapReduce
Road network

Fig 1: A real-time prediction system (RPS) framework.

III. Proposed Algorithm

In this section we introduce some background information about the main components used in this project. A presents the k-NN algorithm as well as its weaknesses to deal with big data classification. Provides the description of the MapReduce paradigm and the implementation usedin this workA. k-NN and weaknesses to deal with big data The k-NN algorithm is a non-parametric method that can be used for either classification and regression tasks.

This section defines the k-NN problem, its current trends and the drawbacks to manage big data. A formal notation for the k-NN algorithm is the following:LetTRbe a training dataset and TS a test set, they are formed by a determined number andof samples,respectively. Each sample xp is a tuple (xp1,xp2...,xpD,ω), where,xpfis the value of the f-th feature of the p-th sample.This sample belongs to a class ω, given by xωp, and a D- dimensional space. For the TRset the classωis known, while it is unknown for TS. For each sample x test contained in the TSset, the k-NN model looks for the k closest samples in theTRset. To do this, it computes the distances between x test and all the samples of TR. The Euclidean distanceis commonly used for this purpose. Thekclosest samples(neigh1,neigh2, ...,neigh) are obtained by ranking (ascending order) the training samples according to the computed distance.

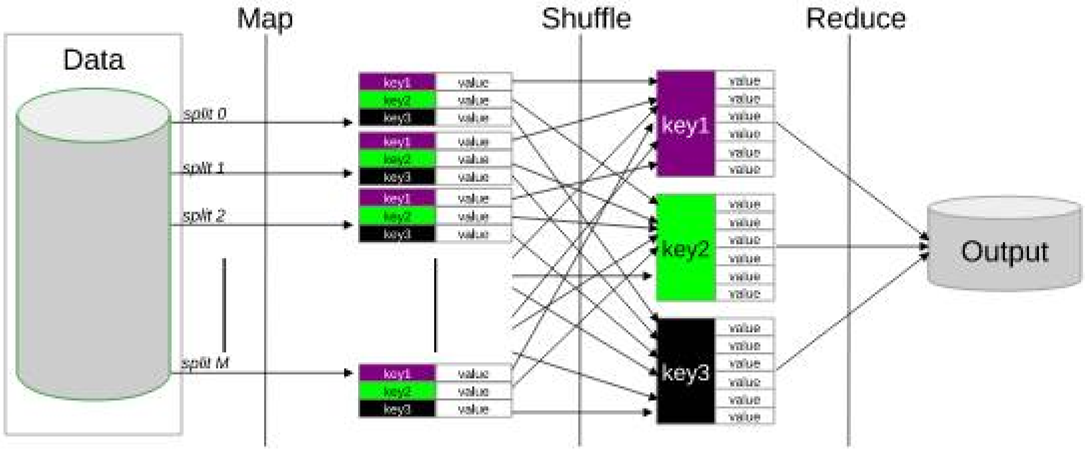

Fig.2. The MapReduce workflow

VI. CONCLUSION

We focused on the real-time prediction with bigtraffic flow data and thus proposed a new MapReduce-based nearest neighbor approach for traffic flow prediction usingcorrelation analysis (TFPC) on a Hadoop platform. To save memory consumption and reduce the computational costs ofbig calculations, TFPC was carried out in a real-time prediction system (RPS) composed of the ODT module and theOPP module. In particular, to enhance the robustness of realtime applications with very large training samples, a parallelk-nearest neighbor optimization classifier (ParKNNO) was built to model traffic flow correlations in ODT, and a novel prediction calculation method was put forward to generate traffic flow prediction in OPP. Furthermore, we evaluated the performance of TFPC on accuracy, speedup, scaleup andsizeup using the LOO-CV method by an empirical study. The results demonstrated that our approach was superior to other comparable methods in terms of accuracy which can be enhanced 90.07% in the best case, and significantly improved the efficiency and scalability of traffic flowprediction.

REFERENCES

[1] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, ``Data mining with big data,"IEEE Trans. Knowl. Data Eng., vol. 26, no. 1, pp. 97107, Jan. 2014.

[2] Y. Lv, Y. Duan, W. Kang, Z. Li, and F.-Y. Wang, ``Traffic flow predictionwith big data: A deep learning approach," IEEE Trans. Intell. Transp. Syst.,vol. 16, no. 2, pp. 865873, Apr. 2015.

[3] Q. Shi and M. Abdel-Aty, ``Big Data applications in real-time trafcoperation and safety monitoring and improvement on urban expressways,"Transp. Res. C, Emerg. Technol., vol. 58, pp. 380394, Sep. 2015.

[4] H. Hu, Y. Wen, T.-S. Chua, and X. Li, ``Toward scalable systems for bigdata analytics: A technology tutorial," IEEE Access, vol. 2, pp. 652687,2014.

[5] J. Zhang, F.-Y. Wang, K. Wang, W.-H. Lin, X. Xu, and C. Chen, ``Datadrivenintelligent transportation systems: A survey," IEEE Trans. Intell.Transp. Syst., vol. 12, no. 4, pp. 16241639, Dec. 2011.

[6] X.-W. Chen and X. Lin, ``Big data deep learning: Challenges and perspectives,"IEEE Access, vol. 2, pp. 514525, 2014.

[7] Z. Zhou, W. Gaaloul, P. C. K. Hung, L. Shu, and W. Tan, ``IEEE accessspecial session editorial: Big data services and computational intelligencefor industrial systems," IEEE Access, vol. 3, pp. 30853088, 2015.

[8] Z. Zhao, W. Ding, J. Wang, and Y. Han, ``A hybrid processing system forlarge-scale trafc sensor data," IEEE Access, vol. 3, pp. 23412351, 2015.

[9] E. I. Vlahogianni, B. B. Park, and J.W. C. van Lint, ``Big data in transportation and traffic engineering," Transp. Res. C, Emerg. Technol., vol. 58,p. 161, Sep. 2015.

[10] Y. Xia, L. Zhang, and Y. Liu, ``Special issue on big data driven intelligenttransportation systems," Neurocomputing, vol. 181, pp. 13, Mar. 2016.

[11] T. T. Tchrakian, B. Basu, and M. O'Mahony, ``Real-time traffic flowforecasting using spectral analysis," IEEE Trans. Intell. Transp. Syst.,vol. 13, no. 2, pp. 519526, Jun. 2012.

[12] E. I. Vlahogianni, J. C. Golias, and M. G. Karlaftis, ``Short-term trafficforecasting: Overview of objectives and methods," Transp. Rev., vol. 24,no. 5, pp. 533557, Sep. 2004.

[13] B. L. Smith, B. M.Williams, and R. K. Oswald, ``Comparison of parametricand nonparametric models for trafc ow forecasting," Transp. Res. C,Emerg. Technol., vol. 10, no. 4, pp. 303321, Aug. 2002.

[14] C. Chen, Y. Wang, L. Li, J. Hu, and Z. Zhang, ``The retrieval of intradaytrend and its inuence on trafc prediction," Transp. Res. C, Emerg.Technol., vol. 22, pp. 103118, Jun. 2012.