ISSN 2348 – 9928
IJAICT Volume 7, Issue 3, March 2020 Doi:10.46532/ijaict-2020010 Published on 05 (3) 2020

SPATIOTEMPORAL TRAFFIC ANALYSIS USING BIG DATA

H. Anandakumar Assistant Professor
Abishek Sailesh UG Scholar
C. Muthumeenal UG Scholar
S. Visalakshi UG Scholar
K. Muthumani UG Scholar

**Department of Computer Science and Engineering,
Akshaya College of Engineering and Technology,
Coimbatore, Tamilnadu, India**

Abstract *- In collaborated online technique traffic prediction methods is proposed with distributed context aware random forest learning algorithm .The random forest is ensemble classifier which learns different traffic and context model form distributed traffic patterns. One major challenge in predicting traffic is how much to rely on the prediction model constructed using historical data in the real-time traffic situation, which may differ from that of the historical data due to the fact that traffic situations are numerous and changing over time. The proposed algorithm is online predictor of real-time traffic, the global prediction is achieved with less convergence time .The distributed scenarios (traffic data and context data) are collected together to improve the learning accuracy of classifier. The conducted experimental results on prediction of traffic dataset prove that the proposed algorithm significantly outperforms the existing algorithm.*

Keywords*— Spatiotemporal, Data mining, Traffic Prediction, Clustering, Apriori, Sequence pattern mining.*

I. INTRODUCTION

Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the strictures of your database architectures. To gain value from this data, you must choose an alternative way to process it. The hot IT buzzword of 2012, big data has become viable as [1] cost-effective approaches have emerged to tame the volume, velocity and variability of massive data. Big data processing is eminently feasible for even the small garage start-ups, who can cheaply rent server time in the cloud.

The value of big data to an organization falls into two categories: analytical use, and enabling new products. Big data analytics can reveal insights hidden previously by data too costly to process, such as peer influence among customers, revealed by analysing [2] shoppers' transactions, social and geographical data. Being able to process every item of data in reasonable time removes the troublesome need for sampling and promotes an investigative approach to data, in contrast to the somewhat static nature of running predetermined reports. The past decade's successful web start ups are prime examples of big data used as an enabler. It's no coincidence that the lion's shares of ideas and tools underpinning big data have emerged from Google, Yahoo, Amazon and Face book. The emergence of big data [3] into the enterprise brings with it a necessary counterpart: agility. Successfully exploiting the value in big data requires experimentation and exploration. Whether creating new products or looking for ways to gain competitive advantage, the job calls for curiosity and an entrepreneurial outlook.

II. RELATED WORK

System architecture

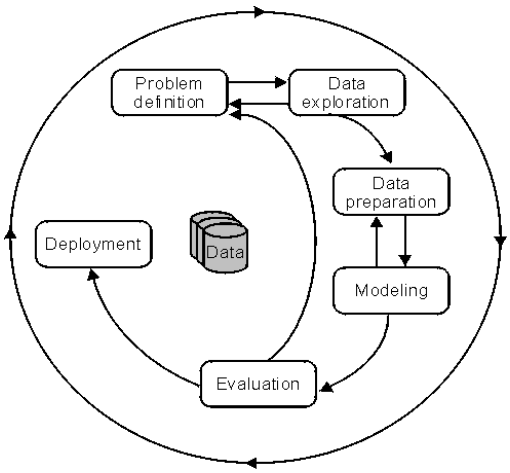

Fig 1 : Data mining process

Figure 1 illustrates the phases, and the iterative nature, of a data mining project. The process flow shows that a data mining project does not stop when a particular solution is deployed. The results of data mining [4] trigger new business questions, which in turn can be used to develop more focused models.

A data mining project starts with the understanding of the business problem. Data mining experts, business experts, and domain experts work closely together to define the project

ISSN 2348 – 9928
Doi:10.46532/ijaict-2020010 Published on 05 (3) 2020

objectives and the requirements. The project objective is then translated into a data mining problem definition.

2.1 Data Exploration

Domain experts understand the meaning of the metadata. They collect, describe, and explore the data. They also identify quality problems of the data. A frequent exchange with the data mining experts and the business experts from the problem definition phase is [5][6] vital. In the data exploration phase, traditional data analysis tools, for example, statistics, are used to explore the data.

2.2 Data Preparation

Domain experts build the data model for the modelling process. They collect, cleanse, and format the data because some of the mining functions accept data only in a certain format. They also create new derived attributes, for example, an average value. In the data preparation phase, data is tweaked multiple times [7][8] in no prescribed order. Preparing the data for the modelling tool by selecting tables, records, and attributes, are typical tasks in this phase. The meaning of the data is not changed.

2.3 Modelling

Data mining experts select and apply various mining functions because you can use different mining functions for the same type of data mining problem. Some of the mining functions require specific data types. The data mining experts must assess each model. In the modelling phase, a frequent exchange with the domain [9] experts from the data preparation phase is required. The modelling phase and the evaluation phase are coupled. They can be repeated several times to change parameters until optimal values are achieved. When the final modelling phase is completed, a model of high quality has been built.

2.4 Evaluation

Data mining experts evaluate the model. If the model does not satisfy their expectations, they go back to the modelling phase and rebuild the model [10][11] by changing its parameters until optimal values are achieved. At the end of the evaluation phase, the data mining experts decide how to use the data mining results.

2.5 Deployment

Data mining experts use the mining results by exporting the results into [12][3] database tables or into other applications, for example, spreadsheets.

III. PROPOSED WORK

K-Means Clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. The problem is computationally difficult (NP-hard); however, there are efficient heuristic algorithms that are commonly employed and converge quickly to a local optimum. These are usually [14] similar to the expectation-maximization algorithm for mixtures of Gaussian distributions via an iterative refinement approach employed by both algorithms. Additionally, they both use cluster centers to model the data; however, k-means clustering tends to find clusters of comparable spatial extent, while the expectation-maximization mechanism allows clusters to have different shapes.

The algorithm has a loose relationship to the k-nearest neighbor classifier, a popular machine learning technique for classification that is often confused with k-means because of the k in the name. One can apply the 1-nearest neighbor classifier on the cluster centers obtained by k-means to classify new data into the existing clusters. This is known as nearest centroid classifier or Rocchio algorithm.

Given a set of observations ($\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_n$), where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k ($\leq$ n) sets $\mathbf{S} = \{S_1, S_2, \ldots, S_k\}$

Advantages of proposed work

- Prediction accuracy is high
- Global prediction helps road network to handle heavy traffic.
- Coordination among distributed entities helps to handle dynamically changing traffic
- It helps to predict the accident rate and fatal injury

IV. PROBLEM STATEMENT

The existing approaches for traffic prediction aim at predicting traffic in specific traffic situations, e.g. either typical conditions or when accidents occur the existing approaches used for traffic prediction deploy models learned offline (i.e. they rely on a priori training sessions) or they are retrained after long periods and thus, they cannot adapt to (learn from) dynamically changing traffic situations. Most previous work is

ISSN 2348 – 9928
Doi:10.46532/ijaict-2020010 Published on 05 (3) 2020

based on empirical studies and does not offer rigorous performance guarantees for traffic prediction.

V. MODULE DESCRIPTION

Dataset collection and Pre process

Dataset collection is a set of raw data collected from source. Preprocessing the data is a process carried out using String Tokenizer

Clustering

K means clustering aims to partition n observations into k clusters in which each observation belongs to cluster with nearest mean , serving as a prototype of the cluster. This results in partitioning of data space into cells.

Prediction

Naive bayesian is a probabilistic classifier based on applying bayes' theorem with strong (naïve) independence assumptions between the features

Rule generation

Sequence pattern Mining and Association rules is a procedure meant to find frequent patterns, correlations, association, and structures from dataset. Apriori algorithm identifies the frequent individual items in the dataset.

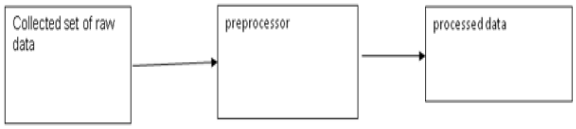

Fig 2 : Dataset collection and Preprocess

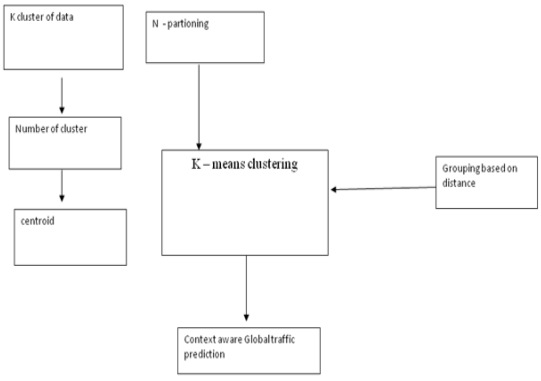

Fig 3 : Clustering

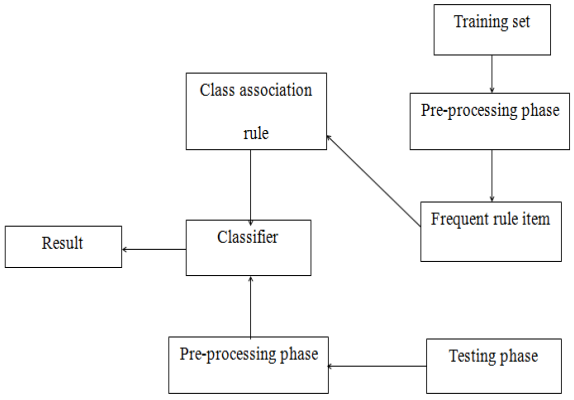

Fig 4 : Prediction

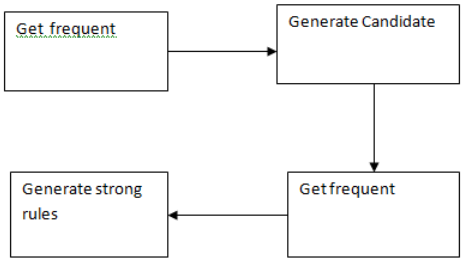

Fig 5 : Rule generation

Scope for Future Work

In future, we plan to extend the current framework to distributed scenarios where traffic data is gathered by distributed entities and thus, coordination among distributed entities are required to achieve a global traffic prediction goal.

VI. CONCLUSION

Both qualitative and quantitative approaches have been used to measure the impact of an accident on road networks and various machine learning techniques have been applied to predict the typical traffic conditions and the impact of accidents, including Naive Bayesian classifier, Decision Tree classifier and Nearest Neighbour classifier. The ensemble of algorithms like bagging and boosting have been developed based on online learning techniques. Ad boost is another standard ensemble algorithm which is represented by Prediction with expert advice and the weight update schemes. The existing approaches for traffic prediction aim at predicting traffic in specific traffic situations, e.g. either typical

ISSN 2348 – 9928
Doi:10.46532/ijaict-2020010 Published on 05 (3) 2020

conditions or when accidents occur the existing approaches used for traffic prediction deploy models learned offline (i.e. they rely on a priori training sessions) or they are retrained after long periods and thus, they cannot adapt to (learn from) dynamically changing traffic situations. Context-aware adaptive traffic prediction is used for predict traffic pattern. This method to overcome the difficulty of dynamically partition the entire context space into multiple smaller context subspaces and maintain and update the sample mean reward estimates for each subspace.

References

[1] Yap, K. S. (2011, April). A new multi agent system based on online sequential extreme learning machines and Bayesian Formalism. In Networking, Sensing and Control (ICNSC), 2011 IEEE International Conference on (pp. 74-79). IEEE.

[2] Xu, J., Tekin, C., & van der Schaar, M. (2013, October). Learning optimal classifier chains for real-time big data mining. In Communication, Control, and Computing (Allerton), 2013 51st Annual Allerton Conference on (pp. 512-519). IEEE..

[3] Dong, C., & Zhou, B. (2013, December). An ensemble learning framework for online web spam detection. In Machine Learning and Applications (ICMLA), 2013 12th International Conference on (Vol. 1, pp. 40-45). IEEE.

[4] Tekin, C., & van der Schaar, M. (2013, October). Distributed online big data classification using context information. In Communication, Control, and Computing (Allerton), 2013 51st Annual Allerton Conference on (pp. 1435-1442). IEEE.

[5] Sidhu, P., Bhatia, M. P. S., &Bindal, A. (2013, December). Empirical Support for Weighted Majority, Early Drift Detection Method and Dynamic Weighted Majority. In Machine Intelligence and Research Advancement (ICMIRA), 2013 International Conference on (pp. 623-627). IEEE.

[6] A. Roshini and H. Anandakumar, "Hierarchical cost effective leach for heterogeneous wireless sensor networks," Advanced Computing and Communication Systems, 2015 International Conference on, Coimbatore, 2015, pp. 1-7.doi: 10.1109/ICACCS.2015.7324082

[7] S. Divya, H. A. Kumar and A. Vishalakshi, "An improved spectral efficiency of WiMAX using 802.16G based technology," Advanced Computing and Communication Systems, 2015 International Conference on, Coimbatore, 2015, pp. 1-4.doi: 10.1109/ICACCS.2015.7324098

[8] M. Suganya and H. Anandakumar, "Handover based spectrum allocation in cognitive radio networks," Green Computing, Communication and Conservation of Energy (ICGCE), 2013 International Conference on, Chennai, 2013, pp. 215-219.doi: 10.1109/ICGCE.2013.6823431

[9] H. Anandakumar and K. Umamaheswari, Supervised machine learning techniques in cognitive radio networks during cooperative spectrum handovers, Cluster Computing (2017), 1–11. doi: 10.1007/s10586-17-0798-3

[10] E.R.Sparks,A.Talwalkar,V.Smith,J.Kottalam,X.Pan,J.E .Gonzalez,M.J.Franklin,M.I.Jordan,T.Kraska,MLI:anA PIfordistributedmachinelearning,in:Proceedings of IEEE 13th International Conferenceon DataMining, Dallas, TX,USA, December 7-10,2013,2013,pp.1187–1192, doi:10.1109/ICDM.2013.158.

[11] B.Larsen,C.Aone,Fast and effective text mining using linear time document clustering,in: Proceedings of the Fifth ACMSIGKDD International Conferenceon Knowledge Discovery and datamining, KDD'99, ACM, NewYork,NY, USA,1999,pp.16–22, doi:10.1145/312129.312186.

[12] Suriya M,Sugandhanaa M,Vaishnavi J,Dhivya Bharathy P,"A Survey on Cognitive Handover between the terrestrial and Satellite Segment", 2016 IJAICT Volume 2, Issue 11, March 2016 Doi:01.0401/ijaict.2016.11.03 Published on 05 (04) 2016

[13] Vamsi krishna tumuluru,Ping Wang,Dusit Niyato,wei song"performance analysis of cognitive radio spectrum access with prioritized traffic" IEEE Communications Letters (Volume: 18, Issue: 7, July 2014)

[14] G.Ezra sastry,S.Tamilarasan,P.Kumar",Dynamic Resource Allocation in Cognitive Radio Networks". International Journal of Computer Science Trends and Technology (IJCS T) – Vol 4 Iss 3, May - Jun 2016