IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 2, February 2021

DOI 10.17148/IJARCCE.2021.10227

TRAFFIC DATA PREDICTION IN INTELLIGENT TRANSPORTATION SYSTEM USING m-KNN ALGORITHM AND PRINCIPLE COMPONENT ANALYSIS

P.Pavithra[1], R.Vadivel[2]*

Department of Information Technology, Bharathiar University, Tamil Nadu India[1]

Department of Information Technology, Bharathiar University, India[2]

Abstract: Nowadays, the capabilities of roads and transportation systems have not evolved in a way that is efficiently copes with the increasing number of vehicles and growth of population. Traffic congestion is becoming the issues of the entire globe. The traffic congestion issues have some other indirect overseen issues such as noise, pollution and increase travelling time. This project aims to explore and review the data mining and machine learning technologies adopted in research and industry to attempt to overcome the direct and indirect traffic issues on humanity and societies. The study is focusing on the traffic management approaches that were depended on data mining and machine learning technologies to detect and predict the traffic only. Using data mining technology in traffic management provides a powerful analysis and processing function of mass traffic data and directs drivers and systems to make better decisions. Knowledge mining and discovery is an emerging area in traffic management systems focuses on using and analyzing large amount of traffic data to be used for traffic control, route guidance, or route programming. This study is important to the traffic research communities, traffic software companies, and traffic government officials. Additionally, this study will draw general attention to a new traffic management proposition approach.

Keywords: Data Mining, machine learning, Decision TreeTraffic management, KNN.

1. INTRODUCTION

Traffic congestion is a major challenge for our day to day life. Traffic congestion is a state where the speed of the vehicles is slower than the actual free stream speed on a road and highway. This is after effect of expanded traffic volume. When traffic is stopped for a specific time interval then congestion situation arises. This can be solved by implementing a well-planned process that handles the congestion in a smart manner. The first step to solving congestion problem is to understand the basic reasons for congestion in various congestion-tackling phases. It can also be done by developing the relevant congestion indicators for monitoring and utilising the existing infrastructure and creating additional capacity using new technology.

Data mining tools [7] used to analyze and extract information from large sets of data are generally classified as “data mining” tools. This paper describes research that is devising a procedure for developing, implementing and monitoring traffic signal timing plans using available data mining tools. The hypothesis premise of the research is that the data collected by signal control systems can be used to improve system design and operations for the current methods of traffic control. The data-mining tool that serves as the foundation in this research for signal plan development is hierarchical cluster analysis, while classification may be used for monitoring plan effectiveness. This paper offers a background on signal timing plan development, with consideration of system state definitions and offering a procedure for improved traffic control through the use of Hierarchical Cluster Analysis. The case study shows that the sensor data provided by ITS holds valuable information regarding the behavior of traffic, capable of automatically generating TOD intervals for transitioning between timing plans as well as providing appropriate volume data for plan development during these automatically generated TOD intervals.

2. RELATED WORKS

As an approach to enhance the prediction performance of tra_c crash severity, many researcherstried to employ ML models. A k-means clustering algorithm wasused to cluster the crash dataset into three clusters. The analysis results revealed that the predictionaccuracy was improved significantly after clustering. Moreover, when compared with the predictionperformance of OP, the developed NN model was found to be superior with an accuracy of 74.6%.

A comparison was conducted between the proposed CNN model and nine commonstatistical and ML models such as k-nearest neighbor (KNN), decision trees (DT), SVM, and NN basedon crash severity prediction performance. The

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 2, February 2021

DOI 10.17148/IJARCCE.2021.10227

comparison results revealed that CNN outperformedthe other models in traffic crash severity prediction with an average F1score of 84%. Many comparative studies were conducted worldwide to compare the performance of differentmodels when used for traffic crash severity. The three developed models were compared based on their predictionaccuracy. The results revealed the prediction accuracies were not significantly different. Moreover,it was found that variable reduction was effective in enhancing the prediction accuracy.

The prediction accuracy of SVM was compared with that of the OP model. It was found that theaccuracy of SVM (48.8%) was higher than that of the OP model (44.0%). Although the accuracy wasfound to be very low, the authors did not apply any dimension reduction technique such as principalcomponent analysis (PCA) on the crash dataset to overcome the problems of correlation between theinput variables. A comparison was conducted between theproposed model and other models such as NN, SVM, and k-NN based on prediction accuracy. It wasfound that the proposed model outperformed the other models with an accuracy of 87%.

Based on this information, the trauma centers would be able to prepare for appropriateand prompt medical treatment. Hence, this study aims to apply PCA, which is seldom used in suchstudies, on the crash dataset and to investigate its effect on crash severity prediction performance of twocommonly used ML models for crash severity prediction (as found in the literature), namely MLP-NNand SVM models.

3. EXISTING SYSTEM

Recently, social networks and media platforms have been widely used as a source of information for the detection of events, such as traffic congestion, incidents, natural disasters (earthquakes, storms, fires, etc.), or other events. Twitter streams to detect earthquakes and typhoons, by monitoring special trigger keywords, and by applying an SVM as a binary classifier of positive events (earthquakes and typhoons) and negative events (non-events or other events). Focus on the detection of fires in a factory from Twitter stream analysis, by using standard NLP techniques and a Naive Bayes (NB) classifier.

DISADVANTAGES

- Event detection from social networks analysis is a more challenging problem than event detection from traditional media like blogs, emails, etc., where texts are well formatted.
- The main difficulty encountered in dealing with problems of text mining is caused by the vagueness of natural language.
- Feature selection is particularly important, since one of the main problems in text mining is the high dimensionality of the feature space.
- SUMs are unstructured and irregular texts, they contain informal or abbreviated words,misspellings or grammatical errors.
- SUMs contain a huge amount of not useful or meaningless information.

4.PROPOSED SYSTEM

We propose an intelligent system, based on text mining and Naive bayes algorithms, for real-time detection of traffic events from Twitter stream analysis. The system exploits available technologies based on state-of-the-art techniques for text analysis and pattern classification. These technologies and techniques have been analyzed, tuned, adapted, and integrated in order to build the intelligent system. Determining the most effective among different state-of-the-art approaches for text classification. The chosen approach was integrated into the final system and used for the on-the-field real-time detection of traffic events.

4.1 ADVANTAGES

- Proposed system may approach both binary and multi-classclassification problems. As regards binary classification, we consider traffic-related tweets, and tweets not related with traffic. System could work together with other traffic sensors.
- ITS monitoring systems for the detection of traffic difficulties, providing a low-cost wide coverage of the road.
- It performs a multi-class classification, which recognizes non-traffic, traffic due to congestion or crash, and traffic due to external events.

IJARCCE

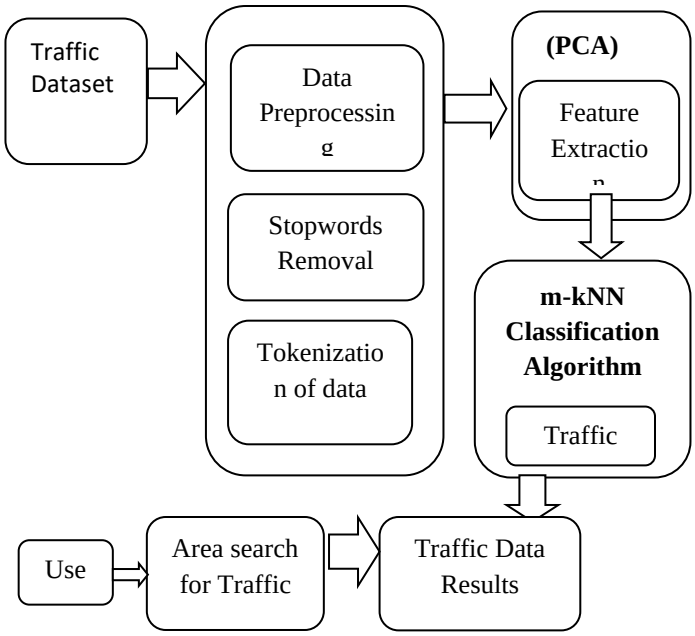
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 2, February 2021

DOI 10.17148/IJARCCE.2021.10227

5. SYSTEM ARCHITECTURE



6. MODULE DESCRIPTION

6.1 USER CONTENT POSTING

- This module is used to design the system based on social network to handle on user posts.
- It has a login page for users and for administrator to handle the traffic reporting system.
- User post their own information or status in our designed social network.

6.2 DATA PREPROCESSING

- In order to extract only the

text of each raw tweet and remove all meta-information associated with it.

- The tokenize removes all punctuation marks and splits each SUM into tokens corresponding to words (bag-of-words representation).
- Stop-word filtering consists in eliminating stop-words, i.e., words which provide little or no information to the text analysis**.**

6.3 DATA CLASSIFICATION

- The system continuously

monitors a user post and analyse the traffic upto the time.

- Naïve Bayes Algorithm used to classify the taffic and normal posts.
- Identify the traffic related post and retrieve the traffic details within a moment.

6.4 TRAFFIC SEARCH

- The user can search the traffic by giving the query as location name.
- This modules retrieves the traffic report and the traffic related messages related to corresponding query.
-

MODIFIED K NEAREST NEIGHBOUR ALGORITHM

Traffic data is classified using the classification algorithm named Modified K Nearest Neighbour Algorithm (m-KNN). The main idea of the presented method is assigning the classlabel of the queried instance into K validated data training points. In other hand, first, the validity of all data samples in the train set is computed. Then, a weighted KNN is performed on any test samples. The following shows the pseudo code of the MKNN algorithm.

Pseudo-code of the MKNN Algorithm:

```
Output_label := MKNN ( train_set , test_sample )
Begin
For i := 1 to train_size
Validity(i) := Compute Validity of i-th sample;
End for;
Output_label:=Weighted_KNN(Validity,test_sample);
```

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 2, February 2021

DOI 10.17148/IJARCCE.2021.10227

Return Output_label ;
End.

In the MKNN algorithm, every training sample must be validated at the first step. The validity of each point is computed according to its neighbors. The validation process is performed for all train samples once. After assigning the validity of each train sample, it is used at the second step as impact or weight of the points in the ensembles of neighbors which the point is selected to attend. To validate a sample point in the train set, the H nearest neighbors of the point is considered. Among the H nearest neighbors of a train sample x, validity(x) counts the number of points with the same label to the label of x. Eq. 1 is the formula which is proposed to compute the validity of every points in train set.

$$Validity(x) = \frac{1}{H}\sum_{i=1}^{H} S(lbl(x), lbl(N_i(x))) \quad (1)$$

Where H is the number of considered neighbors and lbl(x) returns the true class label of the sample x. also, Ni(x) stands for the ith nearest neighbor of the point x. The function S takes into account the similarity between the point x and the ith nearest neighbor.

Weighted KNN is one of the variations of KNN method which uses the K nearest neighbors, regardless of their classes, but then uses weighted votes from each sample rather than a simple majority or plurality voting rule. Each of the K samples is given a weighted vote that is usually equal to some decreasing function of its distance from the unknown sample. For example, the vote might set be equal to 1/(de+1), where de is Euclidian distance. These weighted votes are then summed for each class, and the class with the largest total vote is chosen. This distance weighted KNN technique is very similar to the window technique for estimating density functions. For example, using a weighted of 1/(d+1) is equivalent to the window technique with a window function of 1/(de+1) if K is chosen equal to the total number of training samples. In the MKNN method, first the weight of each neighbor is computed using the 1/(de+1), where is a smoothing regulator and here is selected to 0.5. Then, the validity of e that training sample is multiplied on its raw weight which is based on the Euclidian distance. In the MKNN method, the weight of each neighbor sample is derived according to below equation,

W(i) = Validity (i) * (1/(de+1))

Where W(i) and Validity(i) stand for the weight and the validity of the ith nearest sample in the train set. This technique has the effect of giving greater importance to the reference samples that have greater validity and closeness to the test sample. So, the decision is less affected by reference samples which are not very stable in the feature space in comparison with other samples. In other hand, the multiplication of the validity measure on distance based measure can overcome the weakness of any distance based weights which have many problems in the case of outliers. So, the proposed MKNN algorithm is significantly stronger than the traditional KNN method which is based just on distance.

7. EXPERIMENTAL RESULTS

Here we compare our proposed m-kNN (k Nearest Neighbour) algorithm with existing methods, Bayesian Network and Random Forest Algorithm.

Accuracy

Accuracy is determined as the overall correctness of the model and is computed as the total actual classification parameters $(T_p + T_n)$ which is segregated by the sum of the classification parameters $(T_p + T_n + F_p + F_n)$. The accuracy is computed as like :

$$\text{Accuracy} = \frac{T_p + T_n}{(T_p + T_n + F_p + F_n)}$$

Where T_p - True positive
T_n-True negative
F_n-False negative
F_p-False positive

Algorithm	Accuracy
RF	**87.42**
BN	**89.24**
m-kNN	**95.67**

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 2, February 2021

DOI 10.17148/IJARCCE.2021.10227

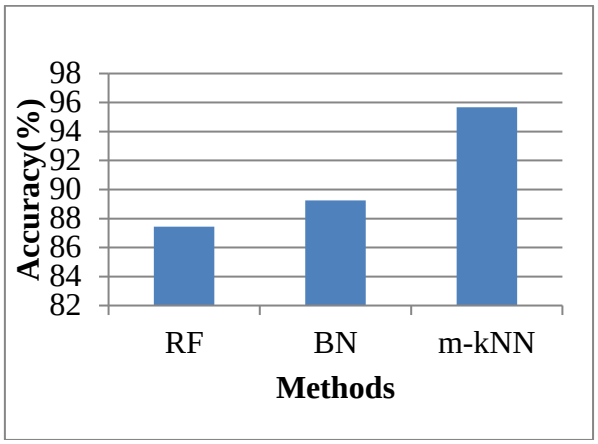


Figure 7.1 Accuracy comparison

From the above Figure 7.1, it can be observed that the comparison metric is evaluated using existing and proposed method in terms of accuracy. For x-axis the algorithms are taken and in y-axis the accuracy value is plotted. The existing RF and BN methods provides lower accuracy whereas the proposed m-kNN. The result proves that the proposed system attains greater utility results using m-kNN with PCA. Thus the proposed m-kNN with PCA is superior to the previous the RF and BN methods for the given transaction databases.

Precision

Precision is discusses as the ratio of the true positives opposite to both true positives and false positives result for imposition and real features. It is distinct as given below

$$\text{Precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

Algorithm	Precision
RF	**90.23**
BN	**93.47**
m-kNN	**96.54**

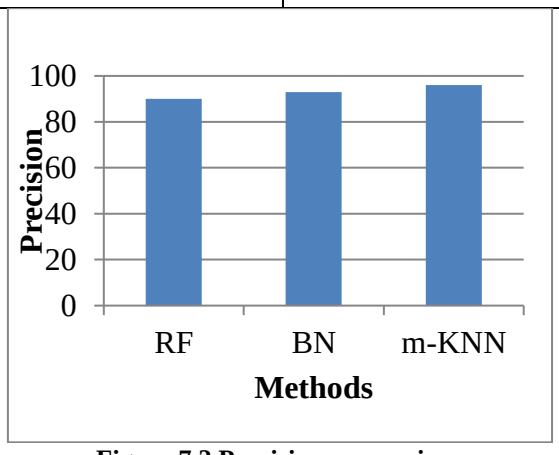


Figure 7.2 Precision comparison

From the above Figure 7.2, it can be observed that the comparison metric is evaluated using existing and proposed method in terms of precision. For x-axis the algorithms are taken and in y-axis the precision value is plotted. The existing methods provides lower precision whereas the proposed system provides higher precision for the given speech sample input.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 2, February 2021

DOI 10.17148/IJARCCE.2021.10227

Recall

Recall value is computed on the root of the data retrieval at true positive forecast, false negative. Generally it can be decided as

$$Recall = \frac{T_P}{T_{P+F_N}}$$

Algorithm	Recall
RF	**88.29**
BN	**91.61**
m-kNN	**94.5**

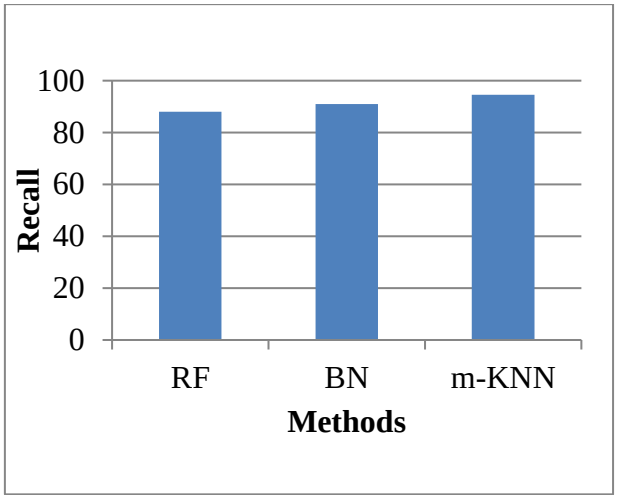


Figure 7.3 Recall comparison

From the above Figure 7.3, it can be observed that the comparison metric is evaluated using existing and proposed method in terms of recall. For x-axis the algorithms are taken and in y-axis the recall value is plotted. The existing methods provide lower recall whereas the proposed system provides higher recall for the given sample input. Thus the proposed m-kNN with PCA algorithm is superior to the previous the RF and BN methods.

8. CONCLUSION

System employed m-KNN (Modified k Nearest Neighbour algorithm) as a classification model and PCA (Principle Component Analysis) is used for Feature Extraction. Road traffic prediction is a critical component in modern smart transportation systems. The proposed method which considerably improves the performance of m-kNN method employs a kind of preprocessing on train data. It adds a new value namedValidity to train samples which cause to more information about the situation of training data samples in the feature space. The validity takes into accounts the value of stability and robustness of the any train samples regarding with its neighbors. Applying the weighted KNN which employs validity as the multiplication factor yields to more robust classification rather than simple KNN method, efficiently. The traffic detection system was employed for real-time monitoring of several areas of the road network, allowing for detection of traffic events almost in real time, often before online traffic news web sites.

9. FUTURE ENHANCEMENT

Descriptive mining methods have aided in generating a real-time information system, identifying traffic patterns, developing travel speed calculation model, and investigating parking decisions. Predictive data mining techniques have helped ininferring the network topology, finding traffic bottlenecks, solving the multi-objective location inventory problem, constructing two data reduction algorithms, and predicting short-term traffic flow in heterogeneous conditions. Increasing the potential of the evolutionary model is rather challenging. An effective evolutionary approach without drawbacks will certainly help in developing enhanced ITS.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 2, February 2021

DOI 10.17148/IJARCCE.2021.10227

REFERENCES

[1] (Mar. 2015). *TomTom*. Accessed: Oct. 11, 2018. [Online]. Available:https://corporate.tomtom.com/news-releases
[2] D. Schrank, B. Eisele, and T. Lomax, ``TTI's 2012 urban mobility report powered by INRIX traf_c data," Texas A&M Transp. Inst. and Texas A&M Univ. Syst., Texas, TX, USA, Tech. Rep. 1, 2012.
[3] J. Raj, H. Bahuleyan, and L. D. Vanajakshi, ``Application of data mining techniques for traf_c densityestimation and prediction," *Transp. Res. Procedia*, vol. 17, pp. 321_330, Dec. 2016.
[4] S. Sundaram, S. S. Kumar, and M. D. Shree, ``Hierarchical clustering technique for traf_c signal decision support," *Int. J. Innov. Sci., Eng. Technol.*, vol. 2, no. 6, pp. 72_82, Jun. 2015.
[5] S. Anand, P. Padmanabham, A. Govardhan, and R. H. Kulkarni, ``An extensive review on data mining methods and clustering models for intelligent transportation system," *J. Intell. Syst.*, vol. 27, no. 2, pp. 263_273, 2018.
[6] J. Zhang, F.-Y. Wang, K. Wang, W.-H. Lin, X. Xu, and C. Chen, ``Datadriven intelligent transportation systems: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 4, pp. 1624_1639, Dec. 2011.
[7] J. Lopes, J. Bento, E. Huang, C. Antoniou, and M. Ben-Akiva, ``Traf_c and mobility data collection for real-time applications," in *Proc. 13th Int. IEEE Annu. Conf. Intell. Transp. Syst.*, Madeira, Portugal, Sep. 2010, pp. 216_223.
[8] K. Miller, M. Miller, M. Moran, and B. Dai, ``Data management life cycle," Texas A&M Transp. Inst., College Station, TX, USA, Tech.Rep. 1, Mar. 2018.
[9] Z. Diao*et al.*, ``A hybrid model for short-term trafic volume predictionin massive transportation systems," *IEEE Trans. Intell. Transp. Syst.*,vol. 20, no. 3, pp. 935_946, Mar. 2019.
[10] K. Kumara, M. Paridab, and V. Katiyar, ``Short term traffic flow prediction for a non urban highway using arti_cial neural network," in *Proc2nd Conf. Transp. Res. Group India*, Agra, India, 2013, pp. 755_764.
[11] R. Ke, Z. Li, J. Tang, Z. Pan, and Y. Wang, ``Real-time traffic flowparameter estimation from uav video based on ensemble classifier andoptical flow," *IEEE Trans. Intell.Transp. Syst.*, vol. 20, no. 1, pp. 54_64,Jan. 2019.
[12] R.Ke, Z. Li, S. Kim, J. Ash, Z. Cui, andY.Wang, ``Real-time bidirectional traffic flow parameter estimation from aerial videos," *IEEE Trans. Intell.Transp. Syst.*, vol. 18, no. 4, pp. 890_901, Apr. 2017.

AUTHORS PROFILE:

P. Pavithra received Bachelors Degree in ComputerApplication in the year 2019 form Syed AmmalArts and Science College,Ramanathapuram, Tamilnadu, affiliated by KaraikudiAlagappa University. She is currently pursuing a Master Degree in Information Technology from 2019 to 2021, at Bharathiar University, Coimbatore, Tamilnadu. she area of interest is Robotics and Multimedia.

R. Vadivel is an Assistant Professor in the Department of Information Technology, Bharathiar University, Tamilnadu, India. He is received his Ph.D. degree in Computer Science form ManonmaniamSundaranar University in the year 2013. He is obtained his Diploma in Electronics and Communication Engineering form State Board of Technical Education in the year 1999, B.E., Degree in Computer Science and Engineering from Periyar University in the year 2002, M.E., degree in Computer Science and Engineering from Annamalai University in the year 2007. He had published over 40 journals papers and conferences paper both at National and International level. His areas of interest include Computer Networks, Network Security, Information Security, etc.,