International Innovative Research Journal of Engineering and Technology

ISSN NO: 2456-1983

International Conference on Emerging Innovation in Engineering and Technology

ICEIET-2017

Predictive Analytics for Traffic Flow Forecasting Using Enhanced K-Nearest Neighbours Algorithm

Dr.Santhi Baskaran[1], S.Lakshmi@Vaishnavi[2], K.Manisha Selva[3], K.Keerthana[4], K.Revathi[5]

[1]Professor/IT, Pondicherry Engineering College
[2]Student Member/IT, Pondicherry Engineering College
[1]santhibaskaran@pec.edu, [2]lakshmivaishnavi@pec.edu, [3]manishaselvak@pec.edu, [4]keerthanak@pec.edu , [5]revathikumar@pec.edu

Abstract

Traffic flow forecasting plays an important role in route guidance and traffic management. Traffic flow prediction is an important precondition to lessen traffic congestion in large-scale urban areas. k-Nearest Neighbour (KNN) is one of the most important methods in traffic flow forecasting, but some disadvantages prevent the widespread application. For traffic flow prediction, the proposed work is concentrated on reducing the time complexity as well as improving the accuracy of prediction. By using the clustering mechanism, the time complexity of the algorithm is reduced. By twofold clustering, the data to be analysed by the algorithm is segregated and hence the accuracy is improved. For improving the accuracy of prediction we use a multivariate approach. We also provide a route guidance with traffic flow, which adds novelty to the concept. To implement the concept, we use publicly available London traffic flow dataset. The concept can be further investigated by considering real-time traffic flow data.

Keywords— k-Nearest Neighbour (KNN), Multivariate approach, Shortest route, Traffic flow forecasting, Twofold Clustering.

I. INTRODUCTION

Data mining is the process to discover actionable information from huge sets of data. This process is usually defined as searching, analysing and sifting through large amounts of data to find relationships, patterns, or any significant statistical correlation. The process uses mathematical analysis to derive patterns and trends that exist in data. Predictive and Descriptive are the two main kinds of models in data mining. Predictive model is used to forecast explicit values which are based on patterns determined from known results [11]. The predictive analytics which is used to increase revenues through improved marketing and to reduce costs through detecting and preventing waste. From this technology, organizations of all types are measurable payoffs. Descriptive model describes patterns in existing data that are generally used to create meaningful subgroups [12].

Traffic flow is the study of the movement of individual travellers and infrastructure between two points and the interactions they made with one another, with the aim of understanding and developing an optimal transport network which results in efficient movement of traffic and minimal traffic congestion problems. The estimation of road traffic flow becomes more important particularly these days to our daily life than ever before which is because of rapid increase in vehicle numbers and urban development. Road traffic state information is most important for the successful deployment of Intelligent Transport System (ITS) applications [13]. Traffic control systems for large traffic networks have attracted much attention, recent times. One of the challenges of traffic control system is the prediction of the traffic. The capability to forecast traffic volume in an operational setting is identified as a critical need for ITS. In particular, traffic volume forecasts will support dynamic traffic control. We need to look for the efficient and effective methods that are able to estimate the traffic for any point of time in the future.

Traffic predictions are very important as they enable us to detect potential traffic jam spots. Based on the information provided from a traffic prediction system we could able to provide certain traffic control methods to avoid the traffic jams. One of the most important applications of traffic control systems is the control of road network traffic [13][4].

Traffic has both spatial and temporal features. The traffic on a road is influenced by traffic on nearby roads and the flow on a road section which is correlated with previous flows on the same section. The most studied methods consider the single section forecast and only seldom take into account the relation among links of a road network. They only consider the temporal dimension, but ignore the spatial one, the most common approach is univariate (data from a single sensor are taken into account) and does not consider seasonality (apart from an implicit distinction between working days and holidays). In univariate approach, which in general terms assumes that the variable of interest is influenced by a single factor only. On the other hand, the multivariate approach assumes that the response variable is influenced by multiple factors. Generally, the road network has an underlying graph structure therefore the natural choice for traffic flow prediction problem in the multivariate framework could be Bayesian Networks (BN).BN involve in both probability theory and graph theory which are suitable for dealing uncertainty and complexity[4][5].

II. LITERATURE SURVEY & REALTED WORK

Various techniques has been proposed and research is done in the area of Advanced Traffic Management System (ATMS). Also many advanced methods have been introduced for real time traffic state estimation. Some research work in the area of ATMS is illustrated below:

The nearest neighbour (NN) [2] technique is very simple, efficient and effective in the field of pattern recognition, text categorization, object recognition etc. The main advantage is its simplicity. The structure less method overcomes memory limitation and structure based techniques reduce the computational complexity. The main disadvantage is that it is easily fooled by irrelevant attributes.

The k-nearest neighbour algorithm [5] is a non-parametric machine learning algorithm generally used for classification. It is also known as instance based learning or lazy learning. KNN algorithm can also be adapted for regression, for estimating continuous variables. The standard KNN method suffers from the curse of dimensionality that is the neighbourhood of a given point become very sparse in a high dimensional space, resulting in high variance. Thus in high dimensional data, "nearest" becomes meaningless. Another drawback is over-fitting, as it occurs when a model is more complex, such as having many parameters relative to the number of observations.

Improved KNN in [8] is a correlation-based K-nearest neighbour algorithm. This new algorithm makes data classification based on the correlation calculation, and uses a modified probability to improve the computational speed and prediction accuracy. When it comes to processing massive high-dimensional data sets, one shortcoming of the traditional K-nearest neighbour algorithm is the time complexity of making Classification.

KNN (k-nearest neighbour) [6] is an extensively used classification algorithm owing to its simplicity, ease of implementation and effectiveness. It is one of the top ten data mining algorithms widely applied in various fields. KNN has few shortcomings affecting its accuracy of classification. It has greater memory requirements as well as high time complexity.

KNN is considered [9] as one of the most important methods in short-term traffic forecasting. In this paper, they have used four tests to find the key factors of the KNN method, which inspires to the future research to improve the method but some disadvantages limit the widespread application.

The study in [4] applies Artificial Neural Network (ANN) for short term prediction of traffic flow using past traffic data. Results show that Artificial Neural Network has consistent performance even if time interval for traffic flow prediction was increased and produced good results even though speeds of each category of vehicles were considered separately as input variables. The drawback here is ANN can be only applied for short term traffic flow prediction with mixed traffic conditions.

The k-nearest neighbour (KNN) model [1] is an effective statistical model applied in short-term traffic forecasting that can provide reliable data to guide travelers. This study proposes an improved KNN model to enhance forecasting accuracy based on spatiotemporal correlation. It achieves multistep forecasting of performance in time-varying traffic conditions. The disadvantage is that accuracy will decrease when the time-varying methods are used.

There are completely different methods and solutions existing for traffic flow prediction. But KNN is better than other solution is that prediction is made for a new instance by searching through the entire training set for the K most similar instances. It summarizes the output variable for those K instances. For regression this might be the mean output variable whereas in classification this might be the mode class value. Also, no learning of the model is required and all of the work happens at the time prediction is requested. KNN makes no assumptions about the structure of the problem being solved. Moreover, KNN makes predictions just in time by calculating the similarity between an input sample and each training instance.

III. PROPOSED WORK

K-Nearest Neighbours is one of the simplest algorithm used for classification. The proposed work is focused on predicting the traffic flow using enhanced KNN algorithm. It uses a distance function called Euclidean distance to find out the nearest neighbours. As KNN is a data mining technique, it has a wide range of applications in classification and prediction. KNN is more effective if the training data is large. In our proposal, large traffic dataset is used and hence KNN is more effective.It can not only guarantee the efficiency but also improve the accuracy.

For traffic flow prediction, the work is concentrated on reducing the time complexity as well as improving the accuracy of prediction. By using the clustering mechanism, the time complexity of the algorithm is reduced. By twofold clustering, the data to be analysed by the algorithm is segregated and hence the accuracy is improved. Initially the dataset is pre-processed to fill the missing values and it is divided into training data(80%) and testing data(20%).

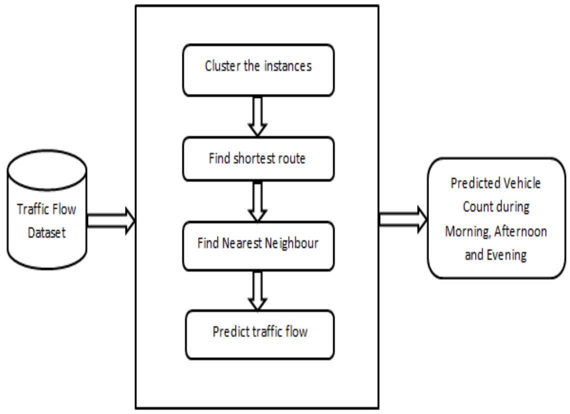

Figure 1 **System Architecture**

Univariate KNN

The big dataset collected for transport needs to be imported into the IDE to make it useful for predicting the traffic data. The dataset available in the .csv (**C**omma **S**eperated **V**alues) format needs to be fetched and converted to the format suitable for processing. Then Univariate KNN is applied to the traffic data for predicting the traffic flow in terms of number of vehicles. In this method, the variables used for prediction are Day, Time of the day (Morning / Afternoon / Night). The pseudocode for univariate KNN:

UKNN(S,P,x) // **S** : Sample training data , **P** : class labels of **S**, **x** : unknown sample
for i=1 **to** n **do**
  Compute distance d(S_i ,x)
end for
Compute set I containing indices for the k smallest distances d(S_i ,x)
return majority label for {P_i where i ϵ I}

A. *Univariate KNN with Clustering*

The functionalities of the Univariate KNN are improved in this module by using clustering mechanism. For Clustering we use K-means Algorithm. First the entire data is clustered on the basis of day into 7 clusters. Then the KNN algorithm is applied to predict the traffic flow. By using the clustering approach, the execution time of the predicting mechanism is reduced with increased accuracy. In this method, the variables used for prediction are Day, Time of the day (Morning / Afternoon / Night). The pseudocode for univariate KNN with clustering :

UKNNC(S,P,x,k) // **S** : Sample training data, **P** :class labels of **S**, **x** : unknown sample, **k** : no. of clusters
Select **k** points as initial centroids
repeat
  Form **k** clusters by assigning all points to closest centroid
  Recompute centroid of each cluster
until The centroid doesn't change
for i=1 **to** n **do**
  Compute distance $d(S_i, x)$
end for
  Compute set I containing indices for the k smallest distances $d(S_i, x)$
  return majority label for {P_i where i ϵ I}

B. *Multivariate KNN with Twofold Clustering*

The functionalities of the Univariate KNN with clustering is improved in this module by considering multiple variables/columns/features and also using twofold clustering mechanism. For Clustering we use K-means Algorithm. First the entire data is clustered on the basis of day into 7 clusters. Secondly the result of the first clustered data is again clustered on the basis of road category into 4 clusters. Then the KNN algorithm is applied to predict the traffic flow. By using the concept of twofold clustering, the execution time of the predicting mechanism is still reduced with increased accuracy. In this method, the variables used for prediction are Day, Morning Vector, Afternoon Vector, Night Vector and Road Category. The pseudocode for multivariate KNN with twofold clustering:

MKNNCC(S,P,x,k1,k2) // **S** : Sample training data , **P** : class labels of **S**, **x** : unknown sample **k1** : no. of outer clusters , **k2** : no. of inner clusters
Select **k1** points as initial centroids
repeat
  Form **k1** clusters by assigning all points to closest centroid
  Recompute centroid of each cluster
until The centroid doesn't change
Select **k2** points as initial centroids from the k1 clustered instances
repeat
  Form **k2** clusters by assigning all points to closest centroid
  Recompute centroid of each cluster
until The centroid doesn't change
for i=1 **to** n **do**
  Compute distance $d(S_i, x)$
end for
Compute set I containing indices for the k smallest distances $d(S_i, x)$
return majority label for {P_i where i ϵ I}

C. *Shortest Route Prediction*

To predict the shortest path from a given source to a destination, we use dijkstra algorithm. The user enters the source and destination. The algorithm traverses through the graph formed for the given source and destination. There may be multiple paths available for a given source and destination. The path with the shortest link length is suggested as the shortest route and the other possible paths are given as alternate routes. Also the traffic flow is predicted for the shortest route during morning, afternoon and evening. The pseudocode for shortest route prediction is:

SRP(G,s,d) // **G** : Graph, **s** : Source City, **d** : Destination City
Initialize the cost of each city to ∞
Initialize the cost of the source city to 0
while there are unknown cities left in the graph
  Select an unknown city b with the lowest link length
  Mark b as known
end while
for each city a adjacent to b
  a's length=min(a's old length, b's length+length of (b, a)
end for

IV. RESULTS AND ANALYSIS

The KNN algorithm is a data mining algorithm which is used to classify and predict the new instance. The algorithm takes as input a dataset as training data and also an unknown sample data. The algorithm calculates the distances and votes on majority labels to predict the new instance. The input is a dataset containing a large number of records and the output is the predicted traffic flow (number of vehicles).The input is based on data collected from the Highways in London. The sample dataset consists of the traffic flow details in London city during morning, afternoon and evening. The implementation of the proposed work is done in java language on Netbeans IDE.

International Innovative Research Journal of Engineering and Technology
ISSN NO: 2456-1983

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Year	CP	Region	LocalAuth	Road	RoadCate	Easting	Northing	StartJunct	EndJuncti	LinkLengt	Morning v	Afternoor	Night Vec	Day
2	2000	6000	London	Barnet	M1	TM	522170	189100	A406	M1 spur	2.55	766	950	2435	Sunday
3	2000	6001	London	Barnet	M1	TM	520150	194300	A2	A4	4.23	2184	2332	5376	Monday
4	2000	6013	London	Hillingdor	M4	TM	508900	178400	A4	A3	1.8	1481	3082	5161	Tuesday
5	2000	6071	London	Tower Ha	A12	TU	538180	181800	A13	A11	1.3	540	1072	2000	Wednesday
6	2000	6075	London	Islington	A1	TU	530600	185870	A503 Cam	A503 Seve	0.12	71	48	63	Thresday
7	2000	6076	London	Barnet	A1	TU	524500	189330	A406	A598 Rege	0.31	149	239	444	Friday
8	2000	6077	London	Barnet	A1	TR	520760	195000	Courtland	A411 Barn	1.12	385	412	759	Saturday
9	2000	6094	London	Southwarl	A2	PU	532700	179360	A3A2198	A201 A10C	0.56	203	122	156	Sunday
10	2000	6095	London	Lewisham	A2	PU	537000	177000	A2 Amersl	A2209	0.37	146	158	279	Monday
11	2000	6096	London	Greenwicl	A2	PU	540600	177000	A2213	A102	0.12	51	49	134	Tuesday
12	2000	6097	London	Bexley	A2	TR	550000	174100	A223	LA Bound	0.81	494	826	1589	Wednesday
13	2000	6107	London	Southwarl	A3	PU	532300	179600	A3202	A2	0.25	165	65	76	Thresday
14	2000	6108	London	Lambeth	A3	PU	531400	178000	A23	A3204	0.56	658	694	1207	Friday
15	2000	6109	London	Wandswo	A3	PU	528400	175200	A205	A3216	0.37	235	158	185	Saturday
16	2000	6111	London	Wandswo	A3	TU	525000	174600	A205	A218	0.19	142	102	150	Sunday
17	2000	6119	London	Westmins	A4	PU	531060	181090	Aldwych	LA Bound	0.19	157	24	26	Monday
18	2000	6120	London	Westmins	A4	PU	528900	180220	Hyde Park	St James S	0.56	773	821	1407	Tuesday
19	2000	6121	London	Kensingto	A4	PU	526000	178870	Earls Cour	Thurloe Pl	1.06	1200	445	549	Wednesday
20	2000	6122	London	Hounslow	A4	TU	519000	178500	M4 jn2	A406	0.68	216	205	351	Thresday
21	2000	6123	London	Hounslow	A4	TU	511000	176600	A312	A30	0.37	58	83	126	Friday
22	2000	6150	London	Barnet	A5	PU	519000	191950	A410	A5100	0.93	83	60	129	Saturday
23	2000	6151	London	Brent	A5	PU	522850	186950	A406	A407	1.24	391	448	558	Sunday
24	2000	6152	London	Westmins	A5	PU	527500	181240	Sussex Gd	Seymour S	0.37	211	228	271	Monday
25	2000	6188	London	Enfield	A10	TU	532600	192000	LA Bound	A406	0.56	134	193	260	Tuesday

Figure 2 **Input Screenshot**

The input dataset consists of the following variables/columns/features :

Year – Traffic volumes are shown for each year from 2000 onwards.

CP (count point) – a unique reference for the road link that links the AADFs to the road network

Road – This is the road name (for instance M25 or A3).

StartJunction – The road name of the start junction of the link

EndJunction – The road name of the end junction of the link

MorningVector – The traffic flow (number of vehicles) during morning

AfternoonVector – The traffic flow (number of vehicles) during afternoon

NightVector – The traffic flow (number of vehicles) during night

Day – The day of the week on which traffic is predicted

RoadCategory – the classification of the road type

The performance of the different implementations of the KNN algorithm are plotted in the form of a graph.

In this paper, two parameters are used for evaluating the performance,

- Time of execution
- Accuracy

A. Time of execution for Traffic Flow Prediction on different days of the week using 3 methodologies of k-Nearest Neighbours algorithm

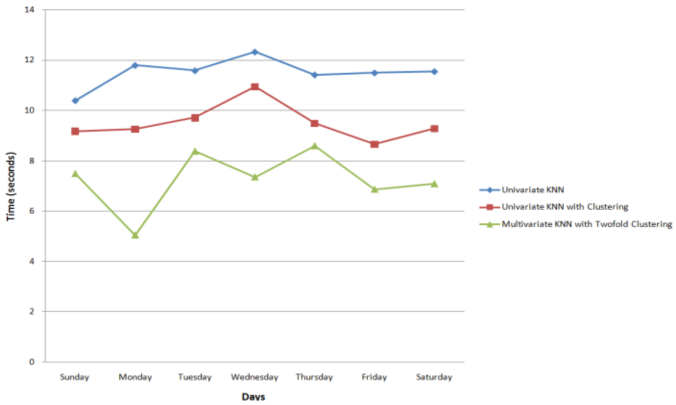

Figure 3 **Time of Execution Graph**

An algorithm is said to be more effective, if it consumes less time for execution. From the above graph, it is observed that the order of time of execution for Multivariate KNN with Twofold Clustering is found to be less than the Univariate KNN with Clustering by 23.6% and time of execution for Univariate KNN with Clustering is found to be less than the Univariate KNN by 17.4%. As the clustering occurs twice in a group of instance, time required to search the data is reduced, so that the execution time gets much reduced. The order of time of execution is

Univariate KNN > Univariate KNN with Clustering > Multivariate KNN with Twofold Clustering

B. Accuracy of Traffic Flow Prediction on different days of the week using 3 methodologies of K-Nearest Neighbours algorithm

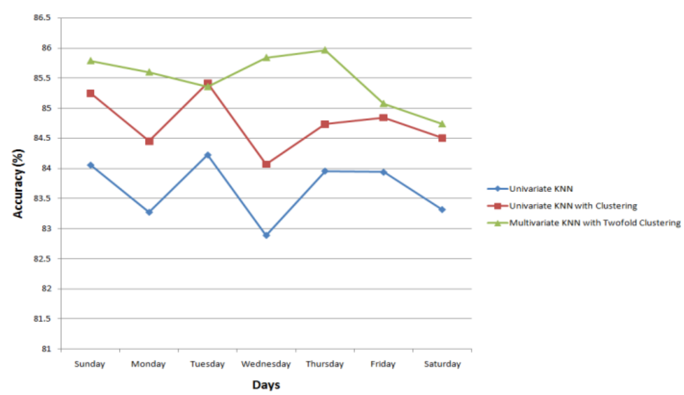

Figure 4 **Accuracy Graph**

As accuracy refers to the nearness of a measured value to an expected value it is observed that from the above graph, accuracy for Multivariate KNN with Twofold Clustering is found to be higher than the Univariate KNN with Clustering by 0.86% and accuracy for Univariate KNN with Clustering is found to be higher than the Univariate KNN by 1.3%. It is because of implementation of twofold clustering technique in our multivariate approach, as the traffic data to be analyzed by the algorithm is isolated, so that only relevant instances are considered for analytics and hence the accuracy is increased. The order of accuracy is

Multivariate KNN with Twofold Clustering > Univariate KNN with Clustering > Univariate KNN

V. CONCLUSION AND FUTURE WORK

Traffic flow forecasting is an important part in Intelligent Transportation system (ITS). The forecasting results can provide the travellers with useful information that helps the travellers to choose better routes and acquire route guidance, so as to lessen the travel time and avoid traffic jams. With the advent of computer technology, the size of data increases gradually and how to enhance the algorithm's accuracy more effectively appears to be particularly important. In this paper, to improve the forecasting accuracy, we have proposed the concept of multivariate KNN with twofold clustering. The experimental results proved that the proposed algorithm can further enhance the accuracy in predicting the traffic flow. Moreover time complexity of the algorithm is reduced. Also we have found the shortest route and alternate routes for a given source and a destination. However, the future work can be done by predicting the traffic flow using real-time data which can be collected through wireless sensors, GPS enabled mobile phone equipped vehicles.

REFERENCES

[1]Pinlong Cai, Yun Peng Wang, "A spatiotemporal correlative k-nearest neighbour model for short-term traffic multistep forecasting," *IEEE Transactions on Intelligent Transportation Systems* 62,pp.21–34,2016.

[2]Nitin Bhatia Vandana, "Survey of Nearest Neighbour Techniques," *International Journal of Computer Science and Information Security (IJCSIS),* vol.8, No. 2, 2010.

[3]Filmon G.Habtemichael, "Short-term traffic flow rate forecasting based on identifying Similar traffic patterns," *IEEE Transaction on Intelligent Transportation systems* 66, pp.61-78, 2016.

[4]Kranti Kumar, M. Parida, V.K, "Short term traffic flow prediction for a non-urban highway using Artificial Neural Network," *Social and Behavioral Sciences* 104, pp.755 764, 2013.

[5]Minakshi Sharma Suresh Kumar Sharm, "Generalized K-Nearest Neighbour Algorithm- A Predicting Tool," *International Journal of Advanced Research in Computer Science and Software Engineering* Vol.3,pp.1-4 Issue 11, November 2013.

[6]Shweta Taneja, Charu Gupta, "An Enhanced K-Nearest Neighbour Algorithm Using Information Gain and Clustering," *IEEE transactions on Advanced Computing & Communication Technologies*,pp. 325-329,2014.

[7]Hosein Alizadeh,"A New Method for Improving the Performance of K Nearest Neighbour using Clustering Technique," *Journal of Convergence Information Technology* Vol.4, No.2, June 2009.

[8]Xinran Li Chenhui Xiang, "Correlation-based K-Nearest Neighbour Algorithm," *IEEETransactions on Intelligent Transportation systems* 2012, pp.185-187.

[9]Jing-ting Zhong, "Key Factors of K-nearest Neighbours Nonparametric Regression in Short-time Traffic Flow Forecasting," *Industrial Engineering and Engineering Management* (IEEM),pp.9-12,2014.

[10]DawenXia, BinfengWang, "A distributed spatial–temporal weighted model on MapReduce for short-term traffic flow forecasting," *Neurocomputing*, pp.246–263, December 17 2015.

[11] https://msdn.microsoft.com>library

[12] https://www.fhwa.dot.gov>tft<chap2

[13] www.investopedia.com/predictive-analytics.asp

Author Biography

Dr.Santhi Baskaran

She received her B.E. degree in CSE from Pondicherry University, M.Tech. degree in CSE from University of Madras and Ph.D degree in CSE from Pondicherry University. She is working as Professor in the Department of Information Technology, Pondicherry Engineering College. She is a Life member of ISTE.

S.Lakshmi@Vaishnavi

She is pursuing her B.Tech degree in the Department of Information Technology in Pondicherry Engineering College from Pondicherry University.

K.Manisha Selva

She is pursuing her B.Tech degree in the Department of Information Technology in Pondicherry Engineering College from Pondicherry University.

K.Keerthana

She is pursuing her B.Tech degree in the Department of Information Technology in Pondicherry Engineering College from Pondicherry University.

K.Revathi

She is pursuing her B.Tech degree in the Department of Information Technology in Pondicherry Engineering College from Pondicherry University.