Traffic Congestion Detection By Mining GPS Data

Suhas Prakash Kaklij
Department of Information Technology
Siddhant College of Engineering, Sudumbare
Pune, India
Email: kaklijsuhas@gmail.com

Prof. Sonali Rangdale
Department of Information Technology
Siddhant College of Engineering, Sudumbare
Pune, India
Email:sonali_rangdale@rediffmail.com

***Abstract*- GPS data is available in large amount, also for devices having GPS a large amount data is being gathered over time. The mining of this huge data is to assist in discovery of the locations which face regular traffic congestion. User will have prior knowledge of such locations which helps in deciding whether or not to opt for that route. Avoidance of such path will also assist in reduction of congestion in such locations. Also observed that work done till now in this field does not give very precise results. The reason behind this is the no proper algorithm are selected and distinguished between on road and off road traffic. To consider all this we proposed this system. This system will be applied over GPS data i.e. data coming from verity of devices like mobile phones, tablets etc. In the technique with this system, these GPS data will be first cauterized using the K-means clustering algorithm. The clusters obtained are filtered out. On further processing these clusters a mining method of Naive bayes algorithm is used for mining for traffic Congestion detection and prediction**

Keywords-K-Means,Traffic Congestion,Traffic Jam Predicion,

I. INTRODUCTION

Road network is biggest network widely used for Transportation .Each city has its Road network. Road are used to travel from one point to another destination point. It is used for daily transport not only for the people but for goods and many more. The biggest problem now a days people facing is Traffic Congestion. People are not able to complete their work due this traffic problem. Also we observed that traffic congestion is of dynamic nature, it is not static. Means traffic congestion is variable as time passes. In current IT world we have lots of traffic information available with us in different format. By using this we can get the flow of traffic information with respect to location and time. This traffic information is important not only for current status of traffic but it can helps to analyze and predict upcoming traffic patterns. We can collect such information by processing GPS data. With availability of 2G and 3G enabled GPS devices, huge datasets are being collected with an average error of 2-15m [2]. Using many of correction strategies such as map-based correction given in [2], this error can further be decrease. This is real time data which gives an opportunity to mine the traffic patterns of particular location .We can analyze such data to get the traffic congestion patterns which in turn helps to detect the location where traffic congestion is possible .Also we can predict possible traffic congestion.

A. Related Work

There is a massive amount of work undertaking in field of analyzing traffic patterns. H. Inose et al. in 1967, as given in [11], proposed how traffic signals are work systematically. Its work proposes for the minimization of delay time of vehicles and allocating preferential offsets to the optimum tree in a road network. In 2002, Ashbrook et al., as given in [10], projected user substantial locations and user activities using GPS data. Their work divided city in to various clusters using K-means clustering which further resulted into a Markov Model. Thus, their work focusses on analyzing user GPS data to mine user-significant locations. As per 2010, Lipan et al. in [5], mined traffic patterns from GPS data gathered from public transport. Their work focuses on monitoring bus schedules. Association rules are made on clusters in which each cluster has its own average speed. In 2011 [2] Mandal K and his team used probe vehicle technique for traffic congestion monitoring, system as whole tries to monitor the traffic flow pattern and then detect the congestion. As given in [3], Yao et al. proposed a speed pattern model which guesses traffic conditions and speed pattern using machine learning. In 2013[1] Anand Gupta and his team proposed a framework for traffic congestion detection focusing more on algorithm which reduces conflict of data for traffic Jam and Traffic signal. These works have given significant and helpful results. However, to the best of the authors' findings, not much emphasis has been given to detection and prediction of traffic congestion with appropriately handling of on road & off road data as well as the Conflict between the Traffic signal and Traffic Jam .Also no proper selection of mining and clustering algorithm

B. Motivation

Detecting traffic jam based on simple rules, such as using a probe vehicle technique, velocity-based approach, and fuzzy logic might not handle the problem stated previously with great effect due to the following reasons

1. Proper Clustering algorithms not selected.
2. Suitable mining methods are not used
3. Not able to segregate on road and off road traffic

This are basic motivation factors for developing such TCD –framework

C. *Contribution*

To achieve the points mention in the motivation, we proposed a framework called TDC- In which T stand for “Traffic”, C for “Congestion”, and D for “Detection”. This framework is hybrid approach of two different top methods namely K-means and Naive Bayes which together helps to give more accurate result for Traffic congestion detection. Whereas framework proposed by Anand Gupta and it team more emphasis given on algorithm which reduces conflict of data for traffic Jam and Traffic signal

D. *Organisation*

To explain the framework - TCD, We organized paper as follows: Section-II explains the structure of the framework and the algorithms associated with it. Section-III Shows Algorithm used in framework Section IV Shows how Experiment is to be carried out Section V shows the expected result as obtained. Section-VI concludes the paper describing

II. PRAPOSED FRAMEWORK

Description of DTC Framework (Refer to Fig 2): The logs files of GPS that are being collected overtime comprise of GPS data coming from Mobiles phones, Notepad devices or GPS enable devices .Data might be in different forms like for users who might be in a motionless state like sitting or in motion such as walking. The initial GPS data collected from mobile devices have log with information of Device ID, Location details like <Lat, Long>, Time related data like Time and date is sent through first stage Data generation module to Data importing module where we load this attributes to database .Many time GPS data available in two different form one having speed information as attribute and one without speed attribute. In our data generation module we do not have speed attribute generated, so we can calculate the it by using Haversine Formula in fig 1

In this formula Φ1, Φ2 are the latitudes of two given points and ʎ1, ʎ2 are their corresponding longitudes. Now we have stored the attributes to data base as <Device Id, Latitude, Longitude, Time, Speed, and Day>

$$d = 2r \arcsin(\sqrt{sin^2\left(\frac{\phi 2-\phi 1}{2}\right)+\cos(\phi 1)\cos(\phi 2)\,sin^2\left(\frac{\text{ʎ}2-\text{ʎ}1}{2}\right)})$$

Fig.1. Haversine formula to calculate distance between two points

All this information stored is used to detect the location of the particular device id. Once Location is detected for device id then by using location information and the speed of the device id we can categories it into On Road and Off Road. The logic to get On Road and Off Road data is simple .We have Device ID with its location including the <latitude, longitude> ,We also have the <latitude, longitude> of the Road . So coordinates which are not in the data set of Road data coordinates all such devices are categorized as Off Road data and Device Id which are in the range of the Road <latitude, longitude> data is categorized as on Road data to avoid the confusion between data set . Now proceed with On Road data and ignored the Off Road data. The on road data is further inputted to Filtering process in which data is process in 2 parts .In part 1, analysis of data is done where data is analyzed from different perspective so that it helps to determine the transportation medium of each Device ID .In this process first threshold average speed of each device is decided so according to data available along with the speed and distance information average speed is calculated and the transportation medium is decided for each device id.

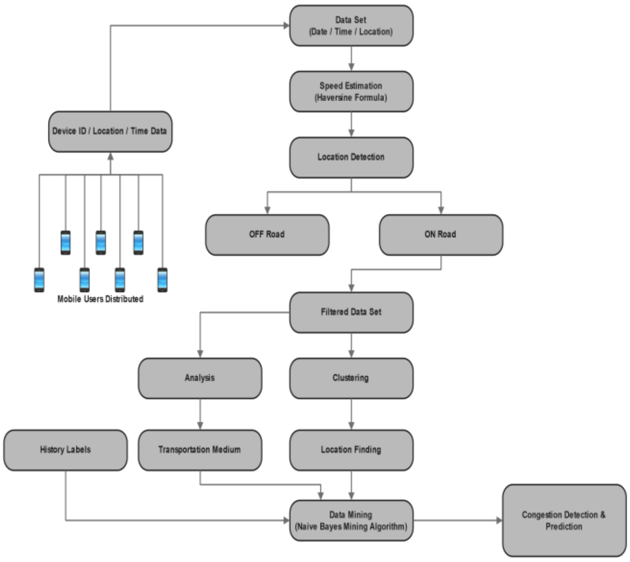

Fig. 2. TCD Framework

In part 2, with the help of latitude, longitude, average speed and a unique Cluster-ID divide a city map into clusters of different sizes. We have chosen the most efficient, faster K Means clustering algorithm. The GPS raw data available with all the above parameters are applied to this K-Means clustering .The one of the main property for selection of the K-Means is whatever clusters are resulted those are non-hierarchical and they do not overlap, also the K-Means produces tighter clusters than hierarchical clustering [7]. Once clustering is do with help of the clusters location of each cluster is detected [6] The output of both part 1 & 2 is inputted to mining process where data is process by Naive Bayes Algorithm [4] for traffic congestion detection. This is more advance over the J48 decision tree classifier [8]. We also have the Historical data available in addition to part 1 & 2 data so that efficient traffic congestion prediction is done

III. ALGORITHM

1. Clustering K-Means Algorithm:

K-means is used to solve the familiar clustering problem. In this a given data set is classify in to specified number of clusters. Main focus is to allocate k number of centroid for each cluster. As different location causes different result, these centroids should be placed in a crafty way. The good choice is to place each centroid as much far as possible. Now choose each point associated to a given data set and position it to the closest centroid. When all points are covered, the first step is completed. Now we need to re-calculate k new centroids as center of the clusters resulting from the last step. Once we have these k new centroids, again previous data set points and closest new centroid needs to tie together. We may notice that the k centroids keep changing their location one by one all changes are done. This algorithm focuses on minimizing an objective function, the objective function [6]

$$J = \sum_{j=1}^{k}\sum_{i=1}^{n}\left\|x_i^{(j)} - c_j\right\|^2 \qquad (1)$$

Where $\left\|x_i^{(j)} - c_j\right\|$ = distance measure between a data

point Xi and the cluster center Cj, *n* is an indicator of the distance of the data points from their respective cluster centers.

Algorithm steps:

1: Assign K points considering objects which are going to clustered .These are the first points of centroid
2: Check the group which has closest centroid and assign object to that group for checking the latest centroid
3: Check all objects are considered for allocation or not, once done then every time recalculate K-centroid Position.
4: Repeat Steps 2 and 3 until no more changes are done. This yields a separation of the objects into groups from which the metric to be minimized can be calculated.

Consider that we have n sample feature x1, x2, ..., xn all from the same class, and we are aware of that they fall into k compact clusters, k < n. Assume mi be the mean of the vectors in cluster i. If the clusters are well detached, we can use a minimum-distance classifier to separate them. We can say that x is in cluster i if || x - mi || is the minimum of all the k distances. This suggests the below method for finding the k means:

Make initial prediction for the means m1, m2... mk

```
1: Continue till there are no changes in any mean
2: Use the projected means to classify the samples into clusters
3: For i from 1 to k
4: Swap mi with the mean of all of samples for cluster i
5: end_for
6: end_until
```

2. Naive Bayes Algorithm:

A naive Bayes classifier consider that the absence (or presence) of a specific feature of a class is unrelated to the presence (or absence) of any other feature, given the class variable. The Naive Bayes classifier is created on Bayes' theorem with independence assumptions between predictors. A Naive Bayesian model is easy to construct, with no complex iterative parameter estimation which makes it mostly useful for very large datasets [4]

The naive Bayesian classifier Algorithm Steps:

Step 1.Consider D be a training set of tuples and their associated class labels. Every tuple is characterized by an n-dimensional attribute, X=(x1, x2… xn), describing n measurements made on the tuple from n attributes, respectively,A1,A2...An.

Step 2. Assume that there are m classes, C1, C2… Cm. Given a tuple, X, the classifier will expect that X belongs to the class having the greatest posterior probability, conditioned on X. That is, the Naïve Bayesian classifier expects that tuple x belongs to the class Ci if and only if

P (Ci|X) > P (Cj|X) for every 1≤ j≤m and j ≠ i

Thus we try to maximize P(Ci|X). The class Ci for which P (Ci|X) is maximized is called the maximum posteriori hypothesis. By Bayes' theorem

$$P(Ci|X) = \frac{P(X|Ci)P(Ci)}{P(X)} \qquad (2)$$

Step 3: Now P(X) is constant for all classes, only P (X|Ci) P (Ci) need be maximized. If the class prior probabilities are not known upfront, then it is commonly supposed that the classes are equally likely, that is, P (C1) =P (C2) =P (Cm), and we would therefore maximize P (X|Ci). Else, we maximize P (X|Ci) P (Ci).

Step 4: Known data sets with many attributes, it would be extremely expensive to compute P (X|Ci). In order to decrease computation in evaluating P (X|Ci), the naïve assumption of class conditional independence is made. This supposes that the values of the attributes are conditionally independent of each another shows (3) and (4), given the class label of the tuple (i.e., that there are no dependence relationships among the attributes).

Thus,

$$P(X|Ci) = \prod_{K=1}^{N} P(xK|Ci) \qquad (3)$$

$$= P\ (X1|Ci) \times P\ (X2|Ci) \times\ldots \qquad (4)$$

So the predicted class label is the class Ci for which P (X|Ci) P (Ci) is the maximum.

(a) Consider that if Ak is categorical, then P (Xk|Ci) is the number of tuples of class Ci in D showing the value xk for Ak, divided by |Ci, D|, where |Ci, D| is the number of tuples of class Ci in D. Now using the values of the Ci and Xk we can check for relation between then with appropriate selection of parameters

(b) A continuous-valued attribute is typically supposed to have a Gaussian distribution with a mean μ and standard deviation σ as per (5) and (6)

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{(x-\mu)^2}{2\sigma^2}} \qquad (5)$$

So

$$P\ (x_k|C_i) = g\ (x_k, \mu_{ci}, \sigma_{ci}) \qquad (6)$$

Step 5: So to predict the class label of X, P (X|Ci) P (Ci) is calculated for each class Ci. The classifier expects that the class label of tuple X is the class Ci if and only if

$$P\ (X|C_i)\ P\ (C_i) > P\ (X|C_j)\ P\ (C_j) \qquad (7)$$

So the predicted class label is the class Ci for which we result P (X|Ci) P (Ci) is the maximum.

IV. EXPERIMENT

The experiment will be carried out on Lenovo Intel® Core™ i5 Processor, 6 GB RAM, Windows 7 64 bit OS.Complete Experiment will be carried out on Single node Hadoop framework. Database used is MongoDB .Planning to consider 300+ dataset from the different GPS enables phones. Data set which don't have speed attributes for them speed is calculated using the Haversine formula as shown in Table I.

TABLE I. TABLE AFTER PROCESSING WITH HAVERSINE FORMULA

Longitude	Latitude	Distance (km)	Time	Speed(km/hr)
57.45879126	22.4157892	0.05563	11:21:46	28.6162
57.45924545	22.4161793	----	11:21:53	----

This data is then inputted to location detection module to detect the exact location of the device. Now the on road and off road traffic is detected. After that filtering of the data set carried out this further provides input to the K-means clustering algorithm which results in to different clusters. Meanwhile the same dataset is processed through the Transport medium segregation module .Once transport medium is identified ,the clustered data and Transport medium segregated data is provide as input to the Naïve bayes mining algorithm . It also uses the historical data as input to do the prediction. Hadoop Distributed File System (HDFS) will be used in the experiment, so this framework can handle the huge amount of data. Also the map-reduced programming method used which help to reduce the time of execution.

V. EXPECTED RESULT

The experiment on framework will give the adequate results. We will check all real-time information available from GPS and the prediction done by the framework. The result obtained by the framework will be plotted on graph. In this "Actual "plotted on x-axis and "Predication" Plotted on Y-axis .To avoid the obscuration and to spread the data we will add the 65% jitter in figure. We are expecting the prediction up to 95%.

There is chance of mispredictions due the few cases like traffic Signal and frequent change in transport mode that is to even 6%.So the overall expected accuracy is 89% .It is seen that mostly mispredicution happen due to the traffic signal .So It is expected that Framework will predict and detect the area where the congestion is happen with maximum accuracy

VI .CONCLUSTION

In the present paper, innovative framework is proposed to detect frequent traffic congestion areas using the data coming in from different kinds of GPS enabled devices like mobiles. With expected correctness up to 89%. Different approach to detect & predict congestion is mention. Different modules are used to segregate on road and off road traffic .Framework has capability to segregate traffic medium. All this features help to reduce the drawbacks of previous framework. The best clustering and mining algorithm used to achieve maximum accuracy. The framework is flexible to different cities by changing the city-dependent thresholds. Also use of Hadoop framework gives capability to handle the traffic big data with improved performance.

ACKNOWLEDGMENT

Author would like to take this occasion to express our profound gratitude and sincere regard to my Project Guide Prof. Sonali Rangdale, for her guidance, valuable feedback and constant help throughout the duration of the project. Her valuable suggestions were of immense help throughout my project work. Her perceptive criticism kept me working to make this project in a much better way. Working under her was an extremely knowledgeable experience for me.

REFERENCES

[1] A.Gupta, S. Choudhary, S. Paul;" DTC: A Framework to Detect Traffic Congestion by Mining versatile GPS data" In the 1st international Conference Emerging Trends & Application in Computer Science (ICETACS 01), pp.97-103, Sept 2013

[2] K.Mandal "Road Traffic Congestion Monitoring and Measurement using Active RFID and GSM Technology" In Intelligent Transportation Systems (ITSC), 14th International IEEE Conference, pp. 1375 - 1379, Oct 2011

[3] Y. H. Ho, Y. C. Wu, M. C. Chen, T.J. Wen, Y.S. Sun ; "GPS Data Based Urban Guidance", In the Proceedings of International Conference of Advances in Social Networks, pp. 703-708, Kaohsiung, Taiwan, July 25-27, 2011

[4] Duan Wei ,Lu Xiang-yang "Weighted Naive Bayesian Classifier Model Based on Information Gain " Intelligent System Design and Engineering Application (ISDEA), International Conference on Oct-2010

[5] F. Lipan and A. Groza; "Mining traffic patterns from public transportation GPS data", In the Proceedings of the 6th International Conference on Intelligent Computer Communication pp. 123- 126, Cluj-Napoca, Romania, August 26 - 28, 2010.

[6] Shi Na , Liu Xumin , Guan Yong ," Research on k-means Clustering Algorithm: An Improved k-means Clustering

Algorithm” Intelligent Information Technology and Security Informatics (IITSI), Third International Symposium on April 2010

[7] G. Nathiya, S. C. Punitha, M. Punithavalli; ”An Analytical Study on Behavior of Clusters Using K Means, EM and K* Means Algorithm”, In International Journal of Computer Science and Information Security, Volume 7, Issue 3, 2010

[8] V.P. Bresfelean; ”Analysis and Predictions on Students’ Behavior Using Decision Trees in Weka Environment”, In the Proceedings of Information Technology Interfaces (ITI ’07), Cavtat / Dubrovnik, Croatia, pp.51-56, June 25-28, 2007

[9] M. Modsching, R. Kramer, K.T. Hagen; ”Field trial on GPS Accuracy in a medium size city: The influence of built up”, In the Proceedings of 3rd Workshop on Positioning, Navigation and Communication (WPNC 06), pp. 209-218, Hannover, March 16, 2006

[10] D. Ashbrook, T. Starner; ”Learning significant locations and predicting user movement with GPS”, In the Proceedings of Sixth IEEE Interna- tional Symposium on Wearable Computers (ISWC 02), pp. 101-108, Seattle, WA, 2002

[11] H. Inose, H. Fujisaki, T. Hamada; ”Theory of road-traffic control based on macroscopic traffic model”, Electronics Letters, Volume 3 , Issue 8, pp. 385- 386, 1967.