RESEARCH ARTICLE OPEN ACCESS

Decision Tree Classification for Traffic Congestion Detection Using Data Mining

R.Sujatha[1], R.Anitha Nithya[2], S.Subhapradha[3], S.Srinithibharathi[4]

[1,2,3,4]Computer Science, Sri Krishna College Of Technology, Coimbatore, Tamilnadu, India

Abstract:

Traffic congestion could be a challenge in Gauteng province of Republic of South Africa and it's a negative impact on the economy of this province therein services and merchandise aren't being rendered on time. Holdup affects the standard of lives of Gauteng residents and guests alike. This information was used for constructing the traffic flow prediction models. To propose a way to spot road holdup levels from rate of mobile sensors with high accuracy . The slippery windows technique to capture the consecutive moving average velocities, that was referred to as a moving pattern .Then road users' judgments and connected info were learned utilizing a decision tree model classification algorithmic program. The evaluations disclosed that the decision tree model .The achieved Associate in Nursing overall accuracy as high as 91% with a precision as high as 94%.

Keywords: Classification, Decision tree, Holdup.

I. INTRODUCTION

This article provides you associate introduction regarding holdup detection and their causes and effects. The most plan behind this project is to find and predict holdup with high accuracy. Holdup could be a challenge in Gauteng province of South Africa and it's a dire impact on the economy of this province therein services and merchandise aren't being rendered on time. The dearth of promptness from staff attributable to traffic jams conjointly affects the economy negatively. It's calculable that each month firms lose close to R1.1 billion within the sort of paying salaries. This loss excludes alternative prices ensuing from the cancellation of conferences. Statistics show that V-E Day of employees' conferences area unit delayed owing to holdup. Holdup affects the standard of lives of Gauteng residents and guests alike, leading to longer than necessary driving time. Additionally ,fuel emissions from vehicles contribute to greenhouse gases. Moreover, holdup will forestall emergency services from reaching people whose lives could be in peril. Among the causes of holdup area unit incidents like automobile accidents, roadblocks, and lane closure attributable to road constructions.

I. PROBLEM FORMULATION

The Most traffic congestion detection methods use numerical statistics on data, to the contrary, approach [3] uses data cubes on historical floating car data to detect and identify recurrent traffic congestion. Historic traffic data is first pre processed and mapped to the road network .Then the decision tree Classification is performed by J48 Algorithm. This algorithm tries to divide the large data into smaller sets until the most homogeneous sets (classes) are generated. In the division process, each attribute is compared to a defined value(s) and separated accordingly. The information gain can be used to measure which attribute is the best predictor. Regression algorithm is implemented in order to achieve high accuracy.

A. Classification Of Traffic Pattern

Analyses on *the dynamics of traffic flow,* starting from intersection flows *to* network-wide *flow* propagation, need correct info on time-varying native traffic flows. To effectively confirm the flow performance measures and consequently the congestion indicators of segmental road items, the power to method such knowledge in real time is out of the question. During this article, a dynamic approach to specify flow pattern variations is projected chiefly concentrating on the incorporation

of neural network theory to produce time mapping for *traffic density* at *the* same time *in* conjunction with a large *traffic flow* model. To upset the noise and therefore the wide scatter of raw flow measures, a filtering is applied before modelling processes [1]. Filtered knowledge are dynamically and at the same time input to processes of neural *density* mapping and *traffic flow* modelling. *The* classification *of flow* patterns over *the* elemental diagram, which is dynamically premeditated with the outputs of the flow modelling sub method, is obtained by considering the density live as a pattern indicator. Densities are mapped by chosen neural approximation technique for every simulation time step considering expressly the flow conservation principle. At the same time, mapped densities are matched over the elemental diagram to specify this corresponding flow pattern [1]. The approach is promising in capturing fulminate changes on flow patterns and is receptive be utilised at intervals a series of intelligent management methods together with particularly non repeated congestion result detection and management.

B. Analyzing Highway Flow Pattern

Historical *traffic* patterns may be used for *the* prediction *of traffic* flows, *as* input for gross *traffic models,* for *the* imputation *of* missing or incorrect information *and as a* basis for *traffic* management situations. This paper investigates the determination of historical traffic patterns by means that of Ward has stratified bunch procedure. Days were clustered before and once a pre-classification into operating days and non-working days, mistreatment 2 completely different definitions *of a* daily *traffic* profile [2]. *The* results *of the* bunch once pre classification square measure clearly higher than before classification .Moreover, operating days square measure easier to classify into distinctive , perennial traffic patterns than non-working days. Finally, bunch on the premise of *quarter-hour* traffic flows *resulted in* an exceedingly higher *classification* of operating days *than* the ballroom dancing bunch that used the daily traffic flow, peak flows, peak times and ratios. The bunch *on* the *premise* of *quarter-hour* traffic flows *resulted in* an exceedingly classification into 5 clusters that show distinct daily flow profiles and square measure representative for the times among the clusters. The day of the week and vacation periods square measure determinative for the cluster a operating day is assessed to. For a Dutch route location, daily traffic profiles square measure sorted by means that of a stratified Ward's bunch procedure mistreatment varied approaches. It is terminated that a pre-classification into operating days and non-working days considerably improves the bunch result. Since the variation among the clusters ensuing from *the* bunch while not pre-classification *is* giant, *the* typical daily *flow* profiles *of* those clusters are not representative for all days among the clusters. This massive variation is because of large variations between daily flow profiles of operating days and non-working days. A pre-classification into operating days and non-working days a filter the most important variations go into advance and allows the bunch rule to sight smaller variations. These results correspond to the results of [6] that showed *traffic* forecasts *were* higher just in case *of a* pre-classification into day-groups and conditions. Secondly, operating day's square measure easier to classify than non-working days. The variation among the non-working day clusters *is* somewhat giant *and the* average daily *traffic* profiles so do not represent *the* daily traffic profiles of all days among the clusters. Moreover, the clusters square measure already comparatively tiny, thus a classification into additional clusters isn't any choice. In all probability, non-working days show less fastened activity patterns compared to operating days. Additional information from contract years can in all probability result in a far better bunch result. Particularly since the cluster a non-working day is assessed to, appears to be addicted to the season. Moreover, for operating days, the bunch on the premise *of* quarter-hour *traffic* flows resulted *in an* exceedingly higher classification than *the* ballroom dancing bunch. The latter resulted in an exceedingly classification into too several tiny clusters that don't represent perennial *traffic* patterns [6]. *The* classification into these tiny clusters *is* perhaps because *of the* very fact that operating days show extremely similar daily *traffic* profiles. Bunch on *the* premise *of a* number *of* the separate options resulted in an exceedingly classification into multiple clusters,

whereas in fact just one cluster exists. Finally, on the premise of the resultant classification of operating days it's terminated that four sorts of operating day scan be distinguished: (1) Mondays, (2) core week days, (3)Fridays and (4) days among vacation periods. These results correspond to the results of [2], World Health Organization additionally found variations between Mondays, core week days and Fridays and partially to the results of [3], World Health Organization categorised Fridays in an exceedingly separate class from the opposite week days. The resultant operating day patterns may be used as input for gross traffic models and *as a* basis *for* traffic management *situations.* Moreover, once predicting traffic flows *on* the premise of historical information, a pre-classification into vacation day, Mondays, core weekdays and rides may be created. Finally, these patterns may be accustomed sight and replace incorrect information and to impute missing information.

C. Identifying Urban Traffic Congestion Pattern

With *the* increasing quantity of *traffic* info *collected* through floating automobile *information,* it is extremely fascinating to seek out meaning traffic patterns like congestion patterns from the accumulated large historical dataset. It is but difficult thanks to the large size of the dataset and therefore the complexness and dynamics of traffic phenomena. A unique floating automobile information analysis methodology supported information cube for congestion pattern exploration is projected during this paper [3]. This methodology is very different from ancient ways that rely solely on numerical statistics of *traffic information. The* read of *the* event *or* spatial-temporal progress *is* customized *to* model *and* live *traffic* congestions. *Consistent* with *a* multi-dimensional analysis framework, *the traffic* jam event is initial known supported spatial-temporal connected relationship of slow-speed road section. Then, it's aggregate by *a* cluster vogue *to* induce *the approach* pattern on *a* unique level *of* detail *of* spatial-temporal dimension. Aggregate location, period and period time for perennial and necessary congestions area unit accustomed represent the congestion pattern. We have a tendency to evaluate our ways employing a historical traffic dataset collected from concerning 12000 taxi-based floating cars for one week in a very giant geographic area. *Results* show *that* the tactic will effectively determine *and* summarize the *congestion* pattern with economical computation and reduced storage price. Traffic congestion events occur once traffic demand is larger than the out there road capability. It is a dynamic spatial-temporal method. Congestion sometimes starts from one road section, then expands on the road and influences the close roads. As time passes by, those full fragments shrink slowly, eventually scale back their coverage and eventually disappear. During this method, full road section area unit spatially shut and temporally approximate. At an equivalent, traffic jam is expounded with several factors, and frequently shows distinction at totally different location or in numerous period and similarity or repetition at similar conditions [3].

D. Detection Of Potential Traffic Jam Based Of Traffic Characteristics Data Analysis

Traffic jam is one among massive and sophisticated issues that happens in several big cities round the world. The advances in technology have enabled pursuit-moving objects as if vehicles travel road networks. That knowledge square measure known as spatio-temporal data [4]. By victimisation data processing techniques, analysis on traffic characteristics may be conducted to sight potential space of traffic congestion. This data is especially helpful for transportation bureau to create applicable call concerning traffic policy. During this paper, we tend to propose a technique to investigate knowledge of traffic characteristics that consists of cluster traffic characteristic knowledge, ranking method, and analysing the cluster with relation to the traffic congestion criteria. A mental image is additionally applied to administer higher understanding of the analysis results. The experiments show that data processing of spatio-temporal knowledge may be wont to analyse traffic characteristic, nonetheless it may neither establish the supply of the traffic congestion nor the full routes. Though there is no mechanical phenomenon knowledge, we tend to square measure still ready to sight of potential traffic congestion supported traffic characteristic knowledge. To do so,

we tend to propose a technique consists of (i) cluster traffic characteristic knowledge, (ii) ranking method, and (iii) analysing the clusters with relation to the traffic congestion criteria .Finally, mental image is applied for example the potential space of traffic congestion. The results of our experiment show that our projected technique couldn't establish supply of the traffic congestion. This is often chiefly as a result of our knowledge has not contained any data concerning vehicle mechanical phenomenon. Yet, it couldn't establish full routes. With this supported result, our future work can concentrate on analysing alternative styles of spatio-temporal knowledge to support Indonesia government.

E. Traffic Jam Detection Using Flock Mining

The widespread use of GPS devices on cars allows the gathering of time-dependent positions of vehicles and, hence, of their movements on the road network. it's doable to investigate such Brobdingnagian assortment of knowledge to appear for crucial state of affairs on the traffic flow[5]. The offline analysis of traffic congestions represents a difficult task for urban quality managers. This type of research will be employed by the traffic planner to predict future areas of traffic congestions, or to enhance the accessibility to specific attraction points in an exceedingly town. Several traffic systems adopt ad-hoc sensors like cameras, induction loops, and magnetic sensors to watch the standing of the traffic flows: these systems square measure terribly costly for installation and maintenance, and those they square measure restricted to the native observance of the road arcs wherever they're put in. On the contrary, the utilization of GPS knowledge to see the traffic conditions needs low installation prices (a half for the installation on the vehicle)and it allows to just about observance the whole road network. We gift associate degree innovative tool that exploits the information *collected* from *GPS* enabled *cars* to notice *the* occurrences *of* traffic jams on the road network. The detection of potential traffic jams is predicated on the invention of slowly moving flock patterns, i.e., a collection of objects slowly moving along for a minimum *quantity* of time [5]. The tool *has* been integrated within the *M-Atlas system* exploiting *the* implementation *of* the T-Flock *algorithmic* rule provided by the system. *To the most effective* of our information, *this* is often the primary *system* that *uses* GPS knowledge, *combined* with flock mining, to notice traffic congestions. Most of the approaches accessible within the literature for traffic analysis square measure supported aggregation of spacial or temporal knowledge specializing in predefined areas. It's necessary to signifies that this tool doesn't give real time analysis, however instead it permits the analysis of the historical knowledge. It will highlight the tight *integration* of the *spatio-temporal and* data processing *tools of* the *M-Atlas system* and the graphical computer program that assists the DM analyst in driving his/her analysis.

F. Visual Traffic Jam Analysis

During this work, we *have a tendency to* gift AN *interactive* system *for visual analysis of* urban holdup supported GPS trajectories .*For* some trajectories we *have a tendency to* develop ways *to* extract *and* derive holdup info. *When* cleanup the *trajectories,* they are matched *to a* road network. Later on, traffic speed on every road phase is computed and holdup events area unit mechanically detected. Spatially and temporally connected events area unit concatenated in, so-called, holdup propagation graphs. These graphs kind a high-level description of a holdup and its propagation in time and area. Our system provides multiple views for visually exploring and analysing the traffic condition of an outsized town as an entire, on the amount of propagation graphs, and on road phase level. Case studies with twenty-four days *of taxi* GPS *trajectories* collected *in* Beijing demonstrate *the* effectiveness *of* our system. We use twenty-four days of *taxi GPS* trajectories *in Beijing* and *a corresponding* street network from Open Street Map. In an exceedingly information driven approach, we have a *tendency to clean* the *GPS trajectories from detector* errors and *fix* apparent errors within the *road* network. With the *cleansed* information, *we* will accurately map the *driving* trajectories *to* the *road* network and *later* on, cipher road speeds. When estimating free flow speed on *every* road phase, we have *a tendency to mechanically* observe holdup events at roads

supported relative low road-speed detection [6]. The concatenation of those events in propagation graphs shows however, a holdup propagates each in area to adjacent roads and in time. supported the *automated* computing results, we have a *tendency to then* build a *visible* interface for interactive exploration of the detected holdup info each intimately on a road phase similarly as on a better level in an exceedingly spacial read on a map and in an exceedingly little multiples read with propagation graphs. We have a *tendency to support* the *analysis* by economical filtering of *area,* time, size and *topology,* and *providing* structured visualizations *of the graphs* through sorting by size and *similarity.* Finally, we offer variety of *case studies* that demonstrate the *effectiveness of our* system. Our system will give users with insights from multiple levels and views.

G. Traffic Congestion Identification

Effective estimation for holdup is that the key step for holdup warning and control. This paper proposes associate in nursing opposition-based *reinforcement learning* theme *for* holdup identification supported fuzzy c-means bunch algorithmic program. *The actions* and opposite *actions* square measure distributed within *the* framework *of reinforcement learning to* update the Q-value for rushing up the training method particularly in exploration mode[9]. The new distinct congestion parameters square measure extracted from the shots and so distinct actions area. Moreover, the cluster centres of the arduous c-means bunch square measure used because the initial price of the fuzzy c-means algorithmic program to accelerate the convergence speed. With the numerical experiments on traffic police work video exploitation the projected technique, holdup detection is created in an efficient manner .In this paper, Associate in Nursing OBRL theme supported FCM bunch algorithmic program for the matter of holdup identification has been proposed [9]. *The states* and *actions* square measure distinct *to* use *in* RL. *The* initial cluster centres of FCM algorithmic program square measure determined by the arduous cluster centres generated by HCM bunch algorithmic program to accelerate the convergence speed. Bias data is employed to pick out action policy in RL and RL is employed to guide bias learning method. This chapter is concerning existing system, main supply and basic theories of the project. Chapter 3 deals with the systematic style and conjointly consists details concerning the projected system. In conveyance networks, it is expected that there will be restricted access to Associate in nursing infrastructure network that may be supported by margin base stations. Such access is restricted in its nature for 2 reasons. First, the readying of the *infrastructure* is *predicted to be* slow and *progressive* resulting in *wide* areas wherever there is no access to the infrastructure. Second, a whole readying is predicted to be thin thanks to price. The coverage offer by a margin base station could also be about 200-300m whereas margin base stations could also be placed each kilometre approximately. Consequently, not all vehicles are going to be connected to the infrastructure in any respect times. to get access to safety or different forms of data, it becomes necessary to believe vehicle to vehicle communications.

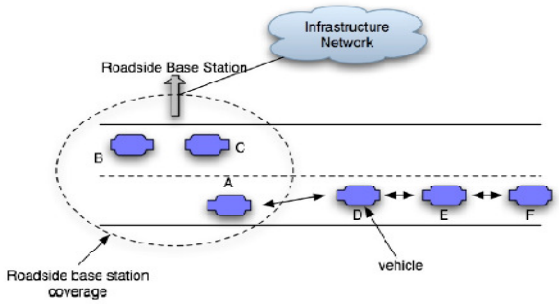


Figure 1: Vehicular network architecture

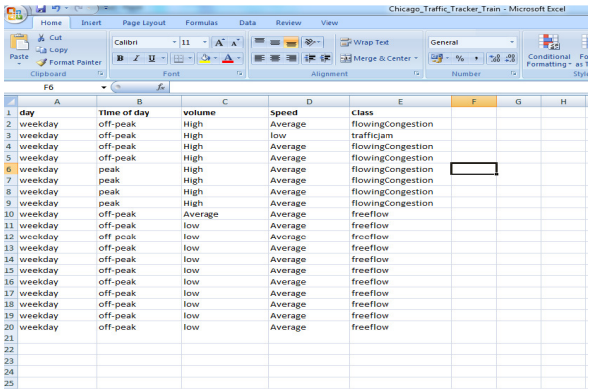
As shown in Figure1, vehicles A, B, and C have access to a roadside infrastructure, which has limited coverage. These vehicles can obtain information from the roadside base station. However, vehicles D, E, and F have no communications with the fixed infrastructure. For instance, Vehicle F will have to rely upon information from vehicle E, which in turn has obtained information that has passed through vehicles A and D. Note that this scenario immediately creates issues that are not necessarily important in other kinds of networks in terms of how to disseminate information and how to assure the security of information. Note also, that vehicles

that are in the range of a roadside infrastructure may be connected to the infrastructure for extremely small durations of time because of small coverage and high vehicular speeds. Therefore, the amount of information that can be pulled from the infrastructure is necessarily limited. It is also possible that vehicles move into the range of the roadside infrastructure with some information obtained from cooperating vehicles they have encountered. The issue then becomes one of updating the information, enhancing the reliability or relevance of information or obtaining information that complements that already available to the vehicle.

III.PROBLEM SOLUTION

1. Pre Processing:

The decreased a group of attributes by concentrating solely on the vehicle rate and the moving pattern of a vehicle, which might infer very different levels of congestion. Then, we tend to applied three steps to arrange the data: 1) smoothening out instant rate, 2) extracting moving pattern of a vehicle victimisation slippery windows technique, and 3) equalisation the distribution of sampling information on every congestion level. Numeric information is real or number numbers, whereas nominal information area unit listing variables that have sure price. Begin Interval, finish Interval, *Detector,* and Link attribute in information *input* area is used for unit nominal information varieties. *Detector and* Link have the correct information kind because of each of them area unit attributes with unchanged price. Starting Interval and ending Interval area unit attributes that represent quantity. These 2 attributes ought to have numeric attributes. Therefore, we would like to alter the illustration of the info to create it numeric. Weka has several filters that area unit useful in pre processing the info. The propose technique to implement new filter algorithmic rule referred to as resample filter supported supervised instance based mostly pre processing. Produces *a* random subsample of *a dataset. The dataset should* have *a nominal category* attribute. *The filter* is created to *keep* up the *category* distribution *within* the *subsample, or to bias the category* distribution toward an identical distribution.



	A	B	C	D	E	F	G	H
1	day	Time of day	volume	Speed	Class			
2	weekday	off-peak	High	Average	flowingCongestion			
3	weekday	off-peak	High	low	trafficjam			
4	weekday	off-peak	High	Average	flowingCongestion			
5	weekday	off-peak	High	Average	flowingCongestion			
6	weekday	peak	High	Average	flowingCongestion			
7	weekday	peak	High	Average	flowingCongestion			
8	weekday	peak	High	Average	flowingCongestion			
9	weekday	peak	High	Average	flowingCongestion			
10	weekday	off-peak	Average	Average	freeflow			
11	weekday	off-peak	low	Average	freeflow			
12	weekday	off-peak	low	Average	freeflow			
13	weekday	off-peak	low	Average	freeflow			
14	weekday	off-peak	low	Average	freeflow			
15	weekday	off-peak	low	Average	freeflow			
16	weekday	off-peak	low	Average	freeflow			
17	weekday	off-peak	low	Average	freeflow			
18	weekday	off-peak	low	Average	freeflow			
19	weekday	off-peak	low	Average	freeflow			
20	weekday	off-peak	low	Average	freeflow			
21								
22								
23								
24								
25								

Figure 2: Pre-processed Datasets

In figure 2, shows the pre processed datasets where all missing data and noisy values are re-sampled using re-sample filter.

2. Classification By Decision Tree

A decision *tree* may be *a classifier that uses a* tree-like graph within *which the* learned functions are diagrammatical by the choice tree. A decision tree may be a supervised prediction technique *because of the* dependent *attribute* and *the* tally *of* categories *(values)* are given. Decision *tree is* simple *for* humans to know. Decision *tree learning* has been applied in issues like predicting medical patient diseases, loan candidates by their probability of defaulting on payments and predicting traffic flow volumes .Decision *tree predict* instances *by sorting them down the tree from the foundation* to *some leaf node that provides* the *prediction* of *the instance.* Every *node within* the *tree specifies* a *check* of *some attribute* of *instance,* and every *branch dropping* from *that node corresponds to at least* one *of the potential* values for this *attribute. Every* leaf *node within* the *tree specifies* associate degree analysis *of associate* degree attribute of *associate* degree instance and *every* branch dropping from *the node represents the values* of *this attribute.* Decision tree *represent a disjunction of conjunctions of constraints on the attribute* values of *instances and* every path *from the tree root node* to *a leaf node* corresponds *to a conjunction of attribute tests are undefined.*

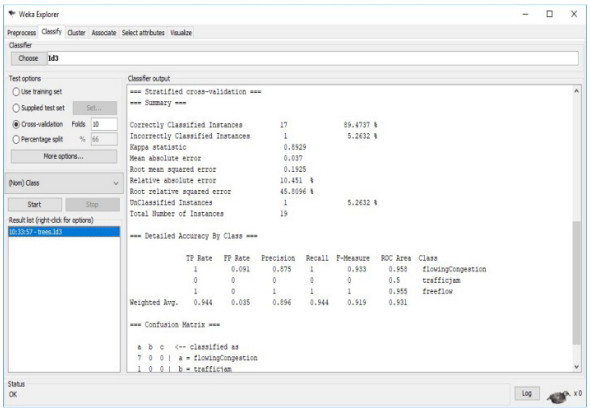


Figure 3: Classified Data

In figure 3, shows the classification by decision tree based on three traffic congestion levels .

3. Prediction Analysis

The information gain will *be* accustomed live *that attribute is that the best predictor. This* may be done by first process *the measure* employed in info gain referred *to* as *the Entropy the decision tree is* much quicker, once trained, compared *to the* ANN. *The* decision *tree* typically throws away input options that it finds not helpful, whereas the Neural Network can use it unless feature choice is finished. It means ANN offers a rather higher performance than call trees. Similarly, BN does not throw away any attributes, unless this can be done throughout pre processing; Thus, BN can beat call trees.

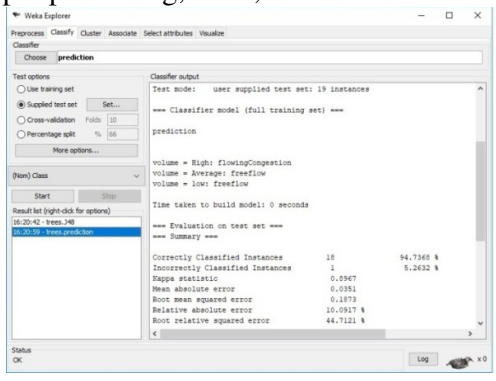


Figure 4: Prediction of traffic congestion

In figure 4, represent overall prediction of traffic Congestion level.

IV CONCLUSION

The information gain will *be* accustomed live *that attribute is that the best predictor. This* may be done by first process *the measure* employed in info gain referred *to* as *the Entropy the decision tree is* way quicker, once trained, compared *to the* ANN. *The* decision *tree* typically throws away input options that it finds not helpful, whereas the Neural Network can use it unless feature choice is finished. It means ANN offers a rather higher performance than call trees. Similarly, BN doesn't throw away any attributes, unless this can be done throughout pre processing; thus, BN can beat call trees

This document is a template. An electronic copy can be downloaded from the conference website. For questions on paper guidelines, please contact the conference publications committee as indicated on the conference website. Information about final paper submission is available from the conference website.

REFERENCES

1. *E. Chung, "Classification of traffic pattern," In Proc. of the 11th World Congress on ITS, pp. 687-694, 2003.*

2. *W. Weijermars and Eric Van Berkum, "Analyzing highway flow patterns using cluster analysis," In Proceedings of the 8th International IEEE Conference on Intelligent Transportation Systems, pp. 308-313, 2005.*

3. *Lin Xu, Yang Yue and Qingquan Li, "Identifying Urban Traffic Congestion Pattern from Historical Floating Car Data," In Proceedings of Social and Behavioral Sciences, Vol. 96, pp. 2084– 2095, 2013.*

4. *A. Amelia and P. Saptawati, "Detection of potential traffic jam based on traffic characteristic data analysis," In International Conference on Data and Software Engineering (ICODSE), pp.1-5, 2014.*

5. *R. Ong, F. Pinelli, R. Trasarti, M. Nanni, C. Renso, S. Rinzivillo and F. Giannotti, "Traffic Jams Detection Using Flock Mining," In European Conference (ECML PKDD-11), pp. 650-653, 2011.*

6. *Z. Wang, Min Lu and X. Yuan, J. Zhang and Van de Wetering H, "Visual Traffic Jam Analysis Based on Trajectory Data," In IEEE Transactions on Visualization and Computer Graphics, Vol. 19, No. 12, pp. 2159 – 2168, 2013.*

7. *A. Gupta, Netaji Subhas, S. Choudhary and S. Paul, "DTC: A framework to Detect Traffic Congestion by mining versatile GPS data," In 1st International*

Conference Proceedings of Emerging Trends and Applications in Computer Science (ICETACS), pp. 97 – 103, 2013.

8. *E Florido, O Castaño, A Troncoso and F Martínez-Álvarez, "Data Mining for Predicting Traffic Congestion and Its Application to Spanish Data," In 10th International Conference on Soft Computing Models in Industrial and Environmental Applications Volume 368 of the series Advances in Intelligent Systems and Computing, pp. 341-351. 2015 .*

9. *Y. Yang, Zhiming Cui, Jian Wu a Guangming Zhang and X. Xian, "Fuzzy C-means Clustering and Opposition-based Reinforcement Learning for Traffic Congestion Identification," In Journal of Information & Computational Science, Vol. 9, No. 9, pp. 2441-2450, 2012.*