ISSN No:-2456-2165

Clustering of India States using Optimized K Means Algorithm

Nisha K
Computer Science and Engineering
SDMCET
Dharwad , India

Basavaraj B Vaddatti
Computer Science and Engineering
SDMCET
Dharwad , India

Abstract:- In today's world we have a positive trend of better infrastructure with huge amount of roads expansion and variety of vehicles moving around the globe. As the number of vehicles is increasing, more number of commuters involved there is a probability of increased number of accidents which can happen. In this paper Canopy K Means algorithm is implemented on the data provided by government of India and then states are classified into Low, Medium and High Accident Zones. Unlike normal k means algorithm the proposed method select the optimized centers so that better classification results are obtained. During the clustering process instead of working on all the attributes of accident data sets top level attributes are found out based on standard deviation.

Keywords:- *K Means, Eigen Value, Clustering.*

I. INTRODUCTION

The use of scientific technology helps in the field of traffic, transportation and motor vehicle. It also helps in improving the life style along with economy of the country. This has also led to increase number of accidents. As per statistics on a daily trend there are around 3000 people who will get killed, around 15,000 people get disabled and 140,000 people get injured. Due to these accidents there is economy loss of around 1.88 billion [1]. This makes it important to have strategy to improve safety. When the capital is limited then the number of traffic accidents can be reduced. If the black spots are identified and managed then road safety can be improvised. For countries which fall under developed state for example United States the identification of hot spots was done at an early stage and good results are obtained in which each result will contain accident number[2,3,4], accident rate[5,6,7], quality control [8] and other metrics. The theoretical system was formulated by Liu [9] with the help of normal distribution model applied on black points. Statistical methods which can compute the frequency of accidents and then apply gamma distribution model is provided by Pei and Ding [10].

An optimized way of identifying the accident hotspots is provided by chen in [11]. Whenever there is a conflict in the hot spot identification then a method has been proposed by Luo and Zhou[12] to resolve such hot spots classification.

All the traditional methods of identifying the hotspots makes use of statistics and threshold value but will fail to find the relation between close accidents. In this paper k means clustering algorithm along with optimized center identification is performed. The data sets are taken for various attributes state wise and then preprocessing is done to remove rows which have all attributes as zeros and then missing attributes are replaced with mean value. After that standard deviation value computation is done to find the top attributes. Based on the values of top attributes under each state is taken then the states are clustered into medium, low and high accidents based on k means. The entire workflow for clustering is provided in Fig1.

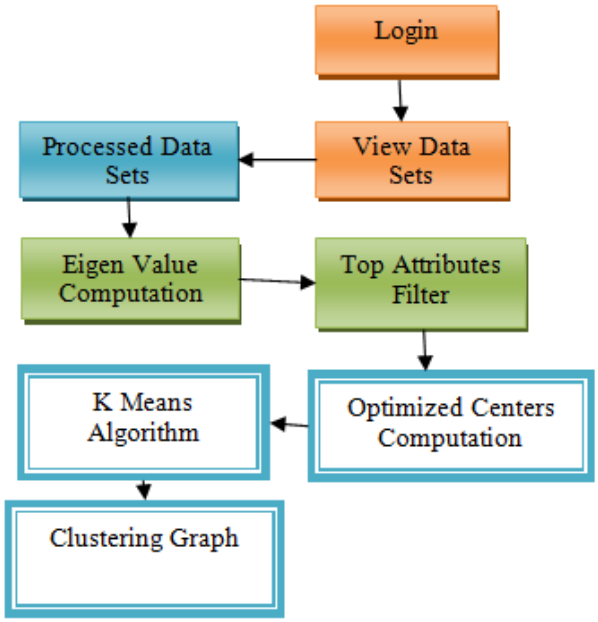

Fig 1:- Workflow for Clustering

Fig1 shows the workflow for clustering of states into Low, Medium and High Accidents. As shown in the fig first the user will register into the application. After that the user will be able to view the data sets which will contain missing values for the attributes. The data sets will act as an input for the preprocessing which will fill the missing attributes with mean value. After that the Eigen value computation is done. The top attributes are found out by sorting the attributes based on descending order of Eigen values. The optimized centers are computed for the three clusters and then for each row the distance is computed. The row will be assigned a cluster based on the lowest distance. The clustering graph indicates how many states

will be of Low Accident, Medium Accident and High Accidents.

II. BACKGROUND

The amount of increase in motor vehicles and also the increase in the road network is becoming one of the major causes of road accidents [1]. This will act as a serious problem for public health and safety across all the nations.

India has a high number of accidents as per the records of the mortality rate database (WHO, 2002). The work identifies the accident black spots on Sarkhej-Gandhinagar Highway (NH 147). Quantum GIS(QGIS) is used to perform the mapping and to perform geospatial analysis.

After that mobility data is analyzed and then visually represented.

The road authorities perform a road safety audit [2]. The audit process has a series of stages namely feasibility, design layout, design details and in-service. In order to increase the safety of road it is important that a formal process is executed to ensure effectiveness and the team performing such a task must be experienced and also should have good amount of training in the safety audit.

Interactive Highway Safety Design Model (IHSDM) [3] contains a package of software analysis tools which can measure the safety and operational values for various geometric design patterns for highways. The tool is updated with relevant design policy data and then decision support tool is used each time to compare high way designs with the relevant design in order to check whether safety and operational performance is satisfactory. The tools can be used by various spans of users like highway project managers, designers, reviewers who perform the review on traffic and safety.

There is dependence between number of vehicle kilometers and number of fatalities along with systematic components [4]. When the execution of the algorithm was done on Netherlands then the dependency had a shift in the time period. The safety actions taken by the society makes its very effective to have the safety factor under control. Whenever there is increase in the traffic volumes then immediately safety effects must also be estimated.

The regression to mean effect is measured before and after the accidents under various assumptions namely computation of mean value of accident rates between sites which includes the amount of populations, frequency of accidents [5]. The work computes the validity of assumptions and generates recommendations in order to improve the quality of results.

The prediction of road safety values [6] from accident prediction models has issues related to statistics which required lot of attention. The modeling of accidents can be performed with the help of Poisson and negative binomial regression and two statistical issues are dealt. The first issue is to check whether explanatory variables should be included in accident prediction model. For each application the model will be different in which few of them avoid over fitting. The second includes a way to fit the model to the data. The second issue is outlier analysis in which a procedure is executed which find the influential outliers. The negative binomial regression model is developed in order to to have a detailed application procedure and then binomial model is applied to urban cities.

The road accident analysis provides major categories of accidents like head-on collisions, accidents during left turn, accidents on truck and accidents of single vehicle [7]. When the analysis of police reports, accident site investigations, vehicle and users involved is done then major factors for vehicle accidents were increased speed, drunk driving, driving with illegal drugs injected in body. In a similar fashion when the major factors of left turn accidents where found out as attention while taking left turn and amount of time that is required to complete the left turn. The major factors for truck accidents are speeding and incomplete search of visual information.

The traffic accidents have become a major issue around the world [8]. The major reasons for traffic accidents include driving behavior with are directly linked to infrastructure and traffic conditions. Machine learning techniques can be used to find the risk of road accidents on a zone. The machine can be trained by using historical data and then prediction function can be called to find the probability class label for the new value. The disadvantage of this approach is that there is a limited coverage of traffic data. Dynamic Traffic Assignment (DTA) make use of temporal features of accident data sets and then finds the Black spots on the network

III. PROPOSED METHOD

In the proposed method first the data sets are taken from government websites. The details of the proposed method under various phases is described as below

A. Data Formation

In this phase the data is collected from the authorized sources in the format of CSV. After that a database was created along with a table to hold the datasets. The datasets are converted from csv into a table using Toad for My SQL software. The various attributes which are taken into consideration are injured bazaar, injured_ bridge,injured_bus_stop,injured_enchroch,injured_hospital, injured_inside_village,injured_near_cinema_hall,injured_n ear_factory,injured_near_school_college, injured_ office_ complex, injured_open_area,injured_ pedes_ cross,injured_petrol_pump,injured_ religious_ place, injured_ residential_area,kiled_bus_stop, killed_bazaar, killed_bridge, killed_enchroch, killed_hospital,killed

_inside_village,killed_near_cinema_hall,killed_near_factory,killed_near_school_college,killed_office_complex,killed_open_area,killed_pedes_cross,killed_petrol_pump, killed_religious_place, killed_residential_ area, near_ petrol_pump,state, total_bazaar,total_bridge,total_ bus_ stop, total_enchroch, total_hospital, total_near_cinema_ hall, total_near_factory,total_near_or_inside_village, total_office_complex,total_open_area,total_pedes_cross,total_religious_place,total_residential_area,total_school_college_near_accidents.

B. Data Preprocessing

This module is responsible for finding the rows from the data sets which has all the attributes as zeros and then removes all the rows which has all attributes as zeros. The attributes which are having zeros in a state row then the value will be replaced with the mean value of other states. Consider that a specific state has value of zero for people killed near factory then the value zero will be replaced with the mean value of killed near factor of other states.

C. Top Attributes Computation

If L is the total number of attributes then for each of the attributes the following steps are done.

- All the unique attributes are retrieved. Consider that it is represent as a set {A1,A2,……An} where Ai is the ith attribute
- For each of the attribute set of values across the rows are obtained. For instance for attribute Ai the values across the rows will be {vi1,vi2,……vil}. Where vij is the value of ith attribute across the jth row.
- The total summation is obtained for the set of values of the attribute j is obtained using equation (1)

$$TS(j) = \sum_{i=1}^{N_l} vi, j \qquad (1)$$

Where Nl = is the number of attribute values

- The average value for the attribute Aj is obtained using equation (2)

$$Avg(Aj) = TS(Aj) / \sum_{i=1}^{Na} TS(Ai) \qquad (2)$$

Where, TS(Aj) is the total summation for attribute Aj and TS(Ai) is the total summation for attribute Ai and Na is the number of unique attributes under consideration

- Find the Square root of each Avg(Aj) . Like this for the L attributes {Sqr(A1),Sqr(A2),…….,Sqr(Al)}
- The square root values are sorted in the descending and top L/2 attributes are taken out.

D. Optimized Center Computation

The Optimized Center Computation is done using the following steps

- Find the maximum value for each of the top attribute across states is found out i.e max
- The cluster1 center is defined as max/2
- The cluster2 center is defined as (max/2-max/3)
- The cluster3 center is defined as (max/2-max/5)
- The steps from a to d are repeated for all the top attributes.

E. K Means State Level Clustering

The K Means state level clustering will take each row of the data set and then compute the distance by taking n attributes and then with respect to each of cluster center the distance is computed. Consider that those distances with respect to center are dc1,dc2 and dc3. Where, dc1 is the distance of data point with respect to cluster center 1, dc2 is the distance of data point with respect to cluster center 2 and dc3 is the distance of data point with respect to cluster center 3. After that the minimum distance is found out among {dc1, dc2, dc3} and then the data point is labeled as LOW, MEDIUM or HIGH ACCIDENT ZONE.

IV. RESULTS

The entire implementation was done with the help of Spring Framework for the back end along with Ext JS and Angular for the front end. This section describes the views which are related to data sets processing, finding the top attributes, and optimized center computation and k means clustering

A. View Data Sets

The user will be able to login into the application and then view all the attributes in the format of Ext JS grid for the accident related data sets. These are raw data sets of accidents data across the 31 states of india. Only the partial attribute columns are shown as there are 24 attributes.The user can perform the check or uncheck of those remaining attributes in order to see them by using the flexibility of Ext JS Grid.

Data Set Accidents Output

State	Injured Near School or College	Total Accidents Village
Punjab	240	1109
Rajasthan	884	3817
Sikkim	0	0
Tamil 0du	4500	5808
Telanga0	897	3398
Tripura	56	180
Uttarakhand	81	204
Uttar Pradesh	1524	3709
West Bengal	557	890
A & N Islands	9	61
Chandigarh	6	27
D & N Haveli	10	25
Daman & Diu	4	6
Delhi	0	0
Lakshadweep	0	0
Puducherry	76	116

Fig 2:- Partial Attributes

The user after login can see the raw data sets. As shown in the Fig 2 there are 3 columns namely state, number of people injured near school or college is the 2nd column and then the 3rd column is the total accidents in the village. Like this all the attributes are shown in the format of the grid.

B. Execute Processing

The admin clicks on perform processing so that unwanted rows are removed from the data sets as well as the missing attribute values are replaced with appropriate values for all the states with the mean value. The preprocessed data sets is represented in the format of Ext JS grid as shown in the Fig 3.

State	Injured Near School or Colleg	Total Accidents Village
Tamil 0du	4500	5808
Telanga0	897	3398
Tripura	56	180
Meghalaya	23	17
Mizoram	4	52
0galand	677.37	52
Orissa	607	1122
Himachal Pradesh	169	318
Gujarat	974	3219
Goa	85	517
Aru0chal Pradesh	48	1866.29
Assam	348	792
Bihar	517	2050
Chhattisgarh	579	3444

Fig 3:- Partial Attributes after Processing

Fig 3 shows the attribute values after processing. If any attribute value is zero then it is replaced by mean of other values of the same attribute.

C. Top Attribute Computation

The standard deviation is computed for all the attribute values. As shown in the fig for all the 24 attributes the standard deviation value is displayed along with the attribute name.

Attribute Weight Computation

Name	Value
total_school_college_near_accidents	0.220891926954313
total_residential_area	0.313608312433553
total_religious_place	0.183619800379356
total_pedes_cross	0.212895494016307
total_office_complex	0.198026376720325
total_open_area	0.562108117622524
total_near_or_inside_village	0.373498118038709
total_near_factory	0.223504087859466
total_near_cinema_hall	0.165846138389752
total_hospital	0.182370961897526
total_enchroch	0.110405727693839
total_bus_stop	0.228446919560948
total_bazaar	0.267884697070485
total_bridge	0.179658612989917

Fig 4:- Weight Computation of Attributes

Fig 4 shows the weight computation of attributes. As shown in the fig 4 there are two columns.The 1st column is the name of attribute and then 2nd column is the value of weight for the respective attribute.

D. Top L/2 Attributes

The fig shows the top attributes among the set of L attributes. In the implementation 24 attributes are taken from the data sets and the top 12 attributes are found out so those can be used for clustering technique.

Attribute Weight Computation

Name	Value
total_open_area	0.562108117622524
total_near_or_inside_village	0.373498118038709
total_residential_area	0.313608312433553
total_bazaar	0.267884697070485
total_bus_stop	0.228446919560948
total_near_factory	0.223504087859466
total_school_college_near_accidents	0.220891926954313

Fig 5:- Top Weight Computation of Attributes

Fig 5 shows the top weight after sorting the weight computed for all the attributes. As shown in the fig5 the top seven attributes are taken out among fourteen top attributes.

E. Optimized Center for Top Attributes

K Means X Center Computation

Name	Cluster1 Center	Cluster2 Center	Cluster3 Center
injured_inside_village	7260	9680	11616
injured_near_factory	2245	2993.33333333333	3592
injured_near_school_college	2250	3000	3600
injured_bus_stop	2764.5	3686	4423.2
injured_bazaar	2972.5	3963.33333333333	4756
injured_open_area	11459.5	15279.3333333333	18335.2
injured_residential_area	5374	7165.33333333333	8598.4

Fig 6:- Optimized X center for Top Attributes

Fig 6 shows the cluster center values for x axis for the top seven attributes. The first column is the name of top attributes. Second column is Cluster1 Center is the center of first cluster. Third column Cluster 2 Center is the center of second cluster and then fourth column is Cluster 3 Center is the center of third cluster. The x attributes are the attributes representing the injured accidents number.

K Means Y Center Computation

Name	Cluster1 Center	Cluster2 Center	Cluster3 Center
killed_residential_area	891.5	1188.66666666667	1426.4
killed_inside_village	1342.5	1790	2148
killed_near_school_college	502	669.333333333333	803.2
killed_open_area	2720	3626.66666666667	4352
killed_bazaar	1017	1356	1627.2
kiled_bus_stop	670	893.333333333333	1072
killed_near_factory	668	890.666666666667	1068.8

Fig 7:- Oprimized Y Center for Top Attributes

ISSN No:-2456-2165

Fig 7 shows the cluster center values for y axis for the top seven attributes. The first column is the name of top attributes. Second column is Cluster1 Center is the center of first cluster. Third column Cluster 2 Center is the center of second cluster and then fourth column is Cluster 3 Center is the center of third cluster. The y attributes are the attributes representing the killed accidents number.

F. K Means Output

This section describes the clustering of states into MEDIUM, HIGH and LOW ACCIDENTS by computing the distance between the center of three clusters and then computing the minimum distance and then Cluster Name is the name of the cluster.

K Means Extended Computation

Cluster Name	Distance C1	Distance C2
MEDIUMACCIDENT	13979.9909960629	13725.0294474487
MEDIUMACCIDENT	8401.84677615582	5956.41195127849
HIGHACCIDENT	14896.2888162018	20155.9759157352
HIGHACCIDENT	13360.047950887	18640.8546209663
HIGHACCIDENT	13531.6043856595	18796.4863560542
HIGHACCIDENT	10954.1205603188	16202.6691936853
HIGHACCIDENT	15191.5278444928	20479.2025235359
HIGHACCIDENT	7417.14131252735	12564.1404003616
HIGHACCIDENT	12781.2558166246	18060.6137861738
HIGHACCIDENT	5586.76903496108	10776.6879265694

Distance C3	Min Distance	State
14932.9133768331	13725.0294474487	Kerala
6903.29885779256	5956.41195127849	Madhya Pradesh
24372.5452370429	14896.2888162018	Aru0chal Pradesh
22867.8893936454	13360.047950887	Assam
23016.0206273804	13531.6043856595	Bihar
20416.167978345	10954.1205603188	Chhattisgarh
24709.438494632	15191.5278444928	Goa
16745.117960767	7417.14131252735	Gujarat
22287.0461685707	12781.2558166246	Harya0
14978.0991691202	5586.76903496108	Andhra Pradesh

Fig 8:- K Means Computation

Fig 8 shows the K Means Computation. As shown in the Fig 8 the first column is the clustering label for the state which can be LOWACCIDENT, MEDIUMACCIDENT and HIGHACCIDENT. The second column is the distance with respect to cluster 1, the third column is the distance with respect to cluster 2, the fourth column is the distance with respect to cluster 3,fifth column is the minimum distance is the lowest distance with respect to the three clusters and sixth column is the state.

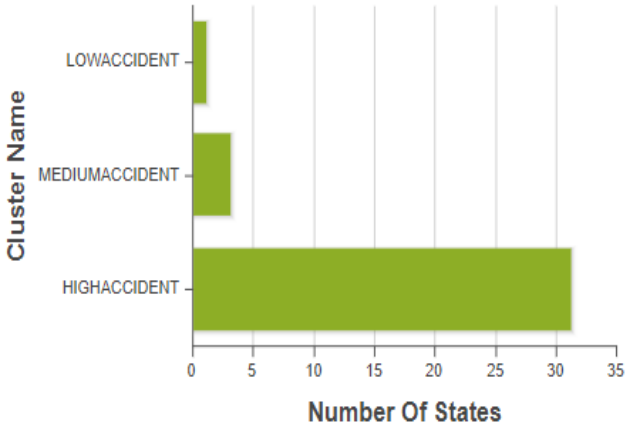

Fig 9:- K Means Graph

Fig 9 shows the k means graph. As shown in the fig there are 31 states which are having HIGHACCIDENT, 3 states are clustered as MEDIUMACCIDENT and then 1 state is classified as LOWACCIDENT.

V. CONCLUSION

In this paper first an overview of the ecommerce system is presented. The paper presents various algorithms in which the transaction logs are maintained for the user purchases. The K Means classification algorithm is applied to classify books into Low Selling, Medium Selling and High Selling Stock. RFM algorithm is applied in order to classify the customers into HIGH LOYAL, LOW LOYAL and MEDIUM LOYAL customers. The weighted page rank algorithm computes purity, connectivity, user id page frequency, recursive weight and page weight is computed for each user. For a HIGH LOYAL customer MEDIUM selling Stock is recommended on the best page. For a MEDIUM LOYAL customer LOW Selling Stock is recommended on the best page.

REFERENCES

[1]. Wenjie Chen, "Analysis of Road Traffic Accident Black Spots [J]", Journal of Chinese People's Public Security University, vol. 2, no. 2, pp. 83-88, 2004.
[2]. "Road safety Audit: A New Tool for Accident Prevention", ITEJournal, vol. 66, no. 2, pp. 1-6, 1995.
[3]. RA. Krammes, "Interactive Highway Safety Design Model: Design Consistency Module", Public Roads, vol. 77, no. 5, pp. 12-17, 1997.
[4]. S. Oppe, "Development of Traffic Safety Global Trends and Incidental Fluctuations", Accident Analysis and Prevention, vol. 23, no. 1, pp. 58-60, 1991.
[5]. C.C. Wright, C.R. Abbess, D.F. Jarrett, "Estimating the Regression-to-Mean Effect Associated with Road Accident Black Spot Treatment: Towards a more Realistic Approach",
[6]. Ziad A. Sawalha, Traffic Accident Modeling Statistical Issues and Safety Applications, pp. 119-129, 2002.

[7]. Lotte Larsen, "Methods of Multidisciplinary in-depth Analyses of Road to Add Traffic Accidents", Journal of Hazardous Materials, vol. 111, no. 3, pp. 115-122, 2004.
[8]. Yuzeng Liu, Research on Intelligent Investigation and Countermeasures of Traffic Accident Black Spots [D], 2005.
[9]. Yuzeng Liu Research on Intelligent Investigation and Countermeasures of Traffic Accident Black Spots [D], 2005.
[10]. Yulong Pei, Jianmei Ding. Improvement in the Quality Control Method to Distinguish the Black Spots of the Road. The 6th Conference of the Eastern Asia Society for Transportation Studies. 2006:2106-2113.
[11]. Shuang Chen. Comparative Study on Identification of Road Accident Black Spots [J]. Shandong traffic science and technology. 2005, (3):7-9.
[12]. Shigui Luo, Wei Zhou. Survey Way of Road Traffic Conflict Technique [J]. Journal of Changan University (Natural Science Edition).2001, 18(1):65-68.