Road Accident Perusal Using Machine Learning Algorithms

K.M. Umamaheswari, Ishit Singh and Ankit Aman

Abstract--- *Road Accidents have been a major concerning issue in India. Approximately 5 Lakh citizens are reported to be the victims of road accidents in a year. The consequences of these accidents are not only monetary losses, but non-monetary losses as well. The families of the victim are affected as well. Accidents might also lead to blockage of roads, which would result in difficulties for a citizen to follow his common routine, such as travelling to work. The high volume of road traffic in India due to the enormous population is also a reason why a single road accident might lead to a big ruckus. This concern is quite prevalent in India and hence, needs to be redressed immediately. To counter this problematic situation, road accident data will be analysed and mined and algorithms like K-means++ method and Apriori algorithm will be used to analyze the dataset provided by the Indian Government. We intend to determine the factors causing these accidents corresponding to every region in India, as well as the severity of each factor. In this process, we are going to determine that K-Means++ provides better results as compared to the standard K-Means. The obtained results can be used to plan preventive measures, ensuring the reduction in road accidents. Using this, causes of accidents can be ranked considering various parameters in order to estimate plans to reduce road accidents.*

Keywords--- *Road Accident Data Analysis, Data Mining, K-Means, Apriori Algorithm.*

I. INTRODUCTION

Road accidents are a major disaster, causing the loss of life, property and economy of a country. As our country grows economically, so too do road traffic accidents. Apart from the obvious loss, this increase in road traffic accidents also hinders the economic development of the country. According to the World Health Organization, 1.25 million people die in road accidents every year. This causes about $ 500 billion a year to the global economy. It is estimated that monetary losses in India will increase to about 2 billion crores annually. This includes the money the patient and his family spent in the hospital.

Causes of Road Accidents

The main reasons behind road accidents are sleep deprivation and alcohol consumption while driving. Excessive psychological, social and economic stress on drivers can also kick at high speeds on congested roads leading to road accidents. Over speeding, overtaking and overloading are the reasons for the increase in road traffic accidents. It has been scientifically proven that the number of accidents increases as speed increases.

K.M. Umamaheswari, Assistant Professor, Dept of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India. E-mail: umamahek@srmist.edu.in

Ishit Singh, UG Student, Dept of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India. E-mail: ishitg4@gmail.com

Ankit Aman, UG Student, Dept of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India. E-mail: ankitaman25@gmail.com

DOI: 10.37200/IJPR/V24I5/PR201839
Received: 15 Feb 2020 | Revised: 07 Mar 2020 | Accepted: 23 Mar 2020

Data mining has proven to be a reliable technology for analyzing road hazards and delivering productive results. Most road accident data analytics use data mining techniques, focusing on identifying factors that influence the severity of an accident. Data mining can be described as a novel technique for extracting hidden and previously unknown information from large amounts of data. Many data mining techniques, such as clustering, classification, and association rule mining, are widely used by researchers in other countries in road accident analysis.

II. RELATED WORKS

In [1], we can see that the two method, K-Means and Hybrid K-Means and the performance of each algorithm is studied. The algorithm is applied on Indian dataset, take from Indian government website(data.gov.in). The paper has concluded that K-Means hybrid has better performance in relation to noise in dataset, as compared to the standard K-Means algorithm.

In [2], Data mining techniques have been applied to study the dataset obtained from Fatality Analysis Reporting System (FARS). Data mining techniques like Apriori Algorithm and Naive-Bayes technique have been applied to obtain information on roadway traffic. This paper has realized that natural disaster's dataset is not complete and required more data.

From [3] we find out that the paper has studied and identified the factors for increasing and decreasing accident rates in the state of Gujarat in India using Cophenetic correlation coefficient. The dataset has been obtained from GVK_EMRI, Gujarat. This paper has presented a data mining clustering techniques that can be used to cluster the hourly counts of road accidents in Gujarat.

In [4], paper has understood the role of behavioural factors of the drivers that play a prominent role in road accidents in states of USA using K-Means clustering method. The dataset has been obtained from National Highway Traffic Safety Administration(NHTSA) website having Fatality Analysis Reporting System(FARS). This paper generated results for each factor and realised who had proper license and which factor affected road accidents the most. In [5], specific locations of frequent occurrence of road accidents have been analysed by the paper to identify the features pertaining to it using k-means clustering. The dataset was gathered from Ministry of Road Transport and Highways (MoRTH). This paper exposed the various factors associated with road accidents in these locations.

In [6], the link between recorded accidents' factors and accident severity has been explored in Dubai using Apriori and Predictive Apriori Algorithm. The dataset was obtained from Dubai Traffic Department, UAE. Paper explored more associations between accident factors and severity when applying Apriori and generated more number of rules when applying Predictive Apriori.

In [7], relationship between Susceptible Elements of Improvement(ESMs), number of crashes and hazardous sections has been studied through the dataset of Andalusia Regional Government and Spanish National Government dataset. The paper identified characteristics of roads, defined through ESMs which have greater influence in crashes.

In [8], segmentation is performed on road data using K-Means clustering method. The dataset was gathered from Indian government Website(data.gov.in). The paper used K-Means clustering to cluster the data based on attribute accident type, road type, light condition and road characteristics.

DOI: 10.37200/IJPR/V24I5/PR201839
Received: 15 Feb 2020 | Revised: 07 Mar 2020 | Accepted: 23 Mar 2020

In [9], data mining methods are used to create a model that smooths out heterogeneity of the dataset obtained from Indian government Website(data.gov.in). The paper uses cluster analysis to determine the accident prone states and territories of India.

In [10], factors influencing the accidents have been found and factors which is more accident prone is identified with Info Gain Attribute Evaluator using WEKA Tool applied on Road accident dataset trained with decision logic. Paper evaluated attribute importance based on info gain attribute evaluator approach to get awareness on which factors are accident oriented.

In [11], frequent patterns causing road accidents are mined from collected datasets from Indian government Website (data.gov.in). Paper found associations among road accidents and predicted the type of accidents for existing as well as for new roads.

III. PROPOSED WORK

The proposed system is to use K-Means ++ for clustering. K-means clustering is basically a method from signal processing to vector quantization, which is known for cluster analysis in data mining. The purpose of K-mean clustering is to divide n observations into n clusters in which each observation is averaging close to the cluster, which acts as a model of the cluster. It divides the data space into Voronoi cells. K-means at least in the cluster-version (least squared Euclidean distance), but not the usual Euclidean distance, which becomes a more difficult Weber problem: this means that the squared errors are optimized, whereas the geometric medieval Euclidean only represents the distance. K- means ++ can be used to initialize centroid values. The K-Means ++ algorithm ensures the initial initiation of centroids and improves the clustering quality. The results of Standard K-Means and K-Means ++ are evaluated.

Benefit: The K-Means algorithm is mainly used for data clustering and is very easy to use. The K-Means algorithm typically measures large datasets and is easily adapted to new examples.

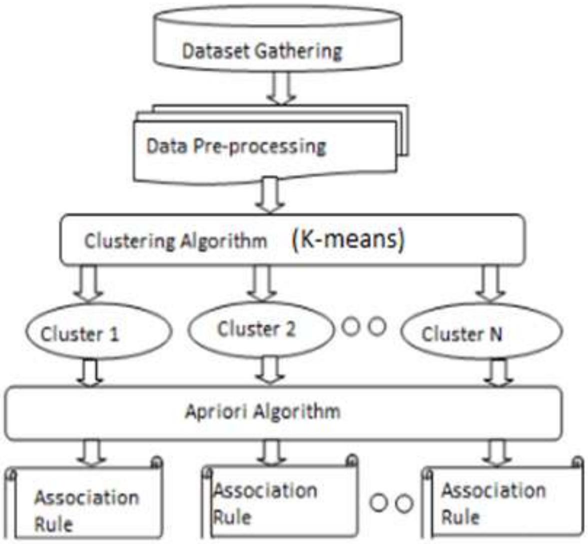


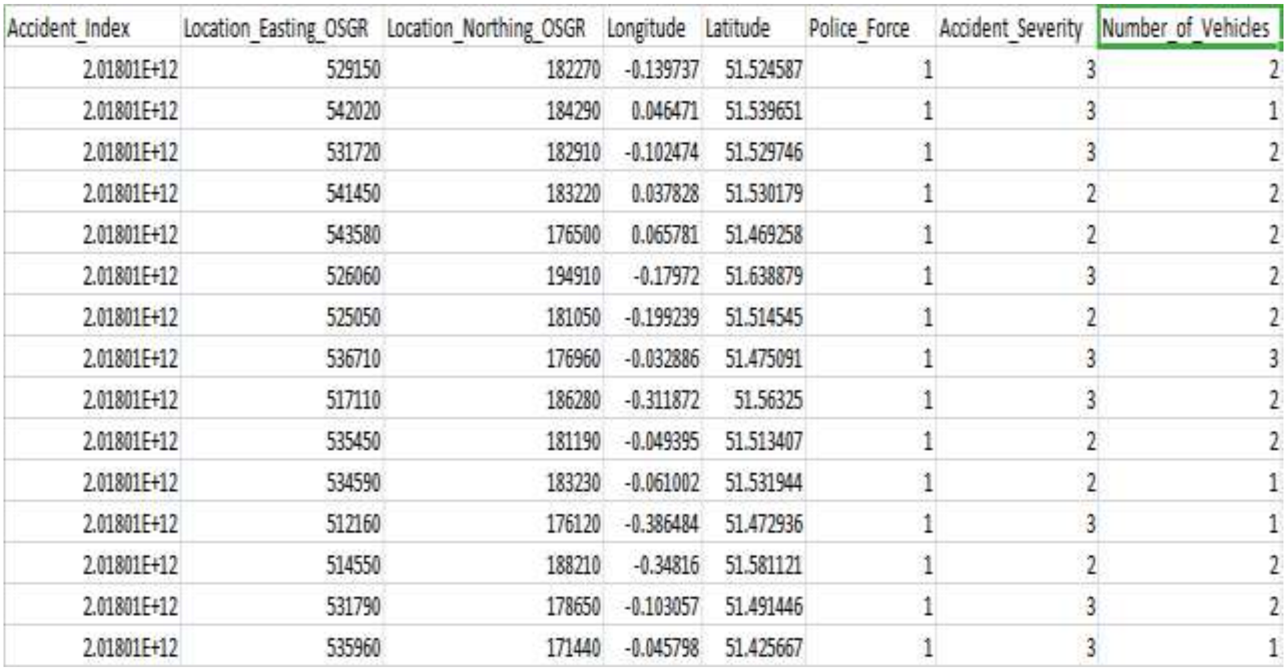
Figure 1: Arhcitectural Diagram

DOI: 10.37200/IJPR/V24I5/PR201839
Received: 15 Feb 2020 | Revised: 07 Mar 2020 | Accepted: 23 Mar 2020

IV. RESULTS ANALYSIS

i. Used Dataset

The dataset is available freely on the website of the Government of UK(data.gov.uk). The dataset provides road safety data of the each year since 1979 as well as the types of vehicles involved and the consequential casualties.



Accident_Index	Location_Easting_OSGR	Location_Northing_OSGR	Longitude	Latitude	Police_Force	Accident_Severity	Number_of_Vehicles
2.01801E+12	529150	182270	-0.139737	51.524587	1	3	2
2.01801E+12	542020	184290	0.046471	51.539651	1	3	1
2.01801E+12	531720	182910	-0.102474	51.529746	1	3	2
2.01801E+12	541450	183220	0.037828	51.530179	1	2	2
2.01801E+12	543580	176500	0.065781	51.469258	1	2	2
2.01801E+12	526060	194910	-0.17972	51.638879	1	3	2
2.01801E+12	525050	181050	-0.199239	51.514545	1	2	2
2.01801E+12	536710	176960	-0.032886	51.475091	1	3	3
2.01801E+12	517110	186280	-0.311872	51.56325	1	3	2
2.01801E+12	535450	181190	-0.049395	51.513407	1	2	2
2.01801E+12	534590	183230	-0.061002	51.531944	1	2	1
2.01801E+12	512160	176120	-0.386484	51.472936	1	3	1
2.01801E+12	514550	188210	-0.34816	51.581121	1	2	2
2.01801E+12	531790	178650	-0.103057	51.491446	1	3	2
2.01801E+12	535960	171440	-0.045798	51.425667	1	3	1

Figure 2: Dataset 1

ii. K-Means Algorithm

K-means algorithm is an iterative algorithm that partitions the dataset into *K* predefined different non-overlapping subgroups known as clusters where each data point belongs to only one cluster. It tries to make the data points inside each cluster as similar as possible while keeping the clusters as different as possible. It allocates data points to a cluster to minimise the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster). The less difference in a cluster, the more similar the data points are in the same cluster.

To determine the number of cluster, elbow method is used.

Elbow Method Procedure

- Determine k – means for various different values of k.
- Find sum of squared error (SSE) for each value of k.
- Plot the curve for SSE vs Number of cluster.
- The elbow point is the number of cluster to be considered.

For our dataset, we get the number of cluster as 3.

DOI: 10.37200/IJPR/V24I5/PR201839
Received: 15 Feb 2020 | Revised: 07 Mar 2020 | Accepted: 23 Mar 2020

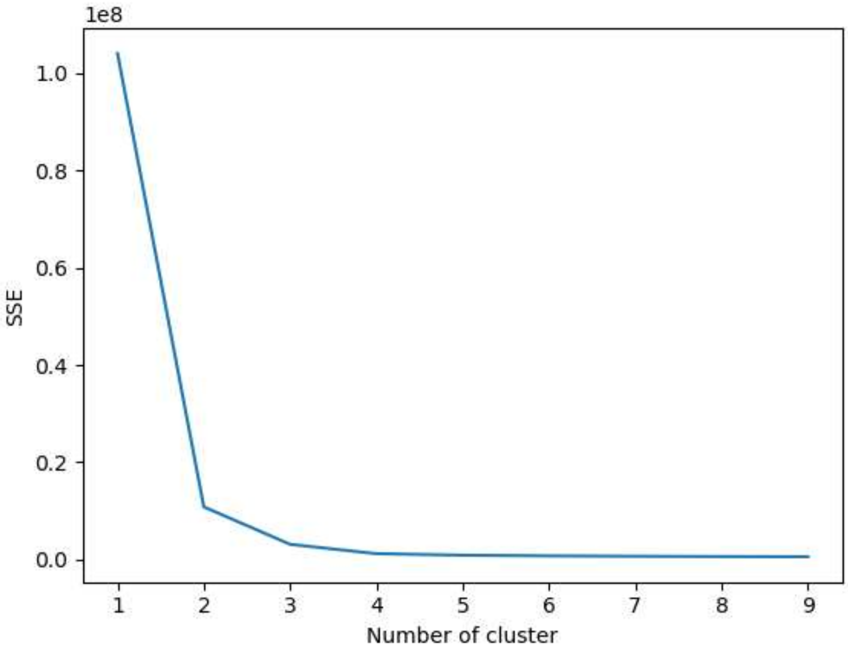


Figure 3: Elbow Method Procedure

The way k-means algorithm works to cluster the dataset into k - clusters is as follows:

- The number of cluster is determined by elbow method.
- The Euclidean distance between each object and cluster centroid is measured.
- According to this distance measure each object gets allocated to that cluster which shows homogeneity.
- Calculate new centroid for that cluster after the object gets allocated to it.

iii. K-Means++

K-Means++ algorithm ensures a smarter initialization of the centroids and improves the clustering quality. It is the same algorithm as the standard K-Means, apart from the initialization. This makes K-Means++ same as K-Means coupled with smarter initialization of centroids.

iv. Apriori Algorithm

Apriori algorithm is used for finding frequent itemsets and association rule learning over relational databases. It proceeds by identifying the frequent individual items in the database and expands them to larger itemsets as long as those itemsets appear often in the database. The frequent item sets determined by Apriori can be used to dictate association rules which highlight general trends in the database.

v. Association Rule

Association rule mining is often used to create set of rules that define the basic patterns in a dataset. The correlation of the two attributes of an accident data is determined by the frequency at which they occur in the data set. Rule A → B shows A occurs first, then B also occurs.

DOI: 10.37200/IJPR/V24I5/PR201839
Received: 15 Feb 2020 | Revised: 07 Mar 2020 | Accepted: 23 Mar 2020

Pseudo code for Apriori algorithm

Lk = {Frequent item set with size k}

Ck = {Candidate item set with size k} L1 = {Frequent 1 item sets} K=1;

While (Lk-1≠ϕ) then Ck+1= candidates generated from Lk For each transaction t ϵ D do Increment the count of candidate in Ck+1 that also contained in t Lk+1= candidate in Ck+1 with minimum support K=K+1; Return UkLk. In association rule mining, to evaluate the quality of rules various interesting measures are used.

V. Conclusion And Future Work

In this paper, various data mining techniques have been applied on the raw dataset used. This paper's main goal is to use K-Means++ method for clustering of data. Apriori Algorithm is applied for association of the various itemsets.

Also the comparison between K-Means and K-Means ++ is performed and the following figure shows how the K-Means++ method is more efficient than the standard K-Means.

This paper aims to focus on factors that lead to road accidents and analyse them.

It aims to generalize the proposed method for clustering and give analysis for future work.

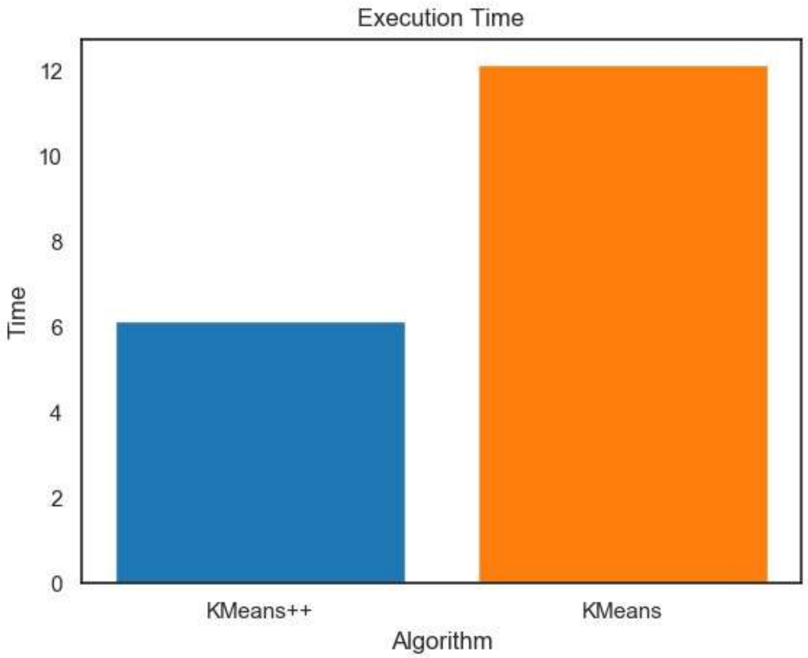


Figure 4: Comparison of K-Means++ with K Means

References

[1] Inderpreet Kaur, Ashish Kumar Luhach, Pooja, "Mining Of Road Accident Data Using K-Mode Clustering And Improved Apriori", *International Journal of Computer Science and Information Security (IJCSIS), Vol. 15, No. 4, April 2017*

DOI: 10.37200/IJPR/V24I5/PR201839
Received: 15 Feb 2020 | Revised: 07 Mar 2020 | Accepted: 23 Mar 2020

[2] Liling Li, Sharad Shrestha, Gongzhu Hu, "Analysis of Road Traffic Fatal Accidents Using Data Mining Techniques" *IEEE SERA 2017, June 7-9, 2017*

[3] Sachin Kumar, Durga Toshniwal, "Analysis of hourly road accident counts using hierarchical clustering and cophenetic correlation coefficient (CPCC)“, *Journal of Big Data, 3(1). doi:10.1186/s40537-016-0046-3*

[4] Helen WR, N.Almelu, S.Nivethitha,"Mining Road Accident Data Based on Diverted Attention of Drivers", *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS).*

[5] Sachin Kumar, Durga Toshniwal, "A data mining approach to characterize road accident locations", *Journal of Modern Transportation, 24(1), 62–72.*

[6] Amira A. El Tayeb, Vikas Pareek, Abdelaziz Anwar, "Applying Association Rules Mining Algorithms for Traffic Accidents in Dubai", *International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-5 Issue-4, September 2015*

[7] Luis Martin, Leticia Baena, Laura Garach, Griselda Lopez, Juan de Ona,"Using data mining techniques to road safety improvement in Spanish roads." *Procedia - Social and Behavioral Sciences, 160, 607–614.*

[8] Priyanka A.Nandurge, Nagaraj V.Dharwadakar, "Analyzing Road Accident Data using Machine Learning Paradigms", *2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC).*

[9] Ayushi Jain, Garima Ahuja, Anuranjana, Deepti Mehrotra, "Data Mining Approach to Analyse the Road Accidents in India", *2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO).*

[10] Suwarma Gothane, Dr. M. V. Sarode, "Analyzing Factors, Construction of Dataset, Estimating importance of factor generation of association rules for Indian Road Accident", *2016 IEEE 6th International Conference on Advanced Computing*

[11] Poojitha Shetty, Sachin P C,Supreeth V Kashyap, Venkatesh Madi, “Analysis of road accidents using data mining techniques”, *International Journal of Engineering & Technology, 7(3.10), 40.*

DOI: 10.37200/IJPR/V24I5/PR201839
Received: 15 Feb 2020 | Revised: 07 Mar 2020 | Accepted: 23 Mar 2020