IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Traffic Collision Analysis and Prediction

Abhijeet Madhukar Tote

Student
Department Of Computer Engineering,
Matoshri College Of Engineering & Research Centre, Eklhare, Nashik, India

***Abstract*—The increase in road fatalities comes as bad news. This cannot be stopped but can be controlled. The Accidents may cause due to driver's health, Driver's feelings, vehicle speed, climate condition, traffic conditions, road conditions, etc. Analysis and prediction on Traffic collision has gained importance in these days. The big dataset is generated every year. The proposed system works on analysis and prediction of road accidents information data using machine learning algorithms and its efficient execution. For analysis same type of accidents are clustered together using EMM algorithm and for prediction association mining is performed using Top-k Improved Association Rule Mining (IARM) algorithm. This algorithm finds the top k rules from dataset. The generated association rules are then provided to the Congestion control using Machine Framework (CCMF) and Traffic Congestion Analyzer using Mapping (TCAMP) algorithms to generate predictions. The Results are compared in terms of accuracy and efficiency with existing systems.**

Keywords*— *Road accidents, association rules , clustering, EMM, feature extraction, top k

I. INTRODUCTION

The increase in road fatalities comes as bad news. This has a negative social and economic impact. This cannot be stopped but can be controlled. World Health Organization (WHO), announces the statistical study every year. 60 million people get injured and 1.60 million people died consistently due to road accidents. A center for Disease Control and Prevention (CDCP) announces the economical impact of road accidents. The accidents cause loss of 100 billion every year The Accidents may caused due to driver's health, Driver's feelings, vehicle speed, climate condition, traffic conditions, road conditions, etc. Analysis and prediction on Traffic collision has gained importance in these days. The big dataset is generated every year. Enormous actions have been taken to enhance Road safety. Conventional strategies can't be utilized in such frameworks as the information produced is in huge volumes.

There is need of such system that analyzes this data in automatic faction. The machine learning algorithms are required for processing. By Analyzing the generated data the common accident types can be clubbed together. It will helpful for statistical analysis. The cause of accident can be predicted by machine learning algorithm. The possibility of accident can be predicted. Such predictions help to add more preventive measures. The requirement in the domain of road accidents motivates to develop a new system. The fusion of existing machine learning algorithms and big data analysis techniques can generate a better analytical report and prediction set.

The accidents may cause due to various parameters such as: climate condition, road conditions, drivers mental status, driver experience, etc. Analysis of accident cases based on such parameters helps to predict the future accident style at certain location or at certain cities. This helps to create preventive measure accordingly.

Machine learning algorithm helps to analyze data using one or more parameters. The clustering algorithm clubs together the similar type of accidents. The association mining rule technique finds the patterns in a dataset. The pattern mining technique helps to extracts the dependency of two or more type of attributes that may cause and accidents. The prediction can be done on the resultant effect of an accident in the form of number of injuries and death count. For example: Driver age=50, whether condition = fog, driver drink = Not checked, Road surface=Dry may cause accident with injury.

The proposed system works on analysis and prediction of road accidents information data using machine learning algorithms and its efficient execution. For analysis same type of accidents are clustered together using EMM algorithm. The Top k Association Rule Mining (IARM) algorithm extracts top k rules from dataset. As compared to the existing IARM algorithm, this algorithm generates top k rules in efficient manner. The top-k technique removes the minimum support dependency The generated association rules are then provided to the Congestion control using Machine Framework (CCMF) and Traffic Congestion Analyzer using Mapping(TCAMP) algorithms to generate predictions

II. RELATED WORK

Sarkar S ,et. al.[2] proposes a prediction system. This system predicts the possible incident in steel plant based on the related stored data. It predicts the occurrence of injury cases and their probable causes. It uses text mining technique with 3 different classification algorithms: Support Vector Machine SVM, Random Forest RF and Maximum entropy Max Ent. These classifiers generate better results in binary and multi-class prediction model. An Ensemble approach is proposed for multi-class prediction model. The Ensemble approach improves accuracy.

Flight crash investigation system is proposed by Sharma S, et.,al.[3]. This system focuses on the flight crash investigation and analysis using data mining techniques. It finds the ground/abroad fatality rate. The clustering algorithm K-Means is used along with the cosine similarity measure. The clustering results group the similar crash information in one group. Similarity finds the relation among

different texts of crashes.

A railroad accident investigation reports generation system is proposed by Williams T, et.,al.[4]. This technique generate a statistical analysis report based on Similarity found among different texts of railroad Accidents. In this technique, the text form published articles is analyzed using data mining techniques. The dataset of published articles contains railway accidents. Using LDA technique topics are extracted from text. The K means clustering is applied on extracted text. The clustering results show that there is recurring themes in many major accidents. After grouping the accidents data the main causes of accidents are extracted and relation among multiple accidents is identified.

Sarkar S,et.,al,[5] Works on prediction of occurrence of accidents and its outcome such as injury, near miss, fatal death, property damage, etc. . And finds the inter relationship of factors causing accidents. The two machine learning algorithms : support vector machine (SVM) and artificial neural network (ANN) are used. The parameter passed to this algorithm is optimized using genetic algorithm and particle swarm optimization. After these algorithms association rules are extracted with the help of decision tree algorithm with PSO and SVM. This technique extracts the root cause behind the injury.

Verma A[6] proposes a analysis tool for steel plant incident data. It explores the hidden factors and patterns form the description. It also identifies the anomaly in incidence reporting. The data is in case report text format. It uses singular value decomposition (SVD) and expectation-maximization (EM) algorithm. This paper only focuses on grouping of similar type of accident data.

Williams T. et.al[7] proposes a system to analyze road accidents based on text mining techniques such as: probabilistic topic modeling and k-means clustering. The system works on major accidents that are occurred in the same fashion. Parameters those are analysed in the system are: track defects, grade crossing accidents, wheel defects, and switching accidents.

Ghazizadeh et.al[8] proposes a technique to analyze national highway complaint dataset. The data is analyzed at 2 levels: fatal incident and injury. latent semantic analysis (LSA) and hierarchical clustering technique is used to cluster complaints.

F. Abdat[9] proposes a system that analyzes recurrent accidents caused due to Movement Disturbance. This is called as Occupational Accident with Movement Disturbance (OAMD) scenarios. The dataset is in the form of descriptive text from. A Bayesian Network (BN)-based model is used to extract informative text. Then Most Probable Explanation (MPE) is extracted by creating clusters based on the similarity measurements.

Babu S.N., et.al.[1] works on road accidents prediction theory based on various parameters like driver-age, experience, vehicle type, whether condition, road conditions, etc. The system uses Congestion control using Machine Framework (CCMF) and Traffic Congestion Analyzer using Mapping(TCAMP) algorithms for prediction on road accidents. Initially the system forms clusters based on the clusters the association rules are extracted. Using association rules predictions are performed. Lots of parameters are used for mining accidental data. The attribute reduction will be useful for accuracy improvement.

III. ANALYSIS AND PROBLEM FORMULATION

Lot of work has been done on road accident analysis using text mining technique. Most of the system generates a statistical analysis by clustering algorithm. The rule extraction helps to analyse the patterns and helps to predict the future occurrence of accidents. There is need to generate analytical report from the accidental database efficiently and generate a road accident predictions by considering more than one factor at a time that causes accidents.

The problem statements can be defined in three folds:

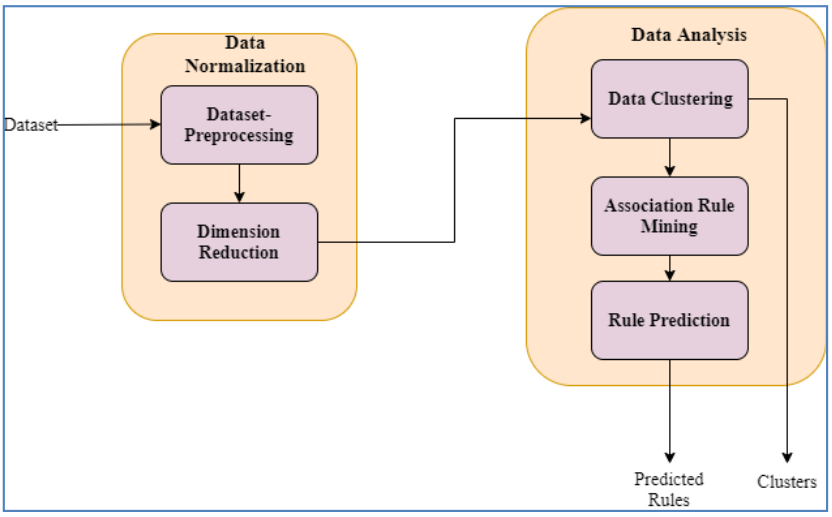
1. Generate an analytical report from accidental database.
2. Generate predictions on accidents using association rule mining.
3. To improve system efficiency for processing.

IV. PROPOSED METHODOLOGY

A. Architecture

Following fig.1 describes the architecture of the system. The accident information dataset is input to the system. The Prediction rules and clusters are the output of system. System mainly categorized in 2 sections: Data normalization and data analysis. Data normalization is treated as a preprocessing step whereas in data analysis phase actual data mining techniques are applied such as data clustering, association rule mining and rule prediction algorithms.

FIGURE 1 : SYSTEM ARCHITECTURE



B. System Working:

The data processing is mainly described in following 4 sections:

1. Data Preprocessing:

In data preprocessing the noisy data is removed. The dataset contains raw information of all type of accidents. The unwanted information from dataset is removed.

The dataset may contain missing values. The missing values are filled using binning and linear regression technique.[7]

2. Dimension reduction:

A Wrapper technique: CFS subset evaluator is used to reduce the dimension count. The reduced dimension count helps to improve system efficiency. This finds the optimal feature set based on the classifier performance.

3. Data Clustering

The whole dataset of road accidents is initially divided in number of clusters based vehicle type and then the cluster is divide in subgroups using parameter like : time, drivers experience, climate etc. The Enhanced Expectation-Maximization algorithm is used for data clustering.[9] this clustering algorithm focuses on grouping of similar members based on probability distribution.

4. Top –K Association Rule Mining

From the clustered data association among data is extracted using Top-k Improved Association Rule Mining (IARM) algorithm[3]. This technique extracts the strong association using support value. The rule are extracted based on vehicle class and road accident parameters. This technique generates top k association rules. The top k technique removes the dependency of support value. Lot of rules are get filtered at early stage and improves the system efficiency.

5. Rule Prediction

Congestion control using Machine Framework(CCMF) and Traffic Congestion Analyzer using Mapping(TCAMP) algorithms are used for prediction.

The generated clusters of EM algorithm based on vehicle type are given to the input to CCMF algorithm. This algorithm finds the clusters of relevant parameters. The cluster is in the form of tree. Using the generated clusters and association rules, the TCAMP algorithm predicts the possibility of road accidents.

C. Algorithms

1. CCMF algorithm

Input: Training dataset T and attributes/parameters.P
Output: Multiple clusters based on parameters of vehicle type using decision Tree.
Processing:

1. If (count(T) is NULL)
   Stop
2. Else if (count(P) is NULL)
   Stop
3. Else if (|T| OR |P|) is 1
   only parameter is considered in the dataset and one node is formed a parent node.
4. Else
   i. For p1 ε P and P ε T
   ii. If (p1 ε VTk)
   iii. split(T) = p1;

2. TCAMP Algorithm

Input Traffic data set
Output predicting the accidents
Processing:

1. Apply pre-processing using binning and linear regression
2. Apply dimensionality reduction using PCA
3. Create data clusters using Enhanced Expectation-Maximization
4. apply Association rules mining using IARM.
5. For vehicleclass (V)
6. For Dc(i) to MAX do
7. Apply Association rule on parameter P(i).
8. Display Prediction set PS.

D. Mathematical Modeling:

The System S can be defined in set theory form as,
S = {I, O,F} where,
I = {I1,I2}, Set of inputs
I1= Traffic data set
I2 = parameters for analysis

O= {O1, O2, O3}, Set of outputs
O1 = Clusters of accidental data
O2 = Association rules
O3 = Accidental condition predictions

F= {F1, F2, F3, F4, F5, F6, F7, F8, F9, F10, F11 }, Set of Functions
F1 = Upload data
F2 = Data Preprocessing
F3 = Binning and linear regression
F4 = Dimension reduction
F5 = Data Clustering
F6 = Enhanced Expectation-Maximization algorithm
F7 = Association Rule Mining
F8 = Improved Association Rule Mining (IARM) algorithm
F9 = Rule Prediction
F10 = Congestion control using Machine Framework
F11 = Traffic Congestion Analyzer using Mapping
F12 = View Result

V. RESULT AND ANALYSIS

The system is implemented on Ubuntu 18.04 with 4GB RAM and I3 processor. The system is implemented in java using jdk 1.8 environments. For single node Hadoop implementation Hadoop-3.1.1 is used.

A. Dataset:

The data is downloaded from government website[10]. This website contains road accident information from different states, with specific year, and with respect to various parameters.

The parameters of dataset are enlisted in the following table:

TABLE 1 : DATASET PARAMETER DESCRIPTION

Sr. No.	Attributes	Possible values
1.	Number of Vehicles	numeric
2.	Time (24hr)	numeric
3.	1st Road Class	1,Motorway 2,A(M) 3,A 4,B 5,C 6,Unclassified
4.	Road Surface,Road Surface Desc	1,Dry 2,Wet / Damp 3,Snow 4,Frost / Ice 5,Flood (surface water over 3cm deep)
5.	Lighting Conditions,	1,Daylight: street lights present 2,Daylight: no street lighting 3,Daylight: street lighting unknown 4,Darkness: street lights present and lit 5,Darkness: street lights present but unlit 6,Darkness: no street lighting 7,Darkness: street lighting unknown
6.	Weather Conditions	1,Fine without high winds 2,Raining without high winds 3,Snowing without high winds 4,Fine with high winds 5,Raining with high winds 6,Snowing with high winds 7,Fog or mist – if hazard 8,Other 9,Unknown
7.	Type of Vehicle	1,Pedal cycle 2,M/cycle 50cc and under 3,Motorcycle over 50cc and up to 125cc 4,Motorcycle over 125cc and up to 500cc 5,Motorcycle over 500cc 6,[Not used] 7,[Not used] 8,Taxi/Private hire car 9,Car 10,Minibus (8 – 16 passenger seats)

		11,Bus or coach (17 or more passenger seats) 12,[Not used] 13,[Not used] 14,Other motor vehicle 15,Other non-motor vehicle 16,Ridden horse 17,Agricultural vehicle (includes diggers etc.) 18,Tram / Light rail 19,Goods vehicle 3.5 tonnes mgw and under 20,Goods vehicle over 3.5 tonnes and under 7.5 tonnes mgw 21,Goods vehicle 7.5 tonnes mgw and over 22,Mobility Scooter 90,Other Vehicle 97,Motorcycle - Unknown CC
8.	Casualty Class	1,Driver or rider 2,Vehicle or pillion passenger 3,Pedestrian
9.	Casualty Severity	1,Fatal 2,Serious 3,Slight
10.	Sex of Casualty	1,Male 2,Female
11.	Age of Casualty	Numeric

B. Performance Measures:

The system performance is measured in terms of :

1. Time: The time required for processing is captures for analysis of prediction system with IARM and top k IARM
2. Memory: The memory required for processing is captured for analysis.

C. Results:

1. Cluster Generation time:

Using EM clustering technique the similar accident entries are grouped together. Following table shows the time required for clustering the different data sizes. As we increase the data size, time required for processing also increases.

Data Count	Processing Time in Millisec.
2000	850
4000	1154
6000	2648
8000	3317

2. Pattern Predicted:

The following table contains the rules extracted from the dataset by using **TCAMP** algorithm. The rule defines the occurrence of one or more parameters together that lead to what kind of Casualty Severity occurs.

Sr. No.	Rules	Casualty Severity
1.	Darkness: street lights present but unlit, DAFog or mist	Fatal
2.	Darkness: no street lighting, road Wet / Damp	Fatal
3.	Darkness: street lights present but unlit, road Wet / Damp	Serious
4.	Daylight: street lights present, road Wet / Damp	Slight

3. Rule extraction with different support value:

For the existing system[1], the minimum support value is required for rule discovery. For larger minimum support value, very few rules are extracted whereas for less minimum support value too many rules are extracted. There is dependency between support value and number of rule extracted. To remove this dependency, top-k rule extraction technique is used. This removes the minimum support dependency and executes the TCAMP algorithm in efficient manner.

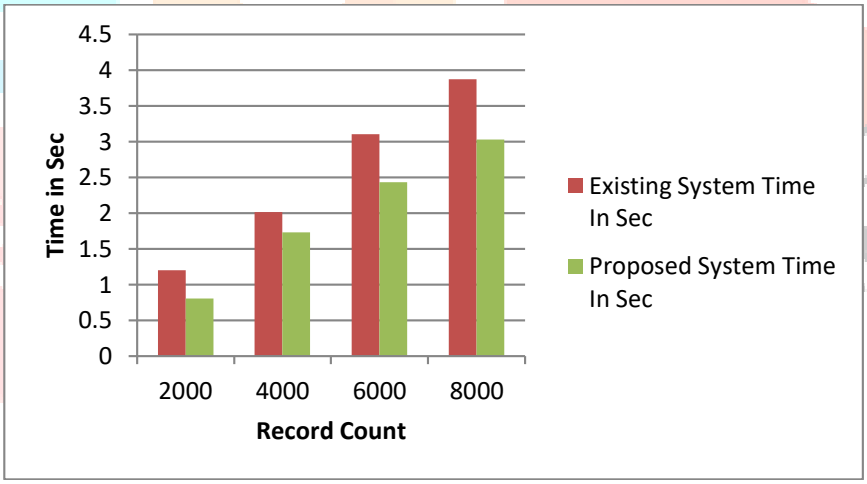
Data Count	Minimum support value	Items found
2000	0.2	360
4000	0.4	194
6000	0.6	73
8000	0.8	26

4. Time of execution for TCAMP algorithm:

Following table shows the comparative analysis of existing and proposed system in terms of time required for processing. The proposed system requires less time as compared to the existing system.

Data Count	Existing System Time In Sec	Proposed system Time In Sec
2000	1.2	0.8
4000	2.01	1.73
6000	3.1	2.43
8000	3.87	3.03

Following graph shows the time analysis results for existing and proposed system The proposed system is executed efficiently as compared to the existing system. The proposed system does not have the dependency of user defined threshold value for rule extraction and prediction process. The existing system is run with min-support value as 0.5. The required for processing increases as we increase the dataset size.



VI. CONCLUSIONS

The accident analysis includes the road, railway, airline and manufacturing plant accidents. In the study of analysis of accident system, the causes of accidents and the implications are studied. Based on the analysis of a prediction can be estimated. The prediction helps to define prevention measures. The proposed system analyzes the road accident dataset and finds the clusters form the data. The road accident may caused due to various parameters like climate, road condition, driver status ,etc. The system extracts top k association rules from the data using IARM algorithm. Based on the rules and accident type road accident predictions are done using CCMF and TCAMP algorithm. During system execution following points are noticed:

- The number of rules depends upon the given threshold. Top k count removes the dependency.
- Time required for proposed system is less than the time required for existing system due to extraction of only top k association rules.

In future the system can be tested on different dataset like flight crashes, industrial accident analysis, etc.

VII. REFERENCES

[1] S.N., Tamilselvi J., "Generating road accident prediction set with road accident data analysis using enhanced expectation-maximization clustering algorithm and improved association rule mining", Journal Europeen des Systemes Automatises, Vol. 52, No. 1, pp. 57-63, April 2019. https://doi.org/10.18280/jesa.520108

[2] Sarkar S, Pateshwari V, Maiti J. (2017). Predictive model for incident occurrences in steel plant in India. In ICCCNT 2017, IEEE, pp. 1-5. http://dx.doi.org/10.14299/ijser.2013.01

[3] Sharma S, Sabitha AS. (2016). Flight crash investigation using data mining techniques. In Information Processing (IICIP), 2016 1st India International Conference on. IEEE, pp. 1-7. http://dx.doi.org/10.14299/ijser.2013.01

[4] Williams T, Betak J, Findley B. (2016). Text mining analysis of railroad accident investigation reports. In 2016 Joint Rail Conference. American Society of Mechanical Engineers V001T06A009-V001T06A009. http://dx.doi.org/10.14299/ijser.2013.01.

[5] Sarkar S, Vinay S, Raj R, Maiti J, Mitra P. (2018). Application of optimized machine learning techniques for prediction of occupational accidents. Computers & Operations Research (Elsevier), pp. 343-348. http://dx.doi.org/10.1145/3075564.3078884

[6] Verma A, Maiti J. (2018). Text-document clustering based cause and effect analysis methodology for steel plant incident data. International Journal of Injury Control and Safety Promotion, 1-11. http://dx.doi.org/10.1080/17457300.2018.1456468

[7] Williams T, Betak J, Findley B., "Text mining analysis of railroad accident investigation reports", In 2016 Joint Rail Conference. American Society of Mechanical Engineers V001T06A009-V001T06A009. http://dx.doi.org/10.14299/ijser.2013.01

[8] Ghazizadeh M, McDonald AD, Lee JD., "Text mining to decipher free-response consumer complaints: Insights from the nhtsa vehicle owner's complaint database", Human Factors 56(6): 1189-1203. http://dx.doi.org/10.1504/IJFCM.2017.089439

[9] Abdat F, Leclercq S, Cuny X, Tissot C.,"Extracting recurrent scenarios from narrative texts using a bayesian network: Application to serious occupational accidents with movement disturbance", Accident Analysis & Prevention 70: 155-166. http://dx.doi.org/10.1016/j.aap.2014.04.004

[10] Dataset: https://data.gov.in/dataset-group-name/road-accidents