Analysis and Prediction of Bangalore Traffic South Road Accidents

Ramya V[#1]
Department of Information Science and Engineering
Dayananda Sagar College of Engineering
Bengaluru Karnataka 560078
[1]*ramy.viju@gmail.com*

Abstract- Data mining is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cut costs or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Traffic accidents cause enormous losses for a country and plenty of national assets drain away every year due to it. The rapid proliferation of Global Position Service (GPS) devices and mounting number of traffic monitoring systems employed by municipalities have opened the door for advanced traffic control and personalized route planning. But the complexity of traffic accident analysis has brought many difficulties to traffic management and decision-making. Most state of the art traffic management and information systems focus on data analysis and very little has been done in the sense of prediction. This paper provides details about how road accidents and traffic data can be analysed and used to predict the probability of an accident to occur. To start with, the analysis has been done on the Bangalore city traffic considering five traffic stations of the south region of the city – Basavanagudi, Kumaraswamy layout, Banashankari, Jayanagar and Chamarajapet.

____*****____

I. INTRODUCTION

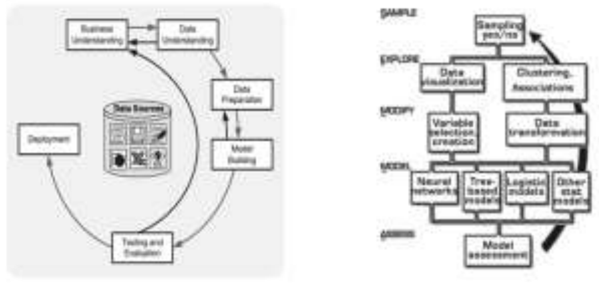
Analysis refers to breaking a whole into its separate components for individual examination. Data analysis is a process for obtaining raw data and converting it into information useful for decision-making by users. Data is collected and analyzed to answer questions, test hypotheses or disprove theories. More precisely , Data analysis can be defined as- Procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data.

In order to systematically conduct data mining analysis, a general process is usually followed. There are some standard processes –

1) CRISP-DM (Cross-Industry Standard Process for Data Mining) is an industry standard process consisting of a sequence of steps that are usually involved in a data mining study

2) SEMMA (sample, explore, modify, model, assess) well-known methodology developed by the SAS Institute

While each step of either approach isn't needed in every analysis, this process provides a good coverage of the steps needed, starting with data exploration, data collection, data processing, analysis, inferences drawn and implementation. The following diagrams show the steps involved in each of these approaches.



CRISP-DM process

Schematic of SEMMA
(original from SAS Institute)

Our project is a combination of both of the above mentioned approaches CRISP-DM and SEMMA. The process followed includes the following main steps – business understanding, data pre-processing, sampling, cross validation, exploration, feature selection, data modelling, regression analysis and assessment.

II. RELATED WORKS

This section provides a background to the research through a review of some of the literature which are central to the scope of this paper.

According to Sachin Kumar and Durga Toshniwal in "Analyzing Road Accident Data Using Association Rule Mining" [1], data mining techniques can been used to analyze the data provided by EMRI (Emergency Management research Institute) in which the accident data is first clustered using K-modes clustering algorithm and further association rule mining technique is applied to identify circumstances in which an accident may occur for each cluster. An Apriori algorithm has then been applied on every cluster using WEKA3.6 to generate association rules.

The paper "Data Integration and Clustering for Real Time Crash Prediction" [2] presents a methodology to develop a classifier which uses historical as well as real-time data for any given road and outputs the probability of accident on any given time on that road. By using clustering method, a Bayesian network (BN) is constructed by mining the database using Kernel Density Estimation (KDE). This BN can act as a classifier to predict the probability of accidents for any given street. Data has been extracted from several simulation runs by PARAMICS micro-simulator.

As most of the current studies did not pay enough attention to the time factor when studying the relationship between traffic state and crashes on highways, the paper "Evolution of Traffic Flow Analysis under Accidents on Highways Using Temporal Data Mining" [3] proposed a method to construct time-series data using traffic flow data when accidents happened. This paper proposes a model to describe traffic flow evolution when accidents happen on highways. A Discrete Fourier Transform was implemented to turn the time series from time domain to frequency domain. By clustering analysis, the traffic dynamics can then be studied. The newly developed method provides a better insight into the evolution of traffic flow on highways and the impact of highway crashes.

In "Application of Spatial Data Mining in Accident Analysis System" [4], a system based on data mining of Geographical Information System (GIS) has been proposed. A three-layer architecture displays the whole process of accident data extracting, preprocessing and mining and it applies spatial data mining to GIS. The first layer is data storage layer which includes databases. The second layer is business logic layer, including spatial data mining function module and accident analysis module. A design module can inquire about, add and delete spatial data, attribute data and accident data. The accident analysis module is used to identify and display accident black-spots on the map according to a certain accident rate and accident level. The third layer is the interactive interface. ArcGIS Engine and C# have been used to develop the system.

In his paper, Andreas Gregoriades [5] proposed a method which is split into two phases, (A) the development of the microscopic simulation model and (B) the development of the agent-based monitoring system phase. During Phase A, a preliminary micro-simulation model of a road network in Cyprus is developed using statistical data. Based on these a topology of the BBN model was created. Phase B of the method addresses the development of an agent-based monitoring system to track changes in traffic volumes, densities, behaviour and speed. This technique can test whether the model would generate predictions similar with the known scenario outcomes.

"Large Imbalance Data Classification Based on MapReduce for Traffic Accident Prediction" [6] suggests using Hadoop framework to process and analyze big traffic data efficiently. Based on this, the predicting system first preprocesses traffic big data and analyzes it to create data for the learning system. The imbalance of created data is corrected by a sampling method. To improve accuracy, corrected data is classified and analysis is applied. Using MapReduce, processing performance for generating training data set and conducting classification can be improved. The overall system is based on Hadoop, and which implements data pre-processing, learning data creation, over sampling by Hive. The cluster and classification analysis is operated by Mahout. First, input the accident and traffic data to Hive and process them with Hadoop.

III. BUSINESS UNDERSTANDING

The key element of a data mining study is to know what the study is for. Goals in terms of things such as "What types of end-users are interested in the product?" or "What are the typical profiles of our end-users?" .Then a plan for finding such knowledge needs to be developed, in terms of those responsible for collecting data, analyzing data and reporting. In similar lines, we scrutinized the data requirements, identified the sources of data and decisions were made on choosing the method and tools for analysis.

IV. DATA UNDERSTANDING

The first stage of the data mining process is to select the related data from many available databases to correctly describe a given business task. Data sources for data selection can vary. For project, data was collected from three sources –

1) Basavanagudi Traffic Police
2) District Commissioner of Police (Hebbal)
3) UK government accidents dataset

The data type can be categorized as quantitative and qualitative data. Quantitative data is measurable using numerical values. It can be either discrete (such as integers) or continuous (such as real numbers). Qualitative data, also known as categorical data, contains both nominal and ordinal data. Nominal data has finite non-ordered values, such as gender data which has two values: male and female. Ordinal data has finite ordered values.

Once relevant data are selected, data pre-processing should be pursued.

V. DATA PRE-PROCESSING

Data pre-processing involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent and/or lacking in certain behaviours or trends and is likely to contain many errors. Data pre-processing is a proven method of resolving such issues. Data pre-processing prepares raw data for further processing. This step consists of attribute selection, data cleaning and data transformation.

Data Field Selection: Data gathered from different sources was consolidated, mapped and scrutinized. Some of the data that is not pertinent to the data mining exercise was ignored.

Data Cleaning: Data cleaning is a process used to determine inaccurate, incomplete or unreasonable data and then improve the quality through correcting of detected errors and omissions. Generally data cleaning reduces errors and improves the data quality. Some entries were clearly invalid, caused by either human error or the evolution of the problem reporting system. Those errors that were correctable were corrected. If all errors detected for a report were not corrected, that report was discarded from the study.

Data Transformation:

Data transformation converts a set of data values from the data format of a source data system into the data format of a destination data system. In this process, few attributed were transformed into required formats for example, the attribute "Hour" was converted into 24-hour format. Also, the attribute values were hard-coded for better representation in the training dataset.

The initial set of attributes identified is presented in the following table.

Accident index	Location easting OSGR	Location northing OSGR
Longitude	Latitude	Police force
Accident severity	No. Of vehicles	NO. Of casualties
Date	Day of week	Time
District local authority	Highway local authority	1st road class
1st road no.	2nd road class	2nd road no.
Road type	Speed limit	Junction detail
Junction control	Pedestrian crossing human control	Pedestrian crossing physical facilities
Light conditions	Weather conditions	Road surface conditions
Special conditions at site	Carriageway hazards	Urban or Rural area
Did police officer attend scene of accident	LSOA of accident location	Station
Age of victim	Road name	Driver type
Pedestrian	Gender	Vehicle to road ratio
Vehicle manoeuvre	Day/Night	Vehicle type
Hit and run	With/without helmet	Driving license suspension

With these attributes identified, only a few of them qualified to be relevant to our data mining goal. With the selected set of attributes, the training dataset was populated.

The attributes considered in the training dataset are presented in the following table.

ATTRIBUTE	TYPE	POSSIBLE VALUES
Station	Numeric	1-Chamarajapet 2-Basavanagudi 3-Jayanagar 4-BSK 5-K.S.Layout
Pedestrian	Nominal	Y or N
Driver type	Numeric	2-Cyclist 3-M.Drivers 4-M/c Riders 5-Pillian riders 6-Passengers 7-Occupants 8-others
Gender	Nominal	M or F
Road type	Numeric	1-National highway 2-State highway 3-Other roads
Accident severity	Numeric	1- killed 2- severely injured 3- slightly injured
Age of victim	Numeric	1-Below 6yrs 2- (6-18) 3- (18-30) 4- (30-50) 5-Above 50
Special conditions at site	Numeric	0-None 1-Auto traffic signal out 2-Auto signal defective 3-Road sign/marking defect 4-Road works 5-Road surface defective 7-Mud
Casualties	Numeric	Any positive value
Junction control	Numeric	1-Authorised person 2-Auto traffic signal 3-Stop sign 4-Givway/uncontrolled
Junction details	Numeric	0-Not a junction 1-Roundabout 2-Mini-roundabout 3-T-junction 5-Slip road 6-Crossroads 7-More than 4 arms 8-Other junction
Speed limit	Numeric	As specified by government
Road name	Numeric	Coded with respect to stations
Day/Night	Nominal	D or N
Road surface conditions	Numeric	1-Dry 2-Wet or Damp 7-Mud
Light conditions	Numeric	1-Daylight 4-Darkness:lights lit 5-Darkness:lights unlit 6-Darkness:no lighting 7-Lighting unknown
Weather conditions	Numeric	1-Fine no high winds 2-Raining no high

		winds 4-Fine+high winds 5-Raining+high winds 7-Fog or mist 8-Other 9-Unknown
Pedestrian crossing physical facilities	Numeric	0-No facilities 1-Zebra crossing 4-Non-junction light crossing 5-Pedestrain phase at signal 7-Subway
Day	Numeric	1-Monday 2-Tuesday 3-Wednesday 4-Thursday 5-Friday 6-Saturday 7-Sunday
Vehicle type	Numeric	1-BMTC Bus 2-KSRTC Bus 3-Factory Bus 4-Private Bus 5-Lorry 6-Car 7-Taxi 8-Jeep 9-Auto rickshaw 10-Motor cycle 11-Scooter 12-Moped 13-Tempo 14-Van 14-Maxi Cab 15-Un-known vehicle 16-Tractor 17-Tanker 18-Cycle 19-Bullock cart 20-Others
Vehicle maneuver	Numeric	1-Reversing 2-Parked 3-Waiting to go-held up 4-slowing or stopping 5-moving off 6-U-turn 7-Turning left 8-waiting to turn left 9-Turning right 10-Waiting to turn right 11-Changing lane to left 12-Changing lane to right 13-Overtake moving vehicle 14-Overtake static vehicle 15-Overtaking nearside 16-Left hand bend 17-Right hand bend 18-Going ahead other
Hour (24hr format)	Numeric	1- (0-6) 2- (6-9) 3- (9-12) 4- (12-15) 5- (15-18) 6- (18-21) 7- (21-24)

VI. SAMPLING

Sampling is a commonly used approach for selecting a subset of the data objects to be analysed. The key principle for effective sampling is – using a sample will work almost as well as using the entire dataset if the sample is representative. In turn, the sample is representative if it has approximately the same property (of interest) as the original set of data. This is where a portion of a large data set is extracted for optimal cost and computational performance. In this process, partitioned data sets were created for better accuracy assessment. The following were performed:

1) Training – used for model fitting
2) Cross Validation – used for assessment and to prevent over fitting
3) Test – used to obtain an honest assessment of how well a model generalizes

VII. CROSS VALIDATION

Cross-validation, sometimes called rotation estimation, is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set. It is mainly used in settings where the goal is prediction and one wants to estimate how accurately a predictive model will perform in practice. In a prediction problem, a model is usually given a dataset of known data on which training is run (training dataset) and a dataset of unknown data (or first seen data) against which the model is tested (testing dataset). The goal of cross validation is to define a dataset to "test" the model in the training phase (i.e., the validation dataset), in order to limit problems like over fitting, give an insight on how the model will generalize to an independent dataset.

One round of cross validation involves partitioning a sample of data into complementary subsets, performing analysis on one subset (called the training set) and validating the analysis on the other subset (called the testing set).To reduce variability, multiple rounds of cross-validation are performed using different partitions and the validation results are averaged over the rounds.

k-fold cross-validation:

In *k*-fold cross-validation, the original sample is randomly partitioned into *k* equal sized subsamples. Of the *k* subsamples, a single subsample is retained as the validation data for testing the model, and the remaining *k* – 1 subsamples are used as training data. The cross-validation process is then repeated *k* times (the *folds*), with each of the *k* subsamples used exactly once as the validation data. The *k* results from the folds can then be averaged or combined to produce a single estimation.10-fold cross-validation is commonly used, but in general *k* remains an unfixed parameter.

In Weka a number of algorithms such as Random Forest, J48, Naïve Bayes and Lib SVM were used for performing cross validation. A brief description of these algorithms is presented below.

- ***Random forest:*** operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.
- ***J48:*** is the implementation of ID3 algorithm developed by Weka project team and it is used to generate univariate decision trees.
- ***Naïve Bayes:*** probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features.
- ***Lib SVM:*** library for support vector machines (are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis).

A comparative study of the above algorithms is summarized in the following table.

Attribute Removed	J48	Random Forest	Naïve Bayes	Lib SVM
Pedestrian	100	100	99.867	93.085
Driver Type	100	100	100	94.1489
Day	100	100	99.867	94.813
Hour	100	100	99.867	94.414
Accident Severity	100	100	99.734	92.42
Age of Victim	100	100	100	93.75
Gender	100	100	99.867	93.218
Vehicle Type	100	100	98.138	91.356
Causalities	100	100	98.138	91.356
Road Type	100	100	99.867	93.085
Speed Limit	100	100	99.867	93.218
Junction Detail	100	100	100	93.617
Junction Control	100	100	99.867	93.085
Pedestrian Crossing	100	100	99.867	94.946
Light Condition	100	100	99.867	94.414
Weather condition	100	100	99.867	95.744
Road Surface Conditions	100	100	99.867	94.016
Special conditions	100	100	99.734	95.079
Vehicle Manoeuvre	100	100	99.867	97.207
Day/Night	100	100	99.867	93.218
Road Name	100	100	99.867	98.67
Pedestrian + Driver type	100	100	100	92.952
Pedestrian + Day	100	100	99.867	93.351
Day+Hour	100	100	99.867	94.964
Light + Weather conditions	100	100	99.867	96.143
*Driver type + age + Vehicle type + Accident severity + Causalities	69.813	61.037	65.292	56.64
*+Speed Limit	69.813	61.968	65.292	56.11
**+Road Type	69.813	61.569	65.292	54.92
**+Gender	71.01	60.638	65.292	54.787
**+Hour	71.1436	61.835	65.292	54.255
**+Day/Night	71.276	60.904	62.109	57.812
*+Day/Night	68.75	60.904	60.546	57.812
**+Hour	70.312	62.633	60.937	58.984

**+Gender	71.276	61.037	66.223	58.593
*Driver type + age + Vehicle type + Accident severity + Causalities + Day/Night + Hour + Gender + Road type	69.531	62.367	66.09	58.203
*+Speed Limit	69.531	60.904	62.109	57.812
Driver type + age + Vehicle type + Accident severity + Causalities + Gender + Speed limit	68.359	61.569	60.937	58.203

VIII. EXPLORATION

This is where unanticipated trends and anomalies are searched in order to gain a better understanding of the data set. After sampling the data, the next step is to explore them visually or numerically for inherent trends or groupings. Exploration helps refine and redirect the discovery process. In our project, this process was carried out using visualisation technique provided by Weka.

IX. FEATURE SELECTION

Feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features (variables, predictors) for use in model construction.

Feature selection techniques are used for three reasons:

- ➢ simplification of models to make them easier to interpret by users
- ➢ shorter training times
- ➢ enhanced generalization by reducing over fitting (formally, reduction of variance)

The goal of feature selection is to choose a subset Xs of the complete set of input features so that the subset Xs can predict the output Y with accuracy comparable to the performance of the complete input set X, and with great reduction of the computational cost.

Feature Selection was performed using XLminer. This add-in provides three algorithms Chi-Square, Mutual Information and Gain Ratio for feature selection.

- ➢ ***Chi-Square:*** x^2 test is used in statistics, to test the independence of two events. More specifically in feature selection it is used to test whether the occurrence of a specific term and the occurrence of a specific class are independent. The Chi-Square formula is as follows:

$$\chi^2 = \sum \frac{(\text{Observed Value} - \text{Expected Value})^2}{(\text{Expected Value})}$$

The top 4 attributes ranked by Chi-Square test are:

- Vehicle_manoeuvre
- Road_name
- Special_conditions_at_site
- Day

- ➢ ***Mutual Information:***

Mutual information (MI) of two random variables is a measure of the mutual dependence between the two variables. More specifically, it quantifies the "amount of information" obtained about one random variable, through the other random variable.

The top 4 attributes ranked by Mutual information are:

- Vehicle_manoeuvre
- Road_name
- Special_conditions_at_site
- Day

- ➢ ***Gain Ratio:*** Information gain ratio is a ratio of information gain to the intrinsic information. It is used to reduce a bias towards multi-valued attributes by taking the number and size of branches into account when choosing an attribute.

The top 4 attributes ranked by Gain ratio are:

- Vehicle_manoeuvre
- Special_conditions_at_site
- Junction_details
- Road_name

X. DATA MODELLING

In this process, models that explain patterns in the data are constructed and variable combination that reliably predicts a desired outcome are searched. Modelling techniques in data mining include artificial neural networks, decision trees, rough set analysis, support vector machines, logistic models, and other statistical models – such as time series analysis, memory-based reasoning and principal component analysis. With respect to our application, decision trees and support vector machines are used for Data Modelling.

XI. REGRESSION ANALYSIS

Regression analysis is a statistical process for estimating the relationships among variables. It includes many techniques for modelling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables (or 'predictors'). More specifically, regression analysis helps one understand how the typical value of the dependent variable (or 'criterion variable') changes when any one of the independent variables is varied, while other independent variables are held fixed.

In our case, dependent variable is the prediction and independent variables are vehicle_manoeuvre, special_conditions_at_site and day.

The regression analysis was performed using Weka and the following formula was obtained:

Prediction = (-0.0329*day+ -0.0349*special_conditions_at_site+ -0.031*vehicle_manoeuvre+0.8254)

XII. ASSESSMENT

In this final step, the usefulness and the reliability of findings from the data mining process are evaluated. A common means of assessing a model is to apply it to a portion of data set put aside (and not used during the model building) during the sampling stage. If the model is valid, it should work for this reserved sample as well as for the sample used to construct the model. Similarly, the model can be tested against known data.

For assessment of the final training dataset of the project, Weka was used.

XIII. CONCLUSION

Statistics from Bangalore City Traffic Police exposes how the causes behind crashes and fatalities are largely neglected while preparing plans to improve road safety which leads to a greater requirement for a well documented road safety analysis. With these analysis results, a Predictive Model for Accident Management System was designed and implemented which facilitates the analysis of road traffic accident data and ascertains the causes of such accidents; suggest precautionary strategies for preventing or controlling accidents for the benefit of road users by providing reports on various factors that lead to such accidents, draw attention of the authorities in providing a better view of road accidents' statistics and thereby help to take up necessary measures and also helps in identifying accident zones.

XIV. REFERENCES

[1] Sachin Kumar , Durga Toshniwal ,"Analyzing Road Accident Data Using Association Rule Mining", IEEE 2015 International Conference on Computing, Communication and Security (ICCCS)

[2] Elahe Paikari,Mohammad Moshirpour, Reda Alhajj, Behrouz H. Far, " Data Integration and Clustering for Real Time Crash Prediction", IEEE IRI August 13-15, 2014, San Francisco, California, USA

[3] An Shi,Zhang Tao, Zhang Xinming, Wang Jian, "Evolution of Traffic Flow Analysis under Accidents on Highways Using Temporal Data Mining", 2014 Fifth International Conference on Intelligent Systems Design and Engineering Applications

[4] Wang Jinlin, Chen Xi, Zhou Kefa, Wang Wei,Zhang Dan ,"Application of Spatial Data Mining in Accident Analysis System", 2008 International Workshop on Education Technology and Training

[5] Andreas Gregoriades, "Towards a user-centered Road Safety Management method based on Road Traffic Simulation", 2007 Winter Simulation Conference

[6] Seoung-hun Park ,Young-guk Ha, "Large Imbalance Data Classification Based on MapReduce for Traffic Accident Prediction", 2014 Eighth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing

[7] Garima R. Singh, Snehlata S. Dongre, "Crash Prediction System for Mobile Device on Android by Using Data Stream Minning Techniques", 2012 Sixth Asia Modelling Symposium

[8] Eyad Abdullah, Ahmed Emam, "Traffic Accidents Analyzer Using Big Data", International Conference on Computational Science and Computational Intelligence, 2015

[9] Lokesh Hebbani, "Road Safety Scenario in India Problems & Solutions", 5th Foundation Day Lecture CiSTUP, IISC January 10, 2014

[10] Costabilea. J., Walla, J., Vecovskia, V & Baileya, "The rapid deployment of an effective road safety countermeasure through a smart phone application- The story of Speed Adviser", Proceedings of the Australasian Road Safety Research, Policing & Education Conference November,2014