International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 3 | ISSN : 2456-3307

Traffic Management Using Big Data Analytic Tool

Saranya Krishna , Sharanya K S , Shwetha K S , Dr. Jitendranath Mungara
ISE Department, New Horizon College of Engineering, Bengaluru, Karnataka, India

ABSTRACT

The analysis of the large-scale data of transportation and accidents has many potentials and it can give very useful insights from the hidden relationship of data. Accidents datasets are used to find main causes of traffic accidents which mainly causes traffic causality and congestion. The vehicular causality dataset used to study human behavior effect on causing traffic accidents. This paper uses nine attributes and two classification algorithm to analyze and predict the possibilities of traffic accidents in python coding environment. Python has a large and comprehensive standard library . pandas is a software library written for the PYTHON programming language for data manipulation and analysis. The random forest algorithm and naïve Bayes algorithm predicts the possibilities of outcomes being a sign of any accident. Checking for the accurate results from algorithms rules and policies are made which is submitted to practioners and decision makers for further road safety measures.

Keywords : Big Data, data mining, PYTHON

I. INTRODUCTION

Traffic safety is one of the main priorities of any governments. Traffic management stands as important issue to be considered for better transportation. The after effect of traffic accident has affected people one or the other way, such as the traffic being slowed down due to an accident or collision in the same road, which may lead in to a congestion. Almost every day, people face traffic accident in one way or the other leading congestion in one or more lanes. Traffic data is located at the core of ITS. Leveraging the big traffic data provides a very good platform to develop Intelligent Transport System (ITS). Although, the traffic data is very huge and rich, there is a lack of useful information. Big data ecosystem has the ability to store, manipulate, analyze and mine large accident datasets and can help us in gaining insight to enhance roadway safety and crashes. Traffic safety aims at reducing the risk of killing seriously injuring of passengers using transportation system and reducing risk of damaging vehicles and transportation infrastructure. Maintaining safety requires a regular monitor and check on different transport elements such as drivers, vehicles, roads, traffic signals and deployed safety processes. To avoid traffic accidents and hence improve traffic management system ,this project gives analysis and predictions on certain constraints .The significance of this paper is presented and addressed by [1] large size of past accidents data is mined and analysed to test robustness and effectiveness .[2] the two big datasets considered are accidents dataset which consists of 146322 examples and the casualties dataset which consists of 194477 examples .[3] The traffic accidents dataset helps us to find out main causes of accidents which causes traffic and congestion and casualty dataset helps in analysing human behaviour effect on causing traffic accidents [4] such dataset is provided to python environment that is widely used high-level programming language for general purpose programming. [5]The dataset is converted into training set and testing dataset by python packages and analysed through the classification algorithms [6]python packages such as scikit, panda and numpy are used for analysing dataset, manipulation and computations required.

II. METHODS AND MATERIAL

1. Related Work

Big Vehicular Traffic Data Mining: Towards Accident and Congestion Prevention by Hamzah Al Najada and

Imad Mahgoub[1] using BigData Techniques. They utilized the UK department for transportation traffic datasets repository which consists of past accident dataset and vehicular causality dataset for the year 2014. Accident dataset consists of 1,46,322 records and causality dataset consists of 19,477 records. To find out the cause of accident numbers of classification algorithms are used and their performance is compared. They used two data mining tools such as Weka and H20. Weka could not handle the large size of datasets to do the classification, but it helps to make feature selection for our data. They performed data pre-processing for datasets which helps to clean the data and it will convert data into ARFF format in order to be able to use them on Weka.

For the sake of examining the data quality he performed preliminary tests on the full datasets, we have considered accident severity attributes as the testing value for the two datasets in these preliminary test. We considered five classifiers Naive Bayes C4.5, Random Forest, Adaboost, Bagging [2]. The quality measure used in the preliminary test are Accuray (ACC) and Area Under the ROC curve (AUC). On comparing their performance AUC gives better interpretation to the data results.

The dataset consists of many attributes this does not mean all of them are needed and they give better results. Hence by using feature selection technique we decreased no of attributes, the processing time and increase the prediction accuracy. Therefore used naive Bayes as the classifier to select the feature in our proposed approach. After applying feature selection technique on both the datasets they selected 9 attributes for the accident dataset and 8 for causality dataset. From the preliminary test we had earlier using Weka an H20 data mining tools. Naive Bayes (NB) and decision tree classifier C4.5 they give best results.

They conducted analysis for both the datasets that is accident analysis and human behaviour and impact analysis. This analysis extracted patterns and findings can assist decision makers and practitioners to build the transportation system intelligently and develop new rules. Their analysis also showed that the human behaviour has strong impact on traffic flow and safety decision.

Their feature direction is to utilize the result obtained so far and predicting a traffic in real time to prevent roadway accident and congestion.

2. Proposed System

Our main aim of this project is to analyse the previous accident as well as vehicular causality records. This analysis done from the collected datasets will be given to the decision makers so that they can formulate rules and policies in order to enhance road safety, decrease the number of accident and avoid traffic congestion. Here we have considered two data sets. First the previous accident datasets with nine attribute and the vehicular causality with six attribute. The attributes for accident dataset is number of vehicle, junction details, day of week, speed limit, weather condition, road type, light condition, junction control, urban and ruler areas. the attributes for vehicular causality are: vehicle reference, causality reference, gender causality, age of reference, pedestrian movement, causality severity. We have used Random Forest algorithm and naive Bayes algorithms to test the accuracy of the result.

Random forest algorithm: Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing the multitude of decision trees at training time. A polling process is carried out in each node of the tree polls. Based on the votes of each node probability is calculated. In particular, trees that are grown very deep tend to learn highly irregular patterns: they over fit their training sets, i.e. have low bias, but very high variance.

Naïve Bayes algorithm: Naïve Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes theorem with strong independence assumption between the features. It is used for text categorization, the problem of judging documents as belonging to one category or other with word frequencies as the features. It is highly scalable, requiring a number of variable in a learning problem. The naïve Bayes equation is given by:

Posterior = (prior * likelihood) / evidence

Further we have coded in python using python packages like pandas, numpy, scipy and calcuated result like in which condition there is a higher

probability of accident. The nine attributes and two classification algorithm are used to analyze and predict the possibilities of traffic accidents in python coding environment. One of the most important features of python is its rich set of libraries for data processing and analytics tasks. python is getting more popularity in big data due to its easy to use features which supports big data processing.

Pandas: It is a Python package providing fast, flexible, and expressive data structures designed to make working with "relational" or "labelled" data both easy and intuitive. It aims to be the fundamental high level building block for doing practical, real world data analysis in Python. It has the broader goal of becoming the most powerful and flexible open source data analysis or manipulation tool available in any language. It Ordered and unordered the time series data. The indexing is achieved by using pandas.

Scipy : The SciPy library depends on the NumPy, which provides convenient and fast N-dimensional array manipulation. The SciPy library is built to work with NumPy arrays and provides many user-friendly and efficient numerical routines such as routines for numerical integration and optimization.

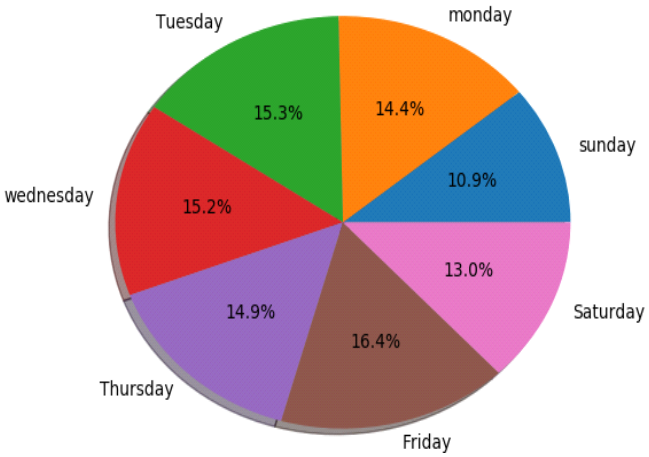
Matplotlib : Matplotlib tries to make easy things easy and hard things possible. You can generate plots, histograms, power spectra, bar charts, error charts, scatter plots, etc., with just a few lines of code. Matplotlib is a plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms.

III. RESULTS AND DISCUSSION

A. Accidents data Analysis

From the analysis of the accident's data and by discovering the hidden relationships, we can extract the main causes of accidents that lead mainly to traffic congestions. The most important features which are strongly relevant and influential to the prediction results are shown in Fig 1. According to fig 1 prediction most of the accident occur during weekend.

- 64.92% of all the accidents happen on residential areas where the speed limit is 30 kmh. Most of the accidents don't happen on highways as most people think.
- 65.8% of all the accidents occur in Urban area. That means most of the accidents are occurring in urban area
- In uncontrolled junction or a give way junction 50% of the accidents take place there
- People would think that most of the accidents take place in bad weather, especially when there is snow or fog. Our analysis shows that 81% of the accidents happens when the weather is fine and has no winds. Taking into consideration that the data we used belongs to UK where the bad weather and fog dominate over the year.
- The same issue with the weather condition, people might think that most accidents occur when there is no light but our analysis showed that 73.8% of the total number of accidents happened in daylight .
- On Fridays, when the weekend starts, policies and speed limits should be changed in Urban areas and single carriage ways because Fridays have the highest percentage of accidents of 16.4 %.

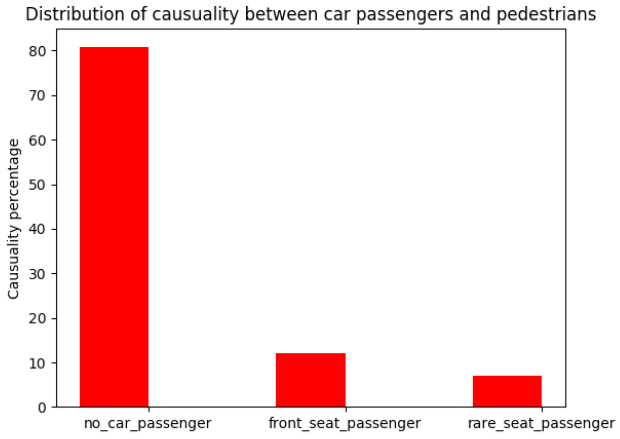


The training accuracy is 86.11 and testing accuracy is 83.69 for random forest . The training accuracy using naïve bayes is 82.83. Fig 1.1 shows that the most of the accidents occur during weekends. Matplotlib is used to produce publication quality figures like histograms,barcharts etc.

B. vehicular causality data Analysis

In this work, we are concerned with mining the driver's data, since human actions motivated by achieving various purposes, cause side effects on our environment, whether they are intentional or not. By studying the human behaviour and impact, we can design new rules to the drivers and the passenger depending on their age, sex, type of passenger and many more individual or combined characteristics. We can also monitor and evaluate the impact of human

activities on the environments that surround us, especially on the roads to achieve the highest safety. Big Data mining would be a good solution for such a problem because human road elements (i.e. pedestrians and drivers) are heterogeneous and inconsistent, thus, no mathematical model can fit this problem. Previous studies in the domain of traffic flow and incidents, have studied, modelled, and simulated human behaviour and impacts by using mathematical modelling techniques . None has studied this behaviour in traffic flow and incidents using historical data mining. Mathematical models can achieve good results, but the issue will still be challenging because the heterogeneous nature of human behavior could not be generalized mathematically. Moreover, the interaction between the human road elements would be random, as well as the large amount of possibilities that would emerge from mathematical models. Big Data mining could assist in studying the human behavior, impact, and interaction, by knowledge discovery from historical data, and this can be easily applied to future cases.



The percentage is 59.2% male to 40.8 % female casualties. Furthermore, fatal accidents happen to male drivers and passengers more than female. Figure 8 shows the distribution for fatal and slight casualties over the sex of the casualty. This piece of info can assist insurance companies in certain areas, depending on studying drivers' behavior to design different policies for male and female. The same thing could be utilized by examining different factors, all of these depend on the content of the analyzed data. It is important to take into consideration the location of humans, because location is a key indicator of human mobility .Fig 1.2 gives the information that most accidents are caused by no car passenger.

IV.CONCLUSION

Thousands of people die in traffic crashes yearly. People lose their lives every day and more people are injured every hour. Fortunately, the existence of the Big Data of traffic crashes, as well as the availability of Big Data analytics tools can help us gain useful insights to enhance the road safety and decrease traffic crashes. In our study we use PYTHON environment to evaluate two classifiers on two big workbench datasets. The used classifiers are: Naive Bayes, Random Forest. From our experiments Naive Bayes gave the optimum results, with the lowest computation time . Our analysis and the extracted patterns and findings can assist decision makers and practitioners to enhance the transportation system intelligently and develop new rules. This study revealed some common misconceptions about road incidents. Our analysis showed that the human behaviour has strong impact on the traffic flow and safety decisions. Our results revealed that driver's attributes such as age and sex could be predicted correctly up to 70% by providing other attributes for an accident or casualty.

V. FUTURE WORK

- Our future directions is to use the results obtained so far in developing a traffic real-time mining system to prevent roadway accidents and congestions.
- There is a need to design an intelligent traffic cloud by making use of cloud computing to solve the problems related to real-time.
- Usage of PLCs and SCADA system in intelligent transportation system for smooth traffic flow can be considered for further work on traffic issue.
- Designing a promising traffic management system to provide smooth traffic flow in non-recursive congestion situation can be an interesting issue for future research.

VI. REFERENCES

[1]. M. Chong, A. Abraham, and M. Paprzycki, "Traffic accident data mining using machine learning paradigms," in Fourth International Conference on Intelligent Systems Design and Applications (ISDA'04), Hungary, ISBN, vol. 1047219710, 2004, pp. 415-420

[2]. I. H. Witten and E. Frank, Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2005

[3]. S. Krishnaveni and M. Hemalatha, "A perspective analysis of traffic accident using data mining techniques," International Journal of Computer Applications, vol. 23, no. 7, pp. 40-48, 2011.

[4]. M. OpenCourseWare, R for machine learning, 2015. [Online]. Available: http://ocw.mit.edu/courses/sloan-school-of-management/ 15-097-prediction-machinelearning-and-statistics-spring-2012/ lecture-notes/a

[5]. H. Ibrahim and B. H. Far, "Data-oriented intelligent transportation systems," in Information Reuse and Integration (IRI), 2014 IEEE 15th International Conference on. IEEE, 2014, pp. 322-329.

[6]. Nejdet DOGRU, Abdulhamit SUBASI, "Comparision of clustering techniques for accident detection", Turkish Journal of Computer Sciences, 2015.

[7]. F Guo, R Krishnan and J W Polak, Short-term traffic prediction under normal and incident conditions using singular spectrum analysis and the k-nearest neighbour method, Centre for Transport Studies, Imperial College London, London SW7 2AZ, fangce.guo07@imperial.ac.uk

[8]. Abdulrahman Abdullah Alkandari, Meshari Aljandal, Theory of Dynamic Fuzzy Logic Traffic Light Integrated System with Accident Detection and Action,@2015 IEEE