ANALYSIS FOR STATUS OF THE ROAD ACCIDENT OCCURANCE AND DETERMINATION OF THE RISK OF ACCIDENT BY MACHINE LEARNING IN ISTANBUL

Halil İbrahim BÜLBÜL
Gazi University, Faculty Education, Department of Computer and Instructional Technologies
Ankara, TURKEY
bhalil@gazi.edu.tr

Tarık KAYA
Gazi University, The Institute of Informatic, Department of Computer Education
Ankara, TURKEY
tarikaya@gmail.com

Yusuf TULGAR
Netdatasoft Software Company
Ankara, TURKEY

Abstract— **The traffic has been transformed into the difficult structure in points of designing and managing by the reason of increasing number of vehicle. This situation has discovered road accidents problem, influenced public health and country economy and done the studies on solution of the problem. Large calibrated data agglomerations have increased by the reasons of the technological improvements and data storage with low cost. Arising the need of accession to information from this large calibrated data obtained the corner stone of the data mining. In this study, assignment of the most compatible machine learning classification techniques for road accidents estimation by data mining has been intended.**

Keywords- Machine learning, data mining, classification techniques, road accident.

I. INTRODUCTION

One of the most complicated and difficult daily needs is overland transportation. Turkey has above 18 million vehicles recorded on Turkish National Police. Every year over 1 million vehicles are added to traffic averagely. 1.2 million people have died and over 50 million people have been injured in road accidents in the world every year. Studies on traffic have executed that road accidents and death-laceration ratio will increase. Design and control of traffic by advanced systems come in view as the important need [1].

Assumptions on the risks in traffic and the regulations and interventions in the end of these assumptions will reduce the road accidents. An assumption system which will be prepared with available data and new risks will be advantageous.

Data mining concept had been come up with by increasing and storage of data in the digital stage. Data mining involves the studies which will discover information from systematic and purposeful data structures obtained from disordered and meaningless data.

Machine learning which is sub-branch of artificial intelligence supplies learning of computer taking advantage of data warehouses. Assumption abilities of computer systems have advanced in the event of machine learning. Utilization of machine learning is a widespread and functional method for taking authentic decisions by using experience. Machine learning is able to attain extract information from data and use statistical methods.

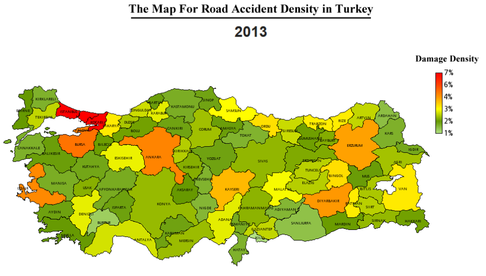
The studies on the analysis of road accident with tree-based models frequency on autobahn [2], classification of road accident with regression method [3], the connection between road accident and structure of the road, weather conditions [4], classification of road accidents data with grouping method [5] had a part in the literature for classification of road accidents.

In this study, machine learning algorithm, which is much acceptable for road accident assumption by the data mining method attaining from road accident data was studied to be found.

II.EVALUATION OF ROAD ACCIDENTS

Developed countries have constituted road accidents information data base to analyze and estimate of road accidents and their consequences. Also, in our country Traffic Insurance Information Center (TRAMER), founded in 2003, was incorporated with Insurance Information and Supervision Center (SBM) in 2008 for all insurance data. As a sub-center of SBM, TRAMER studies on storage of data and constitution of statistical results for road accidents.

It is observed that 854.000 road accidents happened in general of the country in 2013 on the study by TRAMER. %34 of these accidents happened in Istanbul. As seen in the sample of



978-1-5090-6167-9/16 $31.00 © 2016 IEEE
DOI 10.1109/ICMLA.2016.185

Istanbul that is the highest ratio for the population and vehicle in traffic, the rate of accident increases proportional with the ratio of traffic denseness and the vehicle number in traffic [6].

Figure 1 : The Map For Road Accident Density in Turkey

Istanbul where has high density of accident has been considered according to above results in this study. 500 road accident data, that was chosen from the Accident Report in SBM database randomly in 2013. The information, out of personality data, about the vehicle type, the location and the time of accident on the Accident Report was taken into consideration. The raw database for the Accident Report was constituted by reaching the information about temperature and rainfall in the moment of accident from the record of Meteorological Service together with the above mentioned data.

III. DATA MINING AND MACHINE LEARNING

Data mining has been carried out in the fields of retail selling, insurance, health, and banking which comprise from wide data warehouses. For example; to identify purchasing habits of customers, to determine risk management and swindling, to anticipate disease risk can be denoted.

Data warehouses are; theme focused: data for the same conditions and entities is connected with each other, are; integrated: data and data base more than one are knocked together, concerning to the specific time, are; unchangeable: data in the data warehouses are indelible and new data cannot put into data warehouses [7].

Data for data mining must be prepared as in following stages:

- Data collecting
- Data cleaning
- Data integrating
- Data decreasing
- Data transforming [8].

Data cleaning is the stage that remove missing and noise. Also incorrect and extreme data can be removed in this stage. To regain the missing data, the entry for missing data can be taken out, can be completed manually, the same data can be entered, and the missing data can be estimated by regression method.

Data need to restructure according to model and algorithm in the study. Algorithm has only 0 and 1 data as well as other digital values. To be transformed any data to acceptable data structure according to the algorithm, to be studied, is called as data transform stage [9].

Machine learning is a branch of artificial intelligence which maintains that digital systems constitute models and estimate the results which will occur in the new conditions by using the data has been gained.

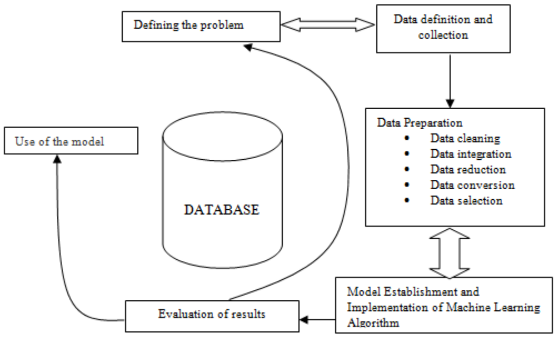


Figure 2 : Machine Learning Process [13]

IV. DATA MINING MODELS

Data mining models can be classified as in followings [10]:

- Value estimation (classifying)
- Clustering
- Association rule

A specified condition has been estimated by using available data in the value estimation model.

The clustering model has been supported by forming specified groups according to similarity of data. Clustering algorithms has been taken into consideration in two groups as hierarchical clustering and non-hierarchical clustering.

Association rule model has been supported by the principle that the rules of forming conditions have been specified all together. It is used for defining purchase trend of customers in retail shopping extensively.

The classification method, above mentioned, will be used for the analysis of the conditions for road accident occurrence and defining accident risks, in this study. To find the algorithm, which reach the best result, will be tried by controlling the data with different algorithms [10].

V. ALGORITHMS FOR DATA MINING CLASIFICATION

a- Decision Tree Classification Algorithms; provide to classification of a tree typed condition according to answers. The algorithm consists of two stages which are occurrence of tree and trimming of branches resulted for error with noisy data. It is important to determine which characteristics will be close with the root and how the data will be divided. it is aimed to find correct and small nodal at the end of this stage. The algorithms of decision tree which is used at most are CART (Simple Cart) and C4.5 (J48).

b- Statistical Classification Algorithms: the data is classified according to their types and used in probability calculus. This type of algorithm estimates which identified categories the new data is concerned. In literature, Bayesyen (Naive Bayes), OneR and Regression algorithms are used vast in statistics.

c- Classification Algorithm for Distance: analyze distance and similarities of each data from one another. The class of new data is estimated by being chosen the closest neighbor in specified number in sample cluster. K closest neighboralgorithm (IBk) is the type of algorithm which is used at most [12].

VI. APPLYING OF MACHINE LEARNING ALGORITHMS

1. Identification of the problem

The studies on reasons and measures of the road accidents are carried out. Classical methods such as raising the level of control, knowledge and consciousness and manufacturing the secured vehicles have been used for getting under control of road accidents. It is seen that using the warnings related to road accidents will be benefitted to control the accidents by pre-estimate in order to increase effects above mentioned methods which are troublesome and costly.

2. Obtaining of Data

According to statistical data in Turkey 37 road accidents of per cent have been happened in Istanbul. The study by SBM shows that sample is sufficient. In this study 500 accident sample was chosen from the data for road accidents in Istanbul. The data in this study was been taken from the official reports related to the road accidents recorded in TRAMER which is only foundation that connect and preserve the reliable data. The information about weather conditions was obtained from Turkish State Meteorological Service according to the information about location and time for the accident obtaining from TRAMER records. Vehicle type, time zone and weather conditions for the accident were used in the evaluation process.

3. Data Preprocessing

The first step, required to pay attention, is data cleaning in the study on data preprocessing. In this step, the missing data have been completed. It is required that the sample data cluster has not been had missing and false information. Also, the data for no accident class is presented in order to be done classification and estimation. The data for the moments had no accident is endless in real life. The class for missing data was completed by using data for the accident and data cleaning methods. The random method was chosen in completion of data. The data obtained with this method was provided not to be in the class had an accident. The records of the situations had no accident was produced with this method. In the finally, 500 data for no accident class was produced in contrast to 500 data for had an accident class. Totally, a data set of 1000 records was produced.

The data conversion, which is a procedure that data converse to acceptable format for using in classifying algorithm, was made after data cleaning in the cluster. The structure of new data cluster after the data conversion has been seen on Table 1.

Qualities		
Before Data Conversion		After Data Conversion
Vehicle1	Car, Truck, Van, Minibus, Bus, Shared ;Taxi, Taxi, Tanker, Tractor	Car (yes, no) = (1, 0)
Vehicle2	Car, Truck, Van, Minibus, Bus, Shared ;Taxi, Taxi, Tanker, Tractor	Car (yes, no) = (1,0)
Date and Time	Date and Time	Time (0-6), (6-12), (12-18), (18-24)= (0,1,2,3)
Temperature	Centigrade Degree	(<-5), (-5 - +5), (+5-+15), (+15-+25), (>25)= (0,1,2,3,4)
Falling	Yes, No	(1,0)
Accident	Yes, No	(1,0)

Table 1: Data Preprocessing

4. Modeling

After data preprocessing, modeling is made. AdaBoost, CART, C4.5, Naive Bayes, OneR, IBk, which are preferred at most among classifying algorithms, will be used. The model which has highest accuracy may be chosen according to the results of these algorithms. WEKA open source code software was used in order to be discovered the results of data by testing with algorithms. WEKA, which is used in academic researches, in the areas of education and industry, has been build up for data analyses and forecasting modeling. WEKA may be used with an interface which shows algorithm and tools as visual [12]. Data preprocessing is made by filter packages in Weka structure. Resample filter is used for forming different samples of data cluster. It has been tried to be found the best results with algorithms on different samples by applying resample filter of data cluster [14].

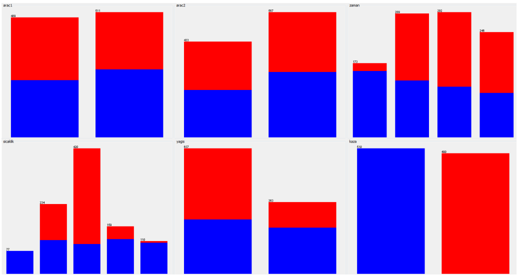


Figure 3 : Distribution of Data Set According to The Accident Classification

In the modeling step, ROC Area value, Precision, Recall, F-Criterion, Kappa Statistics and Performance Percentage (Accuracy) may be analyzed as performance criterions of algorithms. Kappa statistic between 0,6 and 0,8 indicates good agreement, whereas a kappa between 0,8 and 1 indicates perfect agreement. The value of ROC Area and F-Criterion that is calculated according to the sensitivity and certainty expected nearly 1. The percentage of performance of 80% and above indicates that chosen algorithm is rhythmic with the data set [15].

The algorithms for specified targets and comparisons of these algorithms accuracies have been shown in Table 2.

	Classification Algorithms					
	AdaBoost	**CART**	**C 4.5**	**Naive Bayes**	**OneR (Regression)**	**IBk(Knn, K=1)**
The Number of True Samples	782	815	810	802	730	813
The Number of False Samples	218	185	190	198	270	187
ROC Area value	0.841	0.846	0.848	0.872	0.733	0.860
Precision	0.788	0.817	0.812	0.804	0.751	0.816
Recall	0.782	0.815	0.810	0.802	0.730	0.813
F- Criterion	0.781	0.815	0.810	0.802	0.725	0.813
Kappa Statistic	0.565	0.630	0.620	0.605	0.463	0.627
Performance Percentage	78.2	81.5	81	80.2	73	81.3

Table 2: Algorithm Results

IV. CONCLUSIONS AND SUGGESTIONS

Road accidents and losses in specific times and locations are the open sores of developing societies. The controlling and arranging of traffic with advance systems have arisen as an important requirement in parallel with increasing the road accidents. Even simple precautions and natural and coincidental warnings prevent the road accidents. It is acknowledged that the estimation of risks and the arranging and interventions to be made as the results of the risks will diminish these accidents.

The use of machine learning is a functional and extensive method in order to manage to be taken accurate decisions with the experience. Evaluations and estimations with machine learning will bring scientific approach and opinion to the problem. Authentic accidents data attained regularly may be used with machine learning. The required precautions against potential accidents may be taken by reckoning the risks for road accidents.

When we have analyzed the models in the end of this study, the best results have been taken from CART, Ibk, C4,5 and Naïve Bayes algorithms, respectively, according to the performance percentage, Kappa statistic and F-criterion. In ROC Area value, the best results have been taken from Naive Bayes, Ibk, C4,5 and Cart algorithms, respectively. Naive Bayes CART and Ibk algorithms, which give the best three results in the model practice, may be used in the studies on road accidents. These algorithms may be preferred for the improvement of the accident estimate system. While ROC Area value gives in Naïve Bayes of 0,872, CART of 0,846 and Ibk of 0,860, the percentage of performance gives the results in Naïve Bayes of 80,2%, CART 81,5% and Ibk of 81,3%. Also these results supports that these algorithms are efficient for the estimates of road accidents.

The web, mobile and desktop software may be prepared for the accident risk estimate by using these algorithms. The support for more confident traffic ambiance may be given by providing these programs to use of personal and institutional.

REFERENCES

[1] Peden, M. (2004). World report on road traffic injury prevention. Geneva: World Health Organization

[2] Chang, L. Y., & Chen, W. C. (2005). Data mining of tree-based models to analyze freeway accident frequency. Journal of Safety Research, 36(4), 365-375.

[3] Tesema, T. B., Abraham, A., & Grosan, C. (2005). Rule mining and classification of road traffic accidents using adaptive regression trees. International Journal of Simulation, 6(10), 80-94.

[4] Maze, T. H., Agarwai, M., & Burchett, G. (2006). Whether weather matters to traffic demand, traffic safety, and traffic operations and flow. Transportation research record: Journal of the transportation research board, 1948(1), 170-176.

[5] Murat, Y. Ş., & Şekerler, A. (2009). Trafik Kaza Verilerinin Kümelenme Analizi Yöntemi ile Modellenmesi. İMO, Teknik Dergi, 4759-4777.

[6] http://www.sbm.org.tr/tr/haberler/sayfalar/kaza-say%C4%B1s%C4%B1-azal%C4%B1yor!.aspx

[7] Inmon, W. H. (2005). Building the data warehouse. John wiley & sons.

[8] Zhang, S., Zhang, C., & Yang, Q. (2003). Data preparation for data mining. Applied Artificial Intelligence, 17(5-6), 375-381.

[9] Silahtaroğlu, G. (2008) Data Mining With Concepts And Algorithms, İstanbul**:** Papatya Publishing.

[10] Ozkan, Y. (2008). Data Mining Methods. İstanbul: Papatya Publishing.

[11] Quinlan, J. R. (1986). Induction of decision trees. Machine learning, 1(1), 81-106.

[12] Tapkan, P., Özbakır, L., & Baykasoğlu, A. Weka İle Veri Madenciliği Süreci ve Örnek Uygulama.

[13] Bulbul, H. I., & Unsal, O. (2011, December). Comparison of Classification Techniques used in Machine Learning as Applied on Vocational Guidance Data. In Machine Learning and Applications and Workshops (ICMLA), 2011 10th International Conference on (Vol. 2, pp. 298-301). IEEE.

[14] Bouckaert, R. R., Frank, E., Hall, M., Kirkby, R., Reutemann, P., Seewald, A., & Scuse, D. (2013). Weka Manual For Version 3-7-8.

[15] Göker, H., "Üniversite Giriş Sınavında Öğrencilerin Başarılarının Veri Madenciliği Yöntemleri İle Tahmin Edilmesi", Yüksek Lisans Tezi, Gazi Üniversitesi Bilişim Enstitüsü, Ankara, 2012.