The Use of Data Mining Techniques in Analysing Traffic Accidents (An application on Khartoum State)

Mozamel M. Saeed

Department of Computer Science, Prince Sattam Bin Abdulaziz University, Al-Kharj, Saudi Arabia

Abstract

This paper presents a sample of mining algorithms in data represented in "One R, J48, Naïve Bayesian" to know its optimal which pertains to analysis of traffic accidents in Khartoum State occurring through years 2007 – 2016.It is important to note that 389931 record was analysed by the statistical reports structure to reach the mining stage in data in order to creating a mechanism that is capable of studying the elements which smartly play a significant part in traffic accidents for connection. The range of relation designation between them, and its significance in traffic accidents percentage is implemented on Weka program to apply algorithms in data and ,accordingly, the presentation of the results together with analysis since the results showed that the performance of J48 algorithm generally is of more qualifications and surpasses than the other algorithms in accidents data group. It spent 0.02 seconds and the rate of error in the sample was 0.02 through its implementation and assisted in the prediction of data.

The paper concludes with the implementation of J48 classification algorithms for the production of the decision tree through Weka to the point of the existence of classification of cars' accidents which occurred according to time and harm. It also shows that the harm damage rate is of the highest reaching 301394 at the rate of 77.3%, and that 2012 and 2011 accidents were the highest of all years at the rate of 11.3%, and the rate of the lowest accidents in 2007 was 7.4%.

Keywords — *Traffic Accidents, Data Mining, Khartoum State, Data Mining Algorithms, classification.*

I. INTRODUCTION

Since the inception of Sudan, it has witnessed a civilized uprising that was comprehensive and fast and hasn't been restricted to one field only. This resulted in the increase and expansion of the urban centers. The comprehensive development through which the society passed, basically the economic one, assisted in the creation of a net of roads connecting all urban and rural regions.

This development has also participated in the increase of the citizens' incomes which helped them to own a car or more. The transference movement caused by vehicles, the most utilized one in the Sudanese society, resulted in the up rise of traffic accidents which greatly harms the society socially and economically.

In accordance with World Health Organization (WHO), the world annually witnesses an estimation of 1.25 million of death due to traffic accident. As the report shows the youth represents 48% of traffic accidents death, and that the male gender is the most dominant victims of traffic accidents, representing 73% of total death of road accidents. Approximately 50 million of others are exposed to non-fatal wounds. As a result, many of them are disabled, which gives road safety hazard a priority to war, AIDS, Malaria and the other violence incidences as a global health problem. Besides, traffic accidents wounds cause grate economic losses of personnel, their families and the whole country. These losses result in medication costs, minimization of personnel production who die or are exposed to disability due to wounds, and the family members who are obligated to be absent from work or school for the care of the injured ones. In most countries traffic accidents cost 3% of the total local production.[1],[2]

II. LITERATURE SURVEY

Globally, the developing countries record the highest rate of traffic accidents which cover all kinds of vehicles and expose man's life to death and wounds caused by negligence of traffic rules and regulations. Data mining techniques of the latest technology could be used to resolve this issue. Many professional researchers from different places in the world have made their contribution with regard to this field. Where as in Sudan, such researches are rare and hardly observed.

Sachin Kumar et.al. [3], suggested the application of k-means algorithm and ARM technique to resolve traffic accident complicated problems. The researcher divided the different accidental prone location in three different categories that were high, moderate and low frequency, to extract the hidden information behind the data set and adopt some preventive procedures according to accident location.

Ehab [4], Using exploration algorithms to analyse traffic accidents in Syria that aimed, relatively , to utilize updated techniques in databases field for the analysis of traffic accidents and supply decision makers with indicators. In data mining field, the researcher discovered that ,according to the tests , the best method of CART algorithm is the values classification of traffic accident statistics, with the presence of better results for GRI algorithm than the previous one and thus revealing additional essential rules of high reliability .The study recommended the establishment of a national center for the support of the decision.

Sachin et.al. [5], proposed setting a framework to control traffic accident in India where its scored(11,574) during 2009 - 2014 by the use of K-modes clustering technique and association rule mining. The analysis of the results, using a combination of this technique, concluded with his concept that the result will be more efficient if no segmentation is performed.

M. Sowmya [6], showed the work on traffic accident data produced by Hong Kong government transport department in 2008. That study applied Naive Bayes, J48, AdaBoostM1, PART and Random Forest classifiers for predicting classification accuracy to analyse the performance. The classification accuracy on the test results revealed the three cases as accidents, vehicles and casualties.
Shanti et.al. [7], used classification algorithm for vehicle's collision patterns.

The classification algorithms applied to this data set were C4.5, C-RT, CS-MC4, Decision List, ID3, Naïve Bayes and RndTree. The results achieved proved that RndTree techniques were more better and accurate than the other algorithms in collision cases whose fatality rate increased in road accidents.

III. METHODS

A. *Data Mining*

This rapid growth, huge data capacity gathered and the great saved various databases are an efficient tool to assist in benefiting of data collected through identification of objective and useful information. Data mining is considered to be one of the solutions to analyse a huge capacity of data and change it in to positive information and knowledge. Data mining is defined as the extraction or mining knowledge from a huge capacity of data. Some other terms might have a typical or slightly different meaning, like data knowledge mining from data, knowledge extraction, data or pattern analysis. The use of data mining functions is to specify the type of patterns found in data mining tasks. Generally, it is classified into two classes: descriptive and predictive. The descriptive data mining tasks characterizes data general feature in the database, where the predictive one practices based on the current data to make predictions. Data mining or generally known as knowledge discovery in database means extracting or "mining" knowledge from huge capacity of data. [8]

B. *Algorithms Used in the Study*

There are three basic techniques in which algorithms for data mining is based:

- Classification: Assigning entities to one or more of the classification categories such as classification of spam messages by message title and content.

- Clustering: Dividing data into clusters with a meaning, such as classifying a set of traffic accidents by the temporal or spatial time of the incident.

- Rules of association: The discovery of an important relationship hidden in a large data set and represented by intensive relationships in the form of coupling rules.

The work in this paper is deal with, through three
different types of bisection algorithms. We will try to
review the advantages of each one separately and as
follows:

1) Naive Bayesian Classifier

A Naive Bayesian classifier is a simple probabilistic classifier based on applying Bayesian theorem (from Bayesian statistics) with strong (Naive) independence assumptions.

Benefits

- It is fast, highly scalable model in structure and scoring.
- Scales linearly with the number of predictors and rows.
- Establishment of the process for Naive Bayes as parallelized.
- Induced classifiers are easy to interpret and robust to irrelevant attributes.
- Uses evidence from many attributes. The Naïve Bayes can be used for both binary and multiclass classification problems.

2) J48 Decision Tree Classifier

The J48 workbook is a simple C4.5 decision tree. Working on creating a tree-bi-tree approach to decision-making is the most useful classification of the problem. This technique is established as a model tree for the classification process. Once the tree is established, it is applied to all rows () in the database and displays the results in the classifier for them.

Benefits

- Gains a balance of flexibility and accuracy.
- Limits the number of possible decision points.
- Distinguished by high accuracy.

3) One R

A classification algorithm that generates a single rule for each indicator in the data, and then chooses a smaller error base than the sum of the errors used to create a prediction base, and to construct a repeating table for each prediction, based on the target.

It has also been shown that it produces less precise rules among the algorithms and is classified in the production of rules that are simple to interpret.

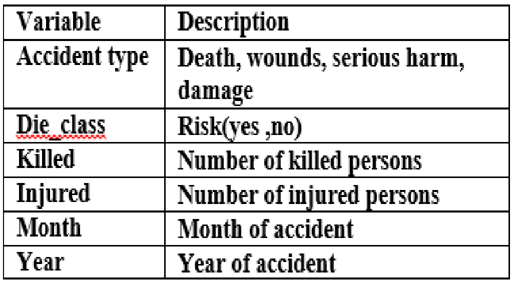
Benefits

- Finds the most common layer.
- Sets the category to this value for prediction.
- Calculates the total error of the rules of each prediction.
- Chooses a prediction with the minimum overall error.

IV. DATASET COLLECTION

This study used datasets which are produced by Khartoum state traffic police [9]. They are also intended to be a nationally representative probability sample from the annual estimation of 1.6 million accident reports in Khartoum state. The dataset for the study includes traffic accident records for 2007 - 2016, a total number of 389931 record.. According to the variable definitions for dataset, this dataset contains only the records of accidents programs, and doesn't include the passengers' information. It also covers labels which are referred to in table (1).

TABLE 1. VARIABLE DEFINITIONS USED IN DATASET



Variable	Description
Accident type	Death, wounds, serious harm, damage
Die_class	Risk(yes ,no)
Killed	Number of killed persons
Injured	Number of injured persons
Month	Month of accident
Year	Year of accident

A. *Data Preparation and Process*

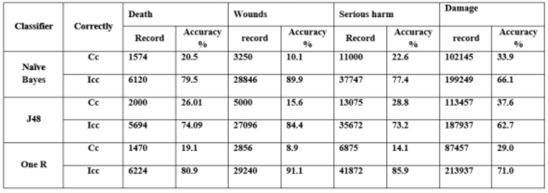
This work deals with the performance of three classification algorithms, which are "Naive Bayesian, J48 and One R ". Khartoum Province traffic department conducted the production of datasets for 2007 – 2016, which is recorded in two different scenarios:

- Accident damage.
- Accident time.

1) *Accidents Damage*

The total records set used is 389931recod, which includes severe death, wounds, serious harm and damage. By the use of these algorithms, it can be classified correctly or incorrectly for all attributes. These attributes are described in table (2).

TABLE 2. CORRECTLY (CC) AND INCORRECTLY (ICC) ACCIDENT DATASET CLASSIFICATION



Classifier	Correctly	Death		Wounds		Serious harm		Damage	
		Record	Accuracy %	record	Accuracy %	Record	Accuracy %	record	Accuracy %
Naïve Bayes	Cc	1574	20.5	3250	10.1	11000	22.6	102145	33.9
	Icc	6120	79.5	28846	89.9	37747	77.4	199249	66.1
J48	Cc	2000	26.01	5000	15.6	13075	28.8	113457	37.6
	Icc	5694	74.09	27096	84.4	35672	73.2	187937	62.7
One R	Cc	1470	19.1	2856	8.9	6875	14.1	87457	29.0
	Icc	6224	80.9	29240	91.1	41872	85.9	213937	71.0

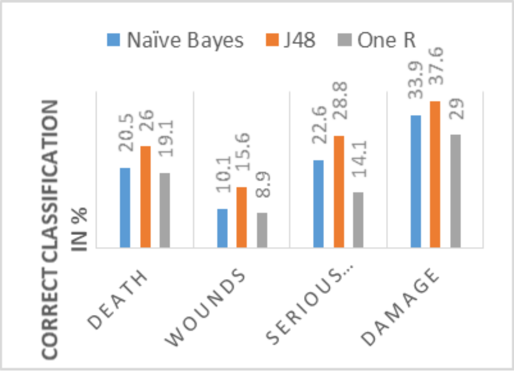


Figure 1. Comparison of Classifiers Naive Bayes, J48, One R Accidents Datasets

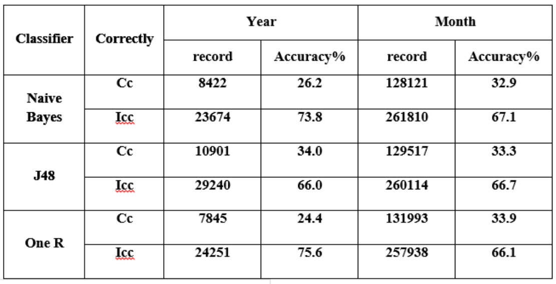
Figure (1) shows accidents datasets graph which represents 20.5% of death for Naïve Bayes, 20.83% for J48, 11.95% for One R, wounds 10.1% for Naïve Bayes, 15.6% for J48, 8.9% for One R, serious harm 22.6% for Naïve Bayes, 28.8% for J48, 14.1% for One R and damage 33.9% for Naïve Bayes, 37.6% for J48, 29% for One R.

It is noted that J48 classification algorithm represents the highest percentage compared with the other classification algorithms. Based on that, J48 algorithm surpasses all the other algorithms performance in the accident dataset.

2) *Accident Time*

The total records set used is 389931recod which includes the attributes year and month. These attributes are described in table (3).

TABLE 3. ACCIDENT TIME CLASSIFICATION



Classifier	Correctly	Year		Month	
		record	Accuracy%	record	Accuracy%
Naive Bayes	Cc	8422	26.2	128121	32.9
	Icc	23674	73.8	261810	67.1
J48	Cc	10901	34.0	129517	33.3
	Icc	29240	66.0	260114	66.7
One R	Cc	7845	24.4	131993	33.9
	Icc	24251	75.6	257938	66.1

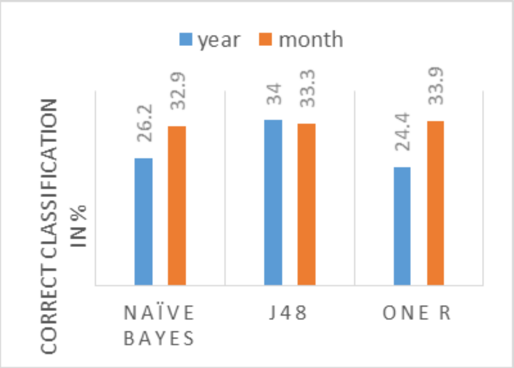


Figure 2. Comparison of Classifiers Naive Bayes, J48, One R Accident Time Datasets

Figure (2) shows accident time dataset graph which represents 26.2% for Naïve Bayes, 34% for J48, 24.4% for One R per year, based on this J48 algorithm reflects the highest percentage when compared with the other classification algorithms. Pertaining to the month, it represents 32.9% for Naïve Bayes, 33.3% for J48, 33.9% for One R , thus there isn't any significant difference between Naive Bayes, J48 and One R.

The analysis process above shows that the performance of J48, generally surpasses the other algorithms in classification of dataset rapidly, makes its implementation necessary as it will be shown later.

V. APPLICATION OF ALGORITHM J48

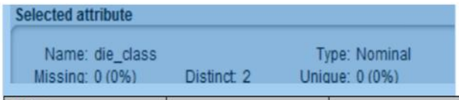
Implementation process of the algorithm will be conducted through classification and prediction as follows:

A. Classification

1) Classification According to Damage Caused by Accidents:

Traffic accidents which causes damage recorded the highest percentage at 77.3%, while 12.5% was due to serious harm, 8.2% results from wounds, and 2% of total accidents was death.

TABLE 4. CLASSIFICATION OF DAMAGE CAUSED BY ACCIDENTS



Selected attribute

Name: die_class Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

Effect	count	%
Death	7694	2.0
wounds	32096	8.2
Serious harm	48747	12.5
Damage	301394	77.3

2) Classification by year:

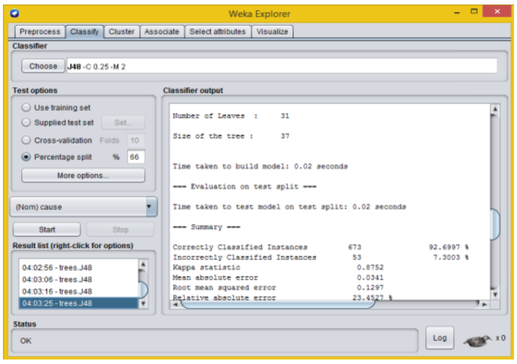
The highest percentage of accidents in 2012 and 2011 was 11.3%, while the lowest percentage of accidents was 7.4% in 2007. Table (5) clarifies this:

TABLE 5. CLASSIFICATION ACCIDENTS BY YEAR

Year	Count	%
2007	28734	7.4
2008	34289	8.8
2009	37712	9.7
2010	39046	10.0
2011	44115	11.3
2012	44254	11.3
2013	41686	10.7
2014	41155	10.6
2015	40202	10.3
2016	38738	9.9

B. The Prediction

After entering the data that was specified as factors influencing the prediction of the J48 algorithm, it was predicted by a percentage of 92.7% (due to the huge capacity of data), and was wrongly predicted at 7.3%. The number of nodes in the tree was 31, and the total size was 37. The implementation of this process took 0.02 seconds, while the error rate in the model was 0.02. Figure (3) clarifies this:



Classifier: J48 -C 0.25 -M 2

Test options:
○ Use training set
○ Supplied test set
○ Cross-validation Folds 10
● Percentage split % 66

(Nom) cause

Result list (right-click for options)
04:02:56 - trees.J48
04:03:06 - trees.J48
04:03:16 - trees.J48
04:03:25 - trees.J48

Classifier output

```
Number of Leaves  :     31

Size of the tree :      37


Time taken to build model: 0.02 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.02 seconds

=== Summary ===

Correctly Classified Instances         673               92.6997 %
Incorrectly Classified Instances        53                7.3003 %
Kappa statistic                          0.8752
Mean absolute error                      0.0341
Root mean squared error                  0.1297
Relative absolute error                 23.4527 %
```

Status: OK

Figure 3. The Implementation of J48 Algorithm to Specify Accident Damage

The results of the implementation of algorithm show that it recorded the best result by studying the linear correlation between the expected values and the observed ones. Figure (4) clarifies this:

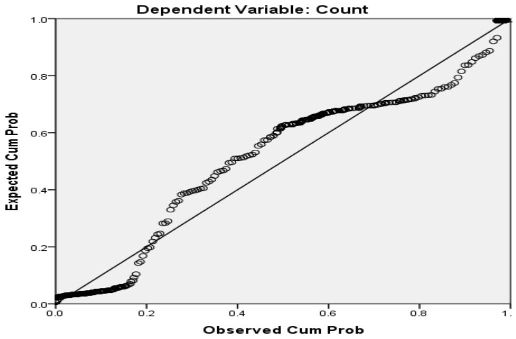


Figure 4. Linear Correlation Between Expected & Observed Values

VI. CONCLUSIONS

This paper reflected one of the most essential issues that is related to man's life, and it represented the traffic accidents, through implementation of the study, in Khartoum State for the period 2007-2016, with 389931 record that covers analysis and sampling. However, confirmation of the conduct of "one R, J48, Naïve Bayesian" algorithms was done for prediction of classification precision and its discovery in the test of both (the harm that results of the accident and its time).

Implementation process showed the surpass of J48 algorithm compared with the others while algorithm proved its qualification much faster in data classification, in addition to the fact that it assisted in data prediction to support decision makers in order to limit traffic accident and maintain the human life.

REFERENCES

[1] World Health Organization, http://www.who.int/ar/news-room/ fact-sheets/detail/road-traffic-wounds

[2] Road Traffic Accident Statistics available at:

http://www.td.gov.hk/en/road_safety/road_traffic_accident_statistics/2016/index.htm.

[3] Sachin Kumar and Durga Toshniwal, "A data mining approach to characterize road accident locations", J. Mod. Transport, 24 (1):62 -72 DOI 10.1007/s40534-016-0095-5, 2016.

[4] Ehab Eldebaja , "The Use of Exploration Algorithms for the Analysis of Traffic Accidents (Syria)" Tishreen University Journal for Research and Scientific Studies - Engineering Sciences Series Vol . (73) No. (2), 2105.

[5] Sachin Kumar and Durga Toshniwal, "A data mining framework to analyse road accident data", Journal of Big Data 2:26 DOI 10.1186/s40537-015-0035-y, 2015.

[6] M. Sowmya and .P. Ponmuthuramalingam, "Analyzing the Road Traffic and Accidents with Classification Techniques", International Journal of Computer Trends and Technology (IJCTT) - volume 5 number 4 - Nov 2013.

[7] S. Shanthi, "Classification of Vehicle Collision Patterns in Road Accidents using Data Mining Algorithms", International Journal of Computer Applications (0975 - 8887) Volume 35 - No.12, December 2011.

[8] Han, Jiawei and Kamber, Micheline.," Data Mining: concepts and Techniques. San Fransisco", Morgan kufman Publisher, 2006.

[9] Khartoum status report on road safety: time for action, 2017.

[10] Pasko Konjevoda and Nikola Stambuk, "Open-Source Tools for Data Mining in Social Science," Theoretical and Methodological Approaches to Social Sciences and Knowledge Management, pp.163-176.