Empirical Comparative Study of Selected Data Mining Classification Algorithms

Olutayo, V.A. PhD
Department of Computer Science
Joseph Ayo Babalola University
Ikeji-Arakeji, Osun State, Nigeria
E-mail – vaolutayo@jabu.edu.ng

ABSTRACT

We noted that, digital data acquisition methods and storage technology have allowed storage of data being stored in different types of databases. With this advancement in database technologies, the need to extract useful information from the database had increased. Achieving this goal requires using the right mining algorithms on the type of the available data in order to have an informed results. A large number of algorithmic (i.e. Id3 Tree, MLP, RBF, FT, Clustering etc.) solutions exist; but until now, little or no empirical research has been done on comparing their efficiency especially on different forms of data. This paper therefore, focusses on the data mining classification algorithms to ascertain which of these data mining classification's algorithmic solutions will scale well or perform better for the different forms of data available in a database. The data were organized into continuous and categorical data. The two forms of data (i.e. continuous and categorical data) were analysed using the selected classification algorithms. Sensitivity analysis was performed and irrelevant inputs were eliminated. The performance measures used to determine the performance of the techniques include Mean Absolute Error (MAE), Confusion Matrix, Accuracy Rate, True Positive, False Positive and Percentage correctly classified instances. Empirical results reveal that, among the machines learning paradigms considered, Id3 Tree approach performed better on categorical data than continuous data while Artificial Neural Networks performed better on continuous data than categorical data with a lower error rate and higher accuracy rate.

Keywords – Data mining, classification algorithms, Categorical and continuous data, Performance comparison

CISDI Journal Reference Format
Olutayo, V.A. (2019): Empirical Comparative Study of Selected Data Mining Classification Algorithms. Computing, Information Systems, Development Informatics & Allied Research Journal. Vol 10 No 4, Pp 13-22. Available online at www.cisdijournal.org. DOI Affix - https://doi.org/ 10.22624/AIMS/CISDI/V10N4P2

1. INTRODUCTION

Analyzing, interpreting and making maximum use of the data is difficult and resource demanding due to the exponential growth of many business, governmental and scientific database [1, 2, 3]. It is estimated that the amount of data stored in the world's database grows every twenty months at a rate of 100%. This fact shows that we are getting more and more exploded by data/information and yet ravenous for knowledge. Data mining therefore appears as a useful tool to address the need for sifting useful information such as hidden patterns from databases [3]. In today's world, where the accumulation of data is increasing in an alarming rate, understanding interesting patterns of data is an important issue to be considered to adjust strategies, to make maximum use of it, and find new opportunities [4]. Organizations keeping data on their domain area takes every record as an opportunity in learning facts. But the simple gathering of data is not enough to get maximum knowledge out of it. Thus, for an effective learning, data from many sources must first be gathered and organized in a consistent and useful manner [5].

Data warehousing allows the enterprise to recognize what it has noticed about its domain area. The data must also be analyzed, understood, and turned into actionable information. This is the point where the application of data mining is needed. Although it is difficult to define precisely and delimit the range and limits of such scientific disciplines, many scholars try to indicate the basic tasks of data mining. In line with this, Hand [1] defines Data mining as the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner.

Data mining can also be seen as a combination of tools, techniques and processes in knowledge discovery. In other words, it uses a variety of tools ranging from classical statistical methods to neural networks and other new techniques originating from machine learning and artificial intelligence in improving database promotion and process optimization. Six basic functions or activities of data mining are classified into directed and undirected data mining [5,6] Specifically classification, estimation and prediction are directed, where the available data is used to build a model that describes one particular variable of interest in terms of the rest of the available data. Affinity grouping or association rules, clustering, description and visualization on the other hand are undirected data mining where the goal is to establish some relationship among all variables.

Up to recent time, the only analysis made on data to get meaning out of it, is simple statistical manipulation that has no power to show all the necessary information content of a given data. But data mining technology, on the other hand has the greatest potential in identifying various interesting patterns for enabling organizations to control data resources for strategic planning and decision-making in their domain area. [7]

2. ANALYSIS OF DATA SET USED

Classification is often seen as the most useful (and lucrative) form of data mining. Although every pattern recognition technique under the sun has been used to do classification (e.g. rough sets, discriminant analysis, logistic regression, multilayer perceptron, decision trees etc.) Therefore, the heart of any data mining classification algorithm is relevant and historical data of the domain in consideration [2,3]. The selection of inputs is the most important aspect of creating a useful prediction, as it represents all of the knowledge that is available to the model to base the prediction on. Also, coding or converting original variable data to one form or the other is very important because, it will help to better analyze such variable data using one of the classification algorithms [8].

However, the purpose of this study, is to run different data mining algorithms which includes, neural networks, decision trees and prisms rule on the various forms of data i.e. (continuous form and categorical form), in order to know the particular algorithm that actually suit for any form of data in terms of reduction in running time as well as a better performance. Variable data set could be of two forms: continuous and categorical. A continuous variable has numeric values such as 0, 1, 2, 3.14, -5, etc. Examples of continuous data are: blood pressure, height, weight, income, age, and probability of illness. On the other hand, Categorical variable data has values that function as labels rather than as numbers. For example, a categorical variable for gender might use the value 1 for male and 2 for female. As another example, marital status might be coded as 1 for single, 2 for married, 3 for divorced and 4 for widowed. This work will unravel whether running these forms of data on classification algorithms will affect their performance rate.

This study therefore, used data of accident records on the first 40 kilometres from Ibadan to Lagos, which were collected from the Nigeria Road Safety Corps. The data sample covered a period of twenty four month. The expected output of the data, is to compare the results of the prediction of the cause of accident and accident prone location on the Lagos – Ibadan express way on the selected classification algorithms. In order to archive this, the data sample were cleaned by filling in missing values; smoothing noisy data, identifying or removing outliers, and resolving inconsistencies.

Finally, the cleaned data are transformed into a format suitable for data mining. The data were then transformed to both continuous and categorical format (the continuous and categorical data are given in Table 2 and Table 3 respectively) after which the selected classification algorithms were later used for the analysis. The data format used is shown in Table 1.

Table 1: Continuous and Categorical Variable Format Used in the Modelling

S/N	Variable Name	Description	Variable Format	Variable Type
1.	Vehicle Type	Small cars	1	Categorical
		Heavy Vehicle	2	Categorical
2.	Time of the day	Morning	1	Categorical
		Afternoon	2	Categorical
		Evening	3	Categorical
		Night / Midnight	4	Categorical
3.	Season	Wet	1	Categorical
		Dry	2	Categorical
4.	Causes	Wrong Overtaking	A	Categorical
		Careless Driving	B	Categorical
		Loss of Control	C	Categorical
		Tyre Bust	D	Categorical
		Over Speeding	E	Categorical
		Obstruction	F	Categorical
		Pushed by another vehicle	G	Categorical
		Broken Shaft	H	Categorical
		Broken Spring	I	Categorical
		Brake Failure	J	Categorical
		Road problem	K	Categorical
		Unknown Causes	L	Categorical
		Robbery Attack	M	Categorical
5.	Spot at which the Accident Occurred	Location	LOC1 LOC2 LOC3	Categorical

In the modelling process, the output variable is the location, critical study of the accident data showed that the locations can be divided into three distinct regions tagged region A, region B and region C, meaning we have three outputs. Where, First location 1 – 10km is Region A or location 1 Between 10km – 20km is region B or Location 2 Above 20km is region C or Location 3.

Table 2: Transformed Data Sample (Continuous)

S/N	TYPE	TIME	SEASON	A	B	C	D	E	F	G	H	I	J	K	L	M	LOC3	LOC2	LOC1
1	2	3	1	0	0	0	1		0	0	0	0	0	0	0	0	1	0	0
2	1	2	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0
3	2	3	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	
4	2	2	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0
5	1	1	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
6	1	3	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
7	2	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
8	1	2	2	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0
9	1	2	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0
10	2	2	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0
11	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0
12	2	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0
13	1	2	2	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
14	1	2	2	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0
15	2	3	2	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0
16	3	2	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0
17	2	2	1	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0
18	1	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
19	2	1	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0
20	2	2	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0
21	2	2	2	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0
22	0	1	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0
23	2	3	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0

Table 3: Showing Categorical Data for Decision Tree Analysis

Type	Time	Season	Wrong-Overtaking	Careless driving	Loss-Of-Control	Tyre burst	Over speeding	Tree-Obstruction	Pushed-By-A-Car	Broken-Shaft	Broken-Spring	Brake-Failure	Road-Problem	Unknown-Causes	Robbery-Attack	Location
Haevy Vehicle	Evening	Wet	False	False	False	True	False	False	False	False	False	False	False	False	False	Location3
Small Car	Afternoon	Wet	False	False	False	False	False	False	False	False	False	False	False	False	True	Location3
Haevy Vehicle	Evening	Wet	False	False	False	True	False	False	False	False	False	False	False	False	False	Location2
Haevy Vehicle	Afternoon	Wet	False	False	False	True	False	False	False	False	False	False	False	False	False	Location2
Small Car	Morning	Dry	False	True	False	False	False	False	False	False	False	False	False	False	False	Location1
Small Car	Evening	Wet	True	False	False	False	False	False	False	False	False	False	False	False	False	Location2
Haevy Vehicle	Morning	Wet	False	True	False	False	False	False	False	False	False	False	False	False	False	Location1
Small Car	Afternoon	Dry	False	False	False	False	True	False	False	False	False	False	False	False	False	Location2
Small Car	Afternoon	Wet	False	False	False	True	False	False	False	False	False	False	False	False	False	Location2
Haevy Vehicle	Afternoon	Wet	False	False	True	False	False	False	False	False	False	False	False	False	False	Location2
Motocycle	Morning	Wet	False	False	False	False	False	False	False	False	False	False	True	False	False	Location2
Haevy Vehicle	Morning	Wet	False	False	False	True	False	False	False	False	False	False	False	False	False	Location2
Small Car	Afternoon	Dry	False	False	False	True	False	False	False	False	False	False	False	False	False	Location1
Small Car	Afternoon	Dry	False	False	False	False	True	False	False	False	False	False	False	False	False	Location2
Haevy Vehicle	Evening	Dry	False	False	False	True	False	False	False	False	False	False	False	False	False	Location2
Haevy Vehicle	Afternoon	Wet	False	False	False	True	False	False	False	False	False	False	False	False	False	Location2
Haevy Vehicle	Afternoon	Wet	False	False	False	True	False	False	False	False	False	False	False	False	False	Location3
Small Car	Morning	Dry	False	False	False	False	False	False	False	False	False	False	False	False	False	Location3
Haevy Vehicle	Morning	Dry	False	False	True	False	False	False	False	False	False	False	False	False	False	Location2
Haevy Vehicle	Afternoon	Wet	False	False	False	True	False	False	False	False	False	False	False	False	False	Location2
Haevy Vehicle	Afternoon	Dry	False	False	False	True	False	False	False	False	False	False	False	False	False	Location2
Small Car	Morning	Dry	False	False	False	False	False	False	False	False	False	False	False	True	False	Location3
Haevy Vehicle	Evening	Wet	False	False	False	True	False	False	False	False	False	False	False	False	False	Location2
Small Car	Evening	Dry	False	False	False	False	False	False	False	False	False	False	False	True	False	Location2
Haevy Vehicle	Morning	Wet	False	False	False	False	False	False	False	False	False	False	False	True	False	Location1
Haevy Vehicle	Morning	Dry	False	False	False	False	False	False	False	False	False	False	False	True	False	Location2
Haevy Vehicle	Morning	Dry	False	False	False	False	False	False	False	False	True	False	False	False	False	Location1
Haevy Vehicle	Afternoon	Dry	False	False	False	False	True	False	False	False	False	False	False	False	False	Location2
Small Car	Morning	Dry	False	False	False	False	True	False	False	False	False	False	False	False	False	Location3

3. METHODS

In this research, the datasets were organized into both categorical data and continuous data. The decision tree was given the categorical data and the artificial neural networks was equally given continuous data, this is to allow the data mining algorithms to run on different data type, so as to determine which performs better. The major step required to obtain result of the research involved analysis of the data using WEKA. WEKA is a collection of machine learning algorithms and data processing tools. It contains various tools for data pre-processing, classification, regression, clustering, association rules and visualization. There are many learning algorithms implemented in WEKA including Bayesian classifier, Trees, Rules, Functions, Lazy classifiers and miscellaneous classifiers. The algorithms can be applied directly to a data set. WEKA is also data mining software developed in JAVA it has a GUI chooser from which any one of the four major WEKA applications can be selected. For the purpose of this study, the Explorer application was used.

4. Experimental Setup, Analysis and Results

Artificial Neural Networks Analysis

In the case of ANN based modelling, the hyperbolic activation function was used in the hidden layer and the logistic activation function otherwise known as sigmoid in the output layer. Models were trained with BP (100 epochs, learning rate 0.01) and SCGA (500 epochs) to minimize the root mean square and mean absolute error. For each output class, both multilayer perception (MLP) and Radical Basis Function Neural Networks (RBF) were used to determine the better networks.

4.1 Radial Basis Function Performance Analysis

The RBF model was experimented with using different number of hidden neurons, and the model with highest classification accuracy for the correctly classified instances was determined. From the result analysis, the RBF model achieved training and testing performance of 54.73% and 40.56% respectively with 0.3478 of mean absolute error.

Table 4: Detailed Accuracy by Class

Class	Roc Area	TP rate	FT rate	Precision	Recall	F- measure
Location (3)	0.716	0.294	0.096	0.476	0.294	0.364
Location (2)	0.517	0.744	0.694	0.598	0.744	0.663
Location (1)	0.568	0.25	0.108	0.35	0.25	0.292
Weighted Avg.	0.572	0.547	0.446	0.523	0.547	0.524

Table 5: Confusion Matrix

	Predicted Category		
Actual Category	Location (3)	Location (2)	Location (1)
Location (3)	10	23	1
Location (2)	10	64	12
Location (1)	1	20	7

4.1 Multilayer Perception Performance Analysis

For the case of MLP model, the model achieved training and testing performance of 78 correctly classified instances representing 52.70% and 28 representing 45.20% with mean absolute error of 0.3479 and root mean square error of 0.5004.

Table 6: Detailed Accuracy by Class

Class	Roc Area	TP rate	FT rate	Precision	Recall	F- measure
Location (3)	0.529	0.158	0.5	0.529	0.514	0.719
Location (2)	0.628	0.581	0.6	0.628	0.614	0.493
Location (1)	0.214	0.133	0.273	0.214	0.24	0.564
Weighted Average	0.527	0.399	0.515	0.527	0.52	0.558

Table 7: Confusion Matrix

	Predicted Category		
Actual Category	Location (3)	Location (2)	Location (1)
Location (3)	18	15	1
Location (2)	17	54	15
Location (1)	1	21	6

A confusion matrix provides detailed information about how data rows are classified by the model. The matrix has a row and column for each category of the target variable (Location). The categories shown in the first Column are the actual categories of the target variable. The categories shown across the top of the table 5 cells are the predicted categories. The numbers in the cells are weights of the data rows with the actual category of the row and the predicted category of the column. The numbers in the diagonal cells are the weights for the correctly classified cases where the actual category matches the predicted category the off-diagonal cells have misclassified row weights. For both RBF and MLP, the confusion matrix showed that the model gave a good performance on location2. Further investigation into the input data especially on the importance of variables revealed that Tyre burst has the highest value of all the sixteen variables, followed by loss of control and over speeding.

4.2 Sensitivity and Specificity Report

The sensitivity and specificity report is used for classification problems where the target variable has two categories. For these types of analyses, one category of the target variable is called the 'positive' category, and the other is called the 'negative' category. These are the parameters used to measure the performance and the accuracy rate of the models.

That is,

TP-represent True positive
FP-represent false positive and
ROC- represent receive operating characteristic curve for the model
ROC is also called the "C statistic".

4.3 Decision Tree Performance Analysis

Several numbers of setups of decision tree algorithms have been experimented and r the best results obtained is reported for my data set.

Each class was trained with entropy of fit measure, the prior class probabilities parameter was set to equal, the stopping option for pruning was misclassification error, the minimum n per node was set to 5, the fraction of objects was 0.05, the maximum number of nodes was 100, surrogates was 5, 10 fold cross-validation was used, and generated comprehensive results. The best decision tree result was obtained with Id3 with 115 correctly classified instances and 33 incorrectly classified instances which represents 77.70% and 22.29% respectively. Mean absolute error was 0.1835 and Root mean squared error was 0.3029.

4.3.1 Decision Tree Performance Analysis on Id3

Table 8: Detailed Accuracy by Class

Class	TP rate	FT rate	Precision	Recall	F- measure	Roc Area
Location (3)	0.688	0.069	0.733	0.688	0.71	0.942
Location (2)	0.897	0.361	0.78	0.897	0.834	0.888
Location (1)	0.517	0.025	0.833	0.517	0.638	0.95
Weighted Average	0.777	0.232	0.78	0.777	0.769	0.912

Table 9: Confusion Matrix

	Predicted Category		
Actual Category	Location (3)	Location (2)	Location (1)
Location (3)	22	10	0
Location (2)	6	78	3
Location (1)	2	12	15

4.3.2 Decision Tree Performance Analysis on Function Tree (FT)

Table 10: Detailed Accuracy by Class

Class	TP rate	FT rate	Precision	Recall	F- measure	Roc Area
Location (3)	0.625	0.086	0.667	0.625	0.645	0.869
Location (2)	0.77	0.361	0.753	0.77	0.761	0.736
Location (1)	0.586	0.101	0.586	0.586	0.586	0.832
Weighted Average	0.703	0.25	0.702	0.703	0.702	0.783

Table 11: Confusion Matrix

	Predicted Category		
Actual Category	Location (3)	Location (2)	Location (1)
Location (3)	20	12	0
Location (2)	8	67	12
Location (1)	2	10	17

4.4 Discussion

In the case of neural networks based modelling, two types of algorithms were used: Multilayer perceptron MLP and Radial Basis Function (RBF). Models were trained with 500 epochs to minimize the root mean square and mean absolute error. For the RBF model, different numbers of hidden neurons were experimented and report the model with highest classification accuracy for the correctly classified instances. From the result, RBF model achieved training and testing performance of 54.73% and 40.56% respectively with 0.3478 of mean absolute error and 0.4484 of root mean square error. Also from the detailed accuracy by class and from confusion matrix of the result, RBF attained an accuracy rate of 0.547.

For the MLP model, the model achieved training and testing performance of 78 correctly classified instances representing 52.70% and 28 incorrectly classified instances representing 45.20% with mean absolute error of 0.3479 and mean square error of 0.5004. From the detailed accuracy by class, MLP attained an accuracy rate of 0.399. In the case of Decision Tree Performance analysis, the, dataset were experimented with two algorithms. They are Id3 and FT (function tree). For Id3 algorithm, there are 115 correctly classified instances and 33 incorrectly classified instances which represent 77.70% and 22.29% respectively. Mean absolute error was 0.1835 and Root mean squared error was 0.3029. Also for functional tree algorithm (FT), total number of tree size was 5 with 105 correctly classified instances representing 70.27% and 44 incorrectly classified instances representing 29.73%.

From the detailed accuracy by class and confusion matrix, Id3 attained accuracy rate of 0.777 and FT attained accuracy rate of 0.703.Finally, comparing the techniques from the result analysis shows that Decision Tree performs better than the Neural Networks based on the error report, number of correctly classified instances and accuracy rate generated.

Table 10: Summary of Performance

Performance Measure	**Neural Networks**	**Decision Tree**
Mean absolute error rate	**RBF Networks** 0.3478	**Id3** 0.1835
Correctly classified instances %	54.73	77.70
Accuracy rate	0.547	0.777
Mean absolute error rate	**MLP Networks** 0.3479	**FT** 0.2519
Correctly classified instances %	52.70	70.27
Accuracy rate	0. 399	0.703

On the attribute selection, using chi-squared attribute evolution with ranking method, tyre burst which represent attribute 7 has the highest value of 13.7826 followed by broken-shaft with 11.1 and loss of control with 10.8756 Also, Location two has the highest record of accidents with tyre burst being the major cause of accident on the highway.

5. CONCLUSION

In this paper, various data mining algorithms have been used on various forms of data in order to empirically analyse the performance of these algorithms, on road accident data set. The location is between the first 40 kilometres along the Ibadan-Lagos Express road. The work used Multilayer Perceptron (MLP) and Radial Basis Function (RBF) on the part of Neural Networks and Id3 and Function Tree algorithms for Decision Tree. Results shows that the Id3 tree algorithm performed better with higher accuracy rate, while Radial basis function performed better than multilayer perceptron in terms of time used in the building of the model and number of correctly classified instances. Finally, our experiments showed that, Decision Tree techniques outperformed Artificial Neural Networks with a lower error report and with a higher number of correctly classified instances and better accuracy rate generated. It is however suggested that, data in categorical form are better analysed with decision algorithm while data in continuous form are better analysed with Neural networks algorithms.

6. DIRECTION FOR FUTURE WORK

This work can be expanded by using different larger databases other than road accidents databases and more data mining classification algorithms can be employed.

REFERENCES

[1] Hand, D., Mannila, H., and Smyth, P. 2001, 'Principles of Data Mining". The MIT Press.

[2] Abdelwahab, H. T. & Abdel-Aty, M. A., Development of Artificial Neural Network Models to Predict Driver Injury Severity in Traffic Accidents at Signalized Intersections. *Transportation Research Record1746*, Paper No. 01-2234.

[3] Bedard, M., Guyatt, G. H., Stones, M. J., & Hireds, J. P., The Independent Contribution of Driver, Crash, and Vehicle Characteristics to Driver Fatalities. *Accident analysis and Prevention*, Vol. 34, 2002, pp. 717-727.

[4] Buzeman, D. G., Viano, D. C., & Lovsund, P., Car Occupant Safety in Frontal Crashes: A Parameter Study of Vehicle Mass, Impact Speed, and Inherent Vehicle Protection. *Accident Analysis andPrevention*, Vol. 30, No. 6, pp. 713-722, 1998.

[5] Martin, P. G., Crandall, J. R., and Pilkey, W. D., 2002, "Injury Trends of Passenger Car Drivers In the USA". *Accident Analysis and Prevention*, 32, pp. 541-557.

[6] Kweon, Y. J., and Kockelman, D. M. 2003, "Overall Injury Risk to Different Drivers: Combining Exposure, Frequency, and Severity Models". *Accident Analysis and Prevention*, 35, pp. 441-450.

[7] Olutayo, V.A. & Eludire, A. A. (2014) Traffic Accident Analysis Using Decision Trees and Neural Networks. *I.J. Information Technology and Computer Science,* Volume 02, pp. 22-28 Published Online in MECS (http://www.mecs-press.org/) DOI: 10.5815/ijitcs.2014.02.03

[8] Ossiander, E. M., & Cummings, P., Freeway speed limits and Traffic Fatalities in Washington State. Accident Analysis and Prevention, Vol. 34, 2002, pp. 13-18.