International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-1, November 2019

Performance Evaluation of Various Machine Learning Techniques Applied on UCI Data set

Nita Pankaj Shende, G.V.S.Rajkumar

Abstract: ***Data mining techniques are used in vast fields one of them is healthcare analysis. The present research is aimed to do the experimental analysis of multiple data mining classification /prediction techniques using three different machine learning classification and prediction tools over the online healthcare datasets. In this research, we have analyze different data mining classification and prediction techniques have been tested on four different online healthcare datasets. The standards used are a percentage of accuracy and error rate of every applied classifier technique. The experimental analysis are performed using the 10 fold cross-validation technique. Best suitable classification technique for a particular online dataset is selected based on the highest classification accuracy and the least error rate as performance measurement indicators.***

Keywords: Healthcare, Data Mining 10 fold Cross-validation, Classification techniques

I. INTRODUCTION

Millions of users around the world area unit exploitation social network sites daily. The analysis of social network information has many potentials in globe applications like friend recommendation system, product recommendation for e-commerce and suspect identification in anti-terrorism. It has been determined that social network information sizes are growing apace. It ends up in a necessity of machine learning algorithms which can scale apace with the number of examples within the information set. The procedure complexness of such algorithms will increase after the square measure applied to massive information sets. There's a necessity to check such machine algorithms whose coaching time remains the same although the data size will increase. Most learning algorithms square measure medium-scale they assume that information is often held on memory and maybe scanned repeatedly. Thus there's a necessity of corporal punishment machine learning algorithms on the distributed atmosphere which can classify information accurately. The social network information is of an unstructured kind. There's a necessity of technique which can classify this information accurately at intervals less time. With huge datasets in social Networks (e.g. weblogs), it'd take days to coach a machine learning formula.

Revised Manuscript Received on November 30, 2019.
***** Correspondence Author
Nita Prakash Shende*, Ph.D, Pursuing, Department of CS & Systems Engineering, GITAM University, Andhra Pradesh, India.
Dr. G.V.S. Raj Kumar, Ph.D, Department of CS & Systems Engineering, Andhra University, Andhra Pradesh, India.

Developing, testing, and deploying such a system which might be preventive. Because the information size is incredibly giant, for coaching this massive dataset distributed cloud computing atmosphere is used. [8][9]

A. Machine Learning Algorithms

In 1959, Arthur Samual has created a comment concerning machine learning that "Machine Learning is that the field of study that offers computers the power to be told while not being expressly programmed." In 1997, Tom Mitchell gave a good definition of machine learning. It's "An element is said to learn from past experience say E concerning some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E." Machine Learning is a type of Artificial Intelligence that makes computers able to learn without programmer support. Machine Learning (ML) plays a vital role in a wide range of critical applications like data mining, image processing, robotics, etc. Machine learning is a sub-branch of artificial intelligence and machine learning.

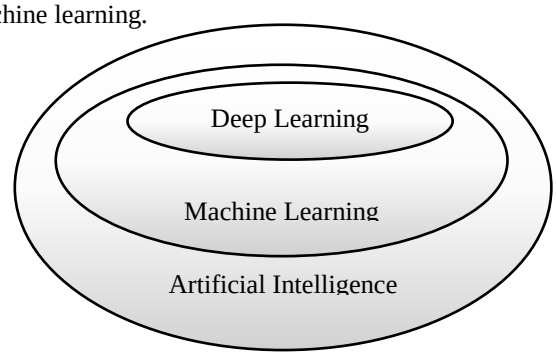


Fig: 1 Machine learning super type

Types of Machine Learning Algorithms

i) Supervised Learning
ii) Unsupervised Learning
iii) Reinforcement Learning

i) Supervised Learning: Target variable is predicted from the set of independent variables. Using set of independent variables, function is generated to map input to desired output. All data is labelled and the algorithms learn to predict output from the input data. Output datasets are provided to train the machine. Examples of supervised learning algorithms are regression, decision trees, random forest, KNN etc. [8][9]

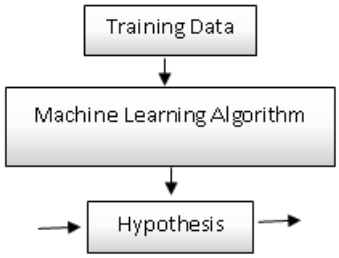


Fig 2: Supervised Learning

ii) Unsupervised Learning: In this type of learning will be no target outcome variable to be to predict. It is used to model structure or pattern of data to learn more about data. In this learning, input variable is provided without output variable

Unsupervised algorithms are further divided into two types

1) Clustering: Clustering is a technique in which inherent groups of data are discovered. K-means is a clustering algorithm

2) Association: It is used to discover rules which describe large portion of data. Apriori is an association mining algorithm.

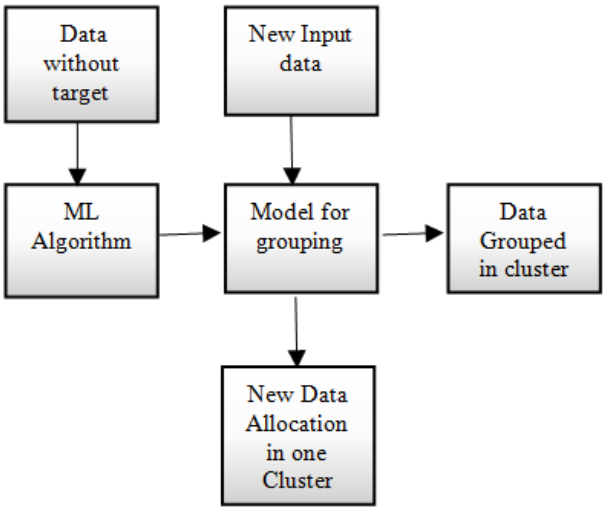


Fig 3: Unsupervised Learning

iii) Reinforcement Learning: Reinforcement Learning is used to train the machine for specific decisions. Machine is allowed to train itself continually using trial and error. It learn from past experiences and tries to take accurate decision. Example of this is reinforcement learning.

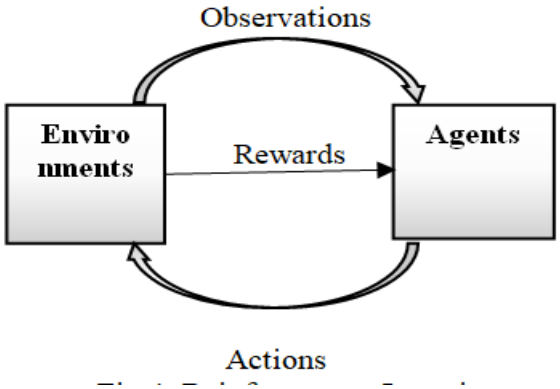


Fig 4: Reinforcement Learning

B. List of Common Machine Learning Algorithms

Commonly used machine learning algorithms are as given below

1. Linear Regression: It is used to predict value of variable based set of continuous variables. Relationship between dependent and independent variables is established or form fitting best possible line. This best line is called as regression line and its equation is given by Ax+b=y

2. Logistic Regression: It is a classification algorithm. It is used to estimate value of discrete variable based on a set of independent variables. It predicts probability of occurrence of event by fitting data to log it function. As it predicts probability, its output lies between 0 and 1.

3. Decision Tree: It is a supervised algorithm used for classification. It is used for classification of continuous as well as categorical data. It is used to divide population into two or more sets based on most significant attribute.

4. Support Vector Machine: Basic purpose is to classify the data. Here, each data element is plotted as a point in n-dimensional space. Data item value will denote coordinate of the point. Line will be drawn to divide data items into two linearly separable groups.

5. Naïve Bayes: It is a classification algorithm. It is based on Bayes theorem. It is easy to build and useful large set of data. It assume that value of particular variable is independent of any other variable of feature set. Bayes theorem provides way to find posterior probability

6. KNN (K Nearest Neighbor): this technique can be used for classification as well as regression purpose. It stores all available cases and classify new cases by taking votes from k neighbors. Class being assigned to new case is calculated by using distance function. Distance function can be Euclidean and hamming distance. Euclidean function can be used for continuous variables while hamming can be used categorical values.

7. K-Means: It is a type of unsupervised learning. It solves clustering problem. It is used to classify given data through different clusters.

8. Random Forest: It is a classification and regression method. In random forest, collections of trees are used. It is implemented by creating multiple decision trees at training time and output will be class which is mode of class's individual trees.

9. Dimensionality Reduction Algorithms: Day by day data is increasing due to voluminous online transactions, ecommerce usage etc.

More details of data are available. Data contains large amount of features. So, it becomes difficult to create robust model. In such cases, dimensionality reduction algorithms are used. These algorithms can be combined with other classification algorithms like decision trees, random forest etc. 10. Gradient Boosting and AdaBoost, Gradient Boosting and AdaBoost is used to make accurate prediction from plenty of data.

It combines multiple weak classifiers to build one strong classifier.

II. REVIEW CRITERIA

A lot of research has done on machine learning algorithm and performance evaluation techniques on them we present summary of existing research papers:

Paper No	Approach/Techniques, Data Sets used	Results & Accuracy
1	ANN,LDA,NB,SVM	LDA shown highest accuracy of 81.61 followed by SVM of 80.65[1]
2	SVM, Random Forest,KNN,NB, Softmax Performance analysis is done on small as well as large data set.	Small dataset-Naïve Bayes-70.80 Large dataset-Random forest-72.36[2]
3	THE USER - OBJECT MATRIX Concept Is Used	Highest implementation is done in python with the use of pandas and numpy[3]
4	Supervised learning-NB,NLP,SVM,Ada Boost,Bagging,DT,Random Forset,J48 Unsupervised Learning-KNN,RBF,K-Means	NB,&MLP shows good accuracy compared to other techniques.[4]
5	SVM,KNN,RF,LR	SVM shows highest accuracy.[5]
6	Zerorule,NB,MLP,SVM,C4.5,Bagging,Boosting	ALARM based network structure methodology is used which is used in scientific studies [6]
7	ID3,C4.5,CART,KNN,NB, SVM	SVM shows highest accuracy of 88% followed by KNN 86%[7]
8	Artificial neural Network(ANN), Support Vector Machine(SVM)	SVM and DNN techniques results are better on voice dataset. Parameter Tuning accuracy was 98.6% SVM shows 99.87% accuracy with ANN [8]
9	NB and DT are used, in this research article, performance Analysis is done on hospital real time data.	Naïve Bayes Gives highest accuracy of 94% [9]
10	Overview of all data mining techniques was discussed	Hybrid predication model is proposed for cost reduction and hospital quality improvement [10]

III. RESULT

This section presents the results and analysis of the experiments conducted for this study. The experiment research methodology was explained in section 3. Different classification methods were applied for the experiments on the four different healthcare datasets taken from the database of UCI. The overview of the selected datasets for this work is shown in Table 1.

Table: I Dataset Description

Dataset	Classes	Attribute	Instances
PIMA Indian Diabetes	2	9	767

Technique Applied	Accuracy Rate	Error Rate
C4.5	84.54	15.4
ID3	77.23	22.68
SVM	96.67	3.18
kNN	80.3	19.66
Prototype NN	63.56	36.61
CRT	78.54	21.58
LDA	78.43	21.7

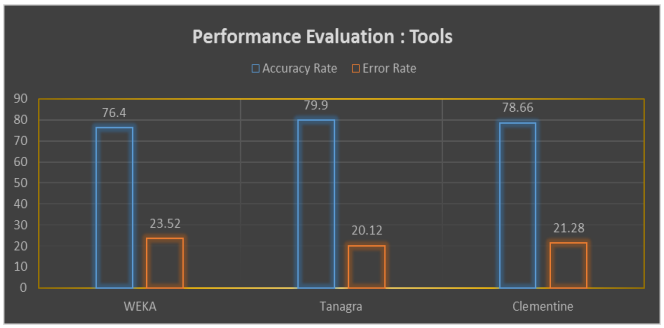
In t in this analysis of WEKA, Tanagra and Clementine Data Mining Machine Learning Tools is used to achieve the objectives proposed. For classification methods, the percentage of accuracy rate and error rate is used as the calculation criteria for evaluation. Such parameters imply that a high accuracy rate value and a low error rate value for a classification technique applied to a dataset indicate for experiments, the data is firstly divided into training data and testing data. The training set is used to construct the validation classifier and test set. In this analysis, 66% and 34% respectively are the percentages used for training and testing results. The participating classification techniques are then implemented using the 10-fold cross validation method to produce the classifiers through the machine learning tools described above. The results are finally reported in terms of

Table 2: Results obtained in WEKA

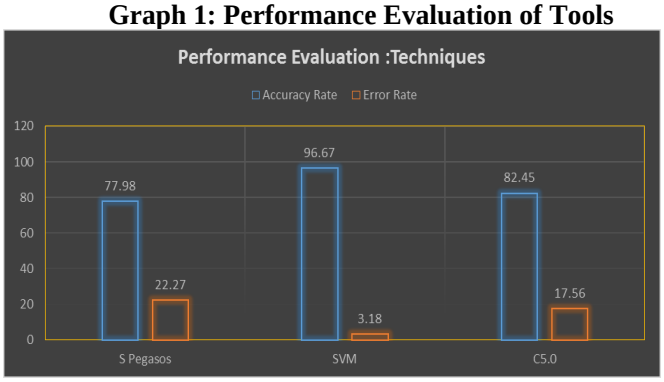
Technique Applied	Accuracy Rate	Error Rate
Bayes Net	74.46	25.34
Naïve Bayes	75.34	23.5
J48	73.9	25.9
MLP	75.78	24.08
SMO	77.45	22.66
Logistic	77.22	22.78
LMT	77.87	22.53
S Pegasos	77.98	22.27
FT	77.56	22.66

TABLE 3: RESULTS OBTAINED IN CLEMENTINE

Technique Applied	Accuracy Rate	Error Rate
NN-RBFN	78.23	21.74
C5.0	82.45	17.56
C&RT	81.34	18.65
QUEST	76.17	23.32
CHAID	77.56	22.65
LDA	76.65	23.18
Logistic	78.25	21.89



Graph 1: Performance Evaluation of Tools



Graph 2: Performance Evaluation of machine Learning Techniques

IV. CONCLUSION

From above performance analysis we conclude that both supervised and unsupervised techniques can be applied on UCI or real time data set and accuracy can be improved for prediction.

REFERENCES

1. A Comparative Study of Supervised Machine Learning Algorithms for Stock Market Trend Prediction, Indu Kumar Etl, ICICCT, IEEE, 2018.
2. Application of machine learning in recommendation systems, Agata Nawrocka etl, IEEE, 2018.
3. Comparative Performance Analysis Of Machine Learning Techniques For Software Bug Detection, Saiqa Aleem,Etl, Itcs, Cst,2015.
4. A Machine Learning Approach for the Classification of Cardiac Arrhythmia, Prajwal Shimpi etl, ICCMC,2017.
5. Performance Evaluation of the Machine Learning Algorithms Used in Inference Mechanism of a Medical Decision Support System, Mert Bal etl, Scientific World Journal,2014.
6. Machine Learning Techniques for Data Mining: A Survey, Seema Sharma etl,IEEE,2013.
7. Performance Analysis of Machine Learning Algorithms for Gender Classification, Laxmi Narayana Pondhu etl, ICICCT ,IEEE,2018
8. A Framework For Decision Making & Quality Improvement By Data Aggregation Techniques On Private Hospitals Data, Syed Ahmed Yasin, Dr.P.V.R.D.Prasad Rao,ARPN,July ,2018.
9. Analysis Of Single And Hybrid Data Mining Techniques For Prediction Of Heart Disease Using Real Time Dataset, Syed Ahmed Yasin, Dr.P.V.R.D.Prasad Rao,IJET,2018.

AUTHORS PROFILE

Nita Prakash Shende, She has completed M.E from Mumbai University in 2014.Her area of interest is computer vision and machine learning, She is pursuing PHD from GITAM University.

Dr. G.V.S. Raj Kumar has completed his Ph.D. in CS & Systems Engineering from Andhra University. He is working GITAM UNIVERSITY as Professor. His areas of interest are image processing, Communication Network and security.