IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-6, Issue-12, Dec 2018 E-ISSN: 2347-2693

A Comparative Study of Supervised Machine Learning Algorithm

D.Sathiya[1*], S. V. Evangelin Sonia[2]

[1,2]Computer Science and Engineering, Sri Shakthi Institute of Engineering and Technology, Coimbatore, India

[*]*Corresponding Author: sathiyads.90@gmail.com*

Available online at: www.ijcseonline.org

Accepted: 22/Dec/2018, Published: 31/Dec/2018

Abstract--- Machine Learning is a process which begins with observations of data to make better decisions of new data in future. Machine Learning algorithms divides as Supervised Machine Learning, Unsupervised Machine Learning, Semi-Supervised Machine Learning and Reinforcement Learning. In this paper, we focus on Supervised Machine Learning Algorithms especially its error rates. A Supervised learning algorithm analyses the training data and produces a classifier (conditional function), which can then be used for mapping test sets. We compare the various Supervised Machine Learning algorithms in terms of its error rates in this paper.

Keywords: Supervised Machine Learning, Classifier, Error Rate

I. INTRODUCTION

Machine Learning is a subset of artificial intelligence, which learn from past experience effectively and produces a predictor which intern predicts the new decisions. This consist of making accurate predictions, completing a task, etc. There are several applications of Machine Learning, sales and marketing, transportation, computer vision, etc.

The learning always requires some observations or data points. Some of the Machine Learning use cases are Recognizing and finding faces in images, Classifying articles in categories like sports, politics, entertainment, recognizing handwritten characters, Natural Language Processing, Medical Diagnosis of Diseases

Machine learning has developed based on the ability to use computers to probe the data for structure, even if we do not know what the structure looks like. The test for model is a validation on new data, not on a theoretical test that proves a null hypothesis. Because machine learning uses an iterative approach to learn from data, thus the learning can be easily automated. Passes are run through the data until a healthy pattern is found [1].

Machine learning algorithms are often categorized as supervised, unsupervised and reinforced Learning

Supervised Learning: Learning from a labelled data is Supervised Learning. The system is trained on a set of data points which are pre-defined training examples. [2] This is done to facilitate the system to find a better prediction (performance measure) on a new test data set. Supervised ML algorithms uses the mapping function

$$Y = f(x)$$

The goal is to approximate the mapping function so well that when we have new input data (x), we can predict the output variables (Y) for that data.

Unsupervised Learning: Learning from a non-labelled data is Unsupervised Learning. That is, the training dataset doesn't have well defined relationships and patterns.

The basic difference between the both the algorithms is, in supervised learning, a portion of output dataset is given to train the model, in order to generate the desired outputs. But, in unsupervised learning no such dataset is provided for learning, rather the data is clustered into classes.

Reinforced Learning: Learning and updating the parameters of model based on the feedback of the output is Reinforcement Learning. Here, dataset is divided into two sets, training set and test set. The program is trained using the well-defined training dataset and is then tuned using feedback from the results of test dataset.

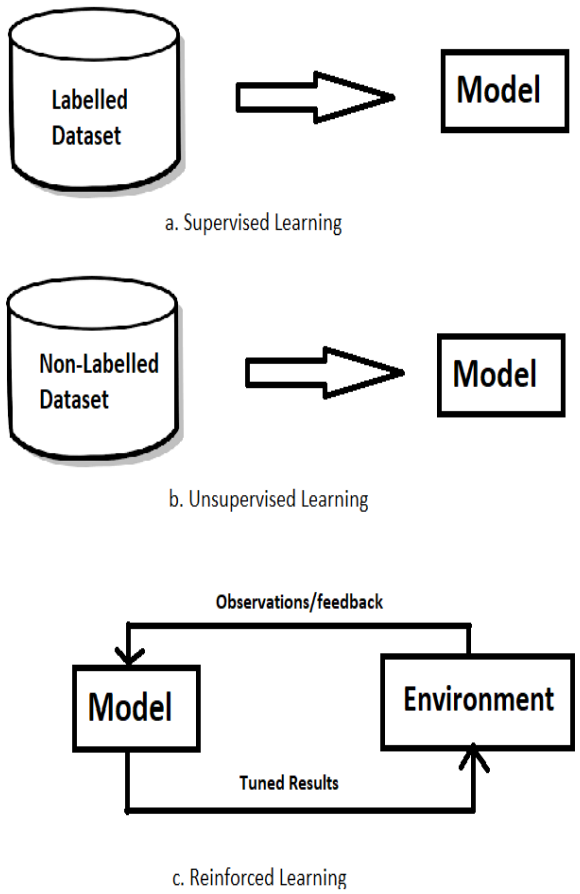

Fig. 1: Classification of Machine Learning Algorithms

ML is perfectly intended for accomplishing the accessibility hidden within Big data. ML handovers on the guarantee of extracting importance from big and distinct data sources through outlying less dependence scheduled on individual track as it is data determined and spurts at machine scale [3].

Here, we identified some popular Supervised Machine Learning algorithms such as, Linear Regression, Naive Bayes Classifier, Support vector Machine and Decision Tree and analysed the papers which discussed about error rates and some improvement techniques.

The remaining part of this work is arranged as follows: Section 3 presents the data preparation which is the first step of machine learning; Section 4 discusses about supervised machine learning algorithm and results of each algorithms and Section 5 gives the overall conclusion.

II. DATA PREPARATION

Before starting with the algorithm, the important step in machine learning is the dataset preparation. This step is called the pre-processing. Data pre-processing is important because the data in the dataset will be raw with impurities. Impurities can be missing data, irrelevant data etc. [4] Commonly it is called as outliers.

III. SUPERVISED MACHINE LEARNING ALGORITHM

Classification algorithms trains machines how to group together data by particular criteria, they group data together based on their characteristics

A. Linear Regression

Linear Regression is the Linear model, where the variables involved are purely considered to be independent of each other. Inputs are x, independent variables and Output is y, dependent variable.

$Y = b_0 + b_1x_1$

Y – Dependent variable

b_0, b_1 – constant

x_1 . co-efficient

Linear regression is used mainly to predict the numerical values. This type is very sensitive to outliers, it cannot resist the outliers and probably it results in wrong data predictions. [5] If there is error in test set and training set, there will be high error rate. Based on the number of errors, error rate will vary.

B. Naive Bayes Classifier

The Naive Bayes Classifier technique is based Bayesian theorem and is particularly suited when the dimensionality of the inputs is high. The naive Bayes algorithm does by making an assumption of conditional independence over the training dataset**.** This drastically reduces the complexity to 2n.

The assumption of conditional independence states that, given random variables X, Y and Z, X is conditionally independent of Y given Z, if and only if the probability distribution governing X is independent of the value of Y given Z.

In other words, X and Y are conditionally independent given *Z* if and only if, given knowledge that *Z* occurs, knowledge of whether *X* occurs provides no information on the likelihood of *Y* occurring, and knowledge of whether *Y* occurs provides no information on the likelihood of *X* occurring.

The naïve Bayes classifier is thought to be one of the most effective classification algorithms today, challenging with

more modern and sophisticated classifiers. Despite being based on raw assumption that all variables in the set are independent, given the output class, the classifier provides proper results. clearly shows about the reduction in error rates. In the naïve Bayesian network, the error rate has been dropped from 25% to 5%. In the hierarchical Bayesian network, there was a drop in 15% error rate [6].

C. Support Vector Machine (SVM)

A support vector machine classifies inputs as belonging to one of two different classes of outputs. It does by calculating whether or not the input is above or below a classifying hyperplane. In two-dimension space, the classifying hyperplane is a line, so an input is classified based on whether the input is above or below a certain line. The input to a support vector machine must be a point in space of numerical information. The hyperplane has two components. [7] Support vector machines will classify inputs even when the inputs are not linearly separable that is there is no hyperplane that classifies perfectly. In such cases, the classifying hyperplane of a support vector machine will have minimal error. Then, the classifying hyperplane of a support vector machine will be in between the two classes of data as much as possible.

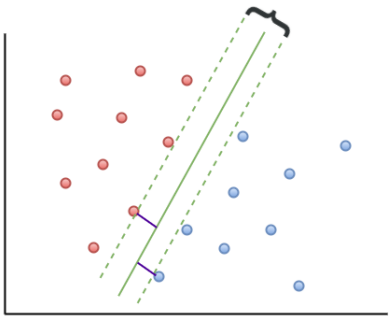

Fig. 2. Support Vector Machine Hyperplane

In SVM, K-fold Cross validation may decrease the error rate. [8] The best ways to combine the k classifiers appear to be the retraining solution and the randomly chosen model. The former performs better in practice, while the second one allows predicting the classifier error rate on unobserved data.

D. Decision Tree

Decision trees are supervised learning algorithms. Decision trees are assigned to the information based learning algorithms which uses different measures of information gain for learning. [9] Decision trees are used for issues where continuous but also categorical input and target features. The main idea of decision trees is to find the descriptive features which contain the most information regarding the target feature and then split the dataset along the values of these features that the target feature values for the resulting sub datasets are as pure as possible. The descriptive feature leaves the target is said to be the most informative. The process of finding the most informative feature is done until the stopping criteria is accomplished, then finally end up in so called leaf nodes. The leaf nodes contain the predictions which will make for new query instances presented to the trained model. This is possible since the model has learned the underlying structure of the training data and hence given some assumptions makes predictions about the target feature value of unseen query instances.

A decision tree contains of a root node, interior nodes, and leaf nodes which are then connected by branches.

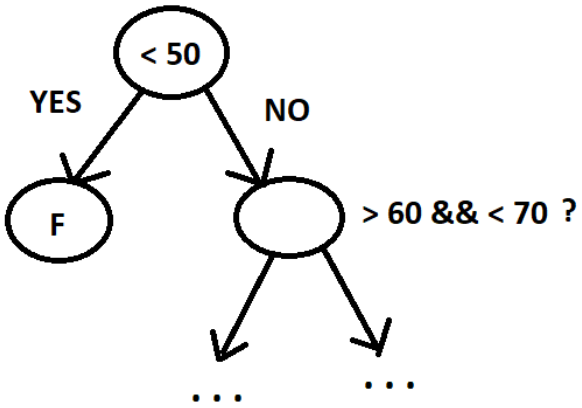

Fig. 3. Decision Tree Sample

The maximum value for the average classification error rate of 51.31% &53.80% obtained for c4.5 and WDT algorithms.

IV. CONCLUSION

Based on the observations from several papers, most of the classification algorithms are sensitive to the outlier, and can affect the error rates in various aspects. Each paper followed a different technique to get rid of outliers. We conclude based on the error rates observed in the papers. Error rate in Regression is strongly based on the outlier detection and correction measures. In SVM, choosing the k value may alter the error rates in higher extent and we observe that Naive Bayes Classifier have minimum error rate when compared with all the above algorithms and can be used for quite large data sets, only when the data pre-processing is done wisely.

REFERENCES

[1] Iqbal Muhammad and Zhu Yan, "Supervised Machine Learning Approaches: A Survey", ICTACT Journal on Soft Computing, Vol. 05, Issue. 03, pp. 946-952, 2015.

[2] Shai Shalev-Shwartz and Shai Ben-David, "Understanding Machine Learning: From Theory to Algorithms", Cambridge University Press, 2014.

[3] Osisanwo F. Y, Akinsola J.E.T, Awodele O, Hinmikaiye J.O., Olakanmi O., Akinjobi J., "Supervised Machine Learning Algorithms: Classification and Comparison", International Journal of Computer Trends and Technology, Vol. 48, Issue. 3, pp. 128-138, 2017.

[4] V. Ilango, R. Subramanian, V. Vasudevan, "A Five Step Procedure for Outlier Analysis in Data Mining" European Journal of Scientific Research, Vol. 75, Issue. 3, pp. 327-339, 2012.

[5] Barbara D. Klein & Donald F. Rossin, "Data Quality in Linear Regression Models: Effect of Errors in Test Data and Errors in Training Data on Predictive Accuracy", Vol. 2, Issue. 2, Informing Science, 1999.

[6] ISSN 1450-216X Vol.75 No.3 (2012), pp. 327-339 © Euro Journals Publishing, Inc. 2012

[7] http://www.europeanjournalofscientificresearch.com European Journal of Scientific Research

[8] ISSN 1450-216X Vol.75 No.3 (2012), pp. 327-339© Euro Journals Publishing, Inc. 2012

[9] http://www.europeanjournalofscientificresearch.com European Journal of Scientific Research

[10]ISSN 1450-216X Vol.75 No.3 (2012), pp. 327-339 © EuroJournals Publishing, Inc. 2012

[11] http://www.europeanjournalofscientificresearch.com European Journal of Scientific Research

[12] Felipe Schneider Costa, Maria Marlene De Souza Pires and Silvia Modesto Nassar, "Analysis Of Bayesian Classifier Accuracy" Journal of Computer Science, Vol. 9 Issue. 11, pp. 1487-1495, 2013.

[13] Hyunjung Shin and Sungzoon Cho, "Neighborhood Property–Based Pattern Selection for Support Vector Machines", Neural Computation, Volume 19 Issue 3, pp. 816-855, 2007.

[14] DavideAnguita , Alessandro Ghio, Sandro Ridella , and Dario Sterpi, "K–Fold Cross Validation for Error Rate Estimate in Support Vector Machines", Proceedings of The 2009 International Conference on Data Mining (DMIN), pp. 291-297, 2009.

[15] A. Sheik Abdullah, S. Selvakumar, P. Karthikeyan and M. Venkatesh, "Comparing the Efficacy of Decision Tree and its Variants using Medical Data", Indian Journal of Science and Technology, Vol. 10 Issue. 18, pp 01-08, 2017.

