Journal of Advances in Science and Technology
Vol. 17, Issue No. 1, March-2020, ISSN 2230-9659

Algorithms for Predicting using Machine Learning

Jyotsna Tiwari[1]*, Dr. Monika Tripathi[2]

[1] Research Scholar, Shri Krishna University, Chhatarpur, M.P.

[2] Professor, Shri Krishna University, Chhatarpur, M.P.

Abstract - ***Given the vast amount of real-world statistics that are easily accessible and the growing popularity of analytics, selecting the best prediction algorithm is crucial. Even though there are a number of forecasting models that are regularly used for predictive analytics, it may be challenging to decide which algorithm is optimal for a certain real-world dataset & research topic. The three most well-known machine learning and predictive analytics algorithms are discussed in this article in addition to the implementation outcomes on real datasets. These algorithms were evaluated and compared using performance comparison metrics such time training, accuracy, sensitivity, specificity, accuracy, the area under the curve and error.***

Keywords - Machine learning, Predictive analytics, K nearest neighbor, SVM, Naive Bayes.

---------------------------------x---------------------------------

1. INTRODUCTION

Daily use of social networking sites is in the millions of people worldwide. Applications in the actual world for social network data analysis include suspect identification in counterterrorism, e-commerce product recommendations, and buddy referral systems. Social network data volumes have been shown to be expanding quickly. Machine learning methods that can scale quickly with number of instances in the data set become necessary as a result. When used on huge data sets, these methods become more computationally difficult. It is important to research machine learning algorithms where training time stays constant as data size grows. The majority of learning algorithms were medium sized and work on the assumption that data may be frequently read and stored in memory. Therefore, it is necessary to run machine learning algorithms in a distributed environment in order to effectively identify data. Unstructured data of the like is found in social networks. A method that can properly identify this data in a short amount of time is required.[1]

It would take days to build a machine learning system using the vast datasets found in social networks (such as web logs). Such a system would obviously be too expensive to develop, test, and implement. Due to the scale of the data, a distributed cloud services environment may be employed for training this enormous dataset.[2]

This branch of research aims to offer computers the capacity to learn without having to be explicitly programmed, as stated by Arthur Samual in 1959. In 1997, Tom Mitchell defined machine learning in a clear and concise manner. When a computer program's performance on a task increases as a function of its experience, it is said to have learned from that experience.[3]

The term "machine learning" refers to the ability of a computer to learn on its own, without the assistance of a programmer. Data mining, image processing, robotics, and other key applications benefit greatly from machine learning (ML).[4]

Algorithm Types for Machine Learning

1. SupervisedLearning
2. UnsupervisedLearning
3. ReinforcementLearning

2. ALGORITHMS FOR MACHINE LEARNING

The following is a list of frequently used machine learning algorithms.

1. Linear Regression
2. Logistic Regression
3. Decision Tree

4. SVM

5. Naïve Bayes

6. KNN

7. K-Means

8. Random Forest

9. Dimensionality Reduction Algorithms

10. Gradient Boost& Adaboost

1. Linear Regression

Variable values are predicted using a continuous collection of variables. Fitting the optimal line establishes the relationship between the variables variables. Known as a regression line, this is the most useful line, and the equation for it is provided by Ax+b=y

2. Logistic Regression

It's a way of sorting things into different categories. When a collection of independent variables is used to estimate a discrete variable's value, it is called a Bayesian model. A logit function is used to forecast the likelihood of an event occurring. Its output ranges from 0 to 1, reflecting its ability to forecast likelihood.

3. DecisionTree

It is a classification method that is supervised. For categorical and continuous data categorization, it is employed. Divide the population into three or more organizations depending on the most important characteristic.

4. Support Vector Machine

It's used to categorise information. An n-dimensional plot is drawn for each piece of data in this procedure. The coordinates of the point are represented by the data item value. In order to separate the data into two linearly distinct groups, a line will be drawn.

5. Naïve Bayes

It's a way of sorting things into different categories. In this model, Bayes theory is used as the basis. Large data sets are simple to create and beneficial to have. It assumes that the value of a specific variable is unaffected by the values of other variables in the set. Feature values are recorded as a vector of feature values, and class labels are picked from a limited set for each issue occurrence. The Bayes theorem gives a method for determining the likelihood of an event occurring in the future.

$P(c \mid x)$ from $P(c)$,$P(x)$ and $P(x|c)$.

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

P(c|x) is given the property x, the likelihood of class c

P(c)is for class c: posterior probabilities

P(x|c) is Probability is the possibility that predictor x will occur given class c.

P(x)is Predictor's previous likelihood x.

6. KNN(Knearest Neighbour)

Use it for both classification and regression. In order to categorise a new situation, it takes into account the opinions of k neighbours. The distance function is used to assign a class to a new case. Euclidean and hamming distances may be used for the distance function. Euclidean and hamming functions may be used to both continuous and categorical data.

7. K-Means

It's a kind of unguided discovery. The clustering issue is solved. It is used to categorise a set of data into several clusters.

8. RandomForest

Uses both classification and regression. Collections of trees are utilised in random forests. Each tree provides a categorization for a new item based on its properties. Multiple decision trees are trained at the same time to produce a single class that is the sum of the modes of the trees.

9. Dimensionality Reduction Algorithms

The amount of data generated by online transactions, e-commerce, etc. is expanding on a daily basis. More information is available about the data. There are a plethora of features packed within the data. As a result, building a solid model becomes more challenging. The employment of dimensionality reduction methods is common in these situations. Decision trees, random forests, and more may be used in conjunction with these techniques. On the basis of correlation matrix, missing values ratios, etc.[5]

10. Gradient Boosting and Ada Boost

Accurate predictions may be made using Gradient Boosting and AdaBoost. One powerful classifier is built by combining many weak ones.

3. THE CLASSIFICATION AND PREDICTION ALGORITHMS OF MACHINE LEARNING

Algorithms for classifying and forecasting fresh examples are used in this process. Based on an established training set of observations, this is done. Pattern recognition is used in classification, for example. This sort of machine learning is supervised. In computer science, a classifier is a programme that uses classification.[6]

Binary and multiclass classifications are the two most used classification methods. Multiclass classification differs from binary classification in that it includes both binary and multiclass classifications. Figure 1 shows the classification and prediction system in action.

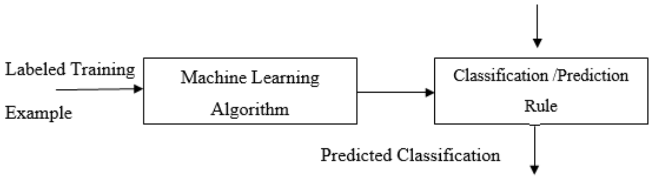


Figure 1: Classifier for Typical Data

The quality of a classifier's output relies on the data's attributes. For every issue, there is no one-size-fits-all algorithm.

3.1 The Advantages of Machine Learning

- Human-created rules and models might be inaccurate, while MLA are generally more accurate.
- Human experts or programmers are not required for this task.
- Automated approach for finding hypotheses that explain data
- Inexpensive as well as adaptable
- Almost any set of training data may be used with it.

3.2 Machine Learning on the Cloud

Distributed machine learning is becoming more crucial in the age of Big Data. Sequential MLA outclass as data grows in volume, variety, and velocity. Large-capacity CPUs and RAMs are already commonplace in computers because to the fast advancement of technology.

It is not possible to exploit the processing capability of contemporary computers with sequenced machine learning methods. Because of this, distributed MLA must be designed to take use of the whole computing capacity and decrease the amount of time necessary to train the machine learning algorithms.[7]

3.3 Framework for Distributed Machine Learning

Mahut is a distributed machine learning platform. HADOOP is used as a foundation for this application.

In Mahut, you'll find a variety of libraries for machine learning algorithms. Mahut has created methods for classification, grouping, and mining. Large amounts of data are processed with it. another machine learning framework that is distributed is Spark. Real-time big data analysis is the primary application. Several classification, clustering, and pattern mining techniques are included in MLib, an Apache Sparks library. A general-purpose execution graph engine and high-level Java, Scala, and Python APIs are included. High-level tools such as Spark SQL (MLlib), GraphX , and SparkStreaming are also supported.[8]

4. MATERIAL AND METHODS

As the test data is processed by the machine learning algorithms, the features are automatically retrieved. Features and expected output are included in the training set. Class labels are used as input to the function, which returns either a continuous value or an estimation.[9]

Implementation of the proposed work relies on Scikit Learn, a Python Machine Learning module. Using it, you may use algorithms that use supervised or unsupervised teaching methods. An additional tool that works on Apache Spark is the Jupyter notebook. Jupyter notebooks using machine learning methods are implemented using the Spark framework.[10]

Orange is open source software that aids in the loading and transformation of information. Data mining, deep learning, predictive analytics, and statistical modelling all benefit from this technique. Using this programme, the suggested model's algorithms may be compared to see how well they perform.[11]

5. RESULT

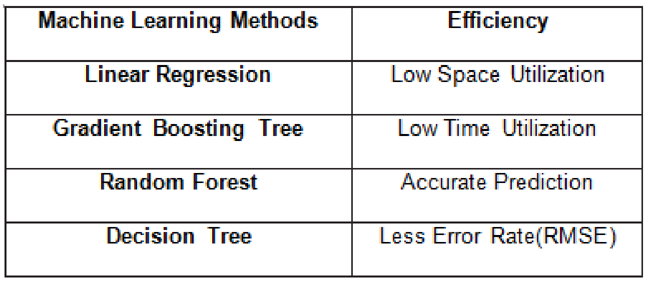
Linear Regression, Random Forest, Decision Tree, and Gradient Boosting tree techniques are employed in the prescribed model to predict the outcome. compares the outcomes of the different algorithms for learning.[12]

Table 1: A Comparison of Machine Learning Methods for Prediction

Year	Linear Regression	Random Forest	DecisionTree	Gradient Booster
1906	23.8106	19.045	15.825	12.697
1919	24.2519	19.045	15.825	19.2367
1916	24.4107	19.045	15.825	19.2367
1921	24.2019	18.9079	24.08	19.2367
1926	25.451	18.9079	24.08	19.0267
1931	23.4775	19.04	24.08	19.0267
1936	24.2686	19.04	24.169	18.967
1941	24.6349	19.103	24.169	19.28
1946	23.8023	18.62	23.963	19.28
1951	24.5434	19.317	23.963	19.28
1956	24.1936	19.317	24.0795	29.33
1961	23.8855	19.317	24.0795	29.33
1966	24.7349	19.317	24.2567	19.225
1971	23.719	19.988	24.2567	19.3716
1976	23.744	19.456	24.276	19.3716
1981	24.4518	19.456	24.267	19.3716
1986	24.2436	19.456	23.8818	19.337
1991	25.2511	19.2493	23.8818	19.554
1996	24.6349	19.08	24.1435	19.554
2001	24.3019	19.08	24.1435	19.554
2006	23.5327	19.457	24.517	19.475
2011	24.7825	19.251	23.486	19.147
2016	25.2378	19.249	24.143	19.357
2021	24.5673	19.364	24.0245	19.264

On top of the constructed model, a variety of machine learning techniques are used, each with a unique set of advantages and disadvantages. provides an overview of the learning techniques' efficiency.[13]

Table 2 : Machine Learning Methods' Efficiency



Machine Learning Methods	Efficiency
Linear Regression	Low Space Utilization
Gradient Boosting Tree	Low Time Utilization
Random Forest	Accurate Prediction
Decision Tree	Less Error Rate(RMSE)

Various regression models were used to analyse the efficiency of the process and get the results. The comparative examination of learning algorithms is a rather impressive piece of work. This approach to discovering regression models is also consistent with the prediction of datasets employing adaptive tools.[14]

6. CONCLUSION

Machine Learning approaches are used to keep track of tasks involving factors linked to time and space while the data analysis is being carried out. For future predictions, this model uses the Spark framework to optimise Machine Learning methods in a distributed environment to reduce the overall execution time complexity. Machine learning efficiency is increased by 70% when the Spark framework is used on top of traditional MapReduce methods. Each learning method is described in detail, along with the specific attributes that make up each one, and the results of the research are compared to the model that is offered. Below you'll find a collection of model metrics.

REFERENCES

1. Arroyo, J., & Maté, C. (2009). Forecasting histogram time series with k-nearest neighbours methods. International Journal of Forecasting, 25(1), 192-207.

2. Barga, R., Fontama, V., Tok, W. H., & Cabrera-Cordon, L. (2015). Predictive analytics with Microsoft Azure machine learning. Berkely, CA: Apress. [

3. Deepak, E., Pooja, G. S., Jyothi, R. N., Kumar, S. P., & Kishore, K. V. (2016, August). SVM kernel based predictive analytics on faculty performance evaluation. In 2016 International Conference on Inventive Computation Technologies (ICICT) (Vol. 3, pp. 1-4). IEEE.

4. Fernando, Z. T., Trivedi, P., & Patni, A. (2013, August). DOCAID: Predictive healthcare analytics using naive bayes classification. In Second Student Research Symposium (SRS), International

Conference on Advances in Computing, Communications and Informatics (ICACCI'13) (pp. 1-5).

5. Kelleher, J. D., Mac Namee, B., & D'arcy, A. (2015). Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies. MIT press.

6. Kendale, S., Kulkarni, P., Rosenberg, A. D., & Wang, J. (2018). Supervised Machine-learning Predictive Analytics for Prediction of Postinduction Hypotension. Anesthesiology, 129(4), 675-688.

7. Lin, J., & Kolcz, A. (2012, May). Large-scale machine learning at twitter. In Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data (pp. 793-804).

8. Mishra, N., & Silakari, S. (2012). Predictive analytics: A survey, trends, applications, oppurtunities & challenges. International Journal of Computer Science and Information

9. Nirbhay Bhuyar , Samadrita Acharya , Dipti Theng, 2020, Crop Classification with Multi-Temporal Satellite Image Data, International Journal of Engineering Research & Technology (IJERT) Volume 09, Issue 06 (June 2020).

10. Nithya, B., & Ilango, V. (2017, June). Predictive analytics in health care using machine learning tools and techniques. In 2017 International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 492-499). IEEE.

11. Rajeshkanna, A., Preetha, V., & Arunesh, K. (2019, March). Experimental Analysis of Machine Learning Algorithms in Classification Task of Mobile Network Providers in Virudhunagar District. In International Conference on E-Business and Telecommunications (pp. 335-343). Springer, Cham.

12. Ratner, B. (2017). Statistical and Machine-Learning Data Mining: Techniques for Better Predictive Modeling and Analysis of Big Data. CRC Press.

13. Satpute, P. C., & Theng, D. P. (2013, April). Intellectual climate system for monitoring Industrial environment. In 2013 Third International Conference on Advanced Computing and Communication Technologies (ACCT) (pp. 36-39). IEEE.

14. Shin, S. J., Woo, J., & Rachuri, S. (2014). Predictive analytics model for power consumption in manufacturing. Procedia Cirp, 15, 153-158.

Corresponding Author

Jyotsna Tiwari*

Research Scholar, Shri Krishna University, Chhatarpur, M.P.