iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** VI **Month of publication:** June 2017

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

Volume 5 Issue VI, June 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Analysing Machine Learning Algorithms & their Areas of Application

Manish Kumar Singh[1], Prof. G S Baluja[2], Dr. Dinesh Prasad Sahu[3]
[1]*Baluja Institute of Technology & Management*
[2]*TTE, The Government of NCT and Delhi*
[3]*Jawaharlal Nehru University*

Abstract*: *Machine learning has become one of the most envisaged areas of research and development field in modern times. But the area of research related to machine learning is not new. The term machine learning was coined by Arthur Samuel in 1952 and since then lots of developments have been made in this field. The data scientists and the machine learning enthusiasts have developed myriad algorithms from time to time to let the benefit of machine learning reach to each and every field of human endeavors. This paper is an effort to put light on some of the most prominent algorithms that have been used in machine learning field on frequent basis since the time of its inception. Further, we will analyze their area of applications.
Keywords:* *AdaBoost, Linear Regression, Decision Tree, Naive Bayes Algorithm, KNN, SVM

I. INTRODUCTION

The principle idea of machine learning field is to enable machine to make generalization from its past experience. Generalization in the context of machine learner implies the machine's ability to work effectively on vivid new and unseen tasks through the experience gained by it from the programmed dataset. The learning machine has to be devised around a space where the training examples are pumped into from a set of unknown probability distribution of general nature so that the machine becomes able enough to generate accurate predictions effectively for new and unfamiliar situations/tasks.

Computational learning theory is basically that branch of theoretical computer science that deals with the computational analysis and the performance of algorithms related to machine learning field. Machine learning theory never guarantees about the performance of such algorithms. This is due to the finite nature of training data sets and their uncertain future of getting well organized to obtain any relevant information. However, the performance of the machine learning algorithms have probabilistic bound and their analysis in such a manner is quite a common practice. For instance, the generalization error can be quantified through bias-variance decomposition where the generalization in the context of machine learning is worked upon to obtain best performances by allowing the complexity match of the associated hypothesis with that of the complexity of the function build around the given data. Moreover, the given two results can be checked for generalization in the context of machine learning that:

A. If the complexity of the hypothesis is less as compared to that of the function, it can be said that the given algorithm model of machine learning under-fits the corresponding data; As a result, the complexity of the algorithm model of the machine learning increases in response that will be a clear indication of decrement of the corresponding training error.

B. If the complexity of the hypothesis is more as compared to that of the function, it can be said that the given algorithm model of machine learning over-fits the corresponding data; As a result, the generalization will be found to be poorer.

The theorists of the computational learning branch also do study of the feasibility and the time complexity of the machine learning algorithms in addition to the study of performance bounds. A machine learning algorithm is said to be feasible when the algorithm gives performance in polynomial time. A machine learning algorithm has two kinds of associated time complexity results – (a). Positive results of time complexity indicates that a set of functions considered are going to be learned by the machine in polynomial time; whereas (b). Negative results of time complexity indicates that a set of functions considered are not going to be learned by the machine in polynomial time.

This article puts a light on the various algorithms discussed under the machine learning field. The section 2 of this article analyzes the machine learning algorithms while the section 3 of this article discusses the area of applications of each algorithm. Finally, the section 4 wraps up the article with suitable conclusion.

II. ANALYSING MACHINE LEARNING ALGORITHMS

There are ten algorithms in machine learning field that works wonderfully to obtain best results in the corresponding area of application. These ten algorithms are obtained through the fundamental learning techniques of Supervised Learning, Unsupervised Learning and Reinforcement Learning.

A. Gradient Boosting & Adaboosting Algorithm

These are the most preferred machine learning algorithms in the present era of machine learning for the reason that they boost the task the of making highly accurate predictions in situations like the handling of data with massive loads. Boosting is considered as an ensemble machine learning algorithm that involves the combination of the myriad base estimators' predictive power to enhance robustness. In other words, it does the combination of several average or weak predictors for building a strong predictor. The most amazing thing about boosting algorithms is that they are frequently used in the Data Science competitions including AV Hackathon, Kaggle and CrowdAnalytix. The data scientists use Python language in the combination of R codes to generate best results through the boosting algorithms.

B. Linear Regression

The linear regression algorithm establishes an relationship between dependent and independent variables after fitting the variables to a line. The line so obtained is known as the line of regression that can be represented through a linear equation of the form:

$$Y = a * X + b$$

Where, X is Independent variable, Y is dependent variable, a is slope and b is intercept. The value of a & b coefficients can be derived through the minimization of the sum of the squared difference of the distance between the regression line and data points.

C. Decision Tree

This is considered as one of the most prominent machine learning algorithms that are quite useful to do classification of both continuous and categorical dependent variables. Using supervised machine learning technique, decision tree algorithm simply splits a given population in terms of at least two homogeneous sets on the basis of the most significant independent variables or attributes.

D. Logistic Regression

This machine learning algorithm is useful to do the estimation of discrete values (generally in the form of binary values of 1 or 0) through a set of obtained independent variables. It is quite beneficial for the prediction of the probability of an event that can be obtained by the fitting of data to a logit function. The methods like eliminate features, interaction terms, non-linear model and regularize techniques are used to do an improvisation of logistic regression algorithm.

E. K-Nearest Neighbors (KNN) Algorithm

Likewise SVM algorithm, KNN algorithm is also applicable in classification and regression based applications. However, it is basically used to solve classification related problems as far as Data Science Industry is concerned. KNN algorithm works by storing all kind of available cases and it further does the classification of any additional case only after considering the vote of a majority of its k-neighbours. The case so obtained is then assigned to that class only with which it shares a maximum similarity. For the measurement of this thing, a distance function is considered. But, KNN algorithm has major limitations associated with it like (a).it is a computationally quite expensive method, (b). Only normalized variables are considered for the reason that the algorithm could be otherwise biased by the higher range variables, and (c). Pre-processing of data is always required.

F. Support Vector Machine (SVM)

This machine learning algorithm makes use of supervised learning technique and is basically used in both classification and regression based applications. Under this algorithm, in an n-dimensional space containing n features, raw data are plotted as points in it, followed by tying of each feature's value to the corresponding coordinate. As a result, the data classification becomes easy. Note that lines obtained under this algorithm are termed as classifiers which are useful for splitting of data to make data plotting on graph conveniently.

G. Random Forest Algorithm

This algorithm makes use of the concept of combining a set of decision trees to obtain a random forest. The classification of each tree is made for the reason to do the classification of a new object on the basis of its attributes for which the tree is supposed to "vote" for the corresponding class. The forest has the responsibility to select the classification that has the maximum votes among all trees in the forest. The plantation and then growing of tree is done if:

1) The training set has number of cases as N out of which only random selection of a N sample is done where the sample is considered as the training set for doing the activity of growing the tree.
2) M variable are there such that M is less than N
3) No pruning of tree is done so as to allow the growing of each tree to the largest possible extent.

H. Naive Bayes Algorithm

This is again one of the most prominent machine learning algorithms where Naive Bayes classifiers are used to represent a series of probabilistic classifiers on the basis of applying of Byes theorem with the help of naive (i.e. strong) independence assumptions made for each feature. Even if any of the features are unrelated, a Naive Bayes classifier simply considers all the properties independently while doing the calculation of the probability of a specific outcome.

I. K-Means Algorithm

Using unsupervised learning technique and used for solving clustering related problems, this algorithm classifies data sets into a specific number of clusters (that number is generally termed as K) in the manner that all other data considered within the cluster are always homogeneous, while other data outside the clusters are heterogeneous.

K-means clustering algorithm works by picking 'K' number of points, usually termed as centroids, for all such obtained clusters and any new centroids are further created on the basis of already existing cluster numbers. The determination of closest distance for each data point is done with respect to newly created centroids and this task is performed again and again unless the centroids are changed.

J. Dimensionality Reduction Algorithms

These machine learning algorithms are meant to derive relevant information, variables and significant patterns out of a huge raw data collected as Big Data from the areas like government agencies, research organizations, and corporate houses. Some of the famous dimensionally reduction algorithms are Factor Analysis, Decision Tree, Random Forest and Missing Value Ratio.

III.AREAS OF APPLICATION OF MACHINE LEARNING ALGORITHMS

With reference to the previous section, machine learning algorithms have wide areas of application. The important applications corresponding to each machine learning algorithm are discussed in this section.

A. Gradient Boosting & Adaboosting Algorithm

These algorithms are quite useful in the case when data collected has a massive load that put the challenge before these algorithms to do highly accurate prediction and generates better results. Hence, they find their usages in competitions like AV Hackathon, Kaggle, and CrowdAnalytix.

B. Linear Regression

It is the one of the most powerful statistical technique to provide deep insights into the behaviour of consumer, business understanding and determination of factors that influences profitability.

C. Decision Tree

It is a quite beneficial algorithm to find applications in classification and identification of many hidden patterns. For example, a decision tree is useful in tracing out the location of natural minerals.

D. Logistic Regression

Since this algorithm is useful for estimating discrete values (in terms of 0 or 1) from a set of independent variables, it finds application in image segmentation and categorization, handwriting recognition, geographic image processing, health care industry to check whether a person is ailing from depression, etc.

E. K-Nearest Neighbors (KNN) Algorithm

As able to solve classification and regression problems, this algorithm finds application in recommender systems, concept search for searching of documents of semantically similar nature.

F. Support Vector Machine (SVM)

This algorithm finds application in the area of text and hypertext categorization, image classification, handwriting recognition.

G. Random Forest Algorithm

This algorithm is quite beneficial in machine fault diagnosis, diabetic retinopathy & biological science.

H. Naive Bayes Algorithm

This algorithm has great application in sex classification to determine whether a person is male or female on the basis of measured features like weight, height, foot size. Another application of this algorithm is credit scoring in the e-lending platform.

I. K-Means Algorithm

This algorithm finds applications in loyalty segmentation, branch geo segmentation, customer-need segmentation category segmentation etc.

J. Dimensionality Reduction Algorithms

These algorithms find application in areas like government agencies, research organizations and corporate houses for the reason that these areas generate a huge data that can be worked on by only these algorithms comprehensively to derive relevant information, variables and significant patterns out of the generated data.

IV. CONCLUSIONS

Machine learning has lots of algorithms to analyze and implement to gain the benefits of them in various fields of human endeavors. All such algorithms work directly or indirectly making the use of idea presented by the fundamental learning techniques of supervised learning, unsupervised learning and reinforcement learning. The major algorithms that are used on a frequent basis in machine learning field include Gradient Boosting and AdaBoosting, Linear Regression, Decision Tree, Logistics Regression, KNN, SVM, Random Forest, Naive Bayes, K-Means and Dimensionally Reduction algorithms. We analyzed the working of each algorithm and also discussed their applications in this paper. We can expect more such algorithms still to be developed in future to derive more benefits out of them.

V. ACKNOWLEDGMENT

We would like to acknowledge all the authors, researchers and the data scientists whose works mentioned in the references had helped us a lot to come up with this paper to analyze the machine learning algorithms. We would also like to say thanks to the journal – IJRASET and it's most respectable Editor-in-chief which has given a space to our paper in their esteem journal.

REFERENCES

[1] The 10 Algorithms Machine Learning Engineers Need to Know, simplilearn. Web. https://www.simplilearn.com/10-algorithms-machine-learning-engineers-need-to-know-article
[2] Machine learning. Wikipedia. Web. https://en.wikipedia.org/wiki/Machine_learning
[3] Bishop, C. M. (2006), Pattern Recognition and Machine Learning, Springer, ISBN 0-387-31073-8
[4] Mohri, Mehryar; Rostamizadeh, Afshin; Talwalkar, Ameet (2012). Foundations of Machine Learning. USA, Massachusetts: MIT Press. ISBN 9780262018258.
[5] Alpaydin, Ethem (2010). Introduction to Machine Learning. London: The MIT Press. ISBN 978-0-262-01243-0. Retrieved 4 February 2017.
[6] Honglak Lee, Roger Grosse, Rajesh Ranganath, Andrew Y. Ng. "Convolutional Deep Belief Networks for Scalable Unsupervised Learning of Hierarchical Representations" Proceedings of the 26th Annual International Conference on Machine Learning, 2009.
[7] Lu, Haiping; Plataniotis, K.N.; Venetsanopoulos, A.N. (2011). "A Survey of Multilinear Subspace Learning for Tensor Data" (PDF). Pattern Recognition. 44 (7): 1540–1551. doi:10.1016/j.patcog.2011.01.004.
[8] Yoshua Bengio (2009). Learning Deep Architectures for AI. Now Publishers Inc. pp. 1–3. ISBN 978-1-60198-294-0.
[9] Jump up ^ A. M. Tillmann, "On the Computational Intractability of Exact and Approximate Dictionary Learning", IEEE Signal Processing Letters 22(1), 2015: 45–49.
[10] Aharon, M, M Elad, and A Bruckstein. 2006. "K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation." Signal Processing, IEEE Transactions on 54 (11): 4311–4322
[11] Goldberg, David E.; Holland, John H. (1988). "Genetic algorithms and machine learning". Machine Learning. 3 (2): 95–99. doi:10.1007/bf00113892.
[12] Michie, D.; Spiegelhalter, D. J.; Taylor, C. C. (1994). Machine Learning, Neural and Statistical Classification. Ellis Horwood.

Volume 5 Issue VI, June 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

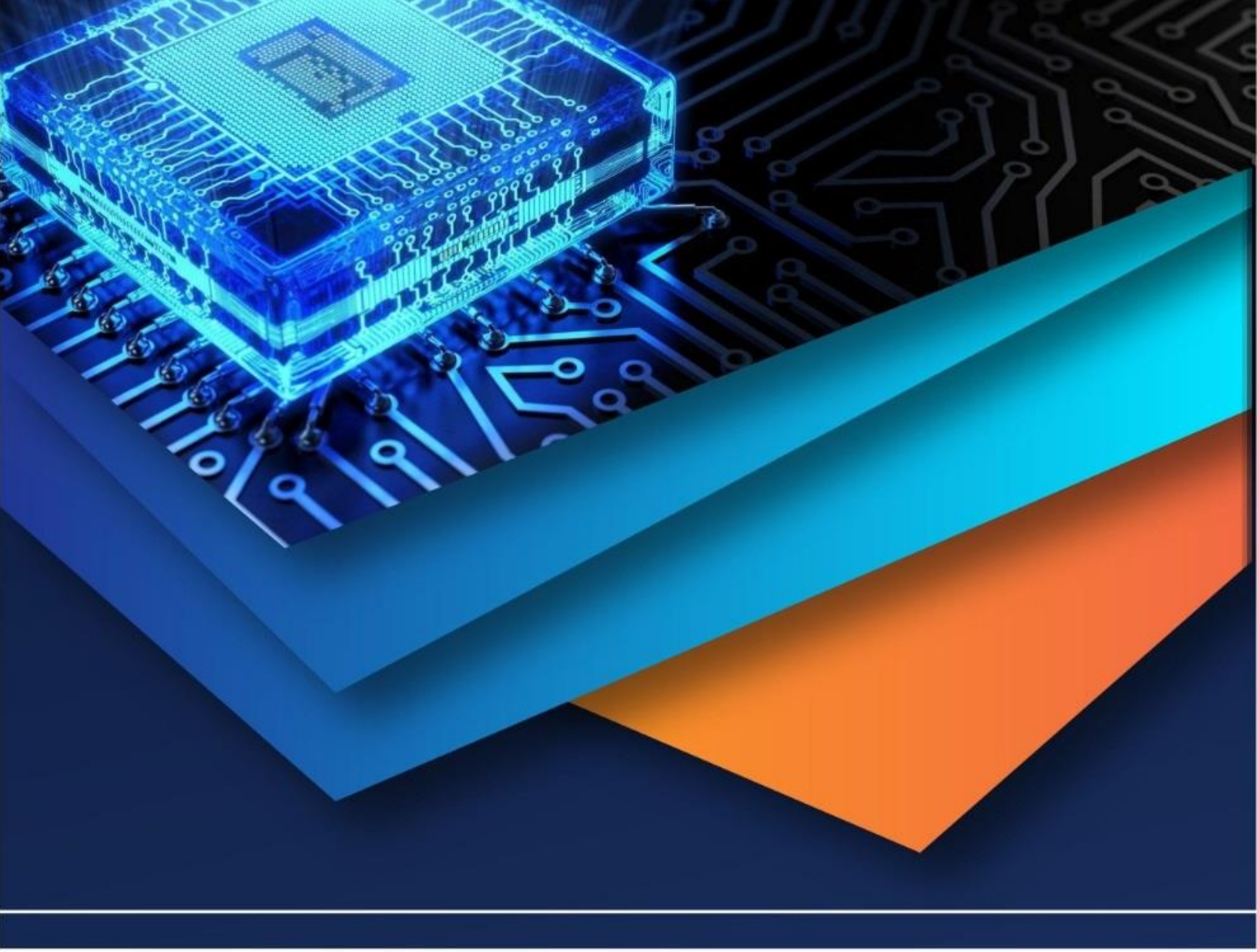
[13] Zhang, Jun; Zhan, Zhi-hui; Lin, Ying; Chen, Ni; Gong, Yue-jiao; Zhong, Jing-hui; Chung, Henry S.H.; Li, Yun; Shi, Yu-hui (2011). "Evolutionary Computation Meets Machine Learning: A Survey" (PDF). Computational Intelligence Magazine. IEEE. 6 (4): 68–75. doi:10.1109/mci.2011.942584.
[14] Bassel, George W.; Glaab, Enrico; Marquez, Julietta; Holdsworth, Michael J.; Bacardit, Jaume (2011-09-01). "Functional Network Construction in Arabidopsis Using Rule-Based Machine Learning on Large-Scale Data Sets". The Plant Cell. 23 (9): 3101–3116. doi:10.1105/tpc.111.088153. ISSN 1532-298X. PMC 3203449 Freely accessible. PMID 21896882.
[15] Urbanowicz, Ryan J.; Moore, Jason H. (2009-09-22). "Learning Classifier Systems: A Complete Introduction, Review, and Roadmap". Journal of Artificial Evolution and Applications. 2009: 1–25. doi:10.1155/2009/736398. ISSN 1687-6229.

ABOUT AUTHORS

Manish Kumar has published two research papers in the international journals on web mining – chiefly “Web Mining: Penning an Era of Information Age" and “Understanding How Crucial Hidden Value Discovery In Data Warehouse Is?" and he has also published four articles in international journal. His areas of interest include Big Data Analytics and Machine Learning.

GS Baluja is a famous Indian author of Computer Science field who has authored numerous books so far - chiefly Data Structure Through C, Data Structures Through C++, Object Oriented Programming Using C++, Java Programming & much more. He has done B.E (Com. Sci.) from Marathwada University and M. Tech from IIT, Delhi.

Dinesh Prasad Sahu received the Master degree (Computer Science & Application) M. Tech (Computer Science & Application) from Jawaharlal Nehru University, New Delhi, India. Currently, He is doing Ph.D. (Computer Science & Engineering) under the guidance of Dr. Karan Singh, from Jawaharlal Nehru University, New Delhi, India & is working in the school of Computer & Systems Sciences, Jawaharlal Nehru University, New Delhi. His primary research interests include parallel and distributed system and Grid Computing. He has published 3 papers in proceedings of peer-reviewed Conferences.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY