International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/IJSRCSEIT

Conceptual Review on Machine Learning Algorithms for Classification Techniques

T. Mohana Priya[1], Dr. M. Punithavalli[2], Dr. R. Rajesh Kanna[3]

[1]Research Scholar, Bharathiar University, Coimbatore, Tamil Nadu, India

[1]Assistant Professor, Dr. SNS Rajalakshmi College of Arts and Science, Coimbatore, Tamil Nadu, India

[2]Professor, Department of Computer Applications, Bharathiar University, Coimbatore, Tamil Nadu, India

[3]Professor and Head, Department of Information Technology, Dr.N.G.P. Arts and Science College, Coimbatore, Tamil Nadu, India

Article Info
Volume 7, Issue 1
Page Number: 215-222

Publication Issue :
January-February-2021

Article History
Accepted : 10 Feb 2021
Published : 18 Feb 2021

ABSTRACT

Machine leaning is a ground of recent research that officially focuses on the theory, performance, and properties of learning systems and algorithms. It is a extremely interdisciplinary field building upon ideas from many different kinds of fields such as artificial intelligence, optimization theory, information theory, statistics, cognitive science, optimal control, and many other disciplines of science, engineering, and mathematics. Because of its implementation in a wide range of applications, machine learning has covered almost every scientific domain, which has brought great impact on the science and society. It has been used on a variety of problems, including recommendation engines, recognition systems, informatics and data mining, and autonomous control systems. This research paper compared different machine algorithms for classification. Classification is used when the desired output is a discrete label.

Keywords : Machine Learning, KNN, ANN, Naive Bayes, Classification

I. INTRODUCTION

Machine learning techniques have been widely adopted in a number of massive and complex data-intensive fields such as medicine, astronomy, biology, and so on, for these techniques provide possible solutions to mine the information hidden in the data. Machine learning is used to teach machines how to handle the data more efficiently. Sometimes after viewing the data, we cannot interpret the pattern or extract information from the data. However, as the time for big data is coming, the collection of data sets is so large and complex that it is difficult to deal with using traditional learning methods since the established process of learning from conventional datasets was not designed to and will not work well with high volumes of data. For instance, most traditional machine learning algorithms are designed for data that would be completely loaded into memory, which does not hold any more in the context of big data.

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E". Learning is used when a pattern exists in the given data which we can't pin down manually. It is because of the Machine Learning that the present century has been witness to the landmark discoveries like Autonomous Helicopter, Handwriting Recognition, Natural Language Processing (NLP), Computer Vision, Speech Recognition, Recommendation Systems, Decision Support Systems (DSS), email filtering, detection of network intruders or malicious insiders working towards a data breach, optical character recognition etc. Machine Learning is at the root of the tree and has lots of branches and sub branches.

II. CLASSIFICATION OF MACHINE LEARNING TECHNIQUES

Machine learning algorithms can be broadly classified into the following four categories:

1. Supervised learning
2. Semi-supervised Learning
3. Unsupervised Learning
4. Reinforcement Learning

Supervised learning

Supervised learning is very common in classification problems, where the aim is to classify the given input data or image into predefined output classes. For example: Handwritten Digit and character recognition, classification of animals or objects depending upon the input image, medical image classification for diagnosis of various diseases, etc-. In supervised learning, there is some supervision while the algorithm is being trained. Here, labeled data is used for training. Different types of supervised learning approaches are as follows: *f*

- Neural Networks
- Naive Bayes Classifier
- Decision Trees
- Linear Regression

Unsupervised learning

In unsupervised learning, no element of supervision is involved. The computer is trained with unlabeled data. Unsupervised learning plays a vital role in those cases where the human expert doesn't know what to look for in the data. The common types of supervised learning approaches are: Association Rules, K-means clustering,etc.

Semi-supervised Learning

Semi-supervised Learning lies between the two approaches mentioned above. In this approach a combination of labeled and unlabelled data is used for training.

Reinforcement Learning

Reinforcement Learning involves the mechanism of reward and punishment for the process of learning. In this type of learning, the objective is to maximize the reward and minimize the punishment.

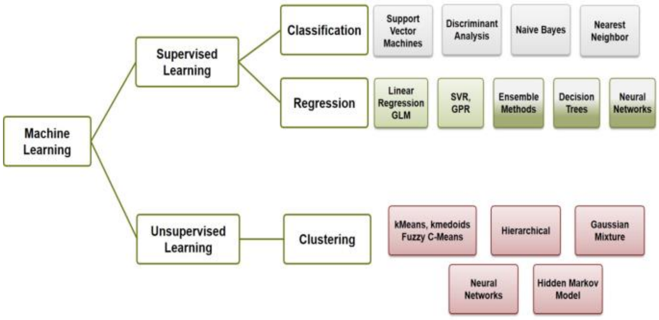

Figure 1 : Machine learning Algorithms

III. RELATED WORK

In this paper we reviewed various articles which is related to machine learning classification algorithm for predict the dataset values for specific problem. In the machine learning model, a classifier was given training with already classified examples and it learns the rules for classification during this training phase.

Lertworaprachaya et al., 2014 [1] proposed a new model of Decision tree algorithm for compose decision trees using interval valued fuzzy membership values. Most existing fuzzy decision trees do not consider the concerned associated with their membership values; however, precise values of fuzzy membership values are not always possible. Because of that, the authors represented fuzzy membership values as distance to model concerned and employ the look ahead based fuzzy decision tree induction method to construct decision trees. The authors also measured the significance of different neighborhood values and define a new parameter unkind to specific data sets using fuzzy sets. Some examples are provided to establish the effectiveness of their approach.

Pan et al., 2015 [2] proposed a novel K-nearest neighbour establish structural twin support vector machine (KNNSTSVM). By applying the intra-class KNN method, different weights are given to the samples in one class to enhance the structural information. For the other class, the expendable constraints are deleted by the inter-class KNN method to speed up the coaching process. For large scale problems, a fast clip algorithm is further introduced for increase of rate.

Xu et al., 2014 [3] addressed a novel boosting algorithm called UAdaBoost which possibly would better the classification performance of AdaBoost with Universum data. UAdaBoost determine a function by minimizing the loss for labeled data and Universum data. The cost function is discount by a greedy, stage wise, functional gradient procedure. Each training stage of UdaBoost is fast and efficient. The standard AdaBoost weights labeled samples over training iterations while UAdaBoost gives an explicit weighting program for Universum samples as well. Also the authors described the practical conditions for the effectiveness of Universum learning. These conditions are based on the analysis of the distribution of ensemble forecasting over training samples. By their experimental results the authors declare that their method can obtain superior performances over the standard AdaBoost by selecting proper Universum data.

Kalai Magal. R, Shomona Gracia Jacob[4] proposed a new “Improved Random forest” algorithm to enhance the classification accuracy for software defect predication. The algorithm worked by incorporating with best feature selection algorithm and the Random Forest to give better accuracy. Correlation based Feature Subset Selection (CFS) algorithm selects the optimal subset of features. The optimal features were then fed as a part of Random Forest classification to give better accuracy in software defect.

L Jiang et al.[5] developed Naïve Bayes(NB) is considered to be a relatively simple machine learning technique based on probability models Bayesian theorem. This classification technique analyses the relationship between each and the class for each instance to derive a conditional probability for the relationships between the feature values and the Class. The conceptual framework for NB is based on joint probabilities of features and Classes to estimate the probabilities of a given document belonging to a given Class

Cover T et.a.[6] proposed K-nearest neighbor is a classical instance-based learning algorithm in which a new case is classified based on the known class of the nearest neighbor, by means of a majority vote. This type of algorithm is also called lazy learning because there is no model building step and the entire computing procedure is performed directly during the prediction. All the cases need to be available during the prediction.

Elder J [7] stated that a linear classifier achieves this goal by making a classification decision based on the value of the linear combination of the features. Linear models for classification separate input vectors into classes using linear decision boundaries [6].The goal of classification in linear classifiers in machine learning, is to group items that have similar feature values, into groups. A linear classifier is often used in situations where the speed of classification is an issue, since it is rated the fastest classifier.Also, linear classifiers often work very well when the number of dimensions is large, as in document classification, where each element is typically the number of counts of a word in a document. The rate of convergence among data set variables however depends on the margin. Roughly speaking, the margin quantifies how linearly separable a dataset is, and hence how easy it is to solve a given classification problem.

Neocleous C et.al [8] discus about Artificial Neural Network (ANN) which depends upon three fundamental aspects, input and activation functions of the unit, network architecture and the weight of each input connection. Given that the first two aspects are fixed, the behavior of the ANN is defined by the current values of the weights. The weights of the net to be trained are initially set to random values, and then instances of the training set are repeatedly exposed to the net. The values for the input of an instance are placed on the input units and the output of the net is compared with the desired output for this instance. Then, all the weights in the net are adjusted slightly in the direction that would bring the output values of the net closer to the values for the desired output.

Taiwo, O. A. [9] stated multilayer perception classifier in which the weights of the network are found by solving a quadratic programming problem with linear constraints, rather than by solving a non-convex, unconstrained minimization problem as in standard neural network training. Other well known algorithms are based on the notion of perception. Perception algorithm is used for learning from a batch of training instances by running the algorithm repeatedly through the training set until it finds a prediction vector which is correct on all of the training set. This prediction rule is then used for predicting the labels on the test set.

E. Sarac and et al [10] introduced that increase in the amount of information on the Web has caused the need for accurate automated classifiers for Web pages to maintain Web directories and to increase search engine performance. Every tag and every term on each Web page can be considered as a feature there is a need for efficient methods to select the best features to reduce the feature space of the WPC problem. The aim is to apply a recent optimization technique, namely the firefly algorithm (FA) to select the best features for Web page classification problem. The firefly algorithm (FA) is a metaheuristic algorithm, inspired by the flashing behavior of fire flies. Using FA to select a subset of features and to evaluate the fitness of the selected features J48 classifier of the Weka data mining tool is employed.

I.Anagnostopoulos, et al.[11] suggested a system to identify and categorize web pages, based on information filtering. The system is a three layer Probabilistic NN (PNN) having biases and radial basis neurons in the middle layer and competitive neurons in the output layer. This is an eCommerce area study domain. Thus, PNN hopes to identify eCommerce web pages to classify them to respective type based on a framework describing commercial transactions fundamental transactions on the web.

IV. Discussions on Machine Learning Algorithms

In the following table-1, we have compared different machine learning algorithms and listed below:

Table 1 : Interpretation of various machine learning classification algorithms

Algorithms	**Advantages**	**Disadvantages**
The *k-Means* method	• Relatively efficient • Can process large data sets.	• Often terminates at a local optimum. • Applicable only when mean is defined. • Not applicable for categorical data. • Unable to handle noisy data. • Not suitable to discover clusters with non-convex shapes.
k-Nearest Neighbor (*k*-NN) classifier	• Nonparametric • Zero cost in the learning process • Classifying any data whenever finding similarity measures of any given instances • Intuitive approach • Robust to outliers on the predictors	• Expensive computation for a large dataset • Hard to interpret the result • The performance relies on the number of dimensions • Lack of explicit model training • Susceptible to correlated inputs and irrelevant features • Very difficult in handling data of mixed types.
Support vector machine (SVM)	• Can utilize predictive power of linear combinations of inputs • Good prediction in a variety of situations • Low generalization error • Easy to interpret results	• Weak in natural handling of mixed data types and computational scalability • Very black box • Sensitive to tuning parameters and kernel choice • Training an SVM on a large data set can be slow • Testing data should be near the training data
Decision Trees	• Some tolerance to correlated inputs. • A single tree is highly interpretable, • Can handle missing values. • Able to handle both numerical and categorical data. • Performs well with large datasets.	• Cannot work on (linear) combinations of features. • Relatively less predictive in many situations. • Practical decision-tree learning algorithms cannot guarantee to return the globally-optimal decision tree. • Decision-tree can lead to over fitting.
Logistic regression	• Provides model logistic probability • Easy to interpret • Provides confidence interval • Quickly update the classification model to incorporate new data	• Does not handle the missing value of continuous variables • Sensitive to extreme values of continuous variables

Naïve Bayes	• Suitable for relative small training set • Can easily obtain the probability for a prediction • Relatively simple and straightforward to use • Can deal with some noisy and missing data • Can handles multiple classes	• Prone to bias when increasing the number of training sets • Assumes all features are independent and equally important, which is unlikely in real-world cases. • Sensitive to how the input data is prepared.
Neural networks	• Good prediction generally • Some tolerance to correlated inputs • Incorporating the predictive power of different combinations of inputs	• Not robust to outliers • Susceptible to irrelevant features • Difficult in dealing with big data with complex model

V. Interpretation and Views

Linear regression predictions are continuous values, logistic regression predictions are discrete values after applying a transformation function.

Naïve Bayes algorithm is used to calculate the probability of an outcome given the value of some variable, that is, to calculate the probability of a hypothesis (h) being true, given our prior knowledge(d).

The k-nearest neighbors algorithm can be applied to both classification and regression problems. In fact, within the Data Science industry, it's more widely used to solve classification problems. It’s a simple algorithm that stores all available cases and classifies any new cases by taking a majority vote of its k neighbors.

Decision Tree algorithm is One of the most popular machine learning algorithms in use today, this is a supervised learning algorithm that is used for classifying problems. A decision tree is used to classify future observations given a body of already labeled observations.

A collective of decision trees is called a Random Forest. To classify a new object based on its attributes, each tree is classified, and the tree “votes” for that class.

Boosting is an ensemble learning algorithm that combines the predictive power of several base estimators to improve robustness.

Support Vector Machine is a process of classification in which is plot raw data as points in an n-dimensional space. The value of each feature is then tied to a particular coordinate, making it easy to classify the data.

VI. CONCLUSION

Machine learning generates a lot of buzz because it's applicable across such a wide variety of use cases.In machine learning, classification is a supervised learning approach in which the computer program learns from the data input given to it and then uses this learning to classify new observation. Choosing an algorithm is a critical step in the machine learning process, so it’s important that it truly fits the use case of the problem.

VII. REFERENCES

[1]. Y. Lertworaprachaya, Y. Yang, R. John, “Interval valued fuzzy decision trees with optimal neighborhood perimeter,” Applied Soft Computing, vol. 24, pp. 851-866, 2014

[2]. X. Pan, Y. Luo, Y. Xu, “K-nearest neighbor based structural twin support vector machine,” Knowledge Based Systems, vol. 88, pp. 34-44, 2015

[3]. J. Xu, Q. Wu, J. Zhang, Z. Tang, “Exploiting Universum data in AdaBoost using gradient descent,” Image and Vision Computing, vol. 32, pp. 550-557, 2014

[4]. Kalai Magal. R, Shomona Gracia Jacob, “Improved Random Forest Algorithm for Software Defect Prediction through Data Mining Techniques”, International Journal of Computer, pp.0975–8887, Vol.No.23, May2015

[5]. L Jiang, D Wang, Z Cai, and XYan: Survey of improving naive bayes for classification.Advanced Data Mining and Applications 2017:134-145.

[6]. Cover T, Hart P. Nearest neighbor pattern classification. IEEE Transactions on Information Theory. Available from: http://ieee.org/document/1053964/. Accessed in 2017 (Feb 20).

[7]. Elder, J. (n.d). Introduction to Machine Learning and Pattern Recognition, IEEE, Available at LASSONDE University EECS Department York website: http://www.eecs.yorku.ca/ course_archive/ 2011-12/F / 4404, 2015

[8]. Neocleous C. & Schizas C. Artificial Neural Network Learning: A Comparative Review. In: Vlahavas I.P., Spyropoulos C.D. (eds) Methods and Applications of Artificial Intelligence,. Springer, Berlin, Heidelberg, doi: 10.1007/3-540-46014-4_27 pp.300-313,2012

[9]. Taiwo, O. A. Types of Machine Learning Algorithms, New Advances in Machine Learning, Yagang Zhang (Ed.), ISBN: 978-953-307 -034-6, InTech, University of Portsmouth United Kingdom. Pp 3 –31, 2014

[10]. Klassen, “A frame work for search forms classification” In Systems, Man, and Cybernetics (SMC), 2012 IEEE International Conference on , PP.1029-1034 Seoul, Korea, 14-17 Oct. 2012

[11]. Anagnostopoulos, C. Anagnostopoulos, V. Loumos and E. Kayafas, “Classifying Web pages employing a probabilistic neural network”, Probabilistic Software, IEEE ,Vol.151, No.3,June 2014, PP.139-150

[12]. M. Nayrolles and A. Hamou-Lhadj, "BUMPER: A Tool for Coping with Natural Language Searches of Millions of Bugs and Fixes," Software Analysis, Evolution, and Reengineering (SANER), IEEE pp. 649-652. 2016

[13]. T. Cheng, K. Chakrabarti, S. Chaudhuri, V. Narasayya and M. Syamala, "Data services for E-tailers leveraging web search engine assets," Data Engineering (ICDE),IEEE, 2013, pp. 1153-1164

[14]. Jebaraj Ratnakumar, “An Implementation of Web Personalization Using Web Mining Techniques”, Journal Of Theoretical And Applied Information Technology, 2014 JATIT

[15]. Tsuyoshi, M and Saito, K., “Extracting User"s Interest for Web Log Data”, IEEE 2016, pp. 343-346, ISBN: 0-7695-2747-7

[16]. R.Rajeshkanna and Dr.A.Saradha “ Multipath Load Balanced Congestion Control Routing Techniques in Mobile Adhoc Network”, International Journal of Scientific Research and Development,Vol.2015 Issues 5, ISSN:2321-0613

[17]. Lidong Wang Data Mining, Machine Learning and Big Data Analytics International Transaction of Electrical and Computer Engineers System. 2017, 4(2), 55-61. DOI: 10.12691/iteces-4-2-2ublished online: July 24, 2017

[18]. https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/

Cite this article as :

T. Mohana Priya, Dr.M.Punithavalli, Dr. R. Rajesh Kanna, "Conceptual Review on Machine Learning Algorithms for Classification Techniques", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 7 Issue 1, pp. 215-222, January-February 2021.

Journal URL : https://ijsrcseit.com/CSEIT217128