DOI: http://dx.doi.org/10.26483/ijarcs.v9i4.6281

ISSN No. 0976-5697

Volume 9, No. 4, July – August 2018

International Journal of Advanced Research in Computer Science

SURVEY REPORT

Available Online at www.ijarcs.info

A STUDY ON MACHINE LEARNING ALGORITHM IN MEDICAL DIAGNOSIS

Ms.D.Pavithra
Department of Computer Science and Engineering
Dr.N.G.P. Institute of Technology
Coimbatore,
Tamil Nadu, India

Dr.A.N.Jayanthi
Department of Electronics and
Communication Engineering,
Sri Ramakrishna Institute of Technology,
Coimbatore, Tamilnadu, India

Abstract : Machine learning is a method of optimizing the performance criterion using the past experience. It builds the mathematical model by using the theory of statistics, as the main task is to infer from the samples provided. The algorithm uses computational methods to get the information directly from the data. They are mainly used in medical diagnosis for making critical decisions, as the data in the medical field is huge and the accuracy of the diagnosis depends on considering the huge data of the patients. ML improves the accuracy of the diagnostic of the disease. It also provides automatic learning techniques for predicting the common patterns from the realistic data. There are different ML algorithms, the appropriate method has to be chosen based on their performance. This paper focuses on the use of different machine learning algorithms like Support Vector Machine, Naïve Bayesian, J48, Random Forest etc. for accurate medical diagnosis.

Keywords: Machine Learning; Support Vector Machine; Random forest; Naïve Bayesian; Mathematical Model.

I. INTRODUCTION

Machine Learning is a discipline of Artificial Intelligence which is the simulation of human intelligence by computer systems. It is the combination of computer science and statistics. Computer Science mainly focuses on solving the problems and identifying whether the problems are solvable at all stages. The idea of statistics is data modelling, hypothesis and measuring the reliability. Machine learning is a paradigm that learn from past experience for the improvement of future performance. The main aim of this field is automatic learning methodologies. Learning means the modification or improvement to the algorithm based on previous experiences without any involvement of human. ML mainly focuses on the development of the programs that uses data to learn themselves. Machine Learning provides algorithms and tools to make the system work intelligently. It is mainly used for the problems without deterministic solution where there is no specific models for the problem. Algorithms are developed based on diverse disciplines and it is used mainly for accuracy, speed and customizability. Machine Learning also contributes in medical diagnosis for disease prediction, data analysis, therapy planning, etc. Machine Learning integrate computer system with the medical field for the efficient diagnosis and the quality treatment by the medical experts.

Medical diagnosis is a critical and important task by different intelligent systems . Any medical treatment starts with screening, diagnosis, treatment, and frequent monitoring. Now days medical field rely on computer technology, ML are used to identify the abnormalities at early stage. Accurate diagnosis is very important for deciding right therapy at the early stage. But in many cases it is very difficult for the expert to identify the level of the patient. With the clinical records, the ML methods can be used to produce a descriptive analysis of clinical features. Machine Learning algorithms is widely used in diagnosis of various diseases like diabetics, heart problems, cancer. Among various algorithms, more frequently used algorithm is support vector machine and decision tree. Basically there are many types of ML algorithms:

Supervised Learning algorithms are trained using a training set of data, based on which the results are predicted. The aim is to predict the output value for the given input vector. The output can be a continuous value or a discrete value. Continuous value is for regression problem and the discrete value is for classification problems. The training data set contains the sample input and output values. Popular supervised learning algorithms are classification and regression. Based on the training dataset, the output for the new input value is predicted[1]. Generally, supervised learning is of two types, parametric models and non-parametric models. In the parametric models, the predictive function is the combination of fixed number of parametric. First stage is the learning stage using the training dataset. After this stage the training data can be discarded as the prediction for new input is based on the learned parameters. Linear regression and classification are some of the parametric models. The most successful parametric model is the neural networks. In nonparametric models, the number of parameters is dependent on the training dataset. The training data set is maintained for the prediction. The most commonly used nonparametric models are support vector machines and nearest neighbor algorithm. These algorithms can be used for both regression and classification

Unsupervised Learning algorithms predicts the results based on the similarities between the input[2]. It contains only a set of input vectors. Clustering is the commonly used unsupervised learning. Clustering is the method of grouping the similar data based on their distance. General property of clusters is that the intra cluster similarity should be high and the inter cluster similarity should be low[3]. The clusters depends on the input data values. Social network analysis, genetic clustering, market analysis are some of the common applications of unsupervised learning.

Semi supervised learning is a combination of both supervised and unsupervised learning. Semi supervised learning are mainly used for supervised learning applications where labeled data are not available or expensive to obtain. It is between labeled and unlabelled data. The oldest form of this algorithm is the self-training model[4]. This is an repetitive process where initially the labeled data values are used with supervised learning and next the unlabelled data values are labeled with the previous known values and finally all the data values are used for prediction. Basically, the semi-supervised learning models can be categorized under four classes: i) generative model ii) model where the decision boundary is in low-density region iii) graph-based model iv) two-step model with unsupervised learning followed by supervised learning.

Reinforcement learning explore the given test data and find the correct output using evaluative feedback. The main features are delayed reward and trial-and-error. This algorithm follows the Markov Decision Process (MDP). It is intimated when the result is wrong, then the algorithm explores the possibilities to find the right result. This algorithm is mainly used in finance, robotics, inventory management.

Deep learning is another form of machine learning where there is a high-level abstraction. It uses various processing layer with linear and nonlinear transformations.

There is a large amount of data for clinical assessment of a particular patient. These big data has to maintained and considered during the diagnosis process. There is in need of proper management to extract and process the data efficiently and effectively [5]. This can be done using the machine learning algorithms. Data are divided and analyzed in the ML classifiers [6]. This paper focus on different machine learning algorithms used in the medical diagnosis and their advantages and disadvantages. Next, the considerations that are to be focused when choosing the algorithm. Finally the conclusion.

II. MACHINE LEARNING ALGORITHMS

Identifying the right algorithm for the medical diagnosis is a part of science and art. The knowledge of different classes of machine learning helps to identify the best algorithm that suits the accurate diagnosis of the disease. This section focuses on explaining the different machine learning algorithms that are commonly used for medical diagnosis.

A. Naïve Bayes Classifier

Naïve Bayes is a supervised learning method that is based on Bayes' Theorem after Thomas Bayes. This is a simple classifier that assigns labels to data which are previously unknown. These class labels are a some finite set. The classifier assumes the presence of a feature is independent of other feature. It considers each feature is contributing separately. This model is mainly used for large data sets . this classifier mainly apply Bayes theorem for the classification and regression problems. The Bayes theorem states that:

$$P(m|n) = (P(n|m) * P(m)) / P(n) \qquad (1)$$

Where P(m|n) is posterior probability where the probability of hypothesis m given the data n. P(n|m) is the probability of data n given that the hypothesis m was true. P(m) is the probability of hypothesis m being true. This is called the prior probability of m. P(n) is the probability of the data [7].

Naive Bayes is an algorithm for binary and multi-class problems. It is the most successful known algorithms for learning to classify text documents, spam filtering in the text documents. This algorithm combined with collaborative filtering provides better performance in accuracy and coverage [8].

Merits of Naïve Bayes Classification

- Highly scalable, simple and easy to implement
- Can be used in probabilistic prediction with less training data
- If the NB conditional independence assumption holds, then it will converge quicker than discriminative models like logistic regression.
- Not sensitive to irrelevant features.

Challenges in Naïve Bayes Classification

- Incomplete training data
- It cannot consider continuous variables
- Assumes that all the attributes are mutually independent

B. Support Vector Machine

Support Vector Machines are based on the decision planes that define decision boundaries. A decision plane is one that separates a set of objects having different class memberships. It is a supervised learning model which analyze the data before classification. Classification is identifying the data to which it belongs. It performs the classification by constructing hyper planes in a multidimensional space that separates cases of different class labels. This plan is called as optimal hyper plane. The objective of the algorithm is to find the optimal plane that can be controlled when the condition is true, which is used to separate the two different classes. SVM supports both regression and classification tasks and can handle multiple continuous and categorical variables.

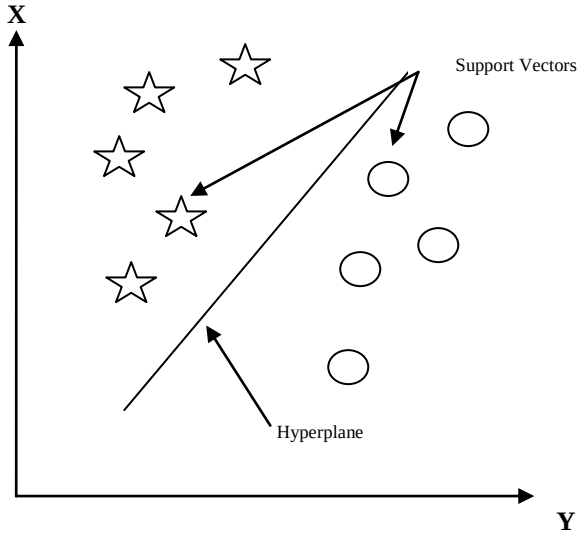

Figure 1: Support Vectors in SVM

The data points near the hyperplane are called as Support Vectors[9], changing or altering the data points would change the hyperplane. The hyperplane is a line that linearly separates and classifies a set of data. Distance between the hyperplane and the data point near the hyperplane on either side is known as the margin[9]. The aim of SVM is to identify the exact hyperplane with the greatest margin between the plane and any point within the training set, so that there is a greater chance of classifying new data correctly.

SVM is mainly used in face recognition, text and hyper text categorization for both inductive and transductive models, bioinformatics for cancer detection, recognize hand written characters, Generalized predictive control

Merits of Support Vector Machine

- Accurate classifiers
- Less overfitting
- It can be more efficient because it uses a subset of training points

Challenges in Support Vector Machine

- SVM cannot be extended to multi-class
- SVM cannot handle large sets if using kernels
- Long training time on large sets of data

C. Neural Network

Neural network is based on the structure and functionalities of biological neurons. The NN is based on the human brain that stimulates the learning process. The term Neural Network is referred as it is based on the neuron of the nervous system which is used to process and transmit the information. Generally, neurons have dendrite that collects the information, soma that process the information, and finally the axon which provides the output after processing. The soma are the processing nodes that has its own knowledge about the input and the rules which are already programmed in it. These nodes are interconnected, where each node in each level will be connected to every other node in other level. There can one or multiple nodes in the axon i.e., output layer which provides the solution. These NN are adaptive, i.e., they can modify according to the input they receive. The basic learning model depends on the weights of the input stream. The input with higher weight will contribute the right answer.

NN is initially fed with large amount of input for learning or training. Each node decides what to send as the output to the next level node based on the input received by that node. Generally, NN is defined in terms of depth, number of layers between the input and output layers. They can have hidden nodes or hidden layers also.

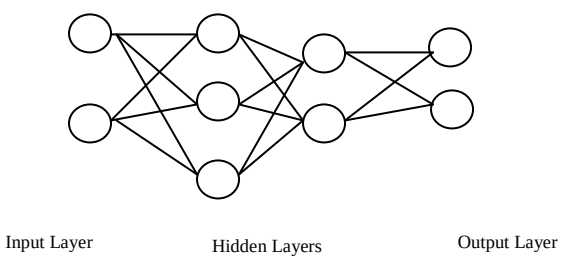

Figure 2: Layers in Neural Network

Artificial Neural Network (ANN) was developed as a broader research of neural network with the combination of artificial intelligence. ANN is group of neurons that are interconnected between each other for processing the information. The connection between them is called as Synapse. The variants of ANN were FeedForward and Feedback [10].

Neural Networks and Artificial neural networks (ANNs) are mainly used for solving more complex problem. Its adaptive nature changes its structure according to the internal and external flow of information. It is widely used in forecasting, Image Processing and Character recognition

Merits of Neural Network

- NN can learn and model non-linear problems
- It has the ability to identify the unseen relationship on unseen data for generalization
- Does not restrict on the input variables

Challenges in Neural Network

- Requires large amount of data
- Fine tuning of the architecture is required for better performance
- More computationally expensive

D. J48 Algorithm

Decision tree algorithm is based on branching methodology which produces maximum outcomes based on the condition. It form a tree like structure, where the node belongs to the test made on the attribute, the branches are the possible outcomes and leaf represents the particular class label. The rules are mentioned through the branches [11]. This is based on the supervised learning algorithm which is used for classification. Input can be in the form categorical or continuous variables. This algorithm initially construct a decision tree with the actual value, after which it splits the tree until a decision is made for the prediction for the given value. This algorithm can be used for both the classification and regression problems. This is more fast and accurate machine learning algorithm as it spits the data into different distinct groups.

J48 is a type of decision tree algorithm and it is a variation of ID3 algorithm. The variants include the identification of missing values, pruning the decision tree after construction, considering the continuous range values, deriving the rules automatically, etc. when compared to other classification algorithms, J48 generate rules for the classification of data. The data with the same will be labeled with the corresponding label name. Based on the test values of the attribute, the values are calculated. Next, the gain value is also calculated. With the selection criteria, all the attributes are compared and the best attribute is selected for branching [12]. After constructing the complete tree, pruning is performed. The main goal of pruning is to remove the classification errors and do the generalization. This algorithm is more flexible and accurate when compared to other decision tree algorithms [13].

Merits of J48 Algorithm

- Generates understandable rules automatically
- Considers the missing values
- Can work for continuous value ranges
- Pruning the decision tree is performed

Challenges in J48 Algorithm

- Tree structure prone to sampling
- Splitting of tree is difficult

E. Random Forest Algorithm

Random forest algorithm was proposed by Leo Breiman in the 2000's[14]. it is also known as random decision forests. This algorithm is an improvement of the decision tree algorithm. It is mainly used to build predictive model for both classification and regression problems. It generates decision trees which grows randomly in selected subspaces. The main idea is to generate small decision tress in random subsets of data, where each decision tree have different data trends and a biased classifier. Each tree is formed in random with input variables which is used to split. Then, the best split is calculated based on the training set features. General methods use multiple learning models but in random forest, it creates a model for entire forest to arrive the best solution.

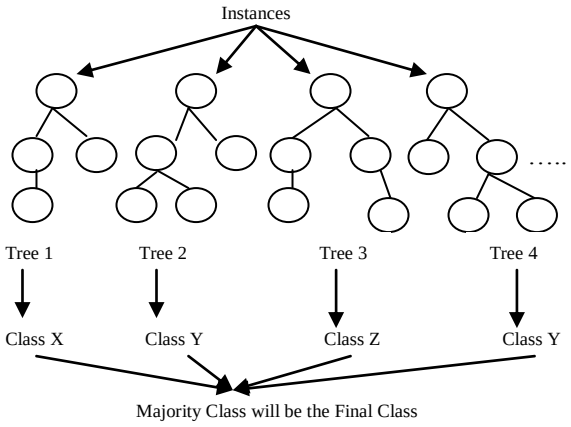

Figure 3: Simplified Random Forest

In the random forest algorithm, there are two stages, first is the random forest creation and then the prediction from the random forest classifier. In the random forest creation, select k features from the total of m features where k<m. Among the k features, select the node d by the best split point. Then, split the daughter nodes using the best split. Repeat the above said procedure until the specified number of nodes is reached. Similarly repeat the steps for building the forest at random subspace. After random forest creation, take the test features and use the rules to predict the outcome for generated decision trees. Calculate the votes for each predicted target. The final prediction will be the high voted predicted target.

Merits of Random Forest Algorithm

- They are fast and easy to implement
- Accurate solution
- Handle large data without overfitting
- Does not restrict on the input variables

Challenges in Random Forest Algorithm

- More complex
- Require more computational resources
- Predictive process is little time consuming for large decision trees

III. CONSIDERATIONS FOR CHOOSING THE ALGORITHM

Machine Learning does not provide one solution or one approach for all the problems. There are several factors to be considered for choosing the right algorithm for the problem. Some problems are specific and it requires unique approach. Whereas, some problems are open and need trial and error approach. Algorithms like supervised learning, classification, regression are open. These algorithms can be used in anomaly detection and even for building general predictive models. In this section, we will discuss the factors that are to be considered for choosing the right algorithm.

A. Accuracy

Getting accurate solution is not necessary always. There are problems which needs only an approximate solution .It is based on the problem. If that is the case, we can reduce our processing time by considering only the approximate methods. The main advantage of approximate methods are they tend to avoid overfitting.

B. Training Time

The number of hours to train the model depends on the algorithm. Training time is closely related to accuracy, both relies on each other. Certain algorithms are highly sensitive to the number of data points. When the time is limited, we can choose an algorithm especially for the large data set.

C. Linearity

Most of the machine learning algorithm uses linearity. Basically, linear classification problems assumes that the classes can be identified using a simple straight line. It includes regression and support vector machine. The linear regression algorithms assumes the data follow the straight line. These assumptions sometimes bring the accuracy down. The type of data should be considered for selecting the algorithm.

D. Number of Parameters

The parameters are the key points to turn when setting up the algorithm. There are many factors that affect the algorithm's behavior like, error tolerance, number of iterations, options between variants, etc. Training time and accuracy may sometimes be quite sensitive for getting right solution. But, algorithms with more number of parameters require more trials to find the correct combination for accurate results.

E. Number of Features

Certain data types contains more feature than the data points, mainly with the genetic and textual data. When the features to be considered are high, the performance of the algorithm gets

down in some cases, as it takes more time for training the algorithm. Support vector machine algorithms are well suited for such problems.

IV. CONCLUSION

The statistical models cannot handle categorical data with missing values and large data points. They do not provide accurate results for large dataset. That is the reason for introducing the Machine Learning algorithms. ML is used in many applications from image detection to disease diagnosis. The main goal of the ML researchers is to provide efficient learning model for better performance. The efficiency mainly focuses on the high accuracy of prediction for critical problems. It is completely data driven and have the capability of examining the large amount of data. These algorithms are more accurate and does not prone to errors. ML algorithms are provided with priority for the development of devices that are mainly used in self-monitoring, self-diagnosis and self-repairing. The statistics and the computer knowledge contributes the ML for developing different learning models for the prediction of solutions for the complex problems. This paper provides the different machine learning algorithms that are commonly used in medical diagnosis of different diseases like diabetes, liver failure, autism identification, heart problems, etc.

V. REFERENCES

[1] J. Han, J. Pei, and M. Kamber, Data mining: concepts and techniques. Elsevier, 2011.

[2] S. Marshland, Machine Learning an Algorithmic Perspective. CRC Press, New Zealand, 6-7, 2009.

[3] I. Witten, E. Frank, and M. Hall. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, San Mateo, CA, 3rd edition, 2011

[4] O. Chapelle, B. Scholkopf, and A. Zien, Semi-supervised learning. MITPress, 2006.

[5] M. Rambhajani, W. Deepanker and N. Pathak, "A Survey on Implementation of Machine Learning Techniques for Dermatology Diseases Classification." In International Journal of Advances in Engineering & Technology , 8, 194-195, 2015.

[6] I. Kononenko, "Machine Learning for Medical Diagnosis: History, State of the Art and Perspective" in Journal of Artificial Intelligence in Medicine , 1, 89-109, 2001.

[7] K. Vembandasamy, R. Sasipriya and E. Deepa, "Heart Diseases Detection Using Naive Bayes Algorithm" in International Journal of Innovative Science , Engineering & Technology, 2, 441-444, 2015.

[8] Ajinkya Kunjir, Basil Shaikh, "A Survey on Machine Learning Algorithms for Building Smart Systems" in International Journal of Innovative Research in Computer and Communication Engineering, 5, 1052- 1058, January 2017.

[9] Types of Machine Learning Algorithms, Taiwo Oladipupo Ayodele, University of Portsmouth, United Kingdom.

[10] Sunpreet Kaur, Sonika Jindal, "A Survey on Machine Learning Algorithms" in International Journal of Innovative Research in Advanced Engineering, 3, 6-14, 2016

[11] Prerna Kapoor, Reena Rani, "Efficient Decision Tree Algorithm Using J48 and Reduced Error Pruning" in International Journal of Engineering Research and General Science, 3, 1613- 1621, 2015.

[12] Y. Bengio. Learning deep architectures for AI. Foundations and Trends in Machine Learning, 2:1–127, 2009

[13] Gaganjot Kaur, Amit Chhabra, "Improved J48 Classification Algorithm for the Prediction of Diabetes" in International Journal of Computer Applications , 98, 13-17, 2014.

[14] Gerard Biau, "Analysis of a Random Forests Model" in Journal of Machine Learning Research, 13, 1063-1095, 2012.