Original Research Article

J. Computing & Int. Systems (2020) 99-105

SACRED HEART RESEARCH PUBLICATIONS

Journal of Computing and Intelligent Systems

Journal homepage: www.shcpub.edu.in

ISSN: 2456-9496

A Review on Classification of Machine Learning Algorithms

K. Shanthi [#1], R. Suguna [#2]

Received on 07th APR 2020, Accepted on 08th JUL 2020

Abstract *— - Machine learning comes under artificial intelligence that is it's an application of AI. In the current world a large amount of data available, it is very important to analyze and extract useful information. Machine learning is used to design the algorithm to perform a task without relying on patterns and explicit instruction. In Machine learning, computer programs are built on the sample data based on the mathematical model which is known as training data, the main aim is computer should learn automatically from the data without human support and take decision accordingly. The fields where machine learning algorithm used are computer vision, malware detection, email filtering, bioinformatics, intrusion detection, Information retrieval and so on where it is difficult to develop a conventional algorithm for performing an effective task. The primary objective of this paper is to highlight the merit and demerit of the classification algorithm used in many applications, which helps in deciding which machine learning algorithm can be used to fulfil the requirement of the application.*

Keywords- ***Machine learning, Classification, Decision Trees, SVM, KNN, Naïve Bayes Classification, Linear Regression***

1. INTRODUCTION

Machine Learning is a system in which the algorithm is built to automatically process the data and predict the output accurately using statistical analysis. Machine learning is used in many applications, Classification of texts or documents like Filtering spam message and Speech recognition, Computer vision tasks like image recognition and face detection, Medical diagnosis, search engines, information extraction systems and Web page ranking. Machine Learning algorithm is classified into three

1. Supervised Learning
2. Unsupervised Learning
3. Reinforcement Learning.

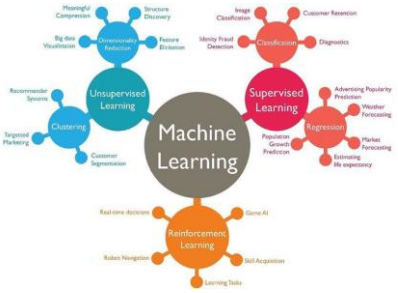

Fig .1 Types of Machine Learning

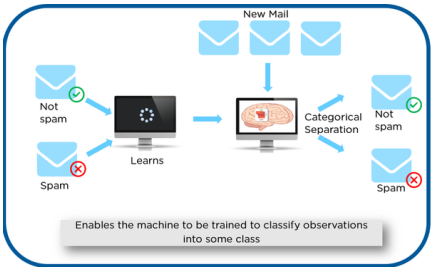

Fig. 2 Training of Machine

1.2 Types of Supervised learning

Classification In Classification, labeled data is used to predict the new data based on the trained data which is in non continuous form (defined label). The graph is non linear. Various classification algorithms of supervised learning are

Corresponding author: E-mail: [1]softshanthi@gmail.com, [2] sugunarajasekar@gmail.com

[1] *Research Scholar, Department of Computer Science, Theivanai Ammal College for Women, Thiruvalluvar University,Villupuram, TamilNadu,605602, India .*

[2]*Assistant Professor, Department of Computer Science, Theivanai Ammal College for Women, Thiruvalluvar University,Villupuram, TamilNadu,605602, India*

1. Logistic Regression/Classification
2. Naive Bayes
3. K-Nearest Neighbors
4. Decision Tree
5. Support Vector Machine

Regression: In Regression labeled data is used to make predictions in a continuous form. The graph is linear. This technique is used in forecasting the weather, time series modeling, process optimization. Various regression algorithms are , ElasticNet Regression, Linear Regression, Lasso Regression, Ridge Regression, Polynomial Regression.

2. LITERATURE SURVEY

Kwang In Kim, in his research paper, used SVM method in text classification. SVM receives gray-scale value as input, which is capable of incorporating the feature extraction scheme within its own architecture. The performance was high when using SVM classifier in text classification. [8]

Liang, J in his research SDTC (Secure Decision tree classification) for cloud-assisted online diagnosis services a new scheme was proposed. This scheme changes the secure outsourced decision tree classification problem to a secure search problem and for protecting the data it uses symmetric key encryption. SDTC scheme secures the user data and it is faster than the linear classification speed. Classification is done in microseconds. [2]

Baygin. M, in his research paper Naïve Bayes approach is used to classify the Turkish document. On the internet, many documents are uploaded for the users they are not categorized according to the user's interest. This classifier is used to classify the document and has achieved a performance rate of 92%.[5]

Zou, X., Hu , in his research paper, the algorithm of the binary classification is optimized and concluded if the value of n is larger in the sigmoid function the iteration will be reduced. In this paper, the prediction is made whether certain cars are accepted by the customers using the binary classification method.[11]

3. LOGISTIC REGRESSION / CLASSIFICATION

Logistic Regression is an extension of the linear model for classification problem used in Machine Learning Algorithm for two-class classification. Logistic Regression is used to predict the value of the dependent variable using a set of independent variables. The output of Linear regression is continuous but in the case of Logistic Regression output is constant that is it can be 0 or 1, Yes or No, instead of giving exact value 0 or 1 its output value lies between 0 and 1.Applications where Logistic Regression is used are Online Fraud Detection, Tumour Malignant, Cancer Detection etc.

3.1 Sigmoid Function: It is also called as Logistic function, takes in a real number and produces 'S' shaped curve which predicts the value as 0 or 1. The value is 1 if the curve moves to positive infinity and its value is 0 if the curve moves to negative infinity. Based on the output of the sigmoid function the value is predicted, if the output is more than 0.5 the value is predicted as 1 or else the value is 0.

The Logistic function is

$$logistic(\eta) = \frac{1}{1 + \exp(-\eta)}$$

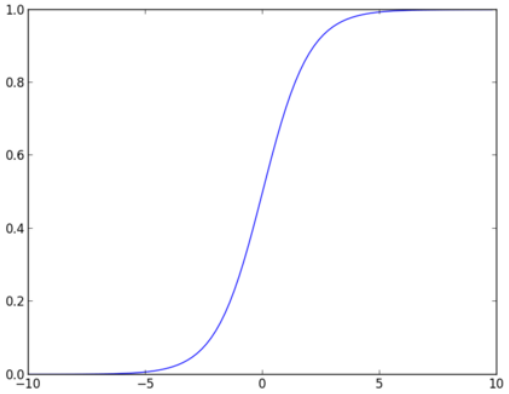

Fig. 3 Logistic Regression

3.2 Types of Logistic Regression

1. Binary - Dependent variable has two possible value either 0 or 1
2. Multinomial – Dependent variable has three or more possible unordered types of values.
3. Ordinal - Dependent variable has three or more possible ordered types values.

4. NAIVE BAYES ALGORITHM

Naive Bayes classifiers are a collection of classification algorithms which shares a common principle which is based on Bayes' theorem. It is a simple, fast, accurate and reliable algorithm, easy to build for very large data sets. It is used in many applications like natural language processing (NLP), Real time prediction, Multi class prediction, Recommendation System etc.

4.1 The Bayes Rule

Bayes theorem works based on conditional probability. The theorem predicts the value, based on prior knowledge of conditions that have might already occurred.

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)}$$

where

P(A) – Probability of occurrence of A.

P(B) - Probability of occurrence of B.

P(A|B) - Probability of A given B.

P(B|A) - Probability of B given B.

4.2 Naïve Bayes Classifier

Naïve Bayes predicts the membership probabilities for each class using the Bayes Theorem. The class which has the highest probability is the most likely class called Maximum A Posteriori (MAP).

The MAP for a hypothesis is:
MAP(A) = max(P(A|B))
MAP(A) = max((P(B|A)*P(A))/P(B))
MAP(A) = max(P(B|A)*P(A))

where P(B) is evidence probability which is used to normalize the result. Presence and absence of features do not influence the features of others, so in Naive Bayes classifier, all the features are considered unrelated to each other. In real dataset calculation of hypothesis is complicated because we have multiple features, to solve this we use feature independence approach.

P(A|Multiple Evidences) = P(B1| A)* P(B2|A) ……*P(Bn|A) * P(A) / P(Multiple Evidences)

Types of Naïve Bayes Classifier

1. Multinomial Naive Bayes – useful in document classification problem.
2. Bernoulli Naive Bayes - similar to the multinomial naive bayes but the predictors are boolean variables.
3. Gaussian Naive Bayes – useful when the predictors take up a continuous value.

5. K-NEAREST NEIGHBOR

K-Nearest Neighbor is a non-parametric algorithm used for regression and classification problem. It stores the training dataset when new dataset appears the data is classified into the category which is much similar to the new data, so it is called as lazy learner algorithm (it doesn't learn from the training dataset). When a new dataset arrives prediction is made by searching the entire training dataset. Choose a value for K, it can be any integer. Distance is calculated between the new data and each row of training dataset using methods like Euclidean, Manhattan, Hamming distance or Minkowski Distance. Sort the data in ascending order and choose the K nearest neighbours. The number of data in each category is counted and the new data is assigned to the category where large number of neighbour is present.It is used in many applications like Text Mining, Medicine, Agriculture and Finance.

Euclidean distance - used to calculate the difference between two points in Euclidean space.

Hamming Distance - used to calculate the distance between binary vectors.

Manhattan Distance - used to calculate the distance between real vectors using the sum of their absolute difference.

Minkowski Distance - Generalization of Euclidean and Manhattan distance.

5.1 Choosing the K value

There is no structured method to find the value for K. If we choose a smaller value for K the local estimate tends to be very poor owing to the data sparseness and the noisy, ambiguous or mislabelled points. If we choose larger value easily makes the estimates over smoothing and the classification performance degrades with the introduction of the outliers from other classes. We can use cross-validation to choose the value of k, validation dataset is formed by selecting a small portion from the training dataset, from this dataset different possible values of k is chosen. Value of k= sqrt(N) where N is the number of samples in the training dataset. In figure 4 we need to classify the new data with black dot into blue or red class. Suppose if we take the value of K=3 the three nearest data point are found. Among the nearest point, the maximum is red so the black dot is assigned to the red class.

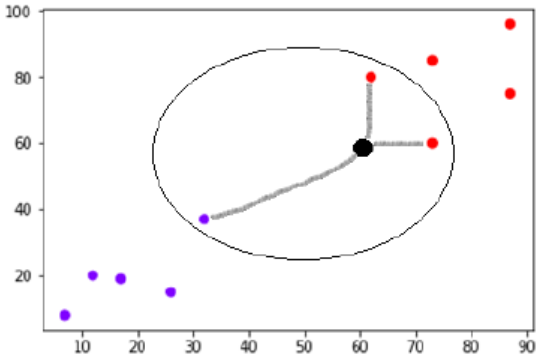

Fig. 4 - K-NEAREST NEIGHBOR

5. DECISION TREE

The Decision tree is a predictive model used in statistics, data mining and machine learning. It is most widely used falls under the category of supervised learning. Decision tree is used in both, classification and regression; it is a non-parametric supervised learning method. The data set is split based on condition. In Decision, tree training model is first created which is used to predict the new value based on simple decision rules. It's built like a flowchart in which the internal node represents the attributes, the branch represents the decision rule and the leaf node represents the outcome. Topmost node is known as the root node.

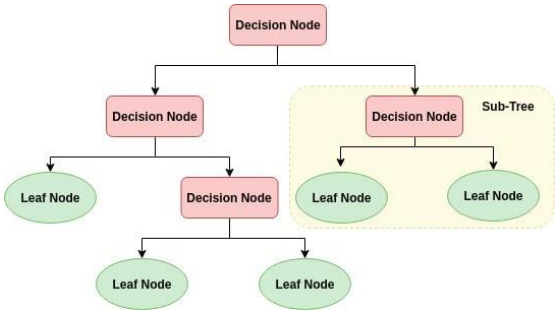

Fig. 5 – Decision Tree

6.1 Working of Decision Tree

In the Decision tree, the outcome variable is continuous as well as categorical. The Decision tree can be built using many algorithms like ID3, C4.5, CART, CHAID, and MARS. The commonly used algorithm is ID3 which uses top-down greedy search approach. In the decision tree, the difficult task is the selection of attributes. Attribute selection can be done using these two methods

1. Information Gain
2. Gini Index

Information Gain: If a node in the decision tree is used to partition the training data into smaller data the value of the entropy changes. The changes in the entropy are the Information Gain. Information gain for the set S is the effective change in the entropy deciding on a particular attribute A.

$$IG(S,A) = H(S) - \sum_{i=0}^{n} P(x) * H(x)$$

Where

IG(S, A) - Information gain by applying feature A.

H(S) – Entropy for the entire set.

H(x) – Entropy after applying feature A.

P(x) - Probability of the event x.

Entropy is the measure of uncertainty of a random variable, the value is 0 when all members belong to the same class and 1 if half to one class half belong to other class.

$$Entropy(S) = \sum_{x \in X} p(x) \log_2 \frac{1}{p(x)}$$

In ID3 algorithm the following steps has to be followed to build the decision tree. The first step is to create root node, then entropy is calculate for the current state H(S) and with respect to the attribute x, H(S,x). The attribute which has the maximum value of IG(S,x) is selected and it is removed from the set of attributes. The process is repeated until the the decision tree has all leaf nodes.

Gini Index is a metric measured to identify randomly selected incorrect element. Attribute with lower Gini value is selected.

Gini = 1- Σ (Pi) 2

6. SUPPORT VECTOR MACHINE

In both classification and Regression we can use Support Vector Machine. In n- dimensional space the training data sets are plotted as a point with the value of each feature being the value of a particular coordinate. In SVM hyperplane is used to differentiate between two classes. There are many possibilities to choose the hyper-plane but we intent is to find the plane with maximum margin i.e. maximum distance between data points of both classes. The application where SVM used are Protein Structure Prediction, Intrusion Detection, Handwriting Recognition,

detecting Steganography in digital images, Breast Cancer Diagnosis.

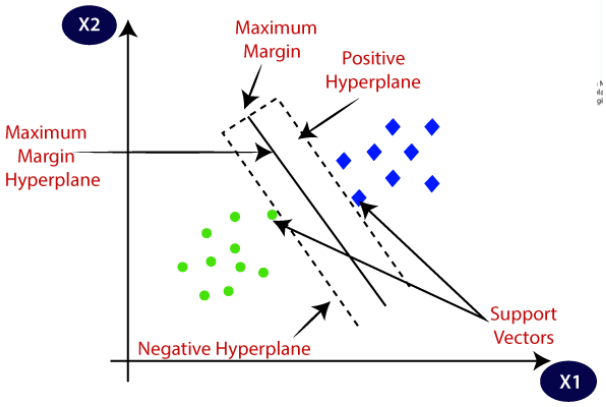

Fig .6 - Hyperplane

SVM is of two types

1. Linear SVM – Single straight line is used to classify the linearly separated dataset.
2. Non- Linear SVM – Single straight line cannot be used to classify the non-linearly separated data.

6.2 Working of SVM

6.2.1 Linear SVM

Consider two data set (green and blue) with two features x1 and x2. Plot the dataset in 2 D plane and multiple lines can be drawn to separate the two datasets. SVM algorithm finds the best hyperplane, it also finds the closest point of the line from both the dataset called a support vector. The distance between the vector and hyperplane is called margin. The hyperplane with maximum margin is called the optimal margin.

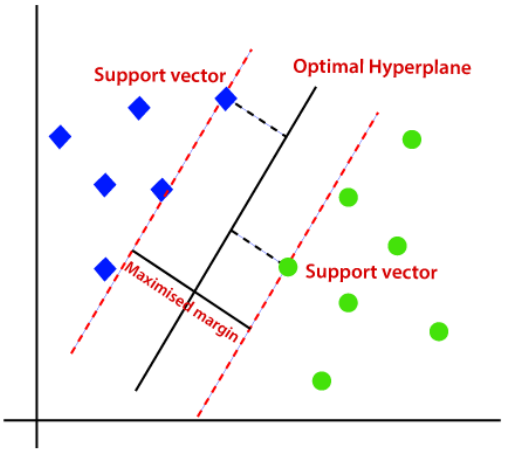

Fig.7 – Linear SVM

6.2.2 Non - Linear SVM

For non-linear dataset we cannot draw a single straight line, to separate the data we add a third dimension Z ($z=x^2+y^2$).

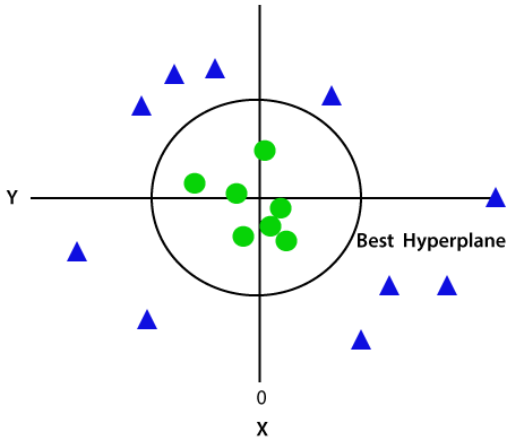

Fig.8 – Non Linear SVM

Kernel Trick: When we are moving from 2 Dimensional spaces to more and more dimensions, computation within the space becomes more expensive. Kernel trick allows us to operate in original space without computing the coordinates of the data in a higher dimensional space.

7. PROS AND CONS OF CLASSIFICATION ALGORITHM

Method	Pros	Cons
LOGISTIC REGRESSION / CLASSIFICATION	• Easy to implement and train the model. • It is more robust. • Using stochastic gradient descent the model can be easily updated with new data.	• It can be used to predict discrete functions only. • We should not use when number of observations are lesser than the number of features. • In the real world it is not possible to assume the linearity between the dependent variable and the independent variables.
NAIVE BAYES ALGORITHM	• Easy, Fast, efficient and highly scalable. • Handles continuous and discrete data. • Small amount of data is enough to train the model.	• It assumes all the attributes are mutually independent but in real life some attributes are independent.

	• Speed is very high when training large amount of data and when processing queries. • Suitable for text classification • Easily handle missing values.	• Zero frequency can occur ie if an attribute occurs in test data set but not in training data set, then will assign 0 prediction which is not possible. • Chance of loss of accuracy
K-NEAREST NEIGHBOR	• Simple and flexible to implement. • It handles multiclass cases. • Does not require training before making a prediction.	• Prediction of K value is difficult. • It needs more storage space. • It does not work efficiently for large dataset. • Computation cost is high.
DECISION TREE	• It can be used in both binary and multiclass problem. • Execution speed is high. • It can be used in both classification and regression. • Simple and easy to understand. • It handles both continuous and categorical variables. • Scaling of data and normalization is not required for the data. • Missing values in the data do not affect the process in the decision tree.	• It is expensive. • Small changes in the data will affect the structure of the decision tree. • More time is needed to train the model. • If there are many class labels, the tree may grow complex and it may lead to over fitting.
SUPPORT VECTOR MACHINE	• It can be used in both binary and multiclass problem. • Execution speed is high. • It can be used in both classification and regression. • Simple and easy to understand. • It handles both continuous and categorical variables. • Scaling of data and normalization is not required for the data. • Missing values in the data do not affect the process in the decision tree	• It is expensive. • Small changes in the data will affect the structure of the decision tree. • More time is needed to train the model. • If there are many class labels, the tree may grow complex and it may lead to over fitting.

8. CONCLUSION

This paper reviews the machine learning algorithm which is most frequently used to solve regression, clustering problems and classification. The advantages and disadvantages of these algorithms, its working procedure and its application have been discussed. Mainly machine learning approaches are interested in considering steadily increasing outputs and accessibility of existing particular challenges on research quality improvement. Machine learning can have applied in systematic reviews on the complex field such as abstract screening process and quality improvements. This paper makes a clear idea which algorithm can be used in which application and none of the algorithm is more influential than the other in different scenarios.

9. REFERENCES:

[1] Kedar Potdar, Rishab Kinnerkar, "A Comparative Study of Machine Algorithms applied to Predictive Breast Cancer Data", International Journal of Science & Research, Vol. 5, Issue 9, pp. 1550-1553, September 2016.

[2] Liang, J., Qin, Z., Xiao, S., Ou, L., & Lin, X. (2019). Efficient and Secure Decision Tree Classification for Cloud-Assisted Online Diagnosis Services. IEEE Transactions on Dependable and Secure Computing, 1–1.

[3] Neocleous, C. & Schizas, C., (2002), Artificial Neural Network Learning: A Comparative Review, LNAI 2308, pp. 300–313, Springer-Verlag Berlin Heidelberg. [87] Nilsson, N.J. (1965).

[4] Zheng, Z. (1998). Constructing conjunctions using systematic search on decision trees. Knowledge Based Systems Journal 10: 421–430.

[5] BAYGIN, M. (2018). Classification of Text Documents based on Naive Bayes using N-Gram Features. 2018 International Conference on Artificial Intelligence and Data Processing (IDAP).

[6] Sonal S. Ambalkar, S. S. Thorat2, "Bone Tumor Detection from MRI Images using Machine Learning: A Review", International Research Journal of Engineering & Technology", Vol. 5, Issue 1, Jan -2018.

[7] Baik, S. Bala, J. (2004), A Decision Tree Algorithm for Distributed Data Mining: Towards Network Intrusion Detection, Lecture Notes in Computer Science, Volume 3046, Pages 206 – 212.

[8] Kwang In Kim, Keechul Jung, Se Hyun Park, & Hang Joon Kim. (2002). Support vector machines for texture classification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(11), 1542–1550.

[9] Bouckaert, R. (2004), Naive Bayes Classifiers That Perform Well with Continuous Variables, Lecture Notes in Computer Science, Volume 3339, Pages 1089 – 1094.

[10] D. Pelleg, A. Moore (2000): "X-means: Extending K-means with Efficient Estimation of the Number of Clusters"; ICML '00 Proceedings of the Seventeenth International Conference on Machine Learning, pp. 727-734.

[11] Zou, X., Hu, Y., Tian, Z., & Shen, K. (2019). Logistic Regression Model Optimization and Case Analysis. 2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT).