To Study Various Machine Learning Technique

Atul Kumar[1], Sunila Godara[2]

Department of Computer Science and Engineering
Guru Jambheshwar University of Science and Technology

Assistant Professor Department of Computer Science and Engineering
Guru Jambheshwar University of Science and Technology

Abstract--We all are acquainted with all the four generation of computer which had before happen. With the change in time and advancement many more equipment and programming changes come. Now in the phase of fifth era need of individuals prompts to the Artificial intelligence. Which is mainly concern with the concept of Machine learning means learning by doing. Machine learning mainly perform the main role in pattern recognition Process. There are many fields in which machine learning process take place which are like credit card fraud detection, search engines, speech recognition, Natural language processing, Number plate recognition etc. Here our aim of this paper is to study about various existing machine learning techniques furthermore about their pros/cons.

Keywords--Decision Tree, K-Nearest Neighbors (KNN), Artificial Neural Network (ANN), Support Vector Machine (SVM), and K-means.

I. INTRODUCTION

Machine Learning is currently the hotspot field of Artificial Intelligence. Machine Learning usually refers to the learning by doing. Machine Learning is the discipline of computer science which deals with the development and Study of various algorithms which can learn by self and predicts results based on the provided tanning set and input data. Means machine learning algorithms are helpful in decision making process because these methods contains the pre stored data and also learn from input data so on bases of having information they support the decision making process[2].

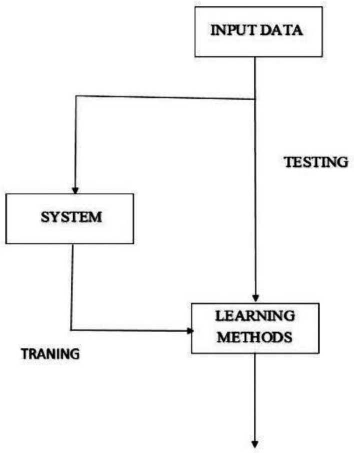

Figure 1: Learning System Model [3].

As we can see the above a simple model of learning system in which input data or Sample data is entered and the data is checked through the System and goes to the learning methods where end results is prepared. Where the result will take place by the reference of trained data already taken by system and new methods will be learned by new input or sampled data [1].

Environment change take place time to time so we need a machine which can learn by itself according to the environment need and redesign problem can be solved. Advancement according to the time will go on to the system and improved knowledge for further use will be available. Through which the system by machine learning process will provide the improved performance [2].

Applications of machine learning process:-

The various applications of machine learning process are:-

1. Page Rank: - When we need an information on internet we go to the search engine and enters the query. The result relevance to our searched query is shown this all takes with the help of machine learning.
2. Automatic translation: - When the need of translation of some text data or words takes place into a language then the system contains some set of predefined document based on those documents and data the new data is being translated.
3. Name Identity Recognition: - When there is some need to searching some data can be like places, titles etc. from document in that case also machine learning also help to solve the problem.
4. Speech Recognition: - In case of speech recognition also system having set of stored training set as the voice of no. of persons then recognition take place based on the training set and learning from the input sample.
5. Recognition of handwriting:-System having the set of data or trained set when query is entered then data related to entered query match result is shown and machine learning take place from the sample data.

Based on the Nature of learning Machine Learning Process can be further classified into Two broad categories:-

1. Supervised Learning.
2. Unsupervised Learning.

These categories of learning take place base on the content given like input data is given or not, Output is define of not, is there any particular method to follow or not or based on learning only by doing at all.

Supervised Learning:-Supervised Learning is the process of learning by mapping the input data to output data. Where the sample input data set and desired goal are defined by the teacher. Main aim is to learn by doing where input and output are given and data process take place on the predefined labelled data set which act as the supervision [1].

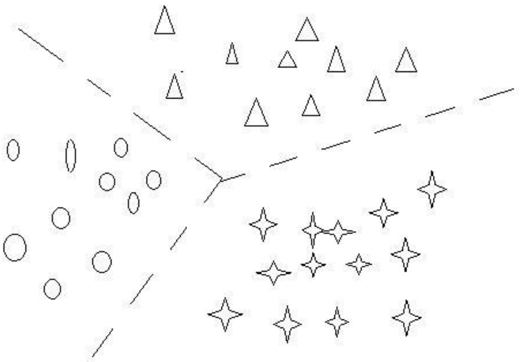

Figure 2: Supervised Learning

There are the various supervised learning algorithm like decision tree, Support Vector Machine (SVM), K – Nearest Neighbor algorithm, Backpropagation Neural Network, Naïve Bayesian Classification etc. which are further discussed.

Unsupervised Learning: -Unsupervised Learning is the process of learning where no labels are present and no targeted output is given. Input data set is given the process is to examine the input data set and grouping the most similar data.Class labels of data sets are unknown [1]. Mainly Clustering is the process of Unsupervised Learning.

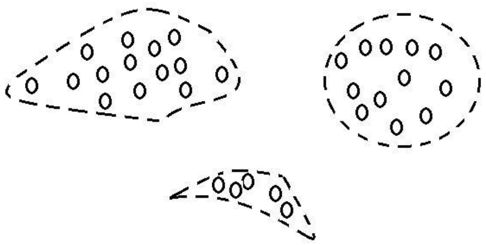

Figure 3: Unsupervised Learning [3].

II. DECISION TREE

A Supervised learning technique which represent the tree like structure which predicts the targeted values or goal by learning simple decision rule. In the Decision tree every branch or segment is known as Nodes. The node at the top level is called Root node and the node at the lowest level or the node having no descendant except root are called leaf node. The node situated between the Root node and leaf node is considered as internal node which represent the test on attribute value. Where the branch comes from the internal node is the result of test. Leaf node here identify the class label and classification rule is associated with each leaf node to root node [2] [5] [9].

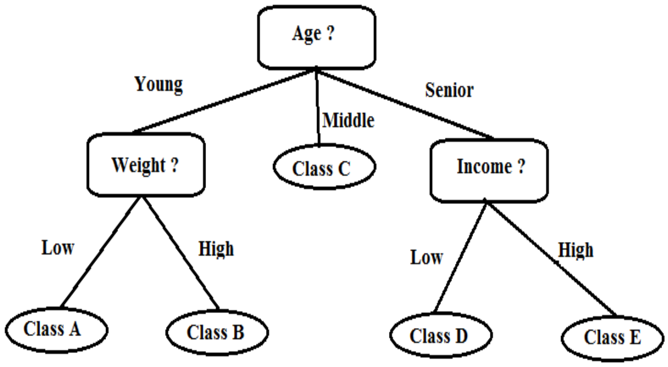

Figure 4: General Decision Tree [9]

Here from the above fig. 4 we can see the overall process of Decision tree. Where Age, Weight and Income are the test on data while outcomes from these test like young, middle, senior, low and high are the branches. Class A, Class B Class C, Class D and Class E are the different class represent the result.

Where tree can be used for the classification and prediction purpose by if then rule we can have like:-

Age(Z,“Youth”) And Weight(Z,“Low”) ---> Class(Z,“A”)

Age(Z,“Middle”) ---> Class(Z,“C”)

Age(Z,“Youth”) And Weight(Z,“Low”) ---> Class(Z,“A”)

Age(Z,“Income”) And Weight(Z,“Low”) ---> Class(Z,“D”) [1].

Attribute Selection Measures in Decision tree for selecting the splitting criteria for data partition are like information gain, gain ratio, and gini index. There No of tree algorithms exists which uses these attribute selection measures for selecting the splitting criteria are:-

ID3(IterativeDichotomiser 3) :- ID3 was developed by Ross Quinlan in 1986.The algorithm used to generate the decision tree which can use be used for further prediction or other purpose.ID3 use the information gain age the attribute selection measure to select the splitting attribute. It does not guarantee that it will provide the optimal solution where the result can reside on local optimum. When the construction of decision tree take place the attribute having largest information gain is selected as splitting attribute[8]. Let see how the information gain is calculated when node N represents the tuple of partitions then the entropy of D is calculated as

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i),$$

[1]

After that we needed to calculate the InfoA (D) is the expected information required to classify a tuple from D.

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j).$$

[1]

Now the information gain can be achived by the above two difference lke

$$Gain(A) = Info(D) - Info_A(D).$$

[1]

When information gain is achived for each attribute comarison is made between the the values of attributes get and the attribute having highest information gain is selected as spiliting attribute.

C4.5(ID3 Sucessor):- One of the most commonly used algorithms in the machine learning C4.5 is the sucessor of ID3 and used to generate the decision tree.C4.5 generates the if than rules from the output of the ID3 alogithm.Where the if than rules are generated which provides the result based on the condition true or false. The Correctness of each rule is examined to determine the order in which they should be applied.Where the part or leaf is need to be discarded called tree pruning wile C4.5 provide Postpurning.By eleminating the rule precondition pruning can be doneif the correctness of rule get improved without it.While C4.5 also use the same concept of spliting attribue. Information gain is measured as it was measured in ID3 concept and the attribute having highest Information gain value which will be selected as the splitting attribute [4].

CART(Classiification & Regression Tree):- Classification and regression Tree is similar to C4.5 but it is diffirent in case of that it supports numerical target variables and it does not compute if else rules sets. Where as the CART Construct the decision tree using the concept of largest information gain [1].

III. K-NEAREST NEIGHBORS(KNN)

K-nearest neighbors is a supervised learning algorithm which contains the available and classify the new sample data based on the similarity measures. K -Nearest Neighbors consists the training data sets and when new case is given as input classification take place on those stored training data sets. Where the output of the process depends on whether KNN used for classification or regression.

KNN in Classification: - When the KNN process take place in the classification the input data is classified based on the majority voting of the neighbours of input data. Where K in KNN process is the positive integer value. When K=1 the input case is assigned directly to the single nearest neighbor. While the varying value of k affect the overall result of the process means when the value of k increase the class to which input case is assigned can change and different result can take place [9].Let see the process through diagram.

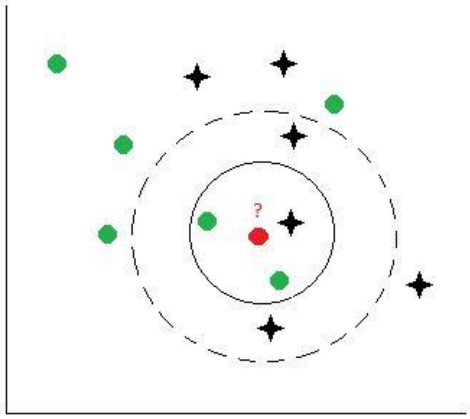

Figure 5:KNN Used for Classification

Here we can understand the process of KNN easily if we see into the above diagram we have red input case and it needs is to classify. Then according to KNN process if K is taken 1 then the data will be assigned to class of star which is nearest. In case the neighbours are increased or value of k is increased to 3 the class of data becomes to green circle. And when it comes to k=5 then class change to star but the process gain the good accuracy when k take place between 3-10. In case of K=1 the process can give bad result of for influence of noise can be noted.

KNN in Regression: - When KNN process in regression takes place then the Output value is the average value of k-nearest neighbours. If the K is taken 1 in regression the outcome contains the value of nearest neighbours. Example if there is an input case(X) to process and its nearest neighbour's value is 2Y. When the k is taken 1 in that case

X = 2Y

When there is another second nearest neighbour with value 2Z and K is taken to in that case

X = (2Y+2Z)/2

Distance Measures many KNN used is Eucliden distance some time city block distance can be taken into account but mostly Eucliden distance is used as distance measures. Eucliden distance D can be calculated like

$$\mathbf{D} = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

[1]

While KNN take place for the predictions then average of the k nearest neighbour is the output. As it compared to regression KNN Approach which use voting of nearest neighbours and KNN classification values 1, 3, 5 are avoided not to ties. Then the KNN predictions approach provide the good result by averaging the value of nearest neighbours.

IV. ARTIFICIAL NEURAL NETWORK(ANN)

Artificial Neural Network is the supervised learning process which is inspired by the biological neuron system. To perform the specific task it contains the highly interconnected processing elements. Large no. of interconnected elements are generally known as Neurons. There are many fields where the working of Artificial Neural Network take place like forecasting the sales control to industrial, risk management ,validation of data etc. but currently are used to simulate complicated relationship [6].A simple model of artificial Neural Network

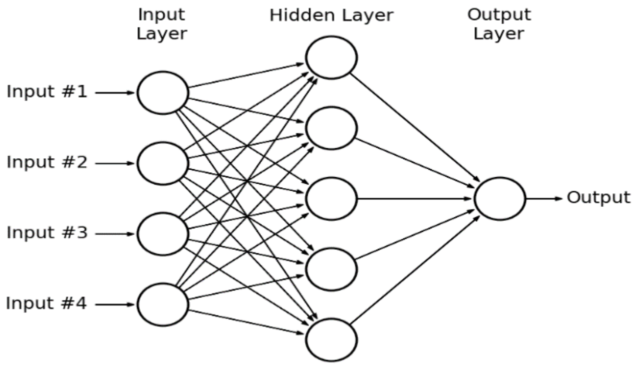

Figure 6: Artificial Neural Network [7]

If we see the simple processing of artificial neural network it consists of three layers which are Input layer, Hidden Layer, Output Layer as we can see in the blow figure. Where Input layer provide the functioning of input the data or values and from where data is feed forwarded to the next layer hidden layer. Hidden layer accept the data from the Input layer along the path from which they are connected. Hidden layer process the data while there no. of hidden layer can take place for solving the particular task. Where data after process is fed forward to the output layer after processing there output is matched to the target. If the match is not take place the weight is adjusted at each processing element [7].

On basis of the architecture of neural networks mainly to type's takes place which are:-

- Feed Forward Network
- Feedback Network

Feed Forward Network: **-** Feed Forward Artificial Neural Network Allow the flow of data only in one direction. There is no way to send the data back. As the data goes for process from input to hidden layer and hidden to output no feedback take place just the input given and processed output is achieved at the end. This type of Artificial Neural Network mainly approach to pattern recognition process.

Feedback Network: **-** Feedback Network allow the network to give the feedback and learn from the experience. Means learn by doing and experience is the main function of Feedback networkif we see the process of feedback Network input goes to input layer then according to weight associated it goes to hidden layer from there after process signal goes to output layer. When the output is matched with targeted output than good if the output does not match then learning take place and weight is adjusted at each processing element. This task improve the performance of existing system which will provide the good decision or result in future. Means overall process is learn by doing and experience. Back propagation Neural Network works on this principle [6] [9].

V. SUPPORT VECTOR MACHINE(SVM)

Support vector machine is the supervised learning method mainly used for regression and classification process. Which contains the trained cases or data where each case belongs to one or more categories. The main task is to classify the new data based on the previous experience and trained data SVM model provides the plane which best separate the given data sets. There are some existing methods which finds the separating hyper plane but not the optimal One. Where SVM provide the Optimal Solution which best separate the given data set. It maximize the margin around the separating hyper plane [1] [8] [10].

In case of liner Support vector machine:-

Where the data sets are linearly separable. Main aim is to maximize the width or the margin which best separate the data set for this purpose the **Margin=2/|w|** is to maximize.

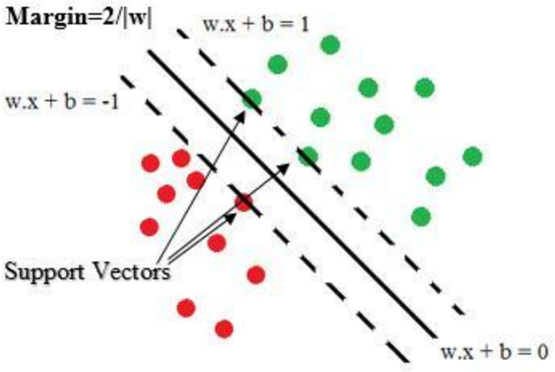

Figure 7: Support vector machine [3]

Here the problem of optimization take place. After solving the optimization problem we have to maximize the function which we have got.

$$f(\mathbf{x}) = \Sigma\alpha_i y_i \mathbf{x_i}^{\mathbf{T}}\mathbf{x} + b$$

In case on Non linear Support Vector Machine :-

Where data is Nonlinearly seprable in that case kernal trick take place which map the original featre space to the higher dimentional space.

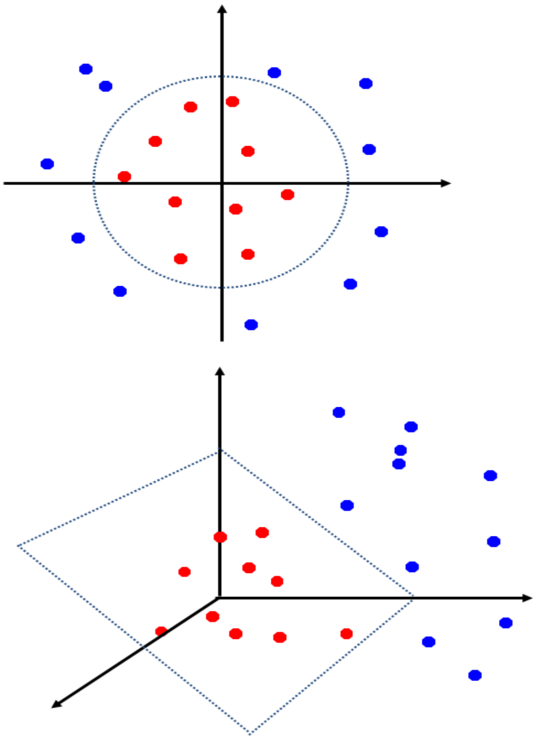

Figure 8: Non Linear SVM Using KernalMethod [3]

As we can see in the above diagram the complete non liner data is mapped to higher dimension using kernel method and easily separated after that process [9].

VI. CLUSTERING

Clustering is the unsupervised learning method in which the object having the similar properties are group together. Clustering is the process of grouping the object of similar property within the group and having different from another group. It's mainly works on the principle of high similarity within the group and dissimilar to the other group. Clustering can also be preferred for outlier detection values which are far away from the actual vales of group. There are various categories in which the clustering methods can be divided which are like partitioning method, Grid Based method, Hierarchical Method, Density Based Method, etc. [8].

- Partitioning Clustering:-First partition are created and then evaluation take place by some criteria.
- Hierarchical Clustering: - Hierarchical decomposing of data set is taken into account using some criteria.
- Density Based Clustering: - Clustering of objects depends upon the density low or high. And also depend on the connectivity function.
- Grid Based Clustering: - All the objects are mapped to the grid like structure and then all the operations are performed on grid and faster computation is achieved.

Here we mainly discuss to clustering techniques k-means and hierarchical clustering.

K-Means: - K-means clustering method comes under the category of Partitioning methods of clustering. In this Clustering technique first the k or we can say no. of cluster are need to be taken manually. Then after that all process take place. Let see the steps of K-Means how does it works [1] [12]:

1) First choose the k points randomly which will act as initial Centroids of the k clusters.
2) Then objects are assigned to the nearest centroid as we have above chosen.
3) Then Computer the Centroid of the each cluster as a new centroid.
4) If the all process not looking sufficient data properly clustered then again go to step 2.

Here we can see the Process of K-Means through simple diagram.

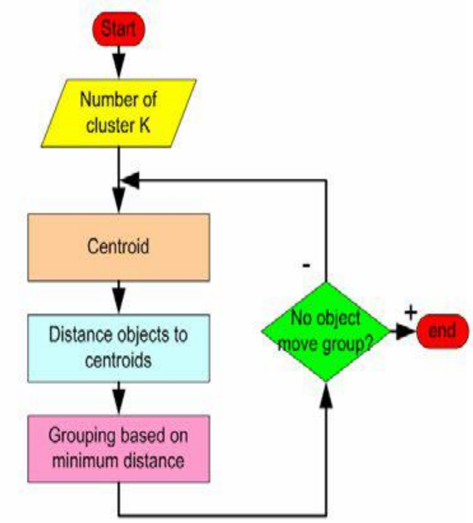
Start
Number of cluster K
Centroid
Distance objects to centroids
Grouping based on minimum distance
-
No object move group?
+
end

Figure 9: K-means flow chart [11]

Here see the process of K-means clustering method in above figure.

VII. CONCLUSIONS

The main objective of our paper is to study about various machine learning techniques. There are the various existing techniques like decision Tree, Neural Network, Support Vector machine (SVM), Clustering etc. These machine learning techniques open the door for researchers mainly in the field of Pattern recognition. Each techniques has own advantage and disadvantage while the technique here studied can be adopted to Number Plate Recognition process. Based on the Image Intensity, Weather Condition in which image captured can adopt different technique for achieving good result in Number Plate Recognition.

REFERENCES

[1] Jiawei Han, Micheline Kamber and Jian Pei Book "Data Mining: Concepts and Techniques," 3rd .edition 2012.

[2] Nils J. Nilsson Book "Introduction to Machine Learning," November 1998.

[3] W. L. Chao, J. J. Ding, "Integrated Machine Learning Algorithms for Human Age Estimation", NTU, 2011.

[4] Solvator Ruggieri, "Efficient C4.5," *IEEE Transactions on Knowledge and Data Engineering*, VOL. 14, No. 2, March/April 2002.

[5] Amany Abdelhalim, Issa Traore, "A New Method for Learning Decision Trees from Rules*," IEEE International Conference on Machine Learning and Applications*, 2009.

[6] Z. Huang, Q.B. Li, H. Yuan, "Forecasting of ionosphere vertical TEC 1-h ahead using a genetic algorithm and neural network," SCIENCE DIRECT *Advances in Space Research*, PP. 1775–1783, 2015.

[7] W. Huang, H. Hong, G. Song, K. Xie, "Deep Process Neural Network for Temporal Deep Learning*," International Joint Conference on Neural Networks (IJCNN),* July 2014.

[8] A. Chaudhary, S. Kolhe, Rajkamal, "Machine Learning Techniques for Mobile Devices: A Review," *Int. Journal of Engineering Research and Applications*, Vol. 3, Issue 6, November/December 2013.

[9] M. Kumari, S. Godara, "Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction," *International Journal of Computer Science and Technology*, Vol. 2, Issue 2, June 2011.

[10] Chih-Wei Hsu and Chih-Jen Lin, "A Comparison of Methods for Multiclass Support Vector Machines," *IEEE Transactions on Neural Networks*, VOL. 13, NO. 2, March 2002.

[11] Pallavi , Sunila Godara, "A Comparative Performance Analysis of Clustering Algorithms," *International Journal of Engineering Research and Applications (IJERA)*, Vol. 1, Issue 3, pp.441-445.

[12] Shraddha Shukla and Naganna S., "A Review On K-means Data Clustering Approach*," International Journal of Information & Computation Technology*, ISSN 0974-2239 Volume 4, pp. 1847-1860, Number 2014.

[13] Sunila Godara, Vanita Rawal, Megha Ranolia, "Analysis on Feature Extraction of Periocular Region (Soft biometrics) using LBP, PCA, ICA & Gabour filters," *International Journal of Computer Trends and Technology (IJCTT)*, VOL. 4, Issue 6, June 2013.

Atul Kumaris pursuing his M.Tech degree in Computer Science & Engineering from Guru Jambheshwar University of Science & Technology, HISAR.He received his B.Tech degree from Maharshi Dayanand University, ROHTAK.

Ms Sunila Godara received MSc and MTech degree in Computer Science & Engg from Guru Jambheshwar University of Science & Technology, HISAR. She is working as Assistant Professor in Deptt of Computer Sc. & Engg, Guru Jambheshwar University of Science & Technology, HISAR. She has published more than 19 papers in national and international journals and conferences. Her research areas are Data Mining and Database Management System.