Machine Learning's Various Algorithms and Choice of Suitable Algorithm for Suitable: A Review

[1]Rakesh Kumar Mittal, [2]Shalini Bhadola, [3]Kirti Bhatia
[1]M.Tech Student, [2]Assistant Professor, [3]Assistant Professor
Computer Science and Information Technology, Sat Kabir Institute of Technology & Management, Bahadurgarh, Haryana, India

Abstract — Artificial Intelligence has the power to change the working environment of the world. It is an important technique of business, science & research, education, engineering, IT, and human development in these days. The sub-set of artificial intelligence is Machine learning. Machine learning is the main focused area of artificial intelligence. It is a technique that makes computer intelligent. It's a learning technique for the machine. The machine learns by its experience. The most important ingredients of experience are data set. Experience enriches the data set. In practice, the data set is the basic building block of the machine learning and experience add on new data to its previous data set. Dataset helps in finding the pattern and it leads to prediction. The algorithm plays the most important role in learning and it is an integral part of it. The algorithm is made in a way that it becomes intelligent by the experience so it does not require any explicit program.

Machine learning has two types, supervised learning and unsupervised learning. In supervised learning, a label is given to data. On the bases of that label, it predicts the result. Supervised learning uses the classification algorithm to grouping the data. Classification algorithm, for example, Naïve Bias, Decision Tree, etc.. In unsupervised learning no label is provided to dataset. Clusters are created for the data set on the basis of features of the dataset. An important algorithm for unsupervised learning is k-means and Principal Component Analysis. All the algorithms are good but an important point is the selection of a correct algorithm for the correct situation. If the choice of algorithm is good then it will work fine and give the accurate result. The choice may be done after the study of the past performance of the algorithm.

Keywords: Machine-learning, Supervised-learning, Unsupervised-learning, K-means, Naïve Bayes, Decision-tree, Classification, Clustering

I. Introduction

Machine learning is a process for creating a model which can solve certain problem automatically without the need of specific programming. Machine becomes intelligent day by day through the data it receives as input. The basis of machine learning is its algorithm which makes machine intelligent. It is a methodology of detecting patterns from data set. These patterns are used to predict the future possible data. Application of this to perform decision making in uncertainty, for example business decision making on the basis of previous data. [1]

It is different from the traditional computing in the sense that a normal computing just store data in the database and retrieve them from database when user demands for it [2].

Machine learning is the field of artificial intelligence in which machine or an automated system learns by its experience. Machine learns from data. Meaning of learning is, a computer program or algorithm learn from its experience. Experience is taken for some particular task. By increasing experience performance should be increase for that particular task [3].

Training and testing is the important part of machine learning. In machine learning first we train the system and develop a model, this is called training. After training, model is tested with some sample data or test data and compares the actual result with the desired result. If the results are same then it is called ideal model.

II. Types of Machine Learning

Machine Learning is a very vast field. As per M. Welling and M. Bowels, we can categories it in various ways as depicted in the picture below:

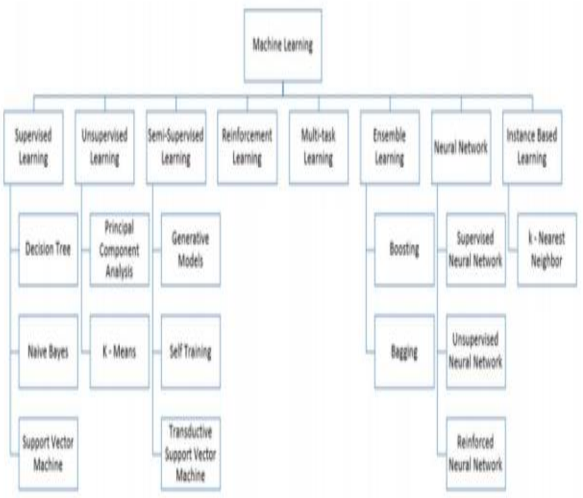

Figure 1: Various Categorization of Machine Learning [4][5]

In this paper we are going to study the following two branches of machine learning: i) Supervised Learning, ii) Unsupervised Learning

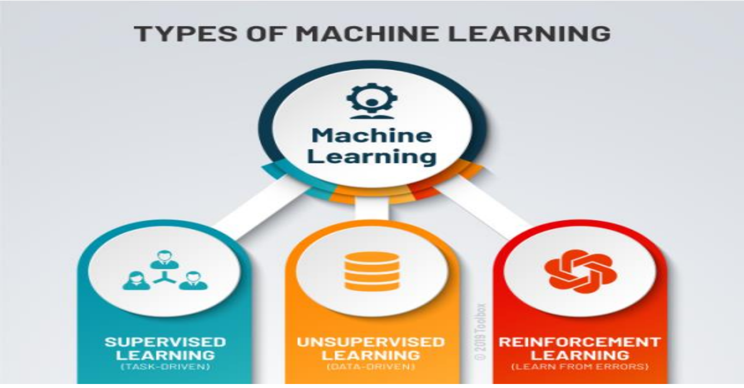

Figure 2: Types of Machine Learning [6] Supervised Learning

Supervised Learning is a type of machine learning which have labeled training data which consisting of training examples. Supervised learning consists of two things, first is input object and second is output value.[7] For example a picture of apple is labeled as "Apple" and a picture of knife is labeled as "Knife" as a training example in a machine. On the basis of this training example when someone input the similar object e.g. apple or knife, machine give the output as the label value in training example.

Algorithm gives the label to the unseen object, to do so algorithm generalizes the training data to unseen object [8].

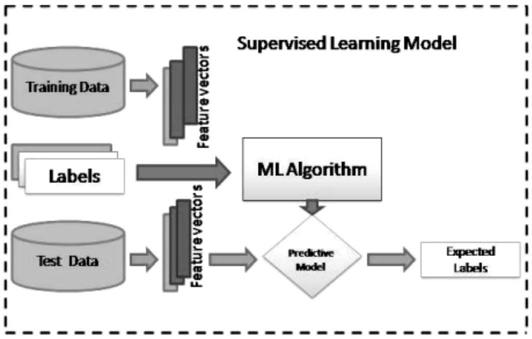

Figure 3: Supervised learning model [2]

III. Classification

Classification process performs on a data set which contains two objects, attributes and classes. Classification function relates the value of attribute and its class [2].

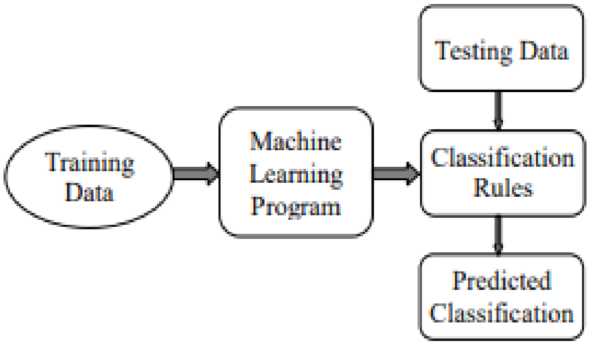

Figure 4: Classification Architecture [2]

A. Steps to develop the Supervise Machine learning [9][10]

1. Select the Data set for training. For example if you want to make a machine for handwriting detection it can be letters, words or whole sentence.

2. Select the training set Gather a training set. The training set must be the representative of the real-world so in supervised learning set of input object and set of and corresponding outputs object are gathered. It can be gathered from human experts.
3. Determine the feature of input function. Input object should be strongly represented. Feature vector are created. Feature vector is a collection of number of features. Number of feature should contain enough information to accurately predict the output.
4. Select the corresponding algorithm.
5. Now run the selected learning algorithm on the gathered training set.
6. Evaluate and test the learned function.

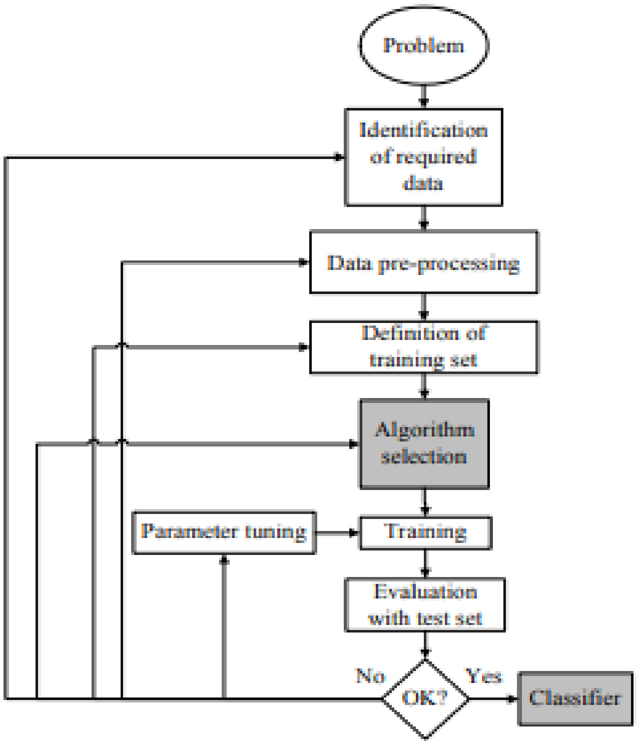

Figure 5: The Process of Supervised Learning [11]

B. Selection of Algorithm for Supervised Learning

In selection of algorithm main issue is accuracy. Base of evaluation is accuracy and accuracy can be determined by: **Accuracy = Number of correct classifications / Total number of test cases**

We can also perform statistical comparison. To perform statistical comparison, we can run two different algorithms on the training set to estimate the difference in accuracy for each case of classifier [9].

IV. Important algorithm for the Machine Learning

A. Decision Tree

Decision Tree algorithm is mainly used for classification purpose. This algorithm groups the attributes on the basis of their values. When we use the word tree, it means it is uses nodes and branches. Nodes represent the attributes and branches represent value, which nodes require [10]. Example of decision tree is given in the figure below:

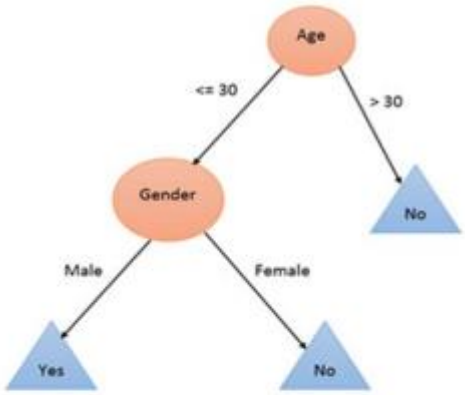

Figure 6 : Decision Tree[11]

B. Naive Bayes Algorithm:

Naïve Bayes algorithm assumes that features are statistically independent. Every value of particular features is independent to value of other feature. It's treated as a good classifier algorithm. Naïve Bayes theorem is based on Bayes theorem [12][13]. It can be express in the probability i.e.:

$P(A|B) = P(B|A)P(A) / P(B)$

Here:P is Probability

P(A|B) : Probability of event B occurring given that A has already occurred

P(A) : Probability of event B occurring

P(B) : Probability of event A occurring

$P(y|X) = P(X|y) P(y) /P(X)$

Where X is weather condition = (High, Low, Slow)
y is the class label Fit or Unfit as per the data set e.g. = y=(Fit)

C. Support Vector Machine

Support Vector Machine (SVM) is machine learning tool that gives solution for classification as well as regression. It is developed by Vapnik at AT & T Bell Laboratory. It gives prediction method based on statistical learning framework. Statistical Learning Framework is also called VC theory developed by Chervonekis and Vapnik. Suport Vector Machine developed hyperplanes to separate the two or more classes. There may be many possible hyperplane but the objective of SVM to find the best hyperplane. Best hyperplane, which have maximum margin, means maximum distance between two or more classes. Maximum distance develop more confidence in developing future data points.[14]

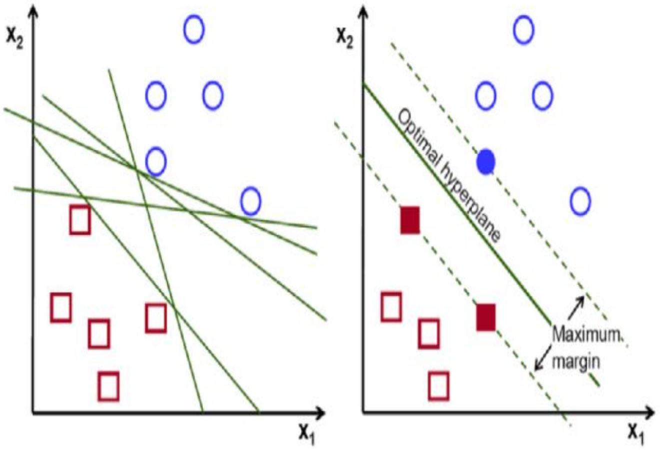

Figure 7: A. Maximum Possible Hyperplane B. Optimal Hyperplane [15][16]

D. Unsupervised Learning

Unsupervised learning is learning algorithm which does not have pre-defined label. It does not have label so it does not require human expert for labeling. We remain unsure about the output. Algorithms of unsupervised learning finds the patterns from the data and give output as required answer. We do not interfere when algorithm is learning. It is unsupervised because no human supervision is required. [17]

E. Types of Unsupervised Learning : There are two types of unsupervised learning i) Principal Component Analysis and ii) Cluster Analysis.

F. *Principal Component Analysis*

Principal component analysis (PCA) is the solution of problem arises with large data set which is difficult to interpret. Principal component analysis reduces the dimensionality of the large database so that interpretability increases. By increasing the interpretability it also minimize the information loss in database. Its objective is to reduce the dimensionality and preserving the maximum variability in the data set. [18]

G. Cluster Analysis

In clustering we group similar items together. Using a clustering algorithm we can group the data points in a specific group. Data points which are member of similar group should have similar features and property. And data points in dissimilar group should have highly dissimilar property and features. Clustering has no criteria of grouping, only user decides the criteria. Criteria are decided as per the user requirement.

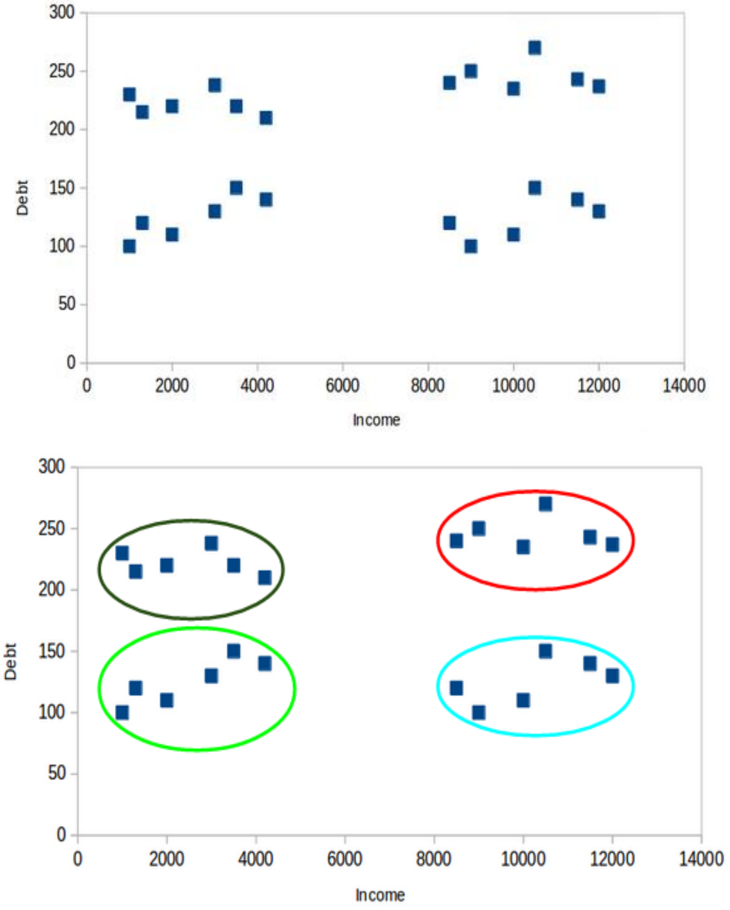

Figure 9: Un-clustered Data and Clustered Data

H. Clustering Algorithm K-means

It is a clustering types of algorithm based on centroid. Purpose of this algorithm is to minimize the distance of data point from the centroid. This algorithm minimizes the distances in between the data point to its centroid. Every cluster has a centroid. Centroid is like prototype of the cluster which is nearest mean of cluster. Centroid can also be called cluster center.

Objective is to convert the number of observation into number of partition, of data set, called cluster. Here number of observation is denoted by n and number of cluster is denoted by K. That's why its name is K-means. [20]

I. How K-means works

Start

Step 1: Decide the number of cluster, k

Step 2: Select the k random points as centroid from the data set. Values other than centroid, is called observed value.

Step3: Now decide the cluster for all of the observed value. To do this uses the Euclidian Distance(ED) between the observed value and distance value.

For example: there is two attribute in data set: Age and Income Two clusters are required

ED of k1 =	$\sqrt{(Xo\text{-}Xc)^2 + (Yo\text{-}Yc)^2}$	Resultant value
ED of k2 =	$\sqrt{(Xo\text{-}Xc)^2 + (Yo\text{-}Yc)^2}$	Resultant value

Here Xo is the observed value of Age, Xc is the centroid value of Age, Yo is the observed value of Income, Yc is the centroid value of Income. K1 is the cluster 1 and K2 is cluster 2(Suppose there are 2 clusters).

Step4: Compare the resultant value, k1 and k2. Which cluster has less resultant value; put the observed value i.e. Xo and Yo in that cluster.

Step5: Find the new centroid for the cluster in which observed value is inserted by :(Old Centroid value + New Centroid value)/2

Keep the centroid value for the clusters same as previous value of centroid which have no change in the cluster data.

Step6: Repeat the Step 3 to 5.

Stop

V. Conclusion

In this paper, we have studied the important algorithm of supervised learning and unsupervised learning. Depending upon the nature of work & condition, we decide the particular algorithm.

For example, if we need to made character recognition application in this case we require supervised machine learning algorithm. In the same way, in case of speech recognition and spam detection, also require supervised learning. If we talk about the particular algorithm of supervised learning, the Naïve Bayes Algorithm of supervised learning is most suitable for Spam filtration because in the Naïve Bayes algorithm we consider every feature of model-independent to each other. If we compare this consideration with spam filtering, each word in a text is independent from other words. So that here Naïve Bayes Algorithm is most suitable. With the same consideration, this algorithm can also be applied to the Text Classification, Sentiment Analysis, and Recommendation System, etc.

Support Vector Machine (SVM) is most suitable for Text categorization [22]. Support Vector Machine doesn't require feature selection. It has the capability to generalize the dataset in high dimensional feature space. This makes the task of text categorization very easy. It is a robust technique as compared to other techniques. It has the ability to set the parameter automatically. These features make the SVM suitable for the text categorization. Other than text categorization SVM is also a good technique for Hypertext categorization, Satellite Data Classification, Reorganization of Hand Written Data, Biological Sciences to classify proteins, Classification of Images, etc.

Unsupervised learning uses clustering for prediction. It is the most important technique of unlabeled learning. Clustering finds its application in the fields of Customer Segmentation, Image Segmentation, Recommendation, and Documents Clustering. Customer Segmentation is used in marketing. The recommendation of a particular product to particular customer segments is the bases of digital marketing. Customers can be divided as per there taste, choice, like/unlike, monthly/annual spending and this can be done by K-means algorithm of unsupervised learning. K-means algorithm makes the predefined number of group/cluster (k) as per the similarity of behavior or features.

Principal Component Analysis (PCA) is another technique for unsupervised learning. PCA is used to reduce the dimensionality of the dataset without losing the information. PCA can avoid the problem of overfitting. For example, to predict the Gross Domestic Production (GDP) of any country for a particular year requires lots of information and databases from the last 10 years. Like the GDP of the last 10 years, inflation rate, employability rate, stock prices, IPO occurred, etc. This required lots of variables, have to understand the relationship between each variable. This can lead the overfitting the model. There is a need to consider every variable but focus on a few of them. To do this we have to reduce the dimensionality of the data set. [23]. Reducing the dimensions from the data set is called dimensionality reduction.

When we work on machine learning no algorithm is good or bad. Every algorithm works fine when we choose the correct one in suitable situations. Before the selection of a particular algorithm, we have to study the experience of the researcher and accuracy of results of the real-time use. Past performance always remains the important ingredient of choice of algorithm.

References:

[1] Kevin P. Murphy, 2012, The MIT Press Cambridge, Massachusetts London, England, Machine Learning A Probabilistic Perspective

[2] Iqbal Muhammad and Zhu Yan2, 2015, School of Information Sciences and Technology, Southwest Jiaotong University, China, SUPERVISED MACHINE LEARNING APPROACHES: A SURVEY

[3] Tom Mitchell, McGraw Hill, 1997, Machine Learning.

[4] M. Welling, "A First Encounter with Machine Learning"

[5] M. Bowles, "Machine Learning in Python: Essential Techniques for Predictive Analytics", John Wiley & Sons Inc., ISBN: 978-1-118- 96174-2

[6] https://www.potentiaco.com/what-is-machine-learning-definition-types-applications-and-Examples/

[7] Stuart J. Russell, Peter Norvig (2010) Artificial Intelligence: A Modern Approach, Third Edition, Prentice Hall ISBN 9780136042594.

[8] Mehryar Mohri, Afshin Rostamizadeh, Ameet Talwalkar (2012) Foundations of Machine Learning, The MIT Press ISBN 9780262018258.

[9] Mehryar Mohri, Afshin Rostamizadeh, Ameet Talwalkar (2012) Foundations of Machine Learning, The MIT Press ISBN 9780262018258.

[10] James Cussens, "Machine Learning", IEEE Journal of Computing and Control, Vol. 7, No. 4, pp 164-168, 1996.

[11] S.B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques", Informatica 31 (2007) 249-268.

[12] L. Rokach, O. Maimon, "Top – Down Induction of Decision Trees Classifiers – A Survey", IEEE Transactions on Systems.

[13] Pouria Kaviani, Mrs. Sunita Dhotre, "Short Survey on Naive Bayes Algorithm", International Journal of Advance Engineering and Research Development Volume 4, Issue 11, November -2017

[14] Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2008). The Elements of Statistical Learning : Data Mining, Inference, and Prediction (PDF) (Second ed.). New York: Springer. p. 134.

[15] https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

[16] Pradhan, Sameer S., et al. "Shallow semantic parsing using support vector machines." Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004. 2004.

[17] Hinton, Geoffrey; Sejnowski, Terrence (1999). Unsupervised Learning: Foundations of Neural Computation. MIT Press. ISBN 978-0262581684.

[18] Ian T. Joliffe, Jorge Cadima, 13 April 2016, Principal component analysis: a review and recent developments, https://doi.org/10.1098/rsta.2015.0202.

[19] Dhara Patel, Ruchi Modi , Ketan Sarvakar, Volume III, Issue IX, September 2014, IJLTEMAS ISSN 2278 – 2540, A Comparative Study of Clustering Data Mining: Techniques and Research Challenges

[20] Youguo Li, Haiyan Wu, 2012 International Conference on Solid State Devices and Materials Science, A Clustering Method Based on K-Means Algorithm

[21] Alexandros Kalousis1, Jo˜ao Gama, and Melanie Hilario,

[22] Thorsten Joachims, Text Categorization with Support Vector Machines: Learning with Many Relevant Features

[23] Matt Brems, April 18, 2017, A One-Stop Shop for Principal Component Analysis, https://towardsdatascience.com/