International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES)
Impact Factor: 5.22 (SJIF-2017), e-ISSN: 2455-2585
Volume 5, Issue 06, June-2019

An Inclusive Study of Machine Learning Classification Algorithms

Rachana Patel

[1]*Department of computer Science & Applications, Charotar University of Science & Technology,*

Abstract— Using machine learning we can develop applications using algorithms which makes computer to learn by own without being explicitly programmed. Supervised learning emphasis on labelled data to train the computer to obtain the desire output. This paper mainly presents the various supervised machine learning classification algorithms like decision tree, Support Vector Machine, Naive Bayes, Neural Network study in depth. Moreover, this paper presents the literature review of application domain like medical, education, agriculture and many more where these classification algorithms are already applied.

Keywords— Decision tree; K-Nearest Neighbour; Neural Network; Naive Bayes; Support vector machine

I. INTRODUCTION

Machine learning makes computers to learn naturally form experience like humans and animals. Algorithms of machine learning follow computational theory which learn information from data given to it without relying on predetermined computation model. Even though, it is the field of computer science, it completely different compare to traditional computational methods in which algorithms are explicitly programmed to solve particular problem. Machine learning algorithms try to learn the natural inner patterns from the given training data which makes better insight and help humans to take accurate decisions and better predictions. The performance of the algorithm gradually improves as the number of training samples increase. These algorithms widely used nowadays in medical field to diagnosis diseases more precisely, to accurate stock market prediction and weather forecasting, media sites to provide the options of songs and movies of your choice as recommendation out of millions of songs, super markets and malls owners to understand the purchasing behavior of their customers. Figure 1 represents the various categories of the machine learning algorithms. However, this paper mainly focuses on study of various classification algorithms of supervised learning and its applications.

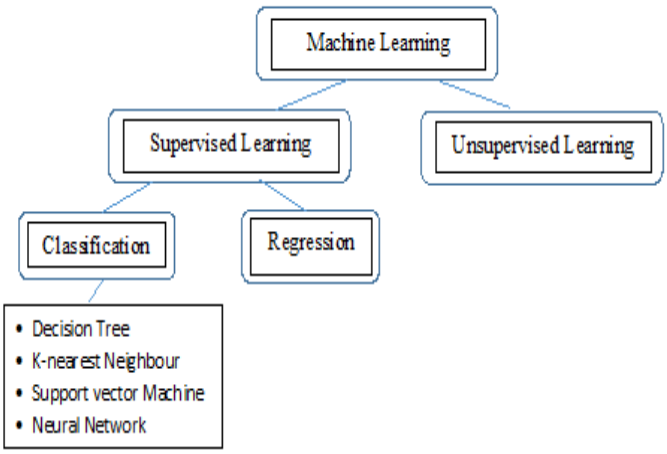

Figure 1 Machine Learning Algorithm Categories

II. TYPES OF MACHINE LEARNING ALGORITHMS

There are mainly the categories of machine learning algorithms 1) supervised learning 2) unsupervised learning and 3) reinforcement learning.

1) Supervised Learning:

Supervised learning algorithms require external assistance. The training data set provided to such algorithms have output variable which needs to be accurately predicated or classified according to the given problem. It is like a teacher supervising learning process, for the given data training set, as he or she knows the correct answer and gradually improves the predication and classification by change the parameter values for the algorithms. In such types learning, input variable(x) is mapped to output variable(y) an algorithm.

$$Y=f(X)$$

The major goal is to make the most efficient mapping function which accurately predict output (Y) variable for the unseen input data (X).

2) Unsupervised Learning:
Unlike supervises learning unsupervised learning have only input variable(X) and corresponding output variable(Y) is not available. It is mainly used to expose the underlying structure of the data or distribution of data which provides better understanding about the given data. It is called unsupervised learning because there is external assistance or correct output is available, algorithms themselves determine and make interesting structure for given data presentable [1].
3) Reinforcement Learning:
It is type of dynamic programming which trains algorithm my means of punishment or reward. The algorithm or agent used in reinforcement learning learns by interacting with its environment. The agent will receive reward or penalties based on their action. Unlike other machine learning algorithm, reinforcement learning algorithm work on its own without being explicitly specified how to do it [2]. An agent could be a self-driving car or a computer program which playing a chess, which will be rewarded according to its performance. Agent gradually increase the its rewards and decreases its penalty by dynamic programming.

III. LITERATURE REVIEW

Prof. Kanak Saxena presented an Efficient Heart Disease Prediction System using Decision Tree [3]. They have designed the model which take patients' health record and based on that rules can be discovered which predict the risk level of heart disease. The rules can be customized or organized based on the requirement. The performance of the classification system is high and predicts the risk level of heart disease precisely.

Yevhenii Udovychenko proposed the Heart Disease Recognition by k-NN Classification of Current Density Distribution Maps [4]. They have used k-Nearest Neighbour classification algorithm. The algorithm is designed for binary classification and classification task can be optimized by selecting number of neighbours for the classifier.

Dana AL-Dlaeen[5] presented work on Decision Tree Classification to Assist in the Prediction of Alzheimer's Disease. Model presented in paper helps the doctors to predict the status of disease from available patients' data. The decision tree classifier used in the paper and based on entropy or information gain it each branch level decision has been made.

Mohammed Aashkaar [6] in their work presented Performance Analysis using J48 Decision Tree for Indian Corporate world. Authors used the data companies' data available in the government portal and used Weka tool for the analysis purpose. They presented confusion matrix and results based on experimentation they carried out.

B. Giraldo [7] presented Support Vector Machine Classification Applied on Weaning Trials Patients. They highlighted method which computes variances in respiratory pattern variability of patients on weaning trials. They have used 35 features obtained from the respiratory flow signal and used Support Vector Machine for classification.

Argyro Kampouraki [8] comes up with Robustness of Support Vector Machine-based Classification of Heart Rate Signals. To classify the heart rate signal from the database which consist twenty records of the youngsters age ranges from 21 to 34 and twenty elderly age ranges from 68 to 85, support vector machine is used.

U Ravi Babu [9] in their work presented Handwritten Digit Recognition Using K-Nearest Neighbour Classifier. They came up with new approach i.e. off-line handwritten digit recognition using structural features. Author has used the MNIST images and used k-nearest classifier for classification. The object is classified which has majority of vote.

Anand Shanker Tewari [10] presented an Opinion Based Book Recommendation Using Naive Bayes Classifier. They have collected and categorize reviews given on the book. In their work they proposed book recommendation system using opinion mining and Naïve Bayes classifier to highlight top ranking books. The probabilistic Naive Bayesian classifier performs extremely well while working with textual data.

IV. CLASSIFICATION ALGORITHMS

Classification is considered as supervised learning approach in machine learning in which learning by computer program is done from the data input provided to it. Based on this learning it classify the new unseen observations. The input data set may be bi-class or multi-class. In this section I am discussing various classification algorithms along with its merits and demerits.
1) Support Vector Machine
Support Vector Machines follows the idea of decision planes which defines clear decision boundaries. A decision plan distinguishes between a collection of objects which belongs to different class memberships. The classification task which demands clear identification of boundaries between different class to which object belongs in known as hyperplane classifier and support vector machine is best suited for such kind of tasks [11]. As shown in the figure 2 the object falls into either class GREEN or RED. The boundary between GREEN and RED is specified using separating line, which indicates objects left to line are RED and right to line are GREEN. Any new object falling right to the line is classified as GREEN and left to line as RED.

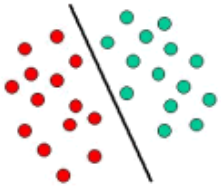

Figure 2 SVM hyperplane which separates two classes

However, all classification problems are not as simple as linear classifier, some are complex demands more complex structure to make meaningful separation. So, for such kind of tasks support vector machines are considered as the handy solution. As shown in figure it rearranges original objects using mathematical function i.e. kernels. The rearrangement of object is known as transformation or mapping [18]. By, rearranging the objects we eliminate the need of drawing complex curve to separate the classes and all we need to identify the optimal line which separates classes.

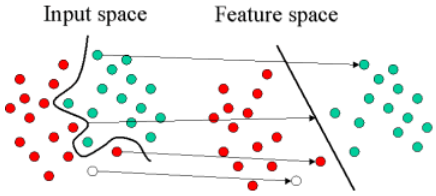

Figure 3 Rearrangement of object to identify optimal line

Classification SVM Type-1
Training in this type involves the minimization of the error function

$$\frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi_i \quad (1)$$

Constraints:

$$y_i \left(w^T \phi(x_i) + b \right) \geq 1 - \xi_i \text{ and } \xi_i \geq 0, i = 1, \ldots, N$$

Classification SVM Type-2
Unlike Classification Type-1, it minimizes the error function

$$\frac{1}{2} w^T w - v\rho + \frac{1}{N} \sum_{i=1}^{N} \xi_i \quad (2)$$

Constraints:

$$y_i \left(w^T \phi(x_i) + b \right) \geq \rho - \xi_i, \xi_i \geq 0, i = 1, \ldots, N \text{ and } \rho \geq 0$$

2) Artificial Neural Network
Artificial Neural Network concept is taken from the human brain neuron system which plays a vital role in human body. It is interconnection of huge number of neurons, which are capable of performing parallel processing for any decision in our body and represents best parallel processing practice [12]. Similar to the human brain ANN computing system comprises large number of interconnected units which allows communication between them. These units are often referred as artificial neurons which can operate in parallel. There are broadly two ANN categories: Feed forward network and Feedback network [13].
a) Feed forward network:
This type of network contains nodes in the layer which non-recurrently connected with the nodes of the previous layers. It follows one directional signal flow from input to output. There is single layer feed forward network in which only single weighted output layer is connected fully with the input layer. However, in multilayer feed forward network, there is existence of multiple weighted layers referred as hidden layer between input and output layer.

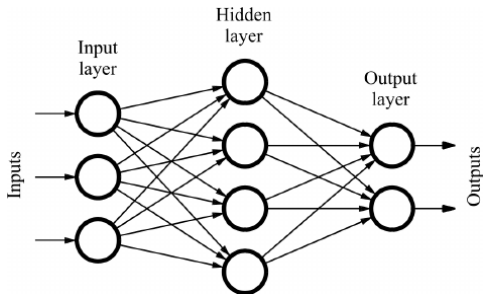

Figure 4 Feed Forward Neural Network

b) Feedback network:
This is referred as recurrent neural network in which the signal flow in both the directions with appropriate feedback path using loops. It is considered as the dynamic network as state in this network keep changing until it achieves equilibrium point.

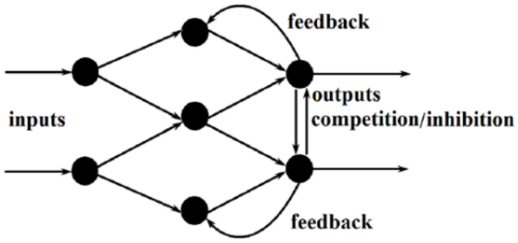

Figure 5 Feedback Neural Network

3) Decision Tree Classification
This classifier is used to build a model in the form of tree structure. The dataset is divided into number of small subsets and at the same time decision tree is incrementally developed. It is like flowchart like structure where test on particular attribute is represented by inside node and each branch represents test outcome and each leaf node represents a class label. For any given test tuple x, attribute values are tested for given decision tree. From root to leaf a path is traced which to predict the class of the tuple. Decision tree can be easily converted to classification rules [14]. This predicative modeling concept is used in Machine Learning, Data Mining and Statistical Analysis. It uses entropy and decision gain to construct a decision tree.

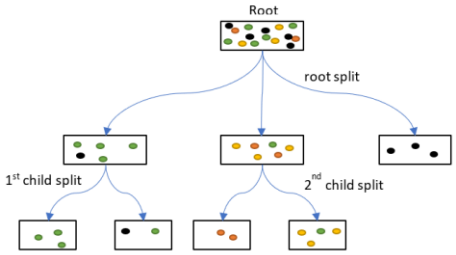

Figure 5 Decision Tree Splitting

Entropy:
Entropy is the degree or amount of uncertainty in the randomness of elements

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i \tag{3}$$

Information Gain:
With respect to independent variable it measures the relative deviation in the entropy. To construct a decision tree it is required to find the attribute the returns the highest information gain [15].

$$Gain(T, X) = Entropy(T) - Entropy(T, X) \tag{4}$$

4) Naive Bayes
It is derived from the Bayesian decision theory. As formulation of it makes some naïve assumptions, so it is called naïve. By considering features as independent of given class it simplifies the learning process [16]. It is highly scalable in nature and require same number of predicators and variable during the problem learning. The values of these variables are found by solving the equating optimization problems.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{5}$$

It was firstly used by text retrieval community in 1960s and became popular method for textual problem like text categorization, document belong to one category or the other using word frequencies as the feature. It could be compatible with more advance algorithm like support vector machine using proper preprocessing of data [17]. All classifier under this category assumes that particular feature is independent of the other feature for the given class variable.

V. CONCLUSIONS

Using classification techniques, we able to understand the way in which data can be grouped and we are able to identify the class of data when new data set is available. This paper presents the detail study of various supervised machine

learning classification techniques. Out of all Decision Tree and Naïve Bayes are frequently used classifiers. Paper has presented the various application domain like business, agriculture, education where classification algorithms can be applied. Classification is widely used technique in many fields and considered as one of the hot research area.

REFERENCES

[1] S. Shwartz, Y. Singer, N. Srebro, "Pegasos: Primal Estimated sub - Gradient Solver for SVM", Proceedings of the 24th International Conference on Machine Learning, Corvallis, OR, 2007
[2] C.M. Bishop, Pattern recognition and machine learning, Springer, New York, 2006.
[3] Purushottam, Prof. (Dr.) KanakSaxena, Richa Sharma, "Efficient Heart Disease Prediction System using Decision Tree", ISBN:978-1-4799- 8890-7/15/$31.00 ©2015 IEEE.
[4] YevheniiUdovychenko, Anton Popov, IllyaChaikovsky, "Ischemic Heart Disease Recognition by k-NN Classification of Current Density Distribution Maps", 978-1-4673-6534-5/15/$31.00 ©2015 IEEE.
[5] Dana AL-Dlaeen, Abdallah Alashqur, "Using Decision Tree Classification to Assist in the Prediction of Alzheimer's Disease", 978-1-4799-3999-2/14/$31.00©2014 IEEE.
[6] Mohammed Aashkaar, Purushottam Sharma, Naveen Garg, "Performance Analysis using J48 Decision Tree for Indian Corporate world", 978-1-4673-8819-8/16/$31.00 ©2016 IEEE.
[7] B. Giraldo, Member, IEEE, A. Garde, C. Arizmendi,Member, IEEE, R. Jané, Member, IEEE, S. Benito, I. Diaz, D. Ballesteros , "Support Vector Machine Classification Applied on Weaning Trials Patients" 1-4244-0033-3/06/$20.00 ©2006 IEEE.
[8] Argyro Kampouraki, Christophoros Nikou כ and George Manis, "Robustness of Support Vector Machine-based Classification of Heart Rate Signals". 1-4244-0033-3/06/$20.00 ©2006 IEEE.
[9] U Ravi Babu, Dr. Y Venkateswarlu, Aneel Kumar Chintha "Handwritten Digit Recognition Using K-Nearest Neighbour Classifier", 978-1-4799-2876-7/13 $31.00 © 2013 IEEE DOI 10.1109/WCCCT.2014.7.
[10] Anand Shanker Tewari, Tasif Sultan Ansari, Asim Gopal Barman, "Opinion Based Book Recommendation Using Naive Bayes Classifier" 978-1-4799-6629-5/14/$31.00c 2014 IEEE
[11] Ruixi Yuan & Zhu Li & Xiaohong Guan & Li Xu "An SVMbased machine learning method for accurate internet traffic classification". Springer Science + Business Media, LLC 2008, Volume 12, Number 2, 149-156, DOI: 10.1007/s10796-008-9131-2
[12] Vidushi Sharma, SachinRai and AnuragDev , "A Comprehensive Study of Artificial Neural Networks", International Journal of Advanced Research in Computer Science and Software Engineering Research Paper,Volume 2, Issue 10, October 2012 ISSN: 2277 128X.
[13] Ms. Sonali. B. Maind and Ms. PriyankaWankar, "Research Paper on Basic of Artificial Neural Network", International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 2 Issue: 1 96 – 100.
[14] Li, Linna, and Xuemin Zhang. "Study of data mining algorithm based on decision tree." In Computer Design and Applications (ICCDA), 2010 International Conference on, vol. 1, pp. V1-155. IEEE, 2010.
[15] Quinlan, J. Ross. "Induction of decision trees." Machine learning 1, no. 1 (1986): 81-106.
[16] Maron, M. E. (1961). "Automatic Indexing: An Experimental Inquiry" (PDF). Journal of the ACM. 8 (3): 404–417. doi:10.1145/321075.321084
[17] Rish, Irina (2001). An empirical study of the naive Bayes classifier (PDF). IJCAI Workshop on Empirical Methods in AI.
[18] William S Noble, "What is a support vector machine?", Nature Biotechnology 24, 1565 - 1567 (2006) doi:10.1038/nbt1206-1565, 12 Dec 2006