International Journal of Future Generation Communication and Networking
Vol. 13, No. 3, (2020), pp. 4069–4075

Supervised Machine Learning: Predispositions, Practices and Perspectives

Pradeep Verma[1], Dr. Poornima Tyagi[2]

[1]*Department of Computer Science, (Himalayan Garhwal University), Uttrakhand, India*
pradeepiilm@gmail.com
[2]*Department of Computer Science, (Himalayan Garhwal University), Uttrakhand, India*
poornima.tyagi@gmail.com

Abstract

Technology is growing rapidly in today's fast moving world and everything is now expected to be done automatically by machine or bots instead of being driven manually by humans be it at office, home or outside markets. In every field, the intention of human is to reduce manual efforts and drive fastest results with high accuracy and efficiency.

Machine learning is one such growing technology which is being adopted now a days in every field for improving efficiency. It is a subset of Artificial Intelligence (AI) which learns automatically by analysing data, repeatedly by identifying patterns, without any human intervention.
Supervised , Unsupervised and Reinforcement Learning are the three major types of techniques used in Machine Learning. In this paper we have explained various Supervised Machine Learning techniques.

Keywords: *Supervised Machine Learning, Classification, Regression, Unsupervised Learning, Reinforcement Learning.*

I. INTRODUCTION

In this fast moving world where automation is the primary focus to bring in efficiency and accuracy with minimal human intervention, the technology plays a vital role. Artificial Intelligence (AI) and Bots technology is replacing human beings in every aspect of life and the machines are meant to be learning and improving themselves through self learning implicit model which learns through its own past experiences and repeated tasks for regular improvement in its performance. This aspect of AI is termed as Machine Learning (ML).

Machine Learning (ML) is a subset of Artificial Intelligence (AI) technique for analysing data for prediction to take better and faster decisions. Machine learning is a technical concept which aims in deriving
hidden patterns where computers are trained with the help of training data for finding these hidden patterns

themselves, from large dataset, and to build a model which can be tested later through the tested data set for accuracy. The training data set keeps on improving the results by reiterating the process and re-analysing the received output and keeps on improving itself after performing the tasks. Hence, we can understand that ML is an implicit learning environment which learns by doing to improve its overall performance after every task and enriches its experience for building improved model.

II. OBJECTIVE

The objective of this paper is to familiarize with the various types of Supervised Machine Learning techniques through a comprehensive study of the techniques and its related methods.

ISSN: 2233-7857 IJFGCN

III. REVIEW OF LITERATURE

Machine Learning Algorithm helps in predictions. Few of the real life application of Machine Learning can be understood through the following research papers:

[1] Gianey, H. K., & Choudhary, R. in their research paper have compared the various supervised machine learning techniques on a Germany based dataset to identify credit risk. The goal of the research was to test the various supervised machine learning techniques on the data set and find the result of each technique and compare them. The outcome as stated in the paper are very straight forward and clearly states that any particular supervised technique cannot be considered as best fit for classification over any data set since it depends on multiple factors like nature of the data set, problem statement and selection of attributes from the dataset. The conclusion of the paper after applying different supervised machine learning techniques is that in absence of availability of defined algorithm to achieve the desired output, it should be left on the machine to analyse the data set and decide itself the best suitable algorithm to be applied to achieve the desired outcome.

[2] Dhankhad, S., Mohammed, E. A., & Far, B. in their research paper have attempted to identify the best possible algorithm suitable for the imbalanced data set to identify Credit-Card fraud. The data set they have used for their analysis is an imbalanced data set comprising of real-world transactions. The authors have applied the various Supervised Machine Learning algorithms over the data set and has compared the same with the super-classifier implemented by them using the Ensemble Learning Method. Post implementation of various techniques and algorithms, the authors have come to the conclusion that the data set considered by them is highly imbalanced and for such kind of data set under-sampling for class distribution must be done in order to reduce bias of data set.

[3] Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. have compared three approaches (Logistic Regression (LR), Random Forests (RF) and Support Vector Machines (SVM)) of Supervised Machine Learning (SML) and have attempted to identify the better approach for identifying Credit-Card Fraud in their research paper. The data set considered them for study is an international Credit-Card transactions dataset based on real life transactions.

In this paper the authors have explained that they have used the data under-sampling technique to eliminate the data imbalance of the training data set and then have applied the above-identified three SML techniques on the training data set. Through this testing of all the three algorithms, the authors have come to conclusion that RF and SVM techniques have the capability of selecting attributes on their own and are capable of performing well in high-dimension data set.

[4] Saravanan, R., & Sujatha, P. in their research paper have studied and explained the various machine learning techniques including Supervised Machine Learning(SML), Unsupervised Machine Learning(USML), Semi-supervised Machine Learning(SSML) and Reinforcement Learning(RL) on four parameters i.e.Aim, Methodology, Advantages and Dis-advantages. The authors have then explained the various classification methods of SML in detail and then have come to the conclusion that SML require human guidance and involvement for preparing model with the help of training data set where as in USML such human guidance and involvement is not required. Further SML can be used for preparing classification model for real-life use in future.

IV. METHODOLOGY

Machine Learning algorithms can be broadly classified into three major categories:
1. *Supervised Learning*, is an machine learning algorithm which uses labelled data to map the given input with the desired output. In this technique, the value of both input and output data is given and our task is to create a model through experience for mapping input to their corresponding output. [5].

The complete input data set is divided into two types: training data set and testing data set. The proportion of training data set is much higher than the tested data set. The objective of the algorithm is

ISSN: 2233-7857 IJFGCN

to build a model on the basis of the training data which is capable to predict the exact input data to its corresponding output. Once the model has been created through the training data set, the tested data set is then applied on the said model for prediction of desired output. [6]

Supervised learning is further classified into two groups:

A. *Classification,* is a method where the objective is to classify the dataset into classes i.e. Yes/No, Fraud/Not-fraud, Male/Female/Transgender to name a few.

In this method, the classification algorithms classifies as well as maps the data set with the defined classes. Each of the class, at the back-end in the algorithm, is classified and mapped with a numeric number which is auto-generated by the algorithm. The result of the algorithm is produced in the form of categorical classes i.e. Yes/No OR Male/Female/Transgender and not in the form of the auto-generated numeric numbers assigned to these classes.

Identifying Fraud Detection, Image Classification, Customer Retention and Diagnostic are few of the real world examples where Classification technique is frequently used.[7]. Few of the Classification techniques are given below:

i) *K-Nearest Neighbors (KNN):* This is a simple yet important classification algorithm which works on similarity of the data attributes. The algorithm classifies data attributes on the basis of value of nearest available data points and clusters the new data point into the group having similar data attributes. In this technique, the values of two clusters would always vary and will not be the same. Various methods used in kNN for identifying the data value and the closest cluster are Euclidian, Manhattan or Minkowski, the equations of which are given below [8] .

$$\text{Euclidean} - \sqrt{\sum_{i=1}^{k}(x_i - y_i)} \qquad (1)$$

$$\text{Manhattan} - \sum_{i=1}^{k}|x_i - y_i| \qquad (2)$$

$$\text{Minkowski} - \left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q} \qquad (3)$$

Whenever a new (unclassified) data point is introduced, its distance from all existing data points basis the similarity of the attributes is calculated with the help of any of the above equations (as applicable) and the new unclassified data point is assigned to the nearest cluster having similar attributes.

ii) *Support Vector Machine (SVM):*
A Support Vector Machine (SVM) is one of the popular supervised technique broadly used for classification, though we can apply SVM in both Regression as well as Classification. The prime goal of SVM is to classify the given data set in ideal hyper-plane. SVM uses the mathematical vector space concept to find the maximum far away limit (decision-boundary) from the hyper-plane where the data set of two different classes should rest. [9]

ISSN: 2233-7857 IJFGCN

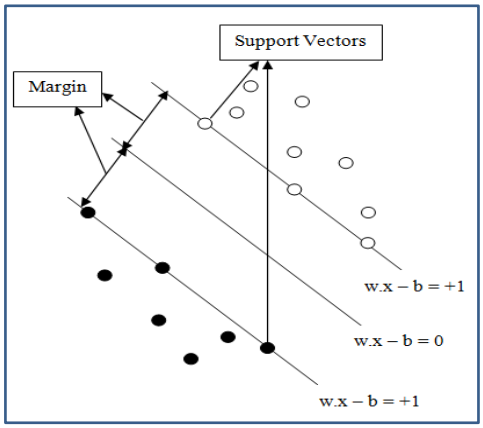

Fig1: Support Vector Machine [10]

iii) Logistic regression(LR):
Logistic regression is a popular binary classification technique. In this technique, the complete data set is classified into binary values i.e. 0 and 1. To classify and determine the binary values Sigmoid function (S shaped curve) is used along with logistic regression [11]. It is widely used in industries and education field for prediction.

Example : To check if a new joinee will stay in the organization beyond 6 months or will leave, prediction on probability of employee retention beyond 6 months will fall in binary values [0,1]. If the predicted probability is less than 0.5 then the person will stay else will leave.

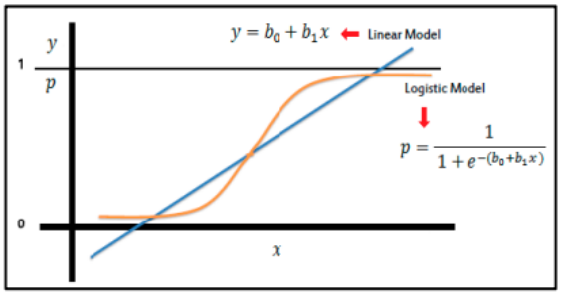

Fig.2: Logistic regression & Sigmoid Function[12]

iv) Decision Tree (DT) Classification:
Decision Tree is a powerful SML technique used for predictions as well as classification. It classifies the data set based on the data attribute value. The algorithm has a "Tree" like structure comprising of multiple nodes and branches. Each node has an attribute and the node which gives best suitable information is classified as "Root Node". Gaining further information from the root note and finding out the suitable data attribute, further internal nodes and branches are created till the time the leaf node (which holds the label of the class) is arrived at. The number of branches depends upon the test performed on the attributes of the internal nodes.[13]

ISSN: 2233-7857 IJFGCN

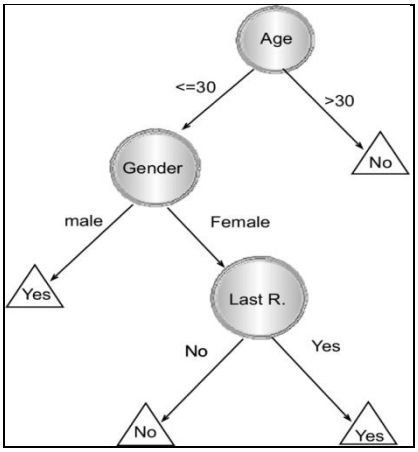

Fig3: Decision Tree Classification [14]

v) Gaussian Naive Bayes(GNN) algorithm for classification:
This Supervised Machine Learning classification technique is an extension of "Naive Bayes algorithm" that uses Guassian Distribution (also named as 'Normal Distribution') works on Joint Probability Function. By this algorithm we capture the values of mean, variance and standard deviation. The data value output, basis the similarity of the classes, is then represented in a 'Bell' shaped curve (Guassian PDF)[15]. Following equation defines the Guassian PDF (Probability Density Function):

$$N(\mu, \sigma^2)(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (4)$$

Where, μ=mean; σ2=variance.
B. *Regression*, is another Supervised Machine Learning classification method where the output is always in the form of a numeric value. It is a predictive technique used for establishing relationship between independent and dependent variables. Regression is mostly used for predictions (eg. Predicting population growth, future cost of a house, predicting an advertisement's popularity), forecasting (eg. Stock market, weather forecasting) and time series modelling to name a few[16].

One of the popularly used Regression Algorithm is Linear Regression.

i) Linear Regression :
Linear Regression is a frequently used modelling technique for carrying out predictions where a relationship between the one or more than one independent variable (input data) and one dependent variable (output data) is established through the best-fit straight line known as "Regression Line".
The value of independent variable can be discrete or continuous where as the value of dependent variable is always continuous.[17]

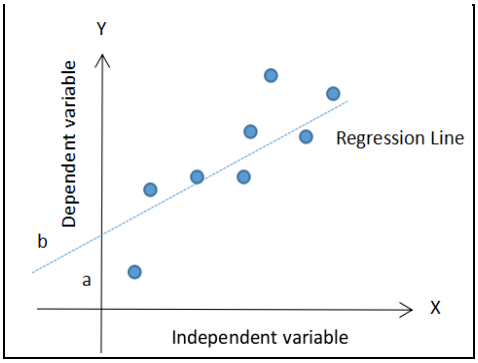

Fig4: Linear Regression

ISSN: 2233-7857 IJFGCN

The equation to solve the linear regression problem is:

$$Y = a + bX + e \quad (5)$$

where, Y is the dependent (output) variable, X is independent (input) variable, a is the intercept and b is the slope of the line and e is error.

2. *Unsupervised Learning*, is an machine learning algorithm where the input data is unlabelled and the output is not known. The input data set implicitly performs the test on itself to find out similar nature data sets and makes clusters. The main challenge of unsupervised learning is to find out hidden pattern and segregate input data basis their attributes. [18]

3. Reinforcement Learning, is a machine learning algorithm where no information related to output is given. In this type of algorithm, the agent is put in to an environment and learns to perform in the environment by doing some actions. These actions either attract reward or punishment to the agent. The ultimate goal of Reinforcement learning is to learn itself within the environment and maximize its rewards through experience. [19]
As compared to other two algorithms, the reinforcement learning is the most difficult learning algorithm for implementation and is a subset of deep learning.

V. CONCLUSION

The objective of this paper was to do a comprehensive study of Supervised Machine Learning techniques and its related methods. Through the study, we have come to a conclusion that the each of the technique has its unique features and application and hence we cannot classify any one technique suitable for any particular data set.

A comparative of the Supervised Machine Learning techniques, defined in this paper, basis our study on three parameters i.e. accuracy, speed of learning with respect to number of attributes and instances and data retrieval cost is as follows:

Techniques	Accuracy	Speed	Retrieval Cost
KNN	Good	High	Expensive
SVM	High	Medium	Expensive
DT	Good	High	Medium
GNN	Low	Good	Medium

Table 1 : Comparison of SML Techniques

All the techniques are equally important and commonly used basis the data set and features selection however, the accuracy of result can be improved by implementing two or more classification algorithm together under suitable conditions.

REFERENCES

1. Gianey, H. K., & Choudhary, R. (2018). Comprehensive Review On Supervised Machine Learning Algorithms. Proceedings - 2017 International Conference on Machine Learning and Data Science, MLDS 2017, 2018-January, 38–43. https://doi.org/10.1109/MLDS.2017.11
2. Dhankhad, S., Mohammed, E. A., & Far, B. (2018). Supervised machine learning algorithms for credit card fraudulent transaction detection: A comparative study. Proceedings - 2018 IEEE 19th International Conference on Information Reuse and Integration for Data Science, IRI 2018, 122–125. https://doi.org/10.1109/IRI.2018.00025

ISSN: 2233-7857 IJFGCN

3. Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. Decision Support Systems, 50(3), 602–613. https://doi.org/10.1016/j.dss.2010.08.008
4. Saravanan, R., & Sujatha, P. (2018). Algorithms : A Perspective of Supervised Learning Approaches in Data Classification. 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), Iciccs, 945–949.
5. Annina Simon, Mahima Singh Deo, Mahima Singh Deo, S. Venkatesan, D.R. Ramesh Babu,D.R. Ramesh Babu, "An Overview of Machine Learning and its Applications", International Journal of Electrical Sciences & Engineering (IJESE);Vol1, Issue 1; 2015 pp. 22-24
6. S.B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques", Informatica 31 (2007) 249-268
7. Nath, A., Agarwal, S., & Ghosh, A. (2016). Classification of Machine Learning Algorithms. International Journal of Innovatice Research in Advanced Engineering, 3(March), 6–11
8. Kalaivani, P., & Shunmuganathan, K. L. (2014). An improved K-nearest-neighbor algorithm using genetic algorithm for sentiment classification. 2014 International Conference on Circuits, Power and Computing Technologies, ICCPCT 2014, 1647–1651. https://doi.org/10.1109/ICCPCT.2014.7054826
9. Muhammad, I., & Yan, Z. (2015). Supervised Machine Learning Approaches: a Survey. ICTACT Journal on Soft Computing, 05(03), 946–952. https://doi.org/10.21917/ijsc.2015.0133
10. Pradhan, A. (2012). SUPPORT VECTOR MACHINE-A Survey. 2(8), 82–85
11. Machine Learning Mastery. Available at: http://machinelearningmastery.com/logistic regression-for-machine-learning/ [Accessed 12 Aug. 2017].
12. Gianey, H. K., & Choudhary, R. (2018). Comprehensive Review On Supervised Machine Learning Algorithms. Proceedings - 2017 International Conference on Machine Learning and Data Science, MLDS 2017, 2018-Janua, 38–43. https://doi.org/10.1109/MLDS.2017.11
13. Peng Ye, "The decision tree classification and its application research in personnel management", Proceedings of 2011 International Conference on Electronics and Optoelectronic, 2011, pp. 1-4
14. L. Rokach, O. Maimon, "Top – Down Induction of Decision Trees Classifiers – A Survey", IEEE Transactions on Systems
15. Mitchell, T: Machine Learning. McGraw-Hill1997; ISBN-13: 987-1-25-909695-2.
16. Thomas G. Dietterich, "Machine-Learning Research", AI Magazine Volume 18 Number 4 (1997)
17. Dasgupta, A., Sun, Y. V., König, I. R., Bailey-Wilson, J. E., & Malley, J. D. (2011). Brief review of regression-based and machine learning methods in genetic epidemiology: The Genetic Analysis Workshop 17 experience. Genetic Epidemiology, 35(SUPPL. 1), 5–11. https://doi.org/10.1002/gepi.20642
18. TM Mitchell, "The discipline of machine learning", (Carnegie Mellon University,School of Computer Science, Machine Learning Department, 2006)
19. Leslie Pack Kaelbling, Michael L. Littman, Andrew W. Moore "Reinforcement Learning: A Survey", Journal of Artificial Intelligence, Research 4 (1996) 237-285, May 1996.

ISSN: 2233-7857 IJFGCN