SUPERVISED MACHINE LEARNING ALGORITHMS: A COMPARATIVE STUDY

Sangeeta Vaibhav Meena
Assistant Professor
Department of Computer Science
S. S. Jain Subodh P.G. (Autonomous) College, Jaipur, Rajasthan, India

***Abstract*: -** Machine learning provides computers with the ability to learn and improve them from the past knowledge instead of being explicitly programmed. In real life there are many applications of machine learning such as virtual assistance, self-driving cars, email spam classification, image and speech recognition, cancer tumour cells identification, sentiments analysis and many more. Machine learning can be applied through supervised learning, unsupervised learning and reinforcement learning. Supervised machine learning aims is to build model that make likelihoods grounded on evidences in the presence of uncertainty which takes known set of input and response data. This paper focuses on classification and regression algorithms that play a vital role in supervised machine learning, whose goal is to assign a class to an observation from a finite set of classes. The different algorithms discussed are Support Vector Machines, Naïve Bayes, K- Nearest Neighbor, Linear Regression, Decision Trees, Artificial Neural Networks, Random Forest and Logistic Regression. The aim of this paper is to provide a comparative analysis of different supervised machine learning algorithms and provide in depth knowledge by comparing these algorithms on different performance parameters. This paper provides new dimensions in the field of machine learning by strengthening the basis of classification and regression algorithms.

Keywords: - *Artificial Neural Networks, K-Nearest Neighbor, Linear Regression, Machine Learning, Naïve Bayes, Supervised Learning, Support Vector Machines, Random Forest*

1. INTRODUCTION

Machine learning enables computers to learn as human beings learn in their life that is learning from experience. This learning process can be through some guidance as in the case of supervised learning or can be done by observing as in the case of unsupervised learning or it can be through encouragement or punishment as in the case of reinforcement learning [1]. In spite of being explicitly programmed, machine learning has the ability to improve system's behaviour based on its experience. There are several machine learning tools in market such as Shogun, Apache Mahout, Scikit-Learn, TensorFlow, Apache Spark Mlib etc. performs various applications of Machine Learning – Classification, Regression, Clustering, Dimensional Reduction, Associations.

Classification is the process of prediction a class or category of a new observation from a set of predefined categories. Classification comes under the category of supervised learning where inputs are provided with labels and algorithm learns to predict an output by generating a function that maps inputs to outputs [2]. When we are talking about classification the predictive model predicts a discrete class label output for example, bank evaluates whether a customer pay loan or not before disburse a loan. This can be done by considering some factors such as customer's income, saving, financial history, age etc. Here in this example there are two possible outcomes- pay loan or not pay loan. This class of classification is called as binary. Multi class captures more than two categories, applications such as video – audio categorization, bioinformatics, text analysis; etc. comes under this category. Classification is a vital feature to separate huge datasets into classes for the purpose of decision making, dimensionality reduction, rule generation, pattern recognition, data mining etc. [3]. In contrast to classification, regression problem is when the output variable is not discrete but in continuous form such as height, weight, salary. A regression model tries to fit the datasets with the best hyper plane which goes through the points.

The paper is organized as follows: section 2 presents an overview of supervised learning algorithms, we discussed eight supervised machine learning algorithms in this paper; section 3 illustrates metrics for evaluating algorithms; section 4 discussed the various factors that affects the performance of the algorithms; finally in section 5 we draw some conclusions.

2. OVERVIEW OF SUPERVISED LEARNING ALGORITHMS

Supervised Learning problems can be further grouped into two types of problems: Regression and Classification problems. Both problems have the same goal as to construct a model that predicts the value of the dependent attributes from the attribute variables. The difference between the two is that the dependent variable in Regression is numerical while Classification works on categorical data. The commonly used algorithms which are discussed in this paper are as follows.

2.1. Naïve Bayes Classifier

Naïve Bayes is a simple but powerful algorithm based on Bayes theorem that works on conditional probability that somewhat will happen, given that somewhat else has already happened. The algorithm works with the assumption that every pair of features is independent to each other. Bayes Theorem calculates the posterior probability P(h|x), from P(h), P(x) and P(x|h). It determines

the likelihood of a particular hypothesis given some observed evidence by the use of a prior probability over hypothesis [5]. It is generally useful for very large data sets, performs quickly if the independence assumption actually holds and requires less training data. Some applications of Naive Bayes in real world like recommendation systems like Amazon, Netflix; classify a news article as technology, politics, entertainment, sports, etc.; check whether a piece of text expressing positive or negative.

2.2. Linear Regression

Linear regression is a predictive analysis algorithm which is used to estimate real values like total sales, cost of a product, and number of calls etc., based on continuous variables [6]. Linear regression establishes relationship between dependent and independent variables by fitting a best line, known as regression line. This regression line is represented by a linear equation Y=a*X+b. Here, Y is a dependent variable, X is an independent variable, a is the slope and b is intercept. The value of a and b is derived based on minimizing the sum of squared difference of the distance between the regression line and the data points. Linear regression is of two types: Simple Linear regression where there is one independent variable and Multiple Linear regression where there is more than one independent variable.

2.3. Logistic Regression

Logistic regression is the simplest non-linear classifier with a linear combination of parameters and nonlinear function for binary classification. Logistic regression is a machine learning algorithm for classification which is based on predictive learning model. It is a statistical method that is used to measure the result with a dichotomous variable (binary values like 0/1, pass/fail, true/false, yes/no) by analysing a data set in which there are one or more independent variables. The objective of logistic regression is to find the best fitting model to describe the relationship between the dependent and a set of independent variables [7]. In this algorithm, the probabilities defining the possible results of a single sample are modelled using a logistic regression. Logistic regression produces a logistic curve where values are limited between 0 and 1.

2.4. Decision Tree

Decision Trees constructs classification models in the form of a tree structure. It is a simple algorithm where feature importance is clearly stated and relations can be easily viewed. The tree starts with a root node then a series of branches whose intersections are called decision nodes and finally end up on leaves nodes which represents a decision or classification [8]. Decision tree creates a sequence of rules that are used to classify future observations when there is a given body of already labelled observations. The tree is constructed in such a way that each node must divide the set of initial observations in a best possible way in order to formulate the rule. Decision tree is used to handle both the categorical as well as numerical data. The decision tree's accuracy is heavily depends upon the decision of making strategic splits. There are many algorithms to decide to split a node into one or more nodes. Some commonly used algorithms in decision tree are Gini Index, Chi-Square, Information Gain, and Reduction in Variance.

2.5. Random Forest

Random forests are an ensemble learning method for classification, based on a multitude of decision trees. It works by creating an assembly of decision trees at training time and outputting the class that is the mode of the classes of the individual decision trees. A number of decision trees on various sub-samples of datasets are fitted and averaged in order to improves the predictive accuracy of the model and also avoids the pitfalls of over-fitting [4]. The samples in the data sets are drawn with replacement but the sub-sample size is at all times the same as the original input sample size. The steps for constructing random forest are as follows-

1. Take a number X of observations from the starting dataset.
2. Take a number K of the M variables available.
3. Create a decision tree on this dataset.
4. Repeat Steps 1 to 4 for N times so as to obtain N trees.

2.6. K-Nearest Neighbour Algorithm

K-Nearest neighbour is a non-parametric technique which does not mark any postulation regarding data distribution. It is easy to understand and interpret and is useful in the field of pattern recognition, statistical estimation, data mining, intrusion detection etc. The concept of this algorithm is that a new problem instance is classified on the basis of K nearest samples. This K is a positive number that can be selected on the basis of some good heuristic technique. To find the optimal value of K, training and validation curves are plotted. The higher value of K minimizes the effect of the noise on the classification. A commonly used weighted scheme is giving each and every neighbour a weight of 1/d where d is the distance to the neighbour. Distance metric for discrete variables is mainly calculated by Hamming distance while for continuous variables, Euclidean distance is used [9]. This algorithms benefit is ease of interpretation and low calculation time compared with other algorithms for classification but needed a lot of memory space to store all of the data.

2.7. Support Vector Machine

Support vector machine is classification as well as regression technique where each data point is plotted as a point in space in order to finds a hyperplane, a decision frontier that maximizes the distance between the closest support vectors of separate classes [10]. Test inputs are then mapped into the same space in order to predict their category based on which side of the hyperplane they lay down. For non-linear decision boundaries SVM use kernel, functions which take low dimensional input space and transform into higher dimensional space in order to convert not separable problem into separable problem. SVM are mainly used

in pattern recognition problems for example, hand writing recognition, stock marketing forecasting, email spam filtering etc. SVM renders robustness to even high dimensionality where each dimension represents a feature. It able to separate classes fast with high accuracy, less overfitting and less amount of memory required [11].

2.8. Artificial Neural Networks

Artificial neural networks or simply saying neural networks are considered as robust classifiers. Neural network is an information processing system that is inspired by the biological nervous system. It is consists of a large number of highly interconnected processing elements called neurons. An artificial neural network is configured through a learning process for a specific purpose such as data classification, object detection or pattern recognition. Like the biological system, learning in neural networks involves adjustment to the synaptic connections that exists between the neurons [12]. All the classification tasks depend upon the labelled datasets involves supervised learning where person transfer their knowledge to the dataset, so that a neural network can learn the correlation between data and labels. Neural network model can be defined as feed-forward where a unit feeds its output to all the units of next layer but no cycle is involved means no feedback loops. The backpropagation algorithm involves two steps: starting with feed-forwarding the values, then calculate the error and propagate it back to the previous layers.

Table 1: Strengths and weaknesses of Supervised Learning Algorithms

S.No.	Algorithm	Strengths	Weaknesses
1	Naïve Bayes	Naive Bayes classifiers are extremely fast compared to more sophisticated methods. It requires a small amount of training data to estimate the necessary parameters.	Naive Bayes performs less as compared to other classifiers and known to be a bad estimator.
2	Linear Regression	Linear regression is simple to understand, Good interpretability and Space complexity is very low.	Linear Regression Is Limited to Linear Relationships, sensitive to Outliers, prone to overfitting of the data.
3	Logistic Regression	Logistic regression is designed for classification purpose, and is mostly used for understanding the influence of several independent variables on a single resultant variable.	Logistic Regression works only when the predicted variable is binary, it assumes that all predictors are independent of each other and data is free of missing values.
4	Decision Tree	Decision Tree can handle both categorical and numerical data. It is simple to understand and visualise, and requires little data preparation.	Decision trees can be unstable as small variations in the data might result in a completely different tree being generated. It can create complex trees that do not generalise well.
5	Random Forest	Random forest classifier is more accurate than decision trees in most of the cases. It helps in reducing over-fitting.	Random Forest are slow in real time prediction and faces difficulties in implementation.
6	K- Nearest Neighbor	KNN algorithm is simple to implement, effective if training data is large and also robust to noisy training data.	In KNN the computation cost is high to determine the value of K as it needs to compute the distance of each instance to all the training samples.
7	Support Vector Machines	SVM uses a subset of training points in the decision function, memory efficient and effective in high dimensional spaces.	SVM does not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation.
8	Neural Networks	ANN is easy to use, with few parameters to adjust. A neural network learns and its reprogramming is not needed. It is applicable to a wide range of problems in real life.	High processing time is required if the neural network is large. Faces difficulties in knowing how many neurons and layers are necessary. Learning can be slow.

3. Evaluation Metrics for Machine Learning Algorithms

Evaluation metrics elucidate the performance of a model and also helps in better understanding of its working [14]. Capability to discriminate among model results is an important aspect of evaluation metrics. In this paper we discussed some commonly uses metrics for evaluating classification algorithms.

3.1. Classification Accuracy

Classification Accuracy means the ratio of number of correct predictions to the total number of input samples. If there is equal number of samples belonging to each class then only it works well. Classification accuracy is good enough but gives the false alarm of achieving high accuracy. Problem arises when the cost of wrong classification of the minor class samples are very high.

3.2. Logarithmic Loss

Logarithmic Loss or Log Loss works well for multi-class classification problems. Its mechanism is to penalise the false classifications. In logarithmic loss, the classifier must allot probability to each class for all the given samples. It has no upper limit and exists between (0, ∞). Higher accuracy occurs only when log loss is near to 0, while lower accuracy when log loss is away from 0.

$$logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} log(p_{ij})$$

3.3. Area Under Curve

Area under curve (AUC) is a metric broadly used for evaluation in binary classification problem. Area under curve classifier is equal to the probability that the classifier will rank a randomly chosen negative example lower than a randomly chosen positive example.

Two basic terms used in AUC are:

True Positive Rate: True Positive Rate is related to the quantity of positive data points that are correctly considered as positive, with respect to all positive data points. True Positive Rate is defined as TP/ (FN+TP).

$$\text{True Positive Rate} = \frac{\text{True Positive}}{\text{False Negative} + \text{True Positive}}$$

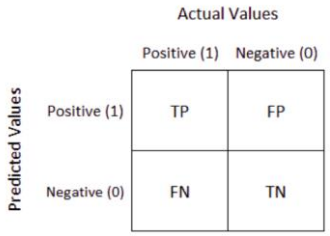
False Positive Rate: False Positive Rate is related to the quantity of negative data points that are mistakenly considered as positive, with respect to all negative data points. False Positive Rate is defined as FP / (FP+TN).

$$\text{False Positive Rate} = \frac{\text{False Positive}}{\text{False Positive} + \text{True Negative}}$$

Both False Positive Rate and True Positive Rate have values in the range 0 and 1.

3.4. Confusion Matrix

Confusion matrix gives output in the form of a matrix which describes the complete performance of the model.



Four terms used in confusion matrices are:-

True Positives: - Cases where we predicted YES and also the actual output was YES.

True Negatives: - Cases where we predicted NO and also the actual output was NO.

False Positives: - Cases where we predicted YES but the actual output was NO.

False Negatives: - Cases where we predicted NO but the actual output was YES.

3.5. F1 Score

F1 Score is the mean between precision and recall and is used to measure the test's accuracy. The range of F1 score is between 0 and 1. It tells us how precise and robust is the classifier. Precise means how many instances it classifies correctly and by robust means it does not miss a significant number of instances. F1 score efforts to discover the balance between precision and recall.

$$F1 = 2 * \frac{1}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}}$$

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Precision: - Precision is the number of correct positive results that is divided by the number of positive results predicted by the classifier.

Recall: - Recall is the number of correct positive results that is divided by the number of all relevant samples including true positive and false negative.

3.6. Mean Absolute Error

Mean Absolute Error measures how far are the predictions from the actual output. It is the average of the difference between the predicted values and the original values. Here the limitation is that it doesn't give us any idea of the direction of the error, it means we are not in a situation to judge whether we are over predicting the data or under predicting the data.

3.7. Mean Squared Error

Mean Squared Error is pretty similar to Mean Absolute Error, with only the difference is that Mean Square Error takes the average of the square of the difference between the predicted values and the original values. In this metric, more focus is on the larger errors as the effect of larger errors become more pronounced then smaller error. The advantage of Mean Square Error over the Mean Absolute Error is that in the former it is easier to compute the gradient while latter requires complicated linear programming tools to compute the gradient.

4. FACTORS AFFECTING THE QUALITY OF ALGORITHMS

Learning algorithms are evaluated on the basis of their ability to correctly predict the class of the observations. Classification not only means placing the observation at the correct category but also not to place them in the incorrect category [19]. Several factors play a vital role in the quality of the algorithms. Some of the factors discussed in this paper are:-

4.1. Size of observations

The number of observations is an important factor that affects the performance of an algorithm. If the model uses fewer observations then there is more difficult to analyse while more observations led to the need for high computer memory and take longer time for analysis.

4.2. Data normalization

Normalization or scaling plays an important role in machine learning algorithms especially in classification. Algorithm may be biased towards a specific set of attributes if the data set is not normalized and this result into poor accuracy values [15]. Some of the scaling methods that can be used are Min-Max normalization or Zero mean normalization, Decimal Scaling and vector normalization.

4.3. Noise handling

Noise is the common problem that affects the quality of the algorithm. It can be defined as the random error in the data set or variance in a measured variable [16]. There are two types of data noise: attribute noise or class noise. When there are erroneous values in the independent attributes of a data-set then attribute noise occurs. Class noise occurs when there are incorrect values in the dependent attributes. Noise filtering, polishing and boosting methods are used to remove data noise.

4.4. Estimation methods

The classification accuracy of any algorithm should not be judged in a single experiment. An effective way is to use cross validation which practices a leave one out kind of technique and averages the accurateness of all the iterations.

4.5. Diversity in training data

Machine learning algorithms works on two set of datasets- training and testing datasets. Classification model built using training datasets in order to predict a class of unknown dataset. In order to classify correctly the training dataset consists of instances which are different from one another and no two instances provide same kind of information. This would help the classifier to learn more different types of patterns.

4.6. Attribute redundancy

Some attributes may not provide much information to the classifier means they do not form important entries in the feature vector. The performance of the classifier can be speed up if these attributes can be removed. Dimensionality reduction can be done using algorithms like Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA) in order to reduce the feature space effectively. Removal of attribute can be done using information such as Mutual information of attributes, Information gain and Chi-square ratio.

4.7. Problem of Outliers

The data points that are inconsistent with the majority of the data values are called the outliers [17].These are the points which are tough to classify as they do not hold the similar feature properties with the most of the data. Univariate, Multivariate and Minkowski error are some of the good methods which remove outliers.

4.8. Underfitting and Overfitting

Underfitting refers to the situation in which an algorithm cannot model the training data and also cannot generalize to new data. An under-fitted classification model is not suitable as it will provide poor pperformances on the training data. It is easy to detect given a good performance metric and can be resolved by trying alternate machine learning algorithms. Overfitting refers to the situation in which an algorithm models the training data too well. It learns the details of data as well as noise in the training data to the extent that it badly affects the performance of the model on new data. The quantity of training data has an important

role as poor quantities of training data could lead to overfitting [18]. Overfitting problem is more likely with the nonlinear and nonparametric models.

Table 2: Comparison of Supervised Learning Algorithms

S. No.	Algorithm	Probem Type	Generative or Discriminative?	Training Speed	Prediction Speed	Capable to handle irrelevant features (separates signal from noise)?	Features need Scaling?	Performs well with small number of observations?	Learns feature interactions automatically?	Average Predictive Accuracy	Loss Function
1	Naive Bayes	Classification	Generative	Fast	Fast	Yes	No	Yes	No	Lower	$-\log P(X, Y)$
2	Linear Regression	Regression	Discriminative	Fast	Fast	No	No (unless regularized)	Yes	No	Lower	Square Loss: $(\hat{Y} - Y)^2$
3	Logistic Regression	Classification	Discriminative	Fast	Fast	No	No (unless regularized)	Yes	No	Lower	$-\log P(Y \mid X)$
4	Decision Trees	Either	Discriminative	Fast	Fast	No	No	No	Yes	Lower	Either $-\log P(Y \mid X)$ or Zero-one Loss
5	Random Forest	Either	Discriminative	Slow	Moderate	Yes (unless noise ratio is very high)	No	No	Yes	Higher	Mean Squared Error
6	K-Nearest Neighbors	Either	Discriminative	Fast	Depend on n	No	Yes	No	No	Lower	Zero-one Loss
7	Support Vector Machines	Either	Discriminative	slow	Fast	Yes	Yes	No	Yes	Higher	Hinge Loss
8	Neural Networks	Either	Discriminative	Slow	Fast	Yes	Yes	No	Yes	Higher	Sum Squared Error

5. CONCLUSION

This paper provides a better understanding of supervised learning algorithms by discussing their strengths and weakness; and also provides a comparative analysis of the algorithms on the basis of various parameters. The important question when dealing with supervised learning algorithms is not whether a learning algorithm is superior to another one or not, but emphasis is on which conditions a particular algorithm can significantly outperform others on a given problem. Some problems are very specific and require a unique approach while some other problems are very open and need a trial and error approach. The likelihood of incorporating two or more algorithms together in order to solve a problem should be investigated so that the strength of one algorithm can complement the weakness of other.

REFERENCES

[1] Pradeep, K. R., & Naveen, N. C. (2017). A Collective Study of Machine Learning (ML) Algorithms with Big Data Analytics (BDA) for Healthcare Analytics (HcA). International Journal of Computer Trends and Technology (IJCTT), 47(3), 149-155.

[2] Osisanwo, F. Y., Akinsola, J. E. T., Awodele, O., Hinmikaiye, J. O., Olakanmi, O., & Akinjobi, J. (2017). Supervised machine learning algorithms: classification and comparison. International Journal of Computer Trends and Technology (IJCTT), 48(3), 128-138.

[3] Zhang, C., Liu, C., Zhang, X., & Almpanidis, G. (2017). An up-to-date comparison of state-of-the-art classification algorithms. Expert Systems with Applications, 82, 128-150.

[4] Wu, D., Jennings, C., Terpenny, J., Gao, R. X., & Kumara, S. (2017). A comparative study on machine learning algorithms for smart manufacturing: tool wear prediction using random forests. Journal of Manufacturing Science and Engineering, 139(7), 071018.

[5] Bouckaert, R. R. (2004, December). Naive bayes classifiers that perform well with continuous variables. In Australasian joint conference on artificial intelligence (pp. 1089-1094). Springer, Berlin, Heidelberg.

[6] Seber, G. A., & Lee, A. J. (2012). Linear regression analysis (Vol. 329). John Wiley & Sons.

[7] Ng, A. Y., & Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In Advances in neural information processing systems (pp. 841-848).

[8] Song, Y. Y., & Ying, L. U. (2015). Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry, 27(2), 130.

[9] Yoon, J. W., & Friel, N. (2015). Efficient model selection for probabilistic K nearest neighbour classification. Neurocomputing, 149, 1098-1108.

[10] Nayak, J., Naik, B., & Behera, H. (2015). A comprehensive survey on support vector machine in data mining tasks: applications & challenges. International Journal of Database Theory and Application, 8(1), 169-186.

[11] Meyer, D., & Wien, F. T. (2015). Support vector machines. The Interface to libsvm in package e1071, 28.

[12] Neocleous, Costas, and Christos Schizas. "Artificial neural network learning: a comparative review."Hellenic Conference on Artificial Intelligence. Springer, Berlin, Heidelberg, 2002.

[13] https://www.dataschool.io/comparing-supervised-learning-algorithms/

[14] Jeni, L. A., Cohn, J. F., & De La Torre, F. (2013, September). Facing Imbalanced Data--Recommendations for the Use of Performance Metrics. In 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (pp. 245-251). IEEE.

[15] Jayalakshmi, T., & Santhakumaran, A. (2011). Statistical normalization and back propagation for classification. International Journal of Computer Theory and Engineering, 3(1), 1793-8201.

[16] Wang, H., Zhang, Q., Jiao, L., & Yao, X. (2016). Regularity model for noisy multiobjective optimization. IEEE transactions on cybernetics, 46(9), 1997-2009.
[17] Aggarwal, C. C. (2015). Outlier analysis. In Data mining (pp. 237-263). Springer, Cham.
[18] Jabbar, H., & Khan, D. R. Z. (2015). Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study). Computer Science, Communication and Instrumentation Devices.
[19] Caruana, R., & Niculescu-Mizil, A. (2005). An empirical comparison of supervised learning algorithms using different performance metrics. Technical Report TR2005-1973, Cornell University, 2005. Available at http://www. cs. cornell. edu/ alexn.