An Insight into Supervised Machine Learning: Review of Classifiers and Performance Metrics for Classifiers

Salvi Jain[1], Deepika Singh Pantola[2]

1College of Computing Sciences and Information Technology, Teerthanker Mahaveer University, Moradabad

2College of Computing Sciences and Information Technology, Teerthanker Mahaveer University, Moradabad

[1]salvijain620@yahoo.com

[2] deep.16feb84@gmail.com

***Abstract*— Supervised machine learning involves the process of constructing a concise model (i.e. classifier) by analysing the training data. The resulting model is then used to correctly determine class labels for the new examples. In literature, there exist a number of classifiers proposed by researchers that can be used to categorize the object's classes whose class label is unknown such as Decision Trees, Genetic Algorithms, Neural Networks, Naive Bayes, k-nearest neighbour, Support Vector machines. This paper provides an extensive literature review for the different classification models and also discusses various metrics for evaluating the performance of the classification models. The selection of any model is based on the performance measures such as speed, predictive accuracy, Robustness, Scalability, Interpretability, Simplicity. So it becomes very essential to select model that is satisfying these criteria in order to get more accurate result.**

***Keywords*— Supervised machine learning, Classification, Classifiers, Classifier's approaches, Performance metrics.**

I. INTRODUCTION

Machine learning(ML) can be considered as a type of artificial intelligence as the algorithms involves in this help computers to act more intelligently by generalizing rather than only storing and retrieving data items. Classification in data mining is a machine learning task which deals with identifying the class**,** a particular instance belongs to. In other words, the classification's goal is to predict the target class for each case in the dataset. Classification is also known as decision problem in critical areas of research such as science, industry and commerce. It is broadly categorized into two categories: supervised machine learning or pattern recognition and unsupervised machine learning which is sometimes known as clustering. Here our main focus will be on supervised machine learning. In this, all data is labelled and the algorithms learn is used to predict the output from the input data.

Before any procedures like selecting tasks, learning material or advice, learner's current situation must be clarified. For this purpose, we need a classifier-a model predicting the class value from other explanatory attribute. A classifier is a supervised function or more specifically a machine learning tool where the target attributes are nominal. It is used after the learning process to classify new data by giving them the best target attribute. This is a process in which prediction is being used. Classifiers can be designed manually, but at present, it is more advisable to learn from the real data. The idea is as follow: First, a classification method is to be selected. Second, we require a sample of data, where all class values are known. The sample data is divided into two parts: a training set and a test set. The training set is provided to a learning algorithm, through which a classifier is derived. Then that classifier is tested with the test set, where all the class values are hidden. If most of the cases are classified correctly in the test set, we assume that it works accurately on the future data. On the other hand, if there are too many errors, it can be assumed that it was a wrong model. In that case, a better model can be built after modifying the data, altering some conditions in learning algorithm or by using any other classification method.

II. LITERATURE SURVEY

The Literature Survey includes various approaches to learn Classifiers. In other words, this includes various Classification techniques.

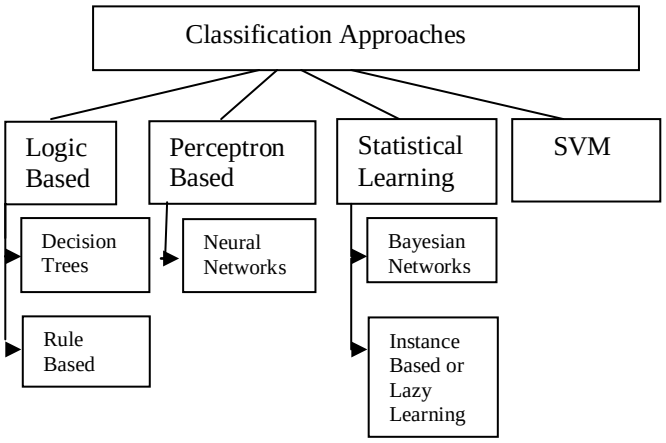

Fig. 1 State of art for classification Techniques

A. LOGIC BASED ALGORITHMS

This section describes the two groups of classifiers for logic based learning algorithms which includes Decision Trees based classifiers and Rule-Based Classifiers.

1) *DECISION TREES*: It is a simple and widely used classification technique proposed in 1998, as in [1]. Decision tree classifiers have some series of questions regarding the attributes of the test data. Each time, an answer is received, a follow-up question is asked until a conclusion about the class label of the test data is reached.

Building an optimal decision tree is a very big challenge in decision tree classifier because of the exponential size of the search space. In literature, there are a number of decision tree based algorithms such as C4.5, as in [2], Rainforest, as in [3].These algorithms employs a greedy strategy by making a series of locally optimum decisions about which attribute to be used for partitioning the data.

2) RULE BASED CLASSIFIERS: We can induce rules from decision trees by creating a separate rule for each path from the root to a leaf in a tree. Also training data can directly produce some rules. This can be done with the help of some rule-based algorithms. The goal is to construct the smallest rule-set, consistent with the training data. Rule-based classifiers makes use of IF-THEN-ELSE condition for classification.

Two techniques are usually followed in this:

2.1 Inductive Logic Programming: Inductive Logic Programming(ILP) is a discipline which is concerned with the investigation of inductive construction of first-order clausal theories from information already existing. Its methodology consists of various steps: First, a "model-theory" is formed. In this, problem specifications are organized in a semantic way. Second, a generic ILP algorithm is produced. Third, all inference rules and operators used in that are produced, resulting in a "proof-theory". Fourth, justification with the help of either probabilistic support or logical constraints on the hypothesis language. Computational learning theory and issues of predicate invention are a topic of concern in ILP. ILP applications formed under two categories: first scientific discovery and knowledge acquisition and other programming assistants.

2.2 Genetic Algorithms: These are the methods for solving constrained as well as unconstrained optimization problem based on natural selection. At each step, they select individuals at random from the current population to be parents and produces children for their next generation. Over successive generations, the population evolves towards an Optimal solution. They uses three types of rules to create next generation from the current population:

- Selection rules select the individuals, called parents, contributing to the population at the next generation
- Crossover rules combine two parents to form children at the next level

- Mutation rules apply random changes to individual parents to form children

B. PERCEPTRON BASED TECHNIQUES

A perceptron can be defined as a single layer neural network, that has a "step function" as its activation function. Mathematically, If x^1 through x^n are input feature values and w^1 through w^n are connection weights/prediction vector, then perceptron computes the sum of weighted inputs: $\sum_I x^i w^i$ and output goes through an adjustable threshold: if the sum is above threshold, output is 1; else it is 0. The most common way the perceptron algorithm is used for learning from a batch of training instances is to run the algorithm repeatedly through the training set until it finds a prediction vector correct on all of the training set. This prediction rule is then used for predicting the labels on the test set.

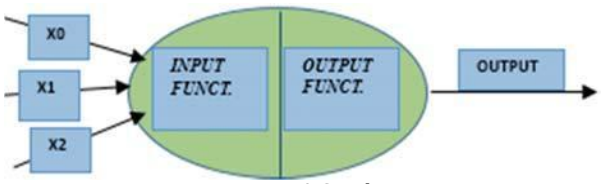

Fig. 2 A Simple Perceptron

NEURAL NETWORKS: This work is proposed, as given in [4]. Artificial Neural Networks(ANN) are a family of models inspired by biological Neural Network and are used to estimate or approximate functions depending on a large number of inputs, generally unknown. These are systems of interconnected neurons which exchange messages between each other. The connections have numeric weights that can be tuned based on experience.

Multilayer perceptron is one of the most important models of artificial neural networks. It is a feed forward supervised type neural network. The multilayer perceptron has a hidden layer and can deliver outputs with more than two classes. The hidden layers should be designed in such a way that it contains sufficient neurons to understand input features and generate three different classes of output. Hidden layers should be as minimum as possible for better working and output. This concept is defined, as in [5]-[6].

Back propagation is used in Multilayer perceptron to optimize the weights. The neural network processes a group of known pairs of input-output. Input whose correct output is known is feed to the network. Using layers in it, input is processed and output is generated. This output is compared to the already known output and then error is calculated. Every neuron in the network has contributed to this error.

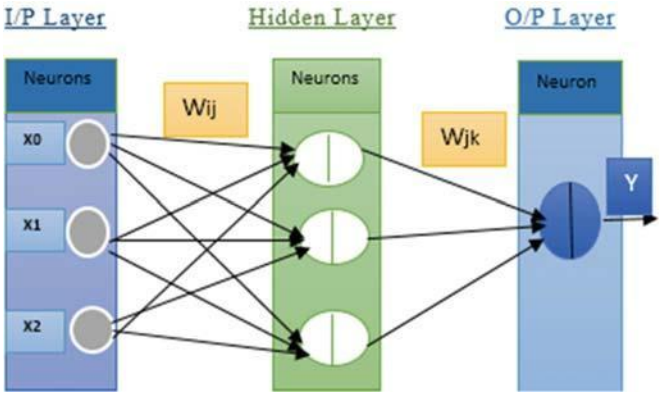

Fig. 3 A Simple Feed Forward Neural Network

C. STATISTICAL LEARNING ALGORITHMS

Statistical learning approaches provides a probability that an instance belongs in each class, rather than simply a classification. These are characterized into two categories: Bayesian Networks and instance-based methods.

1) BAYESIAN NETWORKS: The Naïve Bayes Classifier technique is based on Bayesian Theorem. It is generally used when the dimensionality of the inputs is high. It is capable of calculating the most possible output based on the input. We can add new raw data at runtime and have a better probabilistic classifier. Bayes classifier assumes that the presence(or absence) of a particular attribute of a class is unrelated to the presence(or absence) of any

other feature when the class variable is given. A comprehensive book is given by [7].

2) INSTANCE-BASED LEARNING OR LAZY ALGORITHMS: Instance based learning algorithms are lazy learning algorithms [8], as they delay the generalization or induction process until classification is performed. One of the most important algorithm of this category is k-Nearest Neighbor algorithm.

k- Nearest Neighbor is based on the principle that if any two instances have similar properties within a dataset, then one instance will be in close proximity with the other. If the instances are tagged with a classification label, then the value of unclassified label instance can be determined by observing class of its nearest neighbor. It locates the k nearest instances to the query instance and determine its class by identifying the single class label which is occurring more frequently. For more accurate result, several algorithms uses weighted schemes also [9].

D. SUPPORT VECTOR MACHINES

Support Vector Machines can be used for classification, regression and outlier's detection. In general, real world problems or domains involve non-separable data in which it is very difficult to separate positive data from negative instances in the training set because no hyper plane exists. One solution to this problem is that mapping can be done onto a higher dimensional space and define a separating hyper plane there. The higher-dimensional space is called the transformed feature space, as opposed to the input space occupied by the training instances [10].

It is very important to choose an appropriate kernel function, since they defines the transformed feature space in which training set instances will be classified.

S.N	Classification Techniques	Advantages	Disadvantages
1.	Decision Trees	Easy to interpret, explain; non-parametric	No Online Learning; easily overfit
2.	Rule-Based Classification	Easy to interpret, explain and generate; Can classify new instances; Can handle missing values and numeric attributes	Infinite chaining; Modification of knowledge base difficult; High Computational Cost; Complex Domains
3.	Artificial Neural Networks	Few parameters need to be adjust; No need of linearly separable data; Reprogramming not needed	Speed and size requirement is more; High Processing time; Difficult to predict neurons and layers
4.	Naives Bayes classifier	Easy to use; Needs less training data; Fast implementation	Can't learn interactions between features
5.	k- Nearest Neighbour	Robust to noisy training data; Effective for large training data	High computational cost; Need to determine parameter k
6.	Support Vector Machines	Good in high-dimensional space; Memory efficient; Versatile	Speed and size issue; Support vector approach issue

TABLE 1 COMPARATIVE ANALYSIS OF CLASSIFICATION ALGORITHMS

III. PERFORMANCE METRICS

A Receiver Operating Characteristics(ROC) graph is a technique for visualizing, selecting and organizing classifiers based on their performance. Reference [11] demonstrated the value of ROC curves in evaluating and comparing algorithms. Recent years have seen an increase in the use of ROC graphs in the machine learning community, as it was realized that simple classification accuracy is often a poor metric for performance measure [12]-[14]. They worked well with domains including skewed class distributions and unequal classification error cost.

ROC Curve is a two-dimensional depiction of classifier performance. It is a plot of True positive Rate(TPR) against False positive Rate(FPR) depicting relative trade-offs between benefits(true positives) and costs(false positives). For comparing Classifiers and depicting performance, it is very essential to reduce ROC performance to a single scalar value. A common method is to calculate the area under the ROC Curve, abbreviated AUC [15]. Since the AUC is a portion of the area of the unit square, its value will always between 0 and 1.0. However, because random guessing produces the diagonal line between (0,0) and (1,1), which has an area of 0.5, no realistic classifier should have an AUC less than 0.5. The AUC has an important statistical property: AUC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance equivalent to Wilcoxon test of ranks [16]. The UC Curve is also equivalent to the Gini Coefficient [17], which is twice the area between the diagonal and the ROC Curve. Reference [18] point out that Gini+1=2*AUC.

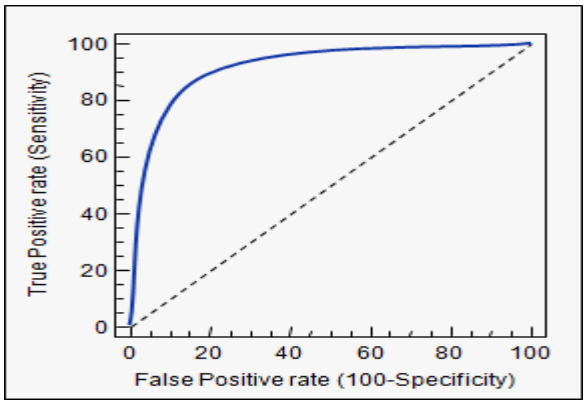

Fig. 4 ROC graph

Another important performance metric is a Confusion matrix. A confusion matrix contains information about actual and predicted classifications done by a classification system. The performance of such systems is evaluated with the help of data provided in its matrix.

IV. CONCLUSIONS

Supervised Machine learning are being applied in different domains. This paper describes some of the approaches to that. Each approach has got some features as well as some disadvantages. The goal of classification result integration algorithm is to generate more certain, precise and accurate result. Comparing superiority of one algorithm with other is not the only concern. Apart from this, it should also be considered that under what conditions and application problem, a particular method can significantly outperform other. It is impossible for a single classifier to attain best possible accuracy, however ensembling of classifiers can minimize this problem to a very large extent. As the data is increasing day-by-day, we believe that this big data will bring lot of opportunities for machine learning algorithms.

ACKNOWLEDGMENT

Firstly, I would like to express my sincere gratitude to our department, College of Computing Sciences and Information Technology, Teerthanker Mahaveer University, Moradabad and all the senior faculties who are involved in making this conference possible. Then, I would like to thank Ms. Deepika Singh Pantola, Assistant Professor, CCSIT, TMU, whose comments greatly improved the research work. Finally, my colleagues for their support and motivation.

REFERENCES

[1] Murthy SK (1998) Automatic construction of decision trees from data: a multi-disciplinary survey. Data Min Knowl Disc 2:345–389

[2] Quinlan JR (1993) C4.5: Programs for machine learning. Morgan Kaufmann, San Francisco

[3] Gehrke J, Ramakrishnan R, Ganti V (2000) RainForest—a framework for fast decision tree construction of large datasets. Data Min Knowl Disc 4(2–3):127–162

[4] Zhang (2000) provided an overview of existing work in Artificial Neural Networks (ANNs).

[5] Camargo LS, Yoneyama T (2001) Specification of training sets and the number of hidden neurons for multilayer perceptrons. Neural Comput 13:2673–2680

[6] Kon M, Plaskota L (2000) Information complexity of neural networks. Neural Netw 13:365–375

[7] Jensen F (1996) An introduction to Bayesian networks. Springer

[8] Mitchell T (1997) Machine learning. McGraw Hill

[9] Wettschereck D, Aha DW, Mohri T (1997) A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. Artif Intell Rev 10:1–37

[10] S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques", Informatica, Vol. 31, No. 3, pp. 249-268, 2007.

[11] Spackman, K.A., 1989. Signal detection theory: Valuable tools for evaluating inductive learning. In: Proc. Sixth Internat. Workshop on Machine Learning. Morgan Kaufman, San Mateo, CA, pp. 160–163.

[12] Provost, F., Fawcett, T., 1997. Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. In: Proc. Third Internat. Conf. on Knowledge Discovery and Data Mining (KDD-97). AAAI Press, Menlo Park, CA, pp. 43–48.

[13] Provost, F., Fawcett, T., 1998. Robust classification systems for imprecise environments. In: Proc. AAAI-98. AAAI Press, Menlo Park, CA, pp. 706–713.

[14] Provost, F., Fawcett, T., Kohavi, R., 1998. The case against accuracy estimation for comparing induction algorithms. In: Shavlik, J. (Ed.), Proc. ICML-98. Morgan Kaufmann, San Francisco, CA, pp. 445–453. Available from: <http://www.purl.org/NET/tfawcett/papers/ICML98final.ps.gz>.

[15] Bradley, A.P., 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recogn. 30 (7), 1145–1159.

[16] Hanley, J.A., McNeil, B.J., 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143, 29– 36.

[17] Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and Regression Trees. Wadsworth International Group, Belmont, CA.

[18] Hand, D.J., Till, R.J., 2001. A simple generalization of the area under the ROC curve to multiple class classification problems. Mach. Learning 45 (2), 171–186.