A Review on Machine Learning Classification Techniques

Akash Sirohi¹, *Dr. Arun Malik², Dr. Isha³

¹Research Scholar, Lovely Professional University
², ³Associate Professor, Lovely Professional University

ABSTRACT:

Now a days everyone shift toward technology and want to do all of their work with the help of technology, which helps to save their time and make work easy. Machine Learning plays an important role in this, which solve the complex problem easily and make easy for humans because the Machine learning algorithm learns once and then operate automatically. Machine learning have many approaches to solve the problem. In this paper, some techniques or algorithms of Classification are discussed.

Keywords: Machine Learning, Algorithms, Techniques, Classification.

INTRODUCTION

Now a days, machine learning used mainly in several fields like business, medical, engineering and so on. Learn the machine with several data samples to automate the particular task by the machine itself is called machine learning. Machine learning is primarily used to predict and to recommend. Machine learning has three types of learning i.e. supervised learning, in which input is divided into train and test data and it has train class based on which it classifies the new instances[11], it is used for classification and regression, another one is Unsupervised learning, in which there is no train output but it uses previous knowledge to identify the data[11], it is also used for clustering, and the last one is Reinforcement learning, it classifies the data such that output comes positive[11], it uses the back propagation technique to learn. There are some techniques of machine learning which are used for classification and come under supervised learning. Figure 1 displays the number of papers published in the field of machine learning from 2010 to 2019.

Decision Tree

A decision tree is a supervised learning technique that is used for classification. It classifies the new instance on the basis of the existing class. It is a data flow diagram type tree structure in which rectangles denote the internal node and ovals denoted the leaf node i.e. class [10]. This technique is used widely like in business, manufacturing, computational biology and so on, because it is easy to understand and easy to apply rather than other classifiers. It has three main algorithms i.e. ID3, CART, and C4.5. It classifies the sample with the help of gain information for ID3, Gini index for CART, and Gain ratio for C4.5. Formulae of decision tree and decision
tree is shown in figure 2 and figure 3 respectively. Figure 4 represents the number of paper published on decision tree from 2010-2019.

(i) Information gain:

\[
\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{v \in \text{Values}(A)} \left( \frac{|S_v|}{|S|} \right) \text{Entropy}(S_v)
\]

(ii) Gain Ratio:

\[
\text{Gain Ratio}(S, A) = \frac{\text{Gain}(S, A)}{\text{Split Information}(S, A)}
\]

\[
\text{Split Information}(S, A) = \sum_{i=1}^{c} \left( \frac{|S_i|}{|S|} \right) \log_2 \left( \frac{|S_i|}{|S|} \right)
\]

(iii) Gini value:

\[
\text{Gini}(D) = 1 - \sum_{j=1}^{n} p_j^2
\]

where \( p_j \) is relative frequency of class j in D.

Fig 2. Formulae for Decision Tree [2]

Fig 3. Decision Tree [11]

Fig 4. Paper Published on Decision Tree
Random Forest

As the name says Forest i.e. collection of trees, so Random forest is the combination of multiple decision trees [3]. It is also used for classification and comes under supervise learning techniques. It has higher accuracy than the decision tree because it combines multiple decision tree results to increase accuracy. Figure 5 represents the random forest structure.

![Random Forest Diagram](image)

**Figure 5.** Random Forest [12]

Support vector Machine

SVM comes under supervising learning. SVM can be used for issues like classification and regression, and it is based on statistical theory. The SVM can classify linear and non-linear data. It considers boundary data rather than all dataset [7]. It places a margin between the classes in such a way that margin is at maximum distance from classes which minimizing the classification error [11]. Working of SVM and paper published on SVM from 2010-2019 is demonstrated in figure 7 and figure 8 respectively.

![Number of Paper Published](image)

**Figure 6.** Paper published on Random forest

**Number of Paper Published**

- **X-axis:** Years (2010-2019)
- **Y-axis:** Number of Paper Published
LITERATURE REVIEW

In Ibomoije Domor Mienye et al. [1] authors studied the performance of the decision tree prediction and also attempted to find ways to improve the performance of the decision tree, decision tree has ID3, CART and C4.5 methods in which ID3 and C4.5 are powerful among them, and made a comparison between different tree-related algorithms. They concluded that the improved algorithms are better and that users accept them.

S.B.Kotsiantis [2] tells that for the classification models Decision tree techniques are mostly used because of it easy to understand. He tries to describe the various issues present in the decision tree and research gap. he concludes that for simplify the decision tree, there are some steps i.e. size controlling by editing, updates the space of tree, updates the algorithms, limits the search algorithm by deleting cases or several features and, change the structure. he also gives some disadvantages i.e. need more storage, lots of computations and decreased comprehensibility.

Hong Zhao and Xiangju Li [3], proposed a cost-sensitive decision tree algorithm for solving the minimal cost classification problem, they called it the BDADT algorithm which is based on a weighted class distribution having batch deleting attribute mechanism. They use 20 UCI data set for their experiment. They conclude that their algorithm gives better results than the existing algorithm. they also conclude that for medium and large data sets there are many calculations that increase the computational time so in the future, the speed of techniques can be increased by decreases the calculation.

In Weiwei Lin et al.[4] authors said classification of the big data is the biggest problem and cannot be solved by some classification algorithms like SVM and logistic regression that's why they proposed a random forest algorithm with parallel computing. They use the data set from an insurance company settled in china, they used some methods i.e. F-measure and G-mean to check its performance. They conclude that their proposed algorithm performs better than any other classification algorithm in terms of performance and accuracy. They
also conclude that with the help of KNN, the learning process can be reduced and they also gave some future scope like apply deep learning with it make it more accurate.

**In Min Zhu et al.**[5] authors classified the medical imbalance data with help of random forest algorithms with class weight. They used UCI data for their testing purpose. They conclude that their proposed algorithm performs better than any other i.e. better classifier with high accuracy. They take class weight according to empirical error and they also conclude that their algorithm classifies the AUC, F1 and Recall better than any other.

**In Bin Dai et al.**[6] authors used the random forest algorithm for detection and diagnose breast cancer, because machine learning techniques mostly used for diagnosing the diseases. They used the multiple eigenvalues and results of multiple decision trees to improve the accuracy and they also combine the many weak classifiers to produce better classification results. They conclude that their proposed algorithm performs better than any other and in the future, they can diagnosis the other diseases also.

**In Jagath Sri Lal Senanayaka et al.**[7] authors tried to detect the bearing fault earlier with the help of a support vector machine, because of fault in bearing cause the failure in electric motor and generator. They trained their algorithm with the help of two features i.e. time and frequency domain. With the help of this algorithm, harmful causes can be rectified easily. They conclude that with the help of their algorithms the failure can be detected 2.5-3.5 days ahead.

**In Mohammad Ali Ghorbani et al.**[8] authors proposed a hybrid algorithm of SVM and FFA(Firefly Algorithm) i.e. SVM-FFA to predict the FC and PWP with the help of soil data set. FC is the soil field capacity and PWP is Permanent Wilting Point. They used 215 soil samples from different places. They also compare their proposed algorithm result with SVM and ANN models on the basis of root mean square error, correlation coefficient, and relative root mean square error.

**In Jing Yan et al.**[9] authors trying to differentiate the Adenocarcinoma cells from other cells without labouring means try to automate this function with the help of Support Vector Machine according to confocal Raman spectra. They use 242 lung Adenocarcinoma cells and 231 normal cells for testing. They also try to control cost sensitivity error by correlation analysis to achieve the best accuracy. They used polynomial function and Gaussian radial basis function with SVM to classify the samples. They conclude that their proposed algorithm gives 97.5% accuracy, 100% sensitivity and 95% of specificity and give 85% accuracy with less number of sample.

<table>
<thead>
<tr>
<th><strong>Author</strong></th>
<th><strong>Problem discussed</strong></th>
<th><strong>Journal</strong></th>
<th><strong>Technology used</strong></th>
<th><strong>Description</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Ibomoiye Domor Mienye et al.</td>
<td>A review of improved decision tree-based algorithm prediction performance</td>
<td>Elsevier</td>
<td>Decision tree-ID3, CART, C4.5</td>
<td>C4.5 improved version of ID3, C4.5 better than pervious and user accept it.</td>
</tr>
<tr>
<td>S.B.Kotsiantis</td>
<td>Decision trees: a recent overview</td>
<td>Springer</td>
<td>Decision tree</td>
<td>Various issue n decision tree, current research point</td>
</tr>
<tr>
<td>Hong Zhao and Xiangju Li</td>
<td>A cost sensitive decision tree algorithm based on weighted class distribution and the mechanism</td>
<td>Elsevier</td>
<td>Decision tree-C4.5</td>
<td>Proposed BDADT algorithm, conclude proposed algorithm perform better</td>
</tr>
<tr>
<td>Author(s)</td>
<td>Title</td>
<td>Journal</td>
<td>Methodology</td>
<td>Summary</td>
</tr>
<tr>
<td>-----------</td>
<td>-------</td>
<td>---------</td>
<td>-------------</td>
<td>---------</td>
</tr>
<tr>
<td>Weiwei Lin et al.</td>
<td>An ensemble for large data analysis of the random forest algorithm</td>
<td>IEEE</td>
<td>Random Forest</td>
<td>Use F-measure and G-mean to check accuracy, conclude proposed algorithm perform well.</td>
</tr>
<tr>
<td>Min Zhu et al.</td>
<td>Weights Class Random Forest Processing Algorithm Imbalanced Medical Data</td>
<td>IEEE</td>
<td>Random Forest</td>
<td>Classify the minority and majority classes of medical data by assign class weight instead of weight for every single data.</td>
</tr>
<tr>
<td>Bin Dai et al.</td>
<td>Using Random Forest Algorithm for Breast Cancer Diagnosis</td>
<td>IEEE</td>
<td>Random Forest</td>
<td>Predict the breast cancer, combine results of multiple decision tree to increase the accuracy.</td>
</tr>
<tr>
<td>Mohammad Ali Ghorbani et al.</td>
<td>Use of vector supporting firefly algorithm machines to predict the field capacity and permanent wilting point</td>
<td>Elsevier</td>
<td>Support Vector Machine</td>
<td>Predict FC and PWP with a proposed hybrid algorithm i.e. SVM-FFA. Compare this model with SVM and ANN and conclude that it was better.</td>
</tr>
<tr>
<td>Jing Yan et al.</td>
<td>Confocal Raman sensing focused on a vector help for the identification of pulmonary adenocarcinoma cells</td>
<td>IEEE</td>
<td>Support Vector Machine</td>
<td>Distinguish the Lung Adenocarcinoma cell from normal cell with SVM by using two different kernels i.e. polynomial function and Gaussian radial basis function.</td>
</tr>
</tbody>
</table>
CONCLUSION

Nowadays machine learning used widely to automate everything which can reduce the labor work and save time and perform the complex task easily. Machine learning also helps in the medical field by prediction or diagnosis of the diseases, so this paper review some techniques of machine learning which are mostly used i.e. Decision tree, Support Vector Machine, and Random Forest.

REFERENCES


