IJCSE International Journal of Computer Sciences and Engineering Open Access

Survey Paper Vol. 9, Issue.3, March 2021 E-ISSN: 2347-2693

Different Types of Multi Class Classification Algorithms: A Study

Johnsymol Joy[1*], Rakhi Krishnan[2], Ziyad Nazeer[3]

[1,2,3]Dept. of Computer Application, Saintgits College of Applied Sciences, M G University, Kerala, India

**Corresponding Author: johnsymoljoy07@gmail.com, Tel.: 8921240480*

DOI: https://doi.org/10.26438/ijcse/v9i3.4144 | Available online at: www.ijcseonline.org

Received: 20/Mar/2021, Accepted: 23/Mar/2021, Published: 31/Mar/2021

***Abstract*—** Classification is a crucial aspect of machine learning. Multi class Classification has an important role in the classification. It is an on-going research in machine leaning field. In this paper we will come to know about multi class classification algorithms. We will see different algorithms like Decision Tree, SVM, Random Forest, Naive Bayes etc. This is clear cut representation of the basic advantages and limitations regarding different types of classification algorithms and the various measures for implementing results. Nowadays, these types of algorithms are playing a substantial role among different program sequences so as to improve the quality of classification.

Keywords— Classification, Decision tree, Naive Bayes, SVM

I. INTRODUCTION

Machine learning which is in the field of computer science is evolved from pattern recognition and computer learning theory in Artificial Intelligence. Machine learning gives computers the ability to learn without explicitly programmed. It teaches the computer to give solutions automatically to a problem. Machine learning mainly deals with the learning and building of algorithms which can be learnt and made predictions for the data set. As we know for some simple jobs assigned to the computers, it's possible for us to program algorithms and give instructions to the computer on how to execute all the steps required to solve a problem. But it will be difficult and challenging for humans to program algorithms for more advanced jobs, so it will be more effective to help the machine to develop its own algorithms for advanced tasks. Machine learning is nowadays used in wide range of applications. Some machine learning applications include speech recognition, image recognition, traffic prediction, DNA sequence classification, financial analysis, online fraud detection etc.

II.CLASSIFICATION

In machine learning, classification refers to the prediction to the problem for some input variables. It specifies to which class the data element belongs to and it is best when the output is finite and discrete. For example, we can classify an email program as spam and legitimate. [1]Classification and prediction are the two forms of data analysis which is used to describe models and extract data class or to predict future data trends. In machine learning, the classification problem is encountered in various areas, such as medicine to identify a disease of patient, in industry etc.

Classification is classified into two:

- Binary Classification
- Multi Class Classification

Binary classification refers to classifying the given data sets into two class labels. Example of binary classification is email classification.

Popular Algorithms in binary classification are:

- Logistic regression
- K-Nearest Neighbors
- Decision Tree
- Naïve Bayes

Multi class Classification refers to classifying the given data sets into more than two class labels. Examples of multi class classification are face classification, plant species classification etc.

Popular Algorithms in multi class classification are:

- Decision tree
- Random Forest
- SVM
- Naive Bayes
- Logistic Regression

MULTI CLASS CLASSIFICATION ALGORITHM

Classification is a two-step procedure; first procedure is learning and the second procedure is prediction. In learning procedure, the model is developed using the given data. In prediction procedure, [2] model is used to predict the response for the given data set. Multi class classification is a crucial problem in machine learning. In multi class accuracy and performance depends on voting and prediction of new class data.

© 2021, IJCSE All Rights Reserved

III.DECISION TREE

Decision tree algorithm is supervised learning algorithm, unlike others it can be used for solving classification and regression problems. [6] Decision tree is used to create a training model which can be used to predict the target value or class by simply learning decision rules. Decision tree is built in a tree structure. In decision tree we start with the root and compare the root attributes with the record attributes. After the comparison several branches are divided and we jump to next node.

Example: Suppose we have a problem to predict if a person can a can buy a car. Then we can classify them based on some categories like age, car type, class etc.

Some Terminologies used in Decision Tree are root node, sub nodes, leaf nodes, branches etc. Root node represents the most significant feature. [10] A Decision Tree always starts with a root and it get further divided several roots. Each node is an attribute. The root is divided into several sub nodes and this process is called splitting. Splitting of sub nodes into several sub nodes, and each sub node is called decision node. Leaf nodes represent different classes.

Decision tree uses several algorithms to split a node into several sub nodes. The selection is based on the available target variable and then split into the resultant sub node. Sum algorithms include ID3, C4.5, CART, CHAID, MARS.

Advantages of Decision [7] Tree includes its simplicity, it is simple to understand and easy to perform, performs well with large datasets, easy to interrupt and visualize, can handle both numerical and categorical data. Disadvantage of Decision Tree includes its instability, it is unstable if small variations occur, complex at times, if some classes dominate it may create biased tree.

IV. SUPPORT VECTOR MACHINE (SVM)

SVM is a supervised learning algorithm. It uses data and recognize pattern for classification and regression. It is used vastly in data mining, pattern [8] recognition, analysis etc. SVM is mainly designed for binary classification but it can be used in multi class classification also by breaking down multi class problem into series of binary classification. This technique is known as Divide and Conquer.

SVM works well in classes which have clear separation, it is more effective in high dimensional spaces and also it is memory efficient. But SVM is not suitable for large data sets. It does not execute well when target classes are overlapping. There is no probabilistic description for classification. Support Vector works for both classification and regression data values but it is mainly used for classification.

Most widely used multi class methods in Support vector machine are as follows:

- One Vs One Machine
- One Vs Rest
- Directed Acyclic
- Graph SVM

Support vector machines are mostly accurate as they can make correct predictions by model. When it comes to precision it measures conditional probability of making true positive decision when the given decision is positive. It makes a sensitive approach of probability where the positive subject is actually classified as positive. SVM are more specific when comes to measuring a negative subject is classified as negative.

Basic advantage of SVM is that it is more convenient in classes which have separation. SVM's are memory efficient and works more efficiently in high dimensional spaces. When it comes to handling large data set these are not suitable and they do not perform any probabilistic explanation or classification and are not good at target class overlapping.

V. RANDOM FOREST CLASSIFIER

Random forest is a classification algorithm. This can be used for multiclass classification. The "forest" it builds, is an ensemble of decision trees, usually trained with the "bagging" method. [9] The general idea of the bagging method is that a grouping of learning models increases the overall result. One of the big advantages of random forest is that it can be used for both classification and regression problems. Let's look at random forest in classification, since classification is sometimes considered the building block of machine learning.

The random forest algorithm selects random samples from the dataset and it will create a decision tree for each selected sample. Then each decision tree output a prediction and the voting will be performed for every predicted result. At last, the algorithm will select the most **voted prediction.**

VI. NAIVE BAYES

Naive Bayes algorithm is an algorithm that are based on Bayes theorem and it is a collection of classification algorithms. [4] Naive Bayes have been used for text classification and text analysis for machine learning problem. It is easy to build and useful for large data sets.
Naive Bayes text classification helps us to determine the conditional probabilities of occurrence of any events. We can increase the accuracy of Naive Bayes by:

- Handle missing data
- Use other distribution
- Remove redundant features
- Use less data
- Segment the data
- Use a generative model

Naive Bayes is another feature that are based on supervised learning algorithm. It is generally used for classification tasks, and in some cases Naïve Bayes can be used for regression.

Advantages are handles continuous and discrete data. It works fasts and can be used to make predictions. It performs less training data compared to other models. It is highly scalable. Disadvantages of Naive Bayes are it assumes all the predictors are independent, sometimes happens in real life. The probability outputs are not to be taken too seriously. The algorithms face the zero-frequency problem then the probability will be unable to make a prediction.

Mostly Bayesian network classifiers are more popular for their supervised classification paradigm. A well-known classifier of this kind is the Naive Bayes classifier. This classifier is probabilistic based on Bayes theorem which considers Naive (Strong) independent assumption. [5] This classifier was introduced onto the text retrieval community. Till now this remains as the most popular method for text categorization and problem of judging documents which belong to one category or the other with word frequency as its feature. A notable advantage of Naive Bayes is that it only requires a small amount of training data to estimate the parameters necessary for classification. Mainly Naïve Bayes is a conditional probability model. Despite its simplicity and strong assumption, the Naïve Bayes has proven to work satisfactorily in any domain. The Naive Bayes technique provides practical learning algorithms, prior knowledge and observed data which can be combined. The basic idea is to find the probability of categories given in a text document by using the joint probability of words and categories. This is based on the assumption of word independence.

VII. LOGISTIC REGRESSION

Logistic regression is a type of machine learning classification algorithm which is used to analyse a dataset in which there are one or more independent variables that determine the outcome and categorical dependant variable. In this method the output is being transformed using the logistic sigmoid function to return a possible value whereas, linear regression uses continuous numerical method to determine the output. Basically, sigmoid functions are a type of mathematical function used to map the predicted values to possibilities. It maps any real value into another real value within a range of 0 and 1. A Logistic Regression has a prescribed limit which is within zero and one. Regressions cannot go beyond this limit so it forms an S shaped curve. In that context the S shaped curve can be indicated as a sigmoid function or a logistic function. The logistic regression forms are of three types which are as follows:

a) Binary Logistic regression (two possible outcomes).
b) Multi nominal Logistic regression (three or more categories without ordering).
c) Ordinal Logistics Regression (three or more categories with ordering).

Further, logistic regression models are more complex (known as sigmoid function or logistic function) instead of linear function. [3] This limits the cost function between zero and one. Logistic Regression is also used for solving classification problems. It is being mentioned as a significant machine learning algorithm because it has the ability to provide the possibility and classify new data using continuous and discrete dataset. These are used to classify observations using different types of data and can easily determine the most effective variables used for classifications. The basic idea is to use the probable threshold value either 0 or 1. A value above the threshold tends to be 1 and the value below the threshold tends to be 0.

Logistic Regression is easier to implement, interpret, and very efficient to train. In logistic regression there are no assumptions about distribution of classes in feature space. Logistic Regressions provide a measure of how to appropriate a predictor and the direction of its association. Multiple classes can be easily extended through logistic regression. Logistic regressions are mostly known for their rapid classification of unknown record. Regressions can interpret model coefficient as an indicator feature importance. These are not useful when the number of observations is lesser than number of features otherwise it may lead to over fitting. Logistic regression can only predict discrete functions and dependant variable bound to a discrete number set.

VIII. CONCLUSION

Multi-class classification is undoubtedly the most common machine learning task. Usually multiclass classification has wide range of applications. It is a machine learning classification task that consists of more than two classes. In this paper we reviewed five multi class classifiers- SVM, Decision Tree, Random Forest, Logistic Regression and Naive Bayes. Classification plays a crucial role in data mining and many Classification algorithms are used to solve difficult problems. Classification is a machine learning technique which is used for data instances to predict future data trends. This paper mainly focuses on different Multi Class Classification algorithms-SVM, Decision Tree, Random Forest, Logistic Regression and Naive Bayes.

REFERENCES

[1]Ms. Prajakta C. Chaudhari, Prof. Dr. S. S. Sane, “Review on Multilabel Classification Algorithms”, IJSRD - International Journal for Scientific Research & Development| **Vol. 3, Issue 11, 2016** | ISSN (online): 2321-0613

[2]. Ruchida Sonar, Dr. P.R. Deshmukh “Multiclass Classification: A Review on International Journal of Computer Science and Mobile Computing” , **Vol.3 Issue.4, pg. 65-69. April- 2014.**

[3]. G. Malik, M. Tarique, " Machine Learning Techniques For Multi-class Classification International Journal of Advancements in Research & Technology", **Volume 3, Issue 2, February-2014** 6ISSN 2278-7763

[4]. Radhika Kotecha, Vijay Ukani, Sanjay Garg."An empirical analysis of multiclass classification techniques in data mining", IEEE conference on analysis of multiclass classification techniques, 2375-1282.

[5]. Mahendra Sahare, Hitesh Gupta, "A Review of Multi-Class Classification for Imbalanced Data", International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970) **Volume-2 Number-3 Issue-5, 160, September-2012.**

[6]. Bhaskar N. Patel, Satish G. Prajapati and Dr. Kamaljit I. Lakhtaria, " Efficient Classification of Data Using Decision Tree" Bonfring International Journal of Data Mining, **Vol. 2, No. 1, March 2012**.

[7]. Mr. Brijain R Patel, Mr. Kushik K Rana , "A Survey on Decision Tree Algorithm", | Volume 2, Issue 1

[8] Eesha Goel , Er. Abhilasha, "Random Forest: A Review: International Journal of Advanced Research in Computer Science and Software Engineering" , **Volume 7, Issue 1, January 2017.**

[9] Dr.R. Saravana Kumar2, "A REVIEW OF MULTI-CLASS CLASSIFICATION ALGORITHMS", Ph.D. Thesis, Rochester Institute of Technology (RIT Scholar Work), (2017) March.

[10] L. Breiman, ―Random Forests,‖ Machine Learning, **vol. 45, no. 1, pp. 5–32, 2001**

AUTHORS PROFILE

Mrs.Johnsymol Joy, pursed master of Technology from School of Computer Sciences, M.G University,Kerala. Currently working as Lecturer in BCA Department ,Saintgits College of Applied Sciences, Mahatma Gandhi University, Kerala, India.

Ms.Rakhi Krishnan, Final Year BCA Student, BCA Department, Saintgits College of Applied Sciences, Mahatma Gandhi University, Kerala, India

Mr.Ziyad Nazeer, Final Year BCA Student, BCA Department ,Saintgits College of Applied Sciences, Mahatma Gandhi University, Kerala,India