High Dimension Multi Class Algorithms (HDMCA) For Classification And Prediction: An Analysis Of Different Algorithms, Performance Measures And Datasets.

V.Shobana, Dr. K.Nandhini

Abstract: The dissection of a particular disorder is a needy one in the modern environment of our living. Our modern living which is not of much physical activity prone to many disorders in our body. Healthcare is a paramount part where so many research designs and proposals are made. Most of the algorithms entangled in healthcare, lets in for a number of results each of which was finer in their own way. The line of work carried out in this research is of applying a dataset for most leading algorithms and the best one is chosen for the next stage. The work is administered through the big data tool R and the results are compared across different metrics. From the results the top two sustaining algorithms are chosen for the forthcoming part.

Keywords: multiclass classification, random forest, SVM, LDA, kNN, CART, big data, HDMCA

1 INTRODUCTION

Algorithms based on data mining are very much useful in the field of healthcare for prediction of disorders. Classification and prediction are the two forms of data analysis. Of the entire classification algorithms decision tree takes a predominant position in predicting the target variable. Decision tree induction or decision tree is a flowchart like tree structure used to predict and classify data more precisely. Many researchers have formulated so many hypothesis using decision trees and they all have shown good results. The dataset taken in this research is thyroid data. The thyroid disorder is one of the most common endocrine disorders which is common worldwide especially in women. Most of the women around the globe suffer from this particular disorder and if they are not treated on time, it will result in serious issues. For predicting thyroid disorders various researches has been done worldwide. The main objective of this research is to compare some of the well known decision tree algorithms such as LDA, CART, kNN, SVM and random forest. These algorithms are performed on the data set using big data open source tools. The performance metrics such as ACC, MAE, PRE, REC, FME and kappa statistic are used to measure the performance of the algorithms.. These algorithms are compared with the above mentioned performance metrics. Each algorithm performs in their own way and the results are compared for the best one.

II. LITERATURE REVIEW

Comparison of algorithms forms an important part in the field of machine learning research (Kibler and Langley ‘1998).

The thyroid dataset analysis started in the year 1984 by Breiman et.al[1984], followed by Cestnik et.al in 1987,Quinana in 1988 and 1989. In the year 1992, Wray Buntine et al. [1] compared the decision tree for several datasets such as breast cancer, hypothyroid, iris etc., the decision tree implementation applied in his work was proposed originally by David Harper, Chris Carter, and other students at the University of Sydney from 1984 to 1988. Bruntine then proposed new splitting rules for decision tree induction. In his article [2] he used CART and ID3 algorithms and compared their performance. J. Huang et al. [3] compared NaiveBayes, SVM and decision trees with AUC and accuracy as performance metrics. The performance metric AUC (Area under Curve of ROC) exhibits several desirable properties compared to other metrics. In 2008, Keles. A et al. [4] formulated an expert system which was based on Neuro-Fuzzy classification for thyroid disorders, with an accuracy of 95.33%. Yuwei Hao et.al [5] generated as MsaDtd (Decision Tree based on MS-Apriori) approach that follows association rule mining and turns out with an accuracy of 87.21%.Maysanjaya et al. [6] used Multilayer Perceptron method to identify the type of thyroid (normal, hypothyroid, hyperthyroid) using WEKA tool. The accuracy of the prediction was 96.74%. A. Tyagi et al. [7] proposed a analysis of thyroid dataset using kNN, SVM, ANN and Decision Trees. The result shows SVM as the best predictor with 99.63%. Many researchers have so many findings with thyroid data set in their own way. Our work is to focus on the decision tree algorithms such as LDA, SVM, CART, kNN and Random Forest. These algorithms are implemented using the open source framework R studio and the results are compared.The framework for comparison of the decision tree algorithms goes in this way. The proposed methodology is shown in Figure.1.

- *V.Shobana1, Dr. K.Nandhini*
- *Research Scholar, Department of Computer Science, Chikkanna Government Arts College, Tirupur, India. 1 Assistant Professor, Department of Computer Science, Dr. N. G. P. Arts and Science College, Coimbatore, India.*
- *Assistant Professor, Department of Computer Science, Chikkanna Government Arts College, Tirupur, India. email: shobana484@gmail.com, krishnandhini@yahoo.com*

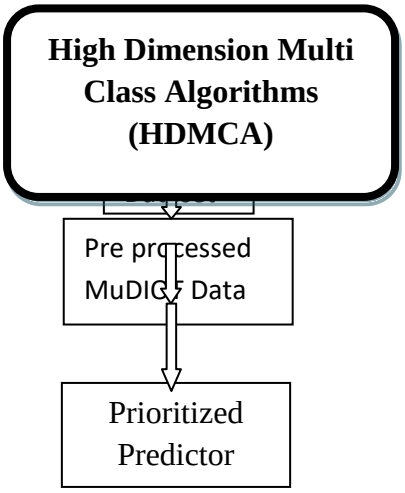

***Fig.1**. Proposed Methodology*

III. METHODOLOGY

In our previous work the thyroid data was pre-processed using MuDIOT (Multi Class Data Imbalance Oversampling Techniques). The thyroid data set suffers from class imbalance problems. It was overcome by random oversampling and under sampling techniques. The decision tree algorithms play a vital role in predicting the target value data and many researches focused on decision trees yields better performance. The methodology formulated implements the thyroid data set across the five algorithms and results are compared. Each algorithm performs in its own way yielding better performance. The MuDIOT data is taken for analysis across five algorithms which is performed in multiclass problems. The algorithms are classified as linear, nonlinear and advanced algorithms.

(i)Linear Algorithms

The linear algorithms taken in this case is LDA (Linear Discriminant Analysis) which finds the probability that each new set belongs to one class and makes predictions. The LDA method when applied to multiple classes it is known as Multiple Discriminant Analysis(MDA).In our dataset the target variable is Multiclass and MDA method is used instead of LDA. The class which gets the highest probability is the output class and the predictions are made. The probabilities are calculated using Bayes theorem. It estimates the probability of the output class (k) by using input (x) and using the probability of each class and the probability of the data belonging to each class

$$P(Y=x|X=x) = (PIk * fk(x)) / sum (PIl * fl(x)) \quad (1)$$

Where PIk refers to the base probability and PIk = nk/n.

(ii) Non linear Algorithms

The nonlinear algorithms discussed here are CART (Classification and Regression Trees) and kNN (k Nearest neighbors). CART is a predictive decision tree algorithm where the target variable is predicted using the features selected for prediction. CART is discovered by Breiman et al. [8] in the year 1984.It uses Gini as the impurity index. When the data is categorical, binary or nominal classification tree is built and when the data is continuous regression is taken into account. It handles both numeric and categorical data.

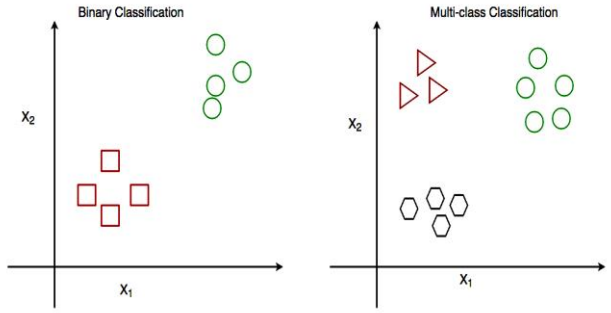

***Fig.3**. Binary and Multiclass Classification.*

Most of the machine learning algorithms concentrates on supervised learning. It is defined as having have input variables (x1) and an output variable (x2) and an algorithm to find the mapping function between input and output x2 = f(x1) . The goal of this function is to approximate the mapping function when new input data (x1) is available a prediction is made to find the output variable (x2) for that data. It is shown in Fig.3 for both binary and multiclass classification. Here in our work the data set taken is of multiclass. The model accuracy can be calculated using the following parameters and the repeated k fold cross validation is taken in our case. The model accuracy can be calculated using the following metrics.

- Data Split
- Bootstrap
- k-fold Cross Validation
- Repeated k-fold Cross Validation
- Leave One Out Cross Validation

kNN or k-nearest neighbours is the uncomplicated form of classification algorithm which does not depends on the type and structure of data. The distance between two samples be the Euclidean distance between their feature vectors. kNN can be used to solve both classification and Regression problems. If the k value is decreased, the prediction becomes less stable and inversely if k value is increased it becomes more stable. On the other hand the algorithm becomes weaker as there are more number of independent variables.

(iii) Advanced Algorithms.

The other two algorithms taken for classification and prediction are SVM and Random Forest. SVM (Support vector machine) a well organized and competent method when the feature vector is of high dimension. It uses 'one vs. all' approach for classifying the data and is similar to multiclass logistic regression. The algorithm works on this way; it generates a hyper plane and classifies the data into several classes. Random Forest on the other hand is a colossal of decision trees. The fact that random forest is so powerful because, "A tremendous number of comparably uncorrelated models (trees) functioning as a group will surpass any of the unique models". Random forests uses bagging and feature randomness to build distinct trees and creates an distinctive forest of trees, which when grouped gives more accurate than that of any unique trees. Archana Chaudhary et.al [9] proposed an improved random forest algorithm which gives better results across variety of datasets. In our work also among our five algorithms random forest out performs among the five yielding high

accuracy. The algorithm is measured across various parameters and the results are shown in Table.1.

IV RESULTS AND DISCUSSIONS

The pre processed data is applied on the discussed algorithms and the results are discussed below.

Model	*Accuracy*	*Kappa*
LDA	*92.88*	*36.78*
CART	*95.84*	*94.14*
kNN	*93.75*	*44.91*
SVM	*95.83*	*72.32*
RF	*99.30*	*1.00*

Table1. Accuracy and Kappa values of five algorithms

The five algorithms accuracy and kappa values are given in Table 1. By observing the values of the table its clear that CART and Random Forest turns to be good for both accuracy and Kappa. Furthermore it can be shown in the form of a figure with confidence level 0.95 as shown in Fig.4

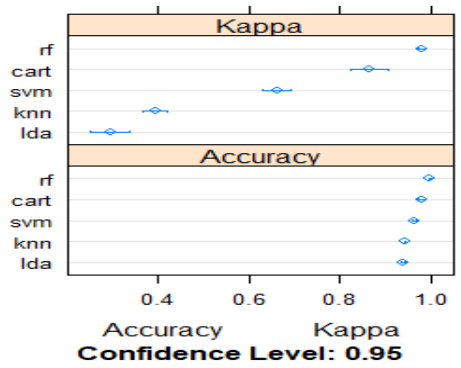

***Fig.4**. Accuracy and Kappa Graph.*

Also the model is built using all the five algorithms and the results are shown in the form of graph as in Fig.5. It is clear from the graph that RF and CART gives the highest accuracy compared to other algorithms. Here the graph is plotted with various models vs accuracy

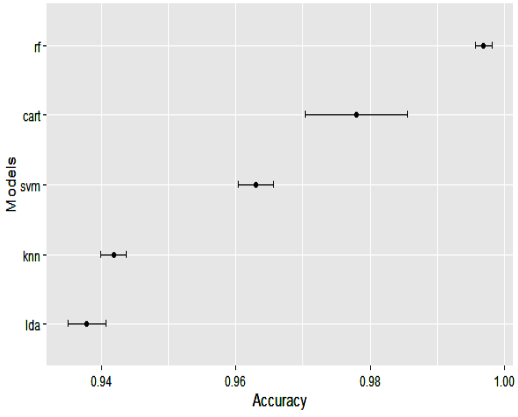

***Fig.5**. Accuracy of algorithms.*

The data preprocessed in our previous work is taken for model building and a resampling is done with repeated k fold cross validation and the results are given in Table.2

pre-processed data

Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 5184, 5184, 5184, 5184, 5184, 5183, ...

Resampling results across tuning parameters:

mtry	*Accuracy*	*Kappa*
2	*99.56*	*96.94*
6	*99.68*	*97.79*
10	*99.68*	*97.77*

***Table 2**.Results of repeated CV*

The results are stable and good for mtry value=6. Accuracy was used to select the optimal model using the largest value. The final value used for the model was mtry = 6.To get better accuracy from the model it is necessary to fine tune the different mtry parameters for different values. It is obvious from the above results that the dataset gives the better accuracy for random forest and CART.

V CONCLUSION

The work presented a comparitive analysis of our dataset across five algorithms. The results of the algorithms are compared against performance metrics such as accuracy, kappa statistic, sensitivity, specificity and F measure with 10 fold cross validation. The CART and random forest approach coins out good for our MuDIOT data with accuracy greater than 95%. The mtry with different values are applied to both the algorithms and the algorithm is fine tuned for better results. The work can be further improvised by tuning the algorithms with different parameters which might yield different results.

REFERENCES :

[1] Buntine, W. & Niblett, T. Machine Learning (1992) 8: 75. https://doi.org/10.1023/A:1022686419106.

[2] Buntine, W. Stat Comput (1992) 2: 63. https://doi.org/10.1007/BF01889584.

[3] J. Huang, J. Lu and C. X. Ling, "Comparing naive Bayes, decision trees, and SVM with AUC and accuracy," Third IEEE International Conference on Data Mining, Melbourne, FL, USA,2003,pp.553-556.doi: 10.1109/ICDM.2003.1250975.

[4] Keles, Ali & Keles, Ayturk. (2008). ESTDD: Expert system for thyroid diseases diagnosis. Expert Systems with Applications. 34. 242-246. 10.1016/j.eswa.2006.09.028.

[5] Hao Y., Zuo W., Shi Z., Yue L., Xue S., He F. (2018) Prognosis of Thyroid Disease Using MS-Apriori Improved Decision Tree. In: Liu W., Giunchiglia F., Yang B. (eds) Knowledge Science, Engineering and Management. KSEM 2018. Lecture Notes in Computer Science, vol 11061. Springer, Cham.

[6] May Sanjaya, I Made & Nugroho, Hanung Adi & Setiawan, Noor Akhmad. (2015). A Comparison of Classification Methods on Diagnosis of Thyroid Diseases. 10.1109/ISITIA.2015.7219959.

[7] A. Tyagi, R. Mehra and A. Saxena, Interactive Thyroid Disease Prediction System Using Machine Learning Technique, 2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC), Solan Himachal Pradesh, India, 2018, pp. 689-693. doi: 10.1109/PDGC.2018.8745910.

[8] Breiman, L., Friedman, J., Olshen, R., Stone, C.: Classification and Regression Trees. Wadsworth and Brooks, Monterey (1984).

[9]. Archana Chaudhary, Savita Kolhe, Raj Kamal, An improved random forest classifier for multi-class classification, Information Processing in Agriculture,Volume 3, Issue 4,2016.

[10] R. Katuwal and P. N. Suganthan, Enhancing Multi-Class Classification of Random Forest using Random Vector Functional Neural Network and Oblique Decision Surfaces, 2018 International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, 2018, pp. 1-8.

[11]. José-Ramón Cano, Pedro Antonio Gutiérrez, Bartosz Krawczyk, Michał Woźniak, Salvador García,Monotonic classification: An overview on algorithms, performance measures and data sets, Neurocomputing,Volume 341,2019,Pages 168-182,ISSN 0925-2312.

[12]. Baloochian, H. & Ghaffary, H.R. Multiclass Classification Based on Multi-criteria Decision-making, Journal of classification, April 2019, Volume 36, Issue 1, pp 140–151.

[13] Wang, Q., Nguyen, TT. Huang, J.Z. et al. An efficient random forests algorithm for high dimensional data classification, Advances in Data Analysis and Classification, December 2018, Volume 12, Issue 4, pp 953–972.