International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 4
DOI: https://doi.org/10.17762/ijritcc.v11i4.6447
***Article Received**: 07 February 2023 **Revised**: 03 March 2023 **Accepted**: 16 March 2023*

Randomized Ensemble Approach with ID3 Algorithm For the Prediction datasets with Imbalance Problem

Sasirekha R, Kanisha B
Department of Computing Technology, School of Computing
SRM Institute of Science and Technology
Chengalpattu, India.
sasirekharajeshkumar@gmail.com
kanishab@srmist.edu.in

Abstract— Nowadays, it is significant to make accurate prediction model by handling imbalance problem. When the larger dataset has been used in the prediction model, that data should be classified into classes which gives '0' and '1' to indicate negative and positive results. While classifying this target value, the larger number of instances can reside in one class and the remaining lower number of instances can be stored in another class. Because of this unequal distribution of data, the machine can be biased and there is high possibility to give wrong predictions. An inaccurate Dataset leads to misprediction. Hence, the imbalanced prediction dataset has been taken. This paper gives a proper information on Randomized ensemble approach with ID3 classifier for the imbalanced prediction dataset.

Keywords- Minority-class; Majority-class; Class Imbalance Problem; ensemble; ID3.

I. INTRODUCTION

Nowadays, machine learning techniques are used to give accurate predictions. It is advisable to keep dataset balanced one to make exact predictions. Understanding the concept of imbalanced dataset is quite tough to learn. When a dataset tend to classify into classes it can be classified in to majority and minority groups. The larger number of data stored in one group called majority group and the remaining lower number of data residing in a another group called minority group. Consider an example to understand about the majority and minority class, if the dataset consists 1000 instances while classifying those instances into positive and negative results certain number of results will resides on negative and only few can give positive. On this situation an imbalance problem occurs and which can lead a model to give wrong predictions. To overcome the unequal distribution of a dataset certain methodologies has been used. Nowadays, Machine learning techniques are used in the field of disease prediction to identify disease earlier to prevent themselves from severity. In that kind of larger datasets, the number of samples should be same in all the groups. if there is any inconsistency or misclassification occurs then it can be handled by the sampling methodologies. In the upcoming sections a clear study about the imbalanced problem has been explained with sample experiments.

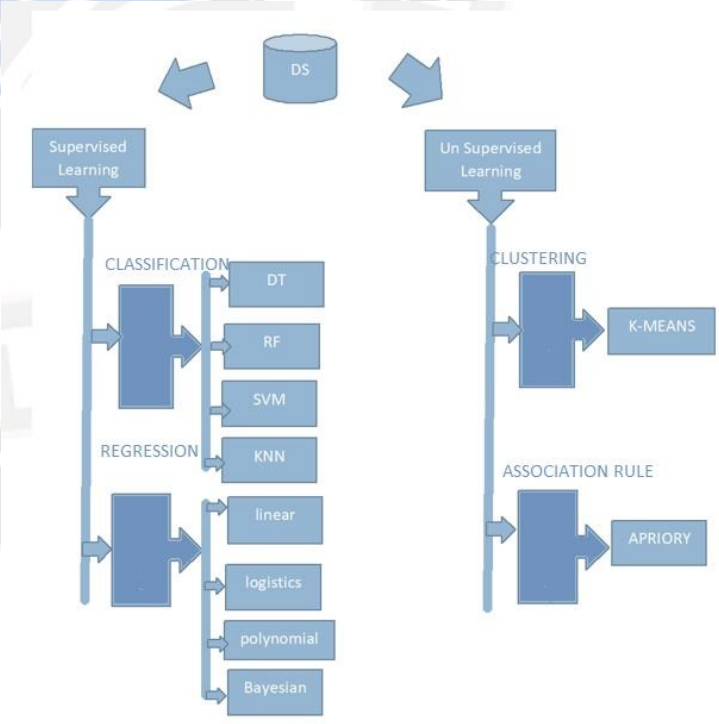


Figure. 1 Curriculum Learning

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 4
DOI: https://doi.org/10.17762/ijritcc.v11i4.6447
***Article Received**: 07 February 2023 **Revised**: 03 March 2023 **Accepted**: 16 March 2023*

II. RELATED WORKS

A. Curriculum learning

In curriculum learning, the environment in which application chosen to work has been considered. Basically, an environment belongs to supervised or unsupervised. Supervised learning[1] When a machine trained with a training dataset that is called supervised learning. Under supervised learning there are 2 familiar works learning methodologies are there. Namely classification and regression. Classification is considered as a widely using methodology to classify anything under a label. Using da labelled dataset a machine can be trained in the classification. In regression analysis an statistical datasets can be taken and analysed to make predictions. Which can give continuous variable as its predictor. Linear and non linear regression used to give the prediction based on the linearity of the taken data. Simple linear regression uses a single data as it's input to make predictions while the multi linear regression uses multiple input variables to make predictions.

Unsupervised learning

Unlabelled set of data can be used for the machines to make outcome of a model. Normally an unsupervised learning includes clustering and association rule based methodologies to bring out the outcomes of a taken application.

Clustering approach is used to group a data based on its characteristics. For example consider a basket which consists of fruits like apple banana orange etc. Using clustering algorithm the fruits can be identified using its characteristics like shape colour and grouped separately.

Association rule is widely used to find out the relationship among the variables, how strongly one variable is related to another. Example- supermarket dataset. How frequently a person can buy the couple of items. Like bread and butter.

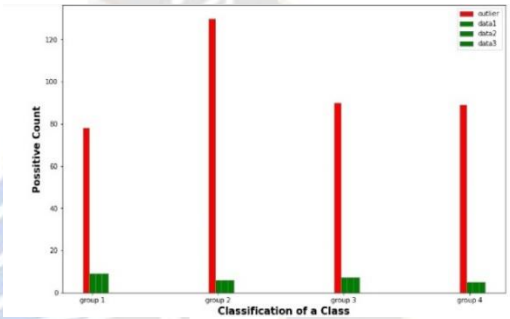
B. Study on Imbalanced problem

When a class tend to classify, it should be distributed equally to make prediction accurately. But when there is an occurrence of an unequal distribution of dataset is known as imbalanced data[2]. which can lead a model to give wrong predictions. Consider an example. A dataset with 10000 instances which is used to predict cancer It can be classified as 9900 negative results i.e., Non cancer. and the remaining 100 instances gives positive results, i.e, cancer. When the distribution of data tend to unequal that means number of positive cases is very low when compare to negative cases. In that case, machine can be biased and it can give wrong predictions. Because of majority of negative predictions, a cancer positive sample can be predicted as a non cancer sample. These kind of wrong predictions can make the severity for the people who is having cancer. Hence, it is necessary to handle these kind of imbalanced datasets to give accurate prediction.

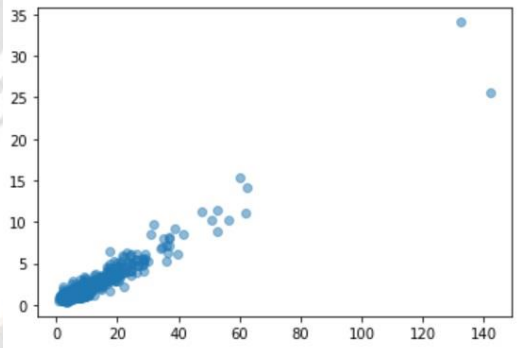
Other kinds of possibilities for the imbalance problem includes an outlier and mistakes recorded during observation.

C. Outlier analysis

An outlier [3] is a data point which is not included in a particular pattern. It's like an odd man out but it is not simple to remove. Because before trying to remove an outlier certain measures should be done. Variance and correlation of the data should be measured to make an decision whether an outlier is associated with any of the instances are not. It is not associated with any of the data then it can be removed. If it is relevant to the working application then it can be handled by considering outlier handling mechanisms.



(a)



(b)

Figure 2. (a) and (b) shows Structure of outlier in 2D view

D. Sampling methodologies

Resampling approaches are used to do the resample of data by analysing the dataset based on the need of the application. Resampling approaches are used to handle the imbalance problem. Unequal distribution of dataset problem can be handled by the resampling approaches. There are 2 widely used approaches are there, that are under sampling and the oversampling approaches.

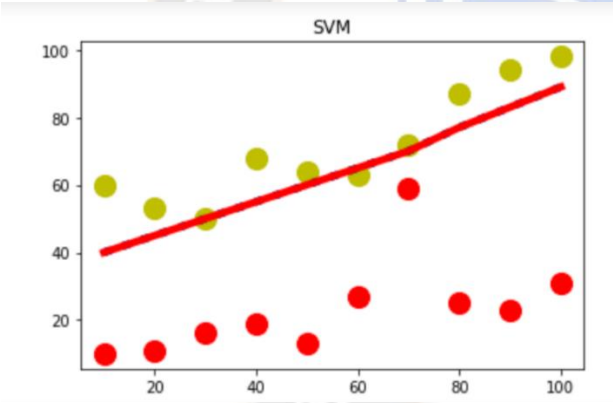
International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 4
DOI: https://doi.org/10.17762/ijritcc.v11i4.6447
***Article Received**: 07 February 2023 **Revised**: 03 March 2023 **Accepted**: 16 March 2023*

When the dataset tend to classify into 2unqual distribution of data then the larger number of data resides in a group called majority class and the remaining fewer data residing in another group called minority class. In under_sampling approach the data in the majority call can be reduced by removing the unwanted data from it. In oversampling approach the data in the minority class can be increased by adding the copy of existing data to it. Smote is considered as a widely used methodology to do oversampling in the minority class. In this paper a proper study on the sampling mechanism has be done properly.

Another methodology used to handle imbalance is ensemble approach. In ensemble approach a taken majority class can be divided into many equal parts to keep the data equal to the minority class. By splitting the data in several equal parts the data can be balanced to give exact predictions.

E. Classfier analysis

Classifier used to classify the dataset to implement the trained model. Certain model react good for training dataset and certain model can be good for test data. SVM [6] has its own way to split the data into 2 parts using margin commonly called as hyper plane.



While KNN can classify a data by means of identifying the neighbouring behaviour. Random forest is also a tree based approach which can be give results better than KNN.

TABLE I. STUDY ON ALGORITHMS

#Ref	observations	
	Algorithms proposed	***Pros& cons***
[6]	SVM	Predict New data by the margins. Positive and negative observations on taken.
[8]	KNN	A targeted node value can be calculated by considering the nearest nodes.
[9]	NAÏVE BAYES	Calculates the edge node.
[10]	RANDOM FOREST	Better than KNN for the prediction datasets but its not suitable for the larger datasets

III. METHODOLOGY PROPOSED AND RESULTS OBTAINED

A. Randomized ensemble to handle imbalanced problem in larger datasets

An Imbalanced problem is not a simple term to solve. Its like an bottle neck portion of the prediction model. It could not be handled blindly. It is necessary to keep dataset balanced to ensure the level of accuracy in the field of prediction mechanisms. First, the dataset can be taken from the std repository. Then the imbalance problem will be identified by means of unequal distribution of data.

Algorithm for randomized ensemble

Step 1: Distribution of data into xMax(majority clas) and xMin(minority class).

Step 2: Split xMax into yMaxTree. yMaxTree can be calculated as Total number of xmax/xMin.

Step 3: ‘Z’ is final Prediction from yMaxTree.

During the step 1, the distribution of data is takes place. The larger number of data resides in one group called majority class denoted by xMax. The remaining Lesser number of data residing in another group called minority class denoted by xMin.

In step 2, the data in the majority class can be divided into several equal parts as the number of data equal to the number of data in the minority class. For example is there 1000 instances in the dataset. Out of 1000 instances 900 going to give negative(majority class) and the remaining 100 instances going to give positive results. Then using Randomized ensemble[5] method the majority class (900 instances) can be divided in to 9 sets of 100 data(9*100) to make the number of instances equal to minority class to give best prediction.

$$yMaxTee = xMax/xMin \quad (1)$$

$$Z = Sum(yMaxTree) \quad (2)$$

B. Ensemble with ID3 Algorithm for prediction

The decision tree algorithm is used to make decisions. It consists of decision nodes which can give predictions based on the taken dataset. ID3 [7] algorithm is a decision tree based algorithm which can use entropy and the information gain to give exact prediction. After balancing a data ID3 algorithm used to give accurate prediction which can increase precision level.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 4
DOI: https://doi.org/10.17762/ijritcc.v11i4.6447
***Article Received**: 07 February 2023 **Revised**: 03 March 2023 **Accepted**: 16 March 2023*

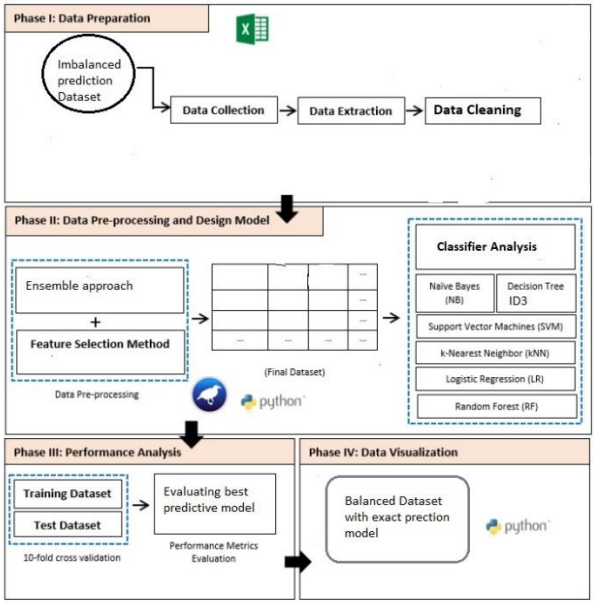


Figure 3. Framework of Proposed model

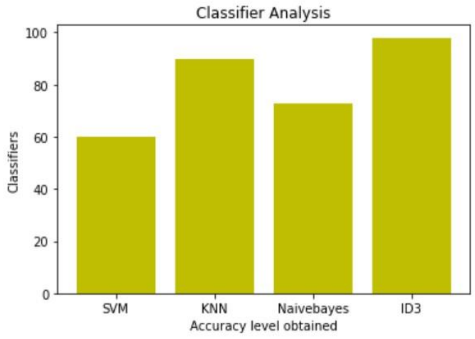


Figure 4. Accuracy level of classifier. Accuracy estimated Using weka tool.

IV. CONCLUSION

After learning about the important machine learning classifiers, ID3 algorithm has been chosen to give accurate prediction and the imbalance problem in the larger datasets can be handled using randomized ensemble approach. Finally, this paper concludes, that the randomized ensemble with Id3 algorithm gives more accurate prediction in a mean time when compare to traditional methodologies.

REFERENCES

[1] G Li Y-F, Guo L-Z, Zhou Z-H (2021) Towards safe weakly supervised learning. IEEE Trans Pattern Anal Mach Intell 43(1):334–346

[2] Sasirekha, R., Kanisha, B., Kaliraj, S. (2022). Study on Class Imbalance Problem with Modified KNN for Classification. In: Hemanth, D.J., Pelusi, D., Vuppalapati, C. (eds) Intelligent Data Communication Technologies and Internet of Things. Lecture Notes on Data Engineering and Communications Technologies, vol 101. Springer, Singapore. https://doi.org/10.1007/978-981-16-7610-9_15.

[3] S. R and K. B, "KNN Based Peak-LOF for Outlier Detection," *2022 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)*, Chennai, India, 2022, pp. 1-5, doi: 10.1109/ICSES55317.2022.9914212.

[4] Abraham B, Nair MS. Computer-aided detection of COVID-19 from X-ray images using multi-CNN and Bayesnet classifier. Biocybern Biomed Eng. 2020;40(4):1436-1445. doi:10.1016/j.bbe.2020.08.005.

[5] Ihya, Rachida & Namir, Abdelwahed & Filali, Sanaa & Aitdaoud, Mohammed & Guerss, Fatima zahra. (2019). J48 algorithms of machine learning for predicting user's the acceptance of an E-orientation systems. SCA '19: Proceedings of the 4th International Conference on Smart City Applications. 1-8Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[6] B. Cao, Y. Liu, C. Hou, J. Fan, B. Zheng and J. Yin,"Expediting the Accuracy-Improving Process of SVMs for Class Imbalance Learning," in IEEE Transactions on Knowledge and Data Engineering, vol. 33, no. 11, pp. 3550-3567, 1 Nov. 2021, doi: 10.1109/TKDE.2020.2974949

[7] B. A. Al-Hameli, A. A. Alsewari, M. Khubrani and M. Fakhreldin, "Accuracy and performance analysis for classification algorithms based on biomedical datasets," *2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM)*, Pekan, Malaysia, 2021, pp. 620-624, doi: 10.1109/ICSECS52883.2021.00119.

[8] R. Gomes da Silva, J. de Oliveira Liberato Magalhães, I. R. Rodrigues Silva, R. Fagundes, E. Lima and A. Maciel, "Rating Prediction of Google Play Store apps with application of data mining techniques," in *IEEE Latin America Transactions*, vol. 19, no. 01, pp. 26-32, January 2021, doi: 10.1109/TLA.2021.9423823.

[9] H. Goto, Y. Hasegawa, and M. Tanaka, "Efficient Scheduling Focusing on the Duality of MPL Representatives," Proc. IEEE Symp. Computational Intelligence in Scheduling (SCIS 07), IEEE Press, Dec. 2007, pp. 57-64, doi:10.1109/SCIS.2007.357670.

[10] V. K. Gupta, A. Gupta, D. Kumar and A. Sardana, "Prediction of COVID-19 confirmed, death, and cured cases in India using random forest model," in *Big Data Mining and Analytics*, vol. 4, no. 2, pp. 116-123, June 2021, doi: 10.26599/BDMA.2020.9020016.