JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 37, 1097-1108 (2021)
DOI: 10.6688/JISE.202109_37(5).0008

Mgini – Improved Decision Tree using Minority Class Sensitive Splitting Criterion for Imbalanced Data of Covid-19

PRATIKKUMAR A. BAROT+ AND HARIKRISHNA B. JETHVA
Department of Computer Engineering
Gujarat Technological University
Gujarat, 382424 India
+*E-mail: pratikabarot@gmail.com*

In the time of COVID-19, medical facilities struggling to fight against the pandemic. Most of the countries face a tough time fighting against this virus outbreak. Even developed countries are struggling to deal with this virus outbreak. Common problem countries face is a lack of medical staff and medical equipment. Machine learning has the potential to play an important role in a different area of medical facilities. With the help of the machine learning model, an effective diagnostic tool can be built which helps in the time of scarcity of medical staff. However medical data is imbalanced and this skew nature of data prevent machine learning algorithm from achieving high accuracy. To deal with this problem of imbalanced data, we proposed a modified decision tree algorithm that uses a minority sensitive Gini index called Mgini. In an imbalanced dataset of COVID-19, it is important to focus on the reduction of overall misclassification cost instead of trying improvement in accuracy value. Mgini is useful splitting criteria when the misclassification cost of the minority sample is huge as compared to the majority class. The use of this proposed new Gini index as a splitting criterion in the decision tree reduces the misclassification cost. Mgini based decision tree has higher accuracy and low misclassification cost as compare to the traditional Gini index based CART algorithm. Our proposed cost-sensitive approach improves imbalanced data classification without the use of data level sampling techniques.

Keywords: COVID-19, imbalanced data, CART, Mgini index, cost-sensitive learning, medical machine learning

1. INTRODUCTION

Imbalanced data have unequal class distribution. It has a majority class and minority class. Data difficulties like data overlapping and small disjunction make the imbalanced data learning more challenging [7, 9]. Improved imbalanced data learning increases the use of machine learning in domains like the medical and judicial systems. Recently, the outbreak of COVID-19 realizes the importance of imbalanced data learning.

COVID-19 is a respiratory illness commonly known as coronavirus. SARS-CoV-2-Severe acute respiratory syndrome coronavirus 2 is the strain of coronavirus disease [6]. As per the report from the World Health Organization (WHO) [6], more than 200 countries are affected due to the COVID-19 virus. The outbreak of COVID-19 was declared as a pandemic by the WHO.

COVID-19 exposes the scarcity of medical facilities even in developed countries. To date, the USA is a worst-hit country by the COVID-19 even though it has the world's best medical facilities. They struggled due to a lack of medical equipment which falls short in

Received August 19, 2020; revised September 11, 2020; accepted September 24, 2020.
Communicated by Maria José Sousa.

the rapid spread of coronavirus. Because of a lack of testing and diagnosis, the real extent of the spread remains unknown. Particularly in the case of COVID-19, this is a dangerous situation and it needs immediate attention.

Credit rating agency – Moody's predict India's growth rate at zero percent in 2020-21 due to COVID-19 [4]. COVID-19 adversary affects the gross domestic product (GDP) of the world. As per the united nation [11], the GDP of the world is reduced by one percent, and it may decrease further if the economy lockdown continues in the third quarter of 2020. As per David Muchlinski *et al.* [8], a reduction in GDP is one of the main causes of civil war onset.

Considering all the adverse aspects of COVID-19, it is very important to control it. To control the spread of COVID-19 it is required to diagnose it well in advance and with good prediction accuracy. Thus, effective machine learning-based medical diagnosis algorithm is in demand in current pandemic time.

Early detection of COVID-19 minimizes the risk of virus spreading. COVID-19 spreads very fast and has an exponential growth rate. Thus more and more testing and accurate early detection is the key to mitigating the outbreak of corona. When countries facing a scarcity of medical support systems and medical staff, a machine learning algorithm could be very helpful for medical diagnosis [13-15].

The outbreak of COVID-19 proves the importance of accurate machine learning tools in the medical domain. Many types of research are going on for accurate prediction of positive COVID-19 cases. However medical data are mostly imbalanced. It contains more patients diagnosed with a negative result as compared to the number of patients with a positive report. Even in the case of COVID-19, although it has a rapid and widespread impact, the numbers of positive cases are less as compared to the number of negative cases. For such imbalanced data traditional machine learning algorithms perform poorly and give more weightage to the accurate classification of majority class [9]. Due to this, it is important to consider the imbalanced nature of data while selecting a machine learning algorithm for prediction. The existing machine learning algorithm works well for balanced data [7]. In the case of imbalanced data, they are biased towards the majority class [7].

The CART algorithm is a widely used algorithm for balanced data. It uses the Gini index as a splitting criterion [5]. However, due to the use of the Gini index, the CART algorithm is biased towards the majority class. The Gini index is skew sensitive and it selects the splitting criteria which are biased towards the majority class [18]. The Gini index calculation assumes that the resultant partitions have equal class distribution and this results in the construction of a biased decision tree. For imbalanced data, the traditional biased decision tree misclassifies the minority class instances. To handle class imbalance we proposed a new minority sensitive Gini index for the decision tree algorithm called Mgini.

In the medical domain, there is a huge difference between the misclassification costs of classes. In the COVID-19, false classification as a negative case results in huge misclassification costs because this disease has an exponential growth rate. In a pandemic situation, false classification as a positive case has noticeable misclassification costs. However, it is much smaller than the earlier one.

We proposed a minority sensitive decision tree algorithm (MiDT) for imbalanced data classification. The MiDT algorithm uses the minority sensitive Gini index for the selection of splitting criteria. The effectiveness of our proposed algorithm is verified on COVID-19, breast cancer, and mammography datasets. The empirical result of MiDT shows improved accuracy as compared to the other traditional classification algorithms. In our empirical study with the Covid-19 dataset, the MiDT shows the least misclassification cost in comparison to the decision tree and SVM algorithm.

In this paper, Section 2 contains related work, Section 3 gives information about the dataset and dataset preprocessing. Section 4 explains the new minority sensitive Gini index called Mgini. Finally, Section 5 is about the experiment and result discussion. The major contribution of this paper is the minority sensitive Gini index for imbalanced data. Using Mgini we proposed a cost-sensitive model to reduce the misclassification cost of COVID-19.

2. RELATED WORK

Recently many researches are going on to predict COVID-19. These researches are either of image-driven or data-driven. In image-driven studies, most of the researches are based on radiographic image processing where they analyze the chest x-ray for the detection of COVID-19 [1-3]. Most of the image-driven researches used a neural network for the diagnosis of COVID-19 from the image dataset.

Many data-driven researches have been performed for COVID-19 [16, 17, 19, 20]. Most of these researches are to predict the COVID-19 spread trend, to predict the death rate and recovery rate, for prediction of the future requirement of the hospital and supporting staff [20]. However, not much study has been performed to propose a data-driven machine learning model for accurate diagnosis of COVID-19 from clinical data.

To propose a machine learning model for the medical dataset we need to handle data imbalance. However, traditional classification techniques do not perform well for the imbalanced data and specifically for the minority class. As per the study of Venkatesan *et al.*, J48 shows the best result as compared to CART, AD TREE, and BF TREE for breast cancer detection [16]. As per Bartosz *et al.* [13], skewed distribution of class is a challenging problem of machine learning. The authors use a boosting technique to propose a machine learning model of medical diagnosis.

Much researches have been performed to handle class imbalance [7, 9, 14, 18]. Most of the researches mainly focuses on the data sampling method [7, 9]. However, the sampling method suffers from a loss of information and over-fitting [9]. Due to this, the use of the sampling method should be avoided especially for medical data.

Under-sampling removes majority class instances which results in information loss and poor prediction accuracy [9]. Because of the huge misclassification cost associated with positive COVID-19 patients, it is required to ensure zero information loss. Oversampling increases the number of minority samples and thus the risk of biasing of the machine learning model towards minority samples also get increases [7]. In pandemic time, when administration struggling with a lack of hospitals, doctors, and medical support staff the false prediction of negative cases also puts extra load on administration. Because of these drawbacks associated with the sampling technique, the machine learning model which handles imbalanced data without sampling techniques is desirable [7].

Isra *et al.* [10], proposed a predictive data mining technique for MERS-CoV infections. They used naïve Bayesian and J48 for prediction. In their study, J48 gives a better result as compared to naïve Bayesian. Ikram *et al.* [18], proposed a new asymmetric entropy measure for decision tree called AECID (Asymmetric entropy for classifying imbalanced data).

As per the study of Fan Wu *et al.* [17], the COVID-19 virus is very much related to a group of SARS-like coronaviruses. Positive COVID-19 patients carry similar symptoms as other flu diseases like influenza. Thus a machine learning model that handles imbalanced data and diagnose COVID-19 with good accuracy needs to build.

3. DATASET

The symptoms of the COVID-19 are mostly the same as other flu caused by respiratory viruses [19]. So the separation of COVID-19 from other normal flu-like influenza is a difficult task. We used a clinical dataset from Kaggle for our proposed model. This dataset is collected from Hospital Israelita Albert Einstein, in São Paulo, Brazil. This hospital was inaugurated in 1971 and it is one of the best hospitals in Latin America. Dataset has 123 attributes and 5644 instances. The dataset contains data that is collected during a visit of patients to the hospital who comes for Covid-19 tests like SARS-CoV-2 RT-PCR and additional laboratory tests. In this dataset, 90% of instances belong to the majority class (negative case) and 10% instances belong to the minority class (positive case). Patient identity and other confidential information are in encoded form.

3.1 Data Pre-Processing

This dataset has a huge number of missing values. From total 5644 instances, only 607 sample contains some meaningful clinical information while the rest of the samples has no clinical data. Data distribution is very sparse. Only patient id, age, gender, and the patient admitted to a regular ward or ICU is available for all 5644 instances. However, this data is not meant for the diagnosis of positive cases. So we removed all samples which do not have any clinical importance.

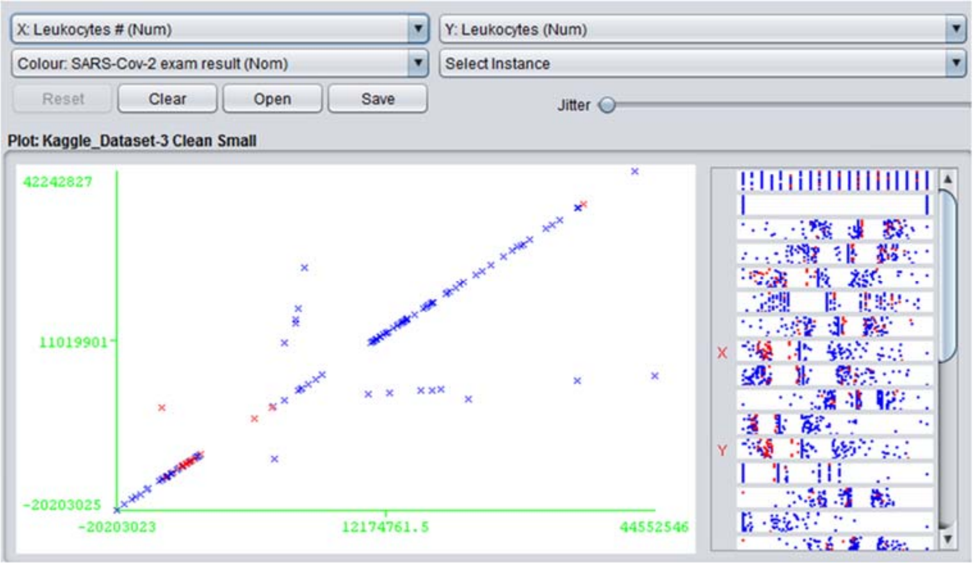
Even in the remaining 607 instances, most of the features out of 123 features has a missing value. For building effective diagnosis machine learning models we mainly focus on clinical data. Thus we removed the patient ID and his/her admission to the normal ward or ICU.

All urine related attributes have no useful clinical information and almost 99% value is missing. So we removed all urine related features from the dataset. Other attributes which have more than 60% of missing values are also removed. Finally, we select 30 attributes for our experimental study. Missing value and noisy data are replaced with the mean value of the attributes. To handle missing value and noise we use class-specific mean value. This ensures that clinical data retain their class-specific characteristics in it.

After data pre-processing, the dataset has 30 attributes and 607 instances. It has 519 negative cases and 88 positive cases. We make sure that all minority cases with valid clinical data remain in the final dataset.

3.2 Statistical Analysis of Dataset

The COVID-19 dataset is sparse. We clean the dataset, fill the missing value, and remove noise from it as mention in the previous data preprocessing section. Finally, the dataset has 30 attributes. To ensure that data is properly cleaned and to avoid the problem of over-fitting the attributes which are strongly related should be removed. We used the WEKA tool to study about data characteristics.



X: Leukocytes # (Num)

Y: Leukocytes (Num)

Colour: SARS-Cov-2 exam result (Nom)

Select Instance

Plot: Kaggle_Dataset-3 Clean Small

Fig. 1. Relationship of Leukocytes # and Leukocytes.

Fig. 1 shows that Leukocytes # and Leukocytes are strongly and positively related so we removed one of the features from those two features.

Similarly, Basophils and Basophils# are also weakly and positively related. So we removed one of the features from those two features as well.

4. MINORITY-SENSITIVE GINI INDEX – MGINI

Gini index is used to measure the impurity of the data. Gini index is considered a good measure for symmetric class distribution [18]. However, for asymmetric class distribution, we need a splitting criterion that considers class imbalance during the selection of splitting criteria. Gini index itself is an asymmetric measure. In the case of imbalanced data, it favors the majority class. We proposed a cost-sensitive Gini index as a splitting criterion called a minority sensitive Gini index (Mgini).

Especially for accurate prediction of the COVID-19 the splitting criterion should be able to give more importance to the accurate prediction of minority class (positive case).

$$Gini = 1 - \sum_{j} p_j^2 \tag{1}$$

Eq. (1) is used to calculate the Gini value. P_j is the probability of class j. CART algorithm selects the feature as a splitting criterion which has the lowest Gini value.

Mgini is an adapted measure. It is minority sensitive for the imbalanced data when the misclassification cost of the minority class is more than the misclassification cost of the majority class. For balanced data, it reflects the performance of the Gini index.

$$Mgini = 1 - \sum_{i=1}^{c} (P_i \div \delta_j * P_i \div \delta_j),\ j = i - 1 \tag{2}$$

P_i: Probability of i^{th} class. δ_i: Misclassification factor of i^{th} class. The misclassification factor is derived from misclassification costs. [Mc_1, Mc_2] is a misclassification cost vector for the binary class dataset.

$$\delta_i = \log(Mc_j),\ Mc_j > 1,\ j = i - 1 \tag{3}$$

For balanced data, if we consider equal misclassification cost for both the classes then the Mgini will work the same as the Gini index. If misclassification cost is not available then for binary class imbalanced data, Mc_j for the i^{th} class is derived from the class distribution weight Wc_i as shown in Eqs. (4) and (5):

$$Wc_i = |c_j|/|c_i|,\ j \neq i,\ j = 1 - i. \tag{4}$$

Where C_i and C_j are the cost of i^{th} and j^{th} class respectively.

$$Mc_i = Wc_i * 10^k \tag{5}$$

Where k is the smallest integer value such that min (Wc_i) > 1.

By using the minority sensitive Gini index we build a decision tree called the MiDT tree.

5. PROPOSED MODEL AND EXPERIMENTAL ANALYSIS

We have used Python 3.7 on windows 10 with 4GB RAM for the implementation of our model. Algorithm 1 shows steps for the construction of the MiDT tree and algorithm-2 shows detailed steps for the computation of the Mgini. The process of MiDT is summarized in Fig. 2.

```
Algorithm 1: MiDT
Input: Dataset (D)
Output: MGTree
Begin
D = CleanData(D)
For each features value
begin
  If MGini > ComputeMGini(D)
  begin
    Mgini = ComputeMGini(D)
    FV = Splitting Criteria  --FV is splitting criteria
  end
end
STree = SplitData(D, FV) --STree is sub-trees
MGTree.append(STree)
For each DS in STree    --Ds is data subset of subtree
```

```
begin
      If reached to stopping criteria
      begin
   Continue
        else
      begin
        MGTree(DS)
      end
end
Return MGTree     --Return built MGTree
End

Algorithm-2: ComputeMgini
Input: Dataset (D), CW
--CW is class weight vector. This is optional.
--CD vector contains class size in terms of number of instances
Output: MGini
Begin
CD = class distribution from dataset  --CD is class distri. vect.
Pi = | CDi | / | D |
If CW is empty
  $CW_i = CW_j / CW_i$, j<>i, j=i-1  --CW is relative weight.
For each $CW_i$ from CW
begin
$MC_i=WC_i * 10^k$   --MC is misclassification cost vector
  $\delta_j$ = log($MC_i$i), $MC_i > 1$, j= i – 1
  ProbSum = ($P_i/\delta_j$) * (Pi /$\delta_j$)
end
Mgini = 1 - ProbSum
Return Mgini
End
```

5.1 Steps of Our Proposed Approach

The MiDT algorithm uses Mgini as a splitting criterion. Due to the use of weighted misclassification factors, the Mgini selects the splitting criteria which alleviate the problem of biasing towards the majority class. The main steps of our proposed algorithm are described below:

Step 1: Pre-process the dataset as explained in Section 3.1.
Step 2: Class distribution ratio is derived from the class distribution.
Step 3: Misclassification factors for both the classes are calculated as per the Eq. (3).
Step 4: From the misclassification factor the Mgini is calculated as per the Eq. (2) for each possible attribute-value pair.
Step 5: The feature-value pair which has the least Mgini value is selected as a splitting criterion and the dataset is split according to that.
Step 6: The process mentioned in Steps 4 and 5 is repeated for the subsets of the dataset.

We used 10-fold cross-validation for training and testing. We had evaluated our model in terms of accuracy and misclassification costs. The total misclassification cost is calculated using Eq. (6).

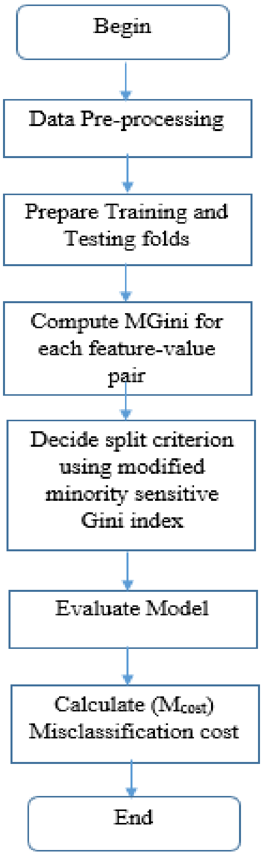

Begin
Data Pre-processing
Prepare Training and Testing folds
Compute MGini for each feature-value pair
Decide split criterion using modified minority sensitive Gini index
Evaluate Model
Calculate (Mcost) Misclassification cost
End

Fig. 2. Flowchart of proposed MiDT algorithm.

6. RESULT AND DISCUSSION

Table 1 shows the results of the J48, CART, MiDT, and SVM for the COVID-19 dataset. Results indicate that the value of the majority biased accuracy parameter is best in the CART algorithm. However, MiDT gives the best ROC value. ROC is a good performance evaluation parameter for the imbalanced data [7, 9]. For the COVID-19 data, the correct prediction of positive cases is very important. The prediction accuracy of positive cases is 67.04% in the case of MiDT. Which is best among other machine learning algorithms we used for our experiment.

The majority portion of the COVID-19 dataset is covered with missing values and there is some noise as well. The result of machine learning can be further improved with a noise-free COVID-19 dataset.

To investigate more about the performance of MiDT we tested it with the biopsy data on breast cancer patients and mammography datasets. Breast cancer has 458 majority instances and 241 minority instances. Table 2 shows the result. For the breast cancer data, the result is far better as compared to the COVID-19 data and the MiDT shows the most optimal minority class accuracy in comparison to the J48 and CART algorithms.

Table 1. Result comparison of MiDT with J48, SVM, and CART.

Parameters	J48	MiDT	CART	SVM
Accuracy	90.77	93.255	93.751	85.5
ROC	0.797	0.82	0.81	0.5
Minority Accuracy	62.5	67.04	63.63	0.0
Majority Accuracy	96.56	98.07	99.03	1

Table 2. Result for breast cancer dataset.

Algorithm	Accuracy	MinorityAcc.	MajorityAcc.
J48	94.84	92.9	95.9
MiDT	94.26	94.6	94.3
CART	94.24	94.1	94.3

For the breast cancer dataset, J48 gives an accuracy of 94.84% which is the best among all three classifiers. However, for the minority class, it is 92.9%, which is the least among the three algorithms. The CART gives a majority accuracy of 94.3% and minority accuracy 94.19%. MiDT shows the optimal result for minority class as compared to the CART and J48 algorithms.

Table 3 shows the performance of J48, CART, and MiDT algorithms for the mammography dataset.

Table 3. Result for mammography dataset.

Algorithm	Accuracy	Minority Acc.	Majority Acc.
J48	98.56	54.6	99.615
MiDT	98.53	56	99.56
CART	98.1	22.3	99.29

From these results, we discover that accuracy measure does not give real performance evaluation of the imbalanced data. For the mammography dataset, J48 gives an accuracy of 98.56% which is best among all three classifiers. However, for the minority class, it is only 54.6%. The CART gives a majority accuracy of 99.29% and minority accuracy of 22.3%. However, it still manages to show the overall accuracy of 98.1% which hides the low accuracy of the minority class. The MiDT shows the optimal result as compared to the CART algorithm. As compared to the CART algorithm which uses the Gini index, the Mgini based MiDT almost double the accuracy rate of the minority class.

We did a cost analysis for the COVID-19 dataset. If we consider misclassification cost as an evaluation parameter then MiDT is the best performer among all four classifiers used for the COVID-19 data. The cost analysis in terms of misclassification cost is given in Table 4. To make a clear difference between the misclassification cost of two classes of COVID-19 and with consideration of the aftermath of each class, we consider a misclassification cost of 95 for the wrong classification of the COVID-19 positive case and misclassification cost of 5 for the wrong classification of the negative case as a positive case. In the COVID-19, if the negative case is misclassified as the positive case then it just requires a few more medical tests and engagement of medical staff to look after the patient, which has a nominal cost in comparison to the large costs associated with the misclassification of a positive case. Total Misclassification cost (TMC) is calculated using the formula given in Eq. (6).

$$TMC = MC_{min} * (100 - MinorityAcc) + MC_{maj} * (100 - MajorityAcc) \qquad (6)$$

MC_{min} is a misclassification cost of the minority class and MC_{maj} is a misclassification cost of the majority class. MiDT ranked 1st as it has the lowest misclassification cost. MiDT reduces misclassification cost by 12% as compared to the 3rd ranked algorithm and approximately 10% as compared to the 2nd ranked algorithm. In comparison to the SVM, the MiDT has approximately 1/3rd of the misclassification cost.

Table 4. Misclassification cost analysis of MiDT, J48, CART, and SVM.

Algorithm	Minority Acc.	Majority Acc.	Total Misclassi. Cost	Rank
J48	62.5	96.56	3579.7	3
MiDT	67.04	98.07	3140.85	1
CART	63.63	99.03	3460	2
SVM	0.0	1.0	9995	4

As shown in Table 4, the two algorithms – CART and J48 have the best accuracy however they lagging behind MiDT in the reduction of misclassification cost. SVM shows the highest misclassification cost and MiDT is most optimal in the reduction of misclassification cost.

7. CONCLUSION

In this time of the pandemic, we realize the importance of a machine learning algorithm for effective medical diagnosis from imbalanced medical data. Due to the imbalanced class distribution, the performance of the traditional decision tree algorithms suffer. The existing splitting criterion of the decision tree is biased towards the majority class that makes the decision tree to favor the majority class. To handle the class imbalance, we proposed a cost-sensitive decision tree (MiDT) that uses a minority sensitive Gini index (Mgini) as a splitting criterion. Empirical results show that the proposed approach works well even for an imbalanced dataset of COVID-19.

Machine learning algorithms should try to reduce misclassification costs instead of focusing on the improvement of accuracy when there is a huge difference between misclassification costs. In the case of the imbalanced data, the overall accuracy alone cannot give true performance evaluation as it is dominated by the majority class accuracy. The total misclassification cost should be considered for the true performance evaluation instead of the overall accuracy when the accurate prediction of the minority class is more important.

Our proposed MiDT algorithm reduces misclassification costs by 10% as compared to the traditional decision tree classifiers such as the CART and J48. In the comparison of the SVM, it shows the improved result for the minority class prediction. For the imbalanced data of Covid-19, the SVM misclassified all minority class instances. The MiDT shows the unbiased result as compared to the SVM for the Covid-19 dataset. The misclassification cost of the MiDT is less than half of the SVM that shows the importance of the MiDT for imbalanced data. The performance of the MiDT algorithm also verified using breast cancer and mammography datasets. As compared to the CART algorithm, for the mammography dataset the MiDT almost double the accuracy of the minority class. MiDT improved imbalanced data classification and reduces misclassification cost without any requirement of data level sampling. In the future, the MiDT can be tested upon other datasets and its performance can be improved further by tuning the misclassification factor.

REFERENCES

1. H. Y. F. Wong, H. Y. S. Lam, and A. H. Fong, "Frequency and distribution of chest radiographic findings in COVID-19 positive patients," *Radiology*, Vol. 296, 2020, pp. E72-E78.
2. M. Y. Ng, E. Lee, and J. Yang, "Imaging profile of the COVID-19 infection: radiologic findings and literature review," *Radiol Cardiothorac Imaging*, 2020, pp. e200034.
3. A. Bernheim, X. Mei, and M. Huang, "Chest CT findings in coronavirus disease-19 (COVID-19): relationship to duration of infection," *Radiology*, 2020, pp. 685-691.
4. Moodys' Prediction of Growth Rate of India: https://www.moodys.com/researchandratings/country/india/-/0420C9/0420C9/-/-1/0/-/0/-/-/-/-/-/-/en/global/pdf/-/rra
5. J. Han, M. Kamber, and J. Pei, *Data Mining Concepts and Techniques*, 3rd ed., Elsevier, 2012.
6. Covid-19 Pandemic Information, https://www.who.int/emergencies/diseases/novel-coronavirus-2019?gclid=Cj0KCQjwzN71BRCOARIsAF8pjfjPxrand4CxLPc2iNdzfz6Tut1ZsA4LkWLU7DmS8VbvDQdYd5I2nYoaAmN-EALw_wcB.
7. P. A. Barot and H. B. Jethva, "Statistical study to prove importance of causal relationship extraction in rare class classification," *Smart Innovation*, *Systems and Technologies*, Springer, Vol. 1, DOI 10.1p0ra0t7ik/9a7b8a-r3o-t3@19g-m63a6il7.c3o-3m_51
8. D. Muchlinski, D. Siroky, J. He, and M. Kocher, *Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data*, Oxford University Press, 2015.
9. J. Stefanowski, *Dealing with Data Difficulty Factors While Learning from Imbalanced Data*, Springer International Publishing, Switzerland, 2016.
10. I. Al-Turaiki, M. Alshahrani, and T. Almutairi, "Building predictive model for MERS-CoV infections using data mining techniques," *Journal of Infections and Public Health*, Vol. 9, 2016, pp. 744-748.
11. https://www.un.org/sustainabledevelopment/blog/2020/04/covid-19-likely-to-shrink-global-gdp-by-almost-one-per-cent-in-2020/.
12. Bulletin of the Atomic Scientists, https://thebulletin.org/2020/04/covid-19-and-the-doomsday-clock-observations-on-managing-global-risk/.
13. B. Krawczyk, M. Galar, L. Jelén, and F. Herrera, "Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy," *Applied Soft Computing Journal*, 2015, http://dx.doi.org/10.1016/j.asoc.2015.08.060.
14. B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Progress in Artificial Intelligence*, Springer, 2016, pp. 221-232.
15. R. B. Hegde, K. Prasad, H. Hebbar, *et al.*, "Development of a robust algorithm for detection of nuclei and classification of white blood cells in peripheral blood smear images," *Journal of Medical Systems*, Vol. 42, 2018, p. 110.

16. E. Venkatesan and T. Velmurugan, "Performance analysis of decision tree algorithms for breast cancer classification," *Indian Journal of Science and Technology*, Vol. 8, 2015, pp. 1-8.
17. F. Wu, S. Zhao, B. Yu, *et al.*, "A new coronavirus associated with human respiratory disease in China," *Nature*, 2020, pp. 265-269.
18. I. Chaabane, R. Guermazi, and M. Hammami, "Enhancing techniques for learning decision trees from imbalanced data," *Advances in Data Analysis and Classification*, Vol. 14, 2020, pp. 677-745.
19. L. Wynants, V. Calster, B. Ben, *et al.*, "Prediction models for diagnosis and prognosis of covid-19 infection: Systematic review and critical appraisal," *BMJ*, Vol. 369, 2020, No. m1328.
20. L. Li, *et al.*, "Propagation analysis and prediction of the COVID-19," *Infectious Disease Modelling*, Vol. 5, 2020, pp. 282-292.

Pratikkumar A. Barot received the B.E. degree From H.N. G.U. and M.E. degree in Computer Engineering from Gujarat Technological University, India. He is pursuing the Ph.D. in Computer Engineering from Gujarat Technological University, India. His research interests include machine learning, data mining, and algorithm design.

Harikrishna B. Jethva currently works at the Head of Department, Department of Computer Engineering, Government Engineering College, Patan, Gujarat, India, His research interest in machine learning, neural network, theory of computation, compiler design, soft computing and algorithms. He is Ph.D. Supervisor in Gujarat Technological University. In addition, he is a Board of Study member in many universities.