Combining Hybrid Approach Redefinition-Multiclass Imbalance (HAR-MI) and Hybrid Sampling in Handling Multi-Class Imbalance and Overlapping

Hartono*a,b,*, Erianto Ongko*c

*a Department of Computer Science, Universitas IBBI, Medan, 20114, Indonesia
*b Department of Computer Science, Universitas Potensi Utama, Medan, 20241, Indonesia
*c Department of Informatics, Akademi Teknologi Industri Immanuel, 20114, Medan, Indonesia

Corresponding author: *hartono@ibbi.ac.id

Abstract—The class imbalance problem in the multi-Class dataset is more challenging to manage than the problem in the two classes and this problem is more complicated if accompanied by overlapping. One method that has proven reliable in dealing with this problem is the Hybrid Approach Redefinition-Multiclass Imbalance (HAR-MI) method which is classified as a hybrid approach that combines sampling and classifier ensembles. However, in terms of diversity among classifiers, a hybrid approach that combines sampling and classifier ensembles will give better results. HAR-MI provides excellent results in handling multi-class imbalances. The HAR-MI method uses SMOTE to increase the number of samples in the minority class. However, this SMOTE also has a weakness where an extremely imbalanced dataset and a large number of attributes will be over-fitting. To overcome the problem of over-fitting, the Hybrid Sampling method was proposed. HAR-MI combination with Hybrid Sampling is done to increase the number of samples in the minority class and at the same time reduce the number of noise samples in the majority class. The preprocessing stages at HAR-MI will use the Minimizing Overlapping Selection under Hybrid Sampling (MOSHS) method, and the processing stages will use Different Contribution Sampling. The results obtained will be compared with the results using Neighbourhood-based under-sampling. Overlapping and Classifier Performance will be measured using Augmented R-Value, the Matthews Correlation Coefficient (MCC), Precision, Recall, and F-Value. The results showed that HAR-MI with Hybrid Sampling gave better results in terms of Augmented R-Value, Precision, Recall, and F-Value.

Keywords— Class imbalance; multi-class dataset; multi-class imbalance; hybrid approach; HAR-MI.

I. INTRODUCTION

The problem of class imbalance has become one of the most exciting data mining problems [1]. The class imbalance has become one of the most interesting research issues regarding data mining, machine learning, and knowledge discovery[2]. This problem occurs because most of the real-world dataset is in an imbalanced state and if it is not handled properly it will cause a class with a small number of samples to become unrepresented and reduce the level of accuracy[3]. In general, the approach to solving class imbalance problems can be divided into 3 (three), namely: data-level, algorithm-level, and hybrid[4]. The data-level approach focuses on efforts to change the distribution of data through a process of over sampling or under-sampling. Oversampling was carried out on the minority class and under-sampling was carried out on the majority class[5]. On the other hand, the algorithm-level approach does not change the distribution of data, but focuses on classifier efforts to pay more attention to minority classes by applying bagging, boosting, or through the ensemble process of existing classifiers[6].

Hybrid Approach is an approach that combines Data-Level and Algorithm-Level[7]. In terms of diversity and classifier performance, a hybrid approach that combines sampling and classifier ensembles will give good results[8]. The Hybrid Method is good at dealing with the binary-class imbalance and multi-class imbalance problems[9]. Multi-class imbalance problems are more difficult to handle than binary-class imbalance, and usually, multi-class balance problems do not stand alone but are accompanied by overlapping[10]. This problem becomes even more challenging if the minority classes are in overlapping conditions[11].
To minimize the impact of multi-class imbalance which is accompanied by overlapping, the preprocessing process has a very significant effect[12]. For this problem, the feature selection method is often used at the preprocessing stage, so the effort to apply the preprocessing stage in the hybrid approach is a wise choice[13]. One of the hybrid approach methods that was applied to preprocess and gives satisfactory results in this problem is the Hybrid Approach Redefinition-Multiclass Imbalance (HAR-MI)[14].

As with most hybrid approach methods, HAR-MI also uses the oversampling method for minority classes by using SMOTE in the feature selection process at the preprocessing stage. One of the Feature Selection methods that provide excellent results in handling overlapping is Minimizing Overlapping Selection under SMOTE (MOSS)[15], even though this oversampling process often causes overfitting[16]. Besides, other problems that are often found in the application of SMOTE are overgeneralization and noise[17]. The use of Minority Over-Sampling Techniques (M-SMOTE) and Edited Nearest Neighbor (ENN), which are a type of Hybrid Sampling, has yielded very satisfying results [18].

It would be interesting if there is a method that combines multi-class balance handling followed by overlapping and at the same time paying attention so that the sampling process does not overfit. This study will combine the use of HAR-MI with Hybrid Sampling. This study's results will be compared with Neighborhood-based under-sampling, which is one of the best methods of handling multi-class imbalance and overlapping[19].

II. MATERIALS AND METHOD

A. Hybrid Approach

The pseudocode of the Hybrid Approach is as follows[20].

Input: $D_T = \{x_1, x_2, ..., x_n\}$/Training Dataset  
$N =$ Number of Classifier  
Output: Classification Prediction $P$

Method:

Step 1 Preprocessing using Preprocessing Method

Step 2 For $i = 1$ to $N$

i. Apply Machine Learning Classification Algorithm on The Attributes of $D_T$

ii. Obtain Classification Prediction $P_i$ from machine learning classification algorithm

End For

Step 3 For $i = 1$ to $n$

Apply processing using bagging, boosting or sampling

End For

B. Hybrid Sampling

The pseudocode of the Hybrid Sampling using M-SMOTE and ENN is as follows[18].

Input: Dataset $S$, Minority Samples $S_{min}$, Majority Sample $S_{maj}$

Output: Final Dataset $S'$

Create global variable $G_{max}$, create array $Eva_{min}, Eva_{maj}, Eva$

Step 1: If $G_{MCC} = 0$

Processing $M - SMOTE$ for $S_{min}$

Processing $ENN$ for $S_{maj}$

End If

Step 2: Calculate $Eva_{min}$ using MCC  
Calculate $Eva_{maj}$ using MCC  
Calculate $Eva$ using MCC

Step 3: If $Eva_{min} < Eva$ or $Eva_{maj} < Eva$  
$G_{MCC} = G_{MCC} - 1$  
End If

Step 4: if $G_{MCC} < 0$  
Terminate and Output Final Dataset $S'$

else  
Return to Step 1  
End If

C. Augmented R-Value

Augmented R-Value states how much overlapping occurs. The greater the Augmented R-Value, the greater the overlapping[21].

$$R_{aug}(D[V]) = \frac{\sum_{i=1}^{k-1} |C_{k-1} - |C_i|}{\sum_{k=1}^{k} |C_i|}$$ (1)

Where $C_0, C_1, ..., C_{k-1}$ are $k$ class labels with $|C_0| \geq |C_1| \geq ... \geq |C_{k-1}|$ and $D[V]$; Dataset D containing predictors in set V. Larger $R_{Aug}$ is higher overlap degree of a dataset.

D. Classifier Performance

Classifier Performance was measured using the Matthews Correlation Coefficient (MCC), Precision, Recall, and F-Value. This classifier performance measurement is carried out based on the confusion matrix shown in Table I[22].

| Predictive Positive Class | Predictive Negative Class |
|---------------------------|---------------------------|
| Actual Positive Class     | True Positive (TP)        | False Negative (FN) |
| Actual Negative Class     | False Positive (FP)       | True Negative (TN) |

The Matthews Correlation Coefficient (MCC), Precision, Recall, and F-Value calculations can be seen in the following equation[18].

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TN \times FN)(TN \times FP)(TN \times FN)(TP \times FP)}}$$ (2)

$$Precision = \frac{TP}{TP + FP}$$ (3)

$$Recall = \frac{TP}{TP + FN}$$ (4)

$$F - Value = \frac{2 \times Precision \times Recall}{Precision + Recall}$$ (5)

E. Proposed Method / Algorithm

The research stages can be seen in Fig. 1.
F. Preprocessing Using Minimizing Overlapping Selection under Hybrid Sampling (MOSHS)

The pseudocode of the preprocessing stage is as follows.

1: \( X \) — matrix with \( p \) predictors: \( X = \{x_1, x_2, ..., x_p\} \), class label: \( y \)
2: For All Samples in Minority
3: \( \text{Hybrid Sampling the Minority Class using } m - \text{SMOTE} \)
4: End For
5: Create NewMinority
6: For All Samples in Majority
7: \( \text{Hybrid Sampling the Majority Class using ENN} \)
8: End For
9: Create NewMajority
10: For All Samples in NewMinority and NewMajority
11: \( \text{Preprocessed Dataset} \)
12: End For

G. Processing Using Different Contribution Sampling (DCS)

The pseudocode of the processing stage is as follows.

1: For \( i = 1 \) to Number of Instance in Preprocessed Dataset
2: Add Preprocessed Dataset to \( S_i \)
3: B — SVM will do for Classifying \( S_i \)
4: Determine the Majority Class
5: Determine the Minority Class
6: For All Instance in Majority Class
7: \( \text{NewSVSets} \) will form by checking and delete the noise in SVSets
8: \( \text{NewNSVSets} \) will form by Multiple Hybrid Sampling
9: End For
10: For All Instance from \( \text{NewSVSets} \) and \( \text{NSVSets} \)
11: Create an instance for majority class
12: End For
13: For All Instance in minority class
14: \( \text{SMOTEBoost Process for SVSets and Create SMOTESets} \)
15: End For
16: For All \( \text{SMOTESets and NewNSVSets} \) do

III. RESULTS AND DISCUSSION

A. Dataset Description

The multi-class imbalanced datasets used in this study were sourced from the KEEL Repository[23]. The dataset used can be seen in Table II.

| Dataset          | #Ex | #Atts | Distribution of Class | IR  |
|------------------|-----|-------|-----------------------|-----|
| Contraceptive    | 1473| 9     | 629/333/511           | 1.89|
| Flare            | 1066| 11    | 147/211/239/95/43/331 | 7.70|
| Car              | 1728| 6     | 384/69/1210/65        | 18.62|
| Thyroid Disease  | 720 | 21    | 17/37/666             | 39.18|
| Red Wine Quality | 1599| 11    | 10/53/681/638/199/18  | 68.10|
| Page-Blocks      | 5473| 10    | 4913/329/28/88/115    | 188.72|

Table II shows that the dataset used has various imbalance ratios, ranging from low, medium, and high imbalance ratios. Likewise, the number of samples also varied.
B. Testing Result

The first test was conducted to obtain Augmented R-Value and MCC values. The test results can be seen in Table III.

| Dataset          | HAR-MI with Hybrid Sampling | Neighborhood Based Under-sampling |
|------------------|-----------------------------|----------------------------------|
|                  | Augmented R-Value | MCC   | Augmented R-Value | MCC   |
| Contraceptive    | 0.327            | 0.97  | 0.337            | 0.91  |
| Flare            | 0.357            | 0.83  | 0.359            | 0.82  |
| Car Evaluation   | 0.367            | 0.85  | 0.373            | 0.81  |
| Thyroid Disease  | 0.379            | 0.81  | 0.381            | 0.79  |
| Red Wine Quality | 0.411            | 0.75  | 0.415            | 0.71  |
| Page-Blocks      | 0.436            | 0.73  | 0.437            | 0.71  |

Based on Table III, it can be seen that for the Augmented R-Value, the results obtained by HAR-MI with Hybrid Sampling are better than the Neighborhood-based under-sampling. The greater the Augmented R-Value, the greater the tendency for overlapping to occur. The MCC value provided by HAR-MI with Hybrid Sampling is also better than that obtained by Neighborhood-based under-sampling. The second test was conducted to obtain Precision, Recall, and F-Value. The test results can be seen in Table IV.

| Dataset          | HAR-MI with Hybrid Sampling | Neighborhood Based Under-sampling |
|------------------|-----------------------------|----------------------------------|
|                  | Precision | Recall | F-Value  | Precision | Recall | F-Value  |
| Contraceptive    | 0.88      | 0.97   | 0.92     | 0.78      | 0.89   | 0.83     |
| Flare            | 0.85      | 0.88   | 0.87     | 0.81      | 0.87   | 0.84     |
| Car Evaluation   | 0.84      | 0.89   | 0.86     | 0.76      | 0.73   | 0.75     |
| Thyroid Disease  | 0.87      | 0.76   | 0.81     | 0.85      | 0.71   | 0.77     |
| Red Wine Quality | 0.82      | 0.81   | 0.81     | 0.82      | 0.72   | 0.77     |
| Page-Blocks      | 0.78      | 0.77   | 0.77     | 0.77      | 0.69   | 0.73     |

Based on Table IV, it can be seen that based on the Precision, Recall, and F-Value values the results given by HAR-MI with Hybrid Sampling are better than the results obtained by Neighborhood-based under-sampling.

C. Statistical Tests

To validate the results of the study, a statistical test was conducted to measure performance using the Wilcoxon Signed-Rank Test[24]. The statistical test results can be seen in Table V.

D. Discussion

Based on the test results and Statistical Tests, it can be seen that in terms of overlapping the HAR-MI method with Hybrid Sampling gives better results compared to MCC between HAR-MI with Hybrid Sampling and Neighborhood-Based Under-sampling. However, in general, the results obtained in overlapping handling are good, where the Augmented R-Value obtained is not too high. Augmented R-Value is very dependent on the imbalance ratio; the higher the value of the imbalance ratio, the higher the overlapping that occurs. There is a significant difference for Augmented R-Value and MCC between HAR-MI with Hybrid Sampling and Neighborhood-Based Under-sampling based on statistical tests.

As for the MCC value, the results given by HAR-MI with Hybrid Sampling are still better and there is a tendency that the more classes there are, the lower the MCC value obtained. As for the Precision, Recall, and F-Value values, the results obtained show that HAR-MI with Hybrid Sampling is also better than MCC between HAR-MI with Hybrid Sampling
and Neighbourhood-Based Under-sampling. The results obtained show that the higher the imbalance ratio, the value of Precision, Recall, and F-Value obtained also decreases.

Based on the results of statistical testing with the Wilcoxon Signed-Rank Test, it was found that for Augmented R-Value, the P-Value is 0.0355223, the P-Value for MCC is 0.0355223, the P-Value for Recall is 0.0312500, and the P-Value for F-Value is 0.0340604. This means that for Augmented R-Value, MCC, Recall, and F-Value, there is a significant difference between HAR-MI results with Hybrid Sampling and Neighbourhood-Based Under-sampling. As for Precision, although HAR-MI results are better than Neighbourhood-Based Under-sampling but based on the test results with the Wilcoxon Signed-Rank Test, there is no significant difference as indicated by the P-Value obtained > 0.05, where the P-Value obtained is 0.0625000.

IV. CONCLUSION

Based on the results in Tables III, IV, and V, it can be seen that in terms of handling multi-class imbalance and overlapping, the results obtained using HAR-MI with Hybrid Sampling give better results compared to Neighbourhood-Based Under-sampling. The results obtained show that HAR-MI with Hybrid Sampling excels at all test values such as Augmented R-Value, MCC, Precision, Recall, and F-Value.

This shows that for handling multi-class imbalance, Hybrid Sampling, which can avoid over fitting, also gives better results compared to Under-sampling or Over Sampling. Future Research can pay attention to the handling of multi-class imbalance accompanied by overlapping in a state of high yield ratio and datasets with a large number of classes and many attributes.

ACKNOWLEDGMENT

A Grant from the Ministry of Education and Culture and the Ministry of Research and Technology of the Republic of Indonesia funded this research.

REFERENCES

[1] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanxue, and G. Bing, “Learning from Class-Imbalanced Data: Review of Methods and Applications,” Expert Systems with Applications, vol. 73, pp. 220–239, May 2017.

[2] A. Guzmán-Ponce, J. S. Sánchez, R. M. Valdovinos, and J. R. Marcial-Romero, “DBG-US: A two-stage under-sampling algorithm to face the class imbalance problem,” Expert Systems with Applications, p. 114301, Nov. 2020, doi: 10.1016/j.eswa.2020.114301.

[3] B. Liu and G. Tsoumakas, “Dealing with class imbalance in classifier chains via random under-sampling,” Knowledge-Based Systems, vol. 192, p. 105292, Mar. 2020, doi: 10.1016/j.knosys.2019.105292.

[4] J. M. Johnson and T. M. Khoshgoftaar, “Survey on deep learning with class imbalance,” J Big Data, vol. 6, no. 1, p. 27, Mar. 2019, doi: 10.1186/s40537-019-0192-5.

[5] P. Shamsoleini, M. Zarapoor, L. Shen, A. H. Sadka, and J. Yang, “Imbalanced data learning by minority class augmentation using capsule adversarial networks,” Neurocomputing, Jul. 2020, doi: 10.1016/j.neucom.2020.01.119.

[6] W. Hou, X. Wang, H. Zhang, J. Wang, and L. Li, “A novel dynamic ensemble selection classifier for an imbalanced data set: An application for credit risk assessment,” Knowledge-Based Systems, vol. 208, p. 106462, Nov. 2020, doi: 10.1016/j.knosys.2020.106462.

[7] F. Rayhan, S. Ahmed, A. Mahhub, R. Jani, S. Shatabda, and D. M. Farid, “CUSBoost: Cluster-Based Under-Sampling with Boosting for Imbalanced Classification,” in 2017 2nd International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS), Dec. 2017, pp. 1–5, doi: 10.1109/CSITSS.2017.8447534.