Bulletin of Electrical Engineering and Informatics
Vol. 10, No. 3, June 2021, pp. 1718~1728
ISSN: 2302-9285, DOI: 10.11591/eei.v10i3.3057

Hybrid approach redefinition-multi class with resampling and feature selection for multi-class imbalance with overlapping and noise

Erianto Ongko[1], Hartono[2]
[1]Department of Informatics, Akademi Teknologi Industri Immanuel, Indonesia
[2]Department of Computer Science, Universitas IBBI, Indonesia
[2]Department of Computer Science, Universitas Potensi Utama, Indonesia

Article Info

Article history:

Received Jul 24, 2020
Revised Dec 4, 2020
Accepted May 1, 2021

Keywords:

Class imbalance
Classifier performance
Multi-class
Noise
Overlapping

ABSTRACT

Class imbalance and overlapping on multi-class can reduce the performance and accuracy of the classification. Noise must also be considered because it can reduce the performance of classification. With a resampling algorithm and feature selection, this paper proposes a method for improving the performance of hybrid approach redefinition-multi class (HAR-MI). Resampling algorithm can overcome the problem of noise but cannot handle overlapping well. Feature selection is good at dealing with overlapping but can experience a decrease in quality if there is a noise. The HAR-MI approach is a way to deal with multi-class imbalance issues, but it has some drawbacks when dealing with overlapping. The contribution of this paper is to suggest a new approach for dealing with class imbalance, overlapping, and noise in multi-class. This is accomplished by employing minimizing overlapping selection (MOSS) as an ensemble learning algorithm and a preprocessing technique in HAR-MI, as well as employing multi-class combination cleaning and resampling (MC-CCR) as a resampling algorithm at the processing stage. When subjected to overlapping and classifier performance, it is discovered that the proposed method produces good results, as evidenced by higher augmented r-value, class average accuracy, class balance accuracy, multi class g-mean, and confusion entropy.

Corresponding Author:

Erianto Ongko
Department of Informatics
Akademi Teknologi Industri Immanuel
Jalan Jendral Gatot Subroto No. 325 Medan, Indonesia
Email: eriantoongko@gmail.com

1. INTRODUCTION

The class imbalance occurs when a class has a significantly smaller number of instances than other classes, as determined by the imbalance ratio (IR), which is the ratio of a class with a significantly smaller number of instances (minority class) to a class with a significantly larger number of instances (majority class) [1] and basically machine learning algorithms work optimally if each class has a number of instances that are not much different [2]. This problem is one of the causes of the low accuracy of classification problems and also causes important information contained in the minority class can not be obtained due to better coverage on the majority class [3]. Handling of multi-class imbalance has greater difficulty compared to two-class problems, especially when it comes to accuracy and difficulty of training data on large datasets with high imbalance ratios [4]. Another thing that escapes attention is the overlapping problem, where several classes overlap with other classes. The overlapping problem has far more impact on accuracy compared to class imbalance [5]. Overlapping conditions can increase the accuracy of one class by decreasing the accuracy of another class. For example, although the overlapping regions have a high concentration of minority classes, the classification results can also provide low accuracy because some instances associated with majority classes are eliminated [6].

The overlapping problem in multi-class can be overcome by using feature selection method which is very effective in dealing with overlapping problems [7]. The reason why feature selection is effective in dealing with overlapping is because of its ability to eliminate uninformative predictors and reduce dimensionality of feature space [8]. However, on the other hand with the noise, the performance given by feature selection can decrease [9] and noise basically has an influence on classification performance [10]. Noise handling in general uses the method of resampling, but often encounters obstacles if there is a state of overlapping [6]. This is an obstacle when handling multi-class imbalance and at the same time also faces other obstacles in the form of overlapping and noise. A number of studies have discussed handling class imbalance accompanied by overlapping or noise. Koziarski *et al.* [11] has proposed the a multi-class combined cleaning and resampling (MC-CCR) method which has the ability to overcome the noise problem but has obstacles in dealing with overlapping. The feature selection method on the other hand has been used by a number of researchers in dealing with overlapping problems, such as: [12] that has proposed density based feature selection and [13] that has proposed rough-set-based feature selection algorithm for imbalanced data (RSFSAID) algorithm.

The ensemble learning approach, especially the hybrid ensembles, is very commonly used in overcoming multi-class imbalance problems [14]. The hybrid ensembles approach has a good ability in handling multi-class imbalances accompanied by overlapping, but what needs to be considered is in situations where the imbalance ratio is high and also conditions that contain noise and overlapping [15]. Xie *et al.* [16] has stated a number of things that need to be considered in improving performance on hybrid ensembles, such as the use of the right selection method on the noise label and also the right sampling at the processing stage. Research conducted by [17] and [18] shows that using the appropriate feature selection method at the preprocessing stage in the hybrid ensembles can provide a good result in handling class imbalance and overlapping. Noise handling in hybrid ensembles can be overcome by choosing the right sampling method at the processing stage [19], [20].

The hybrid ensembles approach that combines the application of preprocessing by using feature selection and sampling at the processing stage is hybrid approach redefinition-multiclass imbalance [21]. The hybrid approach redefinition-multi class (HAR-MI) approach will be combined with the resampling algorithm in the processing stage and feature selection in the preprocessing stage in this analysis. Selection under no sampling [22], [23] and selection under synthetic minority over-sampling technique (SMOTE) [24] are two feature selection approaches that can be used to overcome overlapping problems. According to research conducted by [5] the minimizing overlapping selection under no-sampling (MOSNS) and minimizing overlapping selection under SMOTE (MOSS) methods provide very satisfactory results in dealing with overlapping. Both methods have a similar level of efficiency.

Noise handling by using resampling at the processing stage, there have been a number of studies that have been done. The research conducted by [25] uses the SMOTE Sampling method in handling noise but this research has problems with overlapping classes and also has problems in accuracy. The same thing is also found in using the SMOTE oversampling with edited nearest neighbors (ENN) method [26]. The MC-CCR system is one sampling method that produces excellent results.

HAR-MI method which has good ability in overcoming multi-class imbalance problems but the result would be worse if there are overlapping between class and noise. In the preprocessing step, the HAR-MI system employs the random balance ensemble method, which will be paired with the MOSS method for the preprocessing stage, as well as the MC-CCR method for processing step. The findings will be compared to those obtained using the neighbourhood-based undersampling process, which is one of the best techniques for dealing with class imbalance and overlapping in multi-class imbalanced conditions [18]. Augmented r-value, class average accuracy, class balance accuracy, multi class G-mean, and uncertainty entropy were used to compare these results.

2. RELATED WORKS

2.1. Augmented R-value for multi-class

Each class's R-value shows how much of an instance overlaps the area. R-value has a strong correspondence with classifier performance, according to research conducted by [27]. As can be seen in (1), [23] has suggested a method for calculating this.

$$R_{aug}(D[V]) = \frac{\sum_{i=0}^{k-1}|C_{k-1-i}|R(C_i)}{\sum_{i=0}^{k-1}|C_i|} \tag{1}$$

Where $C_0, C_1, \ldots, C_{k-1}$ are k class labels with $|C_0| \geq |C_1| \geq \cdots \geq |C_{k-1}|$ and $D[V]$: Dataset D restraining predictors in set *V*. A higher R_{Aug} indicates a higher overlap degree.

2.2. Confusion matrix

For the general classification results the classification results can be grouped into 4 (four) groups, namely; true positive (TP), true negative (TN), false positive (FP), and false negative (FN) and can be presented in the confusion matrix as can be seen in Table 1 [28].

Table 1. Confusion matrix for a classification problem

		Predicted (Classified) as	
		Positive	Class Negative
Actually (Really is)	Positive Samples	TP	FN
	Negative Samples	FP	TN

2.3. Classifier performance

The following parameters are used to determine the classifier performance.

- Class average accuracy with *C classes* can be calculated using (2) [29].

$$AvAcc = \frac{1}{C}\sum_{i=1}^{C}\frac{tp_i+tn_i}{tp_i+tn_i+fp_i+fn_i} \tag{2}$$

Where *C* is number of class with *TP, TN, FP, and FN* are the result of predicted (classified) that was obtained from confusion matrix.

- Class balance accuracy for any C^k, confusion matrix, class balance accuracy is defined as [30].

$$CBA = \frac{\sum_i^k \frac{c_{ii}}{\max(c_{i.},c_{.i})}}{k} \tag{3}$$

According to confusion matrix for class balance accuracy as can be seen in Table 2.

Table 2. Confusion matrix for class balance accuracy

		Predicted			
		Class 1	Class 2	Class 3	Total
Actual	Class 1	C_{11}	C_{12}	C_{13}	$C_{1.}$
	Class 2	C_{21}	C_{22}	C_{23}	$C_{2.}$
	Class 3	C_{31}	C_{32}	C_{33}	$C_{3.}$
	Total	$C_{.1}$	$C_{.2}$	$C_{.3}$	N

- The G-mean was proposed as the geometric mean of recall (R) values of two groups by multi class G-mean (mGM) [31]. Sun *et al.* [32] To apply this measure to multiple-class situations, define the G-mean as the geometric mean of each class's recall values.

$$R_i = \frac{c_{ii}}{\sum_{j=1}^{k} c_{ij}} \tag{4}$$

$$mGM = \sqrt[C]{\prod_{i=1}^{C} R_i} \tag{5}$$

- Confusion entropy (CEN). Wei *et al.* [33] proposed using the confusion entropy to determine classifier efficiency. According to the confusion entropy, the misclassification information includes both how the samples with true class label cl_i were misclassified to the other N classes and how the samples from the other N classes were misclassified to class cl_i.

$$CEN = \sum_j P_j CEN_j \tag{6}$$

where

$$P_j = \frac{\sum_k(C_{j,k}+C_{k,j})}{2\sum_{k,l}C_{k,l}} \tag{7}$$

$$CEN_j = -\sum_{k=1,k\neq j}^{N+1}(p_{j,k}^{i}log_{2N}p_{j,k}^{i} + p_{k,j}^{i}log_{2N}p_{k,j}^{i} \tag{8}$$

2.4. Minimizing overlapping selection under SMOTE

The MOSS algorithm is being as [5].

```
X - matric with p predictors: $X = [x_1, x_2, \ldots, x_p]$; class label: y
Over-sampling the Positive Samples using SMOTE; merging the generated instances with original ones to get updated X-matrix, $X_{new}$ and updated class label $Y_{new}$
$X \leftarrow X_{new}; Y \leftarrow Y_{new}$
Establish sparse regularization path $\hat{\beta}(\lambda, \alpha)$ according to (9)
Compute the optimal $(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)^T$ via the (10)
Select those feature with $\hat{\beta}_j \neq 0\ for\ j = 1, 2, \ldots, p$
```

In (9) shows the sparse collection that would be used to create sparse regulatization [34].

$$C_a(\beta) = \frac{1}{2}(1-\alpha)\|\beta\|_2^2 + \alpha\|\beta\|_1 \tag{9}$$

In (10) shows the loss penalties that will be used to determine the optimum $\hat{\beta}_j$.

$$Loss = -\frac{1}{n}\sum_{i=1}^{n}(y_i\beta^T x_i - \ln(1+\beta^T x_i)) \tag{10}$$

The method of handling overlapping has started at the preprocessing stage by adding MOSS, as seen in the previous pseudocode. The MOSS process begins with the provision of *p predictors* and *class labels y*. The oversampling process in the minority class will be carried out using SMOTE, and then the sparse regularization process will be carried out using sparse collection, which can be measured using (2), and then loss penalties will be calculated using (3).

2.5. Multi-class combined cleaning and resampling

The MC-CCR algorithm is being as [11].

```
Input: $x^{(c)}$denotes a subcollection of observations belonging to class in the Set of Observations x.
Parameters: Each sphere has an energy budget for expansion, and the p-norm is used to calculate distance
Output: Observations that have been translated and oversampled X
Function MC-CCR (X, energi, p);
  C←Collection of all classes;sorted by the number of associated observations in a descending order
  for i ← 1 to |c|
    $n_{classes}$ ← number of classes with high number of observations than $C_i$
    if $n_{classes}$ > 0 then
      $X_{min} \leftarrow X^{(ci)}$
      $X_{maj} \leftarrow \emptyset$
      for j ← 1 to $n_{classes}$
        add $\left\lfloor\frac{|X^{(c_i)}}{n_{classes}}\right\rfloor$ randomly selected from $x^{(j)}$ to $x_{maj}$
      end
      $X'_{maj}$, S ← CCR ($X_{maj}$, $X_{min}$, energy, P)
      $X^{(C_i)} \leftarrow X^{(C_i)} \cup S$
      Substitute observation used to construct $X_{maj}$ with $X'_{maj}$
    end if
  end
  return X
```

The MC-CCR method is used to eliminate noise at the processing stage. It should be noted that the impact is limited and basically the combined majority observations.

2.6. Hybrid approach redefinition for multi-class imbalance

The algorithm of hybrid approach redefinition for multi-class imbalance is being as [21].

```
Require: Set S of examples $(x_1, y_1)$
Ensure: New set $S'$ of examples with Random Balance Ensemble Method
totalSize ← |S|
Determine k using Dynamic Ensemble Selection
Building the candidate ensemble for Safe, Borderline, Rare, and Outlier
For all samples in Majority and Minority
    Preprocessing Satge using Random Balance Ensemble Method
    New Majority and New Minority of Preprocessing Dataset
End
Determine the Augmented R-Value
For all instances in Preprocessing Dataset
    Determine Majority and Minority Class
    For All Instances in Majority Class
        Biased Support Vector Machine for Determine SV Sets and NSV Sets
    End
    For All Instances in Minority Class
        Biased Support Vector Machine for Determine SV Sets and NSV Sets
    End
End
For All Instances in NSV Sets from Majority Class
    Process Multiple Random Under Sampling
End
For All Instances in SV Sets from Minority Class
    Process SMOTEBoost
End
```

Based on the preceding algorithm, it is clear that the HAR-MI method is divided into 2 (two) major stages: preprocessing and processing. The random balance ensemble method and dynamic ensemble selection are used in the preprocessing stage. It is clear that the preprocessing stages will generate preprocessing datasets, which will then go through processing stages using different contribution sampling. There are biased support vector machine stages in different contribution sampling that will produce SV sets and no scalpel vasectomy (NSV) sets for both the majority and minority classes. NSV sets from majority classes are then processed using multiple random under sampling, while SV sets from minority classes are processed using SMOTE boost.

3. RESEARCH METHOD

Figure 1 shows the stages of this research. According to the previous Figure, the process will start with the preprocessing stage, which employs MOSS. The sparse selection and lasso penalty values are determined first. This stage's output will be a preprocessing dataset, which will then go through processing stages using MC-CCR. The results from HAR-MI with resampling and feature selection will then be compared to the results from neighborhood-based undersampling.

3.1. Preprocessing stage

MOSS will be used to modify the HAR preprocessing stage for multi-class problems. The following algorithm depicts the preprocessing stages.

```
Require: S as set of instances $(X_i, Y_i)$
Ensure: $S'$ is the new Set with MOSS
totalSize ← |S|
Calculate k as the number of Nearest Neighbors
For All Instances of S
    Building the Positive or Minority Class Borderline as $E_0 C_t^+$
    Building the Negative or Majority Class Borderline as $E_0 C_t^-$
End
Creating a candidate ensemble based on k value for safe, borderline, rare, and outlier candidates
Using Equation 9 to Calculate Sparse Selection
Using Equation 10 to Calculate the Loss Penalty
Determine Sparse Regulatization
For All Instances in Positive
    MOSS is used to sample all instances
```

```
Build newMinority
Build newMajority
End
Determine the Augmented R-Value
return $S^{'}$
```

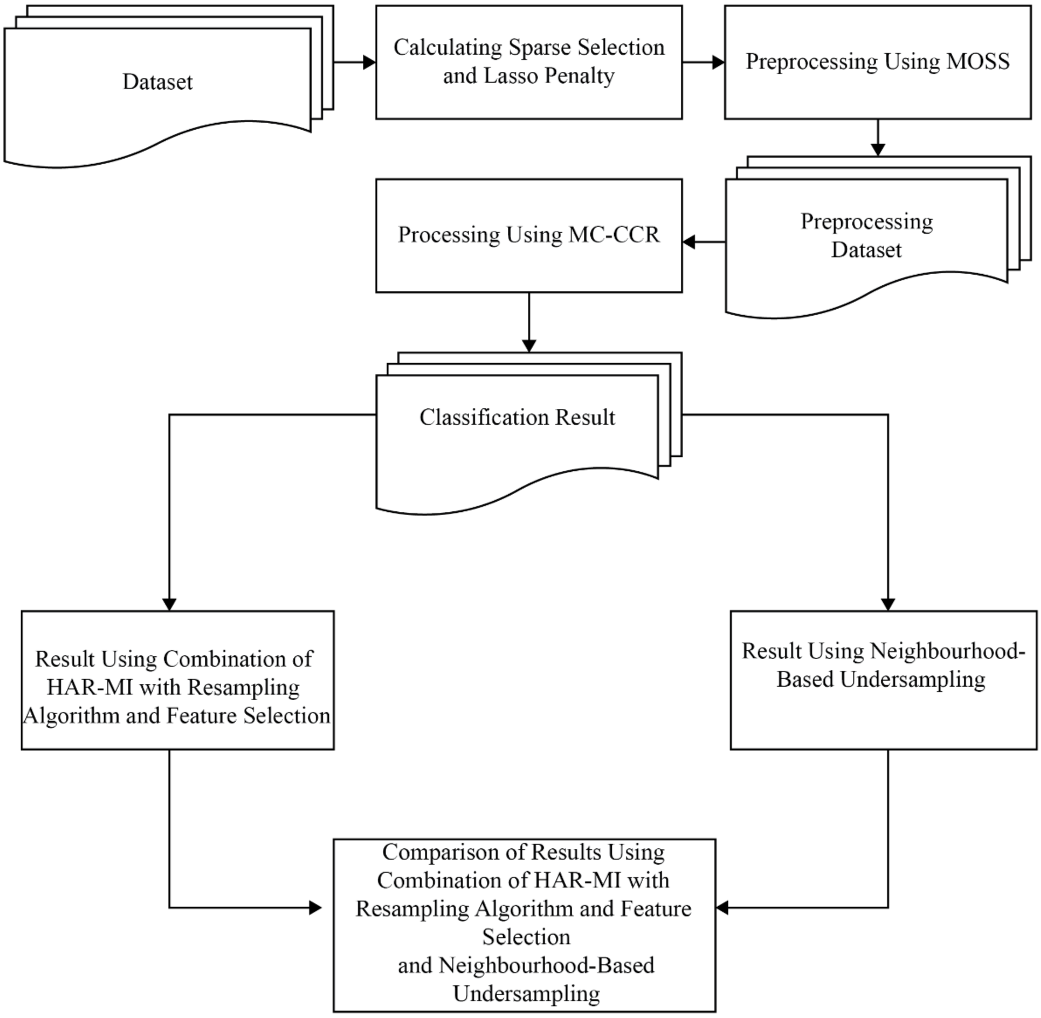
Dataset
Calculating Sparse Selection and Lasso Penalty
Preprocessing Using MOSS
Preprocessing Dataset
Processing Using MC-CCR
Classification Result
Result Using Combination of HAR-MI with Resampling Algorithm and Feature Selection
Result Using Neighbourhood-Based Undersampling
Comparison of Results Using Combination of HAR-MI with Resampling Algorithm and Feature Selection and Neighbourhood-Based Undersampling

Figure 1. Stages of research methods

Based on the preceding algorithm, it is clear that the HAR-MI preprocessing stage will be carried out using one of the feature selection methods, namely MOSS. MOSS is intended to do overlapping handling before entering the processing stage. This MOSS stage begins with determining the value of sparse selection and loss penalty. Furthermore, MOSS will be used to sample each instance in the minority class. The result is a preprocessing dataset which will then be measured in augmented R-value values.

3.2. Processing stage

The following algorithm depicts the processing steps.

```
For all samples in preprocessed dataset
Preprocessed Dataset should be added to $S_i$
Using B-SVM, determine the SV and NSV sets for the majority and minority
For All Sampes in Negative
Checking and removing noise from SV and NSV sets with MC - CCR
End
For All Samples in Positive
Check and remove the Noise in SV Sets and NSV Sets using MC - CCR
```

```
SMOTEBoost Step for SV Sets and Produce SMOTESets
End For
For All SV Sets and NSV Sets from Majority Class do
New NegativeSampleSets
End For
For All SV Sets and NSV Sets from Minority Class do
New PositiveSampleSets
End For
End For
```

According to the previous algorithm, the processing stage begins with the biased support vector machine process to determine SV Sets and NSV Sets for the majority and minority classes. The next step is for each SV Set and NSV Set in the majority class to go through the process of noise removal and resampling using the MC-CCR. On minority classes, the same thing will be done with SV sets and NSV sets. The SMOTEBoost process will be applied to SV sets in the minority class in particular. This whole process will result in a result dataset.

4. RESULTS AND ANALYSIS

4.1. Dataset description

We conducted our experiments using 6 (six) multi-class imbalanced datasets from the knowledge extraction based on evolutionary learning (KEEL) repository, each with a low, moderate, or high. For datasets with a low IR are new-thyroid and balance, datasets with moderate infrared (IR) are flare and car, and dataset with high IR are red wine quality and yeast. Table 3 contains a description of the dataset [35].

Table 3. Description of dataset [35]

Dataset	#Ex	#Atts	Distribution of Class	IR
New-Thyroid	215	5	150/35/30	5
Balance	625	4	288/49/288	5.88
Flare	1066	11	331/239/211/147/95/43	7.7
Car	1728	6	65/69/384/1210	18.61
Red Wine Quality	1599	11	10/53/681/638/199/18	68.1
Yeast	1484	8	463/5/35/44/51/163/244/429/20/30	92.6

Following the selection of the dataset, the next step is to assess the presence of noise. This experiment will use a subset of training examples and randomly replace their labels to generate noise. This experiment will use a noise level of 0.1.

4.2. Testing result

The first test compares the augmented R-value and class average accuracy obtained by using the HAR-MI with resampling algorithm and feature selection. Table 4 shows the test results. According to Table 4, the results obtained by the HAR-MI method with resampling algorithm and feature selection and neighborhood-based undersampling are not significantly different in terms of overlapping. This is indicated by the value of augmented R-value which is not much different. The lower the augmented R-value, the lower the overlapping level. There is a strong relationship between overlapping and accuracy. The lower the overlapping, the better the average class accuracy obtained. It should also be noted that neighborhood-based undersampling tends to have a slight advantage in datasets with a large number of attributes such as flare and red wine quality. The HAR-MI with resampling algorithm and feature selection has the advantage of 4 other datasets. It should be noted that the imbalance ratio has an impact on the results. The higher the imbalance ratio, the more overlapping there will be, and the accuracy obtained will also be lower.

The second test compares the class balance accuracy, multi class G-mean, and confusion entropy obtained by using the HAR-MI method with resampling algorithm and feature selection, as well as neighborhood-based undersampling. Table 5 shows the test results. Based on the Table 4, it is obvious that the number of attributes, the number of classes, and the level of IR all have a significant impact on class balance accuracy. The number of attributes and classes will largely determine the results of class balance accuracy for datasets with similar imbalance ratio levels. This can be seen in the dataset balance results for improved class balance accuracy when compared to the New-Thyroid dataset. When it comes to class balance accuracy, it can be seen that the HAR-MI with resampling algorithm and feature selection method produces better results than neighborhood-based undersampling.

Test results for multi class G-mean show that for both HAR-MI with resampling algorithm and feature selection and neighborhood-based undersampling, the higher the IR, the lower the multi class G-mean value obtained because G-means stated the equilibrium between positive samples and negative samples. The test results for confusion entropy show that the results obtained depend on the number of classes and imbalance ratios. The number of classes determines the results obtained for imbalance ratios that are not significantly different, such as those in the flare and car datasets. In general, the results obtained by the two methods for confusion entropy are not significantly different.

Table 4. Augmented R-value and class average accuracy testing results

Dataset	HAR-MI with Resampling Algorithm and Feature Selection		Neighbourhood-Based Undersampling	
	Augmented R-Value	Class Average Accuracy	Augmented R-Value	Class Average Accuracy
New-Thyroid	0.335	0.972	0.341	0.933
Balance	0.327	0.865	0.345	0.891
Flare	0.367	0.713	0.359	0.721
Car	0.361	0.725	0.373	0.693
Red Wine Quality	0.428	0.687	0.415	0.674
Yeast	0.448	0.623	0.452	0.615

Table 5. Class balance accuracy, multi class G-mean, and confusion entropy testing results for each method

Dataset	HAR-MI with Resampling Algorithm and Feature Selection			Neighbourhood-Based Undersampling		
	Class Balance Accuracy	Multi Class G-Mean	Confusion Entropy	Class Balance Accuracy	Multi Class G-Mean	Confusion Entropy
New-Thyroid	0.913	0.898	0.091	0.897	0.875	0.11
Balance	0.927	0.915	0.105	0.911	0.897	0.12
Flare	0.691	0.715	0.292	0.676	0.687	0.282
Car	0.897	0.875	0.311	0.874	0.815	0.327
Red Wine Quality	0.516	0.467	0.472	0.498	0.465	0.493
Yeast	0.495	0.512	0.526	0.476	0.498	0.523

4.3. Statistical tests

The Wilcoxon signed-rank test is used to perform the statistical test, which is a statistical procedure in order to assess perfromance on the basis of pairwise comparisons [36]. The result for statistical tests can be seen in Table 6.

Table 6. Wilcoxon signed-rank test in order to assess performance

Performance Measurement	P-Value	Hypothesis
Augmented R-Value	0.687500	*H_0 (no significant difference in score between HAR-MI with Resampling Algorithm, Feature Selection, and Neighbourhood-Based Undersampling) is accepted, which means H_1 (significant difference in score between HAR-MI with Resampling Algorithm, Feature Selection, and Neighbourhood-Based Undersampling) is rejected because the p-value is greater than 0.05)*
Class Average Accuracy	0.344118	*H_0 (no significant difference in score between HAR-MI with Resampling Algorithm, Feature Selection, and Neighbourhood-Based Undersampling) is accepted, which means H_1 (significant difference in score between HAR-MI with Resampling Algorithm, Feature Selection, and Neighbourhood-Based Undersampling) is rejected because the p-value is greater than 0.05)*
Class Balance Accuracy	0.03552	*H_0 (no significant score difference between HAR-MI with Resampling Algorithm and Feature Selection and Neighbourhood-Based Undersampling) rejected, which means H_1 (there is a significant difference between HAR-MI with Resampling Algorithm and Feature Selection and Neighbourhood-Based Undersampling in score) Accepted because the p-value is less than 0.05*
Multi Class G-Mean	0.0312500	*H_0 (no significant score difference between HAR-MI with Resampling Algorithm and Feature Selection and Neighbourhood-Based Undersampling) rejected, which means H_1 (there is a significant difference between HAR-MI with Resampling Algorithm and Feature Selection and Neighbourhood-Based Undersampling in score) Accepted because the p-value is less than 0.05*
Confusion Entropy	0.156250	*H_0 (no significant difference in score between HAR-MI with Resampling Algorithm, Feature Selection, and Neighbourhood-Based Undersampling) is accepted, which means H_1 (significant difference in score between HAR-MI with Resampling Algorithm, Feature Selection, and Neighbourhood-Based Undersampling) is rejected because the p-value is greater than 0.05)*

4.4. Discussion

According to the results in Tables 4-6, there is no significant distintion in augmented R-value, class average accuracy, and confusion entropy between HAR-MI with resampling algorithms and feature selection and neighborhood-based undersampling. It indicates that both methods have successfully overcome overlapping with positive outcomes. The confusion entropy obtained is good and this means that the misclassification is spread evenly for all classes. The test results for class balance accuracy and multi class G-mean show that HAR-MI with resampling algorithms and feature selection gives better results compared to neighborhood-based undersampling.

It should be noted that overlapping and accuracy are two interrelated things, the higher the overlapping, the lower the accuracy. The imbalance ratio is the main factor that determines how much overlap there is. The higher the imbalance ratio, the higher the overlapping will be. In terms of overlapping, HAR-MI with resampling algorithms and feature selection has few limitations on datasets with a large number of attributes, in addition to imbalance ratios. The results revealed that, in addition to IR, the number of attributes and the number of classes determined the value of class balance accuracy. The number of classes and the imbalance ratio have a strong influence on multi class G-mean and confusion entropy.

5. CONCLUSION

Based on the test results that for handling multi-class imbalances accompanied by overlapping and noise, the results obtained by HAR-MI with resampling algorithms and feature selection and neighborhood-based undersampling are good. The results obtained by HAR-MI with resampling algorithms and feature selection are generally better than neighborhood-based undersampling. This is indicated by better augmented R-value, class average accuracy, class balance accuracy, multi-class G-mean, and confusion entropy. Although statistically for augmented R-value, class average accuracy, and confusion entropy based on the test results statistically it does not have too significant differences. It should be noted that HAR-MI with resampling algorithms and feature selection and neighborhood-based undersampling has limitations in handling overlapping, where there is a slight decrease in performance in datasets with large numbers of attributes. Imbalance ratio also has a direct relationship with the performance classifier. Future research is expected to be able to handle the decrease in performance in datasets with large number of attributes and also a high IR.

ACKNOWLEDGEMENTS

This study was funded by a fund of the Republic of Indonesia's Ministries of Education and Culture, as well as the Ministry of Research and Technology.

REFERENCES

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

[2] N. Santoso, W. Wibowo, and H. Hikmawati, “Integration of synthetic minority oversampling technique for imbalanced class,” *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 13, no. 1, no. 1, Jan. 2019, doi: 10.11591/ijeecs.v13.i1.pp102-108.

[3] J. Hamidzadeh, N. Kashefi, and M. Moradi, “Combined weighted multi-objective optimizer for instance reduction in two-class imbalanced data problem,” *Engineering Applications of Artificial Intelligence*, vol. 90, p. 103500, Apr. 2020, doi: 10.1016/j.engappai.2020.103500.

[4] C.-M. Vong and J. Du, “Accurate and efficient sequential ensemble learning for highly imbalanced multi-class data,” *Neural Networks*, vol. 128, pp. 268-278, Aug. 2020, doi: 10.1016/j.neunet.2020.05.010.

[5] G.-H. Fu, Y.-J. Wu, M.-J. Zong, and L.-Z. Yi, “Feature selection and classification by minimizing overlap degree for class-imbalanced data in metabolomics,” *Chemometrics and Intelligent Laboratory Systems*, vol. 196, p. 103906, Jan. 2020, doi: 10.1016/j.chemolab.2019.103906.

[6] E. R. Q. Fernandes and A. C. P. L. F. de Carvalho, “Evolutionary inversion of class distribution in overlapping areas for multi-class imbalanced learning,” *Information Sciences*, vol. 494, pp. 141-154, Aug. 2019, doi: 10.1016/j.ins.2019.04.052.

[7] A. Fernández, M. J. del Jesus, and F. Herrera, “Addressing Overlapping in Classification with Imbalanced Datasets: A First Multi-Objective Approach for Feature and Instance Selection,” in *Intelligent Data Engineering and Automated Learning - IDEAL 2015*, Cham, 2015, pp. 36-44, doi: 10.1007/978-3-319-24834-9_5.

[8] A. Wahid *et al.*, “Feature selection and classification for gene expression data using novel correlation based overlapping score method via Chou’s 5-steps rule,” *Chemometrics and Intelligent Laboratory Systems*, vol. 199, p. 103958, Apr. 2020, doi: 10.1016/j.chemolab.2020.103958.

[9] B. Frénay, G. Doquire, and M. Verleysen, “Estimating mutual information for feature selection in the presence of label noise,” *Computational Statistics & Data Analysis*, vol. 71, pp. 832-848, Mar. 2014, doi: 10.1016/j.csda.2013.05.001.

[10] S. Uyun and E. Sulistyowati, “Feature selection for multiple water quality status: Integrated bootstrapping and SMOTE approach in imbalance classes,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 4, no. 4, Aug. 2020, doi: 10.11591/ijece.v10i4.pp4331-4339.

[11] M. Koziarski, M. Woźniak, and B. Krawczyk, “Combined Cleaning and Resampling algorithm for multi-class imbalanced data with label noise,” *Knowledge-Based Systems*, vol. 204, p. 106223, Sep. 2020, doi: 10.1016/j.knosys.2020.106223.

[12] M. Alibeigi, S. Hashemi, and A. Hamzeh, “DBFS: An effective Density Based Feature Selection scheme for small sample size and high dimensional imbalanced data sets,” *Data & Knowledge Engineering*, vol. 81-82, pp. 67-103, Nov. 2012, doi: 10.1016/j.datak.2012.08.001.

[13] H. Chen, T. Li, X. Fan, and C. Luo, “Feature selection for imbalanced data based on neighborhood rough sets,” *Information Sciences*, vol. 483, pp. 1-20, May 2019, doi: 10.1016/j.ins.2019.01.041.

[14] H. Ali, M. N. M. Salleh, R. Saedudin, K. Hussain, and M. F. Mushtaq, “Imbalance class problems in data mining: a review,” *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 14, no. 3, no. 3, Jun. 2019, doi: 10.11591/ijeecs.v14.i3.pp1552-1563.

[15] R. Alejo, R. M. Valdovinos, V. García, and J. H. Pacheco-Sanchez, “A hybrid method to face class overlap and class imbalance on neural networks and multi-class scenarios,” *Pattern Recognition Letters*, vol. 34, no. 4, pp. 380-388, Mar. 2013, doi: 10.1016/j.patrec.2012.09.003.

[16] W. Xie, G. Liang, Z. Dong, B. Tan, and B. Zhang, “An Improved Oversampling Algorithm Based on the Samples’ Selection Strategy for Classifying Imbalanced Data,” *Mathematical Problems in Engineering*, pp. 1-13, 2019, doi: 10.1155/2019/3526539.

[17] F. N. Koutanaei, H. Sajedi, and M. Khanbabaei, “A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring,” *Journal of Retailing and Consumer Services*, vol. 27, pp. 11-23, Nov. 2015, doi: 10.1016/j.jretconser.2015.07.003.

[18] P. Vuttipittayamongkol and E. Elyan, “Neighbourhood-based undersampling approach for handling imbalanced and overlapped data,” *Information Sciences*, vol. 509, pp. 47-70, Jan. 2020, doi: 10.1016/j.ins.2019.08.062.

[19] W. A. Rivera, “Noise Reduction A Priori Synthetic Over-Sampling for class imbalanced data sets,” *Information Sciences*, vol. 408, pp. 146-161, Oct. 2017, doi: 10.1016/j.ins.2017.04.046.

[20] M. Koziarski, B. Krawczyk, and M. Woźniak, “Radial-Based oversampling for noisy imbalanced data classification,” *Neurocomputing*, vol. 343, pp. 19-33, May 2019, doi: 10.1016/j.neucom.2018.04.089.

[21] H. Hartono, Y. Risyani, E. Ongko, and D. Abdullah, “HAR-MI method for multi-class imbalanced datasets,” *TELKOMNIKA Telecommunication Computing Electronics and Control*, vol. 18, no. 2, no. 2, Apr. 2020, doi: 10.12928/telkomnika.v18i2.14818.

[22] R. Tibshirani, “Regression Shrinkage and Selection via the Lasso,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267-288, 1996, doi: 10.1111/j.2517-6161.1996.tb02080.x.

[23] J. Fan and R. Li, “Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties,” *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348-1360, 2001, doi: 10.1198/016214501753382273.

[24] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: synthetic minority over-sampling technique,” *Journal of artificial intelligence research*, vol. 16, pp. 321-357, 2002, doi: 10.1613/jair.953.

[25] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, “A study of the behavior of several methods for balancing machine learning training data,” *SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 20-29, Jun. 2004, doi: 10.1145/1007730.1007735.

[26] D. Angluin and P. Laird, “Learning from Noisy Examples,” *Machine Learning*, vol. 2, no. 4, pp. 343-370, Apr. 1988, doi: 10.1023/A:1022873112823.

[27] S. Oh, “A new dataset evaluation method based on category overlap,” *Computers in Biology and Medicine*, vol. 41, no. 2, pp. 115-122, Feb. 2011, doi: 10.1016/j.compbiomed.2010.12.006.

[28] A. Ali, S. M. Shamsuddin, and A. L. Ralescu, “Classification with class imbalance problem: A Review,” *International Journal of Advances in Soft Computing and Its Application*, vol. 7, no. 3, pp. 176-204, 2015.

[29] P. Branco, L. Torgo, and R. P. Ribeiro, “Relevance-Based Evaluation Metrics for Multi-class Imbalanced Domains,” in *Advances in Knowledge Discovery and Data Mining*, Cham, 2017, pp. 698-710, doi: 10.1007/978-3-319-57454-7_54.

[30] L. Mosley, “A balanced approach to the multi-class imbalance problem,” *Graduate Theses and Dissertations*, Jan. 2013, doi: 10.31274/etd-180810-3375.

[31] M. Kubat, R. C. Holte, and S. Matwin, “Machine Learning for the Detection of Oil Spills in Satellite Radar Images,” *Machine Learning*, vol. 30, no. 2, pp. 195-215, Feb. 1998, doi: 10.1023/A:1007452223027.

[32] Y. Sun, M. S. Kamel, and Y. Wang, “Boosting for Learning Multiple Classes with Imbalanced Class Distribution,” in *Sixth International Conference on Data Mining (ICDM’06)*, Dec. 2006, pp. 592-602, doi: 10.1109/ICDM.2006.29.

[33] J.-M. Wei, X.-J. Yuan, Q.-H. Hu, and S.-Q. Wang, “A novel measure for evaluating classifiers,” *Expert Systems with Applications*, vol. 37, no. 5, pp. 3799-3809, May 2010, doi: 10.1016/j.eswa.2009.11.040.

[34] H. Zou and T. Hastie, “Regularization and Variable Selection via the Elastic Net,” *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301-320, 2005.

[35] J. Alcalá-Fdez, *et al.*, “KEEL: a software tool to assess evolutionary algorithms for data mining problems,” *Soft Comput*, vol. 13, no. 3, pp. 307-318, Feb. 2009, doi: 10.1007/s00500-008-0323-y.

[36] F. Wilcoxon, “Individual Comparisons by Ranking Methods on JSTOR,” *Biometrics Bulletin*, vol. 1, no. 6, pp. 80-83, 1945.

BIOGRAPHIES OF AUTHORS

Erianto Ongko earned his Master's degree (2015) from Universitas Sumatera Utara and Bachelor’s degree (2012) from STMIK IBBI. Both of them are in Computer Science. He serve as a lecturer in Akademi Teknologi Industri Immanuel in Medan, Indonesia. His research topic are in Machine Learning, AI, and Operational Research.

Hartono earned his Doctoral’s degree (2018) from Universitas Sumatera Utara, Master's degree (2010) from Universitas Putra Indonesia YPTK Padang, and Bachelor’s degree (2008) from STMIK IBBI. All of them are in Computer Science. He serve as a lecturer in Universitas IBBI and Universitas Potensi Utama in Medan, Indonesia. His research topic are in Machine Learning, AI, and Operational Research.