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

Erianto Ongko1, Hartono2
1Department of Informatics, Akademi Teknologi Industri Immanuel, Indonesia
2Department of Computer Science, Universitas IBBI, Indonesia
3Department of Computer Science, Universitas Potensi Utama, Indonesia

Article Info

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.

Keywords:
Class imbalance
Classifier performance
Multi-class
Noise
Overlapping

This is an open access article under the CC BY-SA license.

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

Journal homepage: http://beei.org
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. Kozierski 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 (RSFSIAID) 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-multiclass (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.
\[ R_{\text{Aug}}(D[V]) = \frac{\sum_{i=0}^{k-1} |G_i| R(G_i)}{\sum_{i=0}^{k-1} |C_i|} \]  

(1)

Where \( C_0, C_1, ..., C_{k-1} \) are \( k \) class labels with \( |C_0| \geq |C_1| \geq \ldots \geq |C_{k-1}| \) and \( D[V] \): Dataset D restraining predictors in set \( V \). A higher \( R_{\text{Aug}} \) 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>
<thead>
<tr>
<th>Actually (Really is)</th>
<th>Predicted (Classified) as Positive</th>
<th>Class Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive Samples</td>
<td>TP</td>
<td>FN</td>
</tr>
<tr>
<td>Negative Samples</td>
<td>FP</td>
<td>TN</td>
</tr>
</tbody>
</table>

### Table 1. Confusion matrix for a classification problem

#### 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].

\[
\text{AvAcc} = \frac{1}{C} \sum_{i=1}^{C} \frac{tp_i + tn_i}{tp_i + tn_i + fp_i + fn_i} \]

(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 \( C \) classes, confusion matrix, class balance accuracy is defined as [30].

\[
\text{CBA} = \frac{\sum_{i=1}^{C} c_{ii}}{\max_{i,j} c_{ij}} \]

(3)

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

<table>
<thead>
<tr>
<th>Class 1</th>
<th>Class 2</th>
<th>Class 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class 1</td>
<td>C_{11}</td>
<td>C_{12}</td>
</tr>
<tr>
<td>Class 2</td>
<td>C_{21}</td>
<td>C_{22}</td>
</tr>
<tr>
<td>Class 3</td>
<td>C_{31}</td>
<td>C_{32}</td>
</tr>
<tr>
<td>Total</td>
<td>C_{1}</td>
<td>C_{2}</td>
</tr>
</tbody>
</table>

- 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.

\[
R_i = \frac{c_{ii}}{\sum_{j=1}^{C} c_{ij}} \]

(4)

\[
mGM = \left( \prod_{i=1}^{C} R_i \right)^{1/N} \]

(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 \( cl_i \).

\[
CEN = \sum_{j} p_j CEN_j \]

(6)
where
\[
P_j = \frac{\sum_k (G_{j,k}^l C_{j,k})}{\sum_k C_{j,k}}
\]
(7)
\[
CEN_j = -\sum_{k=1}^{N+1} x_j \frac{p^j_{k,j}}{b_2Np^j_{k,j}}
\]
(8)

2.4. Minimizing overlapping selection under SMOTE

The MOSS algorithm is being as [5].

1: X - matic with p predictors: X = [x_1, x_2, ..., x_p]; class label: y
2: 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
3: X ← X_new; Y ← Y_new
4: Establish sparse regularization path \( \tilde{\beta}(\lambda, \alpha) \) according to (9)
5: Compute the optimal \( (\tilde{\beta}_1, \tilde{\beta}_2, ..., \tilde{\beta}_p)^T \) via the (10)
6: Select those feature with \( \tilde{\beta}_j \neq 0 \) for \( j = 1, 2, ..., p \)

In (9) shows the sparse collection that would be used to create sparse regularization [34].
\[
C_a(\beta) = \frac{1}{2} (1 - \alpha) \| \beta \|_2^2 + a \| \beta \|_1
\]
(9)

In (10) shows the loss penalties that will be used to determine the optimum \( \tilde{\beta}_j \):
\[
Loss = -\frac{1}{n} \sum_{i=1}^{n} (y_i \beta^T x_i - \ln (1 + \beta^T x_i))
\]
(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: \( x^{(i)} \) 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 \( X \)

1: Function MC-CCR (X, energi, p);
2: C= Collection of all classes; sorted by the number of associated observations in a descending order
3: for \( i \leftarrow 1 \) to \( |c| \)
4: \( n_{classes} \leftarrow \) number of classes with high number of observations than \( C_i \)
5: if \( n_{classes} > 0 \) then
6: \( X_{min} \leftarrow x^{(i)} \)
7: \( X_{maj} \leftarrow \emptyset \)
8: for \( j \leftarrow 1 \) to \( n_{classes} \)
9: add \( \frac{1}{n_{classes}} \) randomly selected from \( x^{(i)} \) to \( x_{maj} \)
10: end
11: \( X_{maj} \cdot S \leftarrow CCR (X_{maj}, X_{min}, energy, P) \)
12: \( X^{(i)} \leftarrow X^{(i)} \cup S \)
13: Substitute observation used to construct \( X_{maj} \) with \( X_{maj} \)
14: end
15: end
16: 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].

Hybrid approach redefinition-multi class with resampling and feature selection for... (Erianto Ongko)
Require: Set S of examples \((x_i, y_i)\)
Ensure: New set \(S'\) of examples with Random Balance Ensemble Method
1: \(\text{totalSize} \leftarrow |S|\)
2: \(\text{Determine } k \text{ using Dynamic Ensemble Selection}\)
3: \(\text{Building the candidate ensemble for Safe, Borderline, Rare, and Outlier}\)
4: \(\text{For all samples in Majority and Minority}\)
5: \(\text{Preprocessing Satge using Random Balance Ensemble Method}\)
6: \(\text{New Majority and New Minority of Preprocessing Dataset}\)
7: \(\text{End}\)
8: \(\text{Determine the Augmented R-Value}\)
9: \(\text{For all instances in Preprocessing Dataset}\)
10: \(\text{Determine Majority and Minority Class}\)
11: \(\text{For All Instances in Majority Class}\)
12: \(\text{Biased Support Vector Machine for Determine SV Sets and NSV Sets}\)
13: \(\text{End}\)
14: \(\text{For All Instances in Minority Class}\)
15: \(\text{Biased Support Vector Machine for Determine SV Sets and NSV Sets}\)
16: \(\text{End}\)
17: \(\text{End}\)
18: \(\text{For All Instances in NSV Sets from Majority Class}\)
19: \(\text{Process Multiple Random Under Sampling}\)
20: \(\text{End}\)
21: \(\text{For All Instances in SV Sets from Minority Class}\)
22: \(\text{Process SMOTEBoost}\)
23: \(\text{End}\)

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 \((X, Y)\)
Ensure: \(S\) is the new Set with MOSS
1: \(\text{totalSize} \leftarrow |S|\)
2: \(\text{Calculate } k \text{ as the number of Nearest Neighbors}\)
3: \(\text{For All Instances of } S\)
4: \(\text{Building the Positive or Minority Class Borderline as } E_0C_t^+\)
5: \(\text{Building the Negative or Majority Class Borderline as } E_0C_t^-\)
6: \(\text{End}\)
7: \(\text{Creating a candidate ensemble based on } k \text{ value for safe, borderline, rare, and outlier candidates}\)
8: \(\text{Using Equation 9 to Calculate Sparse Selection}\)
9: \(\text{Using Equation 10 to Calculate the Loss Penalty}\)
10: \(\text{Determine Sparse Regularization}\)
11: \(\text{For All Instances in Positive}\)
12: \(\text{MOSS is used to sample all instances}\)
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.

1: For all samples in preprocessed dataset
2: Preprocessed Dataset should be added to $S_i$
3: Using B-SVM, determine the SV and NSV sets for the majority and minority
4: For All Samps in Negative
5: Checking and removing noise from SV and NSV sets with MC - CCR
6: End
7: For All Samples in Positive
8: Check and remove the Noise in SV Sets and NSV Sets using MC - CCR

Figure 1. Stages of research methods
9: SMOTEBoost Step for SV Sets and Produce SMOTESets
10: End For
11: For All SV Sets and NSV Sets from Majority Class do
12: New NegativeSampleSets
13: End For
14: For All SV Sets and NSV Sets from Minority Class do
15: New PositiveSampleSets
16: End For
17: End

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

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>
<thead>
<tr>
<th>Dataset</th>
<th>HAR-MI with Resampling Algorithm and Feature Selection</th>
<th>Neighbourhood-Based Undersampling</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Augmented R-Value</td>
<td>Class Average Accuracy</td>
</tr>
<tr>
<td>New-Thyroid</td>
<td>0.335</td>
<td>0.972</td>
</tr>
<tr>
<td>Balance</td>
<td>0.327</td>
<td>0.865</td>
</tr>
<tr>
<td>Flare</td>
<td>0.367</td>
<td>0.713</td>
</tr>
<tr>
<td>Car</td>
<td>0.361</td>
<td>0.725</td>
</tr>
<tr>
<td>Red Wine Quality</td>
<td>0.428</td>
<td>0.687</td>
</tr>
<tr>
<td>Yeast</td>
<td>0.448</td>
<td>0.623</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>HAR-MI with Resampling Algorithm and Feature Selection</th>
<th>Neighbourhood-Based Undersampling</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Class Balance Accuracy</td>
<td>Multi Class G-Mean</td>
</tr>
<tr>
<td>New-Thyroid</td>
<td>0.913</td>
<td>0.898</td>
</tr>
<tr>
<td>Balance</td>
<td>0.927</td>
<td>0.915</td>
</tr>
<tr>
<td>Flare</td>
<td>0.691</td>
<td>0.715</td>
</tr>
<tr>
<td>Car</td>
<td>0.897</td>
<td>0.875</td>
</tr>
<tr>
<td>Red Wine Quality</td>
<td>0.516</td>
<td>0.467</td>
</tr>
<tr>
<td>Yeast</td>
<td>0.495</td>
<td>0.512</td>
</tr>
</tbody>
</table>

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 performance on the basis of pairwise comparisons [36]. The result for statistical tests can be seen in Table 6.

<table>
<thead>
<tr>
<th>Performance Measurement</th>
<th>P-Value</th>
<th>Hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Augmented R-Value</td>
<td>0.6875</td>
<td>H0 (no significant difference in score between HAR-MI with Resampling Algorithm, Feature Selection, and Neighbourhood-Based Undersampling) is accepted, which means H1 (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)</td>
</tr>
<tr>
<td>Class Average Accuracy</td>
<td>0.3441</td>
<td>H0 (no significant difference in score between HAR-MI with Resampling Algorithm, Feature Selection, and Neighbourhood-Based Undersampling) is accepted, which means H1 (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)</td>
</tr>
<tr>
<td>Class Balance Accuracy</td>
<td>0.0355</td>
<td>H0 (no significant score 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</td>
</tr>
<tr>
<td>Multi Class G-Mean</td>
<td>0.0312</td>
<td>H0 (no significant score 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</td>
</tr>
<tr>
<td>Confusion Entropy</td>
<td>0.1562</td>
<td>H0 (no significant difference in score between HAR-MI with Resampling Algorithm, Feature Selection, and Neighbourhood-Based Undersampling) is accepted, which means H1 (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)</td>
</tr>
</tbody>
</table>

4.4. Discussion

According to the results in Tables 4-6, there is no significant distinction 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


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.