HAR-MI with COSTE in Handling Multi-Class Imbalance

1st Hartono
Department of Engineering and Computer Science
Universitas Potensi Utama
Medan, Indonesia
hartonoibbi@gmail.com

2nd Silvia Lestari
Department of Engineering and Computer Science
Universitas Potensi Utama
Medan, Indonesia
ulandari2796@gmail.com

3rd Andi Rahmadsyah
Department of Engineering and Computer Science
Universitas Potensi Utama
Medan, Indonesia
andijohorr@gmail.com

4th Ridha Maya Faza Lubis
Department of Engineering and Computer Science
Universitas Potensi Utama
Medan, Indonesia
ridhamayafazalubis@gmail.com

5th Muhammad Gunawan
Department of Engineering and Computer Science
Universitas Potensi Utama
Medan, Indonesia
gunawan6451kz@gmail.com

***Abstract*—The class imbalance problem is a serious problem in machine learning. This problem can occur in two-class and multi-class problems. This problem can result in low accuracy and not obtaining information regarding the minority class. Approaches to overcome this problem often use a combination of the Data-Level Approach and Algorithm-Level Approach, which is often referred to as a Hybrid Approach. One of the Hybrid Approach methods to solve the multi-class imbalance problem is the Hybrid Approach Redefinition-Multiclass Imbalance (HAR-MI). The sampling method used in the HAR-MI is the Oversampling method. Oversampling tends to be chosen because Undersampling can eliminate useful information, but Oversampling often causes Over-Fitting. Therefore, this study will modify the HAR-MI approach by using the Complexity-based OverSampling TEchnique (COSTE). COSTE will rank the instances based on the measurement of complexity so that it will provide better results and prevent Over-Fitting. The results showed that HAR-MI with COSTE gave better results than the classic HAR-MI.**

Keywords—Class Imbalance, Hybrid Approach, HAR-MI, Over-Fitting, COSTE

I. INTRODUCTION

The problem of class imbalance is a problem that is difficult to handle, even when using deep learning[1]. Class imbalance is manifested in the number of instances or samples that are unequal between majority and minority classes, and this often causes classification algorithms in machine learning to tend to provide more accurate results for majority classes. And ignore minority classes[2]. Most of the class imbalance problems are binary-class, although the problem sometimes is the multi-class problem[3].

Multi-class imbalance problems tend to be more difficult to handle than binary-class problems[4]. However, various approaches have been put forward to use the principle of decomposition in breaking the multi-class problem into a binary-class problem[5]. One approach that can be used is Dynamic Ensemble Selection[6].

There are several approaches in overcoming class imbalance problems, namely the data-level approach, algorithm-level approach, and cost-sensitive approach[7]. Hybrid Approach is an approach that combines using the Data-Level Approach with the Algorithm-Level Approach[8]. One of the Hybrid Approaches that is effective in overcoming the problem of multi-class imbalance is the Hybrid Approach Redefinition-Multiclass Imbalance (HAR-MI)[9].

The HAR-MI method generally uses the Oversampling method[19]. Oversampling tends to be chosen because Undersampling can eliminate important information[10]. But sometimes, oversampling can lead to over-fitting[11]. This condition led to thinking about the importance of ranking each existing instance or classifier using the Complexity measurement. This is what underlies the Complexity-based OverSampling TEchnique (COSTE) method[12].

This study will use the COSTE approach to replace the SMOTE in HAR-MI, which is used in the preprocessing and processing stages[21]. The preprocessing stage will use Dynamic Ensemble Selection[13], and COSTE and the processing stage will use the Different Contribution Sampling and COSTE[14].

II. RELATED WORKS

A. Hybrid Approach

The pseudocode of the Hybrid Approach is as follows[15].

Input: $D_t = \{x_1, x_2, \ldots, x_n\}$ //Training Dataset
N = Number of Classifier
Output: Classification Prediction P
Method:
Step 1 Preprocessing using Preprocessing Method
Step 2 For $i = 1$ to N do
  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 Preprocessing using Algorithm – Level Approach and Data – Level Approach

  Apply Processing using Algorithm – Level Approach and Data – Level Approach

End For

Based on pseudocode, it can be seen that the stages in the Hybrid Approach are generally divided into 2 (two), namely preprocessing and processing[20]. The preprocessing and processing stages will be carried out using the Data-Level Approach and Algorithm-Level Approach.

B. Dynamic Ensemble Selection

Dynamic Ensemble Selection is used to decompose multi-class problems into binary-class problems. The pseudocode of Dynamic Ensemble Selection is as follows[16].

```
Input:
S {Input Dataset}
K {Neighborhood Size}
Output:
Safe {Set of Safe Samples}
Borderline {Set of Borderline Samples}
Rare {Set of Rare Samples}
Outlier {Set of Outlier Samples}
For All Zi ϵ S do
    Neighbors ← computeNeighbors (Zi, S – {Zi}, K)
    sameClass ← countSameClass (yi, neighbors)
    if sameClass ≥ ⌊0.8k⌋ then
        safe ← safe ∪ {Zi}
    else
        if sameClass ≥ ⌊0.5k⌋ then
            borderline ← borderline ∪ {Zi}
        else
            if sameClass ≥ ⌊0.2k⌋ then
                rare ← rare ∪ {Zi}
            else
                Outlier ← Outlier ∪ {Zi}
            end if
        end if
    end if
end for
```

Based on the pseudocode, it can be seen that Dynamic Ensemble Selection will be used to decompose multi-class problems into binary-class problems. There are 4 classes that will be generated, namely, Safe, Borderline, Rare, and Outlier.

C. COSTE

The pseudocode from COSTE is as follows[12].

```
Input: Dataset N including the minority class instances
N_min and the
majority class instances N_maj
Output: Balanced Dataset N_bal
Array_syn ←
array for storing new synthetic instances

apply the min –
max normalization method to the dataset

for each instance X_i Calculate complexity using
Equation 1
        Complexity_i = Σ_{j=1}^{f} α_j x_j                (1)
rank X_i in the ascending order based on complexity

Calculate the number of new synthetic instances
needed T, T =
number(N_maj) − number(N_min)
if T > number(N_min) − 1 then
    for i = 1,2,…, number(N_min) − 1 do
        new synthetic instance X_new = (X_i + X_(i+1))/2
        add X_new into Array_syn
    end for
    update N_min by merging N_min and Array_syn
    repeat Line 5
else
    for i = 1,2,…, T − 1 do
        new synthetic instance X_new = (X_i + X_(i+1))/2
        add X_new into Array_syn
    end for
    update N_min by merging N_min and Array_syn
end if
return balanced dataset N_bal by merging N_min and N_maj
```

The COSTE method is basically based on calculating complexity from each instance. It is used for ranking each instance. A higher ranking means less complex, and the instance with the highest ranking will be selected to be the classifier.

D. Classifier Performance

The classifier performance is measured by a number of parameters as follows.

1. Confusion Matrix

The Confusion Matrix can be seen in Table 1[17].

TABLE I. CONFUSION MATRIX

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

2. Average Accuracy (**AvAcc**)

AvAcc can be calculated using the following equation[18].

$$TPR = \frac{TP}{TP+FN} \quad (2)$$

$$AvAcc = \frac{\sum_{i=1}^{n} TPR_i}{n} \quad (3)$$

3. Multi-Class G-Measure (**mGM**)

mGM can be calculated using the following equation[18]

$$Recall = TPR \quad (4)$$

$$mGM = \sqrt[n]{\prod_{i=1}^{n} recall_i} \quad (5)$$

III. PROPOSED METHOD

The research stages can be seen in Figure 1.

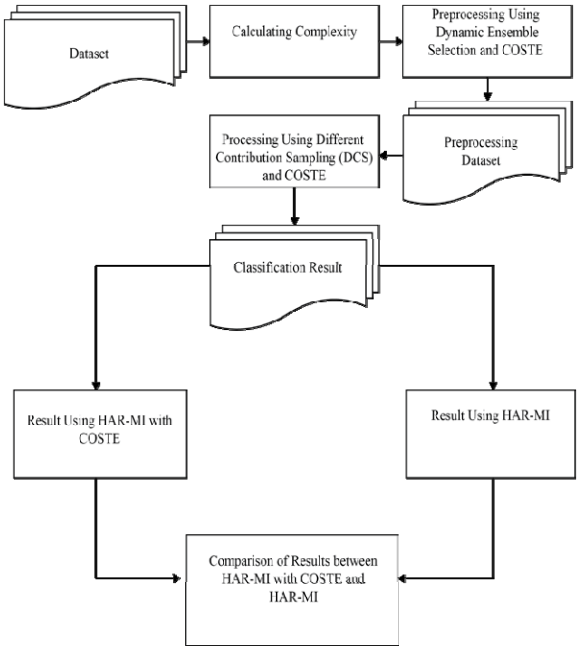

Fig. 1. Stage of Research Method

Based on Figure 1, it can be seen that the process begins with determining the dataset to be used. The dataset used is a multi-class imbalanced dataset. The first thing that will be done is the preprocessing stage, which is carried out using the Dynamic Ensemble Selection and COSTE. Dynamic Ensemble Selection is used to decompose multi-class problems into binary-class problems. This preprocessing stage will produce a preprocessed dataset. This preprocessed dataset will then undergo processing stages using Different Contribution Sampling (DCS) and COSTE. In this section, the Safe and Borderline classes will be grouped into majority class or negative samples, and rare and outlier classes will be grouped into minority class or positive samples. The results obtained will then be compared with the results obtained by the classic HAR-MI.

A. *Preprocessing using Dynamic Ensemble Selection and COSTE*

The pseudocode of the preprocessing stage is as follows.

Input: Dataset N including the minority class instances N_{min} and the
majority class instances N_{maj}
Output: Balanced Dataset N_{bal}

For All $X_i \in S$ do
Determine Safe, Borderline, Rare and Outlier Class
end for
for each instance X_i in Each class Calculate complexity

rank X_i in the ascending order based on Class
Calculate the number of new synthetic instances needed
$T, T = number(N_{maj}) - number(N_{min})$
if $T > number(N_{min}) - 1$ then
for $i = 1, 2, \ldots, number(N_{min}) - 1$ do
new synthetic instance $X_{new} = (X_i + X_{(i+1)})/2$
add X_{new} into $Array_{SYN}$
end for
update N_{min} by merging N_{min} and $Array_{SYN}$
repeat Line 5
else
for $i = 1, 2, \ldots, T - 1$ do
new synthetic instance $X_{new} = (X_i + X_{(i+1)})/2$
add X_{new} into $Array_{SYN}$
end for
update N_{min} by merging N_{min} and $Array_{SYN}$
end if
return balanced dataset N_{bal} by merging N_{min} and N_{maj}

B. *Processing using Different Contribution Sampling and COSTE*

The pseudocode of the processing stage is as follows.

Input: Preprocessed Dataset
Output: Result Dataset
Perform BSVM for Each Class
for All Instance in Safe and Borderline do
Deleting Noise in Sample SVSets
Perform COSTE for Sample in NSVSets
End for
for All Instance in Rare and Outlier do
Deleting Noise in Sample NSVSets
Perform COSTE for Sample in SVSets
End for
for All Instance in SVSets and NSV Sets from Safe and Borderline do
Create Negative Samples
End for
for All Instance in SVSets and NSV Sets from Rare and Outlier do
Create Positive Samples
End for
for All Instance in Negative and Positive Samples do
Create Result Dataset
End for

IV. RESULT AND ANALYSIS

A. *Dataset Description*

The dataset used in this study can be seen in Table II.

TABLE II. DATASET DESCRIPTION

Dataset	#Ex	#Features	Distribution of Class	IR
Cleveland	1728	13	164/55/36/35/13	12.62
Contraceptive	1473	9	629/333/511	1.89
Dermatology	358	33	111/60/71/48/48/20	5.55
Lympography	148	18	2/81/61/4	40.5
Vehicle	846	18	199/212/217/218	1.17
Wine	178	13	59/71/48	1.48

B. Testing Result

The test results can be seen in Table III.

TABLE III. TESTING RESULT

Dataset	HAR-MI with COSTE		HAR-MI	
	AvACC	mGM	AvACC	mGM
Cleveland	0.879	0.913	0.813	0.901
Contraceptive	0.913	0.892	0.921	0.903
Dermatology	0.817	0.821	0.856	0.837
Lympography	0.826	0.821	0.818	0.819
Vehicle	0.842	0.837	0.823	0.821
Wine	0.867	0.873	0.851	0.871

Based on Table III, it can be seen that the Average Accuracy given by HAR-MI with COSTE is better than HAR-MI. Only on the Contraceptive Dataset, the results provided by HAR-MI are better than HAR-MI with COSTE. In general, the number of Features and also the Imbalance Ratio are the factors that most influence the Average Accuracy. The same result is shown by mGM where the results given by HAR-MI with COSTE are better than that of HAR-MI.

C. Discussion

Based on the Average Accuracy (AvACC) test, it can be seen that the number of features and the imbalance ratio greatly influence the results obtained. In general, the results obtained by HAR-MI with COSTE are better than that of the classic HAR-MI. Only the Contraceptive Dataset shows that the classic HAR-MI is better than the HAR-MI with COSTE. The same result is shown by the multi-class G-Measure (mGM), which shows that HAR-MI with COSTE is better than HAR-MI.

V. RESULT AND ANALYSIS

Based on the research results, it can be seen that HAR-MI with COSTE gives better results when compared to classic HAR-MI both in terms of Average Accuracy (AvACC) and multi-class G-Mean (mGM) values. The results obtained are good, but there is a decrease in quality if the existing dataset has a large number of features or has a high imbalance ratio. Future Research should be able to improve the results obtained if there is a dataset with a large number of features and also a high imbalance ratio.

REFERENCES

[1] K. H. Kim and S. Y. Sohn, "Hybrid neural network with cost-sensitive support vector machine for class-imbalanced multimodal data," Neural Networks, vol. 130, pp. 176–184, Oct. 2020, doi: 10.1016/j.neunet.2020.06.026.

[2] A. Anil and S. R. Singh, "Effect of class imbalance in heterogeneous network embedding: An empirical study," Journal of Informetrics, vol. 14, no. 2, p. 101009, May 2020, doi: 10.1016/j.joi.2020.101009.

[3] J. Shin, S. Yoon, Y. Kim, T. Kim, B. Go, and Y. Cha, "Effects of class imbalance on resampling and ensemble learning for improved prediction of cyanobacteria blooms," Ecological Informatics, p. 101202, Nov. 2020, doi: 10.1016/j.ecoinf.2020.101202.

[4] Z.-L. Zhang, X.-G. Luo, S. González, S. García, and F. Herrera, "DRCW-ASEG: One-versus-One distance-based relative competence weighting with adaptive synthetic example generation for multi-class imbalanced datasets," Neurocomputing, vol. 285, pp. 176–187, Apr. 2018, doi: 10.1016/j.neucom.2018.01.039.

[5] M. Galar, A. Fernández, E. Barrenechea, H. Bustince, and F. Herrera, "An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes," Pattern Recognition, vol. 44, no. 8, pp. 1761–1776, Aug. 2011, doi: 10.1016/j.patcog.2011.01.017.

[6] S. García, Z.-L. Zhang, A. Altalhi, S. Alshomrani, and F. Herrera, "Dynamic ensemble selection for multi-class imbalanced datasets," Information Sciences, vol. 445–446, pp. 22–37, Jun. 2018, doi: 10.1016/j.ins.2018.03.002.

[7] A. Fernandez, S. Garcia, F. Herrera, and N. V. Chawla, "SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary," 1, vol. 61, pp. 863–905, Apr. 2018.

[8] F. Rayhan, S. Ahmed, A. Mahbub, 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.

[9] 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, Art. no. 2, Apr. 2020, doi: 10.12928/telkomnika. v18i2.14818.

[10] L. Zhou, "Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods," Knowledge-Based Systems, vol. 41, pp. 16–25, Mar. 2013, doi: 10.1016/j.knosys.2012.12.007.

[11] 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.

[12] S. Feng et al., "COSTE: Complexity-based OverSampling TEchnique to alleviate the class imbalance problem in software defect prediction," Information and Software Technology, vol. 129, p. 106432, Jan. 2021, doi: 10.1016/j.infsof.2020.106432.

[13] V. García, J. S. Sánchez, A. I. Marqués, R. Florencia, and G. Rivera, "Understanding the apparent superiority of over-sampling through an analysis of local information for class-imbalanced data," Expert Systems with Applications, vol. 158, p. 113026, Nov. 2020, doi: 10.1016/j.eswa.2019.113026.

[14] C. Jian, J. Gao, and Y. Ao, "A new sampling method for classifying imbalanced data based on support vector machine ensemble," Neurocomputing, vol. 193, pp. 115–122, Jun. 2016, doi: 10.1016/j.neucom.2016.02.006.

[15] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches," IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 42, no. 4, pp. 463–484, Jul. 2012, doi: 10.1109/TSMCC.2011.2161285.

[16] K. Napierala and J. Stefanowski, "Identification of Different Types of Minority Class Examples in Imbalanced Data," in Hybrid Artificial Intelligent Systems, Mar. 2012, pp. 139–150, doi: 10.1007/978-3-642-28931-6_14.

[17] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," Pattern Recognition, vol. 91, pp. 216–231, Jul. 2019, doi: 10.1016/j.patcog.2019.02.023.

[18] 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.

[19] GUNAWAN, Teddy Surya, et al. Development of control system for quadrotor unmanned aerial vehicle using LoRa wireless and GPS tracking. Telkomnika, 2020, 18.5: 2674-2681.

[20] ROSNELLY, Rika. Combination of Thresholding and Otsu Method in Increasing Results of Identification of Malaria Parasite Type in Thin Blood Smear Image. 2020.

[21] Wanayumini, W., & Harmayani, H. (2019, December). PEMODELAN DIRECT SELLING PRODUK DAN JASA HASIL KUBE DI KECAMATAN KISARAN TIMUR. In Seminar Nasional Multi Disiplin Ilmu Universitas Asahan.