Enhancing Classification Performance of Multi-Class Imbalanced Data Using the OAA-DB Algorithm

Piyasak Jeatrakul
School of Information Technology
Mae Fah Luang University
Chiang Rai, Thailand
piyasak.jea@mfu.ac.th

Kok Wai Wong
School of Information Technology
Murdoch University
Western Australia, Australia
k.wong@murdoch.edu.au

***Abstract*— In data classification, the problem of imbalanced class distribution has attracted many attentions. Most efforts have used to investigate the problem mainly for binary classification. However, research solutions for the imbalanced data on binary-class problems are not directly applicable to multi-class applications. Therefore, it is a challenge to handle the multi-class problem with imbalanced data in order to obtain satisfactory results. This problem can indirectly affect how human visualise the data. In this paper, an algorithm named One-Against-All with Data Balancing (OAA-DB) is developed to enhance the classification performance in the case of the multi-class imbalanced data. This algorithm is developed by combining the multi-binary classification technique called One-Against-All (OAA) and a data balancing technique. In the experiment, the three multi-class imbalanced data sets used were obtained from the University of California Irvine (UCI) machine learning repository. The results show that the OAA-DB algorithm can enhance the classification performance for the multi-class imbalanced data without reducing the overall classification accuracy.**

Keywords - multi-class imbalanced data; classification; artificial neural network; complementary neural network; misclassification analysis; OAA; SMOTE

I. INTRODUCTION

The classification problem has been examined for many years. This problem can basically be divided into two main categories, which are binary classification and multi-class classification problems. For a training set T with n training instances $T = \{(x_1, y_1), (x_2, y_2), \ldots. (x_n,y_n)\}$, where each instance is a member of a problem domain $x_i \in \mathrm{R}^m$ and a class label, $y_i \in \{c_1,c_2, \ldots c_K\}$, where $c_j \neq c_h$ for all $h \neq j$. The multi-class classification is a mapping function between instance X and class label Y where the number of K classes is greater than two, i.e. $f: X \rightarrow Y$, $K > 2$ Generally, the multi-class classification problem is more difficult to handle than the binary classification problem. This is because the number of classes could increase the complexity of the inductive learning algorithm. However, many research studies have simplified the multi-class classification into a series of binary classification in order reduce the complexity of the classifier such as One-Against-All (OAA) [1], One-Against-One (OAO) [2], and All and One (A&O) [3] techniques. By doing this, it is able to efficiently solve the multi-class problem using multi-binary classifiers.

The imbalanced data problem is another significant problem for inductive machine learning (ML). In recent years, many research have shown interest in investigating the class imbalance problem. They have found that the imbalanced data could be one of the obstacles for several ML algorithms [4], [5], [6], [7]. In the learning process of the ML algorithm, if the ratio of the minority class and the majority class is highly different, ML tends to learn the features dominated by the majority class and may recognise little on the features of the minority class. As a result, the classification accuracy of the minority class may be low when compared with the classification accuracy of the majority class. Therefore, in order to address the issue of the minority classes in the imbalanced data set, techniques with special characteristics need to be used to enhance the ML algorithm.

There are two major approaches to deal with an imbalanced data: the data -level approach and the algorithm-level approach [7]. While the data-level approach aims to re-balance the class distribution before a classifier is trained, the algorithm level approach aims to strengthen the existing classifier by adjusting the algorithms to recognise the small class [7]. Although both algorithm-level and data-level approaches have been applied to several problem domains, there are some shortcomings that need consideration. The algorithm- level approach is applicant-dependent or algorithm-dependent [8]. Therefore, it performs effectively only on a certain data set. For the data-level approach, while the under-sampling technique can eliminate useful data from the training set, the over-sampling technique may lead to over-fitting problem in the minority class [4].

When problem domains become more complex such as the classification problem of multi-class imbalanced data, the prior approaches may not be efficiently employed to handle this problem. Some research studies presented that researchers cannot enhance the performance by using techniques from the binary classification to solve the imbalanced data problem in the multi-class classification [9]. The literature has also shown that the re-sampling techniques tend to affect negatively the classification performance of the multi-class imbalanced data

[9]. This is because the under-sampling technique can weaken the learning process if a number of useful instances in each large class are removed. The over-sampling technique, for example Synthetic Minority Over-sampling Technique (SMOTE), also can cause a negative effect because the imbalanced data can hamper the generation of synthetic instances. The synthetic instances generated by SMOTE may be misleading when the small class instances are surrounded by a number of large class instances [9].

Another major issue has been found when the re-sampling techniques for imbalanced data problem are implemented. Each re-sampling technique has a major concern for the minority class rather than the majority class. As a result, the classification accuracy cannot be used to evaluate the performance because the minority class has minor impact on the accuracy when compared to the majority class. There is another reason why the accuracy is less preferable to measure the classification performance. When data balancing techniques are implemented, it can cause a negative effect on the accuracy. While the classification accuracy on the minority class is improved, the accuracy on the majority class tends to decrease. Finally, because of the high ratio of majority class compared to the minority class, the overall accuracy tends to reduce. Generally, for imbalanced data problem, many research studies use alternative measures such as F-measure, G-mean and the area under ROC curve (AUC) rather than conventional classification accuracy [7].

Although not many of the research has covered this problem [9], there are some research studies that have tried to apply data distribution techniques to handle imbalanced data problems in the multi-class classification such as the One Against Higher Order Approach (OAHO) [10] and Multi- IM approach [11]. Although these approaches perform well in several case studies, there are some concerns over these techniques. For the OAHO approach, binary classifiers are ordered and they are constructed as a hierarchy. Therefore, if one of the top classifiers misclassifies the data, the wrongly classified data cannot be corrected by the other lower classifiers [11]. There is a potential risk that the error can affect the overall performance. For the Multi-IM approach, the balancing technique using the random under-sampling over the majority class may affect the ability of classification models. A classifier can eliminate potentially useful data in the majority class that is needed for the training phase [4].

The main objective of this paper is to propose a multi-class classification algorithm with data balancing technique in order to enhance the classification performance of multi-class imbalanced data. The disadvantages above lead to the research focus in this paper, which is how to maintain the overall classification accuracy and enhance the classification performance for the minority class at the same time. Moreover, as mentioned above, it is difficult to handle the multi-class imbalanced data using re-sampling techniques. This also leads to another research focus on how to apply the re-sampling techniques to classify the multi-class imbalanced data in order to obtain satisfactory classification results.

In this contribution, an algorithm named One- Against-All with Data Balancing (OAA -DB) is developed. The One-Against-All (OAA) approach incorporated with the artificial neural network (ANN) classifier has been integrated with the combined re-sampling technique before the experimental data was trained and tested. The OAA approach is selected as a basic technique for this classification because the number of binary classifiers used is less than the other approaches such as One-Against-One (OAO) and All and One (A&O) [2], [3]. The balancing technique is employed by combining Complementary Neural Network (CMTNN) [12] and Synthetic Minority Over-Sampling Technique (SMOTE) [6] in order to balance the class distribution.

II. THE OAA-DB TECHNIQUE

The One-Against-All technique with Data Balancing (OAA-DB) algorithm is proposed to deal with the multi-class classification with imbalanced data. The fundamental principles under this approach are based on the research direction on [10] and [11] which attempt to balance data among classes before performing multi-class classification. The proposed approach combines the OAA and the data balancing technique using the combination of SMOTE and CMTNN. The proposed technique is an extended algorithm from the OAA. It aims to improve the weakness of OAA because OAA has highly imbalanced data between classes when one class is compared with all the remaining classes. Moreover, if OAA uses only the highest output value to predict an outcome, there is a high potential risk that the majority class can dominate the features of the prediction. The concept of codeword which is used in [13] is also applied to this proposed technique in order to define the confidence value of the prediction outcomes. In the following sub-sections, the basic concepts of CMTNN and SMOTE are described. The data balancing technique which combines of CMTNN and SMOTE is then presented, and followed by the algorithm of OAA-DB.

A. Complementary Neural Network (CMTNN)

CMTNN [12] is a technique using a pair of complementary feedforward backpropagation neural networks called Truth Neural Network (Truth NN) and Falsity Neural Network (Falsity NN) as shown in Figure 1.

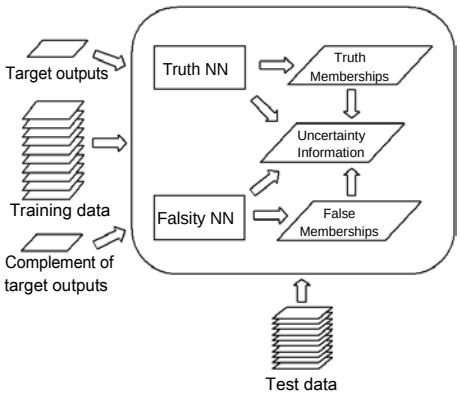

Figure 1. Complementary neural network [14]

While the Truth NN is a neural network that is trained to predict the degree of the truth memberships, the Falsity NN is trained to predict the degree of false memberships. Although the architecture and input of Falsity NN are the same as the Truth NN, Falsity NN uses the complement outputs of the Truth NN to train the network. In the testing phase, the test set is applied to both networks to predict the degree of truth and false membership values. For each input pattern, the prediction of false membership value is expected to be the complement of the truth membership value. Instead of using only the truth membership to classify the data, which is normally done by most convention neural network, the predicted results of Truth NN and Falsity NN are compared in order to provide the classification outcomes [14], [15].

In order to apply CMTNN to perform under-sampling [16], Truth NN and Falsity NN are employed to detect and remove misclassification patterns from a training set in the following steps:

a) The Truth and Falsity NNs are trained by truth and false membership values.

b) The prediction outputs (Y) on the training data (T) of both NNs are compared with the actual outputs (O).

c) The misclassification patterns of Truth NN and Falsity NN (M_{Truth} , $M_{Falsity}$) are also detected if the prediction outputs and actual outputs are different.

$$\text{For Truth NN : If } Y_{Truth\,i} \neq O_{Truth\,i} \quad \text{then } M_{Truth} \leftarrow M_{Truth} \cup \{T_i\} \quad (1)$$

$$\text{For Falsity NN : If } Y_{Falsity\,i} \neq O_{Falsity\,i} \quad \text{then } M_{Falsity} \leftarrow M_{Falsity} \cup \{T_i\} \quad (2)$$

d) In the last step, the new training set (T_c) is constructed by eliminating all misclassification patterns detected by the Truth NN (M_{Truth}) and Falsity NN ($M_{Falsity}$) respectively.

$$T_c \leftarrow T - (M_{Truth} \cup M_{Falsity}) \quad (3)$$

B. Synthetic Minority Over-Sampling Technique (SMOTE)

SMOTE [6] is an over-sampling technique. This technique increases the number of new minority class instances by the interpolation method. The minority class instances that lie together are identified first, before they are employed to form new minority class instances. In Figure 2, it shows how the SMOTE algorithm creates synthetic data. Instance r_1, r_2, r_3, and r_4 are formed as new synthetic instances by interpolating instances x_{i1} to x_{i4} that lie together.

This technique is able to generate synthetic instances rather than replicate minority class instances; therefore, it can avoid the over-fitting problem.

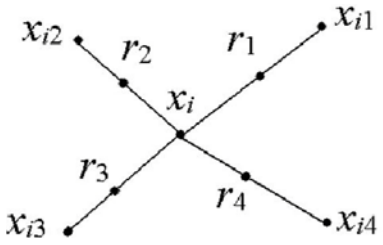

Figure 2. The creation of synthetic data points in the SMOTE algorithm [17]

C. The Combined Technique of CMTNN and SMOTE for Data Balancing

In order to obtain the advantages of using the combination of under -sampling [4] and over-sampling [6] techniques, CMTNN is applied as an under-sampling technique while SMOTE is used as an over-sampling technique. They are combined in order to better handle the imbalanced data problem. The combined technique of CMTNN and SMOTE has been investigated and implemented effectively to the binary classification when handling imbalanced data as demonstrated in [18] and [19]. This data balancing technique can be described by the following steps:

a) The over-sampling technique is applied to the minority class using the SMOTE algorithm. The ratio between the minority and majority class instances after implementing the SMOTE algorithm is 1:1.

b) The under-sampling technique is employed on both classes using the CMTNN under-sampling technique by eliminating all misclassification patterns detected by the Truth NN and Falsity NN.

D. The OAA-DB Algorithm for Dealing with Multi-class Imbalanced Problems

OAA-DB is proposed by integrating the OAA approach and the combined data balancing technique above. A series of binary classifiers using ANN are created before each subset data is trained and tested by each learning model. The steps of the OAA-DB approach are shown as follows:

a) For K-classes of the OAA approach, $f_j(x_i)$ is a mapping function of a binary classifier where j = 1 to K. The outputs of instance i (Y_i) are the results of the map function between each positive class j compared to all other classes.

$$Y_i = \{f_1(x_i), f_2(x_i) , \ldots , f_K(x_i)\} \quad \text{for all } j \text{ from 1 to } K \quad (4)$$

b) For each bit of a codeword

$$cw_i(x_i) = \begin{cases} 1 & \text{if } f_j(x_i) \geq 0.5 \\ 0 & \text{if } f_j(x_i) < 0.5 \end{cases} \quad \text{for all } j \text{ from 1 to } K \quad (5)$$

c). if $cw(x_i)$ contains only one bit of "1" then the class label is c_i with bit "1" else each training set is applied by the data

balancing technique before *K*-binary classifiers are re-trained again

if $cw(x_i)$ after using data balancing contains only one bit of "1"

then the class label is c_j with bit "1"

else the class label for $x_i = c_j$ with $Max(Y_i)$ for all j from 1 to K.

In Figure 3 the flowchart of the OAA-DB algorithm is presented.

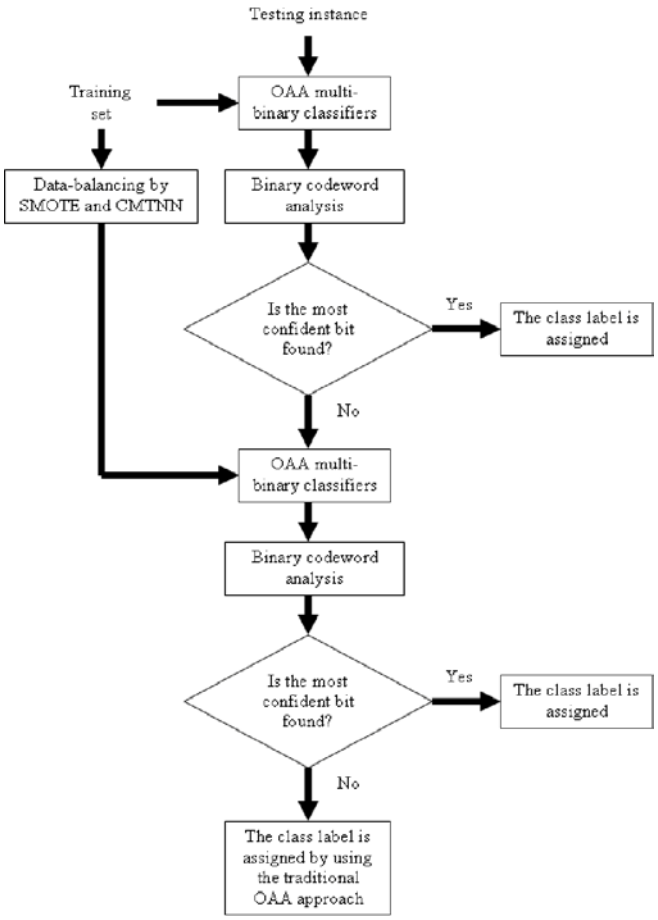

Figure 3. The OAA-DB algorithm

The OAA-DB algorithm starts with using the OAA technique to classify multi-class data. When the *K* outputs of *K*-classes data are produced by multi-binary classifiers, rather than using the highest value to categorise the class label, each *K* output is converted to a binary bit at a threshold equal to 0.5. A binary codeword is represented by the *K* bits class output of each testing instance. If only one bit of the codeword indicates "1", it means that only one class provides the most confidence over other classes. This indicated bit class can be used to label the class. If there is more than one bit of "1" in the codeword, the confidence to provide the class label is still low and the class label is not conclusive at this stage. The combined re-sampling technique of SMOTE and CMTNN will be employed to balance the size of the minority class and majority class. After the training data is balanced, *K* binary classifiers are re -trained again. The codeword method is again used to find the class with the most confident bit. Finally, if there is more than one bit indicating "1", it implies no class with the most confident bit is found, and the conventional method is used. The highest output value of the OAA approach before re-balancing is employed to generate the class label. At this stage, the conventional OAA approach is used to predict the class label rather than using the OAA approach after re -balancing because the OAA-DB algorithm attempts to protect the negative effect of the re-sampling technique. Therefore, this technique aims to improve the performance of the minority class without degrading the overall accuracy.

Moreover, the purpose of the OAA-DB algorithm aims to reduce the ambiguity problem of the OAA approach. This is because the OAA approach consists of *K* binary classifiers and they are trained separately. This can cause the classification boundary to be drawn independently by each classifier as shown in Figure 4. As a result, some regions in the feature space may not be covered (an uncovered region) or they may be covered by more than one class (an overlapped region) [2]. Due to these problems, the OAA approach may not generalise well on the test data. In this case, the confident bit of codeword and the data balancing technique with the OAA-DB algorithm are proposed in order to reduce these problems. The confident bit of codeword can be used to decide a class label with confidence at the overlapped region. The data balancing technique also aims to reduce the problem at the uncovered region.

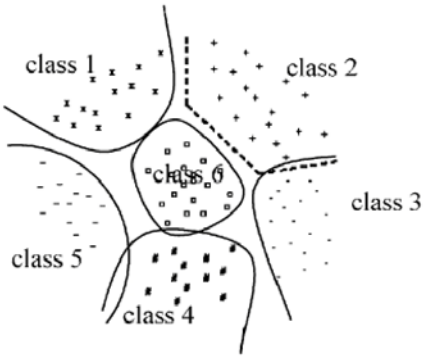

Figure 4. The example of classification boundaries drawn by classifiers trained with the OAA approach [2]

III. EXPERIMENTS AND RESULTS

Three data sets from the University of California Irvine (UCI) machine learning repository [20] are used in the experiment. The data sets for multi-class problems include Balance Scale data, Glass Identification data, and Yeast data. These data sets are selected because they are multi-class imbalanced data sets with different numbers of classes. Each data is described as follows:

- Balance Scale data set was generated to model a psychological experiment. This data set is classified into three classes by having the balance scale tip to the left, tip to the right, or be balanced.

This data contains 625 instances and each pattern is described by four attributes.

- The purpose of the Glass Identification data set is to determine a type of glass. The study of this data set was motivated by criminological investigation. At the scene of crime, the glass may be left as evidence. Thus, an effective classification technique is needed to identify the glass. This data set contains 214 instances associated with six classes. Each instance is composed of nine attributes.
- The purpose of the Yeast data set is to predict the cellular localisation sites of proteins. This data set can be classified into ten classes. It contains 1,484 instances, and each instance is described by eight attributes.

The characteristics of these data sets are shown in Table I. The data distribution of each data set is presented in Table II.

TABLE I. CHARACTERISTICS OF THE EXPERIMENTAL DATA SETS

Name of Data Set	No. of Instances	No. of Attributes	No. of Classes
Balance Scale data	625	4	3
Glass Identification data	214	9	6
Yeast data	1,484	8	10

TABLE II. DATA DISTRIBUTION OF THE EXPERIMENTAL DATA SETS

Name of Data Set	Ratio of Classes (%)				
	C1	*C2*	*C3*	*C4*	*C5*
Balance Scale data	8.00	46.00	46.00	-	-
Glass Identification data	32.71	35.51	7.94	6.07	4.21
Yeast data	31.20	28.91	16.44	10.98	3.44
	C6	*C7*	*C8*	*C9*	*C10*
Balance Scale data	-	-	-	-	-
Glass Identification data	13.55	-	-	-	-
Yeast data	2.96	2.36	2.02	1.35	0.34

In the experiment, after the OAA-DB approach is implemented, the classification performance is then evaluated by the percentage of accuracy and F -measure (F_1). While the accuracy is evaluated for the overall classification performance, F_1 is used to evaluate the classification performance for imbalanced classes. For the purpose of establishing the classification model, each data set is split into 80% training set and 20% test set. This data ratio selected is based on several experiments in the literature which normally confine the experimental data to the test set in the range of 10% to 30% [13], [21], [22]. The cross validation method is applied in order to reduce inconsistent results. Each data set is randomly split ten times to form different training and test data sets. The results of the ten experiments of each data set are averaged to indicate the overall performance of the experimental techniques.

In order to compare the performance of the OAA-DB algorithm with others, OAA, OAO, A&O and OAHO techniques are employed. They are selected because OAA is the basic technique of the OAA-DB algorithm. Furthermore, the OAO techniques have been applied widely to the multi-class classification [2]. The A&O technique is also the combination of OAA and OAO techniques which have provided good results in the literature [10]. Moreover, OAHO is chosen because it is designed specifically for the multi-class imbalanced data. In addition, OAHO has been experimented originally by using ANN as a classifier, which is the same learning model of the OAA-DB algorithm.

Tables III and IV show the classification results of the Balance Scale Data, which comprises of three classes. While Table III shows the performance results in terms of the overall accuracy and macro-F_1, Table IV shows the classification accuracy of each class of each technique.

TABLE III. THE CLASSIFICATION RESULTS OF BALANCE SCALE DATA

Evaluation Measure	OAA	OAO	A&O	OAHO	OAA-DB
Accuracy (%)	92.72	93.36	91.52	94.08	94.56
Macro-F_1 (%)	68.74	82.72	67.66	84.65	85.37

The classification performance in Table III shows that the OAA-DB algorithm outperforms other techniques in terms of accuracy and macro-F_1. While the OAA-DB technique provides the best results (accuracy: 94.56%, macro-F_1: 85.37%), OAHO presents the second best (accuracy: 94.08%, macro-F_1: 84.65%). The OAA-DB algorithm can improve the classification performance for the minority class significantly when compared with the basic OAA. The results of macro-F_1 show the improvement up by 16.63%, from 68.74% for OAA to 85.37% for the OAA-DB algorithm. Furthermore, when the accuracy of each class is compared, the OAA-DB algorithm improves the accuracy of the minority class significantly. The minority class increases up to 60.21% compared with the basic technique, OAA, 9.01%

TABLE IV. THE CLASSIFICATION ACCURACY OF EACH CLASS ON BALANCE SCALE DATA

Class	Ratio of Classes (%)	Accuracy (%)				
		OAA	OAO	A&O	OAHO	OAA-DB
C1	8	9.01	64.43	8.67	68.66	60.21
C2	46	98.11	95.26	97.12	95.08	96.26
C3	46	99.14	95.80	97.69	96.70	97.70
Average		68.75	85.16	67.83	86.81	84.72

In Table IV, although OAHO has accuracy on class one (68.66%) higher than the OAA-DB algorithm (60.21%), it provides lower accuracies on class two and class three, which are the majority classes. While OAHO provides accuracy at 95.08% on class two and at 96.70% on class three, the OAA-DB algorithm shows higher accuracy at 96.26% on class two and at 97.70% on class three. Although the OAA-DB algorithm can improve the classification performance better than OAHO in terms of overall accuracy, and macro-F_1 as shown in Table III provides an average accuracy among classes slightly less than the OAHO algorithm. These are 84.72% and 86.81% performed by the OAA-DB and the

OAHO algorithms respectively. The discussion in Section V will present the reasons why OAHO performs well only with this data set, which consists of fewer numbers of classes, and why the performance results decline when the feature of empirical data sets becomes more complex with a large number of classes.

Table V and VI show the classification results of Glass Identification data, which is composed of six classes. The OAA-DB algorithm still outperforms other techniques in terms of accuracy and macro-F_1.

TABLE V. THE CLASSIFICATION RESULTS OF GLASS IDENTIFICATION DATA

Evaluation Measure	OAA	OAO	A&O	OAHO	OAA-DB
Accuracy (%)	63.26	62.33	62.09	60.93	67.44
Macro-F_1 (%)	44.14	40.62	37.68	49.89	58.15

In Table V, the results show that the OAA-DB technique has higher accuracy than OAA, OAO and A&O by around 4% to 5%. It is also higher than OAHO by around 7%. Furthermore, when the macro-F_1 of the OAA-DB algorithm is compared with OAHO, the macro-F_1 of the OAA-DB algorithm is significantly greater than the macro -F_1 of OAHO by around 8%. When each class is considered in Table VI, the OAA-DB algorithm can produce the improvement on several minority classes including class four, class five and class six. It increases the accuracies up to 81.25%, 72.22% and 81.81% for class four, class five, and class six respectively.

TABLE VI. THE CLASSIFICATION ACCURACY OF EACH CLASS ON GLASS IDENTIFICATION DATA

Class	Ratio of Classes (%)	Accuracy (%)				
		OAA	OAO	A&O	OAHO	OAA-DB
C1	32.7	78.47	83.57	83.25	77.05	78.51
C2	35.51	65.15	62.19	62.89	48.67	62.46
C3	7.94	0.00	2.78	0.00	6.30	1.85
C4	6.07	41.67	22.92	12.50	79.17	81.25
C5	4.21	17.59	11.11	5.56	62.04	72.22
C6	13.55	78.71	71.71	74.14	72.21	81.81
Average		46.93	42.38	39.72	57.57	63.02

In this data set, although the total accuracy of OAHO presents the lowest accuracy (60.93%) compared with others, the macro-F_1 of OAHO is still the second best at 49.89%. It means that although the OAHO technique performs effectively on the imbalanced data problem, it cannot maintain overall accuracy. The inconsistency on these results occurs because the effect of the balancing technique of OAHO has on the overall accuracy. While the balancing technique can enhance the classification performance on the minority class, it can affect the global accuracy as discussed in Section I. In Table VI, although the accuracies of minority classes performed by OAHO increase from 0% to 6.3% (class three, ratio 7.94%), from 41.67% to 79.17% (class four, ratio 6.07%) and from 17.59% to 62.04% (class five, ratio 4.21%) when compared to the OAA technique, the accuracies of the majority classes tends to decrease; for example, the accuracy of the majority class two (ratio 35.51%) decreases from 65.15% to 48.67%. As a result, the global accuracy is reduced because the majority class two has a greater ratio than other minority classes. The decrease of accuracy on the majority class two tends to have more impact on the global accuracy than the increase of accuracy on other minority classes.

In Table VII, the classification results of Yeast data, which contains ten classes, are presented. The OAA-DB technique outperforms the other techniques with the best outcomes of accuracy (60.37%) and macro-F_1 (53.80%). Similar to the previous case, the Glass Identification data, OAHO produces the lowest accuracy at 52.69% which is lower than the other methods by around 6 to 8%. In addition, OAHO performs in third place for the macro-F_1 at 45.33%.

TABLE VII. THE CLASSIFICATION RESULTS OF YEAST DATA

Evaluation Measure	OAA	OAO	A&O	OAHO	OAA-DB
Accuracy (%)	59.87	59.87	58.96	52.69	60.37
Macro-F_1 (%)	44.57	50.47	44.93	45.33	53.80

In Table VIII, when the accuracy of each class is compared between the OAA-DB algorithm and the basic technique, OAA, the OAA-DB algorithm presents better accuracies on five minority classes. These are class three, class seven, class eight, class nine, and class ten. In some classes, the OAA-DB algorithm can increase the accuracies significantly, such as class nine which increases from 10.83% to 43.89%, and class ten which increases from 33.33% to 50.00%.

TABLE VIII. THE CLASSIFICATION ACCURACY OF EACH CLASS ON YEAST DATA

Class	Ratio of Classes (%)	Accuracy (%)				
		OAA	OAO	A&O	OAHO	OAA-DB
C1	31.20	68.80	65.93	66.16	36.38	67.85
C2	28.91	52.48	52.24	51.98	61.43	52.37
C3	16.44	57.92	57.20	56.36	60.47	58.72
C4	10.98	85.01	84.69	85.24	77.83	85.01
C5	3.44	27.28	31.57	30.80	30.98	27.28
C6	2.96	76.67	77.08	73.67	67.59	76.67
C7	2.36	47.28	63.03	56.36	60.05	48.28
C8	2.02	0.00	0.00	0.00	6.25	1.67
C9	1.35	10.83	33.61	10.83	38.61	43.89
C10	0.34	33.33	44.44	33.33	33.33	50.00
Average		45.96	50.98	46.47	47.29	51.17

IV. DISCUSSION

The results in Tables III to VIII show that there are some factors relating to the performance of the results, such as the

size of the training set and the number of classes. Similar to the results as those in [2], the OAO approach performs well when the training data is large while the OAA algorithm provides better results when the size of the training data is small. The results in Tables V and VI show that the OAA algorithm performs better than the OAO approach on the Glass Identification data, which contains fewer training instances, 171 training instances. On the other hand, the OAO algorithm shows better results than the OAA approach on the Balance Scale Data and Yeast data, which contain more training instances at 500 and 1,187 training instances respectively.

Furthermore, in order to discuss why the OAHO approach, which is designed for the multi-class imbalanced data, provides lower performance in terms of the overall accuracy and macro-F_1 when compared with the OAA-DB approach, some disadvantages of OAHO are explained as follows.

Due to the hierarchical structure of the OAHO approach, the misclassification at the upper levels of the OAHO hierarchy cannot be corrected by the lower levels. When the number of classes increases, the number of levels under the OAHO hierarchy needs to be increased as well. As a result, the OAHO could have a high risk of assigning misclassification results at the upper levels. Therefore, the OAHO technique tends to not perform effectively in the problem domains which have a high number of classes. The larger the number of classes contained in a data set, the lower performance can be generated by the OAHO technique. The experiment results indicate that OAHO can improve the overall classification accuracy only on the Balance Scale data set (three -class data) whereas the classification accuracies of the Glass Identification data set (six-class data) and the Yeast data set (ten-class data) are shown as lower than other approaches.

Moreover, the OAHO technique cannot overcome the imbalanced data problem in some test cases. This is because the imbalanced data problem still occurs even though the OAHO technique aims to reduce the effect of this problem by comparing a larger class with a group of smaller classes. In Figure 5, the comparison between classes in the OAHO hierarchy is shown. When c_i is compared with higher order data $\{c_{i+1},\ldots,c_K\}$, there is a possibility that comparison classes are imbalanced. For example, in the Yeast data set, the classifier one performs the comparison between class one (ratio 31.20%) and classes two to ten (ratio 68.80%), and then the classifier two performs the comparison between class two (ratio 28.91%) and classes three to ten (ratio 39.89%). As can be seen, the class imbalance problem still exists by using the OAHO approach in this data set. Consequently, the OAHO technique shows lower performance in terms of macro-F_1 than the conventional approach, OAO, as shown in Table VIII. While the OAHO technique can provide 45.33% of macro-F_1, the OAO technique produces better result at 50.47% of macro-F_1.

In order to explain why the OAA-DB algorithm performs effectively on the multi-class imbalanced data, and why it can increase the overall performance and the performance for minority classes, each technique used in the OAA-DB algorithm has to be discussed. The OAA-DB algorithm combines three major techniques in order to enhance the classification performance. These are the OAA approach, data balancing technique, and the codeword method.

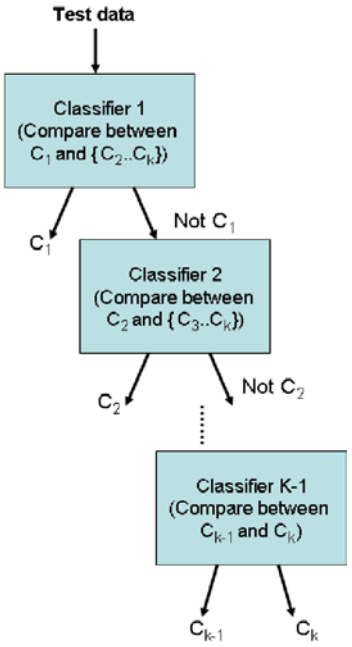

Figure 5. The comparison classes in the OAHO Hierarchy [2]

The OAA approach is first integrated into the proposed algorithm because of its major benefits. OAA can provide some benefits over the OAO and A&O approach, such as using less number of binary classifiers, and shorter total training time [1], [2]. Secondly, the data balancing feature, which combines the re-sampling techniques of SMOTE and CMTNN, can also support the improvement of classification performance for the minority classes. While the SMOTE algorithm is used to increase a number of minority class instances in order to reduce bias toward the majority class, the CMTNN technique is used for under-sampling in order to clean noisy data from the training data. When the training data between classes becomes more balanced, the features in the minority classes can be more recognised by the learning algorithm. As a result, the learning algorithm tends to generalise the accurate prediction for the minority class. Lastly, the OAA-DB algorithm also attempts to reduce the negative effect of the data balancing technique by using the codeword technique. The most confident bit of codeword is used to assign the class label. If the most confident bit cannot be defined, the conventional OAA approach is still employed to assign the class label.

By integrating these three techniques above, the results of the three experimental data sets show that the OAA-DB algorithm performs effectively in each data set. It can enhance the classification performance evaluated by the total accuracy and macro- F_1. This algorithm can enhance the overall performance in terms of the global accuracy and the classification performance for the minority class.

Finally, in order to compare the computational cost of the OAA-DB algorithm with other approaches, the total number of binary classifiers trained in each approach can be considered. For the *K*-class data set, in ascending order, the number of binary classifiers needed for training are *K-1* binary classifiers for OAHO, *K* binary classifiers for OAA, 2*K* binary classifiers for OAA-DB, *K*(*K-1)/2* binary classifiers for OAO, and *K*(*K+1)/2* binary classifiers for A&O. It can be concluded that the OAA-DB approach stands at the medium level of computational cost. While the approaches with high computational cost are A&O and OAO, the techniques with low computational cost are OAHO and OAA.

V. CONCLUSIONS

This paper proposed a technique named as the One-Against-All with Data Balancing (OAA-DB) algorithm to solve the multi-class imbalanced problem. It applies the multi-binary classification techniques called the One-Against-All (OAA) approach and the combined data balancing technique. The combined data balancing technique is the integration of the under-sampling technique using Complementary Neural Network (CMTNN) and the over -sampling technique using Synthetic Minority Over -sampling Technique (SMOTE). The experiment is conducted by using three multi-class data sets from the University of California Irvine (UCI) machine learning repository, that is, Balance Scale data, Glass Identification data, and Yeast data. The results of classification are evaluated and compared in terms of the performance using accuracy and macro-F_1. While the accuracy is used to evaluate the overall performance, macro-F_1 is employed to evaluate the classification performance on the minority classes. The results obtained from the experiment indicated that the OAA-DB algorithm can enhance the classification performance for the multi-class imbalanced data, and it performs better than other techniques in each test case. The OAA-DB algorithm can increase the classification performance of the minority classes and maintain the overall performance in terms of the accuracy.

REFERENCES

[1] M. Aly, "Survey on multi-class classification methods," Caltech, USA, Technical Report, November 2005.

[2] G. Ou and Y. L. Murphey, "Multi-class pattern classification using neural networks," *Pattern Recognition*, vol. 40, pp. 4-18, 2007.

[3] N. Garcia-Pedrajas and D. Ortiz-Boyer, "Improving multiclass pattern recognition by the combination of two strategies," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, pp. 1001-1006, 2006.

[4] 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 Explorations Newsletter*, vol. 6, pp. 20-29, 2004.

[5] J. Laurikkala, "Improving identification of difficult small classes by balancing class distribution," in *Proceedings of the 8th Conference on AI in Medicine in Europe: Artificial Intelligence Medicine*, London: Springer-Verlag, 2001, pp. 63-66.

[6] 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.

[7] Q. Gu, Z. Cai, L. Zhu, and B. Huang, "Data mining on imbalanced data sets," in *International Conference on Advanced Computer Theory and Engineering (ICACTE '08)*, 2008, pp. 1020-1024.

[8] Y. Sun, A. K. C. Wong, and M. S. Kamel, "Classification of imbalanced data: A review," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 23, pp. 687-719, 2009.

[9] Z.-H. Zhou and X.-Y. Liu, "Training cost-sensitive neural networks with methods addressing the class imbalance problem," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, pp. 63-77, 2006.

[10] Y. L. Murphey, H. Wang, G. Ou, and L. A. Feldkamp, "OAHO: An effective algorithm for multi-class learning from imbalanced data," in *International Joint Conference on Neural Networks (IJCNN)*, 2007, pp. 406-411.

[11] A. S. Ghanem, S. Venkatesh, and G. West, "Multi-class pattern classification in imbalanced data," in *Proceeding of the 20th International Conference on Pattern Recognition (ICPR)*, 2010, pp. 2881-2884.

[12] P. Kraipeerapun, C. C. Fung, and S. Nakkrasae, "Porosity prediction using bagging of complementary neural networks," in *Advances in Neural Networks – ISNN 2009*, 2009, pp. 175-184.

[13] P. Kraipeerapun, C. C. Fung, and K. W. Wong, "Multiclass classification using neural networks and interval neutrosophic sets," in *Proceedings of the 5th WSEAS International Conference on Computational Intelligence, Man-Machine Systems and Cybernetics*, Venice, Italy, 2006, pp. 123-128.

[14] P. Kraipeerapun and C. C. Fung, "Binary classification using ensemble neural networks and interval neutrosophic sets," *Neurocomput.*, vol. 72, pp. 2845-2856, 2009.

[15] P. Kraipeerapun, C. C. Fung, W. Brown, K. W. Wong, and T. Gedeon, "Uncertainty in mineral prospectivity prediction," in *the 13th International Conference on Neural Information Processing (ICONIP 2006)*, Hong Kong, 2006, pp. 841-849.

[16] P. Jeatrakul, K. W. Wong, and C. C. Fung, "Data cleaning for classification using misclassification analysis," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 14, no. 3, pp. 297-302, 2010.

[17] J. Wang, M. Xu, H. Wang, and J. Zhang, "Classification of imbalanced data by using the SMOTE algorithm and locally linear embedding," in *The eighth International Conference on Signal Processing*, 2006, pp. 16-20.

[18] P. Jeatrakul, K. W. Wong, and C. C. Fung, "Classification of Imbalanced Data by Combining the Complementary Neural Network and SMOTE Algorithm," in *Neural Information Processing. Models and Applications*. vol. 6444: Springer Berlin / Heidelberg, 2010, pp. 152-159.

[19] P. Jeatrakul, K. W. Wong, C. C. Fung, and Y. Takama, "Misclassification analysis for the class imbalance problem," in *World Automation Congress (WAC 2010)*, Kobe,Japan, 2010.

[20] A. Asuncion and D. J. Newman, "UCI Machine Learning Repository," University of California, Irvine, School of Information and Computer Sciences, 2007.

[21] C. E. Brodley and M. A. Friedl, "Identifying mislabeled training data," *Journal of Artificial Intelligence Research*, vol. 11, pp. 137-167, 1999.

[22] D. Gamberger, N. Lavrac, and S. Dzeroski, "Noise detection and elimination in data preprocessing: Experiments in medical domains," *Applied Artificial Intelligence*, vol. 14, pp. 205-223, 2000.