Jeatrakul, P., Wong, K.W., Fung, C.C. and Takama, Y. (2010) ***Misclassification analysis for the class imbalance problem.*** **In: World Automation Congress, WAC 2010, 19 - 23 September, Kobe, Japan.**

MISCLASSIFICATION ANALYSIS FOR THE CLASS IMBALANCE PROBLEM

PIYASAK JEATRAKUL, MURDOCH UNIVERSITY, AUSTRALIA, p.jeatrakul@murdoch.edu.au
KOK WAI WONG, MURDOCH UNIVERSITY, AUSTRALIA, k.wong@murdoch.edu.au
CHUN CHE FUNG, MURDOCH UNIVERSITY, AUSTRALIA, l.fung@murdoch.edu.au
YASUFUMI TAKAMA, TOKYO METROPOLITAN UNIVERSITY, JAPAN, ytakama@sd.tmu.ac.jp

ABSTRACT

In classification, the class imbalance issue normally causes the learning algorithm to be dominated by the majority classes and recognize slightly the minority classes. This will indirect affect how human visualise the data. Therefore, special care is needed to apply to the learning algorithm in order to enhance the accuracy for the minority classes. In this study, the use of misclassification analysis is investigated for data re-distribution. Several under-sampling techniques and hybrid techniques using misclassification analysis are proposed in the paper. The benchmark data sets obtained from the University of California Irvine (UCI) machine learning repository are used to investigate the performance of the proposed techniques. The results show that the proposed hybrid technique presents the best performance in the experiment.

KEYWORDS: Class imbalance problem, artificial neural network, complementary neural network, classification, misclassification analysis

1. INTRODUCTION

In recent years, many research groups have shown interest in investigating the class imbalance problem. They have found that an imbalanced data set could be one of the obstacles for many Machine Learning (ML) algorithms [1]. Generally, an imbalanced data set occurred when instances in some classes outnumbered the other classes, i.e. the distribution is uneven. The classes which have more instances are called the majority classes while the other classes which have fewer instances are commonly known as minority classes. In the learning process of the ML algorithms, if the ratio of minority classes and majority classes is highly different, ML tends to be dominated by the majority classes and recognize slightly the minority classes. As a result, the classification accuracy of the minority classes may be remarkably low when compared to the classification accuracy of the majority classes. Therefore, in order to take care of the minority classes in an imbalanced data set, techniques are needed to enhance the ML algorithm. This could increase the accuracy of the classification overall.

According to Gu et al. [2], there are two major approaches to deal with imbalanced data sets; data-level approach and algorithm approach. 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 algorithms to recognize the small class. There are three categories of data-level approach. These are the under-sampling technique, the over-sampling technique and the hybrid technique. While the over-sampling technique tries to increase instances of minority classes, the under-sampling technique tries to balance a data set by removing instances of majority classes. In addition, the hybrid technique is a technique that combines different techniques of both under-sampling and an over-sampling.

For the under-sampling techniques, many algorithms have been proposed, for example Random under-sampling [1], Tomek links [3], One-Sided Selection (OSS) [4], Condensed Nearest Neighbor Rule (CNN) [5], Wilson's Edited Nearest Neighbor Rule (ENN) [6], and Neighborhood Cleaning Rule (NCL) [7]. There are also several techniques applied for over-sampling methods such as Random over-sampling [1], and Synthetic Minority Over-sampling Technique (SMOTE) [8].

In order to evaluate the classification performance of an imbalanced data set, the conventional classification accuracy can not be used for this purpose because the minority class has a minor impact on the accuracy when compared to the majority class. Therefore, alternative measures are applied to evaluate the classification performance for this type of problem. Some examples of the evaluation methods are F-measure, Geometric-mean (G-mean), Receiver Operating Characteristic (ROC) curve, and the area under ROC curve (AUC) [2].

There are several studies in the literature addressing the class imbalance problem with the objective to enhance the classification performance of the minority class [1]. For example, Gu et al. [9] proposed sampling methods and random forest based techniques to clean noise from majority instances. Their techniques was compared to several other techniques used to handle the imbalanced data such as Random under-sampling, Random over-sampling, Tomek links, and SMOTE. The AUC was used as a performance measure in order to evaluate the efficiency of each technique. The results showed that their proposed technique provided the best AUC score in five out of eight data sets. Barandela et al. [10] dealt with the class imbalance problem by using

three selection algorithms for under-sampling either the majority class or both classes. These were the classical Wilson's proposal (WE), the Nearest Centroid Neighbourhood (k-NCN) and the Modified Selective (MS) condensing. The geometric mean (G-mean) was used as an indicator for the classifier performance. The results presented that each technique can enhance the classification performance in most test sets. However, they perform well only if they are applied to the majority class. Furthermore, Batista et al. [1] investigated several algorithms to deal with imbalanced data sets. They also proposed hybrid techniques which were the combination of SMOTE and Tomek links, and the combination of SMOTE and ENN. The AUC was evaluated the performance of classifiers. They found that in most experiments, over-sampling techniques presented better results than under-sampling techniques. The hybrid techniques also showed outstanding performance results for experimental data sets with a small number of minority class.

Most of reported research dealing with this problem aimed to increase the classification performance of imbalanced data. They focused on examining the feasibility of re-distribution techniques for handling imbalanced data. The direction of this paper takes an alternate approach by proposing alternative re-distribution techniques using misclassification analysis to enhance the classification performance. Complementary Neural Network (CMTNN) [11] is used as an under-sampling technique in order to re-balance the class distribution. CMTNN is used because of its special feature of predicting not only the "truth" classified data but also the "false" data. For the evaluation, AUC and G-mean are selected as the performance measures. These are good indicators for the class imbalance problem because they try to maximize the accuracy between the minority class and the majority class. The AUC and G-mean are also independent of the imbalanced distribution [4], [10]. Furthermore, the core techniques of classification used in this paper is based on artificial neural networks (ANNs).

In the experiments, four classification data sets from the University of California Irvine (UCI) machine learning repository [12] are used. These include Pima Indians Diabetes, German credit data, Haberman's Survival Data and SPECT heart data. These data sets are selected because they are imbalanced data sets with various ratios between the minority class and the majority class. They are also benchmark data sets which have been commonly used in the literature. Furthermore, the results of other techniques including the Tomek links, ENN and SMOTE are compared to the proposed techniques. These comparison techniques are selected for comparison in this experiment because they have been applied widely to the class imbalance problem.

2 TECHNIQUES AND EVALUATION MEASURES

In this section, the concept of Complementary Neural Network (CMTNN) is described. The two proposed under-sampling techniques based on CMTNN will then be presented.

2.1 Complementary Neural Network (CMTNN)

CMTNN [11] 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 Fig 1.

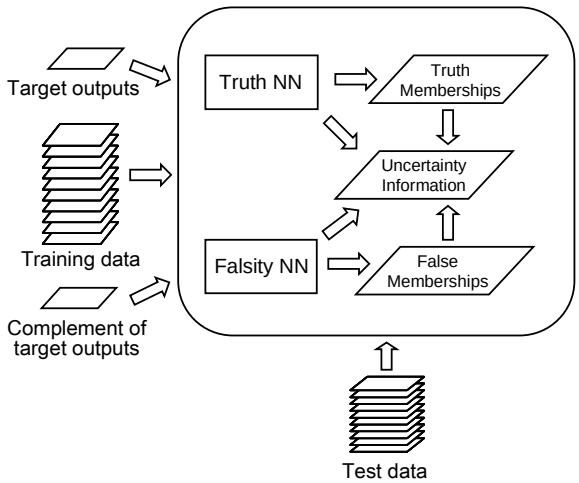

Figure 1*. Complementary neural network [13]*

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. For example, in binary classification problem, if the target output used to train the truth neural network is 0, the complement of this target output to train the falsity neural network will be 1. 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 [13].

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. The difference between the truth and false membership values can also be used to represent uncertainty in the classification [14].

2.2 The Proposed Under-Sampling Techniques

In order to apply CMTNN for under-sampling, Truth NN and Falsity NN are employed to detect and clean misclassification patterns from a training set. There are basically two ways that we can perform under-sampling. The steps of these two techniques are described as follows.

2.2.1 Under-Sampling Technique I

a. The Truth NN and Falsity NN 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.

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

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

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

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

As for training a new neural network classifier, the cleaned data set that removes those misclassification patterns will be used.

2.2.2 Under-Sampling Technique II

a. Repeat the step *a.* to *b.* of cleaning technique I.

b. The new training set (T_c) is cleaned by eliminating all misclassification patterns detected by the Truth NN (M_{Truth}) and Falsity NN ($M_{Falsity}$) respectively.

$$T_c \leftarrow T - (M_{Truth} \cup M_{Falsity}) \quad (4)$$

2.3 The Other Techniques for Comparison

In order to compare the results of the proposed techniques, two under-sampling techniques and an over-sampling technique are used for this purpose. These techniques are described as follows.

2.3.1 Tomek links

Tomek links [3] technique performs by identifying a pair of instance which is belonging to different classes. The nearest neighbour method is used to find a Tomek link pair. Instances in Tomek links pair can be noise or borderline. As an under-sampling technique, only the majority class instances in the Tomek link pairs are eliminated [1].

2.3.2 Wilson's Edited Nearest Neighbor Rule (ENN)

ENN [6] is used as an under-sampling technique by eliminating an instance which belongs to the majority class and at least two of its three nearest neighbors are minority class instances.

2.3.3 Synthetic Minority Over-Sampling Technique (SMOTE)

SMOTE [8] is an over-sampling technique. This technique increases a number of new minority class instances by interpolating method. The minority class instances that lie together are identified by the k-Nearest Neighbor (k-NN) before they are employed to form new minority class instances.

2.4 Evaluation measures

As the widely used measures for the class imbalance problem, Geometric mean (G-mean) and the area under ROC curve (AUC) are applied to evaluate the classification performance in this paper.

2.4.1 Geometric mean (G-mean)

G-mean [2] is defined as the square root of the product of sensitivity and specificity, as shown in equation (5). While sensitivity is the classification accuracy on the minority class instances (the positive instances), specificity is the classification accuracy on the majority class instances (the negative instances). G-mean can balance the performance of machine learning algorithm between a minority class and a majority class.

$$\text{G-mean} = \sqrt{Sensitivity \times Specificity} \quad (5)$$

2.4.2 The Area Under the ROC Curve (AUC)

The Receiver Operating Characteristic (ROC) curve [2] is a two dimensional graph which is the trade-off between sensitivity (true positive rate) and 1-specificity (false positive rate). It can be used to evaluate the classification performance even if the class distribution of minority and majority instances is highly skewed. The AUC is a single scalar which can represent the performance of classification and ranking models.

3 EXPERIMENTS AND RESULTS

Four data sets from UCI machine learning repository [12] are used in the experiment. The data sets for binary classification problems include Pima Indians Diabetes data, German credit data, Haberman's Survival Data, and SPECT heart data.

- The purpose of Pima Indians Diabetes data set is to predict whether a patient shows signs of diabetes.
- The purpose of German credit data set is to predict whether a loan application is “Good” or “Bad” credit risk.
- The purpose of Haberman's Survival data set is to predict whether a patient, who had undergone surgery for breast cancer, survives more than five years or dies within five years.
- The purpose of SPECT heart data set is to predict whether a patient is normal or abnormal on diagnosing of cardiac Single Proton Emission Computed Tomography (SPECT) images

The characteristics of these three data sets are shown in Table I.

Name of data set	No. of instances	No. of attributes	Minority class (%)	Majority class (%)
Pima Indians Diabetes data	768	8	34.90	65.10
German credit data	1000	20	30.00	70.00
Haberman's Survival Data	306	3	26.47	73.53
SPECT heart data	267	22	20.60	79.40

Table I. Characteristics of data sets used in the experiment.

For the purpose of establishing the classification model and testing it, each data set is first split into 80% training set and 20% test set. Furthermore, the cross validation method is used to obtain reasonable results. Each data set will be randomly split ten times to form different training and test data sets. For the purpose of this study, the results of the ten experiments of each data set will be averaged.

For sampling techniques, the two proposed techniques in section 2 are applied differently into six methods. In addition, Tomek links, ENN and SMOTE techniques are used for the comparison. In this paper, nine methods will be implemented on the training sets which are:

a. Under-sampling the majority class using Tomek links technique
b. Under-sampling the majority class using ENN technique
c. Over-sampling the minority class using SMOTE technique
d. Under-sampling both classes using the proposed technique I
e. Under-sampling both classes using the proposed technique II
f. Under-sampling only the majority class using the proposed technique I
g. Under-sampling only the majority class using the proposed technique II
h. Applying the hybrid of method f. and SMOTE
i. Applying the hybrid of method g. and SMOTE

Please take note that for method c, h and i, the ratio between the minority and majority class instances after sampling is 1:1. Finally, after the training sets are processed by the above techniques, new neural network classifiers are trained by the new training sets. The classification performance on the test sets for each data set is evaluated. The results of these each data set are shown in Table II-V.

Techniques	% Accuracy of minority class (Sensitivity)	%Accuracy of majority class (Specificity)	G-Mean	AUC
Original Data	56.70	86.72	70.12	0.8276
a. Tomek links	68.34	78.21	73.11	0.8288
b. ENN	66.26	79.63	72.64	0.8298
c. SMOTE	73.15	75.46	74.30	0.8281
d. Technique I both classes	54.38	88.53	69.38	0.8239
e. Technique II both classes	53.04	88.15	68.38	0.8259
f. Technique I only the majority class	65.31	80.80	72.64	0.8235
g. Technique II only the majority class	71.15	77.29	74.16	0.8292
h. Technique I + SMOTE	76.86	74.26	75.55	0.8332
i. Technique II + SMOTE	75.98	73.11	74.53	0.8300
The best technique	h.	d.	h.	h.
Second best	i.	e.	i.	i.

Table II. The results of each technique on Pima Indians Diabetes data.

Techniques	% Accuracy of minority class (Sensitivity)	%Accuracy of majority class (Specificity)	G-Mean	AUC
Original Data	45.73	89.34	63.92	0.7723
a. Tomek links	61.13	81.27	70.48	0.7793
b. ENN	60.81	82.29	70.74	0.7794
c. SMOTE	66.84	76.44	71.48	0.7777
d. Technique I both classes	48.78	89.82	66.19	0.7784
e. Technique II both classes	45.67	90.32	64.23	0.7766
f. Technique I only the majority class	54.24	84.73	67.79	0.7763
g. Technique II only the majority class	60.69	81.33	70.26	0.7774
h. Technique I + SMOTE	69.25	74.92	72.03	0.7855
i. Technique II + SMOTE	72.91	73.74	73.32	0.7873
The best technique	i.	e.	i.	i.
Second best	h.	d.	h.	h.

Table III. The results of each technique on German credit data.

Techniques	% Accuracy of minority class (Sensitivity)	%Accuracy of majority class (Specificity)	G-Mean	AUC
Original Data	12.29	89.18	33.11	0.5885
a. Tomek links	31.44	85.6	51.88	0.6323
b. ENN	29.25	87.03	50.45	0.6305
c. SMOTE	49.84	68.92	58.60	0.6345
d. Technique I both classes	16.11	93.62	38.84	0.6209
e. Technique II both classes	8.66	95.47	28.75	0.5749
f. Technique I only the majority class	13.02	92.25	34.66	0.5930
g. Technique II only the majority class	24.77	88.33	46.77	0.6378
h. Technique I + SMOTE	56.48	63.73	60.00	0.6452
i. Technique II + SMOTE	57.61	68.42	62.78	0.6770
The best technique	i.	e.	i.	i.
Second best	h.	d.	h.	h.

Table IV. The results of each technique on Haberman's Survival data.

Techniques	% Accuracy of minority class (Sensitivity)	%Accuracy of majority class (Specificity)	G-Mean	AUC
Original Data	45.92	89.34	64.05	0.7590
a. Tomek links	66.41	79.97	72.88	0.8178
b. ENN	62.89	81.97	71.80	0.7895
c. SMOTE	69.48	77.95	73.59	0.8241
d. Technique I both classes	49.07	88.68	65.97	0.7720
e. Technique II both classes	45.90	89.83	64.21	0.8096
f. Technique I only the majority class	57.52	82.93	69.07	0.7763
g. Technique II only the majority class	53.77	85.84	67.94	0.7783
h. Technique I + SMOTE	71.14	76.69	73.86	0.8374
i. Technique II + SMOTE	72.71	75.98	74.32	0.8273
The best technique	i.	e.	i.	h.
Second best	h.	Origin	h.	i.

Table V. The results of each technique on SPECT heart data.

The results in Table II -V show that generally the over-sampling technique, SMOTE, performs better than the under-sampling techniques: Tomek links, ENN and the proposed technique d e f and g. In addition, our proposed hybrid technique h (the proposed technique I + SMOTE) and technique i (the proposed technique II + SMOTE) present the best and second best results in every test set when measured by G-mean, and AUC, while technique c. (SMOTE) ranked the third in these experiments. Furthermore, our hybrid techniques improve the classification accuracy of the minority class (sensitivity) significantly in every data set around 20-45% when compared to that of the original data sets. Moreover, we notice that some techniques which perform better on balancing ratio between the minority and majority classes can present the better results on both G-mean and AUC score. For example, technique f and g, which perform under-sampling on only the majority class, show better performance results than technique d and e, which perform under-sampling on both the minority class and the majority class.

In order to explain why our hybrid techniques (h and i) outperform other techniques, the characteristics of the hybrid techniques need to be discussed. On one hand, SMOTE technique gains the benefits of avoiding the over-fitting problem of the minority class by interpolating new minority class instances rather than duplicating the existing instances [1]. On the other hand, the misclassification analysis using CMTNN can enhance the quality of the training data by removing possible misclassification patterns from the majority class. Furthermore, our hybrid techniques perform re-balancing up to 1:1 ratio between the minority and the majority classes while the other techniques still show a number of imbalanced instances between both classes.

When the hybrid technique h is compared to the hybrid technique i, we found that in most experiments technique i performs better than technique h This is because technique i eliminates all possible misclassification instances from the majority class before the over-sampling technique is applied to the minority class while technique h removes only the possible misclassification instance from the majority class. Therefore, technique h has a high probability to retain bad patterns in the majority class.

For further study, in order to generalize our hybrid technique for the class imbalance problem, other machine learning algorithms such as k-Nearest Neighbor (k-NN), decision tree, and Support Vector Machine (SVM) can be implemented to examine the effect of the different algorithm.

4. CONCLUSIONS

This paper presents the proposed misclassification technique to re-distribute the data in classes to solve the class imbalance problem. This paper uses ANN as the core technique for classification. The CMTNN is applied to detect misclassification patterns. For the proposed technique I, training data is downsized by eliminating only the misclassification patterns discovered by both the Truth NN and Falsity NN. For technique II, the training data is downsized by eliminating all misclassification patterns discovered by the Truth NN and Falsity NN. These two techniques are applied for under-sampling either the majority class or the minority classes. The hybrid techniques are also proposed by combining our under-sampling techniques and the over-sampling technique, Synthetic Minority Over-sampling Technique (SMOTE). After the training sets are re-distributed by several techniques, neural network classifiers are trained by new training data sets. The ANN classifiers are then evaluated and compared in terms of their performances using the widely accepted measures for the class imbalance problem, which are G-mean and AUC. Finally, the results obtained from the experiment indicated that the hybrid technique using the proposed under-sampling technique and SMOTE present the best performance in this experiment.

REFERENCES

[1] 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.

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

[3] I. Tomek, "Two Modifications of CNN," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 6, pp. 769-772, 1976.

[4] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: One-sided selection," in *the Fourteenth International Conference on Machine Learning*, 1997, pp. 179-186.

[5] P. Hart, "The condensed nearest neighbor rule " *Information Theory, IEEE Transactions on*, vol. 14, pp. 515-516, 1968.

[6] D. L. Wilson, "Asymptotic properties of nearest neighbor rules using edited Data," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 2, pp. 408-421, 1972.

[7] 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*: Springer-Verlag, 2001.

[8] 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.

[9] J. Gu, Y. Zhou, and X. Zuo, "Making class bias useful: A strategy of learning from imbalanced data," in *Intelligent Data Engineering and Automated Learning - IDEAL 2007*, 2007, pp. 287-295.

[10] R. Barandela, J. S. Sanchez, V. Garcia, and E. Rangel, "Strategies for learning in class imbalance problems," *Pattern Recognition*, vol. 36, pp. 849-851, 2003.

[11] 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.

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

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

[14] P. Kraipeerapun and C. C. Fung, "Comparing performance of interval neutrosophic sets and neural networks with support vector machines for binary classification problems," in *Digital Ecosystems and Technologies, 2008. DEST 2008. 2nd IEEE International Conference on*, 2008, pp. 34-37.