FUZZY AND SMOTE RESAMPLING TECHNIQUE FOR IMBALANCED DATA SETS

Maisarah Zorkeflee[1], Aniza Mohamed Din[2], and Ku Ruhana Ku-Mahamud[3]

[1]*Universiti Utara Malaysia, Malaysia, maizsarah@gmail.com*
[2]*Universiti Utara Malaysia, Malaysia, anizamd@uum.edu.my*
[3]*Universiti Utara Malaysia, Malaysia, ruhana@uum.edu.my*

ABSTRACT. There are many factors that could affect the performance of a classifier. One of these factors is having imbalanced datasets which could lead to problem in classification accuracy. In binary classification, classifier often ignores instances in minority class. Resampling technique, specifically, undersampling and oversampling are the techniques that are commonly used to overcome the problem related to imbalanced data sets. In this study, an integration of undersampling and oversampling techniques is proposed to improve classification accuracy. The proposed technique is an integration between Fuzzy Distance-based Undersampling and SMOTE. The findings from the study indicate that the proposed combination technique is able to produce more balanced datasets to improve the classification accuracy.

Keywords: imbalanced data, fuzzy logic, fuzzy distance-based undersampling, SMOTE

INTRODUCTION

Data sets are imbalanced if distribution of samples in two classes is unequal. These classes are known as minority and majority classes. Imbalanced data sets can be found in many cases such as credit card fraud detection (Padmaja, Dhuliphalla, Bapi & Laha, 2007), flood prediction (Segretier, Clergue, Collard & Izquierdo, 2012) and stroke prediction (Ou-Yang, Rieza, Wang, Juan & Huang, 2013). Since the size of minority class is lesser than majority class, classifiers will only classify the majority class which will cause high error rate on the minority class (Li, Zou, Wang & Xia, 2013).

Approaches used to handle imbalanced data sets can be categorised as algorithm level approach and data level approach. In the algorithm level approach, existing algorithm is modified in order to recognise instances in minority class (Mahdizadeh & Eftekhari, 2013). The drawback of this approach is its dependency towards classifiers and difficulty to handle (Sahare & Gupta, 2012). At data level approach, data sets are modified by adding instances to minority class or remove instances from majority class. This approach aims to produce balanced data sets (Jeatrakul & Wong, 2012). Data level approach is easier to be used as compared to algorithm level approach because data sets are mended before they are trained by classifiers (Chawla, 2010).

Resampling technique, specifically, undersampling and oversampling techniques are categorised under data level approach. Even though this approach is better than the algorithm level approach, there are several drawbacks found in both techniques. There is a possibility

that useful data might be lost through undersampling technique as it removes data randomly from the majority class while oversampling technique creates overfitting because it adds new instances to minority class (Seiffert, Khoshgoftaar, Van Hulse & Napolitano, 2010). These problems can affect classification accuracy.

Previous works have shown that a combination of undersampling and oversampling technique can improve classification accuracy (Li et al., 2013; Bekkar & Alitouche, 2013). The strategy of combining these techniques is by using their advantages to overcome the deficiency of each technique. This paper proposed a new technique to combine the undersampling and oversampling technique.

This paper is divided into five sections. In Section 2, several related works at data level approach are discussed. Discussion on the proposed integration technique is presented in Section 3. Section 4 provides the details of experiments and discussion of results. Finally, conclusion is provided in Section 5.

RELATED WORK

Binary classification aims to categorise instances in any given sets into two targeted classes. For imbalanced data sets, these two classes are divided into minority and majority classes. Problem occurs when classifiers disregard the minority class which may lead to misclassification. Thus, to overcome this problem, undersampling and oversampling techniques have been developed.

Tomek is introduced to remove instances from majority class and clean the data from noise (Tomek, 1976). Edited Nearest Neighbour Rule (ENN) classifies samples using 3-Nearest Neighbour to form a reference set and any misclassified samples are removed (Wilson, 1972). Distance-based Undersampling (DUS) is a technique that discards instances by calculating the average distance between instances in minority and majority class (Li et al., 2013). Synthetic Minority Oversampling Technique (SMOTE) is the commonly used oversampling technique that creates new synthetic samples to minority class by finding k-nearest neighbour along minority class (Chawla, Bowyer & Hall, 2002).

A combination of SMOTE and Tomek was proposed to oversample the minority class using SMOTE and to remove noise from data sets (Batista, Bazzan & Monard, 2003). There was also a combination of SMOTE and ENN (Batista, Prati & Monard, 2004) which worked similarly as SMOTE and Tomek. However, ENN removes more instances as compared to Tomek. SMOTE-FRST (Ramentol, Verbeist, Bello, Caballero, Cornelis & Herrera, 2012) is a combination of SMOTE and fuzzy rough set theory. This technique applied SMOTE to balance the data set and used fuzzy rough set theory to edit the majority and synthetic instances created by SMOTE.

Improved SMOTE (ISMOTE) and DUS work simultaneously to create balanced data sets (Li et al., 2013). The ratio of new instances created to instances discarded is 1:1. Comparison between combination of ISMOTE and DUS with standalone undersampling and oversampling techniques has been made and the results showed that it performed better than standalone techniques (Li et al., 2013). Complementary Neural Network integrated with SMOTE (Jeatrakul, Wong & Fung, 2010) also gave better performance when compared to single technique.

For evaluation purpose, accuracy is not suitable to be used for imbalanced data sets because the minority class gives smaller impact compared with the majority class. Alternatively, Geometric mean (G-mean) and F-measure are used to evaluate classification performance for imbalanced data sets (He & Garcia, 2009). G-mean is suitable because it is independent to-

wards imbalanced distribution (Jeatrakul, 2010). F-measure is a combination of precision and recall that shows the effectiveness of a classifier (He & Garcia, 2009).

PROPOSED FUZZY DISTANCE-BASED UNDERSAMPLING AND SMOTE

This paper proposed a combination of Fuzzy Distance-based Undersampling (FDUS) and SMOTE techniques to improve the classification performance. The proposed technique is illustrated in Figure 1.

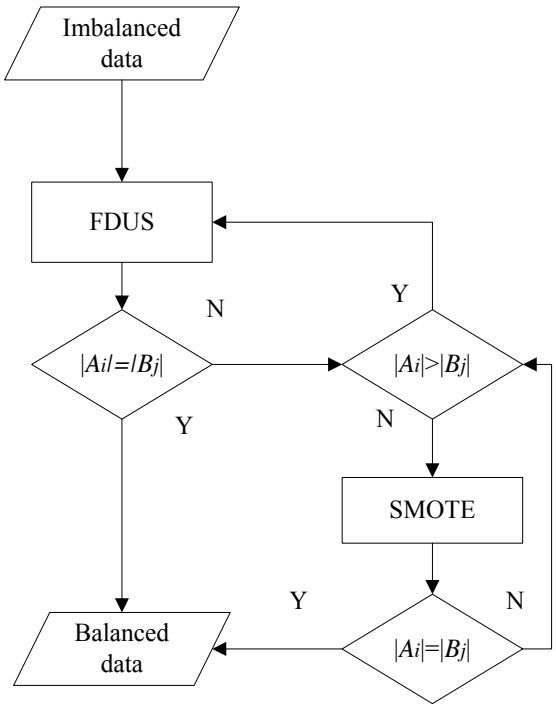

Figure 1. Flowchart of the Proposed FDUS+SMOTE

The flow starts by taking the imbalanced data set as input data. The imbalanced data set is divided into two classes. Initially, let A_i be the majority class and B_j be the minority class. An imbalanced data set is resampled using FDUS technique to produce a balanced data set. But if the number of instances in the majority class, $|A_i|$, is still greater than the number of instances in the minority class, $|B_j|$, then FDUS is repeated. However, if $|A_i|$ has become lesser than $|B_j|$, at this stage, A_i be the minority class, and B_j be the majority class. Then, the data set is resampled using SMOTE. The process is repeated until a balanced data set is produced.

Step 1: Compute distance, d, between instances x in A and y in B.
Step 2: Compute fuzzy logic for d. Triangular and trapezoidal membership function is build. The linguistic variables are defined as 'keep', 'remove temporarily' and 'remove permanently'.
Step 3: Categorise instances in B based on the 'keep', 'remove temporarily' and 'remove permanently' sets.
Step 4: Remove instances in 'remove permanently' set. New data set is generated.

Figure 2. Algorithm of Fuzzy Distance-based Undersampling Technique

FDUS is an undersampling technique that applies fuzzy logic to discard selected samples from the majority class. The algorithm for this technique is provided in Figure 2 where *A* and *B* represent the minority and majority class, respectively.

Step 1: For each instance x in A, find k-nearest neighbour.

Step 2: Select samples randomly from k-nearest neighbour instances and denote as y in P.

Step 3: Find $q=x+difference*gap$ where $difference=x-y$ and gap is random number from 0 to 1.

Step 4: Add q to A. Finally, new data set is created.

Figure 3. Algorithm of SMOTE

The oversampling technique, SMOTE (Chawla et al., 2002) randomly creates new synthetic samples that will be added to the minority class set to create balanced data sets. The algorithm is shown in Figure 3 where *A* is the minority class set. FDUS reduces bias in selecting samples that need to be removed and SMOTE avoids overfitting. These advantages allow the proposed FDUS+SMOTE to create better classification performance.

EXPERIMENT AND RESULT

All experiments were performed using Matlab 2013b. Three imbalanced data sets namely bupa, haberman and pima were selected from UCI machine learning repository (Bache & Lichman, 2013). Table 1 shows the number of instances, number of attributes, number of instances in minority and majority class, and the imbalance ratio between two classes. The imbalanced ratio is defined as the ratio of number of instances in majority class to the number of instances in minority class (Mahdizadeh & Eftekhari, 2013).

Table 1. Characteristics of Data Sets

Data sets	No. of instances	No. of attributes	Minority class	Majority class	Ratio (maj/min)
Bupa	345	7	145	200	1.379
Haberman	306	4	81	225	2.778
Pima	768	8	268	500	1.866

Ten-cross validation is used to split the data sets into 80% training set and 20% testing set. This technique is used to avoid inconsistent results. The proposed Fuzzy Distance-based Undersampling and SMOTE (FDUS+SMOTE) is tested on the imbalanced data sets. Then, the resampled data sets are classified by Support Vector Machine (SVM) and classification performance is evaluated by F-measure and G-mean. The performance of the proposed technique is compared with two combination techniques and two standalone techniques namely SMOTE+Tomek, SMOTE+ENN, FDUS+SMOTE respectively. SMOTE+Tomek and SMOTE+ENN are chosen because they have been widely applied to handle imbalance data sets (Jeatrakul & Wong, 2012). Results are presented in Table 2 and Table 3.

Table 2 shows that the proposed FDUS+SMOTE works better than the other techniques under F-measure evaluation. SMOTE+Tomek and SMOTE+ENN performed better than SMOTE but not FDUS. FDUS produced better result than SMOTE, SMOTE+Tomek and SMOTE+ENN due to its ability to avoid bias in removing instances from majority class. F-

measure shows relation between precision and recall independently. It can be seen that higher proportion of positive instances are correctly classified with higher percentage of F-measure.

Table 2. F-Measure for Each Data Set (%)

Data sets	Bupa	Haberman	Pima
Proposed FDUS + SMOTE	79.69	85.71	81.59
SMOTE+Tomek	61.02	63.27	59.37
SMOTE+ENN	71.43	79.29	66.31
FDUS	88.41	80.00	60.09
SMOTE	40.00	66.02	58.33

Table 3. G-Mean for Each Data Set (%)

Data sets	Bupa	Haberman	Pima
Proposed FDUS + SMOTE	80.70	85.11	65.44
SMOTE+Tomek	71.34	72.28	61.58
SMOTE+ENN	72.88	73.36	69.97
FDUS	79.01	69.28	62.32
SMOTE	52.52	74.47	62.71

In Table 3, the proposed FDUS+SMOTE produced the best G-mean. SMOTE achieved better G-mean than FDUS for Haberman and Pima data sets. SMOTE+Tomek and SMOTE+ENN gave better G-mean than FDUS and SMOTE. Results show that the combination of undersampling and oversampling technique is better than standalone techniques. G-mean represents accuracy of majority and minority class. Higher accuracy of both classes is obtained with higher percentage of G-mean.

Overall, this experiments proved that the FDUS+SMOTE gave good performance result. SMOTE technique has the benefit of avoiding overfitting problem whereas FDUS technique maintains the data quality by reserving useful data. The combination of these advantages created better performance result.

CONCLUSION

This paper proposed a combination between oversampling and undersampling technique to overcome classification problem in handling imbalanced data sets. Results obtained from the experiments showed that FDUS+SMOTE performed better than other resampling techniques and each standalone technique. Performance metric used are F-measure and G-mean that are known as to be suitable to evaluate imbalanced data sets.

ACKNOWLEDGMENTS

The authors wish to thank the Ministry of Education, Malaysia for funding this study under the Long Term Research Grant Scheme (LRGS/b-u/2012/UUM/Teknologi Komunikasi dan Infomasi).

REFERENCES

Bache, K, & Lichman, M. (2013). *UCI Repository of Machine Learning Databases*. Irvine, CA: University of California, School of Information and Computer Science.

Batista, G. E., Bazzan, A. L., & Monard, M. C. (2003). Balancing training data for automated annotation of keywords: A case study. *WOB*, 10-18.

Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, 6(1), 20-29.

Bekkar, M., & Alitouche, T. A. (2013). Imbalanced data learning approaches. *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, 3(4), 15–33.

Chawla, N. V, Bowyer, K. W., & Hall, L. O. (2002). SMOTE : Synthetic Minority Over-sampling Technique, 16, 321–357.

Chawla, N. V. (2010). Data mining for imbalanced data sets: An overview. *Data Mining and Knowledge Discovery Handbook*, pp. 875-886. Springer US.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284.

Jeatrakul, P., & Wong, K. W. (2012). Enhancing classification performance of multi-class imbalanced data using the OAA-DB algorithm. *The 2012 International Joint Conference on Neural Networks (IJCNN)*, 1–8, IEEE.

Jeatrakul, P., Wong, K. W., & Fung, C. C. (2010). Classification of imbalanced data by combining the complementary neural network and SMOTE algorithm. *Neural Information Processing. Models and Applications*, 152-159.

Mahdizadeh, M., & Eftekhari, M. (2013). Designing fuzzy imbalanced classifier based on the subtractive clustering and genetic programming. *Iranian Conference on Fuzzy Systems (IFSC)*, 8–13.

Li, H., Zou, P., Wang, X., & Xia, R. (2013). A new combination sampling method for imbalanced data. *Proceedings of 2013 Chinese Intelligent Automation Conference*, 547-554. Springer Berlin Heidelberg.

Ou-Yang, C., Rieza, M., Wang, H.-C., Juan, Y.-C., & Huang, C.-T. (2013). Applying a hybrid data preprocessing methods in stroke prediction. In Y.-K. Lin, Y.-C. Tsao, & S.-W. Lin (Eds.), *Proceedings of the Institute of Industrial Engineers Asian Conference 2013*, 1441–1449.

Padmaja, T. M., Dhulipalla, N., Krishna, P. R., Bapi, R. S., & Laha, A. (2007). An unbalanced data classification model using hybrid sampling technique for fraud detection. In *Pattern Recognition and Machine Intelligence*, 341-348.

Ramentol, E., Verbiest, N., Bello, R., Caballero, Y., Cornelis, C., & Herrera, F. (2012). SMOTE-FRST: a new resampling method using fuzzy rough set theory. In *10th International FLINS conference on uncertainty modelling in knowledge engineering and decision making*.

Sahare, M., & Gupta, H. (2012). A review of multi-class classification for imbalanced data. *International Journal of Advanced Computer Research*, 2(5), 160–164.

Segretier, W., Clergue, M., Collard, M., & Izquierdo, L. (2012). An evolutionary data mining approach on hydrological data with classifier juries. *2012 IEEE Congress on Evolutionary Computation*, 1–8.

Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2010). RUSBoost: A hybrid approach to alleviating class imbalance. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 40(1), 185-197.

Tomek, I. (1976). An experiment with the edited nearest-neighbor rule. *IEEE Transaction on System, Man and Cybernetics*, 6(6), 448–452.

Wilson, D. L. (1972). Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. *IEEE Transactions on Systems, Man, and Cybernetics*, 2(3), 408–421.