Fuzzy Distance-based Undersampling Technique for Imbalanced Flood Data

Ku Ruhana Ku Mahamud, Maisarah Zorkeflee, and Aniza Mohamed Din

Universiti Utara Malaysia, Malaysia, {ruhana; maizsarah; anizamd}@uum.edu.my

ABSTRACT

Performances of classifiers are affected by imbalanced data because instances in the minority class are often ignored. Imbalanced data often occur in many application domains including flood. If flood cases are misclassified, the impact of flood is higher than the misclassification of non-flood cases. Numerous resampling techniques such as undersampling and oversampling have been used to overcome the problem of misclassification of imbalanced data. However, the undersampling and oversampling techniques suffer from elimination of relevant data and overfitting, which may lead to poor classification results. This paper proposes a Fuzzy Distance-based Undersampling (FDUS) technique to increase classification accuracy. Entropy estimation is used to generate fuzzy thresholds which are used to categorise the instances in majority and minority classes into membership functions. The performance of FDUS was compared with three techniques based on F-measure and G-mean, experimented on flood data. From the results, FDUS achieved better F-measure and G-mean compared to the other techniques which showed that the FDUS was able to reduce the elimination of relevant data.

Keywords: imbalanced flood data, resampling technique, fuzzy distance-based undersampling, fuzzy logic.

I INTRODUCTION

Minority and majority data classes that exist in any imbalanced data sets can be found in many cases including flood prediction. For binary classification, data set is defined as imbalanced if the ratio of two classes is not less than 19:1 (Ding, 2011). The problem that is related to imbalanced data is poor classification performance. Since the size of minority class is lesser than majority class, classifiers will only classify the majority class which leads to poor accuracy because classifiers assume that the distribution of data in both classes is equal (Li, Zou, Wang & Xia, 2013). Hence, to overcome this problem, undersampling and oversampling techniques have been developed.

Random undersampling (RUS) technique is one of the undersampling techniques. RUS removes the instances in majority class randomly until the desired ratio of balanced set is achieved. RUS is easy to be used; however, the random data removal may lead to the loss of useful data (Chairi, Alaoui & Lyhyaoui, 2012). Distance-based Undersampling (DUS) is a technique that discards instances by averaging the distance between instances in minority and majority classes (Li et al., 2013).

In order to estimate class distribution between samples in majority and minority classes, fuzzy logic has been introduced in undersampling technique (Li, Liu & Hu, 2010). The membership function for majority class is based on Gaussian function and α-cut to remove the instances. To deal with large data sets, fuzzy logic is applied to cluster the samples in the majority class to make a selection of which instances are important (Wong, Leung & Ling, 2014). However, the setting of the membership function depends on the calculation of mean value which is sensitive to skewed data sets.

Random oversampling (ROS) randomly duplicates the samples in the minority class. The drawback of ROS is it creates overfitting (Chairi et al., 2012). Synthetic Minority Oversampling Technique (SMOTE) is the commonly used oversampling technique that creates new synthetic samples to the majority class by finding k-nearest neighbour along the minority class (Chawla, Bowyer & Hall, 2002). Results from several experiments conducted showed that undersampling technique produced better classification accuracy than oversampling technique (Bekkar & Alitouche, 2013).

For evaluation purpose, accuracy is not suitable to be used for imbalanced data sets because the minority class has a small impact to the classifier. Instead, Geometric mean (G-mean) and F-measure are used to evaluate the classification performance for imbalanced data sets (He & Garcia, 2009). G-mean is suitable because it is independent towards imbalanced distribution, while F-measure is a combination of precision and recall that shows the effectiveness of a classifier.

In Section 2, an explanation on the proposed undersampling technique is presented. Section 3 describes the performance evaluation and the conclusion is provided in Section 4.

II THE PROPOSED FUZZY DISTANCE-BASED UNDERSAMPLING TECHIIQUE

Fuzzy Distance-based Undersampling (FDUS) technique is an enhancement of Distance-based Undersampling (DUS) by implementing fuzzy logic to the algorithm. Figure 1 shows the algorithm of the proposed FDUS technique to remove instances from majority class.

i. Divide data into majority and minority group
ii. For all data, calculate distance between majority and minority data
iii. Categorize the calculated distance based on fuzzy threshold
iv. Compute fuzzy logic threshold using entropy estimation
v. Remove instances based on trapezoidal and triangular membership functions

Figure 1. Fuzzy Distance-based Undersampling algorithm

Figure 2 illustrates the membership function of the instances. The trapezoidal and triangular membership functions in the figure represent three sets of instances to show the instances that needed to be kept, removed temporarily or removed permanently. Fuzzy logic thresholds are represented as *a*, *b* and *c*.

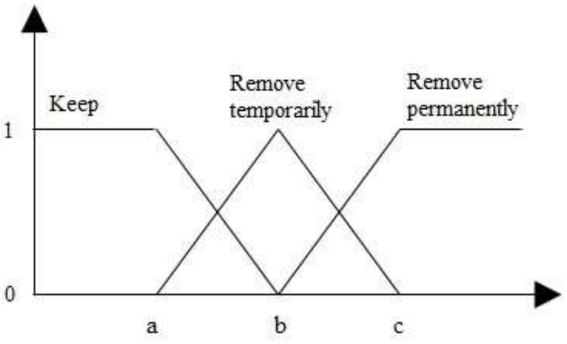

Figure 2. Membership Function of Instances

For instances that belong to the 'keep' set, the instances will remain in the majority class. The 'remove permanently' set represents the instances that will be removed immediately. At this stage, a new majority class is created. For instances that is categorised in 'remove temporarily', the decision of removing the instances will be based on two conditions. These conditions are applicable after considering the size of the new majority class. The first condition is when the number of instances in the new majority class is more than the instances in the minority class. In this case, the instances in the 'remove temporarily' set will be removed immediately. For the second condition, if the number of instances in the new majority class is lesser than the minority class, then the instances will be kept. Finally, new data set with minimal loss of potential data is generated. A balanced data set is produced based on fuzzy thresholds.

III PERFORMANCE EVALUATION

The experiments conducted are designed to minimise the removal of potential data from the majority class. In this paper, data sets from Kaki Bukit, Lubok Sireh, Wang Kelian, Ladang Perlis Selatan and Ulu Pauh from year 2005 until 2013 are used. The pre-processed imbalanced flood data are divided into majority and minority classes. Then, 5-fold cross validation is used to partition the data sets into 4:1 train to test ratio. The proposed FDUS is applied on the training and testing sets and SVM is used for classification. The classification is evaluated by accuracy, F-measure and G-mean. The results from five experiments for each data set are averaged. For comparison purposes, the whole process is repeated using different techniques, which are DUS and SMOTE. Testing is also made on the data sets without applying any undersampling or oversampling technique to analyse whether the use of those techniques are beneficial.

Table 1 is a sample of rainfall and water level data. Rainfall is measured in milimeter (mm), while water level is measured in meter (m) unit. Both rainfall and water level are presented in hourly forms.

Table 1. Sample of Rainfall and Water Level Data

Rainfall Data (mm) for Sungai Pelarit					
	0100	**0200**	**0300**	**…**	**2400**
1/12/13	0	0	0		0
Water Level Data (m) for Wang Kelian					
	0100	**0200**	**0300**	**…**	**2400**
1/12/13	10.21	10.24	10.24	…	10.27

Collected rainfall and water level data are cleaned up from any outliers. For this case, any point that is separated far from other points is considered as outliers. To deal with outliers, the points are corrected by replacing a close approximation point of the remaining values. In order to fill the missing value, interpolation technique is used as described in Equation 1.

$$f(x) = f(x_0) + (x - x_0)\frac{f(b) - f(a)}{b - a} \qquad (1)$$

where *f(x)* = estimation value, *f(x₀)* = value before missing value, x = point of missing value, x_0 =

point of value before missing value, *f(a)* = constant value before missing value, *f(b)* = constant value after missing value, a = constant point before missing value and b = constant point after missing value.

After data cleaning, rainfall and water level data sets are combined. These two attributes will determine the flood occurrence for each catchment area. Table 2 shows the relations of rainfall and water level stage that cause floods (Bedient, Huber & Vieux, 2008). Table 3 presents a sample of flood data set after the rainfall and water level data are combined. The division of no flood and flood classes are done based on Table 2.

Table 2. Causes of Flood (Bedient, Huber & Vieux, 2008)

	Rainfall	Water level stage	Class
Stage	Heavy or very heavy	Warning or danger	Flood
	Heavy or very heavy	Alert	Flood
	Light or moderate	Warning or danger	Flood
	Light or moderate	Alert	No flood

Table 3. Sample of Ulu Pauh Data Set

Date	Time	Rainfall (mm)	Water level (m)	Class
29/3/2009	2.00pm	0	25.67	No flood
29/3/2009	4.00pm	67.30	25.72	Flood
29/3/2009	6.00pm	0.10	28.12	No flood

Table 4 provides details of the flood data sets that include size of the data sets, number of instances in flood class (#Flood), number of instances in no flood class (#No flood), and ratio of majority class to minority class. The imbalanced ratio is defined as the ratio of number of instances in majority class to the number of instances in minority class. Minority and majority classes represent flood and no flood occurrence, respectively.

The results of classification accuracy for no resampling technique, FDUS, DUS and SMOTE are presented in Table 5. FDUS produced the best mean classification accuracy on Kaki Bukit and Ulu Pauh and produced the second best mean classification accuracy on Ladang Perlis. The average of mean classification accuracy and the standard deviation for FDUS are the highest compared to no resampling, DUS and SMOTE. However, even though the standard deviation is ranked as the highest, the value is considered low as stated in Orriols-Puig and Bernado-Mansilla (2009).

Table 6 shows the F-measure for the proposed FDUS and other resampling techniques. FDUS performed the best when it is applied on Wang Kelian and Ulu Pauh data sets compared to the other techniques. For the rest of the data sets, FDUS performed as the second best technique. On average, FDUS gave the best F-measure.

The results of G-mean for flood data sets are summarised in Table 7. The results show that FDUS worked better than DUS and SMOTE for Kaki Bukit, Lubok Sireh, Wang Kelian and Ulu Pauh. On average, FDUS performed as the second best technique after no resampling.

The results of classification accuracy indicated that FDUS allows SVM to classify the data sets correctly specifically on the Kaki Bukit and Ulu Pauh data sets. The classification accuracy is higher on Kaki Bukit and Ulu Pauh data sets because the ratio between majority and minority classes has become smaller when FDUS is applied on the data sets. However, for the other flood data sets, FDUS has lower classification accuracy than no resampling, DUS and SMOTE. This might happen due to other factors such as size, complexity, overlap and small disjuncts (Barua, Islam, Yao, & Murase, 2014).

F-measure determines the exactness of the correctly labelled minority class. Based on Table 6, FDUS appeared as the best technique for two times and second best technique for three times. FDUS is able to adjust the ratio between instances in minority class to instances in majority class to maximize the value of F-measure. High G-mean signifies the accuracy of majority and minority classes is high and the gap between both classes is small. FDUS performed better than DUS and SMOTE. However, FDUS is outperformed by no resampling because the sensitivity and specificity are high. FDUS uses the advantage of fuzzy logic to avoid biasness in choosing the instances that need to be removed from the majority class. Overall, it is apparent that FDUS achieved higher classification accuracy and F-measure and has the highest G-mean.

Table 4. Characteristics of Flood Data Sets

Data sets	Record size	#Flood	#No flood	Ratio (maj:min)
Kaki Bukit	157,775	75	157,700	2102:1
Lubok Sireh	157,775	75	157,700	2102:1
Wang Kelian	157,775	76	157,699	2074:1
Ladang Perlis Selatan	157,775	163	157,612	966:1
Ulu Pauh	157,775	128	157,617	1231:1

Table 5. Classification Accuracy (%) of Standalone Techniques for Flood Data Sets

Resampling technique	No resampling		FDUS		DUS		SMOTE	
Data set	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
Kaki Bukit	99.90	1.87	99.94	0.61	99.67	0.06	99.88	0.19
Lubok Sireh	99.89	0.22	99.80	1.96	99.94	0.56	99.97	0.19
Wang Kelian	99.70	0.34	99.82	1.98	99.96	0.49	99.84	0.20
Ladang Perlis	99.89	0.22	99.94	0.61	99.95	0.63	99.89	0.39
Ulu Pauh	99.60	0.23	99.99	1.22	99.95	0.64	99.89	0.22
Average	99.80	0.58	99.90	1.28	99.89	0.48	99.89	0.24

Table 6. F-measure of Standalone Techniques for Flood Data Sets

Resampling technique	No resampling	FDUS	DUS	SMOTE
Data set				
Kaki Bukit	0.49	0.81	0.85	0.81
Lubok Sireh	0.48	0.84	0.74	0.87
Wang Kelian	0.49	0.85	0.84	0.53
Ladang Perlis	0.65	0.81	0.92	0.79
Ulu Pauh	0.65	0.99	0.87	0.75
Average	0.55	0.86	0.84	0.75

Table 7. G-mean of Standalone Techniques for Flood Data Sets

Resampling technique	No resampling	FDUS	DUS	SMOTE
Data set				
Kaki Bukit	0.99	0.93	0.87	0.90
Lubok Sireh	1.00	0.97	0.82	0.96
Wang Kelian	0.99	0.99	0.96	0.90
Ladang Perlis	1.00	0.93	0.98	0.90
Ulu Pauh	0.99	0.99	0.88	0.90
Average	0.99	0.96	0.90	0.91

IV CONCLUSION

Undersampling technique is chosen to solve the problem of imbalanced data sets, because based on previous research works, the technique performed better than oversampling technique. In this paper, Fuzzy Distance-based Undersampling (FDUS) technique is proposed. FDUS used the advantage of fuzzy logic which is to avoid bias in removing instances in the majority class, and hence minimise the loss of useful data. Based on the experimental results, FDUS produced the best classification accuracy and F-measure on the flood data sets. Based on G-mean value, FDUS is better than DUS and SMOTE but performed lesser than no resampling.

ACKNOWLEDGMENT

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

Barua, S., Islam, M., Yao, X., & Murase, K. (2014). MWMOTE--majority weighted minority oversampling technique for imbalanced data set learning., *IEEE Transactions on Knowledge and Data Engineering, 26*(2), 405-425.

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

Chairi, I., Alaoui, S., & Lyhyaoui, A. (2012). Learning from imbalanced data using methods of sample selection. In *2012 International Conference on Multimedia Computing and Systems (ICMCS)*, 254-257.

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

Ding, Z. (2011). Diversified ensemble classifiers for highly imbalanced data learning and their application in bioinformatics, in *Computer Science Dissertations*, Department of Computer Science, Georgia State University.

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

Li, D.-C., Liu, C.-W., & Hu, S. C. (2010). A learning method for the class imbalance problem with medical data sets. *Computers in Biology and Medicine, 40*(5), 509-518.

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.

Orriols-Puig & Bernado-Mansilla (2009). Evolutionary rule-based systems for imbalanced data sets. *Soft Computing, 13*(3), 213-225.

Wong, G. Y., Leung, F. H., & Ling, S. H. (2014). An under-sampling method based on fuzzy logic for large imbalanced data set. In *2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 1248-1252.