A Conceptual Model of Enhanced Undersampling Technique

Maisarah Zorkeflee, Ku Ruhana Ku-Mahamud, and Aniza Mohamed Din
Universiti Utara Malaysia, Malaysia, {s814594, ruhana, anizamd}@uum.edu.my

ABSTRACT

Imbalanced datasets often lead to decrement of classifiers' performance. Undersampling technique is one of the approaches that is used when dealing with imbalanced datasets problem. This paper discusses on the advantages and disadvantages of several undersampling techniques. An enhanced Distance-based undersampling technique is proposed to balance the imbalanced data that will be used for classification. The fuzzy logic has been integrated in the distance-based undersampling technique to resolve the ambiguity and bias issues.

Keywords: Undersampling technique, fuzzy logic, imbalanced data

I INTRODUCTION

Imbalanced datasets is known as datasets that are distributed unequally where instances in one class is larger than other class. They are known as majority and minority class respectively (Garcia, Sanchez, Mollineda, Alejo & Sotoca, 2007). Flood events data (Wang, Chen & Small, 2013), credit card fraud detection data (Padmaja, Dhulipalla, Krishna, Bapi & Laha, 2007) and oil spill identification data (Brekke & Solberg, 2005) are some of the examples of imbalanced datasets.

Instances in minority class represent cases that are rarely happened such as flood occurrence. Therefore, ignorance towards this class may affect society, economy and environment. The problem that is related with imbalanced datasets referring to binary classification issue is classifier ignores the minority class that lead to decrement of classification accuracy (Mi, 2013).

There are two approaches to deal with imbalanced datasets problem. The first approach is known as algorithm level approach where it addresses modification of existing algorithms (Mahdizadeh & Eftekhari, 2013). The modification of algorithm is done so that minority class can be recognize. Several examples of algorithm-based level approach are adjustment of cost of classes, modification of probabilistic estimation of decision trees and alteration of decision threshold (Garcia et al., 2007). The limitation of this approach is its dependency towards classifier and complicated to handle (Sahare & Gupta, 2012).

The second approach is data level approach and its aim is to adjust the datasets to produce a balanced datasets (Jeatrakul & Wong, 2012). Among these two approaches, data-based level approach is easier to handle and more versatile. The reasons are because datasets are modified to produce balanced data sets before classifier is trained and it is independent towards classifiers as compared to the algorithm-based level approach (Chawla, 2010).

Resampling technique is categorised under data level approach. Removal of data samples from majority class is known as undersampling technique and addition of data samples to minority class is known as oversampling technique (Luengo, Fernandez, Garcia & Herrera, 2011). Several undersampling and oversampling techniques have been proposed in dealing with imbalanced datasets. The most basic oversampling technique is random oversampling technique where it duplicates examples from minority class to balance the datasets (Seiffert, Khoshgoftaar, Van Hulse & Napolitano, 2010). An example of undersampling technique is Distanced-based Undersampling technique that is easy to be used (Li, Zou, Wang & Xia, 2013). It removes samples in majority class based on threshold by averaging the distance between samples in minority class and majority class.

Each of undersampling and oversampling technique has its own advantage and disadvantage. A number of studies showed that undersampling technique provides better classification accuracy than oversampling technique (Bekkar & Alitouche, 2013). Oversampling technique uses all data, but it creates overfitting (Chawla, 2010). In contrast with oversampling technique, undersampling technique decreases the time of training process because the size of data has become smaller. However, it may cause loss of useful data (Chawla, 2010). Useful data is important in delivering information to the users where subsequently it becomes knowledge and can be used to make predictions (Waltz, 2003). Therefore, the enhancement of undersampling technique needs to be focusing on preserving the data as maximum as possible.

Several undersampling techniques have been developed to overcome the problem of useful data removal from majority class. However, there is still lacking in making decision to discard the instances from majority class. The implementation of k-Nearest

Neighbour (k-NN) and mean may cause ambiguity and bias (Napierala & Stefanowski, 2012; Whitley & Ball, 2002). These factors will affect classification accuracy.

In this paper, an enhancement of undersampling technique is proposed. To overcome the mentioned problems, fuzzy logic is utilised due to its ability to overcome ambiguity and bias issues. In Section II, some previous works related to techniques in handling imbalanced datasets are discussed. The discussion on the proposed enhanced undersampling technique is presented in Section III. Conclusion and future work are presented in Section IV.

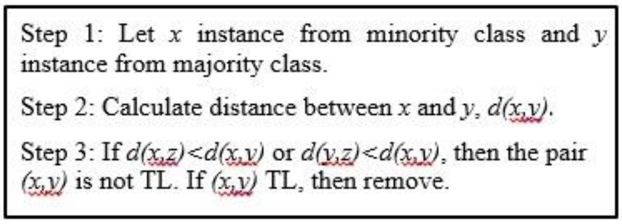
II UNDERSAMPLING TECHNIQUE FOR IMBALANCED DATA

Binary classification aims to categorize elements of given sets to two targeted class. However, inaccuracy of classification occurs when dealing with imbalanced datasets. Imbalanced dataset is a set of two classes that distributed with the ratio of 100 to 1, 1000 to 1 or more (Chawla, Bowyer & Hall, 2002). Class that has higher number of samples is known as majority class while the other is minority class. Decrement of classification accuracy is due to the ignorance of classifier towards instances in minority class. To overcome this problem, oversampling and undersampling techniques are proposed (Chawla, 2010). The aim of these techniques is to do modification to the data instead of the algorithm.

Oversampling technique creates new samples to minority class until required ratio is achieved. The drawback of this technique is it creates overfitting. Hence, produce poor classification performance. Chawla et al. (2002) overcome overfitting by introducing Synthetic Minority Oversampling Technique (SMOTE). Instead of randomly create new samples, SMOTE creates new synthetic samples along k-Nearest Neighbour (k-NN) of minority class. However, oversampling requires high learning time because the size of data is big and the performance of oversampling is lower than undersampling technique (Bekkar & Alitouche, 2013).

Random undersampling (RUS) is one of undersampling techniques that removes instances from majority class randomly. This approach will lead to decrement of classification accuracy due possibility of losing potential useful data (Chairi, Alaoui, & Lyhyaoui, 2012). Seiffert et al. (2010) proposed a repetitive undersampling technique. They generates an ensemble of RUS models in order to get better classification accuracy than RUS. However, due to the randomness, the accuracy of classifier may be improved in a supervised manner (Galar, Fernandez, Barrenechea & Herrera, 2013).

Condensed Nearest Neighbour Rule (CNN) identifies borderline instances (Hart, 1968). However, it includes a big portion of noisy instances (Fitkov-Norris & Folorunso, 2013). To improve CNN, Tomek Links (TL) is introduced (Tomek, 1976). TL chooses samples that are close to the boundary points. This technique not only remove instances from majority class but also clean the data from noise. However, TL has high possibility of discarding potential data because borderline samples can be important in characterising the decision border (Del Gaudio, Batista & Branco, 2013). Figure 1 describes the algorithm of Tomek Links.



Step 1: Let x instance from minority class and y instance from majority class.

Step 2: Calculate distance between x and y, $d(x,y)$.

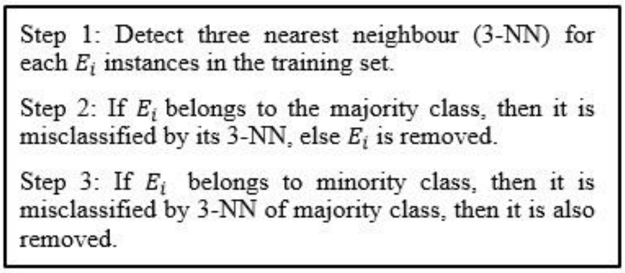
Step 3: If $d(x,z)<d(x,y)$ or $d(y,z)<d(x,y)$, then the pair (x,y) is not TL. If (x,y) TL, then remove.

Figure 1. Algorithm of Tomek Links

One-Sided Selection (OSS) is a combination of two undersampling techniques between TL and CNN (Kubat & Matwin, 1997). Firstly, TL locates noise before CNN is used to identify redundant instances. Then, noise and redundant instances are removed. The weakness of OSS is it requires high learning time (Bekkar & Alitouche, 2013). Reduced Nearest Neighbour (RNN) removes noisy instances while keeping the instances at the border points (Gates, 1971). The disadvantage of RNN is in order to compute the learning set, it requires high learning time.

Wilson’s Edited Nearest Neighbour Rule (ENN) is an enhancement of k-NN to improve 1-NN (Wilson, 1972). Samples are classified using 3-NN rule to form a reference set where three nearest neighbours are identified. Any misclassified samples are removed. ENN removes noise and avoids overfitting (Bekkar & Alitouche, 2012).

Neighbourhood Cleaning Rule (NCL) adapts ENN rule to identify and remove instances from majority class (Laurikkala, 2001). NCL improves ENN by reducing excessive amount of data removal. Figure 2 illustrates steps of NCL. However, the main drawback ENN and NCL is the potential of not producing balanced datasets. This problems due to there is no control to remove patterns of majority class (Garcia et al., 2007).



Step 1: Detect three nearest neighbour (3-NN) for each E_i instances in the training set.

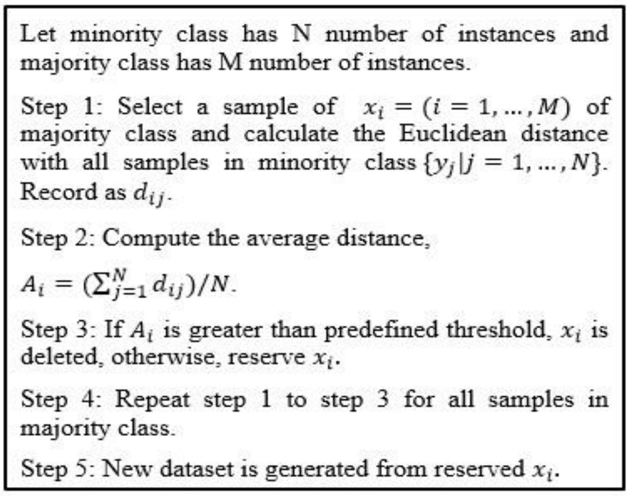
Step 2: If E_i belongs to the majority class, then it is misclassified by its 3-NN, else E_i is removed.

Step 3: If E_i belongs to minority class, then it is misclassified by 3-NN of majority class, then it is also removed.

Figure 2. Algorithm of Neighbourhood Cleaning Rule

Based on reviewed undersampling techniques, k-nearest neighbour identifies the removable of samples in majority class (Zhang, Liu, Gong & Jin, 2011). K-nearest neighbour reduces the bias towards the domination of majority class because the instances in majority class are discarded based on the farthest distance to the *k* nearest neighbour instances in minority class (Garcia, Mollineda & Sanchez, 2008). However, Napierala and Stefanowski (2012) stated that there are cases when the k-nearest neighbours are equally distant from the classified instances that may cause ambiguity.

Distance-based Undersampling (DUS) is a technique of discarding data based on distance calculation (Li et al., 2013). Unlike other undersampling techniques, DUS is easy to handle because it does not consider the boundary samples which are difficult to deal with (Anand, Pugalenthi, Fogel & Suganthan, 2010). The steps of DUS are outlined in Figure 3. However, the drawback of this technique is its biasness towards the majority class. This is due to the implementation of mean in order to identify and ignore the sample of majority class. Mean is very sensitive to skewed data sets, hence, it leads to decrement of classification accuracy (Whitley & Ball, 2002). Therefore, it is not suitable for imbalanced data sets because the result will bias towards the majority class instances.



Let minority class has N number of instances and majority class has M number of instances.

Step 1: Select a sample of $x_i = (i = 1, \dots, M)$ of majority class and calculate the Euclidean distance with all samples in minority class $\{y_j | j = 1, \dots, N\}$. Record as d_{ij}.

Step 2: Compute the average distance,

$A_i = (\sum_{j=1}^{N} d_{ij})/N$.

Step 3: If A_i is greater than predefined threshold, x_i is deleted, otherwise, reserve x_i.

Step 4: Repeat step 1 to step 3 for all samples in majority class.

Step 5: New dataset is generated from reserved x_i.

Figure 3. Algorithm of Distance-based Undersampling

In ambiguity cases, Zadeh (1980) stated that fuzzy logic is suitable to be used as solution. This statement is aligned with other researcher that claimed fuzzy logic has the advantage in solving ambiguity problem (Mahdizadeh & Eftekhari, 2013). The concept of fuzzy logic has been introduced to undersampling technique by Li, Liu and Hu (2010) in order to estimate class distribution between samples in minority and majority class. This approach uses Gaussian function as majority class membership function and α-cut to remove the instances.

Fuzzy set theory aims to reduce complexity without excessive simplification (Singpurwalla & Booker, 2004). In fuzzy set theory, membership function calculates the possible value of essential instances instead of probability in statistic like in Li et al. (2013) to avoid normal distribution assumption (Li, Wu, Tsai & Lina, 2007). Membership function usually be presented in triangle, trapezoidal and Gaussian (Sivanandam, Sumathi & Deepa, 2007). The choice of optimal membership functions have to be considered (Aziz, 2009). In this study, triangular and trapezoidal is used due to their ease of implementation (DeBitetto, 1994).

As pointed out, undersampling technique is one of the techniques to handle imbalanced datasets. Several undersampling techniques have been discussed in this section. To conclude, every technique has its own advantage and disadvantage. Table 1 gives summarization of reviewed undersampling techniques.

Table 1. Undersampling Techniques for Imbalanced Datasets

Technique	Advantage	Disadvantage	Technique to remove instances
RUS	Easy to implement	Lead to big amount of potential data	Randomly remove
RUSBoost (Seiffert et al., 2010)	Less information loss compared to RUS	Randomly choose instances	Repetitive undersampling
CNN (Hart, 1968)	Identify borderline instances	Include noise	k-NN
TL (Tomek, 1969)	Identify instances close to borderline	High possibility remove potential data	k-NN
OSS (Kubat & Matwin, 1997)	Combine two techniques and produce cleaner datasets than CNN and TL	High learning time	k-NN

RNN (Gates, 1971)	Remove noise and keep instances at borderline	High learning time	k-NN
ENN (Wilson, 1972)	Remove noise at boundary points and avoid overfitting	No limitation to remove instances of majority class	k-NN
NCL (Laurikkala, 2001)	Avoid excessive removal of small classes	No limitation to remove instances of majority class	k-NN
DUS (Li et al., 2013)	Easy to implement	Potential to be bias towards majority class	Mean
Fuzzy under sampling (Li et al., 2010)	Avoid wrong assumption of data distribution	Complicated membership function	Fuzzy logic

III PROPOSED ENHANCED DISTANCE-BASED UNDERSAMPLING TECHNIQUE

In this section, the description of the proposed enhanced Distance-based Undersampling (DUS) technique to overcome the imbalanced datasets problem is provided. Fuzzy logic is integrated in DUS to overcome the ambiguity and bias problems.

Figure 4 shows the flowchart of DUS. Imbalanced data sets are divided into two classes and they are denoted as x_i for samples in majority class and y_j for samples in minority class. Then, the distance, d_{ij} between samples in majority class and minority class will be calculated using Euclidean distance followed by the calculation of the mean for the distance which is denoted by A_i. The decision for which samples that need to be removed is based on a predefined threshold as identified in Li et al. (2013). The process is repeated until the distance for all samples are calculated. Finally, balanced data sets are produced.

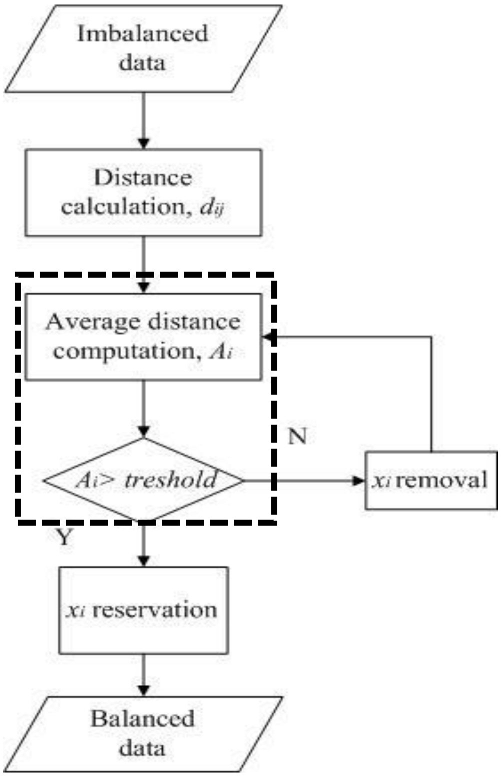
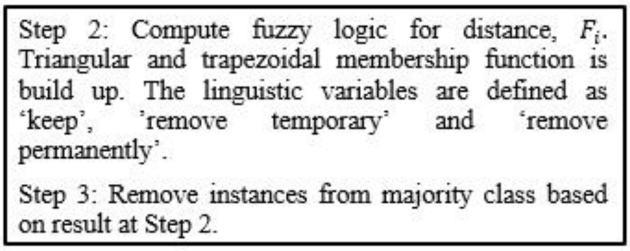


Figure 4. Flowchart of Distance-based Undersampling Technique (Li et al., 2013)

Enhancement of DUS will be done at the dotted box part. The modification of the steps is presented in Figure 5. At step 2, fuzzy logic is introduced to replace the computation of average distance. At this step, membership function is build up. The decision of data removal is based on the membership function as described in Step 3.



Step 2: Compute fuzzy logic for distance, F_i. Triangular and trapezoidal membership function is build up. The linguistic variables are defined as 'keep', 'remove temporary' and 'remove permanently'.

Step 3: Remove instances from majority class based on result at Step 2.

Figure 5. Modification of Distance-based Undersampling

The algorithm for the proposed enhanced undersampling technique is illustrated in Figure 6. The flow starts with removal of any outliers of imbalanced datasets. Then, the imbalanced datasets will be divided into minority and majority class. Let set $k_i = (1, \ldots, p)$ be instances in majority class and $l_j = (1, \ldots, q)$ be instances in minority class.

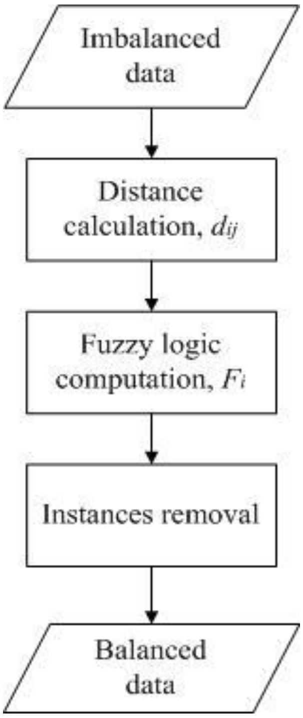

Imbalanced data
Distance calculation, d_{ij}
Fuzzy logic computation, F_i
Instances removal
Balanced data

Figure 6. Flowchart of Proposed Enhanced Undersampling Technique

Distance between all instances in minority and majority class is calculated using Euclidean distance and denoted as d_{ij}. Then, based on the distance, fuzzy logic is computed to categorise the samples into sets of 'keep', 'remove temporary' or 'remove permanently' as illustrated in Figure 7. Based on these categories, the decision of data removal is made. Finally, balanced data sets are produced.

Figure 7 shows the concept of triangular and trapezoidal membership function. The membership function represents instances in the majority class that will be kept, removed temporarily or removed permanently. If the instances belong to the set of 'keep', then the instances will not be discarded. If the instances belong to set of 'remove permanently', the instances will be ignored immediately. At this phase, a new majority class will be produced. For instances that belong to set of 'remove temporary', the decision of removing the instances will be based on two conditions. These conditions are applicable after considering the size of new majority class. The first condition is if the size of new majority class is still bigger than the size of minority class, then the instances in 'remove temporary' set will be ignored. But, if the size of new majority class became smaller than the minority class, then the instances will remain in the majority class.

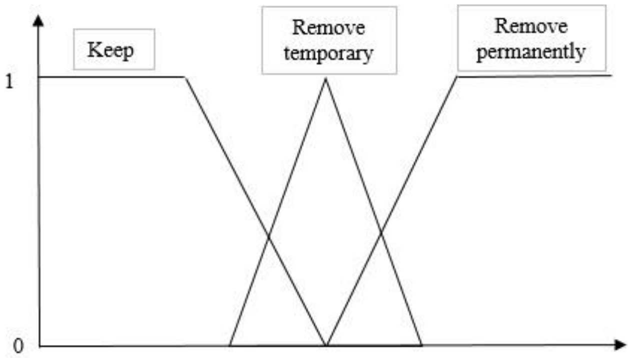


Figure 7. Triangular and Trapezoidal Membership Function

DUS is an easy technique to be implemented when handling with imbalanced datasets. However, it has some lack as mentioned in Section II due to the usage of mean in making decision to discard samples from majority class. Therefore, enhancement of this technique will hopefully improve classification accuracy for imbalanced datasets.

This paper is based on the assumption that fuzzy logic can produce better result on making decision to remove data from majority class. This is due to the advantage of fuzzy logic that is suitable when dealing with bias and ambiguity cases. By proposing fuzzy logic, the tendency of losing useful data will be minimized.

However, the challenge of fuzzy logic implementation to this technique is at the choice to build which type of membership function. In this study, the type of membership function is chosen based on its simplicity as compared to other function such as Gaussian function.

IV CONCLUSION

Datasets are commonly presented in imbalanced distribution which will decrease the classification performance. This implication occurred due to the classifier that neglects the minority class. In most cases, minority class represents important events. Hence, it is important to take into account instances in the minority class.

This study has focused on enhancing the Distance-based Undersampling technique to cater the ambiguity and bias problems. Thus it is hoped that the proposed technique can increase the classification accuracy. Future work will focus on testing the technique with several real imbalanced datasets.

ACKNOWLEDGEMENT

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

Anand, A., Pugalenthi, G., Fogel, G. B., & Suganthan, P. N. (2010). An approach for classification of highly imbalanced data using weighting and under-sampling. Amino acids, 39(5), 1385-1391.

Aziz, A. M. (2009, August). Effects of fuzzy membership function shapes on clustering performance in multisensor-multitarget data fusion systems. In Fuzzy Systems, 2009. FUZZ-IEEE 2009. IEEE International Conference on (pp. 1839-1844). IEEE.

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

Brekke, C., & Solberg, A. H. S. (2005). Oil spill detection by satellite remote sensing. Remote Sensing of Environment, 95(1), 1–13.

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

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

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

DeBitetto, P. A. (1994, July). Fuzzy logic for depth control of unmanned undersea vehicles. In Autonomous Underwater Vehicle Technology, 1994. AUV'94. Proceedings of the 1994 Symposium on (pp. 233-241). IEEE.

Del Gaudio, R., Batista, G., & Branco, A. (2013). Coping with highly imbalanced datasets: A case study with definition extraction in a multilingual setting. Natural Language Engineering, 1–33.

Fitkov-Norris, E., & Folorunso, S. O. (2013). Impact of sampling on neural network classification performance in the context of repeat movie viewing. EANN 2013, Part I, CCIS 383, 213–222.

Galar, M., Fernández, A., Barrenechea, E., & Herrera, F. (2013). Eusboost: Enhancing Ensembles for Highly Imbalanced Data-sets by Evolutionary Undersampling. Pattern Recognition.

García, V., Mollineda, R. a., & Sánchez, J. S. (2008). On the k-NN performance in a challenging scenario of imbalance and overlapping. Pattern Analysis and Applications, 11(3-4), 269–280.

Garcia, V., Sanchez, J. S., Mollineda, R. A., Alejo, R., & Sotoca, J.M. (2007). The class imbalance problem in pattern classification and learning. Congreso Espanol de Informatica (pp. 284–291).

Gates, G. W. (1971). The reduced nearest neighbor rule. IEEE Trans Information Theory, 18(3), 431–433.

Hart, P. E. (1968). The condensed nearest neighboiur rule. IEEE Transactions on Information Theory, 515–516.

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) (pp. 1–8). IEEE.

Kubat, M., & Matwin, S. (1997). Addressing the curse of imbalanced training sets: one-sided selection. Proceedings of the fourteenth conference on machine learning (pp. 179–186).

Laurikkala, J. (2001). Improving identification of difficult small classes by balancing class distribution. University of Tampere, Tech. Rep. A (2).

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, D. C., Wu, C. S., Tsai, T. I., & Lina, Y. S. (2007). Using mega-trend-diffusion and artificial samples in small data set learning for early flexible manufacturing system scheduling knowledge. Computers & Operations Research, 34(4), 966-982.

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

Luengo, J., Fernandez, A., Garcia, S., & Herrera, F. (2011). Addressing data complexity for imbalanced data sets : analysis of SMOTE-based oversampling and evolutionary undersampling. Soft Computing, 15, 1909–1936.

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

Mi, Y. (2013). Imbalanced classification based on active learning SMOTE. Research Journal on Applied Sciences, Engineering and Technology, 5(3), 944–949.

Napierala, K., & Stefanowski, J. (2012). BRACID: A comprehensive approach to learning rules from imbalanced data. Journal of Intelligent Information Systems, 39(2), 335–373.

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 (pp. 341-348). Springer Berlin Heidelberg.

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

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.

Singpurwalla, N. D., & Booker, J. M. (2004). Membership functions and probability measures of fuzzy sets. Journal of the American Statistical Association, 99(467), 867-877.

Sivanandam, S. N., Sumathi, S., & Deepa, S. N. (2007). Introduction to fuzzy logic using MATLAB (Vol. 1). Berlin: Springer.

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

Waltz, E. (2003). Knowledge management in the intelligence enterprise. Artech House.

Wang, D., Chen, P., & Small, D. L. (2013). Towards long-lead forecasting of extreme flood events : a data mining framework for precipitation cluster precursors identification, 1285–1293.

Whitley, E., & Ball, J. (2001). Statistics review 1: Presenting and summarising da-ta. Critical Care, 6(1), 66.

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

Zadeh, L. A. (1980). Fuzzy sets versus probability. Proceedings of the IEEE, 68(3), 421.

Zhang, D., Liu, W., Gong, X., & Jin, H. (2011). A novel improved smote resampling algorithm based on fractal. Journal of Computational Information Systems, 6, 2204–2211.