AN ENHANCED RESAMPLING TECHNIQUE FOR IMBALANCED DATA SETS

MAISARAH BINTI ZORKEFLEE

MASTER OF SCIENCE (INFORMATION TECHNOLOGY)
UNIVERSITI UTARA MALAYSIA
2015

Permission to Use

Abstrak

Set data adalah tidak seimbang apabila sampel data yang terdapat pada satu kelas (kelas majoriti) melebihi kelas selain daripadanya (kelas minoriti). Masalah utama berkaitan dengan data binari tidak seimbang ialah kecenderungan pengelas untuk mengabaikan kelas minoriti. Beberapa teknik persampelan semula seperti pensampelan bawah, pensampelan atas dan gabungan kedua-duanya telah banyak digunakan. Walau bagaimanapun, teknik pensampelan bawah dan pensampelan atas tersebut masih mempunyai kekurangan seperti pembuangan dan penambahan data yang berguna yang menyebabkan masalah ketepatan pengelasan data. Oleh itu, kajian ini bertujuan untuk meningkatkan metrik klasifikasi dengan menambah baik teknik pensampelan bawah dan menggabungkannya dengan teknik pensampelan atas yang telah wujud. Untuk mencapai objektif tersebut, teknik Pensampelan Bawah Berdasarkan Jarak Kabur (FDUS) dicadangkan. Anggaran entropi digunakan untuk menghasilkan ambang kabur untuk mengelaskan sampel di dalam kelas minoriti dengan kelas majoriti kepada fungsi keahlian. FDUS kemudian digabungkan dengan Teknik Pensampelan Atas Minoriti Sintetik (SMOTE) dikenali sebagai FDUS+SMOTE, dilakukan di dalam urutan sehingga data yang seimbang dihasilkan. Kedua-dua teknik, FDUS and FDUS+SMOTE dibandingkan dengan empat teknik yang lain berdasarkan ketepatan klasifikasi, F-ukuran dan G-purata. Berdasarkan keputusan, FDUS mencapai ketepatan klasifikasi F-ukuran dan G-purata yang lebih bagus apabila dibandingkan dengan teknik lain dengan purata masing-masing 80.57%, 0.85 dan 0.78. Ini menunjukkan logik kabur apabila digabungkan dengan teknik Pensampelan Bawah Berdasarkan Jarak mampu mengurangkan penyingkiran data yang berguna. Tambahan, penemuan menunjukkan FDUS+SMOTE menghasilkan prestasi yang lebih baik berbanding gabungan teknik SMOTE dan Pautan Tomek, dan SMOTE dan Penyuntingan Jiran Terdekat pada data penanda aras. FDUS+SMOTE telah mengurangkan pembuangan data yang berguna dari kelas majoriti dan mengelakkan terlebih-padanan. Secara purata, FDUS dan FDUS+SMOTE mampu mengimbangkan data kategorik, integer dan nyata serta membaiki prestasi klasifikasi binari. Selain itu, teknik tersebut menghasilkan prestasi yang baik pada data yang mempunyai saiz rekod kecil yang mempunyai sampel di dalam lingkungan kira-kira 100 ke 800.

Kata kunci: Data tidak seimbang, Teknik persampelan semula, Teknik pensampelan bawah, Teknik pensampelan atas, Logik kabur

Abstract

A data set is considered imbalanced if the distribution of instances in one class (majority class) outnumbers the other class (minority class). The main problem related to binary imbalanced data sets is classifiers tend to ignore the minority class. Numerous resampling techniques such as undersampling, oversampling, and a combination of both techniques have been widely used. However, the undersampling and oversampling techniques suffer from elimination and addition of relevant data which may lead to poor classification results. Hence, this study aims to increase classification metrics by enhancing the undersampling technique and combining it with an existing oversampling technique. To achieve this objective, a Fuzzy Distance-based Undersampling (FDUS) is proposed. Entropy estimation is used to produce fuzzy thresholds to categorise the instances in majority and minority class into membership functions. FDUS is then combined with the Synthetic Minority Oversampling TEchnique (SMOTE) known as FDUS+SMOTE, which is executed in sequence until a balanced data set is achieved. FDUS and FDUS+SMOTE are compared with four techniques based on classification accuracy, F-measure and G-mean. From the results, FDUS achieved better classification accuracy, F-measure and G-mean, compared to the other techniques with an average of 80.57%, 0.85 and 0.78, respectively. This showed that fuzzy logic when incorporated with Distance-based Undersampling technique was able to reduce the elimination of relevant data. Further, the findings showed that FDUS+SMOTE performed better than combination of SMOTE and Tomek Links, and SMOTE and Edited Nearest Neighbour on benchmark data sets. FDUS+SMOTE has minimised the removal of relevant data from the majority class and avoid overfitting. On average, FDUS and FDUS+SMOTE were able to balance categorical, integer and real data sets and enhanced the performance of binary classification. Furthermore, the techniques performed well on small record size data sets that have of instances in the range of approximately 100 to 800.

Keywords: Imbalanced data, Resampling technique, Undersampling technique, Oversampling technique, Fuzzy logic

Acknowledgement

All praise to Allah who gave me patience and strength to complete this study.

I would like to take this opportunity to express my gratitude to my main supervisor, Miss Aniza Mohamed Din for her advice and encouragement. A special thanks to my co-supervisor Prof. Dr. Ku Ruhana Ku Mahamud for her guidance throughout the completion of my study. I would also like to thank the appointed examiners for their valuable critiques to improve my thesis.

To my parents, Zorkeflee Abu Hasan and Badariah Mohd Yusoff, thank you for your love and prayers. To my husband, Zariq Zaquan Razani, thank you for your endless support and encouragement. To my siblings, Zuhaili, Zulhilmi and Mahirah, thank you for your inspirations and words of wisdom to boost up my spirit. I really appreciate all of you.

Last but not least, I wish to thank all my friends especially my labmates, who continuously help, support and motivate me during this journey.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AUC	Area Under ROC Curve
BRACID	Bottom-up induction of Rules and Cases for Imbalanced Data
CNN	Condensed Nearest Neighbour
DID	Department of Irrigation and Drainage
DUS	Distance-based Under-Sampling
ENN	Edited Nearest Neighbor
EUS	Evolutionary Undersampling
FDUS	Fuzzy Distance-based Undersampling
FDUS+SMOTE	FDUS and SMOTE
FN	False Negative
FP	False Positive
G-mean	Geometric mean
ISMOTE	Improved SMOTE
ISMOTE+DUS	ISMOTE and DUS
k-NN	k-Nearest Neighbour
m	meter
mm	millimetre
MSMOTE	Modified SMOTE
NCL	Neighbourhood Cleaning Rule
OSS	One-Sided Selection
RNN	Reduced Nearest Neighbour
ROC	Receiver Operating Characteristics
ROS	Random Over-Sampling
RUS	Random Under-Sampling
SMOTE	Synthetic Minority Over-sampling TEchnique
SMOTE+ENN	SMOTE and ENN
SMOTE+TL	SMOTE and TL
SVM	Support Vector Machine
TL	Tomek Links
TN	True Negative
TP	True Positive

CHAPTER ONE
INTRODUCTION

Data is a set of values of qualitative and quantitative variables in order to deliver information. Often, the distribution of the data sets are imbalanced. This chapter provides some background about imbalanced data sets and the problem related to them. The research objectives, research scope and significance of study are also stated in this chapter.

1.1 Background

Imbalanced data sets occur when the number of samples in one class is low as compared to other classes (Barua, Islam, Yao, & Murase, 2014). In binary classification, the class that contain less instances is known as minority class, and the other class is known as majority class. Examples of imbalanced data sets are flood events (Wang, Chen, & Small, 2013), medical data sets (Dubey, Zhou, Wang, Thompson & Ye, 2014), intrusion detection data sets (Chairi, Alaoui, & Lyhyaoui, 2012), credit card fraud detection (Padmaja, Dhulipalla, Krishna, Bapi, & Laha, 2007), and oil spill identification (Brekke & Solberg, 2005). The issue that is commonly related to imbalanced data is poor classification performance due to the tendency of classifiers to ignore data samples that belong to the minority class (Lin & Chen, 2012; Mangai, Samanta, Das, & Chowdhury, 2010; Mi, 2013). For example, when imbalanced data is classified using Support Vector Machine (SVM), the decision boundary obtained is biased towards the minority class resulting to misclassification (Liu, Yu, Huang, & An, 2011; Bennett & Bredensteiner, 2000). This bias will reduce the performance of SVM with respect to the minority class (Batuwita & Palade, 2013).

Hence, to overcome the problem, several methods have been proposed in algorithm-based and data-based approaches (Chawla, Japkowicz, & Kotcz, 2004; Ganganwar, 2012).

In algorithm-based approach, two possible processes can take place. Either a new algorithm is created or the existing classification algorithm is improved so that it can recognise the minority class (Yang, Fong, Wong, & Sun, 2013). Adjusting the costs of classes to counter the class imbalance, modifying the probabilistic estimation at the tree leaf of decision trees, and altering the decision threshold are some solutions in algorithm-based approach as stated by Ganganwar (2012). However, algorithm-based approach has some disadvantages. For example, it depends on the classifier and is difficult to handle because of the need to correspond the classifier learning algorithm with the application domain (Fitkov-Norris & Folorunso, 2013; Sun, Wong, & Kamel, 2009). In contrast with algorithm-based approach, data-based approach is easier to handle because data sets are modified to produce balanced data sets before the classifier is trained (Chawla, 2010). In addition, the techniques in data-based approach are more versatile because of the independency towards classifiers as compared to the algorithm-based approach (Fitkov-Norris & Folorunso, 2013).

The aim of data-based approach is to modify the ratio of imbalanced data before the data is trained (Chairi et al., 2012; Diamantini & Potena, 2009). The advantage of this approach is its independence towards the classifier (Lopez, Fernandez, Garcia, Palade, & Herrera, 2013). Resampling technique is categorised as a data-based approach and it is divided into undersampling and oversampling techniques. Undersampling technique can be defined as a technique of removing samples from a majority class, while oversampling technique adds samples to the minority class (Luengo, Fernandez, Garcia, & Herrera, 2011). However, these two techniques may lead to a loss of potential data and create overfitting (Chawla, 2010).

Findings showed that undersampling technique provides better classification accuracy than oversampling technique (Bekkar & Alitouche, 2013). However, there is still a lack of approaches of making decisions to remove the instances from the majority class. The implementation of k-Nearest Neighbour (k-NN) and mean in undersampling algorithm may cause ambiguity and bias (Napierala & Stefanowski, 2012; Whitley & Ball, 2001). These factors will cause classification inaccuracy.

The issue with existing techniques in handling with imbalanced data sets is the degree of ambiguity. Fuzzy logic and rough set theory are known as approaches in handling ambiguity (Kanagavalli & Raja, 2011; Verbiest, Ramentol, Cornelis & Herrera, 2012). Fuzzy logic allows to build up membership function to conserve important data and avoid removal of data randomly, while rough set creates lower and upper approximations of a set (Hu, Lin, & Han, 2004; Li, Liu & Hu, 2010). In rough set theory, the concept of ambiguity is based on boundary (Shen & Jiang, 2010).

In order to develop a resampling technique that produce better classification performance, undersampling and oversampling technique are combined (Bekkar & Alitouche, 2013). According to several studies, the combination of the resampling techniques produce better classification accuracy result as compared to standalone techniques because undersampling and oversampling complement each own's advantage and disadvantage (Chawla, Bowyer, Hall, & Kegelmeyer, 2002; Jeatrakul, Wong, & Fung, 2010; Li, Zou, Wang, & Xia, 2013).

In conclusion, imbalanced data sets can cause misclassification accuracy. Many real data sets are presented in imbalanced data forms. Data-based and algorithm-based are two approaches that are implied in handling imbalanced data sets. Among these two approaches, data-based approach is found easier to be implemented. However, there are some lacking in existing resampling techniques that is categorised under data-based approach. Therefore, a research on improving the existing resampling technique for imbalanced data sets is needed to be conducted.

1.2 Problem Statement

In data-based approach, resampling technique is used to adjust the ratio of imbalanced data sets distribution (Naganjaneyulu & Kuppa, 2012). Among the drawbacks of resampling techniques are the undersampling technique may lead to the loss of potential data, while oversampling creates overfitting (Ganganwar, 2012; Liu, Wu, & Zhou, 2009). These problems can lead to classification inaccuracy (Lee & Lee, 2012).

In one of the existing undersampling techniques, the decision of discarding data from the majority class is based on an average (mean) of distance between samples in the minority and majority classes (Li et al. , 2013). Mean is not suitable for imbalanced data because it is sensitive towards skewed data as stated by Whitley and Ball (2001), where it has the tendency to be biased towards the majority class in imbalanced data sets. Besides that, k-NN has been used in making decision of data discarding process (Hart, 1968; Gates, 1971; Wilson, 1972; Kubat & Matwin, 1997). However, the problem in k-NN is the likeliness to have several equal distances between the data samples in minority and majority classes where they can cause ambiguity in choosing the samples that need to be removed (Napierala & Stefanowski, 2012).

Rough set theory and fuzzy logic are two approaches used in dealing with ambiguity problems. However, in rough set theory, the concept of ambiguity is based on boundary, where boundary samples are difficult to deal with classifiers and more likely to be wrongly classified (Anand, Pugalenthi, Fogel, & Suganthan, 2010; Shen & Jiang, 2010). In addition, rough set is not flexible to be used for large data sets (Hu et al., 2004; Shivalkar & Tripathy, 2015).

The combination of undersampling and oversampling techniques produces better classification accuracy than standalone techniques (Jeatrakul et al., 2010). However, the size of samples in the majority class that need to be discarded and the number of new samples that have to be created for the minority class are defined before the data are resampled (Li et al., 2013). The drawback of this approach is the possibility of wrongly choosing the suitable amount of samples that need to be removed or added (Dubey et al., 2014). It will lead to the decrement of accuracy of both majority and minority classes (Li et al., 2013).

Therefore, this study aims to enhance the undersampling technique using fuzzy logic to overcome bias and ambiguity problems. Then, the enhanced undersampling technique is combined with oversampling technique to produce better classification accuracy.

1.3 Research Objectives

The objectives of this study are as follows:

1. To enhance the undersampling technique using fuzzy logic.

2. To integrate the enhanced undersampling with oversampling technique.

3. To evaluate the enhanced resampling techniques.

1.4 Research Scope

Flood disaster causes tremendous damages that affect the society, economy and environment. Flood is caused by excessive amount of rainfall and river water for a certain period of time. However, the rainfall and river water level data are imbalanced. Therefore, this research is focusing on developing the technique to improve the accuracy of classifying a binary class of imbalanced data sets. At the end of this study, an enhanced resampling technique is produced in order to balance the data sets. Support Vector Machine (SVM) is used as a classifier. The data sets for this study are rainfall and river water levels of Perlis from the year 2005 until 2013 that are collected from the Department of Irrigation and Drainage (DID). Some data sets from UCI Machine Learning Repository are chosen as benchmark data.

1.5 Significance of Study

The significance of this study is the enhancement of resampling technique to handle imbalanced data sets. This study aims to produce an enhanced undersampling technique that is able to minimise the loss of potential data in the majority class. By minimising the removal of instances from the majority class, the classification results have become more accurate. Furthermore, the enhancement of resampling technique that involves a combination of the enhanced undersampling and oversampling techniques could improve the classification performance for imbalanced data sets. Moreover, this study is related to the classification of flood data sets. Hence, it is beneficial to the flood disaster management and the local community in order to predict the occurrence of floods.

1.6 Summary

Imbalanced data sets may affect the performance of classifiers because they tend to misclassify the samples in the minority class. Several techniques have been proposed to solve the problem of classifying the imbalanced data. However, there are some issues from the existing techniques that need to be improved. Therefore, this study is focusing on the enhancement of resampling technique to handle imbalanced data sets.

CHAPTER TWO
LITERATURE REVIEW

This chapter provides reviews about imbalanced data sets, techniques to balance the data sets and performance metrics to evaluate the techniques. Data-based and algorithm-based are the approaches that are being used to solve the problems of imbalanced data sets classification. Resampling techniques which consist of undersampling, oversampling and a combination of undersampling and oversampling lie under the data-based approach.

2.1 Imbalanced Data Sets

Imbalanced data sets can be defined as a number of instances in a class (majority class) which outnumbers the other class (minority class), where it can be presented in the ratio of 100 to 1, 1000 to 1 or more (Chawla et al., 2004; Sun, Kamel, Wong & Wang, 2007; Yang & Gao, 2013; Zhang & Wang, 2013). Ding (2011) stated that for binary classification, if the ratio of two classes is not less than 19:1, the data set is defined as imbalanced. Imbalanced data sets exist when rare cases happen and the ignorance towards these cases can affect the society, economy and environment (Sang, Gao & Liu, 2013). Examples for such cases are stroke diagnosis (Ou-Yang, Rieza, Wang, Juan & Huang, 2013), flood prediction (Segretier, Clergue, Collard & Izquierdo, 2012), credit card fraud detection (Padmaja et al., 2007) and oil spill identification (Brekke & Solberg, 2005). Hence, it is important not to overlook the infrequently occurred cases.

Imbalanced data sets can reduce the performance of classifiers because they tend to ignore the instances in minority class and this problem will lead to inaccurate classification accuracy (Chairi et al., 2012; Del Gaudio, Batista & Branco, 2014; Phung, Bouzerdoum & Nguyen, 2009). For example, in classification, flood occurrence is represented as positive instances in minority class and the non-flood occurrence is represented as negative instances in majority class. Therefore, if flood cases are misclassified, the impact is higher than the misclassification of non-flood cases. It can be concluded that more attention should be paid to minority class than majority class.

To overcome imbalanced data sets problem in classification, the proposed solutions can be divided into data-based approach and algorithm-based approach (Sun, Robinson, Adams, Boekhorst, Rust & Davey, 2006; Wang & Yao, 2013).

2.2 Data-based Approach

Data-based approach aims to balance the distribution of instances in both minority and majority classes before a classifier is trained (Jeatrakul & Wong, 2012). The advantage of data-based approach is the independency towards classifiers, hence it is easy to be modified (Fernandez, Lopez, Galar, Del Jesus, & Herrera, 2013; Folorunso & Adeyemo, 2012). The commonly used data-based approach is resampling technique, such as undersampling, oversampling and a combination of both techniques (Li et al., 2013; Sun et al., 2009).

2.2.1 Undersampling Technique

Undersampling technique can be defined when a few samples of the majority class are removed (Mirza, Lin, & Toh, 2013). Undersampling is an efficient technique in balancing the data because it decreases the time of training process due to the removal of instances from the majority class (Liu et al., 2009; Ganganwar, 2012). However, it may reduce the information accuracy since the potential information may be discarded (Gu, Cai, & Zhu, 2009; He, Han, & Wang, 2005). Random Under-Sampling (RUS) is one of the undersampling techniques. RUS removes random instances from the majority class in order to balance the data sets. Although the technique is simple, the data removal may cause a loss of potential data (Chairi et al., 2012).

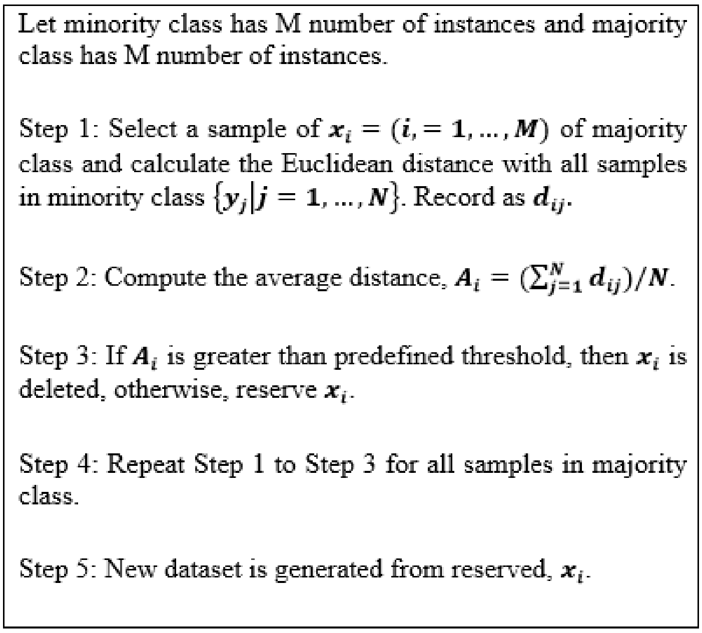
Condensed Nearest Neighbour (CNN) follows the basic approach of nearest neighbour rule to identify the borderline instances (Hart, 1968). The learning process time is less, but it includes a big portion of noisy instances (Fitkov-Norris & Folorunso, 2013). Tomek Links (TL) is an improvement of CNN (Tomek, 1976). Instead of removing the samples from the majority class randomly, TL only chooses samples that are closer to the boundary points. One-Sided Selection (OSS) (Kubat & Matwin, 1997) is a combination of two undersampling techniques; CNN and TL. Removal of examples from the majority class using CNN and TL is implied to remove the noise in order to create a new training set. The drawback of OSS is it requires high learning time (Bekkar & Alitouche, 2013; Jo & Japkowicz, 2004).

Reduced Nearest Neighbour (RNN) (Gates, 1971) removes noisy instances while keeping the instances at the border points. The drawback of RNN is it requires higher learning time to compute the learning set. Wilson’s Edited Nearest Neighbor Rule

(ENN) (Wilson, 1972) is an edited k-NN to improve one nearest neighbour. Wilson classified the samples using three nearest neighbour rules and formed a reference set. Then, the misclassified samples are removed. Neighbourhood Cleaning Rule (NCL) (Laurikkala, 2001) uses ENN rule to identify and remove instances in the majority class. First, three nearest neighbours for each instances in the training set is identified. If the instance belongs to the majority class and it is misclassified by its three nearest neighbour, then the instances are removed. If the instance belongs to the minority class and it is misclassified by three nearest neighbour of the majority class, then it is also removed.

Distance-based Under-Sampling (DUS) (Li et al., 2013) uses Euclidean distance to find the distance between samples before making decisions of discarding the instances in the majority class. Unlike other undersampling techniques, DUS does not consider the boundary samples because classifiers have difficulty in dealing with them (Anand et al., 2010). Hence, DUS is easier to be used for imbalanced data classification. Figure 2.1 shows the algorithm of DUS.

DUS uses mean and includes all samples in order to identify and ignore the sample in the majority class. Since mean is very sensitive to skewed data sets (Mann, 2012; Whitley & Ball, 2001), it is not suitable for imbalanced data sets because the result will be biased towards the majority class instances.



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

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

Step 2: Compute the average distance, $\boldsymbol{A_i} = (\sum_{j=1}^{N} \boldsymbol{d_{ij}})/\boldsymbol{N}$.

Step 3: If $\boldsymbol{A_i}$ is greater than predefined threshold, then $\boldsymbol{x_i}$ is deleted, otherwise, reserve $\boldsymbol{x_i}$.

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

Step 5: New dataset is generated from reserved, $\boldsymbol{x_i}$.

Figure 2.1. Algorithm of Distance-based Under-Sampling (Li et al., 2013)

Overall, the undersampling techniques mainly utilises k-NN to identify the removability of samples in the majority class (Zhang, Liu, Gong, & Jin, 2011). The advantage of k-NN is that k-NN reduces the bias towards the domination of majority class because the instances in majority class are discarded based on the farthest distance to the nearest neighbour instances in minority class (Zhang & Mani, 2003; Garcia, Mollineda, & Sanchez, 2008). However, according to Napierala and Stefanowski (2012), there are cases when the k-NN has equal distance from the classified instances that may cause ambiguity.

According to Zadeh (1980), in ambiguity cases, fuzzy logic is suitable to be used. This statement is aligned with other researchers that claimed fuzzy logic has an advantage in solving ambiguity problems (Ganesh, 2006; Jiang, Deng, Chen, Wu & Li, 2009; Mahdizadeh & Eftekhari, 2013; Sivanandam, Sumathi & Deepa, 2007; Wang, Zhao

& Hao, 2011; Verbiest et al., 2012). Besides fuzzy logic, rough set theory is another approach in handling ambiguity (Kanagavalli & Raja, 2011). In rough set theory, the concept of ambiguity is based on boundary (Shen & Jiang, 2010). The disadvantage is boundary samples is difficult to deal with any classifiers (Anand et al., 2010; Shen & Jiang, 2010). In addition, the drawback of rough set is it is not suitable for large data sets (Hu et al., 2004; Shivalkar & Tripathy, 2015).

Fuzzy set approach has been implemented to solve learning problems for imbalanced data. Fuzzy logic is derived from fuzzy set theory that allows the development of membership function. The aim of fuzzy set is to reduce complexity without simplifying the information excessively (Singpurwalla & Booker, 2004). Membership function is presented in triangular, trapezoidal and Gaussian (Sivanandam et al., 2007), where the choice of optimal membership functions needs to be considered (Aziz, 2009). Membership functions are derived to classify the contribution of instances in both minority and majority classes that is correctly reflected by the prediction error (Visa & Ralescu, 2003).

The derivation of membership functions can be divided into intuition, inference, rank ordering, neural networks, genetic algorithms and inductive seasoning (Ross, 2010). Intuition is based on human's intelligence and one need to be expert in the field of the problem. Inference has similarity with intuition where membership function is formed from facts known. The difference is, this method involves knowledge to perform deductive reasoning. Rank ordering uses polling concept to assign membership values where the preferences are determined by pairwise comparisons. Fuzzy membership functions also can be created by training input data set using neural network. The output of the data points in the trained data sets are membership functions. In genetic algorithm, the membership functions are coded as bit strings. Both neural networks and genetic algorithms are computationally very expensive (Ross, 2010). In developing membership function where the data set has input-output relationships, inductive reasoning is suitable to be used (Sivanandam et al, 2007). Inductive reasoning is performed by entropy minimisation where fuzzy threshold is established.

The concept of fuzzy logic has been introduced by Li et al. (2010) in order to estimate class distribution between samples in minority and majority classes. This technique uses Gaussian function as a majority class membership function and *α-cut* to remove the instances. In Wong, Leung and Ling (2014), undersampling based on fuzzy logic is used for large data sets. The fuzzy logic is applied to cluster the samples in the majority class to make a selection of which samples are important. However, the settings of the membership functions are based from the calculation of mean value where mean is very sensitive to skewed data sets (Mann, 2012; Whitley & Ball, 2001).

2.2.2 Oversampling Technique

Oversampling techniques can be defined as an addition of artificial minority class samples, which is done to balance the size of two classes (Chawla et al., 2002). In contrast to undersampling technique, oversampling technique increases the samples in minority class that lead to increment in training time (Hu, Liang, Ma, & He, 2009). Furthermore, duplicating the samples will cause overfitting that can worsen the prediction (Kim, Baek, & Kim, 2013; Sang et al., 2013). Random Over-Sampling (ROS) is one of the oversampling techniques. ROS randomly duplicates the instances from the minority class to balance the data sets. However, the data replication leads to overfitting (Chairi et al., 2012).

To overcome the overfitting problem, Synthetic Minority Over-sampling TEchnique (SMOTE) (Chawla et al., 2002) has been introduced. SMOTE creates new synthetic instances from the minority class along the line segments that join any or all minority class nearest neighbours. SMOTE has shown as one of the most effective oversampling techniques when it is applied to several applications (Ding, 2011). Besides that, Modified SMOTE (MSMOTE) (Hu et al., 2009) uses a different way from SMOTE in choosing near neighbours. MSMOTE calculates the distance of instances in the minority class and all instances of training data. From the calculation, the decision of creating new instances is based on noise identification. However, MSMOTE does not consider the important feature. Other than MSMOTE, Improved SMOTE (ISMOTE) (Li et al., 2013) is introduced to assign weight vector to the instances in the minority class. A higher weight is assigned to create neighbour instances of the minority class, and a lower weight for neighbour instances of the majority class.

Results of several experiments showed that undersampling techniques produced better classification accuracy as compared to oversampling techniques (Bekkar & Alitouche, 2013; Ganganwar, 2012; He & Garcia, 2009; Liu et al., 2009). However, the combination of oversampling the minority class and undersampling the majority class gave better classification accuracy than standalone techniques as claimed by Chawla et al. (2002) and Li et al. (2013).

2.2.3 Combination of Undersampling and Oversampling Techniques

The combination of undersampling technique and oversampling technique is introduced to overcome the drawback of each technique. The advantage of this approach is it can increase the performance of classifiers (Batista, Bazzan & Monard, 2003; Chawla et al., 2002; Li et al., 2013). There are a few existing works on the combination of undersampling and oversampling techniques such as SMOTE and TL (SMOTE+TL) (Batista et al., 2003). The combination is proposed to oversample the minority class using SMOTE. Then, TL are used to the oversampled class to create better-defined class clusters and to remove borderline and noise instances in the majority class. In this case, instances from both classes are removed to produce a balanced data set with well-defined class.

The combination of SMOTE and ENN (SMOTE+ENN) is quite similar with SMOTE+TL (Batista, Prati, & Monard, 2004). But, ENN removes more instances as compared to TL. So, the data cleaning process is more in-depth. SMOTE+TL and SMOTE+ENN are useful when the data sets have very few instances in the minority class (Batista et al., 2004; Chawla et al., 2004). In Chawla et al. (2002), SMOTE is combined with undersampling technique. The samples in the majority class are randomly removed until the minority class becomes some specified percentage of the majority class. Then, synthetic minority samples are created. The results showed that the combination of SMOTE and undersampling techniques performed better than undersampling alone.

On the other hand, ISMOTE and DUS (ISMOTE+DUS) (Li et al., 2013) work simultaneously to balance the data sets. During the data resampling, the number of instances in the minority class created is the same as the number of instances in the - majority class deleted. ISMOTE+DUS may decrease the classification accuracy because the decision of discarding and adding the samples might be wrong (Dubey et al., 2014). However, the classification accuracy of the combination between ISMOTE and DUS is better than standalone undersampling and oversampling techniques (Li et al., 2013).

Fuzzy set theory has been introduced to both undersampling and oversampling techniques in order to reduce the data in the majority class and generate virtual samples in the minority class (Li et al., 2010). The aim of this technique is to estimate the class distribution to generate balanced data sets. Then, to enhance the classification ability of classifiers, they extend the data attributes by adding corresponding fuzzy class possible values. Figure 2.2 illustrates the flowchart of the combination between undersampling and oversampling techniques where imbalanced data set is balanced by both techniques simultaneously.

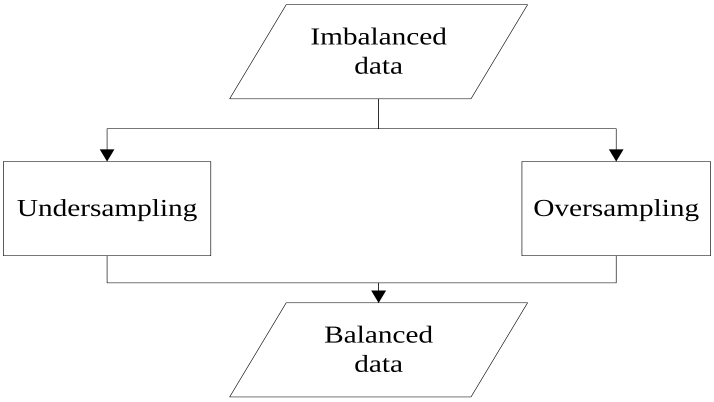


Figure 2.2. Flowchart of Combination of Undersampling and Oversampling Technique (Batista et al., 2003, Batista et al., 2004, Chawla et al., 2002; Li et al., 2010; Li et al., 2013).

2.3 Algorithm-based Approach

Algorithm-based approach aims to create a new algorithm or to improve the existing classification algorithms so it can recognize the positive class (Mahdizadeh & Eftekhari, 2013). The idea of this approach is to discriminate bias in imbalanced class cases, and the advantage of this approach is no modification towards the data sets is done (Garcia, Sanchez, Mollineda, Alejo & Sotoca, 2007; Fernandez et al., 2013). Single classifiers and ensemble of classifiers belong to algorithm-based approach.

2.3.1 Single Classifier

Classification is a task that estimate the correct classes of instances and classifier is an instance to construct algorithm for a specific training set (Rokach, 2009). Support Vector Machine (SVM), Decision Tree and Artificial Neural Network are among the most used classifiers (Ding, 2011). However, due to imbalanced data sets, these classifiers have difficulty in classifying the instances. Classifiers tend to be biased towards the majority class instances.

Decision tree is build when the class label is associated with a leaf which is found by examining the training sets covered by the leaf. Then, the most frequent class label is chosen. In class imbalance problem, the parameters of pruning factor were set to obtain a balance classification (Lee, Yang, Chang & Lee, 2010). The pruning is based on the predicting error. Due to that, there is high probability that new leaf node is labelled as dominant class and some branches predict small (Sun et al., 2007). Furthermore, to avoid overfitting, the decision trees use pruning. However, the pruning technique did not perform well on imbalanced data sets (Liu, Chawla, Cieslak & Chawla, 2010).

Artificial Neural Network increased the weight of the minority class and cost function was introduced while handling with the imbalanced class in the training process (Fu, Wang, Chua & Chu, 2002; Alejo, Garcia, Sotoca, Mollineda & Sanchez, 2007). Artificial Neural Network performed ineffectively because the minority class is not weighted in the networks (Chawla et al., 2002; Carvajal, Chacon, Mery & Acuna, 2004).

SVM is binary classifier that classify based on data points at the minimal distances from the separating hyperplane to the closest points. SVM has successfully applied to classification problems in different domains (Batuwita & Palade, 2013). In imbalanced data sets case, the hyperplane tends to be pushed closer to the positive class. So, SVM creates boundaries to distance the hyperplane from the positive class (Tang, Zhang, Chawla & Krasser, 2009). However, SVM is found to be more effective classifier in dealing with imbalanced data sets than other classifiers (Sun et al., 2007; Tang et al., 2009).

Fuzzy system has been extracted to classify imbalanced data sets where membership functions are created to differentiate the class of minority or majority class (Soler & Prim, 2009). However, Soler and Prim (2009) found that the drawback of this technique is that it is time consuming as it requires many rules because the ideal number of rules needed for the data sets is not defined before the classification.

Fuzzy rule-based classification has been used to distinguish the areas between minority and majority classes (Fernandez, Del Jesus & Herrera, 2009). Nevertheless, fuzzy classifier achieves its optimum classification performance when the training data sets have been balanced at the pre-processing level because it is less sensitive to the learning of imbalanced class distribution (Fernandez et al., 2009).

The finding is similar with Visa and Ralescu (2005) where fuzzy classifier is less sensitive towards imbalanced data. As stated in Visa and Ralescu (2003) and Li et al. (2010), fuzzy logic is suitable to be used in dealing with imbalanced cases. However, problems occur when fuzzy logic is applied in algorithm-based approach because the modification is dependent to the classification algorithm (Dubey et al., 2014).

2.3.2 Ensemble of Classifiers

Ensemble of classifiers is a set of individual trained classifiers and the decisions of classifications are combined to choose the best classifier (Tan, Steinbach & Kumar, 2006). The idea of ensemble-based classifiers is to combine the votes of several classifiers to produce an accurate prediction. Bagging and boosting are the methods of ensemble classifiers. To deal with imbalanced data sets, ensemble classifiers are combined with data-based approach to process the data before learning each classifier (Galar, Fernandez, Barrenechea, & Herrera, 2013).

SMOTEBagging (Wang & Yao, 2009) is an integration of SMOTE and bagging. Synthetic samples in the minority class are created before bagging. Then, bagging is applied to the samples in the majority class. However, the drawbacks of bagging includes the tendency to bootstrap sample for both majority and minority classes, thus, the imbalanced distribution will occur at every iteration (Bekkar & Alitouche, 2013). SMOTEBoost (Chawla, Lazarevic, Hall & Bowyer, 2003) is an integration of SMOTE with boosting. This method creates synthetic instances and indirectly enables weight updating for each iteration. However, SMOTEBoost will worsen the overfitting problem because it will increase the inductive bias due to the increased similarity among the groups of data samples (Kim, 2013).

MSMOTEBoost (Hu et al., 2009) is a modified SMOTE algorithm that integrates with boosting. This method considers the imbalanced data and noise samples but does not consider the differences of important features (Hu et al., 2009). RUSBoost (Seiffert, Khoshgoftaar, Van Hulse, & Napolitano, 2010) is a result of integration between Random Under-Sampling (RUS) with boosting. This method requires less training time as compared to SMOTEBoost, yet sampling with majority class is still time consuming (Ou-Yang et al. 2013). EUSBoost (Galar et al., 2013) is an integration of Evolutionary Undersampling (EUS) with boosting. This method resamples the imbalanced data sets in a supervised manner. However, boosting algorithms may increase generalisation error because new validated data sets might not be trained by the classifiers (Kim, 2013).

To conclude, from the reviewed methods, the disadvantages of algorithm-based approach are its dependency to the classifier, i.e. the modification of algorithms is fully depends on the classifier, not the data set. So, it is difficult to handle because the algorithm might be only suitable for certain domain. For these reasons, data-based approach is more preferable (Fernandez et al., 2013; Bekkar & Alitouche, 2013; Zhong, Raahemi & Liu, 2009).

2.4 Performance Evaluation

In imbalanced data sets cases, only a few performance metrics are suitable to evaluate classification performance. Traditionally, confusion matrix as described in Table 2.1, is used to evaluate classification performance (Chawla et al., 2002). Minority class is represented as positive class and majority class is represented as negative class.

Table 2.1
Confusion Matrix

	Positive prediction	Negative prediction
Positive class	True positive (TP)	False Negative (FN)
Negative class	False Positive (FP)	True Negative (TN)

From Table 2.1, TP represents the number of positive instances that are correctly classified, TN corresponds to the number of negative instances that are correctly classified, FP denotes the number of positive instances that are misclassified and FN refers to the number of negative instances that are misclassified. However, in imbalanced data sets cases, predictive accuracy is not suitable to be used because of its tendency to be biased towards the majority class (He & Garcia, 2009; Goel, Maguire, Li, & McLoone, 2013). There are other alternative performance metrics that are suitable to evaluate classification performance for imbalanced data sets such as Receiver Operating Characteristics (ROC), Area Under ROC Curve (AUC), F-measure and G-mean.

ROC curve is a graph of TP on *x-axis* versus FP on *y-axis* that evaluates the classifier performance (Chawla, 2010). Perfect classification is represented as point (0, 1) and the line *y*=*x* defines the strategy of randomly guessing classes (Garcia et al., 2007). However, it is difficult to describe the result because it is not in a numerical metric (Seiffert et al., 2010). From the ROC curve, the overall performance of a classifier can be measured numerically by computing AUC (Galar et al., 2013). The comparison of which model is better can be described when it has a larger AUC (Tan et al., 2006).

F-measure shows the effectiveness of a classifier (He & Garcia, 2009). It is a combination of recall and precision. Recall measures the instances of positive class that are labelled correctly, while precision describes the number of positive instances that are actually correctly labelled (Goel et al., 2013). Therefore, it is suitable to measure the performance of imbalanced data sets classification.

Geometric mean (G-mean) considers the accuracy of both positive and negative instances (Nguyen, Bouzerdoum & Phung, 2009). Hence, it is suitable for evaluating imbalanced data set classification performance because G-mean maximises and balances the classifier performance (Li et al., 2013). Furthermore, it is independent towards imbalanced distribution (Jeatrakul et al., 2010).

2.5 Summary

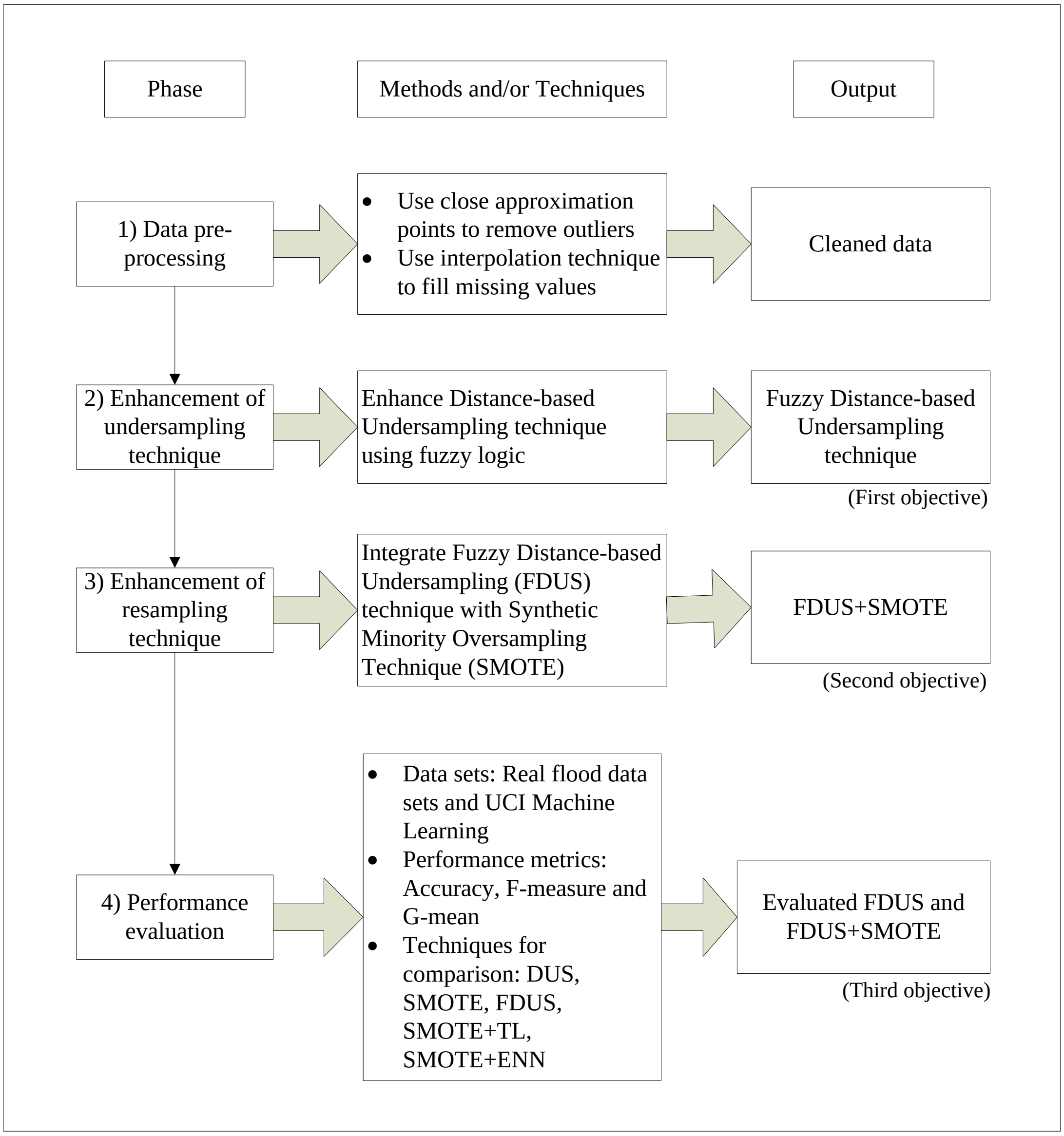
Several techniques have been proposed at data-based and algorithm-based approaches to overcome the matter regarding the classification of imbalanced data sets. Data-based approach is easier to be modified as compared to algorithm-based approach because of its independency towards classifiers. Therefore, this study focuses on the resampling techniques that are categorised under data-based approach. Improvement of the resampling techniques need to be done to increase the classification performance.

CHAPTER THREE
RESEARCH FRAMEWORK AND METHODOLOGY

This chapter details out the research framework and methodological approach for this study. The first section in this chapter focuses on the research framework. The second section explains on the methods to run the experiment for each phase. The content of this chapter is summarised in the final section.

3.1 Research Framework

In this section, the phases that need to be fulfilled to achieve the objectives for this study are described. Figure 3.1 summarises the framework for this research that consists of four phases. The phases are divided into data pre-processing, enhancement of undersampling technique, enhancement of resampling technique and performance evaluation. The output of the second, third and fourth phases will accomplish the first, second and third research objectives, respectively. The first objective is to enhance undersampling technique using fuzzy logic, and the second objective is to enhance resampling technique by combining undersampling and oversampling techniques. The final objective is to evaluate the proposed enhanced resampling techniques.



Phase
Methods and/or Techniques
Output
1) Data pre-processing
Use close approximation points to remove outliers
Use interpolation technique to fill missing values
Cleaned data
2) Enhancement of undersampling technique
Enhance Distance-based Undersampling technique using fuzzy logic
Fuzzy Distance-based Undersampling technique
(First objective)
3) Enhancement of resampling technique
Integrate Fuzzy Distance-based Undersampling (FDUS) technique with Synthetic Minority Oversampling Technique (SMOTE)
FDUS+SMOTE
(Second objective)
4) Performance evaluation
Data sets: Real flood data sets and UCI Machine Learning
Performance metrics: Accuracy, F-measure and G-mean
Techniques for comparison: DUS, SMOTE, FDUS, SMOTE+TL, SMOTE+ENN
Evaluated FDUS and FDUS+SMOTE
(Third objective)

Figure 3.1. Research Framework

3.2 Research Methodology

The details of the four phases that are required to achieve the research objectives are described in this section. At the end of the first phase, the imbalanced data sets are cleaned from any outliers and missing values. Then, in the second phase, an enhanced Distance-based Undersampling (DUS) technique is produced, named Fuzzy Distance-based Undersampling (FDUS) technique. FDUS is combined with oversampling technique in the third phase. The output for the third phase is an enhanced resampling technique.

3.2.1 Data Pre-processing for Flood Data Sets

Rainfall and water level are two of the factors that contribute to flood events. For the purpose of this study, rainfall and river water level data of Perlis are collected from year 2005 until 2013 from the DID. Tables 3.1 and 3.2 are the samples 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 3.1
Hourly Rainfall Data (mm) for Sungai Pelarit

Date	**0100**	**0200**	**0300**	**0400**	**0500**	**0600**	**0700**	**0800**	**…**	**2400**
1/12/13	0	0	0	0	0	0	0	0	…	0
2/12/13	0	0	0	0	0	2	1	0	…	0
3/12/13	0	0	0	0	0	0	0	0	…	0

Table 3.2
Hourly Water Level Data (m) for Wang Kelian

Date	**0100**	**0200**	**0300**	**0400**	**0500**	**0600**	**0700**	**0800**	**…**	**2400**
1/12/13	10.21	10.24	10.24	10.25	10.25	10.27	10.25	10.25	…	10.27
2/12/13	10.25	10.25	10.24	10.25	10.25	10.25	10.25	10.27	…	10.30
3/12/13	10.24	10.25	10.25	10.34	10.36	10.36	10.32	10.29	…	10.29

After all the data sets are collected, the rainfall and water level data sets are cleaned up from outliers. Outliers may exist due to the interruption of data transmission. For this case, any point that is separated far from other points are 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 3.1. Table 3.3 shows the sample of data sets before and after removing the outliers and filling the missing values.

$$f(x) = f(x_0) + (x - x_0)\frac{f(b) - f(a)}{b - a} \tag{3.1}$$

where *f(x) = missing value, f(x*$_0$*) = 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, b = constant point after missing value*

Table 3.3
Sample of Rainfall Data (mm) for Genting Kabu

Number of point	**Date**	**Time (hour)**	**Rainfall (mm) before data cleaning**	**Rainfall (mm) after data cleaning**
1	30/12/2012	0100	0	0
2		0200	0	0
3		0300	8	8
4		0400	14	14
5		0500		19
6		0600	-4	4
7		0700	0	0
8		0800	0	0
9		0900		0.5
10		1000	-1	1

After the data cleaning, rainfall and water level data sets are combined. These two attributes will determine the flood occurrence for each catchment area. Tables 3.4 and 3.5 show the rainfall intensity and water level stages, respectively. Note that each water level station has different water level stages. Table 3.6 shows the relations of rainfall and water level stage that cause floods (Bedient, Huber & Vieux, 2008). Continuous heavy rainfall in two to four hours can cause flash floods. In addition, during the monsoon period, the amount of rain can exceed to hundreds per day. This information is provided by DID.

Table 3.4
Rainfall Intensity

Rainfall (mm)	**Category of storm**
1-10	Light
11-30	Moderate
30-60	Heavy
More than 60	Very heavy

Table 3.5
Water Level Stages of Ulu Pauh

Water level (m)	**Stages**	**Explanation**
26.50	Normal	River level is at normal level.
28.20	Alert	River level is significantly above normal level. DID Flood Operation Room is activated
28.60	Warning	River level is almost to flood level. DID Flood Operation Room is activated.
29.00	Danger	River level can cause flood. Evacuation may be initiated

Table 3.6

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

In this study, data sets from Kaki Bukit, Lubok Sireh, Wang Kelian, Ladang Perlis Selatan and Ulu Pauh from year 2005 until 2013 are chosen based on the catchment areas in Perlis. Table 3.7 presents the sample of flood data set after rainfall and water level data are combined. The division of no flood and flood classes are done based on Table 3.6.

Table 3.7

Sample of Ulu Pauh Data Set

Date	**Time**	**Rainfall (mm)**	**Water level (m)**	**Class**
29/3/2009	12.00pm	0	25.67	No flood
29/3/2009	1.00pm	0	25.67	No flood
29/3/2009	2.00pm	0	25.67	No flood
29/3/2009	3.00pm	0	25.67	No flood
29/3/2009	4.00pm	67.30	25.72	Flood
29/3/2009	5.00pm	51.10	27.95	Flood
29/3/2009	6.00pm	0.10	28.12	No flood
29/3/2009	7.00pm	0	28.19	No flood
29/3/2009	8.00pm	0	28.15	No flood

Overall, Table 3.8 provides the 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 (Mahdizadeh & Eftekhari, 2013). Minority and majority classes represent flood and no flood occurrence, respectively.

Table 3.8
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

3.2.2 Data Pre-processing for Benchmark Data Sets

Besides rainfall and river water level data sets, five imbalanced data sets which are adult, haberman, breast cancer, pima, and bupa are selected as benchmark data from UCI Machine Learning Repository (Bache & Lichman, 2013). Each of the data sets has different characteristics as described in Table 3.9.

The characteristics of the data sets are type of attribute, size of the data sets, number of instances in minority class (#Minority), number of instances in majority class (#Majority), and ratio of majority to minority class. Table 3.9 is ordered based on the descending order of the ratio. A larger ratio means the difference between the number of instances in minority class and the number of instances in majority class is big.

Table 3.9
Characteristics of Benchmark Data Sets

Data sets	**Attribute Type**	**Record Size**	**#Minority**	**#Majority**	**Ratio (maj:min)**
Adult	Categorical, Integer	152	37	115	3.11:1
Haberman	Integer	306	81	225	2.78:1
Breast cancer	Integer	700	242	458	1.89:1
Pima	Integer, Real	768	268	500	1.87:1
Bupa	Categorical, Integer, Real	345	145	200	1.38:1

A significant portion of the processed flood and benchmark data sets is presented in Appendix.

3.2.3 Enhancement of Distance-based Undersampling Technique

The Distance-based Undersampling (DUS) technique has been used in this phase for undersampling technique enhancement. Figure 3.2 demonstrates the flowchart of DUS technique.

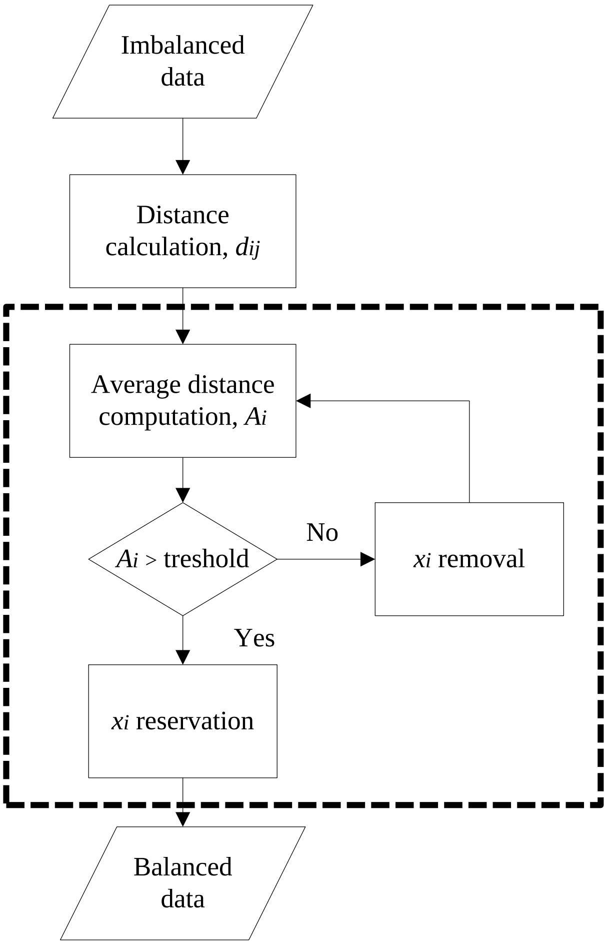


Figure 3.2. Flowchart of Distance-based Undersampling (Li et al, 2013)

This technique starts by taking the imbalanced data sets that are divided into two classes, and they are denoted as $x_i = \{x_1, x_2, \dots, x_n\}$ for samples in majority class and $y_j = \{y_1, y_2, \dots, y_n\}$ as samples in minority class. Then, the distance, d_{ij} between samples in majority class and minority class are calculated using Euclidean distance. The flow continues by computing the mean for the distance and denoted as A_i. Samples that need to be removed are based on predefined threshold. The process is repeated for all samples and balanced data sets are produced. From Figure 3.2, the enhancement of this technique is done at the dotted bordered box. Instead of calculating the mean distance, fuzzy logic is introduced. Hence, Fuzzy Distance-based Undersampling technique is produced.

3.2.4 Enhancement of Resampling Technique

In this phase, resampling technique is enhanced by combining the Fuzzy Distance-based Undersampling (FDUS) technique with Synthetic Minority Oversampling TEchnique (SMOTE). The idea of this combination is adapted from several works as discussed in Chapter 2. Imbalanced data set is balanced by both techniques simultaneously. Undersampling technique removes samples from the majority class, while oversampling technique creates new samples in the minority class. Finally, a balanced data set is produced.

However, in this study, the proposed FDUS and SMOTE are performed in sequence with certain conditions that need to be fulfilled. Firstly, imbalanced data set is divided into majority and minority classes. Then, the ratio between the two classes is adjusted with FDUS. However, if the amount of instances in the majority class has become lesser than the minority class, then SMOTE is used. The resampling process works repetitively until the data set is balanced.

3.2.5 Performance Evaluation

The balanced data sets that are produced after applying the proposed techniques to the imbalanced data sets are classified using Support Vector Machine (SVM). Each imbalanced data set is divided into five partitions using 5-fold cross validation to avoid bias. Testing was performed on FDUS and the enhanced resampling technique. Those techniques will produce balanced data sets. In order to evaluate the performance, SVM has been applied to classify the balanced data sets, and the accuracy obtained from SVM for each test is presented in percentage. A high percentage accuracy reflects the good technique.

Besides classification accuracy, the performance of the proposed techniques are evaluated using F-measure and G-mean as described by Equation 3.2 and Equation 3.3 respectively.

$$F - measure = \frac{2TP}{2TP + FN + FP} \tag{3.2}$$

$$G - mean = \sqrt{\frac{TP}{TP + FN} \times \frac{TN}{TN + FP}} \tag{3.3}$$

Where *TP*=True Positive, *TN*=True Negative, *FP*=False Positive and *FN*=False Negative

The classification accuracy, F-measure and G-mean of the proposed techniques are compared with other techniques. FDUS is compared with DUS and SMOTE. Then, the enhanced resampling technique is compared with other combination techniques (SMOTE+TL and SMOTE+ENN) and standalone techniques (FDUS, DUS and SMOTE).

3.3 Summary

The proposed enhancement of Distance-based Undersampling (DUS) and resampling techniques are presented in this chapter. The enhanced DUS is named as Fuzzy Distance-based Undersampling (FDUS) technique. FDUS technique uses the advantage of fuzzy logic which can reduce bias problems. As a result, FDUS technique can minimise the removal of useful data from the majority class. The main contribution for this study is the enhancement of resampling technique where FDUS technique integrated with SMOTE (FDUS+SMOTE). The proposed resampling technique will be performed in sequence with certain conditions that need to be accomplished.

CHAPTER FOUR

FUZZY DISTANCE-BASED UNDERSAMPLING TECHNIQUE

Chapter Four is divided into three sections. Section 4.1 discusses on the steps to enhance Distance-based Undersampling technique, Section 4.2 describes the experiments and the results of the evaluated enhanced undersampling technique. At the end of this chapter, a summary is provided in Section 4.3.

4.1 Proposed Enhancement of Distance-based Undersampling Technique

Distance-based Undersampling (DUS) technique is enhanced by implementing fuzzy logic to the algorithm. The enhanced technique is named as Fuzzy Distance-based Undersampling (FDUS) technique. Figure 4.1 illustrates the flowchart of the proposed FDUS algorithm.

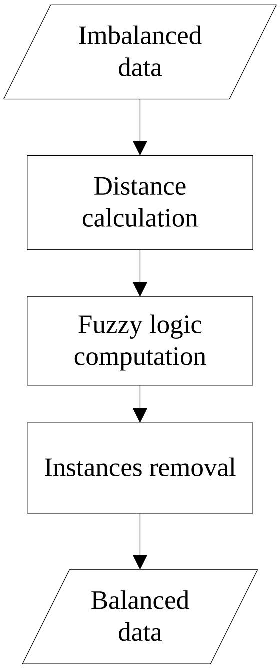


Figure 4.1. Flowchart of Fuzzy Distance-based Undersampling

From Figure 4.1, the flow starts with the division of imbalanced data set into majority and minority classes. Distance between all instances in the majority class, *m* and instances in the minority class, *n* is calculated using Euclidean distance, d_{mn} as shown in Equation 4.1.

$$d_{mn} = \sqrt{(x_m - x_n)^2 + (y_m - y_n)^2} \tag{4.1}$$

where (x_m, y_m) is point in majority class and (x_n, y_n) is point in minority class. Then, fuzzy logic is computed based on the distance to categorise the instances into several sets. In order to categorise the instances into triangular and trapezoidal membership functions, fuzzy thresholds are produced using entropy estimation.

Equations 4.2, 4.3, 4.4 and 4.5 describe the entropy estimation of $p_k(x)$, $q_k(x)$, $p(x)$ and $q(x)$ (Chiang, Shih, Lin & Shih, 2014; Christensen, 1980). Assume thresholds are in the range between x_1 and x_2.

$$p_k(x) = \frac{n_k(x) + 1}{n(x) + 1} \tag{4.2}$$

$$q_k(x) = \frac{N_k(x) + 1}{N(x) + 1} \tag{4.3}$$

$$p(x) = \frac{n(x)}{n} \tag{4.4}$$

$$q(x) = 1 - p(x) \tag{4.5}$$

where

$p_k(x)$ = conditional probabilities that class k sample is in the region $[x_1, x_1+x]$,

$q_k(x)$ = conditional probabilities that class k sample is in the region $[x_1+x, x_2]$,

$p(x)$ = probabilities that all samples are in the region $[x_1, x_1+x]$,

$q(x)$ = probabilities that all samples are in the region $[x_1+x, x_2]$,

$n_k(x)$ = number of class k samples located in $[x_1, x_1+x]$,

$n(x)$ = total number of samples located in $[x_1, x_1+x]$,

$N_k(x)$ = number of class k samples located in $[x_1+x, x_2]$,

$N(x)$ = total number of samples located in $[x_1+x, x_2]$,

n = total number of samples in $[x_1, x_2]$.

From Equations 4.2, 4.3, 4.4 and 4.5, the estimation of entropy is found. Equation 4.6 shows the equation to find minimum entropy.

$$S(x) = p(x)S_p(x) + q(x)S_q(x) \tag{4.6}$$

where

$$S_p(x) = -[p_1(x)lnp_1(x) + p_2(x)lnp_2(x)], \tag{4.7}$$

$$S_q(x) = -[q_1(x)lnq_1(x) + q_2(x)l\square q_2(x)] \tag{4.8}$$

A value of x that gives minimum entropy is the optimum threshold value. Table 4.1 shows the sample of minimum entropy calculation where x is the value of the calculated distance.

Table 4.1
Sample of Minimum Entropy Calculations

x	3.5	**7**	10.2	19.02
p_1	1	1	0.85	0.75
p_2	0.2	0.167	0.28	0.37
q_1	0.33	0.25	0.28	0.33
q_2	0.78	0.87	0.85	0.83
$p(x)$	0.33	0.41	0.50	0.58
$q(x)$	0.67	0.58	0.50	0.41
$S_p(x)$	0.32	0.29	0.49	0.58
$S_q(x)$	0.56	0.46	0.49	0.51
S	0.48	**0.39**	0.49	0.55

Based on the Table 4.1, the minimum S is 0.39. Therefore, the selected *x* is 7, and Figure 4.2 shows its location for membership function. The calculations are repeated to determine the other two thresholds to form trapezoidal and triangular membership function.

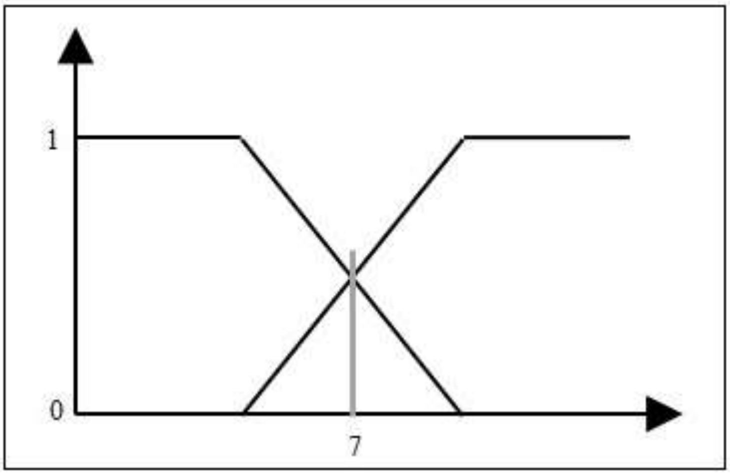


Figure 4.2. Example of Membership Function

The trapezoidal and triangular membership function in Figure 4.3 represents three sets of instances whether the instances need to be kept, removed temporarily or removed permanently. Fuzzy logic thresholds are represented as *a*, *b* and *c*. 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. Balanced data set is produced based on fuzzy thresholds.

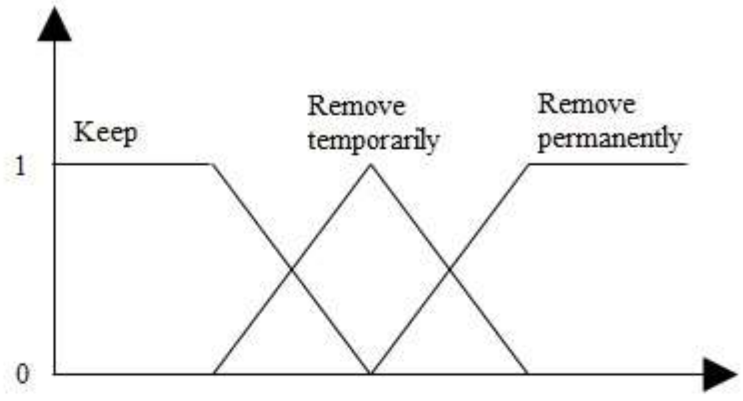


Figure 4.3. Membership Function of Instances

4.2 Experiment and Result

The experiments conducted are designed to minimise the removal of potential data from the majority class by computing the fuzzy logic. The pre-processed imbalanced flood and benchmark data sets 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. After the training and testing sets are applied by the proposed FDUS, SVM is used for classification. The classification is evaluated by accuracy, F-measure and G-mean. The results of the five experiments of each data set will be averaged.

For comparison purposes, the whole process is repeated using different techniques such as DUS and SMOTE. Testing is also made to the data sets without applying any undersampling or oversampling technique to analyse whether the use of those techniques are beneficial.

Table 4.2 and Table 4.3 show the ratio of majority to minority class before and after the techniques have been applied to the flood and benchmark data sets, respectively. From Table 4.2, the ratio after FDUS has been applied to the imbalanced data sets are larger than DUS and SMOTE. Larger ratio indicated the gap between instances in two classes is large. Hence, it shows that FDUS has minimise the removal of potential data.

Table 4.2
Ratio of Majority to Minority Class for Flood Data Sets

Resampling technique / **Data sets**	**No resampling**	**FDUS**	**DUS**	**SMOTE**
Kaki Bukit	2102:1	1865:1	751:1	1043:1
Lubok Sireh	2102:1	1173: 1	757:1	1058:1
Wang Kelian	2074:1	1187:1	781:1	1058:1
Ladang Perlis Selatan	966:1	576:1	341:1	485:1
Ulu Pauh	1231:1	1201:1	434:1	617:1

In Table 4.3, FDUS produced the smallest ratio when it is applied to bupa data set. For adult and pima data sets, FDUS gave the second smallest ratio when compared to DUS and SMOTE. FDUS produced the largest ratio on haberman and breast cancer. The results showed that FDUS has minimised the loss of potential data by removing the instances in majority class based on fuzzy threshold.

Table 4.3
Ratio of Majority to Minority Class for Benchmark Data Sets

Resampling technique / **Data set**	**No resampling**	**FDUS**	**DUS**	**SMOTE**
Adult	3.11:1	2.75:1	2.86:1	1.48:1
Haberman	2.78:1	2.75:1	2.16:1	2.16:1
Breast cancer	1.89:1	1.86:1	1.76:1	1.18:1
Pima	1.87:1	1.22:1	1.13:1	1.39:1
Bupa	1.38:1	1.04:1	1.12:1	3.17:1

Some of the results from previous studies that have been tested on the benchmark data sets are presented in Table 4.4. FDUS is the proposed technique for this research. DUS and combination of Fuzzy Undersampling and Fuzzy Oversampling are techniques categorised as data-based approach while SVM-based Active Learning and Bottom-up induction of Rules and Cases for Imbalanced Data (BRACID) are techniques to solve imbalanced data categorised as algorithm-based approach (Ertekin, Huang, Bottou & Giles, 2007; Li et al., 2010; Li et al., 2013; Napierala & Stefanowski, 2012).

Table 4.4
Benchmark Data Sets Comparison Based On G-Mean

Resampling technique **Data set**	**FDUS**	**DUS**	**Fuzzy Undersampling and Fuzzy Oversampling**	**SVM-based Active Learning**	**BRACID**
Adult	0.86	-	-	0.73	-
Haberman	0.69	0.72	-	-	0.58
Breast cancer	0.91	-	-	-	0.56
Pima	0.65	0.77	0.77	-	0.71
Bupa	0.79	-	0.65	-	-

Results of the experiments are presented in Table 4.5 to Table 4.10 and Figure 4.4 to Figure 4.9.

The results of classification accuracy of no resampling technique, FDUS, DUS and SMOTE are presented in Table 4.5 and Figure 4.4. FDUS produced the best mean classification accuracy on Kaki Bukit and Ulu Pauh. FDUS performed the second best mean classification accuracy on Ladang Perlis. The average result of mean classification accuracy for FDUS is the highest as compared to no resampling, DUS and SMOTE. However, standard deviation of FDUS is the highest. Although the standard deviation is ranked as the highest, the value is considered low as stated in Orriols-Puig and Bernado-Mansilla (2009).

Table 4.5
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

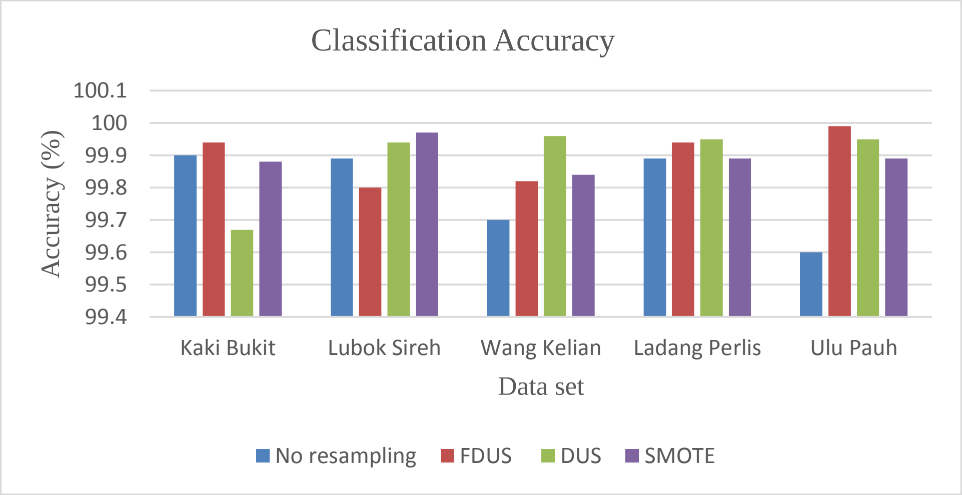


Figure 4.4. Mean Classification Accuracy of Standalone Techniques for Flood Data Sets

Table 4.6 and Figure 4.5 illustrate the F-measure for the proposed FDUS and other resampling techniques. FDUS performed the best when it is applied to Wang Kelian and Ulu Pauh data sets as compared to the other techniques. For Wang Kelian data set, FDUS showed increment of 0.36 than no resampling, 0.32 than SMOTE and 0.10 than DUS. For Ulu Pauh data set, FDUS produced 0.34 better F-measure than no resampling, 0.12 better than DUS and 0.24 better than SMOTE. For the rest of the data sets, FDUS performed as the second best technique. In average, FDUS gave the best F-measure.

Table 4.6
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

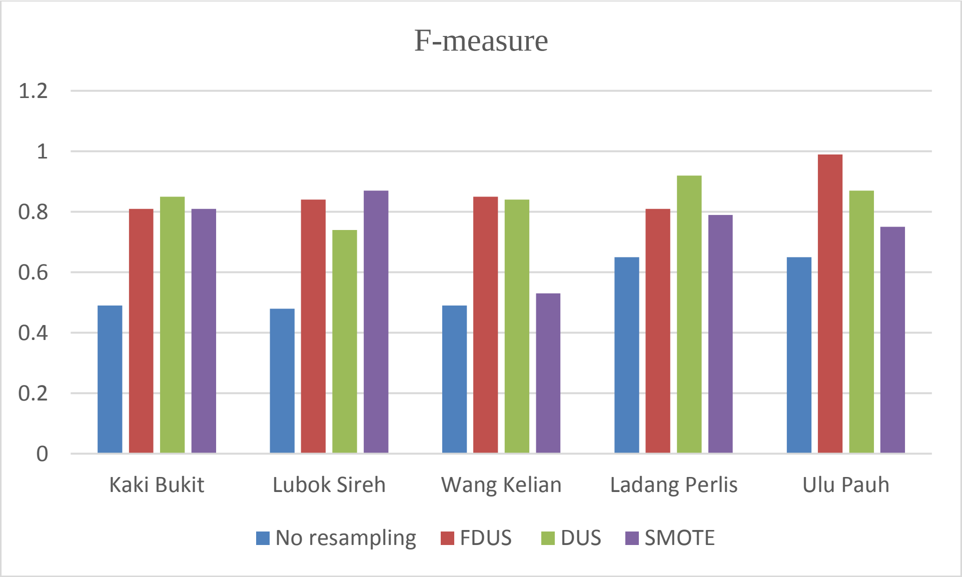


Figure 4.5. F-measure of Standalone Techniques for Flood Data Sets

The results of G-mean for flood data sets are summarised in Table 4.7 and Figure 4.6. The results show that FDUS worked better than DUS and SMOTE for Kaki Bukit, Lubok Sireh, Wang Kelian and Ulu Pauh. For Ladang Perlis, FDUS produced less 0.7 than no resampling and 0.11 than DUS. However, FDUS performed better than SMOTE for 0.09 for Ladang Perlis data set. In average, FDUS performed the second best technique after no resampling.

Table 4.7
G-mean of Standalone Techniques for Flood Data Sets

Resampling technique / **Data set**	**No resampling**	**FDUS**	**DUS**	**SMOTE**
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

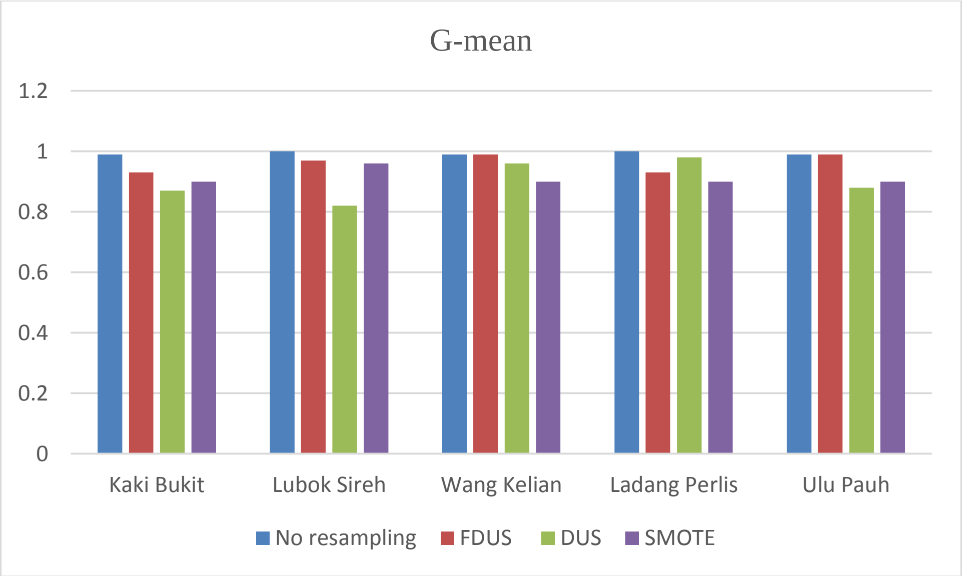


Figure 4.6. G-mean of Standalone Techniques for Flood Data Sets

Table 4.8 and Figure 4.7 describe the classification accuracy of FDUS for benchmark data sets. For three out of five data sets, FDUS gave the best mean classification accuracy as compared to no resampling, DUS and SMOTE. For breast cancer and pima, FDUS performed as the third best technique. The average results show that FDUS has the highest mean classification accuracy and the lowest standard deviation.

Table 4.8
Classification Accuracy (%) of Standalone Techniques for Benchmark Data Sets

Resampling technique	**No resampling**		**FDUS**		**DUS**		**SMOTE**	
Data set	**Mean**	**Standard deviation**	**Mean**	**Standard deviation**	**Mean**	**Standard deviation**	**Mean**	**Standard deviation**
Adult	61.33	0.05	82.61	0.03	75.00	0.03	72.73	0.02
Haberman	75.66	0.10	83.33	0.04	78.90	0.06	75.32	0.14
Breast cancer	93.98	0.04	90.22	0.06	83.15	0.08	94.26	0.04
Pima	67.19	0.04	64.26	0.03	64.91	0.02	59.53	0.02
Bupa	51.16	0.05	82.42	0.05	57.35	0.08	77.10	0.03
Average	69.86	0.06	80.57	0.04	71.86	0.05	75.79	0.05

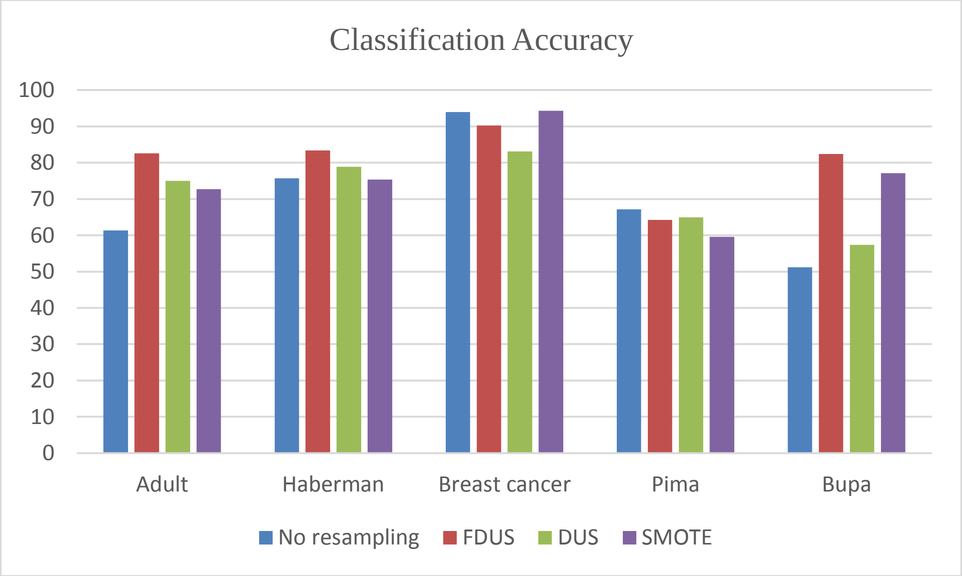


Figure 4.7. Mean Classification Accuracy of Standalone Techniques for Benchmark Data Sets

F-measure for benchmark data sets is presented in Table 4.9 and Figure 4.8. FDUS worked the adult, haberman, pima and bupa data sets. FDUS is positioned as the second best technique for breast cancer data set. Overall, FDUS produce the best F-measure.

Table 4.9
F-measure of Standalone Techniques for Benchmark Data Sets

Resampling technique	**No resampling**	**FDUS**	**DUS**	**SMOTE**
Data set				
Adult	0.12	0.91	0.42	0.24
Haberman	0.35	0.90	0.65	0.66
Breast cancer	0.91	0.92	0.86	0.94
Pima	0.54	0.64	0.55	0.58
Bupa	0.45	0.88	0.61	0.40
Average	0.47	0.85	0.62	0.57

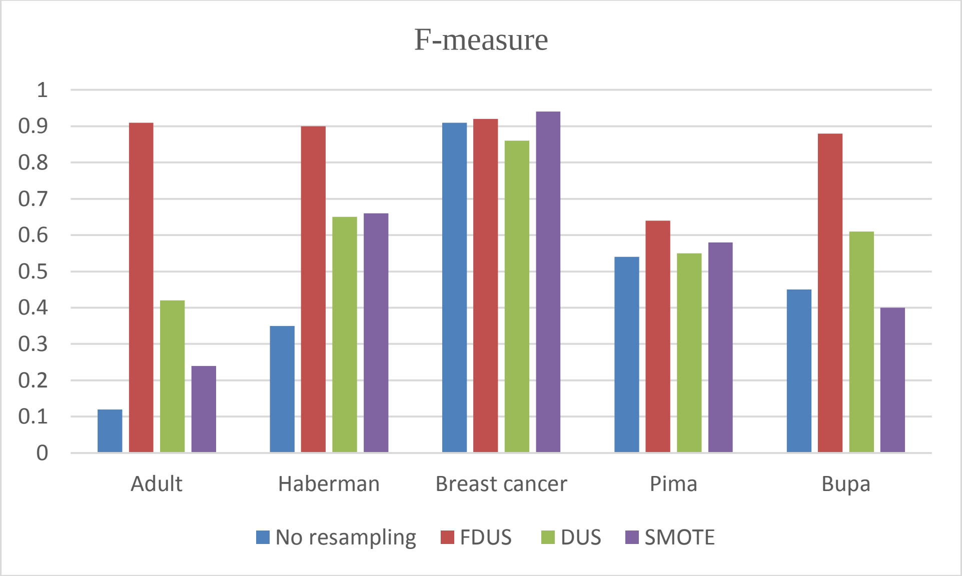


Figure 4.8. F-measure of Standalone Techniques for Benchmark Data Sets

The result of G-mean is depicted in Table 4.10 and Figure 4.9. The results show that FDUS presented the best G-mean on three data sets which are adult, pima and bupa data sets. FDUS is ranked as the third best technique for haberman and breast cancer data sets. In average, FDUS produced the best G-mean as compared to no resampling, DUS and SMOTE.

Table 4.10
G-mean of Standalone Techniques for Benchmark Data Sets

Resampling technique	**No resampling**	**FDUS**	**DUS**	**SMOTE**
Data set				
Adult	0.29	0.86	0.52	0.38
Haberman	0.48	0.69	0.70	0.74
Breast cancer	0.93	0.91	0.84	0.94
Pima	0.64	0.65	0.61	0.63
Bupa	0.51	0.79	0.57	0.53
Average	0.57	0.78	0.65	0.64

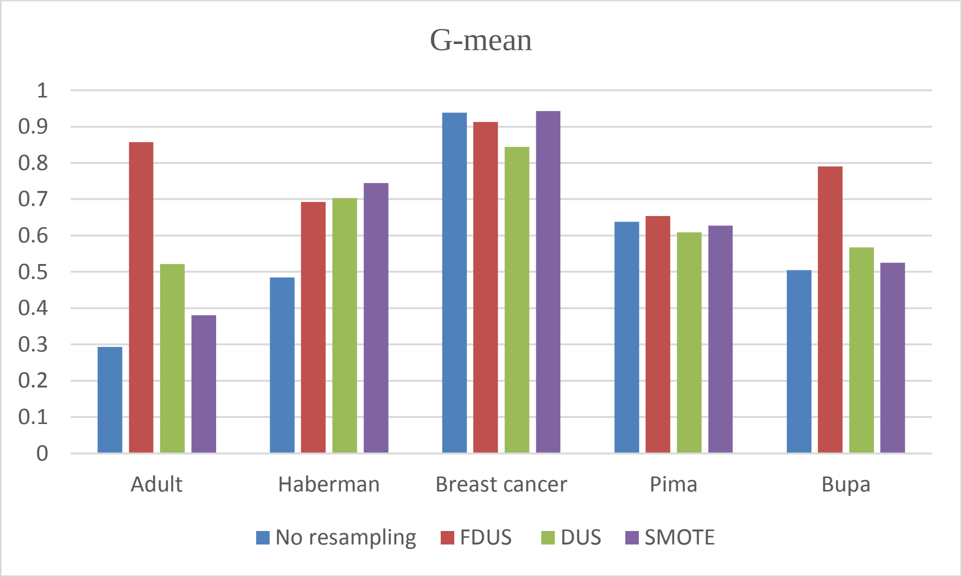


Figure 4.9. G-mean of Standalone Techniques for Benchmark Data Sets

Overall, it is apparent that FDUS achieved higher classification accuracy and F-measure for both flood and benchmark data sets. FDUS achieved the highest G-mean for benchmark data sets and the second best G-mean for flood data sets.

The results of classification accuracy indicated that FDUS allows SVM to classify correctly the data sets specifically on the Kaki Bukit and Ulu Pauh data sets. The classification accuracy is higher on Kaki Bukit and Ulu Pauh data sets because after applying FDUS on the data sets, the ratio between majority and minority classes has become smaller. However, for the other flood data sets, FDUS has lower classification accuracy than no resampling, DUS and SMOTE. This might happened due to other factors such as size, complexity, overlap and small disjuncts (Visa & Ralescu, 2005; Sun et al., 2009; Barua et al., 2014). For benchmark data sets, FDUS performed the best on most of the benchmark data. The classification accuracy of no resampling technique is higher than the accuracy after the data sets are resampled due to the tendency of classifier to ignore the instances in the minority class.

F-measure determined the exactness of the correctly labelled minority class. FDUS performed the best for the maximum number of time for benchmark data sets. For flood data sets, FDUS appeared as the best technique for two times and second best techniques 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 for all flood data sets. However, FDUS is outperformed by no resampling because sensitivity and specificity are high. For benchmark data sets, FDUS has succesfully reduced the number of instances in the majority class and narrowed the difference between both classes. FDUS uses the advantage of fuzzy logic to avoid biasness in choosing the instances that need to be removed from the majority class.

4.3 Summary

This chapter provided detailed descriptions of the enhancement of Distance-based Undersampling technique to produce Fuzzy Distance-based Undersampling technique. Overall, the experimental results from several imbalanced data sets indicated that the proposed Fuzzy Distance-based Undersampling technique provided better classification accuracy, F-measure and G-mean, as compared to Distance-based Undersampling and SMOTE for benchmark data sets. For flood data sets, FDUS outperformed the other techniques based on classification accuracy and F-measure. In term of G-mean, FDUS is ranked as the second best technique. The Fuzzy Distance-based Undersampling technique has utilised the benefit of fuzzy logic, which is to reduce bias.

CHAPTER FIVE

INTEGRATION OF FUZZY DISTANCE-BASED UNDERSAMPLING AND SMOTE

This chapter outlines the enhancement of resampling technique. The chapter begins with the explanation of the steps to enhance the proposed resampling technique in Section 5.1. The experiments and results of the proposed technique is discussed in Section 5.2. Finally, Section 5.3 presents the summary of the chapter.

5.1 Proposed Enhanced Resampling Technique

The proposed enhanced resampling technique is a integration of Fuzzy Distance-based Undersampling (FDUS) and Synthetic Minority Oversampling TEchnique (SMOTE) that work in sequence. Figure 5.1 illustrates the proposed combination of FDUS and SMOTE.

The flow starts by taking the imbalanced data set as input data. The imbalanced data set is divided into two classes. Class that has less instances is known as the minority class, while the other class is known as the majority class. 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 the FDUS process 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 SMOTE algorithm is shown in Figure 5.2. The process is repeated until a balanced data set is produced. Note that, a data set with ratio of 1:1 is perfectly balanced such that $|A_i| = |B_j|$. However, since the data sets used in this research have ratio of as high as 2000:1, it is impossible to achieve the ratio of perfectly balanced. The iteration of this process stopped when the ratio achieve the most balanced ratio after several iterations. The idea is adapted from Garcia et al. (2007) and Weiss and Provost (2003).

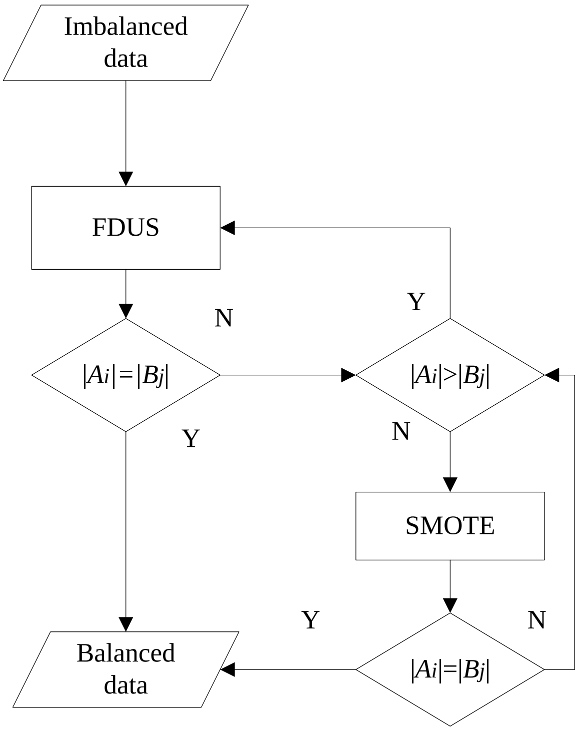


Figure 5.1. Integration of Fuzzy Distance-based Undersampling and Synthetic Minority Oversampling TEchnique

SMOTE is an oversampling technique which randomly creates new synthetic samples to the minority class. Let *M* be the minority class, and *x* be the instances in *M*. For each instance *x* in *M*, k-nearest neighbour is found. Samples are randomly selected from k-nearest neighbour instances, denoted as *y*. Synthetic instances known as *q* are created using Equation 5.1. Then, q is added to M. Finally, a new minority data set is created. SMOTE algorithm is depicted in Figure 5.2

$$q = x + (x - y) \times gap \tag{5.1}$$

where q = Synthetic instances, gap is random number from 0 to 1.

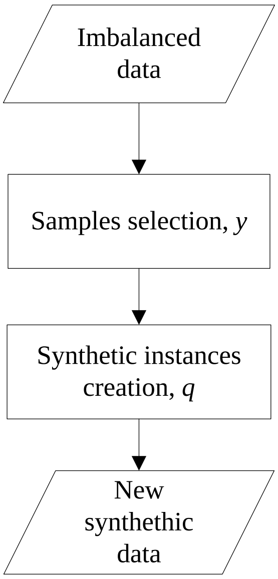


Figure 5.2. Synthetic Minority Oversampling TEchnique (Chawla et al., 2002)

5.2 Experiment and Result

The experiments are carried out to combine the undersampling and oversampling techniques to improve the result produced by the standalone techniques. The flood and benchmark data sets are used for this experiment. The imbalanced data sets are divided into majority and minority classes. Then, these data sets are partitioned using 5-fold cross validation. The data sets are trained and tested on the enhanced resampling technique named as FDUS+SMOTE to produce balanced data sets. SVM is used to classify the data sets, and the classification performance of the proposed enhanced resampling technique is evaluated by accuracy, F-measure and G-mean.

For comparison purposes, the whole process is repeated and replaced by different techniques which are the combination of SMOTE+TL and SMOTE+ENN. The results of the enhanced resampling technique are also compared with no resampling technique, FDUS, DUS and SMOTE.

Table 5.1 and Table 5.2 show the ratio before and after the techniques are applied to the flood and benchmark data sets. From Table 5.1, the smallest ratio produced by FDUS+SMOTE as compared to SMOTE+TL and SMOTE+ENN is after it has been applied on Kaki Bukit. The second smallest ratio produced by FDUS+SMOTE is when it is applied to the other four data sets. Based on conducted experiments, Lubok Sireh, Wang Kelian, Ladang Perlis Selatan and Ulu Pauh data sets fulfilled only the first condition of FDUS+SMOTE which is $|A_i|>|B_j|$ where $|A_i|$ is number of instances in majority class and $|B_j|$ is number of instances in minority class. The final ratio is the best ratio produced after several iteration. For Wang Kelian data set, SMOTE is applied after the data set is processed using FDUS where $|A_i|<|B_j|$. It may happened to Wang Kelian data set due to the distribution of the instances might be overlapping that cause to high number of removal of instances in the majority class.

Table 5.1
Ratio of Majority to Minority Class for Flood Data Sets

Resampling technique Data sets	No resampling	FDUS+SMOTE	SMOTE+TL	SMOTE+ENN
Kaki Bukit	2102:1	321:1	568:1	1043:1
Lubok Sireh	2102:1	45:1	30:1	215:1
Wang Kelian	2074:1	39:1	30:1	334:1
Ladang Perlis Selatan	966:1	60:1	20:1	348:1
Ulu Pauh	1231:1	35:1	10:1	128:1

From Table 5.2, FDUS+SMOTE produced the smallest ratio as compared to SMOTE+TL and SMOTE+ENN on bupa data sets. All benchmark data sets fulfilled the second condition of FDUS+SMOTE which is $|A_i|<|B_j|$. After the data sets are processed by FDUS, $|A_i|$ has become smaller than $|B_j|$. SMOTE is used increase the number of instances in the new A_i to produce the best ratio. Table 5.2 show the best ratio produced by FDUS+SMOTE after several iteration.

Table 5.2
Ratio of Majority to Minority Class for Benchmark Data Sets

Resampling technique Data set	No resampling	FDUS+SMOTE	SMOTE+TL	SMOTE+ENN
Adult	3.11:1	2.33:1	2.80:1	1.35:1
Haberman	2.78:1	2.75:1	2.18:1	2.06:1
Breast cancer	1.89:1	1.66:1	1.12:1	1.06:1
Pima	1.87:1	1.99:1	1.39:1	1.36:1
Bupa	1.38:1	1.19:1	2.61:1	2.05:1

The results of the FDUS+SMOTE, no resampling technique, SMOTE+TL and SMOTE+ENN are presented in Table 5.3 to Table 5.8 and Figure 5.3 to Figure 5.8. The results of the enhanced resampling technique, no resampling technique, FDUS,

DUS and SMOTE are depicted in Table 5.9 to Table 5.14 and Figure 5.9 to Figure 5.14.

Table 5.3 and Figure 5.3 show the results of the classification performance for flood data sets. FDUS+SMOTE gave the best mean classification accuracy when it is used on Ulu Pauh. FDUS+SMOTE performed the second best mean classification accuracy on Lubok Sireh and Wang Kelian and the third best on Kaki Bukit and Ladang Perlis. The average result of mean classification accuracy for FDUS+SMOTE is higher than SMOTE+TL and SMOTE+ENN. The average standard deviation of FDUS+SMOTE is higher than SMOTE+ENN and similar to SMOTE+TL.

Table 5.3
Classification Accuracy (%) of Combination Techniques for Flood Data Sets

Resampling technique	**No resampling**		**FDUS+SMOTE**		**SMOTE+TL**		**SMOTE+ENN**	
Data set	**Mean**	**Standard deviation**	**Mean**	**Standard deviation**	**Mean**	**Standard deviation**	**Mean**	**Standard deviation**
Kaki Bukit	99.90	1.87	99.92	0.02	96.81	0.02	99.97	0.00
Lubok Sireh	99.89	0.22	99.80	0.04	93.31	0.04	88.33	0.04
Wang Kelian	99.70	0.34	99.64	0.03	93.24	0.03	92.96	0.02
Ladang Perlis	99.89	0.22	99.74	0.03	95.53	0.03	99.99	0.00
Ulu Pauh	99.60	0.23	99.77	0.03	96.43	0.03	087.53	0.05
Average	99.80	0.58	99.77	0.03	95.06	0.03	93.76	0.02

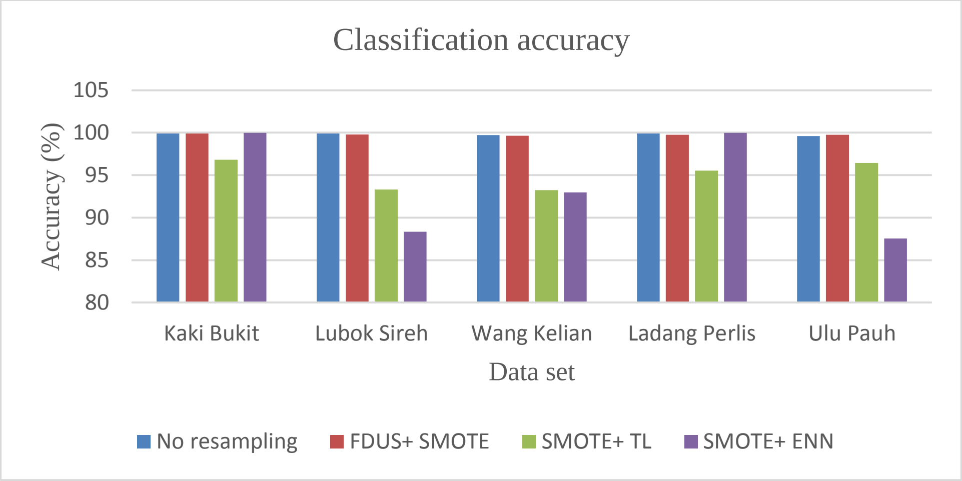


Figure 5.3. Mean Classification Accuracy of Combination Techniques for Flood Data Sets

Table 5.4 and Figure 5.4 show that FDUS+SMOTE produced the best F-measure for all flood data sets. The highest performance is produced on Ladang Perlis data set. FDUS+SMOTE increased the highest F-measure up to 0.47 as compared to no sampling when applied on Lubok Sireh data set. The results show that the biggest gap between FDUS+SMOTE with SMOTE+TL is 0.6 on Wang Kelian data set and FDUS+SMOTE with SMOTE+ENN is 0.62 on Ulu Pauh data set. In average, F-measure of FDUS+SMOTE is the best.

Table 5.4
F-measure of Combination Techniques for Flood Data Sets

Resampling technique **Data set**	**No resampling**	**FDUS+ SMOTE**	**SMOTE+TL**	**SMOTE+ENN**
Kaki Bukit	0.49	0.87	0.86	0.80
Lubok Sireh	0.48	0.95	0.89	0.29
Wang Kelian	0.49	0.93	0.89	0.70
Ladang Perlis	0.65	0.99	0.95	0.88
Ulu Pauh	0.65	0.96	0.95	0.34
Average	0.55	0.94	0.91	0.60

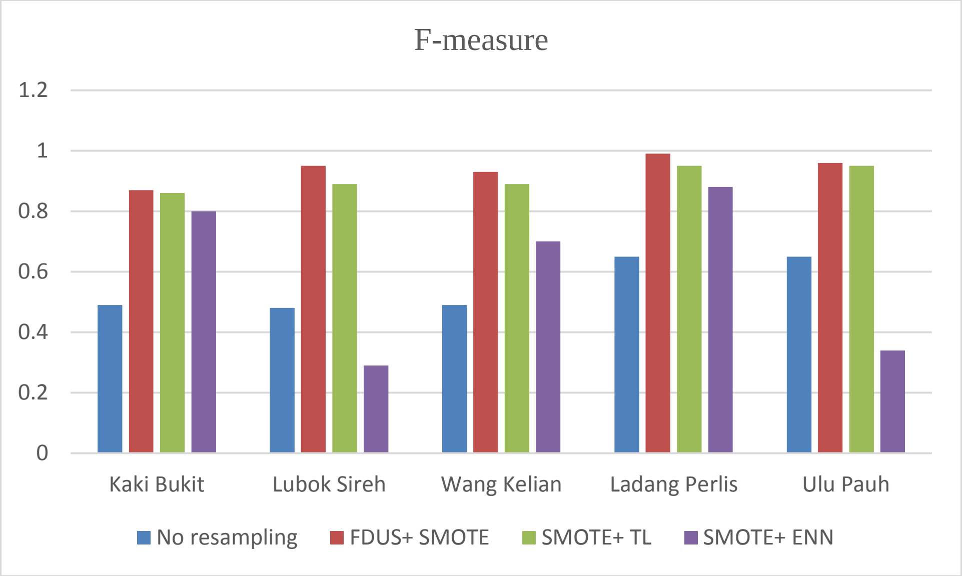


Figure 5.4. F-measure of Combination Techniques for Flood Data Sets

The results of G-mean for flood data sets are shown in Table 5.5 and Figure 5.5. The proposed FDUS+SMOTE performed better than SMOTE+TL and SMOTE+ENN on all data sets. No resampling outperformed all techniques. In average, G-mean of FDUS+SMOTE is the second best.

Table 5.5
G-mean of Combination Techniques for Flood Data Sets

Resampling technique	**No resampling**	**FDUS+ SMOTE**	**SMOTE+TL**	**SMOTE+ENN**
Data set				
Kaki Bukit	0.99	0.95	0.87	0.84
Lubok Sireh	1.00	0.97	0.95	0.78
Wang Kelian	0.99	0.99	0.95	0.93
Ladang Perlis	1.00	0.99	0.96	0.90
Ulu Pauh	0.99	0.99	0.97	0.90
Average	0.99	0.98	0.94	0.87

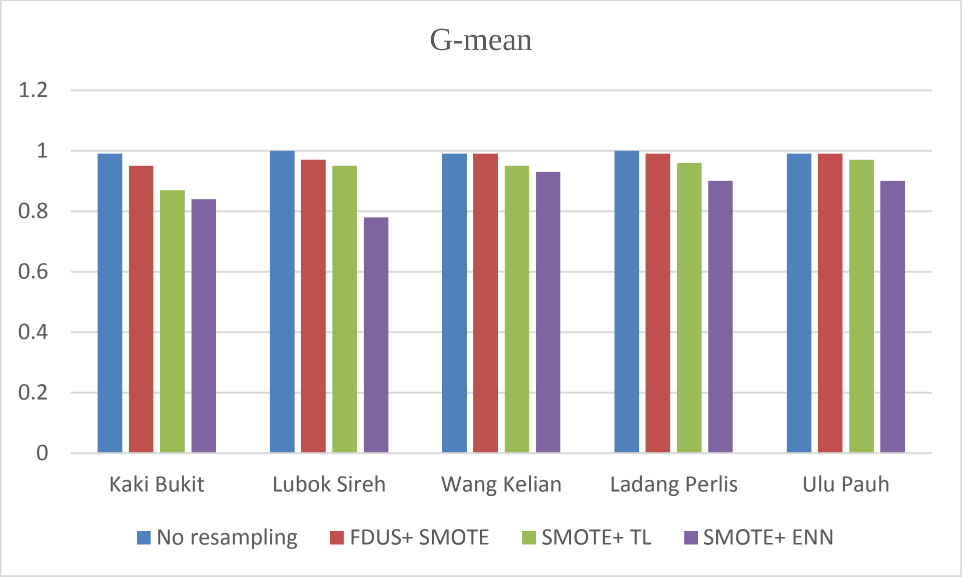


Figure 5.5. G-mean of Combination Techniques for Flood Data Sets

The classification accuracy for benchmark data sets are depicted in Table 5.6 and Figure 5.6. FDUS+SMOTE produced the best mean classification accuracy on adult and bupa data sets. For breast cancer and pima data sets, FDUS+SMOTE performed as the best second technique. Overall, FDUS+SMOTE is ranked as the best techniques. However, the standard deviation is the highest which is 0.13. Although the standard deviation is high, 0.13 is considered low (Orriols-Puig & Bernado-Mansilla, 2009).

Table 5.6

Classification Accuracy (%) of Combination Techniques for Benchmark Data Sets

Resampling technique	**No resampling**		**FDUS+SMOTE**		**SMOTE+TL**		**SMOTE+ENN**	
Data set	**Mean**	**Standard deviation**	**Mean**	**Standard deviation**	**Mean**	**Standard deviation**	**Mean**	**Standard deviation**
Adult	61.33	0.05	93.10	0.09	75.00	0.02	39.06	0.05
Haberman	75.66	0.10	62.96	0.12	77.22	0.06	91.19	0.07
Breast cancer	93.98	0.04	94.38	0.06	94.30	0.05	94.87	0.05
Pima	67.19	0.04	63.25	0.15	60.89	0.04	63.44	0.04
Bupa	51.16	0.05	82.42	0.25	69.64	0.03	72.22	0.04
Average	69.86	0.06	79.22	0.13	75.41	0.04	72.16	0.05

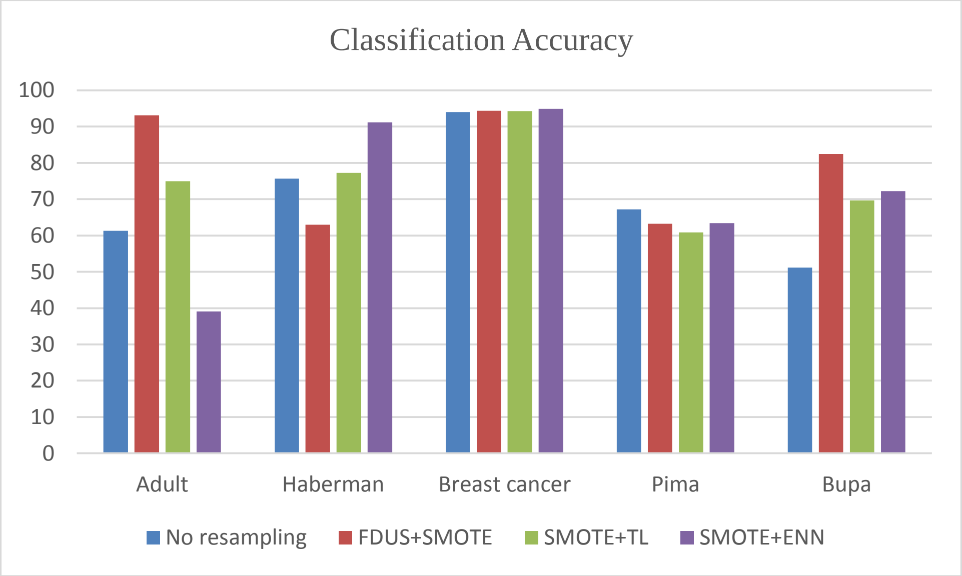


Figure 5.6. Mean Classification Accuracy of Combination Techniques for Benchmark Data Sets

Table 5.7 and Figure 5.7 present the results of F-measure for benchmark data sets. FDUS+SMOTE gave the best performance on all data sets. The highest gap between FDUS+SMOTE and other techniques is shown on adult data set, where FDUS+SMOTE performed 0.8 better than no resampling, 0.58 than SMOTE+TL and 0.66 than SMOTE+ENN. In average, F-measure of FDUS+SMOTE is the best.

Table 5.7

F-measure of Combination Techniques for Benchmark Data Sets

Resampling technique Data set	No resampling	FDUS+ SMOTE	SMOTE+TL	SMOTE+ENN
Adult	0.12	0.92	0.34	0.26
Haberman	0.35	0.86	0.63	0.79
Breast cancer	0.91	0.96	0.95	0.95
Pima	0.54	0.82	0.59	0.66
Bupa	0.45	0.80	0.61	0.71
Average	0.47	0.87	0.63	0.68

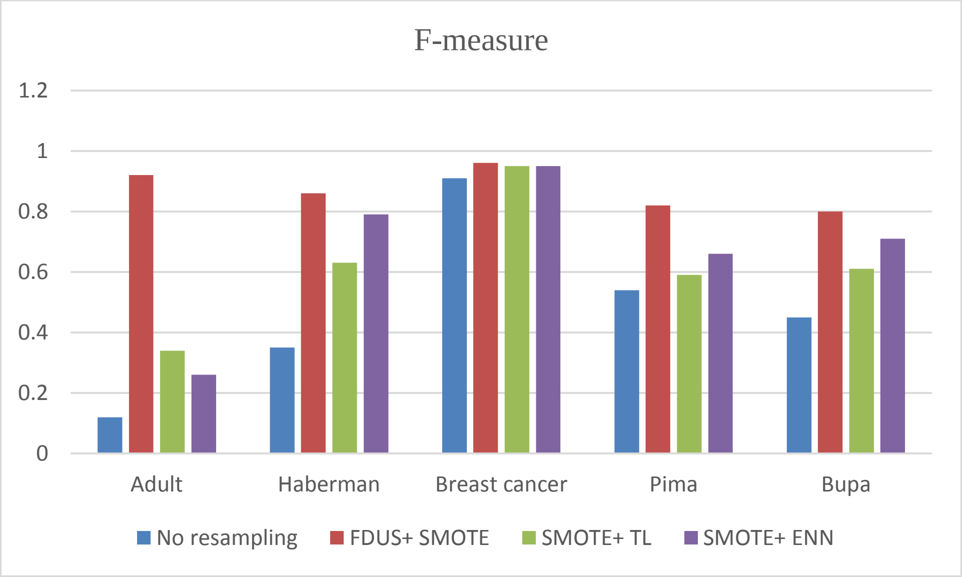


Figure 5.7. F-measure of Combination Techniques for Benchmark Data Sets

The results of G-mean for benchmark data sets are illustrated in Table 5.8 and Figure 5.8. Out of five data sets, FDUS+SMOTE worked the best on four data sets. The highest gap between FDUS+SMOTE and no resampling is 0.37, which can be seen on haberman data set. On adult data set, FDUS+SMOTE showed a significant improvement of 0.46 as compared to SMOTE+TL, and 0.58 as compared to SMOTE+ENN. Generally, FDUS+SMOTE produced the best G-mean.

Table 5.8
G-mean of Combination Techniques for Benchmark Data Sets

Resampling technique	**No resampling**	**FDUS+SMOTE**	**SMOTE+TL**	**SMOTE+ENN**
Data set				
Adult	0.29	0.94	0.48	0.36
Haberman	0.48	0.85	0.72	0.73
Breast cancer	0.94	0.96	0.94	0.95
Pima	0.64	0.65	0.62	0.70
Bupa	0.51	0.81	0.71	0.73
Average	0.57	0.84	0.70	0.69

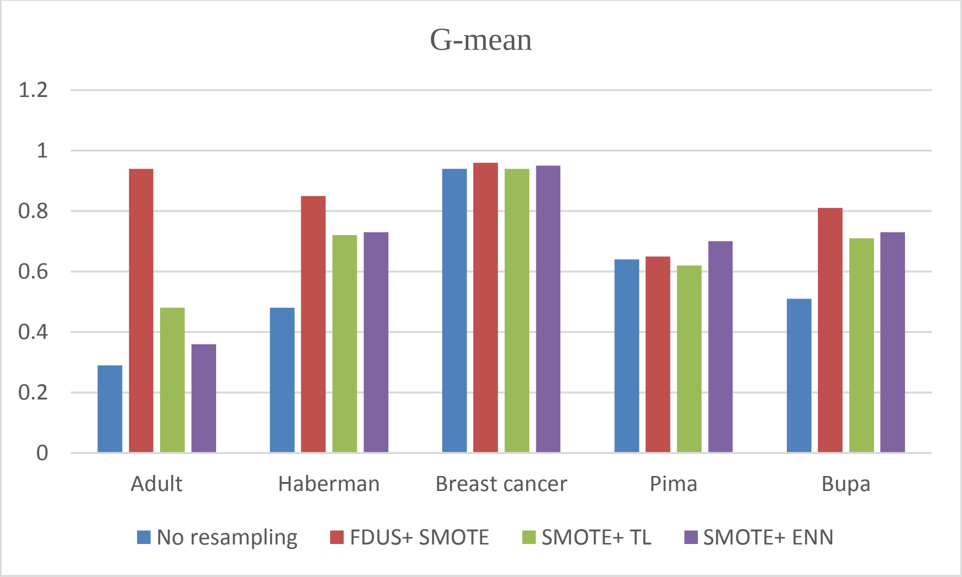


Figure 5.8. G-mean of Combination Techniques for Benchmark Data Sets

Analysing the results, FDUS+SMOTE showed better classification accuracy than SMOTE+TL and SMOTE+ENN on benchmark data sets. FDUS+SMOTE has modified the imbalanced data sets so that SVM can classify the data sets without misclassification However, the results of no resampling technique gave better classification accuracy than FDUS+SMOTE on flood data sets. Even though the classification accuracy of no resampling were better, the results were biased towards the majority class as the classifier tends to minimise the misclassification by classifying all the samples in the majority class.

The results of F-measure of FDUS+SMOTE is higher than the other techniques on both flood and benchmark data sets. The high percentage of F-measure indicated that the percentage of the instances in the minority class that are correctly labelled is high. G-mean of FDUS+SMOTE is also better than SMOTE+TL and SMOTE+ENN. For flood data sets, although the value of F-measure is the highest, the value of G-mean is lower than no resampling. This indicates that FDUS+SMOTE may introduce other complexities or may create new small disjuncts in the instances. However, FDUS+SMOTE has increased the G-mean for benchmark data sets.

The results showed that FDUS+SMOTE has modified the imbalanced data sets to become more balanced than using the other techniques. FDUS+SMOTE used both FDUS and SMOTE's advantages, which are minimising the loss of potential data from the majority class and creating synthetic samples for the minority class to avoid overfitting.

The results of comparisons between the proposed enhanced resampling technique and standalone techniques are presented in Table 5.9 to Table 5.14 and Figure 5.9 to Figure 5.14.

Table 5.9 and Figure 5.9 show the comparison of classification accuracy between FDUS+SMOTE and standalone techniques for flood data sets. The results showed that FDUS+SMOTE has the same mean classification with FDUS when performed on

Lubok Sireh, and performed better than DUS and SMOTE on Kaki Bukit data set. In average, the mean classification of FDUS+SMOTE is lowest as compared to the other techniques. However, FDUS+SMOTE shows the lowest standard deviation.

Table 5.9
Classification Accuracy (%) of FDUS+SMOTE and Standalone Techniques for Flood Data Sets

Resampling technique	**No resampling**		**FDUS+SMOTE**		**FDUS**		**DUS**		**SMOTE**	
Data set	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
Kaki Bukit	99.90	1.87	99.92	0.01	99.94	0.61	99.67	0.06	99.88	0.19
Lubok Sireh	99.89	0.22	99.80	0.03	99.80	1.96	99.94	0.56	99.97	0.19
Wang Kelian	99.70	0.34	99.64	0.03	99.82	1.98	99.96	0.49	99.84	0.20
Ladang Perlis	99.89	0.22	99.74	0.02	99.94	0.61	99.95	0.63	99.89	0.39
Ulu Pauh	99.60	0.23	99.77	0.02	99.99	1.22	99.95	0.64	99.89	0.22
Average	99.80	0.58	99.77	0.02	99.90	1.28	99.89	0.48	99.89	0.24

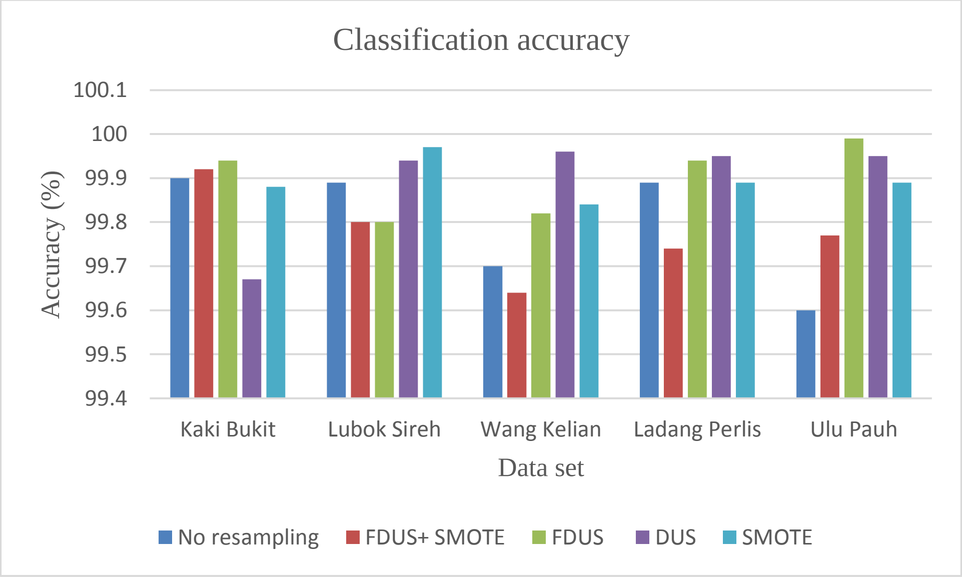


Figure 5.9. Mean Classification Accuracy of FDUS+SMOTE and Standalone Techniques for Flood Data Sets

However, as presented in Table 5.10 and Figure 5.10, FDUS+SMOTE gave better F-measure than the standalone techniques for all flood data sets except for Ulu Pauh. FDUS+SMOTE produced slightly 0.03 less than FDUS for Ulu Pauh data set. FDUS+SMOTE is found to performed better than DUS and SMOTE. In average, FDUS+SMOTE is better than the other techniques.

Table 5.10
F-measure of FDUS+SMOTE and Standalone Techniques for Flood Data Sets

Resampling technique **Data set**	**No resampling**	**FDUS+ SMOTE**	**FDUS**	**DUS**	**SMOTE**
Kaki Bukit	0.49	0.87	0.81	0.85	0.81
Lubok Sireh	0.48	0.95	0.84	0.74	0.87
Wang Kelian	0.49	0.93	0.85	0.84	0.53
Ladang Perlis	0.65	0.99	0.81	0.92	0.79
Ulu Pauh	0.65	0.96	0.99	0.87	0.75
Average	0.55	0.94	0.86	0.84	0.75

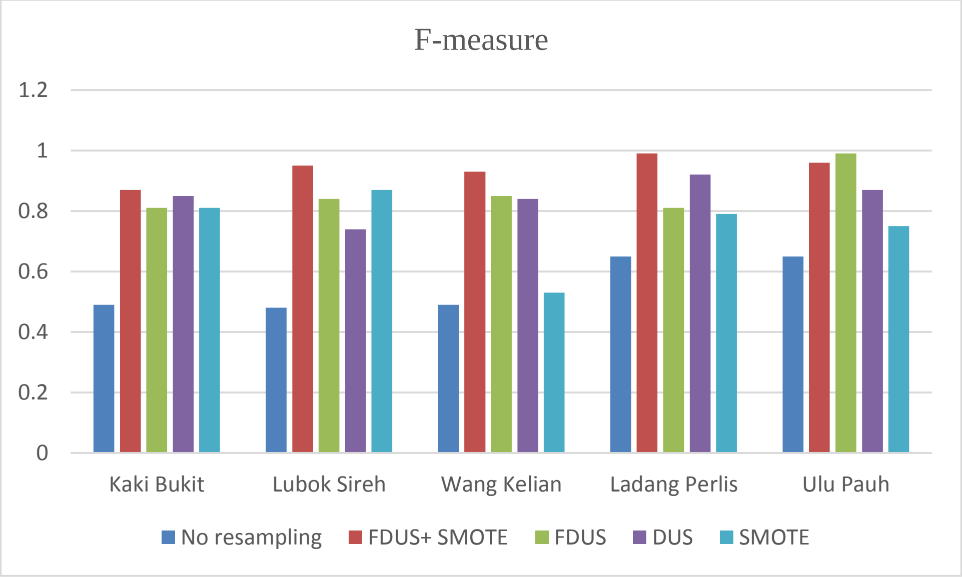


Figure 5.10. F-measure of FDUS+SMOTE and Standalone Techniques for Flood Data Sets

Table 5.11 and Figure 5.11 show that FDUS+SMOTE performed better than standalone techniques in terms of G-mean for four flood data sets. The only exception is on Lubok Sireh data set where FDUS+SMOTE performed similar with FDUS but performed better 0.15 than DUS and 0.01 than SMOTE. Overall, FDUS+SMOTE made the second best G-mean.

Table 5.11
G-mean of FDUS+SMOTE and Standalone Techniques for Flood Data Sets

Resampling technique **Data set**	**No resampling**	**FDUS+ SMOTE**	**FDUS**	**DUS**	**SMOTE**
Kaki Bukit	0.99	0.95	0.93	0.87	0.90
Lubok Sireh	1.00	0.97	0.97	0.82	0.96
Wang Kelian	0.99	0.99	0.99	0.96	0.90
Ladang Perlis	1.00	0.99	0.93	0.98	0.90
Ulu Pauh	0.99	0.99	0.99	0.88	0.90
Average	0.99	0.98	0.96	0.90	0.91

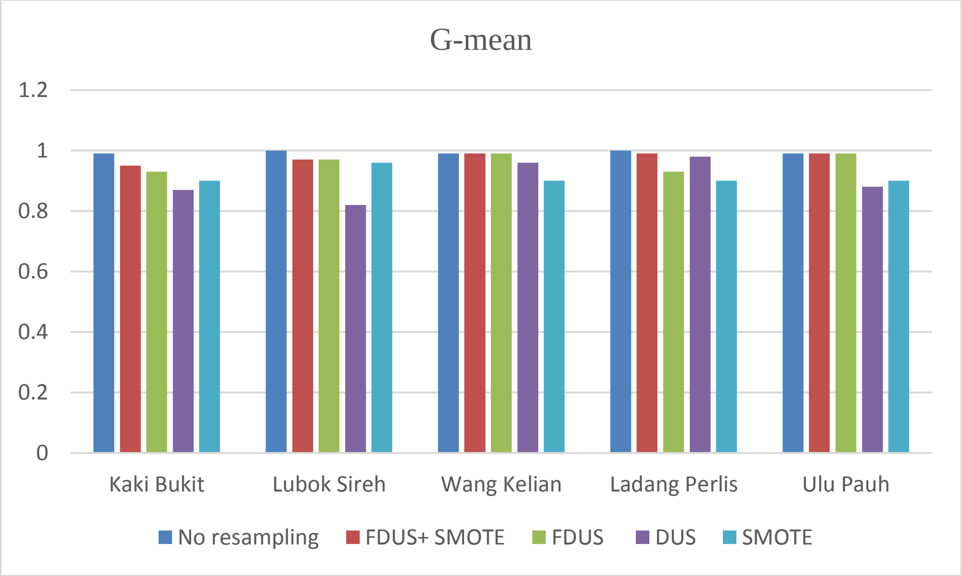


Figure 5.11. G-mean of FDUS+SMOTE and Standalone Techniques for Flood Data Sets

Table 5.12 and Figure 5.12 present the classification accuracy for benchmark data sets of the FDUS+SMOTE and the standalone techniques. Out of five data sets, FDUS+SMOTE outperformed the standalone techniques on two data sets namely adult and breast cancer. FDUS+SMOTE is equivalent to FDUS when it is applied to bupa data set. Overall, FDUS+SMOTE performed better than DUS and SMOTE but not FDUS.

Table 5.12
Classification Accuracy (%) of FDUS+SMOTE and Standalone Techniques for Benchmark Data Sets

Resampling technique	**No resampling**		**FDUS+SMOTE**		**FDUS**		**DUS**		**SMOTE**	
Data set	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
Adult	61.33	0.05	93.10	0.09	82.61	0.03	75.00	0.03	72.73	0.02
Haberman	75.66	0.10	62.96	0.12	83.33	0.04	78.90	0.06	75.32	0.14
Breast cancer	93.98	0.04	94.38	0.06	90.22	0.06	83.15	0.08	75.32	0.14
Pima	67.19	0.04	63.25	0.15	64.26	0.03	64.91	0.02	59.53	0.02
Bupa	51.16	0.05	82.42	0.25	82.42	0.25	57.35	0.08	77.10	0.03
Average	69.86	0.06	79.22	0.13	80.57	0.08	71.86	0.05	75.79	0.05

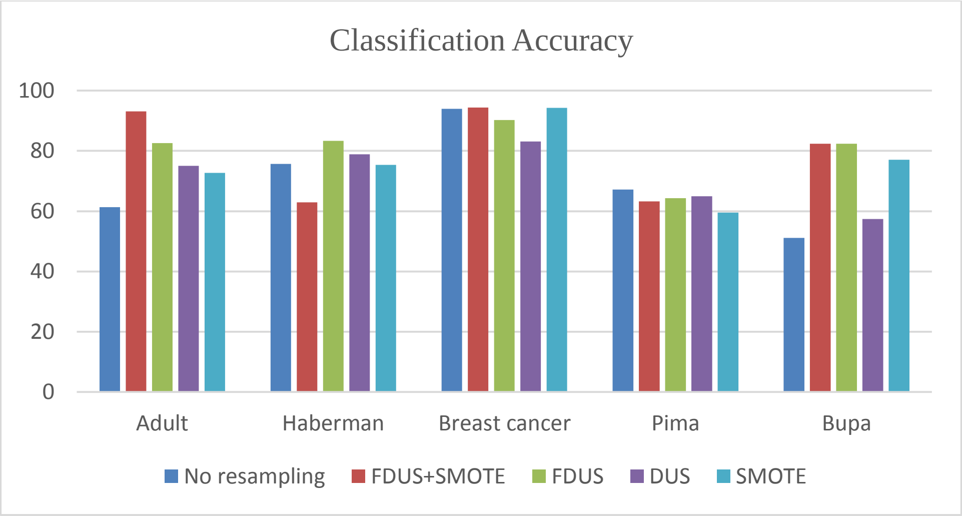


Figure 5.12. Mean Classification Accuracy of FDUS+SMOTE and Standalone Techniques for Benchmark Data Sets

In terms of F-measure, FDUS+SMOTE gave the best performance on three data sets as presented on Table 5.13 and Figure 5.13. For the other two data sets, FDUS+SMOTE performed as the second best technique; outperformed by FDUS by 0.04 on haberman data set and 0.11 on bupa data set. FDUS+SMOTE produced the best F-measure in general.

Table 5.13
F-measure of FDUS+SMOTE and Standalone Techniques for Benchmark Data Sets

Resampling technique	**No resampling**	**FDUS+SMOTE**	**FDUS**	**DUS**	**SMOTE**
Data set					
Adult	0.12	0.92	0.90	0.41	0.24
Haberman	0.35	0.85	0.89	0.64	0.66
Breast cancer	0.91	0.96	0.92	0.8	0.94
Pima	0.54	0.81	0.64	0.54	0.58
Bupa	0.44	0.79	0.88	0.60	0.40
Average	0.47	0.87	0.85	0.61	0.56

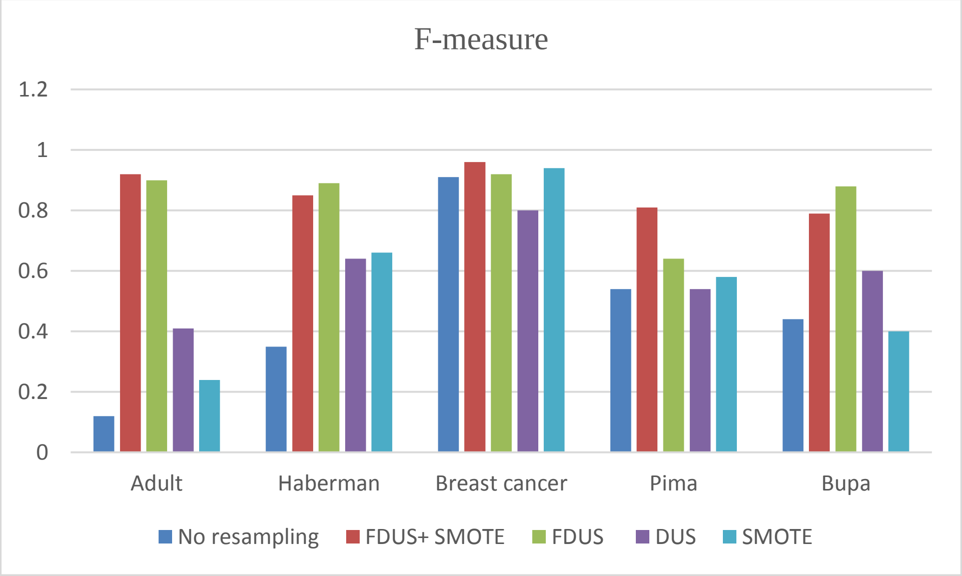


Figure 5.13. F-measure of FDUS+SMOTE and Standalone Techniques for Benchmark Data Sets

Table 5.14 and Figure 5.14 present the comparison between FDUS+SMOTE with FDUS, DUS and SMOTE in terms of G-mean. The results showed that FDUS+SMOTE outperformed the standalone techniques on four data sets. FDUS+SMOTE performed similarly with FDUS on pima data set. Overall, FDUS+SMOTE show the best G-mean.

Table 5.14
G-mean of FDUS+SMOTE and Standalone Techniques for Benchmark Data Sets

Resampling technique **Data set**	**No resampling**	**FDUS+ SMOTE**	**FDUS**	**DUS**	**SMOTE**
Adult	0.29	0.94	0.85	0.52	0.38
Haberman	0.48	0.85	0.69	0.70	0.74
Breast cancer	0.94	0.96	0.91	0.84	0.94
Pima	0.64	0.65	0.65	0.60	0.62
Bupa	0.51	0.81	0.79	0.56	0.52
Average	0.57	0.84	0.78	0.65	0.64

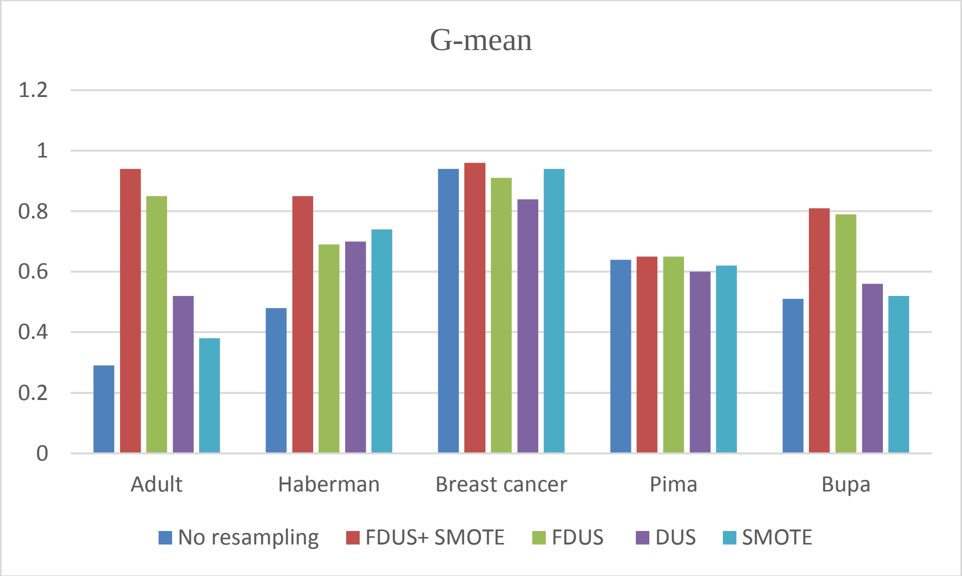


Figure 5.14. G-mean of FDUS+SMOTE and Standalone Techniques for Benchmark Data Sets

From the results of the comparison between FDUS+SMOTE and the standalone techniques, the classification accuracy of FDUS, DUS and SMOTE are better than FDUS+SMOTE for flood data sets. FDUS performed as the best technique for flood data sets. Since the size of the flood data sets is big, which is greater than 10,000 instances, FDUS alone is more suitable to balance the data sets without the need to repeat the process for several times to find the best ratio. It might be the problem related to flood data sets is not related to imbalanced ratio but size, complexity, overlap or small disjuncts (Visa & Ralescu, 2005; Sun et al., 2009; Baru et al., 2014). For benchmark data sets, FDUS+SMOTE performed better than DUS and SMOTE. However, FDUS+SMOTE is outperformed by FDUS on haberman and pima because the ratio after FDUS is applied to the data sets are smaller than FDUS+SMOTE.

Based on F-measure, FDUS+SMOTE performed better than standalone techniques. The results proved that FDUS+SMOTE is able to modify the imbalanced data sets so that the minority class is correctly labelled as compared to standalone techniques. FDUS+SMOTE performed better than standalone techniques when FDUS+SMOTE is applied on imbalanced data sets that have smaller gaps between minority and majority classes. FDUS+SMOTE fully used the advantages of both FDUS and SMOTE to balance the data sets and reduce the ratio of instances in majority and minority classes as presented based on G-mean.

5.3 Summary

The details regarding the proposed enhancement of resampling technique named as FDUS+SMOTE is presented in this chapter. Fuzzy Distance-based Undersampling

(FDUS) is combined with SMOTE to balance the data sets with certain conditions that need to be fulfilled. While FDUS minimises the removal of potential data from the majority class, SMOTE algorithm creates synthetic data in the minority class. Overall, the proposed FDUS+SMOTE provided better results as compared to SMOTE+TL and SMOTE+ENN only for benchmark data sets in terms of classification accuracy, F-measure and G-mean. For flood data sets, classification accuracy and G-mean of FDUS+SMOTE is outperformed by no resampling. However, FDUS+SMOTE produced the best F-measure. The classification accuracy of standalone techniques namely FDUS, DUS and SMOTE is better than FDUS+SMOTE for flood data sets. FDUS+SMOTE outperformed DUS and SMOTE for benchmark data sets but not FDUS. In term of F-measure and G-mean, FDUS+SMOTE is better than FDUS, DUS and SMOTE for both flood and benchmark data sets.

CHAPTER SIX
CONCLUSION

This chapter summarises the study by highlighting the research contribution. This research mainly focuses on the enhancement of resampling technique in order to overcome the problems faced in classifying imbalanced data sets. Section 6.1 describes the contributions of the proposed techniques, and Section 6.2 suggests several future works to improve the proposed techniques.

6.1 Research Contribution

The enhancement of resampling technique is the main contribution for this research. The proposed technique is developed based on two objectives, which are the enhancement of undersampling technique, and the combination of the enhanced undersampling technique with oversampling technique.

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 study, the enhancement of undersampling technique is made on Distance-based Undersampling (DUS) technique. The DUS algorithm is modified by introducing fuzzy logic where triangular and trapezoidal membership functions are constructed. The enhanced DUS named as Fuzzy Distance-based Undersampling (FDUS) technique 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 performed better in terms of classification accuracy, F-measure and G-mean when compared with no resampling, DUS and SMOTE for benchmark data sets. FDUS produced the best classification accuracy and F-measure for flood data sets. Based on G-mean value, FDUS is better than DUS and SMOTE but performed lesser than no resampling.

The enhancement of resampling technique is made by combining FDUS with SMOTE in sequence with conditions that need to be fulfilled. The first condition is where the instances in the majority class is more than the instances in the minority class. Then, the process of resampling using FDUS+SMOTE starts with applying FDUS to the imbalanced data set. The process of undersampling is repeated until a balanced data set is produced. However, after several alterations are made by undersampling and if the amount of instances in the majority class has become lesser than the minority class, only then SMOTE is used to balance the data set. A new balanced data set is produced with the best ratio is produced.

A comparison has been made with other combination and standalone techniques. From the analysis of the results, for benchmark data sets, FDUS+SMOTE has increased the classification accuracy, F-measure and G-mean when compared to other combination techniques. For flood data sets, FDUS+SMOTE is ranked as the best technique in term of F-measure and second best technique in term of classification accuracy and G-mean. Comparison made between FDUS+SMOTE with FDUS, DUS and SMOTE showed that, for benchmark data sets, the F-measure and G-mean is better. In term of classification accuracy, FDUS+SMOTE is ranked as the second best technique outperformed by FDUS. For flood data sets, FDUS+SMOTE produced better F-measure and G-mean than FDUS, DUS and SMOTE. However, FDUS+SMOTE gave less classification accuracy than FDUS, DUS and SMOTE.

From the experiments analysis, the proposed FDUS and FDUS+SMOTE did not improved the results for flood data sets as compared to the other techniques might be due to other factors than imbalanced ratio such as size, complexity, overlap or small disjuncts as has been discussed in Chapter 4 and Chapter 5. Therefore, future work need to be conducted to overcome these problems.

6.2 Future Work

There are two possible works that can be conducted that for future research. First, the analysis of the experiments found out that the performance of the classifier is better when the ratio of the balanced data set is small. However, the optimal balanced ratio could be different for each data set. Hence, further research can be done to find out what is the best ratio of balanced data set for different types of data set. Furthermore, further experiments need to be conducted to analyse other factors than imbalance ratio that affect the performance of classification.

Second, this research only focuses on the problem of imbalanced data that is related to binary class classification. Future research can be conducted on imbalanced data sets which involve multi class classification. FDUS and the enhanced resampling technique (FDUS+SMOTE) should be tested to determine whether these techniques are suitable in handling imbalanced data for multi class classification.

REFERENCES

Alejo, R., Garcia, V., Sotoca, J. M., Mollineda, R. A., & Sanchez, J. S. (2007). Improving the performance of the RBF neural networks trained with imbalanced samples. *Computational and Ambient Intelligence, 4507*, 162–169.

Anand, A., Pugalenthi, G., Fogel, G. B., & Suganthan, P. N. (2010). An approach for classification of highly imbalanced data using weighting and undersampling. *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* (pp. 1839-1844).

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

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

Batista, G. E., Bazzan, A. L., & Monard, M. C. (2003, December). Balancing training data for automated annotation of keywords: a case study. In *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.

Batuwita, R., & Palade, V. (2013). Class imbalance learning methods for support vector machines. *Imbalanced learning: Foundations, algorithms, and applications*, 83-99

Bedient, P. B., Huber, W. C., & Vieux, B. E. (2008). Hydrology and floodplain analysis fourth edition. Prentice Hall.

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

Bennett, K. P., & Bredensteiner, E. J. (2000, June). Duality and geometry in SVM classifiers. In *ICML* (pp. 57-64).

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). Learning from imbalanced data using methods of sample selection. In *Multimedia Computing and Systems (ICMCS),* 254-257. IEEE.

Carvajal, K., Chacon, M., Mery, D., & Acuna, G. (2004) Neural network method for failure detection with skewed class distribution. *INSIGHT, Journal of the British Institute of Non-Destructive Testing*, *46*(7), 399–402.

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

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

Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004). Editorial: special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter*, *6*(1), 1-6.

Chawla, N. V., Lazarevic, A., Hall, L. O., & Bowyer, K. (2003). SMOTEBoost : Improving prediction of the minority class in boosting. *Proceedings Principles Knowledge Discovery Databases* (pp. 107–119).

Chiang, H. S., Shih, D. H., Lin, B., & Shih, M. H. (2014). An APN model for arrhythmic beat classification. *Bioinformatics*, *30*(12), 1739-1746.

Christensen, R. (1980). Entropy Minimax Sourcebook. Vol. 1–4, Entropy Ltd., Lincoln, MA.

Del Gaudio, R., Batista, G., & Branco, A. (2014). Coping with highly imbalanced datasets: A case study with definition extraction in a multilingual setting. *Natural Language Engineering*, *20*(03), 327-359.

Diamantini, C., & Potena, D. (2009). Bayes vector quantizer for class-imbalance problem. *IEEE Transactions on Knowledge and Data Engineering*, *21*(5), 638–651.

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

Dubey, R., Zhou, J., Wang, Y., Thompson, P. M., & Ye, J. (2014). Analysis of sampling techniques for imbalanced data : An n = 648 ADNI study. *NeuroImage, 87*, 220–241.

Ertekin, S., Huang, J., Bottou, L., & Giles, L. (2007). Learning on the border: active learning in imbalanced data classification. In *Proceedings of the sixteenth ACM* *conference on Conference on information and knowledge management* (pp. 127-136). ACM.

Fernandez, A., Del Jesus, M. J., & Herrera, F. (2009). Hierarchical fuzzy rule based classification systems with genetic rule selection for imbalanced data-sets. *International Journal of Approximate Reasoning*, *50*(3), 561–577.

Fernandez, A., Lopez, V., Galar, M., Del Jesus, M. J., & Herrera, F. (2013). Analysing the classification of imbalanced data-sets with multiple classes: binarization techniques and ad-hoc approaches. *Knowledge-Based Systems*, *42*, 97-110.

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.

Folorunso, S. O. & Adeyemo, A. B. (2012). Theoretical comparison of undersampling techniques against their underlying data reduction techniques. *EIECON2012*, 92-97.

Fu, X., Wang, L., Chua, K. S., & Chu, F. (2002). Training RBF neural networks on unbalanced data. *Proceedings of the 9th International Conference on Neural Information Processing*, *2*, 1016–1020.

Galar, M., Fernandez, A., Barrenechea, E., & Herrera, F. (2013). EUSBoost: enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling. *Pattern Recognition*, *46*(12), 3460-3471.

Ganesh, M. (2006). Introduction to fuzzy sets and fuzzy logic. India, ND: Prentice-Hall of India Private Limited.

Ganganwar, V. (2012). An overview of classification algorithms for imbalanced data sets. *International Journal of Emerging Technology and Advanced Engineering*, *2*(4), 42–47.

Garcia, V., Mollineda, R. A., & Sanchez, 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. (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.

Goel, G., Maguire, L., Li, Y., & McLoone, S. (2013). Evaluation of sampling methods for learning from imbalanced data. In *Intelligent Computing Theories (*pp. 92-401). Springer Berlin Heidelberg.

Gu, Q., Cai, Z., & Zhu, L. (2009). Classification of imbalanced data sets by using the hybrid re-sampling algorithm based on isomap. In *Advances in Computation and Intelligence* (pp. 287-296). Springer Berlin Heidelberg.

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

He, G., Han, H., & Wang, W. (2005). An over-sampling expert system for learning from imbalanced data sets. *Neural Networks and Brain, 2005. ICNN&B '05,* (pp. 537–541). Beijing.

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

Hu, S., Liang, Y., Ma, L., & He, Y. (2009). MSMOTE: improving classification performance when training data is imbalanced. *2009 Second International Workshop on Computer Science and Engineering* (pp. 13–17).

Hu, X., Lin, T. Y., & Han, J. (2004). A new rough sets model based on database systems. *Fundamenta Informaticae*, *59*(2), 135-152.

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.

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

Jiang, W., Deng, L., Chen, L., Wu, J., & Li, J. (2009). Risk assessment and validation of flood disaster based on fuzzy mathematics. *Progress in Natural Science*, *19*(10), 1419–1425.

Jo, T., & Japkowicz, N. (2004). A multiple resampling method for learning from imbalanced data sets. *Computational Intelligence*, *20*(1), 18–36.

Kanagavalli, V. R., & Raja, K. (2011). Detecting and resolving spatial ambiguity in text using named entity extraction and self-learning fuzzy logic techniques.

Kim, D.-S., Baek, Y.-M., & Kim, W.-Y. (2013). Reducing overfitting of AdaBoost by clustering-based pruning of hard instances. *Proceedings of the 7th*

International Conference on Ubiquitous Information Management and Communication - ICUIMC '13 (pp. 1–3).

Kim, M. (2013). Geometric mean based boosting algorithm to resolve data imbalance problem. *The Fifth International Conference on Advances Databases, Knowledge and Data Applications* (pp. 15–20).

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, 63-66. Springer Berlin Heidelberg.

Lee, C. Y., Yang, M. R., Chang, L. Y., & Lee, Z. J. (2010). A hybrid algorithm applied to classify unbalanced data. In *Networked Computing and Advanced Information Management (NCM) (*pp. 618-621). IEEE.

Lee, C., & Lee, Z. (2012). A novel algorithm applied to classify unbalanced data. *Applied Soft Computing Journal*, *12*(8), 2481–2485.

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. In *Proceedings of 2013 Chinese Intelligent Automation Conference* (pp. 547-554). Springer Berlin Heidelberg.

Lin, W. J., & Chen, J. J. (2012). Class-imbalanced classifiers for high-dimensional data. *Briefings in bioinformatics*, *14*(1), 13-26.

Liu, X., Wu, J., & Zhou, Z. (2009). Exploratory undersampling for class-imbalance learning, *39*(2), 539–550.

Liu, W., Chawla, S., Cieslak, D. A., & Chawla, N. V. (2010). A Robust Decision Tree Algorithm for Imbalanced Data Sets. In *SDM, 10*, 766-777.

Liu, Y., Yu, X., Huang, J. X., & An, A. (2011). Combining integrated sampling with SVM ensembles for learning from imbalanced datasets. *Information Processing & Management*, *47*(4), 617-631.

Lopez, V., Fernandez, A., Garcia, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences, 250*, 113–141.

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).

Mangai, U. G., Samanta, S., Das, S., & Chowdhury, P. R. (2010). A survey of decision fusion and feature fusion strategies for pattern classification. *IETE Technical Review*, *27*(4), 293–307.

Mann, P. S. (2012). *Introductory Statistics* (8th ed.).Wiley Global Education.

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

Mirza, B., Lin, Z., & Toh, K. A. (2013). Weighted online sequential extreme learning machine for class imbalance learning. *Neural Processing Letters*, 1-22.

Naganjaneyulu, S., & Kuppa, M. R. (2012). A novel framework for class imbalance learning using intelligent under-sampling. *Progress in Artificial Intelligence*, *2*(1), 73–84.

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

Nguyen, G. H., Bouzerdoum, A., & Phung, S. L. (2009). Learning pattern classification tasks with imbalanced data sets. In *P. Yin (Eds), Pattern Recognition* (pp. 193-208). Vukovar, Croatia: In-Teh.

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

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* (pp. 1441–1449). Singapore: Springer Singapore.

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.

Phung, S. L., Bouzerdoum, A., & Nguyen, G. H. (2009). Learning pattern classification tasks with imbalanced data sets. *Pattern Recognition*, 93–208.

Rokach, L. (2009). Ensemble-based classifiers. *Artificial Intelligence Review*, *33*(1-2), 1–39.

Ross, T. J. (2010). Development of membership functions. *Fuzzy Logic with Engineering Applications, Third Edition*, 174-210.

Sang, G., Gao, L., & Liu, Z. (2013). A bias-ensemble learning algorithm for imbalanced data processing imbalanced data-sets classification methods. *Journal of Computational Information Systems*, *9*(5), 2025–2032.

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.

Shen, Q., & Jiang, Y. (2010). Fuzzy sets, rough sets and vague sets. *3rd International Conference onAdvanced Computer Theory and Engineering* (pp. 461–465).

Shivalkar, P. S., & Tripathy, B. K. (2015). Rough Set Based Green Cloud Computing in Emerging Markets.

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. Berlin, Heidelberg: Springer Berlin Heidelberg.

Soler, V. & Prim, M. (2009). Extracting a fuzzy system by using genetic algorithms for imbalanced data sets classification: application on down syndrome detection. In D. A. Zighed, S. Tsumoto, Z. W. Ras, & H. Hacid. (Eds.), *Mining Complex Data* (pp. 23-39). Springer Berlin Heidelberg.

Sun, Y., Kamel, M. S., Wong, A. K. C., & Wang, Y. (2007). Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition*, *40*(12), 3358–3378.

Sun, Y., Robinson, M., Adams, R., Boekhorst, R., Rust, A. G., & Davey, N. (2006). Using sampling methods to improve binding site predictions. *Procs of the 14th European Symposium on Artificial Neural Networks, ESANN 2006* (pp. 533–538).

Sun, Y., Wong, A. K. C., & Kamel, M. S. (2009). Classification of imbalanced data : a review. *International Journal of Pattern Recognition and Artificial Intelligence*, *23*(4), 687–719.

Tan, P.-N., Steinbach, M., & Kumar, V. (2006). Ensemble methods. *Introduction to data mining*, 276–293. United States of America: Pearson Education.

Tang, Y., Zhang, Y. Q., Chawla, N. V., & Krasser, S. (2009). SVMs modeling for highly imbalanced classification. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, *39*(1), 281-288.

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

Verbiest, N., Ramentol, E., Cornelis, C., & Herrera, F. (2012). Improving SMOTE with fuzzy rough prototype selection to detect noise in imbalanced, 169–178.

Visa, S., & Ralescu, A. (2003). Learning imbalanced and overlapping classes using fuzzy sets. *Workshop on Learning from Imbalanced Datasets II (ICML '03)* (pp. 91–104).

Visa, S., & Ralescu, A. (2005, April). Issues in mining imbalanced data sets-a review paper. In *Proceedings of The Sixteen Midwest Artificial Intelligence And Cognitive Science Conference* (pp. 67-73).

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.

Wang, S., & Yao, X. (2009, March). Diversity analysis on imbalanced data sets by using ensemble models. In *Computational Intelligence and Data Mining, 2009. CIDM'09* (pp. 324-331). IEEE.

Wang, X. J., Zhao, R. H., & Hao, Y. W. (2011). Flood control operations based on the theory of variable fuzzy sets. *Water Resources Management*, *25*(3), 777–792.

Wang, S., & Yao, X. (2013). Relationships between diversity of classification ensembles and single-class performance measures. *IEEE Transactions on Knowledge and Data Engineering*, *25*(1), 206–219.

Weiss, G. M., & Provost, F. (2003). Learning when training data are costly: the effect of class distribution on tree induction. *Journal of Artificial Intelligence Research*, 315-354.

Whitley, E., & Ball, J. (2001). Statistics review 1: presenting and summarising data. *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.

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

Yang, H., Fong, S., Wong, R., & Sun, G. (2013). Optimizing classification decision trees by using weighted naive bayes predictors to reduce the imbalanced class problem in wireless sensor network. *International Journal of Distributed Sensor Networks, 2013,* 1–16.

Yang, Z., & Gao, D. (2013). Classification for imbalanced and overlapping classes using outlier detection and sampling techniques. *Applied Mathematics & Information Sciences, 7*(1L), 375–381.

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.

Zhang, I., & Mani, I. (2003). kNN approach to unbalanced data distributions: a case study involving information extraction. In *Proceedings of Workshop on Learning from Imbalanced Datasets*.

Zhang, Y., & Wang, D. (2013). A cost-sensitive ensemble method for class-imbalanced data sets. *Abstract and Applied Analysis, 2013,* 1–6.

Zhong, W., Raahemi, B., & Liu, J. (2009). Learning on class imbalanced data to classify peer-to-peer applications in IP traffic using resampling techniques. In *Neural Networks, 2009. IJCNN 2009,* (pp. 3548-3554). IEEE.