Journal of Theoretical and Applied Information Technology
15th April 2017. Vol.95. No 7

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

TECHNIQUES FOR HANDLING IMBALANCED DATASETS WHEN PRODUCING CLASSIFIER MODELS

[1]ROZIANIWATI YUSOF , [2]KHAIRUL AZHAR KASMIRAN, [3]AIDA MUSTAPHA, [4]NORWATI MUSTAPHA, [5]NOR ASMA MOHD ZIN

[1,5]Faculty of Computer and Mathematical Sciences,UiTM Shah Alam, Selangor, Malaysia
[2,4]Faculty of Computer Science and Information Technology, UPM Serdang, Selangor, Malaysia
[3]Faculty of Computer Science and Information Technology, UTHM Batu Pahat, Johor, Malaysia
E-mail: [1]rozian696@ns.uitm.edu.my, [2]k_azhar@upm.edu.my, [3]aidam@uthm.edu.my, [4]norwati@upm.edu.my, [5]norasmamz@kelantan.uitm.edu.my

ABSTRACT

Imbalanced datasets are a well-known problem in data mining, where the datasets are composed of two classes; the majority class and minority class. A majority class has more instances compared to the minority class. Recent years have brought increased interest in handling imbalanced datasets since many datasets produced are naturally imbalanced. Most existing techniques for classifying data ignore the imbalanced condition, but focused on the accuracy of the model produced where it is biased to the majority class while giving poor accuracy towards the minority class. Although the minority class is something that rarely happens, but in some conditions it will give an important influence to the classifier model. This paper attempts to list all the techniques in handling imbalanced datasets, as well as to compare all the techniques for producing the best classifier model for imbalanced datasets. These techniques have been categorized into sampling, feature selection and algorithmic approaches in the form of a taxonomy for handling imbalanced datasets. The strengths and the weaknesses of these approaches will be discussed in order to identify an appropriate technique that will improve the performance of a classifier model produced. The recent trends in handling imbalanced datasets also will be discussed based on domain and problems exist in dataset.

Keywords: *Imbalanced Data, Sampling, Feature Selection, Cost Sensitive Learning, Classification*

1. INTRODUCTION

The tremendous growth of technology has produced many raw datasets without giving any knowledge. It is thus an opportunity for researchers to produce great knowledge in order to use a dataset without wasting the production of data. However, most datasets are naturally imbalanced and can cause biased knowledge. Imbalanced datasets are normally composed of two classes; the majority (negative) class and minority (positive) class. A majority class has more class instances compared to the minor class [1]. In recent years, the problem of imbalanced datasets has seen an increase in occurrence and has become a major concern due to the performance of the algorithms and models produced that are significantly degraded because of it.

The imbalanced dataset will give a big impact in producing knowledge in real applications such as in the medical field, biological data, text classification, web categorization, risk management and fraud detection [2]. For example, in the medical field the image dataset generally produce imbalanced datasets since normal cases are usually higher compared to abnormal cases. This condition will cause misclassification and will lead to poor mining results. According to [1], most past researches related to medical diagnosis often involve imbalanced datasets as the training set typically has more benign cases compared to malignant cases, and normal cases are more than abnormal cases. For example, a mammography dataset can contain 98% normal pixels and 2% abnormal pixels [3]. Therefore, there is a high potential to predict an abnormal case as a normal case. If there is a misclassification of non-cancerous cells, this may lead to further clinical testing. However, misclassification of cancerous cells will cause to very serious health risks.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

There are many major factors that influence the performance of a classifier model for imbalanced datasets such as the size of datasets, imbalance ratio, characteristics of imbalanced data like small disjunctions, ambiguous boundary between classes, classes overlapping in feature spaces, noisy data and dimensionality data [4]. Normally, the small sample size will influence the process of classifying imbalanced data. The minority data samples are not sufficient to train properly and will cause poor overfitting and poor generalization. Meanwhile, the problem of small disjunction also relates to the minority class. This happens due to the lack of sample of each sub-concept in the minority class [5]. Ambiguous boundary is also caused by the lack of minority data in the boundary area. Therefore, the decision that the boundary produced can be far less than the true boundary and results in poor minority prediction [4]. Besides that, a class overlapping will lead to poor performance for an object in the overlapped area. Class overlapping happens when discriminative rules are constructed to reduce misclassification of data. Producing a classifier model without considering imbalanced datasets is biased towards the major class dataset as well as giving poor accuracy towards the minor class dataset. The minor class instance can be misclassified as a major class instance. Although the minority class is something that rarely happens, but in some conditions, it will give an important influence to the classifier model. Therefore, it is very important to know the condition or the problem occurred on data before producing a classifier model.

In handling imbalanced datasets, there are three major approaches that have been introduced such as sampling, feature selection and algorithmic approaches [6]. The algorithmic approaches consist of cost-sensitive learning and one class learning approach. The combination of these approaches is called the ensemble method. However, all the approaches introduced should focus on the identification of the minor classes since the imbalanced data will cause bias to it. Therefore, the research question is how to choose the best techniques in handling imbalanced datasets in producing classifier model. The strengths and the weaknesses of these approaches are very important to clarify in order to identify an appropriate technique that will improve the performance of a classifier model produced. The condition or problems exist in datasets also will give some influences in producing a good classifier model. So, from this survey paper, the researchers may define the best techniques to produce classifier model based on domain and the condition of dataset. The techniques show in form of taxonomy will make them more easiest to choose the suitable technique based on the strength and weakness of the techniques.

This paper is organized as follows: handling imbalanced techniques will be discussed in Section 2. The strengths and weaknesses of the approaches introduced will be discussed in Section 3. Section 4 will highlight on the trends of techniques in handling imbalanced datasets based on the domain and problems to solve. Finally, the conclusion and discussion are presented in Section 5.

2. HANDLING IMBALANCED TECHNIQUES

2.1. Sampling Techniques

Sampling approach is a technique used without needing to change an algorithm. This is a common practice used in handling imbalanced datasets. This technique modifies data distribution or resize the dataset in order to balance the dataset. In the data sampling approach, different techniques have been tested such as under-sampling and over-sampling, as well as random sampling as shown in Figure1 below.

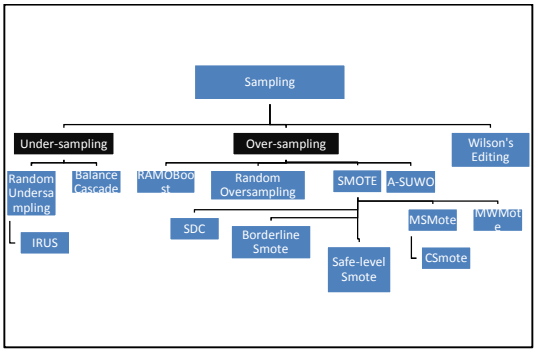

Figure 1: Sampling Techniques

The under-sampling technique is for the majority class where data are randomly removed from the majority class sample. Meanwhile, over-sampling randomly replicates data from the minority class. Both techniques have disadvantages in which under-sampling can cause information loss whereas over-sampling will disturb the data distribution within the class either by overfitting or generating synthetic data points; which do not follow the real class distribution [7]. Recently, from the basic sampling techniques, there are many techniques that have been extended such as random

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

sampling. Random sampling is a combination of under-sampling and oversampling where the data are randomly selected to replicate or to be removed [8]. Random under-sampling samples the data by randomly removing examples from the majority class. Meanwhile, random oversampling is the random replication of data from the minority class. Resampling the data whether to remove or replicate it is based on specific data points such as class boundary and noisy data [7]. However, random sampling will change data drastically from the original data and sometimes will leave outliers in the data. This technique can also cause the loss of information and overfitting.

The IRUS or inverse random under-sampling is produced from random under-sampling. The idea of IRUS technique is to change the ratio of cardinality between the majority and minority classes by under-sampling the majority class repeatedly. Meanwhile, the BalanceCascade technique is focused on a training model to sample the data purposely for classifying a training pattern which is difficult to classify. However, this technique will omit potentially useful data from the majority class [9]. Based on [10], Chawla created a new technique for sampling in c2002 called SMOTE. This technique will add new data of minority class by interpolating between existing minority instances rather than duplicating the original instances. The new instances cause the minority regions of the feature space to be fuller [11]. SMOTE is also employed by the oversampling method for generating synthetic points in the majority class feature space. In recent years, many researchers found that there are many extension for SMOTE technique such as SDC that combines the variant of SMOTE. However, the SMOTE technique is useful only for low dimension datasets [12]. Based on observation, the SDC is much better than SMOTE [13]. Meanwhile, the Modified Smote or MSMOTE is modified from SMOTE. The MSMOTE will divide the data into three parts; safe, border and noise. The MSMOTE is better than SMOTE because it will remove latent noisy instances to create new synthetic instances [9]. The RAMOBoost technique purposely to classify binary classes dataset. [9]. The minority class will be oversample based on adaptive weight adjustment method. Besides that, Safe-level SMOTE will calculate a value for each minority instance called safe level value. Safe level value is defined as the number of other minority among its NN-nearest neighbours [14]. A new technique based on over-sampling has been introduced by [14] known as A-SUWO. This method will clusters the minority objects using a semi-unsupervised hierarchical clustering approach and get the size of each subcluster to oversample using complexity of classification and cross validation. This technique avoid from overlapping between synthetic minority instances with majority instances.

Meanwhile, Wilson's Editing technique reduced a dataset with instance-based learning technique [10]. It will reduce all misclassified data in majority classes. This technique does not require specifying the class distribution. However, this technique will not change the correlation between attributes compared to random oversampling and random under-sampling as SMOTE changes the correlation structure in datasets [10].

2.2 Feature Selection

Feature selection is an important phase in data mining, especially for high-dimensional datasets. Feature selection is a reduction of features [15] which attempts to select an informative feature [16] based on some criteria [17]. These features will represent the whole data without changing the original meaning. Figure 2 below shows the categories of feature selection.

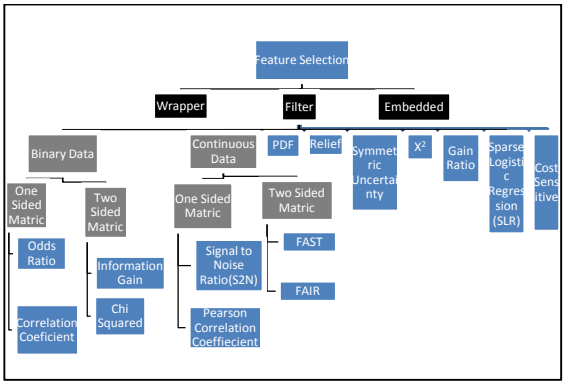

Figure 2: Feature Selection Techniques

The feature selection technique can be categorized into three approaches; wrapper, embedded and filter [18]. The wrapper method selects the features based on the classification of algorithm and must retrain the algorithm with the feature subsets. Meanwhile, the embedded approach is like the wrapper method that relates to the learning algorithm, but the link is stronger than the wrapper method. This method works in conjunction with the learning algorithm to find feature subsets. The filter method used independent criteria to evaluate the subset of features. The

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

independent criteria include distance measures, information measures, dependency measures and soft computing-based measures [19]. Based on [18], the filter based approach is more efficient than the wrapper as this technique does not depend on induction algorithm and requires less computation [20]. However, the wrapper approach will produce a good result, but it depends on the classification algorithms. Sometimes, the features selected are not suitable with other classification algorithms. The wrapper approach normally has high computational cost and risk of overfitting to model [21]. Based on [11], neither the wrapper nor embedded methods have been proposed to handle imbalanced datasets in high-dimensional applications. However, past researchers only applied some heuristic techniques to improve the classifier of imbalanced datasets without selecting relevant features. They applied the embedded method to handle imbalance by embedding backward feature selection with the Support Vector Machine [11].

The feature selection method is necessary in handling imbalanced datasets because imbalance problem commonly comes with high dimensionality of data [22]. Based on [22], this technique is good in handling overfitting and can handle the imbalance problem alone without another technique if it involves high dimensionality data. In selecting the features, highly correlated features s[hould be considered in imbalanced datasets compared to the independent features [23]. There are many ways in handling imbalance by feature selection such as selecting features based on the feature selection metric. [22] proposed a new framework to select features such as selecting positive and negative classes separately. Meanwhile [24] proposed an iterative feature selection and [25] choosing features based on weight for negative and positive features.

For the feature selection metric, it can be measured by one-sided metric or two-sided metric. The one-sided metric measures positive features which indicate the membership in a class. Meanwhile, the two-sided metric selects both positive and negative features scores based on the metric. The negative feature is featured with the lack of membership in a class. Some researchers select the features by filtering strategies to examine the relevance of each feature for classes using the metric. The common filtering strategies are the X^2 statistic, Information gain, Gain ratio, Relief and ReliefF, Symmetric Uncertainty, and chi-square [10], [26]. The X^2 statistic tests the distribution of class. Information gain, gain ratio, and symmetrical uncertainty are measured based on the concept of entropy, which is based on information theory. Information gain metric checks the importance of the attributes of each class. It is about a target class that measures the decrease of the weighted average impurity of the sub attributes compared to the complete attributes. However, it will give the best result with the complete attributes. To avoid from this condition, the gain ratio strategy is applied. To improve the results, symmetrical uncertainty is used to avoid bias toward attributes with more values [26]. Meanwhile, the chi-square tests the association between two attributes, whether it is dependent on each other or not [26]. The Relief and ReliefF strategy can handle noise and multiclass datasets. Based on [10], the X^2 statistic, Information gain, Symmetric Uncertainty and Gain Ratio are highly correlated to each others. The Relief and ReliefF strategy has an average correlation. Correlation is to define which techniques produce the same results to similar data in selecting the best features to represent the data without affecting the condition of data.

[23] introduced a new feature of the selection metric; Feature Assessment by Sliding Threshold (FAST) to handle small samples of imbalanced datasets. Based on [6], this technique is better compared to relief, correlation coefficient (CC) and baseline in selecting features. This technique is a two-sided metric. The FAST is measured by the area under the receiver operating characteristic (AUC) and handled using an even-bin distribution to move the decision boundary of a single feature classifier. However, this is not the best way to obtain a decision boundary. This problem can be handled by gathering the statistics for each boundary from the sample of multiple thresholds [6]. The AUC is good for predictor performance for imbalanced datasets where this score can be used for the feature ranking of attributes in which features with the highest scores are the best features for prediction [6].

In general, the feature selection metric for an imbalanced dataset is considered better when the measurement is separated from the minority class and majority class [23]. However, some of the metrics can handle Boolean data only as they lack continuous data. Based on [22], the metrics for binary data are the chi-square, information gain and odd ratio. Meanwhile, for continuous datasets the Pearson Correlation Coefficient (PCC), Feature Assessment by Sliding Thresholds (FAST), Feature

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Assessment by Information Retrieval (FAIR) and signal to noise correlation coefficient (S2N) are used.

Based on [7] among all features metric, the sparse logistic regression (SLR) has shown higher stability in the selection of various features. Regression is a technique in modelling prediction which investigates dependent and independent variable. There are many techniques from regression such as logistic regression, linear regression, stepwise regression, ridge regression and lasso regression. The SLR is an improvement of logistic regression (LR). This technique can handle very high dimensional data and limited sample size. The features will be selected based on weight in which irrelevant features are given a zero weight. Meanwhile, other features are ranked in a decreasing order of their weight [7]. For example, the WSMOTE algorithm [27]. In this algorithm, the important attribute will be assigned with more weight using the mean and standard deviation of each individual attribute.

Based on [2], they handled imbalance by using unsupervised feature selection based on the filtering technique. This technique is considered for the minority class according to the relation between the distributions of features based on the probability density function (PDF). The features with higher covering areas with the PDF are considered as redundant features and will be removed. The feature selection techniques can also handle imbalanced datasets by using a different framework. Based on [21], the proposed iterative feature selection with different sample data can be employed. This approach identifies a ranked feature list which is effective on the dataset. This technique has been proven better than single iteration while selecting the features. Besides that, a stochastic algorithm Optimal Feature Weighting (OFW) and one- vs-one SVM are also used to find the optimized features from imbalanced and high-dimensional feature space [28]. According to [29], they select a subset of features based on cost sensitive as a feature weight. This technique known as Cost Sensitive Feature Selection using Chaos Algorithm (CSFSG).

2.3 Algorithmic Approaches

An algorithmic approach is a technique in handling imbalance involving changing or creating an algorithm [30] to modify the learning cost, adjusting the probability, threshold and recognition based on one class learning [24]. This approach optimizes the performance of learning algorithm through unseen data. Figure 3 shows show that types of algorithmic approaches consists of Cost Sensitive Learning and One Class Classification.

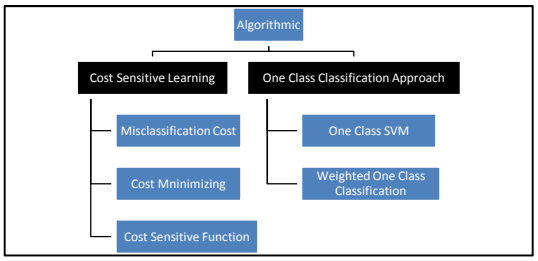

Figure 3: Algorithmic Approaches

2.3.1 Cost sensitive learning

Cost sensitive learning is a technique to define a different cost to misclassification errors such as false negative and false positive patterns [9]. It is also called as reweighting or adjusting the cost of various imbalanced classes [31]. In this technique, the misclassification cost is the important cost to be evaluated and should be minimized. Based on [32], the best metric for evaluating the classifier performance is the total cost. The formula for the total cost is shown in Equation 1:

$$\text{Total cost} = (FN \times CFN) + (FP \times CFP) \quad (1)$$

Many researchers used the algorithmic approach in cost sensitive learning by using different classification algorithms such as Decision Trees and Support Vector Machines [7]. The purpose of algorithmic approach is to optimize the learning performance. There are three categories in implementing cost sensitive learning such as the use of misclassification cost to form dataspace weighting, cost minimizing techniques to combine schemes of ensemble method and combining the cost sensitive function with classifiers [8]. Misclassification cost can be divided to example dependent costs and class dependent costs [29]. Other algorithmic approaches include cost sensitive methods [30], [33], and kernel-based approach such as SVM [30].

2.3.2 One class classification approach

This approach is recognition-based where classes are learned separately [31]. In this approach, training is done only on a sample of one class known as the target class. The purpose of the approach is to create a decision surface that covers all the available data samples. However, all the represented data outside the target concept are

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

labelled as outliers. In handling the imbalance problem, this approach normally focuses on identification of the minority class samples as the target class [34]. The one-class classification is a technique that handles imbalanced dataset through the learning algorithm. This technique has one class target as the normal class whereas outliers is the abnormal class [35]. The distance or similarity between data with the target class will be measured. The threshold value will be identified to evaluate the data.

One of the common one class learning approaches is the one class SVM. This technique attempts to match an object with the target class by measuring the similarity [11]. However this technique only gives good performance when training using the minority sample [23]. Based on [34], they proposed the one class classifier based on weight. This technique will consider the details of minority class samples in the training set. This is because the normal learning algorithm produces a decision boundary that is biased toward the majority class and leads to a high rate of error in the minority class [34]. The calculation of weight for the minority sample is based on the borderline sample, rare sample and outlier's sample. However, this technique is only suitable for the binary classes.

2.4 Ensemble Methods

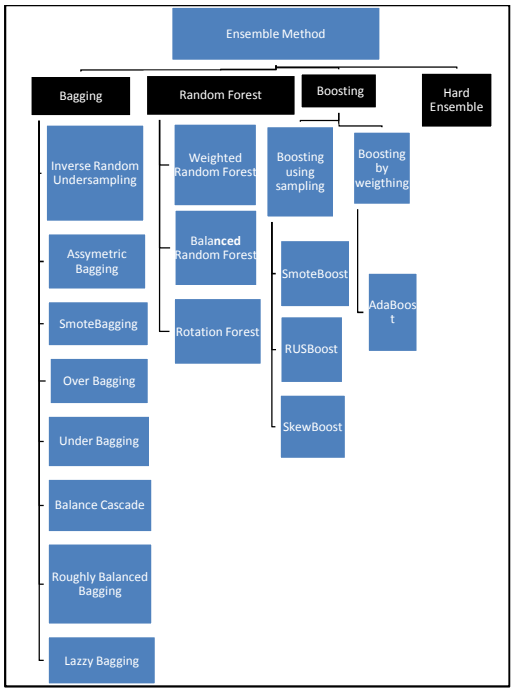

Figure 4: Ensemble Techniques

An ensemble method is a technique to combine or aggregate more than one technique such as two learning algorithms, a classifier with sampling techniques or feature selection and others as in Figure 4 above. Ensembling classifiers is a process of aggregating classifications of diverse classifiers [36]. The most prominent ensemble methods are bagging and boosting. These techniques employ sampling techniques to obtain different training sets for classifiers. Bagging is a process of individual classifier training using different bootstraps of data or by assigning weights to the observations. The bagging technique will reduce the variance of mean squared error (MSE). There are several techniques produced from bagging such as the Inverse Random Undersampling (IRUS), Asymmetric Bagging (AB), SmoteBagging (SB), Over Bagging (OB), Under Bagging (UB), Balance Cascade (BC), Roughly Balanced Bagging (RBB) and Lazy Bagging (LB) [9], [23], [36]–[38]. The IRUS technique will invert the ratio between the majority class and minority class by under-sampling the majority class and oversampling the minority class multiple times. This technique will control the false positive rate. Meanwhile, the AB maintains the size of the minority class, but a partition of equal size is derived from majority class for each bootstrap iteration. The SB generates synthetic observation to the minority class and does bagging to the majority class. The OB technique applies the random oversampling technique to the minority class during each bootstrap iteration. As for the UB, it applies random undersampling. The BC is based on the under-sampling strategy where it investigate the instances of majority class ignored by the under-sampling process. It uses both bagging and AdaBoost. The BC will produce a higher F-Measure, AUC and G-Mean. However, for the RBB, it uses the concept of weightage to balance the data between each class for each bootstrap iteration. The LB technique applies bagging to the nearest points using the nearest neighbouring algorithm. Meanwhile, the RF uses decision trees based on bagging for training. An easy Ensemble technique is produced from the under-sampling technique. This technique will sample the majority class into subsets to train separate classifiers. The results will combine in order to get a decision [9].

Meanwhile, in the boosting technique, classifiers in the ensemble method are trained serially [37]. The first classifier in the ensemble method is trained using bootstrap data and it tests the model produced for the whole set of data. It will determine which sample is correctly predicted and

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

which is not. For incorrect prediction, it will be drawn in the next sample. It means that correctly predicted sample decreases and incorrect predicted sample increases [23]. There are a few techniques that apply the boosting technique such as AdaBoost, SmoteBoost, Random Under Sampling Boost (RUSBoost) and skewBoost. The boosting technique can be divided into two groups based on sampling and by weighting. However, according to [27], the boosting based sampling is better than by weighting because it can be applied on any base learner. The AdaBoost is an iterative boosting by allocating the variants weight to the observations during training time. The weight for misclassified data will increase for each iteration while the weight for correctly classified will decrease [38]. This technique will focus on misclassified data and very efficient ensemble method of classification [12]. Meanwhile, the SmoteBoost is an integration between Smote and AdaBoost [8]. It employs the synthetic sampling for each time boosting. The RUSBoost is a combination between sampling and Boosting. The RUSBoost applies the concept of random under sampling by removing data from the majority class. This technique is similar with AdaBoost where it applies the weightage concept for each time training iteration. This technique is less computational complex and time consuming [9]. The SkewBoost technique is based on SMOTE sampling where it is considered only for majority instances without compromising the accuracy of classifier for the majority instances [27].

Recently, many techniques have been extended from the Random Forest such as Weighted Random Forest, Balanced Random Forest [38] and Rotation Forest [12]. The Rotation Forest will rotate all the subspaces from the original dataset. It will increase the diversity and accuracy of the classifier. However, the ensemble method between classifier methods will cause dependency on the classifier algorithm [7]. Based on [9], they introduced a novelty technique called the HardEnsemble. This technique is a combination between under-sampling and over-sampling techniques. For the over-sampling technique, they applied the CSMOTE technique and for under-sampling, they applied the Reduced Reward Punishment technique.

The ensemble feature selection is a technique where multiple feature selection techniques are combined together. Based on [39], the ensemble methods will give a promising result compared to the single feature selection especially to high dimensional data and small size data. Based on [21], they handled imbalanced datasets by combining feature selection technique with data sampling. The features have been selected for feature rankings such as chi-square, information gain, gain ratio, relief and symmetrical uncertainty. The process of sampling is repeated many times while selecting the features. They produced the best result for highly imbalanced datasets.

3 THE STRENGTH AND THE WEAKNESS OF THE APPROACHES

All the approaches discussed above have their strengths and weaknesses. However, it depends on the condition of data. There are many factors that will influence the classifier model produced from imbalanced datasets such as the ratio of skewness in the dataset and the characteristics of imbalanced datasets such as small disjunctions, small data size, class overlapping, dimensionality of data and noisy data [4]. However, most cases of imbalanced datasets are due to nonsufficient number of minority class instances that will cause improper training. The result produced thus tends to overfitting. Besides that, class overlapping can lead to poor performance over the minority class especially for the objects or data in overlapped areas. Meanwhile, small disjuncts problem happen to minority class where low number of samples for each subconcepts and lack of diversity for minority instances. Despite of that, the consideration of which technique is best for handling a problem of data distribution highly depends on the nature of datasets used for experiment. For the sampling technique, it is suitable for highly imbalanced ratio among datasets [30] and also for large datasets [24]. Based on [14], A-SUWO technique works better compare to other sampling techniques for datasets with higher imbalance ratio. However, the combination of sampling techniques with algorithms can improve the performance compared to using one sampling technique only. This method will improve the accuracy of the minority class, but not for the majority class. Sometimes, this method is biased from the original dataset. This is because it will change the correlation structure extremely. Thus the subset of features produced is not accurate for different samples [40].

To avoid changing the correlation structure, the feature selection method is applied. Feature selection techniques are mostly applied to high dimensional datasets because the imbalance problem is commonly occurred with the issue of high dimensionality of dataset, hence applying the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

feature selection techniques is essential. This technique can reduce susceptibility, overfitting, and reducing storage, memory and processing requirements. Besides that, it will enhance the speed of the training process. However, some of the attributes will be omitted as they will influence the classification of performance. Based on [6], the feature selection technique is better in handling overfitting compared to handling using the classification algorithm. According to [12], feature selection is very important when handling high dimensional imbalanced dataset and can handle overfitting problem caused by oversampling the minority data. Based on past researches, feature selection based on metric measurement for filtering features is the best approach in handling imbalanced dataset. Based on [41], this approach is robust against overfitting compared to the wrapper methods that cannot produce the best feature subset and cause poor classifier performance especially for high dimensional and imbalanced datasets. However, based on [42], the feature selection metrics can be chosen based on the condition of data whether binary data or continuous data. It does not depend on classifiers.

Besides that, the skewness of dataset should also be considered. When the dataset involve continuous data, highly imbalanced and have two classes, then it is better to use the FAST metric in the ratio of 1:5 of minority between majority data [43]. Meanwhile, information gain performs better for a ratio of 1:3 and odd ratio performs well for a ratio of 1:2 or less. The feature selection metrics also perform better through continuous data when using non parametric measurements such as precision recall and ROC. This is because this measurement is used on all possible confusion metrics and can give a threshold to result with the highest performance [41]. Besides that, the number of dataset attributes should also be considered. FAST technique perform better when the data have attributes between 10 and 50. Meanwhile, the S2N can be more effective than FAST when the dataset have more than 100 attributes [42]. However, CSFSG technique can perform better when applied to minority class. According to [22], the feature selection metric is not good enough to handle imbalance because this technique measures each feature independently. The interactions between different features need to be considered in selecting features. Thus it is suitable to involve interaction between features, wrapper and embedded method, but it is costly with high time complexity.

For the cost-sensitive learners, it will handle bias for the minority class by shifting the minority class. This technique is good when the dataset is extremely imbalanced [4]. Meanwhile, the OCC is an interesting solution for real imbalanced datasets especially for the minority class. This technique perform training process on a single class only and can be done through a single model or multiple based classifier. A single model OCC classifier has a difficulty to find a good model due to the small training dataset, high dimensionality of the feature space and the properties of classifier which might produced an overfitting model [1]. However, multiple based classifier on OCC can reduce the risk of producing overfitting model. Furthermore, the OCC can outperform multiclass algorithms and as single-class classifiers are robust in handling many condition embedded in the real data.

The ensemble method is a good technique in handling whatever dataset condition. It is because the ensemble method is a technique that combines many kinds of techniques in handling imbalanced datasets such as the combination of classification methods, sampling with feature selection, sampling with learning algorithms and others. The ensemble technique with algorithmic approaches is good for skewed distributions of data by modifying the classifier [30]. Besides that, the ensemble method can also handle multiclass imbalanced data through Adaboost. However, the Adaboost technique is susceptible to outliers and noisy datasets. Besides that, if there are no constraints on computation time, iteration for many times before ensemble the techniques is better because iteration will not deteriorate the performance and it will be useful to know a priori [9]. It is also known that when ensembles the diversity of classifiers technique, weak learners provide an improvement significantly higher than strong learners due to diversity among weak learners is higher than the strong learners [9]. Besides that, based on [12], the main advantages when combining various classifiers that, it can improve accuracy and reduce the error rate of classifiers compared to using the single classifier technique. Table 1 in the appendix shows the summary of the strength and the weakness of the techniques in handling imbalanced datasets.

4 RECENT TRENDS IN HANDLING IMBALANCED DATASETS BASED ON DOMAIN AND PROBLEMS

The Table 2 in the appendix shows how the previous researchers handle imbalances datasets in different domains. They use different algorithms in

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

handling imbalance that stem from different problems.

Based on the table below, most past research focused on the sampling technique since it is relevant in handling imbalanced datasets without concerning the condition of data. However, this technique will change the correlation of datasets for each sampling technique. Thus the results might not be valid. Recent papers proposed for the ensemble method to handle the limitation of sampling techniques by combining sampling techniques with others. Most researchers combine sampling and feature selection techniques when involving high dimensional datasets like images and microarray datasets. This is because selecting the features of high dimensional imbalanced datasets is crucial when without adding or losing data in balancing the data [44]. However, when it involves noisy datasets and highly imbalanced ratios, most researchers apply the sampling technique to handle these problems.

5 DISCUSSION AND CONCLUSION

The techniques of handling imbalanced datasets explored in this paper are useful for particular conditions of data such as small disjunctions, small data size, class overlapping, high dimensionality and noisy data. All of the techniques investigated can be classified into sampling, feature selection and algorithmic. The algorithmic technique consists of cost sensitive learning and one class classification. Each of the techniques has their own advantages and disadvantages. All the strengths and weaknesses of the techniques have also been explored by researchers and handled by extending the current techniques.

Practically, it has been reported that the sampling technique is mostly used by researchers in handling imbalanced datasets without concerning the condition of data. With no sampling process, a classifier model produced sometimes gives a high accuracy but has a huge gap between sensitivity and specificity. However, under-sampling technique is better than over-sampling because over-sampling can disturb the original class distribution. Besides that, based on this survey, ensembling the feature selection technique with sampling is the best way to handle imbalanced datasets for high dimensional datasets. It is because ensemble the techniques can complement to each others. Feature selection metric technique is the best way in handling imbalanced by using feature selection especially if measurement of feature metric is separated for minority class and majority class. Since the condition of dataset can vary, applying more than one technique with iteration can give better solutions for handling imbalanced datasets. The solutions produced can be selected by using rank aggregation to make sure the best classifier model will be generated. Besides that, the best way in handling imbalanced dataset is not to change the condition of data but find the best feature set that will clear the decision boundary between maximum class and minimum class.

In the future, further discussion on techniques in handling imbalanced datasets on big data will be more challenging since big data having many problems or conditions occurred on datasets.

REFERENCES

[1] B. Krawczyk, G. Schaefer, and M. Wozniak, "Combining One-Class Classifiers For Imbalanced Classification Of Breast Thermogram Features," *Fourth International Workshop on Computational Intelligence in Medical Imaging (CIMI)*, 2013, pp. 36–41.

[2] M. Alibeigi and A. Hamzeh, "Unsupervised Feature Selection Based On The Distribution of Features Attributed to Imbalanced Data Sets," *Int. J. Artif. Intell. Expert Syst.*, vol. 2, no. 1, pp. 14–22, 2011.

[3] A. K. Mohanty, M. R. Senapati, and S. K. Lenka, "A Novel Image Mining Technique For Classification Of Mammograms Using Hybrid Feature Selection," *Neural Comput. Appl.*, vol. 22, no. 6, Feb. 2016, pp. 1151–1161..

[4] W.-J. Lin and J. J. Chen, "Class-Imbalanced Classifiers For High-Dimensional Data.," *Brief. Bioinform.*, vol. 14, no. 1, Jan. 2013, pp. 13–26.

[5] B. Krawczyk, M. Woźniak, and G. Schaefer, "Cost-Sensitive Decision Tree Ensembles For Effective Imbalanced Classification," *Appl. Soft Comput.*, vol. 14, Jan. 2014, pp. 554–562.

[6] H. Pant and R. Srivastava, "A Survey on Feature Selection Methods For Imbalanced Datasets," *Int. J. Comput. Eng. Appl.*, vol. 9, no. 2, 2015, pp. 197–204.

[7] R. Dubey, J. Zhou, Y. Wang, P. M. Thompson, and J. Ye, "Analysis Of Sampling Techniques For Imbalanced Data:

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

An N=648 ADNI Study.," *Neuroimage*, vol. 87, Oct. 2013, pp. 220–241.

[8] E. a. He, H., Garcia, "Learning from Imbalanced Data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, Sept. 2009, pp. 1263–1284.

[9] L. Nanni, C. Fantozzi, and N. Lazzarini, "Coupling Different Methods For Overcoming The Class Imbalance Problem," *Neurocomputing*, vol. 158, Jun. 2015, pp. 48–61.

[10] J. Van Hulse, T. M. Khoshgoftaar, A. Napolitano, and R. Wald, "Feature Selection with High-Dimensional Imbalanced Data," in *2009 IEEE International Conference on Data Mining Workshops*, 2009, pp. 507–514.

[11] S. Maldonado, R. Weber, and F. Famili, "Feature Selection For High-Dimensional Class-Imbalanced Data Sets Using Support Vector Machines," *Inf. Sci. (Ny).*, vol. 286, Dec. 2014, pp. 228–246.

[12] Z. A. O. Seyyedali Fattah, Zalinda Othman, "New Approach For Imbalanced Biological Dataset Classification," *J. Theor. Appl. Inf. Technol.*, vol. 72, no. 1, 2015, pp. 41–57.

[13] S. Zhang, S. Sadaoui, and M. Mouhoub, "An Empirical Analysis of Imbalanced Data Classification," vol. 8, no. 1, 2015, pp. 151–162.

[14] I. Nekooeimehr and S. K. Lai-Yuen, "Adaptive semi-unsupervised weighted oversampling (A-SUWO) for imbalanced datasets," *Expert Syst. Appl.*, vol. 46, 2016, pp. 405–416.

[15] Y. Chen, Y. Lan, and H. Ren, "A Feature Selection Method Base on GA for CBIR Mammography CAD," *2012 4th Int. Conf. Intell. Human-Machine Syst. Cybern.*, Aug. 2012, pp. 175–178.

[16] C. Velayutham and K. Thangavel, "Entropy based unsupervised Feature Selection in digital mammogram image using rough set theory.," *Int. J. Comput. Biol. Drug Des.*, vol. 5, no. 1, Jan. 2012, pp. 16–34.

[17] I. L. Aroquiaraj and K. Thangavel, "Unsupervised Feature Selection in Digital Mammogram Image Using Tolerance Rough Set Based Quick Reduct," *Fourth International Conference on Computational Intelligence and Communication Networks*, 2012, pp. 436–440.

[18] I. L. Aroquiaraj and K. Thangavel, "Mammogram Image Feature Selection Using Unsupervised Tolerance Rough Set Relative Reduct Algorithm," *International Conference on Pattern Recognition, Informatics and Mobile Engineering*, 2013, pp. 479–484.

[19] A. Shakoor, "Soft Computing Based Feature Selection For Environmental Sound Classification," Blekinge Institute of Technology, April 2010.

[20] B. O. Alijla, A. T. Khader, L. C. Peng, M. A. Al-Betar, and W. L. Pei, "Fuzzy Rough Set Approach for Selecting the Most Significant Texture Features in Mammogram Images," *Palestinian International Conference on Information and Communication Technology*, 2013, pp. 51–56.

[21] T. M. Khoshgoftaar, K. Gao, A. Napolitano, and R. Wald, "A Comparative Study Of Iterative And Non-Iterative Feature Selection Techniques For Software Defect Prediction," *Inf. Syst. Front.*, Apr. 2013.

[22] M. Wasikowski, X. Chen, and S. Member, "Combating the Small Sample Class Imbalance Problem Using Feature Selection," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, 2010, pp. 1388–1400.

[23] M. Wasikowski, "Combating the Class Imbalance Problem in Small Sample Data Sets," University of Kansas School of Engineering in, 2009.

[24] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, "Handling Imbalanced Datasets : A Review," *Int. Trans. Comput. Sci. Eng.*, vol. 30, 2006.

[25] J. Wang, J. You, Q. Li, and Y. Xu, "Extract minimum positive and maximum negative features for imbalanced binary classification," *Pattern Recognit.*, vol. 45, no. 3, Mar. 2012, pp. 1136–1145.

[26] H. Wang, B. Raton, and A. F. F. R. Techniques, "A Comparative Study of Filter-based Feature Ranking Techniques,", 2010, pp. 43–48.

[27] S. Hukerikar, A. Tumma, A. Nikam, and V. Attar, "SkewBoost : An Algorithm for Classifying Imbalanced Datasets," in *International Conference on Computer and Communication Technology*, 2011, pp. 46–52.

[28] P. Phoungphol, "A Classification Framework for Imbalanced Data," Georgia State University, 2013.

[29] J. Bian, X. Peng, Y. Wang, and H. Zhang, "An Efficient Cost-Sensitive Feature

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Selection Using Chaos Genetic Algorithm for Class Imbalance Problem," *J. Math. Probl. Eng.*, vol. 2016, 2016.

[30] R. Longadge, S. S. Dongre, and L. Malik, "Class Imbalance Problem in Data Mining : Review," *Int. J. Comput. Sci. Netw.*, vol. 2, no. 1, 2013.

[31] E. Duchesnay, A. Cachia, N. Boddaert, N. Chabane, J.-F. Mangin, J.-L. Martinot, F. Brunelle, and M. Zilbovicius, "Feature selection and classification of imbalanced datasets: application to PET images of children with autistic spectrum disorders.," *Neuroimage*, vol. 57, no. 3, Aug. 2011, pp. 1003–14.

[32] G. M. Weiss, K. Mccarthy, and B. Zabar, "Cost-Sensitive Learning vs . Sampling : Which is Best for Handling Unbalanced Classes with Unequal Error Costs ?," in *DMIN*, 2007, pp. 35–41.

[33] V. López, A. Fernández, and F. Herrera, "On The Importance Of The Validation Technique For Classification With Imbalanced Datasets: Addressing Covariate Shift When Data Is Skewed," *Inf. Sci. (Ny).*, vol. 257, Feb.2014, pp. 1–13.

[34] B. Krawczyk and M. Wo, "Weighted One-Class Classification for Different Types of Minority Class Examples in Imbalanced Data," I*EEE Symposium on Computational Intelligence and Data Mining (CIDM).*, 2014, pp.337-344.

[35] K. Ganesan, U. R. Acharya, C. K. Chua, C. M. Lim, and K. T. Abraham, "One-Class Classification of Mammograms Using Trace Transform Functionals," *IEEE Trans. Instrum. Meas.*, vol. 63, no. 2, pp. 304–311, 2014.

[36] D. Zhu, "A Hybrid Approach For Efficient Ensembles," *Decis. Support Syst.*, vol. 48, no. 3, Feb. 2010, pp. 480–487.

[37] N. V Chawla, "Data Mining and Knowledge Discovery Handbook," 2010.

[38] M. Bekkar, "Imbalanced Data Learning Approaches Review," *Int. J. Data Min. Knowl. Manag. Process.*, vol. 3, no. 4, 2013, pp. 15–33.

[39] T. M. Khoshgoftaar, K. Gao, and N. Seliya, "Attribute Selection and Imbalanced Data: Problems in Software Defect Prediction," in *2010 22nd IEEE International Conference on Tools with Artificial Intelligence*, 2010, pp. 137–144.

[40] T. M. Khoshgoftaar, B. Raton, and J. Van Hulse, "A Novel Feature Selection Technique for Highly Imbalanced Data," 2010, pp. 80–85.

[41] P. G. Kumar and J. B. B. Bell, "Using Continuous Feature Selection Metrics to Suppress the Class Imbalance Problem," *Int. J. Sci. Eng. Res.*, vol. 3, no. 3, 2012, pp. 1–9.

[42] I. Jamali, M. Bazmara, and S. Jafari, "Feature Selection in Imbalance data sets," *Int. J. Comput. Sci.*, vol. 9, no. 3, 2012, pp. 42–45.

[43] X. Chen, "FAST : A ROC-based Feature Selection Metric for Small Samples and Imbalanced Data Classification Problems," in *KDD 08*, 2008, pp. 124–132.

[44] T. Deepa, "A GLFES and DFT Technique for Feature Selection in High-Dimensional Imbalanced dataset," *Indian J. Comput. Sci. Eng.*, vol. 3, no. 2, 2012, pp. 336–343

[45] W. W. Y. Ng, G. Zeng, J. Zhang, D. S. Yeung, and W. Pedrycz, "Dual Autoencoders Features for Imbalance Classification Problem," *Pattern Recognit.*, vol. 60, 2016, pp. 875–889.

[46] H. Yin, K. Gai, and Z. Wang, "A Classification Algorithm Based on Ensemble Feature Selections for Imbalanced-Class Dataset," in *2016 IEEE 2nd International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS)*, 2016, pp. 245–249.

[47] C. Jian, J. Gao, and Y. Ao, "A New Sampling Method For Classifying Imbalanced Data Based On Support Vector Machine Ensemble," *Neurocomputing*, vol. 193, 2016, pp. 115–122.

[48] S. Alshomrani, A. Bawakid, S.-O. Shim, A. Fernández, and F. Herrera, "A Proposal For Evolutionary Fuzzy Systems Using Feature Weighting: Dealing With Overlapping In Imbalanced Datasets," *Knowledge-Based Syst.*, vol. 73, Jan. 2015, pp. 1–17.

[49] J. Lee, N. Kim, and J.-H. Lee, "An Over-Sampling Technique With Rejection For Imbalanced Class Learning," *Proc. 9th Int. Conf. Ubiquitous Inf. Manag. Commun. - IMCOM '15*, 2015, pp. 1–6.

[50] L. Sriram, "Imbalanced Multiclass Data Classification Using Ant Colony Optimization Algorithm," vol. 1, no. 11, 2015, pp. 877–880.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[51] S. Cateni, V. Colla, and M. Vannucci, "A Method For Resampling Imbalanced Datasets In Binary Classification Tasks For Real-World Problems," *Neurocomputing*, vol. 135, Jul. 2014, pp. 32–41.

[52] M. a Mazurowski, P. a Habas, J. M. Zurada, J. Y. Lo, J. a Baker, and G. D. Tourassi, "Training Neural Network Classifiers For Medical Decision Making: The Effects Of Imbalanced Datasets On Classification Performance.," *Neural Netw.*, vol. 21, no. 2–3, 2008, pp. 427–36.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 1: Summarize the strength and the weakness of the techniques in handling imbalanced dataset

Techniques	Strength	Weakness
Sampling	-Suitable for highly imbalanced ratio and large datasets. -Give high accuracy for minority class.	-Change Corellation Structure extremely from original datasets. -Accuracy produced not accurate and always change depends on sampling
Feature Selection	-Reduce susceptibility, overfitting, storage memory and processing. -Enhance speed of training process. -Not rely on classification technique. -Good for continuous dataset -Good for highly imbalanced dataset. -Good for datasets with many attributes.	-Measure features independently without considering interaction between all features
Cost Sensitive Learning	-Produce good result for minority class.	-Train data based on minimum class causes of not accurate and overfitting
Ensembe technique with Iteration	-Produce higher strong learner in handling any condition of data.	-High computation time.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 2: Recent trends in handling imbalanced dataset

Year	Title	Approach	Algorithm	Domain	Problem
2016	Dual autoencoders features for imbalance classification problem [45]	Feature Selection	Dual Autoencoding Features (DAF)	Variety	Overlapping
2016	An efficient cost sensitive feature selection using Chaos Genetic Algrithm for class imbalance problem	Feature Selection	Cost Sensitive Feature Selection Chaos Genetic (CSFSG)	Variety	Large dataset
2016	A classification algorithm based on ensemble feature selections for imbalanced class dataset [46]	Ensemble Method (Feature Selection and Classification)	Ensemble Feature Selection (EFS)	Variety	Low accuracy for minority class.
2016	A new sampling method for classifying imbalanced data based on support vector machine [47]	Sampling	Different Contribution Sampling	Variety	Multiple Imbalanced ratio
2016	Adaptive semi-unsupervised weighted oversampling (A-SUWO) for imbalanced datasets [14]	Sampling	A-SUWO	Variety	Overlapping.
2015	A proposal for evolutionary fuzzy systems using feature weighting: Dealing with overlapping in imbalanced datasets [48]	Feature Selection	Weighting	Variety	Size, highly imbalanced, overlapping
2015	An over-sampling technique with rejection for imbalanced class learning [49].	Sampling	Oversampling generate synthetic data in minority class	Variety	Noisy, overfitting
2015	Coupling different methods for overcoming the class imbalance problem [9].	Ensemble methods	Hard ensemble	Variety	Outliers, imbalanced multi class,
2015	New Approach for Imbalanced Biological Dataset Classification [12].	Ensemble Methods (Sampling and two classification	Ensemble between Smote, Rotation Forest and AdaBoost	Biological Dataset	Variety imbalanced ratio

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

		techniques)			
2015	Imbalanced Multiclass Data Classification Using Ant Colony Optimization Algorithm [50].	Sampling	Under-sampling based ACO algorithm	Microarray data	High dimensionality, size, noise
2015	An Empirical Analysis of Imbalanced Data Classification [13].	Learning Algorithm	Cost Sensitive Learning –SVM	Variety	Imbalanced ratio, size complexity
2014	A method for resampling imbalanced datasets in binary classification tasks for real world problem [51]	Sampling	Combine under-sampling and oversampling techniques	Variety	Imbalanced binary classification
2013	Combining one-class classifiers for imbalanced classification of breast thermogram features [1].	Learning algorithm	One Class Classification (OCC)	Image	One Class Classification
2013	A comparative study of iterative and non-iterative feature selection techniques for software defect prediction [21].	Ensemble method (sampling and feature selection)	Random under-sampling with iterative and non iterative feature selection	NASA datasets	High dimensionality
2012	A GLFES and DFT Technique for Feature Selection in High-Dimensional Imbalanced dataset [44].	Ensemble method (sampling and feature selection)	Granularity learning, fuzzy evolutionary sampling, defuzzification technique	Microarray dataset	High dimensionality
2012	Extract minimum positive and maximum negative features for imbalanced binary classification [25].	Learning algorithm	Extracting minimum positive and minimum negative features	Image	Binary classification
2011	Feature selection and classification of imbalanced datasets: application to PET images of children with autistic spectrum disorders [31].	Feature selection	Hybrid between forward,backward selection. Apply adaptive log likehood penalization, penalization calibration based on randomized data, LOO cross validation	Image	High dimensionality, highly imbalanced
2011	SkewBoost : An Algorithm for Classifying Imbalanced Datasets [27].	Ensemble (sampling and learning algorithm)	WSmote and boosting algorithm	Variety	Imbalanced ratio

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

2010	Using Continuous Feature Selection Metrics to Suppress the Class Imbalance Problem [41].	Feature selection	Pearson Correlation Coefficient (PCC), Signal to Noise Ratio (S2N), Feature Assessment by Sliding Threshold (FAST), Feature Assessment by Information Retrieval (FAIR).	Microarray dataset	High dimensionality
2008	Training neural network classifiers for medical decision making: the effects of imbalanced datasets on classification performance [52]	Ensemble method (learning algorithm, sampling)	Particle swarm optimization, back propagation, under-sampling, oversampling	Image	Small training dataset, large number of features, correlation among features
2008	FAST: A ROC-based Feature Selection Metric for SmallSamples and Imbalanced Data Classification Problems [43]	Feature selection	FAST	Image, microarray dataset	High dimensionality