Indonesian Journal of Electrical Engineering and Computer Science
Vol. 14, No. 3, June 2019, pp. 1560~1571
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v14.i3.pp1560-1571

Imbalance class problems in data mining: a review

Haseeb Ali[1], Mohd Najib Mohd Salleh[2], Rohmat Saedudin[3], Kashif Hussain[4], Muhammad Faheem Mushtaq[5]
[1,2,4,5]Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja, 86400, Johor, Malaysia
[3]Department of Industria Engineering, Telkom University, Bandung, Indonesia

Article Info

Article history:

Received Dec 15, 2018
Revised Feb 14, 2019
Accepted Feb 27, 2019

Keywords:

Classification
Imbalanced data
Machine learning
Majority class
Minority class

ABSTRACT

The imbalanced data problems in data mining are common nowadays, which occur due to skewed nature of data. These problems impact the classification process negatively in machine learning process. In such problems, classes have different ratios of specimens in which a large number of specimens belong to one class and the other class has fewer specimens that is usually an essential class, but unfortunately misclassified by many classifiers. So far, significant research is performed to address the imbalanced data problems by implementing different techniques and approaches. In this research, a comprehensive survey is performed to identify the challenges of handling imbalanced class problems during classification process using machine learning algorithms. We discuss the issues of classifiers which endorse bias for majority class and ignore the minority class. Furthermore, the viable solutions and potential future directions are provided to handle the problems.

Corresponding Author:

Mohd Najib Mohd Salleh,
Faculty of Computer Science and Information Technology,
Universiti Tun Hussein Onn Malaysia,
Batu pahat, Johor, Malaysia.
Email: najib@uthm.edu.my

1. INTRODUCTION

Amount of data is increasing day by day along with disparate distributions in many real time applications. In a dataset, if quantity of specimens present in one classes is more than other class, then this dataset is said to be highly disparate in nature [1], [2]. The major class is used to identify the any imbalance dataset that have more number of specimens, whereas the minor class contains less number of specimens [1]. Oftenly, major class expresses the specimens as negative and minor class expresses the specimens as positive [3], [4]. The amount of majority class specimens dominates the minority class specimens by the class's ratios which can be 100 with 1 and 1000 with 1, etc. The dataset having only two classes is known as binary class, whereas the dataset containing more than two classes is known as multi-class, and both the binary and multi-class datasets suffer from imbalance data problems.

Many real-world domains include imbalance dataset problems, like detecting unreliable telecommunication customers, word pronunciations learning, marking of oil spills in the images of satellite radar, information retrieval, text classification, filtering tasks, revelation of fake telephone calls and most importantly the medical diagnosis. [5]–[7].

In such circumstances, mostly the majority classes bias the classifiers towards themselves and the classifier presents the rates of minority classes classification poorly; eventually, a classifier addresses entirely as majority class and ignores the minority class. To solve problems affiliated with the class imbalance, various techniques have been proposed in literature [8]. This is a thought-provoking and challenging in research topics nowadays [9], where many issues at a time need attention such as multiple classes problem, binary class problem, cost of misclassified class, class overlapping, insignificant disjoints, and size of the

imbalanced datasets. Problems of the binary classes related to imbalance data, received attention, but multi-class imbalance problems having various types of issues are hardly solved. In which the number of majority and minority classes can be one or more than one in the multiple class imbalance problems. Decomposition or any other techniques might be used for multi-class problem, but it still needs consideration. Hence, whenever the data will be disparate in nature, it will be seriously more daring to proceed with the minority class [1].

Owing to the importance of this issue, to solve these problems, there are significant contributions made in developing techniques. These propositions can be categorized into three types according to how they are proceeding with class imbalance, external or data level approach, which is preprocessing of data for rebalancing the class distributions to decrease the disparate distribution effect in classification process [11], [2]. The internal or algorithmic level approach creates or modifies the existing algorithms and takes consequences of minor class into consideration [12]–[14]. And the third one, cost-sensitive approach, that may unite data level and algorithmic level approaches to integrate variety of misclassification cost for every class in learning phase [15], [16].

In external or data level, before the classification process, resampling is performed in datasets to balance the data externally. For example, the specimens of majority class are randomly removed, and specimens of minority class are increased by generating artificial specimens to balance the ratio, or in ideal case, no specimen is created or deleted but choice of specimens to create or eliminate is informed [10]. In algorithmic approach, minority class is taken into consideration and the learner is not allowed to bias for the majority class to overcome the overall cost of misclassification [17]. In cost-sensitive method, we consider all types of costs, and mostly focus on misclassification cost to minimize the total cost in order to make classifier nonbiased [18].

Numerous survey and review papers on imbalanced data problems were published during last decade. Regardless of reasonable work available on handling the imbalanced data set problems, this research study especially focuses more deep survey of class imbalance problems. Following points can summarized as the main objectives of this research.

a) To review the efforts made on imbalanced data to determine how many ideas and solutions are published in this area of research.
b) To follow the research trends in data normalization and distinguishing the consequences of this area.
c) To determine the hurdles and distractions researchers faced by the influences of skewed data.

The remaining paper is organized as follows: Research methodology is given in Section 2. Section 3 briefly describes the issues of imbalanced data in classification problems and performance metrics. Section 4 provides solutions proposed for class imbalance problems. The research gaps are discussed in Section 5. Finally, Section 6 includes the conclusion and potential future directions in this area of research.

2. RESEACH METHODOLOGY

The systematic literature review search can be conducted into two ways: manual [19] and automatic [20]. First strategy was preferred by this study because the second approach shows some drawbacks since automatic search engines which are currently available are not feasible for this kind of study [21]. The manual search from the most relevant sources are commonly used for searching primary studies. This study was conducted into a two-stage search process based on research methodology of [22], in order to compile relevant papers published in last two decades. In primary stage, seven library databases: Springer, Elsevier, IEEExplore, Sage, ACM, Cambridge and Wiley, were employed to search and collect literature, which covers most natural science and social science research fields. To furnish a complete set of search terms to cover application and technical articles on imbalanced data and rare events, a two-level keywords tree is given in Figure 1.

First level of the tree was limited to basics of skewed/imbalanced data, which focused on class imbalance classification, learning from imbalanced data. The second level of tree search terms were divided into two nodes to takeover both practical and technical articles. Keywords for techniques and approaches for the class imbalance classification in data mining, keywords for applications for rare events like fraud detection, cancer medical diagnosis, challenges and their solutions. The primary search yielded 550 papers on imbalanced data domains, which were downloaded and again filtered for next stage. After manual review of paper almost 400 papers were found to be relevant of this study. In second stage of search relevant cross references, through google scholar, these papers are also included in this search review. After second stage theses papers also added into this review and total of 440 papers are included into this study.

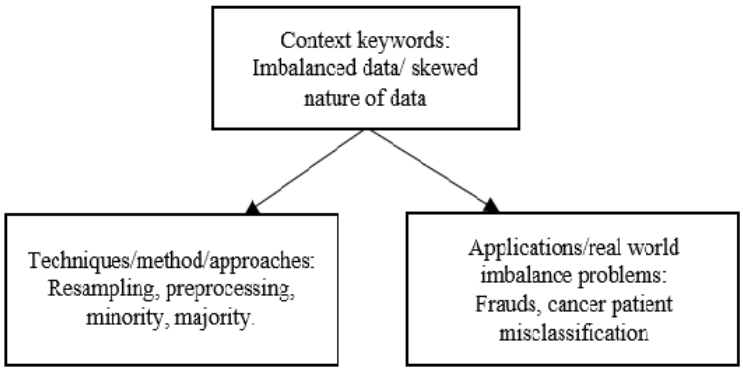

Figure 1. Two level keywords tree for research methodology

3. ISSUES AND PERFORMANCE PARAMETERS IN IMBALANCED DATA PROBLEMS

The class imbalance problems eventuate due to the presence of minor class in data, and on classification, the classifiers show bias behavior towards the majority class. For instance, in a majority class we have 99% specimens and just 1% specimens in minority class, and then accuracy attained by the classification algorithm is 99%, that is from the majority class. Whereas, the minority class is dominated by majority class and that 1% could be some important samples. For example, catching of cancerous cells in medical diagnosis, misclassification of non-cancerous cells may prescribe more clinical tests, but if defected cells of cancer will be misclassified then it poses a very serious health risk. Hence, in unbalanced data classification, due to the design principles of algorithms, the majority class specimens are not or less misclassified. Subsequently, the all-around accuracy of classification in machine learning algorithms mostly misclassifies minority class which will result into more misclassification cost, time and risk evaluation [23].

Rareness comes in the result of erring nature of small disjoints. So, why rareness is a problem, can be understood by knowing the reasons why small disjoints are so error disposed. One reason is that small divisions or disjoints may not show the rare or anomalous cases, yet relatively something else like noisy data. Therefore, it should keep only "meaningful" small divisions. For removing sub-ideas (i.e., disjoints) that are not meaningful, mostly classifier induction systems have some ways of avoiding over fitting. Inductive bias in rare classes also performs a role. So, in the presence of uncertainty (or biasing towards priors class), many induction systems lean to favor more common classes [17]. Distinguishing between two components of the imbalanced datasets is a major issue of imbalanced datasets problems:

- (IR) the imbalance as the ratio of $\frac{number\ of\ minority}{number\ of\ majority}$
- (LI) The lack of information for the minority class.

Both components exist in imbalanced datasets learning problems, but as mentioned before, other factors such as overlapping, over fitting, small disjoints and size of datasets also affect a specific machine learning algorithm. All algorithms suffer from the lack of information (which cannot be learned) but it is important to conclude the ones which do not suffer from imbalance data ratio. For example, a dataset containing 1:99 ratio for minority:majority specimens respectively, *IR* is same for ratio 10:990, but minority class is low esteemed due to poor representation in first dataset and sustain from *LI* lack of information than the second dataset [24].

For performance evaluation, most important consideration is to choose appropriate performance metrics for analysis. To calculate the accuracy rate, we commonly use confusion metrics. But, it shows some flaws in results like it neglects the minority class data and it just lowered the global measures to give the best result like as low error rate [25]. A very applicable visual tool, Receiver Operating Characteristic (ROC) curve demonstrates the diagnostic ability of a learning classification system as its perception threshold is diverse, it interpret the trade-off between the costs (*False Positive rate*) and benefits (*True Positive rate*) [29]. For evaluation of models on average, the Area Under an ROC curve (AUC) gives a measure performance of single classifier [27]. There are also many other metrics used for performance evaluation, some famous and commonly used are Geometric-mean (G-mean), F-measure (Fm), sensitivity, and specificity [25], [28].

4. SOLUTIONS PROPOSED FOR CLASS IMBALANCE PROBLEMS

Owing to the importance of class imbalance problems, there is significant contribution made in developing techniques to solve these problems. These propositions can be categorized into three types according to how they are proceeding with class imbalance, external or data level approach, internal or algorithmic level approach and cost-sensitive approach. Moreover, ensemble learning classifiers are also play a significant role in imbalanced data classification.

4.1. External, Data Level (Preprocessing) Approach

Resampling techniques are used for preprocessing of imbalanced data that can be distributed into three types; random under-sampling method that eliminates specimens of majority class randomly and generates a subset of primary dataset in a way to balance the ratio. It may lead to the loss of potential data due to eliminating some data that can be used in induction process. Random over-sampling method, it increases the quantity of specimens in minority class by replicating the existing specimens randomly and generates a superset of the primary data. But, it can enhance the chances of over fitting due to replication. And lastly, hybrid technique, it unites both sampling methods to balance the ratio [29].

Synthetic minority over sampling technique (SMOTE) [10], in minority class, new specimens are created by interpolation of minority class specimens which reside together. SMOTE selects randomly one of the *k*-nearest neighbors (kNN) of an inferior specimens and creates a duplicate specimen generating value from a random both interpolated specimens. Decision boundaries have been made for minority class to disperse more into the space of majority class. Hereby, this method avoids the over fitting problem but it creates noisy and borderline specimens that may create problems. For such problems faced in SMOTE, some of the filtering based methods are used to avoid noise in imbalanced datasets (i.e. SMOTE-TL and SMOTE-EL) on the other hand, for handling the imbalanced data, original sampling techniques are also modified with neighborhood-balanced bagging (NBBag) [30].

Modified synthetic minority oversampling technique (MSMOTE) [31] is an improved form of SMOTE. By calculation of distances between all specimens in this algorithm, minority class is divided into three groups, latent noise, safe, and border specimens. It rejects the hidden noise spots based on kNN classification method when MSMOTE generates new examples. However, it does nothing for hidden noise instances, also does not prioritize the important characteristics. Extension of SMOTE and Iterative–Partitioning Filter (IPF) is used to handle noises and regular the class boundaries [32]. Modifications of SMOTE go further for more powerful data level techniques, like extension of SMOTE, B1-SMOTE and B2-SMOTE [32] to normalize the imbalanced data.

Majority weighted minority oversampling technique (MWMOTE) [33] selects the specimens of minority class effectively which are difficult-to-learn and then allocates weights appropriately. Moreover, it is able to create accurate artificial examples. Selective preprocessing of imbalanced data (SPIDER) [34] is also a technique which, from majority class, it merges screened complex examples with local oversampling of minority class by coinciding two phases, identification and preprocessing.

A new inverse random under-sampling (IRUS) [35] is used to resolve the imbalanced data problem by using inverse (ratio of unbalance cardinality) approach. It is also significant for multi-label classification. In [36], for handling the problems of imbalanced datasets, radial basis function network (RBFN) is proposed in which strategy of training local weights is used, which is designed by using local and global terms. In local weights training methods, lessened value of imbalance ratio (IR) should be balanced with any technique (i.e. SMOTE or any other) and the major value of IR gives better results.

Authors proposed in [37] a classifier algorithm which is a combination of particle swarm optimization (PSO) and SMOTE, whereas incorporating like logistic regression (LR), C5 decision tree (C5) model, and 1-nearest neighbor search and some famous classifiers. Accuracy indices and G-mean are used as performance metrics for this new set of classifiers for justifying its effectiveness. Experimental results show that PSO + SMOTE + C5 is an efficient hybrid algorithm for 5-year lastingness of breast cancer patients. Another technique very powerful for binary class imbalanced problems is proposed that is fusion of PSO, SMOTE and aided radial basis function (RBF) classifier in [38] and tested by different types of metrics shows combination of SMOTE + PSO-RBF performs well on normal imbalanced datasets, but not satisfactory for the highly imbalanced datasets. Neighbor Weighted K-Nearest Neighbor (NW-KNN) [39] used for sambat online classification and this algorithm is able to classify imbalanced data with optimal value of k neighbor 3. All data level techniques are summarized in Table 1 with pros and cons.

4.2. Internal Algorithmic Approaches

Internal or algorithmic level approach may create or modify the existing algorithms and take consequences of minor class into consideration for handling the imbalanced data. Different types of algorithms are discussed below.

4.2.1. Variants of Support Vector Machine (SVM)

Class imbalance problems occur frequently in DNA microarray data, due to which forecasting performance for minority classes becomes poor. Furthermore, other features, such as high-dimension and high noise, small sample, etc., aggravate this issue. Ant colony optimization (ACO) sampling was proposed in [40], that is a new under-sampling technique based on the ACO idea for handling this problem. Support vector machine (SVM) is a classification technique which is generally used for the imbalanced data. Data level approaches with pros and cons shown is Table 1. SVM and its variants with their pros and cons shown is Table 2.

Table 1. Data level Approaches with Pros and Cons

Data level Methods/Approaches	Pros	Cons
Synthetic minority oversampling technique (SMOTE) [10]	Increase minor class examples to balance	Over fitting
Neighborhood-balanced bagging (NBBag) [30]	Better than current over-sampling bagging extension and competitive to randomly balance bagging.	Costly
Modified SMOTE (MSMOTE) [31]	It reduces the noise.	Does not consider the priorities of important features.
Iterative–Partitioning Filter + SMOTE (SMOTE-IPF) [32]	Addressing the problem of bordered and noise examples in unbalanced data sets.	Small sample size
Majority weighted minority oversampling technique (MWMOTE), [33]	According to Euclidean distance it creates artificial inferior samples.	Multi class imbalance problem.
Selective preprocessing of imbalanced data (SPIDER) [34]	Filter difficult examples.	Complex
Novel inverse random under-sampling (IRUS) [35]	Improve multi-label classification accuracy, Beneficial for irregular learning datasets sizes.	Other different applications to multi-label classification.
Radial Basis Function Networks (RBFN) SMOTE [36]	Higher the IR Value of dataset tends to better result.	More storage Space.
SMOTE + PSO + C5 [37]	Estimate 5-year lastingness of breast cancer patients	It can improve for other cancer datasets.
Combined SMOTE and PSO-based RBF classifiers (SMOTE+PSO-RBF) [38]	Create synthetic specimens for minority class, and RBF give very good performance.	More storage space.

Table 2. SVM and its Variants with Their Pros and Cons

Algorithms/Methods/Approaches	Pros	Cons
Ant colony optimization (ACO) sampling, SVM [40]	To resolve imbalanced data classification problem by ACO algorithm based sample selection process.	Excessive Computational and storage cost.
Mega-trend diffusion and Support vector machine (MTD-SVM) [41]	Enhance the number of samples in minority class.	Synthetic data generation is costly, and MTD technique performs better on small sized datasets.
Adjusted F-measure, SVM with suitable kernel transformation [42]	It manages the cost function and imbalance data with kernel scaling.	Efficient estimation Strategy for parameters and different kernels.
Parallel Selective Sampling (PSS), combined with SVM, (PSS-SVM). [43]	Accurate statistical predictions and low computational complexity	For parallel and distributed computing.
EnSVM and EnSVM+ with additional re-sampling method [44]	Effective than normal SVM	It does not determine the value of k automatically.
Preprocessor SVM with MLP, LR, and RF intelligent algorithm. [45]	It balance data in effective manner and increase the number of samples in minority class.	Not so simpler and faster
Second-order cone programming with SVM (SOCP-SVM) [46]	Due to SVM-LP formulation it gives Robust and improved classification performance.	Only designed for imbalanced data.
Near-Bayesian Support Vector Machine (NBSVM) [47]	It minimizes the misclassification cost by minority class.	Performance metrics

The number of specimens can also be enhanced in minority class by using mega-trend diffusion (MTD) method. SVM and KNN are employed with machine learning methods for creating hybrid MTD-KNN, MTD-SVM and prediction models in the predictor stage [41]. By using the cost analysis, it is marked that this technique MTD-SVM is finest model as compared to Random Forest, Naïve Bayes, and KNN.

In an imbalanced dataset, majority and minority problems are handled in [42], in which SVM is improved with the kernel scaling method. Parallel selective sampling (PSS), combined with the SVM, PSS-SVM [43] classification presented promising results on benchmark datasets, that is far superior from standard SVM because of no convergence. For reducing unbalancing in large data sets, it is able to select data from the majority class. Authors in [44] combined ensembles of SVMs with both under-sampling and over-sampling techniques for improving the prediction performance. Comprehensive experiments showed that this technique is better than individual SVM as well as several other classifiers. The research experimented on selective ensemble EnSVM+ and base model EnSVM with additional re-sampling methods.

Multilayer perceptron (MLP), random forest (RF), logistic regression (LR) and other intelligent machine learning algorithms results are more improved by these approaches in which, SVM trained as the preprocessor for improving the results of intelligent algorithm. Two phases have been taken in balancing approach; in first phase, the SVM tuned the imbalanced data to get improved balance data and this improved data, which is used as the input in the second phase to MLP, RF and LR [45]. Furthermore, these all techniques are summarized in Table 2 with pros and cons.

4.2.2 Clustering

Clustering is used to separate the data into classes, whereas for detecting the minority class samples in data, outlier detection is employed. On the basis of clustering techniques and outliers detection, the similarity based hierarchal decomposition method takes place. It contains two portions in hierarchy construction; one, in which clusters are misclassified, and second, clusters are perfectly classified [48]. Data similarities of labeled subsets at every level are used to make hierarchy and feature subsets as well to build other data based on these different levels. Class overlapping and variety of imbalanced problems can be avoided by this method.

Fuzzy rule based classification systems (FRBCSs) [49] also improved the classification performance. For handling the imbalanced data, this is a useful technique where low and high ratio of imbalanced datasets can be taken by this method by using 2-tuple genetic tuning which also enhances the performance of FRBCSs. Summary of this method is given in Table 3.

4.2.3. Feature Selection

For many machine learning algorithms, feature selection is known to be door step, in such situations when the data is exceptionally high-dimensional. The data which is disparate in nature and high-dimensional in learning, feature selection method is superior to achieve the best possible results. When dealing with small feature sets, the fundamental power of FS metrics is negated. In general, main objective of the feature selection is to permits the classifier for achieving best performance by choosing a subset of *y* features; here *y* is a user-specified parameter. Each feature is valued independently based on a rule by using filters for high dimensional data sets [50].

Existing feature selection measures for imbalanced datasets are not suitable, suggested by Zheng et al [51]. A feature selection frame work was proposed by authors, which chooses the majority and minority classes features separately and combines these features explicitly. Existing measures in this way can be simply converted by considering the features of majority and minority class individually. This method is so far simple but effective for all high dimensional and the imbalanced datasets having limited sample size classes of the imbalanced datasets.

To resolve the problems in imbalance datasets related to feature selection, two different methods were proposed by authors, Decomposition-based and Hellinger distance-based methods. First method was used to tackle the conflicts of the imbalanced class distribution by measuring the distributive differences. In second method, huge classes were isolated in synthetic subclasses and various techniques were utilized for classification process [52]. Summary is given in Table 3.

4.2.4. One Class Learning

In this method, the algorithm identifies such specimens which correspond to that particular class whereas rejects the remaining. In this way, it is helpful for imbalanced data classification. For instance, for minority class, it fetches the samples specifically which belong to its class, ignore the other samples, and does same for the majority class. One class learning for high-dimensional imbalanced datasets, gives better performance than others [50]. Furthermore, this method is based on the rules, the separate and conquer approach is used in rule induction system to build iteratively rules and cover the examples of

previous uncovered training examples, Ripper [53]. From most rare to most common class, it makes rules for each class, up to no negative samples covered; it keeps adding conditions for each rule. The capability of this algorithm provided by Ripper is fairly directed to learn rules for minority class only.

For excessively imbalanced datasets which may be encompass of noisy features and high dimensional space one class learning is efficiently useful. Kowalczyk and Raskutti [5] argue that violent feature selection methods are related to one class learning, however often it can be expensive to apply rather than one class learning, which is more practical. This method is summerized with pros and cons in Table 3.

4.3. Cost Sensitive Learning

Cost sensitivity framework lies between internal and external level approaches. It integrate both approaches, algorithmic level modifications and data level alterations by modifying the learning procedure to accept costs and adding costs to samples respectively [16], [54]. For minimizing the overall error cost of both classes and by assuming higher misclassification cost of minority class, it tends the classifier to bias towards minority class. Moreover, from an example in the aspects of cost sensitivity for minority class of a certain cancer patients medical diagnosis, if we declared cancer patient as positive class (i.e. minority class) and non-cancer, healthy as negative class (majority class) so misclassifying a cancer patient is called "false negative" (that was actually positive but classified as negative) is very sensitive case and expensive as compared to "false positive" (that was actually negative but classified as positive) error i.e. for negative class. In misclassification or delay in correct medical diagnosis and treatment, patient can lose his/her life [18]. For minimizing the misclassification and total test cost, cost matrix for cost sensitive learning and formulation cost also with their improvements are discussed in [18], [56].

For reaching unequal treatment to the classes which is not equally treated by cost-sensitive learning, it preserves the main AdaBoost learning framework [29] and also it introduces cost items simultaneously into weight update formula., Therefore, the common difference in these proposals can be like how they improve the weight update formula. From the boosting family of cost sensitivity, the most representative approaches are AdaC1,AdaC2 and AdaC3[56], CSB1,CSB2 [57], AdaCost[58]. Proposed texonomy for the review of imbalanced class problems in data mining shown as Figure 2.

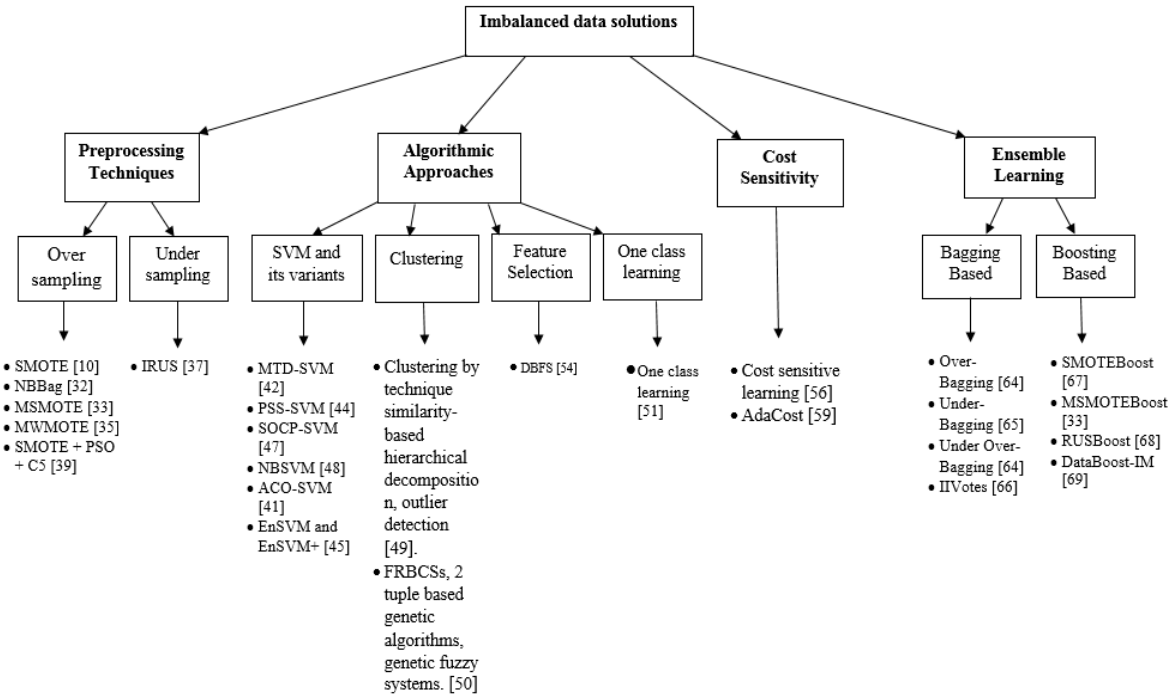

Figure 2. Proposed texonomy for the review of imbalanced class problems in data mining

4.4. Ensemble Learning Algorithms

Composition of several classifiers is used in a manner to get a new classifier so that every one of single classifier performs better [59]. Accuracy of single classifier in machine learning has been increased by

ensembles of classifier as well as single learning classifiers cannot solve imbalance class problem individually, but to deal with this issue, learning algorithms can be specially deliberated [29]. Different techniques and methods can be used to develop ensemble learning algorithms by considering a weak learning algorithm. Like Bagging [60], Boosting [61] and Adaboost [62] are the most famous ensemble learning algorithms.

- Bagging : The concept of bootstrap accumulating to construct ensembles introduced by Breiman [60]. Bagging used for imbalanced data problems in a way to create the balanced datasets distributions of minority and majority samples. Furthermore, four main bagging based ensembles are proposed: Over-Bagging [63], Under-Bagging [64], Under Over-Bagging [63], without compromising the importance of the diversity.
- Boosting : In 1990, Schapire [61] introduced a new term Adaptive Resampling and Combining (ARCing). The researcher certified that Probably Approximately Correct (PAC) learning framework can change a weak learner into a strong learner. Boosting is little variation in bagging like selecting the points which give wrong prediction. SMOTEBoost [65], MSMOTEBoost [31], RUSBoost [66], and DataBoost-IM [67] algorithms are included in boosting based ensembles for training of ensuing classifier directed to minority class.
- AdaBoost [62] was first practicable approach of Boosting, and it is found in top ten data mining algorithms [68]. In imbalanced data, the algorithm biases the learning (the weight), but in AdaBoost whole dataset is used to train each classifier serially and after iteration, the samples which are harder to classify (minority examples) are mainly focused by this approach [62]. AdaBoost.M1 and AdaBoost.M2 [69] are the two of famed modifications that have been used in imbalanced dominions.

For classification of imbalanced data, a novel ensemble technique is also used, that convert an imbalanced dataset into many balanced subsets of original data and number of classifiers with specific classification algorithm are then applied on these multiple subsets. These classifiers for new data give classification results which again united by specific ensemble rule [70].

4.5. Multiclass Imbalance Problem

Binary class imbalance problems are mainly more focused by all efforts done so far. But in multiclass imbalance problems, there is not enough research performed to solve these problems. These unsolved issues found in many real-world applications which have multi-minority, multi-majority classes like one majority and many minority classes or one minority and many majority classes respectively. Both types of classes affect negatively to minority and overall performance [1].

Table 3. Clustering, Feature seletion, One Class Learning, Cost Sensitivity, Ensemble Learning and Multiclass Techniques and Approaches

Algorithms/Methods/Approaches	Pros	Cons
Clustering by technique similarity-based hierarchical decomposition, outlier detection [48]	This technique useful for identification of classes and balancing them.	In duration of training this method, computational complexity is very high.
FRBCSs, 2 tuple based genetic algorithms, genetic fuzzy systems. [49]	Enhance the performance of standard FRBCSs.	Highly imbalanced datasets.
Density Based Feature Selection (DBFS) [71]	To tackle high-dimensional data and the small sized samples problem in imbalanced datasets.	Don't work for multiclass, encounter many problems.
Decomposition-based and Hellinger distance-based methods. [52]	To solve the feature selection issues in the imbalanced datasets.	Only compared with only three trending feature selection methods.
One class learning [50]	Efficient when data is high dimensional, useful for both binary and multiclass imbalanced datasets	Expensive.
Cost sensitive learning [55]	Reduce the total error cost	Define misclassification
AdaCost [58]	Efficient than normal CSL	Complex
Ensemble Learning [59]	It increases the generalization ability and accuracy of single classifier	Alone not sufficient to solve class imbalance problems
A Novel ensemble method for classifying imbalanced data. [70]	No loss of information and remove chances of mistakes	For only binary class imbalanced data.
OAA and OAO Schemes [72] [73]	Improve the coverage of minority class samples	Create ambiguity

For solution of multiclass imbalance problems, most existing methods use decomposition in which it handles each imbalanced binary subtask by employing binary class imbalance techniques. One most common example of imbalanced multiclass problem is protein fold classification, The author Tan et al. [74] put one-

against-one (OAO) [73], one-against-all (OAA) [72] ideas to resolve this problem and improved the prediction of minority class examples, the research built rule-based learners. Without full data knowledge, it trained each individual classifier, which may left uncovered data regions and also cause classification uncertainty by each type of decomposition [74]. Without using class decomposition for multiclass imbalance problem which addresses directly, a cost-sensitive ensemble technique is proposed [75].

A noval adaptive dara structure based oversampling [76] model is proposed which create synthetic samples and Extreme Learning Machine for Ordinal Regression (ELMOP) for multiclass imbalanced data issues.The summary of these methods is shown in Table 3.

5. RESEARCH GAPS

There are corresponding gaps in the in the field of imbalanced data that need attention from the researcher community. Generally, the preprocessing of data in resampling methods is more effective to balance the class data before learning process. Although, many achievements has been made using hybrid sampling techniques. Still, there are some issues that need to be solved; such as, over-fitting, computation cost, lack of consideration of important features, and storage consumption of some techniques. Various methods of SVM in algorithmic level are useful to resolve imbalanced data problems and misclassification cost, but such techniques encounter problems of excessive computational and storage cost, yet some techniques are not simple or fast or specific for the problems. For the purpose of high dimensional data, the feature selection method is efficient to imbalanced class data; however it is inefficient on multi-class datasets, also insignificant for small sized data. Clustering and one-class learning are also specific for some datasets and imbalance problems. Cost sensitive approach is used to reduce the total error cost and efficiently work with boosting learning algorithm for imbalanced class problems, yet it requires additional focus to define misclassification cost. Ensemble learning algorithm handle the over fitting problem and generalization ability of class imbalanced problems, however some ensemble learning techniques alter the data distribution, hard to implement on real world data. Also multi-class dataset imbalanced issues need more improvement. Research gaps are summarized as follows.

- Techniques used for oversampling like SMOTE [10] are suffering through overlapping, noise or overfitting of minority and majority samples due to wrong selection of samples for synthetic generation, this is in also case of under-sampling techniques which remove some potential data from the majority class.
- The hybrid algorithms proposed for classification of imbalanced data like SOCP-SVM [46], MTD-SVM[41] give significant accuracy but complexity of algorithm is increased due to hybrid approaches. Clustering by technique similarity-based hierarchical decomposition [48], outlier detection computational complexity is high in training process. One class learning [50] is significantly expensive as it has complex model which is robust for both binary and multi class datasets, moreover its computational time is long.
- Cost sensitive learning is significant for imbalanced class problems especially in medical real world datasets but it cannot define the misclassification cost specifically.

6. CONCLUSION AND FUTURE WORK

This paper surveyed literature and found the theoretical concepts of imbalance data problems, and presented different challenges and methods to handle imbalance data problems in classification. It is essential to balance the imbalanced class with efficient method by considering the cost factor. The right selection of classifier methods with performance evaluation metrics should be applied in order to accomplish better results. In conclusion, we have found that ensemble learning algorithms handle the over fitting problem and increase generalization ability of class imbalanced problems. The research community presented positive collaboration between the sampling techniques and bagging ensemble learning algorithms; such as, RUSboost, which showed to be the least complex in computation among all significant performers. Furthermore, current novel ensemble learning methods are also effective for both binary. Such as Boosting, which is also useful ensemble learning algorithm; it enhances the performance of weak classifiers. While developing more solutions for imbalanced learning problems, the research community should consider the following summarized directions.

- The nature and structure of samples in minority classes need focus to be improved source of learning complications.
- Future research on learning algorithms should consider new area, research in how learning algorithms are divergent and consider specific method for what type of learning problems.

- Propose solutions for multi-label learning based on particular designed nature of problem.
- Besides altering the data distributions, focus on the nature of imbalanced dataset.
- Propose new techniques for multi-class imbalanced data that consider different relationships between the classes.

ACKNOWLEDGEMENTS

The authors would like to thank Universiti Tun Hussein Onn Malaysia (UTHM) for supporting this research under Postgraduate Incentive Research Grant, Vote No.H334.

REFERENCES

[1] S. Wang and X. Yao, "Multiclass Imbalance Problems : Analysis and Potential Solutions," *IEEE Trans. Syst. Man. Cybern.*, vol. 42, no. 4, pp. 1119–1130, 2012.
[2] N. V Chawla, N. Japkowicz, and P. Drive, "Editorial : Special Issue on Learning from Imbalanced Data Sets," *Sigkdd Explor.*, vol. 6, no. 1, pp. 2000–2004, 2004.
[3] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, 2009.
[4] J. Van Hulse and T. Khoshgoftaar, "Knowledge discovery from imbalanced and noisy data," *Data Knowl. Eng.*, vol. 68, no. 12, pp. 1513–1542, 2009.
[5] B. Raskutti and A. Kowalczyk, "Extreme Re-balancing for SVMs: a case study," *ACM SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 60–69, 2004.
[6] G. Wu and E. Y. E. Chang, "Class-boundary alignment for imbalanced dataset learning," *Twent. Int. Conf. Mach. Learn. (ICML), Work. Imbalanced Data Sets*, no. 1, pp. 49–56, 2003.
[7] R. Yan, Y. Liu, R. Jin, and A. Hauptmann, "ON PREDICTING RARE CLASSES WITH SVM ENSEMBLES IN SCENE CLASSIFICATION," *Landscape*, pp. 21–24, 2003.
[8] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "A comparative study of data sampling and cost sensitive learning," *Proc. - IEEE Int. Conf. Data Min. Work. ICDM Work. 2008*, pp. 46–52, 2008.
[9] X. Wu, "10 Challenging Problems in Data Mining Developing a Unifying Theory of Data Min- ing Scaling Up for High Dimensional Data and High Speed Data Streams," pp. 1–9, 2005.
[10] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.
[11] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data," *ACM SIGKDD Explor. Newsl. - Spec. issue Learn. from imbalanced datasets*, vol. 6, no. 1, pp. 20–29, 2004.
[12] J. R. Quinlan, "Improved estimated for the acccuracy of small disjuncts," *Mach. Learn*, vol. 6, no. 1991, pp. 93–98, 1991.
[13] B. Zadrozny and C. Elkan, "Learning and making decisions when costs and probabilities are both unknown, " *Proc. seventh ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD '01*, pp. 204–213, 2001.
[14] G. Wu and E. Y. Chang, "KBA: Kernel boundary alignment considering imbalanced data distribution," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 6, pp. 786–795, 2005.
[15] A. Freitas, A. da Costa Pereira, and P. Brazdil, "Cost-Sensitive Decision Trees Applied to Medical Data.," *DaWaK*, vol. 4654, pp. 303–312, 2007.
[16] N. V. Chawla, D. A. Cieslak, L. O. Hall, and A. Joshi, "Automatically countering imbalance and its empirical relationship to cost," *Data Min. Knowl. Discov.*, vol. 17, no. 2, pp. 225–252, 2008.
[17] and R. C. A. M. V. Joshi, V. Kumar, "Evaluating Boosting Algorithms to Classify Rare Classes : Comparison and Improvements," *First IEEE Int. Conf. Data Min.*, pp. 257–264, 2001.
[18] C. X. Ling and V. S. Sheng, "Cost-Sensitive Learning and the Class Imbalance Problem," *Encycl. Mach. Learn.*, pp. 231–235, 2008.
[19] S. Keele, "Guidelines for performing Systematic Literature Reviews in Software Engineering," *EBSE Tech. Rep.*, vol. 2.3, no. 01, 2007.
[20] M. Petersen, K. and Feldt, R. and Mujtaba, S. and Mattsson, "Systematic Mapping Studies in Software Engineering," *Proc. 12th Int. Conf. Eval. Assess. Softw. Eng.*, no. February 2015, pp. 68–77, 2008.
[21] P. Brereton, B. A. Kitchenham, D. Budgen, M. Turner, and M. Khalil, "Lessons from applying the systematic literature review process within the software engineering domain," *J. Syst. Softw.*, vol. 80, no. 4, pp. 571–583, 2007.
[22] M. B. J. D. Kannan Govindan, "ELECTRE: A comprehensive literature review on methodologies and applications," *Eur. J. Oper. Res.*, vol. 250, no. 3, pp. 1–29, 2016.
[23] R. Longadge, S. S. Dongre, and L. Malik, "Class imbalance problem in data mining: review," *Int. J. Comput. Sci. Netw.*, vol. 2, no. 1, pp. 83–87, 2013.
[24] A. R. Visa, Sofia, "Fuzzy Classifiers for Imbalanced, Complex Classes of Varying Size," *In Proc. ofthe IPMUConference,Perugia*, p. 393–400., 2004.
[25] and M. K. Han, Jiawei, *Data Mining: Concepts and Techniques*. 2001.
[26] W. B. Yu, Y. chin I. Chang, and E. Park, "A modified area under the ROC curve and its application to marker selection and classification," *J. Korean Stat. Soc.*, vol. 43, no. 2, pp. 161–175, 2014.

[27] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve.," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.
[28] R. Batuwita and V. Palade, "A new performance measure for class imbalance learning. Application to bioinformatics problems," *8th Int. Conf. Mach. Learn. Appl. ICMLA 2009*, pp. 545–550, 2009.
[29] M. Galar, A. Fern, E. Barrenechea, and H. Bustince, "A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches," *IEEE Trans. Syst. MAN, Cybern. C Appl. Rev.*, vol. 42, no. 4, pp. 463–484, 2012.
[30] B. Jerzy and J. Stefanowski, "Neighbourhood sampling in bagging for imbalanced data," *Neurocomputing*, vol. 150, p. 529–542., 2014.
[31] L. M. Hu, Shengguo, Yanfeng Liang, Ying He, "MSMOTE : Improving Classification Performance when Training Data is imbalanced," *Second Int. Work. Comput. Sci. Eng.*, pp. 627–631, 2009.
[32] J. A. Sáez, J. Luengo, J. Stefanowski, and F. Herrera, "SMOTE – IPF : Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering," *Inf. Sci. (Ny).*, vol. 291, pp. 184–203, 2015.
[33] S. Barua, M. Islam, X. Yao, and K. Murase, "MWMOTE — Majority Weighted Minority Oversampling Technique for Imbalanced Data Set Learning," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 2, pp. 405–425, 2014.
[34] J. Stefanowski and S. Wilk, "Selective Pre-processing of Imbalanced Data for," *Data Warehous. Knowl. Discov. (Lecture Notes Comput. Sci. Ser. 5182)*, pp. 283–292, 2008.
[35] M. Atif, J. Kittler, and F. Yan, "Inverse random under sampling for class imbalance problem and its application to multi-label classification," *Pattern Recognit.*, vol. 45, no. 10, pp. 3738–3750, 2012.
[36] A. J. Rivera, C. J. Carmona, and M. J. Jesus, "Training algorithms for Radial Basis Function Networks to tackle learning processes with imbalanced data-sets," *Appl. Soft Comput. J.*, vol. 25, pp. 26–39, 2014.
[37] K. Wang, B. Makond, K. Chen, and K. Wang, "A hybrid classifier combining SMOTE with PSO to estimate 5-year survivability of breast cancer patients," *Appl. Soft Comput. J.*, vol. 20, pp. 15–24, 2014.
[38] M. Gao, X. Hong, S. Chen, and C. J. Harris, "Neurocomputing A combined SMOTE and PSO based RBF classifier for two-class imbalanced problems," *Neurocomputing*, vol. 74, no. 17, pp. 3456–3466, 2011.
[39] A. A. Prasanti, M. A. Fauzi, and M. T. Furqon, "Neighbor Weighted K-Nearest Neighbor for Sambat Online Classification," *Indones. J. Electr. Eng. Comput. Sci. Vol.*, vol. 12, no. 1, pp. 155–160, 2018.
[40] H. Yu, J. Ni, and J. Zhao, "ACOSampling: An ant colony optimization-based undersampling method for classifying imbalanced DNA microarray data," *Neurocomputing*, vol. 101, pp. 309–318, 2013.
[41] A. Majid, S. Ali, M. Iqbal, and N. Kausar, "Prediction of human breast and colon cancers from imbalanced data using nearest neighbor and support vector machines," *Comput. Methods Programs Biomed.*, vol. 113, no. 3, pp. 792–808, 2014.
[42] A. Maratea, A. Petrosino, and M. Manzo, "Adjusted F-measure and kernel scaling for imbalanced data learning," *Inf. Sci. (Ny).*, vol. 257, pp. 331–341, 2014.
[43] A. D. Addabbo and R. Maglietta, "Parallel selective sampling method for imbalanced and large data classification," *Pattern Recognit. Lett.*, vol. 62, pp. 61–67, 2015.
[44] Y. Liu, X. Yu, J. Xiangji, and A. An, "Combining integrated sampling with SVM ensembles for learning from imbalanced datasets," *Inf. Process. Manag.*, vol. 47, no. 4, pp. 617–631, 2011.
[45] M. A. H. Farquad and I. Bose, "Preprocessing unbalanced data using support vector machine," *Decis. Support Syst.*, vol. 53, no. 1, pp. 226–233, 2012.
[46] S. Maldonado and J. López, "Imbalanced data classi fi cation using second-order cone programming support vector machines," *Pattern Recognit.*, vol. 47, no. 5, pp. 2070–2079, 2014.
[47] S. Datta and S. Das, "Near-Bayesian Support Vector Machines for imbalanced data classification with equal or unequal misclassification costs," *Neural Networks*, vol. 70, pp. 39–52, 2015.
[48] C. Beyan and R. Fisher, "Classifying imbalanced data sets using similarity based hierarchical decomposition," *Pattern Recognit.*, vol. 48, no. 5, pp. 1653–1672, 2015.
[49] A. Fernández, M. José, and F. Herrera, "On the 2-tuples based genetic tuning performance for fuzzy rule based classification systems in imbalanced data-sets," *Inf. Sci. (Ny).*, vol. 180, no. 8, pp. 1268–1291, 2010.
[50] M. Wasikowski, X. Chen, and S. Member, "Combating the Small Sample Class Imbalance Problem Using Feature Selection," *EEE Trans. Knowl. DATA Eng.*, vol. 22, no. 10, pp. 1388–1400, 2010.
[51] R. S. Zheng, Zhaohui, Xiaoyun Wu, "Feature Selection for Text Categorization on Imbalanced Data," *Sigkdd Explor.*, vol. 6, no. 1, pp. 80–89, 2004.
[52] L. Yin, Y. Ge, K. Xiao, X. Wang, and X. Quan, "Feature selection for high-dimensional imbalanced data," *Neurocomputing*, vol. 105, pp. 3–11, 2013.
[53] M. Avenue, M. Hill, W. W. Cohen, C. Of, and R. Pruning, "Fast Effective Rule Induction," *Proc. Twelfth Int. Conf.*, pp. 115–123, 1995.
[54] C. X. Ling, V. S. Sheng, and Q. Yang, "Test strategies for cost-sensitive decision trees," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 8, pp. 1055–1067, 2006.
[55] J. Liu, X. Zhou, D. Li, X. Li, Z. Dong, and S. Wang, *Advanced Data Mining and Applications*. 2005.
[56] Y. Sun, M. S. Kamel, A. K. C. Wong, and Y. Wang, "Cost-sensitive boosting for classification of imbalanced data," *Pattern Recognit.*, vol. 40, no. 12, pp. 3358–3378, 2007.
[57] K. M. Ting, "A comparative study of cost-sensitive boosting algorithims," *Sch. Comput. Math.*, pp. 983–990, 2000.
[58] W. Fan, S. J. Stolfo, J. Zhang, and P. K. Chan, "AdaCost: Misclassification Cost-Sensitive Boosting," *Proc. Sixt. Int. Conf. Mach. Learn.*, pp. 97–105, 1999.

[59] J. Kittler, M. Hater, and R. P. W. Duin, "Combining classifiers," *Proc. - Int. Conf. Pattern Recognit.*, vol. 20, no. 3, pp. 226–239, 1998.
[60] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.
[61] R. E. Schapire, "The Strength of Weak Learnability (Extended Abstract)," *Mach. Learn.*, vol. 5, pp. 197–227, 1990.
[62] Y. Freund and R. E. Schapire, "A desicion-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, pp. 119–139, 1997.
[63] S. Wang and X. Yao, "Diversity analysis on imbalanced data sets by using ensemble models," *2009 IEEE Symp. Comput. Intell. Data Min.*, pp. 324–331, 2009.
[64] R. Barandela, J. S. Sánchez, and R. M. Valdovinos, "New Applications of Ensembles of Classifiers," *Pattern Anal. Appl.*, vol. 6, no. 3, pp. 245–256, 2003.
[65] N. V. Chawla, A. Lazarevic, L. O. Hall, and K. W. Bowyer, "SMOTEBoost: Improving Prediction of the Minority Class in Boosting," *Proceeding. Knowl. base Discov. Databases.*, pp. 107–119, 2003.
[66] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "RUSBoost: A hybrid approach to alleviating class imbalance," *IEEE Trans. Syst. Man, Cybern. Part ASystems Humans*, vol. 40, no. 1, pp. 185–197, 2010.
[67] H. Guo and H. L. Viktor, "Learning from Imbalanced Data Sets with Boosting and Data Generation : The DataBoost-IM Approach," *ACM SIGKD Explor. Newsl. - Spec. issue Learn. from imbalanced datasets*, vol. 6, no. 1, pp. 30–39, 2004.
[68] X. Wu *et al.*, *Top 10 algorithms in data mining*, vol. 14, no. 1. 2008.
[69] R. E. Schapire and Y. Singer, "Improved boosting algorithms using confidence-rated predictions," *Mach. Learn.*, vol. 37, no. 3, pp. 297–336, 1999.
[70] Z. Sun, Q. Song, X. Zhu, H. Sun, B. Xu, and Y. Zhou, "A novel ensemble method for classifying imbalanced data," *Pattern Recognit.*, vol. 48, no. 5, pp. 1623–1637, 2015.
[71] M. Alibeigi, S. Hashemi, and A. Hamzeh, "Data & Knowledge Engineering DBFS : An effective Density Based Feature Selection scheme for small sample size and high dimensional imbalanced data sets," *DATAK*, vol. 81–82, pp. 67–103, 2012.
[72] R. Rifkin and A. Klautau, "In defense of one-vs-all classification," *J. Mach. Learn. Res.*, vol. 5, no. Jan, pp. 101–141, 2004.
[73] T. Hastie and R. Tibshirani, "Classification by pairwise coupling," *Ann. Stat.*, vol. 26, no. 2, pp. 451–471, 1998.
[74] A. C. Tan, D. Gilbert, and Y. Deville, "Multi-class protein fold classification using a new ensemble machine learning approach.," *Genome Inform.*, vol. 14, no. July, pp. 206–217, 2003.
[75] Y. Sun, M. S. Kamel, and Y. Wang, "Boosting for learning multiple classes with imbalances class distribution," *Proc. - IEEE Int. Conf. Data Mining, ICDM*, pp. 592–602, 2006.
[76] C. Science, D. Dhanalakshmi, A. S. Vijendran, and A. Info, "Adaptive Data Structure Based Oversampling Algorithm for Ordinal Classification," *Indones. J. Electr. Eng. Comput. Sci. Vol.*, vol. 12, no. 3, pp. 1063–1070, 2018.