Journal of Computer Science

Review

A Wide Scale Classification of Class Imbalance Problem and its Solutions: A Systematic Literature Review

[1]Gillala Rekha, [2]Amit Kumar Tyagi and [1]V. Krishna Reddy

[1]*Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, 522502, Andhra Pradesh, India*
[2]*School of Computing Science and Engineering, Vellore Institute of Technology, Chennai Campus, Chennai, 600127, Tamilnadu, India*

Article history
Received: 18-11-2018
Revised: 07-03-2019
Accepted: 01-07-2019

Corresponding Author:
Amit Kumar Tyagi
School of Computing Science and Engineering, Vellore Institute of Technology, Chennai Campus, Chennai, 600127, Tamilnadu, India
Email: amitkrtyagi025@gmail.com

Abstract: In today's world, most of the data (real world) is present in imbalanced form by nature. This is because of not having efficient algorithms to put this data (i.e., generated data by billion of internet-connected devices (IoTs)) in respective format. Imbalanced data poses a great challenge to (both) data mining and machine learning algorithms. The imbalanced dataset consists of a majority class and a minority class, where the majority class takes the lead over the minority class. Generally, several standard learning algorithms assume the balanced class distribution or equal misclassification costs. If prediction is performed by these learning algorithms on imbalanced data, the accuracy will be high for majority classes, i.e., resulting in poor performance. To overcome this problem (or improving accuracy of deision/prediction-making process), data mining and machine learning researchers have addressed the problem of imbalanced data using data-level, algorithmic level and ensemble or hybrid methods. This article presents a systematic literature review and analyze the results of more than 400 research papers published between 2002-2017 (till June 2017), resulting in a broader and elaborate investigation of the literature in this area of research. Note that extension of this article/work will contain till December 2018 research articles, which will be published in June 2019 (now these more papers/articles did not include due to no. of pages/space issues). The systematic analysis of the research literature has focus on the key role of Data Intrinsic Problems in classification, handling the imbalanced data and the techniques used to overcome the skewed distribution. Furthermore, this article reveals patterns, trends and gaps in the existing literature and discusses briefly the next generation research directions in this area.

Keywords: Class Imbalance Problem, Data Mining, Machine Learning, Data-Level Methods, Algorithmic-Level Methods, Hybrid Methods

Introduction

Imbalanced datasets occur in classification problems where in the datasets are not equally categorized. Generally, Machine Learning (ML) or Data Mining (DM) Algorithms assume an equal class distribution for the data. However, this may not be true for real world situations. The distribution of the data for real world applications is skewed in nature, wherein one class (majority class) is represented much more frequently than the other class (minority class). For learning algorithms, this leads to great difficulty, as they are biased towards the majority class. But at the same time, minority classes may generate useful knowledge. The concept of designing a smart system for handling skewed distribution to overcome the bias is known as learning from imbalanced data (Japkowicz and Stephen, 2002). In the past two decades, this problem is widely addressed by the several research communities. The imbalanced data classification has drawn significant attention from academia and industry (Rokach, 2010). An approach for handling the imbalanced training data for classification problem has been addressed by Anand *et al*. (1993), suggesting a modified neural network algorithm. Since 1993, many methods were developed for handling imbalanced data which focusing on balancing the imbalanced data using preprocessing techniques or modifying the existing classifiers. In general, traditional machine learning algorithms assume that the training data sets are well–balanced with an equal misclassification error cost associated to each of the class (Anand *et al*., 1993). The data intrinsic problems corresponds to a classification learning problem in which one class is represented by a large number of samples called as a majority class (negative class) and another class with few samples called as a minority class (positive class). This leads to majority/minority class imbalance for a given data set. Such skewedly distributed datasets represent as class imbalance problems.

Class imbalance problem has attracted attention of researchers in recent years. If the available training sample size of each class is imbalanced, then an established classification model will tend to allocate the testing sample into the majority class. A proper resampling method together with a power classifier is generally employed for dealing with this problem. Multi-classifier ensembles to outperform single classifier in several experiments. The class imbalance problems are generally encountered in real-world data sets like banking datasets, insurance datasets, medical datasets, network datasets like spotting unreliable telecommunication customers (Ezawa *et al*., 1996), detection of oil spills in satellite radar images (Kubat *et al*., 1998), detection of fraudulent telephone calls (Fawcett and Provost, 1996), information retrieval and filtering tasks (Lewis and Catlett, 1994) and so on. Japkowicz and Stephen (2002) presented a survey on class imbalance problems in classification and also focused on the damage encountered when imbalanced data applied to train the traditional classifier algorithm in 2003. However, the different techniques, approaches, algorithms and metrics to handle class imbalance problem were not addressed in their work. Since 2002, several research articles were published in various reputed journals/ conferences with respect to class imbalance problems. To the best of our knowledge, there is no Systematic Literature Review (SLR) available till now on class imbalance problems. So in our SLR, we present the current approaches and trends with respect to the class imbalance problem using different techniques, algorithms and metrics proposed during the last 16 years (2002-2017). We also analyzed the strengths and weaknesses of different approaches which is used for handling class imbalance problem and addressed the current trends related to class imbalance problems.

This work presents a systematic literature review of the state-of-art research on the class imbalanced problem. We conduct a Systematic Literature Review (SLR) on class imbalance problems to identify, taxonomically classify and systematically compare existing research methods and techniques. The major contribution of this SLR includes identifying the primary objectives of data intrinsic problems in classification and specifies the different existing techniques and approaches used in class imbalance problem.

Hence, the remaining work of this article can be organized as: Section 2 presents the research methodology used in this paper. The classification of current approaches for class imbalance problem is described in Section 3. The results of SLR on class imbalance problem are discussed and analyzed in Section 4. Research implications and future directions including threats (with respect to validity) are investigated in Section 5. In last, Section 6 concludes this work in brief.

Research Methodology

A systematic literature review is a means of evaluating and interpreting all available research relevant to a particular research question, topic area, or phenomenon of interest (Brereton *et al*., 2007). The Systematic Literature Review (SLR) methodology aims at providing an unbiased review as much as possible by being auditable and repeatable. We followed a four-step review process (Fig. 1, also Table 1) which includes defining a research question and their motivation, searching for relevant data, extraction of relevant data and assessing the quality of the data (Brereton *et al*., 2007; Jatoth *et al*., 2017).

This SLR gives a systematic understanding for researchers in data mining and machine learning areas and helps to gain on research implications, solutions and future directions. Further, this SLR presents available approaches, techniques and their constraints for the understanding purpose of the practitioners with respect to respective domain.

Defining the Research Questions and their motivations

In order to conduct the review, we discuss (define) several research questions and their motivations as illustrated in Table 1.

Table 1: Research questions and their motivations

Research Questions	Motivations
RQ1: What are the main research motivations behind class imbalanced problems in classification?	To get insight into class imbalance problems
RQ2: Why class imbalance problems are popular?	To make the balancing of the datasets that are skewed-ly distributed in nature.
RQ3: What are the existing methods and techniques applied for class imbalanced problems?	To identify, study, compare and classify the existing methods and techniques that are used in Class imbalanced problem for classification.
RQ4: What are the existing research issues and future scope for class imbalanced problems?	To understand the research undertaken till now and to find the future directions in this field.

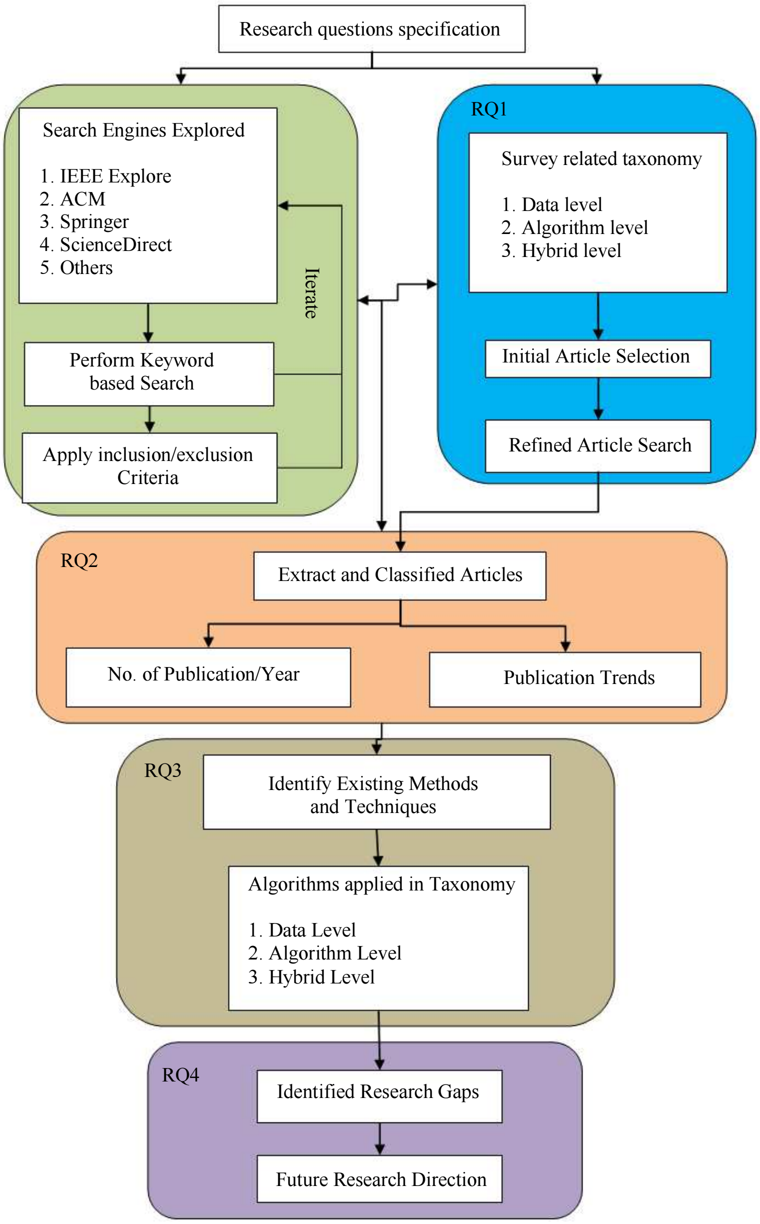

Fig. 1: Overview of Research Methodology (based on (Brereton *et al*., 2007; Jatoth *et al*., 2017))

Searching for Relevant Data

The need for a Systematic Literature Review (SLR) is to identify, classify and compare existing researches in class imbalance problems through a systematic framework. To the best of our knowledge, the class imbalance problems have not been reported using a detailed SLR. The digital databases chosen in this research study/work include IEEE Xplore, ACM, ScienceDirect, SpringerLink and Google Scholar digital libraries. The following search strings are used for searching relavant data (to write this article).

(Class Imbalance Problem for Binary Classification)

AND

(Majority Classes OR Minority Classes OR Skewed Distribution OR Data Intrinsic Problems)

AND

(Sampling methods OR Under-Sampling Methods OR Over-Sampling OR SMOTE methods)

AND

(Data-Level Methods OR Algorithmic Methods OR Hybrid Methods)

AND

(Systematic Literature Review OR Mapping Study OR Systematic Review).

Table 2: List of Online Databases consider for the SLR

S. No	Digital Databases	No. of Papers
1	IEEE	252
2	Science Direct/Elsevier	58
3	Springer	49
4	ACM	11
5	Google Scholar	30
	Total	400

We selected 400 research papers related to class imbalance problems that were published between 2002 and 2017 (Note that from June 2017-December 2018, one more review article will be published in near future as the extension of this work). The different online databases used for extracting the relevant articles are specified in Table 2.

Initial Selection

For initial selection, this work finalized/extracted 400 articles that covers major problems of 'imbalance nature of data' across various digital databases. Then, this work required/checked for the relevant titles for this important study and ignored the remaining papers. This work also ignored the articles from books, thesis, books chapter, white papers, short papers (less than 4 pages) and papers written in non-English languages. Note that this work/article explore all the articles by their titles and abstract by applying the said inclusion and exclusion criteria.

Final Selection

This work further refined the existing search string as mentioned below and selected 281 articles for the study (note that this number will be increased when July 2017-December 2018, many manuscripts which are not covered in this article, will be presented in the extension of this work in near future). This work focused mainly on class imbalance problems in classification techniques.

Title: (Data level methods for Class imbalance problems) Abstract: (Sampling methods OR Under-sampling methods OR Over-sampling Methods OR SMOTE Techniques) Keywords: (Data level OR Class Imbalance Problems OR Sampling Techniques)

AND

Title: (Algorithm level methods for Class imbalance problems) Abstract: (Algorithm Level Classifiers OR Cost-Sensitive Classifiers OR SVM OR Decision Tree OR Fuzzy Techniques OR Neural Networks

OR Co-conventional Neural Network) Keywords: (Algorithm level

OR Cost Sensitive Learning OR Class imbalance problems)

AND

Title: (Ensemble Learning Methods for Class imbalance problems)

Abstract: (Ensemble Methods OR Bagging OR Boosting OR Adaptive Boost OR Hybrid techniques)

Keywords: (Ensemble Methods OR Bagging OR Boosting OR AdaBoost OR Swarm Intelligence Techniques)

Developing and Validating the Review

We developed a procedure for performing the SLR by consulting the several experts (or researchers) who have experience in conducting the systematic review and considered their feedback to improve the scope of research in class imbalance problems, enhance the search strings and filter the criteria related to inclusion and exclusion for the study.

Extractions of Relevant Data

We extracted the relevant data from the digital databases, and discussed as/ in Table 2. The complete description of classification and approaches in class imbalance problems is discussed in Section 3.

Hence, this section discusses the research methodology adopted for class imbalance problem. Now, next section will deal with classification and approaches in class imbalance problems.

Classification and Approaches in Class Imbalance Problems

The class imbalance problems in classification occur when there are significantly lesser samples in one class compared to other class. In recent years, class imbalance problem has attracted the attention of several researchers. Broadly, the class imbalance problem can be addressed at three different levels:

- **Data Level Methods:** At data level, the training dataset is modified to balance the class distribution. These methods apply preprocessing techniques to balance the skewed distribution in the data.
- **Algorithm Level Methods:** These methods directly modify the existing learning algorithms to alleviate the bias towards majority classes and adapt them to mine the data with skewed distribution. These methods provide cost sensitive learning by taking misclassification costs into consideration.
- **Hybrid Methods:** These methods use ensembles of classifiers and also known as ensemble methods. These methods increase the accuracy by training several different classifiers and combine their output to form a single class label

Data Level Methods

In data level, the training dataset is modified to balance the class distribution and make it suitable for a standard learning algorithm. The main aim of data level algorithm is to re-balance the class distribution for the data sets using sampling techniques such as random over-sampling, random under-sampling, and improving sampling techniques, etc. The oversampling and undersampling techniques are the two popular techniques in sampling-based classification for addressing the imbalanced datasets. In the oversampling technique, some samples are added to the minority class to make it balanced when very less information is available for minority class samples. On the other side, in under-sampling technique, some samples of the major class are eliminated to make the dataset balanced.

There are different oversampling and under-sampling techniques that exist in literature such as Random Under-Sampling (RUS) (Xiaoying and Sheng, 2012), Random Over-Sampling (ROS) (Fan and Tang, 2010), Tomek links (Tomek, 1976), Synthetic Minority Over-Sampling Technique SMOTE (Chawla *et al*., 2002), One-Sided Selection (OSS) (Kubat and Matwin, 1997), Condensed Nearest Neighbor Rule (CNN) (Hart, 1968), Neighborhood Cleaning Rule (NCL) (Laurikkala, 2001), SMOTE + Tomek links (Batista *et al*., 2004) etc. Yen and Lee (2009) proposed a cluster-based under-sampling approach where the data form a group/cluster and each cluster has dissimilar characteristics. Chawla *et al*. (2004), the authors presented some of the latest research to address class imbalance problem.

Li *et al*. (2015) used SMOTE technique to balance the microarray data for biological datasets. Xiang and Bai (2013) proposed a re-sampling method in active learning to scarify the true negative rate and to achieve higher true positive rate, which is more important in-person re-identification. In their proposed approach, two re-sampling methods are used to remove samples from the majority class (under-sampling method) and another (over-sampling method) is to add new samples to the minority class. As discussed above and by G.Rekha et al. (2019), we need to give much importance to minority class than majority class. Debowski *et al*. (2012) developed a dynamic sampling framework for multi-class imbalanced data containing many numbers of classes. In their proposed framework, existing sampling techniques such as RUS, ROS and SMOTE are used. Li *et al*. (2014) proposed an oversampling method based on support degree. This technique provides guidance to the people to select minority class samples, to generate new minority class samples. By using support degree, it is possible to identify minority class boundary samples, which in turn produces a number of new samples between the boundary samples and their neighbors and adds the synthetic samples to the original data-set for performing classification.

Further, Dang *et al*. (2015) proposed a novel method, named SPY to solve class imbalance problem. In their method, the label of majority class samples are changed to that of minority class samples in the training data set. The majority class samples are represented as spy samples. As a result, the number of majority class is reduced, whereas, the number of minority class is increased. Li *et al*. (2008) proposed a under-sampling strategy based on Self Organized Map (SOM) clustering to overcome the flaws (in the traditional re-sampling methods) like serious randomness, subjective interference and information loss. Chen *et al*. (2010a) introduced a novel oversampling strategy called Cluster Indexing (CE)-SMOTE based on Cluster Ensemble (CE) to handle imbalanced datasets. In this work, clustering Consistency Index (CI) used to find out the cluster boundaries of minority samples by using k-means algorithm and then oversampled these minority samples to match the original dataset. In this method, CI is calculated for every minority samples and the boundary minority samples with CI lower than the given threshold value are identified. These boundary minority samples are over-sampled to balance the original data set. Koto (2014) proposed three improvements of SMOTE method named as SMOTE-Out, SMOTE-Cosine and Selected-SMOTE. The authors conducted experiments on eighteen different datasets and the results show that the proposed method gave some improvements on balanced accuracy and F1-Score.

Later Van Vlasselaer *et al*. (2013) used SMOTE sampling technique to remove the skewed class distribution in social security fraud. Mahmoudi *et al*. (2014) propose a novel synthetic over-sampling method, called as Diversity and Separable Metrics in Oversampling Technique (DSMOTE). The proposed method includes three steps: (a) Remove abnormal samples from the minority class, (b) select the top 3 minority class samples with respect to a given desired criteria and (c) generating new samples based on those of the selected samples in the previous step. Yeh *et al*. (2016) presented a new over-sampling method to generate synthetic samples for the small minority class data (based on a two-parameter Weibull distribution). Then, Support Vector Machine (SVM) with a polynomial kernel function is used on the available (collected) training dataset to construct the classification model. Zhang *et al*. (2010) developed two strategies to explore the majority class examples ignored by undersampling with the assumption that due to undersampling leading to loss of useful data. In this, in order to achieve good prediction over minority class, they avoid necessary information loss from the majority class (with using both Kmeans algorithm and random sampling approach). To deal with the class imbalance problem, Wang *et al*. (2013a) proposed a new over-sampling technique to generate the synthetic samples for minority class. To balance the data samples, they created a K-Nearest Neighbor (K-NN) graph on the raw positive classes. Then, they build a Minimum Spanning Tree (MST) on the K-Nearest Neighbor graph and generated virtual samples with SMOTE technique on Minimum Spanning Tree to balance the diseases data sets.

Bunkhumpornpat and Subpaiboonkit (2013) proposed a safe level graph method as a guideline tool for selecting an appropriate SMOTE and describes the characteristic of a minority class in an imbalanced dataset. Fu *et al*. (2016) presented an undersampling method based on Principal Component Analysis (PCA) and weighted comprehensive evaluation to improve the dataset's unbalanced condition during the forecast of software fault data. To overcome the problem with the under-sampling method, Jindaluang *et al*. (2014) proposed a cluster-based under-sampling method which uses a clustering algorithm. This algorithm clusters the data in the majority class and selects a number of representative data in many proportions and then combines them with all the data in the minority class as a training set. Ma *et al*. (2011) developed a method Random undersampling Tri-training (RusTri) to handle the class imbalanced in software defect detection data and paid more attention to the data-level approach, i.e., for the enhancement of AUC performance. Kamei *et al*. (2007) experimentally assess the four sampling methods such as random over sampling, SMOTE, RUS and One-Sided Selection (OSS) on four fault-proneness models like Linear Discriminant Analysis (LDA), Logistic Regression analysis (LR), Neural Network and Classification tree by using two module sets of industry legacy software. Note that all four sampling methods enhanced the prediction performance of the Logistic Regression Analysis (LDA) and Logistic Regression (LR) models. Wu and Wang (2009) proposed a cluster-based sampling approach for selecting the representative data as training data to improve the classification accuracy and investigate the effect of under-sampling methods to solve imbalanced class imbalance/distribution problem. Yen *et al*. (2006) presented a spatio-temporal over-sampling method to resolve the problem of class imbalance (in background subtraction). Majumder *et al*. (2016) proposed a novel method, SMOTE and ADAptive SYNthetic Sampling (ADASYN) for balancing the highly imbalanced geometric feature set (extracted from the shoulder pain database). Mathew *et al*. (2015) proposed a Kernel based SMOTE (KSMOTE) method/ technique that creates synthetically minority information in the element space of SVM classifier for class imbalance problem. Pelayo and Dick (2007) explore the use of stratification-based re-sampling approach in software defect prediction. In order to improve the prediction of students final grade different re-sampling techniques including SMOTE, Random Over Sampling (ROS), RUS (Rashu *et al*., 2014) has been used. Shi *et al*. (2016) proposed under-sampling technique to solve the class imbalance problem with (of) the P300 dataset and applied SVM classifier on it, for training the re-sampled this data sets.

Zoric *et al*. (2016) provides an insight on adopting various optimization approaches for SMOTE technique to solve class imbalance problem. Padmaja *et al*. (2007) presents an under-sampling approach called Majority Filter-based Minority Prediction (MFMP) for class imbalance problem. To improve the prediction performance of fault-prone module, Bennin *et al*. (2016) applied over sampling and under sampling approaches and rebalanced both the fault-prone modules and non-fault-prone modules in the training data. Oktavino and Maulidevi (2014) proposed an automatic extraction of information from some particular medium enterprise e-commerce web page with best performance algorithm for text processing and adopted SMOTE technique to balance the data. Das *et al*. (2013a) proposed a novel clustering-based undersampling technique that identifies data regions where minority class samples are embedded deeply inside a majority class.

By removing majority class samples from these regions, ClusBUS preprocesses the data in order to give more importance to the minority class during classification. Farajzadeh-Zanjani *et al*. (2016) proposed a novel sampling techniques based on missing data imputation by means of the Expectation Maximization (EM) and k-NN and applies them to diagnose bearing defects under the class imbalance problems. Researchers proposed an oversampling approach based on the Parzen Window (PW) or Kernel Density Estimation (KDE) (Wang *et al*., 2013a; Bunkhumpornpat and Subpaiboonkit, 2013) for majority class.

Ghazikhani *et al*. (2012a) proposed a non-synthetic oversampling method to handle the class imbalance problem using Support Vector Data Description (SVDD) as a basis for oversampling. The proposed algorithm has two phases. In the first step, SVDD is performed on the minority class to describe it (using a hyper sphere embracing the data), whereas, the second step is oversampling of the support vectors. Hu *et al*. (2009) presented a modified SMOTE for addressing the imbalanced problems based on the SMOTE algorithm. Li *et al*. (2011) applied the over-sampling method called Random- SMOTE for addressing the imbalanced problems. Liu *et al*. (2016a) proposed an improved SMOTE method with K-means algorithm (as an over-sampling method) and RUS method (as under-sampling method) to balance the dataset. Zhou *et al*. (2016) discussed Cost Minimization Oriented SMOTE (CMO-SMOTE) to generate the synthetic minorities that truly help in shift the boundary with specific objectives. The authors also discussed "how the problem of imbalanced learning originates from the Bayesian perspective" and revealed that it is partly triggered by combination of class imbalance and the nature of the class conditional probability density distribution in the data. Blagus and Lusa (2012) implemented SMOTE with different learning algorithms on high-dimensional data (using gene expression data sets). In their proposed method, they used/ applied different learning algorithms like k-NN, LDA methods (Diagonal - DLDA and Quadratic - DQDA), Random Forests (RF), Support Vector Machine (SVM), Prediction Analysis for Microarrays (PAM), Classification And Regression Trees (CART) and Penalized Logistic Regression (PLR) with a quadratic penalty. Al Helal *et al*. (2016) analyzed some of the well established classification algorithms on an imbalanced data set. Re-sampling technique has been used to make the dataset to fit into the feasible region. Then, they applied four different algorithms like Decision Tree (DT) classifer, SVM, K-NN and RF algorithm before and after the resampling a data/ data-set. Mercado-Dıaz *et al*. (2015) presented a comparison of three strategies for managing the imbalance problem such as undersampling, SMOTE and Weighted SVM, in order to analyze the performance of protein function prediction. Mountassir et al. (2012) proposed an under sampling method named as remove similar, remove farthest and remove by clustering to solve unbalanced data sets in supervised sentiment classification in a multi-lingual context. Rosa *et al*. (2014) applied SMOTE to resolve the unbalance between the case and control cases without causing over fitting. Shi *et al*. (2016) used SMOTE to construct new samples through KNN to get a balanced sample. Cateni *et al*. (2011) proposed a hybrid approach (combination of novel oversampling and undersampling procedures) for balancing the original dataset to improve the classification accuracy. Chen *et al*. (2010b) proposed a novel hybrid resampling technique to improve the classification performance using differential evolution (i.e., over-sampling of minority class). Lin *et al*. (2014) proposed a novel method to solve the class imbalance problem. In this work (Drummond and Holte, 2003), authors show that using C4.5 with undersampling establishes a reasonable standard for solving class imbalance problem and also recommended that the least cost classifier is a part of standard, i.e., it provides better results than undersampling for relatively modest costs. Estabrooks *et al*. (2004) the author combined different expressions of the re-sampling approaches to get effective solution for imbalanced data sets.

Table 3: Classification and approaches for data level methods

Author's Name	Techniques used	Applied Algorithms	Metrices used	Data Sets used
Sun and Liu (2016)	SMOTE-NCL	KNN, SVM, DT (C4.5), and Nave Bayes(NB)	AUC (Area Under Curve)	KDDCUP99 data set.
Mustafa and Li (2017)	Maximum Distance based SMOTE	KNN based Fuzzy-rough set	Accuracy	UCI repository
Fan *et al*. (2014)	Re-sampling using the Boundary Ratio	SVM	Recall, Precision, F-measure, G-mean	UCI repository
Chen *et al*. (2010b)	Differential Evolution Clustering	SVM	F-measure, AUC	UCI repository
	Hybrid Resampling SVM (DEC-SVM)	boosted GMM	AUC	UCI repository
Ramentol *et al*. (2015)	Fuzzy Rough NN (FRNN) classifier			
Melillo *et al*. (2013)	CART	C4.5, RF	AUC, Sensivity, Specificity, Precision, F-measure, Accuracy	UCI repository
Cao *et al*. (2013a)	Enhanced Structure Preserving Oversampling (ESPO)	SVM	F-measure, G-Mean, Percision, Recall	UCR time series repository
Babar and Ade (2016)	MLPUS	SVM, MLP, CART	Precision, Recall, G-mean, Accuracy	UCI repository
Cao *et al*. (2011)	SPO (SPO)	SVM	F-measure, G-Mean	UCR time series repository
Cao *et al*. (2014)	ROS	SVM	F-measure	UCR time series repository
Cao and Shen (2016)	Hybrid re-sampling, Twin SVM (TWSVM)	SVM	Fmeasure, G Mean	UCI repository
Chairi *et al*. (2014)	RUC, Sample Selection Technique	SVM	Accuracy	KDD repository
Chen *et al*. (2010a)	CE-SMOTE	C4.5	F-measure, G-Mean, Recall, Precision	UCI repository
Dai (2015)	Inverse RUS	SVM	AUC	CATH databases
Dang *et al*. (2015)	SPY	SVM, K-NN and Radial Function (RF)	Sensitivity, G-Mean	UCI repository
Ghosh *et al*. (2016)	RUS, ROS	SVM, LR	Precision, Recall, F-measure	event sequence data
Makrehchi and Kamel (2007)	Feature Selection	Rocchio Text Classifier	F-measure	real FOAF database
Peng *et al*. (2016a)	RUS	Data Gravitation based Classification (DGC)	Confusion matrix, AUC, Sensitivity, Specificity	KEEL repository
Qing *et al*. (2015)	PCBoost Algorithm	SVM	Accuracy	Material consumption forecasting data set
Zhang *et al*. (2017)	Spatio-Temporal Over-Sampling	NB	F-measure	BMC database
Zhong *et al*. (2009)	RUS, ROS, SMOTE	C4.5, NN	Sensitivity, Specificity, Precision, G-Mean	P2P data set
Zhou *et al*. (2016)	CMOSMOTE	NB	Recall, Precision, G-Mean, F-measure	KEEL repository
Vorraboot *et al*. (2012)	Back Propagation NN (BPNN)	BPNN	TP rate,G- Means, F-measure	UCI repository
Xiang and Bai (2013)	SMOTE	SVM	Accuracy	ETHZ dataset
Van Vlasselaer *et al*. (2013)	SMOTE	RF, LR, NB	AUC, Precision	Belgian social security data set

Table 4: Classification and Approaches for Data Level Methods (Cont...)

Author's Name	**Techniques used**	**Applied Algorithms**	**Metrices used**	**Data Sets used**
Debowski *et al*. (2012)	RUS, ROS, SMOTE	ANN, KNN, C4.5, NB	Accuracy, Recall, G-Mean, F-measure	UCI repository
Rahman *et al*. (2015)	ROS, RUS	ADABOOST, K-NN, SVM-RBF, LR	Accuracy, Sensitivity, Specificity, Precision	Cardiac surgery data set
Xu *et al*. (2013)	RUS, ROS, SMOTE	SVM	Accuracy	BCI Data set
Ren *et al*. (2011)	SMOTE	ANN	Recall, Precision, Specificity, F-measure, Accuracy	DDSM database
Pengfei *et al*. (2014)	Boderline-SMOTE	C4.5	AUC	UCI repository
Triguero *et al*. (2017)	RUS	CHC evolutionary	AUC, G-Mean Algorithm	Evolutionary Big Data Competition ECBDL 14
Li and Deng (2016)	ROS	SVM	F-measure, AUC	RAF-DB
Chen *et al*. (2009)	Sampling Method	C4.5, k-NN	AUC	DIANE database
Lopez *et al*. (2014)	DOBSCV Strategy	C4.5, k-NN, SVM	Accuracy	KEEL repository
Hunt *et al*. (2010)	Sampling Method	GP	Accuracy	UCI repository
Garcia *et al*. (2006)	Evolutionary Under-Sampling (EUS)	k-NN	G-Mean	UCI repository
Garcia *et al*. (2009)	RUS, Evolutionary Algorithm	C4.5	Error rate, Accuracy, G-Mean	UCI repository
Hu *et al*. (2016a)	Dynamic Query-Driven Sample Rescaling	SVM	Specificity, Sensitivity, Precision, Accuracy, Mathews Correlation Coefficient	NUC5tst, nonNUC
Yap *et al*. (2013)	ROS, RUS, Bagging, Boosting	CART, C5, CHAID	Accuracy, Sensitivity, Specificity, Precision	Cardiac surgery data set
Wong *et al*. (2014)	CHC Algorithm	SVM	F-measure, AUC	UCI repository
Wang *et al*. (2014)	SMOTE, PSO	LR, C5, 1-NN	Accuracy, Sensitivity, Specificity, G-Mean	Breast cancer patients
Bader-El-Den *et al*. (2016)	TempC	C4.5	F-measure, G-Mean	KEEL, UCI repository
Saez *et al*. (2015)	SMOTE Iterative-Partitioning Filter	C4.5, PART, KNN, SVM	AUC	KEEL repository
Bhagat and Patil (2015)	SMOTE	RF	Sensitivity, Specificity, Precision, G-Mean, F-measure	UCI repository
Del Rıo *et al*. (2014)	RUS, ROS	RF	G-Mean, F-measure	UCI repository
Chakraborty *et al*. (2017)	Ensemble Learning	Back-Propagation MLP (BPMLP) and RBFNN	Accuracy	UCI repository
Chetchotsak *et al*. (2015)	Growing ring SOM	BPNN, SVM	Precision, Recall, F-measure	UCI repository
Chira and Lemnaru (2015)	EA, Cost-sensitive learning	C4.5	AUC, Accurracy	Benchmark datasets
Yen and Lee (2009)	Cluster-based RUS	ANN	Precision, Recall, F-measure	Census-income database
Das *et al*. (2012)	SMOTE	C4.5, SMO, LogitBoost(LB)	AUC, Accuracy, G-Accuracy	Sensor Events Data Set
Dıez-Pastor *et al*. (2015a)	Ensemble-based Methods	DN+Ba	AUC, F-measure, G-Mean	KEEL repository
Yoon and Kwek (2007)	Clustering, ANN	ANN	Recall, Precision, F-measure	UCI repository
Ryu *et al*. (2016)	SVM	SVM	AUC, Harmonic-Measure	PROMISE repository
Fan *et al*. (2016)	One-sided Dynamic Undersampling Technique	NN	AUC, G-Mean	KEEL repository
Farajzadeh-Zanjani *et al*. (2016)	EMI-OS, kNNI-OS	C4.5	F-measure	CWRU bearing data center
FernaNdez *et al*. (2013)	SMOTE + ENN	DT, SVM, Instance-based learning	Accuracy	KEEL repository

Table 5: Classification and approaches for data level Methods (cont...)

Author's Name	**Techniques used**	**Applied Algorithms**	**Metrices used**	**Data Sets used**
Garcıa *et al*. (2012)	RUS, ROS	k-NN, MLP, SVM, NB, C4.5, RBF	G-Mean	UCI repository
Hamdi *et al*. (2010)	Under-sampling with SOM	DT	AUC	UCI repository
Hinojosa *et al*. (2015)	SMOTE + Tomek link	Iterative fuzzy classification Rules Learning, Multi-Objective EA	AUC	KEEL repository
Hosseinzadeh and Eftekhari (2015b)	Fuzzy-based Undersampling	DT	AUC	KEEL repository
Hu *et al*. (2016b)	SMOTE	DT, AdaBoost, RF	Specificity, AUC	white wines data set
Yuan and Ma (2012)	SMOTE	AdaBoost, GA	G-Mean, Standard Deviation	UCI repository
Liang *et al*. (2014)	RUS	SVM, k-NN, DT, NB, MLP	AUC	UCI repository
Jin *et al*. (2015)	SMOTE	SVM	Precision, Recall, G-Mean, F-measure, Accuracy	1 yuan RMB
Kim *et al*. (2016)	Undersampling using GA	ANN	Sensitivity, Specificity, G-Mean, AUC, H- Measure	Korean Manufacturing Firms, Bankruptcy Data
Bunkhumpornpat and Sinapiromsaran (2017)	Density-based Under-Sampling Technique	C4.5, MLP, RIPPER, NB, k-NN, SVM, LLR, RF	F-measure, AUC	UCI repository
Lim *et al*. (2017)	Cluster-based Synthetic Oversampling with GA	DT	AUC, F-measure	UCI repository
Ma *et al*. (2015)	Partial Least Squares	Kernel based Asymmetric Classifier	AUC	UCI repository
Martin-Diaz *et al*. (2017)	SMOTE	AdaBoost (NB, CART)	AUC, Recall, Precision, Specificity, Accuracy	Vibration and Current Signals

Table 5: Continue

Moreo *et al*. (2016)	Distributional ROS	Linear-kernel SVM	F-measure	Reuters-21578, OHSUMED-S, RCV1-v2.
Napieral and Stefanowski (2015)	Expert Knowledge	ABMODLEM (new argument Based Rule Induction Algorithm)	Accuracy	UCI repository
Park *et al*. (2015)	Decision Boundary Focused Under-Sampling	Gradient Boosting Classifer	Accuracy	YOOCHOOSE
Prachuabsupakij and Soonthornphisaj (2014)	Clustering, SMOTE Sampling	DT	F-measure, AUC	Real dataset
Naganjaneyulu and Kuppa (2013)	Intelligent Under-Sampling (CIL-IUS)	C4.5, BPNN	AUC, F-measure, Precision	UCI repository
Sain and Purnami (2015)	SMOTE, Tomek Links	SVM	AUC, G-Mean, F-measure	UCI repository
Chen *et al*. (2016)	ERUS	C4.5	Recall, G-Mean, AUC	PROMISE Repository
Lee *et al*. (2015)	Modified SMOTE	SVM	Precision, Recall, G-Mean	UCI repository
Ha and Lee (2016)	GA based under-sampling	k-NN, LR, CART, NB	AUC, G-Mean, F-measure	KEEL repository
Purnami and Trapsilasiwi (2017)	SMOTE	SVM	Accuracy	Breast cancer dataset
Zhang and Li (2014)	ROS	C4.5, NB, k-NN	F-measure, G-Mean, AUC	MLRepository
Zhang *et al*. (2012)	Hybrid Sampling Algorithm	SVM, ANN	Accuracy, Sensitivity, Specificity,F-measure, AUC	–
Zhang *et al*. (2010)	Cluster-based Majority Under-Sampling	NB	Recall, Precision, F-measure, G-Mean	UCI repository
Wong *et al*. (2013)	SMOTE and EA	C4.5	Precision, Recall, FMeasure, AUC	UCI repository

Table 6: Classification and Approaches for Data Level Methods (cont...)

Author's Name	Techniques used	Applied Algorithms	Metrices used	Data Sets used
Vorraboot *et al*. (2015)	Hybrid Algorithm	Modified KLM, Modified DBSCAN, RBF	F-measure, G-Mean	UCI and KEEL repository
Varassin *et al*. (2013)	Cluster based Under-Sampling	NB, ADTree, SVM	F-measure	DGSplicer dataset
Thai-Nghe *et al*. (2010)	Sampling Methods	SVM	G-Mean	UCI repository
Li *et al*. (2014)	SDSMOTE	C4.5	F-measure, G-Mean, AUC	UCI Repository, JM1 is from NASA standard data-sets
Cao and Zhai (2015)	Hybrid Re-sampling	SVM	AUC	UCI repository
Ceballes-Serrano *et al*. (2012)	SwarmBoost	SwarmBoost	Sensitivity, Specificity, G-Mean	UCI repository
El-Ghamrawy (2016)	Hybrid Sampling Approach	SVM, KNN	F-measure, ROC, AUC	Real Medical Imbalance Data
Jiang *et al*. (2013)	SBNC-Hybrid Sampling Technique	NB	Misclassification Cost	UCI repository
Koto (2014)	SMOTE Out, SMOTE-Cosine, Selected-SMOTE	SVM	F-measure	UCI repository
Li *et al*. (2008)	Sampling Strategy based on SOM	SVM	F-measure, Precision, Recall	UCI repository
Wang *et al*. (2018)	Sampling methods	Multi-Scale Rotation-Invariant model	Precision, Recall, F-measure	Public Interstitial Lung Disease database
Mahmoudi *et al*. (2014)	DSMOTE	SVM	Accuracy, F-measure, Recall, Precision, G-Mean	IUMS and UCI repository
Oktavino and Maulidevi (2014)	SMOTE	SVM, NB, DT	Accuracy	Indonesia e-commerce
Ruangthong and Jaiyen (2015)	SMOTE	RF, J48	Sensitivity, Specificity, Accuracy	UCI repository
Sarakit *et al*. (2010)	SMOTE	Multinomial NB, DT, SVM	Accuracy	Thai YouTube video clips
Yeh *et al*. (2016)	Over-Sampling Method	SVM	Accuracy, G-Mean, F-measure	UCI repository
Pal and Paul (2017)	SMOTE	Boosted GMM	Accuracy, Precision, Recall, F-measure	KEEL data repository
Hu and Zhang (2013)	Hybrid Approach	DT, NB, KNN, SVM	AUC, F-measure	UCI repository
Xiaoying and Sheng (2012)	RUS, ROS	RF, BN	Precision, Recall, F-measure, V-Measure, AUC	UCI repository
Wang *et al*. (2013a)	ROS	SVM	Sensitivity, Specificity, Accuracy, G-Mean	UCI repository
Bunkhumpornpat and Subpaiboonkit 2013)	Safe Level Graph, SMOTE	NB, C4.5, KNN, RIPPER	AUC, F-measure	UCI Repository
Cateni *et al*. (2011)	RUS	SVM, CART	Accuracy	UCI repository
Fu *et al*. (2016)	PCA	-	Accuracy	NASA-Public fault datasets
Jindaluang *et al*. (2014)	Cluster-based Under-Sampling Method	k-NN, C4.5	Accuracy, Precision, Recall	UCI repository
Kamei *et al*. (2007)	ROS, SMOTE, RUS	LDA, LRA, NN	Recall, Precision, F-measure	Industry Legacy Software
Li *et al*. (2015b)	SMOTE-PSO, SMOTE-BAT	NN, DT	Accuracy	30 Highly skewed sets
Majumder *et al*. (2016)	SMOTE, ADASYN	GMM	Precision, Recall	The UNBC-McMaster Shoulder Pain Expression Archive database
Mathew *et al*. (2015)	Kernel-Based SMOTE	SVM	G-Mean	KEEL repositor

Further, Batista *et al*. (2004) proposed two methods to handle class imbalance problem such as SMOTE + Tomek and SMOTE + ENN and over-sampling method with data cleaning methods in order to produce better-defined class clusters for data sets (with a small number of positive examples). Further, Barandela *et al*. (2004) applied several re-sizing techniques for handling the imbalance issue. Radivojac *et al*. (2004) consider the problem of classification in noisy, high-dimensional and class-imbalanced protein datasets by three-stage machine learning framework consisting of a feature selection stage, a method addressing noise and class-imbalance and a method for combining biologically related tasks through a prior-knowledge based clustering. The list of classification and approaches applied at data level, are included (or presented) in Tables 3-8.

Algorithm Level Methods

In these methods, existing learning algorithms are modified to improve the bias towards majority classes. These dedicated algorithms directly learn from the data using skewed distribution. The modifications include adjusting the costs of the various classes so as to offset the class imbalance, adjusting the probabilistic estimate at the tree leaf when working with decision trees, adjusting the decision threshold and recognition based (i.e., learning from one class) learning. These algorithm level solutions are applied to several classifications techniques including threshold learning and one-class learning (Askan and Sayın, 2014; FernaNdez *et al*., 2013; Raskutti and Kowalczyk, 2004; Van Hulse *et al*., 2007; Cohen, 1995) and cost-sensitive analysis (Domingos, 1999; Elkan, 2001; Margineantu, 2002).

Further, Qiu *et al*. (2015) developed a new adaptive method based on Differential Evolution (DE) for class-imbalanced cost-sensitive learning. In this, they explore the potential information contained in the majority classes and creates an optimal subset with low misclassification costs.

Table 7: Classification and Approaches for Data Level Methods (cont...)

Author's Name	Techniques used	Applied Algorithms	Metrices used	Data Sets used
Wu and Wang (2009)	Hybrid-Sampling Technique	DT	AUC, F-measure	KEEL repository
Yen *et al*. (2006)	Cluster-Based Under-Sampling	BNN	F-measure	DSiE1 OD20
Zhang *et al*. (2014a)	Spatio-Temporal Over-Sampling Method	Kernel Density Estimation	Cost Matrix	CDW2012 database
Zhang *et al*. (2015)	Hybrid Method	DT	Precision, Recall, Accuracy, F-measure, AUC, G-Mean	Tans Hotel reviews data set
Mustafa *et al*. (2015)	Hybrid Method	CART	G-Mean, F-measure, AUC	KEEL repository
Gao *et al*. (2013)	RUSBoost	NB, SVM, MLP, LR, KNN	Accuracy	PROMISE repository
Padmaja *et al*. (2008)	Majority Filter-based minority prediction	DT, KNN, NB, RBF	Accuracy, Recall, Precision, F-measure	UCI repository
Bennin *et al*. (2016)	RUS, ROS, SMOTE, PSMOTE	LR, RVM, k-NN, DT, CART, RF	Normalised Popt	OSS project datasets
Padmaja *et al*. (2007)	SMOTE, RUS	C4.5, NB, k-NN, RBFN	Cost matrix	Insurance dataset
Pelayo and Dick (2007)	SMOTE	C4.5	G-Mean	PROMISE repository
Rashu *et al*. (2014)	SMOTE, ROS, RUS	DT, NB, NN	Accuracy, Precision, Recall, F-measure	Student's records collected from North South University
Sanguanmak and Hanskunatai (2016a)	Hybrid-Sampling Technique	DT	F-measure, AUC	KEEL repository
Chen and He (2009)	SERA	BBagging	Precision, Recall, F-measure, G-Mean	KDD cup 1999 network intrusion dataset
Das *et al*. (2013a)	Clustering-based Undersampline	C4.5, KNN, SVM, NB	TP, FP, AUC, G-Mean	http://ailab.wsu.edu/casas/datasets/prompting.zip
Gao *et al*. (2012)	ROS	RBF	AUC, G-Mean, F-measure	UCI repository
Ghazikhani *et al*. (2012a)	SVM	KNN	F-measure, Recall, Precision	UCI repository
Hu *et al*. (2009)	MSMOTE	C4.5, AdaBoost	Precision, Recall, F-measure	UCI repository
Li *et al*. (2011)	Random- SMOTE	LR	Accuracy	UCI repository
Liu *et al*. (2016a)	SMOTE with K-means	NB	Accuracy, F-measure, G-Mean, Precision, Recall	IPTV datset
Zughrat *et al*. (2015)	Bootstrapping-based Over-Sampling and Under-sampling	Iterative Fuzzy SVM	Accuracy, Sensitivity, Specificity	Rail Manufacturing data from the production line of Tata Steel Europe
Tan and Cai (2010)	AHC Oversampling Method	SVM	Accuracy, G-Mean	China earthquake data center
Prachuabsupakij and Doungpaisan (2016)	SMOTE	NB, DT	Precision, Recall, F-measure, G-Mean	Real-World dataset
He *et al*. (2005)	C-SMOTE	C4.5	F-measure, Recall, Precision	UCI repository

Table 8: Classification and Approaches for Data Level (cont...)

Author's Name	Techniques used	Applied Algorithms	Metrices used	Data Sets used
Demidova and Klyueva (2017)	SMOTE	SVM	Accuracy, Sensitivity, Specificity, F-measure	UCI repository
Ai *et al*. (2016)	Shaped-based Oversampling	C4.5	AUC	KEEL repository
Batuwita and Palade (2010)	ROS	SVM	G-Mean	UCI repository
Blagus and Lusa (2012)	SMOTE	NN, DT	Accuracy, AUC, G-Mean	Breast cancer gene expression data set
Cieslak and Chawla (2008)	RUS, SMOTE	C4.5, SVM	AUC	UCI repository
Das *et al*. (2013b)	wRACOG (wrapper-based Rapidly Converging Gibbs sampler)	DT, SVM, KNN, LR	Sensitivity, G-Mean, ROC	UCI repository
Garcia and Herrera (2008)	EUSTSS	C4.5	G-Mean	UCI repository
Ghazikhani *et al*. (2012b)	Wrapper-based ROS	Fisher Linear Discriminant (FLD), KNN	F-measure, Recall	UCI repository
Al Helal *et al*. (2016)	SMOTE	DT, SVM, KNN, RF	G-Mean	UCI repository
Huang *et al*. (2012)	Random Subspace Method Emsemble method	KNN, Gaussian Classifier (GC), SVM, DT, RF, RSM, AdaBoost	Accuracy	UCI repository
Kocyigit and Seker (2012)	Dynamic Under-Sampling	SVM, LDA, NB	Accuracy, Sensitivity, Specificity, Precision, G-Mean, F-measure	UCI repository
Mercado-Dıaz *et al*. (2015)	SMOTE	SVM	Sensitivity, Specificity, G-Mean	Gene Ontology (GO) consortium
Mountassir *et al*. (2012)	RUS	NB, SVM, KNN	G-Mean	http://www.aljazeera.net
Rosa *et al*. (2014)	SMOTE	MLP, RBF, SVM	AUC	HIV Resistance Drug Database
Sanguanmak and Hanskunatai (2016b)	Hybrid re-Sampling Technique	DT	AUC, F-measure	KEEL repository
Shi *et al*. (2016)	SMOTE	ID3	Precision, Recall, F-measure	Group events territory of China
Thai-Nghe *et al*. (2009)	ROS	DT, BN, SVM	Precision, AUC and F-measure	UCI repository
Wei *et al*. (2014)	Hybrid Approach	C4.5	F-measure, Accuracy	UNIBS dataset
Yang and Gao (2012)	Active Undersampling	BP	F-measure, G-Mean	UCI repository
Zhang and Wang (2011)	Distribution-SMOTE	LR, BP, C5, SVM	Recall, F-measure	UCI repository
Lin *et al*. (2014)	ROS	Case based reasoning	Sensitivity, Specificity, Accuracy	Taiwan University Hospital
Wang *et al*. (2012)	Adaptive Over-Sampling	Cost-sensitive SVM	Sensitivity, Specificity, G-Mean	UCI repository
Dang *et al*. (2013)	Safe-SMOTE	SVM	Sensivity, Specificity	UCI repository
Li *et al*. (2015a)	SMOTE	SVM, KNN, BN, C4.5	Accuracy, AUC	microarray datasets
Qazi and Raza (2012)	SMOTE	C4.5, bagging and SMO	Cost Matrix	KDD Cup 99 dataset
Chawla *et al*. (2002)	SMOTE	C4.5, Ripper and NB	AUC	UCI repository
Drummond and Holte (2003)	Sampling Methods	C4.5	Cost Matrix	UCI repository
Estabrooks *et al*. (2004)	Sampling methods	C4.5	Recall, Precision and F-measure	UCI repository
Batista *et al*. (2004)	ROS + Data Cleaning	C4.5	AUC	UCI repository
Prati *et al*. (2004)	SMOTE + Tomek Links and SMOTE + ENN.	C4.5	AUC and EC	UCI repository
Barandela *et al*. (2004)	RUS, ROS	k-NN	G-Mean	UCI repository
Radivojac *et al*. (2004)	Fishers Permutation Test	LR, DT, k-NN	AUC	Real dataset

Bhowan *et al*. (2009) proposed a Genetic Programming (GP) approach to develop classifiers with high and reasonably balanced minority and majority class accuracy. Kamal *et al*. (2009) proposed three filtering techniques (Higher Weight (HW), Differential Minority Repeat (DMR) and Balanced Minority Repeat (BMR)), to identify important features from biased data collections. Ruangthong and Jaiyen (2015) proposed Rotation Forest-J48 with SMOTE to solve the classification of imbalanced data set. Further, Wong *et al*. (2014) proposed an evolutionary algorithm based an under-sampling method using fuzzy logic for sampling the majority classes and also to reduce the data size. Li *et al*. (2015b) uses meta heuristic optimization algorithms (such as Bat Algorithm (BA) and Particle Swarm Optimization algorithm (PSO)) to optimize the selection for improving the performance of classifiers for imbalanced data. Padmaja *et al*. (2008) proposed a new method for fraud detection, that uses extreme outlier elimination using k-Reverse NN (kRNN) approach.

Zhang *et al*. (2015) proposed a method for unbalanced sentiment classification that combines unbalanced classification method and ensemble learning technique. They applied hybrid method which integrates three different methods such as under-sampling, bootstrap re-sampling and random feature selection to process the data set and trained it using decision tree (as base classifier). Fan *et al*. (2014) suggested that the re-sampling methods are unnecessary (i.e., irrelevant), if the data of each class falls close to the boundary is balanced. They observed that by using standard classifier such as SVM can receive better performance than the re-sampling methods such as SMOTE and RUS. Further, Chen and He (2009) proposed the SElectively Recursive Approach (SERA) framework to address the imbalanced stream data. The Radial Basis Function (RBF) classifier proposed in (Sarakit *et al*., 2010; Gao *et al*., 2012), is applied to rebalanced data set. Hosseinzadeh and Eftekhari (2015a) discussed a fuzzy C-means clustering and Random Forest (RF) methods to construct a high performance classifier by utilizing SMOTE oversampling algorithm for balancing data samples. The authors proposed Simple Hybrid Sampling Approaches (SHSA) to reduce the impact of imbalanced data (i.e., in a data-set). In SHSA, the distance is calculated by using Mahalanobis distance instead of Euclidian distance. Gazzah *et al*. (2015) presented a hybrid approach for training imbalanced datasets using under-sampling the majority (negative) class and over-sampling the minority (positive) class. Xiaoying and Sheng (2012) designed a general rule to select sampling strategy and V-measure metric by considering minority datasets. The benefit of V-measure is paying more attention to accuracy of minority class without hurting more the accurateness of majority classes. Later, Effendy and Baizal (2014) provide combination of sampling techniques and weighted random forest to improve the customer churn prediction model on a sample dataset from a telecommunication industry (situated in Indonesia). Lessmann (2004) applied SVM to solve class imbalance problem. To solve the class imbalance problem using SVM, Shin and Cho (2003) applied informative sampling approach using different costs for different classes and also updated distance of decision boundary. To overcome this problem, Cohen *et al*. (2004) investigate one-class SVMs which can be trained two different classes (based on examples received from a single class) and propose an improvement of oneclass SVMs via a conformal kernel transformation. The impact of both weighted imbalance compensation and sampling techniques are considered and extended the balancing by learning only from one class and ignoring the other (Raskutti and Kowalczyk, 2004). To solve class imbalance problem, Cohen *et al*. (2003) investigate a support vector algorithm in which asymmetrical margins are tuned to improve recognition of rare positive cases. Further, Wu and Chang (2004) proposed a remedy to adjust the class boundary by modifying the kernel matrix, according to the imbalanced data distribution using SVM. The work discussed in (Phua *et al*., 2004) use a single meta-classifier (stacking) to choose the best base classifiers and then combine these base classifiers predictions (bagging) to improve cost savings (stacking-bagging). Phua *et al*. (2004) proposed a new methodology for building decision trees using Consolidated Tree Construction (CTC) algorithm based on resampling techniques. The list of classification and approaches for algorithm level are presented in Table 9-11.

Ensemble (or) Hybrid Methods

Hybrid methods using ensemble algorithms, aim to improve the performance by inducing several classifiers and combine them to obtain a new classifier. AdaBoost (Wang and Yao, 2012), SMOTEBoost (Kaur *et al*., 2016), RUSBoost (Gao *et al*., 2013), DataBoost-IM (Guo and Viktor, 2004), Bagging (Wang and Yao, 2009), OverBagging (Wang and Yao, 2009) and SMOTEBagging (Kaur *et al*., 2016) are some of the popular techniques employed to handle (or solve) class imbalance problem.

Table 9: Classification and approaches for algorithm level

Author's Name	Techniques used	Applied Algorithms	Metrices used	Data Set used
Huang *et al*. (2013)	Cost Sensitive Approach	DT	Macro-average Accuracy	100 volunteers in Beijing, China.
Hu (2012)	L-GEM based RBFNN Classifier	RBFNN	Accuracy	UCI Repository
de Souza *et al*. (2016)	SVM	SVM	Accuracy	-
Tan *et al*. (2015)	Fuzzy Adaptive Resonance Theory (Fuzzy ARTMAP)	Fuzzy	CCR, GMean	UCI repository
Wang *et al*. (2015)	OOB, UOB mode	-	G-Mean	real-world applications
Boonchuay *et al*. (2011)	Modified C4.5	C4.5	Precision, Recall, Specificity	UCI and Statlog repository
Chen *et al*. (2010c)	Cost Sensitive Learning	GA-LVQ	Cost Matrix	Financial statements of French companies
Cheng and Wu (2016)	Weighted Features Cost-Sensitive SVM	SVM	Accuracy, G-Mean, AUC	UCI repository
Duhaney *et al*. (2012)	Cost Sensitive Learning	LR, DT, NB	Arithmetic Mean	Ocean turbine dynamometer
Fan and Tang (2010)	Maximum AUC Linear Classifier	Linear Classifier	AUC	UCI Repository
Haldankar and Bhowmick (2016)	Cost Sensitive Learning	DT, NB, SVM	AUC	UCI repository, Statlog German credit data set
Nikpour *et al*. (2017)	Cost Sensitive Learning	Gravitational fixed radius NN	G-Mean	KEEL repository
Tayal *et al*. (2015)	RankSVM	SVM	Optimized ROC	KDD Cup 1999 dataset
Tan *et al*. (2011)	Fuzzy ARTMAP (FAM)	FAM, Fuzzy ARTMAP Dynamic Decay Adjustment (FAMDDA)	G-Mean	Semiconductor industry
Wang *et al*. (2013b)	Cost Sensitive Learning	Hypernetworks	G-Mean, F-measure, AUC	UCI repository

Table 9: Continue

Xia and Zhang (2014)	Cost Sensitive Learning	SVM	Accuracy	UCI repository, the MNIST database of handwritten digits and the UMIST Face Database
Yongxiong and Li (2014)	Cost Sensitive Learning	C4.5, SVM, KNN, Probabilistic NNa	AUC	319 HVAC duct images
Yu *et al*. (2007)	Cost Sensitive Learning	SVM	Sensitivity, Specificity	UCI repository
Zhao and Shrivastava (2013)	Cost Sensitive Learning	SVM	Sensitivity, Specificity	UCI repository
Ren *et al*. (2011)	FORESTEXTER	SVM	AUC	Reuters-21578, Ohsumed and imbalanced 20 newsgroup
Napierala and Stefanowski (2012)	Bottom-up induction of Rules And Cases for Imbalanced Data (BRACID)	BRACID	Sensitivity, G-mean, F-measure	UCI repository
Shilaskar *et al*. (2017)	Cost Sensitive Learning	SVM, PSO	Accuracy, Sensitivity, Specificity, AUC, Precision, F-measure	Vani,Thyroid, PdA, Cleveland, Audiology, Vertigo, SVD Dataset
Lopez *et al*. (2010)	Cost-Sensitive Learning	Fuzzy Rule Based Classification Systems	AUC	KEEL repository
Lopez *et al*. (2013)	Cost Sensitive Learning	FARC-HD Classifier	AUC	KEEL repository
Alejo *et al*. (2013)	Hybrid Modified Back-Propagation	MLP, NN	AUC, G-Mean, Accurary	Remote sensing datasets: Cayo, Feltwell Satimage and Segment

Table 10: Classification and Approaches for Algorithm Level Methods (Cont....)

Author's Name	**Techniques used**	**Applied Algorithms**	**Metrices used**	**Data Set used**
Fergani and Clavier (2013)	C-SVM	SVM	Accuracy	Openly datasets gathered from three houses
Yala *et al*. (2014)	Modified SVM	SVM	Accuracy	Wireless sensor network
Acskan and Sayın (2014)	SVM	SVM	Specificity, Sensitivity, G-Means	UCI repository
Xu *et al*. (2016a)	Cost Sensitive Learning	LR	G-Mean, F-measure	UCI repository, Mammography dataset
Do *et al*. (2014)	SVM with Stochastic Gradient Descent	SVM	Accuracy	ImageNet 10 and 100, ILSVRC 2010
Bhowan *et al*. (2009)	Cost Adjustment	GP	Accuracy	UCI repository
Qiu *et al*. (2015)	Cost Sensitive Learning	BN Classifiers	Accuracy	KEEL repository
Effendy and Baizal (2014)	Cost Sensitive-Learning	DT	F-measure, Precision, Recall	Customer behavior profile data
Soltani *et al*. (2011)	Cluster-based Undersampling	Hierarchical Clustering	Accuracy, Sensitivity, Specificity, G-Mean	UCI repository
Thach *et al*. (2008)	Cost Senstive Learning	XCS classifier	Precision, Recall, F-measure, G-Mean	UCI repository
Al-Rifaie and Alhakbani (2016)	Hybrid Method	SVM	Accuracy, AUC	UCI repository
Wang *et al*. (2016)	Cost Senstive Learning	ID3, C4.5, CART	Accuracy, F-measure, G-mean, Precision	KPI dataset
Shi *et al*. (2015)	HIOSVM	SVM	F-measure, G-Mean, AUC	UCI repository
Krawczyk *et al*. (2013)	Classifier Ensemble Algorithm	GA	Sensivity, Specificity	UWisconsin dataset, Breast thermogram dataset
Beyan and Fisher (2015)	Hierarchical Decomposition	Hierarchical method	Sensivity, Specificity, G-Mean	KEEL repository
Garcia and Herrera (2008)	EAs for TSS	C4.5	G-Mean	UCI repository
Lessmann (2004)	Parameterization	SVM	F-measure, G-Mean	Real world data
Shin and Cho (2003)	Different Cost	SVM	ROC	Direct Marketing Education
Cohen *et al*. (2004)	Kernel Transformation	one-class SVM	Accuracy, Sensitivity, Specificitynosocomial Dataset	
Raskutti and Kowalczyk (2004)	One-Class Learning	SVM	ROC	KDD 2002 CUP dataset
Cohen *et al*. (2003)	Asymmetrical Margins	SVM	Accuracy, Sensitivity, Specificity, AUC	Public dataset
Wu and Chang (2004)	Kernel-Boundary-Alignment	SVM	AUC	UCI repository
Phua *et al*. (2004)	Cost Senstive Learning	NB, C4.5, BP	Cost Matrix	fraud detection dataset
Perez *et al*. (2004)	Smaller Error rate	C4.5	Cost Matrix	UCI repository
Antonelli *et al*. (2013)	Fuzzy Rule-Based Classifier	Fuzzy Rule-Based Classifier	AUC	KEEL repository
Antonelli *et al*. (2014)	GA, Cost Sensitive Learning	Fuzzy based approach	AUC	KEEL repository
Chen *et al*. (2018)	SMOTE + SVM, SVM Cost-Sensitive, C4.5	S-IDGC Model	AUC, G-Mean	Malware samples, Hiapk market, 91Play market, Baidu market,
Alshomrani *et al*. (2015)	Cost Sensitive Learning	Fuzzy Association Rule-Based Classification, C4.5	F-measure	KEEL repository
Fernandez *et al*. (2010)	Genetic Tuning Approach, Fuzzy rule based Classification Systems	RIPPER, C4.5	AUC	UCI repository

Table 11: Classification and Approaches for Algorithm Level Methods (Cont....)

Author's Name	Techniques used	Applied Algorithms	Metrices used	Data Sets used
Hosseinzadeh and Eftekhari (2015a)	Ensemble Method (fuzzy C-means, Rotation Forest)	C4.5	AUC	KEEL repository
Lin *et al*. (2009)	Meta Imbalanced Classification Ensemble	LDA, SVM	AUC, Accuracy, Specificity, Sensitivity	UCI repository
Li *et al*. (2013)	SVM with Segmentation	SVM	Sensitivity, Specificity, G-Mean	UCI repository
Lu *et al*. (2017)	Ensemble Algorithms, Clonal Selection	C4.5	AUC, F-measure	WEBSPAM-UK2006/2007
Rani and Soujanya (2013)	Two-Tier Classication	C4.5, NB	TPR, TNR, PPV	Approved Drugs
Dash (2016)	Meta-Learning Ensemble Algorithm	C4.5	F-measure, G-Mean, Precision, Recall, Accuracy	http://www.gems-system.org /.Multi-class
Zhu and Wang (2017)	Entropy-based Matrix Learning	Matrix learning	Accuracy, PPV	Real-world sets
Zang *et al*. (2016)	Cost Sensitive Learning	k-NN	Accuracy, Precision, F-measure, G-Mean	KEEL repository
Yu *et al*. (2010)	Cost Sensitive Learning	SVM	Sensitivity, Specificity, G-Mean	UCI repository
Xu *et al*. (2016b)	Maximum Margin	SVM	G-Mean	UCI and KEEL repository
Zhang *et al*. (2014b)	Scaling Kernel-based SVM	SVM	Accuracy, Sensitivity, Specificity, F-measure, G-Mean	KEEL repository

Based on literature survey, the ensemble methods (Breiman, 1996; Joshi *et al*., 2001; Sun *et al*., 2012) like bagging and boosting are the two most popular multi-classifier frameworks and have been applied to deal with the class imbalance problem. Huang *et al*. (2012) introduced a multi-classifier method called Random Subspace Method (RSM) to deal with class imbalance problem. Abolkarlou *et al*. (2014) proposed a novel technique of the ensemble classification for handling the imbalance data using uses hierarchical clustering algorithm for determining the optimal layers. Yuan and Ma (2012) proposed a hybrid method with certain modified weight updating rules to the class imbalance problem. Liu *et al*. (2016b) proposed an improved AdaBoost algorithm to predict user's QoE with the idea of cost-sensitive. Wei *et al*. (2014) designed an ensemble based method called BalancedBoost to improve the performance of models trained on imbalanced datasets. Liu *et al*. (2009) designed a EasyEnsemble and BalanceCascade for class-imbalance learning. These algorithms utilize the major class examples which are basically ignored by under-sampling technique. Both these methods sample the multiple subsets of the major class, train an ensemble from each of these subsets and combine all weak learners in these ensembles into a final ensemble.

Ceballes-Serrano *et al*. (2012) developed a new technique Swarm Boost which combine Boosting, oversampling and sub sampling in optimization criteria to select samples. Jiang *et al*. (2013) proposed three different sampling approaches Sampled Bayesian Network Classifiers (SBNC)-undersampling, SBNC-oversampling and SBNC-hybrid sampling to generate samples from the existing training instances and applied Bayesian network classifiers on the sampled training data. Oliveira *et al*. (2015) proposed a technique for generating classifier ensembles called Iterative Classifier Selection Bagging (ICS-Bagging) to solve class imbalance problem. Ruangthong and Jaiyen (2016) proposed a new hybrid ensemble model based on AdaBoost.M2 and adopt SMOTE algorithm to solve the class imbalance problem in order to predict the probability of term deposit from bank customers. Huang and Zhang (2016) proposed an improved ensemble learning method called AdaBoost with SMOTE to enhance the forecasting performance. Pal and Paul (2017) proposed BoostedGMM model, where SMOTE based oversampling and cluster forest are applied for better clustering. For dealing with class imbalance problems, Hu and Zhang, (2013) proposed a novel hybrid approach which is called as "Clustering-based Subset Ensemble Learning Method". Mustafa *et al*. (2015) discussed a new hybrid machine learning method called Distribution Based MultiBoost (DBMB) for class imbalance problems. Sundarkumar *et al*. (2015) adopted One-Class Support Vector Machine (OCSVM) based under sampling method to predict fraudulent claims in two datasets which are automobile insurance claims and churn prediction in credit card customers.

Further, Braytee *et al*. (2015) present an Artificial Bee Colony (ABC) approach for undersampling technique. In their proposed work, the ABC-Sampling algorithm classifies imbalanced data by identifying the most informative majority samples and is evaluated by training a SVM classifier on the retrieved balanced dataset and evaluating on the test set. Akbani *et al*. (2004) solved the class imbalance problem by applying SMOTE and different error costs (i.e., for SVM algorithm). Later, Krawczyk and Schaefer (2013) examined the samples within the malignant category and determined differing types (kinds) of unbalanced objects. In these method, the authors analyze the performance of progressive ensemble classifiers such as Boosted Support Vector Machine, SMOTEBoost, Pruned Under-Sampling Balanced Ensemble (PUSBE) and Hybrid Cost-Sensitive Ensemble (HCSE). Gao *et al*. (2013) used ensemble learning method called RUSBoost to solve the class imbalance problem. Yongqing *et al*. (2013) proposed an improved SMOTE method incorporating bagging ensemble algorithm called Adaptive SMOTE (ASMOTE) method. ASMOTE method adaptively adjusts the nearest neighbors to solve this critical class imbalance problem.

Further, Wu and Meng (2016) proposed an e-commerce customers churn prediction model which was based on improved SMOTE and AdaBoost. Further, Thai-Nghe *et al*. (2009) adopted over-sampling techniques and cost-sensitive learning for the class imbalance for improving the prediction/ classification. Further, a Modified Hyper Network (MHN) is proposed to deal with this problem, i.e., 'class imbalance problem' (2013b). Also, Cao and Zhai (2015) proposed a hybrid approach for balancing the imbalance dataset in binary classification. In their proposed approach, authors used SMOTE techniques to generate the synthetic data for the minority class and then over-sampling techniques are applied to delete some samples of majority class that are less classified. El-Ghamrawy (2016) presented a Clustered Knowledge Management Development Framework (CKMD) for enhancing the performance of knowledge discovery in imbalanced data. Soltani *et al*. (2011) presented a novel method for gene classification in the extremely imbalanced dataset. Further, Zughrat *et al*. (2015) developed a new Iterative Fuzzy Support Vector Machine (IFSVM) algorithm for imbalanced rail data classification with bootstrapping-based oversampling and undersampling.

The unbiased decision tree model using CART is proposed and applied on IPTV KPI (IPTV KPI Internet Protocol TV) and usercomplaint imbalanced datasets (Wang *et al*., 2016). Note that in general, IPTV is a method of delivering and viewing TV (television) programmes using an Internet Protocol (IP) transmission and service infrastructure. Barandela *et al*. (2003) focuses on the use of combinations of individual classifiers for handling skew distribution and examined with the nearest neighbor classifier. Guo and Viktor (2004) applied DataBoost-IM method to train the imbalanced data set. The DataBoost-IM method was evaluated, in terms of the F-measures, G-mean and overall accuracy, against seventeen highly and moderately imbalanced data sets using decision trees as base classifiers. Matsuda and Murase (2016) presented a single-layered Complex Valued Neural Network (CVNN) to classify imbalanced dataset and SMOTE model used for imbalanced data problem. Hence, the list of classification and approaches for hybrid level are discussed/ presented in Tables 12 and 13.

Table 12: The List of Classification and Approaches for Hybrid Level Methods

Author's Name	Techniques used	Applied Algorithms	Metrices used	Data Sets used
Seiffert *et al*. (2008)	SMOTEBoost	C4.5, NB, RIPPER	ROC and Area under the PRC curve.	UCI repository, Mammography dataset
Wang and Yao (2012)	AdaBoost.NC	AdaBoost	Precision, F-measure, ROC and G-mean	UCI repository
Galar *et al*. (2012)	Ensemble-based Methods	C4.5	Accuracy, ROC	KEEL repository
Kaur *et al*. (2016)	Bagging+SMOTE, RUS	NB, SMO, DT, LR and RBF	ROC, F-measure	GITHUB
Park and Ghosh (2014)	DT	C4.5, CART	ROC	UCI repository, KEEL repository
Patil and Sonavane (2017)	MEMMOT, CMEOT	RF, NB, AdaBoostM1	F-Measure, ROC	UCI/KEEL repository
Kamal *et al*. (2009)	HW: Higher Weight Feature Selection	ReliefF	True Positive Rate (TPR), True Negative Rate	Microarray expression datasets (TNR),
Accuracy				
Peng *et al*. (2016b)	SMOTE-DGC	DGC	ROC, Specificity, Sensitivity, G-Mean	KEEL repository
Sandhan and Choi (2014)	Hybrid Sampling, Bootstrapping	SVM, LR, k-NN, Gaussian Classifier	ROC	structural classification of proteins
Sonak *et al*. (2016)	Hybrid Approach	ANN, K-means, GA	Accuracy	UCI repository
Winata and Khodra (2015)	Adaptive Boosting, Bagging	DT (J48), NB, SMO	Hamming loss, Accuracy, F-measure	LAPOR dataset
Zhang *et al*. (2014c)	Ensemble Method	Modified SVM	Accuracy, Precision, F-measure, G-Mean	KEEL repository
Dwiyanti and Adiwijaya (2017)	RUSBoost	C4.5	F-measure	PT. Telkom Indonesia
Dıez-Pastor *et al*. (2015b)	RB-Boost	k-NN, SVM	ROC	HDDT collection
Abd Elrahman and Abraham (2015)	Hybrid Ensemble Approach	NB, BP, SVM, RF, RBF, C4.5	Sensitivity, Specificity, Precision, F-measure, ROC	UCI repository
Peng and Yao (2010)	AdaOUBoost	SVM	Accuracy	TRECVID data set
Galar *et al*. (2013)	EUSBoost	C4.5	ROC	KEEL repository
Amin *et al*. (2016)	SMOTE, ADASYN, TRkNN, MWMOTE	version2(LEM2), Exhaustive and GA	Accuracy, Recall, Precision, F-measure	Telecom-sector
Abolkarlou *et al*. (2014)	Ensemble Classifier	SVM	Accuracy, G-means, Diversity	KEEL repository
Chaıri *et al*. (2012)	Sample Selection (SS)	MLP	Precision	KDD-CUP-99
Fernandez *et al*. (2017)	RUS, ROS, SMOTE	RF	G-Mean	Big Data UCI repository
Ramentol *et al*. (2012)	SMOTE, Rough-set	C4.5	ROC	UCI repository
Wang *et al*. (2010)	AdaBoost.NC Correlation Learning (NCL)	MLP and C4.5	ROC	UCI repository
Lu *et al*. (2016)	AdaBoost	G-mean Optimized Boosting	F-measure, G-Mean	UCI repository
Oliveira *et al*. (2015)	SMOTE-ICS-Bagging	Iterative Classifier with Bagging	ROC	KEEL repository
Qian *et al*. (2014)	Resampling Ensemble Algorithm	NB	Precision, Recall, G-Mean, F-measure	UCI repository

Table 13: The List of Classification and Approaches for Hybrid Level Methods (Cont..)

Author's Name	Techniques used	Applied Algorithms	Metrices used	Data Sets used
Vinodhini and Chandrasekaran (2017)	Ensemble Algorithm	M-Bagging	ROC, G-Mean	www.cs.uic.edu/ liub/FBS/FBS.html
Tran and Liatsis (2016)	RABOC	One-Class Classification, AdaBoost	Half Total Error Rate	BioSecure DS2, XM2VTS
Abdi and Hashemi (2015)	Boosting, MDOBoost	C4.5	MAUC, G-Mean, Recall	UCI, KEEL repository
Liu *et al*. (2009)	Undersampling	AdaBoost	AUC, F-measure, G-Mean	UCI repository
Huang and Zhang (2016)	AdaBoost, SMOTE	AdaBoost	Precision, Recall, Specificity, Accuracy, F-measure	SKEMPI
Gazzah *et al*. (2015)	Hybrid Approach	SVM	TN, TP	Faces94, SID signature databases
Krawczyk and Schaefer (2013)	Cost-Sensitive	Ensemble Classifier	Sensitivity, Specificity, Accuracy	Breast Thermogram dataset
Shi *et al*. (2015)	Emsemble Method-RUS	SVM	Accuracy	EEG data sets
Yongqing *et al*. (2013)	SMOTE	SVM	Accuracy, Sensitivity, Specificity, G-Mean	UCI repository
Wu and Meng (2016)	Improved SMOTE	AdaBoost	Accuracy, Sensitivity, Recall, Specificity, Precision, G Mean, F Measure	B2C e-commerce site
Farajzadeh-Zanjani *et al*. (2016)	SMOTE	AdaBoost.M1, Bagging	Normalised Popt	CWRU Bearing Data Center
Ruangthong and Jaiyen (2016)	AdaBoost.M2, SMOTE	BN, DT, J48	Sensitivity, Specificity, Accuracy	UCI repository
Liu *et al*. (2016b)	Cost-Sensitive Learning	AdaBoost	F-measure	Jiangsu Telecom
Liu *et al*. (2009)	EasyEnsemble	EasyEnsemble	ROC	UCI repository
Chawla *et al*. (2003)	SMOTEBoost	AdaBoost.M2	Recall, Precision, F-value	KDD Cup-99 dataset
Barandela *et al*. (2003)	Ensemble Learning	1-NN(Multi-Layer Perceptron)	Accuracy	UCI repository
Braytee *et al*. (2015)	ABC	SVM	AUC, F-measure	UCI repository
Jiang *et al*. (2016)	GA-based SMOTE	C4.5	F-measure, G-Mean	KEEL repository
Sundarkumar *et al*. (2015)	SVM based Undersampling	DT, SVM, LR, PNN	Sensitivity, Specificity, Accuracy	Automobile Insurance fraud and Credit card customer churn dataset
Guo and Viktor (2004)	Ensemble Learning	DataBoost-IM	F-measures, G-mean, Accuracy	UCI repository
Matsuda and Murase (2016)	SMOTE	Single-layered complex-valued neural network	Accuracy, Sensitivity	UCI repository
Akbani *et al*. (2004)	SMOTE+SVM	SVM	G-Mean	UCI repository

Table 14: Data Level Techniques with their Strengths and Weaknesses

Techniques	Strengths	Weaknesses
Random Sampling Techniques		
Random under sampling (Xiaoying and Sheng, 2012; Zhang *et al*., 2015; Wu and Meng, 2016; Dai, 2015; Garcıa *et al*., 2012)	1. Uses a non heuristic method that aims to balance class distribution through the random exclusion of majority class examples. 2. Results in improved classifier accuracy.	1. Possibility of discarding potentially useful data. 2. Problem when class overlapping exists. 3. By removing, the samples from the majority class the classifier may miss important concepts pertaining to the majority class.
Random Over Sampling (Xiaoying and Sheng, 2012; Fan and Tang, 2010; Li and Deng, 2016; Del Rıo *et al*., 2014; Chetchotsak *et al*., 2015)	1. A direct approach for randomly replicates minority instances to balance the imbalanced data. 2. Appends data to the original data set.	1. Can increase the likelihood of occurrence of over fitting, since it makes exact copies of existing instances. 2.Hinders learning algorithm.
Data Generation using Synthetic Sampling SMOTE: (Synthetic Minority Oversampling Technique) (Xiang and Bai, 2013; Van Vlasselaer *et al*., 2013; Debowski *et al*., 2012; Xu *et al*., 2013; Amin *et al*., 2016; Mustafa and Li, 2017; Zhong *et al*., 2009; Ren *et al*., 2011; Bhagat and Patil, 2015)	1. Avoids the problem of over fitting of classifiers. 2. Applies Euclidian distance. 3. Very popular due to its simplicity and efficiency. 4. No loss of existing instances. 5. Makes the decision regions larger and less specific.	1. May bring noise in data. 2. Can result in increase in overlapping of classes. 3. May not be very effective for high dimensional data. 4. Does not consider the underlying distribution of the minority class and latent noises in the data set.
M-SMOTE (Modified SMOTE) (Lee *et al*., 2015)	1. Classifies the samples of minority class into three categories such as security samples, border samples and latent noise samples by calculating the distances of all the samples. 2. Near neighbor strategy is used for selecting synthetic data.	1. Does not consider the differences of importance features.
Borderline-SMOTE (Pengfei *et al*., 2014)	1. The borderline examples of the minority class are over-sampled. 2. Strengthen the borderline minority examples.	1. This methods only over-sample the borderline examples of the minority class. 2. Generates synthetic instances in unsuitable locations, such as overlapping regions and noise regions.
Safe-Level-SMOTE (Garcıa *et al*., 2012)	1. The safe level computes by using nearest neighbor minority instances. 2. By synthesizing the minority instances more around larger safe level, we achieve a better accuracy performance than SMOTE and Borderline-SMOTE. 3. Safe Level-SMOTE carefully over-samples a data set.	1. The different definitions to assign safe level would be valuable. 2. Does not classify data sets which have nominal attributes. 3. Does not address automatic determination of the amount of synthetic instances generated by Safe-Level-SMOTE.

Table 15: The data level techniques with their strength and weaknesses

Technique	Strength	Weaknesses
Adaptive Synthetic Sampling Approach ADASYN (Adaptive Synthetic Sampling Approach) (Majumder *et al*., 2016; Amin *et al*., 2016)	1. Reduces the bias introduced by the class imbalance and adaptively shifting the classification decision boundary toward the difficult samples. 2. Assigns weighted distribution for different minority class samples according to their level of difficulty in learning. 3. Provides improved accuracy for both minority and majority classes.	1. Not applicable for Multiple-class imbalanced learning problems. 2. Online learning and evaluation process.
SPIDER: (Selective Pre-Processing of Imbalanced Data for Improving Classification Performance) (Stefanowski and Wilk, 2008)	1. Combines local over-sampling of the minority class with filtering difficult samples from the majority classes. 2. Improves the sensitivity of the minority class while preserving sufficiently accurate recognition of the majority classes.	1. May deteriorate specificity.
One Side Selection (OSS) (Garcıa *et al*., 2012)	1. Simple technique. 2. Data is affected but not the classifier model.	1. Time consuming
T-Link (Tomek link) (Garcıa *et al*., 2012; Sain and Purnami, 2015)	1. Data cleaning method. 2. Removes unnecessary overlapping between classes after synthetic sampling. 3. Easy to implement.	1. Removal of informative and significant samples.
Cluster-Based Oversampling (CBO) (Machado and Ladeira, 2011)	1. Efficiently manages the within-class imbalance problem. 2. Applies K-means clustering technique.	1. Possibility of generating redundant samples.

Table 16: The Algorithm Level Techniques with their Strength and Weaknesses

Technique	Strength	Weaknesses
Classification Based on Associations	1. Minimizes the bias of the classifier towards majority.	1. Fails with the uneven class distribution. 2. Needs pre-processing
Support Vector Machines for the nonstandard situation (Chen *et al*., 2010b; de Souza *et al*., 2016; Cao and Shen, 2016)	1. Effective method. 2. Designed exclusively to handle uneven class distribution	1. A good choice of smoothing parameters is needed. 2. Can lead to over fitting.
Weighted k-NN (Lopez *et al*., 2014)	1. Weights are assigned to the respective classes. 2. Increases the value of the geometric mean.	1.Can lead to over fitting. 2. No control on the number of examples to be removed.
Fuzzy Classifier (Tan *et al*., 2015; Ramentol *et al*., 2015; Tan *et al*., 2011)	1. Reduces the complexity of the data by presenting the data to the user in the form of perceivable fuzzy rules. 2. Learns each class separately and mapping of each class into a fuzzy set	1. Pre-processing might need. 2. Strong implications for the interpretation of the fuzzy set.
MetaCost (Domingos, 1999)	1. Cost reduction method. 2. Stratification can be applied without approximation.	1. Can lead to over-fitting. 2. Increases the learning time.
Feature Selection Framework (Cuaya *et al*., 2011; Alibeigi *et al*., 2012)	1. Effective for Imbalance Class Distribution.	1. Cause over fitting. 2. Performance depends on training the method.
z-SVM (Imam *et al*., 2006)	1. Better Recognition Performance. 2. Less Sensitive to the Selection of learning parameters.	1. Error cost are often unknown and based on domain. 2. Can lead to over fitting.
Hierarchical Fuzzy Rule Based Classification (Fernandez *et al*., 2009)	1.Improves the Accuracy. 2.Based on Genetic Rule Selection Process. 3. Simple and Effective.	1. Pre-processing might be needed.
Class Conditional Nearest Neighbor (Kriminger *et al*., 2012) Distribution	1. Single Class Algorithm. 2. Simple Method.	1. Not efficient to learn from widespread class.
Cost-Sensitive SVM (Cao *et al*., 2013b)	1.Effective and Fast Processing.	1. Error cost are often unknown and decision will be based on domain. 2. May lead to over fitting.
Cost-Sensitive Neural network (Cao *et al*., 2013b)	1. Easy to implement. 2. Process Fast.	1. If error cost is not known, additional cost may introduced. 2. If real cost are unknown then it does not work effectively

The strengths and short-comings of class imbalance classification techniques (Data level, Algorithm level and Hybrid methods) are listed in Table 14-18.

Hence, this section discusses the data-level, algorithm-level and hybrid level techniques for handling class imbalance problem. Now, next section will deal with assessing the quality of the data for class imbalance problems.

Table 17: Hybrid Level Techniques with their Strengths and Weaknesses

Techniques	Strengths	Weaknesses
Cost- Sensitive Boosting AdaCost (Seiffert *et al*., 2008)	1. Uses the cost of misclassification to update the training distribution on successive boosting rounds. 2. Reduces both fixed and variable misclassification costs. 3. Noise resistant. 5. Multiple classifiers provide better prediction than single classifier.	1. Works bad with severely imbalanced data sets. 2. Poor selection of cost adjustment factor leads to, poor performance.
CSB1, CSB2 (Seiffert *et al*., 2008)	1. Uses an adjustment functions and consider the costs in the weight update. 2. Not affect by Noisy Data.	1. Poor selection of cost adjustment leads to poor performance. 2.Time consuming.
AdaC1 (Seiffert *et al*., 2008)	1. Variation of AdaCost. 2. Integrates the costs in the weight update formula (with in the exponent part).	1. Poor selection of cost adjustment leads to poor performance. 2.Can lead to over fitting.
AdaC2 (Seiffert *et al*., 2008)	1. Variation of AdaCost. 2. Integrates the costs in the weight update formula (outside the exponent part).	1. Poor selection of cost adjustment may lead to poor performance.
AdaC3 (Seiffert *et al*., 2008)	1. Variation of AdaCost. 2. Integrates the costs in the weight update formula (both).	1. Complexity increases of more classifiers. 2. Time consuming.
RareBoost (Seiffert *et al*., 2008)	1. Makes use of Confusion Matrix.	1. Leads to over fitting.
AdaBoost (Seiffert *et al*., 2008; Wang and Yao, 2012; Patil and Sonavane, 2017; Rahman *et al*., 2015)	1. Accuracy Oriented Algorithm. 2. Provides Overall Accuracy. 3. Standard approach to improve prediction accuracy. 4. Reduces bias and variance in data.	1. Bias towards the majority class. 2. The overall performance of the classifier is ignored. 3. Sensitive to noise.
Data Preprocessing + Ensemble Learning Boosting-Based SMOTEBoost (Seiffert *et al*., 2008)	1. Combines an Oversampling Technique (SMOTE) with AdaBoost. 2. Flexible Approach. 3. Easy to use.	1.Specific algorithm might not be suitable for all types of problems and data sets. 2.Variety class is difficult to achieve. 3. Increase in the use of more classifiers may develop complexity.
MSMOTEBoost (Hu *et al*., 2009)	1. More Diversity in the Training Data. 2. MSmote + Boost.	1. Increase in the use of more classifiers may increase complexity.
RUSBoost (Gao *et al*., 2013; Dwiyanti and Adiwijaya, 2017)	1. Simple and Fast. 2. Less complex than SMOTE Boost Algorithm.	1. Do not solve multi-class instance problems.

Table 18: Hybrid Level Techniques with their Strengths and Weaknesses (Cont..)

Techniques	Strengths	Weaknesses
DataBoost-IM (Guo and Viktor, 2004)	1. Combines Data Generation and Boosting Procedures. 2. Improves the Predictive Accuracies of both the Majority and Minority Classes.	1. Consumes Time. 2. Generates more minority samples than majority samples.
Data Preprocessing + Ensemble Learning Bagging-based		
SMOTEBagging (OverBagging) (Wang and Yao, 2009)	1. Simple. 2. Oversampling is carried out before training the Classifier. 3. Reduces Variance and Overcomes over-fitting.	1. Works only if the base classifier is good. 2. Selection of the classifier will affect the performance.
QuasiBagging (UnderBagging (UB)) (Chang *et al*., 2003)	1. Under Sampling with Bagging. 2. Can obtain more Diverse Ensembles.	1. Ignore of some useful samples. 2. Computational complexity.
Asymetric Bagging (UB) (Tao *et al*., 2006)	1. Same as Under Bagging.	1. Highly depends on the data.
Roughly Balanced Bagging (UB) (Hido *et al*., 2009)	1. Simple Strategy. 2. The Number of Positive Samples is kept fixed.	1. Leads to over fitting. 2. Selection of the classifier will affect the performance.
Bagging Ensemble Variation (UB) (Li, 2007)	1. Under Sampling Approach is used. 2. Improves Stability and Accuracy.	1. Selection of the classifier will affect the Performance. 2. Increases the space.
UnderOverBagging (Wang and Yao, 2009)	1. Makes use of both under Sampling and over Sampling Methods.	1. Increases the space. 2. Loss of Interpretability.
IIVotes (Bl aszczynski *et al*., 2010)	1. Integrates SPIDER Data Pre-Processing technique with IVotes. 2. Not needed to define number of bags. 3. Provides better trade-off between sensitivity and specificity for the minority class.	1. Time Consuming. 2. Increase in space. 3. Biased towards strong rules of majority class.
Hybrid EasyEnsemble (Liu *et al*., 2009)	1. All the classifiers are trained in parallel. 2. Underbagging+Adaboos	1. Increase in space. 2. More Classifiers are assigned to learn at each iteration.
Balance Cascade (Liu *et al*., 2009)	1. Classifiers are Trained Sequentially.	1. Loss of Interpretability.

Assessing the Quality of the Data

This section presents the results of our literature review on class imbalance problems by addressing the research questions (shown in Table 1). During the analysis of the state-of-the-art literature of class imbalance problems, we consider the following research questions:

RQ-I: What are the main research motivations behind class imbalanced problems in classification?

The main research motivation behind class imbalance problems is identified in section 2.3.1.

RQ-II: Why class imbalance problems are popular?

The popularity of class imbalance problems are justified based on addressing the following questions:

RQ-II.I: What is the number of publications per year in the research area?

The distribution of research papers on class imbalance problems from 2002 to 2017 among the 5 databases are shown in Fig. 2, where the major percentage of papers are extracted from the IEEE database.

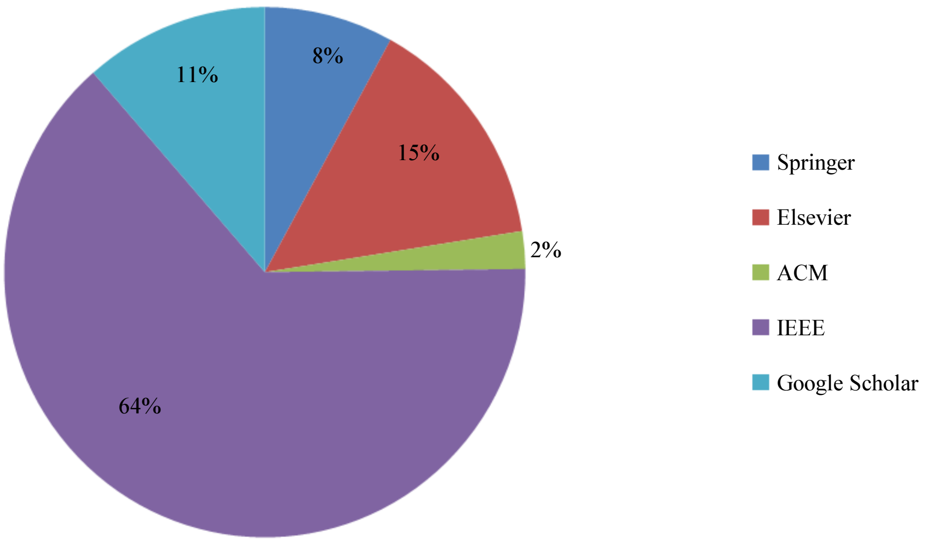

Fig. 2: Distribution of the Papers Relevant Research in Different Publication Databases

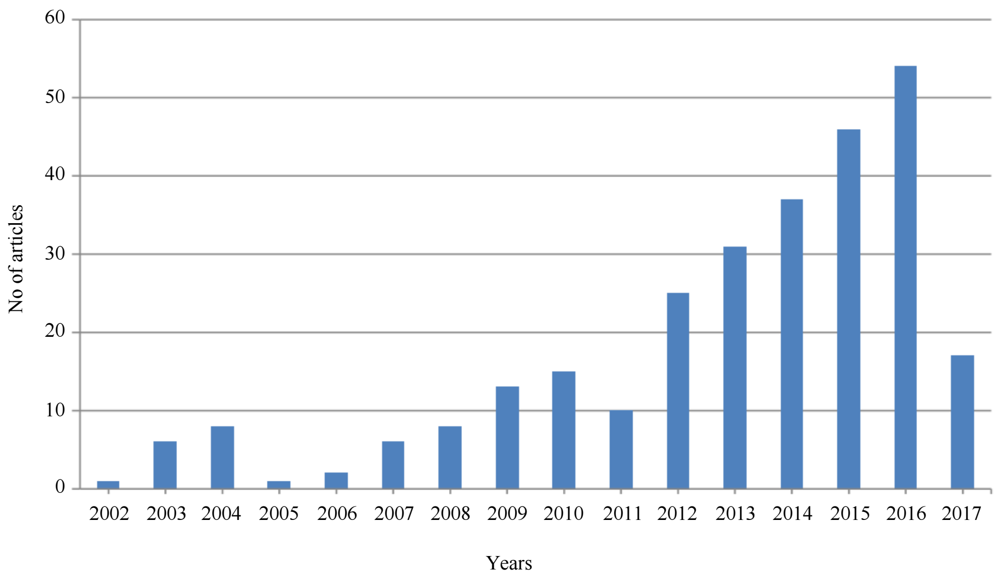

Fig. 3: Numbers of Papers Published in Each Year (up to June 2017)

Table 19: Publication Channels (Journals and Conferences)

Publication Channel	**No. of Publications**
International Joint Conference on Neural Networks (IJCNN).	19
Journal of Neuro computing	7
Journal of Information Sciences	7
Journal of Knowledge-Based Systems	7
International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)	5
International Conference on Machine Learning and Applications (ICMLA).	5
International Conference on Data Mining (ICDM)	6
Transactions on Systems, Man and Cybernetics (IEEE-SMC)	4
International Conference on Fuzzy Systems (FUZZ-IEEE)	4
IEEE Transactions on Knowledge and Data Engineering	4
International Conference on.Systems, Man and Cybernetics (SMC)	3
International Journal on Neural Computing and Applications	3
International Conference on Intelligent Systems Design and Applications (ISDA)	3
Journal of Pattern Recognition	3
International Conference on Tools with Artificial Intelligence (ICTAI).	2
International Conference on Hybrid Intelligent Systems (HIS)	2
Expert Systems with Applications	2
International Conference on Machine Learning and Computing (ICMLC)	2
International Conference on Machine Learning and Cybernetics (ICMLC)	2
Knowledge and information systems	2
Iranian Conference on Electrical Engineering (ICEE)	2
International Conference on. Soft Computing and Intelligent Systems (SCIS)	2
Journal of Biomedical and Health Informatics.	2
Pattern Recognition Letters	2
International Conference on Computer Science and Education (ICCSE)	2
IEEE Transactions on Neural Networks and Learning Systems	2
International eConference on Computer and Knowledge Engineering (ICCKE)	2
International Conference on Advanced Computer Science and Information Systems (ICACSIS)	2
International Conference on Ubiquitous Information Management and Communication (IMCOM)	2
International Conference on Computer and Communications (ICCC)	2
SAI Computing Conference (SAI)	2
International Conference on Internet of Things (iThings)	2
Symposium Series on Computational Intelligence (SSCI)	2
International Conference on Neural Networks and Brain (ICNNB)	1
International Conference on Intelligent Data Engineering and Automated Learning (IDEAL)	1
International Symposium on Empirical Software Engineering and Measurement (ESEM)	1
International Conference on Advanced Computing and Communications Systems (ICACCS)	1
Annual Meeting of the North American Fuzzy Information Processing Society (NAFIPS)	1
International conference on Knowledge Discovery and Data Mining (KDD)	1
IEEE Region 10 Conference TENCON.	1
International Conference on Information Reuse & Integration (IRI)	1
International Conference on Computer Science and Information Technology (ICCSIT)	1
Applied Soft Computing	1
IEEE Congress on Evolutionary Computation (CEC)	1
First International Workshop on Database Technology and Applications (DBTA)	1
International Workshop on Computer Science and Engineering (WCSE)	1
International Conference on Knowledge Discovery and Data Mining (KDD)	1
International Conference on Advanced Information Networking and Applications Workshops (AINA)	1
International Joint Conference on Artificial Intelligence (IJCAI)	1
International conference on Multimedia Information Retrieval (ACM-MIR)	1
International Conference on Information and Financial Engineering (ICIFE)	1
Chinese Control and Decision Conference (CCDC)	1
Engineering Applications of Artificial Intelligence	1
International Conference on Computational Problem-Solving (ICCP)	1
International Conference on Business Intelligence and Financial Engineering (BIFE)	1
International eConference on Computer and Knowledge Engineering (ICCKE)	1
International Conference on Information Technology, Computer Engineering and Management Sciences (ICM)	1
Intelligent Data Analysis	1
International Conference on Computing and Convergence Technology (ICCCT)	1
Journal of Intelligent Information Systems	1
International Conference on Innovative Computing Technology (INTECH)	1
Journal of Personal and Ubiquitous Computing	1
Journal of Computational Biology and Chemistry	1
Symposium of Image, Signal Processing and Artificial Vision (STSIVA)	1

RQ-II.II: What are the publication trends in the research area?

We have shown the number of research papers on class imbalance problems and their year of publications in Fig. 3. From Fig. 3, we observe that the number of papers published is in increasing trend since 2002. Further, we observed that the number of papers published significantly increased in 2016. Most of the papers on class-imbalanced problems are published in IJCNN, ICDM, FUZZ-IEEE, ICMLA and other major conferences and journals (as shown in Table 19, 20, and 21). Among 281 papers, more than 77 papers were published in major journals including Information Sciences, Knowledge-Based Systems, Neurocomputing, Expert Systems with Applications and Applied Soft Computing. A list of distribution of studies per publication channel is shown in Tables 19-21.

RQ-III: What are the existing methods and techniques that are applied for class imbalance problems?

Based on our literature survey, as discussed in Section 3, the class imbalance problems are broadly classified into data level, algorithmic level and hybrid level or ensemble level (Fig. 4). It can be observed that 64% of studies focus on data level techniques, 22% of studies focus on algorithmic techniques and 14% of studies focus on hybrid level. Fig. 4 shows the percentage of data level, algorithm level and hybrid techniques applied for the class imbalance problems.

RQ-III.I: Which types of techniques and approaches have been employed in the research area?

The main techniques used for solving the class imbalance problems are data-level, algorithm level and hybrid methods. The number of research articles published using data level, algorithm level and hybrid methods for each year is presented in Fig. 5. Figure 5 depicts a bar graph indicating the number of research articles published using data level, algorithm level and hybrid method for each year. The Figure is plotted to present the distribution of research attention towards the techniques applied for class imbalance problems in each publication year.

The study shows that the application of data level techniques showed an increasing trend from 2002 over the years. Further, we observed that the number of papers using data level techniques significantly increased in 2016. The study also shows that the application of algorithm level techniques and increased over the years. Based on Fig. 5, we observe that during 2013, 2015 and 2016, major research was contributed towards algorithm level techniques for handling class imbalance problems. Based on Fig. 5, we also observe that hybrid techniques started in 2008 and consecutively increased over the years. Further, we conclude that data level techniques are prominent and highly used for handling class imbalance problems in classification.

RQ-III.II: What are the algorithms applied and metrics used in the research area?

Figure 6 depicts a pie chart indicating the different algorithms used and their percentages for solving the class imbalance problems. From Fig. 6, we observe that the percentage of application of SVM algorithm is 32% for solving the class imbalance problems. The Decision trees closely follow with 31% for solving the class imbalance problem.

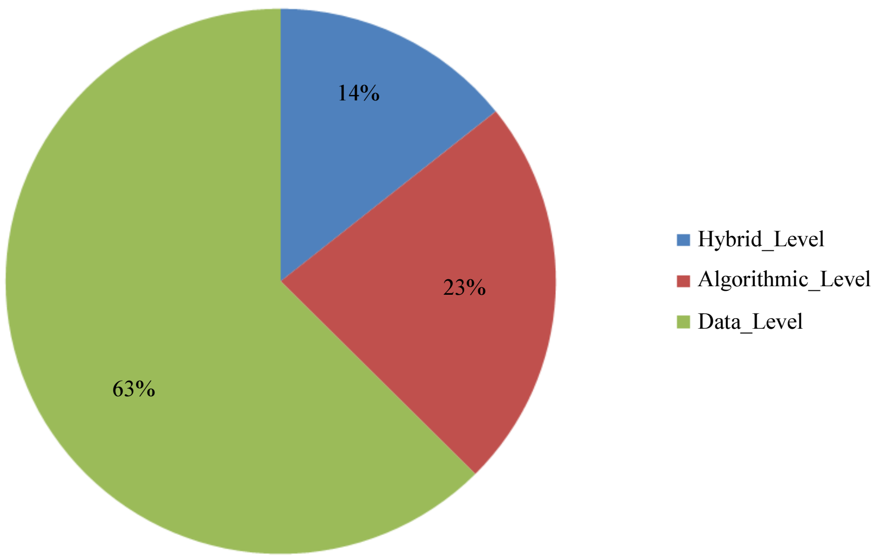

Fig. 4: Percentage of Data level, Algorithmic level and Hybrid Level Techniques applied for Class Imbalance Data Sets

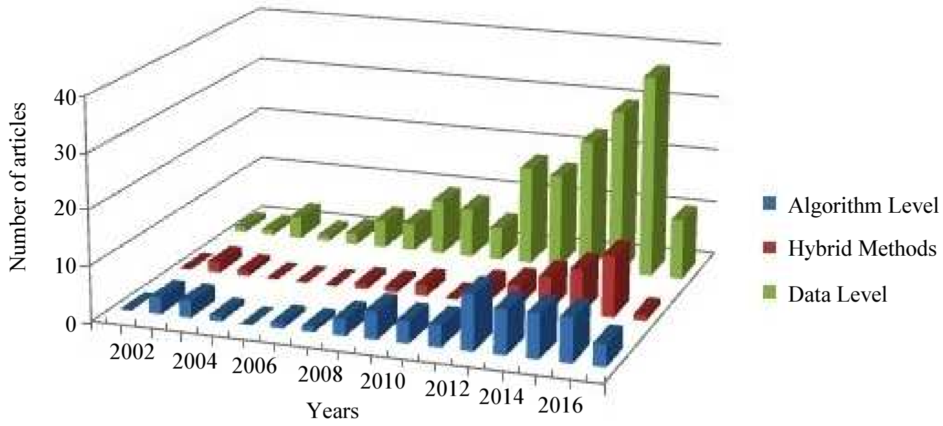

Fig. 5: Distribution of articles related to data level, algorithm level and hybrid techniques for class imbalance problem

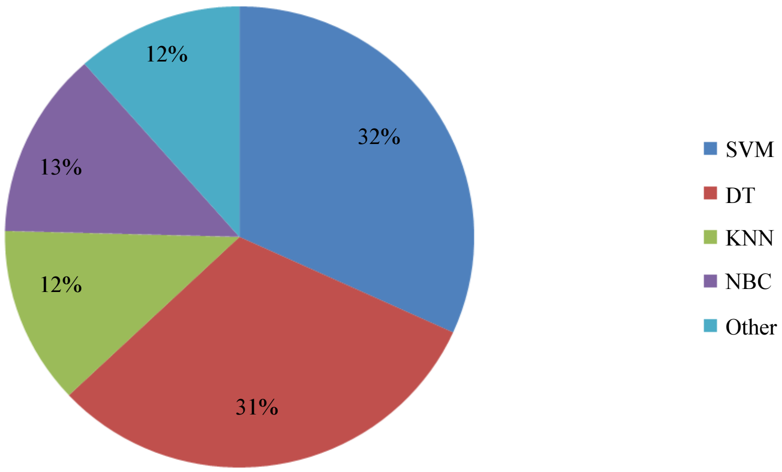

Fig. 6: Application of Different Algorithms and their Percentages

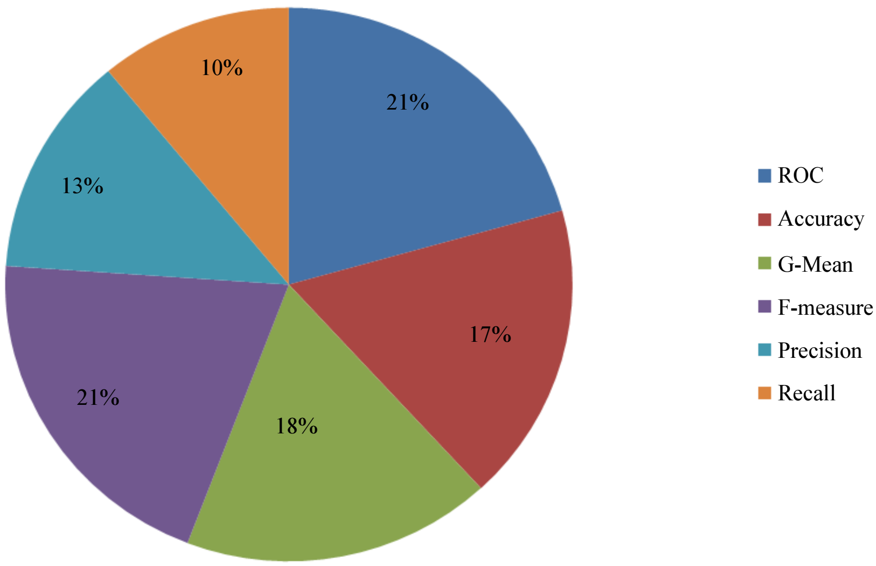

Fig. 7: Percentage of the Different Metrics in Literature

Table 20: Publication Channels (cont....)

Publication	**No. of Publications**
International Conference on Software Engineering and Service Science (ICSESS)	1
International Conference on Biomedical and Health Informatics (BHI)	1
International Symposium on Computational Intelligence and Design (ISCID)	1
International Conference on Computer Modelling and Simulation (UKSim)	1
International Conference on Advances in Social Networks Analysis and Mining (ASONAM)	1
International Conference on Advanced Data and Information Engineering (DaEng)	1
International Workshop on Systems, Signal Processing and their Applications (WoSSPA)	1
Joint IFSA World Congress and NAFIPS Annual Meeting (IFSA/NAFIPS)	1
Progress in Artificial Intelligence	1
International Conference on Contemporary Computing (IC3)	1
Annual Conference of the Industrial Electronics Society IECON	1
International Workshop on Database and Expert Systems Applications (DEXA)	1
International Conference on Mechatronic Sciences, Electric Engineering and Computer (MEC)	1
International Symposium on Communications and Information Technologies (ISCIT)	1
Asian Conference on Pattern Recognition	1
Conference Anthology	1
International Conference on Data Mining Workshops (ICDMW)	1
Chem-Bio Informatics Journal	1
International Symposium on Applied Computational Intelligence and Informatics (SACI)	1
International Conference on Signal-Image Technology and Internet-Based Systems(SITIS)	1
International Conference on Pattern Recognition (ICPR)	1
International Conference on Robotics and Biomimetics(IEEE-ROBIO)	1
International Conference on Big Data and Smart Computing (BIGCOMP)	1
International Conference on Multimedia Computing and Systems (ICMCS)	1
Annals of Operations Research	1
International Conference on High Performance Computing and Applications (ICHPCA)	1
International Journal of Machine Learning and Cybernetics	1
Information Fusion	1
Vietnam Journal of Computer Science	1
International Conference on Identification, Information and Knowledge in the Internet of Things (IIKI)	1
International Conference on Data and Software Engineering (ICODSE)	1
International Conference on Information and Communication Technology (ICoICT)	1
International Computer Science and Engineering Conference (ICSEC)	1
International Conference on Computer and Information Technology (ICCIT)	1
International Conference on Multimedia and Expo (ICME)	1
International Conference on Computer Communication and Networks (ICCCN)	1
International Conference on Biomedical Engineering (BMEiCON)	1
IEEE Transactions on Fuzzy Systems	1
IEEE Transactions on NanoBioscience	1
International Conference on Knowledge and Systems Engineering (KSE)	1
International Congress on Image and Signal Processing (CISP)	1
International Conference on Electrical Engineering and Informatics (ICELTICs)	1
International Conference on Soft Computing in Data Science (SCDS)	1
International Journal of Hybrid Intelligent Systems	1
IEEE International Advance Computing Conference (IACC)	1
Journal of Cognitive neurodynamics	1
International Conference on Intelligent Computer Communication and Processing (ICCP)	1
Latin-American Congress on Computational Intelligence (LA-CCI)	1
International Symposium on Computer Science and Software Engineering (CSSE)	1
International Symposium on Artificial Intelligence and Signal Processing (AISP)	1
Expert Systems with Applications	1
International ACM Recommender Systems Challenge (RecSys)	1
Journal of Procedia Computer Science	1
International Conference on Computational Science and Its Applications (ICCSA)	1
Soft Computing	1
International Conference on Ubiquitous Intelligence and Computing.	1
International Multi-Conference on Systems, Signals and Devices (SSD)	1
International Joint Conference on Computer Science and Software Engineering (JCSSE)	1
International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA)	1
International Conference on Computational Intelligence and Computing Research (ICCIC)	1
International Symposium on Computational and Business Intelligence (ISCBI)	1
Annual Conference on Industrial Electronics Society (IECON)	1
International Conference on Natural Computation (ICNC)	1
International Conference on Innovative Computing Technology (INTECH)	1
Symposium on Signal Processing, Images and Computer Vision (STSIVA)	1
International Conference on Biomedical Image and Signal Processing (ICBISP)	1
International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT)	1
International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT)	1
International Conference on Advances in Electronics, Communication and Computer Technology (ICAECCT)	1
International Conference on Advances in Computing, Communications and Informatics (ICACCI)	1

Table 21: Publication Channels (cont....)

Publication channel	No. of publications
International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)	1
International Conference on Communication and Signal Processing (ICCSP)	1
IEEE Transactions on Multimedia	1
International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)	1
International Conference on Intelligent Computing and Applications (ICICA)	1
International Conference on Soft Computing and Data Mining (SCDM)	1
IEEE Access	1
Journal of Empirical Software Engineering	1
Engineering Applications of Artificial Intelligence	1
International Conference on Industrial Technology (ICIT)	1
World Congress on Intelligent Control and Automation (WCICA)	1
International conference on Research and Development in Information Retrieval	1
Software Quality Journal	1
Arabian Journal for Science and Engineering.	1
Advances in Nature and Biologically Inspired Computing	1
International Computer Symposium (ICS)	1
International Conference on Computer Engineering & Systems (ICCES)	1
IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)	1
International Conference on. Bioinformatics and Biomedicine (BIBM)	1
International Conference on Software Quality, Reliability and Security Companion (QRS-C)	1
International Conference on Signal and Information Processing (IConSIP)	1
International Computer Science and Engineering Conference (ICSEC)	1
International Conference on Service Systems and Service Management (ICSSSM)	1
International Conference on Computers, Software and Applications (COMPSAC)	1
International Conference on Knowledge and Smart Technology (KST)	1
International Wireless Communications and Mobile Computing Conference (IWCMC)	1
International Conference on Science of Electrical Engineering (ICSEE)	1
International Workshop on Computational Intelligence (IWCI)	1
International Conference on Wireless Communications & Signal Processing (WCSP)	1
International Joint Conference on Computer Science and Software Engineering (JCSSE)	1
International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)	1
International Conference on Cloud Computing Research and Innovations (ICCCRI)	1
International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)	1
Conference on Swarm Intelligence and Evolutionary Computation (CSIEC)	1
International Conference on Big Data Computing Service and Applications (BigDataService)	1
IEEE Congress on Evolutionary Computation (CEC)	1
Journal of Complex and Intelligent Systems	1
IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews)	1
IEEE Transactions on Industry Applications	1
Pattern Analysis and Applications	2
Information Processing and Management	1
International Conference on Electrical, Computer and Communication Engineering	1
Mediterranean Conference on Embedded Computing	1
Journal of artificial intelligence research	1
International Conference on Machine Learning	1
European Conference on Principles and Practice of Knowledge Discovery in Databases	1
European Conference on Machine Learning	1
Computational Intelligence	1
ACM Sigkdd Explorations Newsletter	4
International Conference on Artificial Intelligence	1
Brazilian Symposium on Artificial Intelligence	1
Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition, Structural and Syntactic Pattern Recognition	2
Korean Data Mining Conference	1
Intelligent Data Analysis in Medicine and Pharmacology	1
Journal of Biomedical Informatics	1
International Workshop on Pattern Recognition in Information Systems	1

Hence based on the literature survey, we identified the metrics used and calculated the percentage of each metric used to solve class imbalance problems (Fig. 7). It can be observed that 21% focus on ROC and F-Measure metrics and 18% focus on G-Mean. The remaining percent focuses on Precision, Recall and Accuracy Measures. Based on Fig. 7, we conclude that AUC and F-Measure performance metrics provide best result for class imbalance problem.

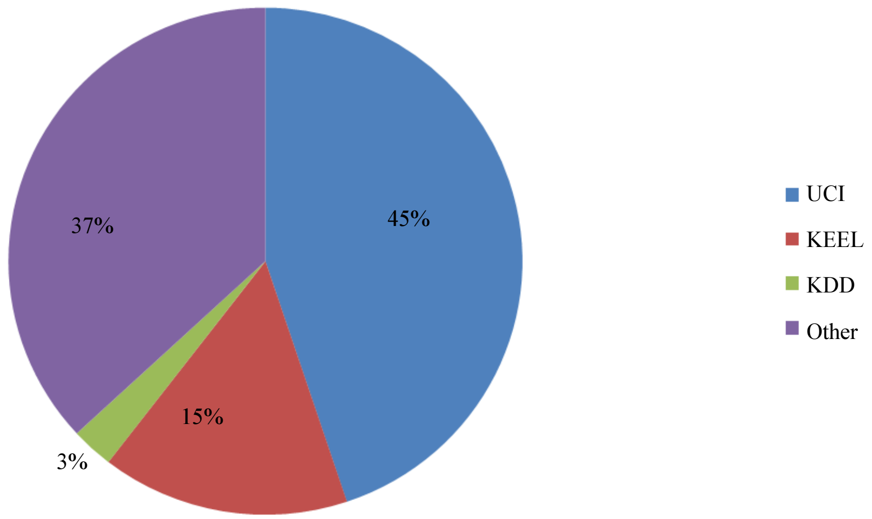

Fig. 8: Distribution of Papers in Each Repository

RQ-III.III: What are the Repositories used for Class Imbalance Problems?

We explore the distribution of all relevant publication with respect to the different repositories used. Figure 8 shows that a large majority of research papers used the University of California, Irvine (UCI) machine learning repository. The other repositories used include Keel, real world data, Promise repository and so on. Hence, this section discusses the assessing the quality of the data for class imbalance problems. Now, next section will deal with research implications and future directions.

Research Implications and Future Directions

In this section, we address the research question RQ4 and discuss the benefits and drawbacks of this SLR. Hence, this section discusses the results of the literature review on class imbalance problem. Now, next section deal with the research challenges and future directions.

Research Challenges and Future Directions

Learning from imbalanced data sets is a vital issue in machine learning. It has been observed that several research works focus on preprocessing of the data before generating a classifier which in turn provides a better solution than other methods. Thus, it allows adding new information or deleting the redundant information to balance the data. However, re-balancing the class distributions artificially does not have significant effect on the performance of the base classifier. Further, it has been observed that several research works are focused on binary class imbalance problem, ignoring multiclass imbalance problem. This may be due to the fact that most of the domain specific applications lead to binary class problems and the complexity to solve multi class problem is much higher than the binary class imbalance problem. It is believed that deriving new techniques to handle multi-class imbalance problems will be the future trend.

We also observe that most of the published works applied the state-of-the art machine learning algorithms like SVM, DT, KNN, NBC. It is also perceived that there is an increase of research on ensemble techniques to solve the problem of class imbalance. Application of two or more techniques, i.e., hybrid approach generally gives the better solution for class imbalance problem. Feature selection method can also be used for classification of imbalance data. The performance of a feature selection algorithm mostly depends on the nature of the problem.

Further, the class imbalance problem is not directly caused by class imbalance, but rather, class overlapping which may yield poor performance of a classifier. To solve the class overlapping issue, many studies have incorporated different sampling techniques during the preprocessing phase. However, these sampling techniques might not be sufficient to solve the overlapping problem. The studies highlighted that overlapping class problem is occurring by redundant features and future scope is to address different feature selection techniques to deal with class overlapping problems.

In Big Data paradigm, a massive amount of information is being generated and developing the effective solutions for processing this data is one of the major challenges in the classification domain. The big data can also be affected by class imbalance problem. Not only the tremendous increase in the volume of data but also the nature of the problem causes challenge to the existing learning algorithms. The big imbalanced data generated from various media like social networks include XML structures, images and video sequences.

Traditional learning algorithms are not designed to work with these specific type of the data. Hence, there is a need for developing an efficient and scalable algorithm for handling large and heterogeneous data. Research works on handling Big imbalanced data are focusing on running machine learning algorithms using MapReduce and Spark. However, the rapid development of Big data computing leads to an increasing demand for advanced machine learning algorithms to solve Big Imbalance problem.

Threats to Validity

The main threat to the validity of this SLR is in construction and evaluation of search string. To avoid bias, we ensure that the process of selection was enhanced by discussions on defining the research questions, inclusion and exclusion criteria and the search strategy. After the discussions, we agreed upon the search strings and the final selection strategy. In our search strategy, we have extracted relevant studies from five search databases using the constructed search string and the obtained studies were filtered using the inclusion and exclusion criteria defined earlier under Section 2.2.1 in Table 3. The search string was constructed to include a maximum number of relevant articles, but there is still a risk of missing some relevant studies due to linguistic barriers and limitation of defined inclusion and exclusion criteria. The key idea in our search strategy is to retrieve the most relevant and available literature studies on class imbalance problems without any bias. To avoid bias, we searched for common terms and combined them in our search strings for identifying the most relevant studies. Due to different perspectives and understanding of inclusion and exclusion criteria by each researcher, we obtained different findings from each researcher. To minimize the bias and increase the reliability, other researchers were called for help to achieve a consensus on the selection of studies. We further ensured the unbias by reviewing the selected studies with respect to their titles and abstracts.

Hence, this section discusses the research implications and future direction for class imbalance problem. Now, next section will deal with concluding remarks.

Concluding Remarks

This SLR described the state-of-art research in handling class imbalance problem for classification, different techniques to handle the imbalanced distribution and future directions. Existing researches have made significant progress towards balancing the skewed distribution of data. This article proposed a comprehensive taxonomy to classify and describe research efforts in this area. We have carefully followed the systematic literature review process, resulting in a broader and elaborate investigation of the literature in this area of research, comprised of 400 papers published from 2002 to 2017 (From July 2017 to December 2018, many manuscripts which are not covered in this article, will be presented in the extension of this work in near future, i.e., not added in this article due to having issue of space/many pages). Based on the results of our literature review, it is obvious that class imbalance problem has received much attention in recently published literature. The research has revealed interesting patterns, trends and gaps in the existing literature and underlined key challenges and opportunities that will shape the focus of future research efforts. We believe the results of our review will help the researchers to advance in this area and wish the taxonomy itself will become useful in the development and assessment of new research directions.

Acknowledgements

This research work is funded by the Koneru Lakshmaiah Education Foundation, and Anumit Academy's Research and Innovation Network (AARIN), India. The authors would like to thank Koneru Lakshmaiah Education Foundation and AARIN, India, an education foundation body and a research network for supporting the project through its financial assistance.

Author's Contribution

Gillala Rekha: Conceived of the work, analyzed the experimental results and drafted the manuscript.

Dr. Amit Kumar Tyagi: Draft the manuscript and designed all schemes discussed in this manuscript/article.

Dr. V. Krishna Reddy: Contributed to the original idea/to write this review article under his supervision.

Conflicts of Interests

The authors declare that there is no conflict of interests regarding the publication of this article/ paper.

Scope of this Work

The scope of this work is limited to the papers selected on the basis of interest from the reputed International Conferences and Symposiums, Springer, ACM, Elsevier, and IEEE Library etc. Several Researches are working on/solving this popular Class imbalance problem on a wide scale, so it is impossible to cite the work of each and every researcher in this field of study.

References

Abd Elrahman, S.M. and A. Abraham, 2015. Class imbalance problem using a hybrid ensemble approach. Int. J. Hybrid Intell. Syst., 12: 219-227. DOI: 10.3233/HIS-160217

Abdi, L. and S. Hashemi, 2015. To combat multi-class imbalanced problems by means of over-sampling and boosting techniques. Soft Comput., 19: 3369-3385. DOI: 10.1007/s00500-014-1291-z

Abolkarlou, N.A., A.A. Niknafs and M.K. Ebrahimpour, 2014. Ensemble imbalance classification: Using data preprocessing, clustering algorithm and genetic algorithm. Proceedings of the 4th International eConference on Computer and Knowledge Engineering, Oct. 29-30, IEEE Xplore Press, Mashhad, Iran, pp: 171-176. DOI: 10.1109/ICCKE.2014.6993364

Ai, X., J. Wu, Z. Cui, X. Xian and Y. Yao, 2016. Enrich the data density of cluster for imbalanced learning using immune representatives. Proceedings of the IEEE International Conference on the Science of Electrical Engineering, Nov. 16-18, IEEE Xplore Press, Eilat, Israel, pp: 1-5. DOI: 10.1109/ICSEE.2016.7806162

Akbani, R., S. Kwek and N. Japkowicz, 2004. Applying support vector machines to imbalanced datasets. Proceedings of the 15th European Conference on Machine Learning, Sept. 20-24, Springer, Pisa, Italy, pp: 39-50. DOI: 10.1007/978-3-540-30115-8_7

Al Helal, M., M.S. Haydar and S.A.M. Mostafa, 2016. Algorithms efficiency measurement on imbalanced data using geometric mean and cross validation. Proceedings of the International Workshop on Computational Intelligence, Dec. 12-13, IEEE Xplore Press, Dhaka, Bangladesh, pp: 110-114. DOI: 10.1109/IWCI.2016.7860349

Alejo, R., R.M. Valdovinos, V. Garcıa and J.H. Pacheco-Sanchez, 2013. A hybrid method to face class overlap and class imbalance on neural networks and multi-class scenarios. Patt. Recognit. Lett., 34: 380-388. DOI: 10.1016/j.patrec.2012.09.003

Alibeigi, M., S. Hashemi and A. Hamzeh, 2012. DBFS: An effective density based feature selection scheme for small sample size and high dimensional imbalanced data sets. Data Knowl. Eng., 81: 67-103. DOI: 10.1016/j.datak.2012.08.001

Al-Rifaie, M.M. and H.A. Alhakbani, 2016. Handling class imbalance in direct marketing dataset using a hybrid data and algorithmic level solutions. Proceedings of the SAI Computing Conference, Jul. 13-15, IEEE Xplore Press, London, UK, pp: 446-451. DOI: 10.1109/SAI.2016.7556019

Alshomrani, S., A. Bawakid, S.O. Shim, A. Fernandez and F. Herrera, 2015. A proposal for evolutionary fuzzy systems using feature weighting: Dealing with overlapping in imbalanced datasets. Knowledge-Based Syst., 73: 1-17. DOI: 10.1016/j.knosys.2014.09.002

Amin, A., S. Anwar, A. Adnan, M. Nawaz and N. Howard *et al*., 2016. Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study. IEEE Access, 4: 7940-7957. DOI: 10.1109/ACCESS.2016.2619719

Anand, R., K.G. Mehrotra, C.K. Mohan and S. Ranka, 1993. An improved algorithm for neural network classification of imbalanced training sets. IEEE Trans. Neural Netw., 4: 962-969. DOI: 10.1109/72.286891

Antonelli, M., P. Ducange and F. Marcelloni, 2014. An experimental study on evolutionary fuzzy classifiers designed for managing imbalanced datasets. Neurocomputing, 146: 125-136. DOI: 10.1016/j.neucom.2014.04.070

Antonelli, M., P. Ducange, F. Marcelloni and A. Segatori, 2013. Evolutionary fuzzy classifiers for imbalanced datasets: An experimental comparison. Proceedings of the Joint IFSA World Congress and NAFIPS Annual Meeting, Jun. 24-28, IEEE Xplore Press, Edmonton, AB, Canada, pp: 13-18. DOI: 10.1109/IFSA-NAFIPS.2013.6608367

Askan, A. and S. Sayın, 2014. SVM classification for imbalanced data sets using a multiobjective optimization framework. Annals Operat. Res., 216: 191-203. DOI: 10.1007/s10479-012-1300-5

Babar, V. and R. Ade, 2016. MLP-based undersampling technique for imbalanced learning. Proceedings of the International Conference on Automatic Control and Dynamic Optimization Techniques, Sept. 9-10, IEEE Xplore Press, Pune, India, pp: 142-147. DOI: 10.1109/ICACDOT.2016.7877567

Bader-El-Den, M., E. Teitei and M. Adda, 2016. Hierarchical classification for dealing with the class imbalance problem. Proceedings of the International Joint Conference on Neural Networks, Jul. 24-29, IEEE Xplore Press, Vancouver, BC, Canada, pp: 3584-3591. DOI: 10.1109/IJCNN.2016.7727660

Barandela, R., R. Valdovinos, J. Sanchez and F. Ferri, 2004. The imbalanced training sample problem: Under or over sampling? Proceedings of the SSPR/SPR: Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition, Structural and Syntactic Pattern Recognition, Aug. 18-20, Springer, Lisbon, Portugal, pp: 806-814. DOI: 10.1007/978-3-540-27868-9_88

Barandela, R., R.M. Valdovinos and J.S. Sanchez, 2003. New applications of ensem1180 BLES of classifiers. Patt. Anal. Applic., 6: 245-256. DOI: 10.1007/s10044-003-0192-z

Batista, G.E., R.C. Prati and M.C. Monard, 2004. A study of the behavior of several methods for balancing machine learning training data. ACM Sigkdd Explorat. Newsl., 6: 20-29. DOI: 10.1145/1007730.1007735

Batuwita, R. and V. Palade, 2010. Efficient resampling methods for training support vector machines with imbalanced datasets. Proceedings of the International Joint Conference on Neural Networks, Jul. 18-23, IEEE Xplore Press, Barcelona, Spain, pp: 1-8. DOI: 10.1109/IJCNN.2010.5596787

Bennin, K.E., J. Keung, A. Monden, Y. Kamei and N. Ubayashi, 2016. Investigating the effects of balanced training and testing datasets on effort-aware fault prediction models. Proceedings of the IEEE 40th Annual Computer Software and Applications Conference, Jun. 10-14, IEEE Xplore Press, Atlanta, GA, USA, pp: 154-163. DOI: 10.1109/COMPSAC.2016.144

Beyan, C. and R. Fisher, 2015. Classifying imbalanced data sets using similarity based hierarchical decomposition. Patt. Recognit., 48: 1653-1630. DOI: 10.1016/j.patcog.2014.10.032

Bhagat, R.C. and S.S. Patil, 2015. Enhanced SMOTE algorithm for classification of imbalanced big-data using random forest. Proceedings of the IEEE International Advance Computing Conference, IEEE Xplore Press, pp: 403-408. DOI: 10.1109/IADCC.2015.7154739

Bhowan, U., M. Johnston and M. Zhang, 2009. Differentiating between individual class performance in genetic programming fitness for classification with unbalanced data. Proceedings of the IEEE Congress on Evolutionary Computation, May 18-21, IEEE Xplore Press, Trondheim, Norway, pp: 2802-2809. DOI: 10.1109/CEC.2009.4983294

Bl aszczynski, J., M. Deckert, J. Stefanowski and S. Wilk, 2010. Integrating selective pre-processing of imbalanced data with ivotes ensemble. Proceedings of the International Conference on Rough Sets and Current Trends in Computing, (CTC' 10), Springer, pp: 148-157. DOI: 10.1007/978-3-642-13529-3_17

Blagus, R. and L. Lusa, 2012. Evaluation of SMOTE for high-dimensional class-imbalanced microarray data. Proceedings of the 11th International Conference on Machine Learning and Applications, Dec. 12-15, IEEE Xplore Press, Boca Raton, FL, USA, pp: 89-94. DOI: 10.1109/ICMLA.2012.183

Boonchuay, K., K. Sinapiromsaran and C. Lursinsap, 2011. Minority split and gain ratio for a class imbalance. Proceedings of the 8th International Conference on Fuzzy Systems and Knowledge Discovery, Jul. 26-28, IEEE Xplore Press, Shanghai, China, pp: 2060-2064. DOI: 10.1109/FSKD.2011.6019836

Braytee, A., F.K. Hussain, A. Anaissi and P.J. Kennedy, 2015. ABC-sampling for balancing imbalanced datasets based on artificial bee colony algorithm. Proceedings of the IEEE 14th International Conference on Machine Learning and Applications, Dec. 9-11, IEEE Xplore Press, Miami, FL, USA, pp: 594-599. DOI: 10.1109/ICMLA.2015.103

Breiman, L., 1996. Bagging predictors. Mach. Learn., 24: 123-140. DOI: 10.1007/BF00058655

Brereton, P., B.A. Kitchenham, D. Budgen, M. Turner and M. Khalil, 2007. Lessons from applying the systematic literature review process within the software engineering domain. J. Syst. Software, 80: 571-583. DOI: 10.1016/j.jss.2006.07.009

Bunkhumpornpat, C. and K. Sinapiromsaran, 2017. DBMUTE: Density-based majority under-sampling technique. Knowl. Inform. Syst., 50: 827-850. DOI: 10.1007/s10115-016-0957-5

Bunkhumpornpat, C. and S. Subpaiboonkit, 2013. Safe level graph for synthetic minority over-sampling techniques. Proceedings of the 13th International Symposium on Communications and Information Technologies, Sept. 4-6, IEEE Xplore Press, Surat Thani, Thailand, pp: 570-575. DOI: 10.1109/ISCIT.2013.6645923

Cao, H., V.Y. Tan and J.Z. Pang, 2014. A parsimonious mixture of Gaussian trees model for oversampling in imbalanced and multimodal time-series classification. IEEE Trans. Neural Netw. Learn. Syst., 25: 2226-2239. DOI: 10.1109/TNNLS.2014.2308321

Cao, H., X.L. Li, D.Y.K. Woon and S.K. Ng, 2013a. Integrated oversampling for imbalanced time series classification. IEEE Trans. Knowledge Data Eng., 25: 2809-2822. DOI: 10.1109/TKDE.2013.37

Cao, P., D. Zhao and O. Zaiane, 2013b. An optimized cost-sensitive SVM for imbalanced data learning. Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, (DDM' 13), Springer, pp: 280-292. DOI: 10.1007/978-3-642-37456-2_24

Cao, P., D. Zhao and O.R. Zaıane, 2013c. A PSO-based cost-sensitive neural network for imbalanced data classification. Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, (DDM' 13), Springer, pp: 452-463. DOI: 10.1007/978-3-642-40319-4_39

Cao, H., X.L. Li, Y.K. Woon and S.K. Ng, 2011. SPO: Structure preserving oversampling for imbalanced time series classification. Proceedings of the IEEE 11th International Conference on Data Mining, Dec. 11-14, IEEE Xplore Press, Vancouver, BC, Canada, pp: 1008-1013. DOI: 10.1109/ICDM.2011.137

Cao, L. and H. Shen, 2016. Combining re-sampling with twin support vector machine for imbalanced data classification. Proceedings of the 17th International Conference on Parallel and Distributed Computing, Applications and Technologies, Dec. 16-18, IEEE Xplore Press, Guangzhou, China, pp: 325-329. DOI: 10.1109/PDCAT.2016.076

Cao, L. and Y. Zhai, 2015. Imbalanced data classification based on a hybrid resampling SVM method. Proceedings of the 2015 IEEE 12th International Conference on Ubiquitous Intelligence and Computing, Aug. 10-14, IEEE Xplore Press, Beijing, China, pp: 1533-1536. DOI: 10.1109/UIC-ATC-ScalCom-CBDCom-IoP.2015.275

Cateni, S., V. Colla and M. Vannucci, 2011. Novel resampling method for the classification of imbalanced datasets for industrial and other real-world problems. Proceedings of the 11th International Conference on Intelligent Systems Design and Applications, Nov. 22-24, IEEE Xplore Press, Cordoba, Spain, pp: 402-407. DOI: 10.1109/ISDA.2011.6121689

Ceballes-Serrano, C., S. Garcia-Lopez, J. Jaramillo-Garzon and G. Castellanos-Domınguez, 2012. A strategy for classifying imbalanced data sets based on particle swarm optimization. Proceedings of the 16th Symposium of Image, Signal Processing and Artificial Vision, Sept. 12-14, IEEE Xplore Press, Antioquia, Colombia, pp: 218-222. DOI: 10.1109/STSIVA.2012.6340585

Chaıri, I., S. Alaoui and A. Lyhyaoui, 2012. Intrusion detection based sample selection for imbalanced data distribution. Proceedings of the 2nd International Conference on Innovative Computing Technology, Sept. 18-20, IEEE Xplore Press, Casablanca, Morocco, pp: 259-264. DOI: 10.1109/INTECH.2012.6457778

Chairi, I., S. Alaoui and A. Lyhyaoui, 2014. Sample selection based active learning for imbalanced data. Proceedings of the 10th International Conference on Signal-Image Technology and Internet-Based Systems, Nov. 23-27, IEEE Xplore Press, Marrakech, Morocco, pp: 645-651. DOI: 10.1109/SITIS.2014.118

Chakraborty, D., S. Saha and O. Dutta, 2017. Weighted bag hybrid multiple classifier machine for boosting prediction accuracy. Proceedings of the International Conference on High Performance Computing and Applications, Dec. 22-24, IEEE Xplore Press, Bhubaneswar, India, pp: 1-6. DOI: 10.1109/ICHPCA.2014.7045346

Chang, E.Y., B. Li, G. Wu and K. Goh, 2003. Statistical learning for effective visual information retrieval. Proceedings of the International Conference on Image Processing, Sept. 14-17, IEEE Xplore Press, Barcelona, Spain, pp: 609-612. DOI: 10.1109/ICIP.2003.1247318

Chawla, N., A. Lazarevic, L. Hall and K. Bowyer, 2003. SMOTEboost: Improving prediction of the minority class in boosting. Proceedings of the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases, (KDD’ 03), Springer, pp: 107-119. DOI: 10.1007/978-3-540-39804-2_12

Chawla, N.V., K.W. Bowyer, L.O. Hall and W.P. Kegelmeyer, 2002. SMOTE: Synthetic minority over-sampling technique. J. Artificial Intell. Res., 16: 321-357. DOI: 10.1613/jair.953

Chawla, N.V., N. Japkowicz and A. Kotcz, 2004. Editorial: Special issue on learning from imbalanced data sets. SIGKDD Explor. Newsl., 6: 1-6. DOI: 10.1145/1007730.1007733

Chen, L., B. Fang, Z. Shang and Y. Tang, 2016. Tackling class overlap and imbalance problems in software defect prediction. Software Quality J., 2016: 1-29. DOI: 10.1007/s11219-016-9342-6

Chen, S., G. Guo and L. Chen, 2010a. A new over-sampling method based on cluster ensembles. Proceedings of the IEEE 24th International Conference on Advanced Information Networking and Applications Workshops, Apr. 20-23, IEEE Xplore Press, Perth, WA, Australia, pp: 599-604. DOI: 10.1109/WAINA.2010.40

Chen, L., Z. Cai, L. Chen and Q. Gu, 2010b. A novel differential evolution-clustering hybrid resampling algorithm on imbalanced datasets. Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, Jan. 9-10, IEEE Xplore Press, Phuket, Thailand, pp: 81-85. DOI: 10.1109/WKDD.2010.48

Chen, N., B. Ribeiro, A.S. Vieira, J. Duarte and J.C. Neves, 2010c. Hybrid genetic algorithm and learning vector quantization modeling for cost-sensitive bankruptcy prediction. Proceedings of the 2nd International Conference on Machine Learning and Computing, Feb. 9-11, IEEE Xplore Press, Bangalore, India, pp: 213-217. DOI: 10.1109/ICMLC.2010.29

Chen, N., A. Vieira and J. Duarte, 2009. Cost-sensitive LVQ for bankruptcy prediction: An empirical study. Proceedings of the 2nd IEEE International Conference on Computer Science and Information Technology, Aug. 8-11, IEEE Xplore Press, Beijing, China, pp: 115-119. DOI: 10.1109/ICCSIT.2009.5234441

Chen, S. and H. He, 2009. SERA: Selectively recursive approach towards nonstationary imbalanced stream data mining. Proceedings of the International Joint Conference on Neural Networks, Jun. 14-19, IEEE Xplore Press, Atlanta, GA, USA, pp: 522-529. DOI: 10.1109/IJCNN.2009.5178874

Chen, Z., Q. Yan, H. Han, S. Wang and L. Peng *et al.*, 2018. Machine learning based mobile malware detection using highly imbalanced network traffic. Inform. Sci., 433-434: 346-364. DOI: 10.1016/j.ins.2017.04.044

Cheng, D. and M. Wu, 2016. A novel classifier-weighted features cost-sensitive SVM. Proceedings of the IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), Dec. 15-18, IEEE Xplore, Chengdu, China, pp: 598-603. DOI: 10.1109/iThings-GreenCom-CPSCom-SmartData.2016.133

Chetchotsak, D., S. Pattanapairoj and B. Arnonkijpanich, 2015. Integrating new data balancing technique with committee networks for imbalanced data: Grsom approach. Cognitive Neurodynam., 9: 627-638. DOI: 10.1007/s11571-015-9350-4

Chira, C. and C. Lemnaru, 2015. A multi-objective evolutionary approach to imbalanced classification problems. Proceedings of the IEEE International Conference on Intelligent Computer Communication and Processing, Sept. 3-5, IEEE Xplore Press, Cluj-Napoca, Romania, pp: 149-154. DOI: 10.1109/ICCP.2015.7312620

Cieslak, D.A. and N.V. Chawla, 2008. Start globally, optimize locally, predict globally: Improving performance on imbalanced data. Proceedings of the 8th IEEE International Conference on Data Mining, Dec. 15-19, IEEE Xplore Press, Pisa, Italy, pp: 143-152. DOI: 10.1109/ICDM.2008.87

Cohen, G., M. Hilario and C. Pellegrini, 2004. One-class support vector machines with a conformal kernel: A case study in handling class imbalance. Proceedings of the Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition, Aug. 18-20, Springer, Lisbon, Portugal, pp: 850-858. DOI: 10.1007/978-3-540-27868-9_93

Cohen, G., M. Hilario, H. Sax and S. Hugonnet, 2003. Asymmetrical margin approach to surveillance of nosocomial infections using support vector classification.

Cohen, W.W., 1995. Fast effective rule induction. Proceedings of the 12th International Conference on Machine Learning, Jul. 9-12, Tahoe City, California, pp: 115-123. DOI: 10.1016/B978-1-55860-377-6.50023-2

Cuaya, G., A. Munoz-Melendez and E. Morales, 2011. A minority class feature selection method. Progress Patt. Recognit. Image Anal. Comput. Vis. Applic. DOI: 10.1007/978-3-642-25085-9_49

Dai, H.L., 2015. Imbalanced protein data classification using ensemble FTM-SVM. IEEE Trans. Nanobiosc., 14: 350-359. DOI: 10.1109/TNB.2015.2431292

Dang, X.T., D.H. Tran, O. Hirose and K. Satou, 2015. SPY: A novel resampling method for improving classification performance in imbalanced data. Proceedings of the 7th International Conference on Knowledge and Systems Engineering, Oct. 8-10, IEEE Xplore Press, Ho Chi Minh City, Vietnam, pp: 280-285. DOI: 10.1109/KSE.2015.24

Dang, X.T., O. Hirose, D.H. Bui, T. Saethang and V.A. Tran *et al*., 2013. A novel over-sampling method and its application to cancer classification from gene expression data. Chem-Bio Inform. J., 13: 19-29. DOI: 10.1273/cbij.13.19

Das, B., D.J. Cook, M. Schmitter-Edgecombe and A.M. Seelye, 2012. Puck: An automated prompting system for smart environments: Toward achieving automated prompting challenges involved. Personal Ubiquitous Comput., 16: 859-873. DOI: 10.1007/s00779-011-0445-6

Das, B., N.C. Krishnan and D.J. Cook, 2013a. Handling class overlap and imbalance to detect prompt situations in smart homes. Proceedings of the IEEE 13th International Conference on Data Mining Workshops, Dec. 7-10, IEEE Xplore Press, Dallas, TX, USA, pp: 266-273. DOI: 10.1109/ICDMW.2013.18

Das, B., N.C. Krishnan and D.J. Cook, 2013b. wRACOG: A Gibbs sampling-based oversampling technique. Proceedings of the IEEE 13th International Conference on Data Mining, Dec. 7-10, IEEE Xplore Press, Dallas, TX, USA, pp: 111-120. DOI: 10.1109/ICDM.2013.18

Dash, S., 2016. A diverse meta learning ensemble technique to handle imbalanced microarray dataset. Proceedings of the 7th World Congress on Nature and Biologically Inspired Computing, (BIC' 16), Springer, pp: 1-13. DOI: 10.1007/978-3-319-27400-3_1

de Souza, J.C.S., S.G. Claudino, R. da Silva Simoes, P.R. Oliveira and K.M. Honorio, 2016. Recent advances for handling imbalancement and uncertainty in labelling in medicinal chemistry data analysis. Proceedings of the SAI Computing Conference, Jul. 13-15, IEEE Xplore Press, London, UK, pp: 217-222. DOI: 10.1109/SAI.2016.7555985

Debowski, B., S. Areibi, G. Grewal and J. Tempelman, 2012. A dynamic sampling framework for multi-class imbalanced data. Proceedings of the 11th International Conference on Machine Learning and Applications, Dec. 12-15, IEEE Xplore Press, Boca Raton, FL, USA, pp: 113-118. DOI: 10.1109/ICMLA.2012.144

Del Rıo, S., V. Lopez, J.M. Benıtez and F. Herrera, 2014. On the use of mapreduce for imbalanced big data using random forest. Inform. Sci., 285: 112-137. DOI: 10.1016/j.ins.2014.03.043

Demidova, L. and I. Klyueva, 2017. SVM classification: Optimization with the SMOTE algorithm for the class imbalance problem. Proceedings of the 6th Mediterranean Conference on Embedded Computing, Jun. 11-15, IEEE Xplore Press, Bar, Montenegro, pp: 1-4. DOI: 10.1109/MECO.2017.7977136

Dıez-Pastor, J.F., J.J. Rodrıguez, C.I. Garcıa-Osorio and L.I. Kuncheva, 2015a. Diversity techniques improve the performance of the best imbalance learning ensembles. Inform. Sci., 325: 98-117. DOI: 10.1016/j.ins.2015.07.025

Dıez-Pastor, J.F., J.J. Rodrıguez, C. Garcıa-Osorio and L.I. Kuncheva, 2015b. Random balance: Ensembles of variable priors classifiers for imbalanced data. Knowledge-Based Syst., 85: 96-111. DOI: 10.1016/j.knosys.2015.04.022

Do, T.N., 2014. Parallel multiclass stochastic gradient descent algorithms for classifying million images with very-high-dimensional signatures into thousands classes. Vietnam J. Comput. Sci., 1: 107-115. DOI: 10.1007/s40595-013-0013-2

Domingos, P., 1999. Metacost: A general method for making classifiers costsensitive. Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (DDM' 99), ACM, pp: 155-164.

Drummond, C. and R.C. Holte, 2003. C4. 5, class imbalance and cost sensitivity: Why under-sampling beats over-sampling. Proceedings of the International Conference on Machine Learning: Workshop on Learning from Imbalanced Datasets II, (LID' 03), Citeseer Washington DC.

Duhaney, J., T.M. Khoshgoftaar and A. Napolitano, 2012. Studying the effect of class imbalance in ocean turbine fault data on reliable state detection. Proceedings of the 11th International Conference on Machine Learning and Applications, Dec. 12-15, IEEE Xplore Press, Boca Raton, FL, USA, pp: 268-275. DOI: 10.1109/ICMLA.2012.53

Dwiyanti, E. and A.A. Adiwijaya, 2017. Handling imbalanced data in churn prediction using rusboost and feature selection (case study: Pt.telekomunikasi Indonesia regional 7). Proceedings of the 2nd International Conference on Soft Computing and Data Mining, (CDM' 17), Springer, pp: 376-385. DOI: 10.1007/978-3-319-51281-5_38

Effendy, V. and Z.A. Baizal, 2014. Handling imbalanced data in customer churn prediction using combined sampling and weighted random forest. Proceedings of the 2nd International Conference on Information and Communication Technology, May 28-30, IEEE Xplore Press, Bandung, Indonesia, pp: 325-330. DOI: 10.1109/ICoICT.2014.6914086

El-Ghamrawy, S.M., 2016. A knowledge management framework for imbalanced data using frequent pattern mining based on bloom filter. Proceedings of the 11th International Conference on Computer Engineering and Systems, Dec. 20-21, IEEE Xplore Press, Cairo, Egypt, pp: 226-231. DOI: 10.1109/ICCES.2016.7822004

Elkan, C., 2001. The foundations of cost-sensitive learning. Proceedings of the 7th International Joint Conference on Artificial Intelligence, (CAI' 99), pp: 973-978.

Estabrooks, A., T. Jo and N. Japkowicz, 2004. A multiple resampling method for learning from imbalanced data sets. Comput. Intell., 20: 18-36. DOI: 10.1111/j.0824-7935.2004.t01-1-00228.x

Ezawa, K.J., M. Singh and S.W. Norton, 1996. Learning goal oriented Bayesian networks for telecommunications risk management. Proceedings of the International Conference on Machine Learning, (CML' 96), Morgan Kaufmann, pp: 139-147.

Fan, Q., Z. Wang and D. Gao, 2016. One-sided dynamic undersampling nopropagation neural networks for imbalance problem. Eng. Applic. Artificial Intell., 53: 62-73. DOI: 10.1016/j.engappai.2016.02.011

Fan, X. and K. Tang, 2010. Enhanced maximum AUC linear classifier. Proceedings of the 7th International Conference on Fuzzy Systems and Knowledge Discovery, Aug. 10-12, IEEE Xplore Press, Yantai, China, pp: 1540-1544. DOI: 10.1109/FSKD.2010.5569339

Fan, Y., Z. Kai and L. Qiang, 2014. A revisit to the class imbalance learning with linear support vector machine. Proceedings of the 9th International Conference on Computer Science and Education, Aug. 22-24, IEEE Xplore Press, Vancouver, BC, Canada, pp: 516-521. DOI: 10.1109/ICCSE.2014.6926515

Farajzadeh-Zanjani, M., R. Razavi-Far and M. Saif, 2016. Efficient sampling techniques for ensemble learning and diagnosing bearing defects under class imbalanced condition. Proceedings of the IEEE Symposium Series on Computational Intelligence, Dec. 6-9, IEEE Xplore Press, Athens, Greece, pp: 1-7. DOI: 10.1109/SSCI.2016.7849879

Fawcett, T. and F.J. Provost, 1996. Combining data mining and machine learning for effective user profiling. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, (DDM' 96), pp: 8-13.

Fergani, B. and L. Clavier, 2013. Importance-weighted the imbalanced data for C-SVM classifier to human activity recognition. Proceedings of the 8th International Workshop on Systems, Signal Processing and their Applications, May 12-15, IEEE Xplore Press, Algiers, Algeria, pp: 330-335. DOI: 10.1109/WoSSPA.2013.6602386

Fernandez, A., M.J. del Jesus and F. Herrera, 2009. Hierarchical fuzzy rule based classification systems with genetic rule selection for imbalanced data-sets. Int. J. Approximate Reason., 50: 561-577. DOI: 10.1016/j.ijar.2008.11.004

Fernandez, A., M.J. del Jesus and F. Herrera, 2010. On the 2-tuples based genetic tuning performance for fuzzy rule based classification systems in imbalanced data-sets. Inform. Sci., 180: 1268-1291. DOI: 10.1016/j.ins.2009.12.014

Fernandez, A., S. del Rıo, N.V. Chawla and F. Herrera, 2017. An insight into imbalanced big data classification: Outcomes and challenges. Complex Intell. Syst.

FernaNdez, A., V. LoPez, M. Galar, M.J. Del Jesus and F. Herrera, 2013. Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches. Knowledge-Based Syst., 42: 97-110. DOI: 10.1016/j.knosys.2013.01.018

Fu, Y., H. Zhang, Y. Bai and W. Sun, 2016. An under-sampling method: Based on principal component analysis and comprehensive evaluation model. Proceedings of the IEEE International Conference on Software Quality, Reliability and Security Companion, Aug. 1-3, IEEE Xplore Press, Vienna, Austria, pp: 414-415. DOI: 10.1109/QRS-C.2016.68

Galar, M., A. Fernandez, E. Barrenechea and F. Herrera, 2013. Eusboost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling. Patt. Recognit., 46: 3460-3471. DOI: 10.1016/j.patcog.2013.05.006

Galar, M., A. Fernandez, E. Barrenechea, H. Bustince and F. Herrera, 2012. A review on ensembles for the class imbalance problem: Bagging-, boosting- and hybrid-based approaches. IEEE Trans. Syst. Man Cybernet., 42: 463-484. DOI: 10.1109/TSMCC.2011.2161285

Gao, K., T. Khoshgoftaar and A. Napolitano, 2013. Improving software quality estimation by combining boosting and feature selection. Proceedings of the 12th International Conference on Machine Learning and Applications, Dec. 4-7, IEEE Xplore Press, Miami, FL, USA, pp: 27-33. DOI: 10.1109/ICMLA.2013.13

Gao, M., X. Hong, S. Chen and C.J. Harris, 2012. Probability density function estimation based over-sampling for imbalanced two-class problems. Proceedings of the International Joint Conference on Neural Networks, Jun. 10-15, IEEE Xplore Press, Brisbane, QLD, Australia, pp: 1-8. DOI: 10.1109/IJCNN.2012.6252384

Garcıa, S. and F. Herrera, 2008. Evolutionary training set selection to optimize C4.5 in imbalanced problems. Proceedings of the 8th International Conference on Hybrid Intelligent Systems, Sept. 10-12, IEEE Xplore Press, Barcelona, Spain, pp: 567-572. DOI: 10.1109/HIS.2008.67

Garcıa, S., A. Fernandez and F. Herrera, 2009. Enhancing the effectiveness and interpretability of decision tree and rule induction classifiers with evolutionary training set selection over imbalanced problems. Applied Soft Comput., 9: 1304-1314. DOI: 10.1016/j.asoc.2009.04.004

Garcıa, S., J. Cano, A. Fernandez and F. Herrera, 2006. A proposal of evolutionary prototype selection for class imbalance problems. Intell. Data Eng. Automated Learn., 2006: 1415-1423. DOI: 10.1007/11875581_168

Garcıa, V., J.S. Sanchez and R.A. Mollineda, 2012. On the effectiveness of preprocessing methods when dealing with different levels of class imbalance. Knowledge-Based Syst., 25: 13-21. DOI: 10.1016/j.knosys.2011.06.013

Gazzah, S., A. Hechkel and N.E.B. Amara, 2015. A hybrid sampling method for imbalanced data. Proceedings of the 12th International Multi-Conference on Systems, Signals and Devices, Mar. 16-19, IEEE Xplore Press, Mahdia, Tunisia, pp: 1-6. DOI: 10.1109/SSD.2015.7348093

Ghazikhani, A., R. Monsefi and H.S. Yazdi, 2012a. SVBO: Support vector-based oversampling for handling class imbalance in k-NN. Proceedings of the 20th Iranian Conference on Electrical Engineering, May 15-17, IEEE Xplore Press, Tehran, Iran, pp: 605-610. DOI: 10.1109/IranianCEE.2012.6292427

Ghazikhani, A., H.S. Yazdi and R. Monsefi, 2012b. Class imbalance handling using wrapper-based random oversampling. Proceedings of the 20th Iranian Conference on Electrical Engineering, May 15-17, IEEE Xplore Press, Tehran, Iran, pp: 611-616. DOI: 10.1109/IranianCEE.2012.6292428

Ghosh, T., D. Sarkar, T. Sharma, A. Desai and R. Bali, 2016. Real time failure prediction of load balancers and firewalls. Proceedings of the IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), Dec. 15-18, IEEE Xplore Press, Chengdu, China, pp: 822-827. DOI: 10.1109/iThings-GreenCom-CPSCom-SmartData.2016.171

Guo, H. and H.L. Viktor, 2004. Learning from imbalanced data sets with boosting and data generation: The databoost-im approach. SIGKDD Explor. Newsl., 6: 30-39. DOI: 10.1145/1007730.1007736

Ha, J. and J.S. Lee, 2016. A new under-sampling method using genetic algorithm for imbalanced data classification. Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication, (IMC' 16), ACM, pp: 95-101.

Haldankar, A.N. and K. Bhowmick, 2016. A cost sensitive classifier for big data. Proceedings of the IEEE International Conference on Advances in Electronics, Communication and Computer Technology, Dec. 2-3, IEEE Xplore Press, Pune, India, pp: 122-127. DOI: 10.1109/ICAECCT.2016.7942567

Hamdi, F., M. Lebbah and Y. Bennani, 2010. Topographic under-sampling for unbalanced distributions. Proceedings of the International Joint Conference on Neural Networks, Jul. 18-23, IEEE Xplore Press, Barcelona, Spain, pp: 1-6. DOI: 10.1109/IJCNN.2010.5596294

Hart, P., 1968. The condensed nearest neighbor rule. IEEE Trans. Inform. Theory, 14: 515-516. DOI: 10.1109/TIT.1968.1054155

He, G., H. Han and W. Wang, 2005. An over-sampling expert system for learning from imbalanced data sets. Proceedings of the International Conference on Neural Networks and Brain, Oct. 13-15, IEEE Xplore Press, Beijing, China, pp: 537-541. DOI: 10.1109/ICNNB.2005.1614671

Hido, S., H. Kashima and Y. Takahashi, 2009. Roughly balanced bagging for imbalanced data. Stat. Anal. Data Mining: ASA Data Sci. J., 2: 412-426. DOI: 10.1002/sam.10061

Hinojosa, C.E., H.A. Camargo and V.Y.J. Tupac, 2015. Learning fuzzy classification rules from imbalanced datasets using multi-objective evolutionary algorithm. Proceedings of the Latin America Congress on Computational Intelligence, Oct. 13-16, IEEE Xplore Press, Curitiba, Brazil, pp: 1-6. DOI: 10.1109/LA-CCI.2015.7435959

Hosseinzadeh, M. and M. Eftekhari, 2015a. Improving rotation forest performance for imbalanced data classification through fuzzy clustering. Proceedings of the International Symposium on Artificial Intelligence and Signal Processing, Mar. 3-5, IEEE Xplore Press, Mashhad, Iran, pp: 35-40. DOI: 10.1109/AISP.2015.7123535

Hosseinzadeh, M. and M. Eftekhari, 2015b. Using fuzzy undersampling and fuzzy PCA to improve imbalanced classification through rotation forest algorithm. Proceedings of the International Symposium on Computer Science and Software Engineering, Aug. 18-19, IEEE Xplore Press, Tabriz, Iran, pp: 1-7. DOI: 10.1109/CSICSSE.2015.7369242

Hu, J., Y. Li, W.X. Yan, J.Y. Yang and H.B. Shen *et al.*, 2016a. KNN-based dynamic query-driven sample rescaling strategy for class imbalance learning. Neurocomputing, 191: 363-373. DOI: 10.1016/j.neucom.2016.01.043

Hu, G., T. Xi, F. Mohammed and H. Miao, 2016b. Classification of wine quality with imbalanced data. Proceedings of the IEEE International Conference on Industrial Technology, Mar. 14-17, IEEE Xplore Press, Taipei, Taiwan, pp: 1712-1217. DOI: 10.1109/ICIT.2016.7475021

Hu, J., 2012. Active learning for imbalance problem using L-GEM of RBFNN. Proceedings of the International Conference on Machine Learning and Cybernetics, May 15-17, IEEE Xplore Press, Xian, China, pp: 490-495. DOI: 10.1109/ICMLC.2012.6358972

Hu, S., Y. Liang, L. Ma and Y. He, 2009. MSMOTE: Improving classification performance when training data is imbalanced. Proceedings of the 2nd International Workshop on Computer Science and Engineering, Oct. 28-30, IEEE Xplore Press, Qingdao, China, pp: 13-17. DOI: 10.1109/WCSE.2009.756

Hu, X.S. and R.J. Zhang, 2013. Clustering-based subset ensemble learning method for imbalanced data. Proceedings of the International Conference on Machine Learning and Cybernetics, Jul. 14-17, IEEE Xplore Press, Tianjin, China, pp: 35-39. DOI: 10.1109/ICMLC.2013.6890440

Huang, H.Y., Y.J. Lin, Y.S. Chen and H.Y. Lu, 2012. Imbalanced data classification using random subspace method and SMOTE. Proceedings of the Joint 6th International Conference on Soft Computing and Intelligent Systems (SCIS) and 13th International Symposium on Advanced Intelligent Systems, Nov. 20-24, IEEE Xplore Press, Kobe, Japan, pp: 817-820. DOI: 10.1109/SCIS-ISIS.2012.6505155

Huang, Q. and X. Zhang, 2016. An improved ensemble learning method with SMOTE for protein interaction hot spots prediction. Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine, Dec. 15-18, IEEE Xplore Press, Shenzhen, China, pp: 1584-1589. DOI: 10.1109/BIBM.2016.7822756

Huang, W., M. Li, W. Hu, G. Song and X. Xing *et al.*, 2013. Cost sensitive GPS-based activity recognition. Proceedings of the 10th International Conference on Fuzzy Systems and Knowledge Discovery, Jul. 23-25, IEEE Xplore Press, Shenyang, China, pp: 962-966. DOI: 10.1109/FSKD.2013.6816334

Hunt, R., M. Johnston, W.N. Browne and M. Zhang, 2010. Sampling methods in genetic programming for classification with unbalanced data. Proceedings of the Australasian Conference on Artificial Intelligence, (CAI' 10), Springer, pp: 273-282.

Imam, T., K.M. Ting and J. Kamruzzaman, 2006. Z-SVM: An SVM for improved classification of imbalanced data. Proceedings of the AI: Australasian Joint Conference on Artificial Intelligence, (CAI' 06), Springer, pp: 264-273. DOI: 10.1007/11941439_30

Japkowicz, N. and S. Stephen, 2002. The class imbalance problem: A systematic study. Intell. Data Anal., 6: 429-449. DOI: 10.3233/IDA-2002-6504

Jatoth, C., G.R. Gangadharan and R. Buyya, 2017. Computational intelligence based QOS-aware web service composition: A systematic literature review. IEEE Trans. Services Comput., 10: 475-492. DOI: 10.1109/TSC.2015.2473840

Jiang, K., J. Lu and K. Xia, 2016. A novel algorithm for imbalance data classification based on genetic algorithm improved SMOTE. Arab. J. Sci. Eng., 41: 3255-3266. DOI: 10.1007/s13369-016-2179-2

Jiang, L., C. Li, Z. Cai and H. Zhang, 2013. Sampled Bayesian network classifiers for class-imbalance and cost-sensitive learning. Proceedings of the IEEE 25th International Conference on Tools with Artificial Intelligence, Nov. 4-6, IEEE Xplore Press, Herndon, VA, USA, pp: 512-517. DOI: 10.1109/ICTAI.2013.82

Jin, O., L. Qu, J. He and X. Li, 2015. Recognition of new and old banknotes based on SMOTE and SVM. Proceedings of the IEEE 12th International Conference on Ubiquitous Intelligence and Computing, Aug. 10-14, IEEE Xplore Press, Beijing, China, pp: 213-220. DOI: 10.1109/UIC-ATC-ScalCom-CBDCom-IoP.2015.53

Jindaluang, W., V. Chouvatut and S. Kantabutra, 2014. Under-sampling by algorithm with performance guaranteed for class-imbalance problem. Proceedings of the International Conference on Computer Science and Engineering, Jul. 30-Aug. 1, IEEE Xplore Press, Khon Kaen, Thailand, pp: 215-221. DOI: 10.1109/ICSEC.2014.6978197

Joshi, M.V., V. Kumar and R.C. Agarwal, 2001. Evaluating boosting algorithms to classify rare classes: Comparison and improvements. Proceedings of the IEEE International Conference on Data Mining, Nov. 29-Dec. 2, IEEE Xplore Press, San Jose, CA, USA, pp: 257-264. DOI: 10.1109/ICDM.2001.989527

Kamal, A.H., X. Zhu, A. Pandya and S. Hsu, 2009. Feature selection with biased sample distributions. Proceedings of the IEEE International Conference on Information Reuse and Integration, Aug. 10-12, IEEE Xplore Press, Las Vegas, NV, USA, pp: 23-28. DOI: 10.1109/IRI.2009.5211613

Kamei, Y., A. Monden, S. Matsumoto, T. Kakimoto and K.I. Matsumoto, 2007. The effects of over and under sampling on fault-prone module detection. Proceedings of the 1st International Symposium on Empirical Software Engineering and Measurement, Sept. 20-21, IEEE Xplore Press, Madrid, Spain, pp: 196-204. DOI: 10.1109/ESEM.2007.28

Kaur, A., K. Kaur and S. Jain, 2016. Predicting software change-proneness with code smells and class imbalance learning. Proceedings of the International Conference on Advances in Computing, Communications and Informatics, Sept. 21-24, IEEE Xplore Press, Jaipur, India, pp: 746-754. DOI: 10.1109/ICACCI.2016.7732136

Kim, H.J., N.O. Jo and K.S. Shin, 2016. Optimization of cluster-based evolutionary undersampling for the artificial neural networks in corporate bankruptcy prediction. Expert Syst. Applic., 59: 226-234. DOI: 10.1016/j.eswa.2016.04.027

Kocyigit, Y. and H. Seker, 2012. Imbalanced data classifier by using ensemble fuzzy c-means clustering. Proceedings of the IEEE-EMBS International Conference on Biomedical and Health Informatics, Jan 5-7, IEEE Xplore Press, Hong Kong, China, pp: 952-955. DOI: 10.1109/BHI.2012.6211746

Koto, F., 2014. SMOTE-out, SMOTE-cosine and selected-SMOTE: An enhancement strategy to handle imbalance in data level. Proceedings of the International Conference on Advanced Computer Science and Information Systems, Oct. 18-19, IEEE Xplore Press, Jakarta, Indonesia, pp: 280-284. DOI: 10.1109/ICACSIS.2014.7065849

Krawczyk, B. and G. Schaefer, 2013. An analysis of properties of malignant cases for imbalanced breast thermogram feature classification. Proceedings of the 2nd IAPR Asian Conference on Pattern Recognition, Nov. 5-8, IEEE Xplore Press, Naha, Japan, pp: 305-309. DOI: 10.1109/ACPR.2013.45

Krawczyk, B., G. Schaefer and M. Wozniak, 2013. A cost-sensitive ensemble classifier for breast cancer classification. Proceedings of the IEEE 8th International Symposium on Applied Computational Intelligence and Informatics, May 23-25, IEEE Xplore Press, Timisoara, Romania, pp: 427-430. DOI: 10.1109/SACI.2013.6609012

Kriminger, E., J.C. Prıncipe and C. Lakshminarayan, 2012. Nearest neighbor distributions for imbalanced classification. Proceedings of the International Joint Conference on Neural Networks, Jun. 10-15, IEEE Xplore Press, Brisbane, QLD, Australia, pp: 1-5. DOI: 10.1109/IJCNN.2012.6252718

Kubat, M. and S. Matwin, 1997. Addressing the curse of imbalanced training sets: One-sided selection. Proceedings of the 14th International Conference on Machine Learning, (CML' 97), pp: 179-186.

Kubat, M., R.C. Holte and S. Matwin, 1998. Machine learning for the detection of oil spills in satellite radar images. Mach. Learn., 30: 195-215. DOI: 10.1023/A:1007452223027

Laurikkala, J., 2001. Improving identification of difficult small classes by balancing class distribution. Artificial Intell. Med.

Lee, J., N.R. Kim and J.H. Lee, 2015. An over-sampling technique with rejection for imbalanced class learning. Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication, (IMC' 15), ACM, pp: 102-108.

Lessmann, S., 2004. Solving imbalanced classification problems with support vector machines. Proceedings of the International Conference on Artificial Intelligence, Jun. 21-24, Las Vegas, Nevada, USA, pp: 214-220.

Lewis, D.D. and J. Catlett, 1994. Heterogeneous uncertainty sampling for supervised learning. Proceedings of the 11th International Conference on Machine Learning, Jul. 10-13, New Brunswick, pp: 148-156.
DOI: 10.1016/B978-1-55860-335-6.50026-X

Li, C., 2007. Classifying imbalanced data using a Bagging Ensemble Variation (BEV). Proceedings of the 45th Annual Southeast Regional Conference, (SRC' 07), ACM, pp: 203-208.
DOI: 10.1145/1233341.1233378

Li, J., H. Li and J.L. Yu, 2011. Application of random-SMOTE on imbalanced data mining. Proceedings of the 4th International Conference on Business Intelligence and Financial Engineering, Oct. 17-18, IEEE Xplore Press, Wuhan, China, pp: 130-133. DOI: 10.1109/BIFE.2011.25

Li, X., B. Zou, L. Wang, M. Zeng and K. Yue *et al*., 2015a. A novel lasso-based feature weighting selection method for microarray data classification. Proceedings of the IET International Conference on Biomedical Image and Signal Processing, Nov. 19-19, IEEE Xplore Press, Beijing, China, pp: 1-5. DOI: 10.1049/cp.2015.0795

Li, J., S. Fong and Y. Zhuang, 2015b. Optimizing SMOTE by metaheuristics with neural network and decision tree. Proceedings of the 3rd International Symposium on Computational and Business Intelligence, Dec. 7-9, IEEE Xplore Press, Bali, Indonesia, pp: 26-32. DOI: 10.1109/ISCBI.2015.12

Li, K., W. Zhang, Q. Lu and X. Fang, 2014. An improved SMOTE imbalanced data classification method based on support degree. Proceedings of the International Conference on Identification, Information and Knowledge in the Internet of Things, Oct. 17-18, IEEE Xplore Press, Beijing, China, pp: 34-38. DOI: 10.1109/IIKI.2014.14

Li, P., P.L. Qiao and Y.C. Liu, 2008. A hybrid re-sampling method for SVM learning from imbalanced data sets. Proceedings of the 5th International Conference on Fuzzy Systems and Knowledge Discovery, Oct. 18-20, IEEE Xplore Press, Shandong, China, pp: 65-69.
DOI: 10.1109/FSKD.2008.407

Li, Q., B. Yang, Y. Li, N. Deng and L. Jing, 2013. Constructing support vector machine ensemble with segmentation for imbalanced datasets. Neural Comput. Applic., 22: 249-256.
DOI: 10.1007/s00521-012-1041-z

Li, S. and W. Deng, 2016. Real world expression recognition: A highly imbalanced detection problem. Proceedings of the International Conference on Biometrics, Jun. 13-16, IEEE Xplore Press, Halmstad, Sweden, pp: 1-6.
DOI: 10.1109/ICB.2016.7550074

Liang, G., X. Zhu and C. Zhang, 2014. The effect of varying levels of class distribution on bagging for different algorithms: An empirical study. Int. J. Mach. Learn. Cybernet., 5: 63-71.
DOI: 10.1007/s13042-012-0125-5

Lim, P., C.K. Goh and K.C. Tan, 2017. Evolutionary cluster-based synthetic oversampling ensemble (eco-ensemble) for imbalance learning. IEEE Trans. Cybernet., 47: 2850-2861.
DOI: 10.1109/TCYB.2016.2579658

Lin, S.C., I.C. Yuan-Chin and W.N. Yang, 2009. Meta-learning for imbalanced data and classification ensemble in binary classification. Neurocomputing, 73: 484-494. DOI: 10.1016/j.neucom.2009.06.015

Lin, Y.B., X.O. Ping, T.W. Ho and F. Lai, 2014. Processing and analysis of imbalanced liver cancer patient data by case-based reasoning. Proceedings of the 7th Biomedical Engineering International Conference, Nov. 26-28, IEEE Xplore Press, Fukuoka, Japan, pp: 1-5.
DOI: 10.1109/BMEiCON.2014.7017371

Liu, Q., X. Wei, R. Huang, H. Meng and Y. Qian, 2016b. Improved AdaBoost model for user's QOE in imbalanced dataset. Proceedings of the 8th International Conference on Wireless Communications and Signal Processing, Oct. 13-15, IEEE Xplore Press, Yangzhou, China, pp: 1-5.
DOI: 10.1109/WCSP.2016.7752682

Liu, R., R. Huang, Y. Qian, X. Wei and P. Lu, 2016a. Improving user's quality of experience in imbalanced dataset. Proceedings of the International Conference on Wireless Communications and Mobile Computing Conference, Sept. 5-9, IEEE Xplore Press, Paphos, Cyprus, pp: 644-649.
DOI: 10.1109/IWCMC.2016.7577132

Liu, X.Y., J. Wu and Z.H. Zhou, 2009. Exploratory undersampling for classimbalance learning. IEEE Trans. Syst. Man Cybernet., 39: 539-550.
DOI: 10.1109/TSMCB.2008.2007853

Lopez, V., A. Fernandez and F. Herrera, 2010. A first approach for cost-sensitive classification with linguistic genetic fuzzy systems in imbalanced datasets. Proceedings of the 10th International Conference on Intelligent Systems Design and Applications, Nov. 29-Dec. 1, IEEE Xplore Press, Cairo, Egypt, pp: 676-681.
DOI: 10.1109/ISDA.2010.5687187

Lopez, V., A. Fernandez and F. Herrera, 2013. Addressing covariate shift for genetic fuzzy systems classifiers: A case of study with FARC-HD for imbalanced datasets. Proceedings of the IEEE International Conference on Fuzzy Systems, Jul. 7-10, IEEE Xplore Press, Hyderabad, India, pp: 1-8. DOI: 10.1109/FUZZ-IEEE.2013.6622396

Lopez, V., A. Fernandez and F. Herrera, 2014. On the importance of the validation technique for classification with imbalanced datasets: Addressing covariate shift when data is skewed. Inform. Sci., 257: 1-13. DOI: 10.1016/j.ins.2013.09.038

Lu, X.Y., M.S. Chen, J.L. Wu, P.C. Chang and M.H. Chen, 2017. A novel ensemble decision tree based on under-sampling and clonal selection for web spam detection. Patt. Anal. Applic., 2017: 1-14. DOI: 10.1007/s10044-017-0602-2

Lu, Y., Y.M. Cheung and Y.Y. Tang, 2016. GOBoost: G-mean optimized boosting framework for class imbalance learning. Proceedings of the 12th World Congress on Intelligent Control and Automation, Jun. 12-15, IEEE Xplore Press, Guilin, China, pp: 3149-3154. DOI: 10.1109/WCICA.2016.7578792

Ma, Y., B.H. Su, S. Zhu, W. Weng and L. Huang *et al*., 2015. Asymmetric classifier based on kernel PLS for imbalanced data. Proceedings of the 10th International Conference on Computer Science and Education, Jul. 22-24, IEEE Xplore Press, Cambridge, UK, pp: 482-485. DOI: 10.1109/ICCSE.2015.7250294

Ma, Y., G. Luo, J. Li and A. Chen, 2011. Combating class imbalance problem in semi-supervised defect detection. Proceedings of the International Conference on Computational Problem-Solving, Oct. 21-23, IEEE Xplore Press, Chengdu, China, pp: 619-622. DOI: 10.1109/ICCPS.2011.6092260

Machado, E.L. and M. Ladeira, 2011. Dealing with rare cases and avoiding overfitting: Combining cluster based oversampling and SMOTE. Department of Computer Science, Brazil.

Mahmoudi, S., P. Moradi, F. Akhlaghian and R. Moradi, 2014. Diversity and separable metrics in over-sampling technique for imbalanced data classification. Proceedings of the 4th International eConference on Computer and Knowledge Engineering, Oct. 29-30, IEEE Xplore Press, Mashhad, Iran, pp: 152-158. DOI: 10.1109/ICCKE.2014.6993409

Majumder, A., L. Behera and V.K. Subramanian, 2016. GMR based pain intensity recognition using imbalanced data handling techniques. Proceedings of the International Conference on Signal and Information Processing, Oct. 6-8, IEEE Xplore Press, Vishnupuri, India, pp: 1-5. DOI: 10.1109/ICONSIP.2016.7857447

Makrehchi, M. and M.S. Kamel, 2007. A text classification framework with a local feature ranking for learning social networks. Proceedings of the 7th IEEE International Conference on Data Mining, Oct. 28-31, IEEE Xplore Press, Omaha, NE, USA, pp: 589-594. DOI: 10.1109/ICDM.2007.26

Margineantu, D.D., 2002. Class probability estimation and cost-sensitive classification decisions. Proceedings of the 13th European Conference on Machine Learning, Aug. 19-23, Springer, Helsinki, Finland, pp: 270-281. DOI: 10.1007/3-540-36755-1_23

Martin-Diaz, I., D. Morinigo-Sotelo, O. Duque-Perez and R.J. RomeroTroncoso, 2017. Early fault detection in induction motors using AdaBoost with imbalanced small data and optimized sampling. IEEE Trans. Industry Applic., 53: 3066-3075. DOI: 10.1109/TIA.2016.2618756

Mathew, J., M. Luo, C.K. Pang and H.L. Chan, 2015. Kernel-based SMOTE for SVM classification of imbalanced datasets. Proceedings of the 41st Annual Conference of the Industrial Electronics Society, Nov. 9-12, IEEE Xplore Press, Yokohama, Japan, pp: 1127-1132. DOI: 10.1109/IECON.2015.7392251

Matsuda, K. and K. Murase, 2016. Single-layered complex-valued neural network with SMOTE for imbalanced data classification. Proceedings of the Joint 8th International Conference on Soft Computing and Intelligent Systems (SCIS) and 17th International Symposium on Advanced Intelligent Systems, Aug. 25-28, IEEE Xplore Press, Sapporo, Japan, pp: 349-354. DOI: 10.1109/SCIS-ISIS.2016.0079

Melillo, P., N. De Luca, M. Bracale and L. Pecchia, 2013. Classification tree for risk assessment in patients suffering from congestive heart failure via long-term heart rate variability. IEEE J. Biomed. Health Inform., 17: 727-733. DOI: 10.1109/JBHI.2013.2244902

Mercado-Dıaz, L.R., J. Navarro-Garcıa and J.A. Jaramillo-Garzon, 2015. A comparison of class-balance strategies for SVM in the problem of protein function prediction. Proceedings of the 20th Symposium on Signal Processing, Images and Computer Vision, Sept. 2-4, IEEE Xplore Press, Bogota, Colombia, pp: 1-5. DOI: 10.1109/STSIVA.2015.7330418

Moreo, A., A. Esuli and F. Sebastiani, 2016. Distributional random oversampling for imbalanced text classification. Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, (DIR' 16), ACM, pp: 805-808.

Mountassir, A., H. Benbrahim and I. Berrada, 2012. An empirical study to address the problem of unbalanced data sets in sentiment classification. Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, Oct. 14-17, IEEE Xplore Press, Seoul, South Korea, pp: 3298-3303. DOI: 10.1109/ICSMC.2012.6378300

Mustafa, G., Z. Niu, A. Yousif and J. Tarus, 2015. Distribution based ensemble for class imbalance learning. Proceedings of the 5th International Conference on Innovative Computing Technology, May 20-22, IEEE Xplore Press, Pontevedra, Spain, pp: 5-10. DOI: 10.1109/INTECH.2015.7173365

Mustafa, N. and J.P. Li, 2017. Medical data classification scheme based on Hybridized SMOTE Technique (HST) and Rough Set Technique (RST). Proceedings of the IEEE 2nd International Conference on Cloud Computing and Big Data Analysis, Apr. 28-30, IEEE Xplore Press, Chengdu, China, pp: 49-55. DOI: 10.1109/ICCCBDA.2017.7951883

Naganjaneyulu, S. and M.R. Kuppa, 2013. A novel framework for class imbalance learning using intelligent under-sampling. Progress Artificial Intell., 2: 73-84. DOI: 10.1007/s13748-012-0038-2

Napieral, K. and J. Stefanowski, 2015. Addressing imbalanced data with argument based rule learning. Expert Syst. Applic., 42: 9468-9481. DOI: 10.1016/j.eswa.2015.07.076

Napierala, K. and J. Stefanowski, 2012. BRACID: A comprehensive approach to learning rules from imbalanced data. J. Intell. Inform. Syst., 39: 335-373. DOI: 10.1007/s10844-011-0193-0

Nikpour, B., M. Shabani and H. Nezamabadi-Pour, 2017. Proposing new method to improve gravitational fixed nearest neighbor algorithm for imbalanced data classification. Proceedings of the 2nd Conference on Swarm Intelligence and Evolutionary Computation, Mar. 7-9, IEEE Xplore Press, Kerman, Iran, pp: 6-11. DOI: 10.1109/CSIEC.2017.7940167

Oktavino, H.F. and N.U. Maulidevi, 2014. Information extractor for small medium enterprise aggregator. Proceedings of the International Conference on Data and Software Engineering, Nov. 26-27, IEEE Xplore Press, Bandung, Indonesia, pp: 1-5. DOI: 10.1109/ICODSE.2014.7062659

Oliveira, D.V., T.N. Porpino, G.D. Cavalcanti and T.I. Ren, 2015. A bootstrap-based iterative selection for ensemble generation. Proceedings of the International Joint Conference on Neural Networks, Jul. 12-17, IEEE Xplore Press, Killarney, Ireland, pp: 1-7. DOI: 10.1109/IJCNN.2015.7280695

Padmaja, T.M., N. Dhulipalla, R.S. Bapi and P.R. Krishna, 2007. Unbalanced data classification using extreme outlier elimination and sampling techniques for fraud detection. Proceedings of the International Conference on Advanced Computing and Communications, Dec. 18-21, IEEE Xplore Press, Guwahati, Assam, India, pp: 511-516. DOI: 10.1109/ADCOM.2007.74

Padmaja, T.M., P.R. Krishna and R.S. Bapi, 2008. Majority Filter-based Minority Prediction (MFMP): An approach for unbalanced datasets. Proceedings of the IEEE Region 10 Conference TENCON, Nov. 19-21, IEEE Xplore Press, Hyderabad, India, pp: 1-6. DOI: 10.1109/TENCON.2008.4766705

Pal, B. and M.K. Paul, 2017. A Gaussian mixture based boosted classification scheme for imbalanced and oversampled data. Proceedings of the International Conference on Electrical, Computer and Communication Engineering, Feb. 16-18, IEEE Xplore Press, Cox's Bazar, Bangladesh, pp: 401-405. DOI: 10.1109/ECACE.2017.7912938

Park, C., D. Kim, J. Oh and H. Yu, 2015. Predicting user purchase in e-commerce by comprehensive feature engineering and decision boundary focused undersampling. Proceedings of the International ACM Recommender Systems Challenge, (RSC' 15), ACM, p: 1-8.

Park, Y. and J. Ghosh, 2014. Ensembles of *alpha*-trees for imbalanced classification problems. IEEE Trans. Knowl. Data Eng., 26: 131-143. DOI: 10.1109/TKDE.2012.255

Patil, S.S. and S.P. Sonavane, 2017. Enriched over_sampling techniques for improving classification of imbalanced big data. Proceedings of the IEEE 3rd International Conference on Big Data Computing Service and Applications, Apr. 6-9, IEEE Xplore Press, San Francisco, CA, USA, pp: 1-10. DOI: 10.1109/BigDataService.2017.19

Pelayo, L. and S. Dick, 2007. Applying novel resampling strategies to software defect prediction. Proceedings of the Annual Meeting of the North American Fuzzy Information Processing Society, Jun. 24-27, IEEE Xplore Press, San Diego, CA, USA, pp: 69-72. DOI: 10.1109/NAFIPS.2007.383813

Peng, L., B. Yang, Y. Chen and X. Zhou, 2016a. An under-sampling imbalanced learning of data gravitation based classification. Proceedings of the 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery, Aug. 13-15, IEEE Xplore Press, Changsha, China, pp: 419-425. DOI: 10.1109/FSKD.2016.7603210

Peng, L., H. Zhang, B. Yang, Y. Chen and X. Zhou, 2016b. SMOTE-DGC: An imbalanced learning approach of data gravitation based classification. Proceedings of the International Conference on Intelligent Computing, (CIC' 16), Springer, pp: 133-144. DOI: 10.1007/978-3-319-42294-7_11

Peng, Y. and J. Yao, 2010. Adaouboost: Adaptive over-sampling and under-sampling to boost the concept learning in large scale imbalanced data sets. Proceedings of the International conference on Multimedia Information Retrieval, (MIR' 10), ACM, pp: 111-118. DOI: 10.1145/1743384.1743408

Pengfei, J., Z. Chunkai and H. Zhenyu, 2014. A new sampling approach for classification of imbalanced data sets with high density. Proceedings of the International Conference on Big Data and Smart Computing, Jan. 15-17, IEEE Xplore Press, Bangkok, Thailand, pp: 217-222. DOI: 10.1109/BIGCOMP.2014.6741439

Perez, J.M., J. Muguerza, O. Arbelaitz, I. Gurrutxaga and J.I. Martın, 2004. Behavior of consolidated trees when using resampling techniques. Proceedings of the 4th International Workshop on Pattern Recognition and Information Systems, (RIS' 04), pp: 139-148.

Phua, C., D. Alahakoon and V. Lee, 2004. Minority report in fraud detection: Classification of skewed data. SIGKDD Explor. Newsl., 6: 50-59. DOI: 10.1145/1007730.1007738

Prachuabsupakij, W. and N. Soonthornphisaj, 2014. Hybrid sampling for multiclass imbalanced problem: Case study of students' performance prediction. Proceedings of the International Conference on Advanced Computer Science and Information Systems, Oct. 18-19, IEEE Xplore Press, Jakarta, Indonesia, pp: 321-326. DOI: 10.1109/ICACSIS.2014.7065824

Prachuabsupakij, W. and P. Doungpaisan, 2016. Matching preprocessing methods for improving the prediction of student's graduation. Proceedings of the 2nd IEEE International Conference on Computer and Communications, Oct. 14-17, IEEE Xplore Press, Chengdu, China, pp: 33-37. DOI: 10.1109/CompComm.2016.7924659

Prati, R.C., G. Batista and M.C. Monard, 2004. Learning with class skews and small disjuncts. Proceedings of the 17th Brazilian Symposium on Artificial Intelligence, (SAI' 04), Springer, Berlin, pp: 296-306. DOI: 10.1007/978-3-540-28645-5_30

Purnami, S.W. and R.K. Trapsilasiwi, 2017. SMOTE-least square support vector machine for classification of multiclass imbalanced data. Proceedings of the 9th International Conference on Machine Learning and Computing, ACM, pp: 107-111.

Qazi, N. and K. Raza, 2012. Effect of feature selection, SMOTE and under sampling on class imbalance classification. Proceedings of the UKSim 14th International Conference on Computer Modelling and Simulation, Mar. 28-30, IEEE Xplore Press, Cambridge, UK, pp: 145-150. DOI: 10.1109/UKSim.2012.116

Qian, Y., Y. Liang, M. Li, G. Feng and X. Shi, 2014. A resampling ensemble algorithm for classification of imbalance problems. Neurocomputing, 143: 57-67. DOI: 10.1016/j.neucom.2014.06.021

Qing, Y., W. Bin, Z. Peilan, C. Xiang and Z. Meng *et al*., 2015. The prediction method of material consumption for electric power production based on PCBOOST and SVM. Proceedings of the 8th International Congress on Image and Signal Processing, Oct. 14-16, IEEE Xplore Press, Shenyang, China, pp: 1256-1260. DOI: 10.1109/CISP.2015.7408074

Qiu, C., L. Jiang and G. Kong, 2015. A differential evolution-based method for class-imbalanced cost-sensitive learning. Proceedings of the International Joint Conference on Neural Networks, Jul. 12-17, IEEE Xplore Press, Killarney, Ireland, pp: 1-8. DOI: 10.1109/IJCNN.2015.7280419

Radivojac, P., N.V. Chawla, A.K. Dunker and Z. Obradovic, 2004. Classification and knowledge discovery in protein databases. J. Biomed. Inform., 37: 224-239. DOI: 10.1016/j.jbi.2004.07.008

Rahman, H.A.A., Y.B. Wah, H. He and A. Bulgiba, 2015. Comparisons of ADABOOST, KNN, SVM and logistic regression in classification of imbalanced dataset. Proceedings of the International Conference on Soft Computing in Data Science, Sept. 2-3, Springer, Putrajaya, Malaysia, pp: 54-64. DOI: 10.1007/978-981-287-936-3_6

Ramentol, E., S. Vluymans, N. Verbiest, Y. Caballero and R. Bello *et al*., 2015. IFROWANN: Imbalanced fuzzy-rough ordered weighted average nearest neighbor classification. IEEE Trans. Fuzzy Syst., 23: 1622-1637. DOI: 10.1109/TFUZZ.2014.2371472

Ramentol, E., Y. Caballero, R. Bello and F. Herrera, 2012. SMOTE-RSB*: A hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory. Knowl. Inform. Syst., 33: 245-265. DOI: 10.1007/s10115-011-0465-6

Rani, T.S. and P. Soujanya, 2013. An ensemble method using small training sets for imbalanced data sets: Application to drugs used for kinases. Proceedings of the 6th International Conference on Contemporary Computing, Aug. 8-10, IEEE Xplore Press, Noida, India, pp: 516-521. DOI: 10.1109/IC3.2013.6612250

Rashu, R.I., N. Haq and R.M. Rahman, 2014. Data mining approaches to predict final grade by overcoming class imbalance problem. Proceedings of the 17th International Conference on Computer and Information Technology, Dec. 22-23, IEEE Xplore Press, Dhaka, Bangladesh, pp: 14-19. DOI: 10.1109/ICCITechn.2014.7073095

Raskutti, B. and A. Kowalczyk, 2004. Extreme re-balancing for SVMS: A case study. ACM Sigkdd Explorat. Newsl., 6: 60-69. DOI: 10.1145/1007730.1007739

Ren, J., D. Wang and J. Jiang, 2011. Effective recognition of MCCS in mammograms using an improved neural classifier. Eng. Applic. Artificial Intell., 24: 638-645. DOI: 10.1016/j.engappai.2011.02.011

Rokach, L., 2010. Ensemble-based classifiers. Artificial Intell. Rev., 33: 1-39. DOI: 10.1007/s10462-009-9124-7

Rosa, R.S., R.H. Santos, A.Y. Brito and K.S. Guimaraes, 2014. Insights on prediction of patients' response to anti-HIV therapies through machine learning. Proceedings of the International Joint Conference on Neural Networks, Jul. 6-11, IEEE Xplore Press, Beijing, China, pp: 3697-3704. DOI: 10.1109/IJCNN.2014.6889659

Ruangthong, P. and S. Jaiyen, 2015. Bank direct marketing analysis of asymmetric information based on machine learning. Proceedings of the 12th International Joint Conference on Computer Science and Software Engineering, Jul. 22-24, IEEE Xplore Press, Songkhla, Thailand, pp: 93-96. DOI: 10.1109/JCSSE.2015.7219777

Ruangthong, P. and S. Jaiyen, 2016. Hybrid ensembles of decision trees and Bayesian network for class imbalance problem. Proceedings of the 8th International Conference on Knowledge and Smart Technology, Feb. 3-6, IEEE Xplore Press, Chiangmai, Thailand, pp: 39-42. DOI: 10.1109/KST.2016.7440523

Ryu, D., O. Choi and J. Baik, 2016. Value-cognitive boosting with a support vector machine for cross-project defect prediction. Empirical Software Eng., 21: 43-71. DOI: 10.1007/s10664-014-9346-4

Saez, J.A., J. Luengo, J. Stefanowski and F. Herrera, 2015. SMOTE–IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering. Inform. Sci., 291: 184-203. DOI: 10.1016/j.ins.2014.08.051

Sain, H. and S.W. Purnami, 2015. Combine sampling support vector machine for imbalanced data classification. Proc. Comput. Sci., 72: 59-66. DOI: 10.1016/j.procs.2015.12.105

Sandhan, T. and J.Y. Choi, 2014. Handling imbalanced datasets by partially guided hybrid sampling for pattern recognition. Proceedings of the 22nd International Conference on Pattern Recognition, Aug. 24-28, IEEE Xplore Press, Stockholm, Sweden, pp: 1449-1453. DOI: 10.1109/ICPR.2014.258

Sanguanmak, Y. and A. Hanskunatai, 2016a. Auto-tuning of parameters in hybrid sampling method for class imbalance problem. Proceedings of the International Conference on Computer Science and Engineering Conference, Dec. 14-17, IEEE Xplore Press, Chiang Mai, Thailand, pp: 1-5. DOI: 10.1109/ICSEC.2016.7859941

Sanguanmak, Y. and A. Hanskunatai, 2016b. DBSM: The combination of DBSCAN and SMOTE for imbalanced data classification. Proceedings of the 13th International Joint Conference on Computer Science and Software Engineering, Jul. 13-15, IEEE Xplore Press, Khon Kaen, Thailand, pp: 1-5. DOI: 10.1109/JCSSE.2016.7748928

Sarakit, P., T. Theeramunkong and C. Haruechaiyasak, 2010. Improving emotion classification in imbalanced YouTube dataset using SMOTE algorithm. Proceedings of the 2nd International Conference on Advanced Informatics: Concepts, Theory and Applications, Aug. 19-22, IEEE Xplore Press, Chonburi, Thailand, pp: 1-5. DOI: 10.1109/ICAICTA.2015.7335373

Seiffert, C., T.M. Khoshgoftaar, J. Van Hulse and A. Napolitano, 2008. Resampling or reweighting: A comparison of boosting implementations. Proceedings of the 20th IEEE International Conference on Tools with Artificial Intelligence, Nov. 3-5, IEEE Xplore Press, Dayton, OH, USA, pp: 445-451. DOI: 10.1109/ICTAI.2008.59

Shi, T., X. Wei and X. Shao, 2016. Mass incidents prediction based on ID3-SMOTE algorithm. Proceedings of the International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, Oct. 13-15, IEEE Xplore Press, Chengdu, China, pp: 115-118. DOI: 10.1109/CyberC.2016.31

Shi, X., G. Xu, F. Shen and J. Zhao, 2015. Solving the data imbalance problem of p300 detection via random under-sampling bagging SVMS. Proceedings of the International Joint Conference on Neural Networks, Jul. 12-17, IEEE Xplore Press, Killarney, Ireland, pp: 1-5. DOI: 10.1109/IJCNN.2015.7280834

Shilaskar, S., A. Ghatol and P. Chatur, 2017. Medical decision support system for extremely imbalanced datasets. Inform. Sci., 384: 205-219. DOI: 10.1016/j.ins.2016.08.077

Shin, H. and S. Cho, 2003. How to deal with large dataset, class imbalance and binary output in SVM based response model. Proceedings of the Korean Data Mining Conference, (DMC' 03), pp: 93-107.

Soltani, S., J. Sadri and H.A. Torshizi, 2011. Feature selection and ensemble hierarchical cluster-based under-sampling approach for extremely imbalanced datasets: Application to gene classification. Proceedings of the 1st International eConference on Computer and Knowledge Engineering, Oct. 13-14, IEEE Xplore Press, Mashhad, Iran, pp: 166-171. DOI: 10.1109/ICCKE.2011.6413345

Sonak, A., R. Patankar and N. Pise, 2016. A new approach for handling imbalanced dataset using ANN and genetic algorithm. Proceedings of the International Conference on Communication and Signal Processing, Apr. 6-8, IEEE Xplore Press, Melmaruvathur, India, pp: 1987-1990. DOI: 10.1109/ICCSP.2016.7754521

Stefanowski, J. and S. Wilk, 2008. Selective pre-processing of imbalanced data for improving classification performance. Proceedings of the 10th International Conference on Data Warehousing and Knowledge Discovery, (WKD' 08), Springer, Berlin Heidelberg, pp: 283-292. DOI: 10.1007/978-3-540-85836-2_27

Sun, Y. and F. Liu, 2016. SMOTE-NCL: A re-sampling method with filter for network intrusion detection. Proceedings of the 2nd IEEE International Conference on Computer and Communications, Oct. 14-17, IEEE Xplore Press, Chengdu, China, pp: 1157-1161. DOI: 10.1109/CompComm.2016.7924886

Sun, Z., Q. Song and X. Zhu, 2012. Using coding-based ensemble learning to improve software defect prediction. IEEE Trans. Syst. Man Cybernet., 42: 1806-1817. DOI: 10.1109/TSMCC.2012.2226152

Sundarkumar, G.G., V. Ravi and V. Siddeshwar, 2015. One-class support vector machine based undersampling: Application to churn prediction and insurance fraud detection. Proceedings of the IEEE International Conference on Computational Intelligence and Computing Research, Dec. 10-12, IEEE Xplore Press, Madurai, India, pp: 1-7. DOI: 10.1109/ICCIC.2015.7435726

Tan, K. and X. Cai, 2010. Prediction of earthquake in Yunnan region based on the AHC over sampling. Proceedings of the Chinese Control and Decision Conference, May 26-28, IEEE Xplore Press, Xuzhou, China, pp: 2449-2452. DOI: 10.1109/CCDC.2010.5498782

Tan, S.C., J. Watada, Z. Ibrahim and M. Khalid, 2015. Evolutionary fuzzy ARTMAP neural networks for classification of semiconductor defects. IEEE Trans. Neural Netw. Learn. Syst., 26: 933-950. DOI: 10.1109/TNNLS.2014.2329097

Tan, S.C., Z. Ibrahim, L.W. Jau, J. Watada and M. Khalid *et al*., 2011. Learning with imbalanced datasets using fuzzy ARTMAP-based neural network models. Proceedings of the IEEE International Conference on Fuzzy Systems, Jun. 27-30, IEEE Xplore Press, Taipei, Taiwan, pp: 1084-1089. DOI: 10.1109/FUZZY.2011.6007330

Tao, D., X. Tang, X. Li and X. Wu, 2006. Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval. IEEE Trans. Patt. Anal. Mach. Intell., 28: 1088-1099. DOI: 10.1109/TPAMI.2006.134

Tayal, A., T.F. Coleman and Y. Li, 2015. Rankrc: Large-scale nonlinear rare class ranking. IEEE Trans. Knowl. Data Eng., 27: 3347-3359. DOI: 10.1109/TKDE.2015.2453171

Thach, N.H., P. Rojanavasu and O. Pinngern, 2008. Cost-xensitive XCS classifier system addressing imbalance problems. Proceedings of the 5th International Conference on Fuzzy Systems and Knowledge Discovery, Oct. 18-20, IEEE Xplore Press, Shandong, China, pp: 132-136. DOI: 10.1109/FSKD.2008.391

Thai-Nghe, N., A. Busche and L. Schmidt-Thieme, 2009. Improving academic performance prediction by dealing with class imbalance. Proceedings of the 9th International Conference on Intelligent Systems Design and Applications, Nov. 30-Dec. 2, IEEE Xplore Press, Pisa, Italy, pp: 878-883. DOI: 10.1109/ISDA.2009.15

Thai-Nghe, N., Z. Gantner and L. Schmidt-Thieme, 2010. Cost-sensitive learning methods for imbalanced data. Proceedings of the International Joint Conference on Neural Networks, Jul. 18-23, IEEE Xplore Press, Barcelona, Spain, pp: 1-8. DOI: 10.1109/IJCNN.2010.5596486

Tomek, I., 1976. Two modifications of CNN. IEEE Trans. Syst. Man Cybernet., 6: 769-772.

Tran, Q.D. and P. Liatsis, 2016. Raboc: An approach to handle class imbalance in multimodal biometric authentication. Neurocomputing, 188: 167-177. DOI: 10.1016/j.neucom.2014.12.126

Triguero, I., M. Galar, H. Bustince and F. Herrera, 2017. A first attempt on global evolutionary undersampling for imbalanced big data. Proceedings of the IEEE Congress on Evolutionary Computation, Jun. 5-8, IEEE Xplore Press, San Sebastian, Spain, pp: 2054-2061. DOI: 10.1109/CEC.2017.7969553

Van Hulse, J., T.M. Khoshgoftaar and A. Napolitano, 2007. Experimental perspectives on learning from imbalanced data. Proceedings of the 24th International Conference on Machine Learning, (CML' 07) ACM, pp: 935-942.

Van Vlasselaer, V., J. Meskens, D. Van Dromme and B. Baesens, 2013. Using social network knowledge for detecting spider constructions in social security fraud. Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Aug. 25-28, ACM, Niagara, Ontario, Canada, pp: 813-820. DOI: 10.1145/2492517.2500292

Varassin, C.G., A. Plastino, H.C. da Gama Leitao and B. Zadrozny, 2013. Under-sampling strategy based on clustering to improve the performance of splice site classification in human genes. Proceedings of the 24th International Workshop on Database and Expert Systems Applications, Aug. 26-30, IEEE Xplore Press, Los Alamitos, CA, USA, pp: 85-89. DOI: 10.1109/DEXA.2013.40

Vinodhini, G. and R. Chandrasekaran, 2017. A sampling based sentiment mining approach for e-commerce applications. Inform. Process. Manage., 53: 223-236. DOI: 10.1016/j.ipm.2016.08.003

Vorraboot, P., C. Lursinsap, S. Rasmequan and K. Chinnasarn, 2012. A modified error function for imbalanced dataset classification problem. Proceedings of the 7th International Conference on Computing and Convergence Technology, Dec. 3-5, IEEE Xplore Press, Seoul, South Korea, pp: 854-859.

Vorraboot, P., S. Rasmequan, K. Chinnasarn and C. Lursinsap, 2015. Improving classification rate constrained to imbalanced data between overlapped and non-overlapped regions by hybrid algorithms. Neurocomputing, 152: 429-443. DOI: 10.1016/j.neucom.2014.10.007

Wang, J., P.L. Huang, K.W. Sun, B.L. Cao and R. Zhao, 2013b. Ensemble of cost-sensitive hypernetworks for class-imbalance learning. Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, Oct. 13-16, IEEE Xplore Press, Manchester, UK, pp: 1883-1888. DOI: 10.1109/SMC.2013.324

Wang, J., Y. Yao, H. Zhou, M. Leng and X. Chen, 2013a. A new over-sampling technique based on SVM for imbalanced diseases data. Proceedings of the International Conference on Mechatronic Sciences, Electric Engineering and Computer, Dec. 20-22, IEEE Xplore Press, Shengyang, China, pp: 1224-1228. DOI: 10.1109/MEC.2013.6885254

Wang, K.J., B. Makond, K.H. Chen and K.M. Wang, 2014. A hybrid classifier combining SMOTE with PSO to estimate 5-year survivability of breast cancer patients. Applied Soft Comput., 20: 15-24. DOI: 10.1016/j.asoc.2013.09.014

Wang, L., J. Jin, R. Huang, X. Wei and J. Chen, 2016. Unbiased decision tree model for user's QOE in imbalanced dataset. Proceedings of the International Conference on Cloud Computing Research and Innovations, May 4-5, IEEE Xplore Press, Singapore, pp: 114-119. DOI: 10.1109/ICCCRI.2016.25

Wang, Q., Y. Zheng, W. Jin and X. Chen, 2018. Multiscale rotation-invariant convolutional neural networks for lung texture classification. IEEE J. Biomed. Health Inform., 22: 184-195. DOI: 10.1109/JBHI.2017.2685586

Wang, S. and X. Yao, 2009. Diversity analysis on imbalanced data sets by using ensemble models. Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining, Mar. 30-Apr. 2, IEEE Xplore Press, Nashville, TN, USA, pp: 324-331. DOI: 10.1109/CIDM.2009.4938667

Wang, S. and X. Yao, 2012. Multiclass imbalance problems: Analysis and potential solutions. IEEE Trans. Syst. Man Cybernet., 42: 1119-1130. DOI: 10.1109/TSMCB.2012.2187280

Wang, S., H. Chen and X. Yao, 2010. Negative correlation learning for classification ensembles. Proceedings of the International Joint Conference on Neural Networks, Jul. 18-23, IEEE Xplore Press, Barcelona, Spain, pp: 1-8. DOI: 10.1109/IJCNN.2010.5596702

Wang, S., L.L. Minku and X. Yao, 2015. Resampling-based ensemble methods for online class imbalance learning. IEEE Trans. Knowl. Data Eng., 27: 1356-1368. DOI: 10.1109/TKDE.2014.2345380

Wang, S., Z. Li, W. Chao and Q. Cao, 2012. Applying adaptive over-sampling technique based on data density and cost-sensitive SVM to imbalanced learning. Proceedings of the International Joint Conference on Neural Networks, Jun. 10-15, IEEE Xplore Press, Brisbane, QLD, Australia, pp: 1-8. DOI: 10.1109/IJCNN.2012.6252696

Wei, H., B. Sun and M. Jing, 2014. BalancedBoost: A hybrid approach for real-time network traffic classification. Proceedings of the 23rd International Conference on Computer Communication and Networks, Aug. 4-7, IEEE Xplore Press, Shanghai, China, pp: 1-6. DOI: 10.1109/ICCCN.2014.6911833

Winata, G.I. and M.L. Khodra, 2015. Handling imbalanced dataset in multi-label text categorization using bagging and adaptive boosting. Proceedings of the International Conference on Electrical Engineering and Informatics, Aug. 10-11, IEEE Xplore Press, Denpasar, Indonesia, pp: 500-505. DOI: 10.1109/ICEEI.2015.7352552

Wong, G.Y., F.H. Leung and S.H. Ling, 2013. A novel evolutionary preprocessing method based on over-sampling and under-sampling for imbalanced datasets. Proceedings of the 39th Annual Conference of the IEEE Industrial Electronics Society, Nov. 10-13, IEEE Xplore Press, Vienna, Austria, pp: 2354-2359. DOI: 10.1109/IECON.2013.6699499

Wong, G.Y., F.H. Leung and S.H. Ling, 2014. An under-sampling method based on fuzzy logic for large imbalanced dataset. Proceedings of the IEEE International Conference on Fuzzy Systems, Jul. 6-11, IEEE Xplore Press, Beijing, China, pp: 1248-1252. DOI: 10.1109/FUZZ-IEEE.2014.6891771

Wu, G. and E.Y. Chang, 2004. Aligning boundary in kernel space for learning imbalanced dataset. Proceedings of the 4th IEEE International Conference on Data Mining, Nov. 1-4, IEEE Xplore Press, Brighton, UK, pp: 265-272. DOI: 10.1109/ICDM.2004.10106

Wu, Q. and M. Wang, 2009. Abnormal BGP routing dynamics detection by sampling approach in decision tree. Proceedings of the 1st International Workshop on Database Technology and Applications, Apr. 25-26, IEEE Xplore Press, Wuhan, Hubei, China, pp: 170-173. DOI: 10.1109/DBTA.2009.175

Wu, X. and S. Meng, 2016. E-commerce customer churn prediction based on improved SMOTE and AdaBoost. Proceedings of the 13th International Conference on Service Systems and Service Management, Jun. 24-26, IEEE Xplore Press, Kunming, China, pp: 1-5. DOI: 10.1109/ICSSSM.2016.7538581

Xia, P. and L. Zhang, 2014. Solving unbalanced problems in similarity learning using SVM ensemble. Proceedings of the International Joint Conference on Neural Networks, Jul. 6-11, IEEE Xplore Press, Beijing, China, pp: 1762-1768. DOI: 10.1109/IJCNN.2014.6889614

Xiang, J.P. and Y. Bai, 2013. Active learning with re-sampling for support vector machine in person re-identification. Proceedings of the International Conference on Machine Learning and Cybernetics, Jul. 14-17, IEEE Xplore Press, Tianjin, China, pp: 597-602. DOI: 10.1109/ICMLC.2013.6890362

Xiaoying, X. and F. Sheng, 2012. A synthesized sampling approach for improving the prediction of imbalanced classification. Proceedings of the IEEE 3rd International Conference on Software Engineering and Service Science, Jun. 22-24, IEEE Xplore Press, Beijing, China, pp: 615-619. DOI: 10.1109/ICSESS.2012.6269542

Xu, G., B.G. Hu and J.C. Principe, 2016a. Robust bounded logistic regression in the class imbalance problem. Proceedings of the International Joint Conference on Neural Networks, Jul. 24-29, IEEE Xplore Press, Vancouver, BC, Canada, pp: 1434-1441. DOI: 10.1109/IJCNN.2016.7727367

Xu, Y., Z. Yang, Y. Zhang, X. Pan and L. Wang, 2016b. A maximum margin and minimum volume hyper-spheres machine with pinball loss for imbalanced data classification. Knowledge-Based Syst., 95: 75-85. DOI: 10.1016/j.knosys.2015.12.005

Xu, G., F. Shen and J. Zhao, 2013. The effect of methods addressing the class imbalance problem on p300 detection. Proceedings of the International Joint Conference on Neural Networks, Aug. 4-9, IEEE Xplore Press, Dallas, TX, USA, pp: 1-5. DOI: 10.1109/IJCNN.2013.6706890

Yala, N., B. Fergani and L. Clavier, 2014. Soft margin SVM modeling for handling imbalanced human activity datasets in multiple homes. Proceedings of the International Conference on Multimedia Computing and Systems, Apr. 14-16, IEEE Xplore Press, Marrakech, Morocco, pp: 421-426. DOI: 10.1109/ICMCS.2014.6911407

Yang, Z. and D. Gao, 2012. An active under-sampling approach for imbalanced data classification. Proceedings of the 5th International Symposium on Computational Intelligence and Design, Oct. 28-29, IEEE Xplore Press, Hangzhou, China, pp: 270-273. DOI: 10.1109/ISCID.2012.219

Yap, B.W., K.A. Rani, H.A.A. Rahman, S. Fong and Z. Khairudin *et al*., 2013. An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets. Proceedings of the 1st International Conference on Advanced Data and Information Engineering, (DIE' 13), Springer, pp: 13-22.

Yeh, C.W., D.C. Li, L.S. Lin and T.I. Tsai, 2016. A learning approach with under-and over-sampling for imbalanced data sets. Proceedings of the 5th IIAI International Congress on Advanced Applied Informatics, Jul. 10-14, IEEE Xplore Press, Kumamoto, Japan, pp: 725-729. DOI: 10.1109/IIAI-AAI.2016.20

Yen, S.J. and Y.S. Lee, 2009. Cluster-based under-sampling approaches for imbalanced data distributions. Expert Syst. Applic., 36: 5718-5727. DOI: 10.1016/j.eswa.2008.06.108

Yen, S.J., Y.S. Lee, C.H. Lin and J.C. Ying, 2006. Investigating the effect of sampling methods for imbalanced data distributions. Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, Oct. 8-11, IEEE Xplore Press, Taipei, Taiwan, pp: 4163-4168. DOI: 10.1109/ICSMC.2006.384787

Yongqing, Z., Z. Min, Z. Danling, M. Gang and M. Daichuan, 2013. Improved SMOTE-bagging and its application in imbalanced data classification. Proceedings of the 13th IEEE Joint International Computer Science and Information Technology Conference, Jan. 1-8, IEEE Xplore Press, China, pp: 1-5. DOI: 10.1109/ANTHOLOGY.2013.6784957

Yongxiong, W. and K. Li, 2014. Feature and weight selection using TABU search for improving the recognition rate of duct anomaly. Proceedings of the IEEE International Conference on Robotics and Biomimetics, Dec. 5-10, IEEE Xplore Press, Bali, Indonesia, pp: 2163-2168. DOI: 10.1109/ROBIO.2014.7090657

Yoon, K. and S. Kwek, 2007. A data reduction approach for resolving the imbalanced data issue in functional genomics. Neural Comput. Applic., 16: 295-306. DOI: 10.1007/s00521-007-0089-7

Yu, T., S. Simoff and T. Jan, 2010. VQSVM: A case study for incorporating prior domain knowledge into inductive machine learning. Neurocomputing, 73: 2614-2623. DOI: 10.1016/j.neucom.2010.05.007

Yu, T., T. Jan, S. Simoff and J. Debenham, 2007. A hierarchical VQSVM for imbalanced data sets. Proceedings of the International Joint Conference on Neural Networks, Aug. 12-17, IEEE Xplore Press, Orlando, FL, USA, pp: 518-523. DOI: 10.1109/IJCNN.2007.4371010

Yuan, B. and X. Ma, 2012. Sampling + reweighting: Boosting the performance of AdaBoost on imbalanced datasets. Proceedings of the International Joint Conference on Neural Networks, Jun. 10-15, IEEE Xplore Press, Brisbane, QLD, Australia, pp: 1-6. DOI: 10.1109/IJCNN.2012.6252738

Zang, B., R. Huang, L. Wang, J. Chen and F. Tian *et al*., 2016. An improved KNN algorithm based on minority class distribution for imbalanced dataset. Proceedings of the International Computer Symposium, Dec. 15-17, IEEE Xplore Press, Chiayi, Taiwan, pp: 696-700. DOI: 10.1109/ICS.2016.0143

Zhang, D., J. Ma, J. Yi, X. Niu and X. Xu, 2015. An ensemble method for unbalanced sentiment classification. Proceedings of the 11th International Conference on Natural Computation, Aug. 15-17, IEEE Xplore Press, Zhangjiajie, China, pp: 440-445. DOI: 10.1109/ICNC.2015.7378029

Zhang, H. and M. Li, 2014. RWO-sampling: A random walk over-sampling approach to imbalanced data classification. Inform. Fus., 20: 99-116. DOI: 10.1016/j.inffus.2013.12.003

Zhang, L. and W. Wang, 2011. A re-sampling method for class imbalance learning with credit data. Proceedings of the International Conference on Information Technology, Computer Engineering and Management Sciences, Sept. 24-25, IEEE Xplore Press, Nanjing, Jiangsu, China, pp: 393-397. DOI: 10.1109/ICM.2011.34

Zhang, X., C. Zhu, H. Wu, Z. Liu and Y. Xu, 2017. An imbalance compensation framework for background subtraction. IEEE Trans. Multimedia, 19: 2425-2438. DOI: 10.1109/TMM.2017.2701645

Zhang, Y., D. Zhang, G. Mi, D. Ma and G. Li *et al*., 2012. Using ensemble methods to deal with imbalanced data in predicting protein-protein interactions. Comput. Biol. Chem., 36: 36-41. DOI: 10.1016/j.compbiolchem.2011.12.003

Zhang, X., Z. Liu, H. Li, X. Zhao and P. Zhang, 2014a. Statistical background subtraction based on imbalanced learning. Proceedings of the IEEE International Conference on Multimedia and Expo, Jul. 14-18, IEEE Xplore Press, Chengdu, China, pp: 1-6. DOI: 10.1109/ICME.2014.6890245

Zhang, Y., P. Fu, W. Liu and G. Chen, 2014b. Imbalanced data classification based on scaling kernel-based support vector machine. Neural Comput. Applic., 25: 927-935. DOI: 10.1007/s00521-014-1584-2

Zhang, Y., P. Fu, W. Liu and L. Zou, 2014c. SVM classification for imbalanced data using conformal kernel transformation. Proceedings of the International Joint Conference on Neural Networks, Jul. 6-11, IEEE Xplore Press, Beijing, China, pp: 2894–2900. DOI: 10.1109/IJCNN.2014.6889420

Zhang, Y.P., L.N. Zhang and Y.C. Wang, 2010. Cluster-based majority under-sampling approaches for class imbalance learning. Proceedings of the 2nd IEEE International Conference on Information and Financial Engineering, Sept. 17-19, IEEE Xplore Press, Chongqing, China, pp: 400-404. DOI: 10.1109/ICIFE.2010.5609385

Zhao, Y. and A.K. Shrivastava, 2013. Combating sub-clusters effect in imbalanced classification. Proceedings of the IEEE 13th International Conference on Data Mining, Dec. 7-10, IEEE Xplore Press, Dallas, TX, USA, pp: 1295-1300. DOI: 10.1109/ICDM.2013.105

Zhong, W., B. Raahemi and J. Liu, 2009. Learning on class imbalanced data to classify peer-to-peer applications in IP traffic using resampling techniques. Proceedings of the International Joint Conference on Neural Networks, Jun. 14-19, IEEE Xplore Press, Atlanta, GA, USA, pp: 3548-3554. DOI: 10.1109/IJCNN.2009.5178804

Zhou, C., B. Liu and S. Wang, 2016. CMO-SMOTE: Misclassification cost minimization oriented synthetic minority oversampling technique for imbalanced learning, Proceedings of the 8th International Conference on Intelligent Human-Machine Systems and Cybernetics, Aug. 27-28, IEEE Xplore Press, Hangzhou, China, pp: 353-358. DOI: 10.1109/IHMSC.2016.160

Zhu, C. and Z. Wang, 2017. Entropy-based matrix learning machine for imbalanced 1670 data sets. Patt. Recognit. Lett., 88: 72-80. DOI: 10.1016/j.patrec.2017.01.014

Zoric, B., D. Bajer and G. Martinovic, 2016. Employing different optimisation approaches for SMOTE parameter tuning. Proceedings of the International Conference on Smart Systems and Technologies, Oct. 12-14, IEEE Xplore Press, Osijek, Croatia, pp: 191-196. DOI: 10.1109/SST.2016.7765657

Zughrat, A., M. Mahfouf and S. Thornton, 2015. Performance evaluation of SVM and iterative FSVM classifiers with bootstrapping-based over-sampling and under-sampling. Proceedings of the IEEE International Conference on Fuzzy Systems, Aug. 2-5, IEEE Xplore Press, Istanbul, Turkey, pp: 1-7. DOI: 10.1109/FUZZ-IEEE.2015.7337850