Journal of Computer Science

Review Articles

An Insight into Rare Class Problem: Analysis and Potential Solutions

[1]Satyam Maheshwari, [2]R.C. Jain and [3]R.S. Jadon

[1]Department of Computer Applications, Samrat Ashok Technological Institute, Vidisha, India
[2]Ex-Director, Samrat Ashok Technological Institute, Vidisha, India
[3]Department of Computer Applications, Madhav Institute of Technology and Science, Gwalior, India

Article history
Received: 25-01-2018
Revised: 17-04-2018
Accepted: 08-05-2018

Corresponding Author:
Satyam Maheshwari
Department of Computer Applications, Samrat Ashok Technological Institute,
Vidisha 464001,
India
Email: satyam.vds@gmail.com

Abstract: The class imbalance problem presents an important challenge to the data mining community, in which the number of examples of one class is more than the others. This problem is characterized by a different distribution of cases between all the classes. In this paper, our goal is to study the various challenges of class imbalance problem and provide a comparative study of the current development of research in learning from imbalanced data. We provide a thorough understanding of the nature of the problem, the methods used for data balancing, the learning objectives and assessment metrics used for getting measurable performance, the stated research solutions and the imbalanced problem in multiple classes. This paper highlights the significant opportunities and challenges in the field and provides potential future research directions in the class imbalance problem.

Keywords: Rare Class, Imbalanced Classification, Data Preprocessing, Cost-Sensitive Learning, Ensemble Learning

Introduction

Classification is a vital task in data mining applications. The classification algorithm is used to train the model to predict the class level of unseen data. The various classification algorithms such as Bayesian network, Decision Tree, Support Vector Machine (SVM) and nearest neighbour have been used to predict the class of unknown data. But all the current classification methods give a comparatively balanced class distribution (Chawla *et al*., 2004).

In a class imbalance problem, the number of instances of one class is substantially more than the other classes, and the class that is of more interest (Minority class or Positive) has a few instances when contrasted with negative (Majority class) cases. At the point when a model framed with the imbalanced dataset, it eventually gives its inclination towards common class, because they are designed to maximize the overall prediction accuracy. Therefore, standard classifiers ignore all minority class instances (treating them as noise or outliner) and lose its classification ability in class imbalance problem. For example, in a dataset whose Imbalance Ratio (IR) is 1:100 (i.e., for each occurrence of the smaller class, there are 100 normal class cases). A conventional classifier may acquire a precision of 99% by the ignorance of uncommon examples, with the classification of all instances as the majority. An exact model is one that can give a higher recognizable proof rate of uncommon cases. Therefore, the class imbalance problem is likewise called the rare class problem.

The presence of class imbalance problem in many real-world data attracts much more growth of attention from the research community; allude to the most challenging problem in data mining area (Yang *et al*., 2006). These issues have been seen in a few fields like as credit card fraud detection (Shen *et al*., 2007), medical diagnosis (Mazurowski *et al*., 2008), detection of oil spills from satellite images (Kubat *et al*., 1998), risk management (Ezawa *et al*., 1996), text classification (Cardie *et al*., 1997), modern manufacturing plants (Segal *et al*., 1994). A lot of research has been done in the class imbalance problem due to its use in various practical applications like Machine learning and Data Mining. To address the class imbalance problem, various techniques have been developed. These methods are divided into three parts: (1) the kind of data and data complexity, (2) the possible solution that can predict/identify the class level of unseen data; (3) the appropriate evaluation metrics to measure the classification performance. Within these suggested groups, the most challenging issue is the second one. The possible solutions of the class imbalance problem can be divided into two categories as data level and algorithm level solutions. At the data level (external method), the data is preprocessed in advance which is the primary motive to remove the effect of skewed class distribution.

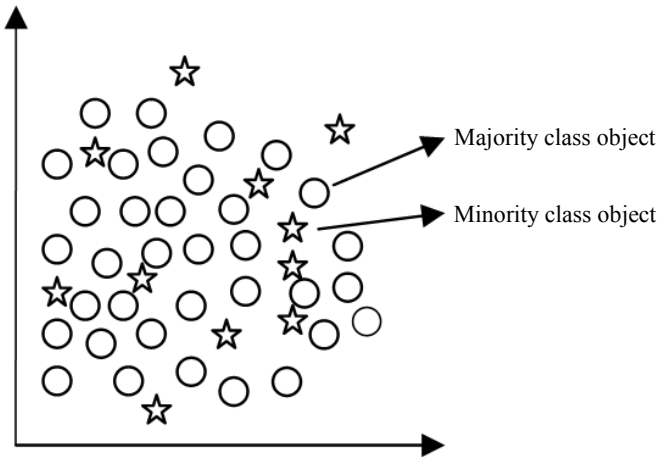

Fig. 1: Imbalanced dataset; stars are out represented by circles

At the algorithm level (internal process), the objective is to create an effective method or change the current one that can bias toward the positive class (Weiss and Provost, 2003). Both the approaches have some drawbacks, some of them are:

1) The data level approach has the drawback of losing some valuable information when majority class samples are under-sampled and over-generalization when minority class samples are over-sampled
2) The disadvantage of the algorithm-based approach is that it requires algorithm specific modification

Fig. 1 demonstrates the dispersion of majority class object and minority class object.

Due to lack of unified framework, we need additional research efforts for the advancement of class imbalance problem. The objective of this paper is to review the two-class imbalance problem and to propose a generalized framework that is appropriate for all types of class imbalance problem.

The organization of the rest of the paper is as follows: In the second section, we discuss the domain of the class imbalance issue. In the third section, we discuss the classification of the imbalance dataset problem, which includes the nature of the problem, data intrinsic characteristics and evaluation criteria for measuring classification performance. The fourth section provides the research objectives of class imbalance learning. The fifth section, exhibit the research solutions for imbalanced learning which incorporate data pre-processing techniques, cost-sensitive learning and various ensemble techniques. In section sixth, we discuss the classification of multi-class imbalance data. Finally, in the seventh section, we present opportunities and challenges for future research in the field and make concluding remarks.

Domain of Class Imbalance Problem

In data mining applications, the class imbalance issue exists in different regions which are of great significance in data mining. The Fig. 2 shows the domain which suffers most due to the class imbalance problem. The following examples briefly illustrate each one:

Medical Diagnosis

All the information about the patient and their medical history are stored in the medical database. The data mining methods applied to these data sets are used to discover the progression and features of certain diseases. This data or information can be used for early identification of diseases. But in the medical domain, disease cases are very infrequent in comparison with typical cases, and the cost of misclassifying will be fatal as possibly influenced patients will be viewed as healthy.

Fraud Detection

Fraud recognition in the money exchange, for example, credit card fraud is costly for each association. Frauds are identified by analyzing the unusual patterns in transaction databases. But, in exchange accumulations, there are more trustworthy clients than a fraudulent transaction. Therefore, it is difficult to find fraud due to rare cases of fraud transactions.

Fault Diagnosis

Due to network-based computer systems, intrusions on systems and networks are overgrowing. Therefore, early detection approaches used to automate and enhance the standard development of fault diagnosis.

Detection of Oil Spills

There are just 2 to 5% of oil flow from the conventional sources. While most of the contamination caused by ships that need to discharge their waste material into the sea. A satellite image-based system could be an

efficient way to find illegal dumping of the waste and could have the significant environmental impact.

The medical diagnostic problem, fraud detection problem and intrusion detection problem are also known as an anomaly detection problem.

Class Imbalance Problem

In this part, we present the nature of the problem; the data intrinsic characteristics, such as lack of density, class overlapping, noisy data, small disjuncts, dataset shift and the evaluation metrics to judge the classifier execution.

Nature of the Problem

The class imbalance issue is the one which has skewed data distribution among its classes. This imbalance form is likewise considered as a between-class imbalance; e.g., 100:1, 1,000:1 and 10,000:1. This problem attracts many research interests from researchers due to several real-world classification problems, like risk management (Huang *et al*., 2006), pollution detection (Lu and Wang, 2008), remote-sensing (Williams *et al*., 2009), fraud analysis (Cieslak *et al*., 2006) and medical diagnosis (Freitas *et al*., 2007; Mazurowski *et al*., 2008).

The significance of the imbalanced problem can be explained with an example from the medical field. Consider the instance of "Mammography dataset" which is used for detection of breast cancer through identification of characteristic masses. By investigating the mammography pictures, collected from a set of distinct patients, the classes that show "positive" or "negative" for a picture illustrative of a "cancerous" or "non-cancerous" patient, individually. In the real world, the non-cancerous patients greatly exceed in the number of cancerous patients, i.e., the "negative" samples outweigh over the "positive" samples. Thus, a classifier is needed that gives a uniform prediction efficiency for both the classes. But in fact, the standard methods have an inclination towards the normal class having accuracies near to 100 percent and the small class having correctness of 0 to 5%. The conventional classifiers overall ignore the small class samples and predict the accuracy close to 100% of the prevalent class.

Therefore, not only between-class imbalance generates an imbalance problem, but also data complexity, such as lack of data, class overlapping, small disjuncts, noisy data and dataset shift also influence the classification accuracy. The Fig. 3 shows the data intrinsic characteristics that also hinder classification performance.

Imbalance Due to Rare Instances

The sample size also plays a vital role in finding the "effectiveness" of a classification model. A data set in which minority class samples are insufficient is known as an imbalance due to rare instances. Japkowicz and Stephen (2002) proposed that when the number of examples of the training set increases, the error rate decreases which are created by imbalanced class distribution. This problem is also referred as *lack of density or lack of information*. When sample size is small, it is tough for algorithms to discriminate rare examples from the prevalent class samples.

Class separability or Overlapping [Fig. 4]

When a sample of one class overlaps on another class, it is known as class overlapping. It is difficult to discriminate in such kind of overlapping classes, and therefore much harder rules are induced to distinguish such examples. The highly overlapped samples decrease the probability of correct classification of the number of minority class instances. Japkowicz and Stephen (2002) proposed that "directly divisible" problems are not easily affected by any measure of imbalance.

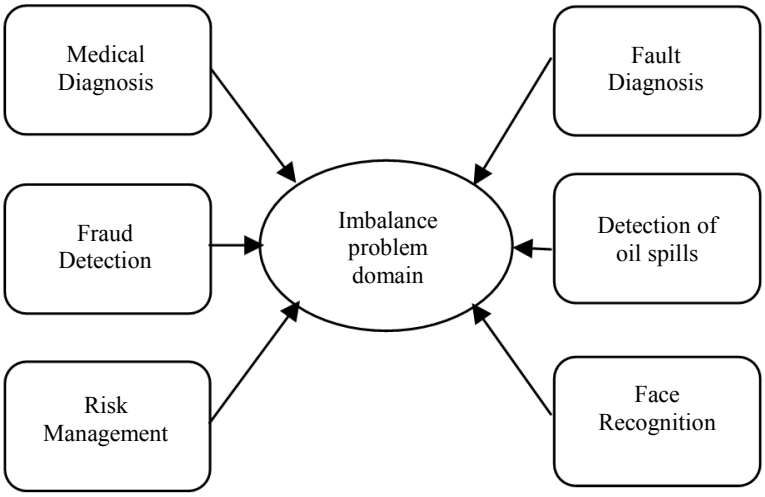

Fig. 2: The domain which affected most due to the Rare Class Problem

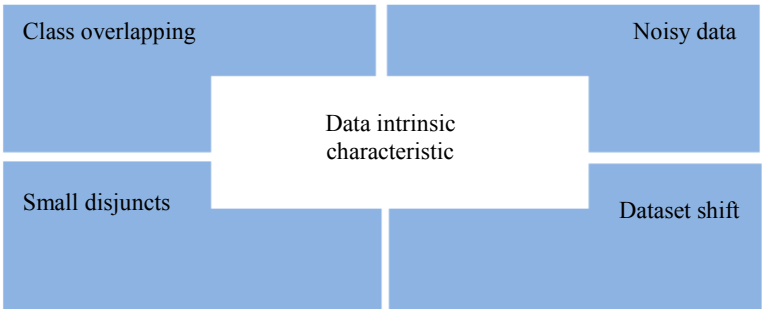

Fig. 3: The data intrinsic characteristics

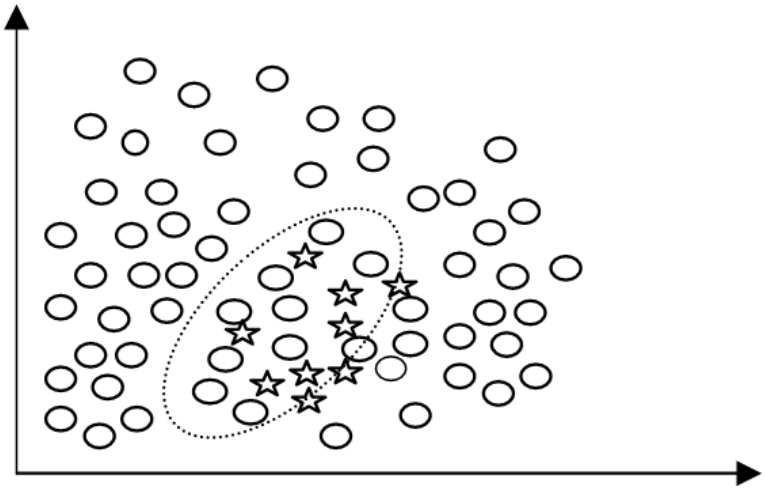

Fig. 4: Class overlapping in imbalanced datasets

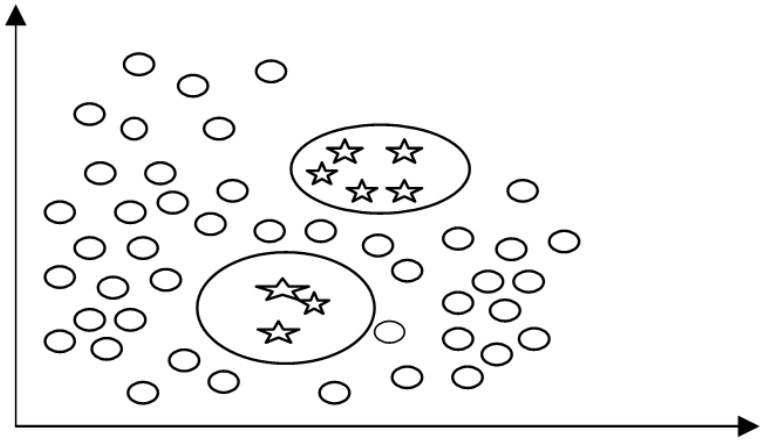

Fig. 5: Small disjuncts in imbalanced datasets

Noisy Data

The presence of the noise significantly affects the minority classes because the small class has few cases. To avoid the noise areas in the learned sub concepts, the over-fitting methods should be applied, such as pruning. The primary disadvantage of this method is that some right minority classes get rejected, so the model ought to be set to give a better overall behaviour for all the class of problem.

Small Disjuncts or Within-Class Concepts [Fig. 5]

The small disjuncts exist in the dataset when the concepts are represented within little groups, where the rare class is framed by sub-concepts (Weiss and Provost, 2003). The complexity of the problem is increased by the presence of the sub-concepts because it is complicated to analyze whether these instances demonstrate actual sub-concepts or are noise examples.

Table 1: Confusion Matrix for a two-class problem

	Predicted Positive		Predicted Negative	
Actual Positive	True (TP)	Positive (FN)	False	Negative
Actual Negative	False (FP)	Positive (TN)	True	Negative

Dataset Shift

The problem of dataset shift occurred because the training and test data have various appropriations. This issue frequently occurred due to sample selection bias. This problem is vital when dealing with the highly imbalanced domain, because the uncommon class is acute to singular classification errors, because of the low number of tests it presents (Moreno-Torres and Herrera, 2010).

Assessment in Imbalanced Domain

The assessment rule plays a crucial role in the evaluation of the classification accuracy and guides the classifier pattern. In the binary-class issue, the confusion matrix (Table 1) exhibits the results of accurately and inaccurately perceived cases of each class. The confusion matrix is used to measure the classification performance of both the prevalent classes and rare classes:

- True Positive Rate: $TP_{rate} = \frac{TP}{TP + FN}$ is the percentage of positive instances correctly classified
- True Negative Rate: $TN_{rate} = \frac{TN}{TN + FP}$ is the percentage of negative instances correctly classified
- False Positive Rate: $FP_{rate} = \frac{FP}{TN + FP}$ is the percentage of negative instances misclassified
- False Negative Rate: $TP_{rate} = \frac{TP}{TP + FN}$ is the percentage of positive instances misclassified

The Equation 1 show accuracy rate, which is the most regularly used trial measures. However, on account of the imbalanced dataset, exactness is no longer an appropriate measure since the rare class has the negligible effect on efficiency as compared to the prevalent class (Joshi *et al*., 2001; Weiss, 2004):

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad (1)$$

The learning objective of classification is: (i) To adjust the recognition capabilities among the two classes; and/or (ii) to enhance the acknowledgement rate of the little class. None of these measures alone is adequate by themselves. Therefore, rather than exactness, other metrics such as Precision, Recall, F-measure and G-Mean are frequently used in the research field to provide extensive evaluations of imbalanced learning problems.

Precision

It is a measure of accuracy, i.e., of the cases labelled as positive, what number of are marked correctly. The precision equation demonstrates that accuracy (Equation 2) is sensitive to data appropriations. However, when it is appropriately used, precision can efficiently measure classification execution in imbalanced learning situations. The shortcoming of precision is that it cannot show that how many positive instances are mislabeled:

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Recall

The recall is a measure of fulfilment, i.e., how many samples of the minority class were marked accurately. The recall equation demonstrates that recall (Equation 3) is not sensitive to data distributions. On the other hand, it doesn't depend on the distribution. The downside of recall is that it doesn't show that what number of occurrences are incorrectly labelled as positive:

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

F-Measure

It is the union of recall and precision which is used for the estimation of the effectiveness of the classification. F-measure is (Lewis and Gale, 1998) used for the unification of precision and recall as an average:

$$F - measure = \frac{2 * Recall * Precision}{Recall + Precision} \quad (4)$$

It provides in-depth observation of the capabilities of the classifier rather than the accuracy measured. However, it is more sensitive to the data distribution. In principle, F-measure is the harmonic mean amongst precision and recall (Tan, Steinbach and Kumar, 2006):

$$F - measure = \frac{2}{1/Recall(R) + 1/Precision(P)} \quad (5)$$

The smaller number of the two is the harmonic mean in F-measure. Hence, if the value of both recall and precision is high, then the estimation of F-measure is likewise high.

G-Mean

It is the ratio of positive and negative accuracy, which estimate the degree of inductive bias. The G-mean (Kubat *et al*., 1998) is defined as:

$$G-mean=\sqrt{TPrate\cdot TNrate} \quad (6)$$

It quantifies the adjusted execution of a learning strategy between positive and negative classes. The difference between arithmetic, harmonic and geometric means are proposed by Tan *et al*. (2006). Though the F-measure and G-mean have improved over exactness measures, there are some shortcomings in specific situations. For example, how might we investigate the execution of various classifiers over a range sample distribution?

ROC Curves

The Receiver Operating Characteristics (ROC) assessment method (Bradley and Bradley, 1997) is used to overcome such issues. In this process, the two-single-column based matrix used for the assessment, i.e., true positive rate (TP_{rate}) and false positive rate (FP_{rate}). These estimations are then connected with the false positive rate on the x-pivot and the true positive rate on the y-pivot. The ROC chart appeared in Fig. 6 gives a visual portrayal of profits and costs of classification with respect to the data dispersion.

The ROC diagram which obtains TP_{rate} = 1 and FP_{rate} = 0 demonstrates the perfect arrangement. Therefore, a classifier is better to another if its relating point in ROC space should be as nearest as possible to the upper left turn in the ROC space. While a classifier whose relating point in ROC point is situated on the diagonal, such as joining the points (TP_{rate} = 0, FP_{rate} = 0), where each case is classified as a negative class and (TP_{rate} = 1, FP_{rate} = 1), where each case is classified as a positive class, is representative of a model that will give a random guess of the class description, i.e., a random classifier. The ROC curve gives a better description of the classification model performance. It is complicated to compare the better model on the basis of ROC curve as one of them has to completely dominate the other model over the entire space (Provost and Fawcett, 1997). The region under a ROC bend (AUC) gives a single measure of a classifier's execution for assessing that which model is superior on average.

ROC observation (Fawcett, 2006) and measurements such as precision, recall and f-measure (Hossin *et al*., 2011) used for estimating the accuracy of learning algorithm of the minority class.

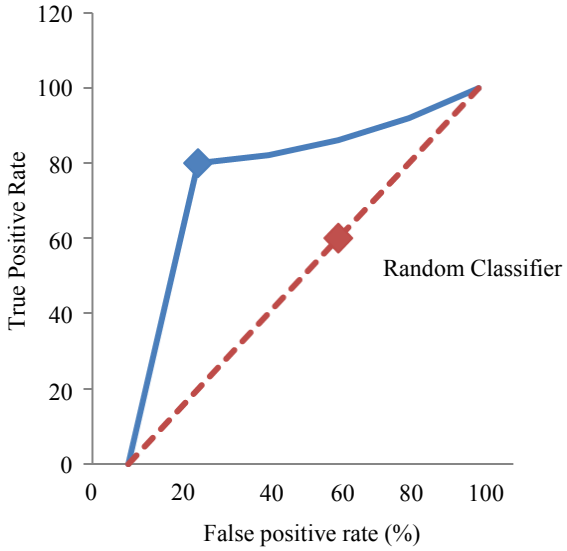

Fig. 6: A Receiver Operating Character (ROC) Plot. The dashed line shows a random classifier, whereas the solid line indicates classifier that is better than the random classifier

Research Objectives

The motivation behind this work is to study the effect of class imbalance issues experienced by the data mining community in a wide range of areas and the use of various methods to enhance classification performance.

In this part, we examine the effect of class irregularity issue in real-world data, and the datasets endure mostly because of a class unevenness issue.

Impact of Class Imbalance Problem on Real-World Data

Rare events are those events, which occurred less frequently as compared to normal events. The developments in learning from class imbalance problem have been motivated by many real-life applications that suffer from skewed data representation. In such cases, the minority class is more important for learning, and hence we need methods for better recognition rates of minority class object.

Some of the real-life issues are identifying malicious attacks, medical diagnostics, fraudulent transactions, sentiment analysis and dealing with exceptional cases in monitoring framework.

Datasets

In this part, we examined the properties of highly skewed datasets which endure more because of the class unevenness issue. We have considered binary datasets from the KEEL dataset repository (Alcalá-Fdez *et al*., 2009; 2011) with various IR; from very imbalanced to low imbalance datasets. Table 2 compresses the qualities of datasets, the number of examples (#Ex.), number of attributes (#Atts.), the class attribute distribution and the IR.

Table 2: The features of the dataset from KEEL collection

Dataset	#Ex.	#Atts	% class (min., maj.)	IR
Abalone19	4174	8	0.77,99.23	128.87
Yeast6	1484	8	2.49,97.51	39.15
Ecoli0137vs26	281	7	2.49,97.51	39.15
Yeast5	1484	8	2.96,97.04	32.78
Yeast1289vs7	947	8	3.17,96.83	30.56
Yeast4	1484	8	3.43,96.57	28.41
Yeast2vs8	482	8	4.15,95.85	23.10
Glass5	214	9	4.20,95.80	22.81
Yeast1458vs7	693	8	4.33,95.67	22.10
Shuttle2vs4	129	9	4.65,95.35	20.50
Glass2	214	9	8.78,91.22	10.39
Yeast1vs7	459	8	6.72,93.28	13.87
Glass4	214	9	6.07,93.28	13.87
Abalone9vs18	731	8	5.65,94.25	16.68
Ecoli4	336	7	6.74,93.26	13.84
Ecoli0147vs56	332	6	7.53,92.47	12.28
Led7digit02456789vs1	443	7	8.35,91.65	10.97

This table is organized according to Imbalanced Ratio (IR) from high to low imbalanced.

Research Solutions

Because of the significance of the imbalanced dataset issue, a large number of solutions are reported in the literature.

These courses of action are arranged into data level, algorithm level, cost-sensitive learning and ensemble learning, depending on how they deal with the issue.

Data Preprocessing Methods

In data preprocessing method (likewise called external approach), the essential target is to rearrange the class allocation by re-sampling the data space (Batista *et al*., 2004; Fernández *et al*., 2008; Napierała *et al*., 2010; Stefanowski and Wilk, 2008). The solution at the data level contains the modification of an imbalanced dataset by some mechanism to provide a balanced distribution. The advantage of this approach is that it is more flexible and its utilization is autonomous of the fundamental classifier. This method can be classified into three categories:

- *Under-sampling techniques* form a subset of the initial dataset by disposing of a portion of the occurrences of prevalent class
- *Over-sampling techniques* built the superset of the initial dataset by copying a portion of the occurrences of the fewer class or creating new ones from the current fewer class samples
- *Hybrid techniques* are the combination of both the above methods, which will expand the extent of the small class and slowly diminish the widespread class

Under-Sampling Methods

Following techniques falls under under-sampling methods:

"*Random Under-Sampling*" (RUS)

Batista *et al*. (2004) is a non-heuristic method which will adjust the class dispersion by taking out the prevalent class samples. The primary disadvantage of this method is that it can eliminate valuable data that could be vital for the learning procedure.

"*Tomek Links*" (TL)

Tomek (1976) can be utilized as under-sampling or as a data-cleaning technique. As an under-sampling method, only samples associated with the dominant class disposed of and as a data- cleaning technique, samples of two classes (prevalent classes and rare classes) are discarded. Tomek Link algorithm work as follows: consider two instance pair (x, y), where x refers to the prevalent class while y refer to the rare class. The distance between these two observations is denoted by $d(x,y)$, a pair (x,y) is called TL if there is no sample z, such that $d(x,z)<d(x,y)$ or $d(y,z)<d(x,y)$. If two examples form a Tomek Links (TL), at that point both of these samples is noise, or the two samples are borderline. The significant advantage of the Tomek Links is that it expels overlapping between classes.

"*One-Sided Selection*" (OSS)

Kubat and Matwin (1997) is an under-sampling method in which the excess, noisy and borderline examples are recognized and expelled from the majority class. OSS is method resulting from the application of Tomek Link (TL) followed by the application of the Condensed Nearest Neighbour (CNN) rule (Hart, 1968). Firstly, TL method is used to remove noise, and borderline majority class samples and then CNN is used to discard samples from the majority class that is redundant and far away from the decision border. The remaining observations, i.e., "safe" majority class as well as minority class instances are used for training.

"*Neighbourhood Cleaning Rule*" (NCL)

Wilson (1972) is an under-sampling method that removes some of the examples from the majority class.

This cleaning algorithm work as follows: For each sample x_i in the training set, its three nearest neighbours are identified. If x_i refers to the prevalent class and the classification given by its three closest neighbours, differ from the original class of x_i, and then x_i is eliminated. If x_i refers to the rare class and its three closest neighbours misclassify x_i, then the nearest neighbours that relate to the prevalent class are excluded. The primary advantage of this method is that it evacuates noisy cases.

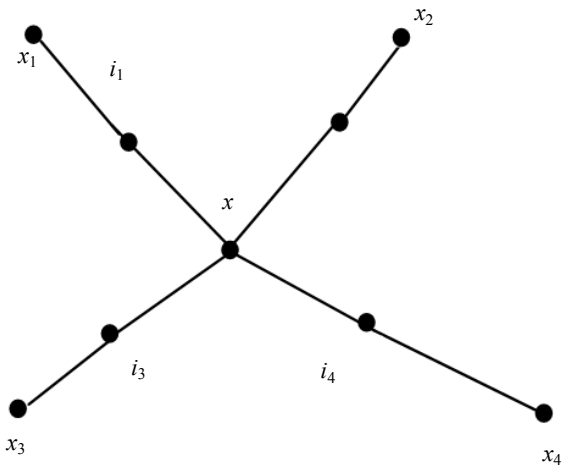

Fig. 7: The figure shows how to generate synthetic data points using SMOTE method

"*Condensed Nearest Neighbour Rule*" (CNN)

Hart (1968) is used to locate a constant subgroup of observations. A subgroup X of Y is constant with Y if applying a 1-NN, X accurately classifies the observations in Y. This algorithm arbitrarily discovers one prevalent class samples and all samples from the rare class and put these cases in X. After that, use a 1-NN over the samples in X to classify the samples in Y. Each misclassified case from Y is displaced to X. The objective of this approach is to remove the samples from the prevalent class that is far-away from the decision border because these instances might be less useful for training.

Over-Sampling Method

This method creates duplicate samples of minority class to balance the ratio of majority class as well as minority class samples. Some of the over-sampling techniques are:

"*Random Over-Sampling*" (ROS)

Batista *et al*. (2004) is a non-heuristic strategy that means to adjust class circulation through randomly replicating minority class samples. The significant inadequacy of this approach is that it can improve the probability of happening over-fitting since it makes precise of existing examples.

"*Synthetic Minority Over-Sampling Technique*" (*SMOTE*)

Chawla *et al*. (2002) is an over-sampling technique in which the primary focus is to develop new minority class instances by interpolating many small class examples that lie together for oversampling the learning set. In this method, the small class is over-sampled by taking each small class instance and introducing synthetic samples along the line segments that join any/all (k) nearest small class neighbours. Depending on the oversampling ratio, neighbours are randomly chosen from the k-nearest neighbours. However, in the SMOTE method, the issue of over-generalization is mainly associated due to the method in which synthetic examples are developed. In Fig. 7, x is the selected points, x_1 to x_4 are selected k-nearest neighbours (in our case 4-NN), and i_1 to i_4 is synthetic data points generated by SMOTE method.

Hybrid Methods

Hybrid methods combine both under-sampling and over-sampling methods. Some hybrid methods are:

- "*SMOTE + Tomek Links* (*TL*)" is used to remove the problem of overfitting associated with SMOTE method. To overcome the overfitting problem and to build a better-defined class cluster, we apply Tomek Links (Tomek., 1976) to the over-examined learning set as a data cleaning strategy. Thus, both prevalent, as well as rare class samples, is discarded to form a balanced training set.
- "*SMOTE + ENN*" (Wilson, 1972) is same as SMOTE + Tomek links. The ENN is used to discard additional samples than the Tomek links, by providing more in-depth data cleaning. ENN is used to remove samples from both majorities as well as minority classes. Thus, any sample that is misclassified by its 3-NN is discarded from the learning set.

Algorithm-Based Methods

At the algorithm level approaches (also called internal), the solution tries to acquire the properties of the existing algorithms which will bias the learning toward to the small classes (Liu *et al*., 2000; Lin *et al*., 2002; Barandela *et al*., 2003a). This technique creates the new method or modifies existing processes, taking into account the significance of minority class samples (Barandela *et al*., 2003b), (Zadrozny and Elkan, 2001; Diamantini and Potena, 2009; Cieslak *et al*., 2012).

The idea behind the algorithm based approach is to select a suitable induction bias that is used to deal with the class imbalance problem. The objective of modifying algorithm is to provide adjustments to the training algorithm, mainly with Decision Tree and SVM. In the decision trees, one method is to change the probabilistic guess at the tree leaf (Zadrozny and Elkan, 2001; Quinlan, 1991); another way is to develop new pruning methods. In SVM, different penalty constant for different classes (Lin *et al*., 2002), or modifying the class boundary based on kernel-alignment (Wu and Chang, 2003) is used.

The major shortcoming of the algorithm-level approach is that it requires knowledge of both the corresponding classifier and the application domain, especially why the training algorithm fails when the class allocation of available data is uneven.

Cost-Sensitive Learning

The cost-sensitive method combines both the data level and algorithm-based approach, assuming greater misclassification cost with examples of the fewer class to the majority class, and thus, seek to curtail the high-cost errors (Chawla *et al*., 2008; Ling *et al*., 2006; Zhang *et al*., 2008).

Cost-sensitive learning biases the classifier toward the small class and therefore the rare class gain importance. The fundamental element of this learning is that it tends to limit the overall misclassification cost. In this way, the cost related to a rare example must be more noteworthy than the cost of misclassifying a predominant one, i.e., Cost(+,-)>Cost(-,+). The accurate classification shows nil penalties, i.e.,C(+,+) = C(-,-) = 0.

The major shortcoming of this approach is that the costs are precisely unknown and we usually tend to utilize approximations or proportions of proportionality. The cost-sensitive learning divided into three types:

1. The first approach based on modifying the training data. In this method, resampling is applied to the original class distribution of the learning dataset by a cost decision matrix using oversampling, undersampling or assigning instance weights. This technique can be clarified by the *interpretation hypothesis* (Zadrozny *et al*., 2003)
2. The second approach changes a specific classifier learning algorithm to develop a cost-sensitive classifier. For instance, in the case of decision tree induction, the tree-building method is used to decrease the misclassification costs. The cost information used to, (i) select the best attribute to split the data (Ling and Li, 2004; Segal *et al*., 1994); and (ii) choose whether a sub-tree ought to be pruned (Bradford *et al*., 1998)
3. The third approach uses Bayes decision theory to attach each example to the class with a lowest expected cost. For instance, a decision tree for a two-class issue attaches the class label of a leaf node, depending on the prevalent class of the training examples that reach the node. A cost-sensitive method selects the class label from the node that decreases the classification cost (Domingos and Metacost, 1999; Zadrozny and Elkan, 2001)

This approach considers that a cost-matrix is accessible for distinct types of errors or examples. However, for a given data set, this matrix is ordinarily unavailable.

In Table 3, the $C(i, i)$ linked with True Positive(TP) or True Negative(TN), treated as the observation that correctly classified in both the cases.

Table 3: Cost matrix

	Predicted Positive '1'	Predicted Negative '0'
Actual Positive '1'	C (1,1) TP	C (0,1) FN
Actual Negative '0'	C (1,0) FP	C (0,0) TN

The rare class considered the minority or positive class, which have more significance for learning. It is more costly to misclassified an actual positive (TP) as negative (FN), to classify an actual negative (TN) as a positive example (FP), i.e., the value $C(0,1)$ assigned to FN is larger than that of $C(1,0)$ related to FP. That is what we analyze from some of the typical cases such as bank scam, medical diagnosis and customer retention.

Ensemble Learning

The basic concept of ensemble learning is to attempt to enhance the execution of single classifiers by inciting various classifiers and then combined their predictions to acquire a new classifier that exceed each one of them. This technique follows the natural human behaviour that tends to look for a few assumptions previously settling on any crucial decisions. The keys to the proper performance of ensembles are "diversity", that is achieved by the combination of ensemble methods and one of the methods, sampling or cost-sensitive learning solutions. The most broadly used ensemble learning methods are Bagging (Breiman, 1996) and AdaBoost (Friedman *et al*., 2000; Schapire, 2002), which are most successful in variance reduction. The ensemble methods are of two types: Bagging and Boosting.

Bagging (*Breiman*, *1996*)

It is bootstrap aggregating to create ensembles. It consists of learning a set of classifiers (every one with an alternate subset, recognized as "bag") with bootstrapped replicas of the initial learning dataset. The random drawing (with replacement) used to form a new dataset so that the initial dataset size maintained. Hence, diversity achieved using resembling method by the use of different datasets. When classifying an unknown sample, all individual classifiers used and a majority or weighted vote is used to deduce the class.

Bagging is the method that is adopted by maximum ensemble methods for imbalanced classification. It is because of its easiness in the combination of data processing methods into Bagging, which is made when each bootstrap copy is calculated.

SMOTEBagging (*Wang and Yao*, *2009*)

In SMOTEBagging, each base classifier is obtained from a random example of learning data. This method combines bagging with SMOTE and over-sampling in each round so that the dataset is entirely balanced. This technique can be developed in two ways: (i)

Bootstrapped replica of the majority class examples; and (ii) using SMOTE and random over-sampling relying upon resampling rate.

UnderBagging (*Barandela et al*., *2003a*)

This method arbitrarily under-samples the dataset in each Bagging round. This methodology is typically connected to the majority class.

Roughly Balanced Bagging (*Hido et al*., *2009*)

It is based on undersampling; however, it doesn't bootstrap a balanced bag. In this method, the quantity of minority samples is kept settled, whereas the number of majority samples is determined using negative binomial distribution.

Boosting (Schapire, 1990): This method needs the entire dataset to learn each classifier serially. However, in each iteration, the algorithm gives more attention to the different examples, with the purpose of correctly classifying those samples that were misclassified in the earlier round. That can be accomplished by equally assigning weights for all samples at the beginning. Then, after each round, incorrect examples get their weights increased; in the country, the weights of correctly classified samples are decreased. Additionally, each classifier assigned another weight depending on its general exactness over the learning set; more certainty allocated to more accurate classifiers. At last, when the new example is presented, every classifier gives a weighted vote (as indicated by its weight) and the majority selects the class mark.

SMOTEBoost (*Chawla et al*., *2003*)

In boosting, the entire training set is utilized to train every classifier. SMOTEBoost introduce synthetic samples using SMOTE method. Since new samples are produced, the weights of the new instances are corresponding to the total number of samples in the new dataset. The weights of the samples from the original dataset are standardized in such a way that they frame an allocation with the new samples.

RUSBoost (*Seiffert et al*., *2010*)

RUSBoost removes examples from the majority class every iteration using the random undersampling procedure. The weights of the remaining samples in the new datasets are normalized to form a balanced distribution.

AdaBoost (*Freund and Schapire*, *1997*)

AdaBoost is most widely used ensemble training method reported to be equipped for bias reduction (Friedman *et al*., 2000). This technique weights each example following its value and put the maximum weight on the examples which are normally misclassified by the past classifiers. AdaBoost try to diminish the bias error as its target is on misclassified examples (Freund and Schapire, 1996). The instance weighting approach of AdaBoost is similar to resampling the data space by integrating both down-sizing and up-sampling. Consequently, this method applies to data-level methods, which apply to most classifications rule without modifying their base learning techniques.

In a given dataset with skewed class distribution, minority class samples are often misclassified. At the point when the AdaBoost method is used, examples of the small class received more weights, and subsequent learning will target of the small class. At first sight, it shows that the AdaBoost method may enhance the prediction accuracy of the rare class. However, some empirical results stated that the improved performance of the rare class is not always guaranteed or up to the mark. The reason behind this is that AdaBoost algorithm is efficiency-driven: Its weighting approach may incline toward the majority class which constitutes more to the general order precision. Subsequently, the issues turn out to be the way to adjust AdaBoost strategy to inclination its boosting approach towards the interested class.

Cost-Sensitive Boosting

The drawback of AdaBoost is that it is an efficiency-oriented method when the class circulation is skewed; this technique inclines the learning toward the majority class since it concentrates more on the general precision. But, the learning objective of the class imbalance problem is to enhance the recognizable rate of the minority class. Therefore, a desirable boosting approach is one who can differentiate distinct kind of examples and hike more weights of those examples which have higher recognition significance.

The Cost-sensitive boosting method keeps the general learning structure of AdaBoost, however, in the meantime recommend the cost items into the weight update method. Based on these, three cost-sensitive boosting strategies AdaC1, AdaC2 and AdaC3 (Sun *et al*., 2007) are implemented with weight update formula.

The Fig. 8 shows the research taxonomy to address the class imbalance problem. These arrangements can be classified into four different categories such as data-level solutions, solutions at the algorithm level, cost-sensitive learning and Ensemble learning. Here, cost-sensitive boosting is like cost-sensitive strategies, but the cost minimization is directed by boosting technique.

The analysis drawn from a comparative study of each reported research solution is shown in Table 4.

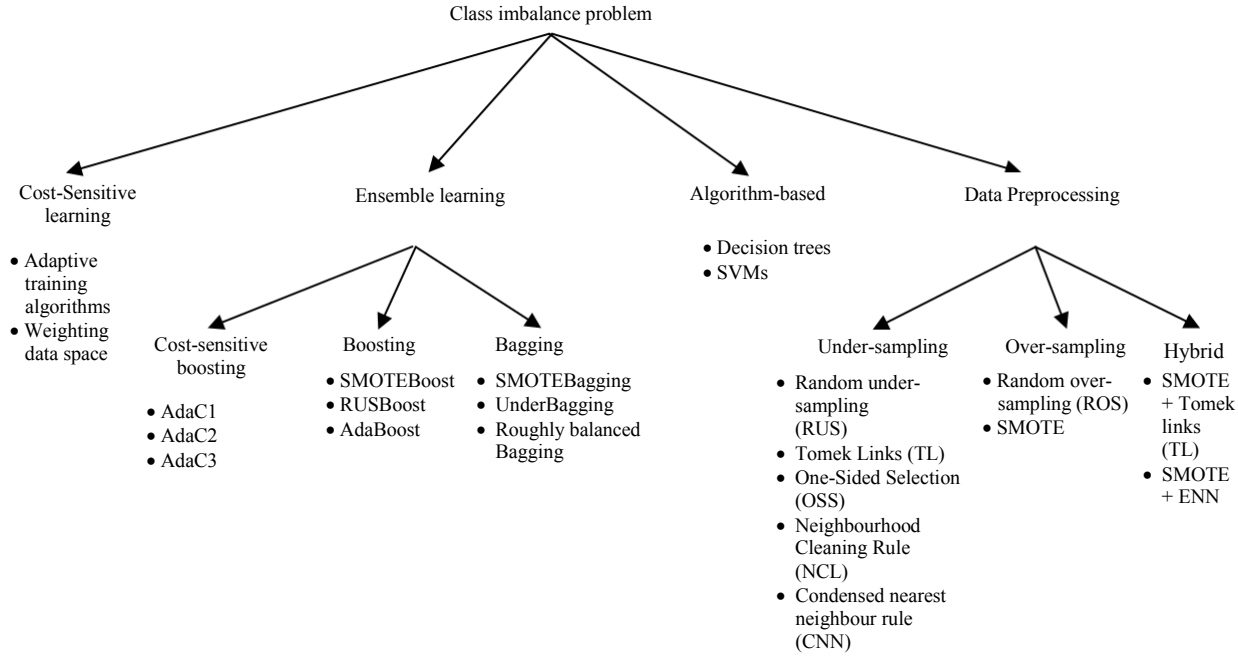

Fig. 8: The research taxonomy to address the class imbalance problem

Table 4: Comparative study of research solutions

Approach	Methods & its description	Algorithms	Advantages	Disadvantages
Data level approach	• Under-sampling - It removes examples of prevalent class	• Random under-sampling • One-sided Selection (OSS) • Neighbourhood Cleaning Rule(NCL) • Tomek Links • Condensed nearest neighbour rule (CNN)	• This method is more adaptable and free from classifier selected, in this way the data should be arranged once for classification • Can be easily implemented	• This method sometimes removes the essential data, which may be useful for learning process
	• Over-sampling – It adds new examples of existing class	• Random over-sampling • Synthetic Minority Oversampling Technique (SMOTE)	• This method creates rules which are precise and furthermore used to enhance the exactness of classification	• May lead to over-fitting • Time-consuming: Introduce additional computational cost
	• Hybrid- It combines both over-sampling and under-sampling methods	• SMOTE + Tomek links • SMOTE + ENN	• This method removes the problem of over-sampling, yet not by chopping down the measure of the dominant classes	• It takes longer training time
Algorithm based method	• This approach creates the new method or modify existing methods	• Decision tree (C4.5) • SVMs	• Effective in certain context	• It requires algorithm-specific modifications • It requires the insight of both the relating classifier learning methods and the application area
Ensemble based learning	• Bagging method –This method enhances the strength and precision of ensemble learning methods	• SMOTEBagging • UnderBagging • Roughly balanced Bagging	• It decreases discrepancy • Its performance is better than individual classifiers • More resilience to noise	• It is more time consuming • The problem of over-fitting
	• Boosting method- It is iterative method, in which after every cycle the weight of misclassified perceptions increments, while weight of accurately classified diminishes	• AdaBoost • SMOTEBoost • RUSBoost	• This method increases the performance of other learning algorithms	• It ignores the general efficiency of the classifier
	• Cost-sensitive Boosting	• AdaC1 • AdaC2 • AdaC3		
Cost-sensitive approach	• It combines both data based approach and algorithm based method when the misclassification cost is huge		• Minimize the cost of misclassification (by biasing the classifier toward the minority class)	• Cost isn't exactly known, need to utilize approximations or ratios of proportionate

Noteworthy Contribution in the Field

Dongre and Malik (2017) reported that data adjusting provide the better solution than other techniques. But Dongre and Malik (2017) suggest that a hybrid approach gives the best solution for class imbalance learning.

Interesting research which highlights a new future direction in imbalanced learning was proposed by Krawczyk (2016). This research analyzed different features of imbalanced learning such as classification, clustering, regression and big data analytics. The author in its future directions focused on the structure and nature of samples in rare classes to gain a better insight into the source of learning difficulties. In our review, we addressed some of the data intrinsic characteristics such as the size of the dataset and the lack of density, class overlapping, presence of noisy data, the problem due to the presence of small disjuncts and the dataset shift problem. The Krawczyk (2016) highlighted all the issues and challenges and provided future research directions in the field and also lead to advancing our understanding of the imbalanced learning system.

But there is still the untouched area that makes imbalanced learning fresh and exciting for the research community and future development.

Classification of Multi-Class Imbalanced Data

In practice, real-world data may have more than two classes, which imply an additional difficulty in the classification performance. Some issues addressed are, the boundaries between the classes may overlap, small sample size or small disjuncts (small class can consist of several subconcepts). These issues are significantly more challenging in the multi-class problems.

These issues can be tackled by developing sophisticated techniques for handling multi-class imbalance problems. In a two-class problem, we have a well-defined relationship between classes: One is considered as the smaller class and other as the prevalent class. The resolutions at the data level modify the class size ratio of the binary classes, either by under-sampling the prevalent class or over-sampling the smaller class and iterate the training method multiple times in search of uniform allocation. But in case of multiple classes, these arrangements are not applicable because of the expanded search area. Similarly, algorithm-based solutions attempt to modify the training methods to incline towards the rare classes. In the multi-class problem, there are many smaller classes exist, and in this position, it becomes difficult to modify according to the training method. Therefore, multi-class imbalance issue will be composed of two simple steps:

1. The initial stage makes out of the detachment of the multi-class imbalance issue into simple binary-class subproblems
2. In the second step, for every subproblem obtained, execute the existing solutions that have been developed to handle the binary-class imbalanced datasets

There are many methods for dealing with multi-class issues:

The One-Versus-One (OVO) and One-Versus-All (OVA) are two common methods to diminish a multi-class classification issue to a set of two-class issues.

One-Versus-One Approach (1 – 1)

In OVO approach (Hastie and Tibshirani, 1998) a classifier is prepared for every match of classes, disregarding the samples that don't have a place with the related classes. When classifying examples, a query is submitted to all binary models, and the choices of these models are consolidated into overall classification (Hüllermeier and Brinker, 2008; Hüllermeier and Vanderlooy, 2010).

One-Versus-All Approach (1 – a)

In OVA approach (Rifkin *et al*., 2004) a classifier among one class and the all other classes are learned (in total “a” classifiers). In OVA approach all the instances of the current class to be considered the minority and the remaining examples the majority.

The problems with these methods are that the two cases increment the training cost and may prompt ties or clashing voting (Duda *et al*., 2001) among the diverse classes.

Conclusion and Future Research

In this work, we investigated the cutting edge research developments on the classification of imbalanced data. In the paper, first we discuss the many examples of application domains that suffer mostly due to class imbalance problem; explain the nature of the class imbalance problem, mainly, we discuss that the imbalance rate by itself does not have the most significant result on classifiers performance, however there are numerous problems that affect for obtaining high classification efficiency for both classes of the problem, i.e., lack of data, the class overlapping, noisy data, small disjuncts and the dataset shift.

We briefly reviewed impact of the class imbalance problem on real-world data; analyzed the properties of highly skewed datasets which suffer most due to the class imbalance problem; provide a comprehensive review of stated research results to this issue, namely, processing of instance, algorithm-based solutions, cost-sensitive learning and ensemble learning with their pros and cons in an attempt to investigate state-of-the-art research directions in subsequent study; analyzed the impact of classification performance in the presence of multiple classes as opposed to only binary-classes.

This review paper mostly emphasizes the research efforts to explore only the solutions of two-class imbalance problem. Therefore, there is an exciting research issue open for future research in the multi-class imbalance problem.

Finally, we emphasize that this work provides the compressive review of current research work and solutions for imbalance datasets with binary classes. Due to huge potential applications, the problem of imbalanced data will attract major consideration in research and scientific communities.

Acknowledgement

The authors would like to thank the reviewers for their valuable comments and suggestions that contributed to the improvement of this work.

Author's Contributions

Satyam Maheshwari: Designed the research plan, organized the study and reviewed the various published articles in the field and contributed to the writing of the manuscript.

R.C. Jain and R.S. Jadon: Coordinated the data analysis and contributed to the writing of the manuscript.

Ethics

The authors confirm that they have read and approved the manuscript and there is no conflict of interest. This work is original and not published elsewhere.

References

Alcalá-Fdez, J., A. Fernández, J. Luengo, J. Derrac and S. García *et al*., 2011. KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. J. Multiple-Valued Logic Soft Comput., 17: 255-287. DOI: 10.1007/s00500-008-0323-y

Alcalá-Fdez, J., L. Sánchez, S. García, M. J. del Jesus and S. Ventura *et al*., 2009. KEEL: A software tool to assess evolutionary algorithms for data mining problems. Soft Comput., 13: 307-318. DOI: 10.1007/s00500-008-0323-y

Barandela, R., R.M. Valdovinos and J.S. Sánchez, 2003a. New applications of ensembles of classifiers. Pattern Analysis Applic., 6: 245-256. DOI: 10.1007/s10044-003-0192-z

Barandela, R., J.S. Sã Anchez, V. Garcã and E. Rangel, 2003b. Strategies for learning in class imbalance problems. Pattern Recognit., 36: 849-851.

Batista, G.E.A.P.A., 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. Newsletter, 6: 20-29. DOI: 10.1145/1007730.1007735

Bradford, J.P., C. Kunz, R. Kohavi, C. Brunk and C.E. Brodley, 1998. Pruning decision trees with misclassification costs. Proceedings of the 10th European Conference on Machine Learning, (CML' 98), Chemnitz, Germany, pp: 131-136. DOI: 10.1007/BFb0026682

Bradley, A.P. and A.P. Bradley, 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit., 30: 1145-1159.

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

Cardie, C. and N. Howe, 1997. Improving minority class prediction using case-specific feature weights. Proceeding of the 14th International Conference on Machine Learning, (CML' 79),. Nashvile, TN., Morgan Kaufmann, pp: 57-65.

Chawla, N.V, K.W. Bowyer, L.O. Hall an W.P. Kegelmeyer, 2002. SMOTE: Synthetic minority over-sampling technique. J. Artificial Intelligence Res., 16: 321-357.

Chawla, N.V., A. Lazarevic, L.O. Hall and K.W. Bowyer, 2003. SMOTEBoost: Improving Prediction of the Minority Class in Boosting. In: Knowledge Discovery in Databases PKDD 2003, Lavrač, N., D. Gamberger, L. Todorovski and H. Blockeel (Eds.),. Springer, Berlin, Heidelberg, ISBN-10: 978-3-540-39804-2, pp: 107-119.

Chawla, N.V., D.A. Cieslak, L.O. Hall and A. Joshi, 2008. Automatically countering imbalance and its empirical relationship to cost. Data Min. Knowledge Discovery, 17: 225-252. DOI: 10.1007/s10618-008-0087-0

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

Cieslak, D.A., N.V. Chawla and A. Striegel, 2006. Combating imbalance in network intrusion datasets. Proceedings of the International Conference on Granular Computing, May 10-12, IEEE Xplore Press, Atlanta, GA, USA, pp: 732-737. DOI: 10.1109/GRC.2006.1635905

Cieslak, D.A., T.R. Hoens, N.V. Chawla and W.P. Kegelmeyer, 2012. Hellinger distance decision trees are robust and skew-insensitive. Data Min. Knowledge Discovery, 24: 136-158. DOI: 10.1007/s10618-011-0222-1

Diamantini, C. and D. Potena, 2009. Bayes vector quantizer for class-imbalance problem. IEEE Trans. Knowledge Data Eng., 21: 638-651. DOI: 10.1109/TKDE.2008.187

Domingos, P. and P. Metacost, 1999. MetaCost: A general method for making classifiers cost-sensitive. Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining, Aug. 15-18, ACM, San Diego, California, USA, pp: 155-164. DOI: 10.1145/312129.312220

Dongre, S.S. and G.L. Malik, 2017. Rare class problem in data mining: Review. Int. J. Adv. Res. Comput. Sci., 8: 1102-1105. DOI: 10.26483/ijarcs.v8i7.4530

Duda, R.O., P.E. Hart and D.G. Stork, 2001. Pattern Classification. 2nd Edn., Wiley, New York, ISBN-10: 0471056693, pp: 654.

Ezawa, K., K. Ezawa, M. Singh and S.W. Norton, 1996. Learning goal oriented bayesian networks for telecommunications risk management. Proceedings of the 13th International Conference on Machine Learning, (CML' 96), pp: 139-147.

Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognit. Lett., 27: 861-874. DOI: 10.1016/j.patrec.2005.10.010

Fernández, A., S. García, M.J. del Jesus and F. Herrera, 2008. A study of the behaviour of linguistic fuzzy rule based classification systems in the framework of imbalanced data-sets. Fuzzy Sets Syst., 159: 2378-2398. DOI: 10.1016/j.fss.2007.12.023

Freitas, A., A. Costa-Pereira and P. Brazdil, 2007. Cost-Sensitive Decision Trees Applied to Medical Data. In: Data Warehousing and Knowledge Discovery, Song, I.Y., J. Eder, T.M. Nguyen (Eds.), Springer, Berlin Heidelberg, ISBN-10: 978-3-540-74553-2, pp: 303-312.

Freund, Y. and R.E. Schapire, 1996. Experiments with a new boosting algorithm. Proceedings of the 13th International Conference on Machine Learning, (CML' 96), The Mit Press, Cambridge, MA, Morgan Kaufmann, Los Altos, CA, pp: 1-9.

Freund, Y. and R.E. Schapire, 1997. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci., 55: 119-139.

Friedman, J., T. Hastie and R. Tibshirani, 2000. Additive logistic regression: A statistical view of boosting. Annals Statist., 28: 337-407. DOI: 10.1214/aos/1016218223

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

Hastie, T. and R. Tibshirani, 1998. Classification by pairwise coupling. Annals Statist., 26: 451-471. DOI: 10.1214/aos/1028144844

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

Hossin, M., M.N. Sulaiman, N. Mustpaha and R.W. Rahmat, 2011. Improving accuracy metric with precision and recall metrics for optimizing stochastic classifier. Proceedings of the 3rd International Conference on Computing and Informatics, Jun. 8-9, Bandung, Indonesia, pp: 105-110.

Huang, Y.M., C.M. Hung and H.C. Jiau, 2006. Evaluation of neural networks and data mining methods on a credit assessment task for class imbalance problem. Nonlinear Analysis: Real World Applic., 7: 720-747. DOI: 10.1016/j.nonrwa.2005.04.006

Hüllermeier, E. and K. Brinker, 2008. Learning valued preference structures for solving classification problems. Fuzzy Sets Syst., 159: 2337-2352. DOI: 10.1016/J.FSS.2008.01.021

Hüllermeier, E. and S. Vanderlooy, 2010. Combining predictions in pairwise classification: An optimal adaptive voting strategy and its relation to weighted voting. Pattern Recognit., 43: 128-142. DOI: 10.1016/j.patcog.2009.06.013

Japkowicz, N. and S. Stephen, 2002. The class imbalance problem: A systematic study. Intelligent Data Anal., 6: 429-449. DOI: 10.1.1.711.8214

Joshi, M.V., V. Kumar and R.C. Agarwal, 2001. Evaluating boosting algorithms to classify rare classes: Comparison and improvements. Proceedings of the 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

Krawczyk, B., 2016. Learning from imbalanced data: Open challenges and future directions. Progress Artificial Intelligence, 5: 221-232. DOI: 10.1007/s13748-016-0094-0

Kubat, M. and S. Matwin, 1997. Addressing the curse of imbalanced training sets: One-sided selection. Proceedings of the 14th International Conferrence 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. Machine Learn., 30: 195-215. DOI: 10.1023/A:1007452223027

Lewis, D. and W. Gale, 1998. Training text classifiers by uncertainty sampling. Proceedings of the17th Annual International Conference Research and Development in Information, (RDI' 98), New York, NY, pp: 73-79.

Lin, Y., Y. Lee and G. Wahba, 2002. Support vector machines for classification in nonstandard situations. Machine Learn., 46: 191-202.

Ling, C.X. and C. Li, 2004. Decision trees with minimal costs. Proceedings of the 21st International Conference on Machine Learning , Jul. 04-08, ACM Press, Banff, Canada, pp: 1-69. DOI: 10.1145/1015330.1015369

Ling, C.X., V.S. Sheng and Q. Yang, 2006. Test strategies for cost-sensitive decision trees. IEEE Trans. Knowledge Data Eng., 18: 1055-1067. DOI: 10.1109/TKDE.2006.131

Liu, B., Y. Ma and C.K. Wong, 2000. Improving an Association Rule Based Classifier. Proceeedings of the European Conference on Principles of Data Mining and Knowledge Discovery, (MKD' 00), Springer, Berlin, Heidelberg, pp: 293-317. DOI: 10.1007/3-540-45372-5_58

Lu, W.Z. and D. Wang, 2008. Ground-level ozone prediction by support vector machine approach with a cost-sensitive classification scheme. Sci. Total Environ., 395: 109-116. DOI: 10.1016/j.scitotenv.2008.01.035

Mazurowski, M.A., P.A. Habas, J.M. Zurada, J.Y. Lo and J.A. Baker *et al*., 2008. Training neural network classifiers for medical decision making: the effects of imbalanced datasets on classification performance. Neural Networks: Official J. Int. Neural Network Society, 21: 427-436. DOI: 10.1016/j.neunet.2007.12.031

Moreno-Torres, J.G. and F. Herrera, 2010. A preliminary study on overlapping and data fracture in imbalanced domains by means of Genetic Programming-based feature extraction. Proceedings of the 10th International Conference on Intelligent Systems Design and Applications, Nov. 29-Dec. 1, IEEE Xplore Press, Cairo, Egypt, pp: 501-506. DOI: 10.1109/ISDA.2010.5687214

Napierała, K., J. Stefanowski and S. Wilk, 2010. Learning from Imbalanced Data in Presence of Noisy and Borderline Examples. In: Rough Sets and Current Trends in Computing, Szczuka M., M. Kryszkiewicz, S. Ramanna, R. Jensen and Q. Hu (Eds.), Springer, Berlin, Heidelberg, pp: 158-167. DOI: 10.1007/978-3-642-13529-3_18

Provost, F. and T. Fawcett, 1997. Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, Aug. 14-17, AAAI Press, Newport Beach, CA, pp: 43-48. Quinlan, J.R., 1991. Improved estimates for the accuracy of small disjuncts. Machine Learn., 6: 93-98. DOI: 10.1007/BF00153762

Rifkin, R., A. Klautau and K. Org, 2004. In defense of one-vs-all classification. J. Machine Learn. Res., 5: 101-141.

Schapire, R.E., 1990. The strength of weak learnability. Machine Learn., 5: 197-227. DOI: 10.1007/BF00116037

Schapire, R.E., 2002. The Boosting Approach to Machine Learning An Overview. In: MSRI Workshop on Nonlinear Estimation and Classification, Schapire, R.E. (Ed.), Berkeley, CA, pp: 149-172.

Segal, R., O. Etzioni, P. Riddle, M. Healy and D. Newman *et al*., 1994. Representation design and brute-force induction in a boeing manufacturing domain. Appears Applied Artificial Intelligence, 8: 125-147.

Seiffert, C., T.M. Khoshgoftaar, J.V. Hulse and A. Napolitano, 2010. RUSBoost: A hybrid approach to alleviating class imbalance. Syst. Humans, 40: 185-197. DOI: 10.1109/TSMCA.2009.2029559

Shen, A., R. Tong and Y. Deng, 2007. Application of classification models on credit card fraud detection. Proceedings of the International Conference on Service Systems and Service Management, Jun. 9-11, IEEE Xplore Press, Chengdu, China, pp: 1-4. DOI: 10.1109/ICSSSM.2007.4280163

Stefanowski, J. and S. Wilk, 2008. Selective Pre-processing of Imbalanced Data for Improving Classification Performance. In: Data Warehousing and Knowledge Discovery, Song I.Y., J. Eder and T.M. Nguyen (Eds.), Springer, Berlin Heidelberg, pp: 283-292. DOI: 10.1007/978-3-540-85836-2_27

Sun, Y., M.S. Kamel, A.K.C. Wong and Y. Wang, 2007. Cost-sensitive boosting for classification of imbalanced data. Pattern Recognit., 40: 3358-3378. DOI: 10.1016/j.patcog.2007.04.009

Tan, P., M. Steinbach and V. Kumar, 2006. Introduction to Data Mining. 1st Edn., Addison-Wesley, Reading, MA, pp: 169.

Tomek, I., 1976. Two Modifications of CNN. IEEE Trans. Syst. Man Cybernet., 6: 769-772. DOI: 10.1109/TSMC.1976.4309452

Wang, S. and X. Yao, 2009. Diversity analysis on imbalanced data sets by using ensemble models. Proceedings of the 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

Weiss, G.M., 2004. Mining with rarity: A unifying framework. ACM SIGKDD Explorat. Newsletter, 6: 7-19. DOI: 10.1145/1007730.1007734

Weiss, G.M. and F. Provost, 2003. Learning when training data are costly: The effect of class distribution on tree induction. J. Artificial Intelligence Res., 19: 315-354.

Williams, D.P., V. Myers and M.S. Silvious, 2009. Mine classification with imbalanced data. IEEE Geosci. Remote Sens. Lett., 6: 528-532. DOI: 10.1109/LGRS.2009.2021964

Wilson, D.L., 1972. Asymptotic properties of nearest neighbor rules using edited data. IEEE Trans. Syst., Man Cybernet., 2: 408-421. DOI: 10.1109/TSMC.1972.4309137

Wu, G. and E.Y. Chang, 2003. Class-boundary alignment for imbalanced dataset learning. Proceedings of the Workshop on Learning from Imbalanced Data Sets, (IDS' 03), Washington, DC, pp: 49-56.

Yang, Q., X. Wu, P. Domingos, C. Elkan and J. Gehrke *et al*., 2006. Challenging problems in data mining research. Int. J. Informat. Technol. Decision Mak., 5: 597-604.

Zadrozny, B. and C. Elkan, 2001. Learning and making decisions when costs and probabilities are both unknown. Proceedings of the 7th International Conference on Knowledge Discovery and Data Mining, Aug. 26-29, ACM, San Francisco, California, pp: 204-213. DOI: 10.1145/502512.502540

Zadrozny, B., J. Langford and N. Abe, 2003. Cost-sensitive learning by cost-proportionate example weighting. Proceedings of the 3rd International Conference on Data Mining, Nov. 22-22, IEEE Xplore Press, Melbourne, FL, USA, pp: 435-442. DOI: 10.1109/ICDM.2003.1250950

Zhang, S., L. Liu, X. Zhu and C. Zhang, 2008. A strategy for attributes selection in cost-sensitive decision trees induction. Proceedings of the 8th International Conference on Computer and Information Technology Workshops, Jul. 8-11, IEEE Xplore Press, Sydney, Australia, pp: 8-13. DOI: 10.1109/CIT.2008.Workshops.51