A Review on Handling Imbalanced Data

Spelmen Vimalraj S
Department of Computer Science
Bharathiar University
Coimbatore, India
spelmen@gmail.com

Porkodi R
Department of Computer Science
Bharathiar University
Coimbatore, India
porkodi_r76@buc.edu.in

***Abstract*--Computational synthesize of the metabolic pathway is take low cost while comparing with the direct trial and error laboratory process. In real world data, more or less all datasets having a skewed distribution of classes. The skewed and the number of instances for certain classes much higher than other classes, this problem is known as the class imbalance problem. Practically this class imbalance problem reduces the classification accuracy because it predicts the minority class instances inaccurately. Class imbalance is an issue encountered by data mining practitioners in a wide variety of fields. The classification of imbalanced data is a new problem that rises in the machine learning framework and it is the major problem raised for the researches and the use of sampling techniques to improve classification performance has received significant attention in related works. In this article the necessity of balancing an imbalanced data is elaborated and the methods proposed by the various authors for to balance the imbalanced data and the evaluation metrics to assess the accuracy and predictive rate of the classification algorithms also have been discussed**

Keywords—Data imbalance, Classification, Oversampling, Undersampling, Hybrid methods.

I. INTRODUCTION

Pathway synthesis is one of the important tasks yet to be improved in the field of data mining. Synthesising a metabolic pathway in imbalanced data is the prevailing problem in the current scenario. Handling the imbalanced data is not an easy task in computational process. The class is said to be a majority class, if the number of instances in a data set for a particular class is higher than the other class. In other terms, the class is said to be minority class, if the number of instances in a database for a particular class is lesser than the other class in the same database. These kinds of data sets are called as imbalanced data sets [1].

The seriousness of imbalance problem in data is explained below: for instance, consider a database having 90 % of instances from the majority class and rest of the instances from the minority class. If all data is predicted as majority class based on the classification rule, the rule acquires 90% accuracy. In this case, the accuracy level is not a proper representation of the classification performance because nothing from the minority class instances are exactly classified. Furthermore, the minority class instances are considered as noisy data and they are eliminated by the classifier or classification algorithm. For improving the accuracy of the classification technique and predicting the accurate result the imbalanced data have to be balanced.

The class imbalance can be inherent property or due to boundaries to obtain data such as cost, confidentiality and large effort [2]. In many real-world applications such as fraud detection (credit card, phone calls, insurance), medical diagnosis, network intrusion detection, fault monitoring, pollution detection, biomedical, bioinformatics, remote sensing (land mine, under water mine)and bioinformatics suffer from these phenomena. Metabolic pathways are referred to a group of biological functions or molecular functions that occurs within the living organism that involves in the growth and development of cells in order to maintain the lifecycle of the particular cell (nelson &cox 2004). This shows that the metabolites are interconnected.

The class imbalance problem in human metabolic pathway stimulates the inflammatory processes in the body and activates the immune system. Leipzig et al, published in the Journal of Allergy and Clinical Immunology, UFZ researchers have been capable to show that this is even applicable to new-born and children under one year of age, and is interrelated with the development of repository disease in childhood. The balance between the metabolites took the responsible for the inflammation process in human body. The class imbalance can be inherent property or due to boundaries to obtain data such as cost, confidentiality and large effort [2].

II. CHARACTERISTICS OF IMBALANCED DATA:

The difficult task remains with the imbalanced dataset is extracting the knowledge from the highly imbalanced datasets. The fundamental problem with the imbalanced dataset is, it has the ability to reduce the performance of the most standard learning algorithms and classification techniques [3]. The imbalanced dataset problems have occurred in various kind of fields and real-world domains such as detection of oil spills in satellite radar images, text

978-1-5386-3702-9/18/$31.00 © 2018 IEEE

classification, spotting telecommunication customer, detection of fraudulent telephone calls, learning word pronunciation, information retrieval and filtering task [4]. The main issues carried out while using the data intrinsic characteristics with in the classification problem are the imbalanced dataset characteristics such as the lack of density in the training dataset, the presence of small disjuncts, the overlapping between classes, the identification of noisy data, and the significance of the border-line instances and the data shift between the training and the test distributions. [5]

A. Small Disjuncts- The big problem with the classification is that the size of samples to be classified. This problem may lead to the lack of information and the lack of information increase the existence of the high dimensional data that is nothing but the large number of features.

B. Over lapping- When dataset having the similar quantity of training data from each class, the overlapping between the classes will occur. This may lead to inference with the probabilities in the overlapping area which makes the classification between two classes as a difficult one.

C. Data shift: - When the training data set follows the various kind of data distribution is the problem named as Data shift. The data shift is the main issue generated due to the wrong selection of bias that normally occurs with the classification problem. Most of the classification techniques that are capable to tackle the datashift[6].

III. METHODS TO HANDLE THE IMBALANCE DATA:

The class imbalance is the problem that occurs frequently in respect to the field of data mining and the data science. In general, there are many methods have been proposed to balance the imbalanced data sets. The most commonly used technique to handle the class imbalance problem are elaborated in this section. Fig.1 Illustrates the methods available to tackle the class imbalance problem and that are categorized into three types. These are namely Data level methods, Algorithmic level methods and Hybrid methods.

A. Data level methods:

To balance the data set the pre-processing step has been employed in the data level approach. The data level methods are often called as external methods because they are try to balance the data by reducing the majority class samples or removing the minority class sampling known as undersampling and oversampling respectively. The data level methods are also classified into three types as oversampling, undersampling and the feature selection respectively.

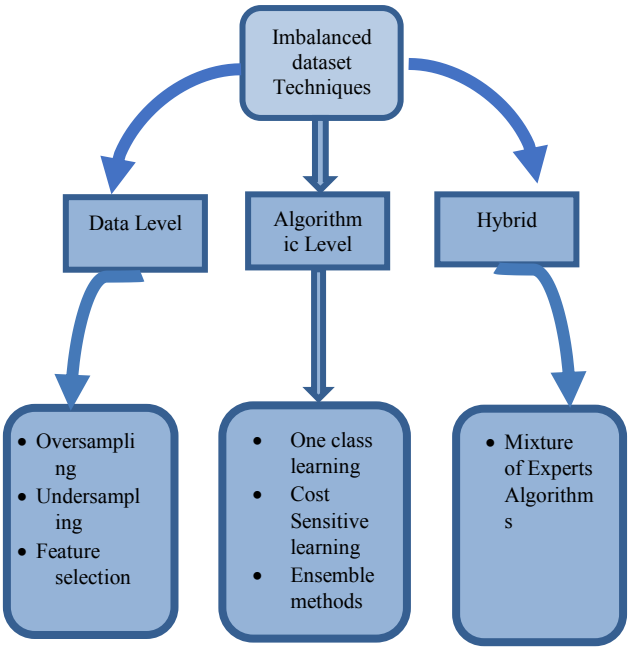

Fig.1 Methods to handle the imbalanced data

*1) Oversampling Methods:*The oversampling and the undersampling are the data level methods that are used to sample and to provide the equitable data distribution among the unbalanced classes. In oversampling the minority class instances are escalated, in order to equate the classes. The most common problem occurred with the oversampling method is nothing but they will not add any new instances or information to the dataset and that may lead to over fitting of classifiers. Whereas in the undersampling method, the method deletes the majority class instances in order to balance the dataset. The most occurring problem with the undersampling is it doesn't consider the information available in the deleted instances. To overcome this problem some of the heuristic approaches have been proposed by the various scientist. [7].

Most importantly the oversampling technique increasing the amounts of minority class but the effect is to find the similar but more specific area in the feature space as the decision area for the minority class. Instead of replacing the minority sample Chawla et al [8] proposed a new technique called Synthetic Minority Oversampling Technique (SMOTE) which creates the "synthetic" example rather than replacing the minority class instances. The Smote is the most commonly used technique and performs better than the random and simple oversampling method. The SMOTE is used for sentence in boundary speech and for the identification of the binding specificity of the regulatory protein and it is used for detecting network intrusions, for detecting the breast cancer, also used in the bioinformatics for miRNA gene prediction, for histopathology, and of photoreceptor-

enriched genes based on expression data [9]. While using the smote based oversampling methods may cause error on the distribution of samples and affect the accuracy of the classifier. The smote cannot really reflect the distribution of original instances in new synthetic instances. This will produce the misclassification of instances by increasing the probability. [10].

The smote algorithm creates the synthetic examples rather than replacing the minority class instances. Haibo et al [11] proposed a novel method namely ADASYN which creates the instances among the two classes in the interior of minority class, instead of creating the synthetic examples. The synthetic data generated for the minority class examples are harder to learn. So the ADASYN method uses the weighted distribution for the minority sample classes. ADASYN enhances the imbalance class learning via two steps: first one is by reducing the bias introduced by the class imbalance. The second one is shifting the classification decision boundary toward the complicated examples. These two are accomplished by dynamic adjustments of weights and adaptive learning procedure [12].

The KNN algorithm classifies the objects based on the closet training examples in the future space. In other terms it is also called as lazy learning or the type of instance-based learning because its function's computations are approximated until the classification. The merits of the KNN algorithm is it is more robust to the noisy data and effective method if the training dataset is too large. The demerits of the KNN algorithm is the value of parameter to be determined that is the number of nearest neighbour could be determined and the distance-based learning is not clear and the computation cost is quiet. [13]

Vladimir N. Vapnik et al proposed a technique based on the Structural Risk Minimization in order to deal with the less-sample classification problems named as SVM. The SVM algorithm can give the better solutions to over learning, dimension disaster and local minimum and nonlinearity. Thus, it has the better generalization capacity [14]. Even though the svm provides the better classification accuracy its performance is lack with the imbalance dataset. To overcome the problem Kai-Biao et al [15] proposed the new technique which is the combined the merits of the Fuzzy C-Means clustering and the SVM.

Decision tree classifiers are the best traditional techniques. Even though it is powerful, it lacks in generalization accuracy for the unseen data. Tin kam ho et al [16] proposed a new technique named as random forest which is an ensemble method which generates the decision trees of classifiers and validate the results of the classifiers. The method combines Breiman's "bagging" idea and the random selection of features. To determine the split the CART is created on a bootstrap sample search across the randomly selected input variables. The major advantage of the random forest is as follows: It overcomes the over fitting problem, It evicts the need for pruning the trees, The importance of the variable and the accuracy is generated, For the outlier data is less sensitive in training data [17]

Fernández-Navarro et al [18] proposed a dynamic over-sampling method for enlightening the imbalance dataset classification by two methods. This technique is incorporated into a memetic algorithm (MA) that reduce the radial basis functions neural networks (RBFNNs). In two stages the training data is resampled in order to handle the class imbalance problem. The first phase consists of balancing the minority class using an over-sampling technique. The memetic algorithm over-sample the data in different phases and provides the new patterns of the minimum sensitivity class. The technique invented by the author is weighed on 13 benchmarking datasets. Also, the proposed technique is compared with other neural network techniques that is intently designed for handle class imbalance problem.

Mazurowskia et al. [19] presented two different neural network methods that is backpropogation and the particle swarm optimization are used are used to examine the use of over sampling and under sampling. According to them, the PSO algorithm was relatively sensitive to handle the class imbalance problem for less number of sample and large number of feature.

José et al. [20] proposed an oversampling technique for dealing with the multi-class imbalance problem and analysis of the class characteristics. The method finds the subsets of significant examples in each class and to deal them with oversampling for each of them independently. This methodology is detecting the four different types of examples in multiclass datasets: safe, borderline, rare, and outliers.

Chao chen et al. [21] proposed two methods based on the random forest technique in order to learn the imbalanced data. The first method, Weighted Random Forest put additional weights on the minority class, thus disciplining more profoundly on misclassifying the minority class. The second method, Balanced Random Forest associates the down sampling majority class method and the ensemble learning idea, artificially fluctuating the class distribution so that classes are signified equally in each tree.

The class imbalance, noise handling and the labelling error has not received significant attention in the field of data mining. Jason et al [22] proposed a new method to overcome this problem, the author took 7 imbalanced data set and tested the datasets with various sampling methods

and algorithms. While comparing with the complex leaners random forest and support vector machine the Naïve bais and nearest neighbour are more robust. From the experimental result the author showed that by reducing the majority class the random undersampling method performs well. With comparison to the complex oversampling techniques like Borderline SMOTE and SMOTE the Wilson Editing method performs well because it targets the mislabelled examples.

Ligang et al [23] analysed the effects of seven sampling methods (both oversampling and undersampling) and five quantitative models with the performance of bankruptcy prediction models. The model is worked on the highly imbalanced dataset. Each model is tested with two different datasets that is highly imbalanced. While comparing with the sampling techniques the support vector machine (SVM) performs well.

Huaxiang Zhang et al [24] proposed a new technique namely Random Walk Over-Sampling (RWO-Sampling) which is used to balance the various class instances by creating the synthetic instances via randomly walking from the data. The proposed method is compared with the alternate algorithms and the KNN algorithm takes much time to calculate the mean and standard deviation and the RWO consumes the less time than the SMOTE to create the synthetic instances.

The existing classification algorithms works better with the majority class and have very poor performance with the minority class. To overcome the problem Antonio Maratea et al [25] proposed a novel technique based on the SVM that is a suitable kernel transformation for the data. By transforming the data, the class boundary is asymmetrically enlarged. And then all the transformed pairs are evaluated with the evaluation metrics like F-measure and AGF. Based on the results those proved that the developed technique has outperform than the C4.5, RIPPER, L2 loss SVM.

*2) Undersampling Methods:*Based on the KNN algorithm Marcelo Beckmann et al [26] proposed new undersampling method. On the basis count of neighbors of each class the instanced were removed in this method to balance the data. The proposed algorithm tested on 33 datasets and compared with the 6 methods namely ENN, SMOTE, NCL and Random Undersampling method. The results compared with the other methods proved that the KNN undersampling method achieved a good accuracy and it act as removing the samples are named as “needle in a haystack” effect. It also act as a cleaning the decision surface, tumbling the class overlapping and also evict the noisy data. According to the results it proved that the KNN undersampling method is best machine learning approach to balance the imbalanced data.

The tomek links is the under-sampling method used for identifying the border line and noisy data [27]. The tomek links are also used for data cleaning that have been used to eliminate the overlapping generated by the sampling methods. In other terms the tomek links are defined as a combination of minimum distanced nearest neighbors of opposed classes [28]. In under sampling method the majority class examples are removed where is in data cleaning method both class examples have been removed.

Yu et al. [29] proposed a heuristic undersampling methods based on the idea of ant colony optimization to address the class imbalance problem. The algorithm begins with the feature selection method to evict the noisy genes in data. On the basis of selection frequency, the significant and informative majority class samples are projected. The proposed method provides the majority balance set which is optimal. The main de-merit of the proposed method is, it took more time while comparing with the simple sampling approaches.

Reducing the majority class samples may cause the loss of data. In order to overcome these problem Yen and Lee [30] proposed a cluster based undersampling technique. Which divides the majority class with k number of clusters and then select the majority class samples with the suitable minority class sample and then it makes the K combined datasets. Finally, all the datasets classified and that gives the best accuracy on the imbalanced datasets.

Orriols et al. [31] stated that the sampling method is the effective technique to handle the imbalanced data and the sampling method uses the supervised learning. The proposed algorithm preserves D (a Variable), that decide which feature could be selected next. The D value is calculated based on the values that already defined. The D value can be updated with the support and ideas from plunging the weights of the features used in classification. By evaluating on the different examples, the classifier’s performance is calculated and then finally the extracted subsets are selected and it will provide the better classification performance. The proposed technique evaluated on different unbalanced binary benchmarking datasets. The dataset features could not be included in this technique. Performance of the technique is evaluated when decreasing the features in dataset increments.

*3) Feature Selection:*Feature selection is one of the challenges prevailed in the imbalanced data. In order to tackle the problem LiuzhiYin et al [32] proposed two new feature selection techniques. The first methods is to decompose the bulk classes into pseudo-subclasses and asses the features with the decomposed data. It decreases the bias rate and the class distribution. The next method is the hellinger distance-based feature selection which does not involve the computation cost with the class information. The proposed approach is compared with

other feature selection method with the help of real world data. Based on the results of evaluation metrics like F-measure and the AUC the proposed method proved that the performance is quite high.

Ilnaz Jamali et al [33] analysed the eight feature selection methods namely Correlation coefficient, Chi-square, Odds Ratio, Signal-to-noise Correlation Coefficient, Information Gain, RELIEF, FAST and FAIR: Feature Assessment by sliding Threshold on the real-world data. Based on the results those proved that which method is suitable for the dataset based on the number of features in the dataset. This will much helpful for the researcher to spend less time on selecting the appropriate feature selection model for the imbalanced dataset.

B.Algorithmic level methods:
The algorithmic level methods are often called as internal approach because it employs the design of new classification algorithm or enhancing the existing algorithms to tackle the bias produced by the imbalanced data. The algorithmic level methods are categorized into ensemble-based methods, threshold methods, one class learning, cost sensitive learning and active learning methods.Most of the classifiers like naïve bayes and some nueral networks provide the score that determines the degree to the instance that is belonging to which class. These ranking methods provide various classification algorithms by adjusting the threshold of an instance belonging to which class.

*1) Learning Methods:*The main aim of the cost sensitive learning is to minimize the misclassification cost and the test costs and the other types of cost or a combination of among these. It is the method which stimulates the models from the unbalanced data class distribution and the impacts by computing and undertaking the imbalance. The cost sensitive learning works or makes decision based on the constructed cost matrix. Cost sensitive learning is a cost sensitive and noise tolerant and it is machine learning system that has two operate stages: the first one is the training stage and the execution stage [34].

Based on some conditions like multi-modality and of the domain space the one class learning solves the discriminative approach that is decision trees and the neural network. Ripper is a rule that iteratively build the rules to cover the uncovered instances. Each rule is developed by combining until no negative instances are covered. The one class learning is specially used on when the dataset is extremely imbalanced and it contains the noisy features. The main drawback is the feature selection is often too expensive to apply. [35].

*2) Ensemble based Methods:*Breiman et al [36] proposed an ensemble-based learning technique that is bootstrap aggregating that is used to construct the ensembles. Classifiers with bootstrapped duplicates the original training dataset. That is new dataset is formed to replace the instances from the original data set with the constructed dataset. Finally, when an unidentified instance is existing in each separate classifier, a majority or weighted vote is to gather the class.

Schapire et al [37] proposed a boosting technique which is known as ARCing, adaptive resampling and combining proved that the weak learner can be triggered in to a stronger learner in the sense of PAC (probably approximately correct) learning framework. The weak learner is nothing but that is slightly better than the random guessing. AdaBoost and the SVM boost are the most commonly used techniques that provide the better accuracy while comparing with other boosting methods to handle the imbalance problem [38].

C. Hybrid Methods:
The hybrid methods are the combination of the data level methods and the algorithmic level with respective combination. The need of hybridization is to overcome the problems with the data level methods and the algorithmic level methods and also to achieve the better classification accuracy.

The major obstacle, typical in medical diagnosis, is the problem of rare positives. Cohena et al. [39] a give solution to this problem via two different approaches that is resampling method using oversampling and under-sampling together with synthetic instances. For tuning the SVM and acquired unbalanced soft margin they presented the class-dependent regularization parameter.

Yong et al [40] proposed the sampling method on the basis of k-means clustering and genetic algorithm which In order to admirably highlight the performance of the minority class presented in the imbalanced data set. The minority class samples were clustered by the k-means clustering and the new samples were obtained from each cluster using genetic algorithm and to carry on the valid confirmation. With the combination of k nearest neighbour algorithm and the support vector machine with the proposed technique have exposed the effectiveness and this could be obtained based on the experimental result.

With the help of various sampling algorithms and the classification techniques Gracia et al. [41] analysed the imbalance ratio effects and the classifiers effect. Random Under Sampling (RUS) and the combination of Wilson's editing (WE) and Modified Selective Subset (MSS) are the undersampling methods and the Synthetic Minority Oversampling Technique and the Gabriel- Graph-based SMOTE are the oversampling methods have been used. The decision acquired from the results shows that the

oversampling technique is performing well because undersampling causes the losses of some important patterns.

Obtaining the better separability from the unbalanced datasets is main challenge in the computerized era. In order to overcome the problem, alberto et al [42] proposed a hierarchical fuzzy rule-based classification (HFRBCS) technique by aggregating the unevenness of the fuzzy partitions on the borderline areas among the classes. This model uses the genetic rule selection procedure in order to get a compact and précised model. HFRBCS which dramatically increase the performance of the classification in overlapping areas in range from minority class to majority class. The SMOTE is used to balance the data before the rule generation phase. From the above things the author proved that the global fuzzy model has been improved.

Class imbalance is the problem to be solved undoubtedly among the datasets in the computer science era. Joonho gong et al. [1] proposed a new technique namely RHSBOOST which is an ensemble-based method in order to deal with the class imbalance problem. The proposed method is a combination of hybrid sampling (Undersampling & ROSE sampling) and AdaBOOST technique. Based on the experimental results they proved that the RHSBOOST generates the reliable & high classification performance.

Most of the prevailing classification techniques be wont not to achieve a better prediction well on minority class instances while the dataset is tremendously unbalanced, because their main motto is to enhance the overall accuracy by not to considering the relative distribution of each class. Yang Liu et al [43] studied the performance of SVMs because it obtained the boundless achievement in vast real applications, within the unbalanced data background. With the experimental analysis, they showed that, from the biased decision boundaries SVMs may suffer and also their prediction performance would be suffered. In order to astound this problem, the author proposed an integrated sampling method which is a combination of two sampling methods (over sampling and under sampling). This technique is used with an ensemble of SVM to enhance the prediction performance.

Galar et al [44] proposed an easiest and most accurate ensemble that is EUSBoost (ensemble construction algorithm), it is based on the RUSBoost. The ensemble method is combination of boosting and the random undersampling algorithms. The main objective of this method is to improve the performance of the classifiers by using the evolutionary undersampling approach. With the results they concluded that the EUSBoost is able to accomplish the state-of-the-art methods based on the ensembles.

To tackle the class imbalance problem and class overlapping on multiclass problem alejo et al [45] proposed a hybrid technique which is a combination of MBP (Modified Back Propagation) and GGE (Gabriel Graph Editing). In this method the new cost function that is based on the MSE in the algorithm that was adapted by the MBP. The proposed technique generates the two effects: (i) during the training process the MBP is used to reimburse the class imbalance and (ii) in the overlapping region the GGE is used to reduce the confusion of the minority class.

To handle the class imbalance problem putthiporn Thanathamathee et al [46] introduced a novel approach based on two concepts. The first concept is, by determining the distance among the class sets with Haudorff distance and recognising all appropriate class boundary data in order to eliminate the imbalanced error dominance effect. The second concept is, by expanding the circulation of training data space to deal with the hidden incoming testing data in advance based on the Bootstraping technique in order to enlightening the testing accuracy. While comparing with other methods the proposed method proved the superiority based on the results.

RAMENTOL et al [47] proposed a technique to address the class imbalance problem named as SMOTE-FRST (sampling based Fuzzy Rough Set Theory). In this method the sampling technique SMOTE is used to resample the data and then the Fuzzy Rough Set Theory is used to handle the class imbalance problem. C4.5 is used as classifier learning algorithm. The proposed method is compared with other method and it works better than the existing SMOTE technique.

Most of the classification algorithms works slowly with the imbalance data because of its large size of features. To address the problem YUN ZHANG et al [48] proposed a new technique that adopt the fast-simple classification and it works with the small number of features. The first classifier is the fast one that identifies most of the features and the second one is for the less quantity of features. This approach effectively increases the speed of the classification algorithm and minimizes the total risk needed for the classification.

MingGao et al [49] proposed a robust and effective algorithm for resolving two-class imbalanced problems, mentioned to as the SMOTE+PSO-RBF, by merging the SMOTE (Synthetic Minority Oversampling Technique) and the PSO (Particle Swarm Optimization) optimised RBF (Radial Basis Function) classifier. The experimental results presented in that study have proved that the proposed SMOTE+PSO-RBF offers an exact modest solution to other prevailing state-of-the-arts approaches

for contesting imbalanced classification problems.

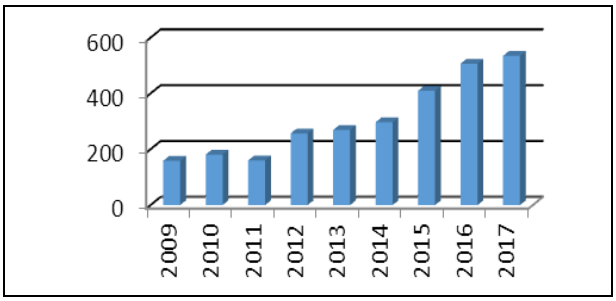

Fig.2 Papers Published from the Year 2009 to 2017

The Fig.2 presents the statistics of research papers published so far in Elsevier publications from the year 2009 to 2017 and its shown steady progress on the imbalanced data classification which paves significant attention to the researchers.

The TABLE I describes the various kinds of techniques and algorithms proposed to handling the imbalanced data as well as the outcomes of the techniques implemented on imbalanced dataset also have been discussed.

TABLE I: Various methods used to solve the imbalanced data

S. No	Authors	Title	Dataset	Algorithms	Method under	Outcome
1.	Joonho Gong et al	RHSBoost: Improving classification performance in imbalance data (2017)	Glass016vs5 – KEEL, Yeast1458vs7 – KEEL, Glass5- KEEL, Yeast2vs8 – KEEL, Yeast4 – KEEL, Yeast1289vs7 – KEEL, Yeast5 – KEEL, Ecoli0137vs26 – KEEL, Yeast6 – KEEL	No treatment (Ovun) AdaBoost SMOTEBoost RUSBoost RHSBoost	Hybrid	1. RHSBoost is proposed to pay attention on the minority class. 2. Boosting technique enhances the accuracy. 3. It achieves the best accuracy on the imbalanced data based on the dominance rank.
2.	José A.Sáez et al	Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets (2016)	Automobile, Ecoli, Flare, Glass, Hayes-roth, Newthyroid, Thyroid, Vehicle, Wine, Winequality, Yeast and Zoo	Over sampling technique: Static SMOTE and adaBoost Classifier: C4.5, SVM, NN	Over sampling	1. Static SMOTE and AdaBoost used to identify the examples safe, borderline, rare, and outliers. 2. Helps to acquire deep knowledge.
3.	Marcelo Beckmann et al	A KNN Unders ampling Approach for Data Balancing (2015)	GlassBWNFP, EcoliCP-IM, Pima, Habermann, EcoliIM, New-Thyroid, EcoliIMU, GlassVWFP, EcoliOM, YeastCyt-Pox, YeastME2, YeastME1, YeastEXC and Abalone19	SMOTE ENN NCL KNN Random Under sampling	Under sampling	1. Majority class instances eliminated based on the amount of neighbors from different classes. 2.KNN-Und method eliminates the majority class instances and also decision surface cleaning, removing the noisy data and reducing the overlapping area at parallel.
4.	HualongYu	ACOSamplin	Colon dataset,	Under	Under	1. ACOSampling controls

	et al	g: An ant colony optimization-based undersampling method for classifying imbalanced DNA microarray data (2013)	CNS (Central Neural System) dataset, Lung cancer dataset and Glioma dataset	sampling based on ant colony optimization	sampling	the lack of information from minority class. 2. The information examples of majority class extracted automatically. 3. The SVM used as a classifier.
5.	R.Alejo et al	A hybrid method to face class overlap and class imbalance on neural networks and multi-class scenarios (2013)	MCayo, MFelt, MSat, MSeg and M92AV3C	MBP GGE	Hybrid	1. SMOTE + GGE outperforms the imbalance well but lack in overlap. 2. MBP + GGE results proved it is better for both.
6.	Mikel Galar et al	EUSBoost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling (2012)	Ecoli (6 datasets), Yeast (8 datasets), Glass04vs5, Shuttle0vs4, Glass4, Page-blocks and Abalone9vs18	Ensemble learners: Bagging and Boosting. Pre-processing algorithms SMOTE, EUS, C4.5	Hybrid	1. Bagging and Boosting together and achieves better classification accuracy. 2. EUSboost – is best suited for imbalanced data sets.
7.	Min gao et al.	A combined SMOTE and PSO based RBF classifier for two-class imbalanced problems (2011)	Pima indian Diabetes dataset, Haberman survival dataset and ADI dataset	SMOTE RBF PSO	Hybrid	1. The RBF classifier has less performance on imbalanced data. 2. To achieve better accuracy SMOTE +PSO+RBF has been combined together.
8.	FranciscoFernández-Navarro et al	A dynamic over-sampling procedure based on sensitivity for multi-class problems (2011)	Hepatitis, BreastC, Haberman, CYTvsPOX, ME2vsOther, Newthyroid, E. coli and Yeast	RBFNN MA SSRBF DSRBF	Oversampling	1. The SSRBF and DSRBF used for preprocessing. 2. MA used for optimization. 3. RBFNN is best rather than all-art-of NN learners.
9.	ZHAI Yun et al [51].	An Effective Over-Sampling Method for Imbalanced Data Sets Classification (2011)	Primadata set, Mammography data set	one-sided selection link and dynamic distribution density-SMOTE	Oversampling	1. OSSLDDD-SMOTE deals with the noisy imbalanced data [52]. 2. The OSSLDDD-method works based on the One-Side Selection Link and Dynamic Distribution of Density. 3. Potential study has been carried on imbalanced data.

10.	Hien M. Nguyen et al	Borderline Over-sampling for Imbalanced Data Classification (2009)	Spect, Glass, Vowel, Yeast, Abalone and Page-blocks	Border-line oversampling method and SVM	Oversampling	1. The borderline instances sampled to extend the minority region. 2. SVM is used as a classifier, 3. Better working with low degree of overlap.
11.	Show-JaneYen et al	Cluster-based under-sampling approaches for imbalanced data distribution (2009)	Census-Income and Overdue Detection	Cluster based undersampling, random selection and NearMiss-2	Undersampling	1. Cluster based undersampling is used with BP to reduce imbalance class distribution. 2. It outperforms than random selection and NearMiss-2.

IV. EVALUATION METRICS:

Evaluation metrics are tied with the machine learning models and techniques. The predictive models are accuracy based one. To can get feedbacks for the predictive models based on the evaluation metrics. After building the classification model it is evaluated with the seven-evaluation metrics namely sensitivity, precision, specificity, geometric mean, f-measure, Mathew correlation coefficient, AUC, ROC [50].

1) Precision- Measures the True positive samples that predicted from all the samples divided by the sum of truly predicted samples and the predicted positive class samples

Prec= TP / (TP + FP).

2) Sensitivity- Sensitivity measures the Truly predicted samples from all the samples divided by the sum of truly predicted class and the false negative instances.

Sens= TP / (TP + FN).

3) Specificity- Specificity measures the negative class sample correctness divided by the sum of negative class samples correctness and the predicted positive samples.

Spec= TN / (TN + FP).

4) Geometric mean- It is the geometric mean of the sensitivity and the specificity.

GM = √ (Sens.Spec).

5) F-Measure- F-measure is the Harmonic mean of the Precision and the Sensitivity.

F- Measure = (2. Prec.Sens)/ (Prec+Sens).

6) Mathew Correlation coefficient- The MCC is used to measure the Superiority of the binary class classification.

MCC = ((TP.TN) – (FP. FN)) / √ (a.b.c.d).

Where a = TP +FP, b= TP +FN, c= TN + FP and d= TN +FN. The MCC values is produced as one by the both high TP and TN and the both low FN and FP values.

7) AUC- By plotting the ROC curve at various threshold the AUC is calculated. Here the X axis is 1-Spec and the Y axis is Sens. If the value produced by the AUC is 0.5 then the model is better working condition than the random guess.

V. CONCLUSION

In this paper the problem with the imbalanced data and the need of balancing the data is discussed in an elaborate manner. And also, the various kinds of methods that have been developed for handling the imbalance problem by various authors have been discussed. From this survey most of the authors have been proved that the SMOTE algorithm performs better than the other state of art algorithms with respect to the class imbalance problem. The imbalance problem exists in many real-world domains like medical diagnosis, fraudulent call detection and telecommunication department. Even though many methods available to handle the imbalance problem in various domains, significant attention yet to be given on the medical diagnosis domain. Thus, the paper suggests much more improvements to be needed on the techniques to optimize the problem of class imbalance.

REFERENCES:

[1]. JoonhoGon et al. RHSBoost: Improving classification performance in imbalance data. Computational Statistics & Data Analysis. Volume 111, July 2017, Pages 1-13.

[2]. S. L. Phung et al, A. Bouzerdoum and G. H. Nguyen. Learning Pattern classification tasks with imbalanced data sets. Pattern Recognition. Pages 193-208.

[3]. Dr.D.Ramyachitra & P.Manikandan. Imbalanced Dataset Classification and Solutions: A Review. International Journal of Computing and Business Research (IJCBR). Volume 5 Issue 4 July 2014.

[4]. Sotiris Kotsiantis et al. Handling imbalanced datasets: A review. GESTS International Transactions on Computer Science and Engineering, Vol.30, 2006.

[5]. VictoriaLópez et al. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Information Sciences. Volume 250, 20 November 2013, Pages 113-141.

[6]. A. Fernández et al. Addressing the Classification with Imbalanced Data: Open Problems and New Challenges on Class Distribution. International Conference on Hybrid Artificial Intelligence Systems (2011). Pages 1-10.

[7]. Zhuoyuan Zheng. Oversampling Method for Imbalanced Classification. Computing and Informatics, Vol 34, No 5

(2015).
[8]. Nitesh V. Chawla et al. SMOTE: Synthetic Minority Over-Sampling Technique. Journal of Artificial Intelligence Research 16 (2002) 321–357.
[9]. Rok Blagus and Lara Lusa. SMOTE for high-dimensional class-imbalanced data. BMC Bioinformatics 2013. 14:106.
[10]. Zhuoyuan Zheng, Yunpeng Cai and Ye Li. Oversampling method for imbalanced classification. Computing and Informatics, Vol. 34, 2015, pages 1017-1037.
[11]. Haibo He et al. ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. Neural Networks, 2008. IJCNN 2008. (IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on: 1-8 June 2008.
[12]. Adnan Amin et al. Comparing Oversampling Techniques to Handle the Class Imbalance Problem: A Customer Churn Prediction Case Study. IEEE Access (Volume: 4) 26 October 2016. Page(s): 7940 – 7957.
[13]. Harshit Dubey and Vikram Pudi. Class Based Weighted K-Nearest Neighbor over Imbalance Dataset. Pacific-Asia Conference on Knowledge Discovery and Data Mining PAKDD 2013. Pages 305-316.
[14]. Guixiong Liu et al. Multi-class Classification of Support Vector Machines Based on Double Binary Tree. Natural Computation, 2008. ICNC '08. Fourth International Conference on: 18-20 Oct. 2008.
[15]. Kai-Biao Lin et al. Imbalance data classification algorithm based on SVM and clustering function. Computer Science & Education (ICCSE), 2014 9th International Conference on: 22-24 Aug. 2014.
[16]. Tin Kam Ho. Random decision forests. Document Analysis and Recognition, 1995. Proceedings of the Third International Conference on: 14-16 Aug. 1995.
[17]. Jehad Ali et al. Random Forests and Decision Trees. IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 3, September 2012.
[18]. FranciscoFernández-Navarro et al. A dynamic over-sampling procedure based on sensitivity for multi-class problems. Pattern Recognition. Volume 44, Issue 8, August 2011, Pages 1821-1833.
[19]. Mazurowskia et al. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. Neural Networks. Volume 21, Issues 2–3, March–April 2008, Pages 427-436.
[20]. José A.Sáez et al. Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets. Pattern Recognition. Volume 57, September 2016, Pages 164-178.
[21]. Chao Chen et al. Using Random Forest to Learn Imbalanced Data. Journal of Artificial Intelligence Research, volume 16, pages 321–357.
[22]. Jason Van Hulse et al. Knowledge discovery from imbalanced and noisy data. Data & Knowledge Engineering. Volume 68, Issue 12, December 2009, Pages 1513-1542.
[23]. YangYong et al. The Research of Imbalanced Data Set of Sample Sampling Method Based on K-Means Cluster and Genetic Algorithm. Energy Procedia. Volume 17, Part A, 2012, Pages 164-170.
[24]. HuaxiangZhang et al. RWO-Sampling: A random walk over-sampling approach to imbalanced data classification. Information Fusion. Volume 20, November 2014, Pages 99-116.
[25]. AntonioMaratea et al. Adjusted F-measure and kernel scaling for imbalanced data learning. Information Sciences. Volume 257, 1 February 2014, Pages 331-341.
[26]. Marcelo Beckmann et al. A KNN Undersampling Approach for Data Balancing. Journal of Intelligent Learning Systems and Applications, 2015, 7, pages: 104-116.
[27]. Nitesh V. Chawla. Data Mining for Imbalanced Datasets: An Overview. Data Mining and Knowledge Discovery Handbook. Pages 853-867.
[28]. Haibo He and Edwardo A. Garcia. Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering (Volume: 21, Issue: 9, Sept. 2009) Page(s): 1263 – 1284.
[29]. HualongYu et al. ACOSampling: An ant colony optimization-based undersampling method for classifying imbalanced DNA microarray data. Neurocomputing. Volume 101, 4 February 2013, Pages 309-318.
[30]. Show-Jane Yen and Yue-Shi Le. Cluster-based under-sampling approaches for imbalanced data distributions. Expert Systems with Applications 36 (2009) 5718–5727.
[31]. Albert Orriols et al. Evolutionary rule-based systems for imbalanced data sets. Soft Computing. February 2009, 13:213.
[32]. LiuzhiYin et al. Feature selection for high-dimensional imbalanced data. Neurocomputing. Volume 105, 1 April 2013, Pages 3-11.
[33]. Ilnaz Jamali. Feature Selection in Imbalance data sets. IJCSI: International Journal of Computer Science Issues, Vol. 9, Issue 3, No 2, May 2012.
[34]. Haibo He and Edwardo A. Garcia. Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering (Volume: 21, Issue: 9, Sept. 2009) Page(s): 1263 – 1284.
[35]. Sotiris Kotsiantis et al. Handling imbalanced datasets: A review. GESTS International Transactions on Computer Science and Engineering, Vol.30, 2006.
[36]. Leo Breiman. Bagging Predictors. Machine Learning. August 1996, Volume 24, Issue 2, pp 123–140.
[37]. Robert E. Schapire and Yoav Freund. Boosting: Foundations and Algorithms. Acm digital library.
[38]. Mikel Galar et al. A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) (Volume: 42, Issue: 4, July 2012) Page(s): 463 – 484.
[39]. GillesCohen et al. Learning from imbalanced data in surveillance of nosocomial infection. Artificial Intelligence in Medicine. Volume 37, Issue 1, May 2006, Pages 7-18.
[40]. YangYong et al. The Research of Imbalanced Data Set of Sample Sampling Method Based on K-Means Cluster and Genetic Algorithm. Energy Procedia. Volume 17, Part A, 2012, Pages 164-170.
[41]. V.García et al. On the effectiveness of preprocessing methods when dealing with different levels of class imbalance. Knowledge-Based Systems. Volume 25, Issue 1, February 2012, Pages 13-21.
[42]. Alberto Fernández et al. Hierarchical fuzzy rule based classification systems with genetic rule selection for imbalanced data-sets. International Journal of Approximate Reasoning. Volume 50, Issue 3, March 2009,

Pages 561-577.

[43]. YangLiu et al. Combining integrated sampling with SVM ensembles for learning from imbalanced datasets. Information Processing & Management. Volume 47, Issue 4, July 2011, Pages 617-631.

[44]. MikelGalar et al. EUSBoost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling. Pattern Recognition. Volume 46, Issue 12, December 2013, Pages 3460-3471.

[45]. R.Alejo et al. A hybrid method to face class overlap and class imbalance on neural networks and multi-class scenarios. Pattern Recognition Letters. Volume 34, Issue 4, 1 March 2013, Pages 380-388.

[46]. PutthipornThanathamathee et al. Handling imbalanced data sets with synthetic boundary data generation using bootstrap re-sampling and AdaBoost techniques. Pattern Recognition Letters. Volume 34, Issue 12, 1 September 2013, Pages 1339-1347.

[47]. E. Ramentol et al. SMOTE-FRST: A new resampling method using fuzzy rough set theory. Proceedings of the 10th International FLINS Conference. Istanbul, Turkey, 26–29 August 2012

[48]. Yun Zhang et al. Parallel classifiers ensemble with hierarchical machine learning for imbalanced classes. Proceedings of the Seventh International Conference on Machine Learning and Cybernetics, Kunming, 12-15 July 2008.

[49]. Ming Gao & Xia Hong. A combined SMOTE and PSO based RBF classifier for two-class imbalanced problems. Neurocomputing. Volume 74, Issue 17, October 2011, Pages 3456-3466.

[50]. SasipornTongman et al. metabolic pathway synthesis based on predicting compound transformable pairs by using neural classifiers with imbalanced data handling. Expert Systems with Applications. Volume 88, 1 December 2017, Pages 45-57.

[51]. ZHAI Yun et al. An Effective Over-Sampling Method forImbalanced Data Sets Classification. Chinese Journal of Electronics. Vol.20, No.3, July 2011.

[52]. Hien M. Nguyen et al. Borderline Over-sampling for Imbalanced DataClassification. Fifth International Workshop on Computational Intelligence & Applications. IEEE SMC Hiroshima Chapter, Hiroshima University, Japan, November 10, 11 & 12, 2009.