International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-5, January 2020

Learning from Imbalanced Data in Classification

Seema S. Yadav, Girish P. Bhole

Abstract: ***Imbalanced data learning is a research area and day by day development is going on. Due to these researchers are motivated to pay attention to find efficient and adaptive methods for real-world problems. Machine learning, as well as data mining, is a field where researchers are finding different methods to solve problems related to imbalanced datasets and also the challenges faced in day to day life. The uneven class distribution in the dataset is the reason behind the degradation of performance in approaches used by data mining as well as machine learning. Continuous advancements of machine learning as well as mining data combining it with big data, a deep insight is required to understand the nature of learning imbalanced data. New challenges are emerging due to this development. Among the two approaches algorithm level and data level, the most popular approach compared to this is the hybrid approach. It is found that there is a bias for the majority class which affects the decision making task and overall accuracy of classification. The ensemble method is an efficient technique to deal with the uneven distribution of data. The aim of the paper is to presents the overview of class imbalance problems, solutions to handle it, open issues and challenges in learning imbalanced datasets. Based on the experiment conducted on one dataset it is found that ensemble technique along with other data-level methods gives good results. This hybrid method can be applied in many real-life applications like software defect prediction, behavior analysis, intrusion detection, medical diagnosis, etc. The paper further provides research directions in learning from the imbalanced dataset***.

Keywords: Class imbalance, classifier, majority and minority class, biasing, sampling, feature selection, imbalanced learning, machine learning, Preprocessing.

I. INTRODUCTION

Data mining in addition to the machine learning field has problems related to uneven classes. Imbalance class problem is that in which one class has more samples and other as less number of samples. The one which has the number of samples is referred to as the majority class. In contrary to this, less number of samples are associated with minority class. The more important and interesting class considered in the number of applications is minority class. The reason for the increase in the imbalance problem is an unusual distribution of class. Occurrences of a small number of examples in a class are recognized to be rare events. An interesting class is one with a small number of instances. Detecting rare events is considered a prediction task This type of event harms society. Such types of events need decision-making responses from humans. Rare events are not often found in daily life, prediction tasks is affected due to imbalanced data. The model performance is degraded during extraction from the skew domains, particularly minority class prediction. In data mining, the Class Imbalance Problem is very important. A clear understanding of this problem, its effects on classifier performance is required. Classification models are built using data mining approaches. This model guides the decision-making activity done by managers, however imbalanced data classification remains a challenge for the traditional classification models. Minor class performance is not so good. The cost associated with major class classification is lesser as compared to misclassification of minor class instance 1) detection of a non-cancer patient as well as cancer patients, in this number of patients suffering from cancer are less. 2) fraud case and un-fraud case in this fraud case are less compared to the un-fraud case. Many real-life applications are affected by class imbalance problem such as in case of diagnosis of the rare disease where patients having rare disease is less than the others, bioinformatics, credit card transaction frauds [38], fraudulent phone calls, fraud in insurance claim, detection of network intrusion and pollution, monitoring faults, remote sensing in land mine and underwater mine, biomedical, Video mining. [13], Cancer malignancy grading [18], etc.

Classifying imbalanced data is challenging to the traditional classification models due to the following reasons:

1) For the balanced training set, the suitable classifier models in data mining are SVM, Logistic regression and decision tree. These models give the suboptimum classification result. Examples that belong to the majority class are covered properly but minority class examples are discarded.

2) The metrics such as accuracy, prediction used globally to measure performance while learning is more inclined to majority class, despite the fact that minority class examples remain unfamiliar though high precision is given by the classifier model used for prediction.

3) The learning model considers the minority class instances as noise and these instances usually overlap with other areas.

4) In imbalanced learning small disjunct, sample size small and high feature dimensions are the challenges, due to which the learning models fail to detect rare patterns.

The paper emphasizes the importance of concepts as well as methods related to learning from class imbalance data. In the following way the paper is organized:

Manuscript published on January 30, 2020.
* Correspondence Author
Seema S. Yadav*, Research Scholar, Computer Engineering, Veermata Jijabai Technological Institute, Matunga, Mumbai, India. Email:ssyadav_p17@ce.vjti.ac.in
Girish P. Bhole, Professor, Computer Engineering, Veermata Jijabai Technological Institute, Matunga, Mumbai, India. Email: gpbhole@ce.vjti.ac.in

General idea about learning problems in class imbalance and the methods designed to address binary imbalanced class data is provided in Section 2. A brief overview of feature selection in the imbalance class problem is provided in Section 3. The description of various performance evaluation metrics used in imbalanced data classes is done in Section 4.
Section 5, provides details about various methods used in problems related to class imbalance. Section 6, provides a comparative analysis of different methods used in class imbalance. Section 7, provides a brief overview of challenges and opportunities in class imbalance. Section 8, provides the application domain of imbalanced data in classification. Section 9, provides the research direction in this area.

II. LITERATURE SURVEY

Learning from Class Imbalance

Section 2. introduces elementary concepts associated with imbalanced learning, besides that the effect of skewed class distribution on recent and emerging applications is mentioned.

2.1 Approaches to tackling imbalanced data

The approaches mainly used to learn from imbalanced data are [1][5][6].

- Data Level methods
- Algorithm Level methods
- Hybrid method

Data Level: Group of examples is modified to balance the distribution in data level methods. This can be done by the addition or deletion of difficult samples. The training set is modified for the standard learning algorithm. The approach that generates new examples and adds it to minority groups is referred to as oversampling ***while*** the approach that removes the examples from the majority groups is called under-sampling. There are standard approaches also that use a random approach to select target samples for preprocessing. Due to this introduction of meaningless samples or removal of important samples/examples takes place. Therefore, the proposed methods were advanced to preserve group structure or else to generate new data as per the underlying distributions. The algorithm provides a way out for cleaning the objects that overlap as well as removes inappropriate samples that affects the performance of learners.

Algorithm-level: In this method, the existing algorithm is directly modified since more preference is given to instances belonging to the majority class. A good understanding is required for the modified learning algorithm. The cost-sensitive approach is the most popular one. The data distribution and the cost of misclassification errors are uneven in many imbalance problems. The cost learning techniques take into account the cost of misclassification; it allocates high cost if positive i.e. minority class is misclassified. In this method given learner is changed to include varying fine for each group of examples into consideration. Thus a higher cost is assigned to a set of objects that is less in number. This approach aims to minimize the cost associated with the misclassification of minority class object. A higher cost is associated with minority class objects. In many real-life applications, there is difficulty in setting real values in the cost matrix. Applying the one class-learning an algorithm-level solution target group can be concentrated and data description can be created. In this way, we focus only on a single set of objects and can eliminate bias towards any group. For more complex problems dedicated approaches are required for using one-class learners.
Hybrid methods: The advantages of two methods i.e algorithm level and data level are combined in a hybrid method. It concentrates on combining these approaches for the extraction of their strong points and decreases their weaknesses. The result of merging classifier ensembles with data-level solutions is, we get an efficient and robust and learner which is more popular. In some of the applications hybridization of cost-sensitive learning plus sampling method is proposed.

2.2 Imbalanced problems in real-life applications

In various applications dealt in real life, we face the problem of uneven distribution of data. There is a lot of progress while learning an imbalanced dataset. The motivation is since there are several circumstances in day to day application where the dataset is uneven. The most significant class is a minor one in such cases, therefore techniques are required for the improvement of its recognition rates. Some of the important issues are prevention of malicious attacks, detection of serious diseases, management of exceptional activities in Facebook, Twitter, etc. the social networks to control uncommon situations when we monitor the systems.
Real-life Applications with imbalanced data
1)Video mining. [13]
2)Text mining. [14]
3)Sentimental analysis. [15]
4)Industrial system monitoring. [16]
5)Behavior analysis. [17]
6)Cancer malignancy grading etc. [18]
7)Software defect prediction. [19]

2.3 Research trends in class imbalance domain

Based on study four main categories have been recognized which is further divided into nine categories. The techniques recommended in the past for dealing imbalance class problem had used 18 different approaches. To tackle the problem more than one approach is used in some techniques. [1]

2.3.1 Categories of class imbalance domain

Various categories of class imbalance domains are represented in Fig 1.
The proposed technique uses the following 18 approaches:

Table I. Approaches used for various domains

1.Nearest Neighbor	10.Rough sets
2.Random principle	11.Kernel function
3.Genetics	12.Geometric mean
4.Clustering	13.Bagging
5.Neural networks	14.Boosting
6.Principal component analysis (PCA)	15.Greedy divide and conquer
7.Support vector machine(SVM)	16.Rotation network

8.Noise filter	17.Immune networks
9.Fuzzy logic	18.Fuzzy rule base

2.4 Binary Imbalanced Classification

One of the progressive branch considered for learning imbalance class problem is a binary classification. Binary imbalanced classification is being introduced since numerous applications in real-life, such as healthy as well as sick patients in medicine, legal and illegal activities in computer safety, or else background object and the target object in computer vision. The structure describes the relationship between classes, i.e. majority and minority. For balancing the classifier distribution to minority class there are several direct methods.

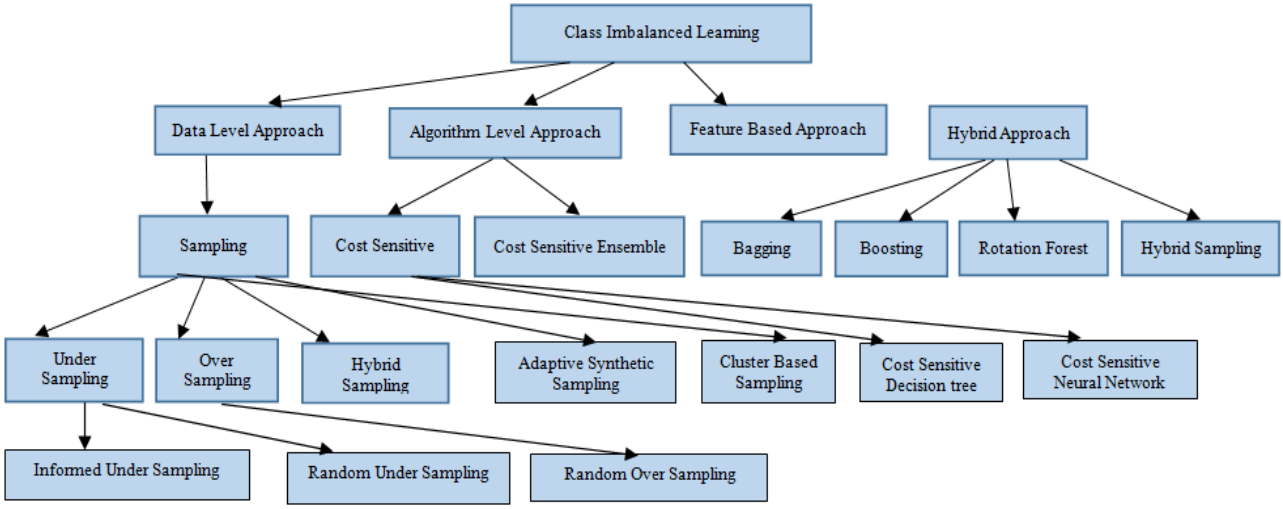

Fig.1 Categories of class imbalance domain

2.4.1 Analysis of classes

Binary imbalanced classification faced learning problems for which imbalance ratio is not the only reason. Even though imbalance is high, the representation of two classes are proper, as well as though they originate from the distribution that does not overlaps, both the classes are well represented and using standard classifier we may obtain worthy classification rates. The reason behind the poor performance is the presence of tough samples inside the class considered as minor. It is recommended from the recent research that minority class samples and their neighbors must be analyzed[4]. The assignment of it to anyone group amongst four. The groups are an outlier, safe, infrequent and marginal. The guidelines that would be well-thought-out to have full understanding besides utilizing the structure of minority class is as follows:

– The proposed classifiers should directly or indirectly include the knowledge related to objects into their training procedure. The classifier should not be biased towards the majority class.

– A similar notion might be used for the preprocessing of data. Owing to the firm construction of minor class, we can concentrate on the selection of important or samples that are difficult. According to samples, the oversampling level can be varied and we can administer the under-sampling method to consider the minority representatives.

– It is required to study the outlier's role or else inappropriate minority class samples. Additionally, it is difficult to conclude that the given object is a real noise or outlier. Removal of noisy or outlier examples results in novel objects classification wrong.

- Labeling methods used to classify the minority objects are kernel methods or KNearest having k value equal to 5. This indicates uniformity in data sharing. The proposed method should be adaptable enough to adjust the size of the studied region as per the chunk sizes.

2.4.2 Extremely imbalanced dataset

This is one of the essential problems associated with class imbalanced data. The imbalance ratio of 1:4 to 1:100 is considered in many scenarios of class imbalance. In the detection of fraud as well as the cheminformatics proportion of imbalance is 1:1000 to 1:5000 [4]. If datasets are exceptionally imbalanced then the classification of it is hardly studied.

Following guidelines must be well-thought-out in understanding the concept of

exceptionally imbalanced classification of datasets:

– The structure of minority class is not clear and its representation is poor if the imbalance is very high. The classification performance is worsened in an application where the preprocessing method (like SMOTE) depends on the relationship between minor objects. The use of random methods is not suitable. Hence, it is required to use methods that authorize minority class to predict class or else reconstruct the structure of it.

– The original problem can be divided into subproblems characterized by less imbalance ratio is one more way to research. The canonical method needs two-way enhancement in procedures (i) splitting up of problem should be meaningful and (ii) Reformation of the originally uneven task.

– A well-organized method to extract features in such type of problems is the third challenge. Transaction done on the internet and data about protein are categorized by high-dimensional and less feature space.

Such problems can appear in social networks and computer vision. Therefore, advanced methods are required for the representation of such data that allows efficient processing of the minority class.

2.4.3 Classifier's output adjustment

To handle the class imbalance, it is required to modify the learning algorithm or training set. It would be misleading if the data distribution is modified deprived of considering an imbalance on the output of classification. According to recent studies, it is noticed that by setting a threshold value on the output of a classifier, good results can be acquired without the use of data resampling.

For additional development of output settlement of classifiers used for imbalanced data following guidelines can be well-thought-out:

– For each class separately, adjustment of output is done. The adjustment value of the parameter is the same for individual objects belonging to the classifier. However, it can be seen that there is no uniformity in minority class and the difficulty level is not the same for the objects within. Hence, new approaches should be developed that considers features of samples classified in addition to this does adjustment of the output of classifier for novel object individually.

– The disadvantage of techniques created on the basis of adjustments of output is that classifiers are overdrive towards minority class, therefore the errors are increased towards the major class. When the new objects are to be classified it is expected that imbalance among classes remains and the assumption is, the compensation of output is not required (as the objects initiate from majority class distribution). In this field, methods that choose indefinite samples as well as alter outputs of objects is essential. It seems that a useful framework is a dynamic selection of classifier between recognized and familiar classifiers.

– An independent method well-thought-out is output adjustment. When the data-level and algorithm-level approaches were previously used, from a general point of view modification of outputs is considered very fruitful. In this way, class balancing on different levels may be achieved and more advanced classifiers can be created. By examining the output compensation, we get a different vision to supervise oversampling and undersampling to determine the stable performance of classes.

2.4.4 Imbalanced learning for ensemble

The class imbalance problem can be handled by the ensemble method. For difficult data, the hybrid method proves to be robust. The robust and highly competitive method proved is the hybridization of random forest, bagging and boosting with sampling and cost-sensitive methods. The heuristic-based method is used in these approaches. The performance of the classifier is difficult to understand in case of uneven classes.

Guidelines required for developing the branch of ensemble learning in case of imbalanced data are as follows:

–Diversity in imbalanced learning requires good understanding. Ensembles in which undersampling is used generally keep the minority class together or it introduces small differences. Thus, if the majority and minority class diversity are very high it will be a drawback.

– What should be the size of ensembles? There is no clear idea regarding this. The size of the ensembles has randomly selected which results in the grouping of the classifiers. It would be more beneficial if the relationship among the number of classifiers required and features of the imbalanced dataset is analyzed to proficiently handle it though we preserve their specific quality. It is necessary to develop ensemble pruning methods specially dedicated to imbalanced problems.

–The majority voting combination method is used in most of the imbalanced ensemble techniques. This is an effective and easy solution in most of the setup. But is it necessary that it is appropriate for imbalanced learning also particularly in case of random methods? We assume that the training of base classifiers is done using sampling methods and their individual qualities are different as it is following the samples having various problems.

III. SELECTING THE FEATURES FROM CLASS IMBALANCED PROBLEMS

The selection of features from an imbalanced dataset remains a serious issue in DM (data mining) and ML (machine learning) [6]. The aim of this method is the selection of essential features, for the improvement of accuracy and classifier performance. An increase in data dimensions and inappropriate features decreases the classifier performance and rises the rate of misclassification particularly in uneven sets. The metrics used for the selection of features are characterized as one-sided if they select features that are only positive else two-sided if they combine positive features with negative features. Likewise, subject to datatype metrics used for the selection of features are characterized as binary or continuous. For example, nominal data and binary data can be handled by odds ratio(OR), Chi-square and Information Gain (IG). But continuous data can be handled by S2N, Pearson correlation coefficient and FAST.

The method used for the selection of features derived from the support vector machine (SVM) is weight vector sensitivity, presented by Nguwi and Cho [10]. Ranking criteria are used by them and they eliminated those features that have less contribution used for increasing the simplification capability of the classifier. Emergent Self-Organizing Map (ESOM) was applied for the clustering of features that are ranked, to make available clusters necessary for unconfirmed classification.

The method used for ranking the features relied on estimating the features of small-sized samples by probability density and imbalanced high dimensional data sets is presented by Alibeigi et al. Selection of features based on density (Density-Based Feature Selection (DBFS)) is used considering the benefit of distribution of features taking place in classes with their associations.

IV. METRICS FOR EVALUATING CLASS IMBALANCED DATASET

Metrics used for evaluation is supposed to be a serious problem. Evaluation Metrics is a pointer for the performance measure of algorithms used in machine learning [6]. Accuracy, as well as error rate, are used as the standard metrics used for evaluation. Since the accuracy is biased towards the class having more samples irrespective of the class having fewer samples, it is not appropriate to use it for handling class imbalance problem and the performance degrades. From the confusion matrix, we can derive metrics that are common for the two-class problem presented in Table

2. Evaluation metrics mostly associated with class imbalance are sensitivity (1), recall (5), precision (4), specificity (2), geometric mean (g-mean) (8), F-measure (6), (7) [11]. The classification performance of each class can be monitored by using sensitivity and specificity.

Table II: Classification confusion matrix

	positive	negative
positive	truepositive	falsenegative
negative	falsepositive	truenegative

ESTIMATED CLASS

REAL CLASS

Assessment of class performance i.e. minority (required to be more) as well as the majority is done by using G-mean and F-measure. One class performance is assessed by the use of precision.

tp – truepositive ; *tn* – truenegative ; *fp* - falsepositive ;
fn- falsenegative ; *Se -sensitivity; Sp -specificity;*
P – precision; A – accuracy;R – recall;

$$Se = \frac{tp}{fp + tp} \quad (1)$$

$$Sp = \frac{tn}{fp + tn} \quad (2)$$

$$A = \frac{tn + tp}{tp + tn + fn + fp} \quad (3)$$

$$P = \frac{tp}{tp + fp} \quad (4)$$

$$R = \frac{tn}{tn + fp} \quad (5)$$

$$F\text{-}measure = \frac{(1+\beta^2)*(R*P)}{\beta^2 * R + P} \quad (6)$$

Where β a constant that is non-negative and is set as 1:

$$F\text{-}measure = \frac{2*P*S}{S+P} \quad (7)$$

$$G\text{-}Mean = \sqrt{tprate + tnrate} \quad (8)$$

In addition to this Area under curve and Receiver operating curve are popular measures in imbalanced class. Representation of the ROC Curve is done by plotting a graph, *fp Vs tp* on *the x-axis* as well as on *y-axis resp*. Employing a receiver operating curve the classifier performance is briefed and visualized.

$$AUC = \frac{tprate + tnrate}{2} \quad (9)$$

Other metrics are cost-sensitive metrics that use the cost curve and cost matrix as the misclassifying error. Cost matrix is used to find the cost associated with the samples to be classified C (i, j) defines cost associated with samples to classify them as class *j* instead *of* class *i*. The total cost can be computed as:

$$Total\ cost = (fn.\ C_{fn}) + (fp.\ C_{fp}) \quad (10)$$

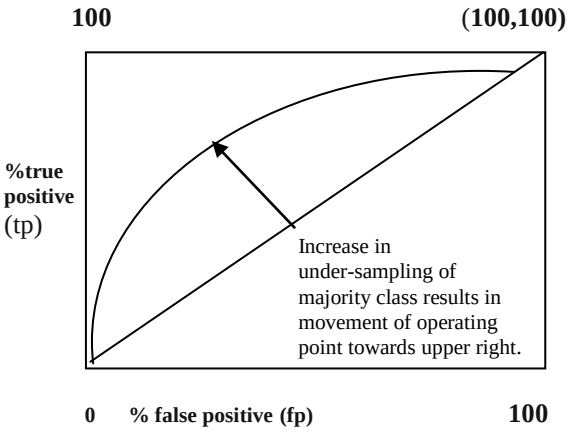

Fig 2. Sweeping out ROC curve through under-sampling.

V. DEALING WITH CLASS IMBALANCE PROBLEM

Imbalanced Data Learning Approaches [2][3]

1. Basic Sampling Methods

- Under sampling method
  - Random under sampling
    - IRUS
  - Balanced Cascade
- Oversampling method
  - Random over sampling
  - SMOTE
    - SDC
    - Borderline Smote
    - Safeline Smote
    - MSMote
      - CSmote
    - MWMote
  - RAMOBoost
  - A-SUWO

2. Novel Approaches to Sampling

- BootOS
- OSS
- Tomek Link
- NCL
- SMOTE
- SMOTE Borderline

3. Methods based on Cost-Sensitive Learning

- Meta-learning sampling
- Meta-learning thresholding
- Meta-learning
- Direct

4. Approaches based on Ensemble Learning

1. Data processing and Ensemble Learning

- Bagging
  - Lazy
  - Under Bagging
    - Quasi
    - Partitioning
    - Asymmetric
    - Ensemble Variation
    - Roughly balanced
  - Over Bagging
    - SMOTE
  - UnderOverBagging
  - Random feature selection
  - IIVotes
- Boosting
  - AdaBoost, SMOTE Boost, Data Boost-IM, RUSBoost
  - , MSMOTEBoost
- Hybrid
  - Easy Ensemble
  - Balance Cascade

2. Methods based on Cost-Sensitive Ensembles

- Cost Sensitive Boosting
  - Rare Boost
  - CSB1
  - CSB2
  - AdaC1
  - AdaCost
  - AdaC3
  - AdaC2

3. Random Forest

- Balanced
- Weighted

5. Methods for feature selection

- Wrapper
- Filter
- PREE
- Embedded

Strategies used for feature selection

1. Common filtering strategies are

- Gain ratio, X^2 statistic, Information gain, Relief and ReliefF, Symmetric Uncertainty, and chi-square.

2. For Binary data

I. One sided metric
   - Odd Ratio
   - Correlation Coefficient

II. Two sided metric
   - Information gain
   - Chi Squared

3. For Continuous data

I. One sided metric
   - S2N ratio (Signal to Noise)
   - Pearson Correlation Coefficient

II. Two sided metric
   - FAST
   - FAIR

6. Algorithm approach

1.Cost Sensitive Learning

- Misclassification cost
- Cost minimizing
- Cost sensitive function

2.One class classification approach

- One Class SVM
- Weighted One Class Classification

7. Algorithm Modifications

- Modifying the distribution reference in the tree
- Balance Entropy
- Advanced splitting criteria

VI. COMPARATIVE ANALYSIS

The benefits and drawbacks of the suggested methods to handle imbalanced class problem as well as comparative analysis of various algorithms are specified in Table III. as well as in Table IV. resp.

Table III: Benefits and shortcomings of the proposed methods to deal with imbalanced class problem. [6]

Approach	Benefits	Shortcomings
1)Sampling i)Under sampling	Implementation is easy. Free from underlying classifier.	Can cause information loss due to elimination of important patterns.
		Time consuming
ii)Oversampling		May lead to overfitting.
2)Cost Sensitive Approach	Reduces misclassification cost .	Misclassification cost are unknown
3)Reorganization based	Improvement in performance if data contains more dimensions .	One class learning is not used to make Naïve Bayes classifier and Decision tree classifier.
4) Approach based on ensemble	Superior performance of classifiers compared to single classifier.	Time required is more. Causes overfitting.

Retrieval Number: E6286018520/2020©BEIESP
DOI:10.35940/ijrte.E6286.018520
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Table IV: Comparative study of algorithms for class imbalance.

Algorithm	Advantage	Disadvantage
AdaBoost [5]	Prediction accuracy is improved of minority class instance.	Overall performance of classifier is ignored.
RUSBoost [6]	Faster ,simpler as compared to SMOTE boost approach.	Multiple class imbalance dataset problem is not resolved by it.
Imbalance in logistic regression [7]	Applied in binary classification.	Performance is affected by outlier.
Linear Proximal SVM [8]	Can deal with dynamic imbalance class problem .	No consideration of sample distribution.
Boosting SVM[9]	SVM classifier performance is upgraded to predict instances of minority class.	Class imbalance is ignored.

VII. CHALLENGES AND OPPORTUNITIES

There are still a number of challenges and opportunities in handling the imbalance data. Following are some of the fields to deal with the challenges: [4]

1. Classification of Multi-instance and Multi-label imbalanced dataset

The assignment of each sample to a set of target label is a multi-label classification. In a multi-instance classification set of bags is trained. Each bag has multiple instances. Bags are labeled not the objects, nor the objects inside bags. Since labels are assigned to bags, not the objects, it does not indicate that only objects appropriate for the given class will be there in the bag. This exists as one of the machine learning problems. Less attention is given to study imbalanced datasets. An approach such as resampling and cost-sensitive is used to deal with uneven distribution of instances and bags inside bags in case of multi-instance classification. A noteworthy development is required in learning multi-instance as well as multi-label imbalanced data. The challenges in learning imbalanced multi-instance as well as multi-label data are:

- There is a requirement of skew-independent classifiers that classify multi-label data without using methods of resampling. The existing multi-label methods should be combined with available multi-label solutions in a multi-class classification domain. Classifiers designed for multi-label classification should be robust to show the same performance as other standard methods used for multi-label classification of balanced data.
- Another direction of research is to study decomposition-based solutions. The standard method is relevant to binary; the significant method is transforming the problem of multiple-label into two-label subproblems. The distribution of labels is balanced when applied individually to each subproblem. Another approach is the use of aggregation.

Sampling approaches that are used currently for learning multi-instance data are based moreover on the level of objects or bags inside bags.

2. Imbalanced data stream learning

The data stream is dynamic in nature. The data reaches the data streams are in online form or batches as the result there are challenges while data distribution is expected uneven. To deal with the changes in data stream as well as stationary streams, methods required should be adaptable enough to handle real life uneven objects.

The following issues are there while learning from an imbalanced data stream. [4]

- There is an assumption that the present work done is maximum on the binary data stream, where correlation among classes might perhaps vary as per time. More samples belong to one class while less in other. The imbalance problem in streaming data is due to various reasons. The issue is dealing with the bias introduced by means of classifiers, in addition to this handling the faded class is one more issue.
- Another challenge is related to obtaining class labels. The assumption is that the class labels are immediately available once the new sample is being classified which would impose higher labeling cost. Present work which is carried out assumes this. The issue here is how to sample an uneven data stream. It is required to develop labeling policies that will consider representatives of minority class as well as adjust the labeling ratio of classes.

3. Learning in Imbalanced Big data

As the complexity level increases in data, open challenges also arise. Imbalance class can also affect the Big data. An efficient and scalable algorithm, methods are required to handle heterogeneous data coming from various areas such as social networks, computer vision, etc. Hadoop and Spark, the computing environments posed additional challenges not developed initially for handling the class imbalance problem. While handling imbalanced big data the issues that need to be analyzed for deep insight are as follows:

- MapReduce is likely to fail when applied to the distributed environment. Methods based on SMOTE oversampling will not work in a distributed environment. To apply SMOTE oversampling on massive data, there is a need for effective implementations and novel global scale, data partitioning methods for data preserving the relation among samples.
- Video sequences, XML structure, graphs, and hyperspectral images are forms of big data. Methods are required for the classification and processing of big data which is in the various forms like graphs, video sequences, hyperspectral images, XML structure, etc. Some restriction is imposed by this datatype on machine learning systems. Numerical standards are used as an alternative for transforming this datatype, preprocessing and learning algorithms are designed to handled massive and skewed data.

4. Imbalanced Regression

A learning imbalance in regression is one more branch to study. There are a number of applications in real life that require prediction of extreme values of target variables and infrequent events for example fault diagnosis, economy, crisis management, etc.

VIII. APPLICATION DOMAINS OF IMBALANCED DATA CLASSIFICATION [1]

- Financial Management
- Chemical and biomedical engineering
- Information Technology
- Energy Management
- Security management
- Infrastructure and industrial manufacturing
- Electronics and communications
- Business Management
- Emergency Management
- Policy, social and education
- Agriculture and horticulture
- Environmental management

IX. EXPERIMENTAL RESULT

A) Experimental setup

i) Dataset- The dataset selected for experiment is credit card fraud.

ii) Objective - The objective is to label fraudulent or genuine for anonymized credit card transactions.

iii) Data set description – Credit card fraud dataset covers transactions done in the month of Sep year 2013 by cardholders of Europe. Transactions happened in 2 days are presented in the dataset. The Credit card dataset is very uneven, the minority class (frauds) reading is 0:172% of total transactions. It is found that there are 492 frauds out of 284,807 transactions.

iv) Methodology –

Step 1) Applied Random under-sampling technique to balance the dataset.
Step 2) The classifier models applied are Random Forest Classifier (RFC), Logistic Regression, Support Vector Machine, Multilayer Perceptron and KNeighbour.

We can get the count of majority and minority class as:
Class 0: 284315; Class 1: 492

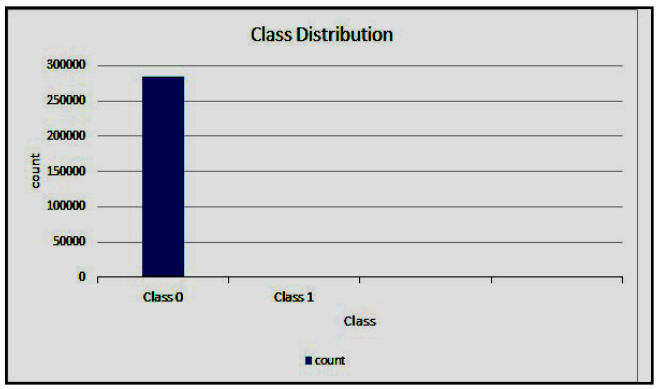

Fig 3. Class count

v) Results

Table V: Performance metrics used for classifiers

Classifier Model	Accuracy	Precision	Recall	ROC-AUC Score
Logistic Regression	0.9992	0.868	0.587	0.917
RFC	0.9995	0.872	0.737	0.938
KNeighbour	0.999	0.824	0.601	0.920
MLPclassifier	0.9991	0.82	0.67	0.938
Linear SVC	0.9992	0.882	0.713	0.926

1)Accuracy measurement for classifiers

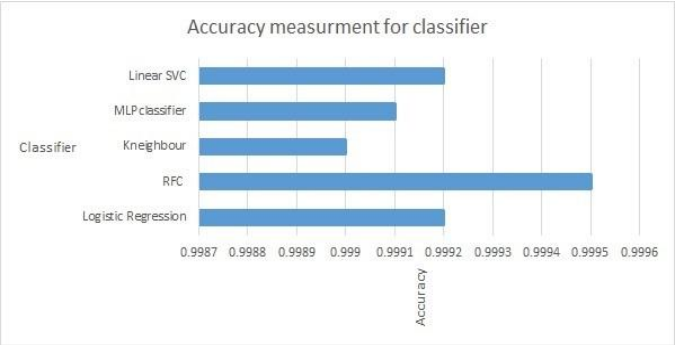

Fig 4. Accuracy measurement for classifiers

2) Precision measurement for classifiers

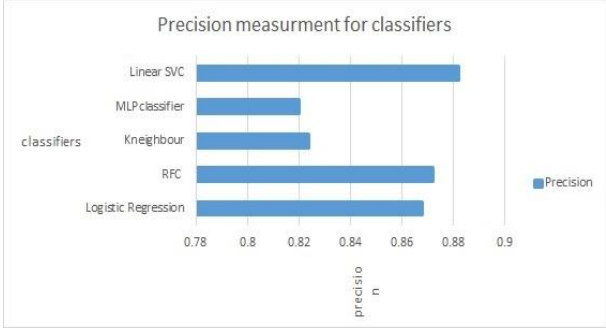

Fig 5. Precision measurement for classifiers

3) Recall measurement for classifiers

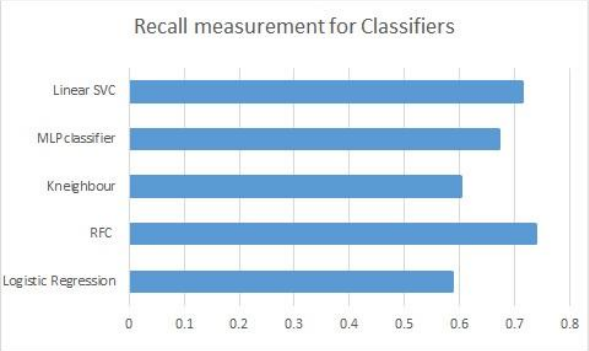

Fig 6. Recall measurement for classifiers

4) Receiver Operating characteristic curve

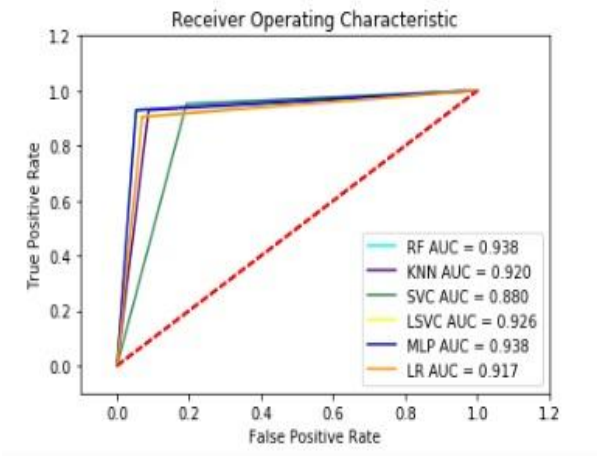

Fig 7. ROC Curve

vi) Conclusion

From the experimental result on the imbalanced dataset it is found that the Random Forest Classifier has more Accuracy, Recall and ROC-AUC score as compared to Logistic Regression, Linear SVC, KNeighbour and MLP classifier model. while applying smote and other classifier model logistic Regression, Random Forest, Linear SVC, KNeighbour, Multilayer Perceptron on it, the accuracy, precision and Recall values which we get will not necessarily be the same in case of other imbalanced datasets.

X. CONCLUSION AND FUTURE RESEARCH DIRECTION

The paper summarizes the class imbalance problem, solutions to handle it, experiment results and the research challenges in real-life applications along with different aspects of imbalanced learning is discussed. Based on the experiment conducted on credit card transaction dataset, it is found that Random Forest has good Accuracy, Recall and ROC-AUC score compared to other classifier model. Despite a lot of work is done on imbalanced learning there are still problems that need to be addressed and new methods need to be developed and existing methods need to be modified due to limitations.

Following directions can be considered for finding solutions for imbalance learning:

1) More emphasis on the nature and construction of minority class instances for a better understanding of the problems.
2) Development of multiple-class imbalanced learning methods.
3) Proposal of novel responses intended for multiple-instance learning as well as learning multiple-label.
4) Development of techniques required to study in-depth individual characteristics of rare samples considering imbalanced regression problem.
5) Analysis of class imbalance in the data stream with shifting distribution.
6) Adapting the imbalanced learning in security management issues.
7) Emergency management: Predicting natural calamities in uneven distribution of data, since natural calamities, are infrequent. However, another kind of emergency event includes incidents related to public health such as outbursts of diseases like malaria, cholera and Ebola, accidents like forest fires, as well as incidents related to social security like terrorist attacks.

Thus it is found that there is still a number of problems related to imbalanced learning that are required to be solved. Attention is required from the researchers in machine learning for the rigorous development of methods.

REFERENCES

1. Guo Haixiang, Li Yijing Jennifer Shang, Gu Mingyun Huang Yuanyue, Gong Bing. (2017, January). “Learning from class imbalanced data: Review of methods and applications”. Elsevier Journal Expert Systems with Applications.Vol 73. pp.220-239.
2. Victoria Lopez, Alberto Fernandez, Salvador Garcia Vasile Palade, Francisco Herrera. (2013, July). “An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics”. Elsevier Journal Information Sciences. pp. 113-141.
3. Mohamed, Bekkar, Dr. Taklit Akrouf Alitouche (2013, July). “Imbalanced Data Learning Approaches Review”. International Journal of Data Mining & Knowledge Process, Academy and research collaboration center (AIRCC), Vol.3, pp.15.
4. Bartosz krawcyk (2016). “Learning from imbalanced data: Open challenges and future directions”. Springer Review.
5. Ronaldo C. Prati, Gustava E.A.P.A. Batista, Maria Carolina Monard (2019). “Data Mining with Imbalanced Class Distribution: Concepts and Methods”,4th International Conference on Artificial Intelligence 2019.
6. Shaza M. Abd Elrahman, Ajith Abraham (2013). “A Review of Class Imbalance Problem”, Dynamic Publisher, Journal of Network and Innovative Computing, Vol.1. pp. 332-340.
7. Shuo Wang, Member, and Xin Yao (August 2012). “Multiclass Imbalance Problems: Analysis and Potential Solutions”, IEEE Transactions On Systems, Man, And Cybernetics—Part B: Cybernetics, Vol. 42, No. 4,
8. Chris Seiffert, Taghi M. Khoshgoftaar, Jason Van Hulse, and Amri Napolitano (2010, January). “RUSBoost: A Hybrid Approach to Alleviating Class Imbalance,” IEEE Transactions On Systems, Man, And Cybernetics—Part A: Systems and Humans, Vol. 40, No. 1.
9. Rukshan Batuwita and Vasile Palade (2010). “Fuzzy Support Vector Machines for Class Imbalance Learning” IEEE Transactions On Fuzzy Systems, Vol 18, No. 3
10. Lei Zhu, Shaoning Pang, Gang Chen, and Abdolhossein Sarrafzadeh (2012). “Class Imbalance Robust Incremental LPSVM for Data Streams Learning”, WCCI 2012 IEEE World Congress on Computational Intelligence. Australia.
11. Benjamin X. Wang and Nathalie Japkowicz (2009), “Boosting Support Vector Machines for Imbalanced Data Sets”, Proceedings of the 20th International Conference on Machine Learning-2009.
12. Rozianiwati Yusof, Khairul Azhar Kasmiran, Aida Mustapha, Norwati Mustapha, Nor Asma Mohd Zin (2017, April). “Techniques for Handling Imbalanced Datasets When Producing Classifier Models” Journal of Theoretical and Applied Information Technology, Vol.95. No 7.
13. Gao, Z., Zhang, L., Chen, M.-Y., Hauptmann, A.G., Zhang, H., Cai, A.-N. (2014). “Enhanced and hierarchical structure algorithm for data imbalance problem in semantic extraction under massive video dataset. Multimedia. Tools Application”. Vol. 68(3), pp.641–657.
14. Munkhdalai, T., Namsrai, O.-E., Ryu, K.H. (2015). “Self-training in significance space of support vectors for imbalanced biomedical event data”. BMC Bioinform. Vol **16**(S–7), S6.
15. Xu, R., Chen, T., Xia, Y., Lu, Q., Liu, B., Wang, X (2015). “Word embedding composition for data imbalances in sentiment and emotion classification”. Cogn. Comput Vol **7**(2), pp.226–240.
16. Ramentol, E., Gondres, I., Lajes, S., Bello, R., Caballero, Y., Cornelis, C., Herrera, F (2016). “Fuzzy-rough imbalanced learning for the diagnosis of high voltage circuit breaker maintenance: the SMOTEFRST-2T algorithm”. Eng. Appl. AI. Vol 48, pp. 134–139.
17. Azaria, A., Richardson, A., Kraus, S., Subramanian, V.S(2014). “Behavioral analysis of insider threat: a survey and bootstrapped prediction in imbalanced data”. IEEE Transaction Computer. Soc. Syst.Vol **1**(2), pp.135–155.

18. Krawczyk, B., Galar, M., Jelen, Herrera, F. (2016). “Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy”. Application Soft Computing. Vol.38, pp.714–726.
19. Siers, M.J., Islam, M.Z. (2015). “Software defect prediction using a cost sensitive decision forest and voting, and a potential solution to the class imbalance problem”. Inform. Syst. Vol.51, pp.62–71.
20. Sun, Y., Wong, A.K.C., Kamel, M.S (2009). “Classification of Imbalanced data: a review”. International Journal of Pattern Recognition Artificial. Intelligence. Vol. 23(4), pp.687–719.
21. Ali, Aida and Shamsuddin, Siti Mariyam and Ralescu, Anca L (2015). “Classification with class imbalance problem: a review". International Journal of Advanced Soft Computing Application, Vol.7, pp.176-204
22. Rout, Neelam and Mishra, Debahuti and Mallick, Manas Kumar (2018). “Handling imbalanced data: a survey”, International Proceedings on Advances in Soft Computing, Intelligent Systems and Applications, Springer, pp.431–443.
23. Tsai, Chih-Fong and Lin, Wei-Chao and Hu, Ya-Han and Yao, Guan-Ting, Francisco Herrera.250 (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics”, Elsevier Journal Information Sciences. pp. 113-141.
24. Fotouhi, Sara and Asadi, Shahrokh and Kattan, Michael W, (2019). “A comprehensive data level analysis for cancer diagnosis on imbalanced data”, Elsevier Journal of biomedical informatics.
25. Salunkhe, Uma R and Mali, Suresh N (2016). "Classifier ensemble design for imbalanced data classification: a hybrid approach", Journal procedia Computer Science, Elsevier, Vol.85, pp.725-732.
26. Wong, Ginny Y and Leung, Frank HF and Ling, Sai-Ho (2018)."A hybrid evolutionary preprocessing method for imbalanced datasets", Elsevier Journal Information Sciences, Vol.454, pp.161-177.
27. Buda, Mateusz and Maki, Atsuto and Mazurowski, Maciej A (2018).,"A systematic study of the class imbalance problem in convolutional neural networks", Elsevier Journal Neural Networks, Vol.106, pp.249–259.
28. Thammasiri, Dech and Hengpraprohm, Supoj and Hengpraprohm, Kairung and Mukviboonchai, Suvimol (2018)."Imbalance Classification Model for Churn Prediction", Elsevier Journal Advanced Science Letters, Vol.24. pp.1348–1351.
29. Xiao, Wendong and Zhang, Jie and Li, Yanjiao and Zhang, Sen and Yang, Weidong (2017)."Class-specific cost regulation extreme learning machine for imbalanced classification", Elsevier Journal Neurocomputing Elsevier, Vol.261, pp.70–82.
30. Lin, Wei-Chao and Tsai, Chih-Fong and Hu, Ya-Han and Jhang, Jing-Shang (2017)."Clustering-based undersampling in class-imbalanced data", Elsevier Journal Information Sciences, Vol.409, pp.17-26.
31. Jian, Chuanxia and Gao, Jian and Ao, Yinhui (2016). "A new sampling method for classifying imbalanced data based on support vector machine ensemble", Elsevier Journal Neurocomputing, Vol.193, pp.115–122.
32. Sun, Zhongbin and Song, Qinbao and Zhu, Xiaoyan and Sun, Heli and Xu, Baowen and Zhou, Yuming (2015). "A novel ensemble method for classifying imbalanced data", Elsevier Journal Pattern Recognition, Vol.48, pp.1623–1637.
33. Loyola-Gonzalez, Octavio and Martinez-Trinidad, Jose Fco and Carrasco-Ochoa, Jesus Ariel and Garcia-Borroto Milton (2016),"Effect of class imbalance on quality measures for contrast patterns: An experimental study”, Elsevier Journal Information Sciences, ,Vol.374,pp.179–192.
34. Saez, Jose A and Krawczyk, Bartosz and Wozniak, Michal (2016). "Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets”, Elsevier Journal Pattern Recognition, Vol.57, pp.164–178.
35. Longadge, Rushi and Dongre, Snehalata (2013),"Class imbalance problem in data mining review”, Vol.1, pp.332-340.
36. Di Martino, Matias and Fernandez, Alicia and Iturralde, Pablo and Lecumberry, Federico (2013). “Novel classifier scheme for imbalanced problems”, Elsevier Journal Pattern Recognition Letters, Vol.34, pp.1146–1151.
37. Ertekin (2013)."Adaptive oversampling for imbalanced data classification”, Springer Journal Information Sciences and Systems, Vol.34, pp.261–269.
38. [38] Shen A, Tong R, Deng Y,(2007),“Application of Classification Models on Credit Card Fraud Detection”, International conference on Service Systems and Service Management, IEEE.

AUTHORS PROFILE

Seema Yadav, ME(CE), is currently a Research Scholar in the Department of CE & IT at Veermata Jijabai Technological Institute, Mumbai and Assistant Professor in the Department of Information Technology at K. J. Somaiya Institute of Engineering and Information Technology, Mumbai, India. She received B.E in Computer Science & Engineering and M. E. in Computer Engineering from Nanded and Mumbai University respectively. She has more than 20 research publications in reputed Journals and Conferences. Her research interests are in the field of Database Systems, Data Mining and Big Data Analytics. She is a member of ISTE and ACM.

Girish P. Bhole, PhD, is a Professor in the Department of Computer Engineering & Information Technology at Veermata Jijabai Technological Institute, Mumbai, India. His research interest includes Distributed Systems, Communication Networks, Cloud Computing, and Big Data Analytics.