Recursive feature elimination technique for Classifying Binary Problem in Imbalanced Datasets

Sachin Kumar Soni
Dept. of CSE & IT
Ravindranath Tagore University
Bhopal, India
Sachinkumarsoni185@gmail.com

Pratima Gautam
Dept. of CSE & IT
Ravindranath Tagore University
Bhopal, India
Pratima_shkl@yahoo.com

Abstract*—imbalanced data Classification with effective form is a major area of the research, as an imbalance of high class gets certainly integral in several applications of real-world, for example, cancer detection, fraud detection, etc. Additionally, imbalanced data of high form have difficulty, where more than few learners exhibit bias in the direction of the class of majority type, and in the major cases, it possibly will discount minority class in total. Imbalance Class has been considered thoroughly over the last few decades by making use of traditional ML models, that is learning of non-deep form. There is very little research work in the field of DL of class imbalance, despite recent advances in DL, along with its growing popularity. Record-breaking outcomes in several diverse domains have been obtained, and it is of great interest to examine the application of deep neural networks for issues involving high-class imbalances. The research is surveyed to improved comprehend the effectiveness of deep learning (DL) as functional to the imbalanced data of classes. In this paper, the Recursive feature elimination technique is applied followed by the Cat boost algorithm on 3 different datasets which Wisconsin, Bupa, and Pima. The results show the accuracy achieved by the proposed algorithm is better as compared to the existing techniques

***Keywords*—class distribution, Imbalanced dataset, imbalance ratio, binary class, SVM, Cat boost**

I. INTRODUCTION

A dataset with an unequal form of distribution among its majority groups as well as minority groups is regarded as having class imbalance, & the magnitude of class imbalance can vary from moderate to extreme in real-world applications (high or extreme). If the groups, e.g., fraud, non-fraud cases, really aren't fairly distributed, a dataset can be called imbalanced. The majority class compensates much of the dataset, while the minority class is mostly called the class of the interest, with a minimal representation of the dataset. Class inequality can be expected with real-world data sets. If the degree of imbalanced data for a class of majority type is severe, then a classification algorithm will generate high overall predictive performance since most instances are expected to be predicted by the model as belonging to the majority class. A model of such type isn't technically useful, as the predicted performance of interest class (that minority class) is far much more significant for professional developers. [1].

He and Garcia [2] indicate that a common point of view maintained by researchers defining imbalanced data as data having such a high form of data imbalance among its 2 classes, suggesting that high form of data imbalance is represented the whenever ratio of the majority class to minority class from 100:1 to 10,000:1 range. Although a given range of the class disparity can be found in Big Data, a high-class imbalance is not a strict concept. From the point of view of successful problem-solving, every class imbalance amount which renders minority class modeling and prediction a dynamic and daunting activity can be called a significant disparity by professional developers [3]. It should be observed that even in the sense of binary classification, we concentrate our survey analysis of written articles on a class imbalance in the big data because usually non-binary classification issues could be described with a series of multiple forms of binary type classification tasks.

High-class disparity, also seen in the big data, makes it very difficult and daunting for a learner to recognize the minority class as a high form of data imbalance creates a partiality in the indulgence of the dominant class. As a consequence, it converts very problematic for the learner to differentiate efficiently among the minority groups and the majority groups, resulting in a comparable task to finding the conceptual needle into the haystack, particularly if imbalance class is severe. This type of biased process of learning may lead to the classification of all cases as majority (negative) class & generate a metric of an unreliably high form of accuracy. In cases where the incidence of false negatives (FN) is comparatively more costly than the false positives (FP), the prediction bias of a learner in service of the class in the majority may have negative implications. [4].

Among many patients of suspicious mole(s) form of pigmentation, for example, some are to have cancer of the melanoma that is class in minority, while many are probable to not have cancer of the melanoma, that is majority class. Here, an FN means that a cancer sufferer is misidentified as having no disease, and this a significant mistake. A false positive, on the other hand, means that a patient with no cancer is classified as having a condition, that is not a significant error-negative. The class imbalance is seen to be quite large throughout this case, thus rendering the class imbalance problem an issue of significant importance within predictive learning. Intrinsic or extrinsic class imbalance found throughout the given dataset may be considered[5], where a class imbalance of intrinsic form represents data distribution of organic form, its characteristics of provided domain & class imbalance of extrinsic form reflects the external resources like cost & storage related to domain data imbalance. An instance of intrinsic class imbalance is the issue of characterizing among 1000 forms of spam emails from 1,000,000 forms of non-spam emails, provided for most emails for non-spam. In comparison, if data transmission & collection is disrupted for the factors of an external type having a unique domain, e.g. lack of the data storage space, data collection laws based on time, etc., a domain where the data gets transmitted and gathered by making use of a consecutive stream that takes to the extrinsic form of data imbalance. It must be remembered that we do not differentiate among published works in this study

paper which concentrate either on intrinsic imbalanced data or extrinsic imbalance data.

II. LITERATURE REVIEW

Japkowicz and Stephen Widely studied elements of the problem of class imbalance, specifically from a deep perspective of learning. Three main variables of the problem were highlighted: definition complexity, size of the training set, & imbalance degree. It has given that low concept type of complexity issues was indifferent to data imbalances, however, models (C5.0 & MLP) poorly performed despite enhanced concept complexity, even though there was a low-class imbalance. Also, it summarized that, given a sufficient amount of a large quantity of training data, a serious problem could've been managed with better performance [6]. Finally, a conclusion whereby over-sampling and techniques of cost-modification are preferable over an under-sampling approach to improve model efficiency for deep learning models.

The property of intrinsic form of the classes that reflect human activities to also be imbalanced form makes it even more important to analyze the subject of AR learning algorithms for imbalance handling, especially because deep learning arrives, which usually requires a larger dataset. Recently, different methods to fix imbalance class for deep learning have been studied by[7]. The study found the various studies involving empirical research on approaching the problem of imbalance class for deep learning is inadequate. A similar survey, however, shows that classical techniques being used in deep learning contexts to handle imbalanced data (e.g. for minority classes it is random over-sampling and cost-sensitive goal function to prevent biased learning against majority classes) showing great results.

Most previous work mostly on the handling of data imbalance for a deep form of NN focuses on task type of computer vision where studied researches are controlled by image classification and therefore not translatable directly to an AR environment. A modified scheme of cost-sensitive learning was given by[8] with positive results compared to conventional cost-sensitive methods and sampling techniques (where the classes in the majority are being under-sampled or classes n the minority are being over-sampled).

This research is focused on the big data, imbalanced datasets of multi-class gained from the images of remote sensing of hyper-spectral. The efficacy of hybrid technique to such datasets is examined, wherein gets cleaned dataset using SMOTE, accompanied by the training of ANN with any of these results, while the output noise of a neural network is analyzed with ENN to remove the output noise; afterward, the resulting dataset is trained with the ANN. The results obtained indicate that when strategies for cleaning are applied to an ANN output in place of input feature vector only, a better classification outcome is achieved. Consequentially, whenever the class imbalance methods are applied to deep learning including big data scenarios, the need to recognize the essence of the classification model is clear. [9].

III. RESEARCH METHODOLOGY

PROBLEM STATEMENT: A dataset is an imbalanced class if the classification categories are not represented approximately equally. The imbalance degree (majority class to minority class size ratio) can be as high as 1:99. In developing classifiers, it is noteworthy that class imbalance is evolving as a significant problem. Furthermore, the class with both the lowest number of incidents is typically the class of concern from the viewpoint of the learning task. This question is of considerable interest as it appears in many real-world classification issues, such as remote sensing, detection of contamination, risk management, detection of fraud, and particularly medical diagnosis. External methods require preprocessing of trained data to create balanced, while internal methods deal with learning algorithm modifications to minimize their intensity to class imbalance.

PROPOSED METHODOLOGY: In this research work, the CatBoost algorithm is applied followed by the Recursive feature elimination technique. In the existing methodology, machine algorithms like logistic regression, NB, and SVM were used along with the Variance ranking feature selection technique but the problem with this technique is that this approach erases features with the variation under a specified cutoff. The awareness of a feature doesn't vary much in itself, it usually has genuine predictive power. Also, Variance ranking is not able to consider the relationship of the features with the variable of the target form. To overcome these problems, the Recursive feature elimination (RFE) technique was applied for making improvements in the previous work. The datasets used in this paper are the Wisconsin dataset which is a breast cancer dataset, the PIMA dataset which is a diabetes dataset, and the BUPA dataset which is a liver dataset. All these datasets are downloaded from the UCI repository.

RFE is prevalent as it is easy to get configure and use and as it's effective at choosing such features (columns) in a training dataset that is much more or utmost related in target variable prediction.

For data with some n features,

- On 1st round 'n-1' models which formed by all features combination leaving one. The feature having the least performance is erased
- On 2nd round models of the 'n-2' type are formed by eliminating other features.

After this process of feature reduction, the CatBoost algorithm is applied for the classification of the dataset as imbalanced or not.

Catboost is Depending on the boosting of gradients. Yandex's latest ML methodology outperforms several current boosting algorithms, such as XGBoost, Light GBMM.

Although deep learning algorithms need a large amount of data and computational resources, for many of these business problems, boosting algorithms are mostly needed. However, it takes hours to train and boost algorithms like XGBoost and occasionally will become frustrated when tuning hyper-parameters.

Base tree structure:

A major difference between CatBoost as well as other boosting algorithms would be that symmetric trees are implemented by CatBoost. This may sound insane, but it helps minimize prediction time, which is critical for environments with low latency.

For various other gradient boosting algorithms Procedure (Light GBM, XG boost).

Step 1. Make in a note of every (or a sample) data points that to train a model of highly biased form.
Step 2. Analyze data point residuals (errors).
Step 3. Train next model with similar data points by having residuals (errors) as the target values.

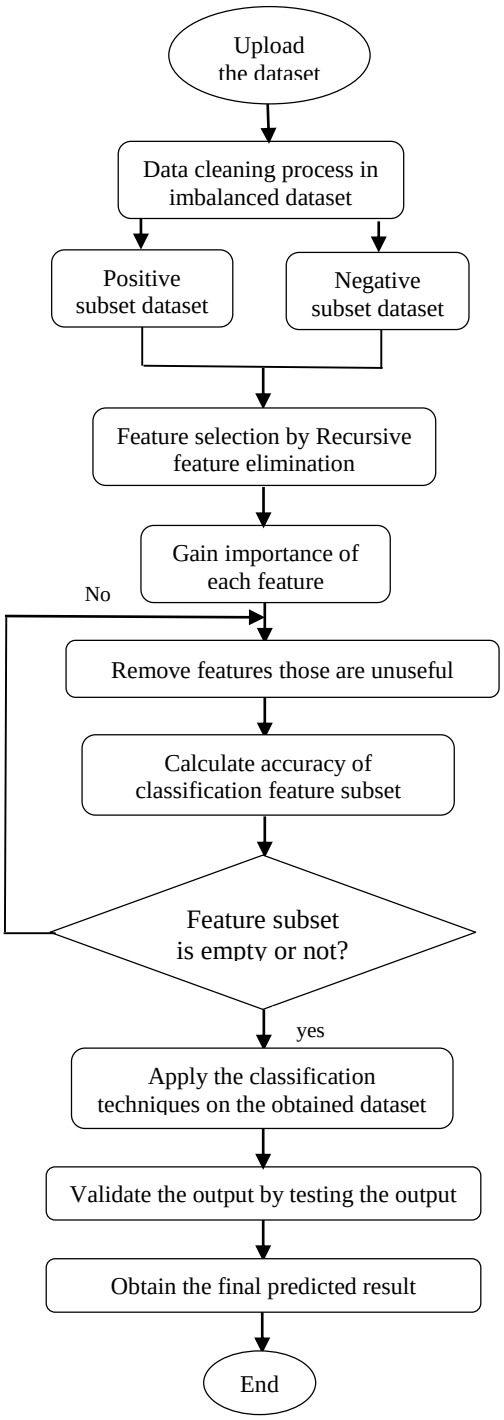
Step 4. Again Repeat the Step 2 & 3 (for no. of n iterations).

This technique is vulnerable to overfitting since by making use of the model which has already been trained on the very same set of data, we measure residuals of every other data point.

CatBoost Procedure

In a very elegant way, CatBoost performs gradient boosting. A description of CatBoost using a toy example is given below.

CatBoost splits a given dataset into randomized permutations and introduces those random permutations with ordered boosting. CatBoost generates four random permutations by nature. We may further avoid overfitting our model with that same randomness. By tuning the bagging-temperature parameter, we could further regulate this randomness.



Data flow diagram of the proposed methodology

Handling Categorical Features.

A very successful vector representation for generalized data is provided by CatBoost. It includes theories of ordered boosting and relates the same as the response coding.

We depict features of categorial form by using mean for data point target values in response coding. By using their class name, we represent the feature value of each data point. This leads to leakage of goals.

CatBoost studies only past data points to calculate the mean value. Below is a detailed explanation with examples.

Categorical Feature Combinations

CatBoost integrates various forms of categorical features. It makes sense to combine two possible features for the maximum amount of time. CatBoost eventually does it for you.

CatBoost finds the best possible feature combinations and considers them as a single feature. Therefore, Catboost is a freaking algorithm when our dataset is in categorical format and is robust for the overfitting problem.

The maximum number of iterations is taken into account in the subdivision of the data set, where the no: are less than thousand as in (Pima, Wisconsin & Bupa) several data was obtained and separated in different (training & test) data by 60 percent and 40 percent ratio, the purpose for taking this near proportion which is to evade a situation with some minority groups in the trained and test data as it divides again in negative & positive form.

Figure 1 is the data flow diagram of the proposed methodology in which the execution steps of the algorithm are briefly visualized.

IV. RESULTS AND DISCUSSION

Figure 2: Result visualization of Wisconsin dataset achieved by Catboost

Figure 3: Result visualization of Bupa dataset achieved by Catboost

F9gure 4: Result visualization of Pima dataset achieved by Catboost

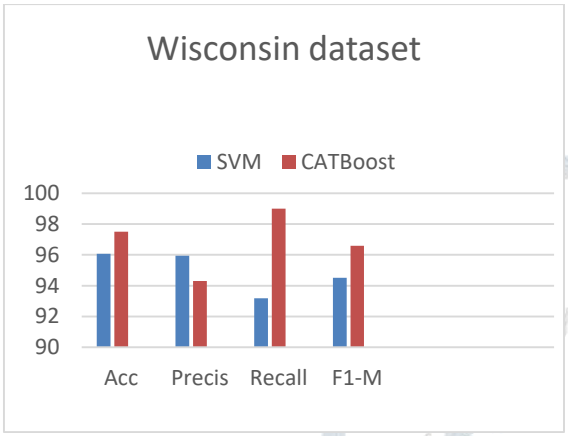


Figure 5 Comparing the performance parameters of the PIMA dataset

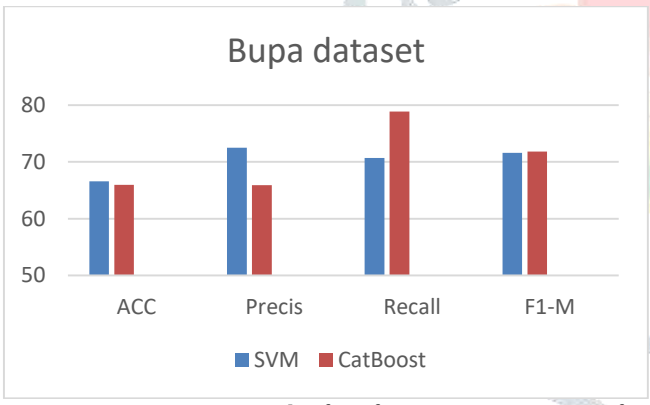


Figure 6 Comparison graph of performance parameters of BUPA dataset

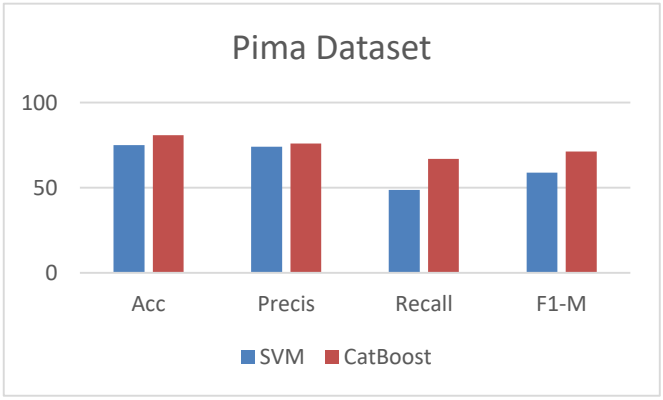


Figure 7 Comparing the performance parameters of the PIMA dataset

The above figures show the comparison graphs of the existing technique and proposed technique which shows that the CATboost has gained a remarkable accuracy as compared to that of the SVM

V. CONCLUSION

A class imbalance has been one of the most difficult issues in different fields, like DM and ML, over the last few decades. The basic state of an imbalanced dataset in which each class associated with a specific dataset is unequally distributed. This paper proposes an efficient feature selection algorithm along with a classification based on the boosting technique, as a response to the imbalanced classification problem, but on many real-world imbalanced datasets, it has achieved very good performance. The findings indicate that in terms of efficacy and efficiency about assessment metrics, i.e. accuracy, recall, F1 measure, and accuracy and recall, the proposed \ approach performs better than the previously proposed model.

REFERENCES

[1] Bauder RA, Khoshgoftaar TM. The effects of varying class distribution on learner behavior for medicare fraud detection with imbalanced Big Data. Health Inf Sci Syst. 2018;6:9

[2] He H, Garcia E. Learning from imbalanced data. IEEE Trans Knowl Data Eng. 2009;21(9):1263–84.

[3] Triguero I, Rio S, Lopez V, Bacardit J, Benítez J, Herrera F. ROSEFW-RF: the winner algorithm for the ECBDL'14 big data competition: an extremely imbalanced big data bioinformatics problem. Knowl Based Syst. 2015;87:69–79

[4] Seliya N, Khoshgoftaar TM, Van Hulse J. A study on the relationships of classifier performance metrics. In: 21st international conference on tools with artificial intelligence (ICTAI 2009). IEEE. 2009. pp. 59–66.

[5] He H, Garcia E. Learning from imbalanced data. IEEE Trans Knowl Data Eng. 2009;21(9):1263–84.

[6] Japkowicz Nathalie, Stephen Shaju. The class imbalance problem: A systematic study. Intelligent data analysis. 2002;6(5):429–49.

[7] Johnson Justin M, Khoshgoftaar Taghi M. Survey on deep learning with class imbalance. Journal of Big Data, 6(1):27, Mar 2019. ISSN 2196-1115.

[8] Khan Salman H, Hayat Munawar, Bennamoun Mohammed, Sohel Ferdous A, Togneri Roberto. Cost-sensitive learning of deep feature representations from imbalanced data. IEEE transactions on neural networks and learning systems. 2017;29(8):3573–87

[9] Erendira Rendón, Roberto Alejo, Carlos Castorena, Frank J. Isidro-Ortega and Everardo E. Granda-Gutiérrez "Data Sampling Methods to Deal With the Big Data Multi-Class Imbalance Problem" Applied Science 2020,