Grenze International Journal of Engineering and Technology, Jan Issue

Addressing Class Imbalance Problem in Chronic Kidney Disease

Pranita Mahajan[1] and Prachi Shahane[2]
[1-2]SIES Graduate School of Technology/Computer Engg, Navi Mumbai, India
Email: Pranita.mahajan@gmail.com, {prachi.shahane}@siesgst.ac.in

***Abstract*—Most classification problems in the healthcare domain perform poorly due to data imbalance. In this paper we did empirical study of literature related to healthcare imbalance data. To perform the experiment we have collected chronic kidney patient's data with de-identification in collaboration with the hospital in Mumbai, India. Data was collected manually and preprocessing highlighted imbalance in collected data. To understand class distribution and improve data balance we studied and implemented various algorithms. First implemented method is oversampling to increase minority observations. To generate new observations, Synthetic Minority Over-sampling Technique (SMOTE) is performed. To avoid outlier effect Adaptive Synthetic (ADASYN) sampling is performed. Majority class was undersampled with undersampling techniques such as random undersampling and variations of near miss methods. At last all results are compared to generate balance data and results are evaluated with Tomeks links metric. This dataset is further used to build an early detection model in Chronic Kidney Disease (CKD). Early detection model is built using clinical features, Gender, Age, Smoking, Alcohol, Birthplace, Start of Dialysis, Frequency of Dialysis, Diabetes, BP, Lipid Profile, Kidney stone, Creatine, Urea, Uric Acid. As the data was collected manually many features had missing values and data depicted imbalanced nature. Before building a classifier model we preprocessed data to avoid biased results in future. Early detection model is then built using Machine Learning algorithms. We compared predictions by Support Vector Machine, Random Forest, Regression model. With our data Random Forest model outperformed with over 93% success rate in classifying the patients with kidney diseases based on performance metrics.**

***Index Terms*— Data Imbalance, Healthcare, Synthetic Minority Over-sampling Technique for imbalanced data (SMOTE), Adaptive Synthetic (ADASYN) sampling, Feature Engineering, Sampling.**

I. INTRODUCTION

The healthcare industry has emerged with some early detection applications and other various healthcare related systems from the clinical and diagnosis data. In recent years, researchers are applying data mining methods such as generalization, characterization, classification, clustering, association, evolution, and pattern matching in healthcare domain. In view of the large amount of medical data being generated, there is growing pressure for improved methods of data analysis and knowledge discovery using appropriate data mining techniques [3].

A. Imbalance Data in Healthcare

When dataset contain highly unequal numbers of samples for classes, it is imbalanced classification problem. For example, in classifying whether a patient is having malignant tumor or not is a binary classification, the

Grenze ID: 01.GIJET.8.1.518

imbalanced classification problem is present when one class has significantly fewer observations than the other class. The class with more samples is called majority class and latter is minority class. With rapid growth of electronic medical records massive patient centric information is available inviting class imbalance problem. Class imbalance is an issue with target class distribution when analyzing data using machine learning algorithms. Healthcare professionals maintain data related to patients which lead to class imbalance problem. In healthcare domain analyzing minority class samples is equally important.

B. Dealing with Imbalance Data

In healthcare applications dealing with minority classes is equally important. If imbalance problem is untreated in early stage analysis, algorithm may learn wrong assumptions and highly compromise the prediction accuracy. In this paper we have discussed various techniques of resampling of datasets to address imbalance problem in healthcare datasets

II. RELATED WORK

As the performance of the majority regular classification algorithms is influenced by data imbalance problem, many researchers have studied and proposed strategies addressing it. This section reviews the literature of rebalancing methods to improve binary and multiple data imbalance problem [3], [4], [5], [6], [7], [8]. We have broadly divided the work in two categories, Data-level rebalancing and algorithm level rebalancing.

A. Data-Level Rebalancing

These rebalancing approaches create samples for minority class to improve balance in imbalanced dataset. It is accomplished during preprocessing phase of data analysis. A popularly used technique removes samples from the majority class (under-sampling) and/or adds more samples from the minority class (over-sampling) [3], [4], [5] .

Random Under-Sampling
The method removes some observations from majority class iteratively until majority and minority class has almost similar number of observations. Undersampling can be a good choice when you have a ton of data -think millions of rows. But a drawback to undersampling is that we are removing information that may be valuable. Under-sampling improves processing time and space complexity by decreasing training dataset. The major issue with under-sampling is, it may remove important samples loosing useful information.

Random Over-sampling
The method randomly duplicates observations from minority class to improve balance in dataset. More balanced training dataset is created by randomly selecting samples of minority class from training set with replacement. This method improves performance of machine learning algorithms such as artificial neural networks, support vector machines and decision trees [1], [4], [8].

Synthetic Minority Over-sampling Technique for imbalanced data (SMOTE)
SMOTE algorithm synthesizes observations by selecting two random observations of minority class. This improves overfitting issue of data imbalance classification. Section 3 shows result of SMOTE on our dataset. It is proved to be a better performer over under-sampling and oversampling methods [7]. SMOTE is preferred for high dimensional data as it creates new samples by overlapping many classes.

B. Algorithm-Level Rebalancing

If classification is performed ignoring data imbalance issue, the model will overfit while testing. This section reviews various algorithmic methods to improve data imbalance.

Bagging
The algorithm creates group of imbalance data by learning individual sample points in dataset. We can combine predictions to improve overfitting in classification. The samples are selected with replacement [1], [2].

Boosting
Unlike bagging, Boosting learns from weak classification models and builds a strong model. As the algorithm learns from mistakes made by every weak classifier, it works best by focusing on wrong predictions [1], [2].

C. Evaluation Metrics for Imbalance Classification

Evaluation metrics used for prediction models cannot be used directly for imbalance classifier evaluation as the dataset is not balanced. From literature we could divide imbalance metrics into three categories [8], [9], [10], [11].

Threshold Metrics

These measures are used with an aim to minimize prediction errors. Measures such as accuracy and F – measure are used to evaluate imbalance classifier. The most popularly used measure is accuracy:

Accuracy = Correct Predictions / Total Predictions

Accuracy can be calculated by using confusion matrix. To calculate F – measure, Precision and Recall are used:

F-measure = (2 Precision * Recall)/(Precision + Recall)*

Where,

Precision = TruePositive / (TruePositive + FalsePositive)

And

Recall = TruePositive / (TruePositive + FalseNegative)

Ranking Metrics

These metric predict probability score of class membership of each observation. Popular measures used are, receiver operating characteristics (ROC) analysis and Area Under the Curve (AUC).

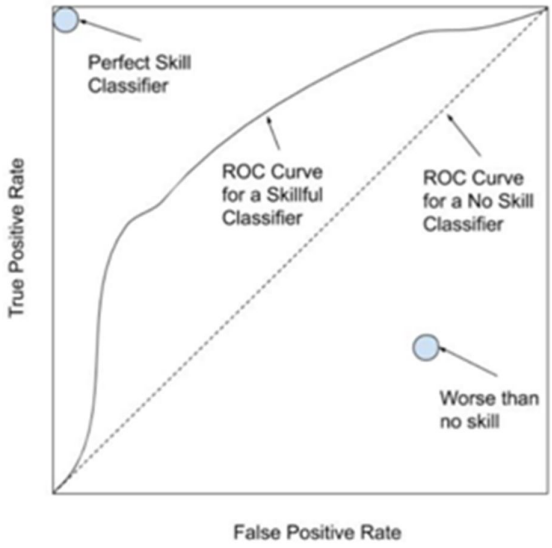

Figure 1. ROC Curve Plot

Figure 1 shows ROC curve plot, this measure performs diagnostic analysis on classifiers by identifying as skill or no skill classifier. As shown in figure the model in upper left corner of plot performs best [9]. If the number of observations in minority class are less then ROC AUC measures performs best.

Probability Metrics

Linear Models such as logistic linear regression and non linear models such as support vector machines are trained with probabilistic approach. Above discussed measures may not justify evaluation of such model. Most popular approach is to calculate log loss, which summarizes the probability distribution. Another popular method is to use BriefScore. It is the mean squared error between the expected probabilities and the predicted probabilities for the positive class.

BrierScore = 1/N * 1/Ni=1n(yhat- Y)^2

Where y is expected value and yhat is predicted values.

To summarize Random resampling is immature method of rebalancing as it may lose important information. Random oversampling may result in overfitting due to duplicating minority observations. SMOTE and Ensemble algorithmic approaches are praised by researchers to improve data imbalance with more precision.

III. Implementation Details

To understand class distribution and improve data balance we studied and implemented various algorithms. First implemented method is oversampling to increase minority observations. To generate new observations, Synthetic Minority Over-sampling Technique (SMOTE) is performed. To avoid outlier effect Adaptive Synthetic (ADASYN) sampling is performed.

A. Dataset

To perform experiment we have collected chronic kidney patient's data with de-dentification in collaboration with the hospital in Mumbai, India. Data was collected manually and preprocessing highlighted imbalance in collected data.

B. Experiment Details

This section details the experimet performed. We studied and implemented various algorithms as explained in following sections . First implemented method is oversampling to increase minority observations.

Exploratory Data Analysis

To understand class distribution we performed exploratory data analysis to measure observations in class. Figure 2 shows imbalance in collected dataset with minority and majority class.

Figure 2 Barplot of observations with CKd and Non-ckd class (x-axis = Class Label, y-axis = Count) Synthetic Minority Over-sampling Technique (SMOTE)

To generate new observations, Synthetic Minority Over-sampling Technique (SMOTE) is performed. Following steps were performed to implement SMOTE.

1. Randomly select 'nonckd' (Minority Class) tuple.
2. Apply k-neares neigbour algorithm. We have used k = 3.
3. The synthetic tuple is then created by generating a convex combination of randomly chosen tuple and nearest neighbour.

We have further extended the work by applying ADASYN. We observed the synthetic generated tuples from 'nonckd' instances were very close to randomly selected tuples creating duplicate instances. This affected the model learning. To avoid this effect Adaptive Synthetic (ADASYN) sampling is performed, it works in a similar manner as SMOTE, however. It generates number of samples in proportion with nearby samples which does not belong to the same class. Following steps were performed to generate new instances of 'nonckd' class.

1. Clculate the ration of 'ckd' to 'nonckd' instances.
   d = # 'nonckd' / # 'ckd'
   Where # 'ckd' – No of minority instances
   # ' nonckd' = No of majority instances
2. As the obtained ration (d = 0.59) below threshold, no of synthetic instances were calculated using
   G = (# 'nonckd' - # 'ckd') * beta
   Where beta is desired balancing level
3. Apply K- neares neighbour algorithm to each 'nonckd' instance and find the 'ckd' class dominance by:
   r = # 'ckd' / k

4. After normalizing ratio (r') obtained in step 3, new no of synthetic instances were calculated by:
   G' = G * r'
5. Step 4 shows adaptive nature of ADASYN algorithm by calculating G'. New minority class instances ('nonckd') are generated from randomly selecting 'nonckd' instance (say 'a') and its nearby 'nonckd' instance (say 'b') by:
   New_instance = a + (b-a) * lambda
   Where a and b : 'nonckd' instances
   Lambda : Constant between (0-1)

In this step we have generated 99 sampples using ADASYN method from the existing minority class instances. Figure 3 shows our improved results.

C. Result

Figure 3 shows improved results after performing SMOTE and and then improving results with ADASYN. We observed highly redudent synthetic samples after applying SMOTE. In this experiment we have improved SMOTE performance by further applying ADASYN method to generate minority class samples with improved spread and variance.

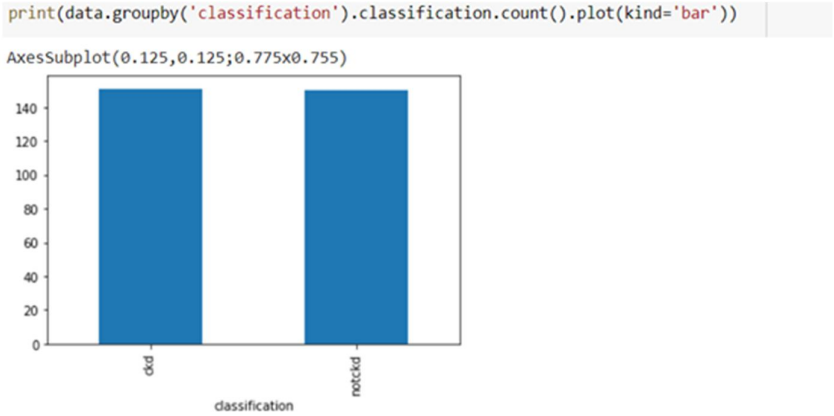

Figure 3. Improved Barplot of observations with CKd and Non-ckd class (x-axis = Class Label, y-axis = Count)

III. CONCLUSIONS

To conclude, data imbalance is crucial issue in healthcare data analysis. Literature shows high impact on prediction accuracy due to overfitting of models. We have discussed various models to perform rebalancing on imbalance data. Our experiment shows improvement after performing data resampling techniques. Readers can perform algorithmic resampling to see check the effects. We have discussed the metrics to evaluate data imbalance methods. Literature shows open challenges of resampling as improvement in overfitting and outlier reduction.

REFERENCES

[1] Fernández A., García S., Galar M., Prati R.C., Krawczyk B., Herrera F. (2018) Data Level Preprocessing Methods. In: Learning from Imbalanced Data Sets. Springer, Cham. https://doi.org/10.1007/978-3-319-98074-4_5

[2] Chapter 3 Imbalanced Datasets: From Sampling to Classifiers, Imbalanced earning: Foundations, Algorithms, and Applications, 2013.

[3] Pranita Mahajan, Dipti P. Rana, "Text Mining in Healthcare", International Journal of Innovative Technology and Exploring Engineering (IJITEE), 9(2), 2019.

[4] Zhao Y, Wong Z S Y, Tsui K L, "A framework of rebalancing imbalanced healthcare data for rare events' classification: A case of look-alike sound-alike mix-up incident detection", Journal of Healthcare Engineering:1–11.

[5] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *jair*, vol. 16, pp. 321–357, Jun. 2002.

[6] G. Douzas and F. Bacao, "Geometric SMOTE a geometrically enhanced drop-in replacement for SMOTE," *Information Sciences*, vol. 501, pp. 118–135, Oct. 2019.

[7] Yang Zhao, Zoie Shui-Yee Wong, Kwok Leung Tsui, "A Framework of Rebalancing Imbalanced Healthcare Data for Rare Events' Classification: A Case of Look-Alike Sound-Alike Mix-Up Incident Detection", Journal of Healthcare Engineering, vol. 2018, Article ID 6275435, 11 pages, 2018.
[8] Anjali S. More, Dipti P. Rana, "An Experimental Assessment of Random Forest Classification Performance Improvisation with Sampling and Stage Wise Success Rate Calculation", Procedia Computer Science, Volume 167, 2020, Pages 1711 1721, ISSN 1877-0509.
[9] Anjali S. More, Dipti P. Rana and Isha Agarwal. (2018) "Random Forest Classifier Approach for Imbalanced Big Data Classification for Smart City Application Domains", *Elsevier, International Journal of Computational Intelligence & IoT* (1) 2: 260-266.
[10] Paula Branco and Luis Torgo and Rita Ribeiro , "A Survey of Predictive Modelling under Imbalanced Distributions", arXiv , 2015.
[11] https://machinelearningmastery.com/tour-of-evaluation-metrics-for-imbalanced-classification, accessed on 4/2/2021.