Learning from Imbalanced Datasets (Supervised and Unsupervised Learning)

Ishwar Venugopal, *(1906084) MSc Artificial Intelligence,University of Essex*

Abstract—Most of the real-life datasets that are encountered for classification problems fall under the category of imbalanced datasets. Most of the traditional classification models that learn from such imbalanced datasets have a tendency to overlook the minority class and produce results that doesn't really depict the actual picture. In this work, we have proposed a new method to train a classification model from datasets with different imbalance ratios. For this purpose, three different publicly available datasets with a low, medium and high imbalance ratios were chosen. A learning mechanism similar to a stratified cross-validation process was applied by combining the techniques of k-means clustering and a Random Forest Classifier. A simple Decision Tree and a simple Random Forest Classifier were chosen as baselines. The results obtained were then compared with these baselines. Permutation tests were carried out to understand the improvement in the results using the new model, as determined by the accuracy scores as well as the Cohen Kappa scores. It was observed that our model out-performed the baselines in cases where the imbalance ratio was either very low or very high.

1 Introduction

The problem of classification is one of the most widely dealt-with machine learning problems. Unequal representation of classes in a classification problem lead to the issue of imbalanced data. For instance, a binary-classification data is said to be imbalanced if the number of instances in one class significantly outnumbers the number of instances present for the other class. This can be extended to multi-class classification problems as well. Either the rarity in such instances of minority classes or the difficulty one may have to endure for obtaining data from such classes, are the major reasons that result in an imbalanced dataset [1]. Classifiers that learn from such imbalanced datasets often tend to overlook the minority class [1], [2].

A large number of real-life datasets fall into this category of imbalanced data. In most of these cases, overlooking the minority classes and obtaining incorrect classification results might come at a very big cost. For instance, a classifier built to detect fraudulent transactions may have to be trained on a dataset which has significantly less number of fraudulent instances as compared to normal transactions [3]. Similarly, improper classification results in a medical dataset can result in very risky decisions [4]. Several methods have been proposed to tackle this problem. Resampling of data and generation of synthetic data points are some widely adopted methods to acknowledge the lower number of instances from a minority class.

The choice of performance metrics for a classifier that learns from an imbalanced dataset, should take into account that it is not merely reflecting the underlying class distribution. Accuracy scores are often misleading in this context [5]. In the current work, we will be looking at the Cohen's Kappa Score as the main performance metric as it is a method of determining the classification accuracy after being normalized by the imbalance of classes in the dataset [6], [7]. Here, we propose a new method for training a classifier that takes into account the imbalanced distribution of classes and without resampling the original dataset. We will be using an approach similar to a cross-validation strategy by combining the techniques of k-mean clustering and a Random Forest classifier. A simple Decision Tree Classifier and a Random Forest Classifier have been used as baselines. The Cohen Kappa Scores of our model has been compared with these baselines. We will also be looking at the accuracy scores for comparison. To understand the effect of the imbalance in data on such classifiers, we will be training our model on three different datasets with different imbalance ratios and compare the results.

An overview of the related works in this field have been presented in Section 2. Section 3 will introduce the methodology used in this work and the subsequent sections have been used to present, discuss and draw conclusions from the results.

2 Background

Most of the works carried out in the context of imbalanced learning can be categorized broadly into Resampling methods and Ensembling methods.

Under-sampling and over-sampling are two very popular resampling techniques with different variations in itself. Under-sampling mainly deals with removing instances from the majority class in order to achieve a balance of classes in the dataset. Yu, Dong-Jun, et al [8] and Prusa, Joseph, et al [9] have implemented random under-sampling methods to alleviate the problem of data imbalance in the domains of biomolecule residue predictions and sentiment analysis, respectively. Mani, Inderjeet, and I. Zhang. [11] studied the effects of under-sampling on the k-nearest neighbour approach to a classification problem involving the extraction of biological protein names. Yen, Show-Jane, and Yue-Shi Lee [10] have also shown to apply a cluster-based under-sampling technique on generated synthetic datasets. Liu, Xu-Ying et al [12] have presented two different algorithms that overcome the deficiency of ignoring majority class

• *E-mail: iv19023@essex.ac.uk*

samples. One of the algorithms train the learner from different subsets of the majority class instances, while the other approach learns sequentially.

On the other hand, over-sampling is mainly aimed at generating synthetic data from known instances of the minority class. Chawla, Nitesh V., et al. [13] proposed an over-sampling method called Synthetic Minority Over-sampling Technique (SMOTE). Later on, there were different works which focused on obtaining useful variations of this model. Han et al [14] have presented two different variations of SMOTE which deal with the over-sampling of only those minority class instances which are near the borderline. Hu, Shengguo, et al [15] have reported better results using an improved version they called MSMOTE, which additionally takes into account the elimination of noise samples by a technique called adaptive mediation.

Apart from resampling the data, there have been different works that focused on modifying the existing classifiers so as to enable it to deal with the imbalanced data. Bootstrap aggregation or bagging is one of the popular ensemble methods. Breiman, Leo [16] has reported that bagging can significantly improve the accuracy in cases where perturbation of the learning set causes significant changes in the predictor constructed. Pino-Mejías, Rafael, et al [17] used a reduced bootstrap resampling process which forced each generated bootstrap sample to have a distinct number of original observations. Boosting is another technique that has been used which focuses on generating a strong classifier from a weak base learner. Chen, Tianqi et al [19] proposed a novel sparsity-aware algorithm using a boosting technique called XGBoost which falls under the category of tree boosting algorithms. Schapire, R. E [18] and Friedman, Jerome H [20] have used other variations of boosting techniques like adaptive boosting and gradient boosting, respectively.

In resampling approaches, even though over-sampling techniques perform better than under-sampling techniques in most cases; it still has an underlying problem of overfitting the data. Boosting techniques have also been effective to a great extent, but still holds a tendency of being sensitive towards noisy data.

3 Methodology

3.1 Datasets

We will first discuss the datasets used for this work. For understanding the effect of imbalance on the results obtained, we chose three different datasets with varied imbalance ratios. The datasets were chosen such that they had a low, medium and high imbalance respectively. We have used datasets having binary classes for the sake of simplicity. The imbalance ratios were calculated as follows:

$$\%Imbalance = (\frac{No.\ of\ instances\ of\ majority\ class}{Total\ no.\ of\ instances}) \times 100 \tag{1}$$

All the code pertaining to the pre-processing of the datasets are available online [1].

3.1.1 Hepatocellular Carcinoma (HCC) Dataset

The Hepatocellular Carcinoma (HCC) dataset had a relatively lower imbalance ratio (61.82%). It is a publicly available dataset [2] containing real-life clinical data of 165 patients diagnosed with Hepatocellular Carcinoma (HCC) Disease. The data types present are multi-variate. Every instance has a corresponding value for 49 different attributes pertaining to the medical condition of the patient. The label determining the survival of the patient after one year is treated as the target variable. The value of 0 to this label corresponds to the death of the patient, and 1 denotes that the patient survives. It had 102 instances where the class value was 1 and 63 instances where it was 0.

3.1.2 Breast Cancer Dataset

This was another publicly available[3] medical dataset which had been widely used in other machine learning works as well [21], [22]. Each instance is characterised by nine different attributes describing the medical condition of patients. The target variable determined whether or not a patient suffered from breast cancer disease. It had a relatively moderate imbalance ratio (0.76%) when compared to the HCC dataset. The class values of 1 and 0 had 196 and 81 instances respectively.

3.1.3 Porto Seguro's Safe Driver Prediction Dataset

This dataset contained information regarding customers and predicted the possibility of the customer initiating an insurance claim in the coming year. This dataset was created by the Brazilian company called Porto Seguro and is publicly available online [4]. It has a very high imbalance ratio of 96.35%, compared to the other two datasets. Class values of 1 and 0 had 21694 and 573518 instances respectively.

3.2 Baselines

The first baseline chosen was a Decision Tree Classifier. The criterion to measure the quality of the split was chosen as entropy with a maximum depth and a minimum sample length of 3 and 5 respectively. A Random Forest Classifier was considered as the second baseline. Both of these baselines were trained on each of the datasets and used for comparing the results obtained with our new model.

3.3 Proposed model

The dataset was first divided into 10 bins with the same imbalance ratio as that of the original dataset. The processes described below were carried out for all permutations of these 10 bins for each of the dataset separately. A combination of clustering and classification techniques were applied to some of these bins as described the subsequent subsections.

[1] https://github.com/ishwarvenugopal/CE888_Data_Science_and_Decision_Making/tree/master/Project/Assignment_1

[2] https://www.kaggle.com/mrsantos/hcc-dataset#hcc-data-

[3] https://archive.ics.uci.edu/ml/datasets/Breast+Cancer

[4] https://www.kaggle.com/c/porto-seguro-safe-driver-prediction%20

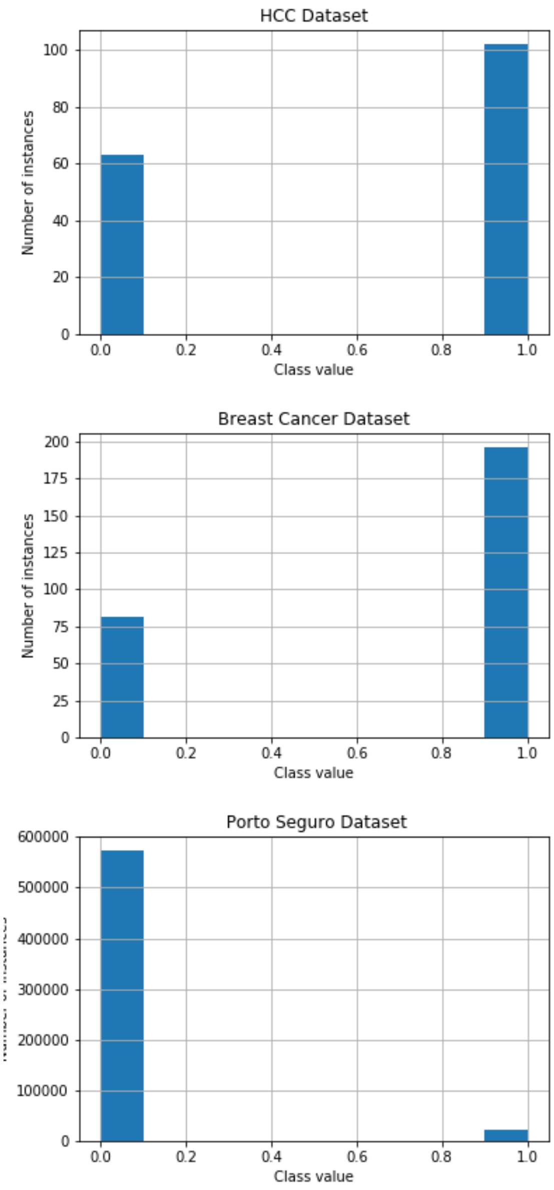

Fig. 1. Imbalance ratios in the datasets

3.3.1 K-means Clustering

A k-means clustering process was carried out on 9 of the 10 bins that were created. The optimum number of the clusters to be created by using the results obtained from a Silhouette method and the Elbow method. The Elbow method produces a plot between the sum of squared distances between data points and the k values. This plot can then be used to manually identify the ideal number of clusters. The point after which the decrease in the target value is linear is considered to be the optimum number of clusters for k-means clustering [24]. Similarly the Silhouette method produces a bar plot corresponding to the Silhouette coefficient to each k-value. The Silhouette coefficient is determined using the 'silhouette_score' function present in the Scikit-learn package [5]. The value of 1 is considered as the best value while -1 is regarded as the worst value. A value of zero corresponds to overlapping clusters [23]. A manual comparison of these plots during each permutation of the 10 bins is carried out to identify the ideal number of clusters in each case [Fig.2].

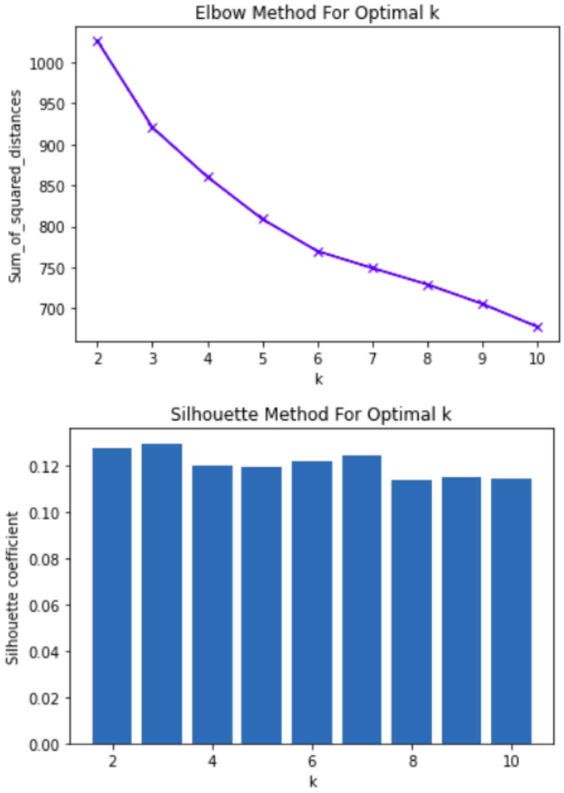

Fig. 2. The plots obtained from Elbow method and Silhouette method in one of the permutation using the Breast Cancer Dataset. It can be observed that the ideal choices can be 6 or 7. The final choice regarding the number of clusters was made based on the Kappa scores obtained by running the process with both the values

3.3.2 Classification

A k-means clustering algorithm was run on 9 of the bins using the optimum value of k selected from the Elbow method and Silhouette method. The number of instances from the minority class was then identified from each of the clusters. For every cluster that contained instances of both the classes, a Random Forest Classifier was trained on the data from that particular cluster.

Now the left-out bin was used as the unseen data to test the model. Each data point in this tenth bin was assigned to it's nearest cluster. Depending on whether that particular cluster contained instances of both the classes or not, a class label was assigned to this data point. If the cluster had just instances of one of the classes, then the data point was assigned to that particular class. In clusters containing instances of both the classes, the Random Forest Classifier trained on the data from that particular cluster was used to predict the class label.

[5] http://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html

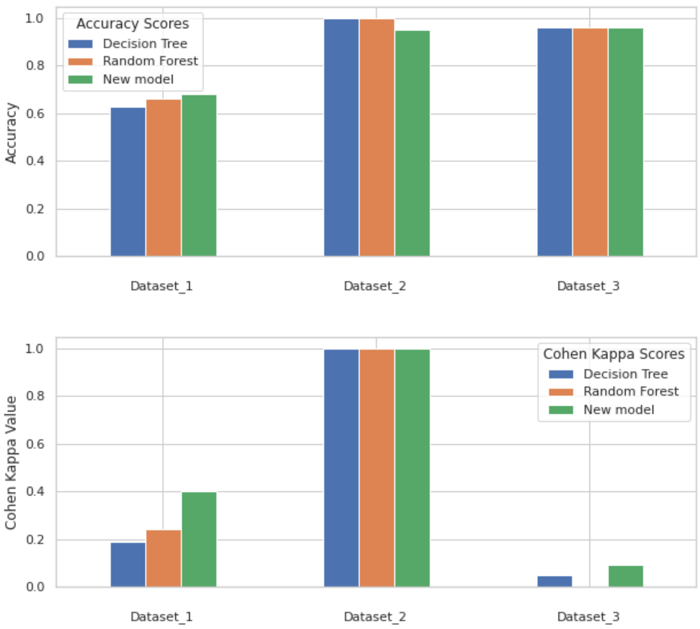

Fig. 3. Comparison between the accuracy scores and the Cohen's kappa scores obtained from each of the models. The results obtained each of the three datasets have been presented here. 'Dataset_1' corresponds to the HCC Dataset, 'Dataset_2' corresponds to the Breast Cancer Dataset and 'Dataset_3' corresponds to the Porto Seguro Dataset.

4 Results

The proposed methodology is similar to a stratified cross-validation technique. Hence, for proper comparison with the baseline models, a cross-validation was performed using both the Decision Tree Classifier as well as the Random Forest Classifier. The accuracy scores and the Cohen's kappa scores were recorded for each of these cross validation processes. Fig.3 denotes the comparison between the accuracy scores and the kappa scores obtained from different models, for each of the datasets. A boxplot for the comparison of the accuracy scores has been presented in Fig.4 and a boxplot for the kappa scores has been presented in Fig.5

To determine the extent to which the results from the new model compared to the baselines, permutation tests were carried out using the performance values obtained during cross validation of the different models on each of the datasets. The p-value obtained was compared for each of these models, on each of the datasets. The results have been combined and presented in Table.1.

5 Discussion

As mentioned earlier, accuracy scores can be misleading in the context of imbalanced datasets. In this work, Cohen's kappa score is considered as the main performance metric. The accuracy scores have also been recorded for comparison. The problem with accuracy scores is evident from Fig.3 and Fig.4. The average value of the accuracy score is 1 (which is the ideal score) for the Breast Cancer dataset. Also for the Porto Seguro Dataset, the accuracy scores are tending towards the ideal value. But in case of the HCC dataset which has a considerably lower imbalance ratio, the accuracy scores still seem to be a pretty reasonable metric for measuring the performance of the classifier. It can be

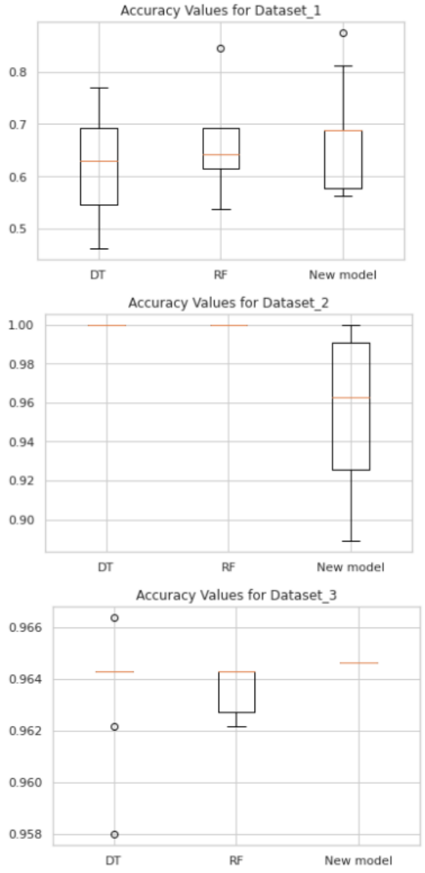

Fig. 4. A boxplot showing the comparison between the accuracy scores obtained from each of the models, for each of the datasets. 'Dataset_1' corresponds to the HCC Dataset, 'Dataset_2' corresponds to the Breast Cancer Dataset and 'Dataset_3' corresponds to the Porto Seguro Dataset.

TABLE 1
Results from the Permutation tests

Models compared	p-value for HCC Dataset		p-value for Breast Cancer Dataset		p-value for Porto Seguro Dataset	
	Accuracy	Kappa	Accuracy	Kappa	Accuracy	Kappa
Decision Tree and New Model	0.1672	0.0039	0.99868	0.0	0.10346	0.01586
Random Forest and New Model	0.36176	0.09408	0.99836	0.50024	0.0	0.32662

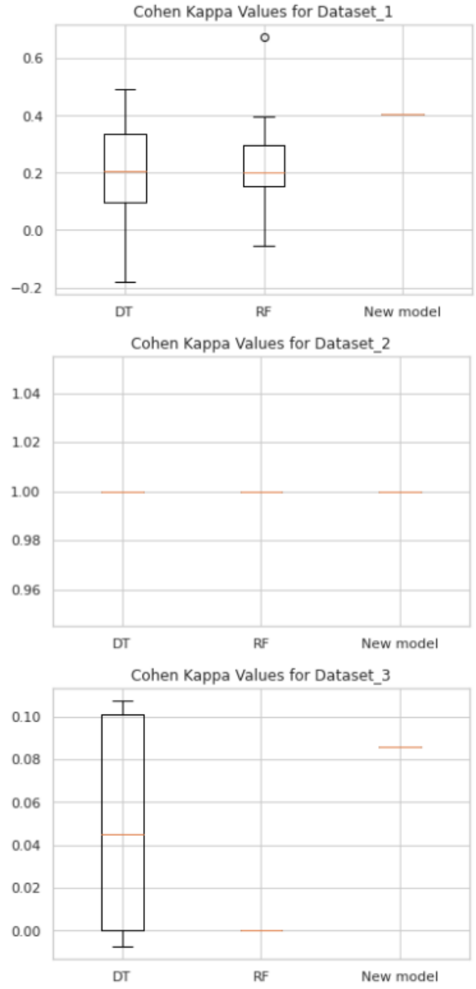

Fig. 5. A boxplot showing the comparison between the Cohen's kappa scores obtained from each of the models, for each of the datasets. 'Dataset_1' corresponds to the HCC Dataset, 'Dataset_2' corresponds to the Breast Cancer Dataset and 'Dataset_3' corresponds to the Porto Seguro Dataset.

observed from the plots for HCC dataset that the accuracy of the new model is relatively higher than both of of the baselines. The p-value further strengthens this inference. For the Breast Cancer Dataset, the accuracy measure of the new model does not provide us much information because the baseline models provide a perfect average accuracy score of 1. This can be because the baseline models have largely overlooked the minority classes. The accuracy scores doesn't seem to have changed much for each of the models trained on the Porto Seguro Dataset which had a considerably large imbalance ratio. As it can be seen from Table.1, the p-value is 0 when the accuracy of Random Forest and the new model is compared. The p-value is lower than that observed for the HCC dataset.

Looking at the Cohen's Kappa scores for the HCC Dataset from Fig.3 and Fig.5, it can be seen that it follows the same trend as the accuracy scores with the new model performing significantly better than the baseline models. The Breast Cancer dataset does not provide much information as all of the models have perfect average Kappa scores of 1. This can be largely because the models couldn't capture the imbalance of the data while making the classification. The results for the Porto Seguro dataset having a relatively high imbalance is very interesting. While the Random Forest outperformed the Decision Tree model in most of the other comparisons, here it seemed to give an average score of 0. The new model seemed to have considerably improved the performance of the classifier.

6 Conclusions

For the HCC Dataset (i.e having a low imbalance ratio), it can be seen that our model considerably improves the accuracy of the classification. Even though the trend in Cohen kappa scores show an improvement in the results for the new model, the p-value indicates that it doesn't significantly change the results obtained by a Decision Tree Classifier.

For the Breast Cancer dataset (i.e having a moderate imbalance ratio), not much information was gained by comparing the accuracy scores of different models. Our model seemed to significantly affect the Cohen Kappa scores when compared to the Random Forest Classifier, as determined by the permutation tests.

For the Porto Seguro Dataset (i.e having a high imbalance ratio), the accuracy measure seems to be the wrong choice for comparison. The kappa values are significantly improved by the new model as compared to the Random Forest Classifier. Considering the Cohen's Kappa Score as the main performance metric for comparison, we can thus infer based on these results that our method works significantly better than a Random Forest Classifier with datasets having a low imbalance ratio or very high imbalance ratios. It doesn't seem to significantly improve the results obtained from a Decision Tree Classifier. This leads us to considering the fact that Decision Tree Classifiers work relatively better than Random Forest Classifiers when it comes to dealing with imbalanced datasets. Our new model improves the performance of the Decision Tree Classifier by a relatively smaller difference.

Future works with regard to this could focus on implementing a Decision Tree classifier (instead of a Random

Forest Classifier) in clusters having instances of more than one class. The results can then be compared with the current results. Also different baselines like SVM or MLPs can be used instead of the baselines chosen in this work.

All the code for the work carried out in this paper is available online at: https://github.com/ishwarvenugopal/CE888_Data_Science_and_Decision_Making/tree/master/Project

References

[1] Kotsiantis, Sotiris, Dimitris Kanellopoulos, and Panayiotis Pintelas.*Handling imbalanced datasets: A review.*, GESTS International Transactions on Computer Science and Engineering 30.1 (2006): 25-36.

[2] Chawla, Nitesh V., Nathalie Japkowicz, and Aleksander Kotcz. *Special issue on learning from imbalanced data sets.* ACM SIGKDD explorations newsletter 6.1 (2004): 1-6.

[3] Akbani, Rehan, Stephen Kwek, and Nathalie Japkowicz. *Applying support vector machines to imbalanced datasets.* European conference on machine learning. Springer, Berlin, Heidelberg, 2004.

[4] Mazurowski, Maciej A., et al. *Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance.* Neural networks 21.2-3 (2008): 427-436.

[5] Akosa, Josephine. *Predictive accuracy: a misleading performance measure for highly imbalanced data.* Proceedings of the SAS Global Forum. 2017.

[6] Jeni, László A., Jeffrey F. Cohn, and Fernando De La Torre. *Facing imbalanced data–recommendations for the use of performance metrics.* 2013 Humaine association conference on affective computing and intelligent interaction. IEEE, 2013.

[7] Raj, Vidwath, Sven Magg, and Stefan Wermter. *Towards effective classification of imbalanced data with convolutional neural networks.* IAPR Workshop on Artificial Neural Networks in Pattern Recognition. Springer, Cham, 2016.

[8] Yu, Dong-Jun, et al. *Improving protein-ATP binding residues prediction by boosting SVMs with random under-sampling.* Neurocomputing 104 (2013): 180-190.

[9] Prusa, Joseph, et al. *Using random undersampling to alleviate class imbalance on tweet sentiment data.* 2015 IEEE international conference on information reuse and integration. IEEE, 2015.

[10] Yen, Show-Jane, and Yue-Shi Lee. *Cluster-based under-sampling approaches for imbalanced data distributions.* Expert Systems with Applications 36.3 (2009): 5718-5727.

[11] Mani, Inderjeet, and I. Zhang. *kNN approach to unbalanced data distributions: a case study involving information extraction.* Proceedings of workshop on learning from imbalanced datasets. Vol. 126. 2003.

[12] Liu, Xu-Ying, Jianxin Wu, and Zhi-Hua Zhou. *Exploratory undersampling for class-imbalance learning.* IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 39.2 (2008): 539-550.

[13] Chawla, Nitesh V., et al. *SMOTE: synthetic minority over-sampling technique.* Journal of artificial intelligence research 16 (2002): 321-357

[14] Han, Hui, Wen-Yuan Wang, and Bing-Huan Mao. *Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning.* International conference on intelligent computing. Springer, Berlin, Heidelberg, 2005.

[15] Hu, Shengguo, et al. *MSMOTE: Improving classification performance when training data is imbalanced.* 2009 second international workshop on computer science and engineering. Vol. 2. IEEE, 2009.

[16] Breiman, Leo. *Bagging predictors.* Machine learning 24.2 (1996): 123-140.

[17] Pino-Mejías, Rafael, et al. *Reduced bootstrap aggregating of learning algorithms.* Pattern recognition letters 29.3 (2008): 265-271.

[18] Schapire, R. E. *A brief introduction to boosting* . IJCAI'99: Proc. of the Sixteenth International Joint Conference on Artificial Intelligence (pp. 1401–1406)." (1999).

[19] Chen, Tianqi, and Carlos Guestrin. *Xgboost: A scalable tree boosting system.* Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016.

[20] Friedman, Jerome H. *Stochastic gradient boosting* Computational statistics data analysis 38.4 (2002): 367-378.

[21] Michalski,R.S., Mozetic,I., Hong,J., and Lavrac,N. *The Multi-Purpose Incremental Learning System AQ15 and its Testing Application to Three Medical Domains*, Proceedings of the Fifth National Conference on Artificial Intelligence, 1041-1045, Philadelphia, PA: Morgan Kaufmann, 1986.

[22] Clark,P. and Niblett,T. Induction in Noisy Domains. In Progress in Machine Learning (from the Proceedings of the 2nd European Working Session on Learning), 11-30, Bled, Yugoslavia: Sigma Press, 1987.

[23] Rousseeuw, Peter J. *Silhouettes: a graphical aid to the interpretation and validation of cluster analysis.* Journal of computational and applied mathematics 20 (1987): 53-65.

[24] Satopaa, Ville, et al. Finding a" kneedle" in a haystack: Detecting knee points in system behavior. 2011 31st international conference on distributed computing systems workshops. IEEE, 2011.