Learning from Imbalanced Datasets (Supervised and Unsupervised Learning)

Vasileios Panagaris, MSc in Data Science, University of Essex, U.K

Abstract—A dataset is named imbalanced when the classification categories are not proportionally equal. Recently a growth of attention has been brought in applying machine learning techniques in "real-world" cases related to imbalanced data problems. Mainly, training classifiers and making predictions becomes harder because they tend to classify all the data into the majority class, which is sometimes the less informative class. This paper reviews the main issues of this problem by introducing the most known sampling techniques used for balancing datasets and the specific metrics for evaluating performance. Three datasets were presented and prepared to perform analysis. Experimental results examined a new proposed method for classifying data points and compares the performance with established baseline models. Ultimately,many different approaches and recommendations to address this problems were introduced and the references cited are covering the major theoretical issues.

Index Terms—Imbalanced Data, classification, Clustering, supervised learning,unsupervised learning.

1 Introduction

Imbalanced data sets is a problem that often researchers encounter in "real world" studies, which are mostly related to classification, such as credit card fraud detection, medical diagnosis and text categorisation. An imbalanced classification problem exists if the distribution of observations across the known classes is biased or skewed. This skeweness reveals the size of the problem, which lies from moderate to severe imbalance, when there is one data point in the minority class or hundreds, thousands examples in the majority class. Therefore, the task of classification becomes harder to implement because the minority class instances are more likely to be misclassified compared to the majority class instances.

Many machine learning algorithms have been used in an effort to build classifiers from imbalanced data and a comprehensive study is shown in [1]. Moreover, accordingly to [2], 527 papers that are related to imbalanced learning were analysed from both a technical and a practical perspective. The most crucial fact, though, is that the performance metrics of the classifiers are influenced by imbalanced data sets and studies have been made in this direction to answer to which metrics are suitable and under which circumstances [3].

2 Background

Researchers have studied widely several approaches to deal with imbalance learning and, especially, with binary classification imbalanced problems. Different machine learning algorithms have been developed recently to tackle this problem, which mostly have been based on sample techniques, cost sensitive learning and ensemble methods [4]. [5] investigates the nature of the problem, the learning difficulties with standard classifier learning algorithms and evaluation measures regarding the classification of imbalanced data. Furthermore, from a technical point of view a generic framework within each algorithm can be placed should be proposed, [6], and, finally , it is crucial to identify which performance metrics are suitable for the evaluation of imbalanced datasets [7].

Many authors have established different approaches to address learning with imbalanced datasets. In order to achieve robust conclusions specific classifiers, such as decision trees, Support Vector Machines (SVM's) and k-Nearest Neighbor (kNN), will be trained and the main goal will be to find the more suitable for preprocessing, cost-sensitive learning and ensemble-based methods [8].Mainly, the issue of imbalance can be resolved through specific data level methods.

Undersampling is a non-heuristic method that focuses on balancing class distribution through the arbitrary eradication of majority class observations [9]. On the other hand, oversampling is a non-heuristic method that has the same purpose as undersampling through the random replication of minority class observations [9]. Additionaly, a feature selection framework has been proposed, which selects features for positive and negative classes independently and after that explicity combines them [10].

Another way for handling imbalance is algorithm level methods, such as cost-sensitive learning, which defines settled and unequal misclassification costs between classes in order to improve classifiers performance when learning form an imbalanced dataset, and, actually, outperforms random resampling [9]. Last but not least, ensemble-based methods is also a way to tackle imbalance and the basic idea is to compose several classifiers from the initial data and then aggregate predictions when unknown instances are presented. [8].

3 Methodology

The goal of this project is to evaluate a new approach for tackling imbalanced datasets, based on a combination

of supervised and unsupervised learning. First of all, two baseline models (Random Forest and Decision Tree) will be established in order to have the initial accuracy achieved in each dataset. Acording to [11] the best method to use for model selection is ten fold stratified cross validation, which divides data in ten folds but keeping the proportion of the target variable of the initial dataset. Mainly,the decision tree and the random forest are implemented with the option of class weighting. Receiver Operating Characteristics (ROC) graph was used to evaluate the performance [12]. Then, data was standardised in order to shift the distribution of each attribute to have a mean of zero and a standard deviation of one, and Principal Component Analysis (PCA) was performed to capture as much variance of the data was possible in two principal components. Furthermore, k-means was performed and the optimal number of k was extracted with the visual representation of the Elbow method. Silhouette method was used to validate the results of Elbow method as an accurate graphical display for partitioning techniques [13]. Finally, a random forest was trained for each cluster and the performance was evaluated on an unseen fold in order to compare it with the baseline established models.

The datasets that will be used in this analysis were found in the web platform Kaggle. After research three imbalanced datasets were chosen and carefully inspected about the variables contained. The first one is created by IBM data scientists for educational purposes and concerns employee's attrition and performance. The target variables indicate if the employee left the company (16% of the population) or continues working in IBM (84% of the population). This dataset does not contain missing values and consists of 35 variables (19 numerical and 16 categorical variables), which handled appropriately.

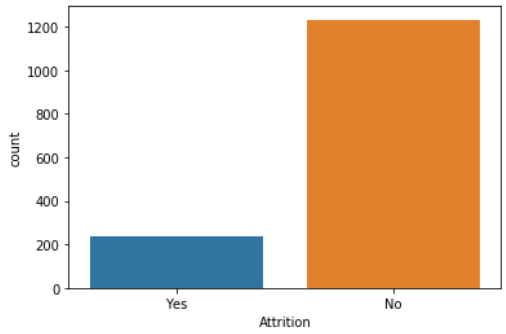

Fig. 1: Employee's attrition

The second data set was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg and concerns diagnosis of breast cancer to determine whether a suspicious lumb is cancerous or not. The target variables is the outcome of diagnosis with 63% of the population characterized as cancer patients and 37% as healthy patients. This dataset does not contain missing values and consists of 6 variables (5 numerical and 1 categorical variables), which handled appropriately.

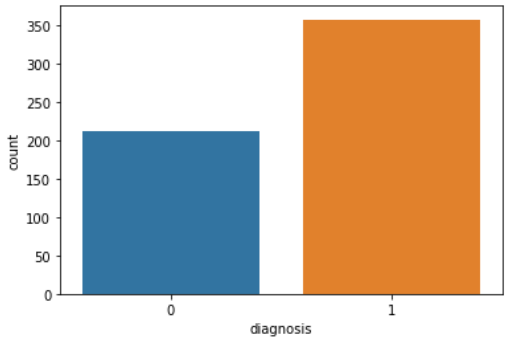

Fig. 2: Diagnosis of breast cancer

The third, and last dataset that will be used for this study, was collected from the Universal Bank and the target variable describes whether the enquiry of a customer for a personal loan was granted (9.6% of the population) or not (90.4% of the population). This dataset does not contain missing values and consists of 14 variables (6 numerical and 8 categorical variables), which handled appropriately.

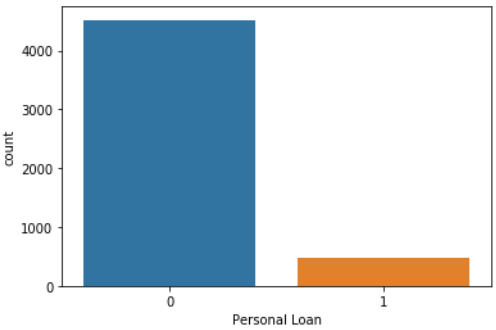

Fig. 3: Personal Loan granted

4 Results

Regarding IBM dataset the first established baseline model (Decision Tree) achieved an AUC=0.73. AUC - ROC curve is a performance measurement for classification problem at various thresholds settings.ROC is a probability curve and AUC represents degree or measure of separability. It tells how much model is capable of distinguishing between classes. An AUC=0.73 it means there is 73% chance that model will be able to distinguish between classes.

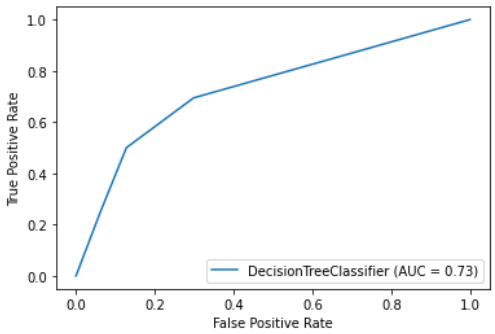

Fig. 4: DecisionTreeClassifier AUC$_{score}$

Figure 5 below, shows the ROC-AUC curve for the second established baseline model (Random Forest). An achieved AUC=0.79 is considered to be a quite good result.

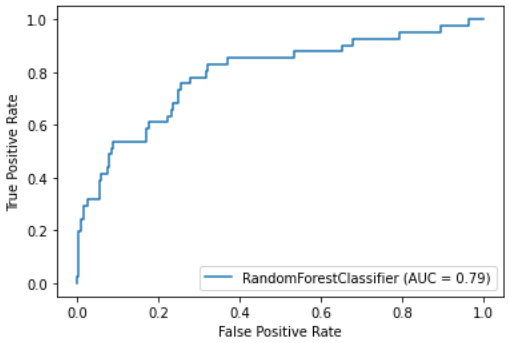

Fig. 5: RandomForestClassifier AUC$_{score}$

From Figure 6 it can be seen from the Elbow method that the optimal number of k to be used for clustering with k-means algorithm is between 3 and 5.

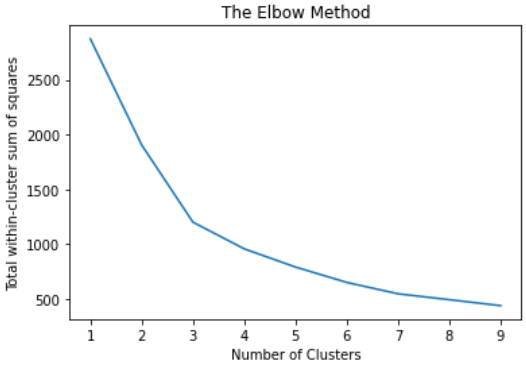

Fig. 6: Elbow Method

Consequently, the silhuette scores were calculated and it can be seen from figure 7 that the highest score is for 3 clusters and it is 0.36. Finally, figure 8 demonstrates data separated in 3 clusters.

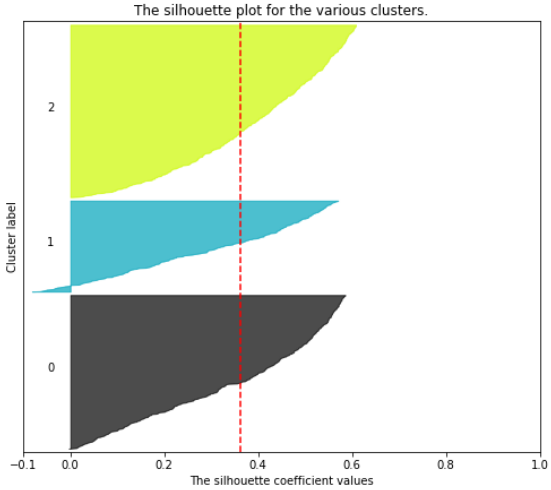

Fig. 7: Silhuette score

Fig. 8: 3-Means clustering

Concerning the breast cancer dataset the first established baseline model (Decision Tree) achieved an AUC=0.91, which means there is 91% chance that model will be able to distinguish between classes.

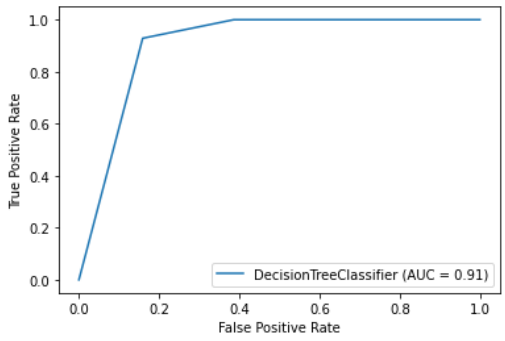

Fig. 9: DecistionTreeClassifier AUC$_{score}$

Figure 10 below, shows the ROC-AUC curve for the second established baseline model (Random Forest). An achieved AUC=0.98 is considered to be an excellent result.

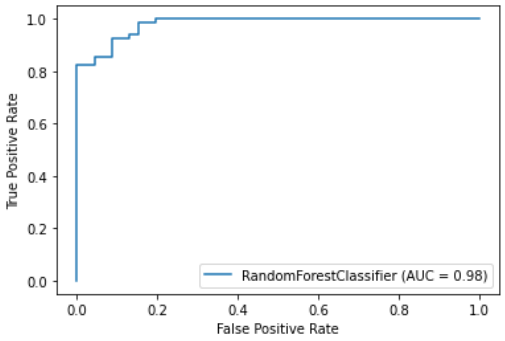

Fig. 10: RandomForestClassifier AUC$_{S}core$

From Figure 11 it can be seen from the Elbow method that the optimal number of k to be used for clustering with k-means algorithm is between 2 and 4.

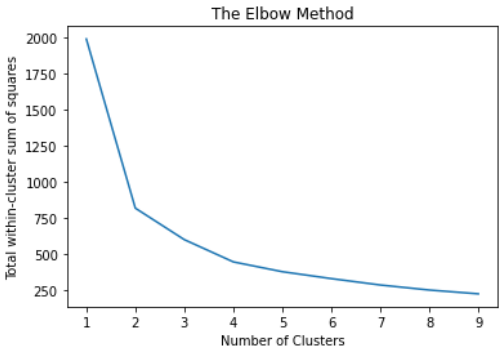

Fig. 11: Elbow Method

Consequently, the silhuette scores were calculated and it can be seen from figure 12 that the highest score is for 2 clusters and it is 0.54. Finally, figure 13 demonstrates data separated in 2 clusters.

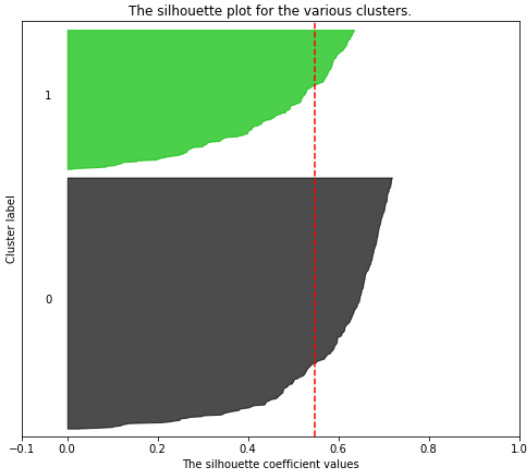

Fig. 12: Silhuette score

Fig. 13: 2-Means clustering

Concerning the universal bank dataset the first established baseline model (Decision Tree) achieved an AUC=0.95, which means there is 95% chance that model will be able to distinguish between classes.

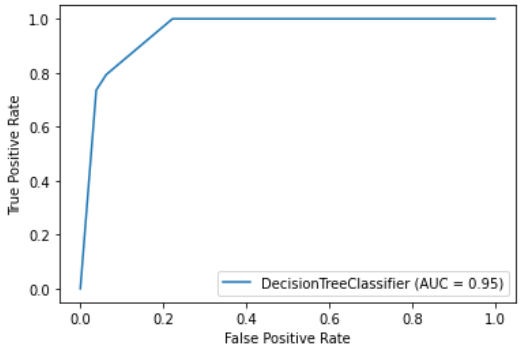

Fig. 14: DecisionTreeClassifier AUC$_{S}core$

Figure 15 below, shows the ROC-AUC curve for the second established baseline model (Random Forest). An achieved AUC=0.97 is considered to be an excellent result.

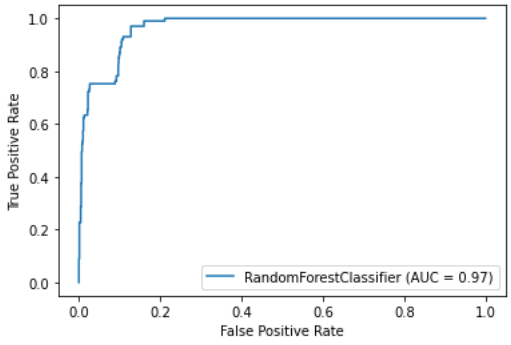

Fig. 15: RandomForestClassifier AUC$_{S}core$

From Figure 16 it can be seen from the Elbow method that the optimal number of k to be used for clustering with k-means algorithm is between 3 and 5.

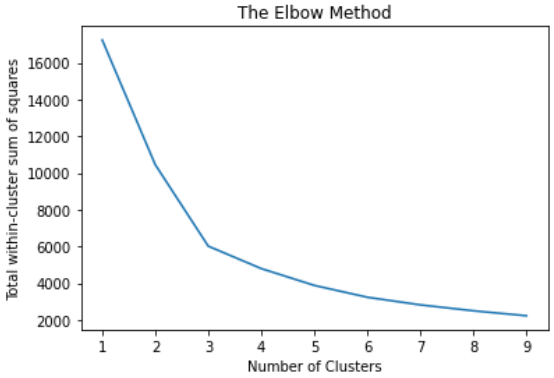

Fig. 16: Elbow Method

Consequently, the silhuette scores were calculated and it can be seen from figure 17 that the highest score is for 3 clusters and it is 0.44. Finally, figure 13 demonstrates data separated in 3 clusters.

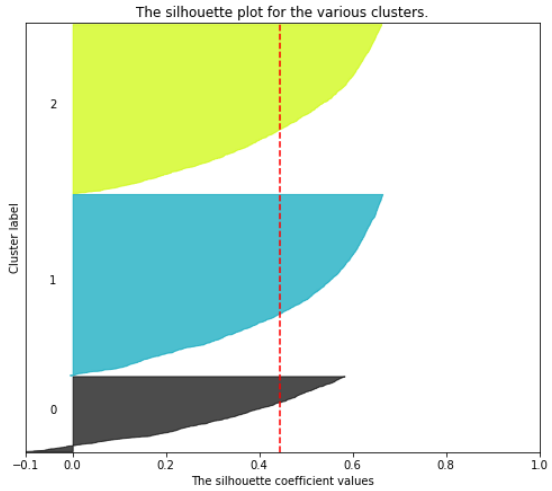

Fig. 17: Silhuette score

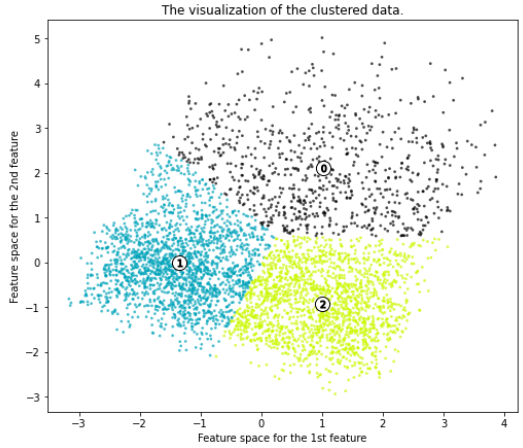

Fig. 18: 3-Means Clustering

In this point, the new method of classifying data points from imbalanced data sets is presented. After deploying k-means with the identified number of clusters for each dataset, a random forest is trained for each cluster. Then, from an unseen fold that was created after the establish of the baseline models, a sample x is assigned to the nearest cluster and, then, is classified accordingly. Finally, a 10-fold validation of this classification helps to compute the average accuracy of this procedure.

5 Discussion

It is crucial to consider how the evaluation of the results will take place. Classification accuracy is not a good metric in applications with class imbalance issues, because it places more weight on the frequent classes than on uncommon classes. Therefore, a classifier cannot perform well on the rare classes. Generally speaking, four criteras are used to evaluate the performance of classifiers when it matters imbalanced data. They are the Minimum Cost criterion (MC), the Maximum Geometry Mean (MGM) criterion of the accuracy on the majority and the minority class, the criterion of Maximum Sum (MS) and the criterion of Receiver Operating Characteristics (ROC) analysis [9]. The most frequent metric, though, is ROC analysis and the associated use of the area under the ROC curve (AUC) to assess overall classification performance. Furthermore, another well known evaluation metric is the F value, which combines precision and recall, and is high when both of them are high. Notably, even if ROC is a popular and strong measure to evaluate performance of binary classifiers, the precision-recall (PRC) plot is highly recommended as the most explanatory tool for visual analysis [14].

In this study after deploying the proposed algorithm the AUC score is compared to the baseline models that have been established in the beginning. The AUC scores for each dataset with the new method is slightly above 0.5, and it is known that a test with no better accuracy than chance has an AUC of 0.5. So, the stratified 10-folg baseline models are significantly better and the new algorithm is not proposed for better results.

6 Conclusion

Previous studies have been shown that prediction performance of classifiers is affected if the appropriate techniques and metrics are not used. Stratification provides us probably one of the best solutions when dealing imbalanced datasets, as minimises the effect that imbalance might have. As a result, the experimental results have been established a better understanding of class imbalance. Ultimately, researchers should consider new directions and take the iniative to develop and propose other approaches to imbalance learning problems, which is a very active research field of machine learning.

REFERENCES

[1] F. Provost, "Machine learning from imbalanced data sets 101," in *Proceedings of the AAAI'2000 workshop on imbalanced data sets*, vol. 68. AAAI Press, 2000, pp. 1–3.

[2] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert Systems with Applications*, vol. 73, pp. 220–239, 2017.

[3] L. A. Jeni, J. F. Cohn, and F. De La Torre, "Facing imbalanced data–recommendations for the use of performance metrics," in *2013 Humaine association conference on affective computing and intelligent interaction*. IEEE, 2013, pp. 245–251.

[4] O. Loyola-González, J. F. Martínez-Trinidad, J. A. Carrasco-Ochoa, and M. García-Borroto, "Study of the impact of resampling methods for contrast pattern based classifiers in imbalanced databases," *Neurocomputing*, vol. 175, pp. 935–947, 2016.

[5] Y. Sun, A. K. Wong, and M. S. Kamel, "Classification of imbalanced data: A review," *International journal of pattern recognition and artificial intelligence*, vol. 23, no. 04, pp. 687–719, 2009.

[6] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert Systems with Applications*, vol. 73, pp. 220–239, 2017.

[7] N. V. Chawla, "Data mining for imbalanced datasets: An overview," in *Data mining and knowledge discovery handbook*. Springer, 2009, pp. 875–886.

[8] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics," *Information sciences*, vol. 250, pp. 113–141, 2013.

[9] S. Kotsiantis, D. Kanellopoulos, P. Pintelas *et al.*, "Handling imbalanced datasets: A review," *GESTS International Transactions on Computer Science and Engineering*, vol. 30, no. 1, pp. 25–36, 2006.

[10] Z. Zheng, X. Wu, and R. Srihari, "Feature selection for text categorization on imbalanced data," *ACM Sigkdd Explorations Newsletter*, vol. 6, no. 1, pp. 80–89, 2004.

[11] R. Kohavi *et al.*, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Ijcai*, vol. 14, no. 2. Montreal, Canada, 1995, pp. 1137–1145.

[12] T. Fawcett, "Roc graphs: Notes and practical considerations for researchers," *Machine learning*, vol. 31, no. 1, pp. 1–38, 2004.

[13] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[14] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets," *PloS one*, vol. 10, no. 3, 2015.