UDC 004.85

THE STUDY OF METHODS FOR CORRECTING CLASS IMBALANCES IN MEDICAL AND PSYCHOLOGICAL DATA FOR THE DEVELOPMENT OF A RANDOM FOREST ALGORITHM

Huliiev N.B.
e-mail: nural.huliiev@nure.ua
Supervisor – Associate Professor O. S. Nazarov, Department of CMIT
Kharkiv National University of Radio Electronics, Department of CMIT
Kharkiv, Ukraine

Random forest is a widely known forecasting. Despite being a powerful mechanism for building relevant models, the algorithm can produce incorrect results and therefore needs improvement. There are many such methods available, but not all of them are necessary in our case, namely in monitoring the development of psychological disorders among people with hypo- and hyperthyroidism. In this experiment, the next step is to choose an approach to eliminate class imbalance among patient medical data, such as undersampling, oversampling, SMOTE, RUSBoost, balanced random forest (BRF), and ADASYN. Based on the data, a linear additive convolution was constructed for decision making according to the criteria of time, accuracy, precision, recall, and f1. Its indicators show that in our case, RUSBoost should be chosen as a way to combat minority classes so that the algorithm produces more accurate results.

In real-world scenarios, datasets are rarely perfectly balanced, and the problem of class imbalance increasingly complicates the classification process, making it a topical area of research. An imbalanced dataset means that one class is represented by a significantly larger number of examples, whilst the others are represented by a significantly smaller number.

Most machine learning algorithms are designed for a roughly uniform distribution of classes. Therefore, in the presence of imbalance, models tend to ‘favour’ the more numerous class, demonstrating high accuracy specifically for it, whilst the less represented class is often classified less accurately or even ignored. Consequently, this negatively affects the overall quality of classification.

Medical data is usually generated based on patient information, which often leads to imbalance. In particular, the number of healthy people in a sample often exceeds the number of patients with a specific condition. Furthermore, the varying prevalence of individual diseases also results in an uneven distribution of classes.

Although smaller classes usually have lower prediction accuracy, in medicine it is precisely these that are of greatest interest.

Therefore, errors in classifying a smaller class can have significantly more serious consequences than errors regarding the majority [1].

The focus of this study is specifically the process of predicting psychological disorders in patients with hypo- and hyperthyroidism based on medical and psychological indicators.

The subject of the study is methods for addressing class imbalance in the Random Forest algorithm.

Despite all of the above, this algorithm has a number of shortcomings. One such drawback is class imbalance, which is the focus of this study. Class imbalance arises when one class significantly outnumbers the others; consequently, such a class is considered the minority. Generally, it is assumed that in classification tasks, the data distribution is balanced. Therefore, in the opposite situation, the class with the smaller sample size is ignored during analysis or is analysed incorrectly [2].

The aim of this study is to determine the best method for addressing data imbalance when implementing the Random Forest algorithm.

We will conduct research and select the most suitable method for addressing class imbalance for the Random Forest algorithm, which has been developed to analyse medical and psychological indicators.

The alternatives considered in the study are as follows: random undersampling, random oversampling, SMOTE, RUSBoost (Random Undersampling + AdaBoost), balanced random forest (BRF), and ADASYN.

The Python code demonstrated that the six algorithms have the following performance metrics (see Table 1):

Table 1 – Numerical characteristics of algorithms

Methods and criteria	Accuracy	Precision	Recall	F1	Time	Convolution
Random under-sampling	0.535754	0.840746	0.535754	0.610383	0,0457	2,59140864
Random over-sampling	0.944794	0.943255	0.944794	0.942957	3,2543	2,43651573
SMOTE	0.941089	0.938677	0.941089	0.939295	13,3661	2,41595685
RUSBoost	0.719155	0.545685	0.719155	0.620273	0,0301	3,19695142
Balanced random forest BRF	0.696184	0.882207	0.696184	0.750664	0,1383	2,25584928
ADASYN	0.938496	0.935228	0.938496	0.936277	13,1897	2,40859136

The best result is achieved by the RUSBoost algorithm [3]. RUSBoost is an algorithm designed specifically for tasks involving unbalanced classes; it is well-suited for handling large datasets, as it is based on the random undersampling of objects from the majority class.

The method combines random reduction of the number of examples from the more represented class with a boosting mechanism. It is an alternative to SMOTEBoost, which uses oversampling, generating new minority samples by interpolating between existing examples and combining this process with boosting. By generating synthetic data, SMOTEBoost increases the training time of the model.

Although this approach has demonstrated effectiveness in a number of applications, primarily for small datasets, as the volume of training data increases, the computational cost of SMOTE rises significantly, which may render it impractical [4].

Furthermore, research findings suggest that RUSBoost may outperform SMOTEBoost, being a simpler and faster method and often delivering better classification performance. To the authors' knowledge, this is the first application of RUSBoost classifiers to the task of hippocampal segmentation [5].

References:

1. Prasetyo E., et al. (2018). Evaluating Ensemble Learning Techniques for Class Imbalance Problem. Scientific Journal of Informatics, 5(2), pp. 184–193. URL: https://journal.unnes.ac.id/journals/sji/article/view/15937/2440.

2. Kaur H., Pannu, H.S., & Malhi, A.K. (2020). Boosting methods for multi-class imbalanced data classification: an experimental review. Journal of Big Data, 7, Article 65. DOI: https://doi.org/10.1186/s40537-020-00349-y.

3. U. Hasanah, A. M. Soleh, and K. Sadik, Effect of Random Under Sampling, Oversampling, and SMOTE on the Performance of Cardiovascular Disease Prediction Models, Jurnal Matematika, Statistika dan Komputasi 21 (2024) 88–102.

4. R. Hidayat, M. A. Syawaludin, and N. Nurmalitasari, Prediksi Churn Pelanggan Multinational Bank Menggunakan Algoritma Machine Learning, Simpatik: Jurnal Sistem Informasi dan Informatika 4 (2024) 89–97.

5. F. Ismail and I. I. Lawanda, Implementasi EDMS dalam Penataan Dokumen di Rail Document System PT. Kereta Api Indonesia (Persero) Daerah Operasi 1 Jakarta, Baca: Jurnal Dokumentasi Dan Informasi 41 (2020) 143–168.