Research Exhibition in Mathematics and Computer Sciences (REMACS 6.0)/
ISBN: 978-629-97440-5-4

APPLICATION OF SMOTE TECHNIQUE FOR IMBALANCED BREAST CANCER DATA SET

Amiratul Alysha Mohd Zulfa and Mohd Hasbullah Mohd Razali
College of Computing, Informatics and Mathematics,
Universiti Teknologi MARA, Perlis Branch
amiratulalysha@gmail.com and hasbullah782@uitm.edu.my

ABSTRACT - Categorization data sets that have uneven class compositions are said to be unbalanced datasets. When a decision system is built to find a rare but significant occurrence, there are often substantial imbalances in real- world domains. The synthetic minority oversampling approach, or SMOTE, is one of the oversampling techniques most frequently employed to address the imbalance issue. By duplicating and adding more minority class samples at random, it seeks to balance the distribution of classes. SMOTE produces fresh minority instances by combining minority examples that already exist. Successful minority case prediction may be hampered by an unbalanced class distribution. This might have a significant impact on industries like medicine. Positive examples that are rare but substantial may be misclassified due to the overwhelming prevalence of the dominant class (negative cases). The data was collected from the UCI website which is Breast Cancer Data set. The algorithm chosen is J48, Random Forest and Logistic based on a little research about the best algorithm in another article The aim of this study is to apply SMOTE technique to an imbalanced breast cancer data set. By using WEKA as a tool with a different percentage of imbalance ratio to get the accurate output. Then, to validate the performance of those algorithms chosen by using Friedman Test with Hommel's Post-Hoc procedure analysis.

Keywords: Breast cancer, SMOTE technique, imbalanced class, Random Forest

1. INTRODUCTION

An essential component of data preparation is balancing the training data. The topic of imbalance has recently received more attention Chawla, N. V. et al., (2002). The safety of a model's training might be put in danger by data imbalance, which happens when classes in a data set are not distributed equally. Breast cancer is the second greatest cause of mortality among women globally, according to Mohammed, S. A., et al., (2020). For some real-world, world-class datasets, including breast cancer databases, balancing datasets is essential. With the SMOTE technique, we will approach the minority class in the imbalanced ratio of breast cancer data sets, which is the objective of this work. According to another study by Ha, T.M., & Bunke, H. (1997), this strategy was influenced by a method that has been successful in handwritten character identification. By applying certain techniques to actual data, they produced extra training data.

2. METHODOLOGY

The Waikato Environment for Knowledge Analysis (WEKA) was used to discretize the original data. The method begins with the preparation of the data which is collected in the UCI website with a CSV file. Next, apply SMOTE on the filter to begin the pre-processing data and change the percentage filter on SMOTE until you get the 40, 45 and 50 percent of Imbalanced Ratio. Then, set the percentage split to 70% before choosing algorithms for classifiers such as J48, Random Forest and Logistic. The result from the classifiers is the percentage of those algorithms chosen to find the most accurate algorithm for balancing the breast cancer data set.

3. RESULTS AND DISCUSSION

According to the Pre-Process Data table conclusion, Random Forest outperforms J48 and the logistic algorithm, the most accurate percentage for balancing was shown by Random Forest algorithm. As a result of the conclusion reached, Random Forest is substantially more accurate in making predictions.

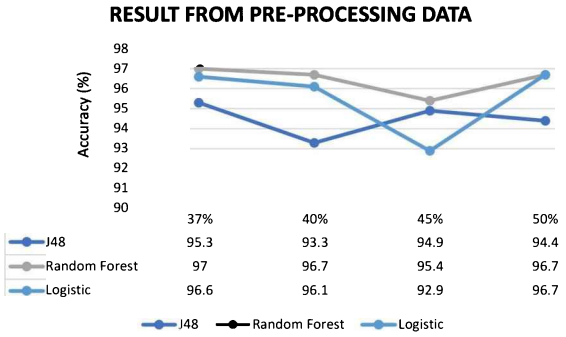

Figure 1. Graph for result from Pre-Processing data

3.1 Friedman Test with Hommel's Post-Hoc Procedure

Table 1. Result for Friedman Test with adjusted P-Hommel

Algorithm	FM. Test	Average rank	Adjusted P_{Hommel}
RF (control)	6	1.5	-
Logistic	P-value =0.049	1.5	1
J48		3	0.06779

Hommel's Procedure may discover minor variations between two procedures. After using Hommel's Procedure, Random Forest algorithm shows no adjusted P_{Hommel} and shows that Logistic algorithm and J48 algorithm has a significant difference.

4. NOVELTY OF RESEARCH / PRODUCT

The purpose of this paper is to solve the problem statement which is biased classifiers for majority class. It is advisable to replicate the same methodology on various classifiers and diverse datasets in order to obtain a well-balanced data set. These algorithms have frequently been utilised in conjunction with SMOTE due to their high efficiency, adaptability, and comprehensibility.

5. CONCLUSION

In conclusion, this study investigated which imbalanced ratio to be used and it is 40%,45% and 50% for computing the algorithm. As a result, Random Forest gets the highest classifiers percentage among other algorithms chosen.

Proven in validation the performance via Friedman Test which is Random Forest does not have a adjusted P_{Hommel} in the result.

REFERENCES

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over- sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.

Ha, T.M., & Bunke, H. (1997). Off-Line, Handwritten Numeral Recognition by Perturbation Method. *IEEE Trans. Pattern Anal. Mach. Intell., 19*, 535-539.

Mohammed, S. A., Darrab, S., Noaman, S. A., & Saake, G. (2020). Analysis of breast cancer detection using different machine learning techniques. In *Data Mining and Big Data: 5th International Conference, DMBD 2020, Belgrade, Serbia, July 14–20, 2020, Proceedings 5* (pp. 108-117). Springer Singapore.