Data Sampling Methods for Imbalanced Classification: A Random Forest Study for Predicting Treatment Switching in Multiple Sclerosis

Jieni Li, MPH[1], Yinan Huang, MS[1], and Rajender R. Aparasu, PhD, FAPhA[1]

[1]University of Houston, College of Pharmacy, Department of Pharmaceutical Health Outcomes and Policy, Houston, TX

Contact Information:
Name: **Jieni Li**
University of Houston
Phone: (713) 743-1239
Email: **jli87@central.uh.edu**

BACKGROUND

- Compared to traditional regression models, the critical advantage of the random forest (RF) algorithm is that it is a very flexible algorithm that can evaluate more predictor variables that are not limited by model assumptions.
- Imbalanced data remains a challenge for utilizing RF algorithms in healthcare research. The imbalanced data might lead to biased prediction using machine learning algorithms.

OBJECTIVE

- This study evaluated different sampling methods in RF models for predicting disease-modifying agents (DMAs) switching among patients with Multiple Sclerosis (MS). Specifically, this study evaluated up-sampling, down-sampling, and synthetic minority over-sampling techniques (SMOTE)

METHODS

Study design and data source

- This study was a retrospective cohort study using the 2009-2019 TriNetx data
- TriNetx is a federated electronic medical record (EMR) data for over 38 million patients from 24 different healthcare organizations in the US.

Sample selection and Outcomes

- Patients were required to have ≥1 outpatient visit and ≥1 prescription in 12 months pre- and 24 months post-index.
- The earliest DMA date was assigned as the index date, and patients receiving DMA other than their index DMA prescription during follow-up were considered as switched.

RF model and sampling methods

- RF models involving 72 baseline variables were trained using 70% of the randomly split data.
- RF classifiers and parameter tuning were implemented among resampled data to train RF models.
- The model performance of different sampling methods, namely up-sampling, down-sampling, and SMOTE, was examined using several composite classification metrics - Area Under the Curves (AUC), accuracy, recall, F-1 score, and G-measure.

RESULTS

Table 1. Model Performance of Different Sampling Methods

	AUC	Accuracy	Specificity	Precision	Recall	F-1	G-Measure	p-value
Up-sampling	0.6513	0.6065	0.6264	0.8893	0.6025	0.7184	0.7320	Reference
Down-sampling	0.6310	0.6221	0.5656	0.8783	0.6336	0.7361	0.7460	0.0252
SMOTE	0.5991	0.8007	0.1236	0.8419	0.9366	0.8867	0.8880	<0.0001

KEY FINDINGS AND CONCLUSIONS

- The analytical sample consisted of 6,097 (84.0%) unswitched and 1,161(16.0%) switched MS patients.
- All sampling methods alleviated the data imbalance problem of DMA switching in the study sample of MS patients.
- Among the three methods, the over-sampling method provided the best AUC for predicting treatment switching in MS.
- However, the SMOTE performed well based on the F-1 score and G measure compared to the other two sampling methods.
- Due to imbalanced data, different composite classification metrics provide a different picture. Therefore, multiple sampling methods should be evaluated based on the extent of imbalance and the need for increasing the performance of composite classification metrics of RF models.

REFERENCES

- Breiman L. Random forests. Machine Learning. 2001;45(1):5-32. doi:10.1023/A:1010933404324
- Liaw A, Wiener M. Classification and regression by random Forest. R news. 2002;2(3):18-22.
- Chawla N v, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research. 2002;16:321-357.
- Chen C, Liaw A, Breiman L. Using random forest to learn imbalanced data. University of California, Berkeley. 2004;110(1-12):24.
- Sokolova M, Japkowicz N, Szpakowicz S. Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation. In: ; 2006:1015-1021. doi:10.1007/11941439_114