ENSEMBLE METHOD FOR TEXT CLASSIFICATION IN MEDICINE WITH MULTIPLE RARE CLASSES

Alessandro Albano [1], Mariangela Sciandra [1] and Antonella Plaia [1]

[1] Department of Economics, Business and Statistics, University of Palermo, (e-mail: alessandro.albano(mariangela.sciandra,antonella.plaia)@unipa.it)

ABSTRACT: The paper presents an ensemble method for text classification in the presence of multiple rare classes in the context of medical record data. Specifically, our study aims to classify clinical notes into multiple disease categories, including rare diseases. The Ensemble method involves combining the predictions of multiple machine learning models to predict the patient's diagnosis more accurately. We used three different machine learning algorithms, namely Support Vector Machine, Random Forest, and Naive Bayes, to generate three distinct models and combine their predictions through an ensemble method. The results demonstrate that the ensemble method improves the classification performance compared to individual models. We evaluated this approach on a dataset of 50,000 clinical notes with multiple rare classes.

KEYWORDS: text classification, ensemble method, machine learning, clinical coding.

1 Introduction

In the field of medicine, text classification is a crucial task for organizing and managing large volumes of medical documents. Proper classification of medical texts can aid in decision-making processes, clinical research, and the development of new treatments. Clinical coding is the task of transforming medical information in a patient's health records into structured codes, and machine learning algorithms have been widely used to classify medical documents automatically. Nonetheless, the accuracy of machine learning methods can be boosted by assembling various methods by combining their outputs. In this paper, we explore ensemble methods for text classification in medicine, specifically dealing with multiple rare classes.
In this paper, we propose using an ensemble method for Clinical coding, i.e., transforming medical records, usually presented as free texts written by clinicians, into structured codes in a classification system like the International

Classification of Diseases (ICD-9) code, involving 18 different labels. Our approach involves fitting multiple machine learning algorithms and combining their predictions to produce a final prediction. Specifically, we use Support Vector Machine, Random Forest, and Naive Bayes and combine their predictions with improving the accuracy of our classification results. Our study adds to the expanding research on clinical natural language processing (NLP), focusing on the specific problem of text classification in the context of medical records with multiple rare classes (imbalanced labels). The literature contains important contributions, such as the work of Alsentzer *et al.* (2019), demonstrating NLP applications in medical research and clinical practice, or the study by Harrison & Sidey-Gibbons (2021) that highlights the potential of NLP models to improve medical NLP tasks. Finally, Wu *et al.* (2022) provides a comprehensive survey of clinical NLP research and applications, specifically focusing on text classification.
The paper is organized as follows. In the next section, we present the experimental setup we used in our study, including the dataset, the machine learning algorithms, and experimental results. Finally, we conclude the paper and discuss future directions for research.

2 Experimental Setup

2.1 Data

In this study, we used the MIMIC-III (Medical Information Mart for Intensive Care III) dataset, a publicly available dataset of de-identified electronic health records of patients admitted to the intensive care unit (ICU) at Beth Israel Deaconess Medical Center between 2001 and 2012. The dataset includes clinical notes such as discharge summaries, progress notes, and nursing notes. The MIMIC-III dataset is widely used in the research community for various tasks, such as predicting patient outcomes, identifying risk factors, and natural language processing.
The clinical notes from patients with different diagnoses, including rare ones (18 total different ones), were preprocessed to remove any personally identifiable information and to extract the relevant text for each diagnosis. Each note was then labelled with its corresponding diagnosis obtaining 50,000 records.

2.2 Machine learning methods

Our study used three machine learning algorithms: i) Support Vector Machine (SVM), a supervised learning algorithm that looks for the optimal hyperplane

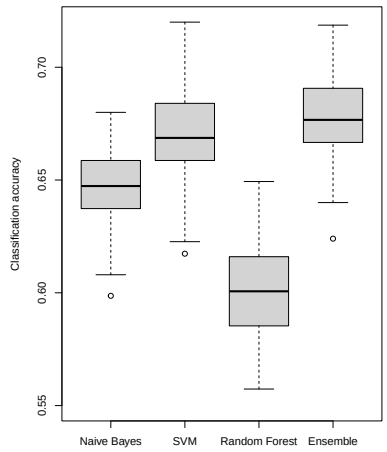
that separates the data into different classes. In our study, we used the linear kernel function to train the SVM model; ii) Random Forest (RF), an ensemble learning algorithm that combines multiple decision trees to improve the accuracy of predictions. RF works by randomly selecting a subset of variables and a subset of data samples to build multiple decision trees. In our study, we used 100 decision trees to build the RF model; iii) Naive Bayes (NB), which is a probabilistic machine learning algorithm that calculates the conditional probability of each variable given a class label and then uses Bayes' theorem to calculate the probability of each class given the variables. In our study, we used the Multinomial Naive Bayes variant to build the NB model.
We then used the Ensemble method to combine the predictions of the three machine learning models and produce a final prediction. Specifically, we used the majority voting method to combine the SVM, RF, and NB model predictions. The majority voting method works by selecting the class label predicted by most of the three models. In other words, if two or more models predict the same class label, that label is selected as the final prediction. If there is no majority, the class label predicted by the model with the iteration-specific highest accuracy is selected as the final prediction.

2.3 Results

The experiments' results (Fig.1a) indicate that the ensemble method achieved better results than individual models in predicting diseases from clinical notes. The median accuracy of the ensemble method was 67.7%, which is higher than the accuracy of individual models such as Naive Bayes (64.7%), SVM (66.9%), and Random Forest (60%), indicating that the ensemble method is more consistent in its predictions.
The results also show that the accuracy of the ensemble method was relatively stable across all quantiles of the accuracy distribution. The ensemble method was able to leverage the strengths of each model and compensate for its weaknesses.

3 Conclusion

In conclusion, our proposed ensemble method for text classification in medicine with multiple rare classes shows promising results for identifying and predicting various diseases from clinical notes. Our approach combines three machine learning algorithms (SVM, RF, and NB) to improve the accuracy of individual models. The results demonstrate that the proposed ensemble method is a



(a) Boxplot of accuracy scores.

	Dev. Std.	Median	Mean
NB	**0.017**	0.647	0.647
SVM	0.020	0.668	0.669
RF	0.020	0.600	0.600
EM	0.019	**0.676**	**0.677**

(b) Summary of accuracy scores.

promising approach for clinical coding, also when dealing with multiple rare classes or imbalanced datasets. Further research can explore the performance of the proposed ensemble method on larger datasets with a broader range of diseases, as well as the potential of incorporating other machine learning algorithms and techniques such as deep learning and active learning. In addition, exploring ways to reduce the computational complexity of the ensemble method without sacrificing performance is also an exciting avenue for future research.

References

ALSENTZER, EMILY, MURPHY, JOHN R, BOAG, WILLIE, WENG, WEI-HUNG, JIN, DI, NAUMANN, TRISTAN, & MCDERMOTT, MATTHEW. 2019. Publicly available clinical BERT embeddings. *arXiv preprint arXiv:1904.03323*.

HARRISON, CONRAD J, & SIDEY-GIBBONS, CHRIS J. 2021. Machine learning in medicine: a practical introduction to natural language processing. *BMC medical research methodology*, **21**(1), 1–11.

WU, HONGHAN, WANG, MINHONG, WU, JINGE, FRANCIS, FARAH, CHANG, YUN-HSUAN, SHAVICK, ALEX, DONG, HANG, POON, MICHAEL TC, FITZPATRICK, NATALIE, LEVINE, ADAM P, *et al.* 2022. A survey on clinical natural language processing in the United Kingdom from 2007 to 2022. *NPJ digital medicine*, **5**(1), 186.