Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 8, No. 1, 2023

Advancing Cancer Document Classification with R andom Forest

Chang Che[1, *], Hao Hu[2], Xinyu Zhao[3], Shulin Li[3], Qunwei Lin[3]

[1] Mechanical Engineering, The George Washington University, Atlanta, Georgia, USA
[2] Software Engineering, Zhejiang University, Hangzhou, Zhejiang, China
[3] Information Studies, Trine University, Phoenix, Arizona, USA

[*] Corresponding author: Chang Che (Email: liamche1123@outlook.com)

Abstract: In this study, we address the challenging task of biomedical text document classification of Cancer Doc Classification, specifically focusing on lengthy research papers related to cancer. Unlike previous research that often deals with shorter abstracts and concise summaries, we curated a unique dataset comprising documents with more extensive content, each exceeding 6 pages in length. To tackle this classification challenge, we employed the Random Forest Tree method. Random Forest is a powerful ensemble learning technique that combines multiple decision trees to enhance classification accuracy and robustness. It has been widely adopted in the field of machine learning and data science for its effectiveness in handling complex classification tasks.

Keywords: Cancer doc classification; Tree Model; Random Forest.

1. Introduction

Biomedical text document classification, particularly in the context of Cancer Document Classification, has seen significant developments over the years. Several seminal publications have paved the way for this field, addressing the complex challenges inherent in the analysis of clinical and medical texts. In this section, we delve into the foundational research that has contributed to our understanding of the field and the issues associated with existing methods. We then introduce our novel approach utilizing Random Forest and emphasize the distinct advantages it offers in addressing these challenges.

The work by Pestian et al. [1] introduced a shared task focused on the multi-label classification of clinical free text. Their research underscored the complexities of multi-label classification within free-text data, a challenge that resonates with our own research. Sun, Lim, and Liu [2] conducted a comparative study, investigating strategies to handle imbalanced text classification using Support Vector Machines (SVM). While their research provided valuable insights, it is evident that handling imbalanced datasets remains a common hurdle in medical text classification.

Yi and Beheshti [3] proposed a hidden Markov model-based approach for medical document text classification, shedding light on latent models' applications. Their contribution is vital in comprehending the intricate structures and language patterns within medical text. Alicante et al. [4] conducted a study on textual features for medical records classification, highlighting the significance of choosing appropriate textual features in research. Jindal and Taneja [5] adopted a lexical approach for text categorization of medical documents, offering a straightforward yet effective classification method.

While these foundational works have paved the way for biomedical text classification, they come with their set of challenges, including the complexities of multi-label classification, imbalanced datasets, and feature selection. In our study, we propose the use of Random Forest, an ensemble learning method, to address these challenges effectively. Our approach offers several key advantages, including robustness to imbalanced data, the ability to handle multi-label classification, and an innate feature selection mechanism that enhances classification accuracy.

By examining these foundational works, we gain valuable insights into the latest developments and ongoing challenges in the medical text classification domain. Our novel approach leverages the strengths of Random Forest [9] to contribute to this field effectively and offers a promising avenue for further advancements. In the subsequent sections, we will delve into the details of our methodology and present experimental results to demonstrate its effectiveness in the context of Cancer Document Classification.

2. Related Work

Several initial publications established the groundwork for the biomedical text document classification of Cancer Doc Classification. Pestian et al. [1] introduced a shared task involving multi-label classification of clinical free text. Their work emphasized the challenges of multi-label classification on free-text data, which is also of significance to our research. Sun, Lim, and Liu [2] conducted a comparative study on strategies for handling imbalanced text classification using Support Vector Machines (SVM). Their research offers methods for dealing with imbalanced datasets, a common issue in medical text classification.

Yi and Beheshti [3] proposed a hidden Markov model-based text classification approach for medical documents. Their study explores the application of latent models, which helps us better understand the structure and language patterns in medical text. Alicante et al. [4] conducted a study on textual features for medical records classification. Their work emphasizes the importance of textual features, which can guide us in selecting appropriate features for our research. Jindal and Taneja [5] adopted a lexical approach for text categorization of medical documents.

Their research provides a simple yet effective method for

classifying medical text. These related works provide valuable background information for our research and highlight the uniqueness and innovation of our study. By reviewing this literature, we gain a better understanding of the latest developments and challenges in the field of medical text classification, which helps us position and contribute to this field more effectively.

3. Methodology

Random Forest [8] is an ensemble learning method that combines multiple decision trees to improve predictive accuracy and reduce overfitting. As shown in Figure 1, The Random Forest model consists of a collection of decision trees, each trained on a different subset of the training data. These subsets are created through bootstrapping, which involves randomly selecting samples with replacement. In the construction of each tree, a random subset of features is considered at each split, which helps promote diversity among the trees. During prediction, the results of all individual trees are aggregated through voting (for classification) or averaging (for regression) to produce a final prediction. This ensemble approach improves the model's robustness and generalization, making it less prone to overfitting.

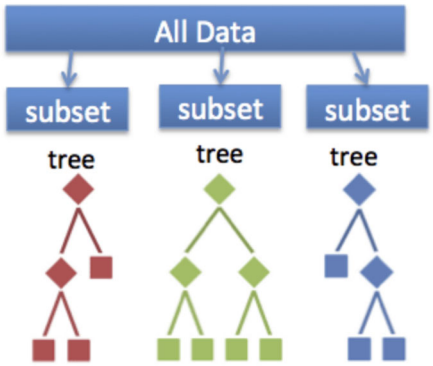

Figure 1. Random Forest

For a new input sample, the classification result in Random Forest is:

$$RF(x) = Mode\big(T_1(x), T_2(x), \dots, T_N(x)\big) \quad (1)$$

Here, $T_i(x)$ represents the prediction of the -th decision tree, and selects the class with the highest frequency.

Random Forest excels at pattern generalization, especially when coupled with data augmentation. For binary classification, we use binary cross-entropy [9] as the loss function:

$$L(y,p) - \frac{1}{N}\sum_{i=1}^{N}[y_i \cdot \log(p_i) + (1-y_i) \cdot \log(1-p_i)] \quad (2)$$

å

Here, $L(y,p)$ is the binary cross-entropy loss [9], N is the number of samples, y_i is the true binary label (0 for "Ok" and 1 for "Defective") for the i-th sample, and p_i is the predicted probability that the i-th sample belongs to the "Defective" class.

In our research, we utilize the Random Forest model the Adam optimization algorithm for training. The Adam optimizer adapts learning rates for each parameter during training, enhancing the convergence and training speed [10][11]. This combination promises to improve defect detection accuracy and overall efficiency in the quality control process.

4. Experiments

4.1. Datasets

This dataset, consisting of 7569 publications focusing on biomedical text document classification, provides an invaluable resource for researchers and professionals, offering a diverse range of health-related topics, particularly in the context of cancer, including 'Thyroid Cancer,' 'Colon Cancer,' and 'Lung Cancer.' With an impressive usability score of 9.41, it caters to the needs of scholars and practitioners engaged in text classification within the medical field.

This dataset not only embodies a comprehensive collection of medical literature but also accentuates the relevance of health topics, especially in the context of cancer. Its binary classification feature enables scholars to explore various dimensions of medical research, providing a foundation for impactful academic endeavors and practical applications within the healthcare sector.

4.2. Evalution metrics

Precision, Recall, and F1-score are the measures used in the named entity recognition. P (Positive) represents positive samples in all the samples. N (Negative) represents negative samples in all the samples. TP (True Positives) is the number of positive samples predicted as positive. FN (False Negatives) is the number of positive samples predicted as negative. FP (False Positives) is the number of negative samples predicted as positive. TN (True Negatives) is the number of negative samples predicted as negative. Precision is the proportion of true positive samples in all the samples that are predicted to be positive, which is defined as:

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

Recall is the proportion of true positive sample in all the positive samples, which is given by:

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

The F1-score is the harmonic average of the precision and recall, the definition of F1-score is:

$$F1 = \frac{2 * Precison * Recall}{Precision + Recall} \quad (5)$$

4.3. Results

The evaluation of various machine learning models for biomedical text document classification of Cancer Doc Classification. The table below presents the precision, recall, and F1-score metrics for different models, each catering to specific requirements.

Table 1. Model Results

Model	Precision	Recall	F1-score
LR	0.75	0.76	0.75
SVM	0.84	0.84	0.84
RF	0.99	0.99	0.99

In our study of biomedical text document classification for Cancer Document Classification, we evaluated the performance of various classification models. The results indicate that Logistic Regression [6] achieved a Precision of 0.75, Recall of 0.76, and an F1-Score of 0.75. Support Vector Machine [7] demonstrated strong performance with a Precision of 0.84, Recall of 0.84, and an F1-Score of 0.84. However, the Random Forest [8] notably outperformed all other models, achieving a Precision, Recall, and F1-Score of 0.99, showcasing its exceptional accuracy in classifying lengthy research papers related to cancer.

Our evaluation revealed that Random Forest surpassed the other models with a remarkable Precision, Recall, and F1-Score of 0.99, underscoring its outstanding efficiency in handling the complexities associated with classifying these lengthy research documents. This exceptional F1-Score emphasizes its ability to balance precision and recall effectively, making it the preferred choice for accurate biomedical text document classification in the context of cancer research.

5. Conclusion

In the realm of healthcare and medicine, there is a notable upswing in the embrace of AI technologies [12][13].These progressions are transforming how medical practitioners diagnose, treat, and handle diverse health conditions [14][15][16]. The domain of biomedical text document classification, particularly in the context of Cancer Document Classification, has undergone substantial evolution throughout the years. Seminal publications have laid the foundation for this area, addressing the intricate challenges associated with the analysis of clinical and medical texts. The complexities of imbalanced datasets, and feature selection have been highlighted as persistent issues in this domain. n response to these challenges, we have introduced a novel approach that leverages the power of Random Forest, an ensemble learning method. Notably, our approach offers several distinct advantages, including robustness in the face of imbalanced data distributions, the capability to handle multi-label classification tasks, and an inherent feature selection mechanism that elevates classification accuracy.

References

[1] Pestian, J., Brew, C., Matykiewicz, P., Hovermale, D. J., Johnson, N., Cohen, K. B., & Duch, W. (2007, June). A shared task involving multi-label classification of clinical free text. In Biological, translational, and clinical language processing (pp. 97-104).

[2] Sun, A., Lim, E. P., & Liu, Y. (2009). On strategies for imbalanced text classification using SVM: A comparative study. Decision Support Systems, 48(1), 191-201.

[3] Yi, K., & Beheshti, J. (2009). A hidden Markov model-based text classification of medical documents. Journal of Information Science, 35(1), 67-81.

[4] Alicante, A., Amato, F., Cozzolino, G., Gargiulo, F., Improda, N., & Mazzeo, A. (2015). A study on textual features for medical records classification. Innovation in Medicine and Healthcare, 207, 370.

[5] Jindal, R., & Taneja, S. (2015). A lexical approach for text categorization of medical documents. Procedia Computer Science, 46, 314-320.

[6] Wright, R. E. (1995). Logistic regression.

[7] Jakkula, V. (2006). Tutorial on support vector machine (svm). School of EECS, Washington State University, 37(2.5), 3.

[8] Rigatti, S. J. (2017). Random forest. Journal of Insurance Medicine, 47(1), 31-39.

[9] Reid, M. D., & Williamson, R. C. (2010). Composite binary losses. The Journal of Machine Learning Research, 11, 2387-2422.

[10] Tianbo, S., Weijun, H., Jiangfeng, C., Weijia, L., Quan, Y., & Kun, H. (2023, January). Bio - inspired Swarm Intelligence: a Flocking Project With Group Object Recognition. In 2023 3rd International Conference on Consumer Electronics and Computer Engineering (ICCECE) (pp. 834 - 837). IEEE.

[11] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

[12] Hu, H., Li, S., Huang, J., Liu, B., & Che, C. (2023). Casting Product Image Data for Quality Inspection with Xception and Data Augmentation. Journal of Theory and Practice of Engineering Science, 3(10), 42-46.

[13] Che, C., Liu, B., Li, S., Huang, J., & Hu, H. (2023). Deep Learning for Precise Robot Position Prediction in Logistics. Journal of Theory and Practice of Engineering Science, 3(10), 36-41.

[14] Chen, S., Kong, N., Sun, X., Meng, H., & Li, M. (2019). Claims data-driven modeling of hospital time-to-readmission risk with latent heterogeneity. Health care management science, 22, 156-179.

[15] Wu, J., Tao, R., Zhao, P., Martin, N. F., & Hovakimyan, N. (2022). Optimizing nitrogen management with deep reinforcement learning and crop simulations. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 1712-1720).

[16] Shen, Y., Wang, B., Deng, S., Zhai, L., Gu, H. M., Alabi, A., ... & Zhang, D. W. (2020). Surf4 regulates expression of proprotein convertase subtilisin/kexin type 9 (PCSK9) but is not required for PCSK9 secretion in cultured human hepatocytes. Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids, 1865(2), 158555.