Coronary Artery Disease Risk Factor Summarization with Longitudinal Electronic Health Record Data

Todd Hartman, Yunzhou Liu, Ziyang Xu, Michelle Nguyen, Adam Charles
Johns Hopkins University, Department of Biomedical Engineering

Abstract

Motivation : Electronic health records (EHR) are widely used by many hospitals to store and organize patient information; however, the crucial information is usually buried within the extensive descriptive text. To fully exploit the utility of EHR, natural language processing (NLP) may aid doctors to summarize the patient history and status.

Methods : Given EHR annotated with coronary artery disease (CAD) risk factors, data was cleaned to unify the structure of every EHR. Three models: rule-based, deep learning and traditional machine learning method were compared for their performance then Naive Bayes algorithm and rule-based algorithm are combined and implemented to group each word in text into categories. Specifically, rule-based algorithm focused on family history while Naive Bayes is applied to the rest of the categories.

Fig.1 General method for modeling and user interface

Results : The evaluation is done on a document level to reflect whether a patient shows any signs of this risk factor. The weighted F1 score of the combined model is 0.916. The result is summarized in a user interface.

Methods

Fig.2 Data preprocessing pipeline

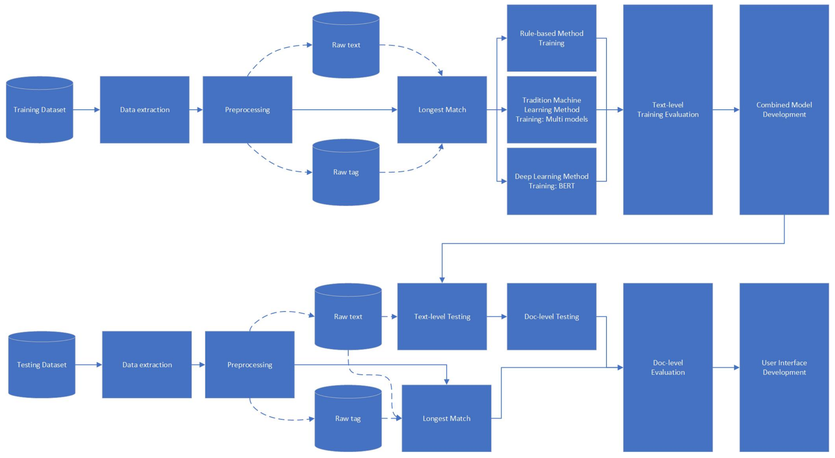

Fig.3 Detailed pipeline for modeling and user interface

The coronary artery disease (CAD) risk factors recognition model and summarization tool aims to track patients' risk factors over time and proposes a system for clinicians to manage patient care with reduced burden.

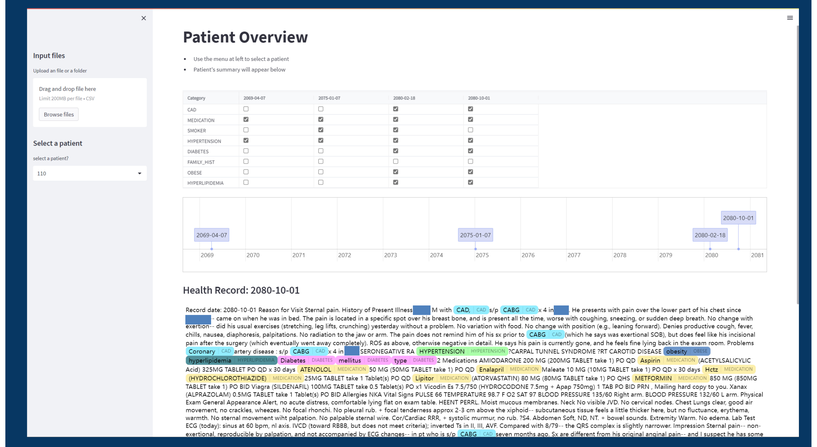

Fig.4 User Interface of
Risk Factor Summarization System *Ver 1.1*

Evaluation

Performance of three different methods on text-level were tested. Deep learning method is less suitable due to the limited data source and was not used. Among classic machine learning models, Naïve Bayes shows better performance and rule-based method is applied to category with few instances.

Machine learning / Deep learning / Rule-based									
type	precision			recall			f1		
Micro ave	**0.8990**	0.6689	0.1941	**0.8990**	0.6470	0.777	**0.8990**	0.6578	0.3106
Macro ave	**0.8832**	0.3872	0.2272	**0.7909**	0.2935	0.7706	**0.8197**	0.3239	0.3509
Weighted ave	**0.8996**	0.6487	NA	**0.8990**	0.6470	NA	**0.8973**	0.6439	NA

Fig.5 Evaluation and Comparison of different models

The final model evaluation is done on the document-level to focus on patient-level statistics. Medication has the best performance due to its unique word bank. CAD is less ideal perhaps due to original tag quality. However, the performance of each category has well supported the document-level risk factor prediction based on patient's health record.

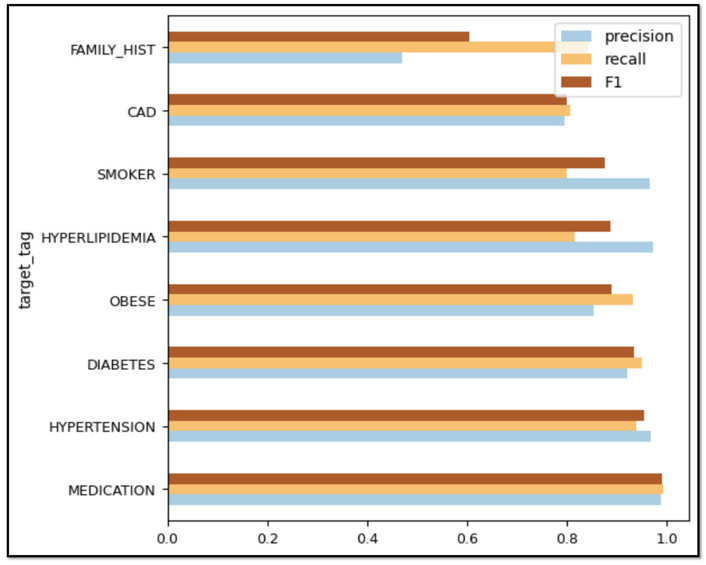

Fig.6 Evaluation of doc-level risk factors

Conclusion

The performance of risk factor prediction is improved by using a multi-model approach. The development of document-level model and user interface may help summarize patients' risk factors over time and help clinicians manage patient care with reduced burden.
In future work, we would aim to improve performance of identifying these risk factors by incorporating additional rules. We would also want to get clinician input on the utility of this tool and design feedback.