Computing Performance Analysis on Clinical Document-level Classification

Wei-Hung Weng, MD[1,2], Kavishwar B. Wagholikar, MBBS, PhD[2,3], Henry C. Chueh, MD[2,3]
[1]Department of Biomedical Informatics, Harvard Medical School, Boston, MA;
[2]Laboratory of Computer Science, Massachusetts General Hospital, Boston, MA;
[3]Department of Medicine, Massachusetts General Hospital, Boston, MA.

Effective computing is critical for developing machine learning-based solutions. In particular, machine learning tasks related to the clinical decision often require real-time modeling and evaluation to provide timely treatment and management. Machine learning based-natural language processing (NLP) techniques allow researchers to construct automated document classifiers to categorize clinical documents based on their content and document-level information without manual processing. Nevertheless, the NLP approach usually creates a higher dimensional feature space, which takes an abundant time to train the classifier if the training corpus is large. There is also the lack of studies on performance analysis on clinical document classification tasks.

We performed document classification tasks based on the medical subdomain of the clinical document using two datasets -- 431 publicly available clinical notes and reports from the Integrating Data for Analysis, Anonymization, and Sharing (iDASH) data repository, as well as 91,237 Massachusetts General Hospital (MGH) clinical notes from the Research Patient Data Registry (RPDR) data repository of Partners HealthCare system. For each dataset, we applied seven clinical feature representation methods and five supervised learning algorithms to construct 35 classifiers. For all classifiers with different size of feature spaces (Table) and algorithms, we evaluated the model performance using performance score (PS), which was defined by $PS = sigmoid(standardized\ F_1\ score) \times [1 - sigmoid(standardized\ processing\ time)]$. The standardization and sigmoid transformation were applied to F_1 score (classifier performance) and processing time (machine performance) to ensure that scales of F_1 score and processing time are consistent and comparable. For all classification tasks, we used the virtual machines on the Partners ERISOne Linux Cluster, with four CPU cores (Intel Xeon 3.47GHz) and 16GB RAM memory.

Table. Dimension of the Feature Space

Feature Representation Method	iDASH	MGH
Bag-of-words (Vocabulary size)	10150	160097
UMLS concepts	4750	25456
UMLS concepts (5 semantic groups)	4531	24457
UMLS concepts (15 semantic types)	3634	18520
Bag-of-words + UMLS concepts	14900	185553
Bag-of-words + UMLS concepts (5 semantic groups)	14681	184554
Bag-of-words + UMLS concepts (15 semantic types)	13784	161949

We found that using the classifiers trained by support vector machine (SVM) with linear kernel yielded the most effective computing performance across two datasets. The appropriate feature representation method, such as using medical concepts with semantic restriction, can additionally augment the performance. We found that the clinical document classification tasks based on the medical subdomain of the document could be done within the acceptable effectiveness, regarding model performance and speed in the scale of ninety thousand documents. We plan to execute and evaluate the same task on a larger analytical platform for scaling up.

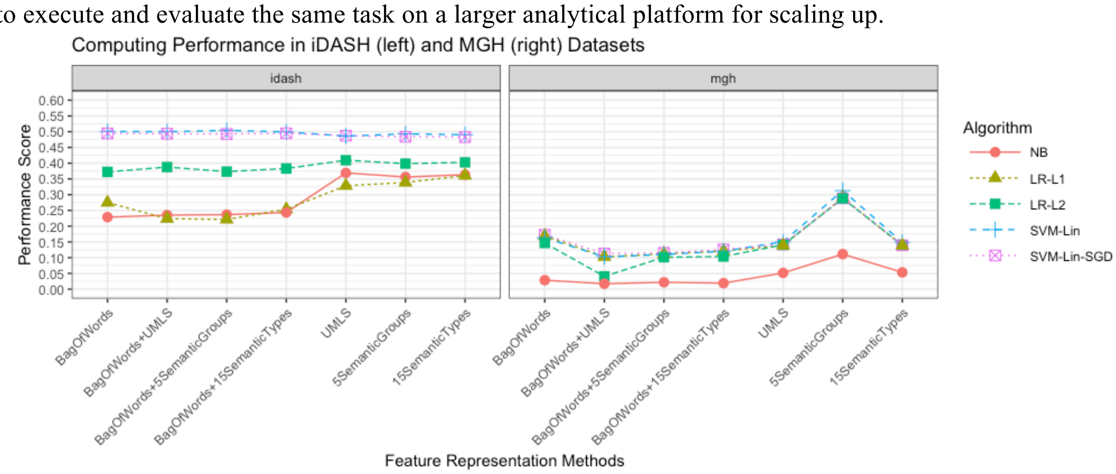

Figure. Comparison of Computing Performance Score within Two Clinical Datasets. Abbreviation: NB: Naïve Bayes, LR-L1: logistic regression with L1 penalization, L2-L2: logistic regression with L2 penalization, SVM-Lin: linear support vector machine, SVM-Lin-SGD: linear support vector machine with stochastic gradient descent