Integrating Structured and Unstructured Patient Data for ICD9 Disease Code Group Prediction

Akshara P*†, Shidharth S*, Gokul S Krishnan, Sowmya Kamath S
Healthcare Analytics & Language Engineering (HALE) Lab, Department of Information Technology,
National Institute of Technology Karnataka, Surathkal, India

ABSTRACT

The large-scale availability of healthcare data provides significant opportunities for development of advanced Clinical Decision Support Systems that can enhance patient care. One such essential application is automated ICD-9 diagnosis group prediction, useful for a variety of healthcare delivery related tasks including documenting, billing and insurance claims. Past attempts considered patients' multivariate lab events data and clinical text notes independently. To the best of our knowledge, ours is the first attempt to investigate the efficacy of integration of both these aspects for this task. Experiments on MIMIC-III dataset showed promising results.

CCS CONCEPTS

• **Applied computing → Health care information systems**.

ACM Reference Format:
Akshara P*†, Shidharth S*, Gokul S Krishnan, Sowmya Kamath S. 2021. Integrating Structured and Unstructured Patient Data for ICD9 Disease Code Group Prediction. In *8th ACM IKDD CODS and 26th COMAD (CODS COMAD 2021), January 2–4, 2021, Bangalore, India.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3430984.3431060

1 INTRODUCTION

Automating International Classification of Diseases, version 9 (ICD-9) code assignment is an important task that can improve healthcare systems by reducing common errors that occur due to manual coding. The large number of distinct combinations of the diagnostic codes and their highly skewed distribution make it a challenging task and therefore have not been effective. However, ICD-9 code group prediction (ICD codes grouped and treated as a multi-task prediction) has been explored recently [1, 2, 5]. In this paper, we adopt a similar approach, by grouping ICD-9 codes into standard categories and using an ensemble model for the prediction task.

2 METHODOLOGY

We used the MIMIC-III dataset [3], extracted patient data and grouped the ICD-9 codes into 20 distinct groups, while discarding records with missing admission ID. A structured dataset consisting of length of stay, gender, age and 480 lab tests results was passed through an ensemble of OneVsRestClassifier and LightGBM classifier [4]. Also, the unstructured radiology notes of the patients in the generated cohort were used to generate word embeddings using a 1000 dimension Word2Vec-CBOW [6] architecture. The embeddings generated were averaged and processed using the ensemble of OneVsRestClassifier and CatBoost classifier [7]. We added class-weights to handle the high class imbalance – if a, b are the class labels and P_a, P_b are their occurrences, then the respective class weights of a, b are given by $\{P_b/(P_b + P_a), P_a/(P_b + P_a)\}$. Finally, the weights of ensembling was heuristically obtained as 0.65 & 0.35 for LightGBM and Catboost.

* Equal Contribution; † Corresponding Author e-mail: akshblr555@gmail.com .

CODS COMAD 2021, January 2–4, 2021, Bangalore, India

ACM ISBN 978-1-4503-8817-7/21/01.
https://doi.org/10.1145/3430984.3431060

3 EXPERIMENTAL RESULTS

In Table 1, the performance observed with reference to the different ensemble models experimented with are tabulated. We used the metrics Area under ROC curve (AUC) and hamming loss to evaluate the models. The proposed ensemble model achieved the best AUC as well as the lowest hamming loss. Our work serves as the benchmark in terms of larger ICD-9 code coverage (obtained from lab events + unstructured radiology notes).

Table 1: Observed ICD-9 Code Prediction Performance

Proposed Models	AUC	Hamming Loss
OneVsRestClassifier+LGBMClassifier+sample-weights	0.672	0.169
Word2Vec+OneVsRestClassifier+CatBoostClassifier	0.667	0.181
Ensemble Model	**0.748**	**0.154**

4 CONCLUSION & FUTURE WORK

In this paper, we explore the integration of structured and text patient data for ICD-9 group prediction. We aim to extend our work by adopting better feature extraction techniques to enhance the prediction performance.

REFERENCES

[1] Tushaar Gangavarapu, Aditya Jayasimha, Gokul S Krishnan, and Sowmya Kamath. 2020. Predicting ICD-9 code groups with fuzzy similarity based supervised multi-label classification of unstructured clinical nursing notes. *Knowledge-Based Systems* 190 (2020), 105321.
[2] T. Gangavarapu, G. S Krishnan, S. Kamath S, and J. Jeganathan. 2020. FarSight: Long-Term Disease Prediction Using Unstructured Clinical Nursing Notes. *IEEE Transactions on Emerging Topics in Computing* (2020), 1–1. https://doi.org/10.1109/TETC.2020.2975251
[3] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data* 3 (2016), 160035.
[4] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 3149–3157.
[5] Gokul S Krishnan and Sowmya Kamath S. 2019. Ontology-driven Text Feature Modeling for Disease Prediction using Unstructured Radiological Notes. *Computación y Sistemas* 23, 3 (2019).
[6] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.
[7] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. 2017. CatBoost: unbiased boosting with categorical features. arXiv:cs.LG/1706.09516