Exploring Embedding Representations for Structured Data in the OMOP CDM

Use-Case for Long Hospitalization Prediction

Jorge Cerejo[1], Bernardo Neves[1,2], Simão Gonçalves[1], José Maria Moreira[1], Nuno A. da Silva[1], Francisca Leite[1]

[1]Hospital da Luz Learning Health, Luz Saúde, Lisboa
[2]Instituto Superior Técnico, Universidade de Lisboa

Introduction

Word embedding techniques have gained influence in the field of medical data analysis, as these techniques enable the representation of medical concepts [1,2].

This study explores the utilization of word embeddings to represent medical codes extracted from a Portuguese tertiary hospital's OMOP CDM database. By representing medical codes into a continuous vector space, we aim to capture inherent relationships between healthcare concepts, potentially enhancing predictions of long hospitalizations (>7days). Predicting the length of stay early can help in better resource allocation and patient care.

Methods

The real-world dataset used encompasses approximately 52M events across 6.5M visits and 500K patients, with an observation period spanning between 2007 and 2023.

We employed the word2vec algorithm to generate medical code embeddings, exploring diverse embedding spaces by varying parameters such as vector and window sizes [3]. Additionally, we expanded our vocabulary using ontology expansion [4].

We use the embeddings as features of machine learning (ML) models for predicting long hospitalization, comparing their performance with count-based features.

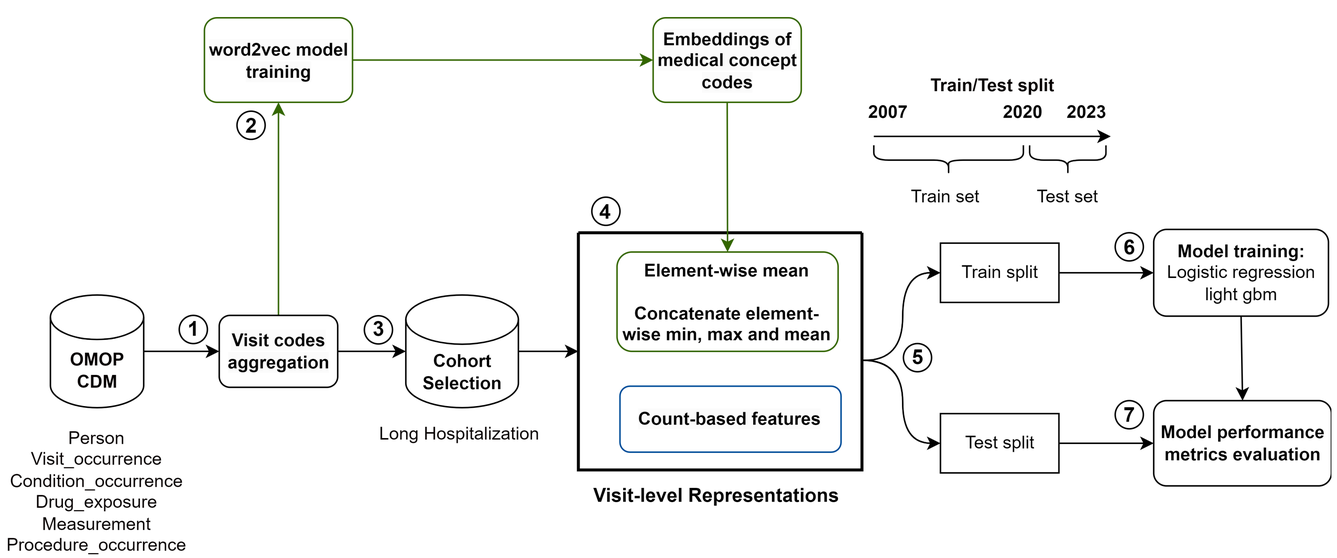
Results

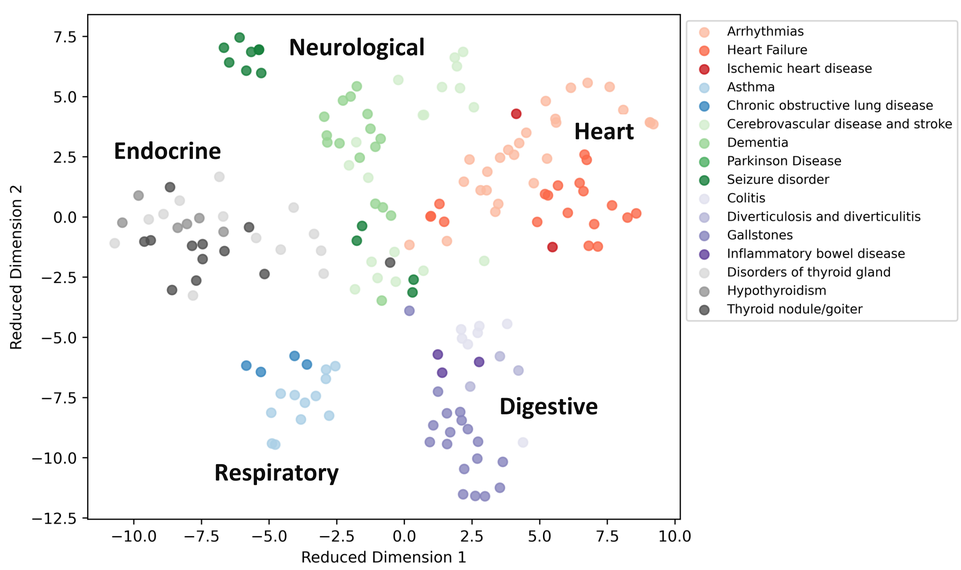
- Ontology expansion significantly increases the event counts of all patients; however, it does not improve predictive performance.
- Smaller embedding dimensions yield suboptimal performance, while dimensions exceeding 100 do not lead to significant improvements.
- Increasing window sizes fails to enhance model performance.
- Concatenation of element-wise min, max and mean representations was show to be the best approach for visit embedding representations.

Conclusions

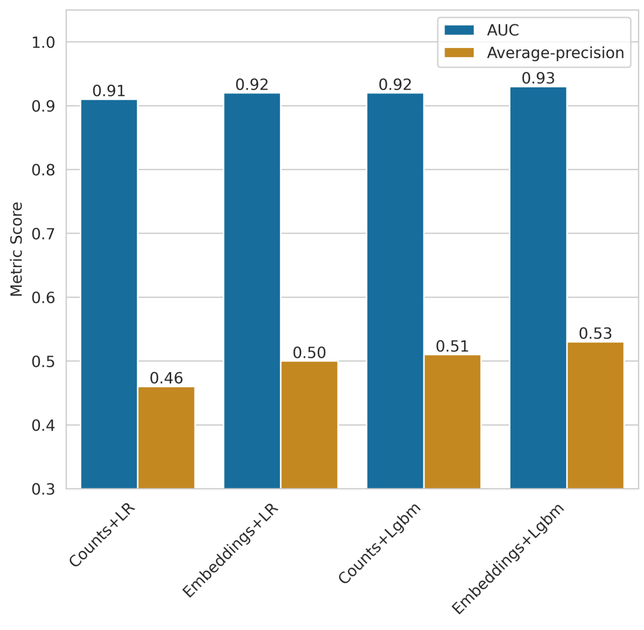
- Our study demonstrates the feasibility of creating embedding representations from structured data in the OMOP CDM.
- Ontology expansion resulted in inferior performance, introducing noise rather than enhancing contextual understanding.
- Explore advanced models with attention mechanisms for ontology expansion [5].
- Incorporate medical code descriptions in embedding creation to improve contextual understanding and code relationships [6].

Embedding representations demonstrate slightly **superior performance** over **count-based features**, while **capturing** data **relationships** and **dependencies.**





Ontology Expansion?		No					No				
Visit representation		Element-wise mean					Concatenation of Element-wise min, max and mean				
Window_size		5	10	20	30	50	5	10	20	30	50
Vector_size	10	0,911	0,912	0,909	0,913	0,911	0,918	0,921	0,918	0,922	0,918
	50	0,918	0,917	0,918	0,918	0,917	0,925	0,925	0,926	0,924	0,925
	100	0,921	0,919	0,917	0,919	0,917	**0,926**	0,925	0,924	0,926	0,924
	300	0,920	0,921	0,920	0,921	0,920	0,926	0,926	0,925	0,926	0,925
	500	0,921	0,921	0,920	0,921	0,920	0,926	0,926	0,925	0,925	0,926
	1000	0,922	0,921	0,920	0,921	0,920	0,925	0,926	0,926	0,926	0,925
Ontology Expansion?		Yes					Yes				
Visit representation		Element-wise mean					Concatenation of Element-wise min, max and mean				
Window_size		5	10	20	30	50	5	10	20	30	50
Vector_size	10	0,871	0,873	0,875	0,878	0,876	0,896	0,897	0,904	0,907	0,907
	50	0,901	0,900	0,902	0,903	0,902	0,911	0,910	0,912	0,915	0,916
	100	0,905	0,906	0,906	0,908	0,907	0,911	0,914	0,918	0,918	0,918
	300	0,911	0,910	0,911	0,911	0,911	0,921	0,920	0,912	0,919	0,920
	500	0,912	0,913	0,913	0,912	0,912	0,923	0,922	0,922	0,922	0,920
	1000	0,914	0,913	0,913	0,913	0,913	0,922	0,923	0,922	0,923	0,921



References

[1] Bajor et al. Embedding Complexity In the Data Representation Instead of In the Model: A Case Study Using Heterogeneous Medical Data. 2018.
[2] Beaney et al. Comparing natural language processing representations of disease sequences for prediction in the electronic healthcare record. Health Informatics; 2023
[3] Mikolov et al. Efficient Estimation of Word Representations in Vector Space. 2013.
[4] Choi et al. GRAM: Graph-based Attention Model for Healthcare Representation Learning. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Halifax NS Canada: ACM; 2017
[5] Hur et al. Unifying Heterogeneous Electronic Health Records Systems via Text-Based Code Embedding. 2021.
[6] Rasmy et al. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. Npj Digit Med. 2021

HOSPITAL DA LUZ LEARNING HEALTH
TRAINING, RESEARCH & INNOVATION CENTER

OHDSI