A Case Study for Predicting in-Hospital Mortality by Utilizing the Hyperbolic Embedding of ICD-9 Medical Ontology

Jiazhen Cao

Department of Computer and Information Science

University of Oregon

Supervisor

Dejing Dou, Ph.D.

In fulfillment of the requirement for the departmental honors of the degree of

Bachelor of Science in Computer Information Science

June 14, 2019

Acknowledgements

I dedicate this thesis to my eximious supervisor Dr. Dejing Dou, Professor Thien Huu Nguyen, my research leader Ph.D. students Qiuhao Lu and Nisansa de Silva, my parents, and my girlfriend who have physically and mentally supported my work.

Abstract

In-hospital mortality prediction is significant for evaluating a patient's severity of illness ahead of the time. The outcome of the evaluation can help physicians to identify which patient is at risk and needs immediate care, it can further increase the efficiency of use of medical resources. In this study, I proposed a method that is similar with the one in our Electronic Health Records (EHRs) research at the CBL Lab and utilized the hyperbolic embedding of ICD-9 medical ontology for the prediction model. The results outperformed the benchmark prediction model and demonstrated that the hyperbolic embedding on ICD-9 is more effective than other graph embedding methods.

Contents

Chapter 1

Introduction

The Intensive Care Unit (ICU) is a place with a high staff to patient ratio that admits people who are severely ill. According to RL et al. [17], the purpose of such a high ratio is for monitoring patients continuously to detect deterioration and correct it if possible before it becomes fatal. Hence, the usage of resources in ICU is expensive and considerable. Predicting patients' in-hospital morality can give valuable information for evaluating the severity of illness. Furthermore, the prediction can help figure out which patient needs immediate attention.

Traditional prediction models are scoring systems [21, 20, 13, 12, 2]. Depends on the system, scores are calculated with variables which are selected by physicians and researchers. Recent models are usually machine learning based [10, 5, 7]. Machine learning approaches demonstrated great effectiveness with improved prediction performance. These prediction models extract information from patients' Electronic Health Records (EHRs) as inputs. EHRs contain important medical features including the diagnosis codes. It's common that current machine learning based approaches utilize only numerical and categorical data, while it's rare that prediction models exploit textual or diagnostic data. I make use of ontological medical concepts that existed on the EHRs to explore and evaluate their effectiveness.

Medical ontology is a model that defines and represents concepts and relations of a certain domain. I mainly focused on the ICD-9 medical ontology. ICD-9 is the 9th revision of the International Statistical Classification of Diseases and Related Health Problems. In a nutshell, the ICD-9 represents diagnoses and their hierar-

chical relationships. One of the reason that I focus on ICD-9 medical ontology is I conduct the experiments on the MIMIC-III data set [11], which takes ICD-9 as the ontology for coding the diagnoses. MIMIC-III is a large data set of information of patients in ICU. Data includes significant signs, notes, observations, code, etc.

The way the diagnoses are used is by representing them in a certain form, and this is what we call representation learning methods. In general, an appropriate data representation brings its related model good performance [3]. Although there are countless methods to represent diagnostic data, using embedding algorithms is more reasonable. The ICD-9 medical ontology is large and each diagnostic code in it can be considered as a distinct feature, so using traditional encoding methods such as ordinal encoding and one hot encoding may cause the final transformed data have too many dimensions. When it comes to embedding method, for textual data, Euclidean space is the common geometry to be used, however, for medical ontology which has explicit hierarchies, through literature reviews, hyperbolic embedding method [15] is a better fit as it can represent the relationships with lower dimensions.

Inspired by the method [9] which add embeddings of note or entities into the benchmark model, I propose a method to exploit the diagnostic information in the EHRs of ICU patients. To evaluate the embedding method I choose, I further compare the performance of the method with different graph embedding methods. The result demonstrates better AUROC (The Area Under Receiver Operating Characteristic Curve) than benchmark with only vital signs, and it supports the claim the hyperbolic embedding on medical ontology is more effective.

Chapter 2

Related Work

2.1   In-hospital Mortality Prediction

The characteristics of the ICU make it a data rich place. Early studies on mortality prediction conducted scoring system development approach by either experts' knowledge driven or data driven. For example, the earliest system APACHE [21] was built based on rules from physicians. Later systems such as the SAPS [13], the SAPS [12], the APS-III [22], and the OASIS [2] are data driven approaches. These scoring systems have metrics for variables they included and calculate a final score based on each score of each variable. The scoring systems have a common limitation that they are biased that they cannot be applied universally and still be precise on an individual at the same time. [1]

Studies had changed directions away from building models, however, With the development and the increasing popularity of machine learning recently. Many successful machine learning based prediction models have been developed today. For instances, Johnson et al. [10] did several approaches by using Logistic Regression, Support Vector Machine (SVM), and Random Forest. Brand et al. [5] developed their Convolutional Neural Network (CNN) based model that was inspired by Suresh et al. [18]. The benchmark [7] I used compared machine learning based models: Logistic Regression and Long Short-Term Memory (LSTM), with traditional scoring systems: SAPS, APS-III, OASIS, and SAPS-II. Although machine learning algorithms often have considerable performance on in-hospital mortality predication, they have a common limitation that only vital signs are used in the

models. Hence, I leverage the diagnostic codes of ICD-9 to discover whether it can improve the performance of the predication model.
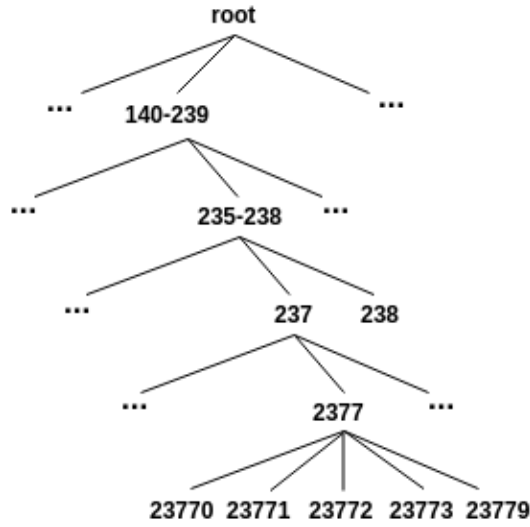
(a) Hierarchical Relationships of a Subtree in ICD-9
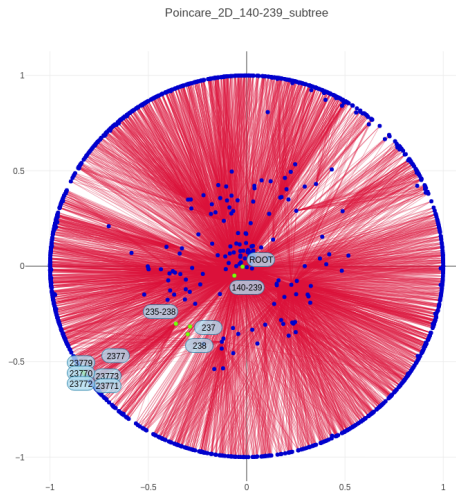
(b) 2-D Hyperbolic Embedding of the Subtree

Figure 2.1 Hierarchical Structure and Visualization of the 2-D embedding of a Subtree of ICD-9

2.2 Hyperbolic Representation and ICD-9

According to Choi et al. [6], proper representation learning has advanced applications in health care analytic. Jin et al. [9] provides a great example that by outperforming the benchmark on standard LSTM model with either embedding

of note or embedding of entities added to the dense layer. It's clear that representation learning has advanced the deep learning usage in the medical domain.

Based on the results of Nickel and Kiela [15], the representation of hierarchies in Euclidean space can results the data into much higher dimensions than the representation in hyperbolic space. The quality of hyperbolic embedding in lower dimensions has been further proved by Sa et al. [16].

As I mentioned before, the major reason for using hyperbolic embedding is because ICD-9 ontology as an explicit hierarchical structure. Here is an example of a subtree of code 140-239 in ICD-9, also see Figure 2.1a.

140-239 Neoplasms:

- 140-149 Malignant Neoplasm Of Lip, Oral Cavity, And Pharynx

- 150-159 Malignant Neoplasm Of Digestive Organs And Peritoneum

- ...

- 235-238 Neoplasms Of Uncertain Behavior

    - 235 Neoplasm of uncertain behavior of digestive and respiratory systems

    - 236 Neoplasm of uncertain behavior of genitourinary organs

    - 237 Neoplasm of uncertain behavior of endocrine glands and nervous system

        * 2370 Neoplasm of uncertain behavior of pituitary gland and craniopharyngeal duct

        * ...

        * 2377 Neurofibromatosis

            · 23770 Neurofibromatosis, unspecified

            · 23771 Neurofibromatosis, type 1 [von recklinghausen's disease]

            · ...

        * ...

    - 238 Neoplasm of uncertain behavior of other and unspecified sites and tissues

- 239-239 Neoplasms Of Unspecified Nature

There is no doubt that this medical ontology is in a hierarchical structure. To show that hyperbolic embedding does represent this structure clearly, I trained a 2-D hyperbolic embedding and plot its visualization, see Figure 2.1b. The root node represent the ontology itself. The node '140-239' is only a little away from the root since coded 140-239 are also very abstract. Nodes '236-238', '237', and '238' are almost on the same level and they are at the middle of the radius. It makes sense because these codes are at the middle of the hierarchical structure. The position of these three nodes may be confusing because from the example above, '237' seems have to be lower than '236-238', however, because the hyperbolic embedding not only learn the relationship between '237' and '236-238', but also consider the relationships between '237' and all the other parent or child nodes, it's acceptable that node '237' is a little bit closer to '140-239' than '236-238'. The most excellent point of this visualization is that all the leaves (e.g. codes 23770-23779) are on the edge of the circle, which means they are at the bottom of the hierarchy.

2.3   MIMIC-III

I conduct the experiment on the MIMIC-III data set [11]. MIMIC-III (Medical Information Mart for Intensive Care III) is a large database comprising about 60000 de-identified admitted ICU patients' data, including demographics, vital signs, physician notes, etc. The benchmark I use also conduct their approaches on this data.

Chapter 3

Methods

3.1 Poincaré Embedding

To be specific, the hyperbolic embedding I use is the Poincaré embedding proposed by Nickel and Kiela [15]. In hyperbolic space of the Poincaré, the distance between two data points is defined as:

$$d(u, v) = \text{arcCosh} \left(1 + 2 \frac{||u - v||^2}{(1 - ||u||^2)(1 - ||v||)^2)} \right) \tag{3.1}$$

Because the Poincaré ball model has a Riemannian manifold structure, similar to what our CBL research conduct, I follow Nickel and Kiela's work to minimize the loss:

$$L = \sum_{(u,v) \in S} \log \frac{\exp^{-d(u,v)}}{\sum_{v' \in N(u)} \exp^{-d(u,v')}} \tag{3.2}$$

by using the RSGD [8] method.

3.2 Mortality Benchmark

I used the benchmark [7] for the extrinsic experiment for evaluating the performance of the embeddings. This benchmark has achieved considerable performance in in-hospital mortality prediction and it has been widely used as a baseline in literature, for example, Jin et al. [9] implemented and compared their methods with models in this benchmark as baselines. I used the standard model (LSTM)

for the experiments. The model is a long short-term memory neural network that takes structured data. The benchmark input cohort consists of 17 features (e.g., Diastolic blood pressure, Glascow coma scale) that are usually considered as important factors in medical fields. Each subject (patients in MIMIC-III are referred to subject) has a series of signals along time stamps. For each time series, the experiment discretizes the signals every 2 hours in a 48 hours window.
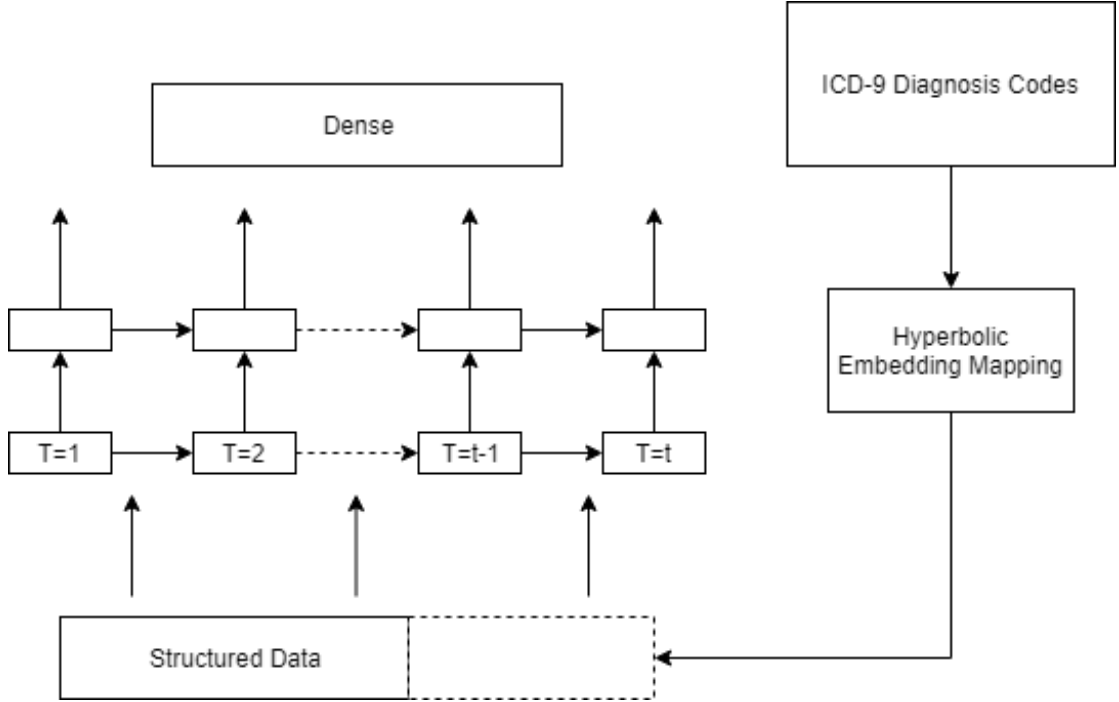

Figure 3.1 Framework

3.3   Utilizing Diagnoses on EHRs

I followed the data processing pipeline provided by the benchmark, to extract all the diagnosis codes from the MIMIC-III data set. All the codes are then mapped to their corresponding embeddings. Since ICU patients always have more than one diagnosis, I took the average of all the diagnoses to generate a single input vector. According to Figure 3.1, the generated vector would be concatenated to the structured input that contains the 17 signals. Eventually, the new input were fed into the standard LSTM model.

3.3.1 Implementation

The implementation follows the steps outlined in Figure 3.1 with one more step beforehand, embed the ICD-9 medical ontology using Poincaré model.

- First, I wrote a script to build a search tree based on the ICD-9 ontology with abstract codes at higher level and detailed nodes at lower level.

- I then constructed an edge list with tuples comprising the relationships between a node with all the other nodes in its branch.

- I implemented the Poincaré model using PyTorch.

- I compared my own implementation with the Poincaré model of an existing python package, Gensim, and chose to use Gensim to generate the embedding.

- By following the benchmark data processing pipeline, training and validation sets of time stamp based ICU stays have been generated.

- Map diagnostic codes to each subject (patient) in the data set.

- Map diagnostic codes of each individuals to the corresponding embedding, since all the subjects have more than one diagnosis code, I form an array of embeddings, take the average of all the embeddings in the array to generate a single vector.

- Finally, I concatenated the corresponding vector to the normalized vital signs for every ICU stay and train the benchmark LSTM model.

Note that the reason I used Gensim package is that the comparison turns out my own implementation is not as good as the public package, there are many possible reasons, for example, the choice of hyper-parameters. Although eventually I used the Gensim package for the experiment, I had a better understanding of this creative representation method.

Chapter 4

Evaluation

In this chapter, I evaluate the method, the Poincaré embedding, and other graph embeddings by checking if embeddings can improve the performance of the benchmark and if Poincaré embedding can outperform other embedding methods.

Embedding	Acc	Pre-0	Pre-1	Re-0	Re-1	**A.R**	A.P
Poincaré	0.8901	0.9153	0.6395	0.9620	0.4312	**0.8769**	0.5914
ComplEx	0.8883	0.9081	0.6520	0.9688	0.3739	0.8720	0.5662
Distmult	0.8892	0.9099	0.6525	0.9677	0.3876	0.8735	0.5777
TransE	0.8880	0.9054	0.6623	0.9720	0.3509	0.8731	0.5799
Rescal	0.8917	0.9129	0.6605	0.9670	0.4106	0.8702	0.5840
Harutyunyan et al.	0.8839	0.9047	0.6281	0.9677	0.3486	0.8607	0.5665

Table 4.1: Prediction Performance of My Method with Different Embeddings

4.1 Setup

This experiment is conducted on MIMIC-III data set [11], I adopted the benchmark data extraction pipeline [7] to generate 42276 ICU stays of 22789 unique de-identified subjects whose age are greater than or equal to 18. The data set has been split to 85% as training set, 15% as validation set. The Poincaré embedding of the ICD-9 embedding was prepared beforehand, see section 3.3.1. Learning rates are all 0.01 and batch sizes are all 32.

4.2   Experiment

I experimented with the Poincaré embedding and several state-of-the-art graph embeddings, ComplEx [19], Distmult [23], TransE [4], and Rescal [14]. All the embeddings have 10 dimensions. The performance of all the models are shown in Table 4.1. In in-hospital mortality prediction, literature demonstrates the AUROC (Area Under the Receiver Operating Characteristics Curve) is the major metric for evaluation. Higher AUROC means better performance. Table 4.1 not only includes the AUROC, but also other metrics for fair comparison.

For all the embeddings experimented, I adopted the same method I proposed, concatenating the embedding to the input of the model. The performance of the benchmark standard LSTM model is at the bottom in the table as a baseline, and performances of other models are above. It's clear that all the embeddings improved the performance of the benchmark model, which indicates that my method does help the model perform better. Furthermore, the model with Poincaré embedding added achieved the best AUROC, which indicates among all these 10 dimensional embeddings, Poincaré embedding is the most effective one.

I did further experiments with Poincaré of 10 dimensions, 100 dimensions, and 300 dimensions. It was surprising that the performance of the Embedding got worse as the dimensions grew. From Table 4.2, the performance of the 100-d Poincaré embedding is slightly lower than it was in 10-d, however, the overall performance of the model is still at the same level as the 10-d Poincaré embedding. Unfortunately, after I concatenated the 300-d Poincaré into the input and trained the model, the performance dramatically dropped. I first suspect that the reason is overfitting, so I plotted all the training curves for visualization. According to Figure 4.1a, Figure 4.1b, and Figure 4.1c, the model converges faster and overfits much more massively as the dimensions grew. For the 100-d Poincaré, the AUROC started to drop after about 50 epochs but it's fine since it has achieved similar performance to the model with 10-d Poincaé embedding, but for the 300-d Poincaré, the AUROC started to drop very early after about 30 epochs and it never reaches an AUROC above 0.863. Many things in this experiment can cause the overfitting, the number of diagnoses of each patients, the model's architecture, and the choice of hyper-parameters. Further experiments are needed to discover

Embedding	Acc	Pre-0	Pre-1	Re-0	Re-1	**A.R**	A.P
Poincaré-10d	0.8901	0.9153	0.6395	0.9620	0.4312	0.8769	0.5914
Poincaré-100d	0.8904	0.9122	0.6531	0.9663	0.4060	0.8745	0.5745
Poincaré-300d	0.8808	0.9071	0.5956	0.9605	0.3715	0.8628	0.5468

Table 4.2: Prediction Performance of My Method with Different Dimensions of Poincaré Embeddings

the reasons.

Consequently, the design of the experiment has met my expectation. The experiment has evaluated my method extrinsically, the hyperbolic embedding of the ICD-9 is outstanding as the benchmark model with itself added outperformed the benchmark model with all the other graph embeddings added by higher AUROC. Besides, all the other metrics listed in the Table 4.1 demonstrate that the overall performance of the model with Poincaré embedding is also excellent. The results proved the effectiveness of my method and it indirectly proved the importance of the diagnostic codes.
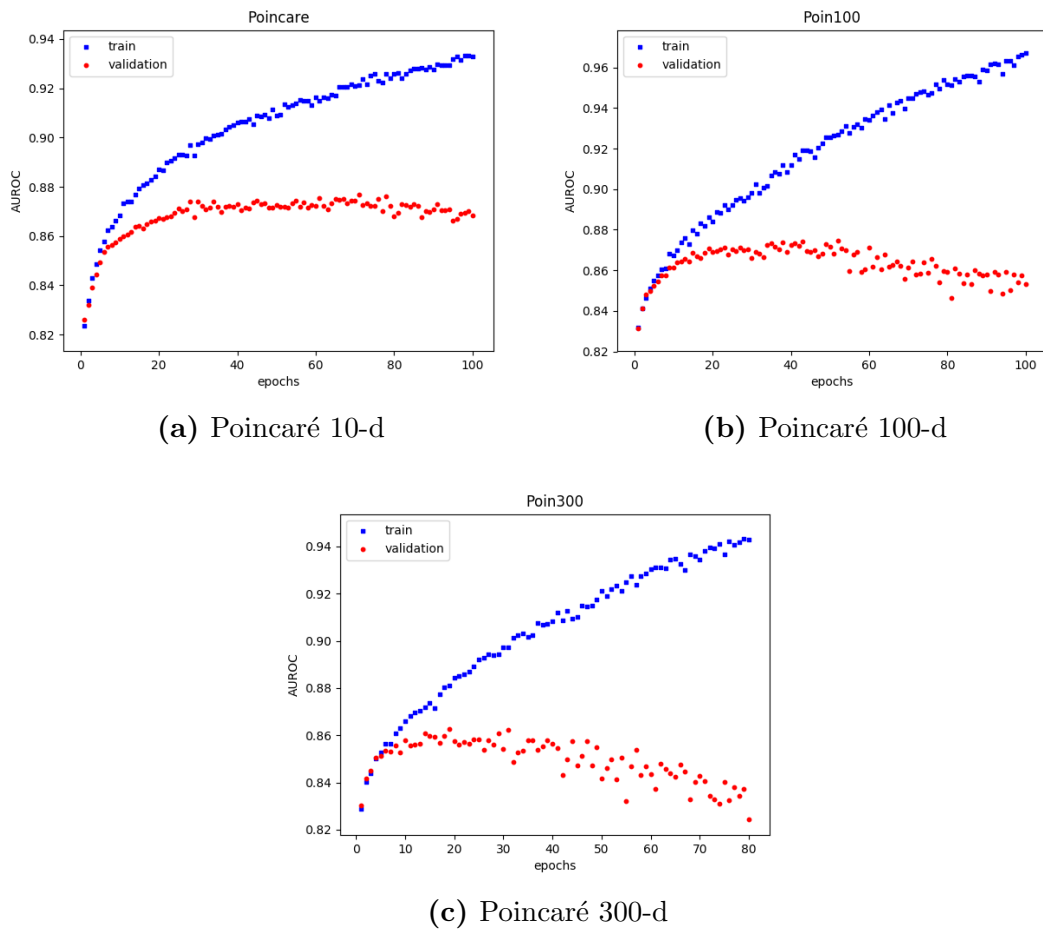
(a) Poincaré 10-d

(b) Poincaré 100-d

(c) Poincaré 300-d

Figure 4.1 The Training Curve of Poincaré Embedding of Different Dimensions

Chapter 5

Future Work

5.1 Overfitting

As my results of the experiments illustrate, the performance of the Poincaré Embedding was getting worse as the dimension grew. From the discussion in Chapter 4, there is no doubt the model with 300-dimension-Poincaré embedding is overfitting. In my opinion, causes may be traced to:

- Sample size: there's only about forty thousand samples in the data set. with a 300-dimension-embedding, it may be easier for such a model with this sample size to overfit.

- Curse of high dimensionality: since in the experiment, I concatenated the embedding to the input vital signs, the 300 dimensional vector can be seen as 300 more features. The large number of features can easily cause the model overfit massively on the training set.

5.2 Lack of Information

Although the Poincaré embedding of the ICD-9 ontology brings noteworthy performance, the embedding I learned is only based on the hierarchical relationships without any descriptions of the diagnosis codes. Hence, I'm seeking for a way to train an embedding with information of an ontology fully utilized.

5.3   Method Improvement

After I discussed the current result of the experiment with Dr.Dou, one thing he pointed out is that my method is different from the method we proposed in our CBL EHR research. The method we proposed is shown below:
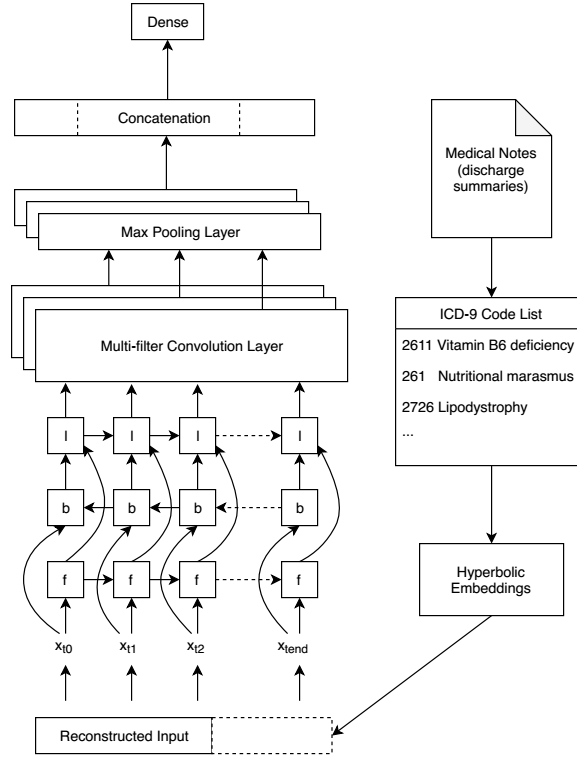

Figure 5.1 Framework of the Method Proposed in the CBL EHR Research

The proposed method extracts latent diagnoses from the textual data in the MIMIC-III data set (i.e. Discharge Note). This method has the potential to outperform the benchmark with higher AUROCs. Since more diagnosis codes can be extracted for each subject (patient), this method may have the potential to further prevent the overfitting of the 300 dimensional Poincaré embedding.

Chapter 6

Conclusions

Conclusions should summarize the problem, the solution and its main innovative features, outlining future work on the topic or application scenarios of the proposed solution.

This study focus on the in-hospital mortality prediction, this task is valuable in the medical domain as it can provide extra information for physicians to evaluate the severity of illness of patients. Since there are already many studies constructed considerable prediction models, I chose a benchmark LSTM model [7] to start with. The data set the benchmark paper and I focused on is the MIMIC-III. This coding ontology used by this data set is the ICD-9 medical ontology, therefore, I trained several hyperbolic embeddings for preparing the experiment. The method I proposed is to make use of the diagnostic codes in the MIMIC-III data set, so I concatenate the trained embedding after the vital signs to generate a new input for the LSTM model. The result of the experiment indicates that the method is effective. I then compared the performance of the LSTM model with Poincaré embedding with other state-of-art graph embeddings. The results showed that the hyperbolic embedding performed better with only 10 dimensions. There were many limitations and errors occurred in this study, the overfitting of the 300 dimensional Poincaré embedding is waiting to be solved, the current embeddings I trained lack of information as only hierarchical structure is used, and the current method is simple as no latent information was extracted from the data set.

References

[1] Knaus William A. "Apache 1978-2001: The Development of a Quality Assurance System Based on Prognosis: Milestones and Personal Reflections." In: *Archives of Surgery* 137.1 (2002), pp. 37–41.

[2] Johnson AE, Kramer AA, and Clifford GD. "A New Severity of Illness Scale Using a Subset of Acute Physiology And Chronic Health Evaluation Data Elements Shows Comparable Predictive Accuracy." In: *Critical Care Medicine* 41.7 (2013), pp. 1711–1718.

[3] Yoshua Bengio, Aaron Courville, and Pascal Vincent. "Representation Learning: A Review and New Perspectives". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 35.8 (Aug. 2013), pp. 1798–1828.

[4] Antoine Bordes et al. "Translating Embeddings for Modeling Multi-relational Data". In: *Advances in neural information processing systems*. 2013, pp. 2787–2795.

[5] Landon Brand et al. "Real Time Mortality Risk Prediction: A Convolutional Neural Network Approach". In: *HEALTHINF*. Jan. 2018, pp. 463–470.

[6] Edward Choi et al. "Multi-layer Representation Learning for Medical Concepts". In: *CoRR* abs/1602.05568 (2016).

[7] Hrayr Harutyunyan et al. "Multitask Learning and Benchmarking with Clinical Time Series Data". In: *CoRR* abs/1703.07771 (2017).

[8] Reshad Hosseini and Suvrit Sra. "An Alternative to EM for Gaussian Mixture Models: Batch and Stochastic Riemannian Optimization". In: *Mathematical Programming* (June 2017), pp. 1–37. ISSN: 1436-4646. DOI: 10.1007/s10107-019-01381-4.

[9] Mengqi Jin et al. "Improving Hospital Mortality Prediction with Medical Named Entities and Multimodal Learning". In: *CoRR* abs/1811.12276 (2018).

[10] Alistair EW Johnson, Andrew A Kramer, and Gari D Clifford. "Data Preprocessing and Mortality Prediction: the Physionet/CinC 2012 Challenge Revisited". In: *Computing in Cardiology Conference (CinC)*. IEEE. 2014, pp. 157–160.

[11] Alistair EW Johnson et al. "MIMIC-III, A Freely Accessible Critical Care Database". In: *Scientific Data* 3 (2016), p. 160035.

[12] Le Gall JR, Lemeshow S, and Saulnier F. "A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study." In: *JAMA* 270.24 (1993), pp. 2957–2963.

[13]   Le Gall JR et al. "A Simplified Acute Physiology Score for ICU Patients." In: *Crit Care Med* 12.11 (1984), pp. 975–977.

[14]   Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. "A Three-Way Model for Collective Learning on Multi-Relational Data." In: *ICML*. Vol. 11. 2011, pp. 809–816.

[15]   Maximillian Nickel and Douwe Kiela. "Poincaré Embeddings for Learning Hierarchical Representations". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., 2017, pp. 6338–6347.

[16]   Christopher De Sa et al. "Representation Tradeoffs for Hyperbolic Embeddings". In: *CoRR* abs/1804.03329 (2018).

[17]   Kane RL amd Shamliyan TA et al. "The Association of Registered Nurse Staffing Levels and Patient Outcomes: Systematic Review and Meta-analysis." In: *Medical Care* 45.12 (2007), pp. 1195–1204.

[18]   Harini Suresh et al. "Clinical Intervention Prediction and Understanding using Deep Networks". In: *CoRR* abs/1705.08498 (2017), pp. 1–16.

[19]   Théo Trouillon et al. "Complex Embeddings for Simple Link Prediction". In: *International Conference on Machine Learning*. 2016, pp. 2071–2080.

[20]   Knaus WA, Draper EA, and Wagner DP. "APACHE II: A Severity of Disease Classification System." In: *Crit Care Med* 13.10 (1985), pp. 818–829.

[21]   Knaus WA et al. "APACHE-acute Physiology and Chronic Health Evaluation: A Physiologically Based Classification System." In: *Crit Care Med* 9.8 (1981), pp. 591–597.

[22]   Knaus WA et al. "The APACHE III Prognostic System. Risk Prediction of Hospital Mortality for Critically Ill Hospitalized Adults." In: *Chest* 100.6 (1991), pp. 1619–1636.

[23]   Bishan Yang et al. "Embedding Entities and Relations for Learning and Inference in Knowledge Bases". In: *Proceedings of the International Conference on Learning Representations (ICLR) 2015*. May 2015, pp. 1–13.
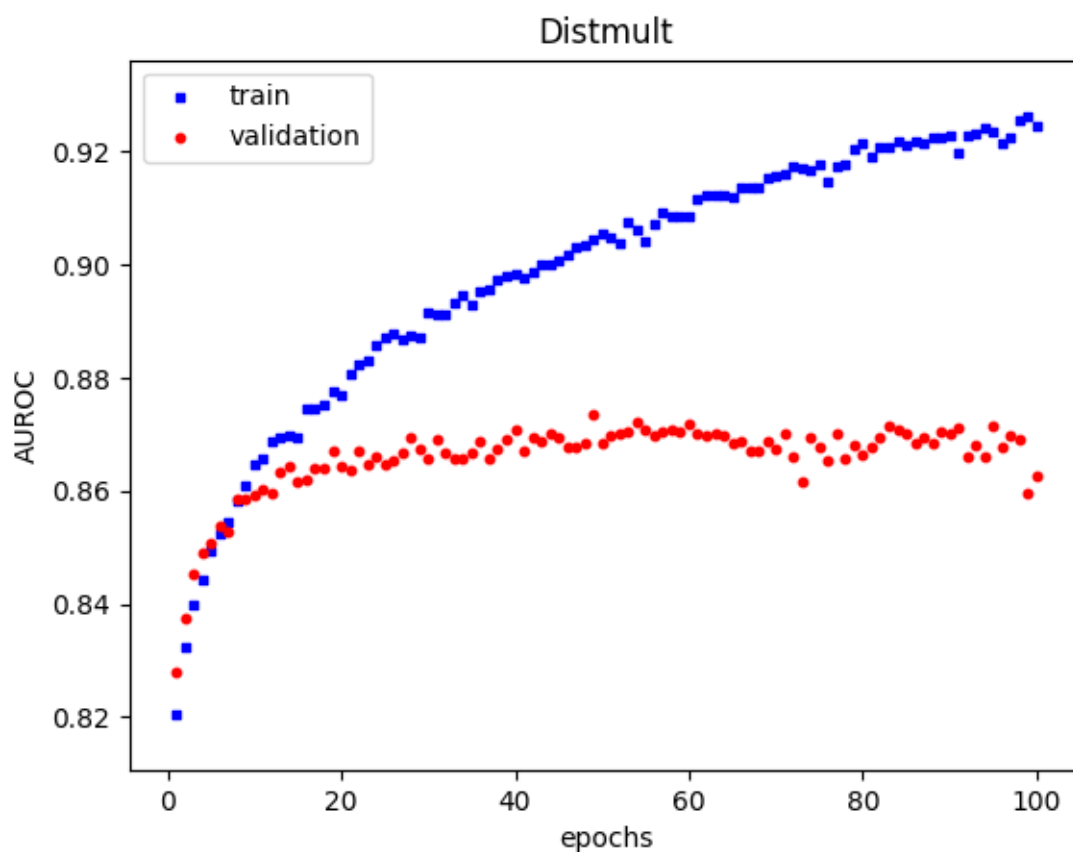
Appendix A

Training Curves of Other Embeddings

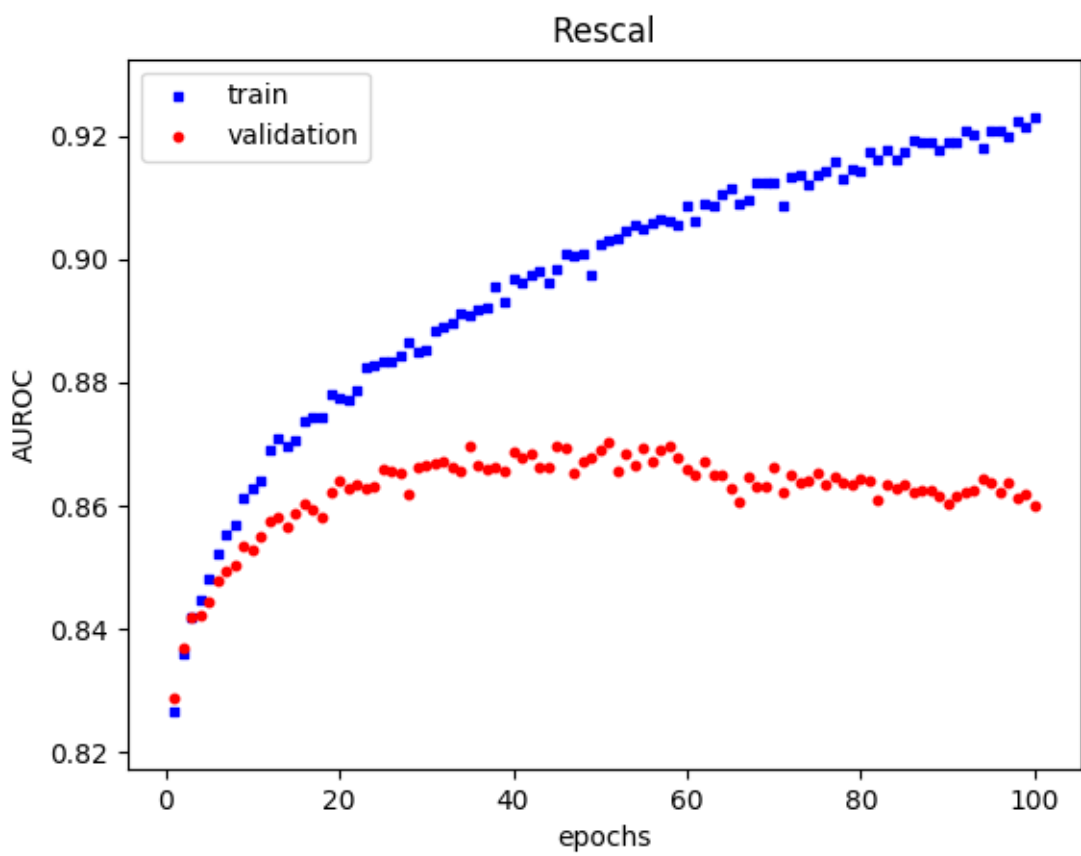**Figure A.1** Training Curve of the Model with Dismult Embedding

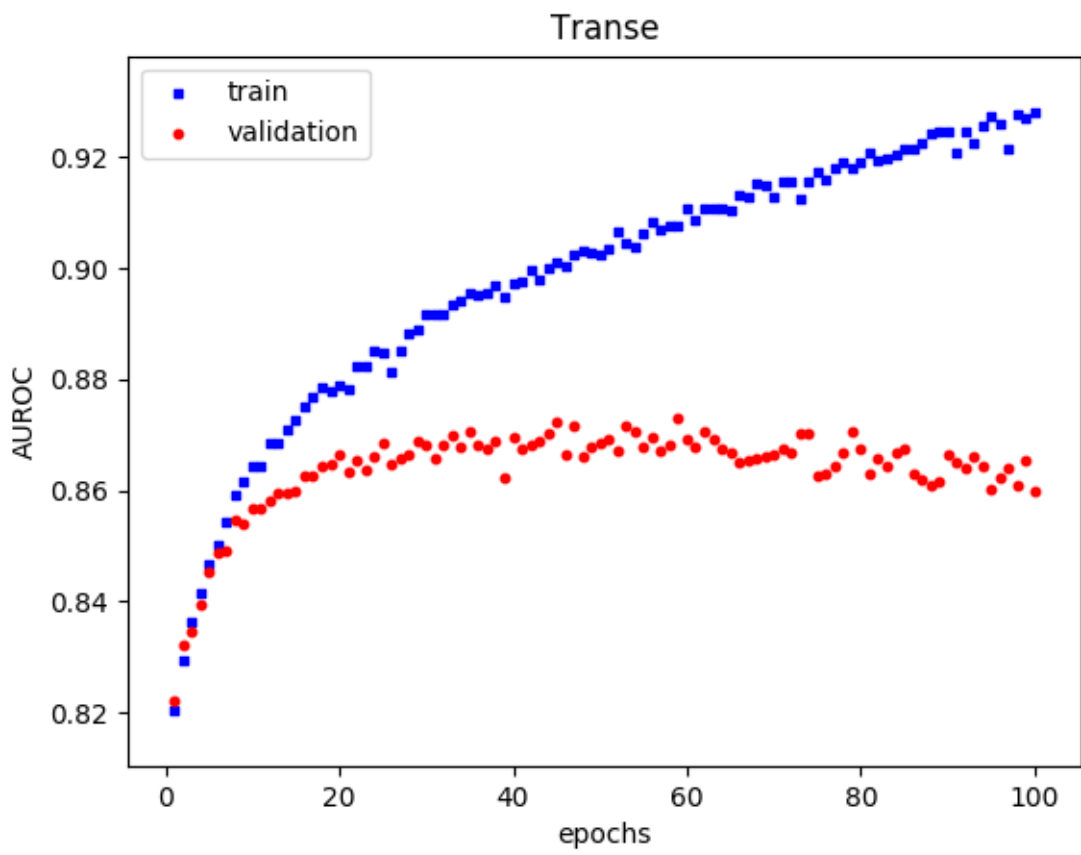**Figure A.2** Training Curve of the Model with Rescal Embedding

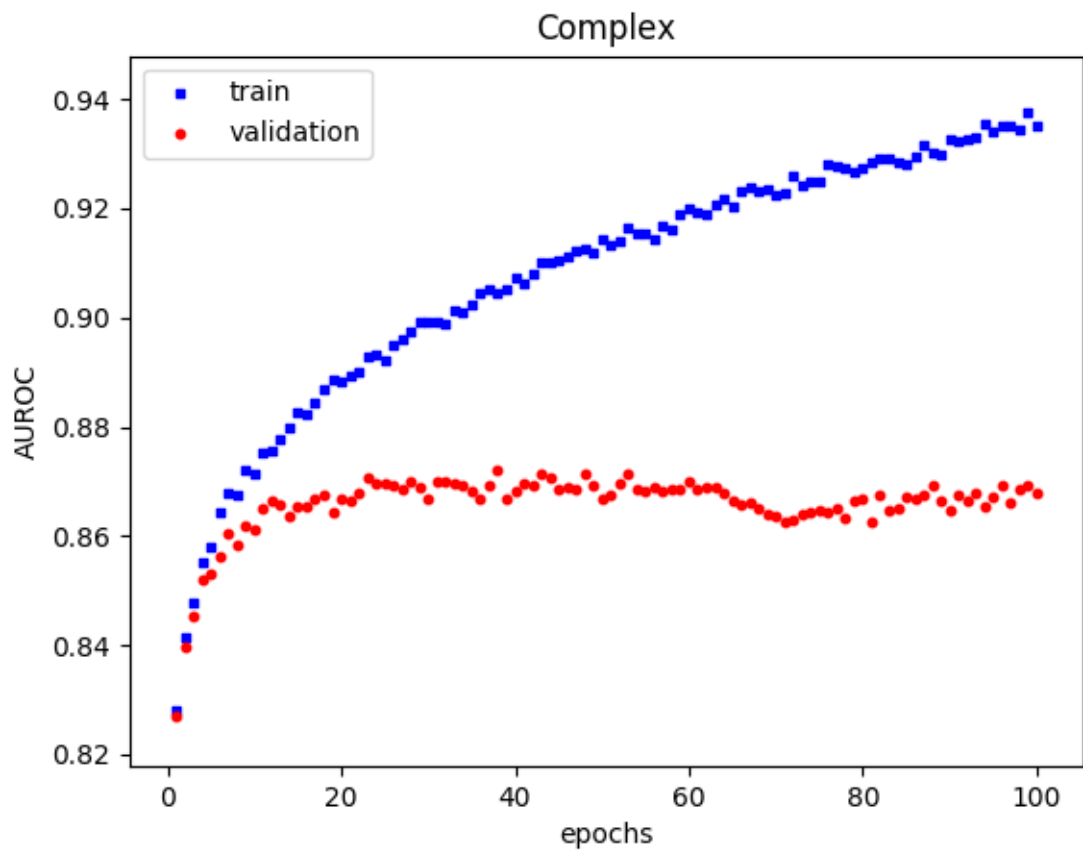**Figure A.3** Training Curve of the Model with TransE Embedding

Figure A.4 Training Curve of the Model with ComplEx Embedding