Deep Entity Embeddings for Cancer Survival Prediction

A.R. Gonçalves[1], A.P de Oliveira[1], P. Ray[1], B. Soper[1], D. Widemann[1], J. Nygard[2], M. Nygard[2]
[1]Lawrence Livermore National Laboratory, [2]Cancer Registry Norway

We used entity embeddings in deep neural networks (DNNs) for the task of cancer survival prediction based on demographical and physiological information. Results showed that the learned embeddings helped to provide more accurate predictions than regular encodings for DNNs, and the learned features boosted the performance of other machine learning methods.

Introduction

- Being able to predict survival time of patients diagnosed with cancer can help unveil important factors to the development of the disease and treatment effect.
- The discrete nature of the associated data makes it harder to train machine learning models based on linear/non-linear function mapping.
- Common approaches are to use categorical variable encodings such as label and one-hot encoding.
- We want to validate whether entity embeddings can help improving the prediction of survival time with machine learning models.

Data

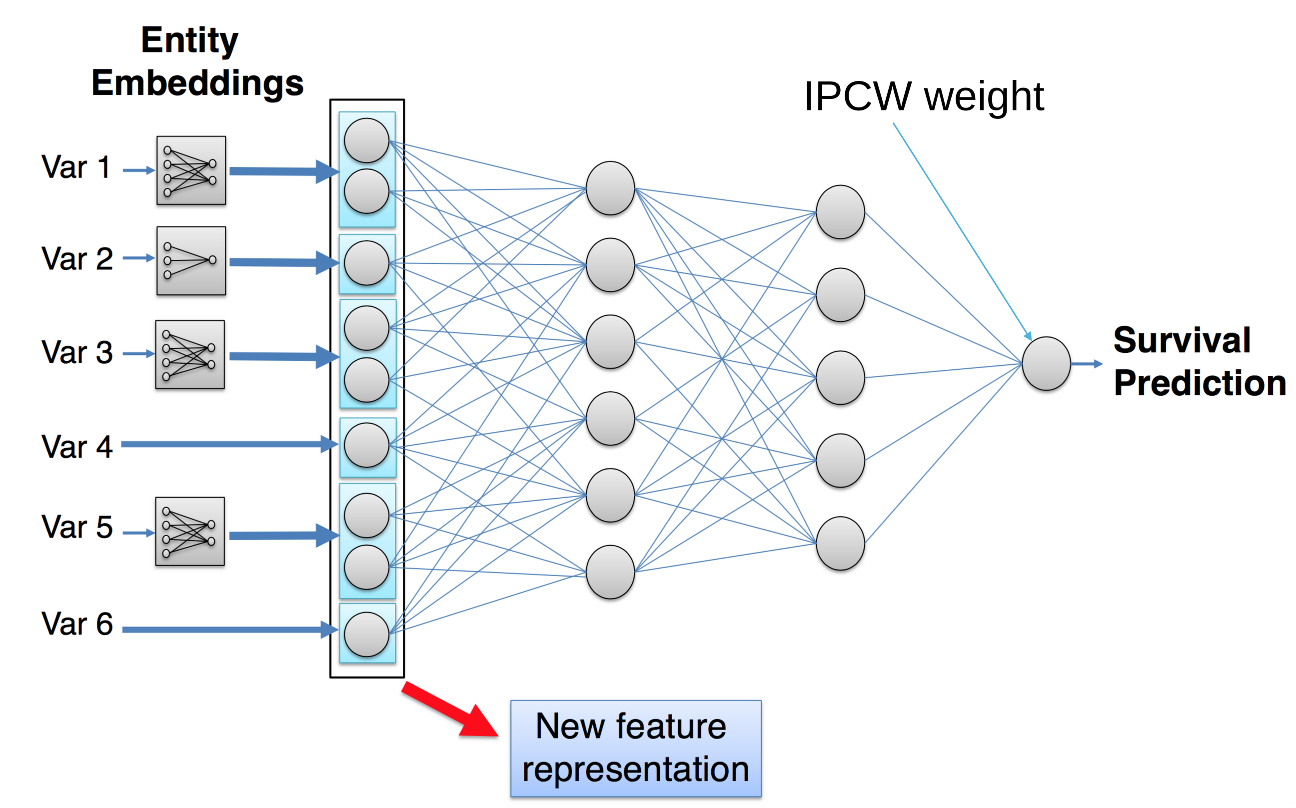
We used public available SEER dataset from National Cancer Institute (NCI) containing cancer incidences in the US from 1973 to 2015. Total amount of data is 27,234 samples. Censoring issued is dealt with IPCW technique.

SEER variable	Description	Type
AGE_DX	Age at diagnosis	Numeric
X_PRIMSITE_1	First two digits of ICD code for anatomical location	Categorical
X_PRIMSITE_2	Third digit of ICD code for anatomical location	Categorical
X_TUMSIZ_COMB_NUM	Tumor size	Numeric
GRADE	Grade	Categorical
SEX	Sex	Categorical
CSLYMPHN	Involvement of lymph nodes	Categorical
DSS1977S	Cancer stage	Categorical
SURGSCOF	Scope of regional lymph node surgery	Categorical
HISTREC	Histology recode, broad groupings	Categorical
DAJCCT	AJCC 'T' component	Categorical
DAJCCN	AJCC 'N' component	Categorical
DAJCCM	AJCC 'M' component	Categorical
DAJCCSTG	AJCC 'stage group' component	Categorical
SURGPRIF	Surgery of primary site, specific	Categorical
X_SURGPRIF_GEN	Surgery of primary site, generic	Categorical

Dataset used in our experiments

Methodology

For each categorical variable we train an embedding. Real value variables are simply passed through to form the new (latent) feature representation, which is then connect to the next layers. The learned features can be used by other regressors.



Experiments and Results

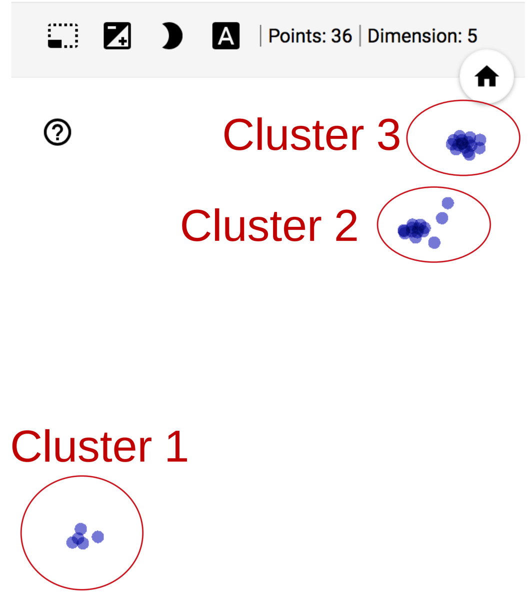
We trained a NN with 2 inner dense layers after the embeddings with 32 and 12 nodes. Total of 3872 parameters. Dropout *p*=0.9 and *selu* activation function. We used Adam optimizer. Embeddings were extracted and used as input features for the regressors. Executed 10 independent runs.

		Ridge Regression		Linear SVR	
	# of feats	RMSE	C-index	RMSE	C-index
Label	16	42.27(0.19)	0.79(.003)	37.90(0.39)	0.78(.003)
One-hot	311	**41.68(0.15)**	0.80(.003)	37.05(0.4)	0.80(.002)
Binary	62	42.00(0.16)	0.80(.002)	37.46(0.39)	0.79(.003)
Embedding	55	41.82 (0.19)	**0.81(.002)**	**36.82(0.4)**	**0.82(.004)**

Predictive performance using traditional encodings and embeddings. Embeddings boosted performance of machine learning methods.

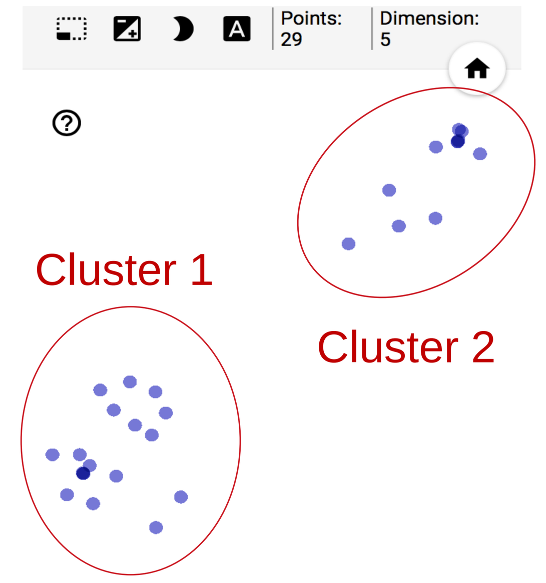
Results: Learned Embeddings

Variable: **Histology Recode—Broad Groupings**



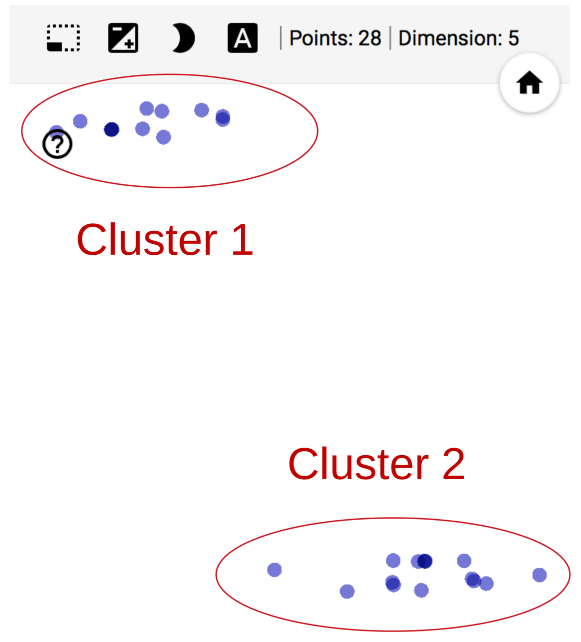
Cluster 1	Cluster 2	Cluster 3
• Code 19	• Code 00	• Code 01
• Code 23	• Code 04	• Code 05
• Code 25	• Code 08	• Code 09
• Code 36	• Code 15	• Code10
• Code 39	• Code 20	• Code 11
	• Code 21	• Code 16
	• …	• …

Variable: **Race/ethnicity**



Cluster 1	Cluster 2
• White	• Black
• Japanese	• American Indian, Alaskan
• Filipino	• Hawaiian
• Laotian	• Korean
• Kampuchean	• Hmong
• …	• …

Variable: **Derived AJCC-6 T**



Cluster 1	Cluster 2
• Ta	• T0
• T2b	• Tis
• T3	• T1
• T3 NOS	• T1a
• T4a	• T1a1
• T4b	• T1a2
• T4b NOS	• T1b
• …	• …

Discussion

- Results showed that entity embeddings are promising mechanisms to boost the training of deep neural networks.
- The new features learned can be used as input for other machine learning regressors.
- In the next steps we want to be able to train embeddings for multiple variables jointly to capture dependence among the variables.

IM Review and Release number: LLNL-POST-752591

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.