Code2Vec: Meaningful Embeddings of Medical Data

David Kartchner, Tanner Christensen, Jeffrey Humpherys, Sean Wade
Computational Medicine Research Group, Brigham Young University

DESCRIPTION

Grouping related disease incidents is a fundamental problem for healthcare and insurance providers. Our research creates a meaningful way to learn numerical representations (called ***embeddings***) of diagnosis and procedure codes that (1) preserves medical meaning, including ostensibly subtle disease patterns and (2) provides a more practical framework for grouping disease incidents. This allows for unsupervised detection of disease comorbidities.

HOW EMBEDDINGS WORK

Vector space embeddings create represent non-numerical data (e.g. words, diagnoses, medications, etc.) that preserves real world meaning.

We use the natural language processing algorithm **GloVe** to find embeddings based on the context in which a particular diagnosis or procedure occurs in an individual's medical history [1].

GloVe leverages the ***Distributional Hypothesis:*** similar words (or codes) appear in similar sentence contexts (or disease history contexts).

Example (Words):

- "The weather forecast says it will rain today."
- "The weather forecast says it will snow today
- Based on the Distributional Hypothesis, "rain" and "snow" are treated as semantically-related words.

Calculating Code Embeddings:
Word/code embeddings are calculated by minimizing:

$$J = \sum_{i,j=1}^{V} f(X_{ij})(w_i^T w_j + b_i + b_j - log(X_{ij}))^2$$

where X_{ij} is the co-occurrence count of word *i* with word *j*, w_i and w_j are the embedded representations of words *i* and *j*, *f* is a weighting function, and b_i and b_j are bias terms.

CLUSTERING RESULTS

- We obtained 50-dimensional embeddings for 20,612 ICD-9 diagnosis and CPT procedure codes.
- We grouped codes into 279 groups using k-means clustering with initial centroids chosen as the means of the corresponding Clinical Classification Software (CCS) disease groupings.

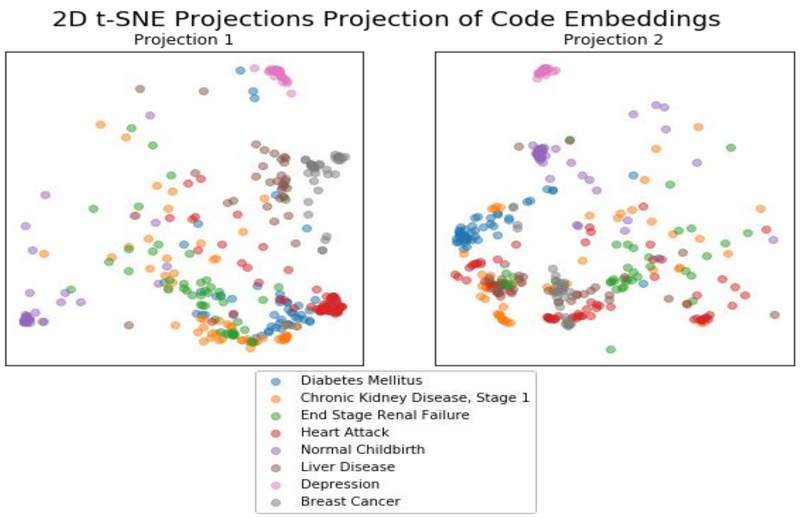

Figure 1: **t-SNE Projections of Selected Code Clusters**

The above plot shows two separate 2-dimensional projection of our code embeddings using t-SNE, which is designed to preserve both local clusters and global separation when reducing dimensionality. We show multiple random projections to better understand our highly dimensional embeddings. Note:

- Disease codes that cluster together occur in the same context
- Diabetes is clustered near heart attack, as well as certain cases of CKD and end-stage renal failure.
- The depression cluster (which contains a variety of mental illnesses) and the childbirth cluster are quite separated from the chronic diseases in both projections.
- Breast cancer and liver disease clusters are located near one another in both plots, potentially reflecting metastatic spread of the cancer to the liver [8].

CLUSTERING RESULTS

	End-Stage Renal Disease Related Code	Code Type	CCS Group
1	Hemodialysis procedure requiring repeated evaluation(s) with or without substantial revision of dialysis prescription	Procedure	58
2	Malnutrition of moderate degree	Diagnosis	52
3	renal dialysis status	Diagnosis	158
4	Conditions due to anomaly of unspecified chromosome	Diagnosis	217
5	End-stage renal disease (ESRD) related services per full month; for patients twenty years of age and older	Procedure	58
6	Osteomalacia, unspecified	Diagnosis	52
7	Renal osteodystrophy	Diagnosis	161
8	End-stage renal disease (ESRD) related services monthly, for patients 20 years of age and older; with 4 or more face-to-face physician visits per month	Procedure	227
9	Secondary hyperparathyroidism (of renal origin)	Diagnosis	161
10	Encounter for extracorporeal dialysis	Diagnosis	158
11	Other and unspecified intracranial hemorrhage following injury without mention of open intracranial wound, with no loss of consciousness	Diagnosis	233
12	Serum screening for cytotoxic percent reactive antibody (PRA); standard method	Procedure	235

Table 1: **Subsample of Stage V Renal Disease (ESRD) Cluster**

Because of the distributional hypothesis, we expect codes in the same cluster to appear in the same context. Note the following about the above:

- Malnutrition is likely associated protein-energy malnutrition, a side effect of dialysis [3].
- Single gene chromosome anomalies have been shown to be a cause of polycystic kidney disease [4].
- Osteomalacia is caused by inability to absorb vitamin D, a symptom of kidney disease [5].
- Intracranial hemorrhage has been associated with polycystic kidney disease [6]. Note that the type our model identified is specifically without an intracranial wound.
- Cytotoxic panel-reactive antibody screenings are performed to assess an individual's suitability for an organ transplant. Kidneys are by far the most frequently transplanted organ [7].

SOURCES

[1] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation." *EMNLP*, vol. 14, 2014, pp. 1532–1543.

[2] World Health Organization et al., "International classification of diseases:[9th] ninth revision, basic tabulation list with alphabetic index," 1978.

[3] Martin K. Kuhlmann, Andreas Kribben, Michael Wittwer, and Walter H. Hörl. Opta—malnutrition in chronic renal failure. *Nephrology Dialysis Transplantation*, 22(3):iii13, 2007

[4] Healthline. Osteomalacia, 2017. Accessed at http://www.healthline.com/health/osteomalacia

[5] F. Hildebrant. Renal medicine 1: Genetic kidney diseases. *The Lancet*, 2010. 375(9722):1287–1295.

[6] S J Ryu. Intracranial hemorrhage in patients with polycystic kidney disease. *Stroke*, 21(2):291–294, 1990.

[7] National Institutes of Health. Organ Donation: Pass it on. *NIH News in Health*.

[8] A Muller, B Homey, H Soto, N Ge et al. Involvement of chemokine receptors in breast cancer metastasis. *Nature*, 410(6824):50-56, 2001.