OJPHI

ISDS 2018 Conference Abstracts

ISDS INTERNATIONAL SOCIETY FOR DISEASE SURVEILLANCE

Exploring the Value of Learned Representations for Automated Syndromic Definitions

Scott Lee*[1], Drew Levin[2], Jason Thomas[1], Patrick Finley[2] and Charles Heilig[1]

[1]Centers for Disease Control and Prevention, Decatur, GA, USA; [2]Sandia National Laboratories, Albuquerque, NM, USA

Objective

To better define and automate biosurveillance syndrome categorization using modern unsupervised vector embedding techniques.

Introduction

Comprehensive medical syndrome definitions are critical for outbreak investigation, disease trend monitoring, and public health surveillance. However, because current definitions are based on keyword string-matching, they may miss important distributional information in free text and medical codes that could be used to build a more general classifier. Here, we explore the idea that individual ICD codes can be categorized by examining their contextual relationships across all other ICD codes. We extend previous work in representation learning with medical data [1] by generating dense vector embeddings of these ICD codes found in emergency department (ED) visit records. The resulting representations capture information about disease co-occurrence that would typically require SME involvement and support the development of more robust syndrome definitions.

Methods

We evaluate our method on anonymized ED visit records obtained from the New York City Department of Health and Mental Hygiene. The data set consists of approximately 3 million records spanning January 2016 to December 2016, each containing from one to ten ICD-9 or ICD-10 codes.

We use these data to embed each ICD code into a high-dimensional vector space following techniques described in Mikolov, et al. [2], colloquially known as word2vec. We define an individual code's context window as the entirety of its current health record. Final vector embeddings are generated using the gensim machine learning library in Python. We generate 300-dimensional embeddings using a skip-gram network for qualitative evaluation.

We use the TensorFlow Embedding Projector to visualize the resulting embedding space. We generate a three-dimensional t-SNE visualization with a perplexity of 32 and a learning rate of 10, run for 1,000 iterations (Figure 1). Finally, we use cosine distance to measure the nearest neighbors of common ICD-10 codes to evaluate the consistency of the generated vector embeddings (Table 1).

Results

T-SNE visualization of the generated vector embeddings confirms our hypothesis that ICD codes can be contextually grouped into distinct syndrome clusters (Figure 1). Manual examination of the resulting embeddings confirms consistency across codes from the same top-level category but also reveals cross-category relationships that would be missed from a strictly hierarchical analysis (Table 1). For example, not only does the method appropriately discover the close relationship between influenza codes J10.1 and A49.2, it also reveals a link between asthma code J45.20 and obesity code E66.09. We believe these learned relationships will be useful both for refining existing syndrome categories and developing new ones.

Conclusions

The embedding structure supports the hypothesis of distinct syndrome clusters, and nearest-neighbor results expose relationships between categorically unrelated codes (appropriate upon examination). The method works automatically without the need for SME analysis and it provides an objective, data-driven baseline for the development of syndrome definitions and their refinement.

Table 1

ICD-10 Code	*Code Description*
J10.1	**Influenza due to other identified influenza virus with other respiratory manifestation**
J11.00	Influenza due to unidentified influenza virus with unspecified type of pneumonia
A49.2	Hemophilus influenzae infection, unspecified site
B33.8	Other specified viral diseases
R19.7	**Diarrhea, unspecified**
A05.9	Bacterial foodborne intoxication, unspecified
J09.X3	Influenza due to identified novel influenza A virus with gastrointestinal manifestations
R19	Other symptoms and signs involving the digestive system and abdomen
J45.20	**Mild intermittent asthma, uncomplicated**
J45.30	Mild persistent asthma, uncomplicated
E66.09	Other obesity due to excess calories
J45.40	Moderate persistent asthma, uncomplicated

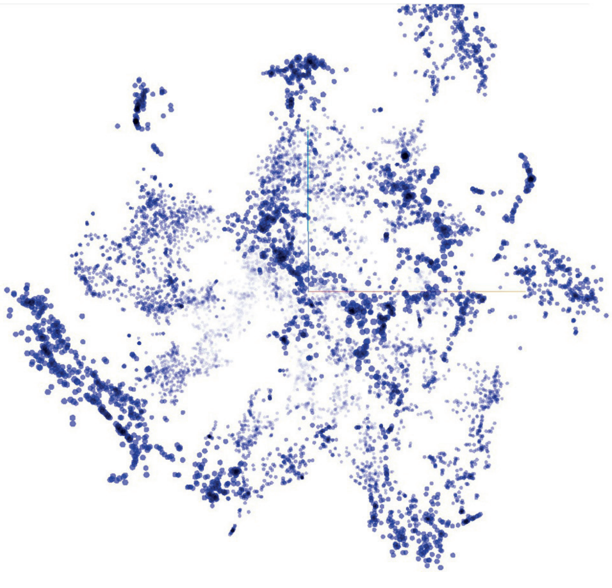

Figure 1: T-SNE visualization of [300 dimensional skip-gram] embedded ICD code vectors. The heterogeneous structure suggests distinct syndrome definitions. Image generated using Google's online TensorFlow Projector.

Keywords

Word embeddings; Deep learning; Syndrome definitions; ICD codes

Acknowledgments

This work was supported by Laboratory Directed Research and Development funding from Sandia National Laboratories. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DENA0003525.

References

[1] Choi Y, Chiu CY-I, Sontag D. Learning Low-Dimensional Representations of Medical Concepts. AMIA Summits on Translational Science Proceedings. 2016;2016:41-50.

[2] Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. InAdvances in neural information processing systems 2013 (pp. 3111-3119).

***Scott Lee**
E-mail: yle4@cdc.gov