**“Medical Concept Representation Learning from Electronic Health Records and its Application on Heart Failure Prediction”**

Edward Choi\*,MS; Andy Schuetz\*\*,PhD; Walter F. Stewart\*\*,PhD; Jimeng Sun\*,PhD

Summarized by-

Fazle Rabby Khan  
ID: 1611089042

CSE-299\_Section 12

Saturday 9:40AM

Submitted to -

FACULTY : INTISAR TAHMID NAHEEN(ITN)

SUMMARY :

The paper is all about to transform heterogeneous clinical data from electronic health records(EHR) into clinically meaningful constructed features using data driven method that rely, in part, on temporal relations among data just to predict heart failure. Data were from Sutter Palo Alto Medical Foundation (Sutter-PAMF) primary care patients.  
Skip-gram applied to Configure the Medical Concept Representation Learning   
By scanning through encounter, medication order, procedure order and problem list records of all 265,336 patients, and extracted diagnosis, medication and procedure codes were assigned to each patient in temporal order. If a patient received multiple diagnoses, medications or procedures at a single visit, then those medical codes were given the same timestamp. The respective number of unique diagnoses, medications and procedures was 11,460, 17,769 and 9,370 totaling to 38,599 unique medical concepts.100-dimensional vectors were used to represent medical concepts, considered 300 was sufficient to effectively represent 692,000 vocabularies in NLP.  
As natural language text can be seen as a sequence of codes, medical records such as

diagnoses, medications and procedures can also be seen as a sequence of codes over time. The total framework maps raw medical concepts into related concept vectors using Skip-gram and validating the utility of the resulting medical concept vectors.

INSIGHTS :  
  
**Theano** , a Python library used for evaluating mathematical expression to implement Skip-gram. Theano can also take advantage of GPUs to greatly improve the speed of calculations involving large matrices. For optimization, Adadelta was used which employs adaptive learning rate. Unlike stochastic gradient descent (SGD), which is widely used for training neural networks, Adadelta does not depend very strongly on the setting of the learning rate, and shows good performance. Using Theano 0.7 and CUDA.7 on an Ubuntu machine with Xeon E5-2697 and Nvidia Tesla K80, it took approximately 43 hours to run 10 epochs of Adadelta with the batch size of 100.

The patient representation is then used to train heart failure prediction models using various classifiers, namely logistic regression, support vector machine (SVM), multi-layer perceptron with one hidden layer (MLP) and K-nearest neighbors classifier (KNN). In the prediction phase, mapping of the medical record of a patient to medical concept vectors was done and generated patient vectors by aggregating the concept vectors afterwards.

**COMPARISON BETWEEN “MEDICAL-CONCEPT-VECTOR” AND “ONE HOT ENCODING”:**

Training of four popular classifiers, namely logistic regression, MLP, SVM, and KNN was done using both one-hot vectors and medical concept vectors. To train the models with medical concept vectors, we converted the medical records to patient vectors.To train the models with one-hot encoding, we converted the medical records to aggregated. one-hot vectors in the same fashion, using one-hot vectors instead of medical concept vectors. But the accuracy level of one-hot-vectors and medical concept vectors were not quite similar. medical concept vectors not only improve performance, but also significantly reduce the training time.

**CONCLUSION:**  
The whole idea is a new way of representing heterogeneous medical concepts as realvalued vectors and constructing efficient patient representation using the state-of-the-art Deep Learning method. The system qualitatively showed that the trained medical concept vectors indeed captured medical insights compatible with our medical knowledge and experience. For the heart failure prediction task, medical concept vectors improved the performance of many classifiers, thus quantitatively proving its effectiveness. Though there are some drawbacks the limitation of methods and possible future works, which include deeper utilization of medical information, combining expert knowledge into our framework, and expanding this type of approach to various medical applications.

1. What is new in this paper?

-> The proposed method is based on skip-gram which popular for learning word representation and an unique classifier is used namely multi-layer perceptron with one hidden layer (MLP).   
**Theano** , a Python library used for evaluating mathematical expression to implement Skip-gram. Theano can also take advantage of GPUs to greatly improve the speed of calculations involving large matrices.

2. What is the unique advantage of the proposed/implemented method?

-> The unique advantage of proposed method is that we can compare between medical-concept-vector and one-hot-encoding. Medical-concept-vector got the highest accuracy.

3. What is your unique contribution in this direction?

-> Since , One-hot-encoding is a vast method, good for big data but less effective when applied on few patients data. It takes more time than Medical-concept-vector.