External Heterogeneous Information Network Embedding for Electronic Medical Records

Edward W Huang
CS512
What is an Electronic Medical Record (EMR)?

● Depends on dataset, but typically contains a patient’s
  ○ Symptoms
  ○ Medical test results (e.g., white blood cell count)
  ○ Medical history
  ○ Prescribed drugs
  ○ Survival information

● More on survival information later
Impact of EMRs

- EMR adoption will save more than $81 billion annually
- Improve patient safety: harmful interactions from multiple drugs
- More data for data mining tasks
- Potential data mining tasks
  - Relationship mining
  - Clinical prediction of survival (CPS)
  - These tasks can further enhance clinical decision-making

Frequent Problems with EMRs

- Missing data
- Semantic mismatch
- These problems make it difficult to perform many data mining tasks
- Notably, tasks relying on determining similarity between pairs of patients is most affected
  - Clustering patients
  - EMR visualization in 2D or 3D space
    - First build a $k$-nearest neighbors graph
Missing data

- Many data mining methods assume availability of all features
  - Assumption usually doesn’t hold
  - Example: doctors do not perform all medical tests on all patients

- Mean imputation
  - Most common method to fill missing values
  - Introduces noise rather than reduce it
Semantic mismatch

- Features that mean the same thing, but occupy different spaces in the vocabulary
  - Example: “halitosis” and “bad breath”
- Similar problem in text mining can be solved with word2vec or other word embedding techniques
  - However, does not translate well to medical records
  - EMR-specific solution: med2vec
    - Relies on sequential information that not all databases have
    - Thus, better suited for chronic disease treatment over many years

Solution: External Information

- In the word2vec CBOW model, we use sliding context windows to infer the meaning of each word.
- In EMRs, missing data essentially means we’re missing a few words in each sentence.
- Solution: use external information to fill in missing context information
  - Molecular interaction networks, such as protein-protein interaction (PPI) networks (protein-protein edges)
  - Known drug-protein targets (drug-protein edges)
  - Known drug-symptom relationships (drug-symptom edges)
- But how can we incorporate all of this external information into EMRs?
Integrating External Information with EMRs

- We wish to create a network from EMR data in order to create a heterogeneous information network.
- Use co-occurrence information to build network:
  - Example: create edge between node $d$ and node $s$ if, for a single patient, $d$ is a drug that is prescribed to treat some symptom $s$.
- With this co-occurrence network, we can mix in PPI network and domain knowledge to create the new HIN.
Data Description

● Lung cancer data set
  ○ 43 patients with squamous-cell lung carcinoma
  ○ 90 patients non-squamous-cell lung carcinoma

● 449 unique features (excluding proteins)

● 133 patient records with high degree of sparsity
  ○ Each record missing an average of 407 out of 449 features
  ○ Patient record matrix: 133x449 matrix

● Each record also contains the patient’s survival information
  ○ Number of days until death
  ○ If no death event, number of days until hospital discharge

● Heterogeneous Information Network
  ○ 11,911 nodes (most of these are proteins)
  ○ 379,715 edges
  ○ 23 edge types
Using HIN to enrich EMRs

● With the HIN, we now wish to fill in the missing “contexts” of the medical records
● Use network embedding technique, ProSNet, which has been shown to be useful in protein function prediction
  ○ word2vec: in sentence, a word’s neighbors should predict the word’s context
  ○ ProSNet: in graph, a node’s neighbors should predict the node’s context
● Get a low-dimensional vector for each node
● Get a pairwise cosine similarity matrix of nodes
● Multiply matrix into patient record matrix
Experiments

- Compare performance of HIN-enriched patient record matrix with the patient record matrix with mean imputation
- Clustering experiment
  - Hospitals are interested in whether a patient will survive
  - With patient survival functions as ground truth, we want to separate patients into two groups
  - Thus, the best clustering is one in which the two clusters have the most different survival functions
- Survival functions can be estimated with Kaplan-Meier curves, then compared with log-rank test
- Low \( p \)-value means the two clusters have significantly different survival rates
- So, the lower the \( p \)-value, the more successful the method was in separating relatively healthy patients from those with short survival times
Squamous-Cell Lung Carcinoma Results

With HIN, $p$-value = 0.001020

With mean imputation, $p$-value = 0.01267
Non-Squamous-Cell Lung Carcinoma Results

With HIN, $p$-value = 0.003652

With mean imputation, $p$-value = 0.02333
Results Discussion

- We showed that the HIN-enriched patient records can improve the performance of clustering over the usual mean imputation method.
- For mean imputation, neither cancer subtype was significantly separated at the 0.01 level, while the HIN-enriched feature matrix was.
- Therefore, the external information (PPI network and domain knowledge) were helpful in solving the issues associated with missing data and semantic mismatch.
Future Work

- Perform clinical prediction of survival
  - Given new patient, predict survival rate by first finding group of patients most similar to him/her, and then extrapolating survival function
  - Can use algorithm similar to $k$-nearest neighbors

- Feature discovery, including novel drug-drug interactions