Feature Engineering to Power Machine Learning Phenotype Development

Xinzhuo Jiang, MS[1,2], Krishna S. Kalluri, MS[1,2], Chao Pang, PhD[1,2], Kai Chen[1,2], Junghwan Lee[1,2], Cong Liu, PhD[1,2], Ruijun Chen, MD[1,2] ,Patrick Ryan, PhD[1,2], Karthik Natarajan, PhD[1,2]
[1]Columbia University Medical Center, New York, NY
[2]OHDSI, New York, NY

Is this the first time you have submitted your work to be displayed at any OHDSI Symposium?

Yes __x___ No_____

Abstract

This study focuses on feature engineering which is a prerequisite for modeling phenotype characterization. Based on concept pairs extracted from the eMERGE Network by reference set group, four types of features were created[1]. Lexical features describe the linguistic similarity of concept pairs. Semantic features measure the distance between two concepts in semantic ontology. Co-occurrence features are generated to show the prevalence of two concepts occurring within the same time window. The last feature group is concept embeddings which computes the dependence of two concepts.

Introduction

As a first study by the Columbia Phenotype group, we focused on identifying pairs of concepts that should belong to the same phenotype based on current state-of-the-art phenotype definitions.
To achieve this goal, concept codes from 53 validated phenotype algorithms from the eMERGE Network[1] were extracted by the reference set group to create positive controls - pairs of concepts belonging to the same phenotype by mapping the drug, measurement and procedure codes to OMOP Standardized Vocabularies concept identifiers and permuting those that belong to one phenotype - and negative pairs were generated using a heuristic that utilizes concepts not belonging to any phenotype. A snapshot of the reference set is shown in figure 1. Besides concept pairs and names, the following features are also included in the reference set: ground_truth, same domain, is_ancestor, min_distance, is_sibling_with_same_parent.

ground_truth	concept_id_1	concept_id_2	concept_name_1	concept_name_2	same_domain	is_ancestor	min_distance	is_sibling_w_same_parent
0	75576	435216	Irritable bowel syndrome	Disorder due to type 1 diabetes mellitus	1	0	9	0
0	80809	378726	Rheumatoid arthritis	Dementia associated with alcoholism	1	0	6	0
1	81064	439770	Pseudopolyposis of colon	Ketoacidosis in type 1 diabetes mellitus	1	0	9	0
0	81097	26942	Felty's syndrome	Hemoglobin SS disease with crisis	1	0	6	0
1	81097	1119155	Felty's syndrome	0.4 ML adalimumab 50 MG/ML Prefilled Syringe [Humira]	0	0		0

Figure 1: Example of reference set

Feature Engineering

Using the reference set provided by reference set group, with positive and negative controls, we generated the features which can be further used in creating a model. In order to generate features, we used several clinical data mining techniques such as creating concept embeddings. We used the reference set as the knowledge base and the condition and procedure, drug and measurement domains in Columbia OMOP instance as the source data. The algorithms were run using Apache Spark to generate new sets of features

which were added to the reference set and used by the modelling team as input data. For better understanding and ease of generating these features, we categorized the newly created features as part of Feature Engineering into four categories as follows.

1. **Lexical Features:**

Lexical Features measures the degree to which the two concepts are similar linguistically. Lexical similarity of 1 means total overlap and 0 means no overlap. In lexical similarity measures we calculated 5 measures for each concept pair including ***Levenshtein_distance, Levenshtein_ratio, Jaro, Jaro_winkler, Fuzz_partial_ratio.***

2. **Semantic Features:**

The similarity score between two concepts is based on the likeness of their meaning or semantic content. We chose these set of features to help better understand the semantic ontology structure and relationship between the concept pairs which will add additional power to modelling work. In this group, 8 different Semantic features were extracted. They are ***Distance_indicator, Semantic_similarity, Lin_measure[2], Jiang_measure[2], Relevance_measure[2], Information_coefficient[2], GraphIC_measure[2], Mica_information_content[3].***

3. **Co-occurrence Features:**

All the co-occurrence matrices are computed based off of domain tables from the latest ohdsi_cumc_deid (inpatient and outpatient data) database available on iNYP. Domain tables include condition_occurrence, procedure_occurrence, drug_exposure, measurement, and observation. We used various time windows to measure the co-occurrence matrices, starting from 60 days to lifetime. The matrices are ***Co-occurrence_60_days/_90_days/_180_days/_360_days/_lifetime***.

4. **Concept Embeddings Features:**

The GloVe algorithm is run on the co-occurrence matrix to generate the concept embeddings[4]. As of now only cooccurrence_lifetime is used for computing the embeddings. Except for ***Lifetime_cooccur_embedding_cosine, 5_year_cooccur_embedding_cosine*** *and* ***Visit_cooccur_embedding_cosine*** *are under development.*

Results

Through feature engineering, four groups and 19 features were created and added to the reference set. In total, we got 23 features in the reference matrix which enriched the predictor variables for training the model. All the features implementation is available on GitHub: https://github.com/cumc-dbmi/phenotype-features.

References

1. Gottesman, Omri, et al. "The electronic medical records and genomics (eMERGE) network: past, present, and future." Genetics in Medicine 15.10 (2013): 761.
2. Deng, Y., Gao, L., Wang, B. and Guo, X. (2015). HPOSim: An R Package for Phenotypic Similarity Measure and Enrichment Analysis Based on the Human Phenotype Ontology. PLOS ONE, 10(2), p.e0115692.
3. Resnik, P. (1999). Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. Journal of Artificial Intelligence Research, 11, pp.95-130.
4. Pennington, J. (2019). GloVe: Global Vectors for Word Representation. [online] Nlp.stanford.edu. Available at: https://nlp.stanford.edu/projects/glove/ [Accessed 21 Jun. 2019].