International Journal of Population Data Science

Journal Website: www.ijpds.org

Using Biomedical Text as Data and Representation Learning for Identifying Patients with an Osteoarthritis Phenotype in the Electronic Medical Record

Meaney, C[1], Widdifield, J[2], Jaakkimainen, L[2], Escobar, M[1], Rudzicz, F[3], and Tu, K[1]

[1]University of Toronto
[2]Institute for Clinical Evaluative Sciences
[3]University Health Network

Introduction

Electronic medical records (EMRs) are increasingly used in health services research. Accurate/efficient identification of a target population with a specific disease phenotype is a necessary precursor to studying the health of these individuals.

Objectives and Approach

We explored the use of biomedical text as inputs to supervised phenotype identification algorithms. We employed a two-stage classification approach to map the discrete, sparse high-dimensional biomedical text data to a dense low dimensional vector space using methods from unsupervised machine learning. Next we used these learned vectors as inputs to supervised machine learning algorithms for phenotype identification.

We were able to demonstrate the applicability of the approach to identifying patients with an osteoarthritis (OA) phenotype using primary care data from the Electronic Medical Record Administrative data Linked Database (EMRALD) held at ICES.

Results

EMRALD contains approximately 20Gb of biomedical text data on approximately 500,000 patients. The unit of analysis for this study is the patient. We were interested in identifying OA patients using solely text data as features.

Labelled outcome information wass available from a random sample of 7,500 patients. We divided patients into training (N=6000), validation (N=750) and test (N=750) cohorts. We learned low dimensional representations of the input text data on the entire EMRALD corpus (N=500,000). We used learned numeric vectors as inputs to supervised machine learning models for OA classification (N=6,000 training set patients).

We compared models in terms of accuracy, sensitivity, specificity, PPV and NPV. The best learned models achieved approximately 90% sensitivity and 80% specificity. Classification accuracy varied as a function of learned inputs.

Conclusion/Implications

We developed an approach to phenotype identification using solely biomedical text as an input. Preliminary results suggest our two-stage ML approach has improved operating characteristics compared to existing clinically derived decision rules for OA classification. Future work will explore the generalizability of this methodology to other disease phenotypes.

http://dx.doi.org/10.23889/ijpds.v3i4.761