International Journal of Population Data Science

Journal Website: www.ijpds.org

IJPDS International Journal of Population Data Science

Swansea University Prifysgol Abertawe

Identify Patients with Congestive Heart Failure through Analyzing Free-Text Clinical Notes

Yann, M[1], Stukel, T[1], Jaakkimainen, L[1], and Tu, K[2]

[1]Institute for Clinical Evaluative Sciences

[2]University of Toronto

Introduction

A number of challenges exist in analyzing unstructured free text data in electronic medical records (EMRs). EMR text are difficult to represent and model due to their high dimensionality, heterogeneity, sparsity, incompleteness, random errors and the presence of noise.

Objectives and Approach

Standard Natural Language Processing (NLP) tools make errors when applied to clinical notes due to physician use of unconventional language, involving polysemy, abbreviations, ambiguity, misspelling, variations, and negation.

This paper presents a novel NLP framework, "Clinical Learning On Natural Expression" (CLONE), to automatically learn from a large primary care EMR database, analyzing free text clinical notes from primary care practices. CLONE's predictive clinical models using text mining and neural network approach to extract features to identify patterns. To demonstrate effectiveness, we evaluate CLONE's ability in a case study to identify patients with a specific chronic condition: congestive heart failure (CHF).

Results

A random selected sample of 7500 patients from Electronic Medical Record Administrative data Linked Database (EMRALD) is used. In this dataset, each patient's medical chart includes a reference standard, manually reviewed by medical practitioners. Prevalence of CHF is approximately 2%. The low prevalence leads to another challenging problem in machine learning: imbalanced datasets. After pre-processing, we build deep learning models to represent and extract important medical information from free text to identify CHF patients through analyzing patient charts. We evaluated the effectiveness of CLONE by comparing the predicted labels with the standard references on a holdout test dataset. Comparing it with a number of alternative algorithms, we improve the overall accuracy to over 90% on a test dataset.

Conclusion/Implications

As the role of NLP in EMR data expands, the CLONE natural language processing framework can lead to substantial reduction in manual processing, while improving predictive accuracy.

http://dx.doi.org/10.23889/ijpds.v3i4.1032