Accepted Manuscript

Novel Approach to Predict Hospital Readmissions Using Feature Selection from Unstructured Data with Class Imbalance

Arun Sundararaman, Srinivasan Valady Ramanathan, Ramprasad Thati

PII: S2214-5796(17)30313-1
DOI: https://doi.org/10.1016/j.bdr.2018.05.004
Reference: BDR 99

To appear in: *Big Data Research*

Received date: 16 October 2017
Revised date: 12 May 2018
Accepted date: 12 May 2018

Please cite this article in press as: A. Sundararaman et al., Novel Approach to Predict Hospital Readmissions Using Feature Selection from Unstructured Data with Class Imbalance, *Big Data Res.* (2018), https://doi.org/10.1016/j.bdr.2018.05.004

Novel approach to predict hospital readmissions using feature selection from unstructured data with class imbalance

Arun Sundararaman, Srinivasan Valady Ramanathan, Ramprasad Thati
Health Analytics Solution Factory, Accenture

A*bstract* - Feature selection for predictive analytics continues to be a major challenge in the healthcare industry, particularly as it relates to readmission prediction. Several research works in mining healthcare data have focused on structured data for readmission prediction. Even within those works that are based on unstructured data, significant gaps exist in addressing class imbalance, context specific noise removal which thus necessitates new approaches readmission prediction using unstructured data. In this work, a novel approach is proposed for feature selection and domain related stop words removal from unstructured with class imbalance in discharge summary notes. The proposed predictive model uses these features along with other relevant structured data. Five iterations of predictions were performed to tune and improve the models, results of which are presented and analyzed in this paper. The authors suggest future directions in implementing the proposed approach in hospitals or clinics aimed at leveraging structured and unstructured discharge summary notes.

Keywords—predictive analytics; Unstructured data; discharge summary; class imbalance; domain related stop words; feature selection; MIMIC III; text mining; N-gram features.

Definitions

AUC : Area under the curve; it is used in classification analysis in order to determine which of the used models predicts the classes best.

Precision: also called positive predictive value is the fraction of relevant instances among the retrieved instances

Recall: also called sensitivity is the fraction of relevant instances that have been retrieved over the total amount of relevant instances.

Specificity: measures the proportion of negatives that are correctly identified as such

F-Score: is a measure of a test's accuracy. It considers both the precision p and the recall r of the test to compute the score

I.INTRODUCTION

Clinical data is getting increasingly complex on different dimensions viz., volume, variety and velocity impacting the quality of care [1]. Unique characteristics of medical data that makes it complex include varied data formats, high level of missing values, privacy of data etc. This increasing complexity is leading to the need for sophisticated methods and techniques such as clinical data mining or predictive analytics for clinical decision support. Predictive analytics is a branch of data mining that supports prediction of outcome and probability of such outcome, based on insights from trends and patterns from historical data. A predictive model seeks to predict the results of a response or dependent variable based on a set of variables known as predictor or independent or criterion variables. Recent advances in predictive mining have given rise to a trend where more and more clinical systems are adopting predictive analytics as part of their clinical decision support modules. In clinical or medical informatics predictive analytics predominantly deals with models to predict patients' health, to support clinicians in diagnostic, therapeutic, or other medical decision tasks [2]. Emphasizing the significance of predictive analytics in evidence based medicine, Sanjeev Sood [3] lists multiple application areas where such techniques are becoming an essential element of clinical systems, early detection of emerging diseases or spotting outbreak of epidemics etc.

The role of discharge summary and insights contained therein in the study of readmission have assumed recent significance. A recent study [4] establishes that high-quality discharge summaries were associated with reduced risk of readmission for patients with heart failure.

Over the years, different data mining techniques have been introduced for application to medical-related fields increasing the complexity of techniques behind the predictive system. Examples of such techniques include, but not restricted to, genetic algorithms or artificial neural networks or fuzzy sets or inductive logic programming [5].

Researchers and practitioners in predictive analytics are advised and encouraged to focus on selecting the most appropriate approach, given the widespread availability of several new computational methods and tools [6]. The need for such novel approaches assumes significance due to the inherent complexities and special characteristics associated with medical data listed in the previous paragraph. Of the several novel possibilities that exist, use of mixed-data for predictions