Smart Health xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Smart Health

journal homepage: www.elsevier.com/locate/smhl

Incentive design for high quality disease prediction model using crowdsourced clinical data

Qinghan Xue*, Mooi Choo Chuah

Department of Computer Science and Engineering, Lehigh University, 27 Memorial Dr W, Bethlehem, PA 18015, USA

ARTICLE INFO

Keywords:
Medical data mining
Hidden patterns
Feature selection
Incentive
Individual rationality
Platform profitability

ABSTRACT

Clinical data mining has great potential for mining hidden patterns in the medical datasets, which can then be used to guide clinical decision making and personalized medicine. While several studies have merged medical data mining techniques with statistical analysis, their proposed mechanisms are excessively complex and are not particularly accurate for individual patients. Therefore, it is essential that a better tool is developed for disease progression and survival rate predictions. In addition, most of the medical datasets are noisy and hence any dataset needs to be cleaned before it is used for predictions. Each dataset may contain many features not all of which are useful for predictions. Therefore, useful feature selection techniques need to be employed before prediction models can be constructed. Furthermore, larger and high quality datasets typically create better prediction models. Thus, in this paper, we explore how data cleaning and feature selection techniques affect the performance of the prediction models. In addition, we develop a new incentive model with individual rationality and platform profitability features to encourage different hospitals to share high quality data so that better prediction models can be constructed. We evaluate our proposed techniques using three datasets and the results show that our proposed methods are more efficient and accurate than several existing prediction models.

1. Introduction

With the growing need for personalized medical treatments to lower healthcare cost, and the ease of storing large scale clinical datasets, researchers are keen on developing data mining methodologies to mine hidden patterns such that they can identify critical factors which affect disease progression and survival rates and use such information to aid the healthcare professionals in making better treatment decisions.

However, discovering knowledge (Dave Smith and Marlow, 2007; Mishra, Das, Mishra, and Mishra, 2010; Roddick, Fule, and Graco, 2003) in healthcare systems is a herculean yet critical task since those available raw medical data are widely distributed, heterogeneous in nature, and voluminous. For example, the Amyotrophic Lateral Sclerosis (*ALS*) is a fatal neuro-degenerative disease with significant heterogeneity and can lead to muscle weakness which gradually impacts the functioning of the body, leading to eventual death. Though the average survival duration is 3–5 years from the onset of symptoms, survival duration is markedly varying for different patients, ranging from several months to over a decade. Because of this heterogeneity, *ALS* clinical trials are inefficient, requiring large numbers of participants and long testing duration to collect sufficient data for the study of survival patterns.

With this dire need for *ALS* prediction tools in mind, a new challenge competition (The DREAM *ALS* Stratification Prize4Life DREAM ALS Stratification Prize4Life Challenge, 2015) is held for better understanding of patient profiles, and seeking mathematical

* Corresponding author.
E-mail addresses: qix213@lehigh.edu (Q. Xue), chuah@cse.lehigh.edu (M.C. Chuah).

https://doi.org/10.1016/j.smhl.2017.12.002
Received 22 June 2017; Received in revised form 12 December 2017; Accepted 13 December 2017

tools to predict the *ALS* progression and survival rate such that better personalized treatment can be procured. The main goal of the challenge is to forecast disease progression of a given patient more accurately in a year's time using his/her three months data.

While clinical data mining technology helps to identify hidden patterns within patients' healthcare data which can aid in developing relevant treatment tools for personalized treatment, many current diagnostic and prognostic tools make decisions based on only a small number of patient characteristics. For example, many cardiologists and critical care physicians believe that the direct measurement of cardiac function provided by Right Heart Catheterization (*RHC*) is enough to guide treatments of certain critically ill patients. However, there are significant limitations to this type of assessment. Since in the catheterization laboratory, hemodynamic variables are typically measured at rest with patients in the supine position, which may not only underestimate the presence and severity of Pulmonary Hypertension (*PH*) but such measurements also do not accurately reflect the true extent of hemodynamic compromise.

These days, many tests were performed on patients and hence clinicians have accessed to huge amount of patients' data but not all measurements are useful for disease progression and survival rate predictions. The collected data sometimes have missing and noisy values. Thus, data mining researchers need to ensure that high quality data is available for training any disease prediction model. For higher accuracy model, a large dataset consisting of sufficient numbers of different categories of patients, e.g., fast, slow and average ALS patients, needs to be available. Therefore, hospitals need to be encouraged to share high quality data such that healthcare data mining researchers can produce better models.

In this paper, we tackle such challenges by developing two tools: first, we propose a data cleaning & feature selection method which allows healthcare data mining researchers to clean up the available large scale dataset and identify relevant features which are critical in predicting the progression and survival rate of any given disease. Secondly, we propose an incentive mechanism which encourages hospitals to share truthfully high quality data which can then be aggregated to generate prediction models with higher accuracy rates. We demonstrate the effectiveness of our approaches using three large datasets from three population-based studies, namely *ALS* (PROACT, 2015), *RHC* (Right Heart Catheterization, 2015) and STAR*D (Nie, Gong, and Ye, 2016). Our experimental results show that our prediction models yield good performance in *ALS* slope, *ALS* & *RHC* survival, and STAR*D relapse predictions. Furthermore, we also prove that our incentive mechanism satisfies two desirable properties: individual rationality and platform profitability.

In summary, our contributions are as follows:

- We develop new learning models by combining data cleaning & feature selection methods with effective machine learning techniques.
- We also design an incentive mechanism to encourage hospitals to share higher quality data, so that the cloud server can generate more accurate models for clinical use.
- We provide extensive experimental results using both $PRO-ACTALS$ and *RHC* datasets to show that our learning models can perform better than some existing prediction tools (e.g., the Cox-survival regression model). In addition, we show that our incentive mechanism has individual rationality and platform profitability properties.

The rest of the paper is organized as follows. Section 2 discusses related work. Section 3 provides brief descriptions of our system model, design challenges and design goals. Section 4 describes how we conduct data cleaning, feature selection and build our prediction models. Section 5 evaluates those models using PRO-ACT *ALS*, *RHC* and STAR*D datasets. Section 6 provides details of our incentive based high-quality ehealth information sharing scheme and presents proofs and evaluation results of our incentive mechanism. Section 7 concludes the paper with discussions of future work.

2. Related work

2.1. Data mining over clinical data

Clinical Data Mining (*CDM*) is the application of data mining techniques using clinical data (Iavindrasana et al., 2009), which involves the conceptualization, extraction, analysis, and interpretation of available clinical data for building practical knowledge and making clinical decisions (Epstein, 2009). For example, the authors in Hoeper et al. (2013) have designed a novel method for pulmonary hypertension predictions using several clinical assessments such that physicians can use these predictions for diagnosis and management of multiple heart diseases.

Instead of diagnosing disease, several studies have been made to address problems such as disease progression prediction, and more recently, survival analysis. Turner et al. (2010) have focused on the patients' diagnosis and tried to investigate the heterogeneity in *ALS*. Meanwhile, Carreiro, Madeira, and Francisco (2013) considered the database as a social network, and try to extract relevant information from related communities (groups of similar patients). In addition, Marwa et al. in Elamin et al. (2015) have presented an algorithm to generate prediction models based on the information gathered on patient's first physician visit. While their mechanism is beneficial, it doesn't use genetic data such as age, weight, gender et al., and requires patients to do more extensive cognitive tests. Thus, instead of only using the information of patient's first visit, Wang, Li, and Fang (2016) have proposed a novel predication method that combines a dimension reduction technique with Support Vector Regression (*SVR*) to predict the progression of *ALS* in the next 3–12 months based the data collected from patients over the most recent 3 months. While their method is robust and accurate, it is difficult to determine the optimal parameter values for *SVR* model with radial base function kernel.

In addition, Taylor et al. (2016) have used different methods including random forest (*RF*), pre-slope, and generalized linear