Journal of Biomedical Informatics 60 (2016) 187–196

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Analyzing health insurance claims on different timescales to predict days in hospital ☆

Yang Xie [a,*], Günter Schreier [b], Michael Hoy [a], Ying Liu [a], Sandra Neubauer [b], David C.W. Chang [a], Stephen J. Redmond [a], Nigel H. Lovell [a]

[a] The Graduate School of Biomedical Engineering, UNSW Australia, Sydney, New South Wales 2052, Australia
[b] AIT Austrian Institute of Technology GmbH, 8020 Graz, Austria

ARTICLE INFO

Article history:
Received 16 April 2015
Revised 5 January 2016
Accepted 5 January 2016
Available online 28 January 2016

Keywords:
Health care
Temporal data mining
Predictive modeling
Health insurance claims

ABSTRACT

Health insurers maintain large databases containing information on medical services utilized by claimants, often spanning several healthcare services and providers. Proper use of these databases could facilitate better clinical and administrative decisions. In these data sets, there exists many unequally spaced events, such as hospital visits. However, data mining of temporal data and point processes is still a developing research area and extracting useful information from such data series is a challenging task. In this paper, we developed a time series data mining approach to predict the number of days in hospital in the coming year for individuals from a general insured population based on their insurance claim data. In the proposed method, the data were windowed at four different timescales (bi-monthly, quarterly, half-yearly and yearly) to construct regularly spaced time series features extracted from such events, resulting in four associated prediction models. A comparison of these models indicates models using a half-yearly windowing scheme delivers the best performance on all three populations (the whole population, a senior sub-population and a non-senior sub-population). The superiority of the half-yearly model was found to be particularly pronounced in the senior sub-population. A bagged decision tree approach was able to predict 'no hospitalization' versus 'at least one day in hospital' with a Matthews correlation coefficient (MCC) of 0.426. This was significantly better than the corresponding yearly model, which achieved 0.375 for this group of customers. Further reducing the length of the analysis windows to three or two months did not produce further improvements.

© 2016 Elsevier Inc. All rights reserved.

1. Introduction

Hospitalization is usually the largest component of health expenditure [1]. Early identification of those with a higher risk of hospitalization could help in making efficient health resource management decisions. There exists work which attempts to predict hospitalizations for specific disease groups based on laboratory data and medical records. Niewoehner et al. [2] developed hospitalization risk indexes for chronic obstructive pulmonary disease (COPD) patients using their spirometry, demographics, and medical history data. Sugimoto et al. [3] discovered that serum intact parathyroid hormone levels obtained in outpatients with heart failure were shown to be an independent predictor of hospitalization. However, limited work has investigated the possibility of developing a general model which is not disease specific, which can set a benchmark against which disease-specific models may be compared. To our best knowledge, the most relevant recent work includes a data mining competition ('Heritage Health Prize') [4] and research done by the authors [5]. Both studies developed models to predict days in hospital for a general population based on health insurance claims.

In our previous work [5], we utilized bagged decision tree classifier models to predict the total number of days spent in hospital in the subsequent calendar year for individuals from a general population, using large-scale health insurance claims data. The proposed method performs well in the general population as well as different demographic sub-populations.

However, information available to perform the prediction comes from two sources; firstly, a list of customer attributes (e.g., customer demographics and insurance enrollment information), and secondly, the history of customer hospital admission claims. The latter can be considered as a time series of unequally

☆ This research was supported under the Australian Research Council's Linkage Projects funding scheme (Project No. LP0883728) and HCF Foundation.

* Corresponding author.
E-mail addresses: yang.xie@unsw.edu.au (Y. Xie), guenter.schreier@ait.ac.at (G. Schreier), n.lovell@unsw.edu.au (N.H. Lovell).

http://dx.doi.org/10.1016/j.jbi.2016.01.002
1532-0464/© 2016 Elsevier Inc. All rights reserved.

spaced events. Furthermore, each hospitalization event can itself be expressed as a sequence of hospital services utilized by the patient during that particular stay [5]. While it is straightforward to directly present the first type of information as features to a classification algorithm, data mining of time series and point processes, such as the hospitalization events described above, is still a developing research area. It is rather challenging to extract relevant information in a useful manner. The relative timing of recorded hospital visits used for prediction seems likely to contain subtle, yet valuable information which may lead to more accurate predictions if it can be harvested properly.

Reviews of methods applicable to a broader class of *temporal data mining* problems [6,7] highlight the following approaches which are applicable to the analysis of point process data:

1.1. Conversion to time series

When point processes occur relatively frequently, they can be converted to an equally-spaced time series. Prediction of future values from such time series is a well studied area (see [8,9]). However, this is not straightforward with the complex data encountered here, and information would be lost in the conversion process.

1.2. Estimation of the similarity between time series data

If a distance or similarity measure can be developed, then non-parametric classification approaches such as *K-nearest neighbors* (KNN) can be used. This returns a quantity based on the set of the closest matches in the training data. An example is *Edit Distance* (the number of alterations required to convert one time series to another) [10,11]. However, such methods may be too computationally costly for very large data sets like in the problem at hand.

1.3. Estimation of the underlying risk process

Poisson and other similar generalized linear models are often used for insurance purposes, and these are useful for determining the distribution of the possible number of claims [12,13]. Poisson regression makes the assumption that the response variable is drawn from a Poisson distribution. The method also assumes the logarithm of the Poisson distribution's expected value can be modeled by a linear combination of selected model parameters. Another approach taken is to model the system as a Cox process, which is a Poisson process where the expected value of the distribution changes over time in terms of the known factors [14]. However, both Poisson and Cox regression frameworks assume the response variable to be non-linearly and monotonically proportional to the explanatory variables. While useful, in exploratory studies like the one presented here, non-linear interactions and correlations between exploratory variables may be too complex for a log-linear model like this to capture, ultimately leading to poorer performance relative to more flexible pattern recognition models.

1.4. Symbolic point data mining

If all scalar values and perhaps time intervals between observations are discretized into categorical variables, the entire time series can be expressed as a sequence of symbols. The key step in such methods is defining a language that can adequately represent the temporal dimension of the data [15]. Most work relies on the use of temporal abstraction (TA) [16] and temporal logic [17], which allow the description of complex temporal patterns and temporal interactions among multiple time series. TA is the first step, which is the process of segmenting and aggregating time series data into explicit and symbolic representations, making it suitable for human decision making or data mining [18]. Next is mining these temporal patterns derived through TA, which is a relatively young research field. Most work mine temporal association rules (TARs), based on Allen's temporal relations, which represent temporal events using before/after chains (e.g., event A precedes/overlaps/finishes-by/contains event B) [17]. There have been several reports of the application of these methods to healthcare data sequences, which comprised hybrid events temporally interacting with each other (e.g., medications and physiological measurements) [15,19,20]. However, in the claim data set used for this study, no medication data, laboratory results or physiological measurements were available, and events were temporally sparse, with the average number of procedure claims per customer per year less than two (Table 2). Given the sparse time series and the lack of temporal heterogeneity between feature time series, further confounded by the vast number of potentially predictive features available in the big data set used here, it was decided not to pursue such an approach in this paper.

1.5. Heuristic time series features

Some features can be invented heuristically based on properties of the time series [21,22]. Indeed in one such study, they were found to give better performance compared to other model-based methods [21]. Our previous reported approach [5] falls into this category, in which most of the medical features extracted from properties of the hospital admission records and procedure claims were aggregated for each customer. This was found to give good performance and is meaningful as a first attempt to solve this problem. However, processing features in this way is expected to lead to a loss of temporal information.

The objective of this study is to develop a method which includes time information more explicitly and evaluate its performance on predicting the number of hospitalization days for a general population. Considering that the density of hospital admission events during a year is relatively sparse or zero for most of the claimants, a mixture of heuristic time series features and windowing was utilized. Specifically, features extracted from medical events are sorted into smaller time intervals to bring in more detailed information about the relative timing of events, which would potentially improve prediction performance. In addition, the time intervals are varied through a set of different scales, i.e., a year, half year, quarter, and two months. Our aim is to explore the impact of varying the temporal resolution on the predictive power of such models. At the same time, a model evaluation and comparison routine is proposed to assess whether the differences in the performance of different time scale models is statistically significant.

2. Methods

2.1. Data set

The data set consisted of 100,000 de-identified customers of the Hospitals Contribution Fund of Australia (HCF), one of Australia's largest combined registered private health fund and life insurance organizations. The 100,000 subjects were randomly selected from the HCF customer database. Only those customers who had enrolled with HCF before 1/1/2011 were included in the selection process. Three consecutive years of data, from 2011 to 2013, were provided for analysis.

These data contain tables of hospital admission administrative records and hospital procedure claims, as well as basic demographic information of customers [5]. Customer demographics include information related to customers, such as sex, age, the type

Download English Version:

https://daneshyari.com/en/article/6927842

Download Persian Version:

https://daneshyari.com/article/6927842

Daneshyari.com