一般口演
# 一般口演27
# データ活用・解析
2017年11月23日(木) 12:45 〜 14:15 D会場 (10F 会議室1002)

[4-D-2-OP27-1] 英国 NHSデータを用いた糖尿病医療費予測

由井 俊太郎[1], 三好 利昇[1], 大崎 高伸[1], 伴 秀行[1], ノーマン ステイン[2], シーラ マコンキンデール[3], マーティン ギブソン[2]（1.（株）日立製作所 研究開発グループ, 2.NorthWest EHealth Limited, 3.NHS Salford Clinical Commissioning Group）

英国では、糖尿病の NHS予算に占める割合が現在では10%であるが、2035年度には少なくとも17%になると予想され、問題が深刻化している。この問題に対し、予防は有効な手段であるが、その定量的な効果を得るには相当の時間がかかる。予防による医療費抑制効果を算出するため、我々は英国マンチェスター地区サルフォード市の NHSデータを用いて医療費予測技術を開発した。
人口24万人のサルフォード市には、家庭医と病院を統合したデータベースがあり、詳細な解析が可能である。私たちは最初の試みでマルコフモデルを用いた技術開発を行ったが、健診が無いため検査値に多くの欠損値があり、またモデルの拡張に労力がかかるという問題があった。
この問題を解決するため、私たちはベイジアンネットワークベースの技術開発を行った。複雑な関係性を記述できるため、欠損値があっても補間できる。さらにネットワークの自動構築も可能になる。
サルフォード市にある11869人のデータを用いて実験した結果、欠損値の問題を解決したため、医療費の予測が可能なデータ数が約2倍になり、予測誤差も5%以内の精度を達成した。これにより、欠損値の補間とネットワークの自動構築により、医療費予測が可能になる事を確認した。

英国 NHS データを用いた糖尿病医療費予測

由井 俊太郎[*1]、三好 利昇[*1]、大崎 高伸[*1]、伴 秀行[*1]、
ノーマン ステイン[*2]、シーラ マコンキンデール[*3]、マーティン ギブソン[*2*4]
*1（株）日立製作所 研究開発グループ、*2 NorthWest EHealth Limited、
*3 NHS Salford Clinical Commissioning Group、*4 Salford Royal NHS Foundation Trust

Future Diabetes Medical Cost Prediction Using UK NHS Data

Shuntaro Yui[*1], Toshinori Miyoshi [*1], Takanobu Osaki [*1], Hideyuki Ban [*1],
Norman Stein [*2], Sheila McCorkindale [*3] and Martin Gibson [*2*4]
*1 Hitachi Ltd. , *2 North West EHealth,
*3 NHS Salford CCG, *4 Salford Royal NHS Foundation Trust

Abstract

In the UK, diabetes is one of the fastest-growing health threats: Currently 10% of the NHS budget is spent on diabetes treatment and this is likely to rise to at least 17% by 2035, which is clearly unsustainable. Prevention could be effective for reducing cost burden though it takes much time to get the quantitative impact. In order to demonstrate the degree to which delaying or preventing the onset of type 2 diabetes could be economically effective, our approach is to establish a cost modelling framework using UK diabetes data in Salford, Greater Manchester. Salford has a population of 242,000 and an integrated electronic record system across both primary and secondary care. Using data above, we propose a new cost prediction method based on a Bayesian network. Experimental results using Salford diabetes data (11869 patients) showed that the amount of data available was twice larger because of missing data inclusion and less leverage of disease expansion, whilst predicted medical cost was £761 in which prediction error achieved less than 5%. It demonstrates that the proposed method can achieve future diabetes medical cost without complicated efforts by interpolation of missing data values and semi-automatic model construction.

Keywords: Decision Support, Cost Prediction, Bayesian Model

1.Introduction

In the UK, diabetes is one of the fastest-growing health threats: Currently 10% of the NHS budget is spent on diabetes treatment and this is likely to rise to at least 17% by 2035, which is clearly unsustainable. Prevention could be effective for reducing cost burden though it takes much time to get the quantitative impact. In order to demonstrate the degree to which delaying or preventing the onset of type 2 diabetes could be economically effective, our approach is to establish a cost modelling framework using UK diabetes data in Salford, Greater Manchester.

Salford has a population of 242,000 and an integrated electronic record system across both primary and secondary care which allowed the opportunity to capture detailed analysis across both spectrums of care at the same time. We had developed the above framework based on a Markov model; however, this framework faces two challenges 1) inability to predict for those whose indicators are missing, and 2) high degree of leverage for model construction.

2.Objective

In order to solve the challenge above, we propose a new cost prediction method based on a Bayesian network. The new method could interpolate significant missing data values because the new model describes the complex relationship between each indication. It also introduces a data-driven approach to model construction because it can achieve automatic model construction by learning from data.

3.Method: Disease Progression Model (DPM)

A DPM is a Bayesian network based technology which describes probabilistic models of complex systems characterised by the presence of multiple indicators.

Our alternative goal of this study is to achieve the following year's prediction using the initial year's data based on a Bayesian network. We start by assuming that we can use two years' data (initial year (y_i), following year (y_f), $y_f > y_i$). Figure 3.1 shows the overview of the network for this study.

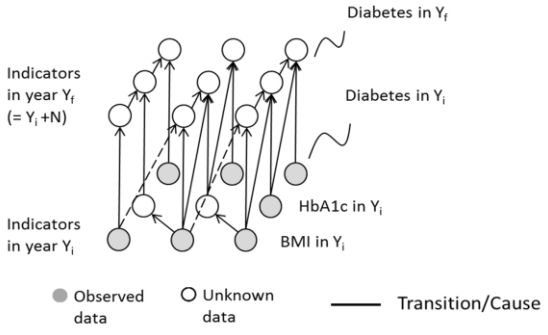

Figure 3.1: Network overview in disease progression

The DPM has three steps (Fig. 3.2). We first extract and discretise the data to set up the Bayesian network (Pre-processing). We then construct two classes of Bayesian network: one class represents the initial year y_i and the other represents the following year y_f (Model Structure Learning). Once the model has been created, we are finally able to

generate the following year's prediction from the model given the initial year's data (Inference).

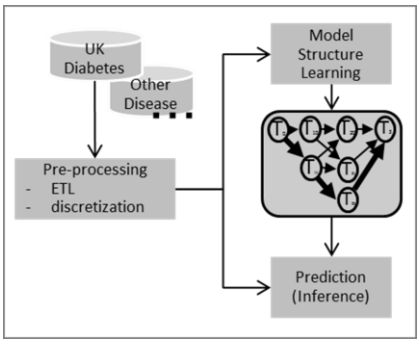

Figure 3.2: DPM approach

4.Evaluation

4.1 Cohort Inclusion Criteria and Data Characteristic

We worked with data from Salford, GM, UK: Salford has a population of 242,000 and an integrated electronic record system called the "Salford Integrated Record (SIR)" across both primary and secondary care which includes demographic, diagnoses, clinical observations, test results, operations and medication history. We developed a Bayesian model using approximately 12,000 records of patients who are diagnosed as IGR/type 2 diabetes in ICD10/Readcodes or who take diabetes medications from SIR (primary/secondary data in Salford) in Fiscal year 2010/11. We have also used cost of primary care, medication and secondary care.

Table 4.1 Cohort data characteristic (initial year)

Number of Patient	Average Age	Average BMI (Kg / m^2)	Average HbA1c (mmol / mol)
11869	62.1	31.6	54.4

4.2 Evaluation Method

We compare the predicted error between the proposed method and the conventional method (Markov model). We also compare the missing rate in order to show how much data we can newly predict. In order to compare with the conventional method, we use exactly the same data set for the proposed method as for the conventional method because the conventional method can deal with the IGR or type2 diabetes data only without missing data.

4.3 Result

Table 4.2 shows the average predicted medical costs and errors in the following year applying the Markov model (conventional method, see Figure 4.1) and the DPM (proposed method) respectively. We found that the predicted error in the DPM was improved from the Markov model. Table 4.3 shows the model cohort characteristic in each stage defined in the Markov model from missing data point of view. The column labelled "Whole data samples" indicates that, for example, there are 1148 IGR patients in Salford available for prediction. The column labelled "Available data samples in Markov model" indicates that, for example, there are 986 IGR patients in Salford whose future medical cost can be predicted using the Markov model. "Available data samples in Markov model" is less than "Whole data samples" because we selected disease progression factors using the multivariable polytomous logistic regression method in the Markov model which means that, in the Markov model, we can only predict the future cost of patients none of whose indicators are missing.

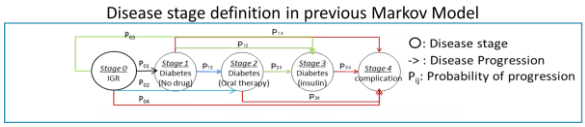

Figure 4.1: Disease stage definition in previous Markov Model

Table 4.2 Average predicted medical costs and errors in the following year (Stage0-3) (£)

Ground truth	Markov model: average (error)	DPM: average (error)
747.23	692.88 (7.3%)	761.00 (1.8%)

Table 4.3 Data availability comparison between Markov model and Bayesian model

	Disease progression factors using Markov model	Available Data samples in Markov model	Whole data samples (no-missing data included)
Stage 0 (IGR)	Fasting Blood Glucose, Age	986	1148
Stage 1 (newly diagnosed diabetic)	HbA1c, Age, Triglyceride, Family History	458	769
Stage 2 (Diabetes with oral agent therapy)	HbA1c, Age, Creatinine, Gender	4324	4911
Stage 3 (Diabetes with insulin)	HbA1c, Age, Creatinine, Family history	452	513
Other	-	0	4528
Total	-	6220	11869

5. Discussion

Result above showed the comparison with the Markov model in order to show the advantage of incorporating missing data. We found that the predicted error in the DPM was improved from the Markov model whilst the Markov model can predict only those patients whose indicators are not missing. In this evaluation, the number of patients who can be predicted in the

Markov model was only 52 % compared with the whole dataset.

Let us consider the advantage of leverage for model construction. Chapter 3 showed that the DPM can achieve automatic model construction by learning from data. This meant the DPM can support a data-driven approach to model construction compared to the Markov model based method. The DPM allows the user to create the model without complicated efforts.

6. Summary

This paper presented a new medical cost prediction method to support decision making for new healthcare improvement. We proposed a new cost prediction method based on a Bayesian network with automatic discretised function. Experimental results using Salford diabetes data (11869 patients) showed that the amount of data available was twice larger because of missing data inclusion and less leverage of disease expansion, whilst predicted medical cost was £761 in which prediction error achieved less than 5%. It demonstrates that the proposed method can achieve future diabetes medical cost without complicated efforts by interpolation of missing data values and semi-automatic model construction.

Reference

[1] State of the nation (England): challenges for 2015 and beyond. Available: https://www.diabetes.org.uk/About_us/What-we-say/Statistics/State-of-the-nation-challenges-for-2015-and-beyond/

[2] N.Hex et al., “Estimating the current and future costs of Type1 and Type 2 diabetes in the UK, including direct health costs and indirect societal and productivity costs,” Diabetic Medicine, 29, pp. 855–862, 2012.

[3] Daphne Koller and Nir Friedman, Probabilistic Graphical Models, The MIT Press.