KICS The Korean Institute of Communications and Information Sciences

Available online at www.sciencedirect.com

ScienceDirect

ICT Express 2 (2016) 150–154

www.elsevier.com/locate/icte

Smartphone-based personalized blood glucose prediction☆

Juan Li*, Chandima Fernando

Computer Science Department, North Dakota State University, Fargo, USA

Received 1 August 2016; received in revised form 6 October 2016; accepted 10 October 2016
Available online 25 October 2016

Abstract

Effective blood glucose control is essential for patients with diabetes. However, individual patients may not be able to monitor their blood glucose level regularly because of all manner of real-life interference. In this paper, we propose a personalized diabetes prediction mechanism that leverages smartphone-collected patient data and population data to drive personalized prediction. Unlike existing predictive models, this model utilizes pooled population data and captures patient similarities, and eventually produces a personalized blood glucose prediction for an individual. We have implemented the proposed model as a mobile application and have performed extensive experiments to evaluate its performance. The experimental results demonstrate that the proposed prediction mechanism can improve the prediction accuracy and remedy the problem of sparse data in the existing approaches.

Keywords: Diabetes; Smartphone; Blood glucose; Prediction; Personalized care

1. Introduction

Diabetes is becoming increasingly common around the world. In order to control blood sugar levels and prevent hypoglycemia, frequent blood glucose (BG) monitoring is needed by diabetes patients and their healthcare professionals. Diabetes control is aided by BG self-monitoring by facilitating the creation of an individualized BG profile. This profile can help healthcare professionals to draw up an individualized treatment plan for a particular patient. Moreover, it can also give diabetes patients and their families the ability to make appropriate day-to-day treatment choices about diet and physical activity, as well as about insulin or other agents [1].

However, frequent and regular monitoring of BG level is difficult and often impractical in a patient's daily life. The frequency with which diabetic patients should monitor their BG level varies from patient to patient. Most experts agree that insulin-treated patients should monitor BG at least four times a day, most commonly while fasting, before meals, and before bed. In some circumstances, however, patients are unable to maintain this frequent monitoring; for example, a patient may be in a meeting, may not have the testing devices with them, or may forget the test. For this reason, accurate prediction of BG level is very important for diabetes self-management.

There has been much research into automatic BG prediction using machine learning algorithms. For example, in the Artificial Pancreas Project [2], BG level is predicted so that the insulin flow can be continuously adjusted to meet patient needs. Despite the existing research, a challenge remains to predict BG level accurately and make personalized recommendations. Existing prediction models can be classified as either population-based prediction (e.g., [3–6]) or patient-based prediction (e.g., [7–10]). A population-based prediction of one patient's situation is based on independent population data. As shown in our experiments, the accuracy of this model in estimating individual risk is relatively low. Patient-based prediction normally requires a large amount of historical patient data to make a usable prediction. Sometimes, this requirement cannot be satisfied (e.g., for a new patient).

To address the aforementioned problems of existing approaches, our work shifts from using either population- or patient-based analysis and prediction to a synthesized population/patient-based analysis and prediction. We propose a personalized data-centric predictive model to predict patients'

* Corresponding author.
E-mail addresses: j.li@ndsu.edu (J. Li), chandima.fernando@ndsu.edu (C. Fernando).

Peer review under responsibility of The Korean Institute of Communications Information Sciences.

☆ This paper is part of a special issue titled Special Issue on Emerging Technologies for Medical Diagnostics guest edited by Ki H. Chon, Sangho Ha, Jinseok Lee, Yunyoung Nam, Jo Woon Chong and Marco DiRienzo.

http://dx.doi.org/10.1016/j.icte.2016.10.001

BG levels automatically based on their daily activity patterns collected from their smartphones, their historical information kept in their smartphones, as well as the historical population data. We leverage population data to drive personalized prediction. In particular, we propose a three-stage evolution model that includes a time-series regression model based on personal history, a pooled panel data (PPD) regression model, and a pre-clustered personalized regression model. A prediction system can choose an appropriate model from the three aforementioned models based on its data used for prediction.

2. Related work

There have been numerous studies of risk prediction related to diabetes management. In 1996, Shanker proposed the use of artificial neural networks to predict the onset of diabetes mellitus among the Pima Indian female population near Phoenix, Arizona [11]. Researchers have developed risk scores and predictive models for diabetes screening based on population studies [12]. More recently, Choi et al. developed two models to screen for prediabetes using an artificial neural network and support vector machine (SVM), and performed a systematic evaluation of the models using internal and external validation [13]. Zecchin et al. quantified the potential benefits of glucose prediction in terms of a reduction in the frequency/duration of hypoglycemia [14]. The Diabetes Support System (4DSS) designed by Ohio University tries to provide intelligent decision support systems for patients with type 1 diabetes (T1D) who are on insulin pump therapy [15].

Different machine learning algorithms have been proposed to predict BG level. Sandham et al. proposed the use of an artificial neural network and neuro-fuzzy systems to predict BG level for expert management of diabetes mellitus [8]. El-Jabali [16] used artificial neural networks to create a dynamic simulation model of T1D. Plis et al. [9] proposed a generic physiological model of BG dynamics to generate informative features for a support vector regression model to predict BG levels. Based on their experiments, this model outperforms diabetes experts at predicting BG levels, but its precision is still relatively low at about 42%. Decision trees (DTs) have been used for analysis and prediction in diabetes management [3,4]. For example, Pociot et al. proposed a novel DT-based analytical method to predict T1D mellitus [3]. Han et al. used DTs to build a model for diabetes prediction from the Pima Indians Diabetes dataset [4]. Sudharsan et al. applied random forest (RF), SVM, k-nearest neighbors, and naïve Bayes to predict hypoglycemia [5].

Based on the source of the training data, the aforementioned prediction models can be classified as either population-based (e.g., [3–6]) or patient-based (e.g., [7–10]). As already mentioned, they suffer from issues of low accuracy and/or sparse data.

3. Methodology

3.1. Data collection

Owing to the near-ubiquitous use of smartphones, we can use them to collect patient information such as BG measurements taken at regular intervals, and the corresponding daily events that impact the BG levels such as insulin, meals, exercise, and sleep. The ambient sensing features of a smartphone can help us collect some user information (e.g., exercise and sleep) automatically. Some other data (e.g., insulin dose) need to be entered manually. Because of its popularity, we mainly use touch-based gestures as the default input method. The collected data are preprocessed and automatically uploaded to the cloud.

3.2. Personalized prediction

We apply a three-stage evolution model to make more accurate and personalized BG predictions.

(1) Time-series prediction model based on patient data

Because smartphone-collected BG measurements have a natural temporal ordering, we can model the problem of predicting BG level as one of forecasting a time series. A time series is a sequence of data points, typically consisting of successive measurements made over a time interval. A time-series dataset differs from a regular one in that there is a natural ordering to the observations in the former. We confine our study to discrete time series. Our BG prediction problem is to use BG measurements up to time t to predict a future BG level at time $t+1$. This BG prediction problem can be modeled as follows:

$$\mathrm{BG}_{t+1} = f(\mathrm{BG}_t, \mathrm{BG}_{t-1}, \mathrm{BG}_{t-2}, \ldots, \mathrm{BG}_{t-n}). \quad (1)$$

For an observed BG series with n points, where t refers to the most recent observation and $t-n$ is the most distant observation, a future BG value at $t+1$ can be estimated with a function f. Function f is known as the model, and is used to obtain an estimated value BG_{t+1}.

(2) Pooled-panel-data regression model

The aforementioned time-series regression model makes predictions based on an individual patient's historical data. If the system has a large amount of historical data for the individual, this model can make better predictions. However, an individual patient may have very limited historical data. In the extreme case, the individual might not have any previous data at all. We understand that the inability to predict with time-series regression models is a small-sample issue. To address this issue, we propose to employ panel data to increase the sample size. A pooled panel data (PPD) regression model takes into account the fact that historical information of all patients enables cross-patient information sharing and thus overcomes issues related to data sparsity.

We choose linear regression to realize our PPD regression model for predicting BG level. Linear regression can model the relationship between the scalar dependent variable BG and the explanatory (or independent) variables. Let BG denote the dependent variable whose values we wish to predict, and let $X_1, \ldots, X_k$ denote the independent variables from which it is to be predicted; the value of variable X_i in period t (or in row t of the dataset) is denoted by X_{it}. For example, in our dataset, X_i could be insulin dose, hypoglycemic symptoms, meal ingestion,

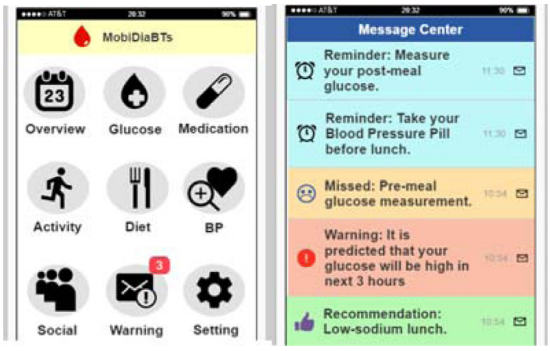

Fig. 1. Screenshots of the prototype system MobiDiaBTs.

or exercise activity. The equation for computing the predicted value of BG_t is

$$BG_t = b_0 + b_1 X_{1t} + b_2 X_{2t} + \cdots + b_k X_{kt}. \quad (2)$$

(3) Pre-clustered personalized regression model

The predictive performance of the PPD model may be influenced by the significant underlying model heterogeneity due to the variety of patients. To address this issue and improve the precision, we segment patients into groups whose prediction models are quite similar. In other words, we cluster patient data in order to group similar patients. The goal of clustering is to remove the influence of patients who have little or no similarity with the testing patient for whom predictions are being made. Clustering is based on similarities between patients, so that patients who are quite different from the testing patient are removed from his/her cluster. Patients in other clusters do not contribute to the prediction score of the testing patient. Thus, removing these patients does not result in loss of information, but effectively reduces the noise caused by these data. Clustering provides an additional benefit by reducing the number of data used for prediction, which both simplifies and improves the prediction results. In order to measure similarity between patients, an appropriate similarity measurement should be applied. The chosen measurement should correspond to the characteristics that distinguish the clusters embedded in the data. Candidate similarity measurement includes Euclidean distance, cosine similarity, Jaccard similarity, and Pearson similarity.

4. Implementation and evaluation

We have implemented the proposed model and integrated it into our prototype mobile application MobiDiaBTs [17]. Fig. 1 shows screenshots of the mobile application running on an iPhone 6 Plus. As shown, the mobile application provides various self-management functions to users. The "Overview" option gives users a summarized view of their historical health data. The "Glucose", "Medications", "Activity", "Diet", and "BP" options provide users with a friendly interface with which to input (or automatically collect) their health-related data. Among these functions, the "Warnings" component calls the proposed BG prediction model and automatically sends warning messages to users based on the predicted BG level.

Table 1
Comparison of different time-series regression approaches.

Metric	SVM	DT	RF
RMSE	68.76	41.06	39.73
MAE	63.097	36.423	37.586
R^2	0.05769283	0.29638461	0.79896364

We also performed experiments to evaluate the performance of the proposed model using the diabetes dataset from the UCI Machine Learning Repository [18]. This dataset includes 70 sets of data recorded on diabetes patients. Among the many criteria for comparing forecasting models, we choose the three most commonly used ones, namely mean absolute error (MAE), root-mean-square error (RMSE), and coefficient of determination (R^2) to evaluate the prediction performance. The MAE is the average error between the predicted BG level and the actual BG level:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |p_i - y_i|. \quad (3)$$

Here, p is the predicted BG level, y is the actual BG level, and n is the number of BG measurements taken.

The RMSE measures the standard deviation of the differences in the predicted BG level ($\hat{y}$) and the actual BG level (y):

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}. \quad (4)$$

The coefficient of determination, denoted as R^2, measures the proportion of the variance of the BG level, which is the dependent variable that can be predicted from the given features (i.e., the independent variables). This is an indicator that shows how well the data fit to the model:

$$R^2 = 1 - \frac{\text{Sum of squares of the residual}}{\text{Explained sum of squares}}. \quad (5)$$

The explained sum of squares is the sum of the squares of the differences of the predicted values and the mean value of the response variable y, the BG level.

The first set of experiments was carried out to analyze the performance of the patient-based regression model. Fig. 2 compares the prediction performance of three different time-series regression approaches, namely the SVM-based approach, the DT-based approach, and the RF-based approach. As illustrated, the accuracy of the SVM-based approach is not ideal. The RF model has the best prediction performance among the three.

Table 1 also lists the comparison of the three approaches using three different error metrics. We can see that both of the DT and RF models perform better than the SVM model in relation to RMSE, MAE, and R^2.

In the second part of the experiment, we studied the impact of sample data size (i.e., frequency of BG measurement) on prediction accuracy. We then verified the performance of the

Table 2
Comparison of pooled panel data (PPD)-based prediction with pre-cluster-based predictions.

Metric	Panel	2-cluster	3-cluster	5-cluster	9-cluster	43-cluster
RMSE	39.438	36.695	33.573	33.074	31.193	27.453
R^2	0.7681	0.8047	0.8014	0.8032	0.8156	0.8883

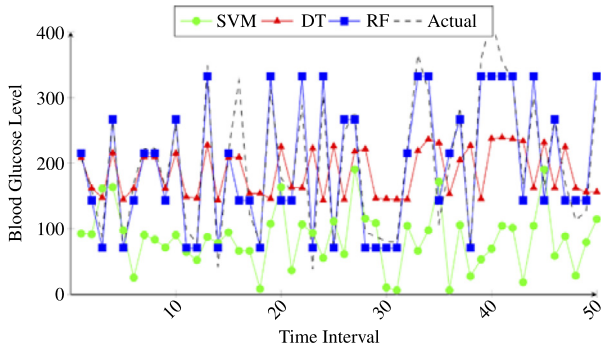

Fig. 2. Predicted blood glucose (BG) levels using support vector machine (SVM), decision tree (DT), and random forest (RF) methods.

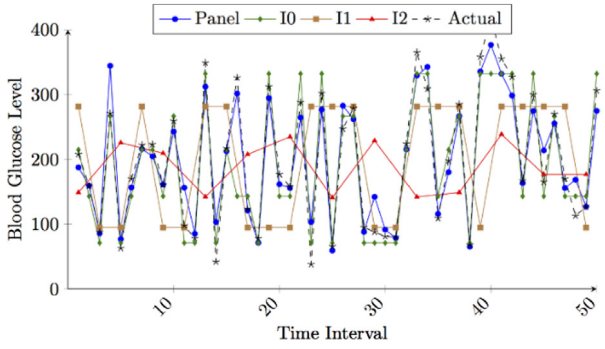

Fig. 3. Predicted BG levels using PPD regression and different-sized personal data (I0: patient-based prediction using all the samples in the dataset, I1: patient-based prediction with half the sample size. I2: patient-based prediction with a quarter of the sample size).

PPD regression model in remedying the data sparsity problem of the patient-based regression mode. Fig. 3 illustrates the predictive performance under different-sized patient data. In the figure, I0 represents patient-based prediction using all the samples in the dataset, I1 represents patient-based prediction with half the sample size, and I2 represents patient-based prediction with a quarter of the sample size. We reduced the sample size by reducing the BG measurement frequency. As shown in Fig. 3, as the frequency of an individual's BG measurements is reduced, the prediction accuracy decreases. Prediction using I2 (a quarter of the sample) deviated from the actual value, making it a useless prediction. Fig. 3 compares the performance of the regression model using PPD and the regression model using different-sized patient data. We can see that PPD can effectively remedy the small-sample-size problem by using the historical information of all the other patients to allow for cross-patient information sharing. This experiment tells us that, in the case of sparse personal data, we can turn to PPD as a solution.

In the third part of the experiment, we aimed to verify that pre-clustered regression would improve the prediction accuracy by providing more personalized prediction. Of course, like PPD-based prediction, this approach would also help to remedy the data sparsity problem associated with patient data. We use hierarchical clustering to group patients based on their similarity. Because of the limitation of the dataset, we can only measure the similarity based on feature patterns specified in the dataset. Table 2 compares the PPD-based prediction performance with that of personalized pre-cluster-based prediction with different cluster sizes. As can be seen from the table, the latter type of prediction is more accurate. The cluster-based prediction captures patient characteristics and can therefore make more personalized predictions. Table 2 also shows that the error rate continues to decrease when we refine the clusters. However, because of the small dataset with only 70 patients, the improvement due to clustering is not significant.

From these experiments, we can see that the proposed pre-cluster-based prediction can improve the predictive accuracy compared to that of PPD-based prediction (e.g., [3–6]). It can also remedy the data sparsity problem suffered by approaches that are based on the data of individual patients (e.g., [7–10]).

5. Conclusions

The global diabetes epidemic is a serious public health challenge. Close BG monitoring is important for diabetes self-management. However, patients are not necessarily able to measure their BG levels regularly. The goal of this study was to develop a personalized BG prediction model to predict patient BG levels accurately and automatically. Our work moves away from population or patient-based analysis and prediction to a synthesis of population and patient-based analysis and prediction. The proposed prediction models have been evaluated by extensive simulation experiments. The experimental results demonstrate that the proposed model improves the prediction accuracy and remedies the data sparsity problem of the existing models.

References

[1] Evan M. Benjamin, Self-monitoring of blood glucose: the basics, Clin. Diabetes 20 (1) (2002) 45–47.
[2] R. Hovorka, Artificial pancreas project at Cambridge 2013, Diabetic Med. 32 (8) (2015) 987–992.
[3] Flemming Pociot, Allan E. Karlsen, Claus B. Pedersen, Mogens Aalund, Jørn Nerup, European Consortium for IDDM Genome Studies, Novel analytical methods applied to type 1 diabetes genome-scan data, Am. J. Hum. Genet. 74 (4) (2004) 647–660.
[4] Jianchao Han, Juan C. Rodriguez, Mohsen Beheshti, Diabetes data analysis and prediction model discovery using rapidminer, in: 2008 Second International Conference on Future Generation Communication and Networking, vol. 3, IEEE, 2008, pp. 96–99.

[5] Bharath Sudharsan, Malinda Peeples, Mansur Shomali, Hypoglycemia prediction using machine learning models for patients with type 2 diabetes, J. Diabetes Sci. Technol. 9 (1) (2015) 86–90.
[6] S. Priya, R.R. Rajalaxmi, An improved data mining model to predict the occurrence of type-2 diabetes using neural networks, in: International Conference on Recent Trends in Computational Methods, Communication and Controls, ICON3C 2012, 2012.
[7] Alberto Pugliese, Mingder Yang, Irina Kusmarteva, Tiffany Heiple, Francesco Vendrame, Clive Wasserfall, Patrick Rowe, et al., The juvenile diabetes research foundation network for pancreatic organ donors with diabetes (nPOD) Program: goals, operational model and emerging findings, Pediatr. Diabetes 15 (1) (2014) 1–9.
[8] W.A. Sandham, D.J. Hamilton, A. Japp, K. Patterson, Neural network and neuro-fuzzy systems for improving diabetes therapy, in: Engineering in Medicine and Biology Society, 1998. Proceedings of The 20th Annual international Conference of The IEEE, vol. 3, IEEE, 1998, pp. 1438–1441.
[9] Kevin Plis, Razvan Bunescu, Cindy Marling, Jay Shubrook, Frank Schwartz, A machine learning approach to predicting blood glucose levels for diabetes management, in: Modern Artificial Intelligence for Health Analytics. Papers from the AAAI-14, 2014.
[10] Salim Chemlal, Sheri Colberg, Marta Satin-Smith, Eric Gyuricsko, Tom Hubbard, Mark W. Scerbo, Frederic D. McKenzie, Blood glucose individualized prediction for type 2 diabetes using iPhone application, in: 2011 IEEE 37th Annual Northeast Bioengineering Conference, NEBEC, IEEE, 2011, pp. 1–2.
[11] Murali S. Shanker, Using neural networks to predict the onset of diabetes mellitus, J. Chem. Inf. Comput. Sci. 36 (1) (1996) 35–41.
[12] Henry S. Kahn, Yiling J. Cheng, Theodore J. Thompson, Giuseppina Imperatore, Edward W. Gregg, Two risk-scoring systems for predicting incident diabetes mellitus in US adults age 45 to 64 years, Ann. Intern. Med. 150 (11) (2009) 741–751.
[13] Soo Beom Choi, Won Jae Kim, Tae Keun Yoo, Jee Soo Park, Jai Won Chung, Yong-ho Lee, Eun Seok Kang, Deok Won Kim, Screening for prediabetes using machine learning models, Comput. Math. Methods Med. 2014 (2014).
[14] Chiara Zecchin, Andrea Facchinetti, Giovanni Sparacino, Claudio Cobelli, Reduction of number and duration of hypoglycemic events by glucose prediction methods: a proof-of-concept in silico study, Diabetes Technol. Ther. 15 (1) (2013) 66–77.
[15] Cindy Marling, Matthew Wiley, Tessa Cooper, Razvan Bunescu, Jay Shubrook, Frank Schwartz, The 4 diabetes support system: A case study in CBR research and development, in: International Conference on Case-Based Reasoning, Springer, Berlin Heidelberg, 2011, pp. 137–150.
[16] A. Karim. El-Jabali, Neural network modeling and control of type 1 diabetes mellitus, Bioprocess Biosystems Eng. 27 (2) (2005) 75–79.
[17] Juan Li, Jun Kong, Cellphone-based diabetes self-management and social networking system for American Indians, in: 2016 18th International Conference on E-health Networking, Application & Services, HealthCom'16, IEEE, 2016.
[18] David Newman, UCI machine learning repository, 2007. http://archive.ics.uci.edu/ml/.