Towards Predictive Modeling for Automatic Carbohydrate Estimation and Early Hyperglycemia Detection in Type 2 Diabetes

Dr. Mudassir Rashid, Pulkita Jain

Introduction: Type 2 diabetes (T2D) is a heterogeneous disease with a significant degree of interpersonal variability that affects an estimated 34 million Americans. T2D is characterized by an increase in resistance to insulin, a decrease in insulin production and secretion, or some combination of these factors. This causes individuals with T2D to have difficulty controlling their blood glucose concentration (BGC) and experience periods of high (hyperglycemia) and low (hypoglycemia) BGC. Prolonged hyperglycemia can lead to chronic and severe health conditions over time. Early hyperglycemia warning systems based on continuous glucose monitoring (CGM) sensors may provide a convenient solution for monitoring and reducing the severity of hyperglycemia episodes. Continuous glucose monitoring sensors and machine learning algorithms can automate the process of meal size estimation, improve the accuracy of the carbohydrate estimations, and reduce the involvement of the subject. The aim of this project was to use Partial Least Squares (PLS) regression to model real-time data from CGM devices and optimize the model parameters for best generalized performance across all subjects.

Methods: A hyperglycemia prediction algorithm based on PLS regression and qualitative trend analysis has been developed. Hyperglycemia prediction proactives estimates in the future when a person's BGC will rise above a certain threshold, which was considered as a tunable parameter for this project, but is sometimes considered as 180 mg/dL. Real-time data from 135 patients was obtained, cleaned to filtered to reduce noise and rectify missing measurements. The data are then modelled through PLS regression. PLS relates regressor and regressed variables by maximizing the covariances between them. PLS builds linear relations between input data and output data and uses these relations to predict future values. PLS has *latent variables* that describe the important underlying features of the data. PLS was selected because it is a powerful tool that is convenient to implement and personalize for each subject. A matrix of past CGM data is used to handle the CGM time series, and the data are split into training and testing data. The training set is used to identify the model parameters and algorithm hyperparameters before evaluating the prediction algorithm on the independent testing set. Usually 80% of the dataset is designated for training and 20% is designated for testing so we used this allocation.

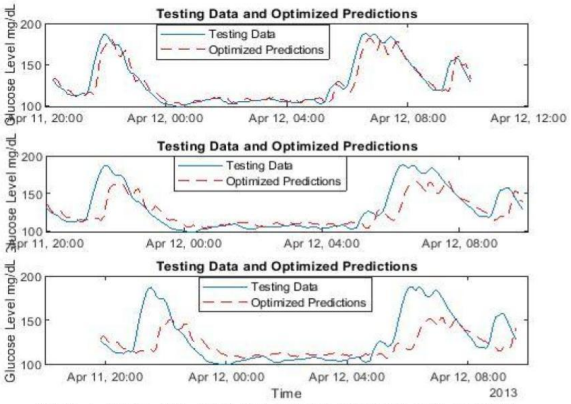

Figure 1. Testing data (blue) and predicted values(red) with optimized number of latent variables. a) FH =15 mins , b) FH =30 mins , c) FH =45 mins

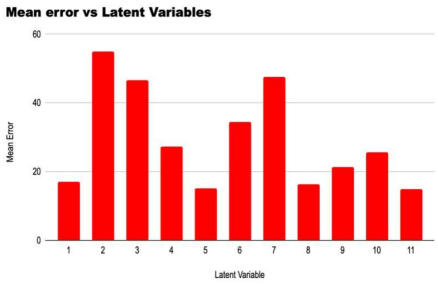

Figure 2. Mean Population Prediction Error vs. Number of PLS Latent Variables. The number of variables used was selected based on the number of lagged values used and the lowest error value.

Results: The two important parameters for our model to function well are the past horizon, defined as the number of past data points used to make future predictions, and number latent variables that are an integral part of the PLS model. These parameters were optimized by explicit enumeration and grid search approach to minimize the mean square error (MSE). We obtained the lowest MSE with the number of latent variables as 5 and past horizon as 7. Figure 2 shows the average MSE with respect to the different latent variables. Using these optimized parameters, we calculated 15-min-ahead, 30-min-ahead and 45-min-ahead predictions using the PLS model. Figure 1 shows these predictions of subject 1 along with the error we obtained. The errors obtained are as follows: 9.3946 for 15-min-ahead, 18.0048 for 30-min-ahead and 23.6090 for 45-min-ahead predictions.

Discussion: These findings suggest that the PLS model fits the real-time data well and demonstrates the associated errors with optimised parameters of past horizon and number of latent variables that will be used in the future research to develop this model further. By predicting accurately further into the future, we will be able to provide proactive early warnings to the user on impending hyperglycemia events. Future research will focus on developing the logic inference that will sound the alarms based on the predictions made in the current work. The logic inference will be developed using the first- and the second-order derivatives of the prediction curve. CGM sensors and machine learning algorithms will be able to automate the process of meal size estimation, improve the accuracy of the carbohydrate estimations, and help regulate glycemia in people with T2D.