Prediction of Glucose Concentration in Blood Plasma with Support Vector Regression Algorithm

Shanthi S

ABSTRACT--- Diabetes Mellitus is due to the disorder of glucose metabolism because of defects in insulin secretion or insulin action. It has become a major health challenge nowadays. Monitoring and regulation of blood glucose is inevitable to avoid diabetic complications. Prediction of near future glucose levels and giving alert for appropriate action could be done by machine learning techniques. This would greatly assist the diabetes patients in the daily management of diabetes. This paper discusses the effectiveness of Support Vector Regression in diabetes management. The methodology has been applied to three different data sets and performance measure is analyzed with Root Mean Square Error values.

Keywords: Diabetes Mellitus, Blood Glucose Prediction, Machine Learning,Support Vector Regression.

I INTRODUCTION

During last few decades, diabetes has become a major disease worldwide and hence an increasing measure of attention has been paid to it because of its social and economic implications. The Diabetes Mellitus (DM) disease is considered as the lack of ability of pancreas in regulating the Blood Glucose (BG) concentration. Higher BG levels,called as Hyperglycemia lead to long term complications in Heart, Kidney and Nervous systems. Lower BG levels known as Hypoglycemia lead to instant effects like attackin the brain activity and loss of consciousness for short term. The diabetic subjects strive hard to maintain their blood glucose levels in the normal range i.e 70 to 120 mg/dL[1]. Risks of Diabetes could be avoided by examining and regulating the BG properly as told by Diabetic Complications and Control Trial group[2]

Machine Learning (ML) is a field of Computer Science that has evolved due to the ability of computers to learn on its own without being programmed precisely [3]. The concepts like pattern recognition and computation learning theory in Artificial Intelligence (AI) leads to the evolution of ML [4]. The algorithms of ML can learn from large body of data and make predictions [5]. A variety of diagnostic and prognostic problems in medical domain could be solved by ML techniques. ML could be used to analyze clinical parameters which help in the assessment of disease progression, planning of therapy and for overall health care management. Findings suggest that ML would act as a guide in a new era in digital health care tools and lead to enhanced health care delivery. Several ML and data mining techniques have been used for the identification and managing of diabetes. This paper briefs about the ML techniques and their concept and how Support Vector Regression (SVR)method could be applied of for the prediction of BG levels ahead of certain prediction horizon.

1.1Machine Learning

The three broad categories of ML are discussed here.The first type is the Supervised learning. In this method,a function would be identified by the system from the categorised training data. Second type is the Unsupervised learning. In this method, the system learns by gathering the structure of uncategorized data. The third type is the Reinforcement learning.In this method, the system learns by interacting or cooperatingwith the dynamic environment. In supervised learning, the system would learn and develop a system with reference of target function.The function thus learnt is a countenance of a model defining the data. The function's objective is to forecast the value of the dependent parametervalue which is the output value, from a set of input variables which are the dependent variables, otherwise could be called as characteristics or features [6].

In supervised learning, learning of tasks could be done by 2 methods; Classification and Regression. The classification models attempt to calculate distinct modules or groups while regression models forecast numerical values. Some of the most common techniques are Decision Tree (DT), K-Nearest Neighbourhood (KNN), Genetic Algorithm (GA), Artificial Neural Network (ANN) and Support Vector Machine (SVM).

The unsupervised learning would reveal the concealed structure of data or relations between variables. In this, the training data would containexamples without any corresponding tags.

In reinforcement learning, a set of techniques would try to learn through straightcollaboration with the situations so as to get the most out ofit [7]. The system has no previousinformation about the actions of environment. It would find through trial and error. This method of learning has been realistic for sovereign system due to independence in relation to surroundings.

1.2 Support Vector Regression

For classification and regression, the standard method in ML is the SVM. Classification problems are most

Revised Manuscript Received on August 14, 2019.

Shanthi S, Professor, Department of Electronics and Communication Engineering, Saveetha University, Chennai, Tamilnadu, India. (Email: sshanthijj@gmail.com)

commonly done with SVMs. SVMs locate a hyperplane that best divides the data set into two classes. SVM outfits a learning algorithm for identifying subtle patterns in complex datasets. SVR is different from SVM in the way of predicting real values rather than categorizing as a class. SVR recognizes the occurrence of non-linearity in the data and provides a capable prediction model. This technique does not depend on any parameters. SVR depend on kernel functions rather than depending on spreading of the fundamentalreliant and autonomous variables. A nonlinear model would be constructedwith the permit of SVR without altering the descriptive values, helping in improvedversion of the resultingexemplary [8].

SVR will not care about the predictions as long as the error is a lesser amount than certain value. This is well-known as Principle Maximal Margin. SVR has been viewed as a convex optimization problem because of this maximal margin idea. The commonly used kernel functions in SVR are as follows; Linear, Polynomial, Sigmoid, Radial Basis Function (RBF). The kernel function transforms data from nonlinear space to linear space. The proposed work applies epsilon insensitive SVM (ε-SVM) regression. In ε-SVM regression, interpreter variables and experimental response values have been incorporated. This method of regression would discover a function f(x) that diverges from 'y' not greater than 'ε' for each preparation point 'x'. The training data 'x' would could contain multivariable set of 'n' remarks with experientialanswer values 'y', to find the linear function

$$f(x) = x'\beta + b \quad (1)$$

and make sure that it is as plane as possible, locate f(x) with the negligible norm value (β'β). Here the weight and bias parameters are β and b respectively. The above problem is devised as a convex optimization problem to curtail

$$J(\beta) = {}^{1}/_{2}\ (\beta'\beta) \quad (2)$$

This function is subjected to all residuals having a value less than ε. It could be expressed in equation form as follows,

For all n $$|Y_n - (x_n'\beta + b)| \leq \epsilon \quad (3)$$

It might be probable that there would not be any such function f(x) existsthat satisfiesall these restrictions for all the points. The variables called as slack variables ξ and ξ* are introduced for each point which deal with these constraints. Hence the objective function becomes the Primal formula,

$$J(\beta) = {}^{1}/_{2}\,(\beta'\beta) + C\sum n\,(\xi + \xi^{*}) \quad (4)$$

Subject to following conditions.

For all n,

$$yn - (xn'\beta + b) < \xi n + \xi n^{*} \quad (5)$$

$$(xn'\beta + b) - yn \leq \xi n + \xi n^{*} \quad (6)$$

$$\xi n^{*} \geq 0 \quad (7)$$

$$\xi n \geq 0 \quad (8)$$

Here the constant 'C' is said to be the box constraint which is a positive numeric value that would act as the penalty control for the observations which lie external to the margin of epsilon. The box constant helps to avoid over fitting. This value of C acts as the trade-off parameter between the flatness of f(x) and the deviations of ε.

II. MACHINE LEARNING IN DIABETES MANAGEMENT – RELATED WORKS

The given set of clinical cases acts as examples for ML techniques. The intelligent systems learn with those examples. These methods would be able to produce a regular description of the quantifiable characters that typify the medical situation. Researches have proved that different diseases could be diagnosed and analyzed with ML algorithms in accurate manner [9][10][11]. The patients could be classified as Diabetic or Non-Diabetic based on the threat factors by the Adaboost and Bagging ensemble ML methods in J48 decision tree [12]. The UCI Repository has been employed to train and test the prediction of Diabetes with Genetic Programming [13]. ANN has been used to predict diabetes–chronic disease [14]. With Linear Discriminant Analysis and Morlet Wavelet Support Vector Machine

(LDA-MWSVM), a system for diabetes diagnosis has been developed [15]. Ant Colony based classification has also been proposed for diabetes diagnosis [16]. A multivariate regression using SVR has been dealt for glucose prediction [17]. The subcutaneous glucose level of Type 1 Diabetes (T1D) has been predicted by multivariate analysis method. SVR technique has been utilized in another work of BG prediction with mobile platform [18]. Literature survey in this field leads to the wrapping up that SVR is appropriate for forecast of BG levels and only few works have been carried out so for. The remaining part of the paper addresses the methodology of applying SVR to the BG data sets and analyzes the performance of SVR.

III. METHODOLOGY

The glucose regulatory system has the intrinsic non linearity and non-stationaritycharacteristics. Hence the nonlinear regression methods could be applied for the efficient forecastof glucose concentration in blood.

3.1Data Sets

Three different data sets have been used to test the proposed methods. The first data set have been obtained from the Glucosim, a web based diabetes simulator developed by Illinois Institute of Technology [19][20]. For the given input conditions, this simulator gives the 24 hours blood glucose dynamics of a T1D Mellitus subject with 1 minute sampling frequency. Timing and dosage of Insulin and Carbohydrate (CHO) food intake, weight of a person and duration of exercise are the considered input parameters. The second data set had been from the glucose control project of the University of California, San Diego [21]. This glucose control project data base had BG values in 5, 10 and 20 minutes sampling interval for a one day period and some with two day period. The third data set include the CGMS data for 25 days culled from a T2D subject [22]. Actually

this data set has the collection of other physiological parameters values also which were intended to be used in modeling of Diabetes in free living conditions. However, the BG dynamics alone has been used for the proposed work.

3.2 Implementation of SVR

The regression algorithm was developed with MATLAB® (2013). The open source package libSVM had been used [23]. SVR for the estimation of BG levels was implemented with epsilon SVR and RBF kernel function. Since the nonlinear time series of BG dynamics would be better tracked by Gaussian distribution, the Gaussian RBF kernel has been utilized. The training process involved the application of Grid search and tenfold cross validation to ensure the minimum error. Setting the accurate values for meta-parameters 'C', 'ε' and kernel parameter 'γ' determines the SVR estimation accuracy.

The parameter C resolves the trade-off amid the model complexity and the degree to which the variations larger than ε are abided. The width of ε-insensitive zone could be controlled by the parameter ε which results in optimum fitting of training data. The number of support vectors used for the construction of regression function depends on the value of ε. Both C and ε have impact on the prediction model complexity. In literatures it has been proved that these parameters could be selected analytically from the training data and level of noise in the training data respectively [24]. For all training samples,

$$C \geq |f(x)| \quad (9)$$

$$\varepsilon \propto (\sigma/\sqrt{n}) \quad (10)$$

where σ is the standard deviation and n is the number of samples. The kernel width

parameter gamma is appropriately selected to reflect the input range of training and testing data. It could be (0.2 to 0.5)*range (x). The model has been trained with these parameters, which is then used for prediction of near future BG values. The data from three data sets have been sampled with five minute frequency for uniformity and applied to the model and trained to predict the 6 step ahead and 12 step ahead predictions i.e., 30 and 60 minutes prediction horizons respectively. Recursive method has been adopted. The glucose metabolism of each specific patient data has been learnt by the implemented SVR. And at the given prediction horizon (PH), the prediction of BG value has been done for each and every individual subject.

IV. RESULTS AND DISCUSSION

The SVR model has been implemented as discussed in previous section and its performance has been analyzed with the metric of Root Mean Square Error (RMSE) between the estimated values and correct values of target. RMSE has been chosen as the metric for measuring the concert of the prediction model, because it is the popular loss function ensured by the researchers. The performance of the prediction model was analyzed in prediction horizons (PH) of 30 minutes and 60 minutes. The effectiveness of the model has also been evaluated in varied data ranges i.e., in normal, hypo and hyperglycemic ranges. The RMSE values of the SVR model for certain datasets selected in random have been given table 1

Data Set	Id	RMSE in mg/dL	
		30 Minutes PH	60 Minutes PH
Data Set Obtained through Glucosim Simulator	# 005	8.67	14.20
	# 020	11.05	16.85
	# 057	6.51	11.78
	# 089	9.53	13.56
	# 100	10.51	14.78
Data Set from Glucose Control Project of UCSD	# 108	7.36	15.09
	# 136	12.25	16.74
	# 168	9.03	13.49
	# 185	11.05	15.48
	# 199	10.62	14.73
Data Set from CGMS	# 210	5.61	10.37
	# 243	7.96	15.62
	# 261	11.35	17.40
	# 285	6.54	11.79
	# 300	9.68	14.63

Table 1.Sample Results of the Proposed Prediction Model

We could observe that the RMSE values are higher in 60 minutes predictions which revealed that the efficiency of the prediction model would depreciate with time. Incorporation of additional physiological features has to be made for making prediction in longer time durations. And the results do not show noteworthy disparity in the varied data sets. The capability of the projected prediction model has also been tested in different data ranges viz Normal (70 – 120 mg/dL), Hypo (< 70 mg/dL) and Hyperglycemia (>120mg/dL) for 30 minutes and 60 minutes PH. The average RMSE values in the three datasets have been specified in table 2.

Data Set	Normal Range		Hypoglycemic Range		Hyperglycemic Range	
	30 Minutes PH	60 Minutes PH	30 Minutes PH	60 Minutes PH	30 Minutes PH	60 Minutes PH
Glucosim Simulator	5.35	8.27	8.56	14.76	6.93	11.23
Glucose Control Project of UCSD	7.62	10.44	9.15	15.68	8.27	10.86
CGMS	6.97	9.82	9.73	17.09	7.18	12.39

Table 2.RMSE Values (in mg/dL) in Normal, Hypo and Hyperglycemic Ranges

The performance of the prediction models was almost similar in normal and hyperglycemic ranges. When the comparisons were made more specifically, the RMSE values at hyperglycemic range were greater than that of normal range. However, all the three data sets exhibited higher error rates in hypoglycemic range. This could be due to the physiological reasons. That is, when the blood glucose level is decreasing, the glucose stored in the liver is pumped out for balance action which leads to sudden increase BG which would be missed by the prediction model.

V. CONCLUSION

Apart from food intake and insulin tolerance, many variables such as constant worry, substantial action, hormonal alterations and phases of growth, therapeutic procedures, sickness, infectivity, tiredness etc... affect the Glucose levels. In short we can say the BG dynamics is subjected to several sources of disturbances. And the impact of these disturbances on BG level is highly related, forceful and non-linear. Discrimination of effects of each input in the BG fluctuations is very much difficult. Accordingly, the proposed model could be improved in such a way that it would be able to get into description of the instantaneous and multiple effects of calorie intake, physical actions, anxiety and their relations. The accuracy of the proposed prediction model could further be increased by optimizing the SVR parameters with suitable technique.

REFERENCES

1. American Diabetes Association: Diagnosis and classification of diabetes mellitus. Diabetes Care, (2011) vol. 34, no. 1, pp. 62–69.
2. Nathan, D.M., Cleary P.A., Backlund, J.Y., Genuth, S.M., Lachin, J.M., Orchard, T.J., Raskin, P., Zinman, B. : Intensive diabetes treatment and cardiovascular disease in patients with type 1 diabetes. New England Journal of Medicine. (2005) vol. 353, no. 25, pp. 2643-2653.
3. Samuel A.: Some studies in machine learning using the game of checkers. IBM J Res Dev. (1959).
4. http://www.britannica.com/EBchecked/topic/1116194/machine-learning.
5. Kohavi, R., Provost, F.: Glossary of terms. Mach Learn.(1998). 30:271–4.
6. Russell, Stuart; Norvig, Peter : Artificial Intelligence: A Modern Approach (2nd ed.). (2003) Prentice Hall.
7. Alpaydin, E.: Introduction to Machine Learning The MIT Press, Cambridge Massachusetts London England (2004).
8. DebasishBasak, Srimanta Pal and Dipak Chandra Patranabis. : Support Vector Regression. Neural Information Processing – Letters and Reviews. (2007). Vol. 11, No. 10.
9. Aishwarya, R., Gayathri, P., Jaisankar, N.: A Method for Classification Using Machine Learning Technique for Diabetes. International Journal of Engineering and Technology (IJET) .(2013). 5, 2903–2908.
10. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., Chouvarda, I.: Machine Learning and Data Mining Methods in Diabetes Research. Computational and Structural Biotechnology Journal. (2017).15, 104–116.
11. DhomseKanchan B., M.K.M.: Study of Machine Learning Algorithms for Special
12. Disease Prediction using Principal of Component Analysis, in: 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication. (2016).IEEE.pp.5–10.
13. Perveen, S., Shahbaz, M., Guergachi, A., Keshavjee, K.: Performance Analysis of Data Mining Classification Techniques to Predict Diabetes. Procedia Computer Science. (2016). 82, 115–121.
14. Bamnote, M.P., G.R.: Design of Classifier for Detection of Diabetes Mellitus Using Genetic Programming. Advances in Intelligent Systems and Computing. (2014). 1, 763–770.
15. Tarik A. Rashid, S.M.A., Abdullah, R.M., Abstract.: An Intelligent Approach for Diabetes Classification, Prediction and Description. Advances in Intelligent Systems and Computing. (2016). 424, 323–335.
16. DuyguÇalisir, EsinDogantekin: An automatic diabetes diagnosis system based on LDA-Wavelet Support Vector Machine Classifier. Expert Syst. Appl. (2011). 38(7): 8311-8315
17. MostafaFathiGanji, Mohammad SanieeAbadeh: A fuzzy classification system based on Ant Colony Optimization for diabetes disease diagnosis. Expert Syst. Appl. (2011). 38(12): 14650-14659.
18. Eleni I. Georga, Vasilios C. Protopappas, Diego Ardigò, Michela Marina, Ivana
19. Zavaroni, DemosthenesPolyzos, Dimitrios I. Fotiadis: Multivariate Prediction of
20. Subcutaneous Glucose Concentration in Type 1 Diabetes Patients Based on Support Vector Regression. IEEE J. Biomedical and Health Informatics. (2013) . 17(1): 71-81
21. Bunescu,R., Struble, N., Marling, C., Shubrook, J., Schwartz, F. : Blood Glucose Level Prediction Using Physiological Models and Support Vector Regression, 12th International Conference on Machine Learning and Applications. (2013). vol. 1, pp. 135–140.
22. Agar, B., Eren, M., Cinar, A.: Glucosim: Educational software for virtual experiments with patients with type 1 diabetes. Proceedings of the Annual Int. Conf. IEEE
23. Engineering in Medicine and Biology.(2005). pp. 845 – 848.
24. 20. Glucosim. Diabetes simulator available from: <http:// 216.47.139.198/glucosim/index.html>. (2006).
25. 21. UCSD.University of California San Diego.Available from <http:// glucosecontrol.ucsd.edu>. (2008).