International Journal of Management, Economics and Social Sciences
Special Issue-International Conference on Medical and Health Informatics (ICMHI 2017)
2018, Vol. 7(S1), pp.132 – 141.
ISSN 2304 – 1366
http://www.ijmess.com

A Comparative Analysis of Data Mining Techniques for Prediction of Postprandial Blood Glucose: A Cohort Study

Huan-Cheng Chang[1]
Pin-Hsiang Chang[2]
Sung-Chin Tseng[3]
[*]Chi-Chang Chang[4]
Yen-Chiao Lu[5]

[1]Dept. of Community Medicine, Landseed Hospital, Taoyuan, Taiwan
[2] Dept. of Healthcare Management, Yuanpei University of Medical Technology, Hsinchu, Taiwan
[3]Div. of Family Medicine, Chiayi Chang Gung Memorial Hospital, Chiayi, Taiwan
[4]School of Medical Informatics, Chung-Shan Medical University/Hospital, Taichung, Taiwan
[5]School of Nursing, Chung-Shan Medical University, Taichung, Taiwan

The use of advanced predictive techniques and reasoning models has greatly assisted clinicians in improving the diagnosis, prognosis, and treatment of diabetes. Although numerous studies have focused on the relationship between abnormal blood glucose levels and diabetes, few have focused on the risk forecasting of postprandial blood glucose levels in patients with diabetes. This work aimed to develop a model for the prediction of postprandial blood glucose levels to screen for undiagnosed diabetes cases in a cohort study. The performance of the proposed model was then compared with those of five other data-mining techniques: random forest (RF), support vector machine (SVM), C5.0, multilayer perceptron (MLP), and logistic regression (LR). The data of 1,438 patients who were admitted to Landseed Hospital, Northern Taiwan, over the period of 2006 and 2013 were collected and used to evaluate the performances of the data-mining techniques. Compared with the 4.5, SVM, MLP, and LR models, the RF model had the best prediction capability for postprandial blood glucose levels in terms of the overall correct classification rate. The results of this study underscore the importance of identifying the preclinical symptoms of abnormal blood glucose levels. The proposed model provides precise reasoning and prediction and can be used to help physicians improve the diagnosis, prognosis, and treatment of patients with diabetes.

Keywords: Data mining techniques, random forest, support vector machine, multilayer perceptron, logistic regression

The global prevalence of diabetes mellitus, commonly referred to as diabetes, has drastically increased (Liu et al., 2017). Consequently, dialysis treatment for diabetic nephropathy has become a large burden on the national health insurance of Taiwan. The early diagnosis of the risk factors related to changes in postprandial blood glucose levels could help prevent or delay diabetic nephropathy. Moreover, early diagnosis may

Manuscript received May 24, 2017; revised June 20, 2017; accepted July 20, 2017. © The Author(s); CC-BY-NC; Licensee IJMESS
*Corresponding author: changintw@gmail.com

improve the outcomes of patients with diabetes, and the regular screening of blood glucose levels and blood pressure can decrease the incidence of diabetes. Screening for undiagnosed diabetes through blood sampling, however, is prohibitive because of the high costs and invasiveness of the technique. Accurate and precise reasoning and prediction models may greatly help physicians improve the diagnosis, prognosis, and treatment of diabetes. Several studies have been conducted to clarify the response of glucose levels in diabetic patients to various stimuli. Several factors affect the postprandial levels of blood glucose. These factors include age, weight, waist girth, white blood red blood cell counts, and globulin, high-density lipoprotein, and urine red blood cell concentrations.

Data-mining techniques have been widely used to predict blood glucose levels. The use of data-mining techniques to construct prediction models for blood glucose levels does not require strong model assumptions and can capture delicate underlying patterns and relationships in empirical data, hence providing promising results for the prediction of blood glucose levels. Although data-mining techniques have been utilized in numerous studies to predict fasting blood glucose and/or postprandial blood glucose levels, few studies have attempted to utilize data-mining techniques to predict or classify postprandial blood glucose as normal or abnormal. Moreover, most existing studies on blood glucose levels in diabetic patients are based on a continuous glucose monitoring system, a device that is installed on the patient for measuring the patient' s blood glucose over specific intervals. To the best of our knowledge, no study has utilized data-mining for the prediction of postprandial blood glucose levels in a cohort study. Therefore, a model for the prediction of postprandial blood glucose levels was proposed and designed in this study. The predictive performance of the proposed model was compared with those of five data-mining techniques.

The five data-mining methods used in this study are random forest (RF), support vector machine (SVM), C5.0, multilayer perceptron (MLP), and logistic regression (LR). RF is an ensemble learning method that grows multiple random tree classifications to generate an overall classification. SVM is based on statistical learning theory and is derived from the structural risk minimization principle for estimating a hyperplane for classification. C4.5 is a non-parametric and fast classification technique that adopts a greedy approach and uses a top-down recursive divide-and-conquer strategy to construct a decision tree. MLP is a neural network

commonly used to solve classification problems and is trained with a backpropagation algorithm. MLP also utilizes a supervised learning technique to transform sets of input data into a desired output. LR is a widely used statistical modeling technique that is a special case of the linear regression model. The major advantage of this approach is that it can produce a simple probabilistic formula of classification. These five data-mining techniques have been used to predict blood glucose levels. However, to the best of our knowledge, these five models have not been used to predict postprandial blood glucose levels in a cohort study.

Tresp et al. (1999) utilized recurrent neural networks and time-series convolution neural networks to predict the blood glucose levels of patients with diabetes. The recurrent neural network combined with the linear error model exhibited excellent performance and outperformed the compartment and time-series convolution neural-network models. Wang et al. (2016) used an improved grey (1, 1) model to predict the postprandial blood glucose levels of patients with type 2 diabetes using limited data. The improved grey model outperformed the autoregressive (AR) model in the prediction of blood glucose levels. Wang and An (2014) applied a least-squares-based AR model to predict blood glucose levels. The model accurately illustrated the changes in blood glucose levels to provide an early warning for the occurrence of low blood glucose. García-Jaramillo et al. (2013) adopted and compared the performance of three interval models in predicting the postprandial blood glucose levels of patients with type 1 diabetes under the conditions of uncertainty and intra-patient variability. The rest of this paper is organized as follows. A brief review of related works is presented in Section 2. RF, SVM, C4.5, MLP, and LR are introduced in Section 3. The experimental results are provided in Section 4, and the conclusion is provided in Section 5.

METHODOLOGY

-Random Forest

Random forest (RF) is a supervised machine learning algorithm which combines classification method based on the un-weighted majority of class votes (Breiman, 2001). In a RF, first, multiple random samples of variables are selected as the training dataset using the bagging procedure. The bagging procedure means

random sampling with replacement which is a meta-algorithm can be used to reduce variance and aids to elude over-fitting synchronously. Then, the tree-type classifiers corresponding to selected samples are constructed in the data training process. A large number tree makes RF from the selected samples. Finally, all classification trees are combined and final classification results are obtained by voting on each class and then choosing the winner class in terms of the number of votes to it. The RF performance measure by a metric called out of bag error calculated as the average of the rate of error in each weak learner. In RF, each individual tree is explored in a particular way. First, given a set of training data N, n random samples with repetition (Bootstrap) are taken as training set by using bagging procedure. Then, for each node of the tree, M input variables are determined, and m variables (<<M) are selected for each node. The most important variable randomly chosen is used as a node. The value of m remains constant. Finally, each tree is developed to its maximum expansion. Please refer to Breiman (2001) for more detail information of RF.

-Support Vector Machine

The basic idea of SVM initially, linearly or non-linearly, map the input vectors into a higher dimensional feature space. Then, SVM seeks an optimized hyperplane to separate two classes in the feature space. A description of SVM algorithm is follows. Let $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, $\mathbf{x}_i \in R^d$, $y_i \in \{-1, 1\}$ is the training set with input vectors and labels. Where, N is the number of sample observations and d is the dimension of each observation, y_i is known target. SVM is to seek the hyperplane $\mathbf{w} \cdot \mathbf{x}_i + b = 0$, where w is the vector of hyperplane and b is a bias term, to separate the data from two classes with maximal margin width $2/\|\mathbf{w}\|^2$, and the all points under the boundary is named support vector. For optimal the hyperplane, SVM is to solve the following optimization problem (Vapnik 2000).

$$\text{Min} \quad \Phi(\mathbf{x}) = \frac{1}{2}\|\mathbf{w}\|^2 \qquad (1)$$

$$\text{S.t.} \quad y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1, i = 1, 2, ..., N$$

As it is hard to solve eq.(1), it is transformed to be dual problem by using Lagrange method. The value of α in the Lagrange method must be non-negative real coefficients. The eq. (1) is transformed into the following constrained form,

$$\text{Max} \quad \Phi(\mathbf{w}, b, \xi, \alpha, \beta) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1, j=1}^{N} \alpha_i \alpha_j y_i y_j \mathbf{x}_i{}^T \mathbf{x}_j \quad (2)$$

$$\text{S.t.} \sum_{j=1}^{N} \alpha_j y_j = 0, \quad 0 \le \alpha_i \le C, i = 1, 2, ..., N$$

In eq. (2), C is the penalty factor and viewed as a tuning parameter which can be used to control the trade-off between maximizing the margin and the classification error. In general, it could not find the linear separate hyperplane in all application data. In the non-linear data, it must transform the original data to higher dimension of linear separate is the best solution. The higher dimension is called feature space, it improve the data separated by classification. The common used kernel function is radial basis function (RBF). It is applied in this study. For more details about SVM, please refer to Vapnik (2000).

-C 4.5

C4.5 classifier is a process for the classification and retrieves useful information in the form of a decision tree. The algorithm adopts a greedy approach in which the decision trees are constructed in a top-down recursive divide and conquer manner on the basis of a training set (Quinlan 1993). C4.5 builds decision trees from a set of training data based on the concept of information entropy. The training data is a set of already classified samples. Each sample is a vector including attributes or features. The training data is augmented with a vector representing the class that each sample belongs to. Each attribute of the data can be used to make a decision. C4.5 examines the normalized information gain that results from choosing an attribute for splitting the data. The attribute with the highest normalized information gain is the one used to make the decision. The algorithm then recurs on the smaller sub-lists. For more details about C4.5, please refer to (Larose 2005).

-Multilayer Perceptron

Multilayer Perceptron (MLP) is gained their popularity due to it is a simple architecture but a powerful problem-solving ability. Back propagation is a general supervised method for iteratively calculating the weights and biases of the MLP. This type of model is termed BPN. BPN uses a steepest descent technique with learning and momentum terms. A BPN topology consists of a number of nodes (neurons) connected by links and consists of three layers: input layer, hidden layer(s) and output layer. The nodes in the input layer

receive input signals from an external source and the nodes in the output layer provide the target output signals. Any layers between input and output layers are called hidden layers. Since one hidden layer network is sufficient to model any complex system with desired accuracy the designed BPN model in this study will have only one hidden layer. A three-layer BPN is used in this study. In a BPN topology, each layer comprises several neurons that are interconnected by sets of weights. The neurons obtain inputs from initial inputs or interconnections and generated outputs using a nonlinear transfer function. BPN uses gradient steepest descent training algorithm to minimize error and adjusts interconnection weights. For the gradient descent algorithm, the step size, called the learning rate, must be specified first. The learning rate is crucial for BPN since smaller learning rates tend to slow down the learning process before convergence while larger ones may cause network oscillation and unable to converge. Please refer Haykin (1999) for more details about MLP.

-Logistic Regression

LR is similar to a linear regression model but is suited to models where the dependent variable is dichotomous. A logistic regression model specifies that an appropriate function of the fitted probability of the event is a linear function of the observed values of the available explanatory variables. In producing the LR equation, the maximum-likelihood ratio was used to determine the statistical significance of the variables. LR is useful for situations in which can be able to predict the presence or absence of a characteristic or outcome based on values of set of predictor variables. LR model for p independent variables can be written as

$$H(Y=1)=\frac{1}{1+e^{-(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p)}} \qquad (3)$$

where $P(Y=1)$ is probability of presence. And $\beta_0,\beta_1,\beta_2,\cdots,\beta_p$ are regression coefficients. There is a linear model hidden within the logistic regression model. The natural logarithm of the ratio of $P(Y=1)$ to $1-P(Y=1)$ gives a linear model in X_i:

$$g(x)=ln\left(\frac{P(Y=1)}{1-P(Y-1)}\right)=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p \qquad (4)$$

The $g(x)$, has many of the desirable properties of a linear regression model. The independent variables can be a combination of continuous and categorical variables.For more details about logistic regression, please refer to Hosmer et al. (2013).

Data Collection

We collected data from LandSeed Hospital for 2006-2013. After excluding some follow-up records (e.g., records for patients age<18), we obtained records for patients who visited the hospital on three separate occasions over the period of 2006 and 2013 for two consecutive years. The patients had normal postprandial blood glucose levels at the first visit and may have abnormal postprandial blood glucose levels at the third visit. The data included 1438 clinical follow-up records. Of these records, 438 patients reported abnormal postprandial blood glucose at the third visit. Previous studies have studied the incidence and risk factors associated with diabetes. In this study, each subject in the dataset contained 29 predictor variables, as shown in Table 1, and the response variable is whether the postprandial level of blood glucose is normal or not. The performances of the five data-mining methods were evaluated using the 10-fold cross-validation method. The data-mining software WEKA, which was developed by Frank et al. (2016), was utilized to develop the RF, C4.5, SVM, MLP, and LR models with default settings for each algorithm.

RESULTS

Sensitivity and specificity are the two important measures in medical/healthcare classification. Sensitivity (also called the true-positive rate) is a measure of the proportion of positives that are correctly identified, and specificity (also called the true-negative rate) is a measure of the proportion of negatives that are correctly identified. The correct classification rate (CCR), sensitivity, and sensitivity were used as the three indexes for judging the performance of the five classification methods.

The classification results for postprandial blood glucose levels (the confusion matrix) predicted by the RF model are summarized in Table 1. From the results presented in Table 1, we can observe that the overall CCR is 82.68%. That is, {1-1} is 923 (CCR of 92.30%) and {2-2} is 266 (CCR of 60.73%). {1-1} represents sensitivity and indicates that a class 1 subject, which is a subject with normal postprandial blood glucose levels, is correctly classified into class 1. {2-2} represents specificity and indicates that a class 2 subject, which is a subject with abnormal postprandial blood glucose levels, is correctly classified into class 2.

Tables 2–5 show the classification results of C4.5, SVM, MLP, and LR, respectively. Table 2 shows that

the overall CCR of the C4.5 method is 76.56% with a sensitivity of 85.90% and specificity of 55.25%. Table 3 depicts that the CCR of the SVM method is 69.61% with a sensitivity of 99.20% and specificity of 2.05%. The CCR, sensitivity, and specificity of MLP model are 75.73%, 83.20%, and 56.68%, respectively, as shown in Table 4. As shown in Table 5, the CCR of the LR method is 74.48% with a sensitivity of 84.00% and specificity of 52.74%.

	Classified Class	
Actual Class	1 (normal)	2 (abnormal)
1 (normal)	923 (92.30%)	77 (7.70%)
2 (abnormal)	172 (39.27%)	266 (60.73%)
	Overall CCR : 82.68%	

Table 1. Classification Results Using RF Model

	Classified Class	
Actual Class	1 (normal)	2 (abnormal)
1 (normal)	529 (85.90%)	141 (14.10%)
2 (abnormal)	196 (44.75%)	242(55.25%)
	Overall CCR : 76.56%	

Table 2. Classification Results Using C4.5 Model

	Classified Class	
Actual Class	1 (normal)	2 (abnormal)
1 (normal)	992(99.20%)	8 (0.80%)
2 (abnormal)	429 (97.95%)	9(2.05%)
	Overall CCR : 69.61%	

Table 3. Classification Results Using SVM Model

	Classified Class	
Actual Class	1 (normal)	2 (abnormal)
1 (normal)	832(83.20%)	168 (16.80%)
2 (abnormal)	181 (41.32%)	257(56.68%)
	Overall CCR : 75.73%	

Table 4. Classification Results Using MLP Model

The summarized results of the five constructed models are shown in Table 6 and were used to evaluate their capability to predict postprandial blood glucose levels. From the data shown in the Table, we can conclude that the RF model has the best capability to predict postprandial blood glucose levels in terms of

the overall CCR. The SVM model has the highest sensitivity of 99.20% but has the lowest specificity of 2.05%. The RF model generated the highest specificity of 60.73% and the second-highest sensitivity of

	Classified Class	
Actual Class	1 (normal)	2 (abnormal)
1 (normal)	840(84.00%)	160 (16.00%)
2 (abnormal)	207 (47.26%)	231(52.74%)
	Overall CCR: 74.48%	

Table 5. Classification results using LR model.

92.30%. The RF model outperformed the five models in specific and general situations, indicating that it has better classification accuracy than the other five approaches. Therefore, the RF model is an effective alternative model for the prediction of postprandial blood glucose levels.

Algorithms	Overall CCR	Sensitivity {1-1}	Specificity {2-2}
RF	**82.68%**	**92.30%**	**60.73%**
C4.5	76.56%	85.90%	55.25%
MLP	75.73%	83.20%	56.68%
LR	74.48%	84.00%	52.74%
SVM	69.61%	99.20%	2.05%

Table 6. Classification Results of the Five Data Mining Models

CONCLUSION

Accurate and precise reasoning and prediction models greatly help physicians improve the diagnosis, prognosis, and treatment of diabetes. We used five data-mining techniques and designed a model for the prediction of postprandial blood glucose levels on the basis of the known risk factors of diabetes. The results showed that the RF approach exhibited the highest classification accuracy out of the five models. Its specificity and overall CCR were higher those of the C4.5, SVM, MLP, and LR models. Therefore, the RF model provides better classification accuracy than the other competing approaches and is an effective data-mining method for the prediction of postprandial blood glucose levels in a cohort study.

REFERENCES

Wang, Y., Wei, F., Sun, C. & Li, Q. (2016). The Research of Improved Grey GM (1, 1) Model to Predict the Postprandial Glucose in Type 2 Diabetes. *BioMed Research International*, *2016*. Retrieved Nov. 1, 2017, from http://doi.org/10.1155/2016/6837052.

Wang, Y. & An, B. (2014). The research least squares based on AR model of glucose prediction. *Advanced Materials Research*, 10(971-973): 284-287.

Tresp, V., Briegel, T. & Moody, J. (1999). Neural-network models for the blood glucose metabolism of a diabetic. *IEEE Transactions on Neural Networks*, 10(5):1204-1213.

García-Jaramillo, M., Calm, R., Bondia, J., & Vehí, J. (2012). Prediction of postprandial blood glucose under uncertainty and intra-patient variability in type 1 diabetes: a comparative study of three interval models. *Computer methods and programs in biomedicine*, 108(1): 224-233.

Yamaguchi, M., Kaseda, C., Yamazaki, K., & Kobayashi, M. (2006). Prediction of blood glucose level of type 1 diabetics using response surface methodology and data mining. *Medical and Biological Engineering and Computing*, 44(6): 451-457.

Georga, E., Protopappas, V., Guillen, A., Fico, G., Ardigo, D., Arredondo, M. T. & Fotiadis, D. I. (2009). Data mining for blood glucose prediction and knowledge discovery in diabetic patients: The METABO diabetes modeling and management system. *Conference proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society*, 2009, 5633-6.

Oviedo, S., Vehí, J., Calm, R., & Armengol, J. (2016). A review of personalized blood glucose prediction strategies for T1DM patients. *International Journal for Numerical Methods in Biomedical Engineering*, 33(6): e2833.

Zarkogianni, K., Mitsis, K., Litsa, E., Arredondo, M. T., Fico , G., Fioravanti, A., & Nikita, K. S. (2015). Comparative assessment of glucose prediction models for patients with type 1 diabetes mellitus applying sensors for glucose and physical activity monitoring. *Medical & biological engineering & computing*, 53(12): 1333-1343.

Fernando, W. C. T. (2016). *Blood glucose prediction models for personalized diabetes management*. Unpublished doctoral dissertation, North Dakota State University, United States.

Vapnik, V. N. (2000) *The Nature of Statistical Learning Theory*. Berlin: Springer.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1): 5-32.

Larose, D. T. (2005). *Discovering Knowledge in Data: An Introduction to Data Mining*. New Jersey: John Wiley & Sons.

Haykin, S. (1999). Neural network: A comprehensive foundation. *Englewood Cliffs*. NJ: Prentice Hall.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. NJ: John Wiley & Sons.

Quinlan, J. R. (1993) *C4.5: programs for machine learning*. San Francisco: Morgan Kaufinann.

Frank, E., Hall, M. A. & Witten, I. H. (2016). The WEKA Workbench. In Witten, I. H., Frank, E., Hall, M. A. & Pal, C. J. (2016), *Data Mining: Practical Machine Learning Tools and Techniques (4th. Ed.)*, San Francisco: Morgan Kaufmann.

Liu, Q., Li, W., Xue, M., Chen, Y., Du, X., Wang, C., Han, L., Tang, Y., Feng, Y., Tao, C. & He, J-Q. (2017). Diabetes mellitus and the risk of multidrug resistant tuberculosis: A Meta-analysis. *Scientific Reports*, 7(1), 1090. Retrieved Nov 1, 2017, from https://www.nature.com/articles/s41598-017-01213-5.

ACKNOWLEDGMENT

This work is supported by the Chung Shan Medical University Hospital and LandSeed Hospital: CSMU-LSH-102-02.