Prediction of Diabetes Mellitus Based on Boosting Ensemble Modeling

Rahman Ali*, Muhammad Hameed Siddiqi*, Muhammad Idris*, Byeong Ho Kang**, Sungyoung Lee*[1],

{rahmanali, siddiqi, idris, sylee}@oslab.khu.ac.kr,
byeong.Kang@utas.edu.au

*Dept. of Computer Engineering, Kyung Hee University, Korea
**Dept. of Computing and Information Systems, University of Tasmania, Australia

Abstract. Healthcare systems provide personalized services in wide spread domains to help patients in fitting themselves into their normal activities of life. This study is focused on the prediction of diabetes types of patients based on their personal and clinical information using a boosting ensemble technique that internally uses random committee classifier. To evaluate the technique, a real set of data containing 100 records is used. The prediction accuracy obtained is 81.0% based on experiments performed in Weka with 10-fold cross validation.

Keywords: Knowledge Acquisition, Reasoning, Ensemble Modeling, Prediction, Diabetes, Boosting

1 Introduction

Diabetes mellitus is a chronic disease that results from absolute or relative deficiency of insulin. At the Ubiquitous Computing Laboratory[2], our group is working on the development of a cloud-based clinical decision support system (CDSS) for chronic disease[1]. Currently, this system predicts diabetes in patients and gives recommendations using standard rule-based reasoning. We extends the system by predicting the diabetes using machine learning approach. Machine learning is an active research area and has extensively been used for different medical applications in general and prediction of diabetes mellitus in specific. A number of ensemble learning techniques have been proposed in literature for the diabetes mellitus summarized in Table 1.

[1] Corresponding author: Sungyoung Lee
[2] http://uclab.khu.ac.kr/

Table 1. Ensemble learning approaches for the prediction of diabetes

S.No	Problem taken into account	Machine Learning-based Ensemble Methods
1	Risk forecasting for diabetes Type-2[2]	Gaussian NB, Log. Regression, K-NN, CART, Random Forests and SVM
2	Predicting the presence of diabetes[3]	SVM and BP NN
3	Predicting glucose level in diabetes data[4]	Linear and Bayesian ensemble model
4	Predicting psychosocial wellbeing of patients[5]	Multi-layer perceptron neural network model
5	Architecture for diabetes prediction system [6]	Fuzzy logic, NN and CBR
6	Predicting Type-2 diabetes[7]	Fisher linear discriminate analysis, SVM and DT
7	Predicting Type-2 diabetes [8]	Random forest and gradient boosting machine

In the literature, we found no such diabetes prediction system that predicts the diabetes from patients personal and clinical data. We propose the ensemble of AdaboostM1 with random committee to predict the diabetes types from patients personal and clinical data. The proposed method is supported by an architecture that integrates data management, learning and prediction components together to consider as an additional contribution of this study.

Rest of the paper is structured as section 2 describes the methodology, Section 3 focus on evaluation of the results and Section 4 concludes the work with future directions.

2 Proposed Methodology

The methodology of the proposed approach is explained with the help of architecture shown in Fig. 1. The main focus of the study is 'diabetes knowledge acquisition and prediction engine' rather than each individual component of the architecture.

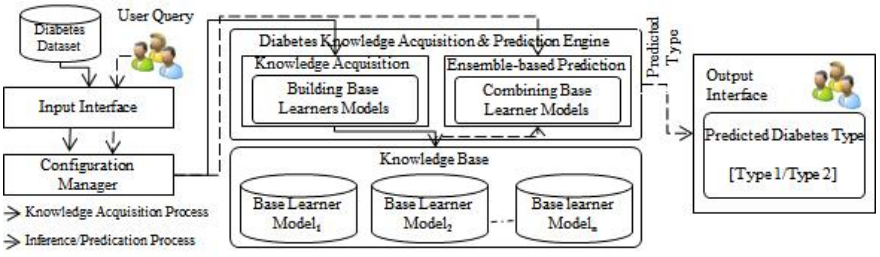

Fig. 1. Knowledge acquisition and reasoning architecture for predicting diabetes types

For the learning process, the diabetes dataset is passed through input interface to the configuration manager for preprocessing and selecting algorithms for learning. Knowledge acquisition component learns the data and stores in the knowledge base.

For the prediction process, user query is passed through the input interface to the configuration manager where an appropriate combination method (function), for combining the predictions of multiple classifiers, is selected. Based on the selection ensemble-based prediction component is activated which loads learned models from the

knowledge base and combines their outputs for the final prediction sent to the output interface.

For ensemble modeling and prediction, we adapted AdaboostM1 algorithm [9] along with random committee. Random committee classifier uses random tree as the base classifier. The boosting technique repeatedly runs random tree over various distribution of training diabetes data and combines the outputs in a single random committee classifier. The final prediction is a straight average of the predictions generated by the individual random tree classifiers.

3 Results Evaluation

We acquired diabetes dataset of 100 patients records with 56 of type-1 and 44 of type-2 from a local hospital. The parameters used for evaluation are: Gender, Total Cholesterol, Triglyceride, Low-density lipoprotein, High-density lipoprotein, Occupational therapy, Physical therapy, Fasting blood sugar, Systolic blood pressure, Diastolic blood pressure, Weight, Height, Hypoglycemia, HbA1c, Diabetes type, vital-Symptoms, Sentiments, and Activities. Table 2 shows the characteristics of the dataset.

Table 2. Characteristics of the diabetes dataset

Dataset	**#attributes**	**#instances**	**Missing values**	**#numeric attributes**	**#nominal attributes**	**#classes/instances per class**	
						Type1	Type2
Diabetes	18	100	No	12	6	56	44

Experiments were performed using Intel Pentium Dual-CoreTM (2.5 GHz) desktop computer with a 4GB RAM in Weka [10] with a 10−fold cross validation scheme for splitting the data into training and testing sets. The results are shown in Table 3 and Table 4.

Table 3. Detailed accuracy (by class) of the proposed ensemble-based approach

Class	**TP Rate**	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**ROC Area**
Diabetes Type 1	0.839	0.227	0.825	0.839	0.832	0.82
Diabetes Type 2	0.773	0.161	0.791	0.773	0.782	0.82
Weighted Avg.	0.81	0.198	0.81	0.81	0.81	0.82

Table 3 shows that the prediction accuracy is 81% which can be further improved by using preprocessing, feature selection and adding more data to the dataset.

Table 4. Confusion matrix for diabetes type 1 and type 2 predictions

Diabetes Type	**Type 1**	**Type 2**
Type 1	47	9
Type 2	10	34

Table 4 shows that the prediction accuracy of class type-1 is higher than the prediction accuracy of class type-2. The accuracy of class type-2 may increase if more data of type-2 is added to the training dataset.

4 Conclusion and Future Work

We have proposed a novel ensemble method (AdaboostM1 with random committee) to predict the diabetes type of patients. This work can be further extended in future by integrating it into our cloud-based CDSS system for chronic disease and adding a feedback mechanism on its top to increase the satisfaction level of the users and overall system.

Acknowledgment. This work was supported by the Industrial Strategic Technology Development Program (10035348, Development of a Cognitive Planning and Learning Model for Mobile Platforms) funded by the Ministry of Knowledge Economy (MKE, Korea).

References

[1] Hussain, M., Khattak, A.M., Khan, W.A., et al., 'Cloud-based Smart CDSS for chronic diseases', Health and Technology, 2013, 3, (2), pp. 153-175

[2] Yukun Chen, M.S., and Warren Clayton, M.D.: 'Type 2 Diabetes Risk Forecasting from EMR Data using Machine learning', AMIA Annual Symposium Proceedings. Vol. 2012. American Medical Informatics Association, 2012, pp. 606–615

[3] Zolfaghari, R.: 'Diagnosis of Diabetes in Female Population of Pima Indian Heritage with Ensemble of BP Neural Network and SVM', International Journal of Computational Engineering & Management, 2012, 15, (4)

[4] Stahl, F.: 'Diabetes Mellitus Glucose Prediction by Linear and Bayesian Ensemble Modeling', Sweden, 2012

[5] Narasingarao, M., Manda, R., Sridhar, G., et al., 'A clinical decision support system using multilayer perceptron neural network to assess wellbeing in diabetes.' The Journal of the Association of Physicians India, 2009, 57, pp. 127-133

[6] Thirugnanam, M., Kumar, P., Srivatsan, S.V., et al., 'Improving the Prediction Rate of Diabetes Diagnosis Using Fuzzy, Neural Network, Case Based (FNC) Approach', Procedia Engineering, 2012, 38, pp. 1709-1718

[7] Chen, H., and Tan, C.: 'Prediction of type-2 diabetes based on several element levels in blood and chemometrics', Biological trace element research, 2012, 147, (1-3), pp. 67-74

[8] Sood, A., Diamond, S., and Wang, S.: 'Type 2 Diabetes Mellitus Classification', Department of Computer Science, Stanford University, 2012

[9] Freund, Y., and Schapire, R.E.: 'Experiments with a new boosting algorithm', in ICML, 1996, Vol. 96, pp. 148-156

[10] Hall, M., Frank, E., Holmes, G., et al., 'The WEKA data mining software: an update', ACM SIGKDD explorations newsletter, 2009, 11, (1), pp. 10-18