Java-Based Diabetes Type 2 Prediction Tool for Better Diagnosis

Shankaracharya, M.Sc., APGDBI,¹ Devang Odedra, M.Sc.,¹ Medhavi Mallick, M.Sc.,¹ Prateek Shukla, BTech,² Subir Samanta, Ph.D.,³ and Ambarish S. Vidyarthi, Ph.D.¹

Abstract

Background: The concept of classification of clinical data can be utilized in the development of an effective diagnosis system by taking the advantage of computational intelligence. Diabetes disease diagnosis via proper interpretation of the diabetes data is an important problem in neural networks. Unfortunately, although several classification studies have been carried out with significant performance, many of the current methods often fail to reach out to patients. Graphical user interface-enabled tools need to be developed through which medical practitioners can simply enter the health profiles of their patients and receive an instant diabetes prediction with an acceptable degree of confidence.

Methods: In this study, the neural network approach was used for a dataset of 768 persons from a Pima Indian population living near Phoenix, AZ. A neural network mixture of experts model was trained with these data using the expectation-minimization algorithm.

Results: The mixture of experts method was used to train the algorithm with 97% accuracy. A graphical user interface was developed that would work in conjunction with the trained network to provide the output in a presentable format.

Conclusions: This study provides a machine-implementable approach that can be used by physicians and patients to minimize the extent of error in diagnosis. The authors are hopeful that replication of results of this study in other populations may lead to improved diagnosis. Physicians can simply enter the health profile of patients and get the diagnosis for diabetes type 2.

Introduction

DIABETES MELLITUS TYPE 2 has been recognized as a continuing health challenge for the 21st century in both developed and developing countries because of modernization of lifestyles as well as increasing urbanization and economic development.¹ It is a global problem of devastating human, social, and economic impact having around 300 million people worldwide living with diabetes.² Considering the importance of this problem, several guidelines have been proposed, and numerous studies have been carried out for prevention of diabetes.^{3,4} However, diabetes type 2 is frequently asymptomatic,⁵ and although detection is improving, the delay from disease onset to diagnosis may exceed 10 years.⁶ To diagnose the diabetes, a physician has to analyze many factors. Undoubtedly, the evaluations of data taken from patients and experts' decisions are crucial in diagnosis. However, factors like inexperience of the expert or his or her fatigue may lead to erroneous diagnosis. Early intervention

with lifestyle modifications or pharmacotherapy has been shown to effectively delay or prevent type 2 diabetes mellitus in adults.⁷ Hence there is a growing need for early detection of diabetes.

Many studies have been carried out using various strategies of computational intelligence, spanning, among others, logistic regression,⁸ support vector machines,⁹ and neuro-fuzzy inference systems.¹⁰ Of all these, artificial neural networks have been extensively explored for detection of diabetes. A comprehensive review in this regard presents interesting details of several such studies and current trends.¹¹ Neural nets try to simulate the human brain's ability to learn. That is, the artificial neural net is also made of neurons and dendrites. Unlike the biological model, a neural net has an unchangeable structure, built of a specified number of neurons and a specified number of connections between them (called "weights"), which have certain values. What changes during the learning process are the values of these weights. While different inputs are being learned, the weight values are changed dynamically

Departments of ¹Biotechnology and ³Pharmaceutical Sciences, Birla Institute of Technology, Mesra, Ranchi, India.

²Centre for Bioinformatics, School of Biotechnology, Shobhit University, Meerut, India.

until their values are balanced, so each input will lead to the desired output. The training of the neural net results in a matrix that holds the weight values among the neurons. Once a neural net had been trained correctly, it will probably be able to find the desired output to a given input that had been learned, by using these matrix values.¹¹ In the present study, the mixture of experts (ME) model¹² of a neural network was trained with the expectation-minimization algorithm and successfully incorporated into a computer interface where medical practitioners can enter the patient data and get the diagnosis.

Procedures

ME system

ME is a supervised learning algorithm that divides a learning task into appropriate subtasks, each of which can be solved by simple expert network. The global output of the ME system is derived as a convex combination of the outputs from a set of N experts, in which the overall predictive performance of the system is generally superior to that of any of the individual experts. The ME error function is based on the interpretation of MEs as a mixture model with conditional densities as mixture components (for the experts) and gating network outputs as mixing coefficients.

ME architecture (Fig. 1) is composed of several expert networks and a gating network. Gating network produces a scalar output from the vector input x . The gating network works on a generalized linear function in which the output for i^{th} input variable is given by

$$g(x, v_i) = \frac{e^{\xi_i}}{\sum_{k=1}^N e^{\xi_k}}$$

where $\xi_i = v_i^T x$ and v_i is the weight vector.

Each expert network produces an output vector for an input vector from the following generalized linear equation:

$$o_i(x) = f(W_i x)$$

where W_i is a weight matrix.

The final output of ME is thus the sum of multiplication of the output from the gating and expert networks:

FIG. 1. General architecture of the mixture of experts.

$$o(x) = \sum_{k=1}^N g(x, v_k) o_k(x)$$

Figure 1 shows the architecture of a ME network, consisting of N expert networks and one gating network, both having access to the input vector x ; the gating network has one output g_i per expert. The standard choices for gating and expert networks are generalized linear models and multilayer perceptrons. The output vector of a ME is the weighted (by the gating network outputs) mean of the expert outputs:

$$y(x) = \sum_{j=1}^m g_j(x) y_j(x)$$

The gating network outputs $g_j(x)$ can be regarded as the probability that input x is attributed to expert j . In order to ensure this probabilistic interpretation, the activation function for the outputs of the gating network is chosen to be the soft-max function:

$$g_j = \frac{\exp(z_j)}{\sum_{i=1}^m \exp(z_i)}$$

where z_i is the gating network output before thresholding. This soft-max function ensures that the gating network outputs sum to unity and are non-negative, thus implementing the soft competition between the experts.

The use of a soft-max function in the gating network and the conditional densities (ϕ_j) guarantee that the distribution is normalized: $\int p(t|x) dt = 1$. This distribution forms the basis for the ME error function, which can be optimized using gradient descent or the Expectation-Maximization (EM) algorithm.

Methods

For prediction of diabetes type 2, the Pima Indian diabetes dataset¹³ was used. It is a collection of medical diagnostic reports of 768 examples from a population living near Phoenix, AZ. This is a popular dataset used in large number of diabetes type 2 studies. Characteristics of the attributes for the population in the dataset are described in Table 1.

We found several records with missing data in one or more columns of the original dataset, which may lead to compromised training efficiency. To achieve better training efficiency, the dataset was preprocessed to impute the records with missing data. The resultant processed dataset

TABLE 1. ATTRIBUTE VECTORS WITH POPULATION CHARACTERISTICS IN THE TRAINING DATASET

Attribute	SD	Mean	Type
Number of times pregnant	3.4	3.8	Continuous
Plasma glucose concentration	32.0	120.9	Continuous
Diastolic blood pressure (mm Hg)	19.4	69.1	Continuous
Triceps skin fold thickness (mm)	16.0	20.5	Continuous
2-h serum insulin ($\mu\text{U}/\text{mL}$)	115.2	79.8	Continuous
Body mass index [weight in kg/(height in m^2)]	7.9	32.0	Continuous
Diabetes pedigree function	0.3	0.5	Continuous
Age (years)	11.8	33.2	Continuous
Class	—	—	0 or 1

consisted of 532 samples. Prior to training, each sample in the training dataset was normalized to the interval between 0 and 1 to exclude the bias caused by different means as well as to account for nonlinearities in the sigmoidal activation functions. The database was divided with 400 samples in the training set and 132 samples in the test set. An ME model comprising two experts and one gating network was created. Each of the expert and gating networks was a multilayer perceptron network with 20 hidden nodes. The weights were updated by a conjugate gradient descent method. A sigmoid activation function was used for each expert network, whereas the gating network operated with a soft-max activation function. Thus, there are two soft competition mechanisms in the mixture of expert architecture: on the basis of the supervised error, expert networks compete for the right to learn the training data, whereas gating networks associated with diverse features compete for the right to select an appropriate expert network as the winner for generating the output.

A form was developed using JDK version 1.6.0 of Java. The Swing library of Java was used to develop the look and feel features in the form. When the user enters data in the form, these values are passed to the MATLAB (MathWorks, Natick, MA) engine. Once the MATLAB engine is ready, MATLAB code is executed from within the Java environment. The output calculated by the MATLAB engine is then displayed on the form.

Results

Various configurations were tested to achieve the best optimum accuracy of the algorithm by the trial and error approach. The number of hidden nodes is associated with the mapping ability of the network. The larger the number of hidden nodes, the more powerful will be the network. However, if the number of hidden nodes is too large, then over-training occurs. Consequently, the generalization of the network may get worse, which may further result in poor performance of the network on test data. Only the performances with significantly good accuracy on test data are discussed here. An accuracy of 70.7% was achieved with an EM iteration of 900 and 20 hidden nodes. An EM iteration of 1,100 and 30 hidden nodes yielded an accuracy of 85.7%. An accuracy of 89.28% was achieved with an EM iteration of 1,000 and 20 hidden nodes. The overall best accuracy of 96.9% was achieved with an EM iteration of 1,100 and 20 hidden nodes. The prediction error during the training process approached almost zero as shown in Figure 2.

Sensitivity, specificity, and total classification accuracy were used as measures to determine the performance of the ME classifier. These three terms can be defined as follow: sensitivity = (number of true-positive outputs/number of actually positive cases); specificity = (number of true-negative outputs/number of actually negative cases); and total classification accuracy = (number of correct decisions/total number of cases).

FIG. 2. Network output of training and prediction error in training. Color graphics available online at www.liebertonline.com/dia

TABLE 2. CONFUSION MATRIX

	<i>Without diabetes</i>	<i>With diabetes</i>
Without diabetes	90	3
With diabetes	1	38

Thus sensitivity, specificity, and total classification accuracy were calculated as 96.7%, 97.4%, and 96.9%, respectively, from the confusion matrix (Table 2).

A form was developed where the user can enter the values of variables (Fig. 3). The input variables are number of times pregnant, plasma glucose concentration, diastolic blood pressure (mm Hg), triceps skin fold thickness, body mass index, diabetes pedigree function, and age. Clicking on "Clear" allows clearing all the field values. When "Submit" is clicked, the diagnosis is displayed in the frame.

Discussion

The basic ethic of medical science is "prevention is better than cure." Automated diagnosis of diseases has always been of interest as an interdisciplinary study among computer and medical science researchers. Brause,¹⁴ by successful application examples, showed that human diagnostic capabilities are significantly worse than that of the neural diagnostic systems. Diabetes is a opportune disease for data mining technology for several factors—there is a huge amount of data, and diabetes is a common disease that costs a great deal of money.¹⁵

Considering the importance of diabetes prediction in medical health care, few attempts have been made in the past to design tools to predict diabetes. One such tool, Diabetes

Personal Health Decisions (PHD),¹⁶ is driven by the Archimedes Model. This model¹⁷ is a health modeling program that brings together a vast range of clinical research data to make predictions about health risk. This tool asks the user to enter information about his or her health history: height, weight, cholesterol levels, blood pressure readings, last dilated eye exam, current medications, and so on. PHD suffers from several limitations. First, this tool is not specialized for diabetes prediction. Moreover, it requires too many input variables and thus adds complexity to the task of making prediction. Another tool, the QDScore¹⁸ diabetes type 2 risk calculator, predicts diabetes based on ethnicity, age, sex, body mass index, smoking status, family history of diabetes, treated hypertension, cardiovascular disease, and current use of corticosteroids. However, the assumption of constant utility over time in this study may have led to an overestimation or underestimation of the true impact of patients' health state on their outcomes.

Although several algorithms for predicting the risk and diagnosis of type 2 diabetes have been developed, no widely accepted diabetes prediction tool has been developed and validated for use in routine clinical practice.¹¹ The tool designed in this study is distinguished from other tools by several features. It uses only eight inputs to make prediction. ME has shown significant improvement in the prediction problems, particularly in the diagnosis problem. This study demonstrates 97% accuracy on the Pima Indian Diabetes dataset, which further confirms the outstanding performance of this unique algorithm. Although training and validation were performed on limited population data, it will be interesting to reproduce these results in other populations from different geographical locations for wider application. Successful development of a tool with a simple interface is a one step

FIG. 3. Graphical user interface for diabetes type 2 diagnosis prediction. Color graphics available online at www.liebertonline.com/dia

further to extend this utility to people. In an era of increasing number of diabetes patients, such attempts, although at the beginning stage, are very important. The authors are highly hopeful that this study will lead to further development of research in this direction to reach out to large numbers of patients.

Conclusions

The current tool is developed on the Pima Indians dataset. Although this dataset provides well-validated data for predicting diabetes diagnosis, it can be argued that models trained on such a dataset may not perform equally well on profiles of patients from other ethnic groups. It is therefore recommended that the models of choice must be trained on a dataset that more closely represents patient profiles of the medical practitioner within a specific geographic region or community. In the future relatively more easy-to-use tools can be developed with variables that are easy to use without compromising the accuracy of the tool. The current study shows successful application of computer technology to diabetes prediction. If the artificial neural network-based prediction approach shows improved medical diagnosis, it may be more widely acceptable and made more readily available to assist in patient care in a larger number of hospitals and clinics. This system can be used in medical camps or in common checkup camps, etc., so that the maximum number of patients will be diagnosed in a short period of time.

Acknowledgments

We are very thankful to the Department of Biotechnology, Birla Institute of Technology, Mesra, India, where this work was performed. We also acknowledge the Department of Agriculture, Government of Jharkhand (grant 5/BKV/Misc/12/2001) and BITSNetSubDIC (grant BT/BI/04/065/04) from the Department of Biotechnology, Government of India for providing infrastructural facilities.

Author Disclosure Statement

No competing financial interests exist.

References

- Lieberman LS: Dietary, evolutionary, and modernizing influences on the prevalence of type 2 diabetes. *Annu Rev Nutr* 2003;23:345–377.
- Zhang Y, Dall T, Mann S: The economic costs of undiagnosed diabetes. *Population Health Manag* 2009;12:95–101.
- Paulweber B, Valensi P, Lindström J, Lalic NM, Greaves CJ, McKee M, Kissimova-Skarbek K, Liatis S, Cosson E, Szendroedi J, Sheppard KE, Charlesworth K, Felton AM, Hall M, Rissanen A, Tuomilehto J, Schwarz PE, Roden M, Paulweber M, Stadlmayr A, Kedenko L, Katsilambros N, Makrilakis K, Kamenov Z, Evans P, Gilis-Januszewska A, Lalic K, Jotic A, Djordjevic P, Dimitrijevic-Sreckovic V, Hühmer U, Kulzer B, Puhl S, Lee-Barkey YH, AlKerwi A, Abraham C, Hardeman W, Acosta T, Adler M, AlKerwi A, Barengo N, Barengo R, Boavida JM, Charlesworth K, Christov V, Claussen B, Cos X, Cosson E, Deceukelier S, Dimitrijevic-Sreckovic V, Djordjevic P, Evans P, Felton AM, Fischer M, Gabriel-Sanchez R, Gilis-Januszewska A, Goldfracht M, Gomez JL, Greaves CJ, Hall M, Handke U, Hauner H, Herbst J, Hermanns N, Herrebrugh L, Huber C, Hühmer U, Huttunen J, Jotic A, Kamenov Z, Karadeniz S, Katsilambros N, Khalangot M, Kissimova-Skarbek K, Köhler D, Kopp V, Kronsbein P, Kulzer B, Kyne-Grzebalski D, Lalic K, Lalic N, Landgraf R, Lee-Barkey YH, Liatis S, Lindström J, Makrilakis K, McIntosh C, McKee M, Mesquita AC, Misina D, Muylle F, Neumann A, Paiva AC, Pajunen P, Paulweber B, Peltonen M, Perrenoud L, Pfeiffer A, Pölönen A, Puhl S, Raposo F, Reinehr T, Rissanen A, Robinson C, Roden M, Rothe U, Saaristo T, Scholl J, Schwarz PE, Sheppard KE, Spiers S, Stemper T, Stratmann B, Szendroedi J, Szybinski Z, Tankova T, Telle-Hjellset V, Terry G, Tolks D, Toti F, Tuomilehto J, Undeutsch A, Valadas C, Valensi P, Velickiene D, Vermunt P, Weiss R, Wens J, Yilmaz T: Take action to prevent diabetes—the IMAGE toolkit for the prevention of type 2 diabetes in Europe. *Horm Metab Res* 2010;42(Suppl 1):S37–S55.
- Jackson DM, Wills R, Davies JH: Public awareness of the symptoms of diabetes mellitus. *Diabet Med* 1991;8:971–972.
- Harris MI, Klein R, Wellborn TA: Onset of NIDDM occurs at least 4–7 yr before clinical diagnosis. *Diabetes Care* 1992;15:815–819.
- Knowler WC, Barrett-Connor E, Fowler SE: Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *N Engl J Med* 2002;346:393–403.
- Abbot RD: Logistic regression in survival analysis. *Am J Epidemiol* 1985;121:465–471.
- Barakat NH, Bradley AP, Barakat MB: Intelligible support vector machines for diagnosis of diabetes mellitus. *Trans Inf Technol Biomed* 2010;14:1114–1120.
- Dogantekin E, Dogantekin A, Avci D, Avci L: An intelligent diagnosis system for diabetes on Linear Discriminant Analysis and Adaptive Network Based Fuzzy Inference System: LDA-ANFIS. *Digit Signal Process* 2009;20:1248–1255.
- Shankaracharya, Odedra D, Samanta S, Vidyarthi AS: Computational intelligence in early diabetes diagnosis: a review. *Rev Diabet Stud* 2010;7:252–263.
- Jordan MI, Jacobs RA: Hierarchical mixtures of experts and the EM algorithm. *Neural Comput* 1994;6:181–214.

13. Frank A, Asuncion A: UCI Machine Learning Repository. University of California, School of Information and Computer Science. archive.ics.uci.edu/ml (accessed February 1, 2011).
14. Brause RW: Medical analysis and diagnosis by neural networks. *Lect Notes Comput Sci* 2001;99:1–13.
15. Cali AM, Caprio S: Prediabetes and type 2 diabetes in youth: an emerging epidemic disease? *Curr Opin Endocrinol Diabetes Obes* 2008;15:123–127.
16. Diabetes PHD Personal Health Decision by American Diabetes Association. www.diabetesarchive.net/phd/profile/start.jsp (accessed February 1, 2011).
17. Schlessinger L, Eddy DM: Archimedes: a new model for simulating health care systems—the mathematical formulation. *J Biomed Inform* 2002;35:37–50.
18. Hippisley-Cox J, Coupland C, Robson J, Sheikh A, Brindle P: Predicting risk of type 2 diabetes in England and Wales: prospective derivation and validation of QDScore. *BMJ* 2009;338:880.

Address correspondence to:
*Shankaracharya, M.Sc., APGDBI
Department of Biotechnology
Birla Institute of Technology
Mesra, Ranchi–835215, India*

*E-mail: shankaracharya@bitmesra.ac.in
shankaracharya2222@gmail.com*