Application of Data Mining Methods in Diabetes Prediction

Messan Komi, Jun Li
Computer Science department of Inner Mongolian University
Hohhot, China
e-mail: messanbathe@yahoo.fr, lijunalex@yahoo.com

Yongxin Zhai, Xianguo Zhang
Computer Science department of Inner Mongolia University
Hohhot, China
e-mail: 1923468541@qq.com, 2595083628@qq.com

Abstract—**Data science methods have the potential to benefit other scientific fields by shedding new light on common questions. One such task is help to make predictions on medical data. Diabetes mellitus or simply diabetes is a disease caused due to the increase level of blood glucose. Various traditional methods, based on physical and chemical tests, are available for diagnosing diabetes. The methods strongly based on the data mining techniques can be effectively applied for high blood pressure risk prediction. In this paper, we explore the early prediction of diabetes via five different data mining methods including: GMM, SVM, Logistic regression, ELM, ANN. The experiment result proves that ANN (Artificial Neural Network) provides the highest accuracy than other techniques.**

Keywords-diabetes; classificaton; data mining; prediction

I. INTRODUCTION

HUMAN body needs energy for activation. The carbohydrates are broken down to glucose, which is the important energy source for human body cells. Insulin is needed to transport the glucose into body cells. The blood glucose is supplied with insulin and glucagon hormones produced by pancreas. Insulin hormones produced by the beta cells of the islets of Langerhans and glucagon hormones are produced by the alpha cells of the islets of Langerhans in the pancreas. When the blood glucose increases, beta cells are stimulated and insulin is given to the blood. Insulin enables blood glucose to get into the cells and this glucose is used for energy. So blood glucose is kept in a narrow range. Diabetes is a chronic disease with the potential to cause a worldwide health care crisis. According to International Diabetes Federation 382 million people are living with diabetes across the whole world. By 2035, this will be doubled as 592 million [1]. However, early prediction of diabetes is quite challenging task for medical practitioners due to complex interdependence on various factors. Diabetes affects human organs such as kidney, eye, heart, nerves, foot etc.

Data mining is a process to extract useful information from large database, as there are very large and enormous data available in hospitals and medical related diabetes. It is a multidisciplinary field of computer science which involves computational process, machine learning, statistical techniques, classification, clustering and discovering patterns.

Recently, Data mining techniques have been widely used in predicting the data like time-series [2, 3]. A number of data mining algorithms have been proposed for early prediction of disease with higher accuracy in order to save human life and reduce the treatment cost [4]. Thus, applying these algorithms to predict diabetes should be done. In our work, we used five different supervised learning methods to conduct our experiment. This paper also aims to propose an effective technique for earlier detection of the diabetes disease. The rest of the paper is organized as follows. In section 2. We will specificity discuss all the methods that we used. The details over our experiment are then presented in section 3. Finally, in section 4, we conclude the paper.

II. METHOD

In this paper, five different data mining methods are used to conduct the early prediction over diabetes, they are including: GMM, ELM, SVM, Logistic regression and ANN. In this section, we will specificity discuss them.

A. *Prediction Using GMM Classifier Algorithm*

A Gaussian mixture model (GMM) is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. One can think of mixture models as generalizing k-means clustering to incorporate information about the covariance structure of the data as well as the centers of the latent Gaussians. The Gaussian Mixture Model Classifier [5] is useful and basic supervised learning classifier algorithm that is able to classify a large group of N-dimensional signals.

A Bayesian Gaussian mixture model is normally stretched to match a set of multivariate normal distributions modeling a vector x with N random variables. A vector of parameters may be modeled using a Gaussian mixture model prior distribution and the vector of estimates is given by:

$$\mathrm{P}(\theta) = \sum_{i=1}^{k} \emptyset i N(\mu i \sum i) \qquad (1)$$

Note that the i^{th} vector component is characterized by normal distributions with weights ∅i , means μi and covariance matrices $\sum \mathrm{i}$

To include this prior into a Bayesian estimation, the prior is multiplied with the known distribution p(Θ|x) which is given by :

$$\mathrm{P}(\theta/x) = \sum_{\mathrm{i}=1}^{\mathrm{k}} \theta\mathrm{iN}(\mu\mathrm{i} \sum \mathrm{i}) \qquad (2)$$

978-1-5090-6238-6/17/$31.00 ©2017 IEEE

Note that new parameter i, μi and $\sum i$ that are updated using the EM algorithm.

Applying GMM classifier has two steps. In training stage, GMM classifier is trained for each class. The probability distribution functions are calculated for each class. In application stage, a class of the maximum probability density function is assigned to this class.

The application of GMM is very wide. For instance, in image processing and computer vision, traditional image segmentation models often assign to one pixel only one exclusive pattern. In fuzzy or soft segmentation, any pattern can have certain "ownership" over any single pixel. If the patterns are Gaussian, fuzzy segmentation naturally results in Gaussian mixtures. In addition, GMM is also used as classifier to help doctor predict disease.

In our experiment, the GMM was established to predict the diabetes. Since we just need to predict whether one patient will suffer from diabetes or not, so two models was established with each model consisting of some Gaussian components. For initialization of the parameters over the model plays a important role on the final accuracy of the model, many experiments with different amount of Gaussian Components were conducted. And the details are listed in Table I.

TABLE I. GMM

The number of Gaussian component	**1**	**2**	**3**	**4**
Final accuracy	**0.64**	**0.81**	**0.75**	**Fail to converge**

The table strongly revealed that: the best result was achieved over the GMM is 81%. When the number of the Gaussian component is 1, the performance of the model is very poor, because the model with 1 Gaussian component is not complicated enough to fit the data. If the number of the component is 3 or bigger than 3, the performance of the model is also not satisfying due to the limitation of the training data amount. In general, the more complicated the mode is, the more training data it need to train the model. Concerning the amount and the complexity of the data, the best performance would be obtained when the model is consisting of 2 Gaussian components. Though, compared with the other models, the GMM model did not provide the best performance, but it is the fastest algorithm to reach the final accuracy.

B. Prediction Using ANN

The artificial neural network is much similar as natural neural network of a brain [6]. Artificial Neural networks (ANN) typically consist of multiple layers or a cube design, and the signal path traverses from front to back. Back propagation is the use of forward stimulation to reset weights on the "front" neural units and this is sometimes done in combination with training where the correct result is known. More modern networks are a bit freer flowing in terms of stimulation and inhibition with connections interacting in a much more chaotic and complex fashion. Dynamic neural networks are the most advanced, in that they dynamically can, based on rules, form new connections and even new neural units while disabling others.

Generally, the artificial neural network is consisting of the layers and network function, the layers of the network are including: input layer, hidden layer and output layer. The input neurons define all the input attribute values for the data mining model. In our work, the number of neurons is 7, since each item in our data set has 7 attributes, including: Glucose, Blood Pressure, Skin Thickness, Isulin, BMI, Diabetes Pedigree Function, and age. For the hidden layer, hidden neurons receive inputs from input neurons and provide outputs to output neurons. The hidden layer is where the various probabilities of the inputs are assigned weights. A weight describes the relevance or importance of a particular input to the hidden neuron. Mathematically, a neuron's network function *f(x)* is defined as composition of other functions g_i *(x),* which can further be defined as a composition of other functions. This can be conveniently represented as a network structure, with arrows depicting the dependencies between variables. A widely used type of composition is the nonlinear weighted sum, where $f(x) = K(\sum_i w_i g_i(x))$, where K (commonly referred to as the activation function) is some predefined function, such as the hyperbolic tangent and sigmoid function. The important characteristic of the activation function is that it provides a smooth transition as input values change, like a small changes in input produces a small changes in output.

The artificial neural networks are applied to tend to fall within the broad categories. Application areas include the system identification and control (vehicle control, trajectory prediction, [15] process control, natural resources management), quantum chemistry, game-playing and decision making (backgammon, chess, poker), pattern recognition (radar systems, face identification, object recognition and more), sequence recognition (gesture, speech, handwritten text recognition), medical diagnosis, financial applications (e.g. automated trading systems), data mining (or knowledge discovery in databases, "KDD"), visualization and e-mail spam filtering.

Artificial neural networks have also been used to diagnose several cancers. An ANN based hybrid lung cancer detection system named HLND improves the accuracy of diagnosis and the speed of lung cancer radiology. [16] These networks have also been used to diagnose prostate cancer. The diagnoses can be used to make specific models taken from a large group of patients compared to information of one given patient. The models do not depend on assumptions about correlations of different variables. Colorectal cancer has also been predicted using the neural networks. Neural networks could predict the outcome for a patient with colorectal cancer with more accuracy than the current clinical methods. After training, the networks could predict multiple patient outcomes from unrelated institutions.

In our paper, we use ANN to predict the diabetes. The best result, 0.89, is obtained when the number of hidden layer is 2, and the number of each hidden neurons is 5. The *f(x)*=Sigmoid was chosen as network function. At the beginning, the parameters were initialized randomly.

Concerning the optimization methods, the SGD (stochastic gradient descend) was chosen. For the sake of the higher accuracy, the other contrast tests were conducted as well. However, due to the limitation of the amount and availability of the training data sets, the range of the number of the hidden layer is set from 1 to 3, and the number of the neurons is from 5 to 10. The details are shown in Table II.

TABLE II. ANN

The number of the hidden layer	1	1	2	2	3
The number of the hidden neurons	5	10	5,5	10,10	5,5,10
Final accuracy	0.80	0.79	0.89	0.82	0.74

C. Prediction Using ELM (Extreme Learning Machine)

Extreme learning machines are feed forward neural network for classification or regression with a single layer of hidden nodes, where the weights connecting inputs to hidden nodes are randomly assigned and never updated. The weights between hidden nodes and outputs are learned in a single step, which essentially amounts to learning a linear model [7]. According to their creators, these models are able to produce good generalization performance and learn thousands of times faster than networks trained using back propagation. The simplest ELM training algorithm learns a model of the form:

$$Y = W_2\sigma(W_1x) \tag{3}$$

where W_1 is the matrix of input-to-hidden-layer weights, σ is some activation function, and W_2 is the matrix of hidden-to-output-layer weights. Generally, the W_1 would be filled with Gaussian random noise and W_2 can be estimated by least-squares fit.

As a learning technique, ELM has demonstrated good potentials to resolving regression and classification problems and also provides a unified learning platform with a widespread type of feature mappings. In addition, ELM can approximate any target continuous function and classify any disjoint regions. From the optimization method point of view, ELM has milder optimization constraints compared to LS-SVM and PSVM; Recently, ELM techniques have received considerable attention in computational intelligence and machine learning communities, in both theoretic study and applications. Some prior work has revealed that ELM method can be applied to many fields and has achieved fantastic results [8, 10]. Inspired by the Tiwari [9], who utilized the ELM method in forecasting daily urban water demand from limited data, and finally demonstrated the superiority of the ELM over other methods. Consequently, we also carry our experiments over ELM model and the best result is fixed in 82% with 10 neurons and *f(x)=sigmoid* in the hidden layer. The results of other experiments with different settings is shown in Table III.

TABLE III. ELM

The number of Gaussian Neurons	5	10	20	40
Final accuracy	0.75	0.82	0.74	0.76

D. Predicting Using Logistic Regression and SVM

In statistics [11] Logistic regression is a regression model where the dependent variable is categorical, namely binary dependent variable—that is, where it can take only two values, "0" and "1", which represent outcomes such as pass/fail, win/lose, alive/dead or healthy/sick. Logistic regression is used in various fields, including machine learning, most medical fields, and social sciences. For example, the Trauma and Injury Severity Score (TRISS), which is widely used to predict mortality in injured patients, was originally developed using logistic regression [11]. Many other medical scales used to assess severity of a patient have been developed using logistic regression [12, 13]. The technique can also be used in engineering, especially for predicting the probability of failure of a given process, system or product. It is also used in marketing applications such as prediction of a customer's propensity to purchase a product or halt a subscription. In economics it can be used to predict the likelihood of a person's choosing to be in the labor force, and a business application is about to predict the likelihood of a homeowner defaulting on a mortgage. Conditional random fields, an extension of logistic regression to sequential data, are used in natural language processing.

In this paper, Logistic regression was used to predict whether a patient suffer from diabetes, based on seven observed characteristics of the patient. Concerning the complexity and variety of the patient, the expected result was failed to be achieved, as the final result is 0.64;

The Support Vector Machine (SVM) was first proposed by Vapnik, and SVM is a set of related supervised learning method always used in medical diagnosis for classification and regression. SVM simultaneously minimize the empirical classification error and maximize the geometric margin. So SVM is called Maximum Margin Classifiers. SVM is a general algorithm based on guaranteed risk bounds of statistical learning theory, so called structural risk minimization principle. SVMs can efficiently perform non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. The kernel trick allows constructing the classifier without explicitly knowing the feature space. Recently, SVM has attracted a high degree of interest in the machine learning research community [14]. Several recent studies have reported that the SVM (support vector machines) generally are capable of delivering higher performance in terms of classification accuracy than the other data classification algorithms. SVM is a technique suitable for binary classification tasks, so we choose SVM to predict the diabetes.

To avoid over fitting due to the same data for the training and testing of SVM-based model, a 10-fold cross-validation strategy was used in the training data set. In this regard, the total data set was partitioned into 10 nearly equal subsets. In each of the 10 steps, 9/10 of the sample was used for training and 1/10 for testing. To evaluate the robustness of the SVM model, we choose various kernel functions in the SVM technique for predicting the diabetes. And our experiments revealed that SVM with RBF showed better performance than with other kernel functions (Liner, Polynomial) in accuracy as the best result is around 0.74 and it is far below our expectation. This is our first study to investigate predictors of disease by means of SVM, and it failed to present good performance. The reason is SVM is well known for its discriminative power for classification, especially in the cases where a large number of features are involved, and in our case where the dimension of the feature is 7. Further studies will be conducted with a larger number of features related to our case.

III. Experiment Result

In section 2, we specificity discuss the data mining methods, and the accuracy to predict the diabetes disease using different techniques is shown in Table IV. Based on the results demonstrated, The ANN method provides highest accuracy of the 0.89 to predict the disease. Compared with other methods and due to the complexity and the variety of the data set, the Logistic regression and SVM are less able to obtain an expected result. In the future work, we will explore the further reason.

TABLE IV. Result

Method	GMM	ANN	ELM	Logistic regression	SVM
Best result	0.81	0.89	0.82	0.64	0.74

IV. Conclusion

The diabetes prediction system is developed using five data mining classification modeling techniques. These models are trained and validated against a test dataset. All five models are able to extract patterns in response to the predictable states. The most effective model to predict patient with diabetes appear to be ANN followed by ELM and GMM. Although not the most effective model, the Logistic regression result is easier to read and interpret, what is more, the training over Logistic regression is very efficient. Although the ANN do outperform other data mining methods, the relationship between attributes and the final result is more difficult to understand.

Although we achieved a fair accuracy over the prediction of diabetes, our study still has several limitations. The primary limitation of this study is its small sample size, which made it very difficult for any of the endpoints to achieve statistical significance. The second limitation was that we did not directly measure medication adherences. Finally, our data was mainly based on patient information. However, this study only illustrates a potential use of the data mining method.

In the medical field accuracy in prediction of the diseases is the most important factor. In the analysis of data mining techniques, ANN classifier gives 89% of highest accuracy using MATLAB tool. Since the diabetes is a chronic disease it has to be prevented before it affects people. In future, the proposed work can be further enhanced and expanded for the disease prediction. For instance, the feature used in this paper can incorporate other medical attributes. It can also consider to use other data mining techniques, like Time Series, Clustering and Association Rules. Moreover, continuous data can be used instead of discrete data as well.

Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grant 61461039.

References

[1] Devi, M. Renuka, and J. Maria Shyla. "Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus." International Journal of Applied Engineering Research 11.1 (2016): 727-730.

[2] Berry, Michael J., and Gordon Linoff. Data mining techniques: for marketing, sales, and customer support. John Wiley & Sons, Inc., 1997

[3] Witten, Ian H., et al. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016.

[4] Emoto, Takuo, et al. "Characterization of gut microbiota profiles in coronary artery disease patients using data mining analysis of terminal restriction fragment length polymorphism: gut microbiota could be a diagnostic marker of coronary artery disease." Heart and vessels 32.1 (2017): 39-46.

[5] Giri, Donna, et al. "Automated diagnosis of coronary artery disease affected patients using LDA, PCA, ICA and discrete wavelet transform." Knowledge-Based Systems 37 (2013): 274-282.

[6] Fatima, Meherwar, and Maruf Pasha. "Survey of Machine Learning Algorithms for Disease Diagnostic." Journal of Intelligent Learning Systems and Applications 9.01 (2017): 1.

[7] Huang, Guang-Bin, Qin-Yu Zhu, and Chee-Kheong Siew. "Extreme learning machine: theory and applications." Neurocomputing 70.1 (2006): 489-501.

[8] Huang, Guang-Bin, Qin-Yu Zhu, and Chee-Kheong Siew. "Extreme learning machine: theory and applications." Neurocomputing 70.1 (2006): 489-501.

[9] Tiwari, Mukesh, Jan Adamowski, and Kazimierz Adamowski. "Water demand forecasting using extreme learning machines." Journal of Water and Land Development 28.1 (2016): 37-52.

[10] Uçar, Ayşegül, Yakup Demir, and Cüneyt Güzeliş. "A new facial expression recognition based on curvelet transform and online sequential extreme learning machine initialized with spherical clustering." Neural Computing and Applications 27.1 (2016): 131-142

[11] Boyd, C. R.; Tolson, M. A.; Copes, W. S. (1987). "Evaluating trauma care: The TRISS method. Trauma Score and the Injury Severity Score". The Journal of trauma. 27 (4): 370 – 378. doi:10.1097/00005373-198704000-00005. PMID 3106646.

[12] Kologlu M., Elker D., Altun H., Sayek I. Validation of MPI and OIA II in two different groups of patients with secondary peritonitis // Hepato-Gastroenterology. – 2001. – Vol. 48, № 37. – pp. 147–151

[13] Kologlu M., Elker D., Altun H., Sayek I. Validation of MPI and OIA II in two different groups of patients with secondary peritonitis // Hepato-Gastroenterology. – 2001. – Vol. 48, № 37. – pp. 147–151

[14] Laura Auria1and Rouslan A. Moro2, “Support Vector Machines (SVM) as a Technique for Solvency Analysis ” .Symp. Computational Intelligence in Scheduling (SCIS 07), ASME Press, Dec. 2007, pp. 57-64, doi:10.1109/SCIS.2007.357670.

[15] Zissis, Dimitrios (October 2015). "A cloud based architecture capable of perceiving and predicting multiple vessel behaviour". Applied Soft Computing. 35: 652–661. doi:10.1016/j.asoc.2015.07.002.

[16] Graves, Alex; and Schmidhuber, Jürgen; Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks, in Bengio, Yoshua; Schuurmans, Dale; Lafferty, John; Williams, Chris K. I.; and Culotta, Aron (eds.), Advances in Neural Information Processing Systems 22 (NIPS'22), December 7th–10th, 2009, Vancouver, BC, Neural Information Processing Systems (NIPS) Foundation, 2009, pp. 545–552