Diabetes Prediction Using Medical Data

Article · January 2017

CITATIONS
34

READS
12,997

1 author:

Asir Antony Gnana Singh Danasingh
Anna University , Tiruchirappalli
69 PUBLICATIONS **636** CITATIONS

SEE PROFILE

Journal of Computational Intelligence in Bioinformatics
ISSN 0973-385X Volume 10, Number 1 (2017) pp. 1-8

http://www.rfgindia.com

Diabetes Prediction Using Medical Data

Dr. D. Asir Antony Gnana Singh, Dr. E. Jebamalar Leavline, B. Shanawaz Baig

Department of Computer Science and Engineering [1&3],
Department of Electronics and Communication Engineering[2],
Anna University, BIT-Campus, Tiruchirappalli, India.
asirantony@gmail.com[1], jebilee@gmail.com,[2] shanawaz05@gmail.com[3]

Abstract

Nowadays, diabetes has become a common disease to the mankind from young to the old persons. The growth of the diabetic patients is increasing day-by-day due to various causes such as bacterial or viral infection, toxic or chemical contents mix with the food, auto immune reaction, obesity, bad diet, change in lifestyles, eating habit, environment pollution, etc. Hence, diagnosing the diabetes is very essential to save the human life from diabetes. The data analytics is a process of examining and identifying the hidden patterns from large amount of data to draw conclusions. In health care, this analytical process is carried out using machine learning algorithms for analysing medical data to build the machine learning models to carry out medical diagnoses. This paper presents a diabetes prediction system to diagnosis diabetes. Moreover, this paper explores the approaches to improve the accuracy in diabetes prediction using medical data with various machine learning algorithms and methods.

Keywords: Medical diagnosis, Medical data analytic, Diabetes disease, Prediction, Neural networks, Machine learning algorithm.

1. INTRODUCTION

Diabetes is the fast growing disease among the people even among the youngsters [1]. Diabetes is caused by the increase level of the sugar (glucose) in the blood. The diabetes can be classified into two categories such as type 1 diabetes and type 2 diabetes. Type 1 diabetes is an autoimmune disease. In this case, the body destroys the cells that are essential to produce insulin to absorb the sugar to produce energy.

This type can be caused regardless of obesity. The obesity is the increase of body mass index (BMI) than the normal level of BMI of an individual [2]. Type 1 diabetes can occur in childhood or adolescence age. Type 2 diabetes usually affects the adults who are obese. In this type, the body resists observing insulin or fails to produce insulin. Type 2 generally occurs in the middle or aged groups [1]. Moreover, there are other causes for diabetes such as bacterial or viral infection, toxic or chemical contents in food, auto immune reaction, obesity, bad diet, change of lifestyles, eating habit, environment pollution, etc. Diabetes leads various diseases such as cardiovascular complications, renal issues, retinopathy, foot ulcers, etc [1].

Data analytic is a process of examining and identifying the hidden patterns from large amount of data for drawing conclusions. In health care, this analytical process is carried out using machine learning algorithms for analysing the medical data to build machine learning models to carry out the medical diagnoses. Machine learning is a type of artificial intelligence (AI) that enables a system to learn by itself and develop the knowledge models to make decision by predicting the unknown data or label of the a given data.

The machine learning algorithms can be roughly categorized into three types namely supervised learning, unsupervised learning and semi-supervised learning. The supervised learning algorithms are used when human expertise does not exist (navigating on Mars), humans are unable to explain their expertise (speech recognition). Solution changes in time series (routing on a computer function) and to solution needs to be adapted to particular cases (user biometrics). The supervised learning algorithms are classified into different types such as probability-based, function-based, rule-based, tree-based, instance-based, etc. The unsurprised learning is the descriptive type learning. This learning is used to describe or summarize the data. The examples of the unsupervised learning algorithms are clustering, association rule mining, etc. The semi-supervised learning is the combination of supervised and unsupervised. This paper presents a diabetes prediction system to diagnosis the diabetics. Moreover, the supervised learning algorithm is used to learn the diabetes data and to develop diabetes predication system for diagnosing diabetes. The accuracy of this prediction system is improved using pre-processing technique.

The rest of this paper is organized as follows: Section 2 reviews the literature. Section 3 presents the diabetes prediction system. Section 4 details the experimental setup and procedure. Section 5 discussed results and discussion and Section 6 concludes the paper.

2. LITERATURE REVIEW

This section reviews various research works that are related to the proposed work. Mohammed Abdul Khaleel et al conducted a survey on data mining techniques on medical data for finding locally frequent diseases. The main focus of this survey is to analyse the data mining techniques required for medical data analysis that is especially used to discover locally frequent diseases such as heart lung cancer, ailments, breast cancer using classification and regression tree (CART) algorithm and

the decision tree algorithms such as ID3, C4.5 [3]. Chunhui Zhao et al presented a system for Subcutaneous Glucose Concentration prediction. This proposed model can predict the type 1 diabetes mellitus [4].

Vaishali Aggarwal et al presented a performance analysis of the competitive learning algorithms on Gaussian data for automatic cluster selection and also studied and analysed the performance of these algorithms and randomized results have been analysed on 2-D Gaussian data with the learning rate parameter kept simple for all algorithms. Algorithms used in their work include clustering algorithm, competitive learning algorithm and frequency sensitive competitive learning algorithm. Supervised learning machine algorithms are used for classification of the Gaussian data [5]. K. Srinivas et al developed applications of data mining techniques in healthcare and prediction of heart attacks. This research used medical profiles such as age, sex, blood pressure and blood sugar and predicted the likelihood of patients getting a heart and kidney problems [6].

M. Durairaj and V. Ranjani discussed the potential use of classification-based data mining techniques such as rule-based methods, decision tree algorithm, Naïve Bayes and artificial neural network (ANN) to the massive volume of healthcare data. In this research, medical problems have been analysed and evaluated such as heart disease and blood pressure [7]. Salim Diwani, et al discussed the applications of data mining in health care. This paper also presented an overview of research on health care application using data mining techniques. Data mining is a technology that is used for knowledge discovery in databases (KDD) and data visualization. Moreover, the medical data in the form of text, and the digital medical images such as X-rays, magnetic resonance imaging (MRI) are used for disease diagnostic processing [8].

Darcy A. Davis proposed individual disease risk prediction based on medical history. This paper also predicts each patient's greatest disease risks based on their own medical history data. Dataset are used for medical coding and collaborative assessment and recommendation engine (CARE) information technique [9]. From this literature, it is observed that the machine learning algorithms place a significant role in knowledge discovery form the databases especially in medical diagnosis with the medical data.

3. DIABETES PREDICTION USING MEDICAL DATA

This section presents the diabetes prediction system for diabetes diagnosis. Figure 1 illustrates the flowchart chart representation of the system model. Initially, the diabetes dataset is given into the data pre-processing module. The pre-processing module removes the irrelevant features from the diabetes dataset and gives the pre-processed dataset with relevant features to the machine learning algorithm. Then, the machine learning algorithm develops a learning model from the pre-processed dataset. This learning model is known as knowledge model. Furthermore, the diabetes is predicted for a person's medical report or data using the learning model.

4. EXPERIMENTAL SETUP AND PROCEDURE

This experiment is conducted using WEKA software [11] with the configuration of computer system 4 GB RAM, Intel(R) Core (TM)2 CPU 1.73 GHz Processor, Windows 7 64-bit operating system. For the conduction of this experiment, the diabetes medical dataset (Pima Indians diabetes dataset) has been collected from University of California, Irvine (UCI) machine learning repository [11]. The sample view of the dataset is illustrated in Figure 2. This dataset contains medical report of 768 persons. This medical report (dataset) includes 8 features of the persons such as number of time pregnant, plasma glucose concentration, blood pressure, skin fold thickness, insulin level, body mass index (BMI), diabetes pedigree function, age, and the results such as whether the person has diabetes (positive) or not (negative).

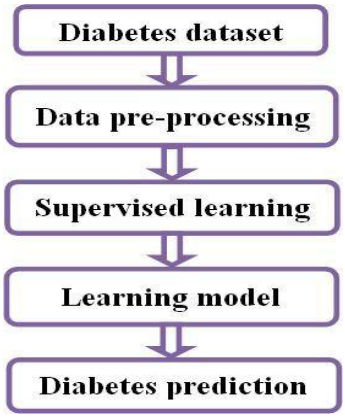

Figure 1 Flowchart representation of diabetes predication system

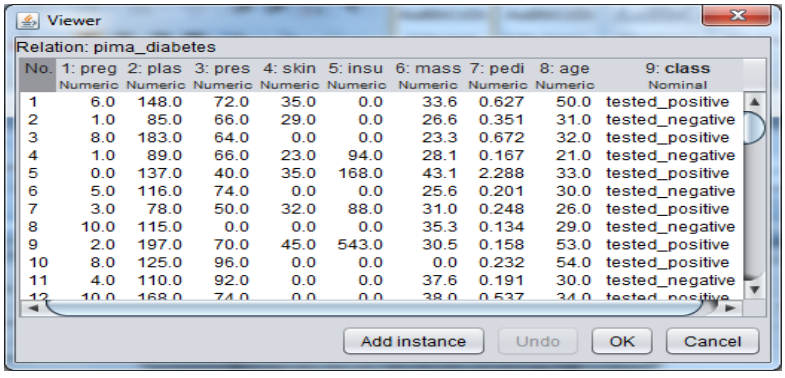

Figure 2 Sample view of dataset

The correlation-based feature selection technique is used for data pre-processing in order to remove the irrelevant features. Three different types of supervised machine learning algorithms namely probabilistic-based naïve Bayes (NB), function-based multilayer perceptron (MLP), decision tree-based random forests (RF). The test methods such as 10-fold cross validation (FCV), use percentage split with 66% (PS), and use training dataset (UTD) as the test dataset are used.

Initially, the diabetes dataset is given into the machine algorithms (NB, MLP, RF) and the accuracy with different test methods (FCV, PS, UTD) is noted in the Table 1. Then, the dataset is given into the correlation-based feature selection to perform the pre-process and the irrelevant feature are removed from the dataset and the dataset is given to the machine learning algorithm (NB, MLP, RF) and the accuracy is noted with different test methods (FCV, PS, UTD) and tabulated in the Table 1. Figure 3 shows the average accuracy of the machine learning algorithms with and without pre-processing technique.

5. RESULTS AND DISCUSSION

Table 1 shows the accuracy of machine learning algorithms (NB, ML, PRF) on the diabetes dataset with respect to different test methods (FCV, PS, UTD) with pre-processing method (WPP) and without pre-processing method (WOPP). Figure 2 shows the accuracy of machine learning algorithms (NB, ML, and PRF) on the diabetes dataset with respect to different test methods (FCV, PS, UTD) with pre-processing method (WPP) and without pre-processing method (WOPP).

From Table 1 and Figure 3, it is observed that for the NB machine learning algorithm, the PS test method produces better accuracy compared to other methods without pre-processing method. Moreover, the pre-processing method increases the accuracy for the NB machine learning algorithm. For the MLP machine learning algorithm, UTD test method produces better accuracy compared to other methods without pre-processing method. Moreover, the pre-processing method increases the accuracy for MLP machine learning algorithm except UTD test method. For the RF machine learning algorithm, UTD test method produces better accuracy compared to other methods without pre-processing method. Moreover, the pre-processing method increases the accuracy for the RF machine learning algorithm except FCV test method. From Figure 4, it is observed that the pre-processing technique produces better average accuracy for NB compared to other machine learning algorithm.

Table 1 Accuracy of machine learning algorithms (NB, ML, PRF) on diabetes dataset with respect to different test methods (FCV, PS,UTD) with pre-processing method (WPP) and without pre-processing method (WOPP).

Test method	**WOPP**			**WPP**		
	NB	**MLP**	**RF**	**NB**	**MLP**	**RF**
FCV	76.30	75.39	75.78	77.47	75.52	74.73
PS	77.01	74.32	78.54	79.69	78.54	80.07
UTD	76.30	80.59	100.00	77.60	76.82	100.00
Average	76.53	76.76	84.77	78.25	76.96	84.93

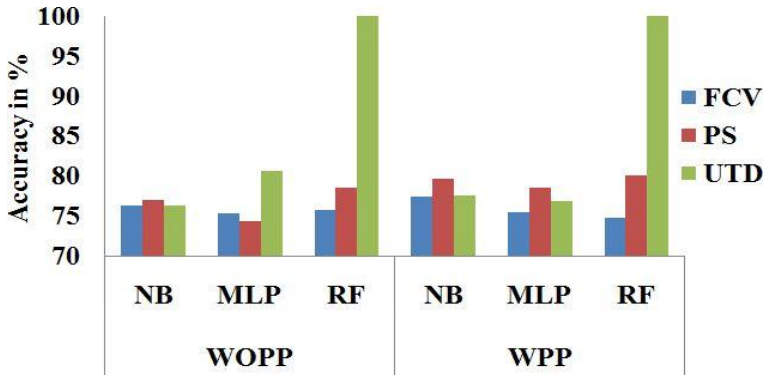

Figure 3 Accuracy of machine learning algorithms (NB, ML, PRF) on the diabetes dataset with respect to different test methods (FCV, PS,UTD) with pre-processing method (WPP) and without pre-processing method (WOPP).

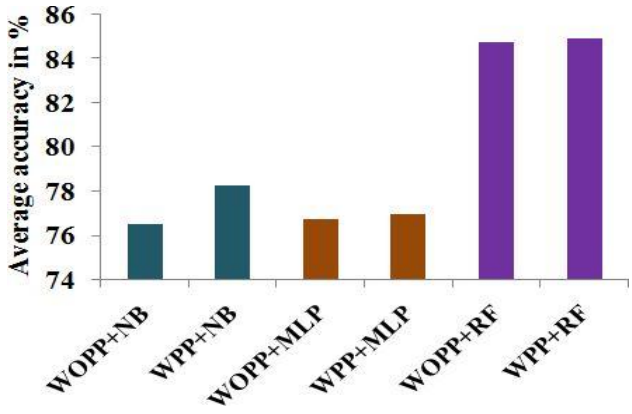

Figure 4 Average accuracy of machine learning algorithms with and without pre-processing

6. CONCLUSION

This paper presented a diabetes prediction system for diabetes diagnosis. In order to develop this system, the dataset is collected from the University of California, Irvine (UCI) repository. Different machine learning algorithm namely probabilistic-based naïve Bayes (NB), function-based multilayer perceptron (MLP), decision tree-based random forests (RF) are used to build the machine learning model to carry out the diagnosis of diabetes. Furthermore, the machine learning model is tested with different testing methods such as 10-fold cross validation (FCV), use percentage split with 66% (PS), and use training dataset (UTD) to evaluate the performance of the machine learning model in terms of accuracy. The pre-processing technique is used to increase the accuracy of the model. From the results, it is observed that the pre-

processing technique increases the accuracy of the machine learning algorithm except two cases. The pre-processing technique produces better average accuracy for NB compared to other machine learning algorithm.

REFERENCES

[1] Kaveeshwar, S.A., and Cornwall, J., 2014, "The current state of diabetes mellitus in India". AMJ, 7(1), pp. 45-48.

[2] Dean, L., McEntyre, J., 2004, "The Genetic Landscape of Diabetes [Internet]. Bethesda (MD): National Center for Biotechnology Information (US);. Chapter 1, Introduction to Diabetes. 2004 Jul 7.

[3] Mohammed, A.K., Sateesh, K. P., Dash G. N., 2013, "A Survey of Data Mining Techniques on Medical Data for Finding Locally Frequent Diseases" International Journal of Advanced Research in Computer Science and Software Engineering, 3(8), pp. 149-153.

[4] Chunhui, Z., Chengxia, Y., 2015, "Rapid Model Identification for Online Subcutaneous Glucose Concentration Prediction for New Subjects with Type I Diabetes", IEEE Transactions on Biomedical Engineering, 62 (5), pp. 1333 – 1344

[5] Vaishali, A., Harsh, K., Anil, K.A, 2016, "Performance Analysis of the Competitive Learning Algorithms on Gaussian Data in Automatic Cluster Selection", 2016 Second International Conference on Computational Intelligence & Communication Technology.

[6] Srinivas, K., Kavihta, R.B., Govrdhan, A., 2010 "Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks" International Journal on Computer Science and Engineering", 2(2), pp. 250-255

[7] Durairaj, M., Ranjani, V., 2013, "Data Mining Applications In Healthcare Sector: A Study", International Journal of Scientific & Technology Research, 2(10), pp. 31-35.

[8] Salim, D., Suzan Mishol., Daniel, S.K., Dina M., Anael S., 2013, "Overview Applications of Data Mining in Health Care: The Case Study of Arusha Region" International Journal of Computational Engineering Research, 3(8), pp. 73 -77.

[9] Darcy, A. D, Nitesh V.C., Nicholas B, 2008, "Predicting Individual Disease Risk Based on Medical History" CIKM '08 Proceedings of the 17th ACM conference on Information and knowledge management, pp. 769-778

[10] Eibe F, Mark A.H, Ian H.W., 2016, "The WEKA Workbench. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition.

[11] Lichman, M., 2013, "UCI Machine Learning Repository" [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.