A Broad Investigation for Diabetes Detection Using Machine Learning Techniques

Akshaya Hotkar, Pranjali Patil Manjushree Shinde, Sonali Bhalerao, Prof. S. A. Hadke

Abstract:

Diabetes mellitus is a group of metabolic disorder which has affected hundreds of millions of people. There have been plenty of researches about diabetes detection, which are mostly based on the Pima Indian diabetes data set. The identification of diabetes is of great significance, concerning its harsh complications. In this paper, we are dealing with most popular techniques Deep Neural Network, Support Vector Machine, used to identify diabetes. In this we are comparing the accuracy of each classifier over several ways of data preprocessors and therefore we modify the parameters to improve their accuracy. We also examine the significance between each feature with the classification result.

IndexTerms **- Component,formatting,style,styling,insert.**

I. INTRODUCTION

Diabetes is a collection of metabolic diseases in which patient or person has high blood sugar due to problem of producing insulin. Glucose is used to by body as a source of energy, and the pancreas produces a hormone called insulin that helps convert the glucose from the food you eat into energy. When the body does not produce sufficient insulin or does not make any at all the glucose does not reach your cells to be used for power. This results in diabetes. Diabetes can affect people of any age or any gender. It can affect people with any lifestyle. It leads to high blood sugar, together with some symptoms including frequent urination, increased thirst, increased hunger and weight loss. A SVM and DNN are the two classifiers used to identify the diabetes. Basically, these techniques are used to get accuracy by performing cross-validation on the Pima Indian data set. Comparing to the previous work, a comprehensive study is made containing a number of common techniques used to diabetes identification, intending to compare their performance and find the best one among them. Through the experiment, there is a comparison of several common and data preprocessors for each of the classifiers used, and find the best preprocessor respectively. Then compare these to reach their approximate maximum accuracy, and particularly analyze how to modify the parameters in DNN (Deep Neural Network). At last, analyze the relevance of each feature with the classification result, and this will help to modify the data set in future studies.

Keyword:

smart health applications , diabetes prediction

Related Work:

The research about using machine learning technique to identify J. W. Smith and his cooperators published a paper about using a so-called 'ADAP' algorithm to identify diabetes. They use the Indian Pima data set of diabetes onset of women as their training and testing data, and the accuracy of their algorithm is about 76%. Though the result it made was not the best, it has inspired many researchers to apply machine learning technique to the identification of diseases like diabetes. Many great results have been made using various algorithms. They especially focused on adopting the algorithm on some particular input data and reached 84.7% on the identified inputs. Kayaer's team used GRNN technique to identify diabetes. They discussed how to build the network and had a similar result as Gail A. The technique Kayaer used was much simplified compared to Gail's, but it was still a complex one regard to the scale of the data set. From all those researches we can see that they all explored diabetes identification through one particular method, and modified and improved it to its best or approximate best. The purpose of our research is to explore a bunch of common machine learning techniques for diabetes identification, and compare them comprehensively.

This system is useful for early prediction of diabetes. The user who will use this system needs to first register into the system. The details will be stored into the database. After registration the user will login into the system. Now the user will enter the details like age, gender etc which is mentioned in the dataset which we are using. The dataset used in the system for machine training is Pima Indian dataset. For prediction support vector machine and DNN algorithm is used. Dataset is trained in the form of .arff file. After prediction the system will provide the solutions according the prediction. The system takes input as the attributes, Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for both classification and regression challenges. However, it is mostly used in classification problems. The algorithm plots the each data item as a point in n-dimensional space with the value of each feature being the value of a particular coordinate. Then, the classification is performed by finding the hyper-plane that distinguishes the two classes very well. DNN is deep neural network which we are using in the system for classification purpose. The dnn is based on neural network. In that there are layers which are made of *nodes*. A node is a place where calculation happens, loosely patterned on a neuron in the human brain, which fires when it encounters sufficient stimuli. Here we are taking attributes as age, pregnancy, glucose, blood pressure, insulin, bmi and diabetes pedigree functions which are nothing but the nodes. In the dnn a node combines input from the data with a set of coefficients, or weights that either amplify or dampen that input, thereby assigning significance to inputs for the task the algorithm is trying to learn. These input-weight products are summed and the sum is passed through a node's so-called activation function, to determine whether and to what extent that signal progresses further through the network to affect the ultimate outcome, say, an act of classification.

Motivation:

Diabetes mellitus has a direct signal of high blood sugar, together with some symptoms including frequent urination, increased thirst, increased hunger and weight loss. Patient of diabetes usually need constant treatment, otherwise, it will possibly lead to many dangerous life-threatening complications. The automatic and early detection for diabetes is the today's need.

Mathematical Model

Mathematical Model:
Mathematical model set theory S= {s, e, X, Y, Φ}
S = Start of the program
1. Register/Login into the system
2. Provide Dataset (Diabetes Health care Data using ARFF File).
E = End of the program

Identify the Diabetes Patient Detection

X = Input of the program= {P, R, Q,Y}
P = ARFF Data
R = Attribute Data File (Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI , Pedigree Function, Age)
Q = Total Number of Classification &Accuracy(Using SVM, DNN Algorithm.)
Y = Output of program = Diabetes Predication Yes Or No

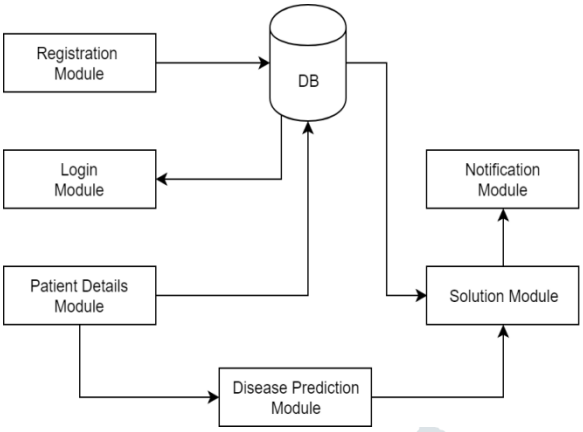
First, users provide data and system work on ARFF File Using Given Algorithm.
Let R be the set of User Data
D = {D1, D2, D3... .....Dn}
Let A be the set of categories Dataset(Attribute)
therefore,
T = {T1, T2, T3...,..... An}
Overall Data is evaluated with the help of these SVM AND DNN Algorithm which basically represents Diabetes Daises.

System Architecture:

Architectural Diagram is graphical representation of the concepts, their principles, elements and components that are part of architecture. This architecture diagram gives us the flow of the algorithm and overall functionality of the system. The user of the system selects one product after which the process starts.

The system architecture describes the overall flow of the system. This system is useful for early prediction of diabetes. The user who will use this system needs to first register into the system. The details will be stored into the database. After registration the user will login into the system using the login.jsp page. Now the user will enter the details like age, gender etc which is mentioned in the dataset which we are using. The dataset used in the system for machine training is Pima Indian dataset. For prediction support vector

machine and Dnn algorithm is used. Dataset is trained in the form of .arff file. The prediction result is shown on the result.jsp page. After prediction the system will provide the solutions according the prediction.



Conclusion:

The most accurate classifier, DNN, however, still have potential to improve further. One way of doing that is to add the number of hidden layers, a couple more of layers can be added. But the cost of that may beyond the benefit it earns, so if we want to improve DNN with more hidden layers, some advanced tricks will get involved such as drop out layer, or more kinds of regularization terms.

ACKNOWLEDGMENT

It gives us great pleasure in presenting the preliminary project report on '**A Broad Investigation for Diabetes Detection Using Machine Learning Techniques '**
I would like to take this opportunity to thank my internal guide for giving me all the help and guidance I needed I am really grateful to them for their kind support. Their valuable suggestions were very helpful.
I am also grateful to HOD for her in dispensable support and suggestions.

Name of Students

[1] Akshaya Hotkar, [2.] Pranjali Patil [3] Manjushree Shinde, [4] Sonali Bhalerao, [5] Kranti Jadhav

.

REFERENCES

[1] Kemal Polat, Salih Gunes, and Ahmet Arslan, "A cascade learning system for classification of diabetes disease: Generalized Discriminant Analysis and Least Square Support Vector Machine," Expert Systems with Applications, vol. 34. 1, January. 2008, pp. 482-487.
https://www.sciencedirect.com/science/article/pii/S0957417406002995

[2] Kayaer K and Yildirim T, "Medical diagnosis on Pima Indian diabetes using general regression neural networks," Proceedings of the international conference on artificial neural networks and neural information processing, 2003, pp. 181-184.
https://pdfs.semanticscholar.org/ef31/2e378325707b371c4727f6b1f9225fc03a9f.pdf

[3] Jack W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes, "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus," Proc. Annu. Symp. Comput. Appl. Med. Care, November 9. 1988, pp. 261-265.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2245318/

[4] Karegowda A. G., Manjunath A. S. and Jayaram M. A., "Application of genetic algorithm optimized neural network connection weights for medical diagnosis of pima Indians diabetes," International Journal on Soft Computing, vol. 2. 2, 2011, pp. 15-23.
https://pdfs.semanticscholar.org/2982/b42d1ca826c758c30ab59e44a5feac7daf4e.pdf

[5] Carpenter G. A. and Markuzon N., "ARTMAP-IC and medical diagnosis: Instance counting and inconsistent cases," Neural Networks, vol. 11. 2, 1998, pp. 323-336.
https://open.bu.edu/bitstream/handle/2144/2316/96.017.pdf?sequence=1

[6] Carpenter G. A. and Markuzon N., "ARTMAP-IC and medical diagnosis: Instance counting and inconsistent cases," Neural Networks, vol. 11. 2,
1998, pp. 323-336.

[7] Wold S., Esbensen K. and Geladi P., "Principal component analysis," Chemometrics and intelligent laboratory systems, vol. 2. 1-3, 1987, pp. 37-52.

[8] Balakrishnama S. and Ganapathiraju A., "Linear discriminant analysis-a brief tutorial," Institute for Signal and information Processing, vol. 18, 1998.

[9] Deng L. and Yu D., "Deep learning: methods and applications," Foundations and Trends in Signal Processing, vol. 7. 3-4, 2014, pp. 197-387.

[10] Lee H., "Tutorial on deep learning and applications," NIPS 2010 Workshop on Deep Learning and Unsupervised Feature Learning, 2010.