Predict And Decision Making For Diabetes

Mitul Patel[1], NeelAmin[2], Prof. Ajaykumar T. Shah[3]
[1, 2] Dept of Computer Engineering
[3]HOD, Dept of Computer Engineering
[1, 2, 3]Alpha College of Engineering and Technology

Abstract- *Now days from health care industries large volume of data is generating. It is necessary to collect, store and process this data to discover knowledge from it and utilize it to take significant decisions. With the help of technology, it is necessary to build a system that store and analyze the diabetic data and predict possible risks accordingly. Predictive analysis is a method that integrates various data mining techniques, machine learning algorithms and statistics that use current and past data sets to gain insight and predict future risks. In this work machine learning algorithm in Hadoop MapReduce environment are implemented for Pima Indian diabetes data set to find out missing values in it and to discover patterns from it. This work will be able to predict types of diabetes are widespread, related future risks and according to the risk level of patient the type of treatment can be provided.*

Keywords- diabetes mellitus, random forest, decision tree, neural network, machine learning, feature ranking.

I. INTRODUCTION

Diabetes is a common chronic disease and poses a great threat to human health. The characteristic of diabetes is that the blood glucose is higher than the normal level, which is caused by defective insulin secretion or its impaired biological effects. The earlier diagnosis is obtained, the much easier we can control it. Machine learning can help people make a preliminary judgment about diabetes mellitus according to their daily physical examination data, and it can serve as a reference for doctors. For machine learning method, how to select the valid features and the correct classifier are the most important problems. Machine learning methods are widely used in predicting diabetes, and they get preferable results. Decision tree is one of popular machine learning methods in medical field, which has grateful classification power. Random forest generates many decision trees. Neural network is a recently popular machine learning method, which has a better performance in many aspects. So in this study, we used decision tree, random forest (RF)and neural network to predict the diabetes.

Neural Network:

Neural network is a math model, which imitates the animal's neural network behaviors. This model depends on the complexity of the system to achieve the purpose of processing information by adjusting the relationship between the internal nodes .According to the connections' style, the neural network model can be divided into forward network and feedback network. In this paper, we used the Neural Pattern Recognition app in MATLAB, which is a two-layer-feed-back network with sigmoid hidden and soft max output neurons.

In neural network, there are some important parts, namely input layer, hidden layer and output layer. The input layer is responsible for accepting input data. We can get the results from the output layer. The layer between the input layer and the output layer is called hidden layer. Because they are invisible to the outside. There is no connection between neurons on the same layer. In this network, the number of hidden layers set to 10, which can get a better performance.

Machine Learning Methods:

Six machine learning classification models have been used for prediction of android applications .The models are available in R open source software. R is licensed under GNU GPL. The brief details of each model is described below.

Decision Trees:

The basic algorithm of decision tree [7] requires all attributes or features should be discretized. Feature selection is based on greatest information gain of features. The knowledge depicted in decision tree can represented in the form of IF-THEN rules. This model is an extension of C4.5 classification algorithms described by Quinlan.

Random Forest (RF):

Random forests [8] are a group learning system for characterization (and relapse) that work by building a large number of Decision trees at preparing time and yielding the class that is the mode of the classes yield by individual trees. Every user should be comfortable for the working of the known as a basic computer and net browser. They must have basic knowledge of English Language. User have to login one time. User can select the desired person by selecting categories. User must have some knowledge of how to use any websites. They have been some create account of basic needs.

Model Validation:

In many studies, authors often used two validation methods, namely hold-out method and k-fold cross validation method, to evaluate the capability of the model According to the goal of each problem and the size of data, we can choose different methods to solve the problem. In hold-out method, the dataset is divided two parts, training set and test set. The training set is used to train the machine learning algorithm and the test set is used to evaluate the model . The training set is different from test set. In this study, we used this method to verity the universal applicability of the methods. In k-fold cross validation method, the whole dataset is used to train and test the classifier. First, the dataset is average divided into k sections, which called folds. In training process, the method uses the k-1 folds to training the model and onefold is used to test. This process will be repeat k times, and each fold has the chance to be the test set. The final result is the average of all the tests performance of all folds . The advantage of this method is the whole samples in the dataset are trained and tested, which can avoid the higher variance). In this study, we used the five-fold cross validation method.

Linear Models (LM):

The Linear Model [10] is numerically indistinguishable to a various regression analysis yet burdens its suitability for both different qualitative and numerous quantitative variables.

II. FUTURE ENHANCEMENT

Machine learning can help people make a preliminary judgment about diabetes mellitus according to their daily physical examination data, and it can serve as a reference for doctors. The earlier diagnosis is obtained, the much easier we can control it. So they can aware before anything happened.

III. CONCLUSION

Diabetes mellitus is a disease, which can cause many complications. How to exactly predict and diagnose this disease by using machine learning is worthy studying. According to the all above experiments, we found the accuracy of using PCA is not good, and the results of using the all features and using mRMR have better results. The result, which only used fasting glucose, has a better performance especially in Luzhou dataset. It means that the fasting glucose is the most important index for predict, but only using fasting glucose cannot achieve the best result, so if want to predict accurately, we need more indexes. In addition, by comparing the results of three classifications, we can find there is not much difference among random forest, decision tree and neural network, but random forests are obviously better than the another classifiers in some methods. The best result for Luzhou dataset is 0.8084, and the best performance for Pima Indians is 0.7721, which can indicate machine learning can be used for prediction diabetes, but finding suitable attributes, classifier and data mining method are very important. Due to the data, we cannot predict the type of diabetes, so in future we aim to predicting type of diabetes and exploring the proportion of each indicator, which may improve the accuracy of predicting diabetes.

IV. ACKNOWEDGMENT

We express our sincere thanks to Prof. Ajaykumar T.Shah Head of Department of Computer Engineering, Alpha College of Engineering and Technology for their Support and guidance for this project and care taken by them in helping us to complete the project work successfully.

REFERENCES

[1] Margaret H. Dunham,-“Data Mining Techniques and Algorithms”, Prentice Hall Publishers.

[2] http://en.wikipedia.org/wiki/Diabetes_mellitus.

[3] Knowledge Discovery in Databases, http://www2.cs.uregina.ca.

[4] For Designing machine learning , python and Others: https://www.w3schools.com/

[5] PardhaRepalli, “Prediction on Diabetes Using Data mining Approach”.

[6] Introducing to machine learning with python: A Guide for Data Scientist Originally published: October 2017