Enhanced System to Achieve Better Classification Performance in Forecasting Diabetes and Breast Cancer

G. Prabaharan, K. Muthupriya and K. Chairmadurai

March 14, 2020

ENHANCED SYSTEM TO ACHIEVE BETTER CLASSIFICATION PERFORMANCE IN FORECASTING DIABETES AND BREAST CANCER

Dr. G. PRABAHARAN, Ph. D.,
prabaharang@gmail.com
Adhiparasakthi Engineering College, Melmaruvathur-603319

K.MUTHUPRIYA M. E-CSE
kmuthupriya1993@gmail.com
Adhiparasakthi Engineering College, Melmaruvathur-603319

K.CHAIRMADURAI M. E-CSE
kchairmadurai@gmail.com
Adhiparasakthi Engineering College, Melmaruvathur-603319

Abstract-- **Machine learning algorithm acts an important part in our life. It is the subset of Artificial intelligence. In the medical theme, Machine learning is used for the recognition and classification of diseases. It can classify breast cancer, diabetes or other diseases more perfectly from datasets. No medicines are invented to prevent diabetes and breast cancer fully. Breast cancer is increasing very promptly between women. The cost of breast cancer medicine is very extreme. Further researches are running on diabetes and breast cancer. Future our model to predict the diseases more perfectly rather than the previous models. Data place an essential act in predicting the diseases using some classification algorithms in provisions to the classifier occurrence for breast cancer and diabetes forecast.**

1. INTRODUCTION

Cancer

Cancer is when cells in the body modify and spread out of organize. Your body is made up of tiny construction blocks called cells. Normal cells grow when your body requirements them, and die when your body doesn't requirement them any longer. Cancer is made up of abnormal cells that spread still although your body doesn’t requirement them. In mainly types of cancer, the abnormal cells requirement to form a lump or mass called a tumor.

Breast Cancer

Breast cancer is cancer that starts in the breast. It occurs when cells in the breast are misused and start to requirement out of organize. The ducts and the lobules are the 2 parts of the breast where cancer is most possible to start. Breast cancer is one of the most familiar types of cancer in women in the U.S. Doctors don't still know accurately what causes it. Once breast cancer occurs, cancer cells can extend to extra parts of the body, assembly it life-threatening. The outstanding news is that breast cancer is often found early, when it's tiny and before it has extend.

Diabetes

Diabetes is a disorder that affects your body’s capability to create or use insulin. Insulin is a hormone. When your body turns the food you eat into force (also called sugar or glucose), insulin is free to assist transfer this force to the cells. Insulin acts as a “key.” Its chemical idea tells the cell to open and get glucose. If you create tiny or no insulin, or are insulin resistant, also much sugar remains in your blood. Blood glucose levels are higher than normal for persons with diabetes.

2. PROPOSED SYSTEM

The proposed system overcomes the above mentioned issues in a capable way. In proposed system, the system analyses a variety of forecast algorithms in diabetes and breast cancer. A relative study of these five algorithms Decision Tree Classifier, K-nearest neighbor, Naive Bayes, Random Forest, Support Vector Machine. The algorithm which produces most excellent forecasts is determined. Predict the diseases more perfectly rather than the previous models.

3. SYSTEM ARCHITECTURE

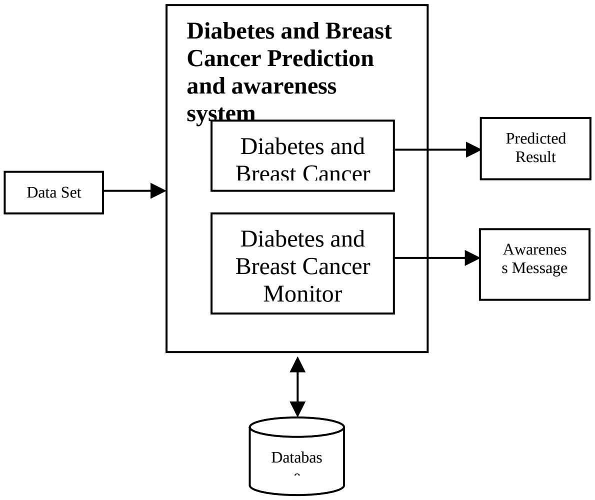
Diabetes and Breast Cancer Prediction and awareness system
Diabetes and Breast Cancer
Diabetes and Breast Cancer Monitor
Predicted Result
Awarenes s Message
Data Set
Databas e

Figure 3.1 System Architecture

In Figure 3.1, the overall system architecture is depicted. We proposed a system which collects the sample data of few usual health check-up that can be done by any individual by their own, and that data sets will be processed in the diagnosis software. Based on the input data, the predefined messages will be displayed to the user as an output. This diagnosis software developed to interfere the input data sets to identify the presence of diabetes and cancer.

The output of the system will be in two form:

1. First is, the diagnostic data that is predicted results and
2. Second part is, awareness message.

The user can consult a doctor with these output whether it is necessary or she can do her work as usual. The processed data will be stored in a database for anytime retrieval.

4. FUNCTIONALITIES OF THE SYSTEM

There are three main functions in the system

1. Dataset collection and processing
2. Analyzing the prediction algorithm
3 .Implementing the best algorithm

Dataset collection and processing

In this module, the data sets are from University of California-Irvine (UCI) machine learning repository. One is Wisconsin cancer data and the other is PIMA Indian diabetes data set.

The cancer data set contains 700 instances with 11 attributes including the class attribute. All the data are numeric and it is a two class classification problem since there are two different values for the class. The class attribute only contains values: 2 and 4. 2 defines as the cancer to be benign and 4 defines to be malignant. The attributes of this data set are as follows:

- ID
- clump thickness (range of value: 1-10)
- Uniform cell size (range of value: 1-10)
- Uniform cell shape (range of value: 1-10)
- Marginal adhesion (range of value: 1-10)
- Single epi cell size (range of value: 1-10)
- Bare nuclei (range of value: 1-10)
- Bland chromation (range of value: 1-10)
- Normal nucleoli (range of value: 1-10)
- Mitoses (range of value: 1-10)
- Class (range of value: 2 and 4)

The diabetes data has 769 instances with 9 attributes including the class attribute. This is also a two class classification problem and the values are numeric. The class attribute contains values: 0 and 1 defines as “NO” and “YES” for diabetes. This data set has missing values. The attributes of this data set are as follows:

- Preg
- Plasma gluc conc
- Bp
- Triceps
- Serum insulin
- Bmi
- Ped func
- Age
- Class (range of value: 0 and 1)

Both of these data sets suffer from class imbalance problem. So one of the class values are more frequent than the other one. Usually, if the class is imbalanced then often classifiers create biased results. So that result is not much reliable and does not make any sense. With imbalanced class problem, classifier always tries to predict the result to be the frequent class value.

Analyzing the prediction algorithm

1. Decision Tree Classifier

In common, Decision tree analysis is an analytical modelling tool that can be applied across many zones. Decision trees can be constructed by an algorithmic approach that can divide the dataset in different ways based on different situation. Decisions tress are the mightiest algorithms that falls under the type of supervised algorithms. They can be used for both classification and regression tasks.

2. K-nearest neighbor

K-nearest neighbors (KNN) algorithm is a type of supervised ML algorithm which can be used for both classification as well as regression analytical problems. However, it is mainly used for classification predictive problems in trade.

The following two properties would define KNN well

Lazy learning algorithm: KNN is a lazy learning algorithm because it does not have a focused training phase and uses all the data for training while classification.

Non-parametric learning algorithm: KNN is also a non-parametric learning algorithm because it doesn’t assume anything about the original data.

3. Naive Bayes

Naïve Bayes algorithms is a classification technique based on applying Bayes' theorem with a tough statement that all the predictors are independent to each other. In least words, the statement is that the presence of an element in a class is independent to the presence of any other element in the same class. For example, a phone may be considered as smart if it is having touch screen, internet feature, high-quality camera etc. Though all these features are dependent on each other, they provide independently to the likelihood of that the phone is a smart phone.

4. Support Vector machine

SVM Support vector machines (SVMs) are mighty still adjustable supervised machine learning algorithms which are used both for classification and regression. But generally, they are used in classification problems. In 1960s, SVMs were first introduced but later on they got refined in 1990. SVMs have their single way of implementation as compared to additional machine learning algorithms. Lately, they are extremely in style because of their ability to handle multiple continuous and categorical variables.

Implementing the best algorithm

5. Random Forest

Random forest is a supervised learning algorithm which is used for both classification in addition to regression. But however, it is mostly used for classification problems. As we know that a forest is made up of trees and additional trees means more robust forest. Equally, random forest algorithm creates decision trees on data samples and then gets the forecast from each of them and finally selects the greatest result by means of voting. It is an ensemble method which is better than an only decision tree because it reduces the over-fitting by averaging the end result.

5. RESULT ANALYSIS

For cancer data, RF performs the best as we can see from the TABLE. However, it is the slowest one as well. KNN has the least accuracy among these three but the fastest one. So, for cancer data, the more the accuracy is, the slower the classification is and vice-versa. This is also true for the diabetes data set. The higher accuracy means the slower build time as it is seen from the diabetes data in Table

Classifiers	Accuracy	Training Time	Prediction Time	F1 Score (Training Set)	F1 Score (Testing Set)
K-Nearest Neighbor	93.85%	0.012	0.022	0.9264	0.9081
Decision Tree	94.74%	0.0033	0.002	1	0.9239
Naïve Bayes	92.60%	0.006	0.003	0.9058	0.9341
Random Forest	95.61%	0.381	0.064	1	0.9101
Support Vector Machine	9.23%	0.97	0.95	0.93	0.82

Table 4.1 Comparison of algorithm

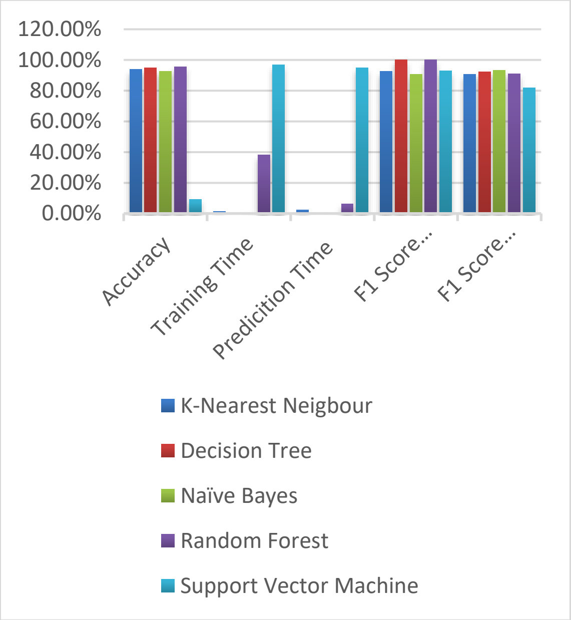

Figure 5.1 Result analysis

5. RELATED WORK

Diabetes mellitus is a group of metallic disorder characterized by steep levels of blood glucose extended over a time. It results the defection in insulin production or improper action of the cells to the insulin created. It is one of the important open health care brave worldwide. Diabetes exists in a body when pancreas does not construct sufficient hormone insulin or the human body is not being able to use the insulin properly. The diagnosis of requirement to generate and process the vast amount of data. Data mining techniques have proven its usefulness and effectiveness in order to estimate the unknown relationships or patterns if exists with such vast data.

The performance Analysis of Machine Learning Techniques to Forecast Diabetes Mellitus. Diabetes mellitus is a common disorder of human body caused by a group of metabolic disorders where the sugar levels over an extended period is very elevated. It affects different organs of the human body which thus harm a huge number of the body's system, in particular the blood veins and nerves. Early forecast in such disease can be controlled and save human life. To achieve the goal, this research work mainly explores a variety of risk factors related to this disease using machine learning techniques. Machine learning techniques provide capable result to extract knowledge by constructing predicting models from diagnostic medical datasets collected from the diabetic patients. Extracting knowledge from such data can be useful to forecast diabetic patients.

6. CONCLUSION

Detecting diseases like breast cancer and diabetes strength be helpful for the patients as well as the doctors. That is how the doctors may discover a way to determine the patients' condition and also if someone is at a high risk of cancer the doctors can decide on the medication and a way of life to help them survive an enhanced life.

REFERENCES

[1].M.Seera and C. P. Lim (2014) "A hybrid intelligent system for medical data classification," Expert Syst. Appl., vol. 41, no. 5, pp. 2239–2249.

[2].DeepikaVerma and Nidhi Mishra (2019) "Analysis and prediction of breast cancer and diabetes disease datasets using data mining classification techniques", 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE) 978-1-5386-9111-3.

[3].VandanaRawat and Dehradun (2019), "A Classification System for Diabetic Patients with Machine Learning Techniques", International Journal of Mathematical, Engineering and Management Sciences Vol. 4, No. 3, 729–744.]

[4].ShraboniRudra , Minhaz Uddin , Mohammed MinhajulAlam(2019) "Forecasting of Breast Cancer and Diabetes Using Ensemble Learning" International journal of computer communication and informatics. vol. 5, no. 3, pp. 424–427.