RESEARCH ARTICLE

International Journal of Computer Communication and Informatics

DOI: 10.34256/ijcci1911

Forecasting of Breast Cancer and Diabetes Using Ensemble Learning

Shraboni Rudra [a], Minhaz Uddin [a], Mohammed Minhajul Alam [a,*]

[a] Department of Computer Science East Delta University, Chittagong, Bangladesh.

*Corresponding Author
nazim@eastdelta.edu.bd
(Mohammed Minhajul Alam)

Received : 15-03-2019
Accepted : 08-05-2019

ABSTRACT: Machine learning algorithm plays an important role in our life. It is the subset of Artificial intelligence. Recently, everyone tries to use AI or try to invent something related to AI for making life easier. In the medical field, Machine learning is used for the recognition and classification of diseases. It can classify cancer, diabetes or other diseases more accurately from datasets. So, we propose a model which is the combination of Support vector machine and Ad boost. This combine method is known as Ensemble learner. In this paper, we are predicting diabetes and breast cancer. We have used SVM for classification purpose then have applied Ad boost for boosting purposes. The number of a diabetes patient is increasing very rapidly. It causes many other diseases like kidney failure; Eye disorder etc. No medicines are invented to prevent diabetes fully. Breast cancer is increasing very rapidly between women. The cost of breast cancer treatment is very high. More researches are running on diabetes and breast cancer. We proposed our model to predict the diseases more accurately rather than the previous models.

Keywords: SVM, Adaboost, Breast Cancer, Diabetes, Orange- tool, Machine Learning.

1. Introduction

Health is wealth. The development of a country mostly depends on the health condition of the people. So we proposed a model to predict the diseases. In this purpose, we used machine learning approaches and predicted that a person is suffering from diabetes and breast cancer. Machine learning is a method to predict any new information based on previous information. It helps us to classify, cluster or boost data. In this paper, we basically used the classification and the boosting method. For classification purposes, we used a support vector machine. We used Adaboost for boosting performance. A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyper plane. In other words, given labeled training data, the algorithm outputs an optimal hyper plane which categorizes new examples. In two dimensional spaces, this hyper plane is a line dividing a plane into two parts wherein each class lay on either side [1]. Adaboost classifier combines the weak classifier algorithm to form a strong classifier. A single algorithm may classify the objects poorly but if we combine multiple classifiers with the selection of training set during each iteration and assigning the right amount of weight in the final voting, we can have a better accuracy score for the overall classifier [2].

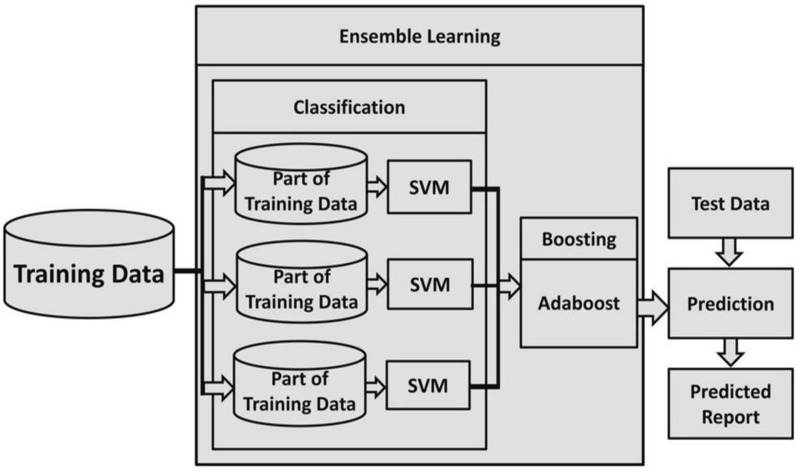

Fig1. The Block Diagram of Our Proposed Model perceptron algorithm.

Diabetes is a disease in which your blood glucose, or blood sugar, levels are too high. Glucose comes from the foods take you to eat. Insulin is a hormone that helps the glucose to get into your cells to give them energy. With type 1 diabetes, your body does not make insulin. With type 2 diabetes, which is the common type, your body does not make or use insulin well. Diabetes can damage your eyes, kidneys, and nerves [3]. Breast cancer can happen when cells in the breast begin to grow out of control. These cells usually form a tumor that can often be seen on an x-ray or can be felt as a lump. The tumor is malignant (cancer) if the cells can grow into (invade) surrounding tissues or spread (metastasize) to distant areas of the body. Breast cancer occurs almost entirely in women, but man can also be affected by breast cancer, too [4]. By using our model based on certain data, we can predict that a person is suffering from diabetes and cancer or not. If we predict this accurately than the person will take treatment timely. As a result, they can improve from diseases as early as possible.

In this paper, Section II contains the related work which is proposed by many author's, Section III describes the Machine Learning Algorithms at orange software which is used in our model, Section IV is a methodology where we describes our proposed model, Section V is an experimental evolution where we describe the description of our data, Section VI is a result analysis step where we describe the result of our model, Section VII we concluded our work.

2. Related Work

C. Kalaiselvi and Dr. G. et al., [5] proposed an Adaptive Neuro-Fuzzy Inference System (ANFIS) to diagnose diabetes. The number of times Pregnant, Plasma Glucose level, diastolic blood pressure(mm Hg), skin rashes and thickness(mm), 2 Hrs Serum Insulin, BMI, diabetes pedigree, age parameters are taken to diagnose diabetes [6]. Proposed K-means, which is also used for data reduction with the J48 decision tree as a classifier for classification. In order to get the experimental result, they used the Pima Indians Diabetes Dataset from the UCI Machine Learning Repository. Dr. Prof. Neeraj, using J48 data mining algorithm to diagnose the breast cancer [7]. Analyzing the decision tree generated by the algorithm using 10-fold cross-validation method to predict the recurrent events, based on the attributes such as node-caps, a degree of malignancy, age, tumor-size, menopause, irradiate they combine the five algorithms such as Nave Bayes, SMO, REP Tree, J48 and MLP algorithms, a classify breast cancer and diabetes. After analysing the performances of all algorithm, found that nave Bayes gives 72.70 percent accuracy on breast cancer dataset and SMO gives 76.80 percent accuracy on diabetes dataset [8]. To diagnose diabetes proposed a learning algorithm which ensemble boosting algorithm with the perceptron algorithm to improve the performance of the perceptron algorithm in the prediction of undiagnosed patients [9]. The proposed method is tested on three different publicly available datasets and compared with the performance of the perceptron algorithm.

3. Machine Learning Algorithms at Orange Software

A. Support Vector Machine

SVM is a machine learning technique that separates the attributes space with a hyperplane, thus maximizing the margin between the instances of different classes or class values. The technique often yields supreme predictive performance results. Orange embeds a popular implementation of SVM of the LIBSVM package. This widget is its graphical user interface. Its estimation accuracy depends on a good setting on cost (C), regression loss epsilon(s) and kernel parameters. In orange, we found two types of SVM based on different maximization of the error function. The Kernel is a function that transforms attribute space into a new feature space to fit the maximum- margin hyperplane. There are also three parameters based on a model. The parameter is g, c, and d. g is the gamma constant in kernel function which recommended value is 1/k where k is the number of the attributes. c for the constant co in the kernel function which default value is zero and d for the degree of the kernel which default value is three [10].

B. Adaboost

The Adobos (short for "Adaptive boosting") widget is a machine-learning algorithm, formulated by "Yoav Freund and Robert Schapire". It can be used with other learning algorithms to boost their performance. It does so by tweaking the weak learners. Adaboost works for both classification and regression. In orange, Adaboost contains three parameters. The first parameter is the number of estimators, the second parameter is learning rate, which determines to what extent the newly acquired information will override the old information (0 = the agent will not learn anything, 1 = the agent considers only the most recent information), the third parameter is fixed seed for random generator which is used to reproduce the result by setting a fixed seed. In orange ad boost contains two types of boosting method which is a classification algorithm and regression loss function [11].

4. Methodology

For the purpose of prediction, a prediction model was defined. The working principle of the proposed model has been shown in Fig.1.In our model; we had taken 75 percent data for training purpose and remaining 25 percent data put for testing purpose.

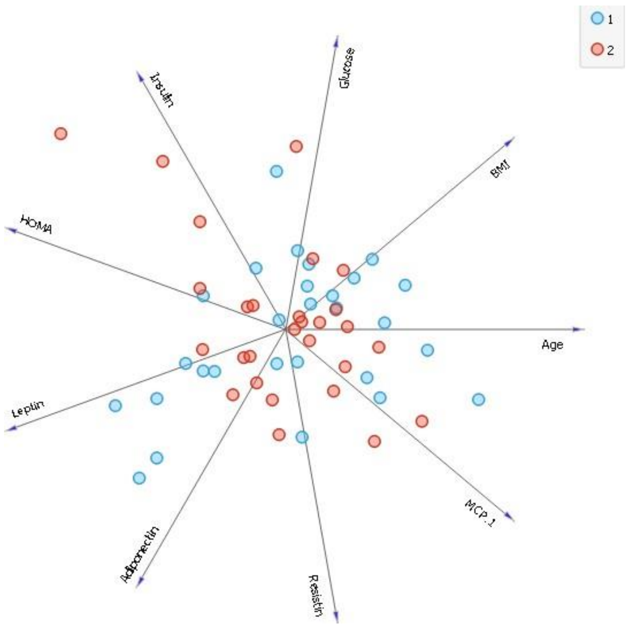

Fig2. The Classification of Cancer Dataset

To get a better result, we need to set the cost, the regression loss epsilon, and the kernel function. In our model, we used the sigmoid kernel function, set the cost of the higher value and the regression loss epsilon lower value. But the overall performance is very poor. So, according to Fig.1, we sent all the classification result in the boosting step. We were needed to apply to boost purpose, we used the Adaboost algorithm. In this case, we also needed to set two parameters. The first one is the number of estimators and the second one is the learning rate. We use test data to predict the result. The predictor function gives the predicted report. At last, we had the main part of our model is ensemble learning. The method of combining classification and boosting algorithms together is known as the ensemble learning. In the ensemble learning, we have two parts. The first part is the classification and the second part is boosting. Before classifying, we had divided the training data into a three sampler part. Then we applied the SVM classification algorithm. This classification algorithm classified the training data properly which we see in the Fig.2 and Fig.3. In Fig.2, we classify the cancer dataset where "1" that means blue circle indicates the healthy people and "2" that means red circle indicates the cancer patient. In Fig.3, we classify the diabetes dataset where "0" that means blue circle indicates the healthy people and "1" that means red circle indicates the Diabetes patient measured the overall performance.

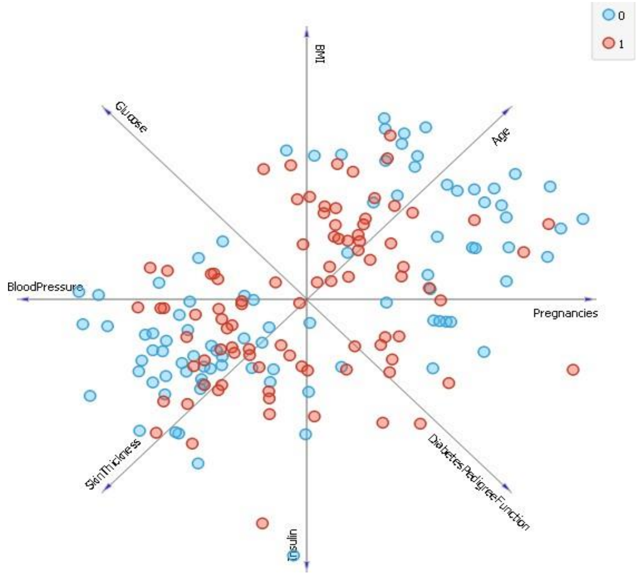

Fig 3. The Classification of Diabetes Dataset

5. Experimental Evaluation

A. Diabetes Dataset

The Pima Indian Diabetes dataset is used for Training and testing our model. We use the database because it is available free of charge for use in non-commercial research. This data- set contains 768 Number of Instances.

Attribute information

- Number of times Pregnant
- Plasma Glucose level
- Diastolic Blood pressure (mm/Hg)
- Skin rashes and thickness (mm)
- 2Hrs Serum Insulin
- BMI
- Diabetes pedigree
- Age

All the input parameters have numeric values. The First parameter is the total number of times the patient Pregnant. The second parameter is plasma glucose Concentration 2 hours in an oral. The third parameter is the diastolic blood pressure value which is measured in mm by Hg. The fourth parameter is skin rashes and thickness, which is measured in mm. The fifth parameter is 2-hours serum insulin test which finds the amount of Insulin creation in the patient's body. The sixth parameter is the patient's body mass index. The seventh parameter is the Diabetes pedigree which is the function value based on the diabetes family Hierarchy. The last parameter is age. The output parameter is classified into two categories. Class value- 0 is interpreted as tested negative for Diabetes; Class Value-1 is interpreted as tested positive for Diabetes [12].

B. Cancer Dataset

The Coimbra breast cancer dataset is used for training and testing our model. We use the database because it is available free of charge for use in non-commercial research. This dataset contains 116 numbers of instances.

Attribute information

- Age (years)
- BMI (kg/m2)
- Glucose (mg/dL)
- Insulin (U/mL)
- HOMA
- Leptin (ng/mL)
- Adiponectin (g/mL)
- Resistin (ng/mL)
- MCP-1(pg/dL)

All the input parameters have numeric values. The First parameter four parameter descriptions is the same as the previous data-set. The fifth parameter is HOMA which is a method for assessing -cell function and insulin resistance (IR) from basal (fasting) glucose and insulin or C-peptide concentrations. The sixth parameter Leptin is is measured in ng by mL. The seventh parameter Adiponectin that is measured in g by mL.The eighth parameter Resistin which is measured in ng by mL.The last parameter MCP-1 Which is measured in pg by dL. The output parameter is classified into two categories. Class value-1 is interpreted as no cancer and class value-2 is interpreted as cancer patient [13].

6. Result Analysis

The first step of our model is a classification. After classifying, we are measuring the performance. For measuring performance, we use five parameters.The parameters are Area under ROC is the area under the receiver-operating curve,

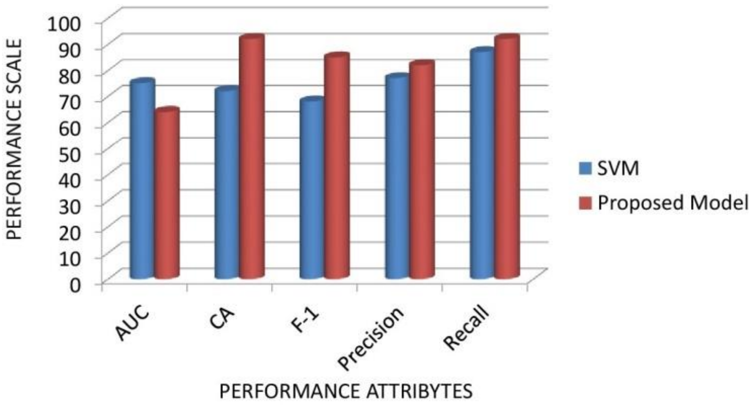

Fig4. Performance Analysis

Table 1. Comparison Of The Proposed Work With The Existing Works [5], [6], [7], [8], [9].

Method	Accuracy	Reference	Diseases
ANFIS with Adaptive KNN	80 %	C. Kalaiselvi (2014)	Diabetes Cancer
K-means &Decision Tree	90.04 %	Wenqian Chen (2017)	Diabetes
J48	75.5 %	Dr Prof.Neeraj (2017)	Cancer
SMO	76.80 %	Deepika Verma (2017)	Diabetes
J48	74.2 %	Deepika Verma (2017)	Cancer
Ensemble Perceptron Algorithm	74 %	Roxana Mirshahvalad (2017)	Diabetes
Proposed Model	92 %	This Study	Diabetes &Cancer

7. Conclusion

To predict diabetes and cancer many pieces of research are going. This paper discusses about a new approach for the prediction of diabetes and breast cancer using medical records of the patient. We use data-set for training and testing purpose. SVM is used to classify the data-set. For increasing the overall performance we apply boosting method after classification. This combined method gives more accuracy, precision, recall, and F-measure. From the classification and boosting rule, we take a decision that who is diabetes positive or negative and who is breast cancer positive or negative. So the method can be used by medical field to predict these diseases. In the future, we try to predict other diseases like lung cancer, brain tumor etc using this method. For the benefit of the medical field, we can use this model present and future as well.

References

[1] Savan Patel, Chapter 2 : SVM (Support Vector Machine)—Theory https://medium.com/machine-learning-101/chapter-2-svm-supportvector-machine-theory-f0812effc72, Accessed (2017)

[2] Emily Coberly, MedlinePlus-based health information prescriptions: a comparison of email vs paper delivery, J. Innov. Health Infor. (2013) 197-205

[3] Marilyn Fenichel, American Cancer Society Changes Breast Cancer Screening Guidelines to Reflect Analysis of Benefits and Harms, J. Nat. Can. Inst. 108(2016)

[4] Wenqian Chen, Shuyu Chen, Hancui Zhang and Tianshu Wu, A Hybrid Prediction Model for Type 2 Diabetes Using K-means and Decision Tree, 8th IEEE Inter. Confe. Soft. Eng. Serv. Sci., (2017) 386-390.

[5] Dr. Prof. Neeraj, Sakshi Sharma, Renuka Purohit, and Pramod Singh Rathore, Prediction of Recurrence Cancer using J48 Algorithm, 2nd Inter. Conf. Comm.Elect. Syst., (2017) 386-390.

[6] Deepika Verma and Dr. Nidhi Mishra, Analysis and Prediction of Breast Cancer and Diabetes disease datasets using Data mining classification Techniques, Inter. Conf. Intell. Sust. Sys. (2017) 533-538.

[7] Roxana Mirshahvalad and Nastaran Asadi Zanjani, Diabetes Prediction Using Ensemble Perceptron Algorithm, 9th Inter. Confe. Comput. Intell. Comm. Net. (2017)190-194.

[8] Roxana Mirshahvalad and Nastaran Asadi Zanjani, Diabetes Prediction Using Ensemble Perceptron Algorithm, 9th Inter. Confe. Comput. Intell. Comm. Net. (2017) 190-194.

[9] Janez Demšar and Blaž Zupan, Orange: Data Mining Fruitful and Fun - A Historical Perspective, Informatica, 37(2013) 55-60

[10] Md. Maniruzzaman, Md. Jahanur Rahman, Accurate Diabetes Risk Stratification Using Machine Learning, J. Med. Sys., 42 (2018) 92

[11] Tuba kiyan, Tulay Yildirim, Breast Cancer Diagnosis ¨ Using Statistical Neural Networks, J. Electrical Electron. Eng., 4 (2004) 1149-1153.

[12] W.H. Wolberg, O. L. Mangasarian, Multisurface Method of Pattern Separation for Medical Diagnosis Applied to Breast Cytology, Proc. Natl. Acad. Sci. U.S.A., 87 (1990) 9193–9196.

[13] Sander Greenland, Stephen J. Senn, J Kenneth,J Rothman , John B. Carlin, Charles Poole, Goodman, and Douglas G. Altman, Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations, Europ. J. Epid. 31 (2016) 337–350.

About The License