Enhancing Performance of Diabetes and Cancer Prediction for a Patient with Higher Accuracy by Combining the Results of Different Classifiers of Machine Learning Techniques

[1]Vikrant Waghmare, [2] Prof. Harish Barapatre, [3] Mahesh Pimpalkar
1, 2, 3 Yadavrao Tasgaonkar Institute of Engineering and Technology,
University of Mumbai.

Abstract:

Cancer and Diabetes are prolonged diseases which have enormous capability to cause a worldwide health care catastrophe. Various conventional methods, based on physical and chemical tests, are available for diagnosing this disease. Machine learning is an evolving scientific field in data science dealing with the ways in which machines learn from experience. An effective way to classify data is through classification or data mining. This becomes very handy, especially in the medical field where diagnosis and analysis are done through these techniques. The various classification models such as Decision Tree, Artificial Neural Networks, Logistic Regression, Association rules and Naive Bayes are used in this system. The proposed system allows the user to make use of these algorithms to predict the risk of diabetes and Cancer in human body. Based on the results of performed experiments, the Random Forest algorithm shows the highest accuracy with the least error rate. The dataset used is the Pima Indians Data Set, which has the information of patients. The aim this project is to develop a system (a mobile application) which can perform early prediction of Diabetes and Cancer for a patient with a higher accuracy by combining the results of different machine learning techniques

Index Terms: Classifiers, ELM, Disease, Healthcare, Prediction

---

Introduction:

Health-care information systems tend to capture data in databases for research and analysis in order to assist in making medical decisions. As a result, medical information systems in hospitals and medical institutions become larger and larger and the process of extracting useful information becomes more complex. Traditional manual data analysis has become inefficient and methods for efficient computer based analysis are needed. To achieve this aim, many approaches to computerized data analysis have been well thought-out and examined. Cancer and diabetes are two critical diseases in our society. Every year numerous people die out of cancer. Data mining represents a significant advancement in the type of analytical tools. It has been proven that the benefits of introducing data mining into medical analysis are to increase diagnostic accuracy, to reduce costs and to save human life.

In proposed system, four popular classifiers for disease risk prediction are studied. These algorithms consists Decision Tree, Artificial Neural Network, Logistic Regression and Naive Bayes.

Objectives

The following are some important Objectives:

1. The main objective of the research is to predict Cancer and diabetes. For cancer it will predict the stage as "Malignant" or "Benign" and for diabetes it will predict as "YES" or "NO". The prediction is based on some of the state of the art machine learning algorithms.
2. The project has another objective as to optimize the performances of these well-established machine learning algorithms.

3. Test the results of existing algorithms like Decision Tree, KNN, Random Forest, perceptron and Native Bayes Theorem.
4. Compare results with Extreme Learning Machine ELM.
5. Extract knowledge from information stored in database and generate clear and understandable description of patterns.
6. Access performance of models by calculating classification accuracy in terms of

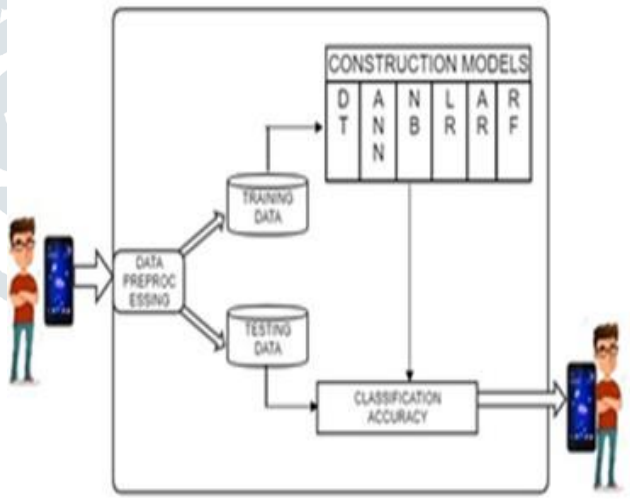
    a. Accuracy
    b. Weighted average precision
    c. Weighted average Recall
    d. Weighted average F-measure

Achievements

This project is to develop a system which can perform early and accurate prediction of cancer and diabetes for a patient with a higher accuracy by combining the results of different machine learning techniques. To design an algorithm for classifiers Decision tree, Artificial Neural Network, Naive Bayes and Association Rule for prediction of diabetes using available Dataset collected from PIMA Indian UCI library. To produce the results from weka tool using same dataset and compare accuracy of both results. The proposed system presents four stages of the process of conceptual framework in the study.

Conceptual Models

The proposed system presents four stages of the process of conceptual framework in the study. The process starts with data manipulation. Next, four models will be investigated for finding a prediction model.



Architecture diagram

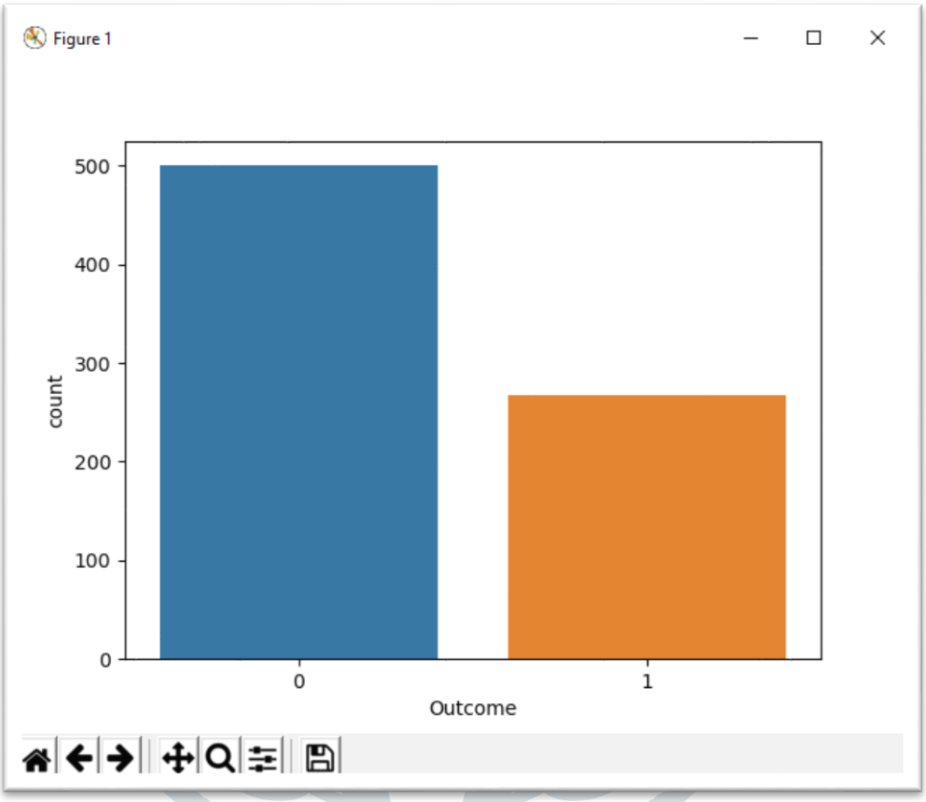
The process starts with data manipulation. Next, four models will be investigated for finding a prediction model. Then, accuracy of each model will be calculated and compared for seeking the best model. Detecting diseases like cancer and diabetes might be helpful for the patients as well as the doctors. From the doctors' perspective, they can help the patients to identify their next step by identifying the vulnerability of cancer or prevalence of diabetes in a patient. The study ends up with creating a web application.

Experimental Setup

The Proposed System for the diagnosis of diabetes disease is divided into two stages as shown in System Architecture. In the first Stage the Feature selection on the disease dataset is done to reduce the feature space dimension and at this stage different sets of features are obtained. In the second stage, ELM classifier is used to classify these feature subsets

and the classification accuracy is evaluated. The fittest set of feature subsets with the best classifier parameters are chosen to get an optimal system. The range of neurons is fixed from 1 to 200. The Process is carried out as below:

Step 1. Different feature subsets are obtained by Feature selection using Genetic Algorithm.
Step 2. Pima Indian Dataset is randomly divided into 10 fold of equal size using k fold cross validation methodology. This is done to maintain the class distribution in each and every fold in the same dataset.
Step 3. First feature subset is fed into ELM to get its Fitness value
Step 4. ELM parameters are initialized within the selected range.
Step 5. Classification is performed by using 10-fold cross validation
Step 6. Classification accuracy in each fold are calculated and the overall accuracy is obtained.
Step 7: Repeat Steps 3 to 6 for all feature subsets.
Step 8: The feature subset with the highest overall classification Accuracy is chosen as the best discriminating subset.
Step 9: Initialize the server after execuation of training mode.
Step 10: Initialize the User Interface module for testing purpose.
Step 11: Enter the value and pridict the result in the form of Yes or No.



Positive and Negative Class of Instances

Conclusion:

In this project, some classification algorithms are experimented. Some optimization attempts are made to improve the algorithms performances. Detecting diseases like cancer and diabetes might be helpful for the patients as well as the doctors. From the doctors' perspective, they can help the patients to identify their next step by identifying the vulnerability of cancer or prevalence of diabetes in a patient. That is how the doctors may find 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 lifestyle to help them live a better life.

References:

1. Han Wu, Shengqi Yang, Zhangqin Huang, Jian He, Xiaoyi Wang, [2018]"Type 2 diabetes mellitus prediction model based on data mining", Informatics in Medicine Unlocked 10 (2018) 100–107. (Base Paper)

2. Stefano Bromuri, Serban Puricel, Rene Schumann,[2016] "An expert Personal Health System to monitor patients affected by Gestational Diabetes Mellitus: A feasibility study", Journal of Ambient Intelligence and Smart Environments 8(2016) 219–237.

3. Gyorgy J. Simon, Pedro J.,[2016] "Extending Association Rule Summarization Tech- niques to Assess Risk of Diabetes Mellitus", IEEE TRANSACTIONS ON KNOWL- EDGE AND DATA ENGINEERING.

4. M. Seera and C. P. Lim,[2014] "A hybrid intelligent system for medical data classi- fication" Expert Syst. Appl., vol. 41, no. 5, pp. 2239–2249, 2014.

5. J. Tang, C. Deng, and G. Huang [2016], "Extreme Learning Machine for Multilayer Perceptron" IEEE Trans. Neural Networks Learn. Syst., vol. 27, no. 4, pp. 809–821, 2016.

6. G. Chandrashekar and F. Sahin, "A survey on feature selection methods" Comput. Electr. Eng., vol. 40, no. 1, pp. 16–28, 2014.

7. W. Yu, T. Liu, R. Valdez [2010], "Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes.," BMC Med. Inform. Decis. Mak., vol.10, p. 16, 2010.

8. American Cancer Society, "Cancer Facts and Figures 2015," Cancer Facts Fig. 2015, pp. 1–9, 2015.

9. J. C. N. Chan, V. Malik, W. Jia,[2009], "Diabetes in Asia: epidemiology, risk factors, and pathophysiology.," JAMA, vol. 301, no. 20, pp. 2129–40, 2009.

10. UCI Machine Learning Repository: Flags Data Set. [Online]. Available: https://archive.ics.uci.edu/ Cancer Wisconsin (Original). [Accessed: 19-Jan-2019].

11. UCI Machine Learning Repository: Flags Data Set. [Online]. Available: http://archive.ics.uci.edu/ Indians Diabetes. [Accessed: 19-Jan-2019].

12. Aishwarya S and Anto S [2014]," A Medical Expert System based on Genetic Algo- rithm and Extreme Learning Machine for Diabetes Disease Diagnosis", International Journal of Science, Engineering and Technology Research (IJSETR), Volume 3, Issue 5, May 2014

13. B.Fayssal, M.A.Chikh [2013],"Design of fuzzy classifier for diabetes disease using Modified Artificial Bee Colony algorithm", Computer methods and programs in biomedicine, No.1, pp.92-103, 2013.

14. R.Yuan, B.Guangchen [2010],"Determination of Optimal SVM Parameters by Us- ing Genetic Algorithm/Particle Swarm Optimization", Journal of Computers, No.5, pp.1160-116, 2010.

15. Senyue Zhang, Wenan Tan, and Yibo Li [2018],"A Survey of Online Sequential Ex- treme Learning Machine", 2018 5th International Conference on Control, Decision and Information Technologies (CoDIT-18),Thessaloniki, Greece April 10-13, 978-1- 5386-5065-3/2018 IEEE