EasyChair Preprint

№ 2332

Improve the performance of cancer and diabetes detection using novel technique of machine learning

Samrudhi R. Kaware and Vinod S. Wadne

January 8, 2020

Improve the Performance of Cancer and Diabetics Detection by Using Novel Technique of Machine Learning

Samrudhi R. Kaware, M.E.Student
Department of Computer
JSPM's ICOER Wagholi
Pune, India.
Email.: samrudhikaware17@gmail.com

Dr. V. S. Wadne, Associate Professor
Department of Computer
JSPM's ICOER,Wagholi,
Pune,India.
Email.:vinods1111@gmail.com

***Abstract*—**These system allows the user to make a use of algorithms to predict the risk of diabetes mellitus in human body. The various classification models such as Decision Tree, artificial Neural Networks, Logistic Regression, Association rules and Naive Bayes are used in this system. Then the Random Forest technique is used to find the accuracy of each model in the system. The dataset used is the Pima Indians Diabetes Data Set, which has the information of patients, some of them have developing diabetes therefore, this project is aimed to create a mobile application for predicting a person's class whether present in of the diabetes and cancer risk or not. In this study, we have analyzed medical data using several classification algorithms in order to optimize classifier performance for cancer and diabetes prediction. For this project we have guidance letter from "Max Care Hospital" under support of medical oncologist Dr. Satish Sonawane (onco-surgeon) to study the medical history of patients and their medical reports.

Keywords: *Decision tree, ANN, Health-Care*

I. INTRODUCTION

Cancer and diabetes are two destructive diseases in our society. Every year numerous people die out of cancer. The agency for healthcare research and quality (AHRQ) says that medical cost for cancer in the year 2011 in the united states was 88.7 billion dollars [1]. And out of various types of cancer, breast cancer has been one of the significant types over the past years [2]. Sometimes, breast cancer is detected at a stage when chances of survival are very low. Computer science can play some role to detect vulnerability of a cancer patient with medical data with the help of machine learning. By manipulating medical data, having attributes of cancer cell, a system can predict if the cancer is benign or malignant [3]. If the cancer is in a benign stage, then taking appropriate measures can help the patient survive and can even heal them completely in some cases. Cancer and diabetes is another disease that kills people slowly. Cancer and diabetes has become prevalent almost all over the world [4]. However, according to a study of asian. Diabetic prevention organization, 60 percent of the whole world's diabetic population are from asia [5].

So, asian people are at high risk. Existence of cancer and diabetes in a patient can be predicted by machine learning. So, in this research, cancer and diabetes is predicted as binary values like 1 or 0 meaning "yes" or "no". The data set that is used for cancer and diabetes includes attributes of the patients' feature that might lead to the existence of cancer and diabetes. Machine learns the attributes and then predicts in "yes" or "no". The main objective of the research is to predict cancer and cancer and diabetes. For cancer it will predict the stage as "malignant" or "benign" and for cancer and diabetes it will predict as "yes" or "no". The prediction is based on some of the state of the art machine learning algorithms. The project has another objective as to optimize the performances of these well-established machine learning algorithms. Some experiments will be performed to see if the algorithms can perform better on a different setup. Performance comparison is checked across different classifiers to understand how they behave with the same data set and how much time does each one take to build a classification model. One challenging prospect of this project is to achieve some techniques to apply curriculum learning on the data set.

For this project we have guidance letter from "Max Care Hospital" under support of medical oncologist Dr. Satish Sonawane (onco-surgeon) to study the medical history of patients and their medical reports.

II. LITERATURE SURVEY

A. *Title :'An Optimization Approach to Improve Classification Performance in Cancer and Diabetes Prediction'*

Author: Mustakim Al Helal , Atiqul Islam Chowdhury , Ashraful Islam , Eshtiak Ahmed , Md. Swakshar Mahmud , Sabrina Hossain.
Year : 2019
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 patient's 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.

B. *Title :Type 2 diabetes mellitus prediction model based on data mining*

Author: Han Wu, Shengqi Yang
Year : 2018.
WEKA toolkit and use the same Pima Indian Diabetes Dataset. Realistic dataset provided by Dr. Schorling was used to test & verify the model.

K-means Algorithm and Logistic Regression:
A novel model based on data mining techniques for predicting type 2 diabetes mellitus (T2DM). Based on a series of pre-processing procedures, the model is comprised of two parts, the improved K-means algorithm and the logistic regression algorithm. The Pima Indians Diabetes Dataset and the Waikato Environment for Knowledge Analysis toolkit were utilized to compare results with the results from other researchers.

C. Title: "An expert personal health system to monitor patients affected by gestational diabetes mellitus: a feasibility study."

Author: serban puricel, rene schumann, johannes krampf
Year: 2016
The graphs and charts provided in this study were of simple nature, as a future development, more elaborated charts and a representation of the data that puts physiological patterns in relation may allow the medical doctors to have a better understanding of the health of the patient, thus improving their ability to effectively modify the treatment plan.

D. Title: An improved electro magnetism-like mechanism algorithm and its application to the prediction of diabetes mellitus.

Author: angelia melani adrian, kun-huang chen
Year: 2015
In this study, 54 uci datasets are used to evaluate the performance of various classification algorithms. These datasets are characterized according to data sizes, features, and classes.

III. PROPOSED METHODOLOGY

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.

A. Extreme Learning Machine (ELM)

Extreme Learning Machines (ELM) are feed forward neural networks for classification, regression, clustering, sparse approximation, compression and feature learning with a single layer or multiple layers of hidden nodes, where the parameters of hidden nodes (not just the weights connecting inputs to hidden nodes) need not be tuned. These hidden nodes can be randomly assigned and never updated (i.e. they are random projection but with nonlinear transforms), or can be inherited from their ancestors without being changed. In most cases, the output weights of hidden nodes are usually learned in a single step, which essentially amounts to learning a linear model. The name "extreme learning machine" (ELM) was given to such models by its main inventor Guang-Bin Huang

i Efficient for multi-layer of ELM, auto-encoder, and feature learning, PCA and random projects are specific cases of ELM when linear neurons are used.
ii Regularization of output weights, ridge regression theories, neural networks generalization performance theories (maximal margin in binary class cases), SVM and LS-SVM provide suboptimal solutions.
iii Homogenous architectures for compression, feature learning, clustering, regression and classification.

IV. SYSTEM ARCHITECTURE

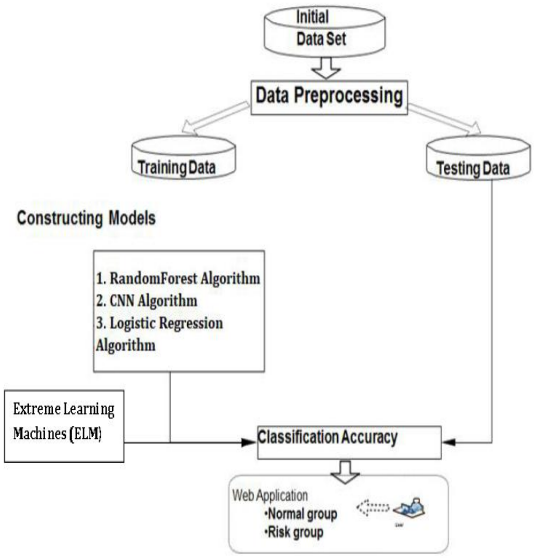


Figure 1: System architecture

I. ALGORITHM

SVM Algorithm
Input: D dataset, on-demand features, aggregation-based features,
Output: Classification of Application
for each application App-id in D do
Get on-demand features and stored on vector x for App-id
x.add (Get-Features(app-id));
end for
for each application in x vector do
Fetch first feature and stored in b, and other features in w.
$h_{w,b}(x) = g(z)$ here $z = (w^T x + b)$
if ($z \geq 0$)
assign g(z)=1;
else g(z)=-1;
end if
end for

Generate decision tree
Check if algorithm satisfies termination criteria
Computer information-theoretic criteria for all attributes
Choose best attribute according to the information-theoretic criteria
Create a decision node based on the best attribute in step 3
Induce (i.e. split) the dataset based on newly created decision node in step 4
For all sub-dataset in step 5, call C4.5 algorithm to get a sub-tree (recursive call)
Attach the tree obtained in step 6 to the decision node in step 4
Return tree
Input: an attribute valued dataset D
Tree={}
If D is "Pure" OR other stopping criteria met then Terminate
End if
For all attribute a Є D do
Compute information the reotic criteria if we split on a
End for
abest = Best attriubute according to above computed criteria
Tree= Create a decision node that tests a best in the root
Dv= Induced sub-Datasets from D based on a best
For all Dv do
Treev=C4.5(Dv)
Attach Treev to the corresponding branch of Tree
End for
Return Tree

II. SYSTEM REQUIREMENTS

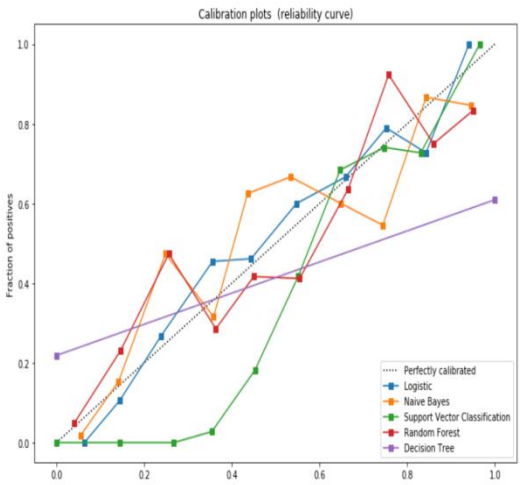
A. Software Requirement

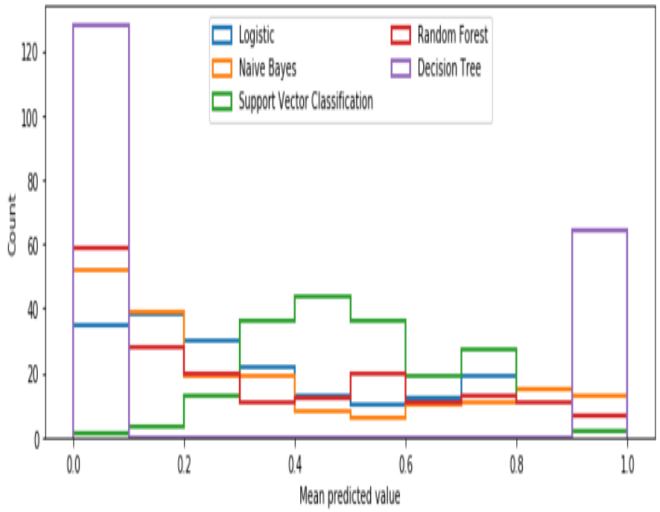
1. Operating System: Windows 7 or above
2. Programming Language: Python 3.7
3. IDE: Python IDLE

B. Hardware Requirement

1 Processor: Pentium Processor Core 2 Duo or Higher
2 Hard Disk: 250 GB (min)
3 RAM: 1GB or higher
4 Processor Speed: 3.2 GHz or faster processor

V. RESULT AND DISCUSSION





II. CONCLUSION

This system aimed to establish an appropriate prediction model for the high risk T2DM group. Based on a number of researchers' experiences, authors proposed a novel model, which consists of double-level algorithms, i.e., the improved K-means and logistic regression algorithms. In order to make a valid comparison with others' results, it was necessary to conduct this model using the WEKA toolkit and use the same Pima Indian Diabetes Dataset. Proper filters were utilized to improve the validity and rationality of the dataset. The proposed model that consisted of both cluster and class method ensured the enhancement of prediction accuracy.

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 difficult. Traditional manual data analysis has become inefficient and methods for efficient computer based analysis are needed. To this aim, many approaches to computerized data analysis have been considered and examined. Data mining represents a significant advance 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 resources.

A. Applicability

This study is also applicable to encourage and promote good health of people. In addition, the diabetes prediction will be created as a simple diagnosis application and will be published by a website. However, this application is only an initial diagnosis. People who found that they are in the diabetes risk group should go to see a doctor for formal diagnosis to prevent themselves from serious diabetes.

B. Abbreviations

Agency for Healthcare Research and Quality (AHRQ), Extreme Learning Machine (ELM), Principal Component Analysis (PCA), Support Vector Machine(SVM).

REFERENCES

1. Mustakim Al Helal, Atiqul Islam Chowdhury , Ashraful Islam , Eshtiak Ahmed , Md. Swakshar Mahmud, Sabrina Hossain "An Optimization Approach to Improve Classification Performance in Cancer and Diabetes Prediction" International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 February, 2019(Base Paper)
2. American Cancer Society, "Cancer Facts & Figures 2015," Cancer Facts Fig. 2015, pp. 1–9, 2015.
3. BCSC, "Types of Breast Cancer," Breast Cancer Society of Canada, 2014. [Online]. Available: http://www.bcsc.ca/p/41/l/506/t/Breast-Cancer-Society-of-Canada ---Types-of-Breast-Cancer.
4. M. Seera and C. P. Lim, "A hybrid intelligent system for medical data classification" Expert Syst. Appl., vol. 41, no. 5, pp. 2239–2249, 2014.
5. J. Tang, C. Deng, and G. Huang, "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, M. Gwinn, and M. J. Khoury, "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.