PREDICTION OF HEART DISEASE USING MACHINE LEARNING ALGORITHM

[1]Rahul Chaurasia, [2]Saksham Gupta and [3]Shipra Singh Siddhu
[1,2,3]Students, Dept. of Computer Science & Engineering, ABES Institute of Technology
Dr. A.P.J. Abdul Kalam Technical University, Uttar Pradesh

Sanjeev Kumar
Assistant Professor, CSE Department
ABES Institute of Technology
AKTU, Uttar Pradesh

Abstract

Data mining techniques have been applied magnificently in many fields including business, science, the Web, cheminformatics, bioinformatics, and on different types of data such as textual, visual, spatial, real-time and sensor data. Medical data is still information rich but knowledge poor. There is a lack of effective analysis tools to discover the hidden relationships and trends in medical data obtained from clinical records. Data mining techniques and machine learning algorithms play a very important role in this area. The researchers accelerating their research works to develop a software with the help machine learning algorithm which can help doctors to take decision regarding both prediction and diagnosing of heart disease. The main objective of this research paper is predicting the heart disease of a patient using machine learning algorithms.

Key words **- heart disease, detection technique, data mining technique, machine learning algorithm, K-nearest neighbor, support vector machine.**

Introduction

Data mining is a process of discovering/extracting the meaningful information from huge amount of data [1]. An extensively accepted formal definition of data mining is given subsequently "Data mining is the non-trivial extraction of implicit previously unknown and potentially useful information about data" [2]. The data mining techniques are very beneficial to predicting the various diseases in the healthcare industry. Disease prediction plays most important role in the data mining. The highest mortality of both India and abroad is due to heart disease. So, it is vital time to check this death toll by correctly identifying the disease in initial stage. The detection of heart disease from "various factors or symptoms is a multi-layered issue which is not free from false presumptions often accompanied by unpredictable effects" [3]. Thus, we can use patients' data that have been collected and recorded to ease the diagnosis process and utilize knowledge and experience of numerous specialists dealt with the same symptoms of diseases.

Heart Disease: -The heart attack occurs when the arteries which supply oxygenated blood to heart does not function due to completely blocked or narrowed.

Various types of heart diseases are [4]
1) Coronary heart disease
2) Cardiomyopathy
3) Cardiovascular disease
4) Ischaemic heart disease
5) Heart failure
6) Hypertensive heart disease
7) Inflammatory heart disease
8) Valvular heart disease

Common risk factors of heart disease include
1) High blood pressure

2) Abnormal blood lipids
3) Use of tobacco
4) Obesity
5) Physical inactivity
6) Diabetes
7) Age
8) Gender
9) Family Generation

Machine Learning is extensively used in diagnosing several diseases like heart [5] and other crucial diseases. Among various algorithms in data modeling, decision tree is known as the most popular due to its simplicity and interpretability [6], [7]. Nowadays more efficient algorithms such as SVM and artificial neural networks have also become popular [8], [7], [9].
The rest of the paper is organized as follows: Section II provides data description; Section III algorithm used; Section IV provided performance comparison. Section V concludes the paper.

PYTHON 3.6.4

It is an open source programming language for our experiment. The Python interpreter and the extensive standard library that are freely available in source or binary form for all major platforms from the Python Web and may be freely distributed. It has efficient high-level data structures and a simple but effective approach to object-oriented programming. Python's elegant syntax and dynamic typing, together with its interpreted nature, make it an ideal language for scripting and rapid application development in many areas on most platforms. The Python interpreter is easily extended with new functions and data types implemented in C or C++ (or other languages callable from C). Python is also suitable as an extension language for customizable applications. The new machine learning algorithms can be used with it and existing algorithms can also be extended with this language.
We have applied following two commonly used classifiers for prediction on the basis of their performance. These classifiers are as follows: Support Vector Machine, K-Nearest Neighbor

II. Dataset Description

We performed computer simulation on one dataset. Dataset is a Heart dataset. The dataset is available in UCI Machine Learning Repository [10]. This dataset was obtained from Cleveland database. This is publicly available dataset in the Internet. Cleveland dataset concerns classification of person into normal and abnormal person regarding heart diseases. Dataset contains 303 samples and 13 input features as well as 1 output feature. A list of all those features is given in Table 1.

Table 1: Features in the Dataset

Feature No.	Feature Name	Description
1	Age	Age in Years
2	Sex	1=male 0=female
3	Cp	Chest Pain Type: 1=typical angina 2=atypical angina 3=non-angina pain 4=asymptomatic
4	Trestbps	Resting blood pressure (in mm Hg)
5	Chol	Serum cholesterol in mg/dl
6	Fbs	Fasting Blood Sugar > 120 mg/dl: 1=true 0=false
7	Resteg	Resting electrocardiographic results: 0 = normal 1 = having ST-T wave abnormality

		2 =showing probable or define left ventricular hypertrophy by Estes 'criteria
8	Thalach	Maximum heart rate achieved
9	Exang	Exercise induced angina: 1 = yes 0 = no
10	Oldpeak	Depression induced by exercise relative to rest
11	Slop	The slope of the peak exercise segment: 1 = up sloping 2 = flat 3= down sloping
12	Ca	Number of major vessels colored by fluoroscopy that ranged between 0 and 3.
13	Thal	3 = normal 6= fixed defect 7= reversible defect
14	Num	Diagnosis classes: 0 = healthy 1= patient who is subject to possible heart disease

III. Algorithm used

K-Nearest Neighbor: - KNN is one of the most simple and straight forward data mining techniques. It is called Memory-Based Classification as the training examples need to be in the memory at run-time [11]. When dealing with continuous attributes the difference between the attributes is calculated using the Euclidean distance i.e. **dist(p,q) = sqrt($(p_1-q_1)^2 + p_2-q_2)^2 + \ldots + p_n-q_n)^2$)** .

A major problem when dealing with the Euclidean distance formula is that the large values frequency swamps the smaller ones. For example, in heart disease records the cholesterol measure ranges between 100 and 190 while the age measure ranges between 40 and 80. So the influence of the cholesterol measure will be higher than the age. To overcome this problem the continuous attributes are normalized so that they have the same influence on the distance measure between instances. KNN usually deals with continuous attributes however it can also deal with discrete attributes. When dealing with discrete attributes if the attribute values for the two instances are different so the difference between them is equal to one otherwise it is equal to zero.

Support Vector Machine: - Support Vector Machine (SVM) is a category of universal feed forward networks like Radial-basis function networks, pioneered by Vapnik. SVM can be used for pattern classification and nonlinear regression. More precisely, the support vector machine is an approximate implementation of the method of structural risk minimization. This principle is based on the fact that the error rate of a learning machine on test data is bounded by the sum of the training-error rate and term that depends on the Vapnik-Chervonenkis (VC) dimension. The support vector machine can provide good generalization performance on pattern classification problem [12].

Optimal Hyperplane for patterns: Consider the training sample where x_i is the input pattern for the ith instance and y_i is the corresponding target output. With pattern represented by the subset y_i= +1 and the pattern represented by the subset y_i= -1 are linearly separable. The equation in the form of a hyperplane that does the separation is:

$$wTx + b = 0 \qquad (3)$$

where x is an input vector, w is an adjustable weight vector, and b is a bias. Thus,

$$wTx_i + b \geq 0 \quad \text{for} \quad y_i = +1 \qquad (4)$$

$$wTx_i + b < 0 \quad \text{for} \quad y_i = -1 \qquad (5)$$

For a given weight vector w and a bias b, the separation between the hyperplane defined in eq. 3 and closest data point is called the margin of separation.

IV. PERFORMANCE COMPARISONS

Algorithm classification	Correct Classification Rate	Mis- Classification Rate
K-Nearest Neighbor(KNN) + Principle Component Analysis(PCA)	**84.44%**	**15.56%**

Support Vector Machine(SVM)	**73.77%**	**26.23%**

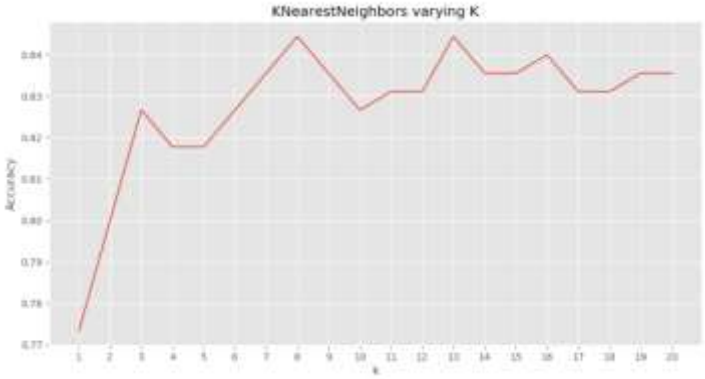

Fig-Accuracy of KNN

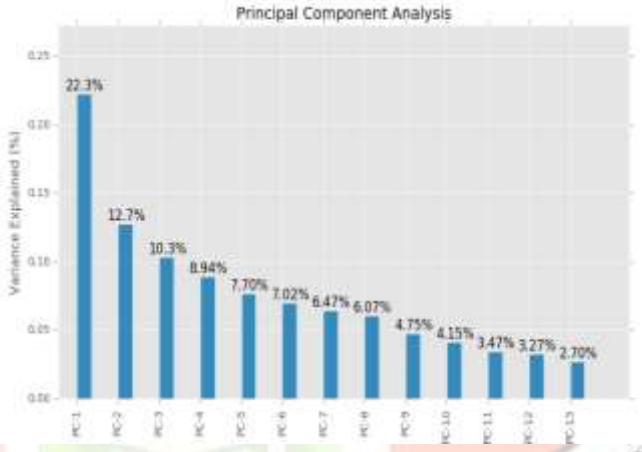

Fig-Principal Component Analysis Result

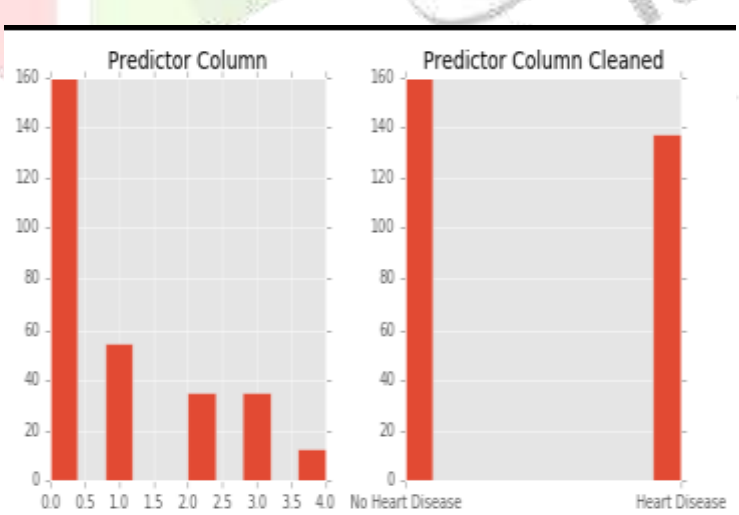

Fig-Principle Component Analysis Column Prediction

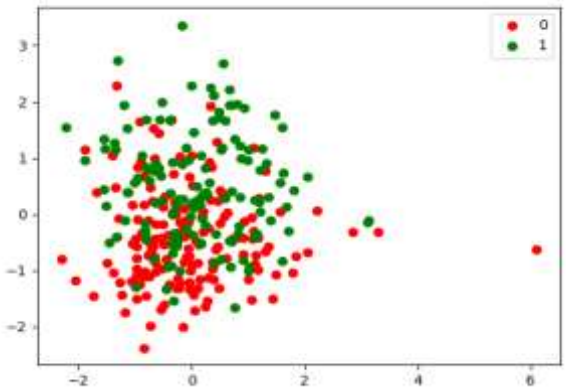

Fig-Classification of SVM

V. Conclusion

In this paper, we carried out an experiment to find the predictive performance of different classifiers. We select two popular classifiers considering their qualitative performance for the experiment. We also choose one dataset from heart available at UCI machine learning repository. In order to compare the classification performance of two machine learning algorithms, classifiers are applied on same data and results are compared on the basis of misclassification and correct classification rate and according to experimental results in table 1, it can be concluded that K-Nearest Neighbor with principle component analysis is best as compared to Support Vector Machine.

After analyzing the quantitative data generated from the computer simulations, moreover their performance is closely competitive showing slight difference. So, more experiments on several other datasets need to be considered to draw a more general conclusion on the comparative performance of the classifiers.

Reference

1. Chaitrali S. Dangre and Dr. Mrs. Sulbha S. Apte, “Improved Study of Heart Disease Prediction System Using Data Mining Classification Techniques”, International Journal of Computer Applications, Vol.47, No.10, pp. 0975 – 888, June 2012.
2. Frawley and Piatetsky-Shapiro, Knowledge Discovery in Databases: An Overview. The AAAI/MIT Press, Menlo Park, C.A, 1996.
3. S. B. Patil and Y. S. Kumaraswamy, "Extraction of significant patterns from heart disease warehouses for heart attack prediction," International Journal of Computer Science and Networks Security, vol. 9, pp. 228-235, 2009.
4. B.L Deekshatulua Priti Chandra “Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm “ M.Akhil jabbar* International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA) 2013.
5. S. Shilaskar and A. Ghatol, “Feature selection for medical diagnosis : Evaluation for cardiovascular diseases,” Expert Syst. Appl., vol. 40, no. 10, pp. 4146–4153, Aug. 2013.
6. C.-L. Chang and C.-H. Chen, “Applying decision tree and neural network to increase quality of dermatologic diagnosis,” Expert Syst. Appl., vol. 36, no. 2, Part 2, pp. 4035–4041, Mar. 2009.
7. A. T. Azar and S. M. El-Metwally, “Decision tree classifiers for automated medical diagnosis,” Neural Comput. Appl., vol. 23, no. 7–8, pp. 2387–2403, Dec. 2013. [10] Y. C. T. Bo Jin, “Support vector machines with genetic fuzzy feature transformation for biomedical data classification.,” Inf Sci, vol. 177, no. 2, pp. 476–489, 2007.
8. N. Esfandiari, M. R. Babavalian, A.-M. E. Moghadam, and V. K. Tabar, “Knowledge discovery in medicine: Current issue and future trend,” Expert Syst. Appl., vol. 41, no. 9, pp. 4434–63,7/2014.
9. A. E. Hassanien and T. Kim, “Breast cancer MRI diagnosis approach using support vector machine and pulse coupled neural networks,” J. Appl. Log., vol. 10, no. 4, pp. 277–284, 12/2012.
10. UCI Machine Learning Repository, Available at http://archive.ics.uci.edu/ml/machinelearningdatabases/ statlog/german/
11. E. Alpaydin, Voting over Multiple Condensed Nearest Neighbors. Artificial Intelligence Review, pp. 115–132. 1997.
12. Christopher J.C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery, Springer, 2(2), pp.121-167, 1998.