Heart Disease and Diabetes Prediction using Data Mining

Tanmay Tamhane[1] Mateen Shaikh[2] Sanjaykumar Boga[3] Mrunal Tanwar[4] A.E. Patil[5]
[1,2,3,4,5]Department of Information and Technology
[1,2,3,4,5]Rajiv Gandhi Institute of Technology, Mumbai, Maharashtra, India

Abstract— The healthcare industry collects huge amounts of health related data which, unfortunately, is not 'mined' to discover hidden information for effective decision making. Discovery of hidden patterns and their relationships often goes undetected. Advanced data mining techniques can help remedy this situation. Our project describes about using data mining techniques, namely Naive Bayesian and WAC (Weighted Associative Classifier).In our system, we will store patients' health record details and update it on that patient's regular visits. The doctor will check the patient's record from time to time. And, based on the results shown by our system, the doctor will come to know the occurrence of a disease in future.

Key words: Naive Bayesian, WAC (Weighted Associative Classifier, Heart Disease, Diabetes, Data Mining

I. INTRODUCTION

The major challenge facing the healthcare industry is the provision for quality services at affordable costs. A quality service implies diagnosing patients correctly and treating them effectively. Poor clinical decisions can lead to disastrous results which is unacceptable. Even the most technologically advanced hospitals in India have no such software that predicts a disease through data mining techniques. There is a huge amount of untapped data that can be turned into useful information.

Medical diagnosis is known to be subjective; it depends on the physician making the diagnosis. Secondly, and most importantly, the amount of data that should be analyzed to make a good prediction is usually huge and at times cannot be managed. In this context, machine learning can be used to automatically infer diagnostic rules from descriptions of past, successfully treated patients, and help specialists to make the diagnostic process more objective and more reliable

Our system can answer complex queries, which conventional decision support systems cannot. Using medical parameters such as age, sex, blood pressure and blood sugar it can predict the likelihood of patients' getting a heart disease or diabetes in future. In enables significant knowledge, for example, patterns, relationships between medical factors related to heart disease and diabetes, to be established. It can serve as a training tool to train nurses and medical students to diagnose patients. It is a web based user friendly system and can be used in hospitals, if they have a data warehouse for their hospital.

II. PROPOSED SYSTEM

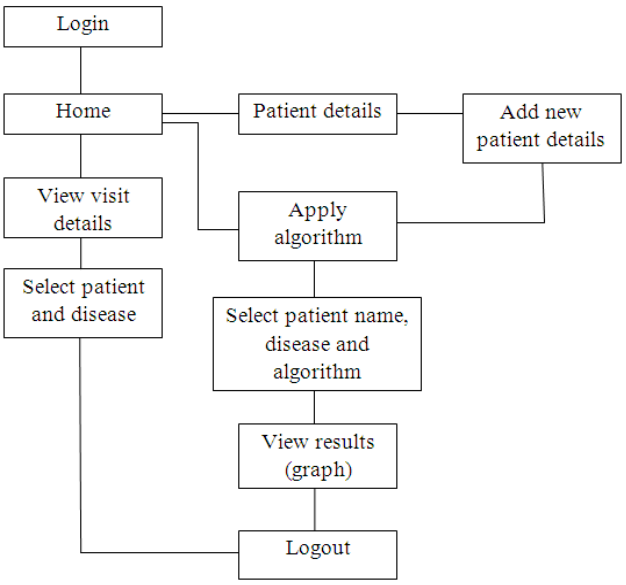

Fig. 1: Block Diagram of the System

A. How it works?

Firstly, Admin logs in to the site. After authentication of the admin he is directed to homepage. From home page he can do different things like view patient details, he can also visit particular patients last visit details and also apply algorithm by selecting the records of a particular patient.

If he chooses to see the patient details then he is shown different patient details, from that he may even add new patient's records. Then if he wishes to apply algorithm to check the disease, he may view the graph of the patient's results from past visits. If the Admin wishes to just view the details of the patient then he may go to view visit details and there he should enter the patient name or id and then he can view the details.

The Admin may directly go to apply algorithm phase there he needs to provide the patient name, disease and algorithm by which he wants to view the result after successful viewing of the result he may log out.

III. EXPERIMENTAL DATA

We have used medical datasets of heart disease and diabetes obtained from UC Irvine Archive of machine learning datasets [1]. There are 14 attributes of heart disease dataset and 9 attributes of Diabetes.

No	Name	Description
1	Age	Age in years
2	Sex	1=male; 0=female
3	Cp	Chest pain type(1=typical angina ;2= atypical angina ;2= non- angina pain; 4=Asymptomatic)
4	Trestbps	Resting blood pressure(in mm Hg on admission to the hospital)

5	Chol	Serum cholesterol in mg/dl
6	Fbs	(fasting blood sugar>120 mg/dl)(1=true;0=false)
7	Restecg	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	ST depression induced by exercise relative to rest
11	Slope	The slope of the peak exercise ST segment(1=upsloping;2=flat;3=downsloping)
12	Ca	Number of major vessels(0-3) colored by flourosopy
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)

Table 1: Heart Disease Dataset

No	Attribute	Description
1	Number of times pregnant	Numerical value
2	Plasma glucose concentration	Glucose concentration in a 2 hours in an oral glucose tolerance test
3	Diastolic blood pressure	In mm HG
4	Triceps skin fold thickness	Thickness of skin in mm
5	2-hour serum insulin	Insulin(mu U/ml)
6	Body mass index	(weight in kg/ height in m)^2)
7	Diabetes pedigree function	A function – to analyze the presence of diabetes
8	Age	Age in years
9	Class	1 is represented as "tested positive for diabetes and 0 as negative

Table 2: Diabetes Dataset

IV. METHODOLOGY

A. *Naive Bayesian Algorithm:*

The Naive Bayesian Classifier technique is based on the so-called Bayesian theorem and is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayesian can often outperform more sophisticated classification methods [2].

Abstractly, naive Bayesian is a conditional probability model: given a problem instance to be classified, represented by a vector $x=(x_1,\ldots,x_n)$ representing some *n* features (dependent variables), it assigns to this instance probabilities.

$$p(C_k|x_1,\ldots,x_n)$$

for each of *k* possible outcomes or classes.

The problem with the above formulation is that if the number of features *n* is large or if a feature can take on a large number of values, then basing such a model on probability tables is infeasible. We therefore reformulate the model to make it more tractable [4]. Using Bayesian' theorem, the conditional probability can be decomposed as

$$p(C_k|\mathbf{x}) = \frac{p(C_k)\ p(\mathbf{x}|C_k)}{p(\mathbf{x})}.$$

In plain English, using Bayesian probability terminology, the above equation can be written as

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}.$$

TP Rate	FR Rate	Precison	Recall	F-Measure	ROC	Class
0.702	0.258	0.702	0.702	0.702	0.726	NO
0.742	0.298	0.742	0.742	0.742	0.726	YES

Table 3: Confusion Matrix of Naïve Bayesian Heart Disease Dataset

TP Rate	FR Rate	Precison	Recall	F-Measure	ROC	Class
0.844	0.388	0.802	0.844	0.823	0.819	Tested negative
0.612	0.156	0.678	0.612	0.643	0.819	Tested positive

Table 4: Confusion Matrix of Naïve Bayesian Diabetes Dataset

B. *WAC (Weighted Associative Classifier):*

It is a concept that uses Weighted Association Rule for classification. Weighted Support and Confidence Framework to extract Association rule from data repository is used by Weighted ARM. The WAC is new Technique to get the significant rule instead of flooded with insignificant relation. Each attribute is assigned a weight ranging from o to 1 to reflect their importance in prediction model. Attributes that have more impact will be assigned a high weight (nearly 0.9) and attributes having less impact are assigned low weight (nearly 0.1) [2].

Once the pre-processing gets over, Weighted Association Rule Mining (WARM) algorithm is applied to generate interesting pattern.

This algorithm uses the concept of Weighted Support and Confidence framework instead of tradition support and confidence. Rules generated in this step are known as CAR (Classification Association Rule) and is represented as X Class label where X is set of symptoms for the disease. Examples of such rules are,

(Hypertension,"yes")→Heart_Disease="yes" And {(Age,">62"),(Smoking_habits,"yes"),(Hypertension,"yes")}→Heart_Disease="yes".[3]

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC	Class
0.685	0.281	0.66	0.686	0.673	0.68	No
0.719	0.814	0.742	0.719	0.73	0.719	Yes

Table 5: Confusion Matrix of WAC Heart Disease Dataset

TP Rate	FP rate	Precision	Recall	F-Measure	ROC	Class
0.868	0.466	0.776	0.868	0.82	0.727	Tested negative
0534	0.132	0.684	0.534	0.6	0.727	Tested positive

Table 6: Confusion Matrix of WAC Diabetes Dataset

V. ACKNOWLEDGEMENT

We wish to express our sincere gratitude to Dr. U. V. Bhosle, Principal and Prof. D.M.Dalgade, H.O.D of Information Technology Department of RGIT for providing us an opportunity to do our project work on "Heart Disease and Diabetes Prediction using Data Mining". This project bears on imprint of many people. We sincerely thank our project guide A.E Patil for his guidance and encouragement in successful completion of our project synopsis. We would also like to thank our staff members for their help in carrying out this project work. Finally, we would like to thank our colleagues and friends who helped us in completing the project synopsis successfully.

VI. CONCLUSION

In this research paper we have used Naive Bayesian algorithm and WAC algorithm to efficiently predict Heart Disease and Diabetes. This will help doctors in diagnosing the patients. The technology used is .NET and MS SQL Server 2008.

VII. FUTURE SCOPE

Following features can be added:

- Suggestion of medicines for the diagnosed disease can be made.
- Add diagnosis of more diseases such as Hepatitis, Tuberculosis etc.
- Making the interface more user friendly.

REFERENCES

[1] UCI Machine Learning Repository http://www.ics.uci.edu/~mlearn/MLRepository.html
[2] Jiawei Han and Micheline Kamber,"Data Mining Concepts and Techniques", 2Ed. M.K Publications.
[3] Jyoti Soni, Uzma Ansari, Dipesh Sharma, Intelligent and Effective Heart Disease Prediction System using Weighted Associative Classifiers, IJSCN,Vol. 3 No. 6 June 2011.
[4] N. Aditya Sundar, P. Pushpa Latha, M. Rama Chandra, Performance analysis of classification Data Mining techniques over Heart Disease Data base, ISSN, Vol.2, May-Jun 2012.

