Rizvi College of Engineering

Department of Computer Engineering

Project Synopsis Report

on

Medical Diagnosis Using Data Mining

Submitted in partial fulfilment of the requirements

of the degree of

Bachelors of Engineering

by

Huzaifa Vakil(70)

Zain Momin(33)

Asad Siddiqui(63)

Sohel Tharani(68)

Guide:

Dr. Anupam Choudhary

University of Mumbai

2018 - 2019

Abstract

The healthcare industry collects a huge amount of data which is not properly mined and not put to the optimum use. Discovery of these hidden patterns and relationships often goes unexploited. Our research focuses on this aspect of Medical diagnosis by learning pattern through the collected data of diabetes, hepatitis and heart diseases and to develop intelligent medical decision support systems to help the physicians. In this project, we propose the use of decision trees C4.5 algorithm and Naïve bayes algorithm to classify these diseases and compare the effectiveness, correction rate among them.

Keywords: Active learning, decision support system, data mining, medical engineering, ID3 algorithm, CART algorithm, C4.5 algorithm.

Department of Computer Engineering

Rizvi College of Engineering,

Off Carter Road, Bandra(W), Mumbai-400050.

Certificate

This is to certify that the project synopsis entitled "Medical Diagnosis Using Data Mining" has been submitted by Huzaifa Vakil, Zain Momin, Asad Siddiqui and Sohel Tharani under the guidance of Prof. Anupam Choudhary in partial fulfillment of the requirement for the award of the Degree of Bachelor of Engineering in Computer Engineering from University of Mumbai.

Certified By

Prof. Anupam Choudhary	Prof. ShiburajPappu
Project Guide	Head of Department

Prof. Anupam Choudhary	Prof. _____
Internal Examiner	External Examiner

Dr. Varsha Shah

Principal

Index

Chapter 1

Introduction

The healthcare industry collects a huge amount of data which is not properly mined and not put to the optimum use. Discovery of these hidden patterns and relationships often goes unexploited. Advanced data mining modeling techniques can help overcome this situation.

There is a huge amount of untapped data that can be turned into useful information. The decision support systems that have been developed to assist physicians in the diagnostic process often are based on static data which may be out of date. A decision support system which can learn the relationships between patient history, diseases in the population, symptoms, pathology of a disease, family history and test results, would be useful to physicians and hospitals. The concept of Decision Support System (DSS) is very broad because of many diverse approaches and a wide range of domains in which decisions are made. DSS terminology refers to a class of computer-based information systems including knowledge based systems that support decision making activities. In general, it can say that a DSS is a computerized system for helping make decisions. A DSS application can be composed of the subsystems. However, the development of such system presents a daunting and yet to be explored task. Many factors have been attributed but inadequate information has been identified as a major challenge. To reduce the diagnosis time and improve the diagnosis accuracy, it has become more of a demanding issue to develop reliable and powerful medical decision support systems (MDSS) to support the yet and still increasingly complicated diagnosis decision process.

The medical diagnosis by nature is a complex and fuzzy cognitive process, hence soft computing methods, such as decision tree classifiers have shown great potential to be applied in the development of MDSS of heart diseases and other diseases. The aim is to identify the most important risk factors based on the classification rules to be extracted.

Chapter 2

Aim and Objective of the project

Aim:

Medical Diagnosis Using Data Mining and to identify the most important risk factors based on the classification rules to be extracted. This will explain how well data mining and decision support system are integrated and also describes the datasets undertaken for this work.

Objective:

1. It helps decision makers utilize data and models in order to identify problems.
2. It solves problems and makes decisions.
3. It incorporates both data and models and they are designed to assist decision makers in semi-structured and unstructured decision making processes.
4. It provides support for decision making.
5. It improves effectiveness, rather than the efficiency of decisions.

Chapter 3

Literature Survey

1. Decision Support System for Medical Diagnosis Using Data Mining
   IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 3, No. 1, May 2011. D.Senthil Kumar, G.SathyadeviandS.Sivanesh
   a. In this we got an idea of medical diagnosis and the decision tree algorithm for effective classification.
   b. We have found 83.184% accuracy with the CART algorithm which is greater than previous research of ID3.

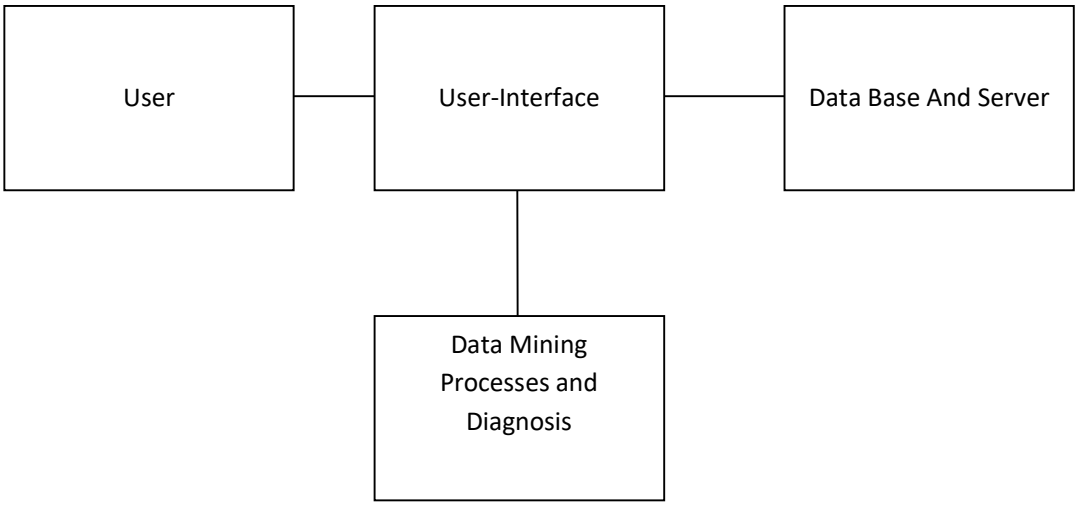
2. USING DATA MINING TECHNIQUES FOR DIAGNOSIS AND PROGNOSIS OF CANCER DISEASE International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.2, April 2012N. Aditya Sundar, P. Pushpa Latha, M. Rama Chandra
   a. The system extracts hidden knowledge from a historical heart disease database.
   b. Classification Matrix methods are used to evaluate the effectiveness of the models. The two models are able to extract patterns in response to the predictable state.
   c. In future the predictor can be used to design a web-based application to accept the predictor variables and Automated system Decision Tree based prediction can be implemented in remote areas like rural regions or country sides, to imitate like human diagnostic expertise for prediction of ailment.

Chapter 4

Proposed System, Methodology and Status of work

4.1 Proposed System

Our system will contain User-interface, Data Base and Data sets.



User: The user will enter all the medical records of the patient to find a suitable diagnosis.

User Interface: It will provide a graphical Interface so that the user can enter the valid parameters related to the patients. The user interface will also provide the diagnosis as output.

Database and Datasets: The database will provide us with the datasets of patients having heart related problems so that the Data Mining algorithms can scan the data for possible diagnosis.

Data Mining Processes and Diagnosis: The algorithms will take inputs from the user and the datasets from the database to find a possible diagnosis.

For the better efficiency we implemented the Naïve Bayes and C4.5 Algorithm.

4.2 C4.5 Algorithm:-

C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan. C4.5 is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier. Authors of the Weka machine learning software described the C4.5 algorithm as "a landmark decision tree program that is probably the machine learning workhorse most widely used in practice to date".

C4.5 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy. The training data is a set S={s1,s2,...} of already classified samples. Each sample si consists of a p-dimensional vector (x{1,i},x{2,i},...,x{p,i}) , where the xj represent attribute values or features of the sample, as well as the class in which si falls.

At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized information gain (difference in entropy). The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then recurs on the smaller sublists.

C4.5 is a software extension of the basic ID3 algorithm designed by Quinlan to address the following issues not dealt with by ID3:

- Avoiding overfitting the data
- Determining how deeply to grow a decision tree.
- Reduced error pruning.
- Rule post-pruning.
- Handling continuous attributes.
- e.g., temperature
- Choosing an appropriate attribute selection measure.
- Handling training data with missing attribute values.
- Handling attributes with differing costs.
- Improving computational efficiency.

For example:-

CLASSIFICATION RULES

Significant rules are extracted which are useful for understanding the data pattern and behavior of experimental dataset. The following pattern is extracted by applying C4.5 decision tree algorithm.

Rules Extracted for Heart Disease Dataset are as follows,

1. Heartdisease(absence):-Thal=fixed_defect,Number_Vessels=0, Cholestoral =126-213.
2. Heart_disease(presence):-Thal=normal,Number_Vessels=0,Old_Peak=0-1.5, Max_Heart_Rate=137-169, Cholestoral=126-213.
3. Heart_disease(absence):-Thal=normal,Number_Vessels=0,Old_Peak=0-1.5, Max_Heart_Rate=137 169,Cholestoral=214-301, Rest=0, Pressure=121-147.

4.3 Naïve Bayes:-

Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of independence between every pair of features. Given a class variable y and a dependent feature vector $x_1$through x_n, Bayes' theorem states the following relationship:

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y)P(x_1, \ldots x_n \mid y)}{P(x_1, \ldots, x_n)}$$

Using the naive independence assumption that

$$P(x_i|y, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = P(x_i|y),$$

for all i, this relationship is simplified to

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y) \prod_{i=1}^{n} P(x_i \mid y)}{P(x_1, \ldots, x_n)}$$

Since $P(x_1, \ldots, x_n)$ is constant given the input, we can use the following classification rule:

$$P(y \mid x_1, \ldots, x_n) \propto P(y) \prod_{i=1}^{n} P(x_i \mid y)$$

$$\Downarrow$$

$$\hat{y} = \arg\max_y P(y) \prod_{i=1}^{n} P(x_i \mid y),$$

and we can use Maximum A Posteriori (MAP) estimation to estimate $P(y)$ and $P(x_i \mid y)$; the former is then the relative frequency of class y in the training set.

The different naive Bayes classifiers differ mainly by the assumptions they make regarding the distribution of $P(x_i \mid y)$.

In spite of their apparently over-simplified assumptions, naive Bayes classifiers have worked quite well in many real-world situations, famously document classification and spam filtering. They require a small amount of training data to estimate the necessary parameters.

Naive Bayes learners and classifiers can be extremely fast compared to more sophisticated methods. The decoupling of the class conditional feature distributions means that each distribution can be independently estimated as a one dimensional distribution. This in turn helps to alleviate problems stemming from the curse of dimensionality.

Chapter 5

Methodology

This system uses the CRISP-DM (cross industry standard process for data mining) methodology to build the mining models. It consists of six major phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. Business understanding phase focuses on understanding the objectives and requirements from a business perspective, converting this knowledge into a data mining problem definition, and designing a preliminary plan to achieve the objectives.

Data understanding phase uses the raw the data and proceeds to understand the data, identify its quality, gain preliminary insights, and detect interesting subsets to form hypotheses for hidden information. Data preparation phase constructs the final dataset that will be fed into the modeling tools. This includes table, record, and attribute selection as well as data cleaning and transformation. The modeling phase selects and applies various techniques, and calibrates their parameters to optimal values. The evaluation phase evaluates the model to ensure that it achieves the business objectives. The deployment phase specifies the tasks that are needed to use the models.

Data Mining Extension (DMX), a SQL-style query language for data mining, is used for building and accessing the models‟ contents. Tabular and graphical visualizations are incorporated to enhance analysis and interpretation of results.

5.1 Datasets

Attribute value to be taken into the project for disease.

Heart disease dataset

No.	Name
1	Age
2	Sex
3	Cp
4	Trestbps
5	Chol
6	Fbs
7	Restecg
8	Thalach
9	Exang
10	Oldpeak
11	Slope
12	Ca
13	Thal
14	Num

Status of Work

1. Synopsis prepared
2. The Paper "MEDICAL DIAGNOSIS USING DATA MINING" accepted and in-progress of publication.

Conclusion

The decision-tree algorithm is one of the most effective classification methods. The data will judge the efficiency and correction rate of the algorithm. We used 10-fold cross validation to compute confusion matrix of each model and then evaluate the performance by using precision, recall, F measure and ROC space. C4.5 algorithm will show the best performance among the tested methods. The results that will be shown here make clinical application more accessible, which will provide great advance in Heart Disease.

The project will be developed on the decision tree algorithms C4.5 and Naïve Bayes classifier towards their steps of processing data and Complexity of running data. The C4.5 algorithm performs better in performance of rules generated and accuracy. This will show that the C4.5 algorithm is better in induction and rules generalization compared to ID3 algorithm. Finally, the results will be stored in the decision support repository. Since, the knowledge base is currently focused on a narrow set of diseases. The approach will validate and it is possible to expand the scope of modeled medical knowledge. Furthermore, in order to improve decision support, interactions should be considered between the different medications that the patient is on.

REFERENCES

1. Decision Support System for Medical Diagnosis Using Data Mining
   IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 3, No. 1, May 2011
   ISSN (Online): 1694-0814 www.IJCSI.org

2. USING DATA MINING TECHNIQUES FOR DIAGNOSIS
   AND PROGNOSIS OF CANCER DISEASE International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.2, April 2012 Shweta.bitdurg@gmail.com

3. PERFORMANCE ANALYSIS OF CLASSIFICATION DATA MINING TECHNIQUES OVER HEART DISEASE DATA BASE
   [**IJESAT]** INTERNATIONAL JOURNAL OF ENGINEERING SCIENCE & ADVANCED TECHNOLOGY Volume-2, Issue-3, 470 – 478 http://www.ijesat.org

Appendix-I: Paper Publication

The Paper ”MEDICAL DIAGNOSIS USING DATA MINING” have been submitted and accepted with Uniqueness of 76% in International Journal of Emerging Technologies and Innovative Research (JETIR), Registration ID: JETIR90293.

Authors: Huzaifa Vakil, Zain Momin, Asad Siddiqui and Sohel Tharani.

Guide: Dr. Anupam Choudhary