Study On the Classification Algorithms for Prediction of Diseases

Dr.P.Radha,R.Divya

Assistant professor, PG and Research Department of computer Application,
Government Arts college (Autonomous), Coimbatore, Tamil Nadu ,India.
Research scholar, Government .Arts College (Autonomous) ,Coimbatore. Tamil Nadu, India.
r.divyarun@gmail.com

***Abstract*—-Data mining is the extraction of information from a large database. Huge amount of data related to healthcare are available but information is not extracted from the huge data. There is a less amount of effective analysis tool to discover the hidden information and trends in data. In health care department data mining is mainly used for disease prediction. There are many data mining techniques present for predicting the disease like association rules, classification, clustering. In this paper, the study is done on the prediction of diseases etc based on the classification algorithms and the best classification algorithm is found. In this paper comparison of accuracy of the prediction of diseases is also done.**

Index Terms- **Data Mining, C4.5, K-NN, SVM, Naïve Bayes.**

I. Introduction

Data mining is a non- trival process of categorizing valid, novel, potentially useful and ultimately understandable patterns in data. Data mining is used to discover knowledge from data bases [1]. In other words knowledge mining from data, knowledge extraction, data/patterns analysis, data archeology, and data dredging .Many people treat data mining as KDD (Knowledge discovery from data, or KDD [10]. The Knowledge discovery process is 1. Data cleaning (to remove noise and inconsistent data), 2. Data integration (where multiple data sources combined) 3.Data selection (where data relevant to the analysis task are retrieved from the data base) 4.Data transformation, 5.Data mining, 6. Pattern evaluation,
7. Knowledge presentation. Medical data mining has been an efficient exploring for the hidden information in data sets of medical field. In health care field in older days data mining is not widely used, but now days, it become most popularly used method. There are two main goals in data mining one is the prediction and another one is description. Prediction includes some variables or fields in the data set to predict unknown or future values of other variables of interest. Description involves on finding the patterns and describing the data that can be understand by people.

With the fast increase in population the number of diseases also increased. In medical field most of the diseases are related symptoms which make difficult to predict the disease for the doctors. For this problem the data mining is used for predicting the disease. In this paper the study is done on the prediction of disease based on the classification algorithms.

II. Existing System

The study on the disease prediction research, there are two types of systems are accessible for this function. The area exact for predicting the particular diseases for eg system mean with predicting the cancer disease only it won't find the other diseases. Another type is the method which concentrates on many disease predictions.

Further the medical prediction method is splitted in to two groups such as the method which uses the symptoms to find the disease and not considering the family medical history, age, etc and another one is considering the family medical history, age etc.

In the existing system the study was based on the prediction of diseases using the classification algorithms and it is compared with efficiency of prediction method. In many papers the researchers used the classification methods to predict the diseases.

III. LITERATURE REVIEW

DOMAIN CERTAIN ALGORITHM FOR PREDICTION

There are many particular diseases prediction algorithms based on the survey of diseases like Diabetes, Breast Cancer, Heart Disease, Cancer and Kidney Disease.

A. DIABETES MELLITUS

Diabetes mellitus is one of the mainly crucial health challenges in both developing and developed countries [2]. The Pima Indian diabetic database at the UCI machine learning laboratory has become a standard for testing data mining algorithms to see their prediction accuracy in diabetes data classification.

In this paper, the study of the comparison is done on the classification algorithms.

METHODOLOGY:

- The basic step is the pre-processing; in this step repeated values are removed.
- C4.5 algorithm is used to find the accuracy.
- The missing values are interchange by means and Medians

© 2016 IJAIR. All Rights Reserved

- After the preprocessing method the SVM algorithm is used to find the accuracy.
- This study includes data preprocessing, prediction, accuracy
- Data Pre-Processing: uses attribute identification and selection; remove redundant values to improve the quality of the data.

Classification: After preprocessing, classification algorithm is used to classify the data. By applying the classification algorithm, it is used to predict the disease and it shows the accuracy of the prediction.

Table 1: [1] Summary of disease prediction for diabetes

Algorithm	Various measures used
C4.5	Accuracy 71.1%
SVM	Accuracy 78%

B. KIDNEY DISEASE

- Prediction of kidney disease is important task in medical field [3].
- The synthetic kidney function test (KFT) dataset have been initiate for study of kidney disease.
- This dataset contains 584 examples and 6 attributes are used in this paper.
-

METHODOLOGY:

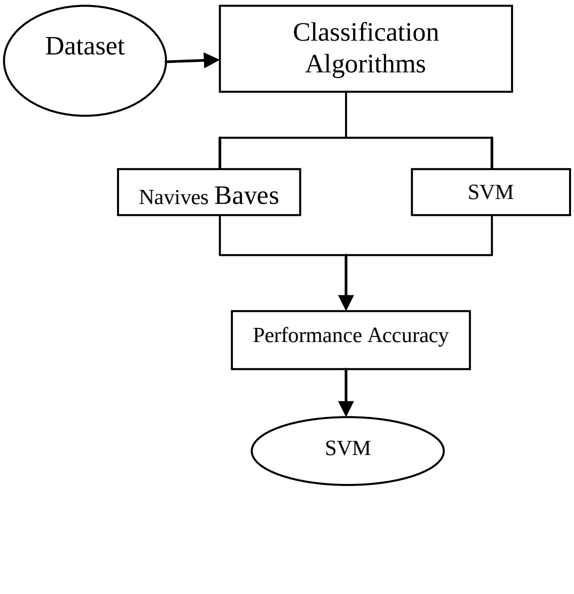

Table 2: Summary of disease prediction for [3] kidney disease

Algorithm	Various measures used
Naïve Bayes	Accuracy 70.96%
SVM	Accuracy 76.32%

C. BREAST CANCER DISEASE

Breast cancer is mainly a heterogeneous disease. Breast Cancer prediction and prognosis are two medical challenges to the research [4] .For this paper UCI dataset has been used .The important type of breast cancer is ductal carcinoma, which starts in the lining of milk ducts. The next type of breast cancer is lobular carcinoma, which starts in the lobules of breast.

Data Mining is the mighty tool and technique to handling this task. In data mining breast cancer research has been one of the main topics in the medical field during the recent years. The classification of breast cancer data is used for the prediction of the result of some other diseases. There is lot of techniques to predict the breast cancer.

METHODOLOGY:

In this paper the C4.5 algorithm is compared with SVM algorithm.

Table 3: Summary of disease prediction for [4] Breast cancer

Algorithm	Various measures used
C4.5	Accuracy 79.1%
SVM	Accuracy 96.74%

D. CANCER

In this paper the author used J48 classification algorithm for finding the breast cancer by using two levels [6]. The First level the prediction is done on the basis of Wisconsin Breast Cancer Dataset (WBCD); the result find from the WBCD is categorized into malignant and benign classes. The next level prediction is based on the pathological and physiological parameters of malignant breast cancer dataset.

METHODOLOGY:
In this paper, the comparison was done on SVM classifier with J48 algorithm. The result was SVM prediction accuracy is better than J48 algorithm.

Table 4: Summary of disease prediction for [6] Cancer

Algorithm	Various measures used
J48	Accuracy-81%
SVM	Accuracy-95%

E. HEART DISEASE

In presence of days, Heart Disease is a major cause of issues in world [9]. All over the world, deaths due to Heart Disease is increasing quickly than from any other disease. It is very difficult to predict the possible difficult regarding Heart Disease in advance. To identify the possible complications, many systems are made that uses clinical data sets for identifications. Some of the systems predict heart disease based on risk factors. A lot of visible risk factors that are common in Heart Disease patients can be used effectively for prediction. System based on risk factors helps not only medical experts for prediction but also warn the patients in advance about the possible presence of heart disease.

Data mining tools used for this system are Support Vector Machine and Genetic Algorithm

METHODOLOGY:
The dataset contains 50 people data collected from different services done by American Heart Association .In this paper SVM weight optimization by Genetic Algorithm was used. This system uses Linear Kernel function for learning and training the SVM Classifier. But the results produced by this were not much better. To get more accurate results the genetic algorithm is used with SVM as optimizer.

SVM CLASSIFIER
A SVM Classifier is used having input space and feature space. The input space is based on the final set of risk factors for each patient. The feature space is obtained by maximizing the margin between the two classes for which training is fast and gives the best output results.

Table 5: Summary of disease prediction for [9] Heart Disease

Algorithm	Various measures used
Genetic Algorithm +Neural network	0.90444
SVM classifier	0.95152

LIVER DISEASE

In this research work, Naïve Bayes and Support Vector Machine (SVM) classifier algorithms are used for predict the liver disease. The liver is the second largest internal organ in the human body, playing a major role in metabolism and many functions in our body [11]
DATASET

Indian Liver Patient Dataset (ILPD) was used and taken from UCI repository. In this dataset there are 576 instances and ten attributes. The attributes are Age, Gender, TB, DB, albumin, A/G Ratio, SGPT, SGOT, and Alkphos.

CLASSIFICATION ACCURACY

Accuracy $\dfrac{TP+TN}{TP+FP+TN+FN}$

Where TP-True positive, FP-False Negative, TN- True Negative, FN-False Negative

TP Rate: It is used to find the high true-positive rate. It is called as Sensitivity.

$$TPR=\frac{TP}{TP+FN}$$

Precision*:* $\dfrac{TP}{TP+FP}$

MULTIPLE DISEASES

In paper [7] Mediquery medical system is proposed for the disease prediction based on symptoms, current medical history, taking different tests.

METHODOLOHY
In this method the diseases are identified by these methods.
In step-1. The disease is predicted based on the symptoms.

In step-2. The diseases are predicted by the recent medical history.

In step-3.The disease is identified by taking the various tests.

IV RESULTS

Results shows that in prediction of diabetes disease SVM classification shows the accuracy rate higher than C4.5. In predicting the breast cancer disease C4.5 shows the less accuracy than SVM. In prediction of kidney disease the Naïve Bayes algorithm is compared with the SVM. The SVM algorithm results the higher accuracy. In prediction of heart disease the author compare the genetic algorithm with the SVM classifier, the best accuracy was SVM classifier.

Table 7: Summary of best classification algorithm on prediction of diseases.

Disease name	ALGORITHM USED	BEST ACCURACY
Diabetes	C4.5,SVM	SVM-78%
Kidney disease	Naïve Bayes,SVM	SVM-76.32%
Breast cancer	C4.5,SVM	SVM-96.74%
Cancer disease	J48,SVM	SVM-95%
Heart disease	Genetical algorithm, SVM	SVM-0.95%
Liver Disease	Naïve Bayes,SVM	SVM-79.66%

V.DATASET CONSIDERED

PIDD-Pima Indian Diabetes Database Several constraints are considered on the selection of these examples from a large database.

UCI Repository-According to the survey of United States in 2014, there are 232,670 females and 2,360 males having this type of cancer.

WBCD-Wisconsin Breast Cancer Dataset

MIAS-Mammography Image Analysis Society

ILPD-Indian Liver Patient Dataset

VI CONCLUSION

In this paper the survey was done on the classification algorithms on the prediction of diseases and the study was done on the data preprocessing methods. After compared on the all the classification papers the SVM algorithm shows the best accuracy result. In health care department by using data mining techniques the disease are predicted the quickly. Prediction is the one of the important factor in the classification.

The survey is done on various data mining classification algorithms. For multiple disease prediction method data mining provide quick prediction result by considering various parameters such as current medical trends, seasonal effect etc.

References

1. Data Mining Concepts and Techniques, Third edition, 3/e Han,et al. Morgan Kaufmann Publishers, An imprint of Elsevier,@2012, by Elsevier Inc.

2. Classification of Diabetes diseases Using Support Vector Machine, V.Anuja Kumari, R.Chithra.,International Journal of Engineering Research and Application (IJERA) Vol.3, Issue 2, March-April 2013,pp.1797-1801.

3. Data Mining Classification Algorithm for Kidney Disease Prediction, Dr.S.Vijayarani, Mr.S.Dhayanand. International Journal on Cybernetics & Information (IJCI) Vol.4, No .4, August 2015.

4. Diagnosis and Prognosis Breast Cancer Using Classification Rules, Miss.Jahanvi Joshi, Mr.RinalDoshiDr.Jigar Patel. International Journal of &Engineering Research and General Science Volume 2, October-November, 2014.

5. A Survey On Disease Diagnosis Algorithms, Aanchal Oswal, Vanchana Shetty,Mustafa Badshah, Rohit Pitre, Manali Vashi, International Journal of Advanced Research in Computer Engineering & Technology(IJARCET) Volume 3 Issue 11, November 2014.

.6. Rajkumar Gaur Grewal Babita Pandey “Two level Diagnosis of Breast cancer using Data mining”, International Journal of computer Applications(0975-8887),Volume 89 –No18,March-2014.

7. Rebeck Carvalho, Rahul Isola, Amiya KumarTripathy “MediQuery-An Automated Decision Support System”, IEEE, ISSN: 1063-7125, Page No-1-6

8. George L. Tsirogiannis, Dimitrios Frossyniotis, Konstantina S. Nikita, and Andreas Stafylopatis“A Meta-classifier Approach for Medical Diagnosis”, G.A. Vouros and T. Panayiotopoulos (Eds.): SETN 2004, LNAI 3025, pp. 154–63, 2004.

9. Classification of Heart Disease using Genetic algorithm and SVM algorithm ,International Journal Of Advanced Research in Computer science and Software Engineering,ISSN:2277 128x.

10. M.A. Hernandez and S.J.Stolfo,” Real-World data is dirty: Data cleansing and the merge/purge problem, “Data Mining Knowl.Discovery, Vol.2, no.1, pp, 9-37, and 1998.

11. Dr.S.Vijayarani and Mr.Dhayanand,”Liver Disease Prediction Using SVM and Bayes Algorithms.