Comparative Study of Data Mining Algorithms in Medical Data

Manjusha K. K
Research Scholar, Dept. of Computer Science
Karpagam University
Coimbatore, India

Sankaranarayanan K
Dean,
Sri Ramakrishna Institute of Technology
Coimbatore, India

Abstract— **Dermatological diseases are the most crucial disease for both men and women. Skin diseases are commonly misdiagnosed with each other. One of the most common forms of medical malpractices globally is an error in diagnosis. In this study, we briefly examine the potential use of classification based data mining techniques such as Rule based, Decision Tree, Naive bayes and Artificial Neural Network to massive volume of healthcare data. Because of the huge amount of data, they are not "mined" to discover hidden information. This is an extension of naive bayes which discover hidden patterns and relationships often go unexploited. Diagnosis of Dermatological Diseases can answer complex "what if" queries which traditional decision support systems cannot. Using generic symptoms such as age, sex, fever, whether, eye symptoms, it can predict the likelihood of patients are used. Aim of the study is to propose a model for early detection and correct diagnosis of the disease which will help the doctor to predict the disease So diagnosis of a single patient can differ significantly if he was examined by different physicians. So we need a standard format for predicting such a disease. Today, automated medical analysis help doctors to diagnose and predict diseases, at a very fast pace. Medical dataset used for this work contain 230 instances with 21 attributes. In this paper we have experimented on data gathered from the southern part of Kerala, India. Weka is used to calculate the accuracy of the dataset we are collected. This paper evaluates three classification algorithms such as Naive Bayes classifier, J48 and Support Vector Machine for the evaluation of some skin patient datasets.**

***Keywords**— Data mining; Dermatology; Naive Bayes; J48 Weka*

I. INTRODUCTION

Data Mining classification techniques are admired in various automatic medical diagnoses tools and it is the process of extracting hidden knowledge from data. Its application includes several fields like banking, insurance and Crime detection including health care. Medical Industry faces many problems due to the increase of types of diseases and their specific management. In addition, the amount of data generated by healthcare transactions is too large, diverse and complex to be analyzed by traditional methods. The application of data mining on medical data can foreground new, useful and potentially lifesaving knowledge. Data mining in medical analysis helps to increase diagnostic accuracy, reduce treatment cost and save human resources [1]. Knowledge discovery in medical databases is a well-defined process and data mining is an essential step. Data mining is, in short, "Knowledge mining from data". Data mining is the process of analyzing data from different views and summarizing it into useful information. Classification algorithms find a set of rules to represent data into classes. It includes two steps; the first step tries to find a model for the class attribute as a function of other variables of the datasets. In the second step the related class of each record is determined by applying formerly designed model on the new and unseen dataset [2]. A popular algorithm based on probability theory is Naive Bayes' algorithms. A predictive model algorithm for classification task is induction of decision trees.

II. BACKGROUND

The large growth of medical databases available in technologically advanced countries has motivated medical researchers in those countries to use data mining for knowledge discovery from these databases. With the steady increase in the volume of stored data, data mining techniques assume an increasingly important role in arriving at patterns and extracting knowledge to provide better patient care and effective diagnostic capabilities. This makes it difficult to analyze the data in order to make important decision regarding patient health. So, it becomes essential to generate a powerful tool for analyzing and extracting important information from this complex data and derive a vital knowledge from it for future reference and research. The analysis of health data can provide a great boost to healthcare by enhancing the performance of patient management tasks. Data mining technologies can provide benefits to healthcare organization for grouping the patients having similar type of diseases or health issues so that healthcare organizations can prescribe the most effective treatments [3]. Data mining applications can be developed to evaluate the effectiveness of medical treatments. By comparing and contrasting causes, symptoms and courses of treatments, data mining can deliver an analysis of the most effective courses of action. For example, the outcomes of patient groups treated with different drug regimens for the same disease or conditions can be compared to determine which treatments work best and are most cost effective. Data mining is vital in the most critical sector of healthcare both in developed and developing countries. Considering the immense benefit it has ushered in, one can envisage the tremendous scope for furthering studies in this field.

A. Decision Tree

Decision trees have been used in analysis of large and complex bulk of data in order to discover useful patterns. The basic decision tree algorithm is called ID3 (Iterative Dichotomizer3). ID3 can handle only discrete values, but the successor C4.5 can handle numeric values. Classification and Regression Trees (CART) approach is suited for analysis of categorical and continuous datasets. J48 is the implementation of ID3 algorithm developed by the WEKA [4]. It can handle different types of data like numeric, nominal, textual data and can also process incorrect or missing values. J48 can be implemented in data mining packages in different platforms and easy to understand because of its presentation. J48 show high performance with small effort.

B. Naïve BayesUnits

Naive Bayes classification is a probabilistic classification based on the Bayes theorem. It shows high speed and accuracy in any dataset.It works on one assumption that is the effect of an attribute values of other attributes. This assumption is caked class condition independence.[5]

The probability of data record X having the class label C_i is

$$\mathbf{P(C_j|X)} = \frac{P(X|C_i)*P(C_i)}{P(X)}$$

The class label C_i with largest conditional probability value determines the category of the data record.

It is very practical when the dimensionality of the inputs is high. The word "Naive" implies the independence between all attributes. Naive Bayes (NB) is a machine-learning method that has been used for over 50 years in biomedical informatics [6]. It requires only small amount of training data to estimate the parameter which is very useful for health care applications [7]. Naive Bayes computes conditional probabilities of the classes given with the instance and select the class with highest posterior [8]. Regardless of this simplified assumption and naive design, naive bayes classifier works well in many complex real world situations. Bayes classification is outperformed by current approaches, like boosted trees or random forests.

C. Dermatological Diseases

This research work was directed towards the prediction and analysis of some commonly seen skin diseases and symptoms focusing on the relationship of symptoms to find important factors and rules that affect skin disorders. The selected conditions have certain common features and this part of the work discusses the eight types of dermatological conditions namely rubella , Kawasaki disease, scarlet fever, fifth disease (erythema infectiosum), no vaccination subitum, (exanthema subitum or roseola infantum), measles, chickenpox and entrovirus.

III. RELATED WORK

Chang et.al [9] conducted five experiments focusing on six major skin diseases and used decision tree of data mining combining with neural network classification methods to construct best predictive model in dermatology. The study predicted and analysed six commonly seen skin diseases namely psoriasis, seborrheic dermatitis, lichens planus, pityriasis rosea, chronic dermatitis and pityriasis rubra pularis. All classification technology could predict the disease with considerable accuracy with neural network model having the highest level of accuracy of 92.62%.

Zeon et. al developed a disease prediction system, DOCAID, for predicting typhoid, malaria, jaundice, tuberculosis and gastroenteritis based on patient symptoms and complaints employing Naive Bayes Classifier algorithm [10]. An accuracy of 91% accuracy in predicting the diseases were reported by the authors.

Theodorali et. al [11] developed prediction model for predicting the final outcome in patients suffering from severe injuries after accident. The analysis included a comparison of data mining techniques using classification, clustering and association algorithms. Using this analysis they obtained results in terms of sensitivity, specificity, positive predictive value and negative predictive value and compared the results between different prediction models.

IV. DATA ANALYSIS SOFTWARE

Weka ("Waikato Environment for Knowledge Analysis") is a popular suite of machine learning software written in Java, developed at University of Waikato, New Zealand [12]. For analysing data we are using Weka 3.7.9. It contains a collection of algorithms for data analysis and predictive modelling, together with GUI for easy access to this functionality. Weka supports several standard data mining tasks like data pre-processing, classification, clustering, association rules, visualization and feature selection. It also offer different test option like Cross validation, using training set, test set, percentage split etc. We have used Naive bayes method, J48 decision tree to perform the mining and classification process.

V. METHODOLOGY

The research work is structured into 3 stages as represented in figure 1. The first stage includes data collection & pre processing and producing training data and analyzing variables. In the second stage we use Weka tool to check the accuracy of the models. The third stage presents explanation of the prediction model.

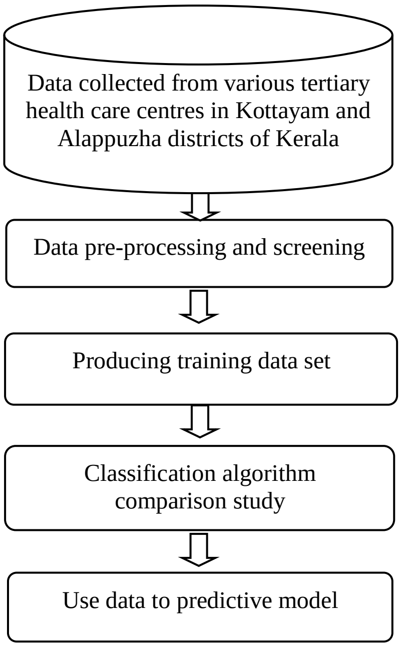
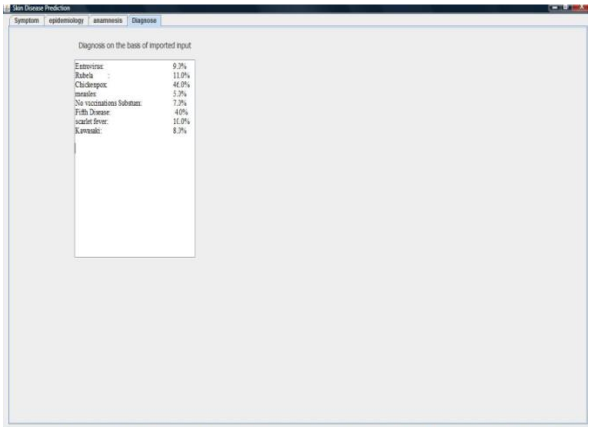

Data collected from various tertiary health care centres in Kottayam and Alappuzha districts of Kerala
Data pre-processing and screening
Producing training data set
Classification algorithm comparison study
Use data to predictive model

Fig. 1. Structure of the research work.

VI. RESULTS AND DISCUSSION

The Graphical User Interface developed in Java Apache-Tomcat-5.5.35 and the results when the diagnosis on the basis of imported input has been shown in the figure 2.

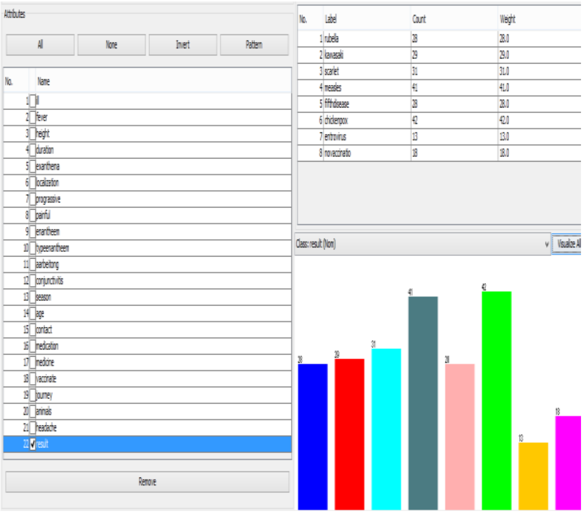


Diagnosis on the basis of imported input

Entrovirus:	9.3%
Rubela :	11.0%
Chickenpox:	46.0%
measles:	5.3%
No vaccinations Substum:	7.3%
Fifth Disease:	4.0%
scarlet fever:	10.0%
Kawasaki:	8.3%

Fig. 2. Graphical user interface

This software gave accurate results, with the consultation of doctors. The doctors can input symptoms into the software and get the most probable disease from the eight diseases. The data was compared using weka software. The results of the experimentation are shown in figure 3.



No.	Name
1	☐ ill
2	☐ fever
3	☐ height
4	☐ duration
5	☐ exanthema
6	☐ localization
7	☐ progressive
8	☐ painful
9	☐ enantheem
10	☐ typeenantheem
11	☐ aarbeitong
12	☐ conjunctivitis
13	☐ season
14	☐ age
15	☐ contact
16	☐ medication
17	☐ medicine
18	☐ vaccinate
19	☐ journey
20	☐ animals
21	☐ headache
22	☑ result

No.	Label	Count	Weight
1	rubella	28	28.0
2	kawasaki	29	29.0
3	scarlet	31	31.0
4	measles	41	41.0
5	fifthdisease	28	28.0
6	chickenpox	42	42.0
7	entrovirus	13	13.0
8	novaccinatio	18	18.0

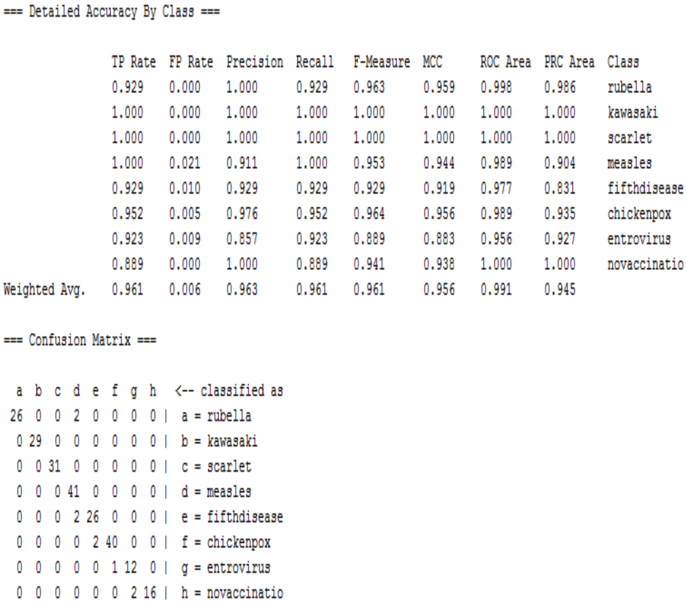
Fig. 3. Experimentation results

The results showed that all the eight diseases depend on all attributes. The attributes fever, season, vaccination, headache and aarbeitong play an important role for the prediction. 197 patients value depends on fever. Vaccination is also an important factor, because most of the disease depends on whether the person is vaccinated or not.

The data set used for the experiment contains 230 instances with 21 attributes and eight class attributes to test and substantiate the difference among classification algorithms. The classification algorithm includes Naive Bayes, J48. After analyzing data with WEKA tool results shows that the highest correctly classified instances is 228(99.13%) by decision tree and naive bayes shows 201(87.4%), correctly classified instances respectively.

The time taken is also an important parameter when we are comparing results. Naive Bayes classification requires only 0.01s and J48 requires 0.03s.

The work was also targeted in comparing the performance of the various algorithms with dataset as given in figure 5. The result showed that Naive bayes produced less precision and true Positive rate as compared with J48 algorithm. J48 is more efficient in all parameter like TP-rate, FP-rate, Precision, Recall and ROC area. Confusion matrix produced by J48 is given in figure 4. Confusion matrix produced by Naive Bayes is given in figure 5.



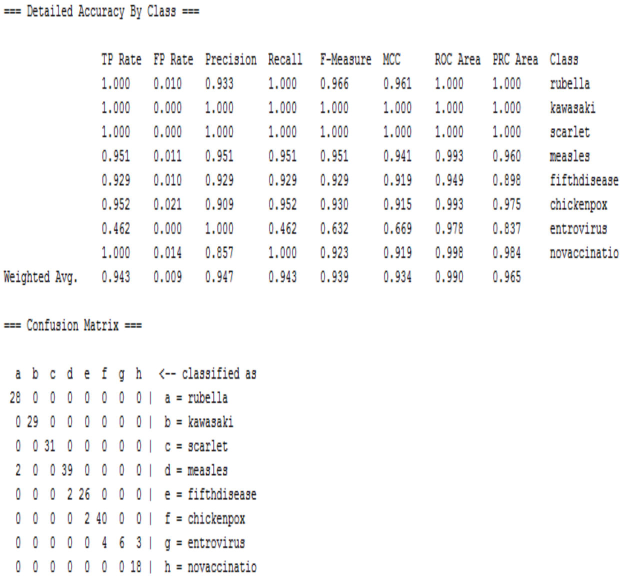
```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.929    0.000    1.000      0.929    0.963      0.959    0.998     0.986     rubella
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     kawasaki
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     scarlet
                 1.000    0.021    0.911      1.000    0.953      0.944    0.989     0.904     measles
                 0.929    0.010    0.929      0.929    0.929      0.919    0.977     0.831     fifthdisease
                 0.952    0.005    0.976      0.952    0.964      0.956    0.989     0.935     chickenpox
                 0.923    0.009    0.857      0.923    0.889      0.883    0.956     0.927     entrovirus
                 0.889    0.000    1.000      0.889    0.941      0.938    1.000     1.000     novaccinatio
Weighted Avg.    0.961    0.006    0.963      0.961    0.961      0.956    0.991     0.945

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h   <-- classified as
 26  0  0  2  0  0  0  0 |  a = rubella
  0 29  0  0  0  0  0  0 |  b = kawasaki
  0  0 31  0  0  0  0  0 |  c = scarlet
  0  0  0 41  0  0  0  0 |  d = measles
  0  0  0  2 26  0  0  0 |  e = fifthdisease
  0  0  0  0  2 40  0  0 |  f = chickenpox
  0  0  0  0  0  1 12  0 |  g = entrovirus
  0  0  0  0  0  0  2 16 |  h = novaccinatio
```

Fig. 4. Confusion Matrix of J48



```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.010    0.933      1.000    0.966      0.961    1.000     1.000     rubella
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     kawasaki
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     scarlet
                 0.951    0.011    0.951      0.951    0.951      0.941    0.993     0.960     measles
                 0.929    0.010    0.929      0.929    0.929      0.919    0.949     0.898     fifthdisease
                 0.952    0.021    0.909      0.952    0.930      0.915    0.993     0.975     chickenpox
                 0.462    0.000    1.000      0.462    0.632      0.669    0.978     0.837     entrovirus
                 1.000    0.014    0.857      1.000    0.923      0.919    0.998     0.984     novaccinatio
Weighted Avg.    0.943    0.009    0.947      0.943    0.939      0.934    0.990     0.965

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h   <-- classified as
 28  0  0  0  0  0  0  0 |  a = rubella
  0 29  0  0  0  0  0  0 |  b = kawasaki
  0  0 31  0  0  0  0  0 |  c = scarlet
  2  0  0 39  0  0  0  0 |  d = measles
  0  0  0  2 26  0  0  0 |  e = fifthdisease
  0  0  0  0  2 40  0  0 |  f = chickenpox
  0  0  0  0  0  4  6  3 |  g = entrovirus
  0  0  0  0  0  0  0 18 |  h = novaccinatio
```

Fig. 5. Confusion Matrix of Naive Bayes

VII. CONCLUSION

A computer aided model has been developed for the analysis of different disease. In this software we are using 21 attributes only, we can extend it with other parameters. This expert system predicts eight different skin diseases and can help doctors to predict the disease confidence. The data for the study was collected from a limited region and future study can be done by collecting data from a wide region which will help to predict the demographic dependence of the disease. The best prediction model was obtained with algorithm developed using J48. Thus, we conclude that this software helps the doctors to clear their confusion when predicting diseases with similar symptoms and helps to take better decision. The expert system also helps to save the time and expense of patients. This can be extended to predict other type of diseases or we can use the same dataset with other data mining techniques. However computer aided diagnosis must be regarded only as one form of supportive measure. Medical practitioner's responsibility and general patient medical care must be given the most priority.

REFERENCES

[1] Yang Guo, Guohua BAi, Yan Hu, "Using bayes network for prediction of type – 2 diabetes", 7th International Conference for Internet Technology and Secured Transactions (ICITST), 2012, London.

[2] Reza Entezarin-Maleki, Arash Rezaei, Behrouz Mimaei-Bidgoli, "*Comparison of classification methods based on the type of attributes and sample size*", Journal of Convergence Information Technology (JCIT), Vol. 4, No. 3, pp.94 – 102, 2009.

[3] Boris Milovic, Milan Milovik, "*Prediction and decision making in heathcare usind data mining*" Kuwait Chapter of Arabian Journal of Business and Management Review, Vol. 1, No. 12, Aug 2012

[4] www.data-mining.business-intelligence. uoc.edu/home/j48-decision-tree.

[5] Prof.M.S.Prasad Babu, Bendi Venkata Ramana, Boddu Raja Sarath Kumar, New Automatic Diagnosis of Liver Status Using Bayesian Classification

[6] Caruana, R. and Niculescu-Mizil, A.: "An empirical comparison of supervised learning algorithms". Proceedings of the 23rd International Conference on Machine Learning, 2006.

[7] www.ic.unicamp.br/~rocha/teaching/2011s2/.../naive-bayes-classifier.pdf

[8] Karpagavalli S, Jamuna K. S, Vijaya M. S, "*Machine learning approach for preoperative anaesthetic risk prediction",* International Journal of Recent Trends in Engineering, Vol. 1, No.2, May 2009.

[9] Chun-Lang, Chih-Hao Chen, "*Applying decision tree and neural network to increase quality of dermatological diagnosis*", Expert System with Applications, Vol. 3, pp. 4035 – 4041, 2009.

[10] Zeon Trevor Fernando, Priyank Trivedi, Abhinandan Patni, Priyal Trivedi, "*DOCAID: Predictive healthcare analytics using naive bayes classsification* ", Second Student Research Symposium (SRS), International Conference on Advances in Computing, Communications and Informatics (ICACCI'13), 22 – 25 August 2013.

[11] Eleni-Maria Theodoraki, Stylianos Katsaragakis, Christos Koukouvinos Christina Parpoula, *"Innovative data mining approaches for outcome prediction of trauma patients",* J. Biomedical Science and Engineering, Vol. 3 pp. 791-798, 2010.

[12] http://www.cs.waikato.ac.nz/ml/weka/index_documentation.html.