MEDICAL DATA MINING : AN EXPERT DIAGNOSTIC METHOD FOR DERMATOLOGICAL DISEASES

Manjusha K.K [1] *Sankaranarayanan . K*[1] *Seena. P*[2]

ABSTRACT

Data mining is becoming fore front in the healthcare industry today, with the key role that it plays in the prediction of diseases based on collated data. Medical diagnosis is an important but complicated task that should be performed with great accuracy and efficiency. Various studies prove that the diagnosis of a single patient can differ significantly if examined by different physicians or by the same physician at different times. Today, automated medical analysis help doctors to diagnose and predict diseases, at a very fast pace. This study addresses dermatological diseases which are largely neglected, but may even prove fatal if left unattended.

Medical dataset used for this work contain 230 instances with 22 attributes. Weka is built in software tool for data mining. Five classification algorithms used are J48 (Decision tree), Naive Bayes' (NB), Multilayer perception (Artificial Neural Network), ZeroR (Rule based) and Multiclass classifier (Support Vector Machine). Prediction of dermatological diseases is very difficult because of the large number of similar disease presentations. In this paper we have experimented on data gathered from the southern part of Kerala, India. The GUI, developed in Java, reveals the chances of different dermatological disease and also finds out the probabilities of occurrence of each disease.

Keywords: Data mining, Dermatology, Diagnosis, Weka.

I. INTRODUCTION

The Health Care 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 is essentially, the process of extracting or mining knowledge from large amounts of data. Data mining is an innovation that can assist physicians in dealing with this large amount of data. Its methods can offer facilities ranging from interpreting complex diagnostic tests to combining information from multiple sources and providing support for differential diagnosis. 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-

[1]Research Scholar, Karpagam University, Coimbatore, India. Tamil Nadu, India.

[1]Dean, Sri Ramakrishna Institute of Technology, Coimbatore, Tamil Nadu, India.

[2]Asst Professor, Dept of Dermatology, Govt Medical College, Kottayam, Kerala, India.

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. Other three algorithms are Artificial Neural Networks, Support Vector Machine and Rules based.

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. Healthcare organizations produce and collect large volumes of information on daily basis, which are often very complex. 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.

Decision tree :

Decision trees are one of the most powerful tool in data mining and knowledge discovery. It has 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.

Naive bayes :

In medical data mining, Naive Bayes classification plays an important role. It is a probabilistic classification based on the Bayes theorem. 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 [5]. It requires only small amount of training data to estimate the parameter which is very useful for health care applications [6]. Naive Bayes computes conditional probabilities of the classes given with the instance and select the class with highest posterior [7]. 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.

Support vector machine :

SVM efficiently perform with both linear and non linear data. Sequential minimal optimization was invented by John Platt in 1998 at Microsoft Research [8]. SMO is an iterative algorithm for solving optimization problem. It normalizes all attributes by default and also replaces all missing values and transforms nominal attributes into binary ones.

ARTIFICIAL NEURAL NETWORKS

Artificial neural network has been used to solve a wide variety of tasks that are difficult to solve using ordinary rules based programming, including computer vision and speech recognition. A multilayer perceptron(MLP) is a feed forward artificial neural network model that maps sets of input data onto a set of appropariate outputs. A MLP consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next node. It utilizes a supervised learning technique called back propagation for training the network [9].

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.

Rubella (German measles) is an infection in which there is a rash on the skin. The main symptom of rubella is a red rash, usually starting on the head and neck. Other possibilities include headache, mild conjunctivitis and runny nose.

Kawasaki disease often begins with a high and persistent fever. Bilateral conjunctival injection was reported by many publications to be the most common symptom after fever. It is an illness that involves the skin, mouth and lymph nodes, and most often affects kids under age 5 and can fully recover within a few days. Untreated, it can lead to serious complications that can affect the heart.

Scarlet fever (Scarlatina) is a disease which causes a distinctive pink-red rash occurs when the bacteria release toxins. Scarlet fever is characterized by sore throat, fever, bright red tongue, red spots on the soft palate, headache etc. It more frequently occurs in the late winter or in early spring.

The name fifth disease (erythema infectiosum) derives from its historical classification as the fifth of the classical childhood skin rashes or exanthems. Their classification is as follows: 1. Measles, 2. Scarlet fever, 3. Rubella, 4. Dukes' disease and 5. Fifth disease. Fifth disease starts with a low-grade fever, headache and cold like symptoms such as a runny or stuffy nose. These symptoms pass, and then a few days later the rash appears. The bright red rash usually begins on the face.

No vaccination subitum (Exanthema subitum or roseola infantum), sixth disease or three-day fever is a disease of children, generally under two years old although it has been known to occur in eighteen-years-olds, whose manifestations are usually limited to a transient rash (exantheem) that occurs following a fever of about three day's duration. Typically the disease affects a child between six months and two years of age, and begins with a sudden high fever. A fever, sometimes as high as 106F, appears suddenly and lasts 3 to 5 days. Other symptoms like irritability, malaise and runny nose may be present at this time.

The classical signs and symptoms of measles (rubeola) include four-day fever and cough, head cold and conjunctivitis along with fever and rashes. The fever may reach up to 104°F. Observation of Koplik's spots is also diagnostic of measles. Small white spots usually develop inside the mouth a day or so. Later measles appear around 10 days after the person become infected. A red and brown spotty rash will develop at the initial stage. Also with red eyes, high temperature (around 104°F), dry cough, Koplik's spots in the mouth and throat.. Risk factors for severe measles include underlying immunodeficiency, Vitamin A deficiency etc.

The initial symptom of chickenpox is also rash of red, itchy spots. Chickenpox (Varicella) is a common illness that causes an itchy rash and will spread quickly and easily from someone who is infected. It also shows red spots or blisters all over the body Season for chickenpox is mainly winter and spring, usually between March and May. It shows blisters on the skin of the infected person. The initial symptom of chickenpox starts with fever up to 102^{0}F, a headache and sore throat. Most people will get chickenpox at some point in their lives if they have not had the chickenpox vaccine. The first symptoms of chickenpox often are a fever up to 102F, a headache and sore throat. The rash usually appears about 1 or 2 days after the first symptoms start. A person with chickenpox is infectious one to two days before the rash appears.

Entrovirus results in respiratory and gastrointestinal symptoms or flu-like symptoms (fever, headache, body ache etc.). Upper respiratory tract symptoms such as runny and stuffy nose, loss of smell or taste, vitamin and mineral deficiencies, etc. are other symptoms.

Related Work

Chang et.al [10] 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 [11]. An accuracy of 91% accuracy in predicting the diseases were reported by the authors.

Theodorali et. al [12] 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.

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 [13]. For analysing data we are using Weka 3.7.9. The Weka 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. The important features of Weka are it is open source and platform independent. It also offer different test option like Cross validation, using training set, test set, percentage split etc. We have used Naive bayes method, SMO, J48 decision tree, Multilayer Perceptron in Neural Network to perform the mining and classification process.

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.

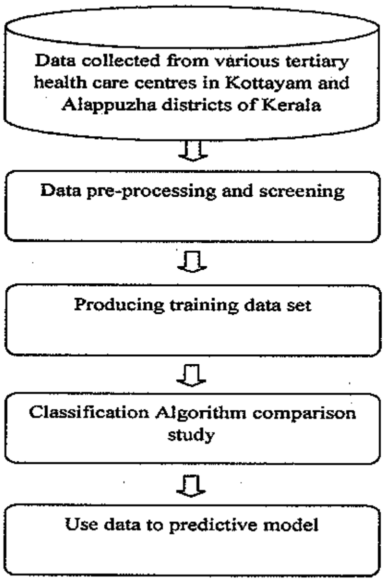
Data collected from various tertiary health care centres in Kottayam and Alappuzha districts of Kerala
Data pre-processing and screening
Producing training data set
Classification Algorithm comparison study
Use data to predictive model

Figure 1: Research structure of the work

Data Collection

Data was collected from various tertiary health care centres in Kottayam and Alappuzha districts of Kerala. The study was developed on the basis of the survey and the questionnaires are prepared and filled out by skin specialist in Kottayam Medical College. After filtering and correcting missing values we got 230 skin disease data and this data is used for prediction of eight skin diseases which show similar symptoms. Among all, 44 pieces of data were chickenpox, 28 rubella, 29 Kawasaki disease, 31 scarlet fever, 41 measles, 27 fifth disease, 12 entrovirus and 18 no vaccination subitum,. In this study 21 factors exist and all are clinical attributes. The attributes are temperature (>90F or <90F), level of fever (90-95, 95-102, >102), duration of fever (number days fever started), type of exanthema (Maculopapules, vesicular, maculopapular rash), progressive exanthema (slowly or quickly), painful exanthema (Y,N), type of enantheem (kopik spot, pharyngitis, blisters, aardbeitong), localization (face, neck, cheek, body), conjunctivitis, seasons (winter, summer, autumn, spring), age (baby, young, elder), prior contact, patient medication (medicine A, medicine B, medicine C), whether patient is vaccinated or not, recent journey etc.

Data Pre-processing

The database of this study contains a total of 230 skin disease cases. After data pre-processing all data was divided into two sections: They were the training data set with 180 and test data set with 50 respectively to verify the accuracy of predictive model. The data is analyzed and implemented in WEKA tool. Data mining finds out the valuable information hidden in huge volumes of data. Weka tool is a collection of machine learning algorithms for data mining techniques, written in Java. It consists of data pre-processing, classification, regression, association rules, clustering and visualization tools. We have used Naive bayes method, SMO, J48 decision tree, Multilayer Perceptron in Neural Network to perform the mining and classification process. We have used 10 folds cross validation to minimize any bias in the process and improve the efficiency of the process.

Results And Disucssion

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.

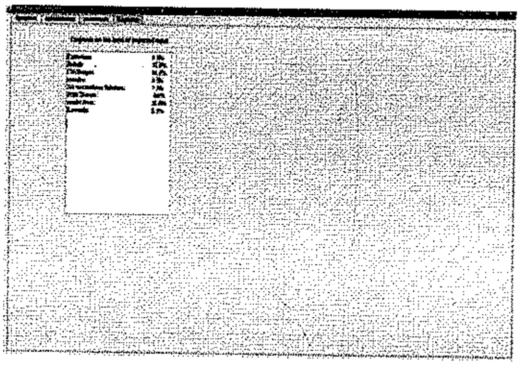

Figure 2 : Graphical User Interface

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.

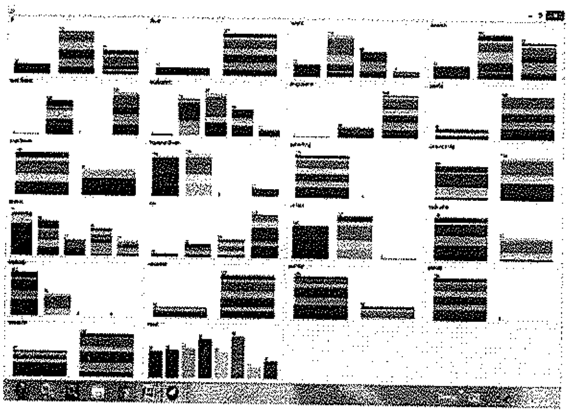

Figure 3: Experimentation results

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, SMO, Multilayer Perception. After analysing data with WEKA tool results shows that the highest correctly classified instances is 228 (99.13%) by support vector machine algorithm. Artificial neural network, decision tree and naive bayes shows 226 (98.26%), 221(96.08%) and 211(91.73%) correctly classified instances respectively (figure 4).

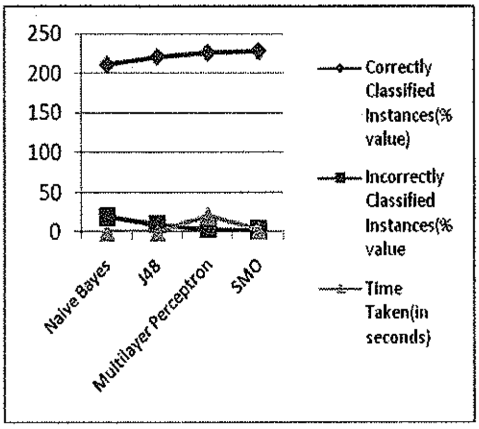

Figure 4: Simulation result of data mining algorithm

The time taken is also an important parameter when we are comparing results. Naive Bayes classification requires only .013s and Artificial Neural Network requires 20.96s as compared with other algorithms.

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 other three algorithms. Support Vector Machine is more efficient in all parameter like TP-rate, FP-rate, Precision, Recall and ROC area. Confusion matrix produced by SMO is given in table 1.

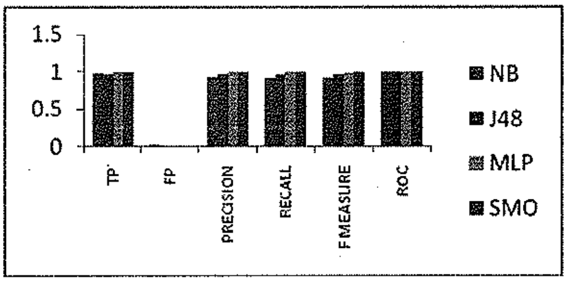

Figure 5: Performance study of data mining algorithms

Table 1 : 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
0	0	0	41	0	0	0	0	d=measles
0	0	0	2	25	0	0	0	e=fifthdisease
0	0	0	0	0	44	0	0	f=chickenpox
0	0	0	0	0	0	12	0	g=entrovirus
0	0	0	0	0	0	0	18	h=novaccinatio

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 support vector machine 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.ic.unicamp.br/naive-bayes-classifier.pdf

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

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

[7] 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.

[8] Platt, John, "*Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines*", CiteSeerX, 1998.

[9] www.en.wikipedia.org/wiki Artificial_neural_network

[10] 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.

[11] 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.

[12] 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.

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

Author's biography

Manjusha K.K, received her BSc. in Mathematics (2004) and Master in Computer Application (2007) from University of Calicut. She is currently working as Asst. Professor in Dept. of Computer Science, Providence Women's College, Calicut and pursuing her Doctoral degree in Computer Science from Karpagam University, Coimbatore, Tamil Nadu. Her research area is data mining of medical data.

K.Sankaranarayanan, presently working as DEAN (PG Studies) at Sri Ramakrishna Institute of Technology Coimbatore, Tamil Nadu, India completed his B.E (Electronics and Communication Engineering) in 1975, and M.E (Applied Electronics) in 1978 from P.S.G.College of Technology, Coimbatore under University of Madras. He did his Ph.D. (Biomedical Digital Signal Processing and medical Expert System) in1996 from P.S.G.College of Technology, Coimbatore under Bharathiar University. He has so far published 58 research papers in National and International Journals and around 60 papers in National and International conferences. His areas of interest include Digital Signal Processing, Computer Networking, Network Security, Biomedical Electronics, Neural Networks and their applications, and Opto Electronics. He has more than 32 years of teaching experience and worked in various Government and self fincing Engineering colleges.