A REVIEW ON PREDICTIVE AND DIAGNOSIS OF DISEASES USING SYMPTOMS

Shital Patil[1], Shubham Hadkar[2], Komal Jadhav[3], Vishakha Pisat[4]
CSE Dept., SRTTC College of Engineering, Savitribai Phule Pune University, Pune, India.

Abstract
Data Mining is one of essential areas of research that is more popular in health organization. Data mining plays a role for uncovering new trends in healthcare organization which helpful for all the patient associated with this field. The healthcare environment is still information rich but knowledge poor. This research investigates parameters like gender and age group that are most likely to be affected by many diseases. There is a lack of effective analysis tools to discover hidden relationship and trends in data. It investigate the sick and healthy factors which contributes to the many diseases in males and females. Data mining approaches are implemented to identify these factors.

Keywords: K- means Clustering, SOM, and Data Mining.

INTRODUCTION

Predictive analytics is the practice of extracting information from existing data sets in order to determine patterns and predict future outcomes and trends. Predictive analytics does not tell us what will happen in the future it will help us in analysis and predicting solutions. The purpose of a prediction algorithm is to forecast future values based on our present records. [1]

Today Data Mining plays a very important role in each and every field. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Today due to increase in population there is also increase in diseases and to predict these diseases and diagnose it is a very much challenging task for the healthcare system. For this it requires a large amount of complex data about various patients their medical history, diseases diagnosis, medical devices and resources. This data is an important resource to be processed to discover the hidden patterns analyse it and make decision.

Data mining activities are divided into three categories:

1. **Discovery:** Includes the process of searching the database to find hidden patterns without a default pre-set.
2. **Predictive Modelling:** Includes the process of discovering patterns in databases and use them to predict the future.
3. **Forensic Analysis:** Includes the process of applying extracted patterns to fined unusual elements.

[2]

Review of Literature

Vikas Chaurasia et al. [3] the objective of this research work is to predict more accurately the presence of heart disease with reduced number of attributes. Originally, thirteen attributes were involved in predicting the heart disease. Thirteen attributes are reduced to 11 attributes. Three classifiers like Naive Bayes, J48 Decision Tree and Bagging algorithm are used to predict the diagnosis of patients with the same accuracy as obtained before the reduction of number of attributes. This study will also work to identify those patients which needed special attention for treatment.

M. Akhil jabbar et al. [4] proposed an efficient associative classification algorithm using genetic approach for heart disease prediction. The main motivation for using genetic algorithm in the discovery of high level

prediction rule is that discovered rules are highly comprehensible, having predictive accuracy and of high interestingness values. Experimental results show that most of the classifier rules help in the best prediction of heart disease which even help doctor in their diagnose decision. It has certain limitation that it uses large number to predict the heart disease using data mining approach. We can reduce the number of attributes to make it less complex and better.

Mrs.G.Subbalakshmi,[5] Mr K. Ramesh, Mr M. Chinna Rao has developed a Decision Support in Heart Disease Prediction System (DSHDPS) using data mining modelling technique, namely, Naïve Bayes to discover the relationship between variables in data in healthcare industry. It is implemented as web based questionnaire application. It can serve a training tool to train nurses and medical students to diagnose patients with heart disease.

Niti Guru et al proposed the prediction of various disease like Sugar, Heart disease, Blood Pressure with the use of neural networks. The Neural Network is tested and trained with 13 input variables such as smoke, Age, obesity, Blood Pressure, Angiography's report and the like. Training was carried out with the aid of back-propagation algorithm. The system identified the unknown data from comparisons with the trained data, whenever unknown data was fed by the doctors and generated a list of probable diseases that the patient is vulnerable to [6].

Shantakumar B.Patil, Y.S.Kumaraswamy [7] proposed a methodology for the extraction of significant patterns from the heart disease warehouses for heart attack prediction has been presented. At start, the data warehouse is pre-processed so that it can be made suitable for the mining process. After the pre-processing gets over, the heart disease data warehouse is clustered with the aid of the Kmeans clustering algorithm, which will extract the data applicable to heart attack from the warehouse. As a result the frequent patterns applicable to heart disease are mined with the aid of the MAFIA (Maximal Frequent Item set Algorithm) algorithm from the data extracted.

Related work

In this paper the previous related work are reviewed. In this paper lists of symptoms are required of every diseases to be predict. Some of the system exist proposed only relevant diseases of heart cancer but these is inaccurate for the person not affected from any diseases listed whereas in this system a list of major as well as minor diseases can be detected through the system and predict the diseases through the system. This system may not give total information about diagnosis that is not relevant to the patient's family records.

Iliad Medical Diagnostic Support Program is an expert diagnostic system which is used to explain the relationships for finding the diseases. This system uses the Bayesian classification to compute the probability for possible diagnosis. In this paper we are using SOM also known as Self Organizing Map. The Self-Organizing Map is one of the most popular neural network models. Self-organizing maps are different from other artificial neural networks as they apply competitive learning as opposed to error-correction learning (such as backpropogation with gradient descent), and in the sense that they use a neighbourhood function to preserve the topological properties of the input space. The SelfOrganizing Map is based on unsupervised learning, which means that no human intervention is needed during the learning and that little needs to be known about the characteristics of the input data. We could, for example, use the SOM for clustering data without knowing the class memberships of the input data. The SOM can be used to detect features inherent to the problem.

The Self-Organizing Map is a two-dimensional array of neurons:

$$M = \{m1,2, \dots , m(pxq)\}$$

One vector used for this system is a neuron also called as codebook vector. This vector has the same dimensions as the input vectors i.e. n - dimensional. Each neuron of the vector is connected to each other in an adjacent way having relation among them form a topology or a structure. In this algorithm a number of neurons and the relation among the topologies are fixed from the beginning. The number of neurons determines the scale or the granularity of the resulting model. The selection of scale affects the accuracy and the generalization capability of the model.

Analysis of Symptoms and diseases

Sample Dataset of some diseases with their symptoms and treatment Table 1

COMMUNICABLE DISEASES

Name	Causative Agent	Symptom	Prevention	Treatment
A.VIRAL DISEASES. Common Cold	Rhinovirus	Drainage Scratchy throat Sneezing Fever Headache	Use of Hand sanitizer's disinfectants and facial tissues.	Decongestants for nasal symptoms, Acetaminophen and ibuprofen for headache and fever.
Rabies(Hydrophobia)	Street Virus	Severe Headache High Fever choking feels thirsty but has fear of water	Complete Immunization of pet dog, cat. Isolation of rabied dogs.	Pasteur treatment
H1N1(Swine Flu)	H1N1 Virus	Chills Cough Fatigue Headache loss of Appetite	**Medications:** Amantadine Zanamivir	H1N1 vaccine
B.BACTERIAL DISEASES. Tuberculosis	Mycobacterium Tuberculosis	Fever, General weakness Loss of appetite Yellowish blood stained sputum	Isolation of TB patients Vaccination by BCG Diagnosed By: Mantoux reaction	Streptomysin, para-amino salicylic acid, rifampicin etc.
Typhoid	Salmonella typhi	Damage to the intestinal wall, Abdominal tenderness.	Personal Cleanliness, Protection of food and water dust and flies.	Ampicillin and chloromycetin
Tetanus(lock jaw)	Clostridium tetani	Painful muscle contraction of neck and jaw. Increased	Active immunization in case of cut/injury. The cut surface should	Human tetanus immunoglobulin neutralizes the toxin Vaccine - DPT

		rigidity of jaws-lock jaw	be kept properly covered and bandaged to avoid contamination	
C. PROTOZOANS **Malaria**	Plasmodium	Nausea High fever with chills Cycle of fever and sweating repeated after 2-3 days	Fill up ditches, ponds and pools with earth to prevent breeding of mosquitoes Cover drains or make underground drain	Quinine, mepacrine, chlor oquinone etc. SF- 66 vaccine.
C.HELIMINTHIS **Ascariasis**	Ascaris lubricoides	Indigestion Appendicitis Acute colic pain	Avoid use of raw vegetables and contaminated water Proper sanitary disposal	Adult worm can be removed by mixture of oil of chenopodiumand tetrac hlorethylene.
FILARIASIS	Wuchereria bancrofti	Periodic attacks of fever	Destruction of mosquitoes Protection against mosquito bites	Diethyl carbamazine, Para melaminyl phenylstibonate

NON COMMUNICABLE DISEASES

Name	Causes	Symptom	Prevention	Treatment
Diabetes	Disturbance in the secretion of insulin from pancreas	Frequent urination Loss of weight Tired vision	Periodic check ups Avoid taking starchy and sugary foods	Sulphonylureas, glipzide, etc.
Heart Attack	Atheroma Hypertension	Severe pain in chest	No smoking Avoid rich animal fat avoid stress	Regular exercise is necessary

THE K-MEAN CLUSTERING ALGORITHM

The K-MEAN CLUSTERING is a method of vector quantization originally developed from the signal processing that is popular for cluster analysis in data mining. K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. K-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori. The main idea is to define k centres, one for each cluster. These centres should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centre. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as barycentre of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centre. A loop has been generated. As a result of this loop we may notice that the k centres change their location step by step until no more changes are done or in other words centres do not move any more. Finally, this algorithm aims at minimizing an objective function know as squared error function given by:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{ci} (\|xi - vj\|)^2$$

Where,

'$\|x_i - v_j\|$' is the Euclidean distance between x_i and v_j.

'c_i' is the number of data points in i^{th} cluster.

'c' is the number of cluster centres.

Conclusion

This paper uses K-Means algorithm and the SOM techniques to determine and predict the diseases using the symptoms. By using these techniques, it improves the overall speed and increase the accuracy of algorithm. This system also reduces the chances of getting misleading of any terms and regards in accurate details of diseases accurately. This system also helps students and other people for easily identifying the diseases which will also saves many lives and give faster and better result.

References

[1] E.W.T. Ngai, Li Xiu, D.C.K. Chau." Application of data mining techniques in customer relationship management: A literature review and classification" Expert Systems with Applications 36 (2009) 25922602.

[2] Golriz Amooee," A Comparison Between Data Mining Prediction Algorithms for Fault Detection
(Case study: Ahanpishegan co.)", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 6, No 3, November 2011.

[3] Vikas Chaurasia, Saurabh Pal," Data Mining Approach to Detect Heart Dieses International Journal of Advanced Computer Science and Information Technology (IJACSIT)", Vol. 2, No. 4, 2013, Page: 56-66, ISSN: 2296-1739 © Helvetic Editions LTD, Switzerland www.elvedit.com.

[4] M.Akhil jabbar, Dr.Priti Chandra, Dr.B.L Deekshatulu " Heart Disease Prediction System using Associative Classification and Genetic Algorithm". ICECIT, 2012.

[5] Mrs.G.Subbalakshmi, "Decision Support in Heart Disease Prediction System using Naive Bayes" ISSN: 0976-5166 Vol. 2 No. 2 Apr-May 2011.

[6] Niti Guru, Anil Dahiya, Navin Rajpal, "Decision Support System for Heart Disease Diagnosis Using Neural Network", Delhi Business Review, Vol. 8, No. 1 (January - June 2007).

[7] Shantakumar B.Patil, Y.S.Kumaraswamy," Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network", European Journal of Scientific Research ISSN 1450-216X Vol.31 No.4 (2009), pp.642-656.

[8] H.R.Warner and O.Bouhaddou," Innovation Review; Iliad A Medical Diagnostic Support Program. Top health Inf.Manage", Vol.14, No.4, 1994.