Classification Rules by Decision Tree for Disease Prediction

Smitha.T
PhD-Research Scholar,
Karpagam University,
Coimbatore
(Asst.ProfessorMCA,Dept,
SNGIST, N.Paravoor, Kerala)

V.Sundaram
Director-MCA,
Karpagam Engineering college
Coimbatore

ABSTRACT
This research paper mainly focuses on a data mining technique, that had an objective of creating a prediction model, using decision tree for predicting the chances of occurances of diseases in an area, particularly slum. This model also identifies different significant parameters which can be used to help for the creation of model.Decision tree is one of the learning algorithm which possess certain advantages which make it suitable for discovering rules for data mining application. In this paper the decision tree has been applied to classify the inhabitants in an area based on the chances of hitting a disease. This paper intended to discover the rules for the disease hit using decision tree algorithm. The paper also explores what rule can act in this area for the future prediction.

Keywords
Clustering , Data Mining, Decision Tree, Prediction.

1. INTRODUCTION
We are in an age often referred to as the information age. In this information age,because we believe that information leads to power and success, and thanks to sophisticated technologies such as computers, satellites, etc., we have been collecting tremendous amounts of information. Initially, with the advent of computers and means for mass digital storage, we started collecting and storing all sorts of data, counting on the power of computers to help sort through this amalgam of information. Unfortunately these massive collections of data stored on disparate structures very rapidly became overwhelming. This initial chaos has led to the creation of structured databases and
database management systems (DBMS). The efficient database management systems have been very important assets for management of a large corpus of data and especially for effective and efficient retrieval of particular information from a large collection whenever needed. The proliferation of database management systems has also contributed to recent massive gathering of all sorts of information. Today, we have far more information than we can handle: from business transactions and scientific data, to satellite pictures, text reports and military intelligence.
Data mining is a system retrieval is simply not enough anymore for decision-making.But confronted with huge collections of data, we have now created new needs to help us make better choices. These needs are automatic summarization of data, extraction of the "essence" of information stored, and the discovery of patterns in raw data.

The goal of data mining is to extract important information from data that was not previously known.[1] Data mining is a technique that has a large number of applications in a wide variety of different fields to recognize certain patterns or trends. Once we've analyzed the information, we can make conclusions and decisions which are based on logic. Once we are able to predict the behavior of something which we are analyzing, we will be able to make strategic decisions that can allow you to achieve certain goals.

Exploration is the first stage. The goal of the exploration stage is to find important parameters and determine their nature. If we have a large number of parameters to consider, we may need to reduce them to a range that is easy to deal with. Second stage is pattern identification, where we are selecting the best one that will allow us to make the best predictions. The third stage is called deployment. We will not want to move to this stage until we have found a consistent pattern from stage 2 that is highly predictive. The method that is used with data mining to make predictions is called modeling. The purpose of data mining is to take the model and place it in a situation where the answer is unknown.[3]

2. OBJECTIVES

1. Analyse the database for the creation of an unsupervised model to identify the most significant parameters of affected area. And to predict the chances of hitting the disease using the supervised classifier model.

2. To build a model for classifying the inhabitants based on disease hit.

2.1 The role of data mining in prediction process

Data mining is the process of analyzing data from different perspectives and summarizing it into a useful information that can be used to predict the trend analysis. Prediction is a task of learning a pattern from examples and using the developed model to predict future values of the target variable.

One of the effective way to create and use a data mining model is to get the user to actually understand what is going on so that an immediate action can take directly.. There are many tools for analyzing the dataIt allows the users to analyze data from different dimensions, categorize it and summarize the identified relationships into different formats and finding correlations or patterns among different fields in large relational databases.

There are different types of analysis. They are data visualization, Rule induction, nearest neighbor method, clustering, generic algorithm, decision tree model etc. T the main advantage of data mining is the ability to turn feeling into facts. Data mining can be used to support or refuse the feelings of people. It can be used to add credibility to the feelings . Data mining can discover unexpected patterns that were not under consideration when the mining process started. Some advantages of data mining algorithm in prediction of contagious diseases are they were able to select the correct parameters, helpful to analyze the facts and reasons behind the hit, to make comparative reports over standard norms , to know the information flow etc.[4].

3. LITERATURE REVIEW

In the selected areas in Cochin, especially in slums, for the last fifteen years , after the rainy season some epidemics are hitting.. The unimprovement of living standards mainly in slums is the major reason for the epidemics . Studies reveals that less hygienist and poverty is the main classes of victims of epidemics. There are different types of epidemics due to different reasons. The main factors which influencing an epidemic are Poor hyginity, Rapid climate change ,Drinking water contamination Unplanned sewage disposal system etc. There are many water borne diseases such as cholera, typhoid, diarria etc. .

After every rainy reasons, through media we can understand that hundreds of people were admits in the hospital due to some contagious disease and some of them were even loss their life. We can see that in this area there are many types of epidemics effecting in different season. The causess of each disease may vary but we can surely conclude that the contamination of drinking water is the first and main amongst the various reasons such as temperature extremities occurring due to the fast and non uniform change in climatic whether condition. Poor health of inhabitants in the slums resulted by poverty and old age, Living conditions like personal hygenity, hygenity of the society, lack of health awareness or insufficient knowledge in observing hygenity**,** percapita income of the inhabitance etc are also some of the reasons behind this calamity. Water-borne disease is, simply, any illness resulting from ingestion of or contact with water. Like food-borne diseases, water-ingestion illnesses are either infections or intoxications. Organisms responsible for infections are mianly bacteria. These organisms usually occur in water contaminated with sewage (eg, especially bird and mammal excrement) or by infected persons or animals. Intoxications may be chemical in nature (eg, copper, lead, insecticide poisonings) and usually occur as a result of metal leaching into water (from pipes or containers) and through the accidental spillage or seepage of chemicals into water supplies. They can also occur through toxins produced by blue-green algae (cyanobacteria), eg, *Anabaena*, *Microcystis* or *Oscillatoria*. These organisms have caused even deaths in southern part of India through drinking pond water,. Illnesses acquired through contact with water are caused by bacteria..

3.1Data selection and analysis

To find the prediction of contagious disease hit in a slum, different types of data were collected from different sources. The slum selected for this research study is "24 murry colony" which is situated at Kochi, Kerala.

There are 24 families with 96 inhabitants are living here. After each rainy season, some contagious disease is hitting in almost all families in this slum. This is happening for the last several years. So all the data have collected from each families about each members. The main parameters were education, income, hereditary factors, area located as slum, drainage facilities, drinking water facilities, toilet facility , waste disposal, electricity, approaches to hospital, roads, educational institutions, livelihood etc and created a database. Based on different parameters classified the inhabitants into different levels and created suitable model based on the selected algorithm.

3.2 Data collection

The row data used in the research were collected from health department, Hospital, Urban Local Body, inhabitants from slum, Doctors from various hospital, health officers, different records from urban local body, on site observation etc. We have collected the same type of data from inhabitants inside the slum as well as outside the slum .

To ensure the consistency of result, missing values were also dealt with. Irreverent records and duplicated data were eliminated to reduce the size of data set. Data synchronization was also carried out.

4. CHOOSING FUNCTIONS OF DATA MINING AND MINING ALGORITHM

The segmentation of inhabitants according to the hit of disease may be viewed as a clustering problem.For applying the cluster technique, the data set was further reduced to include only one colony with hereditary disease history. This is to identify the people who can hit the disease and finally become inconsistent.

Ninety six inhabitants come under this category. Thus from the above data set , the unsupervised model was built only with the records of inhabitants who tends to become insolvent. The unsupervised model was built with k-means clustering algorithm. It aims to break the collected data into separate "clusters" grouped by like characteristics.

It groups data using a top-down approach as it starts with a predefined number of clusters and assigns observation to them. This method is relatively efficient in processing large data set. The problem of predicting inhabitance insolvency may be viewed as classification problem. The distribution of inhabitants are uneven.(90% solvent and 10% insolvent). So with these characteristics the problem is difficult to solve. So a new data set had to be created specifically for data mining function**.**

For the new set, calculated the hirt ratio and eliminated the records whose hit ratio is less. By applying classification technique, reduced the data set and calculated the percentage of insolvency. We can see that insolvency is higher with less population.

Used ID3 algorithm to make decision tree by employing a top-down greedy search through the given sets to test each attribute at every tree node and for classifying solvent and insolvent inhabitants.

Some Attributes With Their Possible Values TABLE-1

NO	ATTRIBUTE	POSSIBLE VALUES
1	INCOME	LOW/AVERAGE/HIGH
	EMPLOYMENT	EMPLOYEED/UNEMPLOYEED
3	ENVIRONMENTAL CONDITION	GOOD/AVG/POOR
4	NO.OF MEMBERS IN AFAMILY	BELOW 5/BETWEEN 6-10/ABOVE 10
5	SANITATION FACILITY	GOOD/FAIR/POOR
6	EDUCATION	EDUCATED/NONEDUCATED

5. PATTERN EVALUATION

The clustering analysis is identified in three different clusters according to the difference in inhabitance disease hit.
Cluster1- High hit ratio
Cluster2- Medium hit ratio
Cluster 3- Zero Hit ratio
From the classification rules[10], the most significant non climatic risk factors were identified as less educated, poor hygenity, less sanitation, population immunity and control activities. The identified climatic risk factors were seasonal climate, rainfall, temperature variation , spread of deadly diseases, water surface temperature , prediction interval etc.
The second objective was to build a classification model for solvent and insolvent inhabitants using supervised learning with the help of variables.

5.1 Reasons to become insolvent

1. Disease due to climatic risk factors
2. Disease caused due to non climatic risk factors.

We can make a decision on the inhabitants according to the way in which they become insolvent.[9]
The third objective was to use the classifier model built to be used for predicting inhabitants insolvency. Predictive accuracy of the model can be calculated as the percentage of test samples that are correctly classified.(95% have been correctly classified).
Thus the model which used classification rules [5] is an effective method for grouping solvent and insolvent inhabitant in this context.

6. CONCLUSION

This research study involved a real life application problem. Two kinds of models are developed.

1. An unsupervised clustering model for identifying the significant characterstics of insolvent customers
2. A supervised classification model for insolvency prediction.

The clustering model allowed us to understand different group behavior for history of disease hit and accordingly take action. The knowledge extracted from the clustering model helped to identify the significant characteristics of insolvent inhabitants which formed a particular cluster.
The supervised classification model was built on a data set. This model allowed predicting the insolvency of inhabitants well in advance so that the action measures can be taken against the insolvent inhabitants.[4]
95% of the prediction accuracy was achieved employing the decision tree classification model in the research.[5]. Overall performance is also good.
This model also identified two types of patients-inhabitants become patient(insolvent)due to the climatic risk factors such as seasonal climate, rainfall data, spread of deadly diseases, water surface temperature, temperature and perception measurement etc and inhabitants who became patients those due to non climatic risk factors such as population immunity and control activities, vector abundance, family history etc. The prediction interval is also a factor for the analysis.

7. REFERENCES

[1] Jaiwei Han;Micheline kamber;Data mining concepts and Techniques;Morgan Kaufmann Publishers.

[2] Fayyad U.M.Piatetsky-Shapiro.G & smith.P" From data mining to knowledge discovery in databases'AI magazine 17(3) pp-37-54.

[3] Ms.Sunu Mary Abraham"User Behaviour BasedClustering and Decision Tree Model for predicting customer insolvency in Telecommunication Business.Karpagam Journal-Jan-2011, Volume 5

[4]K.S.Adekeye and M.A.Lamidi, "Prediction Intervals: A tool for monitoring outbreak of diseases" Intternational journal for data Analysis and information System jan-2011-Vol-3.

[5] Aitchison.J and Dunsmore, Statistical Prediction Analysis: Cambridge University Press.

[6] Waleed Alsabhan and Oualid Ben Ali " A new multimodal approach using data mining: the case of jobseekers in the USA" Intternational journal for data Analysis and information System jan-2011-Vol-3.

[7] Rui Xu , Donald C.and WunschClusteriv ng, Iee Press-2008.

[8] Bori Mirkin(2005) clustering for Data mining Chapman & Hall/Crc.

[9]Apte, C.and Weiss,S.M(1997), " Data mining with Decision Trees and Decision Rules" Future generation computer systems, 13,197-210.

[10] Ch.Ding, X.He"K means clustering via principal component Analysis Proc.of international conference on machine learning(2004),pp.225-232,2004.