ISSN 2250-3153

Knowledge Discovery from Real Time Database using Data Mining Technique

Smitha.T*, Dr.V.Sundaram**

PhD-Research Scholar, Karpagam University, Coimbatore*
Director-MCA, Karpagam College of Engineering**

Abstract- The major aspire of this paper is to make a study and to suggest remedial measures for disease management in certain area with the help of data mining technique-clustering. Using this method, the user can identify the disease incidence and reasons with specific parameters and can obtain different solutions for its control. This paper intends to discover the data mining algorithm in the prediction of contagious disease in an area.

Index Terms- Classification rule, Cluster analysis, Data mining, Knowledge discovery.

I. INTRODUCTION

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information that can be used to predict trend analysis. Data mining can discover unexpected patterns that were not under consideration when the mining process started Prediction is a task of learning a pattern from examples and using the developed model to predict future values of the target variable.[1]

Many application can benefit the by the use of information and knowledge extracted from a large amount of data. Knowledge discovery can be viewed as the process of nontrivial extraction of information from a large database, information which is implicitly presented in the data, previously unknown and potentially useful for the users.

One of the effective ways 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 data.

It allows the users to analyze data from different dimensions, categorize it and summarize the identified relationships into different formats and finding correlations among different fields in large relational databases.

II. OBJECTIVES

The purpose of the study is to develop a data mining solution which make diagnosis of the disease as accurate as possible and helps deciding if it is reasonable to start the treatment on suspected patients without waiting for the exact medical test result or not.

The study also concentrates to identify different parameters to analyze the facts and reasons behind the disease.

III. APPLICATION OF DATA MINING IN TREND ANALYSIS

In data mining the extracted data must be transform and load into the data warehouse system. Then store and manage the data into a multidimensional database system. By providing the data access to an analyst the data can be viewed in a presentable format.

The different types of analysis include data visualization, Rule induction, nearest neighbor method, clustering, generic algorithm, decision tree model etc.[2] 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.

A. *Advantages of data mining algorithm in prediction of contagious diseases*

- Helpful in taking quick decision regarding the chances of hitting
- Systematic and smooth flow of information in functional area.
- Able to select the correct parameters
- Analyzing the facts and reasons behind the disease.
- Quick compilation and analysis of large volume of unstructured data from various sources helps to take timely decision making to better control over the system.
- Comparative reports over standard norms.

IV. PRELIMINARY STUDY

We can see the after the rainy season especially in the month of June to August, some type of epidemics were hitting. The un-improvement of living standards and un-hygienist were is the major reason for the epidemics. Studies reveals that less hygienist and poor living environmental condition are 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 hygienist, Rapid climate change, Drinking water contamination Unplanned sewage disposal system etc. There are many water borne diseases such as cholera, typhoid, diarria etc. .

Through media we can understand that hundreds of people were admitted in the hospitals due to some contagious disease and some of them were even losses their life. There are many types of epidemics affecting in different season. The causes of

each disease may vary, but we can surely conclude that the contamination of drinking water is the first and main among the various reasons such as temperature extremities, climatic difference etc. Poverty and old age, personal unhygenity, unhygenity of the society, lack of health awareness or insufficient knowledge, 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. water-ingestion illnesses are either infections or intoxications. Organisms responsible for infections are mainly bacteria. These organisms usually occur in water contaminated with sewage (e.g., especially bird and mammal excrement) or by infected persons or animals. Intoxications may be chemical in nature (e.g., 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 (cyan bacteria), e.g., *Anabaena*, *Microcystis* or *Oscillatoria*. These organisms have caused even deaths through drinking pond water; Illnesses acquired through contact with water are caused by bacteria.

A. Data Collection and Analysis

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

The sample data were collected with interview with 24 families with 96 inhabitants. After each rainy season, some contagious disease is hitting in almost all families in this area. This is happening for the last several years. So all the data have collected from each family about each member. 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.

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.

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.

B. Building of Data Mining Algorithm

The entire inhabitants in this area are divided into different clusters based on different parameters. The cluster technique, to apply 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 due to that particular parameter, i.e. disease history.[4]

45 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, means chance to become patient. The unsupervised model was built with k-means clustering algorithm. It aims to break the collected data into separate "clusters" grouped by like characteristics... [10]

Clustering is the process of grouping a set of physical or abstract objects into classes of similar objects .Cluster- collection of data objects that are similar to one another within the same cluster & are dissimilar to objects in other clusters. I am also known as data segmentation. [9]

There are different clustering approaches such as Partitioning algorithms: which construct various partitions and then evaluate them by some criterion, hierarchy algorithms that create a hierarchical decomposition of the set of data (or objects) using some criterion, density-based algorithm, based on connectivity and density functions, grid-based algorithm, based on a multiple-level granularity structure ,model-based algorithm in which a model is hypothesized for each of the clusters and the idea is to find the best fit of that model to each other .

It is selected partitioning algorithm and this method is relatively efficient in processing in large data set. The problem of predicting inhabitance insolvency may be viewed as classification problem. The distribution of inhabitants is uneven. (80% solvent and 20% 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.[3]

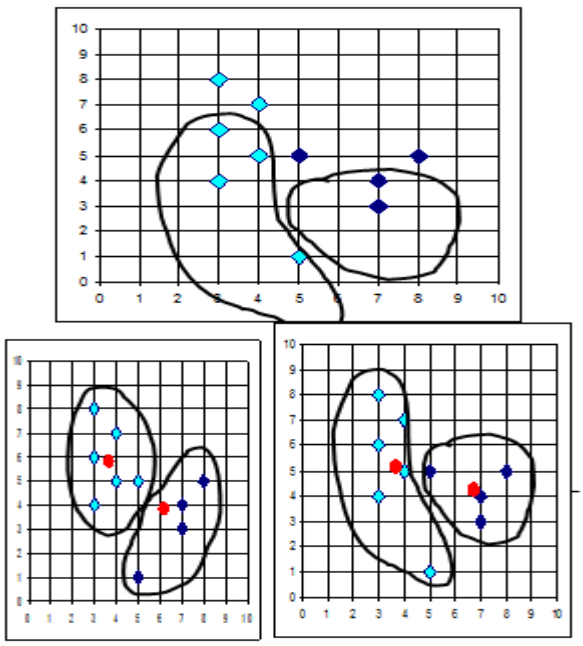
For the new set, eliminated the clusters whose hit ratio is less. By applying classification technique, reduce s the data set and calculate the percentage of insolvency. We can see that insolvency is higher with less population.[8]

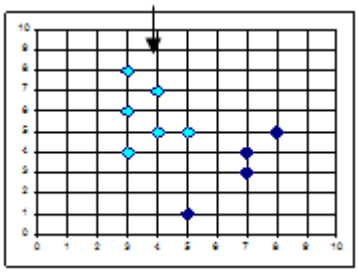
V. PATTERN EVALUATION

Three different clusters were identified based on the disease hit .Cluster A with no disease history, Cluster B with less disease history and Cluster C with high disease history. The inhabitants in this cluster have more tendency to become insolvent.

Different climatic and non-climatic parameters were also identified as less educated, poor hygienist, less sanitation, population immunity seasonal climate, sudden rainfall, temperature variation, spread of deadly diseases, water surface temperature, prediction interval etc.

THE K-MEANS CLUSTERING METHOD





A classification model was also made for and insolvent inhabitants using supervised learning with the help of variables. The reasons to become insolvent was also identified.

A classification rule was made on the inhabitants according to the way in which they become insolvent. The classifier model is used for predicting inhabitant's 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 clustering model is an effective method for clustering solvent and insolvent inhabitant in this context.

VI. CONCLUSION

This research study involved a real life application problem. Two kinds of models are developed .An unsupervised clustering model for identifying the significant characteristics of insolvent customers and 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.

95% of the prediction accuracy was achieved employing the decision tree classification model in the research. Overall performance is also good.[5]

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.

VII. FUTURE ENHANCEMENTS

The same database can be applied with different other data mining techniques and can compare the result for better performance.

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 Based Clustering 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" International 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" International journal for data Analysis and information System jan-2011-Vol-3.

[7] Rui Xu , Donald C.and WunschClustering, 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.

AUTHORS

First Author – Smitha.T. She has acquired her Post Graduate Degree in Computer Application and M.Phil in Computer science from M.K.University.Now doing PhD at Karpagam University under Dr.V.Sundaram. .She has 9 years of teaching experience and 4 years of industrial experience. She has presented many papers, regarding data mining, in national as well as international conferences.Now working as an Asst.Professor–MCA department of Sree Narayana Guru Institute of Science and Technology, N.Paravoor, Kerala. Her area of interest is Data mining and Data Warehousing.
Email id - smithahari2005@gmail.com

Second Author – Dr.V.Sundaram. He is a postgraduate in Mathematics with PhD in applied mathematics.He has 45 years of teaching experience in India and abroad and guiding more than 10 scholars in PhD and M.phil at Karpagam and Anna University. He has organized and presented more than 40 papers in national as well as international conferences and have many publications in international and national journals. He is now working as the Director of MCA Department of Karpagam Engineering College. He is a life member in many associations. His area of specialization includes fluid Mechanics, Applied Mathematics, Theoretical Computer Science, Data mining, and Networking etc.
Email id - Dr_vsundaram@gmail.com