CASE STUDY ON HIGH DIMENSIONAL DATA ANALYSIS USING DECISION TREE MODEL

*SMITHA.T, #DR.V.SUNDARAM

*(MCA, M.Phil)-PhD-Research Scholar,
Karpagam University, Coimbatore;(Asst.Professor-MCA Dept, SNGIST, N.Paravoor, Kerala)

#Director-MCA, Karpagam College of Engineering,
Karpagam University, Coimbatore

ABSTRACT

The major aspire of this paper is to build a model to predict the chances of occurances of disease in an area. This paper mainly concentrating the data mining technique-Decision tree model to identify the significant parameters for prediction process. The decision tree model created with the help of ID3 algorithm.

Keywords: classification, decision tree, data mining, prediction.

1. INTRODUCTION.

The development of information technology has made a global change in the universe. IT applications are today's indispensable tool for any future prediction. We can say that the prediction of a future event by the evaluation of a current event can be successfully done with the help of application of information Technology , especially data mining tools.[1] The use of IT application like data mining software, providing a trend analysis. Introducing data mining Tools and Techniques into prediction of a disease in a slum is surely helpful to find an appropriate solution to eradicate the hazardous situation from the area.

Data mining is a system of searching through large amounts of data for patterns. It is a relatively new concept which is directly related to computer science. It can be used with a number of computer techniques such as pattern recognition and statistics .The goal of data mining is to extract important information from data that was not previously known. 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 you've analyzed the information, you can make conclusions and decisions which are based on logic. Once you are able to predict the behavior of something you are analyzing, you will be able to make strategic decisions that can allow you to achieve certain goals.[2]

Exploration is the first stage, where you are exploring and preparing data. The goal of the

exploration stage is to find important variables and determine their nature. If you have a large number of variables to consider, you may need to reduce them to a range that is easy to deal with. Second stage is pattern identification. Look for patterns and choose the best one that will allow you to make the best predictions. The third stage is called deployment. You will not want to move to this stage until you 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. Modeling is the process of creating a model. The purpose of data mining is to take the model and place it in a situation where the answer is unknown.

1.1 OBJECTIVES:

1. Analysis of the database to build an unsupervised model for the identification of the most significant characteristics of insolvent inhabitants in the selected area using decision tree

2. To predict the chances of hitting the disease using the supervised classifier model for insolvent customers

3. To build a model for classifying the solvent and insolvent inhabitant using the decision tree model..[9]

1.2 DATA MINING AND ITS USES 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 future or 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 data. Data mining tool such as SPSS is one among it. It 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.[4]

In data mining first the data must be extracted, transform and load the transaction data into the and stored in a data warehouse system. Then store and manage the data into a multidimensional database system. Then providing the data access to analyst with the help of a software and presenting the data in a useful presentable format.

There are different types of analysis. They are data visualization, Rule induction, nearest neighbor method, clustering, generic algorithm, decision tree model etc. In database operation , the predicted result will be already known by the user. But 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 . [3] 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 diseases are

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

2. RESEARCH METHODOLOGY AND DATA ANALYSIS

To find the prediction of disease hit in an area, high dimensional data were collected from different sources.. In the monsoon some disease

is hitting in almost all families in this area.This is happening for the last several years. So the relevant data have collected from each families about each members. The main parameters were name, age, education, income,employment, sex, hereditary factors, area , drainage facilities, drinking water facilities, toilet facility , waste disposal, electricity, approaches to hospital, roads, livelihood etc and created a database.

2.1 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.

3 CHOOSING FUNCTIONS OF DATA MINING AND MINING ALGORITHM

The segmentation of inhabitants according to the hit of disease may be viewed as the main problem. For applying the decision tree 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. The main advantage of k-means algorithm are its simplicity to understand the results and its computational efficiency.

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, eliminate the records whose hit ratio is less. By applying classification technique, reduce the data set and calculate the percentage of insolvency. We can see that insolvency is higher with less population.Use 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.[6]

3.1 Why decision tree induction in data mining?

Decision tree is used to classify an unknown sample and to test the attribute values of the sample against the decision tree.It has relatively faster learning speed (than other classification methods)and can be convertible to simple and easy to understand classification rule and can use SQL queries for accessing databases. It has good classification accuracy with other methods.

3.2 Classification by Decision Tree Induction

Decision Tree is a flow-chart-like tree structure. Internal node denotes a test on an attribute.Branch represents an outcome of the test. Leaf nodes represent class labels or class distribution.

Decision tree generation consists of two phases

- Tree construction

At start, all the training examples are at the root. Partition examples recursively based on selected attributes.

- Tree pruning

Identify and remove branches that reflect noise or outliers.

3.3 Algorithm for Decision Tree Induction

- Basic algorithm (a greedy algorithm)
  - Tree is constructed in a top-down recursive divide-and-conquer manner
  - At start, all the training examples are at the root.
  - Attributes are categorical .
  - partitioned recursively based

    Test attributes are selected on the basis of statistical measure such as information gain .
- Conditions for stopping partitioning
  - All samples for a given node belong to the same class
  - There are no remaining attributes for further partitioning – majority voting is employed for classifying the leaf. There are no samples left.

Table 1:Training data set

This follows an example from Quinlan's ID3(Iterative Dichotomiser)

idcode	sex	Disease history	envirnment	education	age	class
1	F	Y	GOOD	AVG	36	N
2	F	N	GOOD	AVG	38	N
3	F	Y	POOR	AVG	42	Y
4	F	Y	AVG	AVG	53	Y
5	F	N	GOOD	GOOD	40	N
6	M	N	GOOD	GOOD	36	N
7	M	N	POOR	AVG	16	Y
8	F	N	POOR	POOR	50	Y
9	F	Y	GOOD	GOOD	25	N
10	M	Y	POOR	POOR	5	N
11	M	N	GOOD	AVG	26	Y
12	F	Y	POOR	GOOD	45	Y
13	M	Y	POOR	POOR	53	Y
14	F	Y	AVG	POOR	42	Y

7

3.4 Attribute Selection Measures

- An attribute selection measure is selecting the splitting criterion that "best" separates a given data partition. For that purpose it is using Information gain (ID3/C4.5)

\Table 2:Attributes with their possible values

NO	ATTRIBUTE	POSSIBLE VALUES
1	ENVIRONMENTAL CONDN	POOR/AVG/GOOD
2	SEX	FEMALE/MALE
3	DISEASE HISTORY	YES/NO
4	AGE	<12, 12-18,19-35,36-50,51-87
5	EDUCATION	POOR/AVG/GOOD
6	INCOME	LOW/AVG/HIGH

3.5 Information gain (ID3/C4.5)

- All attributes are assumed to be categorical.
- Can be modified for continuous-valued attributes .
- Select the attribute with the highest information gain.
- Assume there are two classes, P and N .
- Let the set of examples S contain p elements of class P and n elements of class N .
- The amount of information, needed to decide if an arbitrary example in S belongs to P or N is defined as

$$I(p,n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$$

----------(1)

- Assume that using attribute A a set S will be partitioned into sets $\{S_1, S_2, \ldots, S_v\}$
  - If S_i contains p_i examples of P and n_i examples of N, the entropy, or the expected information needed to classify objects in all subtrees S_i is

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p_i, n_i)$$

----------(2)

  - The encoding information that would be gained by branching on A

$$Gain(A) = I(p,n) - E(A)$$

----------(3)

Class P: disease history= "yes"

Class N: disease history= "no"

I(p, n) = I(8, 6) =0.940 bits

Compute the entropy for disease history:

$$E(dh) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p_i, n_i)$$

----------(4)

$$Gain(dh) = I(p,n) - E(dh)$$

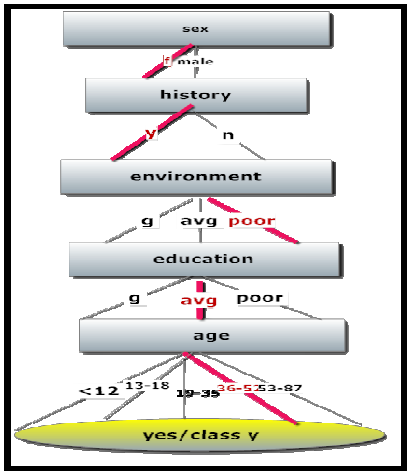
0.246 bits

Similarly,

$$Gain(sex) = 0.029$$

$$Gain(envionmentalcondn) = 0.048$$

Gain of disease history has the highest information gain. So it can be considered as the splitting attribute.



Decision tree –to check splitting attribute(sex)

4. Tree pruning: To Avoid Overfitting in Classification

The generated tree may over fit the training data because Too many branches, some may reflect anomalies due to noise or outliers .Result is in poor accuracy for unseen samples[7]

The main two approaches to avoid overfitting are

Prepruning: Halt tree construction early—do not split a node if this would result in the goodness measure falling below a threshold.So it is difficult to choose an appropriate threshold

Postpruning: Remove branches from a "fully grown" tree—get a sequence of progressively pruned trees.Use a set of data different from the training data to decide which is the "best pruned tree.

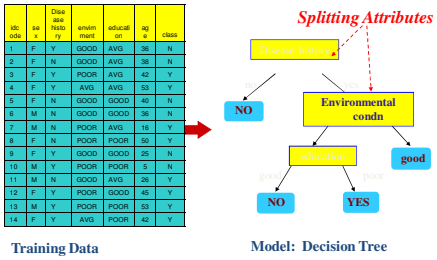


Figure 2:Example of a Decision tree

4.1 Decision tree comparison & results

- The decision tree classifier applied on the data set uses different splitting criteria such as information gain, gain ratio etc.
- The options resulted in different decision trees and the resultant accuracy of each tree when applied to the testing samples also verified.
- The change in selection criteria of best attribute may change the performance of decision tree classifier.[8]

5. PATTERN EVALUATION

The clustering analysis is identified in three different clusters according to the difference in inhabitance disease hit.

Cluster1-No disease history

Cluster2- Less disease history

Cluster 3- High tendency and disease history(This inhabitants have a tendency to become insolvent.)

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 TYPES OF CAUSES TO BECOME INSOLVENTS

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.

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 decision tree model is an effective method for clustering solvent and insolvent inhabitant in this context.[9]

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 mode l allowed predicting the insolvency of inhabitants well in advance so that the action measures can be taken against the insolvent inhabitants.

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.

- Hence from the decision tree we can conclude that mostly female inhabitant with a hereditary history living in a poor environment condition and having an average age of greater than 35 is suffering the disease.

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

REFERENCES:

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

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

[3] 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.

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

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

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

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

[8]Rui Xu , Donald C.and WunschClustering, Iee Press-2008.

[9] Smitha.T*, Dr.V.Sundaram**,** Knowledge Discovery from Real Time Database using Data Mining Technique, International Journal of Scientific and Research Publications, Volume 2, Issue 4, April 2012 1 ISSN 2250-3153.

[10] 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.

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. She has also published articles in international journals..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.

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.