ISSN No. 0976-5697

Volume 3, No. 3, May-June 2012

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Decision Tree of Behavioral Model for Income of Indian Adults at USA

Dr. Manju Mandot[1] and Rajesh Soni*[2]
Associate prof., Dept. of C.S. & I.T.[*1] Research Scholar, Dept. of C.S. & I.T.[2]
J.R.N. Rajasthan Vidyapeeth, Udaipur, India
Manju.mandot@gmail.com[*1]
rajeshrajsoni@indiatimes.com[2]

***Abstract*-**Data mining is the important field of computer science. Decision tree is important tool of data mining. All the tools of data mining is used to extract knowledge from huge amount of data base. In this paper after giving introduction of data mining, decision tree is discussed in detail, how it is constructed. The adult dataset is popular data set. In this there are more than thirty two thousand instances. Using Weka, an important data mining tool, the decision tree is constructed for country India in adult data set. Using that decision tree the knowledge can be extracted.

***Key words*-** Data mining, Decision tree, Adult data set, Visualization, Weka etc.

I. INTRODUCTION

Data Mining is a techniques, it helps to extract important data from data ware house. It is the process of arranging large amount of data and picking the relevant information by the use of certain algorithm. As the amount of data doubling, data mining is becoming an important tool to transform this data into information. The techniques in data mining are the result of a large process of research. Data Mining is supported by 3 technologies. [1]

(i) Massive data collection
(ii) powerful computers
(iii) Sophisticated Data Mining algorithms.

By the use of data mining tools, it's allowing business decision to knowledge driven decision. Data mining tools reduce time consuming process of business question. By the use of data mining tools, we can find hidden pattern, predictive information etc. Data mining software is analytical tool for analyzing data.

Data mining is synonym for knowledge discovery in Database (KDD). Data mining is an essential step in the process of knowledge discovery in the data base.

Data mining algorithms are based on two kinds of learning (i) Supervised learning (ii) Un supervised learning

II. CLASSIFICATION ALGORITHM [2]

The purpose of classification algorithm is to find relationships in the data & how different combinations of variables affect each other. Classification algorithms have 2-step

First Step: - This step also known as learning step. This step creates a model to describe a pre-determined set of data classes or concepts.

Second Step: - In this step the model is used for classification.

A. Decision Tree learning:

In decision tree learning, a set of training example is broken down into smaller & smaller subsets. At the same time an associated tree get incrementally developed.

At the end of learning process, a decision tree covering the training set is returned.

a. Representation of Decision Tree:

(i). Each Internal node corresponding to a set
(ii). Each branch corresponding to a result of test.
(iii). Each leaf node assigns a classification.

Once the tree is developed, a new instance is classified by starting at the root and follows the path dictated by the test result for this instance.

Steps to build decision tree [3]

(i) Firstly test all attributes & select the one that would fiction as the best root.
(ii) In this step break-up training set into subsets based on the branches of the root node.
(iii) Test the remaining attributes to see which one fit best underneath the branch of the root node.
(iv) Continue this step for all other branches until
a. All example of a subset are of are type.
b. There are no examples left
c. There are no more attribute left.

Determination of root

Entropy (E) is given by

$E(S) = - p \log_2 P - (1-P) \log_2 (1-p)$

S - It is a sample of training examples

P - It is proportion of positive examples in S.

E (C(X +, Y -)) i.e. there are x positive training Example and Y negative training example

Information Gain

G(S, A) = Expected reduction in entropy due to sorting on A [4]

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Now we have to calculate information gain G(S, A)

Each branch corresponds to an attribute value node.

III. ADULT DATA SET

From archive.ics.uci.edu/m1/ website which is UC Irvine Machine learning Repository. There is popular data sets name Adult. The donor of this is Ronny Kahavi & Barry Becker. The Adult data set [5] is also known as census income data set. The area of this data set is social. There are 48842 no. of instances, 14 Number of attributes. Missing values are also there.

This data set is result of extraction from the 1994 census data base.

A. Visualization of India data set:

As shown in diagram the minimum & maximum value of age is 34 & 61 respectively. In workclass the 70 instances belong to private. In education 27 instances belong to master & 21 Bachelors. In marital-status 64 instances are in married-civ-spouse. 40 instances are prof-specialty in occupation. In relationship 58 are husband. In race 85 are Asian-Pac-islander. In sex 89 are male. In hours-per-week 8 minimum & 84 maximum. In class 40 instances >50 K & ≤ 50K

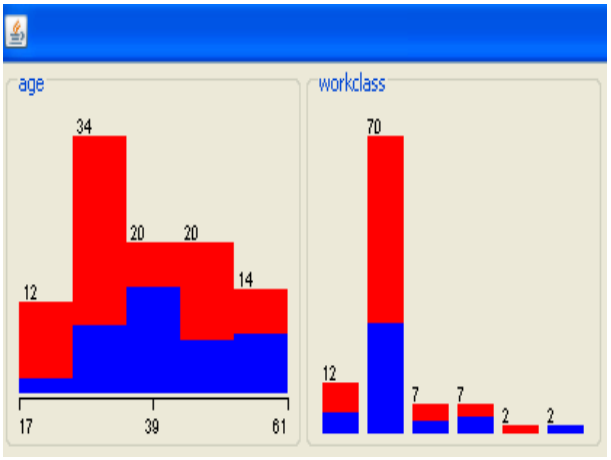

Figure 1 Visualization of age & workclass

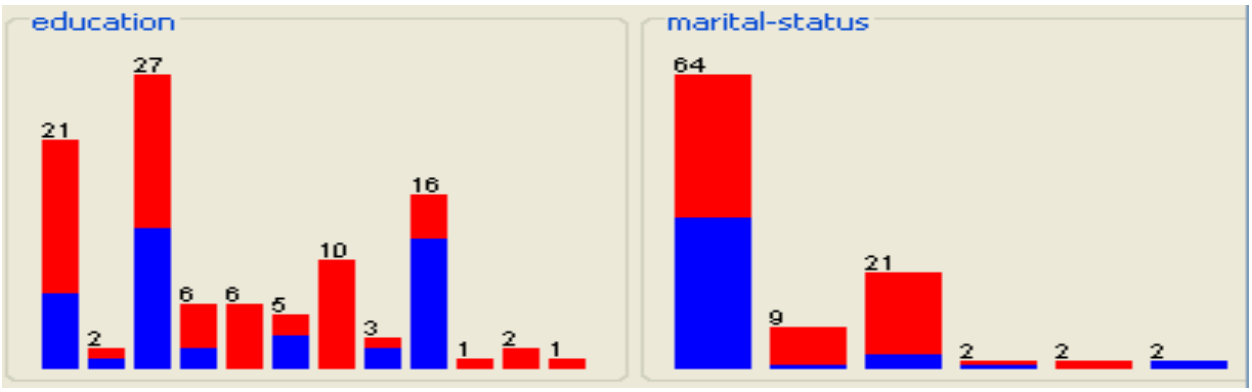

Figure 2 Visualization of education & marital-status

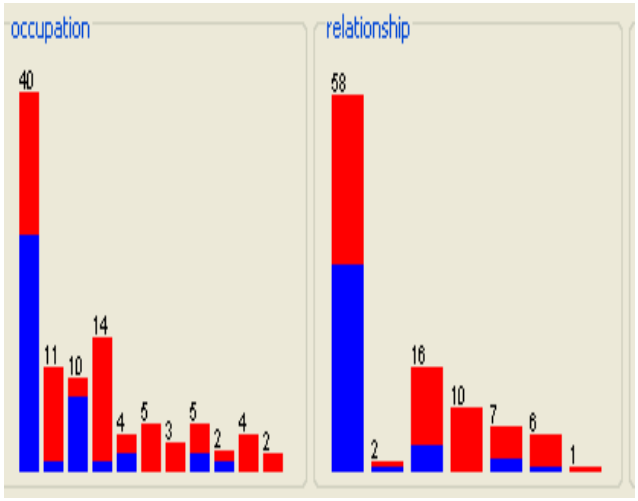

Figure 3 Visualization of occupation & relationship

Table 1 Classification of instances

S.No.	Occupation	No. of Instances
1	Prof-specialty	40
2	Adm-clerical	11
3	Exec-managerial	10
4	Sales	14
5	Other-service	4
6	Craft-repair	5
7	Handlers-cleaners	3
8	Tech-support	5
9	Transport-moving	2
10	Protective-serv	4
11	Machine-op-inspct	2

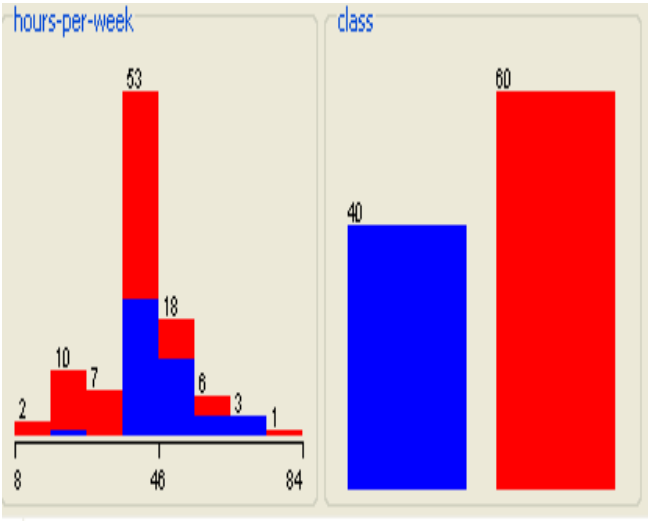

Figure 4 Visualization of hours-per-week & class

B. Decision tree of India:

Weka [6] tool is used for building decision tree. In adults data sets at native country options select India. All the instances belongs India at native country attribute displayed. There are 100 instances of India

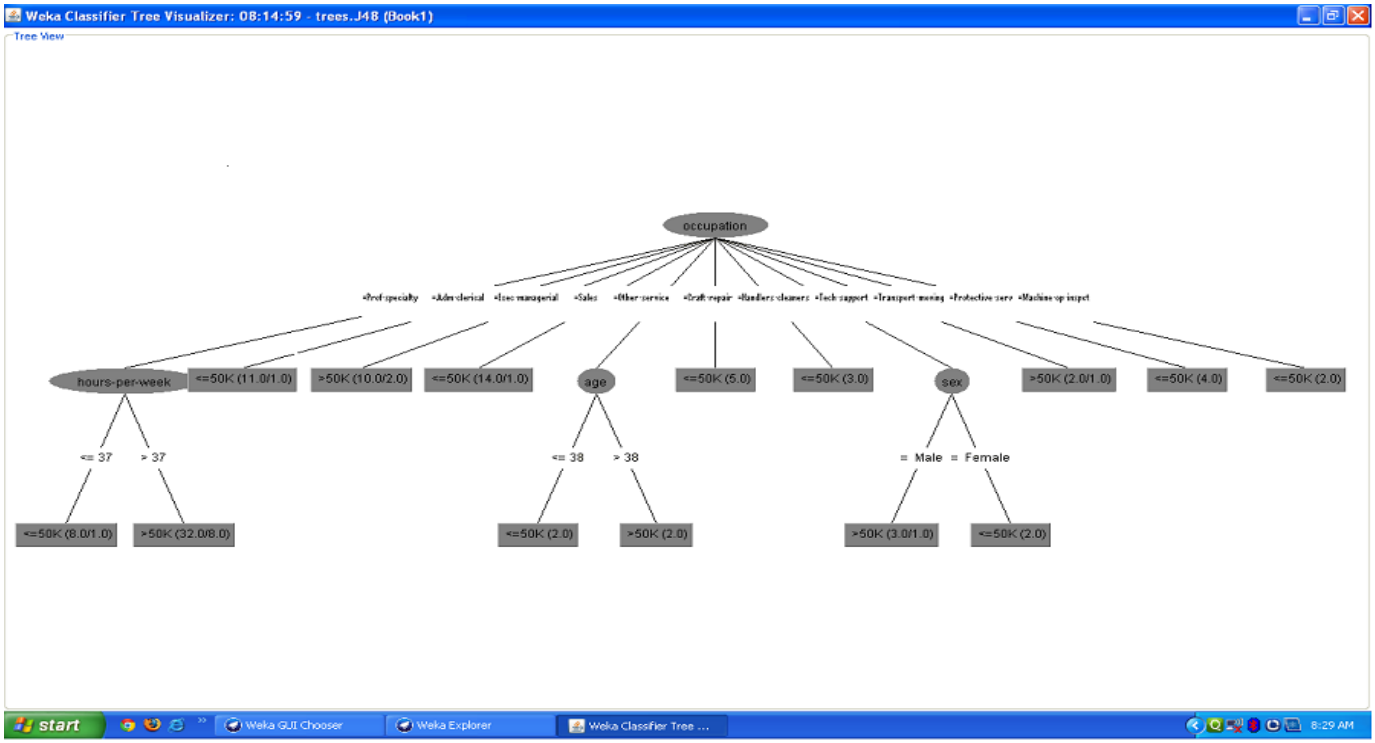

Figure 5 Decision Tree of India (Adult data set)

IV. ANALYSIS

Expected salary for different professional as per their occupation are given table below

Table 2 occupation & salary

S.No.	Occupation	Condition	Class	Remark
1	Prof-specialty	hours-per-week<=37	<=50k	Otherwise >50k
2	Tech-support	Age<=38	<=50k	Otherwise >50k
3	Exec-managerial		>50k	
4	transport-moving		>50k	
5.	Other-service	Age<=38	<=50k	Otherwise >50k

For Adm-clerical, Sales, Craft-repair, Handlers-cleaners, Protective-serv, Machine-op-inspct class is <=50k.

V. CONCLUSION

The decision tree is the important tool for data mining. Based upon this tool knowledge can be extracted from the decision tree of India. The decision can be made about the class or salary of Indian person belongs to different occupation. The businessmen can make planning about his employee for range of class or salary.

VI. REFERENCE

[1]. B. Saikia, D. Bhowmik ,Thesis Study of Association Rule Mining and Different Hiding Techniques ,Department of Computer Science Engineering ,National Institute of Technology ,Rourkela, 2009 ,pp2.

[2]. R. Knoetze, Thesis The Mining and Visualisation OF Application Services Data, Faculty of Science, Nelson Mandela Metropolitan University, 2005, pp32.

[3]. www.bi.snu.ac.kr/Courses/g-video12s/files/ID3.ppt.

[4]. www.ist.temple.edu/~vucetic/cis526fall2007/ping.ppt

[5]. R. Kohavi, B. Becker Silicon Graphics, UCI Machine Learning Repository, 1994,

http://archive.ics.uci.edu/ml/datasets/Adult .

[6]. http://www.cs.waikato.ac.nz//ml/weka