ISSN:2231-0711

Rajesh Soni | IJCSET |March 2013 | Vol 3, Issue 3, 90-93

Visualization of Behavioral Model using WEKA

Rajesh Soni
Lecturer, B.N.P.G.Girls'College, Udaipur

ABSTRACT:-The pictorial presentation is very useful for understanding the dataset. The visualization of India dataset of adult dataset have been done using freeware tool WEKA. Different attributes are presented graphically to understand their characteristics.
Key word: Adult dataset, WEKA, Visualization, India etc.

1. INTRODUCTION

Data Visualization is a science, which used images to represent the data in order to enable the viewer to view it in a better way. Representation of data in graphical form is data visualization. There are large number of conventional ways, these are Bar graphs, Pie charts, Histograms, table etc. The main aim is to present details to viewer. By the use of visualization the understanding the data and their relationship is much simpler.

The related fields of data visualization are

(i)Data acquisition (ii)Data analysis(iii)Data governance(iv) Data mining (v)Data transforms

(i)Data acquisition

Data acquisition means sampling of the real world to generate data that can be manipulated by a computer

(ii)Data analysis

In data analysis there are studying & summarization of data to extract useful information & make decision. Data analysis is closely related to data mining. Difference is that, in data mining there are large data sets.

(iii)Data governance

Data governance contain the people, process & technology to create a consistent & enterprise view of an organization's data in order to

(a) Increase consistence in decision making
(b) Increase confidence in decision making
(c) Decrease risk
(d) Improve data security
(e) Maximize the income generation
(f) Designate accountability

(iv)Data Mining

Data mining is the process to extract knowledge from large amount of data.

(v)Data transforms

It is the process of automation & transformation of both real time & offline data from one format to another

2. DATA VISUALIZATION TECHNIQUES

To represent data graphically there are many techniques, some of them are

(i)Charts (Pie diagrams & bars)
(ii)Graphs
(iii)Maps
(iv) 3-D surface
(v) Images
(vi)Animation

Data visualization techniques can be classified into 5 categories

(i)Geometric techniques
(ii)Icon- based techniques
(iii)Pixel – oriented techniques
(iv)Hierarchical techniques
(v)Graph- based techniques

Geometric technique is the visualization of geometric transformation & projections of data. Example:- Scatter –plot matrices , Landscapes etc.

In icon –based technique there is visualization of data values by mapping each multidimensional data to an icon Example:- Color icons, Tile bars

Pixel – oriented technique used to represent each attribute values of a data item as a colored pixel & display the attribute values belonging to one data item in separate windows.

Hierarchical technique is used for visualization of data using hierarchical partitioning of k-dimension space into 2D or 3D subspaces. Graph base technique uses large graphs to represent information & structure of data sets. The graph can be 2D or 3D.

3. ADULT DATA SET [2]

From archieve.ics.uci.edu/m_1/ website which is UC Irvine machine learning Repository, There is popular data sets name Adult. The donor of this is Ronny Kohavi & Barry Becker. The Adult data set is also known as census income data set. The area of this data set is social. There area 48842 no. of instances, 14 number of attributes. Missing values are also there.

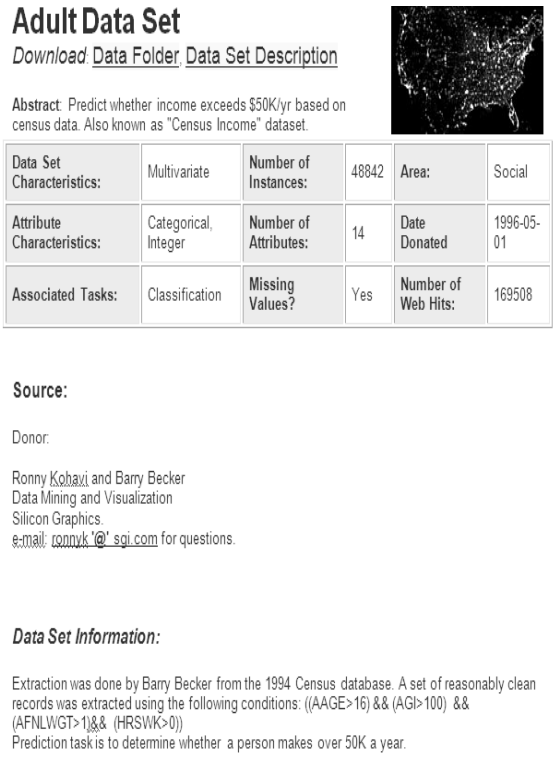
Attribute Information:

Listing of attributes:

- Class : >50K, <=50K.
- Age**:** continuous.
- work class: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked. fnlwgt: continuous.
- Education : Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool. education-num: continuous.
- Marital-status**:** Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
- Occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-

cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.

- Relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried. race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
- Sex: Female, Male.
- Capital-gain: continuous.
- Capital-loss**:** continuous.
- Hours-per-week: continuous.
- Native-country : United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holland-Netherlands.



Adult Data Set

Download: Data Folder, Data Set Description

Abstract: Predict whether income exceeds $50K/yr based on census data. Also known as "Census Income" dataset.

Data Set Characteristics:	Multivariate	Number of Instances:	48842	Area:	Social
Attribute Characteristics:	Categorical, Integer	Number of Attributes:	14	Date Donated	1996-05-01
Associated Tasks:	Classification	Missing Values?	Yes	Number of Web Hits:	169508

Source:

Donor:

Ronny Kohavi and Barry Becker
Data Mining and Visualization
Silicon Graphics.
e-mail: ronnyk '@' sgi.com for questions.

Data Set Information:

Extraction was done by Barry Becker from the 1994 Census database. A set of reasonably clean records was extracted using the following conditions: ((AAGE>16) && (AGI>100) && (AFNLWGT>1)&& (HRSWK>0))
Prediction task is to determine whether a person makes over 50K a year.

Figure 1 Adult Dataset

4. WEKA [1]

Weka is open source and WEKA full form is Waikato Environment for knowledge analysis. WEKA consists of several standard machine learning techniques. Weka can be used to derive knowledge from database that are far too large to be analyzed by hand. WEKA acts as an interface to machine learning schemes & data sets. Weka can be used to produce rules & decisions trees based on the current data set. These rules & decisions trees can be used in application to the domain.

The key features of Weka's are

(i) It provides many different algorithms for data mining & machine learning
(ii) It is open source & freely available
(iii) IT is plat form - independent
(iv) It can be used easily by people who are not data mining specialist
(v) It has kept up to date.

WEKA can be used for data pre-processing, classification regression, clustering, association rules & visualization. Weka website is http://www.cs.waikato.ac.nz/ml/weka

Interfaces

(i) Explorer (ii) knowledge flow (iii) Experimenter (iv) command line interface (CLI)

Weka requires arff file format

In arff format there is requirement of declaration of @ Relation, @ attribute and @ data. @ RELATION declaration associates a name with the dataset

@ RELATION <relation - name>

@ ATTRIBUTE declaration specifies the name & type of an attribute

@ATTRIBUTE < Attribute-name><data type>

Attribute types

(i) Nominal (ii) Numeric (iii) string (iv) Data (v) Relational

(i) Nominal: - This contain one of a predefined list of values

(ii) Numeric: i.e. real or integer number.

(iii) String: - This is enclosed in double quotes

@ Data declaration is a single line denoting the start of data segment

? Used to represent missing values

WEKAs have

(i) 49 data pre processing tools
(ii) 76 classification / regression algorithms.
(iii) 8 clustering algorithm
(iv) 3 Algorithm for finding association rules
(v) 3 graphical user interfaces.

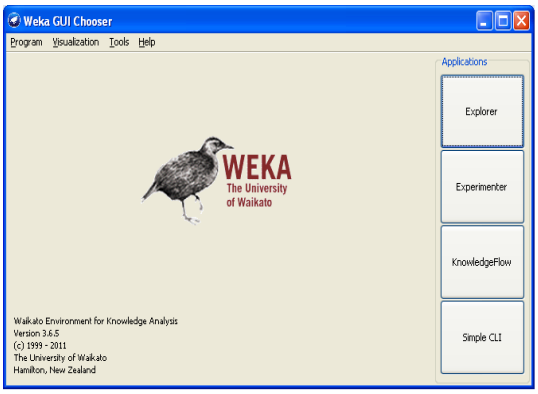


Figure 2 Snapshot of WEKA

5. VISUALIZATION OF COUNTRY INDIA IN ADULTS DATA SET USING WEKA

There are 100 records of country India in Adult data set. The mathematical characteristics & visualization of various attributes have been done using WEKA as shown in figure 3 to figure 12.

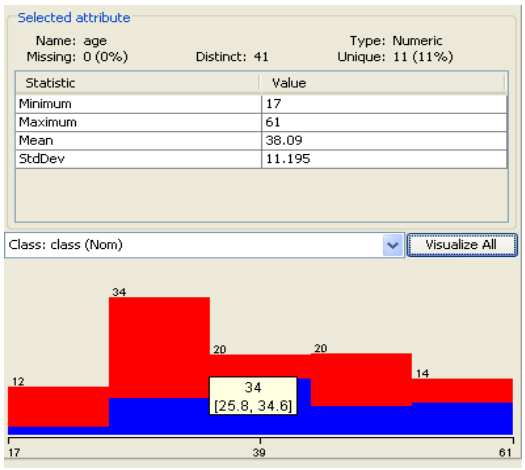


Figure 3 Visualization of attribute age

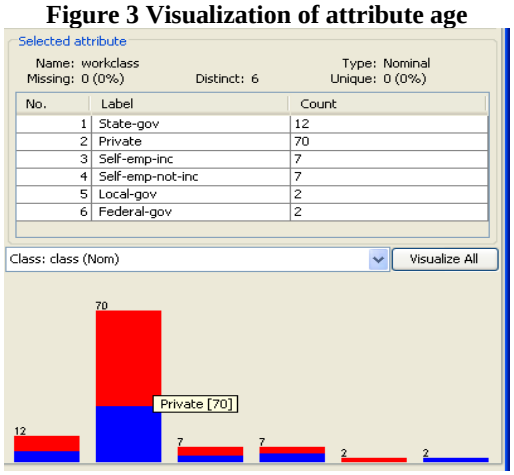


Figure 4 Visualization of attribute work

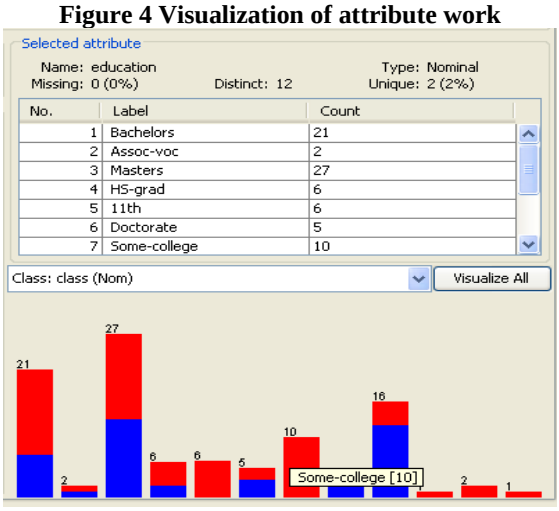


Figure 5 Visualization of attribute education

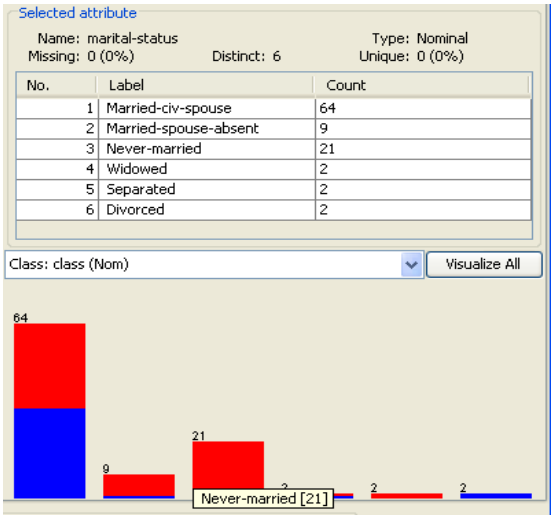


Figure 6 Visualization of attribute marital-status

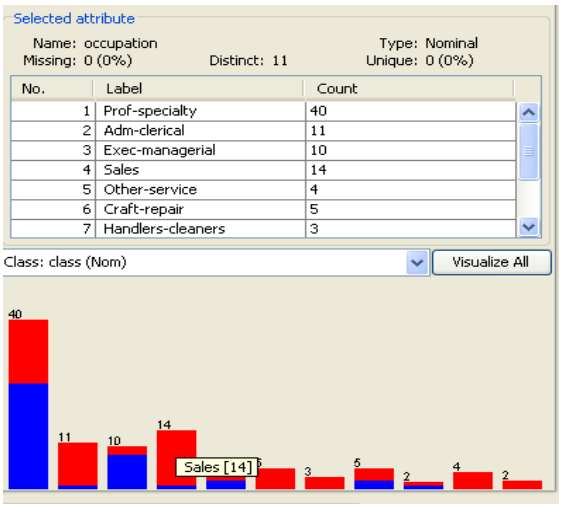


Figure 7 Visualization of attribute occupation

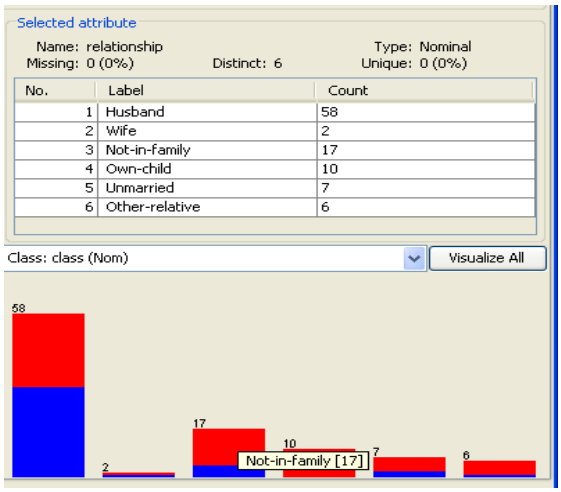


Figure 8 Visualization of attribute relationship

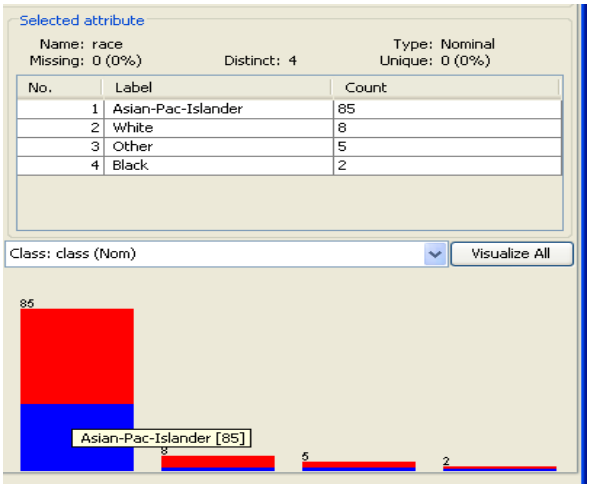


Figure 9 Visualization of attribute race

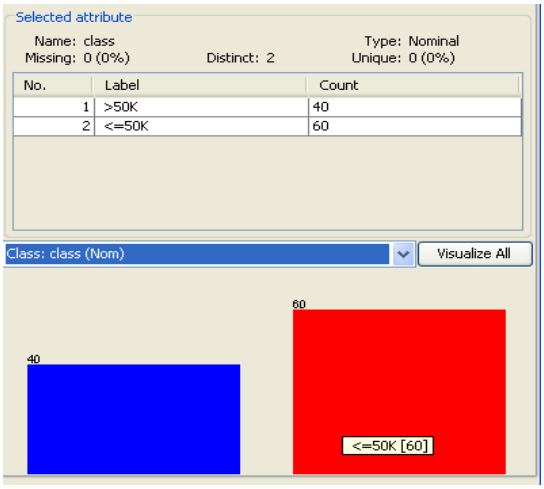


Figure 12 Visualization of attribute class

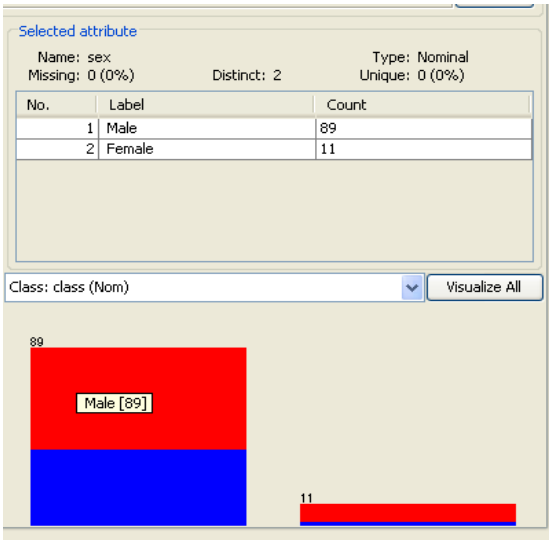


Figure 10 Visualization of attribute sex

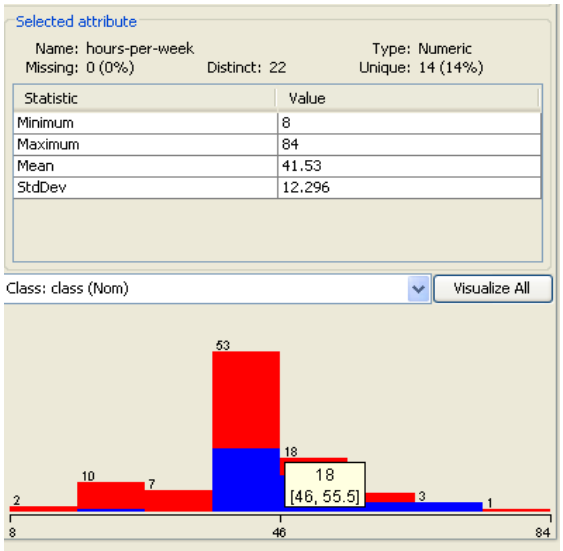


Figure 11 Visualization of attribute hours-per-week

6. ANALYSIS

Attribute	Key characteristics
Age	Minimum value 34 ,Maximum value 61
Workclass	70 instances belong to Private
Education	27 instances belong to Master, & 21 to Bachelor degree
Marital-status	64 instances are in married-civ-spouse
Occupation	40 instances are prof-specialty
Relationship	58 instances are husband
Race	85 instances are Asian-Pac-islander
Sex	89 instances are male & 11 female
Hours-per-week	8 minimum & 84 maximum
Class	In 40 instances >50k ,60 instances <=50k

7. CONCLUSION

The visualization of India data set helps us in very different ways. We can make decision about the average salary of Indians in USA. We can reach at the conclusion about the socio-economic status of Indians in USA.The condition of Indian female in USA.The WEKA is very powerful tool for presenting the mathematical characteristics & visualization.

REFERENCES

1 http://www.cs.waikato.ac.nz//ml/weka
2 .R. Kohavi, B. Becker Silicon Graphics, UCI Machine Learning Repository, 1994, http://archive.ics.uci.edu/ml/datasets/Adult