Volume 3, Issue 5, May 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Developing a Rule Based Model for product analysis Using Frequent pattern analysis

Mr. Ramesh Gupta,
ASET Department,
Amity University,
Noida (U.P)-201303, INDIA,
Research Scholar

Mr. Gaurav Dubey
ASCA Department,
Amity School of Computer Science,
Noida (U.P)-201303, INDIA,
Research Scholar

Mr. Sanjeev Choudhary
ASET Department,
Amity University,
Noida (U.P)-201303, INDIA,
Research Scholar

Abstract - With a huge amount of data stored in databases and data warehouses, it is very important to develop powerful tools for analysis of such data and mining useful knowledge from it. This paper is about a specific case of real estate world dataset to demonstrate how each information step can be completed interactively to get useful and required knowledge in less time/efforts from its data warehouses with human efforts. As we know that "Research is for the people not for yourself" so we are pleased to work for the real estate and hence for the society for the better living. In this paper, a survey of the research in the area of real estate is presented with its findings. Applying Data mining techniques to the real estate industry can be very useful in extracting customer preferences at any given time.

Keywords: Data Mining, Real Estate, Customer preferences, Interactive Knowledge Discovery.

1. INTRODUCTION

The data is collected from 307 customers of real estate projects going on NCR region. This region has seen as a rapid development of Residential and commercial projects during last few years. Data mining technology allows a company to use the mass quantities of data that it has compiled, and develop correlations and relationships among this data to help businesses improve efficiency, learn more about its customers, make better decisions, and help in planning. Data Mining has three major components Clustering or Classification, Association Rules and Sequence Analysis. This technology can develop these Result oriented Analysis.

In this paper we are considering various attributes of a customer buying a flat such as: Age, Occupation, Family Income, Family size, No. of Bedrooms in the flat and area and cost of the flat. we are divided the customer's occupation into three broad categories which are:

i.**Professional**: This may include teachers, doctors, lawyers, IT professionals etc.

ii.**Govt. Employees**: This may include Bureaucrats,PSU Employees, Nationalized Bank Employees, Ex Defense Personnel etc.

iii.**Businessman**: Everybody who is self employed or has his/her own business is included in this category.

This thesis work introduces a research, which aims to develop a data warehouse system according to the requirements of customers in real estate, and use data mining rule based technology to learn useful information and knowledge from the data warehouse system of real estate.

2. RESEARCH METHODOLOGY

Since the process of knowledge discovery begins with data selection, the researcher chose real estate dataset for the data mining experiment. The dataset has been created by survey there are 27 attributes and 307 instances in the dataset.

TABLE 1: Description of attributes of Real-estate Dataset

Attribute Name	**Description**	Category
First_name	First name of customer	Discreate
Last_name	Last name of customer	Discreate
Email_id	Email-id of customer	Discreate
Mobile_no	Mobile no of customer	Continue
Time_contact	Contact timing of customer	Discreate
Age	Age of customer	Continue
Sex	Sex of customer	Discreate
Address	Address of customer	Discreate
City	City of customer	Discreate
State	State of customer	Discreate
Zipcode	Zipcode of city of customer	Continue
Occupation	occupation of customer	Discreate

Occ_address	Occupation address of customer	Discreate
Family_size	Family size of customer	Continue
Date	Contact date of customer	Discreate
Pro_buying	Buying a property	Discreate
Pro_selling	Selling a property	Discreate
Pro_location	Property location	Discreate
Pro_city	Property city	Continue
Pro_price	Property price	Continue
Pro_size	Property size	Continue
Pro_bedroom	Property bedroom	Continue
Pro_persquarefeet	Property per square feet size	Discreate
Annual_income	Annual income of customer	Continue
Pro_loan	Required loan for Property	Discreate
Pro_status	Property available status	Discreate
Pro_builder	Property builder name	Discreate

The data was first converted into desired MS-EXCEL format and cleaned for missing values while maintaining the integrity and intactness of the dataset then convert into text tab delimiter format.

3. DATA ANALYSIS AND RESULTS

The data mining may help in answering several important and critical questions related to present application domain such as: 'What factors are more crucial to find the property?' or 'What is the price of a particular property?'. The role of data mining is not to practice the outcomes but to fetch useful information and knowledge from the real-estate dataset. TANAGRA - a free, open-source, user-friendly software product developed by Ricco Rakotomalala, has been used to conduct the experimentation.
The Table 2 shows the different components of TANAGRA used in the data mining experiments under discussion.

3.1. Working with Tanagra on real estate Dataset:

Open Tanagra then open data set file which is in txt, xls or arff format. We use tanagra for finding the rule based mining analysis of various attributes such as Max, Min, Mean, std. deviation etc. then we find its occupation attribute with respect to age, bedroom, annual income, size of flat, cost of the flat etc. then finally we are able to give description of customer queries preferences in respect of area, size, cost, location of the flat information.

TABLE 2: Components of TANAGRA Used in the Experiments

Tab	Operator (component)	Comment
Data visualization	View dataset	Visualise the current dataset in a grid.
Statistic	Group characterization	Select the grouping values
Statistic	Univariate discrete state	Displaying the discrete values information
Instance selection	Rule based selection	Select a subset of examples based upon a rule.
Instance selection	Discreate select example	Select the discrete values
Feature selection	Define status	Specify the attributes to use

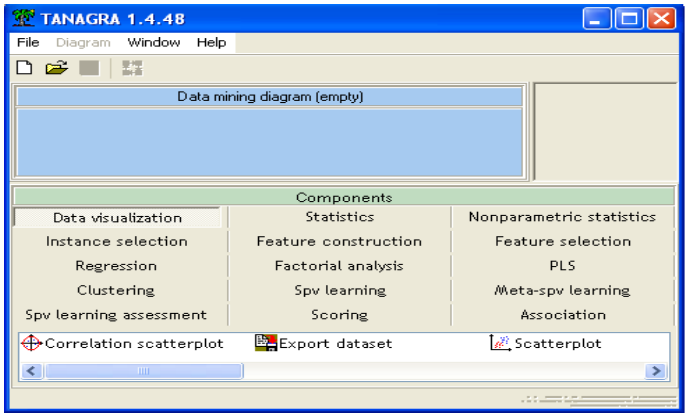

Fig 1. Tanagra Home page

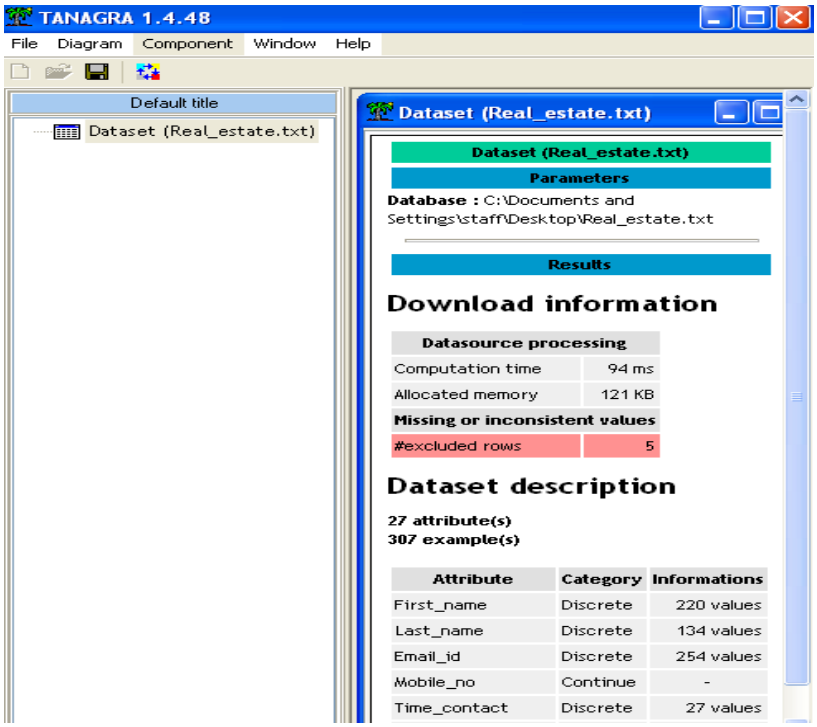

Fig 2. open file information

As we open the file in dataset it will appear as given below, the screen shot given below clearly indicate the open file name Real_estate.txt and also on the task bar of Tanagra. The down load information is on the right side of the screen given in Figure 2 and Figure 3.

Now we select view dataset from data visualization tab and drag it and drop on dataset. And get all display of dataset which is in Figure 3.
Now we select define status from feature selection tab and then drag it and drop to dataset then right click on define status and select parameters from pop up menu we get attributes given in figure 4.
Now select one attribute as input as occupation and four attributes as age, family_size, pro_size, pro_bedroom as target and press OK button. From statistics tab we choose Univariate continuous stat, drag and drop it in define status1. Then we use view command from pop up menu we will get following figure 5.
From example age has min value is 20, max values is 53, Average 36.66, Std dev is 6.5397 and avg. std dev is 0.1784 and similarly family size has 1,9,3.7492,1.0567,0.2818.
pro_bedroom has 1,5,2.7362,0.7830,0.2862 and pro_size has min 500,max 2200,average 1217.4463, std dev 273.1416 and average std dev 0.2244.

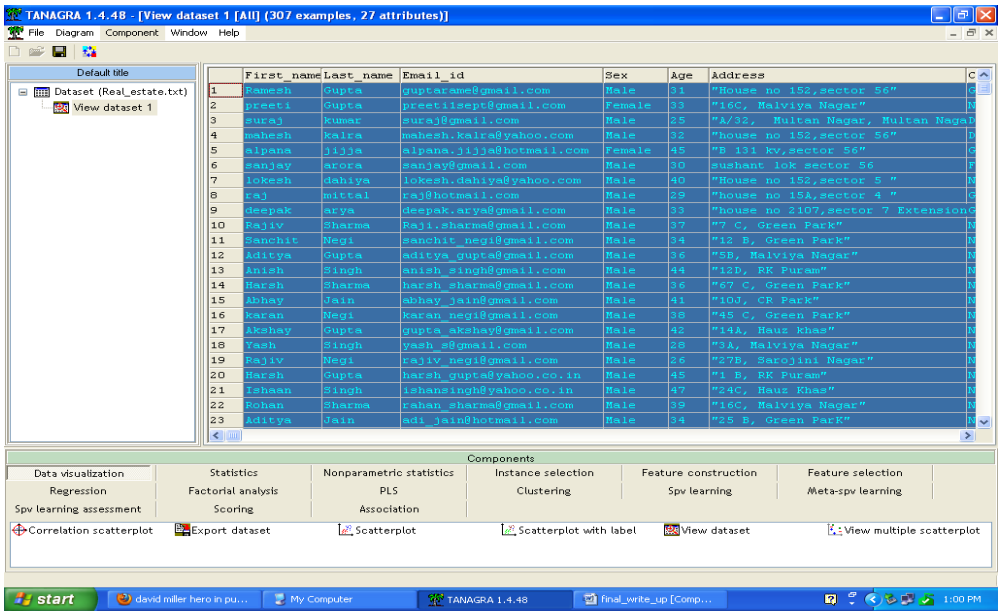

Fig 3. Real estate dataset in Tanagra

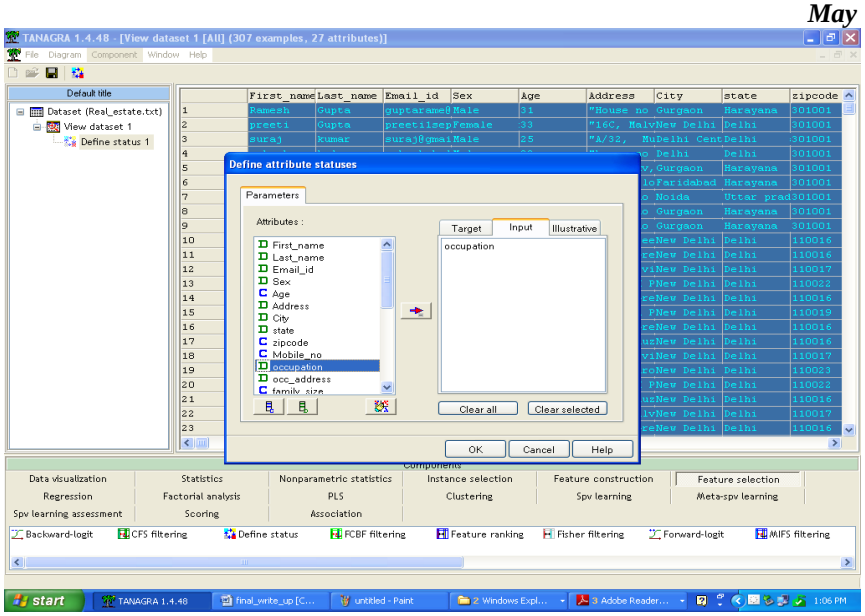

Fig 4. selection of different attribute in Tanagra

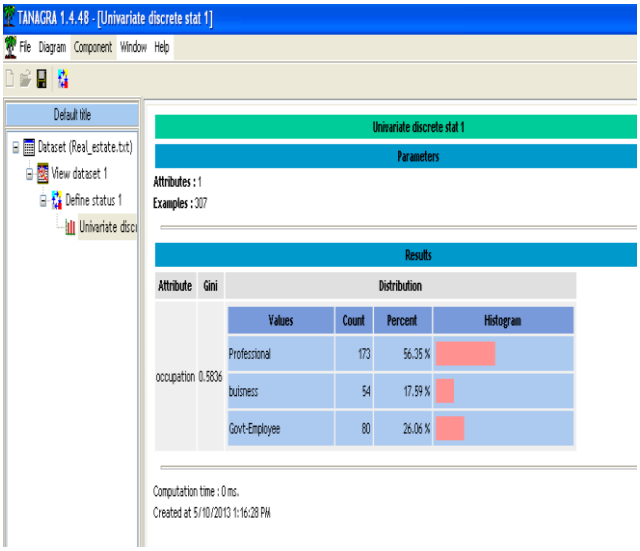

Fig 5. Min, Max, Average, Std-dev, average std dev of different attribute

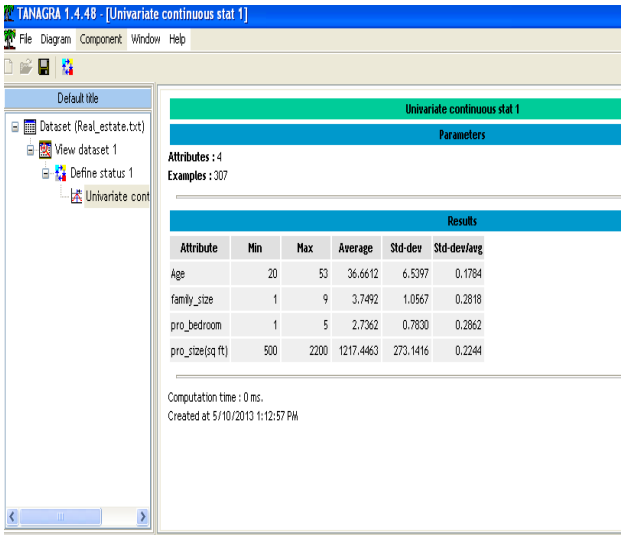

Fig 6. Occupation Distribution

Figure 6 shows the occupation percentages in professional, business, Govt-Employee category.

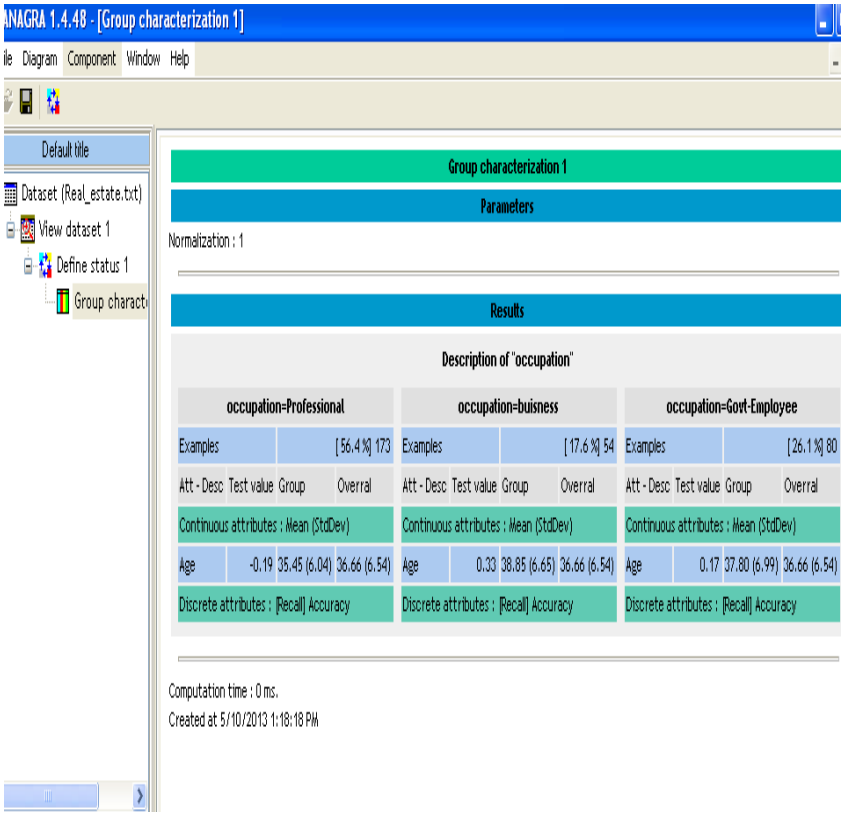

Fig 7. Average age of the customers

In Figure 7, we see that the Professionals decide to buy a flat at very early age as compared to Govt. Employees or Businessmen. This point to the aspirational and non conservative nature of the Professionals and younger generation. They start investing in real estate at much earlier stage than Govt-Employees.

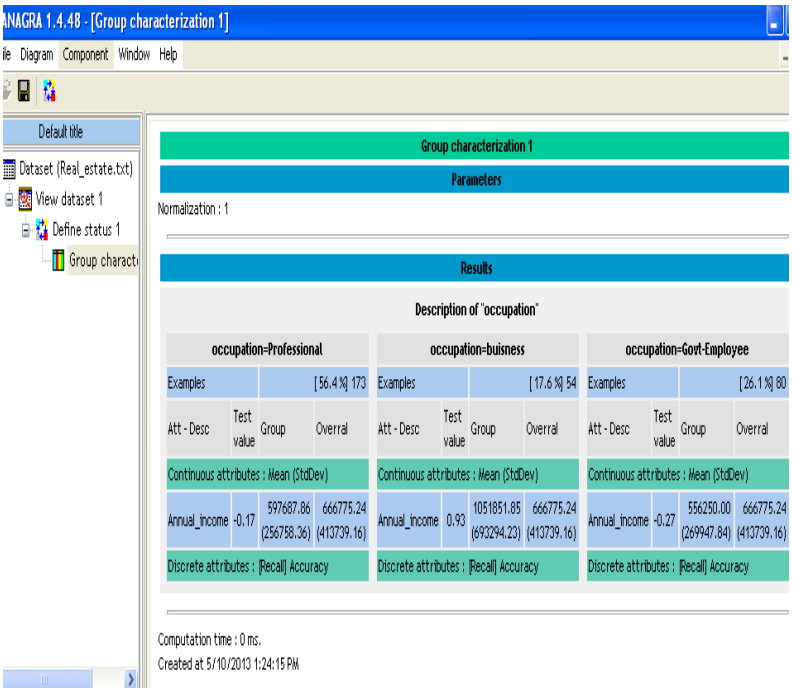

Fig 8. Average Annual Income of the customer

Average annual income, as we see in Figure 8 is expectedly highest among the Businessman because of the nature of work.

Fig 9. Average bedroom (BHK) of the customer

Fig 10. Average family size of the customer

In Figure 9, we see that the businessman buy larger flats in terms of no. of bedrooms. This is due to the fact that they have bigger incomes.

In Figure 10, we see that the average family size for professionals is considerably lower than the other two categories. This is due to two prominent reasons. One being that the professionals buying the flats are younger age. Secondly young professionals are conciously choosing to have smaller families because of the urban lifestyle constraints.

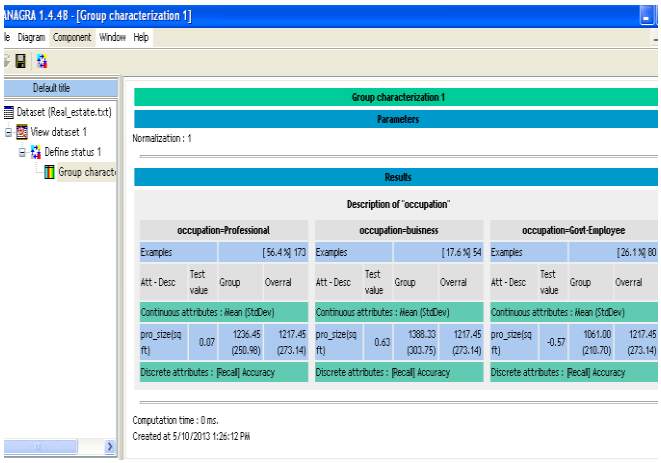

Fig 11. Average size of the flat of customer

Govt-Employees prefer less area of the flat in comparison of two other categories.

Fig 12. Average cost of the flat

Although the professionals purchase flats with lesser no. of bedrooms(Figure) and with lesser area(Figure) as compared to Businessman but the average cost of the flat for professionals is higher as compared to govt. employees. This is mainly due to the fact that professionals have greater tendency in choosing premium specifications and better located projects as compared to govt. employees.

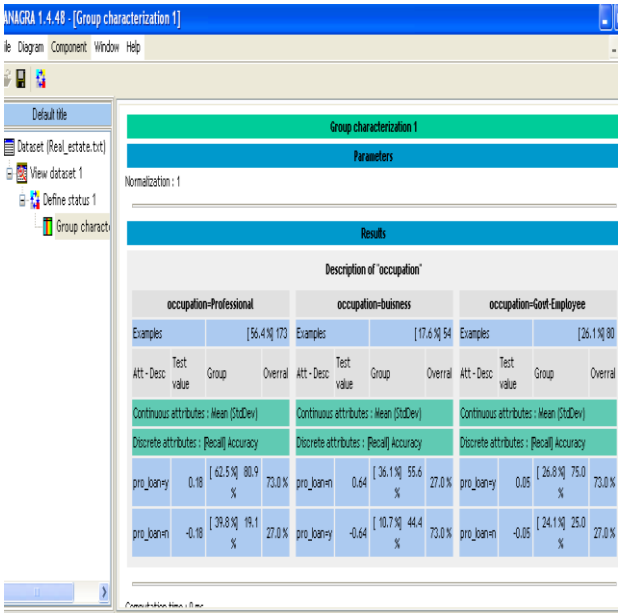

Fig 13. loan percentage taken by customer

Another important aspect of the real estate industry is the loan facility. Most people rely on long term loans to purchase a flat. This is confirmed in the survey where percentage of professionals taking loan is as high as 80.9% (Figure 13). This is expected as the professionals are younger demographic compared to other two occupation so their need for loan and reliance on loan is much higher as compared to other two. Also the percentage of businessman taking loan is much lower.

Fig 14. people taking loan/not taking loan

Maximum people want to take loan. This is true across all occupation whether it is Professionals, businessman, or govt. employees. This is mainly due to reason that loan companies prefer giving loans to younger people so it is difficult for older people to get loan.

3.2 Rule Based Model Analysis

Rule No 1 : Now customer want to retrieve the information from the dataset

If (pro_bedroom=2) and (pro_price >=2500000 and pro_price<=3000000) then 11 instances will come out according to the input which is shown in figure 15 and figure 16

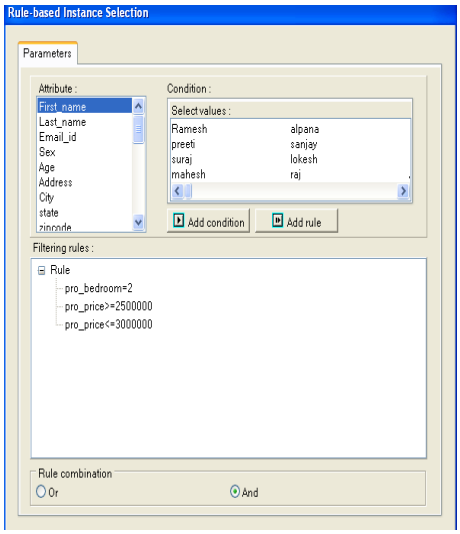

Fig 15. Design of Rule 1 for instance selection

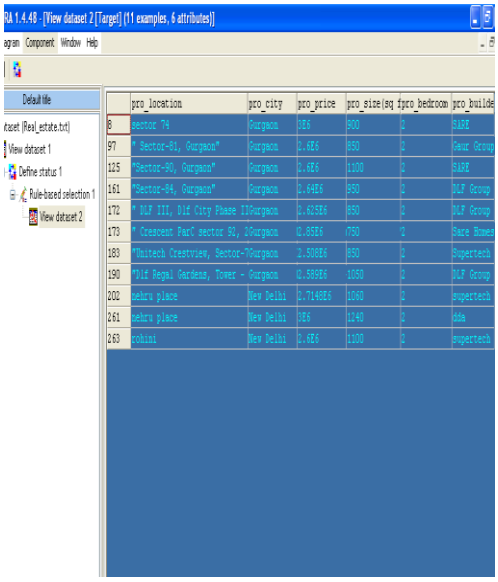

Fig 16. Instance selection based upon Rule 1

Rule 2 :if customer search information according to the builder name =”SARE” then 5 instances will come out according to the figure 18.
If builder name =”SARE” then ?

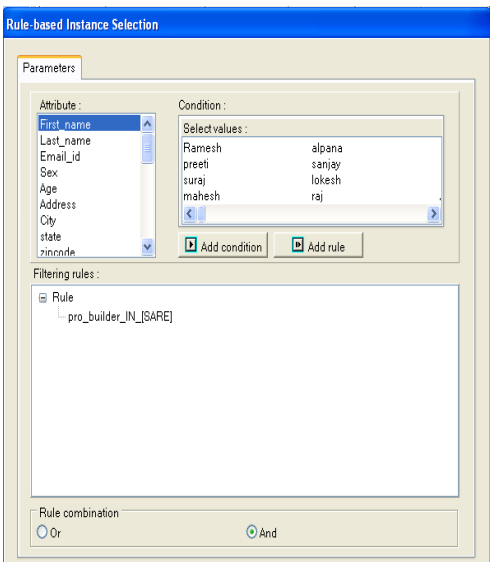

Fig 17. Design of Rule 2 for instance selection

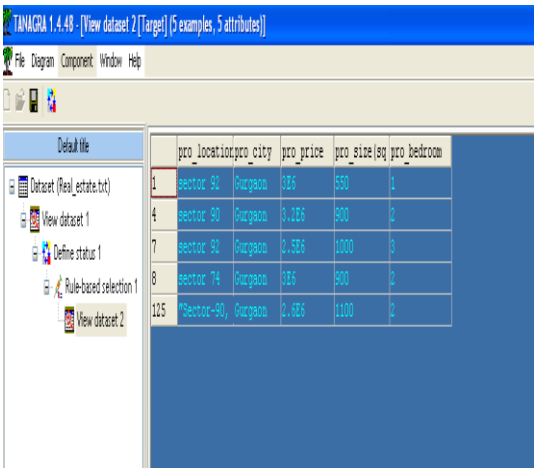

Fig 18. Instance selection based upon Rule 2

Rule3:If(pro_location=”sector92”andpro_city=”Gurgaon” then total 3 outcomes will be come and it shows pro_price ,pro_size,pro_bedroom,pro_builder which are as a target attributes.

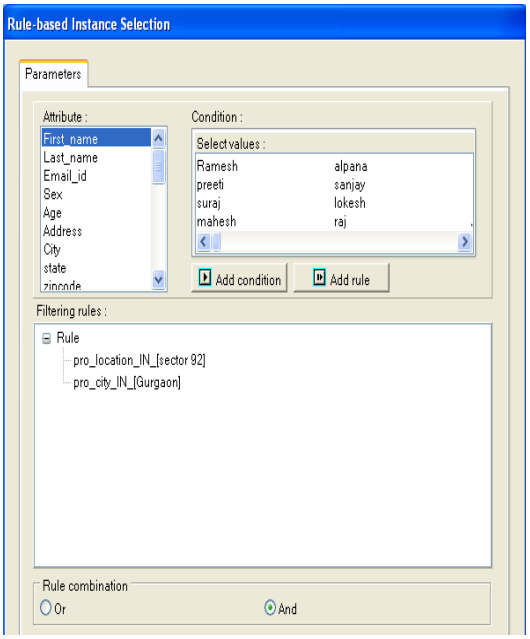

Fig 19. Design of Rule 3 for instance selection

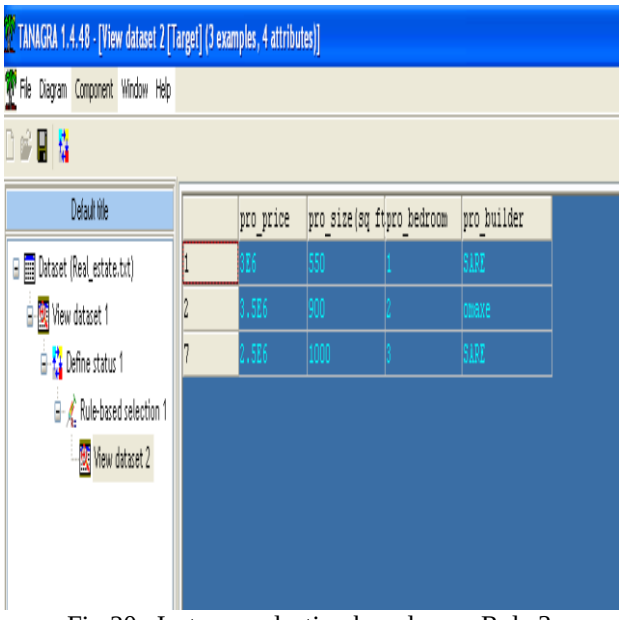

Fig 20. Instance selection based upon Rule 3

Rule 4 : If pro_city=”New Delhi” then total 184 outcomes will be come and it shows pro_price ,pro_size,pro_bedroom,pro_builder,pro_location which are as a target attributes in figure 22.

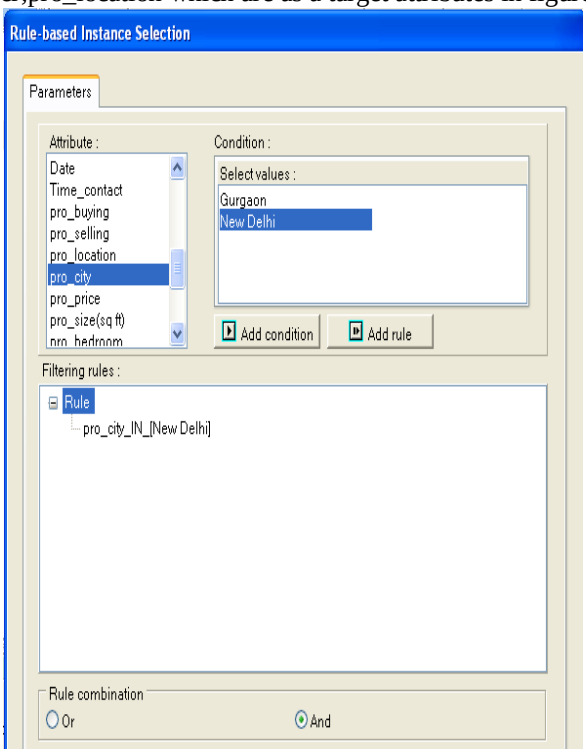

Fig 21. Design of Rule 4 for instance selection

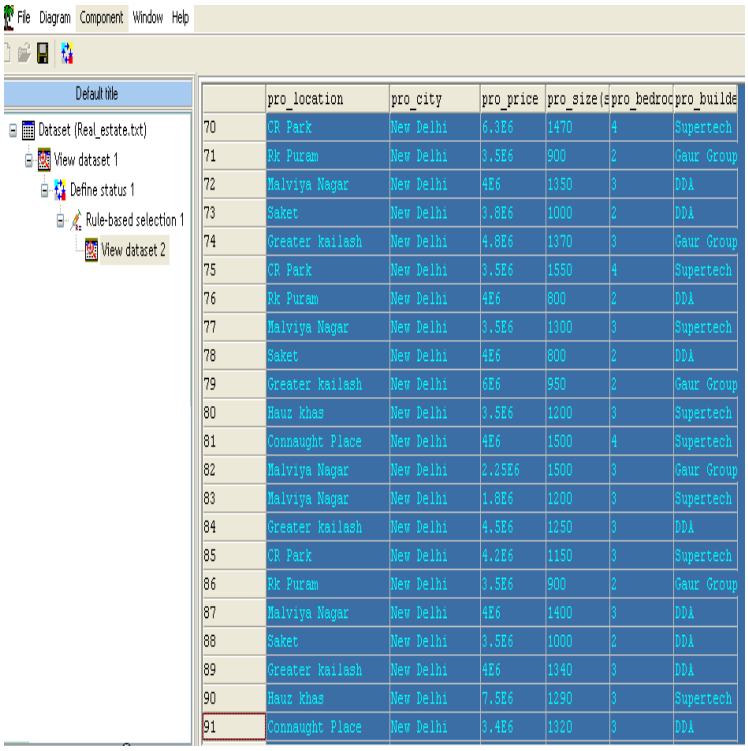

Fig 22. Instance selection based upon Rule 4

4. Conclusion

This paper was intended to provide an indication of the interactive nature of the knowledge discovery process. Majority of the customers are young professionals with smaller families who do not mind paying a bit extra for premium facilities. Rest of the market is divided between Govt-Employees and Businessmen.

Govt-Employees are conservative in their buying preferences whereas businessman tends to the buy the flats with best available specifications a developer can offer. Further researches in this area can even more customer preferences and importance to the customer. This may include features like distance to school, hospital, market, railway station etc.

Since this research would involve even larger and more complex data. Customer can search all the information of the flat by using different- different rules applies in the dataset. So in this paper customer can buy flat with their preferences easily.

References

[1] Ankerest M, Human Involvement and Interactivity of the Next generation's Data Mining Tools. ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. Santa Barbara, CA, 2001.

[2] Han J. and Kamber M., Data Mining: Concepts and Techniques. San Francisco, Morgan Kauffmann Publishers, 2001.

[3] Yeh C., King-Jang Yang and Tao-Ming Ting, Knowledge Discovery on RFM Model using Bernoulli Sequence. Expert Systems with Applications, Vol. 36(3-2), 2009, pp. 5866-5871.

[4] Ye N. and Li X., Application of Decision Tree Classifiers to Computer Intrusion Detection. Real-Time System Security, 2003, pp. 77 – 93.

[5] Masoud F.A.M., Moh'd Belal Al- Zoubi, Salah I.and Ali Al-Dahoud, Fast Algorithms for Outlier Detection. Journal of Computer Science, Vol. 4 (2):, 2008, pp. 129-132.

[6] Rakotomalala R. (2003). Tanagra. http://eric.univyon2.fr/_ricco/tanagra.

[7] B. Mobasher, N. Jain, E.-H. Han, and J. Srivastava, "Web Mining: Patterns from WWW Transactions," Dept. Comput. Sci., Univ. Minnesota, Tech. Rep. TR96-050, Mar. 1997

[8] C. Romero, S. Ventura "Educational data Mining: A Survey from 1995 to 2005", Expert Systems with Applications (33), pp. 135-146, 2007

[9] Tanagra - A free data mining software for teaching and research. http://eric.univ-lyon2.fr/_ricco/tanagra/en/tanagra/.

[10] "Data Mining" Def. www. Dictionary.Com. Date of retrieval: 01/06/2012.

