[26.1.2016]

Data Mining Assignment 3 (Data Cubes)

Done by: Debarati Das (1PI13CS052)

Overview:

Data Cubes are used to represent data along some measure of interest. Although called a "cube", it can be 2-dimensional, 3-dimensional, or higher-dimensional. Each dimension represents some attribute in the database and the cells in the data cube represent the measure of interest. For example, they could contain a count for the number of times that attribute combination occurs in the database, or the minimum, maximum, sum or average value of some attribute. Queries are performed on the cube to retrieve decision support information.

In this assignment we are provided with the bankdata.csv which has information about people in the bank. We remove unnecessary attributes and discretize age in this dataset and then compute data cubes in order to efficiently query the data cube for information we require.

Results:

1. **Determine the average income for :**
    - **a) Inner City Males**
    - **b) Middle Aged Rural Females**
    - **c) Young Suburban People**

- The answer for the first question can be derived from the 2D Data cube of Region and Sex, which would provide the average income to be **Rs. 26538.128188**

- The answer for the second question can be derived from the 3D Data cube of Age, Region and Sex, which would provide the average income to be **Rs. 27482.261538**

- The answer for the third question can be derived from the 2D Data cube Age and Region, which would provide the average income to be **Rs. 16227.501500**
-

2. **In the 2-D cuboids, which cells have a support of less than 5% (i.e. count less than 5% of the total count)?**

   Since 5% of total count(which is 600) is 30.

   The cells in the 2D cuboid which satisfy the support criteria are :

|YOUNG|RURAL	COUNT(category)	28
|YOUNG|SUBURBAN	COUNT(category)	20
|MIDDLE|SUBURBAN	COUNT(category)	22
|OLD|SUBURBAN	COUNT(category)	20