Data Mining Assignment-2

Mudit Pandey 2014A7PS017H
Vedic Sharma 2014B4A7424H

April 29, 2017

Contents

1 Language used

The entire code was written in C++.

2 Pre-processing

We process the data as follows:

1. We have 8 continuous valued attributes and 1 two-valued attribute i.e *the class variable*

2. For each continuous valued attributed, we divide the values into a number of bins.

3. Each bin of an attribute corresponds to one item

4. For each transaction, we first divide the value into bins for that attribute. The bin value will then help us in determining the item number it belongs to.

5. The bin value for each attribute is decided as follows:-

   (a) First we calculate the maximum and minimum value for that attribute

   (b) Next we divide this range into equal sized bins.

   (c) The bin number is then decided by the bin in which the value lies in. For example: Say the maximum and minimum values for an attribute are 0 and 10 respectively. Also assume that we want 5 equal sized bins. Then the various bins are:-

      i. 0-2
      ii. 2-4
      iii. 4-6
      iv. 6-8
      v. 8-10

      Now an attribute value of 3 lies in bin number 2. Attribute value of 6 lies in bin number 4.

      Note: We exclude the upper limit of the range from the bin.

6. Once the bin number is obtained for the value, we give it an item number using the simple formula:-

$$item\ number = Attribute\ number * Number\ of\ bins + Bin\ number \tag{1}$$

3 Compilation and Execution

1. g++ -std=c++11 apriori.cpp
2. ./a.out

4 Support and Confidence Values

We have generated rules for the following:-

1. Support=0.20 Confidence value=0.8

   Number of rules=105
2. Support=0.20 Confidence value=0.85

   Number of rules=68
3. Support=0.20 Confidence value=0.90

   Number of rules=28
4. Support=0.20 Confidence value=0.95

   Number of rules=4
5. Support=0.25 Confidence value=0.80

   Number of rules=65
6. Support=0.25 Confidence value=0.85

   Number of rules=42
7. Support=0.25 Confidence value=0.90

   Number of rules=14
8. Support=0.25 Confidence value=0.95

   Number of rules=3

9. Support=0.30 Confidence value=0.80

   Number of rules=31

The output is generated in a file which list all frequent item sets with their support values and the set of rules with their confidence values.

NOTE: The above results are generated for number of bins = 5