[2.2.2016]

Data Mining Assignment 4 (Association Mining)

Done by: Debarati Das (1PI13CS052)

Overview:

An association rule has two parts, an antecedent (if) and a consequent (then). An antecedent is an item found in the data. A consequent is an item that is found in combination with the antecedent.

Association rules are created by analyzing data for frequent if/then patterns and using the criteria support and confidence to identify the most important relationships. Support is an indication of how frequently the items appear in the database. Confidence indicates the number of times the if/then statements have been found to be true.

In data mining, association rules are useful for analyzing and predicting customer behavior. They play an important part in shopping basket data analysis, product clustering, catalog design and store layout.

In this assignment we are provided with the supermarket.arff which has information about products bought at the supermarket. We remove unnecessary attributes which do not make transactions clear and apply the apriori algorithm in order to efficiently obtain rules for the information we require.

Results:

Task 1
Use the Apriori method to generate association rules as follows:
Metric Type = Confidence, Minimum MetricValue = 0.9, Number of Rules = 10
Question 1: Explain the top 5 rules in simple English.

1. biscuits=t frozen foods=t pet foods=t milk-cream=t vegetables=t 516 ==> bread and cake=t 475 conf:(0.92)

This indicates that out of the 516 transactions involving items on the left hand side, 475 involved the items in the right hand side as well. So that would imply that in 92% of transactions where customers bought biscuits, frozen foods, pet foods, milk cream and vegetables, they also bought bread and cake.

2. baking needs=t biscuits=t milk-cream=t margarine=t fruit=t vegetables=t 505 ==> bread and cake=t 464 conf:(0.92)

This indicates that out of the 505 transactions involving items on the left hand side, 464 involved the items in the right hand side as well. So that would imply that in 92% of transactions where customers bought baking needs, biscuits, milk cream, margarine, fruits and vegetables, they also bought bread and cake.

And similarly for the remaining three rules as well.

3. biscuits=t frozen foods=t milk-cream=t margarine=t vegetables=t 585 ==> bread and cake=t 537 conf:(0.92)

4. biscuits=t canned vegetables=t frozen foods=t fruit=t vegetables=t 536 ==> bread and cake=t 492 conf:(0.92)

5. baking needs=t frozen foods=t milk-cream=t margarine=t fruit=t vegetables=t 517 ==> bread and cake=t 474 conf:(0.92)

Task 2
Generate, using the Apriori method, rules that will predict the attribute total based on other attributes (Hint: Use the car option with a minimum confidence of 0.8. Note that this attribute has two values, namely low and high).
Question 2: Explain the top 5 rules in simple English.

1. baking needs=t biscuits=t sauces-gravy-pkle=t frozen foods=t tissues-paper prd=t 574 ==> total=high 470 conf:(0.82)

This implies out of all the transactions that involved the items on the right hand side, the total was in the category high 82% of the time. So that means, When the customer bought baking needs, biscuits, sauces , frozen foods and tissue paper together, the total of his cost came in the high category (470/574) times.

Similar observation can be made in the cases of 2,3,4,5 listed below.

2. bread and cake=t biscuits=t sauces-gravy-pkle=t frozen foods=t tissues-paper prd=t 600 ==> total=high 491 conf:(0.82)

3. bread and cake=t baking needs=t sauces-gravy-pkle=t frozen foods=t tissues-paper prd=t 620 ==> total=high 506 conf:(0.82)

4. bread and cake=t baking needs=t biscuits=t sauces-gravy-pkle=t tissues-paper prd=t 595 ==> total=high 483 conf:(0.81)

5. bread and cake=t biscuits=t sauces-gravy-pkle=t tissues-paper prd=t vegetables=t 583 ==> total=high 469 conf:(0.8)

Question 3: In what way are the rules generated by this task useful?

1. These rules are helpful in analyzing customer behavior. For example, we can probably find out which set of items are favored by the customer given that there is an indication that they buy another set of items.If a customer tends to buy baby clothing supplies, we can rearrange the store such that the baby food supplies are also placed nearby.
2. What are the subsequent purchases after buying a certain item ? We can answer this question.
3. Since top 5 cases in Question 2 yielded HIGH total, we can also extrapolate that its likely the company was running with profit.