

Program – 10

AIM: To Implement Apriori algorithm in C++

Introduction and Theory

Association rules are if-then statements that help to show the probability of relationships between data items within large data sets in various types of databases. Association rule mining has a number of applications and is widely used to help discover sales correlations in transactional data or in medical data sets.

Association rule mining, at a basic level, involves the use of machine learning models to analyze data for patterns, or co-occurrence, in a database. It identifies frequent if-then associations, which are called association rules.

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

Association rules are created by searching 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 data. Confidence indicates the number of times the if-then statements are found true. A third metric, called lift, can be used to compare confidence with expected confidence.

Association rules are calculated from itemsets, which are made up of two or more items. If rules are built from analyzing all the possible itemsets, there could be so many rules that the rules hold little meaning. With that, association rules are typically created from rules well-represented in data.

Measure 1: Support. This says how popular an itemset is, as measured by the proportion of transactions in which an itemset appears.

The support of an itemset X, $\text{supp}(X)$ is the proportion of transaction in the database in which the item X appears. It signifies the popularity of an itemset.

$$\text{supp}(X) = \text{Number of transaction in which X appears} / \text{Total number of transactions.}$$

If the sales of a particular product (item) above a certain proportion have a meaningful effect on profits, that proportion can be considered as the support threshold. Furthermore, we can identify itemsets that have support values beyond this threshold as significant itemsets.

Measure 2: Confidence. This says how likely item Y is purchased when item X is purchased, expressed as $\{X \rightarrow Y\}$. This is measured by the proportion of transactions with item X, in which item Y also appears.

Confidence of a rule is defined as follows:

$$\text{conf}(X \rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X)}$$

It signifies the likelihood of item Y being purchased when item X is purchased.

Program – 10

This implies that for 75% of the transactions containing onion and potatoes, the rule is correct. It can also be interpreted as the conditional probability $P(Y|X)$, i.e., the probability of finding the itemset Y in transactions given the transaction already contains X .

It can give some important insights, but it also has a major drawback. It only takes into account the popularity of the itemset X and not the popularity of Y . If Y is equally popular as X then there will be a higher probability that a transaction containing X will also contain Y thus increasing the confidence. To overcome this drawback there is another measure called lift.

Measure 3: Lift. This says how likely item Y is purchased when item X is purchased, while controlling for how popular item Y is. A lift value greater than 1 means that item Y is likely to be bought if item X is bought, while a value less than 1 means that item Y is unlikely to be bought if item X is bought.

The lift of a rule is defined as:

$$lift(X \rightarrow Y) = \frac{supp(X \cup Y)}{supp(X) * supp(Y)}$$

This signifies the likelihood of the itemset Y being purchased when item X is purchased while taking into account the popularity of Y .

If the value of lift is greater than 1, it means that the itemset Y is likely to be bought with itemset X , while a value less than 1 implies that itemset Y is unlikely to be bought if the itemset X is bought.

```
Apriori( $T, \epsilon$ )
   $L_1 \leftarrow \{\text{large 1 - itemsets}\}$ 
   $k \leftarrow 2$ 
  while  $L_{k-1} \neq \emptyset$ 
     $C_k \leftarrow \{c = a \cup \{b\} \mid a \in L_{k-1} \wedge b \notin a, \{s \subseteq c \mid |s| = k-1\} \subseteq L_{k-1}\}$ 
    for transactions  $t \in T$ 
       $D_t \leftarrow \{c \in C_k \mid c \subseteq t\}$ 
      for candidates  $c \in D_t$ 
         $count[c] \leftarrow count[c] + 1$ 
     $L_k \leftarrow \{c \in C_k \mid count[c] \geq \epsilon\}$ 
     $k \leftarrow k + 1$ 
  return  $\bigcup_k L_k$ 
```

Applications:

Market Basket Analysis:

This is the most typical example of association mining. Data is collected using barcode scanners in most supermarkets. This database, known as the “market basket” database, consists of a large number of records on past transactions. A single record lists all the items bought by a customer in one sale. Knowing which groups are inclined towards which set of items gives these shops the freedom to adjust the store layout and the store catalogue to place the optimally concerning one another.

Medical Diagnosis:

Association rules in medical diagnosis can be useful for assisting physicians for curing patients. Diagnosis is not an easy process and has a scope of errors which may result in

Program – 10

unreliable end-results. Using relational association rule mining, we can identify the probability of the occurrence of an illness concerning various factors and symptoms. Further, using learning techniques, this interface can be extended by adding new symptoms and defining relationships between the new signs and the corresponding diseases.

Census Data:

Every government has tonnes of census data. This data can be used to plan efficient public services(education, health, transport) as well as help public businesses (for setting up new factories, shopping malls, and even marketing particular products). This application of association rule mining and data mining has immense potential in supporting sound public policy and bringing forth an efficient functioning of a democratic society.

Protein Sequence:

Proteins are sequences made up of twenty types of amino acids. Each protein bears a unique 3D structure which depends on the sequence of these amino acids. A slight change in the sequence can cause a change in structure which might change the functioning of the protein. This dependency of the protein functioning on its amino acid sequence has been a subject of great research. Earlier it was thought that these sequences are random, but now it's believed that they aren't. Knowledge and understanding of these association rules will come in extremely helpful during the synthesis of artificial proteins.

Code

```
1  #include <bits/stdc++.h>
2  #include <map>
3  using namespace std;
4
5  ifstream fin;
6  double minfre;
7  vector<set<string> > datatable;
8  set<string> products;
9  map<string, int> freq;
10
11 vector<string> wordsof(string str)
12 {
13     vector<string> tmpset;
14     string tmp = "";
15     int i = 0;
16     while (str[i])
17     {
18         if (isalnum(str[i]))
19             tmp += str[i];
20         else
21         {
22             if (tmp.size() > 0)
23                 tmpset.push_back(tmp);
24             tmp = "";
25         }
26         i++;
27     }
28     if (tmp.size() > 0)
29         tmpset.push_back(tmp);
```

Program – 10

```
30     return tmpset;
31 }
32
33 string combine(vector<string> &arr, int miss)
34 {
35     string str;
36     for (int i = 0; i < arr.size(); i++)
37         if (i != miss)
38             str += arr[i] + " ";
39     str = str.substr(0, str.size() - 1);
40     return str;
41 }
42 set<string> cloneit(set<string> &arr)
43 {
44     set<string> dup;
45     for (set<string>::iterator it = arr.begin(); it != arr.end();
46 it++)
47         dup.insert(*it);
48     return dup;
49 }
50 set<string> apriori_gen(set<string> &sets, int k)
51 {
52     set<string> set2;
53     for (set<string>::iterator it1 = sets.begin(); it1 !=
54 sets.end(); it1++)
55     {
56         set<string>::iterator it2 = it1;
57         it2++;
58         for (; it2 != sets.end(); it2++)
59         {
60             vector<string> v1 = wordsof(*it1);
61             vector<string> v2 = wordsof(*it2);
62
63             bool alleq = true;
64             for (int i = 0; i < k - 1 && alleq; i++)
65                 if (v1[i] != v2[i])
66                     alleq = false;
67             if (!alleq)
68                 continue;
69
70             v1.push_back(v2[k - 1]);
71             if (v1[v1.size() - 1] < v1[v1.size() - 2])
72                 swap(v1[v1.size() - 1], v1[v1.size() - 2]);
73
74             for (int i = 0; i < v1.size() && alleq; i++)
75             {
76                 string tmp = combine(v1, i);
77                 if (sets.find(tmp) == sets.end())
78                     alleq = false;
79             }
80
81             if (alleq)
82                 set2.insert(combine(v1, -1));
83         }
84     }
85     return set2;
86 }
```

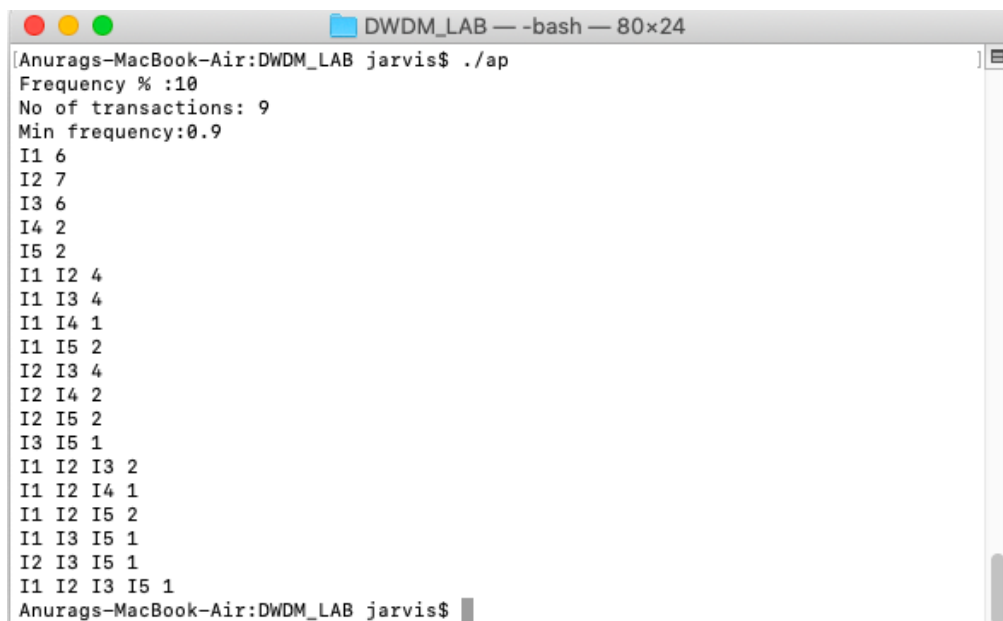
Program – 10

```
87 int main()
88 {
89     fin.open("apriori.in");
90     cout << "Frequency % :";
91     cin >> minfre;
92     string str;
93     while (!fin.eof())
94     {
95         getline(fin, str);
96         vector<string> arr = wordsof(str); //taking data from file ,
97         set<string> tmpset;
98         for (int i = 0; i < arr.size(); i++)
99             tmpset.insert(arr[i]);
100        datatable.push_back(tmpset);
101        for (set<string>::iterator it = tmpset.begin(); it !=
102 tmpset.end(); it++)
103        {
104            products.insert(*it);
105            freq[*it]++;
106        }
107    }
108    fin.close();
109    cout << "No of transactions: " << datatable.size() << endl;
110    minfre = minfre * datatable.size() / 100;
111    cout << "Min frequency:" << minfre << endl;
112    queue<set<string>::iterator> q;
113    for (set<string>::iterator it = products.begin(); it !=
114 products.end(); it++)
115        if (freq[*it] < minfre)
116            q.push(it);
117    while (q.size() > 0)
118    {
119        products.erase(*q.front());
120        q.pop();
121    }
122    for (set<string>::iterator it = products.begin(); it !=
123 products.end(); it++)
124        cout << *it << " " << freq[*it] << endl;
125    int i = 2;
126    set<string> prev = cloneit(products);
127    while (i)
128    {
129        set<string> cur = apriori_gen(prev, i - 1);
130        if (cur.size() < 1)
131            break;
132        for (set<string>::iterator it = cur.begin(); it !=
133 cur.end(); it++)
134        {
135
136            vector<string> arr = wordsof(*it);
137            int tot = 0;
138            for (int j = 0; j < datatable.size(); j++)
139            {
140                bool pres = true;
141                for (int k = 0; k < arr.size() && pres; k++)
142                    if (datatable[j].find(arr[k]) ==
143 datatable[j].end())
```

Program – 10

```
144         pres = false;
145         if (pres)
146             tot++;
147     }
148     if (tot >= minfre)
149         freq[*it] += tot;
150     else
151         q.push(it);
152 }
153 while (q.size() > 0)
154 {
155     cur.erase(*q.front());
156     q.pop();
157 }
158 for (set<string>::iterator it = cur.begin(); it !=
159 cur.end(); it++)
160     cout << *it << " " << freq[*it] << endl;
161
162     prev = cloneit(cur);
163     i++;
164 }
165 }
```

Results & Outputs



```
[Anurags-MacBook-Air:DWDM_LAB jarvis$ ./ap
Frequency % :10
No of transactions: 9
Min frequency:0.9
I1 6
I2 7
I3 6
I4 2
I5 2
I1 I2 4
I1 I3 4
I1 I4 1
I1 I5 2
I2 I3 4
I2 I4 2
I2 I5 2
I3 I5 1
I1 I2 I3 2
I1 I2 I4 1
I1 I2 I5 2
I1 I3 I5 1
I2 I3 I5 1
I1 I2 I3 I5 1
Anurags-MacBook-Air:DWDM_LAB jarvis$
```

Findings and Learnings:

1. We learned about association rule mining and familiarized ourselves with the test.arff dataset.
2. We implemented and tested Apriori in C++.