 ISBN:978-93-5254-640-4

Association Rule Mining using Apriori Technique

Nilambari Kale[1], Tripti Mahabare[2]

[1] Department of computer science, Pratibha College of Commerce & Computer Studies, Chinchwad, pune, India
nilambari.mate@gmail.com

[2] Department of computer science, Pratibha College of Commerce & Computer Studies, Chinchwad, pune, India
tru.neel@gmail.com

***Abstract*: Association rule mining is an important data mining task for many retail business applications. Data mining is defined as to discover hidden information from databases. There are different data mining tasks like classification, prediction, time series analysis, clustering, sequence discovery, association rule, and summarization. Association rule is very useful data mining task. Association rule mining is used in many businesses to discover customer buying patterns or to analyze association between many products though these products are not related to each other. In the last years a great number of algorithms have been proposed with the objective of solving the obstacles presented in the generation of association rules. This paper presents information about association rule mining, techniques used for association rule mining.**

Keywords: Data mining, association rule mining, apriori algorithm.

Introduction

Today's business strategies are different from business strategies which were used before 10 to 15 years. Now a day's every business operations are executed smartly. Business executives are continuously doing some analysis and depends on that they are taking necessary decisions to increase the profit. For such applications data warehousing and data mining is used. In simple terms data warehousing is the technique used to store huge amount of data and data mining is the technique used to extract useful or hidden information from the data.

Pattern mining algorithms can be applied on various types of data such as transaction databases, sequence databases, streams, strings, spatial data, graphs, etc.

Pattern mining algorithms can be designed to discover various types of patterns: subgraphs, associations, indirect associations, trends, periodic patterns, rules, lattices, sequential patterns, etc.

But what is an interesting pattern? There are several definitions. For example, some researchers define an interesting pattern as a pattern that appears *frequently* in a database. Other researchers want to discover *rare patterns*, patterns with a high *confidence*, the top patterns, etc.

Association rule mining is a procedure which is meant to find frequent patterns, correlations, associations, or causal structures from data sets found in various kinds of databases such as relational databases, transactional databases, and other forms of data repositories.

Given a set of transactions, association rule mining aims to find the rules which enable us to predict the occurrence of a specific item based on the occurrences of the other items in the transaction. So in a given transaction with multiple items, it tries to find the rules that govern how or why such items are often bought together. For example, peanut butter and jelly are often bought together because a lot of people like to make sandwiches.

Also surprisingly, diapers and beer are bought together because, as it turns out, that dads are often tasked to do the shopping while the moms are left with the baby.

Association Rule mining is two step process:

1. Find all frequent itemsets from given dataset:
2. Find strong association rule from frequent itemsets.

1. Find all frequent itemsets from given dataset:

Let I={I1,I2…….Im} be an itemset. Let D, the task relevant data, be a set of transactions where each transactin T is an itemset such that T □ I.

Let A be the set of items.

An association rule in an implication of the form A⇒B where A □ I, B □ I , A ≠ ∅, B ≠ ∅ and A∩B=∅.

The rule holds two measures. First is support s, where s is the percentage of transactions that demonstrate the rule.

Second the confidence c, where c is the conditional probability that, given A present in a transition , B will also be present.

Support (A B)=P(A B)

Confidence (A B)=P(B/A)

Rule that satisfies minimum support threshold (min_sup) and minimum confidence threshold (min_conf) is called strong.

A set of items is called itemset. An itemset that contains k items is called k – itemset.

Many algorithms are available to fine frequent itemsets like Apriori algorithm, FP-tree algorithm. The popular and easiest way to use apriori algorithm.

Apriori algorithm works as follows:

APRIOIRI is an efficient algorithm to find association rules (or, actually, **frequent itemsets**). The apriori technique is used for "generating large itemsets." Out of all candidate (k)-itemsets, generate all candidate (k+1)-itemsets.

Apriori Pseudocode

Apriori (T, ε)

$L_1 \leftarrow \{$*large 1-itemsets that appear in more than* ε *transactions* $\}$
$k \leftarrow 2$
while $L_{k-1} \neq \varnothing$
$C_k \leftarrow$*Generate* (L_{k-1})
for transactions $t \in T$
$C_t \leftarrow$Subset (C_k, t)
for candidates $c \in C_t$
$\text{count}[c] \leftarrow \text{count}[c] + 1$
$L_k \leftarrow \{c \in C_k | \text{ count}[c] \geq \varepsilon\}$
$k \leftarrow k + 1$
$\bigcup_k L_k$
return

Example:

Given Transaction Table:

Transaction No.	Items
T1	1, 2, 3, 4, 5, 6
T2	7, 2, 3, 4, 5, 6
T3	1, 8, 4, 5
T4	1, 9, 0, 4, 6
T5	0, 2, 2, 4, 5

and Support or threshold is 3.

Solution**:**

Step 1: Count the occurrence of each item.

Item	Occurrence / Frequency
1	3
2	3
3	2
4	5
5	4
6	3
7	1
8	1
9	2
0	2

Step 2: Remember, the algorithm says, an item is considered to be frequent if it's bought more than the Support/Threshold i.e. 3. Therefore, below is the list of Frequent Singletons.

ISBN:978-93-5254-640-4

Item	Occurrence / Frequency
1	3
2	3
4	5
5	4
6	3

Step 3: We start making pairs out of the frequent itemsets we got in the above step.

ItemPairs
12
14
15
16
24
25
26
45
46
56

Step 4: After getting the frequent Item Pairs, we start counting the occurrence of these pairs in the Transaction Set.

ItemPairs	Occurrence / Frequency
12	1
14	2
15	2
16	1
24	3
25	3
26	2
45	4
46	3
56	2

Step 5: Now again, follow the Golden Rule, and discard non-frequent pairs.

ItemPairs	Occurrence / Frequency
14	3
24	3
25	3
45	4
46	3

Now we have a table with pair of frequent items. Suppose we want to find frequent triplets. We the above table and make all the possible combinations.

Step 6: Make combinations of triples using the frequent Item pairs.

To make triples, the rule is: IF 12 and 13 are frequent, and then the triple would be **123**. Similarly, if 24 and 26 then triple would be **246**.

So, using the above logic and our Frequent ItemPairs table, we get the below triples:

ItemTriples
245
456

Step 7: Get the count of the above triples (Candidates).

ItemTriples	Occurrence / Frequency
245	3
456	2

After, this, if we can find quartets, then we find those and count their occurrence/frequency.

If we had 123, 124, 134, 135, 234 and we wanted to generate a quartet then it would be 1234 and 1345. And after finding quartet we would have again got their count of occurrence

/frequency and repeated the same also, until the Frequent ItemSet is null.

Thus, the frequent ItemSets are:

Frequent Itemsets of Size 1: 1, 2, 4, 5, 6
Frequent Itemsets of Size 2: 14, 24, 25, 45, 46
Frequent Itemsets of Size 3: 245

2. *Find strong association rule from frequent itemsets:*

Once the frequent itemsets from the are found then generate strong association rules from them. This can be done as follows:

a. For each itemset l, generate all nonempty subsets of l.
b. For every non empty subset s of l, output the rule "s=>(l-s)" if support_count(l)/ support_count(s)>=min_conf, where min_conf is minimum confidence threshold.

Let's take above example.
We have got Frequent Itemsets of Size 3is 245.
So, X = {2,4,5}
The nonempty subsets of X are {2, 4}, {2, 5}, {4, 5}, {2}, {4} and {5}.
The resulting association rules are shown below, each listed with its confedence:
Rule1: {2, 4}=>{5}, confidence=3/3=100%
Rule2: {2, 5}=>{4}, confidence=3/3=100%
Rule3: {4, 5}=>{2}, confidence=3/4=75%
Rule4: {2}=>{4, 5}, confidence=3/3=100%
Rule5: {4}=>{2, 5}, confidence=3/5=60%
Rule6: {5}=>{2,4}, confidence=3/4=75%
If min_conf is, say 75% then except Rule5 all other rules are the output.

The main applications of association rule mining:

- Basket data analysis - is to analyze the association of purchased items in a single basket or single purchase as per the examples given above.
- Cross marketing - is to work with other businesses that complement your own, not competitors. For example, vehicle dealerships and manufacturers have cross marketing campaigns with oil and gas companies for obvious reasons.
- Catalog design - the selection of items in a business' catalog are often designed to complement each other so that buying one item will lead to buying of another. So these items are often complements or very related.

Decisions taken using these association rules always beneficial for organization to increase the profit e.g for the rule Rule1: {2, 4}=>{5} organization may give offer for customers like "Buy Item 2 & 4 Get 5 Free".

Conclusion

Association rule mining is very important and popular topic used in businesses to increase the profit. As we have seen in above example to retrieve strong association rule from given dataset is very easy using apriori algorithm. Unlike conventional classification rules, association rules can contain more than one conjunct in the right side of the rule. Accuracy is very important for these type of applications which somehow depends on the accuracy of dataset under consideration. Apriori algorithm suffers from two nontrivial costs: 1) It generate huge no. of candidate sets. 2) It repeatedly scans databases.

References

1. "Association Rules: Problems, solutions and new applications" by María N. Moreno, Saddys Segrera and Vivian F. López
2. "Data Mining concepts & Techniques" , 3rd edition, Jiawei Han|Micheline Kamber|Jian Pei