DATA MINING AND DISCOVERY

7PAM2005-0901-2022

Coursework – Report Assignment

Topic 1: Frequent Itemset Generation and Rule Generation

Name: Hetanshi Kachhiya Patel

Student ID: 21067973

Group Number: 42

Introduction: What is Frequent Itemset Generation and Rule Generation?

Itemset refers to a collection of one or more items. For example: {Notebook, Pen, Pencil}

The number of itemset that contains particular itemset is called **Support Count.**

Support Fraction refers to the fraction of transactions in which an itemset occurs.

$$\text{i.e., } s(X) = \sigma(X)/N$$

Where, σ (X) is support count and N is number of transactions.

An itemset is called frequent if s(X) is greater than some user-defined threshold, *minsup.*

Association Rule: This is an implication expression of the form $X \rightarrow Y$, where $X \cap Y = \emptyset$. The strength of it will be measured in terms of its **support** and **confidence**.

$$\text{Support, } s(X \rightarrow Y) = \sigma(X \cup Y)/N;$$

$$\text{Confidence, } c(X \rightarrow Y) = \sigma(X \cup Y)/\sigma(X);$$

The objective of **Frequent Itemset Generation** is to find all the itemsets that satisfy the *minsup* threshold.

Rule Generation is used to extract all the high confidence rules from the frequent itemsets.

The Apriori Principle:

It uses frequent itemsets to generate association rules. Each frequent k-itemset, can produce up to $2^k - 2$ association rules. If an itemset is frequent, then all its subsets must be frequent. If the itemset is infrequent, then its subsets must be infrequent. So, all subsets need to be pruned immediately.

Apriori algorithm in frequent itemset generation has 2 main characteristics which are first **level-wise algorithm** and second, it employs **generate-and-test** strategy.

This algorithm uses a level-wise approach for generating association rules, where each level corresponds to the number of items that belong to the rule consequent. In Rule generation we do Confidence-Based Pruning.

All rule with items in set of items having

- Support$\geq$*minsup* threshold
- Confidence $\geq$*minconf* threshold

There is dataset of Grocery of store on which I have applied the Apriori Algorithm to extract some useful information from the data. I have chosen this type of data because we can find some similarity between the items bought together and frequently.

	support	itemsets	length
0	0.315789	(BISCUIT)	1
1	0.631579	(BREAD)	1
2	0.421053	(COFFEE)	1
3	0.315789	(CORNFLAKES)	1
4	0.315789	(SUGER)	1
5	0.368421	(TEA)	1

figure 1

	support	itemsets	length
1	0.631579	(BREAD)	1

figure 2

Here is the output of my code. It shows the support, itemsets and length. I first transformed the data using TransactionEncoder and then I applied the apriori algorithm on it with the support count greater than 0.3 and we got the results which is shown in figure 1. It shows there is no support in which the things bought together but many people bought Bread frequently and which has support 0.631579 and which is followed by Coffee and so on.

The second figure shows the results of the rules which I have applied to extract the meaningful information from the last result (figure 1). I have extracted the frequent itemset which has the support greater than 0.5 and length equal to 1.

Then I tried to extract the frequent itemsets in which Bread and Coffee bought together but I got empty result which shows these things are not frequently bought together.

References – 1. https://ebookcentral.proquest.com/lib/herts/detail.action?docID=5720020

The book - Introduction to Data Mining EBook: Global Edition

2. http://rasbt.github.io/mlxtend/user_guide/frequent_patterns/apriori/

Repository Link - https://github.com/hetanshipatel/Data-mining-Coursework