Multilevel Association Rules in Data Mining

Abhishek Kajal
Deptt. of Comp. Sc.
VDIET, Julana, Jind
abhishekkajal@gmail.com

Isha Kajal
Deptt. of Comp. Sc.
CDLU, Sirsa
dearishakajal@gmail.com

Data is the basic building block of any organization. Be it an individual or an organization of any type, it is surrounded by huge flow of quantitative or qualitative data. Data are the patterns which are used to develop or enhance information or knowledge. All the organizations big or small has bulk of data which needs to be stored or retrieved systematically to form information. The repository of data is known as Database. With the advancement in computer science, the database has taken many shapes. According to the applications, starting from the traditional file system to hierarchical, Network, Relational, Object Oriented, Associative, now it has reached to Data Warehouses and Data Marts etc. But every piece of data stored in these databases may not be useful for the organization. Organizations need to filter the useful data from the bulk of data which can be used for decision making, reporting or analysis. These useful patterns or pieces of data are known as interesting patterns.

Organizations are more interested in the interesting data rather than the bulk of data. So they need a systematic and scientific approach to extract meaningful data out of heaps of the data and to find out the relations among these patterns. **Association Rule mining** is the scientific technique to dig out interesting and frequent patterns from the transactional, spatial, temporal or other databases and to set associations , relations or correlations among those patterns (also known as item sets) in order to discover knowledge or to frame information. Associations rules can be applied in various fields like network management, Basket data analysis, catalog design, clustering, classification, marketing etc. Association rules establishes the relationship between different **variables** to analyse the present situation. For e.g to find the relationship between the various items sold at a shopping mall ,the association rule can be applied on the huge amount of data recorded by the Shopping mall. For e.g the rule { Computer , Printer} -→ {UPS} found in the sales data of a mall would indicate that if a customer buys Computer and Printer together, he or she would definitely also buy UPS. This information can be used making the decision regarding keeping the stock of the products as well as to analyse the customer buying habbits and promotional activities for future.

Association rule works on the database of transactions where every transaction contain list of itemset(patterns). Measures of the rule are Support and Confidence. **Support** of rule is proportion of transaction in the data set that contain the itemset to the total number of transactions. The **Confidence** of a rule is ratio of total number of transactions with all the items to the number of transaction with the A item set.For e.g if Dataset *T* is given the an itemset *A* has number of occurences in it. An association rule is the relation ship between two itemsets *A* and *B* . such as

A⇨B

means when A occurs B also occurs .

To illustrate and understand the basic terms we consider a small database of 6 transacions and 3 items.The rule is

{computer,printer} ⇨ {UPS}

Transaction Id	**Itemset**
1	{computer,printer }
2	{computer,printer ,Ups}
3	{computer}
4	{computer,printer ,Ups}
5	{Ups}
6	{computer,printer ,Ups}

This implies that if customer buy Computer and Printer , he tend to buy UPS also . Out of 6 transaction 3 transactions support this rule .In 3 rcords all the three items are brought together.

So the Support of rule denoted as ***Supp(A)*** is proportion of transaction in the data set that contain the itemset to the total number of transactions. In the above example , the itemset {Computer,Printer , UPS} has a support of 3/ 6 = 0.5 since it occurs in 50% of all transactions (3 out of 6 transactions).

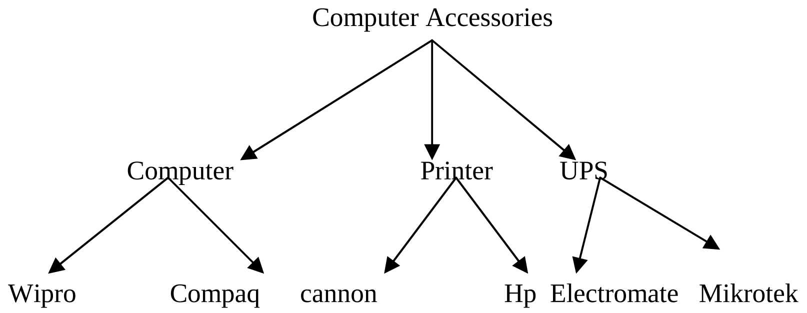
The Confidence of a rule (denoted as conf(A ⇨B)=Ratio of total number of transaction with all the items to the number of transaction with the A item set . for e.g Computer and Printer are purchased 4 times and out of 4 transactions UPS is purchased three times with Computer and Printer i.e A so the

conf(A,B)= ¾ =.75 i.e 75%.

So the association rule is the technique to set the relation between item sets to draw important conclusions. It set the minimum support and confidence threshold and evaluate the frequent itemset and then use the evaluated itemset to frame desired results. Asoociation rule mining can be broadly classified into following categories :

- Boolean or quantitative associations
- Single dimension or multidimensional associations
- Single level or multilevel associations

Above mentioned example is the example of single level aoociation. Our previous work has been focused on mining association rules at a single concept level. But there are applications which need to find association rules at multiple concept level. For eg. Organisation need to find how many customers buy mikrotek Ups with Wipro computers and cannon printers. In multilevel association rules , organisation want to evaluate association between item sets at different level of hierarchies because items often form hierarchies . For e.g



Rules regarding item sets at suitable levels could be relatively functional.it can help organizations to make promotional strategies and help enhancing the sales and setting the future plans. A basic approach to multi level association rule mining is top-down progressive deepening approach. In this approach first find out the frequent item set at the highest level as we do in single level association rule mining , then move to the next level (lower level) to find out the frequent item set at that level of hierarchy and continue moving to the lowest level until no more frequent item set can be found. For eg.

- First find out the freqent item set and mine strong association rule :

  {Computer, Printer} → {UPS} { 20%,60%}

- Then find out the lower level rules :

  { wipro computer, cannon printer} → {UPS} { 6%,50%}

Multiple level association rule mining can work with two types of support- Uniform and Reduced.

1. Uniform Support : In this approach same minimum support threshold is used at every level of hierarchy. There is no need to evaluate itemsets containing items whose ancestors do not have minimum support. The minimum support threshold has to be appropriate. If miniumum support threshold is too high the we can loose lower level associations and if too low then we can end up in generating too many uninteresting high level association rules. For e.g

At Level 1	Computer , Printer
Minimum supp 5%	[support – 10%]
At Level 2	Wipro Computer , Cannon printer
Minimum support 3%	[support 7%]
	Wipro Computer , HP Printer
	[support 3%]

2. Reduced Support : In this approach reduced minimum support is used at lower levels
   a. There are following search strategies:
      i. Independently level by level
      ii. Filtering across the levels by k-itemset
      iii. Filtering across the levels by single item
      iv. Controlled level filtering by single item.

1. Independently Level by Level : This technique is basically based on full breadth search. It is not required to know in advance frequent itemset for pruning. Each node is evaluated at each level , regardless of whether or not its ancestor node is found to be frequent.

2. Filtering across the levels by single item set : In this technique descendants are only checked only if ancestor is found to be frequent. So item at ith level will be only checked if and only if item at i-1th level is frequent.For eg wipro computer is not examined if computer is not frequent.
3. Filtering across the levels by k-itemset : In this approach a k item set at ith level is only examined if and only if its ancestor k item set at i-1 th level is frequent . For eg Wipro Computer , cannon printer will be examined only if Computer and printer are frequent.

Checking for redundancy :

There can be redundancy in some of the rules due to its ancestors associations between items For eg.

Rule 1 : Computer, Printer -→ Ups [support =10% , Confidence =70%]

Rule 2: Wipro Computer , Cannon printer-→ Mikrotek UPS [support =3% , confidence =70%]

In this eg first rule is an ancestor of the second rule . A rule is redundant if its support is close to the expected value, based on the rule’s ancestor.

References:

[1] R. Agarwal, C. Aggarwal, and V. V. V. Prasad. A tree projection algorithm for generation of frequent itemsets. In Journal of Parallel and Distributed Computing (Special Issue on High Performance Data Mining), 2000.

[2] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. SIGMOD'93, 207-216, Washington, D.C.

[3] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. VLDB'94 487-499, Santiago, Chile.

[4] R. Agrawal and R. Srikant. Mining sequential patterns. ICDE'95, 3-14, Taipei, Taiwan.

[5] R. J. Bayardo. Efficiently mining long patterns from databases. SIGMOD'98, 85-93, Seattle, Washington.