International Journal of Information Technology and Knowledge Management
January-June 2012, Volume 5, No. 1, pp. 181-183

COMPREHENSIVE STUDY OF ASSOCIATION RULES IN DATA MINING

Amandeep Mehta and Vivek Chandra

ABSTRACT: 21st century has seen enormous and unprecedented technological developments especially in the fields of Electronics, IT and Computer Science. These technological developments have affected all big or small organizations, directly or indirectly. Any organization, be it business, financial, educational, social or health care, all of them have millions of data or list of values to be stored, maintained, manipulated and used in the form of Databases for future predictions which affect their business transactions and trends. This bulk of data stored in database may not be useful to the organization. Data mining is the technique to extract meaningful or interesting data from the bulk of data. Data mining has many tasks but this paper stresses on Association rule learning which establishes the relationship and correlations among the various attributes stored in database. Association Rules are applicable in doing the market analysis, analysing the Customer trends, social network activities, sales trends, buying habbits, banking transactions, fraud detection, health transactions etc. Association rules establishes the relationship between different variables to analyse the present situation. In this paper we have presented the basic fundamentals about association rule mining. There are many algorithms existing for association rule mining but we have surveyed the working and limitations of Apriori algorithm.

1. INTRODUCTION

21st century has seen enormous and unprecedented technological developments especially in the fields of Electronics, IT and Computer Science. These technological developments have affected all big or small organizations, directly or indirectly. Any organization, be it a business, financial, educational, social or health care, all of them have millions of data or list of values to be stored, maintained, manipulated and used. For e.g. data about Inventory, employees, sales, purchases, customers, incomes, expenses etc. Data may be in the shape of simple structure or complex structure-valued data, such as set and list valued data and data with nested structures. Data may be stored using traditional file system or it may be stored in the form of small database or big Data Warehouse or Data Mart. All this data help organizations to set market strategies, to analyze customer's behaviors, to analyze sales trends, to predict future trends etc. Data may be in any form but it is not useful unless it is changed to meaningful information as these set of values or data sets help the organizations in analyzing their present and predict their future. For e.g. extensive customer database maintained by a company give it a chance to know their existing customers and by analyzing their trends they can very well know their perspective customers also. As Organizations collect data from number of homogeneous or heterogeneous sources so the data gets stored in various shapes and formats. To be able to make this data useful, there is a need to extract meaningful information hidden in data. This requirement of organizations have given birth to a new interdisciplinary field of Computer science i.e. Data Mining. As recorded in Wikipedia, "Data mining[1] is the process of discovering new patterns from large data sets involving methods from statistics and artificial intelligence and database management." Data mining helps us to extract meaning ful information hidden in bulk of data. It allows us to ask the questions: "What interesting information is contained in or may be learned from the database?" Though through simple database queries, we could reach the information we expect. But there may be some information that cannot be reached by simple database queries. This hidden information is what we reveal when implementing data mining techniques. For this reason data mining is also known as "Knowledge Discovery". As mentioned by Jiawei Han and Micheline Kamber in their book "Data Mining–Concepts and Techniques", The abundance of data, coupled with the need for powerful data analysis tools, has been described as a data rich but information poor situation". Because they have lots of data stored but do not have the techniques to extract appropriate meaning according to the need of users. Extracting knowledge manually from data is time consuming, complex and full of errors methods. So the Data mining tools provide us the fast techniques to extract hidden information and draw conclusions from it.

Data Mining has various tasks like Anomoly detection, Association rule learning, Clustering, Classification, Regression, Summarization and lots of research is being done in all thses tasks. But for our paper we are stressing on Association Rule Learning.

[1] Assistant Prof., Dept. of Computer Sc. S.D College, Ambala Cantt, E-mail: mehtaamandeep@yahoo.com.
[2] Head of Department (IT), M.P East Jone Dis. Com, Jabalpur (MP).

2. ASSOCIATION RULES LEARNING

In transactional database data is stored in the form of set of attributes. Association Rule learning also known as Dependency modelling is the technique of finding out relations or associations between these attributes. These relationships among attributes helps organisations to draw various conclusions. For e.g. a Shopping mall may wish to analyse the customer purchasing habbits so as to decide about the group of items to be sold in the mall. Using association rule learning, the mall can decide which products are frequently bought together and use this information for marketing purposes. This is sometimes known as market basket analysis. 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 {Shoes, Socks} ⟶ {Shoe Polish} found in the sales data of a mall would indicate that if a customer buys shoes and socks together, he or she would definitely also buy shoe polish. 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.

The problem of mining association rules was first introduced in Agrawal et al 1993. Association Rules are applicable in doing the market analysis, analysing the Customer trends, social network activities, sales trends, buying habbits, banking transactions, fraud detection, health transactions etc. Association rules establishes the relationship between different variables to analyse the present situation.

2.1. Basic Terms Used in Association Rule Learning

Itemset: set of items (single or multi-valued) invloved in analysis. e.g {shoes, socks, shoe polish} is a item set.

Transaction: Records showing the action (sale, purchase or other) on set of items. Each Record can have Transaction ID associated with it to uniquely identifying the transaction.

Data Set: Set of records.

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

$$A \Rightarrow B$$

means when A occurs B also occurs.

2.2. Measures of Association Rules

Support: Support of rule is proportion of transaction in the data set that contain the itemset to the total number of transactions.

Confidence: 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.

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

$$\{\text{shoe, Socks}\} \Rightarrow \{\text{polish}\}$$

Transaction Id	Itemset
1.	{shoe, socks}
2.	{shoe, socks, polish}
3.	{shoe}
4.	{shoe, socks, polish}
5.	{polish}
6.	{shoe, socks, polish}

This implies that if customer buy Shoes and Socks, he tend to buy Polish 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 {shoe, socks, polish} 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 ($A \Rightarrow B$) = Ratio of total number of transaction with all the items to the number of transaction with the A item set. for e.g shoe and socks are purchased 4 times and out of 4 transactions polish is purchased three times with shoe and socks i.e A so the conf (A, B) = ¾ = .75 i.e 75%.

2.3. Working of Association Rules

Association rule mining is to search out association rules that fulfils the predetermined minimum support and confidence from a given transactional database. The problem is usually divided into two phases.

- First phase is to search those itemsets whose occurrences exceed a predefined threshold in the database; those itemsets are called frequent itemsets.
- The second phase is to generate association rules from those frequent itemsets with the constraints of minimal confidence.

Various efficient algoriths are used to generate association rules like:

1. Apriori
2. Eclat
3. FP-growth algorithm
4. GUHA procedure ASSOC etc.

For our paper we will be discussing Apriori Algorithm.

3. APRIORI

Apriori[2] algorithm is very efficiently used for framing association rules. It is basically used for the databases which contain huge volume of transactions. It is based on bottom up approach. It is used for mining frequent itemset for association rules. algorith is performed in following steps:

1. Find out itemset which occur frequently and has minimum support represented by ith–itemset.
2. For cutting down the itemset any subset of itemset which occur frequently is also treated as frequent itemset and any subset of itemset which does not occur frequently is also considered as non-frequent.
3. To search L_k, a set of candidate k-itemsets is generated by joining L_{k-1} with itself.
4. Use the frequent itemsets to generate association rules.

Pseudocode: (by Agrawal et al at IBM Almaden Research Centre)

Pass 1

1. Generate the candidate itemsets in C_1.
2. Save the frequent itemsets in L_1.

Pass k

1. Generate the candidate itemsets in C_k from the frequent itemsets in L_{k-1}.
   (a) Join L_{k-1} p with L_{k-1} q, as follows:

   insert into C_k

   select $p.item_1$, $p.item_2$, ... , $p.item_{k-1}$, $q.item_{k-1}$

   from L_{k-1} p, L_{k-1}q

   where $p.item_1 = q.item_1$, ... $p.item_{k-2} = q.item_{k-2}$, $p.item_{k-1} < q.item_{k-1}$.

   (b) Generate all (k – 1)-subsets from the candidate itemsets in C_k.

   (c) Prune all candidate itemsets from C_k where some (k – 1)-subset of the candidate itemset is not in the frequent itemset L_{k-1}.
2. Scan the transaction database to determine the support for each candidate itemset in C_k.
3. Save the frequent itemsets in L_k.

Exampe: Assume the user-specified minimum support is 50%

- Given: The transaction database shown below:

TID	A	B	C	D	E	F
T_1	1	0	1	1	0	0
T_2	0	1	0	1	0	0
T_3	1	1	1	0	1	0
T_4	0	1	0	1	0	1

- The candidate itemsets in C_2 are shown below:

Itemset X	supp (X)
{A, B}	25%
{A, C}	50%
{A, D}	25%
{B, C}	25%
{B, D}	50%
{C, D}	25%

- The frequent itemsets in L_2 are shown below.

Itemset X	supp (X)
{A, C}	50%
{B, D}	50%

As shown in the e.g Apriori algorith scans the database and generate 1-itemset frequent patteren which satisfies minimum support threshold set by data miner denoted by L_1, then perform L_1 join L_1 to generate the 2-itemset frequent pattren which satisfies minimum support denoted by L_2. Then perform L_2 join L_2 to generate 3-itemset frequent pattren frequent patren and perform prune step to reduce the size. Subset of Frequent pattren is considered frequent and subset of non-frequent pattren is termed as non-frequent. It keeps on Performing k-itemset, join and prune process till C_k = empty. Then the confidence of found frequent pattren is compared with minimum confidence threshold to frame strong association rules.

4. LIMITATION OF APRIORI

1. If minimum threshold is not properly set, the results will be incorrect.
2. Uniform threshold is used which is not practical.
3. Lots of Iterations are involved.
4. If lots of data contains non-frequent itemset then lots of time is wasted in generating candidate key.
5. High time is taken to scan.

REFERENCES

[1] Wikipedia

[2] R. Agarwal and R. Srikant., "Fast Algorithms for Mining Association Rules in Large Databases", Proceedings of the 20th International Conference on VLDB, pp. 487-499, Santiago, Chile, September 1994.

[3] "Data Mining –Concepts and Techniques", by Jiawei Han and Micheline Kamber.