International Journal of Electronics Communication and Computer Engineering
Volume 3, Issue 4, ISSN (Online) 2249 –071X, ISSN (Print): 2278 – 4209

# An Overview of Association Rule Mining

**K. Lavanya**
M.Tech., Pydah College of Engineering and Technology, Kakinada, A.P., India

**A. Kamala Priya**
Assistant Professor, Dept. of CSE, Pydah college of Engineering and Technology, Kakinada, A.P., India

**P. Suresh Babu**
Assistant Professor, Kaushik College of Engineering, Visakhapatnam, A.P., India

***Abstract* — Data mining is a process concerned with uncovering patterns, associations, anomalies and statistically significant structures in data. Association rule mining is a data mining task that discovers associations among items in a transactional database. Association rules have been extensively studied in the literature for their usefulness in many application domains such as recommender systems, diagnosis decisions support, telecommunication, intrusion detection, etc. Efficient discovery of such rules has been a major focus in the data mining research. This paper presents an overview of association rule mining- positive and negative association rules. Research in association rules mining has initially concentrated in solving the obvious problem of finding positive association rules; that is rules among items that exist in the stored transactions. It was only several years after that the possibility of finding also negative association rules became especially appealing and was investigated.**

***Keywords* — Apriori algorithm, Association rules, Data mining, Itemset.**

## I. INTRODUCTION

Data mining [1] is a process concerned with uncovering patterns, associations, anomalies and statistically significant structures in data. It typically refers to the case where the data is too large or too complex to allow either a manual analysis or analysis by means of simple queries. Data mining consists of two main steps, data pre-processing, during which relevant high-level features or attributes are extracted from the low level data, and pattern recognition, in which a pattern in the data is recognized using these features (Figure 1). Pre-processing the data is often a time-consuming, yet critical, first step. To ensure the success of the data-mining process, it is important that the features extracted from the data are relevant to the problem and representative of the data.

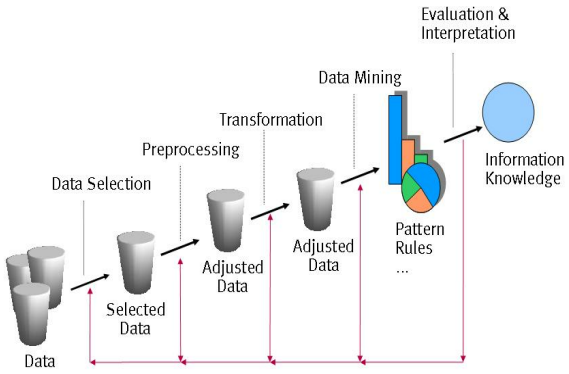


Fig.1 : Data mining process

Depending on the type of data being mined, the pre-processing step may consist of several sub-tasks. If the raw data is very large, we could use sampling and work with fewer instances, or use multi-resolution techniques and work with data at a coarser resolution. Next, noise in the data is removed to the extent possible, and relevant features are extracted. In some cases, where data from different sources or sensors are available, data fusion may be required to allow the miner to exploit all the data available for a problem. At the end of this first step, we have a feature vector for each data instance. Depending on the problem and the data, we may need to reduce the number of features using feature selection or dimension reduction techniques such as principal component analysis (PCA) or its nonlinear versions. After this pre-processing, the data is ready for the detection of patterns through the use of algorithms such as classification, clustering, regression, etc. These patterns are then displayed to the user for validation. Data mining is an iterative and interactive process. The output of any step, or feedback from the domain experts, could result in an iterative refinement of any, or all, of the sub-tasks.

There are four major tasks in Data Mining:

- Classification
- Clustering
- Association Rule Discovery
- Sequential Pattern Discovery

Association rule mining[2][3] is a data mining task that discovers associations among items in a transactional database. Association rules have been extensively studied in the literature for their usefulness in many application domains such as recommender systems, diagnosis decisions support, telecommunication, intrusion detection, etc. Efficient discovery of such rules has been a major focus in the data mining research. This paper presents an overview of association rule mining. This paper is organized as follows: Section 2 presents positive association rules. Section 3 presents negative association rules. Section 4 presents how Apriori algorithm is used to generate association rules. Section 5 presents an algorithm for discovering negative association rules. And finally section 6 presents the conclusion.

## II. ASSOCIATION RULE MINING-DEFINITION (POSITIVE ASSOCIATION RULES)

A formal definition of the positive association rule[4][5] is as follows: Let I={i1,i2,…,im} be a set of literals called items, with m considered to be the dimensionality of the problem. Let D be a set of transactions, where each transaction T is a variable length set of items such that $T \subseteq I$ . Since the quantities of items bought in a transaction are not taken into account, each item is a binary variable

International Journal of Electronics Communication and Computer Engineering
Volume 3, Issue 4, ISSN (Online) 2249 –071X, ISSN (Print): 2278 – 4209

representing if an item has been bought or not. Each transaction is associated with a unique identifier, called its TID. Consider X to be a set of items in I. A transaction T is said to contain X if and only if $X \subseteq T$. Each itemset has an associated measure of statistical significance called support, which is defined as the fraction of transactions in D containing the specific itemset.

An association rule is an implication of the form $X \rightarrow Y$, where $X,Y \subset I$, and $X \cap Y = \emptyset$. X is called the antecedent and Y is called the consequent of the rule. The rule $X \rightarrow Y$ holds in the transaction set D with confidence c if c% of transactions in D that contain X also contain Y. The rule $X \rightarrow Y$ has support s in the transaction set D if s% of transactions in D contain $X \cup Y$. In other words, the support of the rule is the probability that X and Y hold together among all possible presented cases. The confidence of a rule on the other hand is the conditional probability that the consequent Y is true under the condition of the antecedent X. The problem of discovering all association rules from a set of transactions D consists of generating the rules that have a support and confidence greater than a given threshold. The association rules mining task can be decomposed into two main steps:

1. Generate all frequent itemsets that satisfy *minsup*.
2. Generate all association rules that satisfy *minconf* using the frequent itemsets as input.

## III. Negative Association Rules

While the positive association rules handle the existence of items within the transactions, the negative ones deal also with the absence of items. For example a positive association rule would handle all transactions containing beers and chips and would generate all corresponding association rules. An approach dealing with negative association rules on the other hand would also consider rules such as: those that buy beer buy also chips, but not dry fruits (beer $\rightarrow$ chips $\wedge \neg$ dry fruits), or those that buy beer but not wine buy also chips (beer $\wedge \neg$ wine $\rightarrow$ chips). The measures used for determining the strength of a correlation and for pruning the insignificant ones are again the support and confidence metrics.

A generalized negative association rule[4][5] can include various negated and positive items (i.e., items existing and items absent from transactions) either in its antecedent or in its consequent. An example of such a rule would be: $A \wedge \neg C \wedge \neg F \wedge W \rightarrow B \wedge \neg D$. However the obvious insurmountable obstacle created by such contemplation is the number of possible itemsets that would have to be generated and counted as well as the number of rules created, since apart from the existing items we would have to consider all possible absent items and their combinations.

## IV. Apriori Algorithm

One of the most popular data mining approaches is to find frequent itemsets from a transaction dataset and derive association rules. Apriori[6] is the most popular and effective algorithm to find all the frequent itemsets in dataset. Apriori algorithm is given in Algorithm 1. The first pass simply counts item occurrences to determine the frequent 1-itemsets. A subsequent pass consists of two phases. First, the frequent itemsets $F_{k-1}$ found in the $(k-1)$-th pass are used to generate the candidate itemsets $C_k$ using the apriori-gen function. Next, the database is scanned and the support of candidates in $C_k$ is counted. The subset function is used for this counting.

The apriori-gen function takes as argument $F_{k-1}$, the set of all frequent $(k-1)$- itemsets, and returns a superset of the set of all frequent $k$-itemsets. First, in the join steps, $F_{k-1}$ is joined with $F_{k-1}$.

```
insert into Ck
select p.itemset1, p.itemset2, . . . , p.itemsetk−1,
       q.itemsetk−1
from Fk−1 p, Fk−1q
where p.itemset1 = q.itemset1, . . . , p.itemsetk−2 =
      q.itemsetk−2, p.itemsetk−1 < q.itemsetk−1.
```

Here, $F_k$ $p$ means that the itemset $p$ is a frequent $k$-itemset, and $p.itemset_k$ is the $k$-th item of the frequent itemset $p$.

**Algorithm 1** Apriori Algorithm

```
F1 = {frequent 1-itemsets};
for (k = 2; Fk−1 = ∅; k ++) do begin
  Ck = apriori-gen(Fk−1); //New candidates
foreach transaction t ∈ D do begin
  Ct = subset(Ck , t); //Candidates contained in t
foreach candidate c ∈ Ct do
  c.count ++;
end
 Fk = {c ∈ Ck |c.count ≥ minsup };
end
 Answer= ⋃k Fk ;
```

Algorithm 2 shows association rule generation algorithm.

**Algorithm 2** Association Rule Generation Algorithm

```
H1 = ∅ //Initialize
foreach; frequent k-itemset fk , k ≥ 2 do begin
    A = (k − 1)-itemsets ak−1 such that ak−1 ∈ fk ;
    foreach ak−1 ∈ A do begin
     con f = support( fk)/support(ak−1);
    if (con f ≥ mincon f ) then begin
      output the rule ak−1 ⇒ ( fk − ak−1)
        with confidence = conf and support = support( fk );
      add ( fk − ak−1) to H1;
  end
end
  call ap-genrules( fk , H1);
end
Procedure ap-genrules( fk : frequent k-itemset, Hm: set
        of m-item consequents)
if (k > m + 1) then begin
```

International Journal of Electronics Communication and Computer Engineering
Volume 3, Issue 4, ISSN (Online) 2249 –071X, ISSN (Print): 2278 – 4209

*Hm*+1 = apriori-gen(*Hm*);
**foreach** *hm*+1 ∈ *Hm*+1 **do begin**
*con f* = support( *fk*)/support( *fk* − *hm*+1);
**if** (*con f* ≥ *mincon f* ) **then**
**output** the rule *fk* − *hm*+1 ∈ *hm*+1
with confidence = *conf* and support = support( *fk* );
**else**
**delete** *hm*+1 from *Hm*+1;
**end**
**call** ap-genrules( *fk* , *Hm*+1);
**end**

## V. NEGATIVE ASSOCIATION RULE MINING

The most common framework in the association rules generation is the "support-confidence" one. An interesting measure called conviction that adds to the support-confidence framework.

The conviction of a rule is defined as

$$\text{Conv}(X => Y) = \frac{1-\sup(Y)}{1-Conf(X => Y)}$$

Conv(X=>Y) can be interpreted as the ratio of the expected frequency that X occurs without Y (that is X=>¬ Y) if X and Y were independent divided by the observed frequency of incorrect predictions. The range of conviction is 0 to ∞. The following algorithm 3 is used for mining negative association rules.

**Algorithm 3** Mining Negative Association Rules

Input: TDB-Transactional Database
MS-Minimum Support
MC-Minimum Confidence
Output: Negative Association Rules
Method:
NAR<-Φ
Find F1<- Set of frequent 1- itemsets
**for** ( k=2;Fk-1!=Φ; k++)
Ck= Fk-1 ⋈ Fk-1
// Prune using Apriori Property
**for each** i ε Ck, any subset of i is not in Fk-1 then
Ck = CK - { i }
**for each** i ε Ck
s= Support( i);
**for each** A,B (A U B= i )
**if** ( Supp(A -> ¬ B) ≥ MS && Conviction (A->¬ B)≤2.0)
NAR<- NAR U { A-> ¬ B)
**if** ( Supp(¬ A->B) ≥ MS && Conviction(¬ A->B) ≤2.0) then
NAR <- NAR U { ¬ A -> B}
**end**
**end**
**end**

## VI. CONCLUSION

Association rule mining, one of the most important and well researched techniques of data mining. It aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories. Association rules are widely used in various areas such as telecommunication networks, market and risk management, inventory control etc., Association rule mining is to find out association rules that satisfy the predefined minimum support and confidence from a given database.

Typical association rules consider only items enumerated in transactions. Such rules are referred to as positive association rules. Negative association rules also consider the same items, but in addition consider negated items (i.e. absent from transactions). Negative association rules are useful in market-basket analysis to identify products that conflict with each other or products that complement each other.

## REFERENCES

[1] Adrians, P., & Zantige, D. (1996). Data mining. Addison-Wesley.
[2] Agrawal, R., Imielinski, T., and Swami, A. N. 1993. Mining association rules between sets of items in large databases. In roceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 207-216.
[3] Tan, P., Kumar, V.: Interestingness measures for association patterns: A perspective. In: Proc. of Workshop on Postprocessing in Machine Learning and Data Mining. (2000)
[4] Wu, X., Zhang, C., Zhang, S.: Mining both positive and negative association rules. In: Proc. of ICML. (2002) 658–665.
[5] Yuan, X., Buckles, B., Yuan, Z., Zhang, J.: Mining negative association rules. In:Proc. of ISCC. (2002) 623–629.
[6] C. Borgelt. Efficient implementations of Apriori and Eclat. In Proc. of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations (FIMI'03), volume 90 of CEUR Workshop Proceedings, 2003. http://ftp.informatikrwth-aachen.de/Publications/CEUR-WS/Vol-90/borgelt.pdf.

## AUTHORS PROFILE

**K. Lavanya**
Pursuing her M.Tech from Pydah College of Engineering & Technology, Kakinada, A.P., India. Her research interest includes Data Mining, Database Management System and Computer Networks.

**A. Kamala Priya**
Working as Asst. Professor in Pydah College of Engineering & Technology, Kakinada, A.P., India. Her research interest includes Data Mining, Database Management System and Computer Networks.

**P.Suresh Babu**
Working as Asst. Professor in Kaushik College of Engineering, Visakhapatnam, A.P., India. His research interest includes Data Mining, Database Management System and Computer Networks.