ISSN (e): 2250 – 3005 || Volume, 05 || Issue, 03 || March – 2015 ||
International Journal of Computational Engineering Research (IJCER)

Pattern Discovery Using Apriori and Ch-Search Algorithm

Prof.Kumbhar S.L.[1], Mahesh Aivale[2], Kailas Patil[3], Pravin Jadhav[4], Baliram Sonawane[5]

1 Professor, Department Of Computer Engineering, SBPCOE, Indapur, India
2,3,4,5 Students, Department Of Computer Engineering, SBPCOE, Indapur, India

Abstract:
The association rule describes and discovers the relations between variables in large databases. Generally association rules are based on the support and confidence framework. In apriori algorithm, association rules are used for finding minimum support & minimum confidence. But this method can lead to loss of rules i.e. negative rules are lost. Hence Ch-Search algorithm is introduced which uses its strongest rule i.e. commonly used the Coherent rule which promotes information correctness and leads to appropriate decision making among the same item sets. The coherent rule discovers positive as well as negative outcomes, and does not require finding minimum support & minimum confidence frameworks leading to no loss of any rules. The paper describes how the Ch algorithm is used by coherent rule for exact pattern discovery in large databases over apriori algorithm.

Keywords: *Data Mining, Association Rule, Apriori algorithm, Support, Confidence, Coherent Rule, Ch-Search Algorithm etc…*

I. INTRODUCTION

Data Mining is the process of uncovering or discovering hidden and potentially useful information from larger amount of database. Database Mining is normally done using a data ware house, hidden information is not very obvious and not visible directly but rather is interpreted in some manner when it is discovered. Data mining process normally consists of two broad categories such as descriptive information & predictive information. The descriptive information we find patterns that are human interpretable something that human can understand we try to discover that kind of patterns from my set of data.

The Association rules are a simple if/than statement that shows the relations among the variables in large database. Association rules are created by analyzing data for frequent if/then patterns and using the frameworks as like support and confidence to identify the most important relationships. Support indicates that how frequently the items appear in the database. Confidence indicates the several times if /then statements have been found to be true. In data mining, association rules that very useful for analyzing and predicting customer behavior.

Association rule is an implication expression of form X->Y where X and Y are item sets. We can say that for pattern "customers who purchase item A also intend to buy item B together" is represented by above equation A=>B Thus association rules are important for knowing all possible relationships between large numbers of possible combinations of items.

Frequent Pattern: Frequent pattern are pattern (such as item set, sub sequence or structure) that appear in a data set frequently.

1. Item set: Item set is non empty set of items. Its cardinality can be one to any positive number. Any transaction contains item set of constant size(x).

2. Frequent item sets: Frequent item sets are sets which occurs one item at a time. The set of item that have at least a given minimum support. In simple words item sets which occur frequently.
   The item sets which are below threshold are called as infrequent item sets. The frequent and infrequent item sets are subsets of the superset. Both indicate its presence of item set. This distinguishes them from absence of item set which is same item set being absent from the same set of transaction record.
   Consider 3 items in database such as X, Y and Z.
   X transaction contains only item A.
   {A} is presence of item set.

Now Y and Z are said to be absent from this transaction. Thus {B} and {C} are absence of item sets and are represented by - {B} and -{C} resp.

3. Subsequence pattern: If items occur frequently, it is called frequent sub-sequence pattern.
4. Structured pattern: A sub structured such as sub tree, sub graph, or sub lattice that can be combined with item set or sequence if a sub structure occurs frequently, it is called as frequent structured pattern. Association rules are discovered from frequent item sets. The minimum support threshold must be pre-set by user for determining frequent item set. Rules generated by {A} item set that are present or frequent item set are called positive association rules And rules involving absence of item set i.e. {B} and {C} are called negative association rules.

Generation of association rule can be consisting of two sub problem:

1. Finding Frequent item sets whose occurrences exceed a predefined minimum support
2. Deriving Association rules from those frequent item Sets.

II. APRIORI ALGORITHM

Apriori Algorithm was first proposed by R.Agrawal & Shrikant in 1994 in during their research working market basket analysis. The basic idea of the apriori Algorithm was to generate the candidate key & frequent item sets .The algorithm uses two keywords such as support & confidence.

a) Support is nothing but the probability of the buying the product i.e. the number of instances in the given to the total number of records in data-set.

$$\text{Support}=\text{freq}(x, y)/N$$

Where N is the number of records in dataset

b) Confidence is nothing but the relative probability of buying the product i.e. how often the consequent item occurs in combination of antecedent and consequent item.

$$\text{Confidence} = \text{freq}(x, y)/x$$

For example, if a person buys a product such as mobile, he can also buy memory card too. So the shop keeper keeps memory card with mobile, this is support & in confidence a person buys a memory card, he can also buy headphone or battery or cover with it. This relational probability is helpful in keeping your relation in the product so actually this algorithm is useful in finding the frequent pattern.

For smaller database or data-set frequent pattern can be found out easily here, in the example of mobile shop there were only 10 or 20 pieces reducing the transaction at last. But when large database/ & data-set are used such as in big bazaar & mall can have thousands of data items that leads to thousands of transaction. In such cases the database is iterative scanned that is very difficult. Hence apriori type approach must be used to reduce the complexity & computing time.

Two important steps in apriori algorithm are:

1. Find all item set that have minimum support (Frequent item set also large item set)
2. Use frequent item set to generate rules. The candidate generation function F [k-1] and return a super-set (called the candidate) of the set of all frequent K-item set. It has two steps:
   i) Join Step: Generate all possible candidates
      Item set Ck of length k.
   ii) Prune Step: Remove those candidates in Ck
      That cannot be frequent.

2.1 Generate Frequent Item Set

1. Count support of each individual item
2. Create a set F with all individual items with min support.
3. Creates “Candidate Set” C [K] based on F [K-1]
   Check each element c in C[k] to see if it meets min support.
4. Return set of all frequent item sets.

2.2 Generate Candidate Sets

1. Create two sets differing only in the last element, based on some seed set.
2. Join those item sets into c.
3. Compare each subset s of c to F [K-1] – if s is not in F [K-1], delete it.
4. Return final candidate set.

Example of generation candidate
L3= {123,124,134,135,234}
Use the concept self joining = L3* L3

- 1234 from 123 & 124
- 1345 from 134 & 135

Then use concept pruning (removing)

- 1345 is removed because 145 are not in L3.

So candidate set is
C4= 1234

2.3 Apriori Algorithm Examples with Working

Database: D, Minimum Support: 0.5

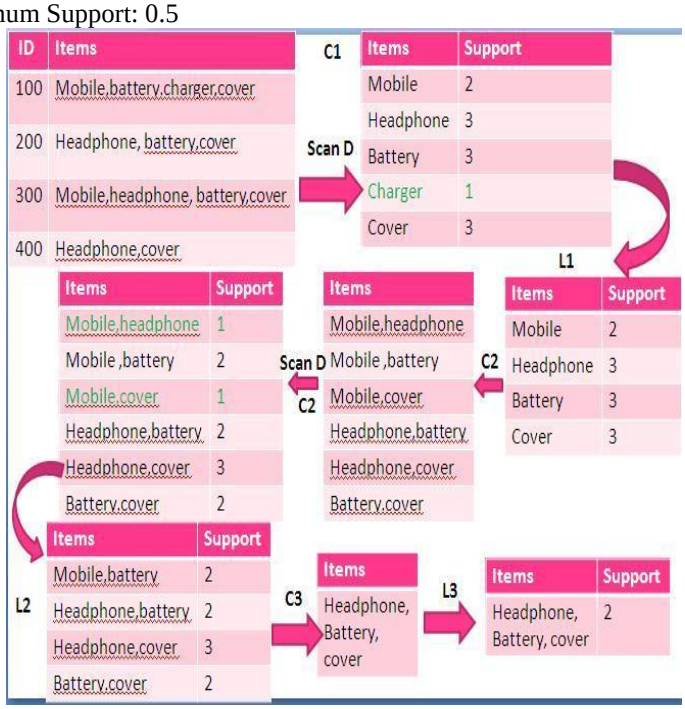

Fig -1: Apriori Algorithm Example

Workout of the example of apriori algorithm: Here is the database of mobile shop with minimum Support 0.5. There are four transactions each has unique ID.

In the first stage scan the entire database as sequential scanned which generates C1 i.e. candidate item set in which support is calculated. Here in the example mobile is occur 2 times, headphones occurs 3 times & so on in database D,then generate the frequent item set are stored in L1 in that charger item set occur one time so it is less than minimum support i.e.

Support=1/4=0.2

When support value < minimum support

Then this entry shall be pruned i.e. remove and then C2 is generated with possible combination of item sets in L1. Then again the entire database is scanned with item sets present with C2 calculate the support value.L1 is calculated using pruning method (support value < minimum support) i.e. L2. C3 consist of possible combination generated by scanning entire database & then again using pruning method (support value < minimum support) i.e.L3 is generated.

III. COHERENT RULE

We propose a system to discover pattern report as coherent rules. Based on the properties of propositional logic and therefore the coherent rules are discovered. In table 1, some item are present that contains relations between a rule on sequence (RHS) C, & a rule antecedent (LHS). The rule antecedent A consist of a combination of item called an antecedent item set X , The rule consequence C consist of a combination of item called an antecedent item set Y.The antecedent item set X may present at represent as X & absence it represent as ¬X, consequence item set Y may represent as Y & absence it represented as ¬Y.

Table -1: antecedent and consequence items set

		A rule consequence (RHS), C		
		Y	¬Y	Total
A rule antecedent (LHS), A	*X*	*Q1*	*Q2*	*A1*
	¬*X*	*Q3*	*Q4*	*A2*
	Total	*C1*	*C2*	*m*

Rules

i. X⇒Y is mapped to propositional logic implication
p→q if and only if Q1>Q2, Q1>Q3 and Q1>Q4.

ii. X⇒¬Y is mapped to propositional logic implication
P→¬q if and only if Q2>Q1, Q2>Q3, Q2>Q4.

iii. ¬X⇒Y is mapped to propositional logic implication
¬p→q *if* and only if Q3>Q1, Q3>Q and Q3>Q4.

iv. ¬X⇒¬Y is mapped to propositional logic implication
¬p→¬q if and only if Q4>Q1, Q4>Q2, and Q4>Q3.

Having mapped each called pseudo implication.

By pseudo implication, that is near by a real implication according to propositional logic. It is not a genuine implication since there are differences pseudo implication is true or false based on comparison of support. Coherent Rules are a pair of antecedent and consequence item sets, X and Y represents using a pair of rules the truth table value for equivalent. E.g. X⇒Y, ¬X⇒¬Y where,

i) X⇒Y is mapped to logic equivalent p≡q if and only if Q1>Q@, Q1>Q3, Q4>Q2 and Q4>Q3

ii) X⇒¬Y is mapped to logic equivalent p≡¬q if and only if Q2>Q1, Q2>Q4, Q3>Q1 and Q3>Q4. And also has ¬X⇒Y is mapped to logic equivalent ¬p≡q if and only if Q2>Q1, Q2>Q4, Q3>Q1 and Q3>Q4.

iii) ¬X⇒¬Y is mapped to logic equivalent ¬p≡¬q if and only if Q1>Q2, Q1>Q3, Q4>Q2 and Q4>Q3

Having mapped each rule is called pseudo implication of equivalent. Coherent rules by using the properties of the positive and negative rules on the condition: Set (positive rule) > Set (negative rule) at preselected consequence item set.

IV. CH-SEARCH ALGORITHM

In Ch-search algorithm there is no need to preset minimum support to find out association rules. Coherent rule are found based on logical equivalence. Also further these rules are utilized as association rule. In Ch-search, it is not needed to create frequent item set and association rule within each item set.

Input: D – a database, Y – a consequence item set
Output: CR – a set of coherent rules

[1] $CR \leftarrow \emptyset$
[2] $I \leftarrow$ find a set of unique items from D
[3] Let $A = I - Y$
[4] $Y.count \leftarrow$ total counts of Y in D
[5] $O_{P(A)} \leftarrow$ virtually map the power sets of A to the indices of a binary system
[6] For each i-th element of the power sets of A in order of O_i,
(i) $X \leftarrow \{P_i : i \in P(A)\}$
(ii) $S(X,Y) \leftarrow XY.count$
(iii) $S(\neg X,Y) \leftarrow Y.count - S(X,Y)$
(iv) if $S(X,Y) > S(\neg X,Y)$,
if equation (2) is met, $CR = CR \cup (X,Y)$
Loop [6] until $i = |P(A)|$
(v) remove all power sets of A having the i-th element
[7] return CR

* For example, given 3 items, the first item set *null* – a member in the power sets of X, item set $X_{i=1\_}$ is indexed using binary number '0', item set $X_{i=2\_}$ is indexed using '1_2', and item set $X_{i=3\_}$ is indexed using '10_2'.

Basically, in apriori algorithm, negative rules are not found. But in case of Ch-search algorithm, negative rules are found and used to implement both positive and negative rules found.

"Patterns are discovered based on generated rules which are more efficient.

i) Positive Rules: Some of the association rules consider only items enumerated with transactions, such rules are called positive association principle.
Ex. mobile =>headphone

ii) Negative Rules: Negative association rules consider the same items, & also negated items.
Ex. ¬mobile =>¬headphone. "[Galphad,2013]

The Algorithm presented in this paper extends the support-confidence framework with correlation coefficient threshold. With finding confident positive rules having a strong correlation, the algorithm discovers negative association rules with strong negative correlation found between the strongest correlated rules, followed by rules with moderate and small strength values. After finding the association rules it was found that patterns are more efficient than the rules. In association rules only those attributes are considered as strongly responsible to find the result.

In case of the patterns all the attributes are considered.

eggs = 1 and aquatic = 1 and predator = 1 and legs = 1 and hair = 0 and feathers = 0 and milk = 0 and airborne=0 and toothed = 0 and backbone= 0 and breaths = 0 and venomous = 0 and fins = 0 and tails = 0 and domestics = 0 == INVERTEBRATE

Patterns are more efficient than rules [4].

V. PROPOSED SYSTEM

Throughout apriori algorithm observed that, there are some issues such as multiple scan from the database, low accuracy, candidate technology process complicated, more space, time, memory etc. Therefore other approach has to be found out which can work on these problems. In theoretically, Ch-Search Algorithm generates correct & valid style, and generate association rule using proposed algorithm instead of any minimum support threshold criteria like apriori algorithm also produce standard association rules using propositional logic & classify the test files using generated rule & pattern & finally look when placed against system result with apriori criteria.

The proposed architecture for apriori & Ch-Search Algorithm:

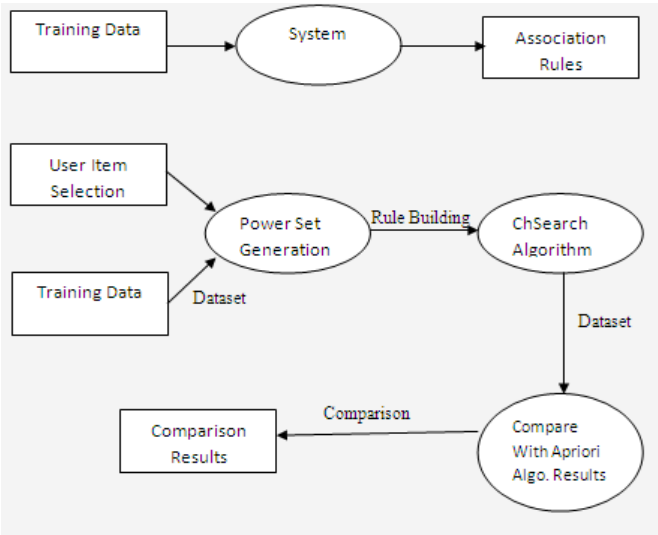

Fig.2: Pattern Discovery using apriori & Ch-Search Algorithm

VI. CONCLUSIONS

In this paper, generalized framework is used to discover association rules that have the proportional logic, & a specified framework (Coherent Rules Mining framework) with a basic algorithm to generate coherent rules from a given data set. The contribution of this work mainly focused on discovery of coherent rules through a complete set of interesting association rules that are implication according to propositional logic. The Search for coherent rules does not require a user to pre-set a minimum support threshold. In contrast an association rule is typically not implication according to propositional logic and infrequent item set a coherent rule mining framework can thus be appreciated for its ability to discover rules that are both implication and complete according to propositional logic from a given dataset.

ACKNOWLEDGEMENTS

This Paper involves many respected hands. We offer our sincere thanks to Prof. Nalawade V.S & Prof. Kumbhar S.L who guided us at each and every stage with his valuable suggestions and guidance while completing this paper. We are also thankful to all the staff members of the Computer Department, for their suggestions and moral support in completion of this paper. We shall always be grateful to them for their splendid and sportive help and for providing necessary facilities. We also like to thank our HOD Prof. More A.S who creates a healthy environment for all of us to learn in best possible way.

REFERENCES

[1] Alex Tze Hiang Sim, "Discovery of Association Rules without a minimum support threshold – Coherent Rules Discovery" March 2009.

[2] Alex Tze Hiang Sim, Maria Indrawan, Samar Zutshi, Member, IEEE, and Bala Srinivasan,"Logic Based Pattern Discovery", IEEE transaction on knowledge and data engineering, Vol.22, No.6, June 2010.

[3] Sharada Narra, Madhavi Dabbiru, "Privacy Preserving Generalized Coherent Rule Mining", International Journal Of advances Research in Computer Science and Software Engineering (IJARCSSE), ISSN: 2277 128X, Vol 3, Issue 10, Oct 2013.

[4] Pratik P. Raut, Sagar D. Savairam, Pravin S. Lokhande, Varsha B. Khese, Prof. M.N. Galphad. "Association Rule Mining by Using New Approach of Propositional Logic", International Journal of Computational Engineering Research, Vol, 03, Issue, 4, April 2013.

[5] Alex Tze Hiang Sim, Samar Zutshi, Maria Indrawan, Bala Srinivasan, "The Discovery of Coherent Rules"