DISCOVERING PATTERN FROM LARGE DATABASE USING COHERENT RULES

Mrs. S.Mahalakshmi

PG Scholar Computer Science Department ,IFET College of Engineering and technology

IFET Nagar,GangaramPalayam,Villupuram-605108

sm2igrow@gmail.com

Abstract:

In the data mining field, association rules have been researched for more than fifteen years; however, the degree to which the support threshold effectively discovers interesting association rules has received little attention. This thesis proposes a new framework for data mining through which interesting association rules called *coherent rules* can be discovered. Coherent rules are those associations that can be mapped to logical equivalences according to propositional logic. Hence, coherent rules can be reasoned as logically true statements based solely on the truth table values of logical equivalence. Discovering coherent rules resolves the many difficulties in mining associations that require a preset minimum support threshold. Apart from solving the issues of a support threshold, the coherent rules found can also be reasoned as logical implications due to the mapping to the truth table values of logical equivalence.

1. INTRODUCTION

An algorithm to discover coherent rules is also presented in this thesis. The algorithm was designed to find the shortest and strongest rule or most effective coherent rules by exploiting the properties of coherent rules. In this paper, we introduced a new rule to rectify the loss of data's. It is used to detect the relationship or associations between specific values in large data set. A priori is a representational algorithm based on this framework and many other algorithms area without this threshold specified, typically, no association rules can be discovered because the procedure to discover the rules will quickly exhaust the available resources domain. Nonetheless, having to constrain the discovery of association rules with a preset threshold, in turn, requires in-depth domain knowledge before the discovery of rules can be automated. The use of min sup generally assumes that:

- The knowledge of interest must have occurred frequently at least equal to the threshold.
- A single threshold is enough to identify the A domain expert can provide the threshold value accurately.
- Knowledge sought by an analyst.

2. Problems Existing in Association Rule Mining:

Existing association rule mining algorithms are based on minimum support threshold in order to generate rules which require domain knowledge; in this case interesting rules are missing for future analysis.

- Different minimum support threshold would result inconsistent mining results even when the mining process is performed on the same dataset.

- Association rule discovered using support and confidence may not be correlated in statistics.

- Some rules may not be interesting due to high marginal probabilities in their consequence item sets.

- Frequently co-occurred association rules even with high confidence values may not be truly related Association rule mining algorithms have high correlation error rate.

3. Issues using a Minimum Support:

The issues with Issues with discovering association rules reverberate around loss of rules and quality of rules discovered. Specifically, if rules are lost, it is misleading to report an incomplete set of rules and at the same time create a sense that all available rules have been found. This situation misleads a decision maker into thinking that only these rules are available which, in turn, will lead a decision maker to reason with incomplete information. For example, it is erroneous to assume that a subset of an incomplete set of rules has the strongest rules. Reasoning with incomplete information while not knowing it may lead to in appropriate conclusion or decisions. Quality wise, association rule mining is known to report on every detail of associations among items but unable to identify in specific the type of knowledge rules required. This is especially true when the association rules required mix between those infrequently and frequently observed rules. A large proportion of rules that fall in-between these frequencies of occurrences quickly collude the results discovered. In fact, rules discovered must not be too rare; otherwise, the mining process could take forever or its reported results are too large and difficult to process.

4. Market Basket Analysis:

A number of functions are used in data mining including, for example, link analysis, prediction and visualization. A link analysis typically discovers the knowledge of "what goes with what" and "what follows what". The latter is called *sequence analysis* and identifies a sequence of events, while the former is known as *affinity analysis* ,Finding these associations helps to describe customers' buying habits. Knowing such associations helps a retailer to devise effective marketing strategies. A promotion to increase the sale of any one item within an association could increase the sales of another item.

5. Loss of Association Rules Involving Frequently Observed Items:

Some frequent association rules are lost due to the heuristics involved in setting a minimum support threshold. Use of a minimum support threshold to identify frequent patterns assumes that an ideal minimum support threshold exists for frequent patterns, and that a user can identify this threshold accurately. Assuming that an ideal minimum support exists, it is unclear how to find this threshold [3]. This is largely due to the fact that there is no universal standard to define the notion of being frequent enough and interesting. The strength value of association rules has been occasionally debated in statistics would result in inconsistent mining results, even when the mining process is performed on the same data set.

6. Loss of Association Rules Involving Infrequently Observed Items:

Some infrequent association rules are actionable. Typically, a data set contains items that appear frequently while other items rarely occur. For example, in a retail fruit business, fruits are frequently observed but occasionally bread is also observed. Some items are rare in nature or infrequently found in a data set. These items are called rare items. If a single minimum support threshold is used and is set high, those association rules involving rare items will not be discovered. Use of a single and lower minimum support threshold, on the other hand, would result in too many uninteresting association rules. This is called the rare item problem defined by according to the latter pointed out that in maintaining the use of a minimum support threshold to identify rare item sets, many users will typically group rare items into an arbitrary item so that this arbitrary item becomes frequent.

7. Implication:

In an argument, the truth and falsity of an implication (also known as a compound proposition) (->) necessarily relay on logic. Each implication, having met specific logical principles, can be identified (for example, one may be a material implication, while the other may be an equivalence).Each has a set of different truth values. This will be explained later. We highlight here that an implication is formed using two propositions p and q. These propositions can be either true or false for the implication interpretation. For example, "apples are observed in a customer market basket" in the same way the in social network the behavior of friends with in the same geographical is same means it is a true interpretation if this has been observed. From these propositions, we have four implications

1. p -> q,
2. p ->ר q,
3. ר p -> q, and
4. ר p ->ר q.

Each is formed using standard symbols "->" and "ר" The symbol "->" implies that the relation is a mode of implication in logic, and "ר" denotes a

false proposition. The example for implications 1 and 2 below:

1. If "apples are observed in a customer market basket," then "bread is observed in a customer market basket" p -> q.
2. If "apples are observed in a customer market basket," then "bread is NOT observed in a customer market basket" p -> ר q. The truth and falsity of any implication is judged by "adding" (^) the truth values held by propositions p and q. In a fruit retail business where no bread is sold, the implication that relates p and q will be false based on the operation between truth values; that is1 ^ 0 =0. The second implication based on the operation will be true because 1 ^ 1=1.Hence, we say that the latter implication p -> ר q is true, but the first implication p -> q is false. Each implication has its truth and falsity based on truth table values alone. There are a number of modes of implication. In the same way the tags that occur frequently with common link so that occurrence of one may lead to click on another tag or occurrence of one tag may prevent the other tag. We highlight two modes of implication and their truth table values in the next two sections.

8. Material Implication:

A material implication meets the logical principle of a contraposition. A contra positive (to a material implication) is written as ר q -> ר p. For example, suppose, if customers buy apples, that they then buy oranges is true as an implication. The contra positive is that if customers do not buy oranges, then they also do not buy apples. In case of tags the same tags with similar properties may occur together and if they don't occur and its implication has the truth values of its contra positive, ר(p ^ ר q), it is a material implication. That is, p and q if ר(p^ ר q).

9. Equivalence:

An equivalence (=) is another mode of implication. In particular, it is a special case of a material implication. For any implication to qualify as equivalence, the following condition must be met p =q ifר(p xor q) where truth table values can be constructed in Table 3. Equivalence has an additional necessary condition. Due to this condition, propositions are now deemed both necessary and sufficient relates with," in short)

10. Truth Table for a material Equivalence:

This is largely due to the fact that there is no universal standard to define the notion of being frequent enough and interesting. The strength value of association rules has been occasionally debated in statistics. In one case, Bobbie doll. A minimum support is set on each item in a data set. Hence, we have a finer granularity ofa minimum support threshold compared to the classic approach. Use of MIS results in association rules being found in which item sets occur infrequently and below a minimum support threshold.

Mapping of Association Rules to Equivalences

Equivalences:	$p \equiv q$	$\neg p \equiv \neg q$
Association Rules:	$X \Rightarrow Y$	$\neg X \Rightarrow \neg Y$

True or False on Association Rules	Required Conditions (to map associations to equivalences)	
T	$X \Rightarrow Y$	$\neg X \Rightarrow \neg Y$
F	$X \Rightarrow \neg Y$	$\neg X \Rightarrow Y$
F	$\neg X \Rightarrow Y$	$X \Rightarrow \neg Y$
T	$\neg X \Rightarrow \neg Y$	$X \Rightarrow Y$

11. Mapping Association Rules to Equivalences

The distinctions between an association rule and an implication is analyzed and highlighted the motivation to map an association rule to an implication. This section explains how to map an association rule to equivalence. A complete mapping between the two is realized in three progressive steps. Each step depends on the success of a previous step. In the first step, item sets are mapped to propositions in an implication. Item sets can be either observed or not observed in an association rule. Similarly, a proposition can either be true or false in an implication. Analogously, the presence of an item set can be mapped to a true proposition because this item set can be observed in transactional records. Having mapped the item sets, an association rule can now be mapped to an implication in a second step. An association rule has four different combinations of presence and absence of item sets having mapped item sets and association rules, now map association rules into specific modes of implication that have predefined truth table values and focus on equivalence. Based on a single transaction record in association rule mining methods.

12. Mapping Using a Single Transaction Record

An item set has two states. In a single transaction record, an item can either be present or absent from the transaction record. It follows then that a proposition can either be true or false. If an item set is observed in a transaction record, it is analogous to having a true proposition. In the same way, item sets are mapped to propositions p and q as follows

Item set X is mapped to p =T, if and only if X is observed.

- Absence of item set X, that is, ר X is mapped to p =F, if and only if X is not observed.
- Item set Y is mapped to q = T, if and only if Y not observed.
- Absence of item set Y, that is, ר Y is mapped to q= F, if and only if Y is not observed.

Each component of an association rule is now mapped to propositions. Using the same mapping concept, an association rule can be mapped to a

true or false implication. An association rule consists of two item sets X and Y. Following the mappings above:

- Item sets X and Y are mapped to p and q = T, if and only if X and Y are observed.

13. Mapping Using Multiple Transaction Records

Previously, item sets have been mapped to propositions p and q if each item set is observed or not observed in a single transaction. In data containing multiple transaction records ,an item set X is observed over a portion of transaction records. This total number of observations is given by the cardinality of the transactions in database D that contain X, known as support:

support; S(X)=IDxI

It support the number of times X which is a support S(X) denotes the number of times X which is observed in the entire data. Similarly, support S(X) denotes the number of times X which is not observed in the entire data.

Association Rules and Supports

Association Rule	Support
$X \Rightarrow Y$	S(X, Y)
$X \Rightarrow \neg Y$	S(X, $\neg Y$)
$\neg X \Rightarrow Y$	S($\neg X$, Y)
$\neg X \Rightarrow \neg Y$	S($\neg X$, $\neg Y$)

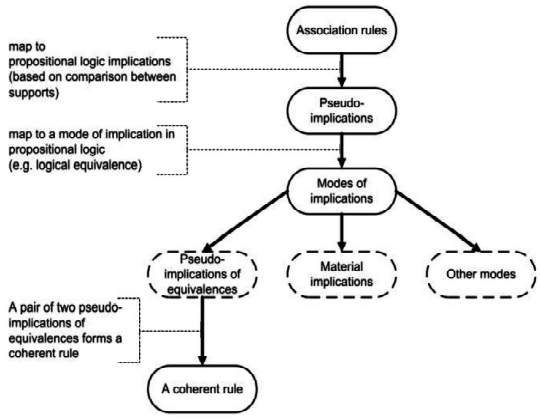
A Pseudo implication is judged true or false based on a comparison of supports, which has a range of integer values. In contrast, an implication is based on binary values. The former depends on the frequencies of concurrences between item sets (supports) in a data set, whereas the latter does not and is based on truth values. Having a minimum support threshold results in the loss of some interesting association rules concerning items infrequently observed in a dataset. In addition, interesting association rules concerning absence of items are also lost. As a result, existing research has focused on three approaches to discover more association rules.

The process to find association rules does not require user knowledge of a minimum support threshold because an implication stays independent from user knowledge. The implication is purely based on logical grounds. . Unlike support threshold settings, these parameters do not require users to have prior knowledge of the context in which the data mining takes place. We have tested our framework on several datasets. The results confirm the strength of coherent rules in finding association rules that can be reasoned logically and in finding association rules that consider both infrequent items and negative associations.

14.Generalized framework of association rules:

The algorithm used to discover coherent rules is also efficient. This was demonstrated by the number of pruning's made to the search space during the discovery process. This study suggests that our framework for discovering coherent rules offers a technique for data mining that overcomes the limitations associated with existing methods and enables the finding of association rules among the presence and/or absence of a set of items without a preset minimum support threshold. The results justify continuing research in this area in order to increase the body of scientific knowledge of data mining and specifically, association rules - and to provide practical support to those involved in data mining activities. The use of association rule mining technique is to describe the associations among items in a database. These associations represent the domain knowledge encapsulated in databases.

Identifying domain knowledge is important because these knowledge rules usually are known only by the domain experts over years of experience. Thus, association rule mining is useful to identify domain knowledge hidden in large volume of data efficiently. The discovery of association rules is typically based on the support and confidence framework where a minimum support (min sup) must be supplied to start the discovery process .A priori is a representational algorithm based on this framework and many other algorithms area without this threshold specified, typically, no association rules can be discovered because the procedure to discover the rules will quickly exhaust the available resources.



A generalized framework of association rules that based on pseudo implications.

Using the concept of pseudo implication, we now can further map association rules to specific modes of implications such as material implications and equivalences. Each follows the same truth values of the respective relations in logic. exchanged both sides, this new equivalence follows also the truth table values of equivalence. This characteristic is not observed in material implications and remains as a "more relaxed" mode of implication. In mining rules from data sets and without requiring

background knowledge of a domain, we need a strong reason to identify the existence of rules.

15. The Differences in Setting Up Thresholds:

There are fundamental differences between the thresholds used in the coherent rules mining framework and the threshold set by a user for the support and confidence framework and also used that are frequently and infrequently observed in a set of transaction records. Based on a single transaction record in association rule mining, Having mapped the item sets, an association rule can now be mapped to an implication in a second step. An association rule has four different combinations of presence and absence of item sets.

id	Content of T_{id}	id	Content of T_{id}
1	i_2	14	$i_2, i_3, i_4, i_5, i_6, i_7$
2	i_6	15	$i_1, i_2, i_3, i_4, i_5, i_6, i_7$
3	i_2	16	$i_1, i_2, i_3, i_4, i_5, i_6, i_7$
4	i_3	17	$i_1, i_2, i_3, i_4, i_5, i_7$
5	i_3	18	i_1, i_2, i_3, i_4, i_6
6	i_3, i_4	19	i_1, i_2, i_3, i_4
7	i_3, i_4	20	i_1, i_2, i_3, i_5, i_6
8	i_3, i_4, i_5, i_7	21	i_1, i_2, i_3, i_5, i_6
9	$i_1, i_3, i_4, i_5, i_6, i_7$	22	i_1, i_2, i_5
10	$i_1, i_3, i_4, i_5, i_6, i_7$	23	i_2, i_5
11	i_1, i_3, i_4, i_5, i_7	24	i_1, i_2, i_4, i_7
12	$i_1, i_2, i_3, i_4, i_5, i_6, i_7$	25	i_2
13	i_2, i_3, i_4, i_5, i_7		

16. Artificial Transaction Records

In the retail domain, the reasons for associations between items are not obvious. Customers have various reasons to buy different items together. Using mapping to logical equivalences, we can discover coherent rules The discovery of coherent rules is useful in application domains where the domain knowledge is not known to a user and its association rules without the need to survey on customers. As a result,
we know that some items are associated together based logical grounds.

.

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

[1] $CR \leftarrow \varnothing$
[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) $X \leftarrow \{I_i : i \in P(A)\}$
(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 $X_{\{\}}$ is indexed using binary number '0', item set $X_{\{1\}}$ is indexed using '1_2', and item set $X_{\{1,3\}}$ is indexed using '10_2'.

A simple search for coherent rules algorithm (Ch Search).

17. Distinct Features of Ch Search

We list some features of Ch Search compared to a priori. Unlike a priori, Ch Search: does not require a preset minimum support threshold.Ch Search does not require a preset a minimum support threshold to find association rules. Coherent rules are found based on mapping to logical equivalences. From the coherent rules, we can decouple the pair for two pseudo implications of equivalences. The latter can be used as association rules with the property that each rule can be further mapped to a logical equivalence. does not need to generate frequent item sets. Ch Search does not need to generate frequent item sets. Nor does it need to generate the association rules within each item set. Instead, Ch Search finds coherent rules directly. Coherent rules are found within the small number of candidate coherent rules allowed through its constraints.

Features of Ch Search

#	Class Attributes	Frequency of Occur-rence	%	Type of Associa-tion
1	*Reptile*	5/ 101	4. 95	Infreq.
2	*Mammal*	41/ 101	40.59	Freq.
3	*Invertebrate*	10/ 101	9.90	Freq.
4	*Insect*	8/ 101	7.92	Freq.
5	*Fish*	13/ 101	12.87	Freq.
6	*Bird*	20/ 101	19.80	Freq.
7	*Amphibian*	4/ 101	3.96	Infreq.

identifies negative association rules. Ch Search, by default, also identifies negative association rules. Given a set of transaction records that does not indicate item absence, a priori cannot identify negative association rules. Ch Search finds the negative pseudo implications of equivalences and uses them to complement both the positive and negative rules found. (Although a priori finds negative association rules if a transaction database is transformed into binary attributes, the rules found typically contradict one another.

18. Coherent Rule Algorithms

1. Collect the data from database.
2.Preprocess the data by prediction techniques
3.Proposed association rule technique (by calculating support and confidence)
4.Then Association tag bundles rules and items are discovered then by using coherent rules based on logic the item sets are discovered and the modules are explained in design phase.

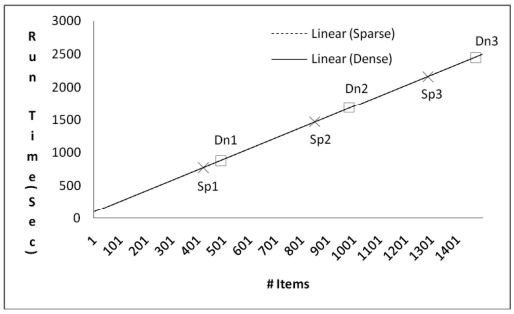
19. CONCLUSION

The association rules are analyzed based on different threshold values and the data items are mined, based on logic the coherent rules are analyzed and the tag based items are discovered. These association rules include item sets that are frequently and infrequently observed in a set of transaction records. In addition to a complete set of rules being considered, these association rules can

also be reasoned as logical implications because they inherit propositional logic properties. Having considered infrequent items, as well as being implicational, these newly discovered association rules are distinguished from typical association rules. These new association rules reduce the risks associated with using an incomplete set of association rules for decision making, as following:

- Our new set of association rules avoids reporting

that item A is associated with item B.



Reporting association graph items

The stronger association between item A and the absence of item B. Using prior association rules that do not consider this situation could lead a user to erroneous conclusions about the relationships among items in a data set. Again, identifying the strongest rule among the same items will promote information correctness and appropriate decision making. The risks associated with incomplete rules are reduced fundamentally because our association rules are created without the user having to identify a minimum support threshold. Among the large number of association rules, only those that can be mapped to logical equivalences according to propositional logic are considered interesting and reported. In this paper, we introduced our contributions in novel frameworks: a generalized framework to discover association rules that have the properties of propositional logic, and a specific framework (Coherent Rules Mining Framework) with a basic algorithm to generate coherent rules from a given data set. The discovery of coherent rules is important because through coherent rules, a complete set of interesting association rules that are also implicational according to propositional logic can be discovered. The search for support threshold. In contrast, an association rule is typically not implicational according to propositional logic, and the many approaches used to output association rules have lost rules involving infrequent item sets. A coherent rules mining framework can thus be appreciated for its ability to discover rules that are both implicational and complete according to propositional logic from a given data set.

REFERENCES

[1] Alex Tze Hiang Sim, Maria Indrawan, Samar Zutshi, Member,IEEE, and Bala Srinivasan,"Logic-Based Pattern Discovery," Ieee transactions on knowledge and data engineering, vol. 22,no.6,2010.

[2] R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases," SIGMOD Record, vol. 22, pp. 207-216, 1993.

[3] C. Longbing, "Introduction to Domain Driven Data Mining," Data Mining for Business Applications, L.Cao,P.S. Yu, C. Zhang, and H. Zhang, eds., pp. 3-10,Springer,2008.

[4] Piatetsky-Shapiro, G. (1991), Discovery, analysis,and presentation of strong rules, in G.Piatetsky-Shapiro & W. J. Frawley, eds, "Knowledge Discovery in Databases,"AAAI/MIT Press, Cambridge, 2009.

[5] Jochen Hipp, Ulrich Guntzer, and Gholamreza Nakhaeizadeh ,"Algorithms for association rule mining" - A general survey and comparison, SIGKDD Explorations, 2(2):1-58, 2000.