 ISSN : 0976-8491 (Online) | ISSN : 2229-4333 (Print)

Innovation of Logic-Based Pattern

[1]Thota. Jamalaiah, [2]N. V. Kiran babu Movva, [3]K. Nageswara Rao
[1,2,3]Dept. of CSE, MIST, Sathupally, AP, India

Abstract
In the data mining field, association rules are discovered having domain knowledge specified as a minimum support threshold. The accuracy in setting up this threshold directly influences the number and the quality of association rules discovered. We propose a framework to discover domain knowledge report as coherent rules. Coherent rules are discovered based on the properties of propositional logic, and no background knowledge to generate them. From the coherent rules discovered, association rules can be derived objectively and directly without knowing the level of minimum support threshold required.

Keywords
Threshold, Synthesizing, Data Mining, Association Rules

I. Introduction

The use of association rule mining technique is to describe the associations among items in a database. These associations represent the domain knowledge en- capsulated in databases. 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 must be supplied to start the discovery process [1]. A priori is a representational algorithm based on this framework and many other algorithms are a priori-like. Nonetheless, having to constrain the discovery of associatetion 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:

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

In this paper, we propose a novel framework to address the above issues by removing the need for a minimum support threshold. Associations are discovered based on logical implications. The principle of the approach con- siders that an association rule should only be reported when there is enough logical evidence in the data. To do this, we consider both presence and absence of items during the mining. By considering this new approach in finding data pattern, a solution toward fulfilling domain-driven data mining requirements [2] can be made. The proposed algorithm suggests a solution in two areas:

It eliminates the need to use different intelligence models and its combinations as suggested in [2] to determine appropriate threshold for the mining algorithms. The proposed algorithm discovers the natural threshold based on observation of data set. Hence, assuming that there are different intelligence models and a way of synthe- sizing it, the proposed algorithm can incorporate it to determine the target item(s) as an expression of business problem that one wants to solve.

It provides a logical underpinning to the discovery process of patterns. Currently, the illustration of the mapping of constraints to the discovery process in this paper is based on support value.

II. Issues Using Minimum Support Threshold

Issues with discovering association rules reverberate around loss of rules and quality of rules discovered. 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 inappropriate conclusion or decisions. This is especially true when the association rules required mix between those infrequently and frequently observed rules.

A. Loss of Association Rules Involving Frequently

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. If the scale in [5] is adopted for mining, then the minimum support threshold set based on [6] would be lower. This case shows that one user's understanding of an ideal strength value may be different from another's. For data mining, different minimum support thresholds would result in inconsistent mining results, even when the mining process is performed on the same data set. Some users will find fewer association rules compared to others who use a lower minimum support threshold. For the latter, association rules associated with frequent items should be discovered but are lost. We consider this situation as a case of losing association rules involving frequent items.

B. 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. These items are called rare items [7-9]. 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. Another practice is to split the data set into two or several blocks according to the frequencies of items, and mine each block using a different minimum support threshold. This approach does not need a preset minimum support threshold but a parameter that determines the actual discovery of frequent item sets [13]. Yun et al. [13] proposed a minimum support threshold called a second support to segregate item sets that occurred infrequently from coincidences, and a minimum relative support, which is the maximum of the proportion of the support of an item set against the support of each item within the item set.

C. Association Rules that are Measured Using Other Measures of Interestingness

A number of researchers have pointed out that association rules are not necessarily interesting even though they may have at least a minimum support threshold and a minimum confidence threshold. Brin et al. [14] show that association rules discovered using a support and confidence framework may not be correlated in statistics. The use of leverage and lift is also fundamental in designing a new measure of interestingness.

Table 1: Truth Table for a Material Implication

p	q	$p \supset q$
T	T	T
T	F	F
F	T	T
F	F	T

III. Generalized Association Rule Mining Framework

We propose a novel association rule mining framework that can discover association rules without the need for a minimum support threshold. This enables the user, in theory, to discover knowledge from any transactional record with- out the background knowledge of an application domain usually necessary to establish a threshold prior to mining. An implication having a rule where the left-hand side is connected to the right-hand side correlates two item sets together. This implication exists because it is true according to logical grounds, follows a specific truth table value, and does not need to be judged to be true by a user.

A. An Implication 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's interpretation. For example, "apples are observed in a customer market basket" is a true interpretation if this has been observed. From these propositions, we have four implications

1. P → q,
2. P →¯q,
3. ¯p → q, and
4. ¯p →¯ q.

Each is formed using standard symbols "→" and "⌐" The symbol "→" implies that the relation is a mode of implication in logic, and "⌐" denotes a false proposition.

Table 2: Truth Table for Equivalence

p	q	$p \equiv q$
T	T	T
T	F	F
F	T	F
F	F	T

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 → ⌐q.

C. Mapping Association Rules to Equivalences 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.

Table 3: 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	$\neg X \Rightarrow \neg Y$	$X \Rightarrow Y$

Each component of an association rule is now mapped to propositions. 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 ¼ T, if and only if X and Y are observed.
- X → Y is mapped to implication p→ q, if and only if both X and Y is observed.
- X → ר Y is mapped to implication p → ¯q, if and only if X is observed and Y is not observed.
- ר X → Y is mapped to implication ¯p → q, if and only if X is not observed and Y is observed.
- ר X → ר Y is mapped to implication ¯p → ¯q, if and only if both X and Y are not observed.

- X → Y is true;
- X → ר Y is false;
- ר X → Y is false; and
- ר X → ר Y is true.

1. 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.

Table 4: Association Rules and Supports

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

X => Y is mapped to an implication p -> q, if and only if
S(X, Y)>S(X, רY)
S(X, Y)>S (רX, Y) and
S(X, Y)>S (ר X, רy)

X=>רY is mapped to an implication p -> רq, if and only if
S(X, רY)>S(X, רY)
S(X, רY)>S (רX, Y) and
S(X, רY)>S (ר X, רy)

רX=>Y is mapped to an implication רp ->q, if and only if
S (רX, Y)> S(X, Y)
S (רX, Y)>S(X, רY) and
S (רX, Y)>S (ר X, רy)

רX=>רY is mapped to an implication רp ->רq, if and only if
S (רX, רY)> S(X, Y)
S (רX, רY)>S(X, רY) and
S (רX, רY)>S (ר X, Y)

Fig. 1, A generalized framework of association rules that based on pseudo implications. are fundamental differences. Pseudo implication is judged true or false based on a comparison of supports, which has a range of integer values.

IV. Coherent Rules Mining Framework

The pseudo implications of equivalences can be further defined into a concept called coherent rules (see fig. 1). Two pseudo implications of equivalences always exist as a pair because they are created based on the same conditions as shown in (2).

Since they share the same conditions, two pseudo implications of equivalences:
X > Y and ¬X =>¬Y

Table 5: Artificial Transaction Records

Frequency of co-occurrences		Consequence, Y	
		$Y = \{i_7\}$	$\neg Y = \neg\{i_7\}$
Antecedent, X	$X = \{i_2\}$	8	5
	$\neg X = \neg\{i_2\}$	3	9

id	Content of T_{id}	id	Content of T_{id}
1	i_7	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_5	18	i_1, i_2, i_3, i_4, i_6
6	i_1, i_4	19	i_1, i_2, i_3, i_4
7	i_1, i_4	20	i_1, i_2, i_3, i_5, i_6
8	i_1, 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		

Table 6:

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 $X_{1_{10}}$ is indexed using binary number '0', item set $X_{1_{10}}$ is indexed using '1_2', and item set $X_{1_{10}}$ is indexed using '10_2'.

V. Finding Coherent Rules in Transaction Records

Each pseudo implication of equivalence is an association rule, which can be further mapped to a logical equivalence. To explain this more formally, we adopt the following notation. Assume that I ¼ fi1 ; i2 ; ... ; in g, a set of items. Let T be a table of transaction records (relational table) such thatT ¼ ft1; t2 ; ... ; tm g. A task-relevant transaction record ti holds a subset of items such that ti I. Assume that we can predetermine the total number of items contained in two independent supersets A ¼ fa1 ; ... ; au g and C ¼ fc1 ; ... ; cv g such that A I; C I and A \ C ¼ .

Table 6, Contingency Table for Antecedent X and Consequence Y fig. 2. A simple search for coherent rules algorithm (ChSearch). Customers have various reasons to buy different items together. Using mapping to logical equivalences, we can discover coherent rules and its association rules without the need to survey on customers. As a result, we know that some items are associated together based on logical grounds. Instead, the support of fi2 g[:fi7 g is the highest. Equation (7) cannot be mapped to logical equivalences since Q1 > Q2, Q1 > Q3, Q4 > Q2, and Q4 > Q3, where:In this particular example, we have chosen i7 as a target item. In real-life situations and business problems, the target item can either be determined by the user or by using a DDM intelligence model (if available).

Table 7: Contingency Table for Antecedent X and Consequence Y

Frequency of co-occurrences		Consequence, Y	
		$Y = \{i_7\}$	$\neg Y = \neg\{i_7\}$
Antecedent, X	$X = \{i_2\}$	7	9
	$\neg X = \neg\{i_2\}$	4	5

Table 8: Total Frequency of Class Attributes

#	Class Attributes	Frequency of Occurrence	%	Type of Association
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.

Given a set of transaction records that does not indicate item absence, a priori cannot identify negative association rules. See a further discussion on this in Section VI.

VI. Experiments

A. Settings

we use the Zoo data set [13] to perform a series of experiments before test it on transaction records for Market Basket Analysis. The Zoo data set has seven classes of animals. It has a spectrum of frequency of occurrences in the transaction records. We show the frequencies of each class in Table 8. If the frequency of occurrence is below 5 percent, we consider the class of animal to be rare. Otherwise, the class is identified as frequent (that is, incurring at least 5 percent of support values).

B. Quality of Coherent Rules

1. The minimum support threshold is set at 5 percent [13-15].
2. The minimum confidence threshold is set at 50 percent [1,14].

- Infrequent association rules that contain less observed classes
- Frequent association rules that have frequently observed classes.

1. Total Number of Rules Discovered

The algorithm ChSearch discovers all 265 coherent rules for mammal. In comparison, the algorithm a priori finds 20,853 association rules (based on a priori implementation release 4.27 by [12]) without the constraints mentioned in Section 6.2. Out of these, 20,588 or 98.7 percent of the association rules can be argued redundant according to logic.

2. Infrequent Rules

Rules involve classes reptile and amphibian especially cases of reptile (1) and amphibian (1) are considered infrequent rules. Theoretically, a priori could not find association rules involving them. Two coherent rules are found by our algorithm:

Table 9:

#	ChSearch	Apriori
1	$milk(1) \Rightarrow mam.(1)$, $milk(0) \Rightarrow mam.(0)$	$milk(1) \Rightarrow mam.(1)$ [S=40.6%, C=100%]
2	$hair(1) \Rightarrow mam.(1)$, $hair(0) \Rightarrow mam.(0)$	$hair(1) \Rightarrow mam.(1)$ [S=38.6%, C=90.7%]
3	$leg(4) \Rightarrow mam.(1)$, $\neg leg(4) \Rightarrow mam.(0)$	$leg(4) \Rightarrow mam.(1)$ [S=30.7%, C=81.6%]
4	$catsize(1) \Rightarrow mam.(1)$, $catsize(0) \Rightarrow mam.(0)$	$catsize(1) \Rightarrow mam.(1)$ [S=31.7%, C=72.7%]
5	$toothed(1) \Rightarrow mam.(1)$, $toothed(0) \Rightarrow mam.(0)$	$toothed(1) \Rightarrow mam.(1)$ [S=39.6%, C=65.6%]
6	Nil	$domestic(1) \Rightarrow mam.(1)$ [S=7.9%, C=61.5%]
7	Nil	$breathes(1) \Rightarrow mam.(1)$ [S=40.6%, C=51.2%]

Table 10: Contingency Table for Breathes (B) and Mammal (M)

Frequency of co-occurrences		Consequence, Y		
		Y='M'	$\neg Y$=$\neg$'M'	Total
Antecedent, X	X='B'	41	39	80
	$\neg X$=$\neg$'B'	0	21	21
	Total	41	60	101

Reported because it is contradicted by other associations rules with similar attribute such as

1. Domesticð0Þ) mam: ð1Þ, [S ¼ 80:5%, C ¼ 37:5%],
2. Domesticð0Þ) mam:ð0Þ, [S ¼ 54:5%, C ¼ 62:5%].

The above two association rules have support values higher than 7.9 percent. That is, if "not domestic" (domestic (0)) is associated with mammal, then it cannot be associated with "not mammal" logically. Hence, these rules are not reported by ChSearch as implicational association rules. In fact, a detail analysis on Zoo data set reveals that there are 41 sets of mammal (such as buffalo, bear, and elephant) but only eight of them are domestic animals. Many more In the next section, we highlight that our approach can discover rules containing infrequent items that may not be discovered based on support and confidence framework.

3. Frequent Rules

We repeat the experiment on the most observed class mammal. We list and compare the shortest rules found by ChSearch and a priori algorithms in Table of these seven shortest association rules.

VII. Conclusions

We used mapping to logical equivalences according to propositional logic to discover all interesting association rules without loss. 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:

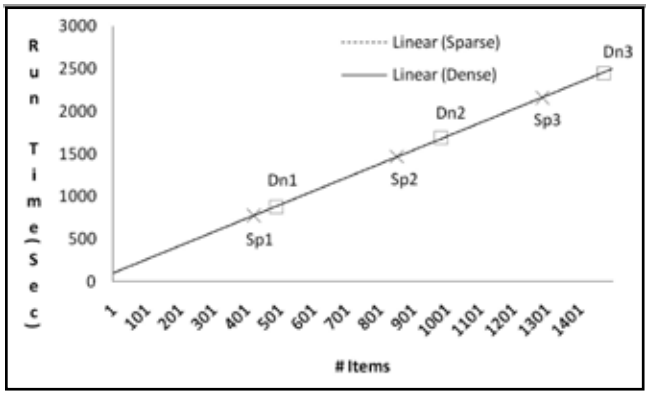

Fig. 1: Search Time on an Increase Complexity on Dense and Sparse Data Sets

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 repo

References

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

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

[3] G.I. Webb, S. Zhang,"k-Optimal Rule Discovery", Data Mining and Knowledge Discovery, Vol. 10, No. 1, pp. 39-79, 2005.

[4] E. Babbie, F. Halley, J. Zaino,"Adventures in Social Research", Data Analysis Using SPSS 11.0/11.5 for Windows", Pine Forge Press, 2003.

[5] C. Frankfort-Nachmias, A. Leon-Guerrero,"Social Statistics for a Diverse Society", Pine Forge Press, 2006.

[6] J.F. Healey, E.R. Babbie, J. Boli,"Exploring Social Issues: Using SPSS for Windows 95, Versions 7.5, 8.0, or Higher", Pine Forge Press, 1999.

[7] B. Liu, W. Hsu, Y. Ma,"Mining Association Rules with Multiple Minimum Supports," Proc. ACM SIGKDD, pp. 337-341, 1999.

[8] Y.-H. Hu,"An Efficient Algorithm for Discovering and Main- tenance of Frequent Patterns with Multiple Minimum Supports", master's thesis, Dept. of Information Management, Nat'l Central Univ., 2003.

[9] Y.S. Koh, N. Rountree, R.A. O'Keefe,"Finding Non-Coincidental Sporadic Rules Using Apriori-Inverse", Int'l J. Data Warehousing and Mining, Vol. 2, pp. 38-54, 2006.

[10] H. Mannila,"Database Methods for Data Mining", Proc. Fourth Int'l Conf. Knowledge Discovery and Data Mining (Tutorial), 1998.

[11] Y.-H. Hu, Y.-L.Chen,"Mining Association Rules with Multiple Minimum Supports: A New Mining Algorithm and a Support Tuning Mechanism", Decision Support Systems, Vol. 42, pp. 1-24, 2006.

[12] W.-Y. Lin, M.-C. Tseng, J.-H. Su,"A Confidence-Lift Support Specification for Interesting Associations Mining", Proc. Sixth Pacific-Asia Conf. Advances in Knowledge Discovery and Data Mining (PAKDD), pp. 148-158, 2002.
[13] H. Yun, D. Ha, B. Hwang, K.H. Ryu, "Mining Association Rules on Significant Rare Data Using Relative Support", J. Systems Software, Vol. 67, pp. 181-191, 2003.
[14] S. Brin, R.Motwani, C. Silverstein,"Beyond Market Baskets: Generalizing Association Rules to Correlations", Proc. 1997 ACM SIGMOD, pp. 265-276, 1997.
[15] G.I. Web,"Association Rules", The Handbook of Data Mining, pp. 26-39. Mahwah, 2003.

T. Jamalaiah pursuing M.Tech in the department of Computer Science and Engineering in MIST sathupally. His research areas data mining and data warehousing.

N. V. Kiran Babu. Movva working as an Associate Professor in the department of computer science and engineering.His research areas network security, data mining and data warewhouseing.

K.Nageswara Rao working as an Associate Professor in the department of computer scien-ce and Engineering. His research areas Computer Networks, data mining and data warehousing.