A Novel Method of Mining Association Rule with Multilevel Concept Hierarchy

S.Prakash,
Research Scholar & Assistant Professor, Department of IT, SasurieCollege of Engineering,TamilNadu, India

M.Vijayakumar
Research Scholar & Assistant Professor, Department of IT, NandhaCollege of Technology,TamilNadu, India

R.M.S.Parvathi,
Principal & Professor, Department of CSE, Sengunthar College of Engineering for women, Tamil Nadu, India

ABSTRACT
In data mining, there are several works proposed for mining the association rules which are frequent. Researchers argue that mining the infrequent item sets are also important in certain applications. Discovering association rules are based on the preset minimum support threshold given by domain experts. The accuracy in setting up this threshold directly influences the number and the quality of association rules discovered. Even though the number of association rules is large, some interesting rules will be missing and the rules quality requires further analysis. As a result, decision making using these rules could lead to risky actions. Here the focus is mainly on mining both the frequent and infrequent association rules which are more interesting and does not have redundant rules. This is based on predefined rules formed using propositional logic and then the predefined rules are processed by comparing with elements in the actual dataset. The association rules which are obtained will not have redundancies and they will be logically correct. Generalized association rules will be obtained if single level mining is performed. These rules can only help in very high level decision making. In order to allow for in-depth decision making, more specific association rules are obtained. Therefore multiple level mining processes is employed here.

KEYWORDS
Association Rule, Multilevel Association Rule Mining, concept hierarchy

1. INTRODUCTION
Data mining refers to extracting or "Mining" knowledge from large amount of data. Association rule mining finds interesting association among a large set of data items. 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 threshold value is used in the discovery process.

The key element that makes association rule mining practical is the minimum support. It is used to prune the search space and to limit the number of rules generated. However, using only a single minimum support implicitly assumes that all items in the database are of the same nature or of similar frequencies in the database. This is often not the case in real-life applications. In the retailing business, customers buy some items very frequently while others rarely. Usually, the necessities, consumables and low-price products are bought frequently, while the luxury goods, electric appliance and high-price products infrequently.

In such a situation, if set minimum support too high, all the discovered patterns are concerned with those low-price products, which only contribute a small portion of the profit to the business. On the other hand, if set minimum support too low, will generate too many meaningless frequent patterns and they will overload the decision makers, who may find it difficult to understand the patterns generated by data mining algorithms. For example, medical applications have many important symptoms and diseases that are infrequent in medical records. If the minimum support threshold is set too high, then finding such rare disease is not possible as there will not be any patterns related to that disease and if minimum support threshold is set too low, it will generate too many meaningless rules.

This problem was solved by extending the existing association rule model to allow the user to specify multiple minimum supports to reflect different natures and frequencies of items. Specifically, the user can specify a different minimum support for each item. Thus, different rules may need to satisfy different minimum supports depending on what items are in the rules. This new model enables users to produce rare item rules but specifying different supports id very difficult and requires in depth domain knowledge.

Multi-Level Association Rule Mining
If mining at multiple concept levels is considered, and each rule is restricted to associate only items at the same level of a concept hierarchy, even though the mining algorithm would search over each level of the entire hierarchy, then the rule which is obtained will be such as "milk -> bread" or "skim milk -> whole-wheat bread". But the requirement is to find a specific milk product to discount, which would boost the sales of all types of bread the most. So mining association rules which span multiple levels for this type of knowledge discovery is needed. The rules will be strong in association of items between the bottom level concepts in the sub tree rooted at milk and the top level concept of the sub tree rooted at bread. This would lead to find rules of the form "Diary land skim milk -> bread", and then the product with the highest probability of causing an increase in bread sales is chosen.
This approach proposes a novel framework to address the issues by removing the need for a minimum support threshold. Associations are discovered based on logical implications. The principle of the approach considers that an association rule should only be reported when there is enough logical evidence in the data. To do this, both presence and absence of items during the mining process is considered. This framework discovers association rules that can be mapped to different modes of logic implications in propositional logic. It also finds the coherent rules considering the item sets across different levels on a well-known domain and compares the rules found to those discovered for a single level.

2. LITERATURE SURVEY

Using 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 [18]. This is largely due to the fact that there is no universal standard to define the notion of being frequent enough and interesting. In this case one user's understanding of an ideal strength value may be different from another user's.

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 [11]. 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 Mannila. Instead of preprocessing the transaction records, using multiple minimum thresholds called minimum item supports (MISs) . Nonetheless, a user needs toprovide an MIS threshold for each item which is difficult. The common aim, however, was to offset heuristics when setting up a minimum support threshold. In all these approaches, we see that state-of-the-art association rule mining has drifted from the original idea of mining frequent patterns alone to considering other patterns as well. Using a minimum support threshold alone cannot identify these patterns specifically.

Brin [6] shows that association rules discovered using a support and confidence framework may not be correlated in statistics. These association rules show item sets co-occurring together, with no implications among them. Scheffer highlights that in many cases, users who are interested in finding items that co-occur together are also interested in finding items which are connected in reality. Having a minimum support threshold does not guarantee the discovery of interesting association rules, as such rules may need to be further processed and quantified for interestingness.

The usage of leverage and lift are good alternatives in mining association rules without relying on pruning a minimum support threshold. The authors in mined arbitrarily top k number of rules using lift, leverage, and confidence without using a preset minimum support threshold. The use of leverage and lift is also fundamental in designing a new measure of interestingness. Among such work, authors in [13] considered lift as also one of the 12 interesting criteria to generate interesting rules called an informative rule set. Authors in devised a conditional probability-like measure of interestingness based on lift (termed as dependence) called the Conditional Probability Increment Ratio (CPIR) and used this to discover required interesting rules accordingly. Apart from the use of lift, leverage, and its derived measure of interestingness, measures of interestingness that consider a deviation from independence have also been used. The authors in [3] used Pearson's correlation coefficient to search for both positive and negative association rules that have a strong correlation. This search algorithm found the strongest correlated rules, followed by rules with moderate and small strength values. The generalized framework is used 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.

Multi-level association rules are used to find the preferences for items that are not covered by the single-level association rules due to the data sparseness. Instead of considering single attribute for the rules multiple attributes are considered for obtaining more specific rules.

Problem Definition

The coherent rules are identified for the domain and based on the support value as well as the logical equivalences according to propositional logic the association rules are filtered out. The rules obtained include item sets that are frequently and infrequently observed in a set of transaction records for a single level. Here the generalized framework is obtained. The generalized framework will not help in-depth decision making as the rules obtained will be high level. Though the rules are logically correct, some scenarios occur in which the support is equal for both the rules that are coherent and here no proper decision can be made.

3. METHODOLOGY

The steps involved in the development of multilevel propositional logic based knowledge discovery involve the following phases:

- Data Preprocessing
- Propositional Logic for Coherent Rule Generation
- Multi-Level concept hierarchy
- Performance measure on generated rules

Data Preprocessing

Data preprocessing is a major step in data mining where all the unnecessary data in the datasets will be removed. Based on the nature of datasets taken for the experiment the preprocessing process will take place. If the dataset contains mostly missing attributes then those should be supplied by taking the previous observations.

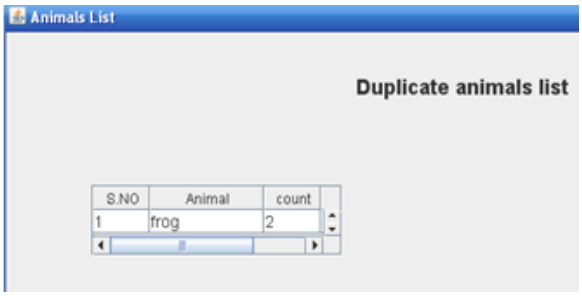


Figure 1Logic Pattern Discovery-useless attributes

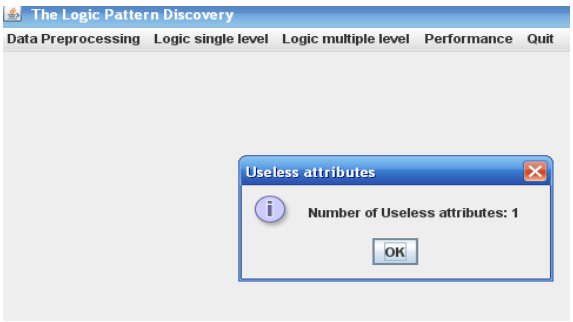


Figure 2 Logic Pattern Discovery-duplicate animals list

In this work mainly removing the duplicate records and useless attributes were performed. Figure 1 shows the logical pattern discovery for useless attributes. These are done to minimize the

amount of data to be processed and also to remove duplicate records if any with different values which may be misleading. The useless attribute is nothing but the attribute which contains unique values. Figure 2 shows logic pattern discovery for duplicate animal list

3.1 Propositional Logic for Coherent Rule Generation

The pseudo implications of equivalences are called coherent rules. Not all pseudo implications of equivalences can be created using item sets X and Y. If one pseudo implication of equivalence can be created, then another pseudo implication of equivalence also coexists. Two pseudo implications of equivalences always exist as a pair because they are created based on the same conditions. Since they share the same conditions, two pseudo implications of equivalences, coexist having mapped to two logical equivalences. The result is a coherent rule that meets the same conditions.

Coherent rules meet the necessary and sufficient conditions and have the truth table values of logical equivalence. The coherent rule consists of a pair of pseudo implications of equivalences that have higher support values compared to another two pseudo implications of equivalences. Coherent rules are defined based on logic. This improves the quality of association rules discovered because there are no missing association rules due to threshold setting. A user can discover all association rules that are logically correct without having to know the domain knowledge. This is fundamental to various application domains. For example, one can discover the relations in a retail business without having to study the possible relations among items. Any association rule that is not captured by coherent rules can be denied its importance. Figure 3 and 4 shows animal support list and rule list respectively

S.NO	type	attrib	count
22	1	1	39
23	1	0	2
24	2	0	20
25	3	0	5
26	4	0	13
27	5	0	4
28	6	0	4
29	6	1	4
30	7	0	10
31	1	0	41
32	2	1	20

Figure 3 Support List

S.NO	type	attrib
15	Mammal	hairPresent
16	Insect	hairPresent
17	Insect	hairAbsent
18	Bird	hairAbsent
19	Reptile	hairAbsent
20	Fish	hairAbsent
21	Amphibian	hairAbsent
22	Invertebrate	hairAbsent
23	Mammal	feathersAbsent
24	Bird	feathersPresent
25	Reptile	feathersAbsent

Figure 4 Rule list

3.2 Multi-Level Concept Hierarchy

In multilevel association rules mining, different propositional logic is used at different concept levels. Discover frequent patterns and strong association rules at the top-most concept level. With this user can find a set of pair-wised frequent items and a set of association rules at each level. The process repeats at even lower concept levels until no frequent patterns can be found. During multilevel association rule mining, the taxonomy information for each (grouped) item is encoded as a sequence of digits in the transaction table. Repeated items (i.e., items with the same encoding) at any level will be treated as one item in one transaction.

In the proposed concept hierarchy model, items may have different propositional logic and taxonomic relationships to discover the large item sets. The propositional logic for an item set is set as the combinatorial sub logic supports of the items contained in the item set, while the propositional logic for an item at a higher taxonomic concept is set as the minimum sub logics of the items belonging to it. Encoding scheme represents nodes in the predefined taxonomies for mining multilevel rules. Figure 5 shows the multiple attribute rule list.

Nodes are encoded with respect to their positions in the hierarchy using sequences of numbers and the symbol. It then filters out unpromising item sets in two phases. In the first phase, an item group is removed if its occurring count is less than the propositional logic. In the second phase, the count of a propositional logic rules is checked to determine whether it is large. The proposed algorithm then finds all the large item sets for the given transactions by comparing the count of each item set with its combinatorial logic.

3.3. Performance Measure On Generated Rules

The employed multilevel concept hierarchy based association rule mining using propositional logic is experimented with synthetic data to evaluate the association rule and coherent rules without having the domain knowledge. The results are also compared with logic based pattern discovery model with single level association rule mining model in terms of number of strong rules and weak rules generated. Effects of non sensitive rules are derived in both the existing and proposed schemes to show the efficiency of novel association rule mining process.

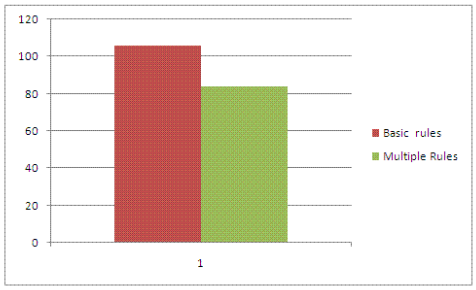


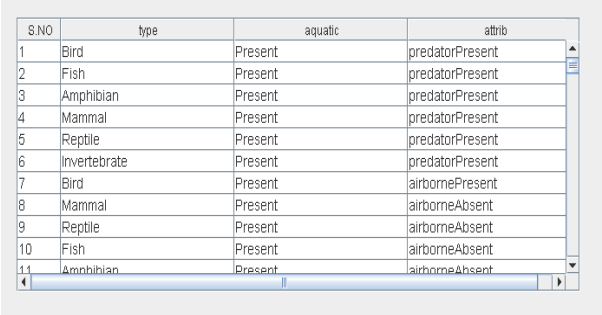
Figure 5 Multiple Attribute Rule List

The models works well with problems involving uncertainty in data relationships, which are represented by multilevel concepts for propositional logic rule derivation. The proposed mining algorithm can thus generate large item sets level by level and then derive concept multilevel association rules from transaction dataset. The results shown in the example implies that the proposed algorithm can derive the multiple-level association rules under different propositional logic in a simple and effective way.

4. SYSTEM IMPLEMENTATION

The proposed system is very easy to implement. In general implementation is used to mean the process of converting a new or revised system design into an operational one. The implementation can be preceded with UCI data repository through Swing in java which supports numerous look and feels, including the ability to create own look and feel. The ability to create a custom look and feel is made easier. The main characteristics of the Swing toolkit are platform independent, customizable, extensible, configurable and lightweight. Comparison between single level association rules and multilevel association rule are shown in figure 6

The implementation phase is less creative than system design. It is primarily concerned with user training, and file conversion. The system may be requiring extensive user training. The initial parameters of the management information system should be modifies as a result of a programming. The system developed is completely menu driven. Further a simple operating procedure is provided so that the user can understand the different functions clearly and quickly.



S.NO	type	aquatic	attrib
1	Bird	Present	predatorPresent
2	Fish	Present	predatorPresent
3	Amphibian	Present	predatorPresent
4	Mammal	Present	predatorPresent
5	Reptile	Present	predatorPresent
6	Invertebrate	Present	predatorPresent
7	Bird	Present	airbornePresent
8	Mammal	Present	airborneAbsent
9	Reptile	Present	airborneAbsent
10	Fish	Present	airborneAbsent
11	Amphibian	Present	airborneAbsent

Figure 6 single level association rules Vs Multilevel association rule in terms of rule generation

5. CONCLUSION

The proposed multilevel association rule mining 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. These association rules reduce the risks associated with using an incomplete set of association rules for decision making. The multilevel coherent rules mining framework is appreciated for its ability to discover rules that are both implicational and complete according to propositional logic from a given data set at various conceptual levels.

This work can be enhanced by taking an unclassified dataset and perform classification before performing the experiment. But this classification should be appropriate in such a way that the rules obtained in the result should also be logically true when they are compared with the original dataset before classification.

6. REFERENCES

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

[2] Alex TzeHiangSim, Maria Indrawan, SamarZutshi and BalaSrinivasan(2010) “Logic Based Pattern Discovery” IEEE Transactions on Knowledge and Data Engineering, VOL. 22, NO. 6, pp.798-811.

[3] Antonie M.-L. and Zaýane O.R. (2004) “Mining Positive and Negative Association Rules: An Approach for Confined Rules” Proc. European Conf. Principles and Practice of Knowledge Discovery in Databases (PKDD ’04), pp. 27-38.

[4] Bing Liu, Minqing Hu, and Wynne Hsu (2000) “Multi-Level Organization and Summarization of the Discovered Rules” Proc. ACM SIGKDD, Aug 20-23.

[5] Blanchard J., Guillet F., Briand H. and Gras R. (2005) “Assessing Rule Interestingness with a Probabilistic Measure of Deviation from Equilibrium” Proc. 11th Int’l Symp. Applied Stochastic Models and Data Analysis (ASMDA ’05), pp. 191-200.

[6] Brin S., Motwani R. and Silverstein C. (1997) “Beyond Market Baskets: Generalizing Association Rules to Correlations” Proc. 1997 ACM SIGMOD, pp. 265-276.

[7] Cao Longbing (2008) “Introduction to Domain Driven Data Mining,” Data Mining for Business Applications, Springer, pp. 3-10.

[8] Choonho Kim and Juntae Kim (2003) ”A Recommendation Algorithm Using Multi-Level Association Rules” Proc. IEEE/WIC Int’l Conf. on Web Intelligence page. 52.

[9] Hellerstein J.L., Ma S., Perng C (2002) ”Discovering actionable patterns in event data” IBM Systems Journal, Vol 41, NO 3.

[10] KanimozhiSelvi C.S. and Tamilarasi A. (2009) ”An Automated Association Rule Mining Technique With Cumulative Support Thresholds”Int. J. Open Problems in Compt. Math, Vol. 2, No. 3.

[11] Koh Y.S., Rountree N. and O’Keefe R.A (2006) “Finding Non-Coincidental Sporadic Rules Using Apriori-Inverse” Int’l J. Data Warehousing and Mining, vol. 2, pp. 38-54.

[12] Laszlo Szathmary, Amedeo Napoli and PetkoValtchev (2007) “Towards Rare Itemset Mining” Proc. of the 19th IEEE ICTAI, Vol. 1,pp. 305-312.

[13] Li J. and Zhang Y. (2003) “Direct Interesting Rule Generation” Proc. Third IEEE Int’l Conf. Data Mining, pp. 155-162.