Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 45 (2015) 143 – 149

Procedia Computer Science

International Conference on Advanced Computing Technologies and Applications (ICACTA-2015)

Removal of duplicate rules for Association Rule Mining from multilevel dataset

A K Chandanan[a],M K Shukla[b]

[a]*Department of Computer Science and Information Technology,JVW University,Jaipur,India*
[b] *Department of Computer Science & Engineering, Sunder Deep Group of Institution, Ghaziabad, India*

Abstract

Association rules are one of the most researched areas of data mining. This is useful in the marketing and retailing strategies. Association mining is to retrieval of a set of attributes shared with a large number of objects in a given database. There are many potential application areas for association rule approach which include design, layout, and customer segregation and so on. The redundancy in association rules affects the quality of the information presented. The goal of redundancy elimination is to improve the quality and usefulness of the rules. Our work aims is to remove hierarchical duplicacy in multi-level, thus reducing the size of the rule set to improve the quality and usefulness without any loss.

Peer-review under responsibility of scientific committee of International Conference on Advanced Computing Technologies and Applications (ICACTA-2015).

*Keywords:*Association Rule,Duplicate Rules,Multi-level and quantative data

1. Introduction

Association rule discovery [4], discover all frequent pattern among all transaction of data attributes. Discovery of frequent pattern is presented in form of rules [5]. The findings are presented in the form of rules between different sets of items, along with measuring metrics i.e. joint and conditional probabilities of the pre-existing and subsequent, the metrics are used to judge a rule's importance in the mining [7]. A closed set is use to present the item set [6] which contains its own boundary (limit points). If you are outside of a closed set, it means you may move a small amount in any direction by this you will be still outside of the set..

- Intersection of any closed set is a closed.

Peer-review under responsibility of scientific committee of International Conference on Advanced Computing Technologies and Applications (ICACTA-2015).
doi:10.1016/j.procs.2015.03.106

- The union of many closed set is a closed.
- The empty set is also a closed.
- The whole set is a closed.

Sequential pattern mining, which is the process of extracting certain sequential patterns whose support exceeds a predefined minimum support threshold value, has been studied widely in the last decade in the data mining community However, and less work, has been done on sequential association rule mining[1][3].

Only in recent years, several prediction models which introduced the concept of sequential association rule mining have been proposed [1], most of which use sequence and temporal constraints in generating association rules. In the classical association rule mining [2], the resulting rule set can easily contain thousands of rules in which many of the rules are redundant and are useless in practical aspects. While in the case of sequential association rule mining, the same set of items with different ordering yields different sequential patterns in sequential pattern mining which makes the number of frequent sequential patterns usually much larger than the number of frequent item sets generated from a dataset of a similar size. When rules are in rapid growth for the set of association rules, especially as we lower the frequency requirements. The larger frequent item sets causes for more the number of rules to be generated for the dataset, many of which are redundant. These existing approaches mainly discuss how to efficiently generate sequential patterns, and do not pay much attention to the quality of the discovered patterns, in particular, all of these approaches suffer from the problem that the volume of the discovered patterns and association rules could be exceedingly large, but many of the patterns and rules are actually redundant and thus need to be pruned.

2. Related Work

One approach to address the quality of association rules is to apply constraints to generateonly those associati on rules that are interesting to users. Both [8] and[9] proposed algorithms that incorporate item constraintsto the pro cess of generating frequent itemsets. Some work has also been done on measuringassociation rules with interestingn ess parameters [15]. These approachesfocus on pruning the association rules to get more general or informativ e association rulesbased on interestingness parameters. The approach proposed in[11][16] integrates various cons traints into the mining process includingconsequent constraint and minimal improvement constraint.The cons equent constraint iused to restrict rules with certain consequent specified by the user.The minimalimprov ement constraint is used to simplify the antecedents of rules based on items'contribution to the confidence and therefore prune association rules that have more specificantecedent but do not make more contribution to the confi dence. Another approach is to usea taxonomy of items to extract generalized association rules [12], i.e., togenerate rul es between itemsets that belong to different abstract levels in the taxonomy,especially between high abstract le vels, aiming to reducing the number of extracted rules from dataset .The approaches mentioned above aim to reduce the number of extracted rules and alsoimprove the “usefulness ” of these rules, but eliminating redudancy in association rules isnot their focus. The approaches proposed in [13] and [14]focus on extracting nonredundant itemsets and association rules.

2.1 Limitation with herariechal data

There are some limitation to remove duplicate rule from multi leve dataset . here we are taking some assumptions as:

Table 2.1 assumptions and Limitations

Assumptions	Limitations
•Data taxonomy belong to tree structure with single root •All node of tree have single parent excluding root •Each edge in tree structure with equal weight value	• Approach is based on effectiveness not in efficiency • Approach deals with hierarchical structure only

2.2 Remove of duplicate rules from multilevel data

A multilevel dataset is one which has an implicit taxonomy or concept tree, like the exampleshown in Fig. 1. The items in the dataset exist at the lowest concept level but are part of ahierarchical structure and organization. Thus for example, 'ME' is an item at thelowest level of the taxonomy but it also belongs to the high level concept category of 'Scie nce' and also the more refined category 'Engg'.Each entry in the hierarchy has oneparent (or immediate superto pic) with a path back to the root possible from any where in thehierarchy. The hierarchy information can be encode d with each topic allowing information abouta given topic's ancestry. For example, 'ME' can be encoded as 1_1_2.

This first digit in the sequence '1' indicates that it belongs to first category in the first level concept. The second secquence digit '1' indicates that in belongs to first category in the second level concept under belonging category with one level upper. The third digit '2' in sequence ponts to second category in the third level concept under the category from above one level from current level and so on. As per the assumption made the order of the siblings in the this taxonomy is not so important. Thus in this structure the node 'ME'is encoded as 1_1_2 but if made to the first node under the 'Engg' then it would be encoded as 1_1_* and the node 'CSE ' would then be encoded as 1_1_1 . the encoding is done in a simple left to right manner, because of the tree nature of the multi-level dataset a defferent approach to finding frequent itemsets is needed as standard Apriori approach does not take the tree structure in to consideration

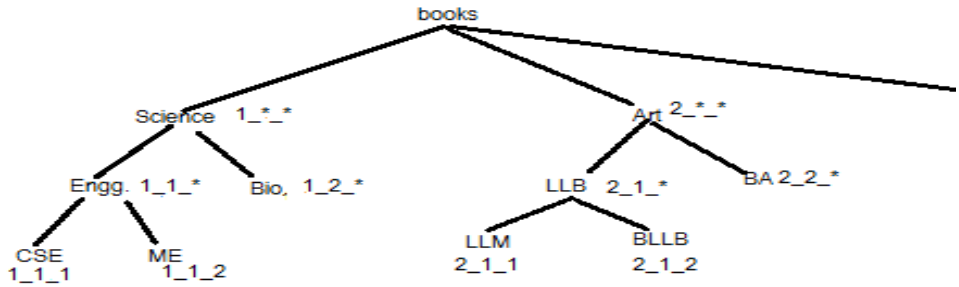

Fig. 1. An Example of Taxonomy Multilevel dataset

2.3 Duplicate rules in multilevel data set

To show the result for multilevel frequent itemset approach in icluding Apprioi approach in explained with using following trasaction table (Table 2.2)

Table 2.2. Simple multilevel transactions.

Trans No	Items in transaction
1	[1_1_1, 1_2_1, 2_1_1, 2_2_1]
2	[1_1_1, 2_1_1, 2_2_2, 3_2_3]

3	[1_1_2, 1_2_2, 2_2_1, 4_1_1]
4	[1_1_1, 1_2_1]
5	[1_1_1, 1_2_2, 2_1_1, 2_2_1,]
6	[1_1_3, 3_2_3, 5_2_4]
7	[1_3_1, 2_3_1]
8	[3_2_3, 4_1_1, 5_2_4, 7_1_3]

This simple multilevel dataset has 3 concept levels with each item from higher level belonging to the lowest level. The item ID in the table stores/holds the hierarchy(tree structure) information for each item of the ataset. Thus the item 1_2_1 belongs to the first category at concept level 1 and for level 2 it belongs to the second subcategory of the first level 1 category. Finally at concept level 3 it belongs to the first subcategory of its parent category at level 2.As can be seen in Table 2.3 the contents of all of the frequent itemsets come from the same level,that is within a given itemset all the items come from the one concept level.

Table 2.3. Frequent itemsets derived from Heirarical Dataset

Single - itemsets pair	Two - itemsets pair	Three -itemsets pair
[1_*_*]	[1_*_*, 2_*_*]	[1_1_*, 1_2_*, 2_2_*]
[2_*_*]	[1_1_*, 1_2_*]	[1_1_*, 2_1_*, 2_2_*]
[1_1_*]	[1_1_*, 2_1_*]	
[1_2_*]	[1_1_*, 2_2_*]	
[2_1_*]	[1_2_*, 2_2_*]	
[2_2_*]	[2_1_*, 2_2_*]	
[1_1_1]	[1_1_1, 2_1_1]	
[2_1_1]		
[2_2_1]		

2.3.1Taxonomy of hierarichal data

The use of frequent itemsets as the basis for association rule mining often results in the generationof a large numb er of rules. This is a widely recognized problem. Recent work has demonstratedthat the use of closed itemsets an d generators can reduce the number of rules generated . This has helped to greatly reduce redundancy in the rule s derived fromsingle level datasets. Despite this, redundancy still exists in the rules generated from multileveldatasets even when using some of the methods designed to remove redundancy. This redundancywe call hierarchical redundan cy. Here in this section we will introduce hierarchical redundancyin multilevel datasets and show that existing appro aches do not remove this type of redundantrules. Hierarchical redundancy only exists in multilevel datasets and exists purely because thereis a hierarchical or taxonomic structure around which a dataset or database is organised. Reliable ApproximateRules approaches generate the basis rules. Thediscovered rules are from multiple levels and can i nclude crosslevel rules, due to crosslevelfrequent itemsets .The ReliableExactRule andReliableApproximateRul e approaches can remove redundant rule, but as we will show, it doesnot remove hierarchy redundancy. The rules give n in Tables 2.4, 2.5, For theseexamples the minimum confidence threshold is set to 0.50 or 50% for the dataset .

Table 2.4. Multi-level association rules on heirarical dataset

No.	Rule Extract on extact Basis Rules	Supp	Conf
1	[1_2_*] ==> [1_1_*]	0.571	1.0
2	[2_2_*] ==> [1_1_*]	0.571	1.0
3	[2_1_1] ==> [1_1_1]	0.428	1.0
4	[2_1_*] ==> [1_1_*, 2_2_*]	0.428	1.0
No.	Rule Extract on Approximate Basis Rules	Supp	Conf
1	[1_1_*] ==> [1_2_*]	0.571	0.666
2	[1_1_*] ==> [2_2_*]	0.571	0.666
3	[1_1_1] ==> [2_1_1]	0.428	0.75
4	[1_1_*] ==> [1_2_*, 2_2_*]	0.428	0.5

5	[1_2_*] ==> [1_1_*, 2_2_*]	0.428	0.75
6	[2_2_*] ==> [1_1_*, 1_2_*]	0.428	0.75
7	[1_1_*] ==> [2_1_*, 2_2_*]	0.428	0.5
8	[2_2_*] ==> [1_1_*, 2_1_*]	0.428	0.75

Finally, if we used a standard Apriori approach for finding the frequent itemsets from thetransaction dataset F closed itemsets and generators we are able togenerate the following rules using the ReliableExactRule and th e ReliableApproximateRule approaches as shown in Table 2.5.

Table 2.5. Exact basis and approximate basis association rules derived from a standard Aprioriapproach.

No.	Rule Extract on extact Basis Rules	Supp	Conf
1	[2_1_1] ==> [1_1_1]	0.374	1.0
No.	Rule Extract on Approximate Basis Rules	Supp	Conf
1	[1_1_1] ==> [2_1_1]	0.374	0.751

As shown in Table 2.5, there is only one exact basis rule and one approximate basis rule that canbe derived from a flat version of the transactional dataset shown in Table 2.2. This shows thathaving a hierarchy or taxonomy increase the number of frequent itemsets and therefore thenumber of association rules that can be derived when a multievel and/or crosslevel approach isused.This redundancy comes purely from the dataset or database having multile con cept levelsthrough a hierarchy or taxonomy. In a flat dataset all of the items are at one single concept leveland thus the items are all unrelated. In a multilevel dataset, topics and / or items can be acrossseveral concept levels. Thus we hav e supertopics (containing smaller more specific topics) andsubtopics . Because of this, topics now haverelations amo ngst themselves. These relations introduce what we call hierarchical redundancy,which we aim to remove from th e basis rule sets.Our example onlyowed exact basisassociation rules, however, this hierarchical redundancy can be found in approximate associationrules.

2.3.2 Non-Redundant Multi-Level Exact Rules

In this section, we propose a new definitions , which helps to determine and eliminate hierarchically redundant association rules from tree structured data set. First of all We will give our definition of hierarchical redundancy, then we apply the definition to existing non-redundant rule extraction approaches. Finally, we write an algorithm to implement these improved approaches. We also detail a recovery algorithm which allows all the exact rules deemed to be hierarchically redundant to be recovered from the hierarchically non-redundant exact rule set for improved approaches[16].

Table 2.6 Definitions

exact rule generation for hierarchical redundancy	***aprox rule generation for hierarchical redundancy***
Let us consider two rules	Let us consider two rules
(a) R1: X1→Y	(a) R1: X1→Y with confidence value C1
(b) R2: X2→Y	(b) R2: X2→Y with confidence value C2
Both (a) and (b) are exactly same rule in dataset Y.R1 is duplicating to R2 if:	Both (a) and (b) are very close to actual association rule which derived for dataset Y. R1 is duplicating to R2 if:
(i)X1 is made up with items such that at least one item present in X1 is inherited is available in X2	(i) X1 is made up with items such that at least one item present in X1 is inherited is available in X2
(ii)X2 is made up with items such that at least one item available in X2 is ancestor of item available in X1	(ii)X2 is made up with items such that at least one item available in X2 is ancestor of item available in X1
	(iii) Non-ancestor item available in X2 all are available

	in item set X1. (iv) the relation between confidence of R1 and R2 : R1(C1) ≤ R2(C2)

Algorithm for no duplicate multilevel rules

From definition 2.1 we can now develop the necessary algorithms to implement ourproposed enhanced approaches for deriving nonredundant exact association rules.For thefollowing algorithms, c is a closed itemset, C is the set of closed itemsets, g is a generator and G is the set of generators.

Input: Set of exact basis rules & set of frequent closed item sets
Output: Set of multi-level association rules that covers the basis set and the hierarchically redundant set.

1. recovered ϵ Ø
2. for all $r \in exactbasis$
3. candidatebasis rules ← Ø
4. determine if any of the items x in the antecedent X of rule
   r: X => Y are the ancestor of any generator g in the list of
   generators G and if so store g in list A
5. determine all of the possible subsets of list A and store as S
6. for all $s \in S$ check to ensure every $x \in X$ for rule r has a
   descendant in s and if not add x to s so that $s \in x$
7. if s has no ancestors in Y & s has no descendants in Y &
   for all items $i \in s$ there are no ancestor-descendant relations
   with item $i' \in s$ & for all items $i \in Y$ there are no ancestor-
   descendant relations with item $i' \in Y$
8. insert {r : s => Y} in candidatebasis rules
9. end loop
10. if for all $x \in X$ test to see that they have a descendant item $i \in A$
    and if not add x to A
11. if A has no ancestors in Y & A has no descendants in Y & for all
    $i \in A$ there are no ancestor-descendant relations with item $i \in A$
    & for all items $i \in Y$ there are no ancestor-descendant relations
    with item $i' \in Y$
12. insert {r : A => Y} in candidatebasis
13. for all c : B => D ∈ candidate basis rules
14. if $B * D = itemset\ i \in closed\ itemset\ list\ C\ \&\ B = g \in Gi$
15. insert {r : B => D, g.supp} in recovered
16. end loop
17. end loop
18. return exactbasis ∗ recovered

Fig. 2.2 Algorithm to recover hierarchically duplicate association rules .

3. Conclusion

Redundancy in association rules mining decreases the speed for rules generation. It cause so many rules to be generated for same set of attribute. Our goal was to remove redundancy in the item set it reduce unnecessary time utilization of algorithm. With use of above designed algorithm redundant rules for the large /multilevel dataset could be removed. By reducing the set of rule improve the quality and efficiency of mining without any loss of

information. We have proposed an approach to remove redundancy using upper level closed frequent item set and with the help of generator.

References

1. Agrawal R, R Srikant . Mining sequential patterns. *Proceedings of the Eleventh International Conference on Data Engineering:* 1995.
2. Kotsiantis S,D Kanellopoulos. Association Rules Mining: A Recent Overview. *GESTS International Transactions on Computer Science and Engineering:* 2006. Vol.32 (1): p. 71-82.
3. Gaul W ,L Schmidt-Thieme . Mining Generalized Association Rules for Sequential and Path Data. *Proceedings of the 2001 IEEE International Conference on Data Mining*: 2001. 3
4. Ceglar A., J F Roddick . Association Mining. *ACM Computing Surveys (C SUR):*2006. Vol 38(2): p. 13. 4
5. Tanna P, Y Ghodasara. Foundation for Frequent Pattern Mining Algorithms Implementation. *International Journal of Computer Trends and Technology (IJCTT):* July 2013 .Vol 4(7). 5
6. http://en.wikipedia.org/wiki/Closed_set 8
7. Raheja N, R Kumar.Optimization of Association Rule learning in distributed database using clustering techniques .*International Journal on Computer Science and Engineering (IJCSE):* Dec 2012. ISSN: 0975-3397, Vol. 4(12) p 1874-1880. 6
8. Ng R T,V. S Lakshmanan, J Han, A Pang. Exploratory mining and pruningoptimizations of constrained association rules . *ACM SIGMOD Record:* 1998. Vol 27(2): p.13-24. 7
9. Srikant R,Q Vu, R Agrawal. Mining Association Rules with ItemConstraints. *Paper presented at the 3rd International Co ference on Knowledge Discovery:* Aug 1997. 9
10. Bayardo R J, RAgrawal,D Gunopulos. ConstraintBased Rule Mining in Large,Dense Databases. *Data Mining and Knowledg e Discovery:* 2000. Vol 4(23): p. 217-240. 10
11. Han J, Y Fu . Mining MultipleLevel Association Rules in Large Databases. *IEEETransactions on Knowledge and Data Engin eering:* 2000.,*Vol 11*(5) : p.798-804 11
12. Zaki M. J. Mining Non-Redundant Association Rules. *Data Mining and Knowledge Discovery* : 2004.Vol 9:p. 223-248. 12
13. Pasquier N, RTaouil, YBastide,GStumme, L Lakhal. Generating a CondenseRepresentation for Association Rules. *Journa l of Intelligent Information Systems:* 2005.Vol 24(1):p. 29-60. 13
14. Chandanan A K, M K Shukla .Review on Redundancy free Association Rule mining. *IJCA Proceedings on International Conference on Communication Technology ICCT (5)*:13-16, October 2013.
15. Shaw Gavin ,Xu Yue , Geva, Shlomo . Deriving Non-Redundant Approximate Association Rules from Hierarchical Datasets. *In: 17th ACMConference on Information and Knowledge Management* ; 26-30 October 2008, Napa Valley, USA.
16. Shrivastava Neeraj, Swati Lodhi Singh.Overview of Non-redundant Association Rule Mining. *Research Journal of Recent Sciences;* Feb 2012. Vol. 1(2): ISSN 2277-2502: p.108-112.