Deriving Non-Redundant Approximate Association Rules from Hierarchical Datasets

Gavin Shaw
Faculty of Information Technology
Queensland University of Technology
Brisbane, Australia
gavin.shaw@qut.edu.au

Yue Xu
Faculty of Information Technology
Queensland University of Technology
Brisbane, Australia
yue.xu@qut.edu.au

Shlomo Geva
Faculty of Information Technology
Queensland University of Technology
Brisbane, Australia
s.geva@qut.edu.au

ABSTRACT

Association rule mining plays an important job in knowledge and information discovery. However, there are still shortcomings with the quality of the discovered rules and often the number of discovered rules is huge and contain redundancies, especially in the case of multi-level datasets. Previous work has shown that the mining of non-redundant rules is a promising approach to solving this problem, with work by [6,8,9,10] focusing on single level datasets. Recent work by Shaw et. al. [7] has extended the non-redundant approaches presented in [6,8,9] to include the elimination of redundant exact basis rules from multi-level datasets. Here we propose a continuation of the work in [7] that allows for the removal of hierarchically redundant approximate basis rules from multi-level datasets by using a dataset's hierarchy or taxonomy.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications – *Data Mining*.

General Terms

Algorithms.

Keywords

Multi-level datasets, redundant association rules.

1. INTRODUCTION & RELATED WORK

Since its introduction in [1], association rule mining is now an important and widely used data mining technique. The aim of this technique is to extract frequent patterns, interesting co-occurrences and associations amongst sets of items in large transactional databases. Traditionally there has been two steps in obtaining association rules: determining the frequent patterns or itemsets and generating the rules from these frequent patterns/itemsets. Often too many association rules containing redundancies are discovered which too often become overwhelming and difficult to comprehend. Through the use of frequent closed itemsets the issue of redundancy can be dealt with by deriving non-redundant association rules [6,8,9,11]. However, this work has only dealt with redundancy in single level datasets. Multi-level datasets (in which the items are not all at the same concept level) contain information at different levels and to obtain it all, techniques that take all the levels into account are needed [2,3,4,5]. Rules derived from multi-level datasets can also have the same issues with redundancy as those from single level datasets. While existing approaches used to remove redundancy in single level datasets [6,9,10] can be adapted for use in multi-level datasets, they still fail to remove all of the redundancies present, namely the redundancy of hierarchy, where one rule at a given level gives the same information as another rule at a different level.

This paper looks into hierarchical redundancy for approximate basis association rules (which have a confidence of less than 1) and proposes a continuation and extension of the work in [7]. From this a more concise non-redundant approximate basis rule set can be derived.

CIKM'08, October 26–30, 2008, Napa Valley, California, USA.
ACM 978-1-59593-991-3/08/10.

2. GENERATING NON-REDUNDANT APPROXIMATE BASIS RULES

Whether a rule is interesting and/or useful is usually determined through the support and confidence values that it has. However, this does not guarantee that all of the rules that have a high enough support and confidence actually convey new information.

For example, the item 1-1-1 (rule 1) is a descendant of the more general/abstract item 1-*-* (rule 2). If we know that rule 2 says 1-*-* is enough to fire the rule with consequent C, whereas rule 1 requires 1-1-1 to fire with consequent C, any item that is a descendant of 1-*-* will cause a rule to fire with consequent C. It does not have to be 1-1-1. Thus rule 1 is more restrictive. Because 1-1-1 is part of 1-*-* having rule 1 does not actually bring any new information to the user, as the information contained in it is actually part of the information contained in rule 2. Thus we consider rule 1 to be redundant.

The exception to this would be if a rule which would normally be considered redundant in fact has a higher confidence value than the rule it is being considered redundant to. Since approximate association rules are measured by their confidence, which indicates their strength, trustworthiness, accuracy and/or reliability, it is important to ensure those rules with a high confidence are kept. Thus rule 1 (1-1-1 ==> 2-2-*, 2-1-1) would normally be considered redundant to rule 2 (1-*-* ==> 2-1-1, 2-2-*) as the antecedent of rule 1 is a descendant of the antecedent of rule 2. However, if the confidence of rule 1 is 0.75, while rule 2 has a confidence of only 0.6, we have more confidence in that rule 1 is correct than rule 2. Because of this, rule 1 should be kept in the approximate basis rule set and should not be considered redundant.

From the previously described details for hierarchical redundancy in approximate basis rule sets, we propose the following definition for hierarchical redundancy in approximate basis association rules.

Definition 1 (Hierarchical Redundancy for Approximate Basis): Let $R_1 = X_1 \Rightarrow Y$ with confidence C_1 and $R_2 = X_2 \Rightarrow Y$ with confidence C_2 be two approximate association rules, with exactly the same itemset Y as the consequent. Rule R_1 is redundant to rule R_2 if (1) the itemset X_1 is made up of items where at least one item in X_1 is descendant from the items in X_2 and (2) the itemset X_2 is entirely made up of items where at least one item in X_2 is an ancestor of the items in X_1 and (3) the other non-ancestor items in X_2 are all present in itemset X_1 and (4) the confidence of R_1 (C_1) is less than or equal to the confidence of R_2 (C_2).

3. RESULTS

As can be seen, the use of our approach has reduced the approximate basis rule set for nearly all cases shown here. In some instances the basis set was only reduced by a few rules, but in other cases there was a more significant reduction in the size of the basis set. For example, in Table 1 for dataset T4 there was a reduction of 1583 rules from 6427 to 4844, which is about 24.6%. Also a reduction of around 25% for dataset H1 was achieved and around 18.1 to 13.7% for dataset T3. For other datasets the reduction is between about 11 to 16%. By using this approach we have successfully reduced the size of the approximate basis without losing any information as all algorithms successfully recover all of the approximate rules.

Table 1. Results for built datasets where ML_T2L1 with cross level add-on is used to extract frequent itemsets.

	Approximate Basis						
Data set	MMA	MMA with HRR	%	RAB	RAB with HRR	%	Approx Rules
H1	36	27	25	35	26	25	68
T1	181	161	11	166	146	12	2047
T2	700	587	16	398	347	12	1447
T3	2546	2085	18	1608	1387	13	4332
T4	6427	4844	24	3415	2970	13	7267

For the above table (Table 1), MMA and RAB refer to existing algorithms presented in [6,10] respectively. MMA with HRR and RAB with HRR are our extended algorithms (based on the algorithms presented in [6,10]) that have been enhanced to remove hierarchical redundancy.

4. CONCLUSION

Redundancy in association rules affects the quality and usefulness of the information presented in a rule set. The goal of redundancy elimination is to improve the quality and use of the rules by reducing the number of rules. Our work aims to remove hierarchical redundancy in multi-level datasets, thus reducing the size of the rule set to improve the quality and usefulness, without causing the loss of any information. We have proposed an approach which removes hierarchical redundancy on top of removing non-hierarchical redundancy through the use of frequent closed itemsets, generators and a dataset's hierarchy/taxonomy.

5. REFERENCES

[1] R. Agrawal, T. Imielinski & A. Swami, 'Mining Association Rules between Sets of Items in Large Databases', in ACM SIGMOD International Conference on Management of Data (SIGMOD'93), Washington D.C., USA, 1993, pp 207-216.

[2] J. Han & Y. Fu, 'Mining Multiple-Level Association Rules in Large Databases', IEEE Transactions on Knowledge and Data Engineering, Vol 11, pp 798-805, Sep/Oct, 1999.

[3] T.-P. Hong, K.-Y. Lin & B.-C. Chien, 'Mining Fuzzy Multiple-Level Association Rules from Quantitative Data', Applied Intelligence, Vol 18, pp 79-90, Jan, 2003.

[4] M. Kaya & R. Alhajj, 'Mining multi-cross-level fuzzy weighted association rules', in 2nd International IEEE Conference on Intelligent Systems, 2004, pp 225-230.

[5] B. Liu, M. Hu & W. Hsu, 'Multi-Level Organization and Summarization of the Discovered Rules', in Conference on Knowledge Discovery in Data, Boston, Massachusetts, USA : ACM Press, 2000.

[6] N. Pasquier, R. Taouil, Y. Bastide, G. Stumme & L. Lakhal, 'Generating a Condensed Representation for Association Rules', Journal of Intelligent Information Systems, Vol 24, pp 29-60, 2005.

[7] G. Shaw, Y. Xu & S. Geva, 'Eliminating Redundant Association Rules in Multi-level Datasets', in Proceedings of the 4th International Conference on Data Mining (DMIN'08), 2008, Las Vegas, USA, To appear.

[8] Y. Xu & Y. Li, 'Mining Non-Redundant Association Rules Based on Concise Bases', International Journal of Pattern Recognition and Artificial Intelligence, Vol 21, pp 659-675, Jun, 2007.

[9] Y. Xu, & Y. Li, 'Generating Concise Association Rules', in 16th ACM Conference on Conference on Information and Knowledge Management (CIKM'07), Lisbon, Portugal, 2007, pp 781-790.

[10] Y. Xu, Y. Li & G. Shaw, 'Concise Representations for Approximate Association Rules', in Proceedings of the 2008 IEEE International Conference on Systems, Man & Cybernetics (SMC'08), 2008, Singapore, To appear.

[11] M. J. Zaki, 'Mining Non-Redundant Association Rules', Data Mining and Knowledge Discovery, Vol 9, pp 223-248, 2004.