Association Rule Representation

SAP DEVELOPER NETWORK

Applies To:

Conceptually Business Intelligence and developed in ABAP.

Summary

Association rule mining forms an important research area in the field of data mining. Usually association rules generate large number of rules, and it is very difficult to identify and deduce useful rules. We described an approach to represent this large number of rules in a simple and understandable manner without loss of any useful information. We implemented this approach on SAP ABAP/4. To show the viability of our approach, the experiments are conducted on credit dataset and an insurance database.

By: K. Sai Krishna

Company: Cognizant Technology Solutions

Date: 27 February 2006

Table of Contents

Introduction

In data mining, association rules describe the interesting relation among different attributes. These rules are of the form A→B where A and B are disjoint subsets. The main goal of association rules is to discover rules that have the support and confidence greater than or equal to minimum support and minimum confidence respectively. The 'Support' is the count of the itemset, which appears with respect to the entire dataset, while 'Confidence' is the count of rule when the consequent occurs with respect to the antecedents. Minimum support and Minimum confidence values are user specified inputs.

Agrawal et al [1] has been proposed an Apriori algorithm in 1993 to generate association rules. Classification of association rules includes Boolean, Quantitative and Fuzzy association rules [2,6,7,10,15]**.** Boolean Association Rules are used to find associations among attributes in a dataset, which have values either 0 or1.

The rest of the paper is organized as follows. In section 2, we present an algorithm to convert large number of rules to less number of rules. In section 3, we present an example for our approach. Experimental results on credit and insurance datasets have been shown in section 4. Finally, we conclude in section 5.

Association Rule Representation

In this section, we present a novel algorithm, to represent a set of association rules efficiently with less number of rules. Flow of methodology is shown in Figure 1. The input for this methodology is set of association rules and the output is set with less number of rules.

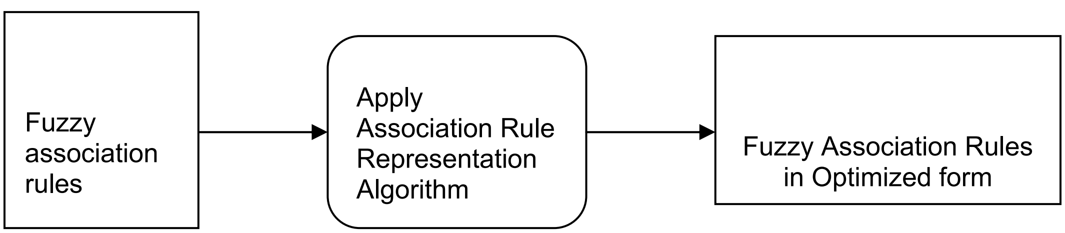

Figure 1. Association rule representation methodology

The algorithm, which converts large number of association rules into less number of rules, is shown below.

Algorithm:

Input: R= Rule set, each rule is of the form X→Y, where X and Y are itemsets.

1. Initialize R, Result set of rules RS=R

2. For all rules in rule set R

    If rules {X→Y, Y→X}∈ R then

        RS =RS - {X→ Y, Y→X}

        RS=RS U {X ↔ Y}

    Endif

End for

3. Rule X→Y is a strong association rule when all the rules, which are subsets of the itemset {X, Y}, exist in the rules set R.

    If X→Y is a strong association rule then

        RS=RS- {All sub rules of itemset {X, Y}}

        Mark rule X→Y in Rule set R as a strong association rule.

    Endif

4. Rule X→Y is a weak association rule when all the rules related to the itemset {X, Y} does not exist in the rules set R. If rule is weak association rule then mark those rules as weak association rule in rule set RS.

5. Rule X→Y is a semi strong association rule when at least one of the rules related to the itemset {X, Y} exist in the rules set. If rule is semi strong association rule then mark those rules as semi strong association rule in rule set RS.

6. Repeat steps 3-5 for all rules in rule set R

7. End

The given algorithm initializes a set of association rules. Each rule is of the form X→Y, where X and Y are itemsets. This algorithm does not modify or remove any knowledge from dataset rather converts it into a form which is easily understandable by users. The resultant of the algorithm is a set which contain less number of association rules, which facilitates the easy reading of all rules. This rule representation algorithm is useful when support and confidence thresholds are less, that is number of rules is more.

The first step of algorithm initializes the rules set R and copies the same rule set into RS. In second step, if rules $A \rightarrow B \in R$ and $B \rightarrow A \in R$ then the algorithm removes these rules from RS and appends it as a single rule $A \leftrightarrow B$. The implementation of this step is easy and results in the reduction of many rules. Third step of the algorithm defines a term called Strong Association Rule. A Rule is a strong association rule when all subset of rules exist in R. Third step marks all strong association rules as strong in RS and removes all subset of strong association rules in RS. Step 4 and 5 marks some of the rules as Semi strong and weak association rules. Semi strong association rules are rules, which have at least one sub rule to corresponding itemset. Number of semi strong association rules is moderate in any kind of association rules. Weak association rules are the rules, which do not have any sub rules to corresponding itemset. These rules are impractical. Sixth step repeats steps 3-5 for all rules in rule set RS. By seeing strong rules, any person can use those rules for business decisions.

An Example

Consider an example association rules shown in Table 1.

Table 1. Example Association rules

1.A1→A2	9.A4→A5
2. A2→A1	10.A2→A4
3. A1→A3	11.A4→A2
4. A1→A4	12.A3→A6
5. A4→A1	13.A6→ A3
6. A5→A3	14.A1, A2→A4
7. A3→A5	15.A1, A7→A3
8. A5→A4	16.A2, A4→A1

Step1: Initializes the rule set and copies into RS. Shown in table 2.

Table 2. Association rules after initialization

1.A1→A2	9.A4→A5	1.A1→A2	9.A4→A5
2. A2→A1	10.A2→A4	2. A2→A1	10.A2→A4
3. A1→A3	11.A4→A2	3. A1→A3	11.A4→A2
4. A1→A4	12.A3→A6	4. A1→A4	12.A3→A6
5. A4→A1	13.A6→ A3	5. A4→A1	13.A6→ A3
6. A5→A3	14.A1, A2→A4	6. A5→A3	14.A1, A2→A4
7. A3→A5	15.A1, A7→A3	7. A3→A5	15.A1, A7→A3
8. A5→A4	16.A2, A4→A1	8. A5→A4	16.A2, A4→A1

Step 2 of the algorithm combine the rules are of the form A→B and B →A into a single rule like A↔B. Rule numbers 1, 2, 4, 5, 6, 7, 8, 9, 10, 11, 12, and 13. After applying rule 2, rule set RS becomes Table 3.

Table 3 Association rules

1.A1↔A2	6. A2↔A4
2. A1→A3	7. A3↔A6
3. A1↔A4	8. A1, A2→A4
4. A5↔A3	9. A1, A7→A3
5. A5↔A4	10. A2, A4→A1

Step3 of algorithm finds strong association rules and removes all sub rules of corresponding itemset. After applying rule 3, rule set RS becomes Table 4.

Table 4. Association rules

1. A1→A3	5. A1, A2→A4 --Strong
2. A5↔A3	6. A1, A7→A3
3. A5↔A4	7. A2, A4→A1 --Strong
4. A3↔A6	

Step 4 and Step 5 of algorithm find semi strong and weak association rules. After applying rule 4 and rule 5, rule set RS becomes Table 5.

Table 5. Association rules

1. A1→A3	5. A1, A2→A4 --Strong
2. A5↔A3	6. A1, A7→A3 --Semi Strong
3. A5↔A4	7. A2, A4→A1 --Strong
4. A3↔A6	

In this example 16 association rules are represented into 7 association rules without any loss of information. This approach is useful when large number of association rules is present. By placing status of association rules as Strong, Semi strong and weak, the user can easily know that strong rules are very much useful and easy to understand.

Experimental Results

We implemented our approach using SAP ABAP/4. We applied the proposed association rule representation algorithm on German credit dataset [23] which consists of 20 attributes. We have done analysis/experiment on 1000 records with eight attributes namely, Status of existing checking account, Duration in month, Credit Amount, Saving account, Present employee since, present residence since, Number of existing credits at this bank, Job details.

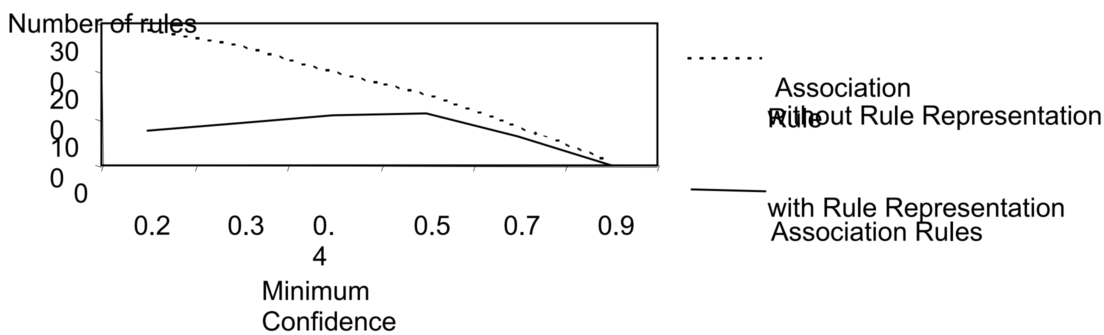

Figure 1. Graph between Minimum Confidence and Number of rules using Association rule representation algorithm

Figure 1 shows the comparison of number of association rules with and without association rule representation with constant minimum support as 0.2 and with various minimum confidence values. The Figure 1 shows that the number of association rules is less with association rule representation. This figure shows that if the minimum confidence is less, then the difference between rules generated with and without association rule representation is more. If minimum confidence is medium, then difference between rules generated with and without association rule representation is very much less.

We applied the proposed association rules algorithm on insurance data [14]. Data set consists of 86 attributes. We have done analysis/experimentation on 5000 randomly selected records with seven attributes namely, customer subtype, Number of houses, average size household, average age, married, average income, number of mobile home policies. The minimum support and confidence value is considered as 0.2 and 0.2 respectively for generating association rules. Figure 2 shows the comparative bar graph representation of association rules generation with and without using association rule representation algorithm. As it is evident from the graph that the rules generated using association rule representation algorithm is almost just 50% or less as compared to the original rule set. Although these rules being much less in number, we are retaining all useful information helping the domain expert in making effective and timely decisions with this less number of interesting rules.

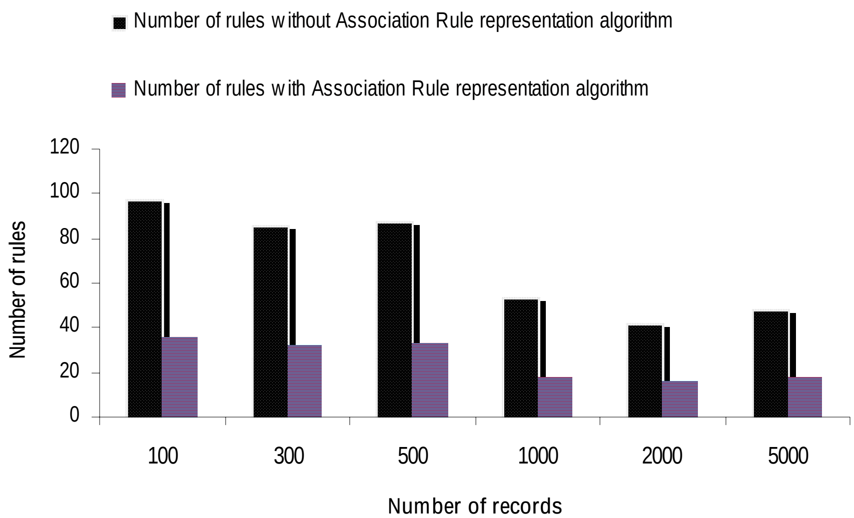

Figure 2. Comparison of Rules with and without Association Rule Representation

Conclusion

The proposed fuzzy association rule representation algorithm is very useful for top management for taking timely decisions by seeing the less number of rules. With the help of new proposed algorithm many fuzzy association rules can be represented into less number of association rules without loss of any information.

References

[1]. Agrawal, R., Imieliński, T., Swami, A.: Mining association rules between sets of items in large databases. In the Proceedings of ACM SIGMOD international conference on Management of Data (1997) 207-216.

[2]. Au, Wai-Ho., Keith, Chan, C.C.: Mining Fuzzy Association Rules in a Bank-Account Database. TFS 11(2) (2003) 238—248.

[3]. Barro, S., Bugarin, A.J., Carinena, P., Diaz-Hermida, F.:A framework for fuzzy quantification models analysis. In IEEE Transactions on Fuzzy Systems, Vol 11 (2003) 89-99.

[4]. Cock, M. De., Cornelis, C., Kerre, E..E.: Elicitation of fuzzy association rules from positive and negative examples. Fuzzy Sets and Systems 149 (2005) 73–85.

[5]. Delgado, M., Marin, N., Vila, Maria-Amparo.:Fuzzy Association Rules: General Model and Applications. IEEE Transactions on fuzzy systems, Vol. 11(2003) 238-248.

[6]. Graaf, J.M.de., Kosters, W.A., Witteman,J.J.W.: Interesting Fuzzy Association Rules in Quantitative Databases, In the 5th European Conference on Principles of Data Mining and Knowledge Discovery), Freiburg, Germany (2001) 140-151.

[7]. Hong, Tzung-Pei., Lin, Kuei-Ying., Wang, Shyue-Liang.: Fuzzy data mining for interesting generalized association rules, Fuzzy Sets and Systems, v.138 (2003) 255-269.

[8]. Han, J., Fu, Y.: Discovery of multi-level association rules from large databases. In proceedings of the 21st International Conference on Very Large Data Bases (1995) 420-431.

[9]. Jain, A. K., Dubes,R. C.: Algorithms for Clustering Data. Prentic Hall, New Jersey, 1988.

[10]. Kuok, C. M., Fu, A., Wong, M. H.: Mining Fuzzy Association Rules in Databases, SIGMOD Record, Vol. 27, No. 1 (1998) 41-46.

[11]. Park, J.S., Chen, M.S., Yu, P.S.: An efficient hash based algorithm for mining association rules. In proceedings of the ACM SIGMOID International Conference on management of Data (1995) 175—186.

[12]. Srikant, R., Agrawal, R.: Mining quantitative association rules in large relational tables. In: H.V. Jagadish and I.S. Mumick, (eds): Proceedings of the ACM SIGMOID International Conference on Management of Data (1996) 1-12.

[13]. Toivonen, H.: Sampling large databases for association rules. In Proceedings of the 22nd International Conference on Very Large Databases (1996) 134-145.

[14]. Van der Putten, P., Someren, M.van. (eds).: The Insurance Company Case. Published by Sentient Machine Research, Amsterdam. Also a Leiden Institute of Advanced Computer Science Technical Report (2000).

[15]. Zhang, W: Mining Fuzzy Quantitative Association Rules, 11th IEEE International Conference on Tools with Artificial Intelligence(2000) 99 – 102.

[16]. Zhang, C., Zhang, S.: Association Rule Mining Models and Algorithms. Springer-Verlag Berlin Heidelberg (2002).

[17]. Ramakrishnan Srikant, Rakesh Agrawal.: Mining quantitative association rules in large relational tables Proceedings of the 1996 ACM SIGMOD international conference on Management of data Montreal, Quebec, Canada Pages: 1 – 12

[18]. Keith C. C. Chan, Wai-Ho Au: Mining Fuzzy Association Rules. CIKM 1997: 209-215

[19]. Tzung-Pei Hong, Chan-Sheng Kuo, Shyue-Liang Wang: A fuzzy AprioriTid mining algorithm with reduced computational time. Applied Soft Computing 5(1): 1-10 (2004)

[20]. Cody Wong, Simon Shiu, and Sankar Pal.: Mining fuzzy association rules for web access case adaptation. Proc. of the Workshop Program at the Fourth International Conference on Case-Based Reasoning (2001)

[21]. Edward R. Omiecinski. Alternative interest measures for mining associations in databases. IEEE Transactions on Knowledge and Data Engineering, 15(1):57-69, Jan/Feb 2003.

[22]. Fu. A. et al., 1998. Finding fuzzy sets for the mining of fuzzy association rules for numerical attributes, in Proceedings of 1st Intl. Symposium on Intelligent Data Engineering and Learning (IDEAL'98), pages 263-268.

[23]. http://www.ics.uci.edu/~mlearn/MLSummary.html

Author Bio

K. Sai Krishna working in Cognizant Technology Solutions.

Disclaimer & Liability Notice