

Mining Negative Association Rules

Tushar Mani

Ph.D. (CSE), NIMS University, Jaipur (Rajasthan), India

Abstract: Association rule mining is a data mining task that discovers associations among items in a transactional database. Typical association rules consider only items enumerated in transactions. Such rules are referred to as positive association rules. Negative association rules also consider the same items, but in addition consider negated items (i.e. absent from transactions). Negative association rules are useful in market-basket analysis to identify products that conflict with each other or products that complement each other. They are also very convenient for associative classifiers, classifiers that build their classification model based on association rules. Many other applications would benefit from negative association rules if it was not for the expensive process to discover them. Indeed, mining for such rules necessitates the examination of an exponentially large search space. In this paper, we propose an algorithm that mines negative association rules by using conviction measure which does not require extra database scans.

I. Introduction

Association rule mining is a data mining task that discovers relationships among items in a transactional database. Association rules have been extensively studied in the literature for their usefulness in many application domains such as recommender systems, diagnosis decisions support, telecommunication, intrusion detection, etc. The efficient discovery of such rules has been a major focus in the data mining research community. From the original *apriori* algorithm [1] there have been a remarkable number of variants and improvements of association rule mining algorithms [2].

Association rule analysis is the task of discovering association rules that occur frequently in a given data set. A typical example of association rule mining application is the market basket analysis. In this process, the behaviour of the customers is studied when buying different products in a shopping store. The discovery of interesting patterns in this collection of data can lead to important marketing and management strategic decisions. For instance, if a customer buys bread, what is the probability that he/she buys milk as well? Depending on the probability of such an association, marketing personnel can develop better planning of the shelf space in the store or can base their discount strategies on such associations/correlations found in the data. All the traditional association rule mining algorithms were developed to find positive associations between items. By positive associations we refer to associations between items existing in transactions (i.e. items bought). What about associations of the type: “customers that buy Coke do not buy Pepsi” or “customers that buy juice do not buy bottled water”? In addition to the positive associations, the negative association can provide valuable information, in devising marketing strategies. Interestingly, very few have focused on negative association rules due to the difficulty in discovering these rules.

Although some researchers pointed out the importance of negative associations [3], only few groups of researchers [4], [5], [6] proposed an algorithm to mine these types of associations. This not only illustrates the novelty of negative association rules, but also the challenge in discovering them.

II. Basic Concept and terminology

This section introduces basic concept of association rules and some related work on negative association rules.

2.1 Association Rules

Formally, association rules are defined as follows: Let $I = \{i_1, i_2, \dots, i_n\}$ be a set of items. Let D be a set of transactions, where each transaction T is a set of items such that $T \subseteq I$. Each transaction is associated with a unique identifier TID . A transaction T is said to contain X , a set of items in I , if $X \subseteq T$. An *association rule* is an implication of the form “ $X \rightarrow Y$ ”, where $X \subseteq I$; $Y \subseteq I$, and $X \cap Y = \Phi$. The rule $X \rightarrow Y$ has *support* s in the transaction set D if $s\%$ of the transactions in D contain $X \cup Y$. In other words, the support of the rule is the probability that X and Y hold together among all the possible presented cases. It is said that the rule $X \rightarrow Y$ holds in the transaction set D with *confidence* c if $c\%$ of transactions in D that contain X also contain Y . In other words, the confidence of the rule is the conditional probability that the consequent Y is true under the condition of the antecedent X . The problem of discovering all association rules from a set of transactions D consists of generating the rules that have a *support* and *confidence* greater than given thresholds. These rules are called *strong rules*, and the framework is known as the *support-confidence framework* for association rule mining.

A negative association rule is an implication of the form $X \rightarrow \neg Y$ (or $\neg X \rightarrow Y$ or $\neg X \rightarrow \neg Y$), where $X \subseteq I$, $Y \subseteq I$ and $X \cap Y = \emptyset$ (Note that although rule in the form of $\neg X \rightarrow \neg Y$ contains negative elements, it is equivalent to a positive association rule in the form of $Y \rightarrow X$. Therefore it is not considered as a negative association rule.) In contrast to positive rules, a negative rule encapsulates relationship between the occurrences of one set of items with the absence of the other set of items. The rule $X \rightarrow \neg Y$ has support $s\%$ in the data sets, if $s\%$ of transactions in T contain itemset X while do not contain itemset Y . The support of a negative association rule, $\text{supp}(X \rightarrow \neg Y)$, is the frequency of occurrence of transactions with item set X in the absence of item set Y . Let U be the set of transactions that contain all items in X . The rule $X \rightarrow \neg Y$ holds in the given data set (database) with confidence $c\%$, if $c\%$ of transactions in U do not contain item set Y . Confidence of negative association rule, $\text{conf}(X \rightarrow \neg Y)$, can be calculated with $P(X \neg Y)/P(X)$, where $P(\cdot)$ is the probability function. The support and confidence of itemsets are calculated during iterations. However, it is difficult to count the support and confidence of non-existing items in transactions. To avoid counting them directly, we can compute the measures through those of positive rules.

III. Related Work In Negative Association Rule Mining

A new idea to mine strong negative rules presented in [15]. They combine positive frequent itemsets [8] with domain knowledge in the form of taxonomy to mine negative associations. However, their algorithm is hard to generalize since it is domain dependent and requires a predefined taxonomy. Finding negative itemsets involve following steps: (1) first find all the generalized large itemsets in the data (i.e., itemsets at all levels in the taxonomy whose support is greater than the user specified minimum support) (2) next identify the candidate negative itemsets based on the large itemsets and the taxonomy and assign them expected support. (3) in the last step, count the actual support for the candidate itemsets and retain only the negative itemsets. The interest measure RI of negative association rule $X \rightarrow \neg Y$, as follows $RI = (E[\text{support}(X \cup Y)] - \text{support}(X \cup Y)) / \text{support}(X)$ Where $E[\text{support}(X)]$ is the expected support of an itemset X .

A new measure called *mininterest*, (the argument is that a rule $A \rightarrow B$ is of interest only if $\text{supp}(A \cup B) - \text{supp}(A) \text{supp}(B) \geq \text{mininterest}$) added on top of the support-confidence framework [17]. They consider the itemsets (positive or negative) that exceed minimum support and minimum interest thresholds as itemsets of interest. Although, [17] introduces the “mininterest” parameter, the authors do not discuss how to set it and what would be the impact on the results when changing this parameter.

A novel approach has proposed in [16]. In this, mining both positive and negative association rules of interest can be decomposed into the following two sub problems, (1) generate the set of frequent itemsets [8] of interest (PL) and the set of infrequent itemsets of interest (NL) (2) extract positive rules of the form $A \Rightarrow B$ in PL, and negative rules of the forms $A \rightarrow \neg B$, $\neg A \rightarrow B$ and $\neg A \rightarrow \neg B$ in NL. To generate PL, NL and negative association rules they developed three functions namely, $\text{fipi}()$, $\text{iipis}()$ and $\text{CPIR}()$.

The most common frame-work in the association rule generation is the “Support-Confidence” one. In [14], authors considered another frame-work called correlation analysis that adds to the support-confidence. In this paper, they combined the two phases (mining frequent itemsets [8] and generating strong association rules) and generated the relevant rules while analyzing the correlations within each candidate itemset. This avoids evaluating item combinations redundantly. Indeed, for each generated candidate itemset, they computed all possible combinations of items to analyze their correlations. At the end, they keep only those rules generated from item combinations with strong correlation. If the correlation is positive, a positive rule is discovered. If the correlation is negative, two negative rules are discovered. The negative rules produced are of the form $X \rightarrow \neg Y$ or $\neg X \rightarrow Y$ which the authors term as “confined negative association rules”. Here the entire antecedent or consequent is either a conjunction of negated attributes or a conjunction of non-negated attributes.

An innovative approach has proposed in [13]. In this generating positive and negative association rules consists of four steps: (1) Generate all positive frequent itemsets $L(P_1)$ (ii) for all itemsets I in $L(P_1)$, generate negative frequent itemsets of the form $\neg(I \cup I_2)$ (iii) Generate all negative frequent itemsets $\neg I_1 \neg I_2$ (iv) Generate all negative frequent itemsets $I_1 \neg I_2$ and (v) Generate all valid positive and negative association rules. Authors generated negative rules without adding additional interesting measure(s) to support-confidence frame work.

A new and different approach has been proposed in [7]. This is simple but effective. It is not using any additional interesting measures and additional database scans. In this approach, it is finding negative itemsets by replacing a literal in a candidate itemset by its corresponding negated item. If a candidate itemset contains 3 items then it will produce corresponding 3 negative itemsets one for each literal.

IV. Discovering Negative Association Rules

The most common framework in the association rules generation is the “support-confidence” one. Although these two parameters allow the pruning of many associations that are discovered in data, there are cases when many uninteresting rules may be produced. In this paper we consider another interesting measure

called conviction that adds to the support- confidence framework. Next section introduces the measure conviction.

- The conviction of a rule is defined as:

$$Conv(X \Rightarrow Y) = \frac{1 - supp(Y)}{1 - conf(X \Rightarrow Y)}$$

conv(X=>Y) can be interpreted as the ratio of the expected frequency that X occurs without Y (that is X=>¬ Y) if X and Y were independent divided by the observed frequency of incorrect predictions. The range of conviction is 0 to ∞

A. Algorithm MPNAR

In this section we propose and explain our algorithm.

Algorithm: Mining Negative Association Rules

Input: TDB-Transactional Database

mins -minimum support

minc-minimum confidence

Output: Negative Association Rules

Method:

```

1. NAR = ∅ /* initially NAR is empty */
2. Scan the database and find the set of frequent 1.itemsets(F1)
3. for (k=2; Fk-1 != ∅; k++)
4. {
5. Ck = Fk × Fk-1 /* generates candidates itemsets */
6. // Prune using Apriori Property
7. for each i ∈ Ck, any subset of i is not in Fk-1 then Ck = Ck - { i }
8. for each i ∈ Ck /* perform database scanning to find support */
9. {
10. s = Support( i);
11. for each A,B (A ∪ B = i )
12. {
13. if ( Supp(X ∩ ¬ Y) ≥ mins && Conviction(X ∩ ¬ Y) ≤ 2.0) /* produces NAR of the form X ∩ ¬ Y */
14. NAR = NAR ∪ { X ∩ ¬ Y } /* produces NAR of the form ¬ X ∩ Y */
15. if ( Supp(¬ X ∩ Y) ≥ mins && Conviction(¬ X ∩ Y) ≤ 2.0) then
16. NAR = NAR ∪ { ¬ X ∩ Y }
17. }
18. }
19. }

```

- Support (¬ X) = 1-support(A)
- Support(X ∪ ¬ Y) = support(X) – support(X ∩ Y)
- Support(¬ X ∩ Y) = support(Y) – support(X ∩ Y)
- Support (¬ X ∪ ¬ Y) = 1 – support - support(X) – support(Y) + support(X ∩ Y)
- The generation of positive rules continues without

V. Experimental Results

We tested our algorithm with [14]. We consider a transactional database contains many transactions. We tested our algorithm with reference [14] with different minimum supports and minimum confidences. Our algorithm is performing well than one in [14].

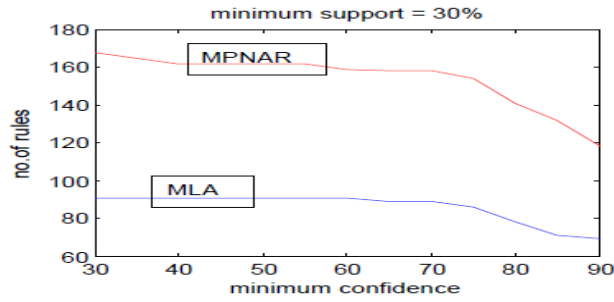


Fig 1: minimum support =30% and different minimum confidences

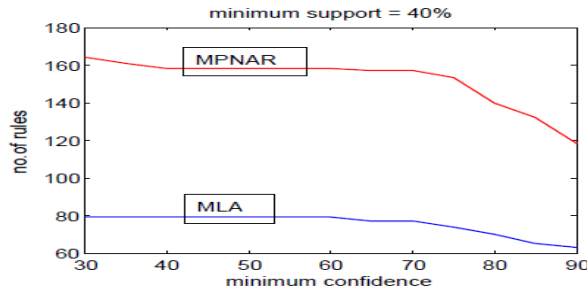


Fig 2: minimum support =40% and different minimum confidences

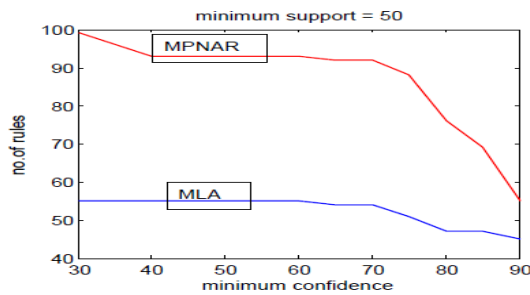


Fig 3: minimum support =50% and different minimum confidences

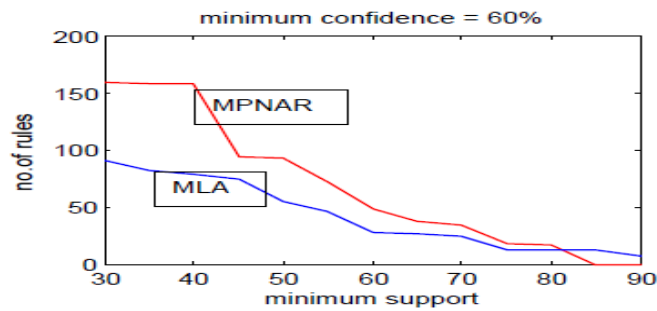


Fig 4: minimum confidence =60% and different minimum supports

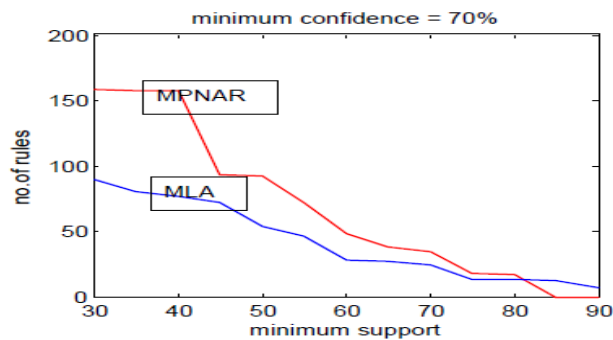


Fig 5: minimum confidence =70% and different minimum supports

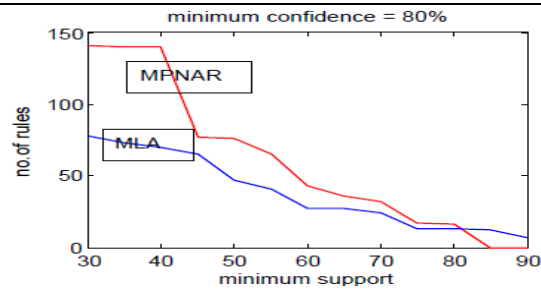


Fig 6: minimum confidence =80% and different minimum supports

VI. Conclusion And Future Work

In this paper we introduced a new algorithm to generate both positive and negative association rules. Our method adds the conviction to the support-confidence framework to generate stronger positive and negative rules. We compared our algorithm with [14] on a real dataset. We discussed their performances on a transactional database and analyzed experimental results. The results prove that our algorithm can perform better than one in [14].

In future we wish to conduct experiments on some other real datasets and compare the performance of our algorithm with other related algorithms such as reference [7] and reference [16].

References

- [1] Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. In: Proc. of SIGMOD. (1993) 207–216
- [2] Goethals, B., Zaki, M., eds.: FIMI'03: Workshop on Frequent Itemset Mining Implementations. Volume 90 of CEUR Workshop Proceedings series. (2003) <http://CEUR-WS.org/Vol-90/>.
- [3] S.Brin, R. Motwani, and C.Silverstein. *Beyond market baskets: Generalizing association rules to correlations*. In ACM SIGMOD,Tucson, Arizona, 1997.
- [4] Wu, X., Zhang, C., Zhang, S.: Mining both positive and negative association rules. In: Proc. of ICML. (2002) 658–665
- [5] Teng, W., Hsieh, M., Chen, M.: On the mining of substitution rules for statistically dependent items. In: Proc. of ICDM. (2002) 442–449
- [6] Savasere, A., Omiecinski, E., Navathe, S.: Mining for strong negative associations in a large database of customer transactions. In: Proc. of ICDE. (1998) 494–502
- [7] B.Ramasubbareddy, A.Govardhan, and A.Ramamohanreddy. Mining Positive and Negative Association Rules, IEEE ICSE 2010, Hefei, China, August 2010.
- [8] Goethals, B., Zaki, M., eds.: FIMI'03: Workshop on Frequent Itemset Mining Implementations. Volume 90 of CEUR Workshop Proceedings series. (2003) <http://CEUR-WS.org/Vol-90/>.
- [9] Teng, W., Hsieh, M., Chen, M.: *On the mining of substitution rules for statistically dependent items*. In: Proc. of ICDM. (2002) 442–449
- [10] Tan, P., Kumar, V.: Interestingness measures for association patterns: A perspective. In: Proc. of Workshop on Postprocessing in Machine Learning and Data Mining. (2000)
- [11] Gourab Kundu, Md. Monirul Islam, Sirajum Munir, Md. Faizul Bari ACN: An Associative Classifier with *Negative Rules* 11th IEEE International Conference on Computational Science and Engineering, 2008.
- [12] Brin,S., Motwani,R. and Silverstein,C., “ *Beyond Market Baskets: Generalizing Association Rules to Correlations*,” Proc. ACM SIGMOD Conf., pp.265-276, May 1997.
- [13] Chris Cornelis, peng Yan, Xing Zhang, Guoqing Chen: Mining Positive and Negative Association Rules from Large Databases , IEEE conference 2006.
- [14] M.L. Antonie and O.R. Za'iane, “*Mining Positive and Negative Association Rules: an Approach for Confined Rules*”, Proc. Intl. Conf. on Principles and Practice of Knowledge Discovery in Databases, 2004, pp 27–38.
- [15] Savasere, A., Omiecinski,E., Navathe, S.: *Mining for Strong negative associations in a large data base of customer transactions*. In: Proc. of ICDE. (1998) 494- 502..
- [16] Wu, X., Zhang, C., Zhang, S.: *efficient mining both positive and negative association rules*. ACM Transactions on Information Systems, Vol. 22, No.3, July 2004,Pages 381-405.
- [17] Wu, X., Zhang, C., Zhang, S.: *Mining both positive and negative association rules*.In: Proc. of ICML. (2002) 658–665
- [18] Yuan,X., Buckles, B., Yuan, Z.,Zhang, J.:*Mining Negative Association Rules*. In: Proc. of ISCC. (2002) 623-629.
- [19] Honglei Zhu, Zhigang Xu: *An Effective Algorithm for Mining Positive and Negative Association Rules*. International Conference on Computer Science and Software Engineering 2008.