# MINING POSITIVE AND NEGATIVE ASSOCIATION RULES

R.SUMALATHA

M. tech Student
CSE Department, JNTU University, Hyderabad
Jyothishmathi Institute of Technology & Science, Karimnagar, AP, India

B. RAMASUBBAREDDY

Associate Professor
CSE Department, JNTU University, Hyderabad
Jyothishmathi Institute of Technology & Science, Karimnagar, AP, India

**Abstract**

Association rule mining is one of the most popular data mining techniques to find associations among items in a set by mining necessary patterns in a large 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 useful for constructing associative classifiers. In this paper, we propose an algorithm that mines positive and negative association rules without adding any additional measure and extra database scans.

**Keywords:** Data Mining, Negative Association Rules, Support, Confidence.

## 1. Introduction

Association rule mining is a data mining task that discovers associations 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. Efficient discovery of such rules has been a major focus in the data mining research. From the celebrated *Apriori* algorithm [1] there have been a remarkable number of variants and improvements of association rule mining algorithms [2]. A typical example of association rule mining application is the market basket analysis. In this process, the behavior of the customers is studied with reference to 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 are chances that customer buys milk as well? Depending on some measure to represent the said chances 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 exist in transactions containing the items bought together. 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. This paper is structured as follows: the next section recalls preliminaries about Association Rules, In Section 3, existing strategies for mining negative Association Rules are reviewed. The proposed algorithm is presented in Section 4 and this is a new Apriori-based algorithm for finding all valid positive and negative association rules. Section 5 contains conclusions and future work.

## 2. Basic Concepts and Terminology Experimental Results

This section introduces association rules terminology and some related work on negative association rules.

### *2.1 Association rules*

Formally, association rules are defined as follows: Let $I = \{i_1, i_2, \ldots i_n\}$ 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. Definition of Negative Association Rule 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 $X \cap Y = \Phi$ (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, $supp(X \rightarrow \neg Y)$, 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, $conf(X \rightarrow \neg Y)$, can be calculated with $P(X \neg Y)/P(X)$, where *P(.)* 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.

## 3. Related Work in Negative Association Rule Mining

We give a short description of the existing algorithms that can generate positive and negative association rules.

The concept of negative relationships mentioned for the first time in the literature by Brin et.al [11]. To verify the independence between two variables, they use the statistical test. To verify the positive or negative relationship, a correlation metric was used. Their model is chi-squared based. The chi-squared test rests on the normal approximation to the binomial distribution (more precisely, to the hyper geometric distribution). This approximation breaks down when the expected values are small.

A new idea to mine strong negative rules presented in [14]. They combine positive frequent itemsets 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[support( X U Y )]-support( X U Y ))/support(X) Where E[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 $supp(A \cup B) - supp(A) supp(B) \geq mininterest$) added on top of the support-confidence framework[16]. They consider the itemsets (positive or negative) that exceed minimum support and minimum interest thresholds as itemsets of interest. Although, [8] 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 [15]. 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 of interest (PL) and the set of infrequent itemsets of interest (NL) (2) extract positive rules of the form A=>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, fipi(), iipis() and CPIR().

The most common frame-work in the association rule generation is the "Support-Confidence" one. In [13], 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 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 [12]. 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 $\neg$ ( I1 I2 ) (iii) Generate all negative frequent itemsets $\neg$ I1 $\neg$I2 (iv) Generate all negative frequent itemsets I1 $\neg$ I2 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.

## 4. Algorithm

In this section we propose and explain our algorithm. Apriori algorithm has two steps, namely, the join step and the prune step. In join step all frequent itemsets of previous level joined itself to obtain candidate itemsets i.e., by joining $L_{k-1}$ to itself i.e.,

$C_k = L_{k-1} \bowtie L_{k-1}$.

In pruning step, for each I ε $C_k$, it applies *Apriori Property (*an itemset is frequent if all its subsets are also frequent*)*. Itemsets satisfying Apriori Property are called as valid itemsets and it is denoted by $PC_K$. $NC_K$ can be obtained by replacing each literal in $PC_K$ by its corresponding negated item. For a valid candidate with n literals it produces n negative itemsets. Support of negative itemsets can be obtained from positive itemsets.

### *4.1 Algorithm: Negative association rules (NAR).*

*Input:*

D:Transactional database,
ms: minimum support,
mc: minimum confidence

*Output:* Positive and Negative Association Rules
*Method:*

(1) $L_1$=frequent-1-positive-itemsets(D)
(2) $N_1$=frequent-1-Negative-itemsets(D)
// complement frequent-1-positive-itemsets(D)
(3) $L=L_1 \cup N_1$;
(4) for (k=2; $L_{k-1} \neq \emptyset$; k++)
(5) {
(6) // Generating $C_k$
(7) for each $l_1, l_2 \varepsilon L_{k-1}$
(8) If($l_1[1]=l_2[1]$^……$l_1[k-2]=l_2[k-2]$^$l_1[k-1]<l_2[k-1]$)
(9) $C_k=C_k \cup$ {l1 [1]…….$l_1$ [k-2], $l_1[k-1]$, $l_2[k-1]$}
(10) end if
(11) end for
(12) // Pruning using Apriori property
(13) for each (k-1)- subsets s of c in $C_k$
(14) If s is not a member of $L_{k-1}$
(15) $C_k=C_k - \{c\}$
(16) end if
(17) end for
(18) $PC_k= C_k$;
(19) for each c in $PC_K$
(20) $NC_k=\{c_k^1 / c_k^1$ is obtained by replacing a literal of c in $PC_k$ by its negation}
(21) //Pruning using Support Count
(22) Scan the database and find support for all c in $PC_k$
(23) $L_k$=candidates in PCk that pass support threshold
(24) Find support for all $c_k^1$ in $NC_k$ from supports of members of $PC_K$ and $L_{k-1}$
(25) $N_k$= candidates in $NC_k$ that pass support threshold
(26) $L= L_k \cup N_k$
(27) }

- Line 1 generates positive-frequent-1-itemsets
- Line 2 generates negative-frequent-1-itemsets by complementing 1-itemsets obtained in Line 1
- Line 8 and 9 generates candidate itemsets $C_k$ using Apriori algorithm
- Line 13-15 pruning candidate itemsets in $C_k$ using Apriori property
- Lines 18, after pruning, the remaining elements are treated as valid candidates and is denoted by $PC_k$.
- Line 19-20, for each literal of this valid candidate, replace the literal with the corresponding negated literal, creates a new negative rule and denoted by $NC_k$. Each valid candidate with n number of literals in the antecedent will generate n new negative itemsets. For example a 3-itemset ABC will give 3 negative items ¬ ABC, A¬ BC, and AB¬ C.
- Line 22-23, prune all items in $PC_k$ using support count and add to $L_k$, set of frequent k-itemsets
- Line 24, find support count of all items in $NC_k$ using $PC_k$ and $L_{k-1}$.
- support(¬ A) = 1- suuport(A)
- support(AU¬ B) = support(A)-support(AU B)
- support(¬ AUB) = support(B)-support(AU B)
- support(¬ AU¬ B) = 1-support(A)-support(B) + support(A U B)
- Line 25, $N_k$ is the set of all elements whose support ≥ minsupp.
- The generation of positive rules continues without disruption and the rich but valuable negative rules are produced as by-products of the Apriori process.

## 5 Experimental Results Conclusion and Future Work

We tested proposed algorithm on a synthetic dataset to study the performance of the algorithm. However, to illustrate the algorithm, we hereunder provided the solution to a typical example.

***5.1. Example*:** Let us consider transactional database with 507 transactions and 12 items. In Graph 1 the Rules of various support count and fixed confidence is shown

.

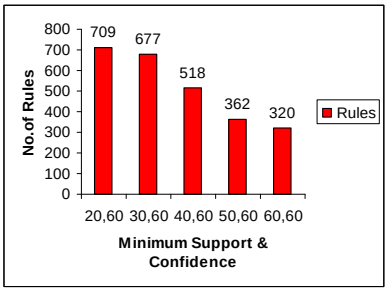


Figure 1. Number of Rules

In the following graph, if support count of items varies then number of Rules also varies for a given fixed confidence

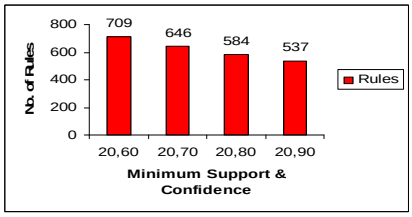


Figure 2. Number of Rules generated for fixed minimum support count varies in confidence

***5.2. Example*:.** Let us consider transactional database with 9 transactions and 5 items. In Graph 1 the Rules of various support count and fixed confidence is shown

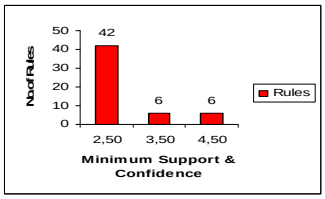


Figure 3. Number of rules generated for fixed confidence.

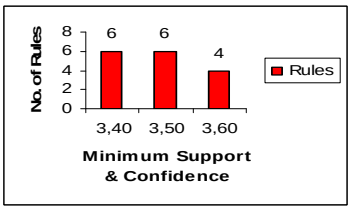


Figure 4. Number of Rules generated for fixed minimum support count varies in confidence

## 6 Conclusion and Future Work

In this paper, we proposed an algorithm that mines both positive and negative association rules. Our method generates positive and negative rules with existing support-confidence framework and no extra scan is required for mining negative association rules. We conducted experiments on synthetic data set. It is producing larger number of negative rules. In future we wish to conduct experiments on real datasets and compare the performance of our algorithm with other related algorithms.

## References

[1] Agrawal , R ; Srikant,R. (1994). *Fast algorithms for mining association rules.* In VLDB, Chile.

[2] Han, J; Pei, J; Yin, Y.(2000). *Mining frequent patterns* without candidate generation. In SIGMOD, dallas,Texas.

[3] Blakeand , C; Merz, C .UCI repository of machine learning databases.

[4] Brin, S; Motwani, R ; Silverstein , C. (1997) .*Beyond market* baskets: Generalizing association rules to correlations. In ACM SIGMOD, Tucson, Arizona.

[5] Thiruvady,D ;Webb,G.(2004). .*Mining negative association rules using grd*. In PAKDD, Sydney, Australia.

[6] Ramasubbareddy,B; Govardhan, A; Ramamohanreddy,A.(2009). Adaptive approaches in mining negative association rules. In intl. conference on ITFRWP-09, India.

[7] Goethals, B., Zaki, M.(2003). *FIMI'03: Workshop on* Frequent Itemset Mining Implementations. Volume 90 of CEUR Workshop Proceedings series. http://CEUR-WS.org/Vol-90/.

[8] Teng, W; Hsieh, M; Chen, M.(2002). *On the mining of* substitution rules for statistically dependent items. In: Proc. of ICDM. 442–449

[9] Tan, P; Kumar, V.(2002). Interestingness measures for association patterns: A perspective.In: Proc. of Workshop on Postprocessing in Machine Learning and Data Mining.

[10] Gourab Kundu, Md; Monirul Islam; Sirajum Munir, Md; Faizul Bari.(2008). ACN: An Associative Classifier with *Negative Rules* 11th IEEE International Conference on Computational Science and Engineering.

[11] Brin,S; Motwani, R; Silverstein,C. (1997).*Beyond Market* Baskets: Generalizing Association Rules to *Correlations,*" Proc. ACM SIGMOD Conf., pp.265-276.

[12] Chris Cornelis; peng Yan; Xing Zhang; Guoqing Chen.(2006).Mining Positive and Negative Association Rules from Large Databases , IEEE conference.

[13] Antonie,M.L; Za¨ıane,O.R.(2004). *Mining Positive and* Negative Association Rules: an Approach for Confined *Rules*, Proc. Intl. Conf. on Principles and Practice of Knowledge Discovery in Databases, pp 27–38.

[14] Savasere, A; Omiecinski, E; Navathe, S.(1998). *Mining for* Strong negative associations in a large data base of *customer transactions.* In: Proc. of ICDE. 494- 502

[15] Wu, X; Zhang, C; Zhang, S.(2004).Efficient m*ining both positive and negative association rules*. ACM Transactions on Information Systems, Vol. 22, No.3,Pages 381-405.

[16] Wu, X; Zhang, C; Zhang, S.(2002). *Mining both positive and* negative association rules.In: Proc. of ICML. 658–665

[17] Yuan,X; Buckles, B;Yuan, Z; Zhang, J.(2002). *Mining* Negative Association Rules. In: Proc. of ISCC.623-629.

[18] Honglei Zhu; Zhigang Xu. (2008). *An Effective Algorithm for* Mining Positive and Negative Association Rules. International Conference on Computer Science and Software Engineering .

[19] Pradip Kumar Bala. (2009).A Technique for Mining Negative *Association Rules* . Proceedings of the 2nd Bangalore Annual Compute Conference .

[20] Jiawei Han, Micheline Kamber . Data Mining: Concepts and Techniques.

Ramatenkhi Sumalatha Received the Bachelor of Technology ,(Computer Science and Information Technology ) From Jawaharlal Nehru Technological university, Hyderabad, in 2003. M.Tech (C.S.E) From JNTU(Jawaharlal Nehru Technological University) , Hyderabad in 2010.