I J C T A, 9(41) 2016, pp. 1112-1118 ISSN: 0974-5572

MINING POSITIVE AND NEGATIVE ASSOCIATION RULES: A SURVEY

Ujwala Manoj Patil and J.B. Patil

Abstract: Data mining is getting increasing acceptance in science and business areas that need to identify and represent certain dependencies between attributes. The kind of knowledge that could be discovered from a database is represented in the form of association rules. All the traditional association rule mining algorithms were developed to find positive associations between items i.e. A⇒B, whereas negative association rule is an implication of the form A⇒¬B, ¬A⇒B, ¬A⇒¬B, where A and B are database itemsets, ¬A, ¬B are negations of database items. Here we review apriori based algorithms to find both positive and negative associations between items. First, we discuss the interestingness measures used in current approaches and then we present and discuss an algorithm in the detail. The review also presents the advantages and limitations of the existing techniques.

Key Words: *Association Rule Mining(ARM), Association Rules (AR), Positive Association Rules (PAR), Negative Association Rules (NAR)*

I. INTRODUCTION

Data Mining is used for extraction of knowledge from large data sets. Data mining is broadly classified in the areas such as Association Rules, Classifications, and Clustering [1, 2]. Association rule mining discovers relationships from the huge amount of data by generating rules. Association rule mining is useful in many application domains like recommender system, decision support, health care, intrusion detection, etc. [2, 3, 4]. Association rule mining was introduced by Agrawal et.al in terms of the *apriori* algorithm [1]. After that there have been a remarkable number of variants and improvements of association rule mining algorithms [5, 6, 7, 8, 9, 10, 11]. Traditional association rule mining algorithms have been developed to find associations between items. This association is in between the items that exist in the transactional database. The associations are of two types, called positive associations and negative associations. The traditional association is called positive associations which consider the presence of the item, i.e A⇒B while another is negative that negates presence of the item i.e. A⇒¬B, ¬A⇒B, ¬A⇒¬B. Positive association rules are useful in decision making, likewise negative association rules also play important role in decision making.

1.1 Contribution of This Paper

The main contribution of this work as follows:

1. We have surveyed the current literature to discover the positive and negative association rules. There are very few papers which discover both positive and negative association rules. Algorithms in those papers are discussed in details by showing extensions of the basic *apriori* algorithm.
2. The main focus of this survey was to understand what and how different interestingness measures are used to discover both positive and negative association rules.
3. We also discussed the advantages and limitations of the existing techniques.

Department of Computer Engineering, R.C.Patel Institute of Technology, Shirpur, Maharashtra, India
Email: patilujwala2003@gmail.com, jbpatil@hotmail.com

The remainder of this paper is organized as follows: section 2 gives basic concepts and terminology involved in association rule mining. Section 3 presents related work to mine both positive and negative association rules. We summarize this paper in section 5.

II. BASIC CONCEPTS AND TEMINOLOGY

2.1 Preliminaries

Suppose $I = \{i_1, i_2. . . i_N\}$ be a set of N distinct *items* and data D is a set transactions over I . Each transaction T contains a set of items $i_1, i_2 \ldots i_k \in I$ *i.e.* $T \subseteq I$. A transaction has an associated unique identifier called *TID*. An *association rule* is an implication of the form $A \Longrightarrow B \ (or\ A \rightarrow B)$, where $A, B \subseteq I$, and $A \cap B = \emptyset$. A is called the *antecedent* of the rule, and B is called the *consequent* of the rule. A set of items (antecedent or consequent) is called an *itemset*. In general, for simplicity, an itemset $\{i_1, i_2, i_3\}$ is sometimes written as $i_1 i_2 i_3$. Let us denote by $|A \cup B|$ the number of transactions that contain both A and B and |D| denote the number of transactions in the database.

Definition 1

Association Rule (AR): The association rule is an implication of the form $A \Longrightarrow B$ where $A \subseteq I, B \subseteq I$ and $A \cap B = \emptyset$ with a support and a confidence above a minimum threshold.

2.2 Association Rule Mining

Association rule mining seeks rules of the form $A \Longrightarrow B$ with support and confidence greater than, or equal to, user-specified minimum support (*ms*) and minimum confidence (*mc*) thresholds respectively, where

—A and B are disjoint itemsets, i.e, $A \cap B = \emptyset$,

— Support $(A \Longrightarrow B)$ = support $(A \cup B)$, and

— Confidence $(A \Longrightarrow B)$ = support $(A \cup B)$/ support (A)

This is referred as Agrawal support-confidence framework. Association analysis can be divided into two steps [1]:

Step 1: Generate all large itemset:

Input- L_1= {large-1 sequences}

Output- maximal sequences in $\cup_k L_k$

For (k=2;L_{k-1};k++) do
begin
C_k=New candidates generated from L_{k-1}
For each user-sequence c in the database do
Increment the count of all candidates in C_k that are contained in c.
L_k = Candidates in C_k with minimum support.
End

Step 2: Generate rules that have minimum confidence.

At the end of step 2, we find all the interesting association rules of the form $A \Longrightarrow B$, are called as a positive association rule.

2.3 Positive and Negative Association Rules

Positive associations are associations between items that are present in transactions which consider the presence of the item. While negative associations are rules that comprise relationships between present and absent items. A rule of the form $A \Longrightarrow B$ is called positive association rule and the rule in other forms

A⟹¬B, ¬A⟹B, and A⟹¬B are called negative association rules. Along with positive association rules (PAR), negative association rules (NAR) are also useful in decision making. Negative association rules are used to find the attributes that are in conflict or complement with each other. Mining of positive and negative rules is very expensive as it has to explore large search space. Till date very few algorithms in the literature have been proposed which use various interestingness measures to find positive as well as negative association rules. We have surveyed the literature to find what interestingness measures [8, 12, 13, 14, 15, 16, 17, 18, 19, 20] are used by various algorithms and how these interestingness measures are used to find positive and negative association rules. Interestingness measures used for negative association rules are computed from the relative information of positive association rules.

Following are some definitions of interestingness measures related to PAR and NAR mining [8, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21]:

Definition 2

Positive Item: Each transaction T contains a set of items $i_1, i_2 \ldots i_k \in I$ *i.e.* $T \subseteq I$, i_1 is an item that is present in a transaction is called positive item.

Definition 3

Negative Item: A negative item is defined as an item that is absent from the transaction. It is represented as $\neg i_k$ *i.e.* $i_k \notin I$.

Definition 4

Positive Support: The rule A ⟹B comes into existence in the transaction D, it has the support level, s, and when and only when the proportion that D contains the transaction A∪B is s, i.e.

$$s = support\ (A \Longrightarrow B) = (A \cup B)$$
$$= |\{T|A \cup B \subseteq T \wedge T \in D\}| / |D|\ , where\ A \cup B\ means\ that\ both\ A\ and\ B\ are\ present.$$

Definition 5

Positive Confidence: The rule A ⟹B comes into existence in the transaction D, it has the confidence level, c, and when and only when the proportion that D contains the transaction A is c, confidence measures the strength of a rule i.e.

$$c = confidence\ (A \Longrightarrow B) = (B/A) = |\{T|A \cup B \subseteq T \wedge T \in D\}| / |T|A \subseteq T \wedge T \in D|\ ,$$
$$where\ A \cup B\ means\ that\ both\ A\ and\ B\ are\ present.$$

Definition 6

Positive Interest: S*uppose* $A, B \subseteq I, A \cap B = \emptyset$, *so we have*

$$\frac{support(A \cup B)}{support(A)support(B)}$$

Definition 7

Positive Piatetsky-Shapiro's (PS) Interest:

A rule is not interesting if its antecedent and consequent are independent.

$$Interest\ (A \Longrightarrow B) = |\ support(A \cup B) - support(A)\ support(B)|$$

Definition 8

Negative Support: *Suppose* $A, B \subseteq I, A \cap B = \emptyset$, *so we have*

$$Support\ (\neg A) = 1 - support\ (A),$$
$$Support\ (A \Longrightarrow \neg B) = support\ (A) - support\ (A \cup B),$$
$$Support\ (\neg A \Longrightarrow B) = support\ (B) - support\ (A \cup B),$$
$$Support\ (\neg A \Longrightarrow \neg B) = 1 - support\ (A) - support\ (B) + support\ (A \cup B).$$

Definition 9

Negative Confidence: *Suppose* $A,B \subseteq I, A \cap B = \emptyset$, *so we have*

$Confidence\ (A \Longrightarrow \neg B) = (support\ (A) - support\ (A \cup B))/support\ (A)$, *i.e* $1 - Confidence\ (A \Longrightarrow B)$

$Confidence\ (\neg A \Longrightarrow B) = (support\ (B) - support\ (A \cup B))/(1 - support\ (A))$

$Confidence\ (\neg A \Longrightarrow \neg B) = (1 - support\ (A) - support\ (B) + support\ (A \cup B))/(1 - support\ (A))$ *i.e.* $1 - Confidence\ (\neg A \Longrightarrow B)$

Definition 10

Negative Interest: *Suppose* $A,B \subseteq I, A \cap B = \emptyset$, *so we have*

$$Interest(A, \neg B) = \frac{support(A \cup \neg B)}{support(A)support(\neg B)}$$

$$Interest(\neg A, B) = \frac{support(\neg A \cup B)}{support(\neg A)support(B)}$$

$$Interest(\neg A, \neg B) = \frac{support(\neg A \cup \neg B)}{support(\neg A)support(\neg B)}$$

Definition 11

Negative Piatetsky-Shapiro's (PS) Interest: *Suppose* $A,B \subseteq I, A \cap B = \emptyset$, *so we have*

$Interest\ (A \Longrightarrow \neg B) = support(A \cup \neg B) - support(A)\ support(\neg B)$

$Interest\ (\neg A \Longrightarrow B) = support(\neg A \cup B) - support(\neg A)\ support(B)$

$Interest\ (\neg A \Longrightarrow \neg B) = support(\neg A \cup \neg B) - support(\neg A)\ support(\neg B)$

Definition 12

Frequent Itemset of Potential Interest: $fipi(I) = support\ (I) \geq ms \wedge A \cup B = I \wedge fipis\ (A, B)$

Where $fipis(A,B) = A \cap B = \emptyset \wedge f(A, B, ms, mc, mi) = 1, ms = min\ support, mc = min\ confidence, \quad mi = min\ interest$

$$f(A, B, ms, mc, mi) = \frac{support(A \cup B) + confidence(A \Rightarrow B) + interest(A,B) - (ms + mc + mi) + 1}{|support(A \cup B) - ms| + |confidence(A \Rightarrow B - mc|) + /interest(A,B) - mi/ + 1}$$

Definition 13

Infrequent Itemset of Potential Interest: $iipi(J) = support\ (J) \leq ms \wedge A \cup B = J \wedge iipis\ (A, B)$

Where, $fipis(A,B) = A \cap B = \emptyset \wedge g(A, \neg B, ms, mc, mi) = 2, ms = min\ support, mc = min\ confidence, \quad mi = min\ interest$

$$g(A, \neg B, ms, mc, mi) = f(A, \neg B, ms, mc, mi) + \frac{support(A) + support(B) - 2ms + 1}{|support(A) - ms| + |support(B) - ms| + 1}$$

Definition 14

Certainty Factor (CF) or CPIR: *Suppose* $A,B \subseteq I, A \cap B = \emptyset$, *so we have*

$$CF(A \Longrightarrow B) = CPIR(B|A) = \frac{support(A \cup B) - support(A)support(B)}{support(A)\big(1 - support(B)\big)}$$
$$= \frac{confidence(A \Longrightarrow B) - support(B)}{support(\neg B)}$$

Definition 15

Person's Ø Correlation Coefficient:

Suppose $A, B \subseteq I, A \cap B = \emptyset$, *for association rule* $A \Longrightarrow B$ *so we have*

$$\emptyset(A \Longrightarrow B) = \frac{support(AB)Support(\neg A\neg B) - support(A\neg B)support(\neg AB)}{\sqrt{(support(A)support(\neg A)support(B)support(\neg B))}}$$

III. RELATED WORK IN POSITIVE AND NEGATIVE ASSOCIATION RULE MINING

In this section, we discuss three well - known algorithms that generate both PAR and NAR.
First, we discuss the algorithm proposed by Xindong Wu et. al [13]. They extend the basic Agrawal et.al *apriori* algorithm [1] which uses support-confidence framework. Along with support-confidence, they used Piatetsky-Shapiro's (PS) interest given in definition 11. The algorithm starts like apriori, and is decomposed into two steps:

1. Generate all frequent and infrequent large itemsets: Itemsets which satisfy user - specified minimum support and minimum interest with $fipis(A, B) = 1$ are declared as frequent itemsets of potential interest i.e positive itemsets. $fipis(A, B)$ is calculated as per definition 12. Itemsets which do not satisfy user - specified minimum support and minimum interest with $iipis(A, B) = 2$ are declared as infrequent itemsets of potential interest i.e negative itemsets. $iipis(A, B)$ is calculated as per definition 13. At the end of step one, they declare all positive large itemsets and negative large itemsets.
2. Generate all possible rules: To generate positive and negative association rules they used CF or CPIR measure given in definition 14. All possible combinations of large itemsets (found in step 1) which satisfy user - specified minimum confidence i.e. CF or CPIR $\geq$ mc.

The good side of Xindong Wu et. al algorithm is that it mines, both PAR and NAR efficiently. While mining PAR and NAR, they use Piatetsky-Shapiro's (PS) interest along with support-confidence, but they do not discuss how to set it and variations of Piatetsky-Shapiro's (PS) interest in the result.

Second, Jingrong Yang et. al [14] developed an algorithm to discover PAR and NAR based on apriori. Again the algorithm is divided into two parts:

1. Generate all frequent large itemsets: This part is same as that of Agrawal et.al *apriori* algorithm [1].
   It finds all large itemsets which satisfy user-specified minimum support. The support of the given itemsets can be computed from definition 4.
2. Generate all possible rules: It generates PAR and NAR from all large itemsets discovered in part 1. Before generating PAR and NAR, it computes the correlation between itemsets. The correlation is computed with the help of Piatetsky-Shapiro's (PS) interest given in the definition 11.
   The correlations between itemsets are either positive, negative or zero.

- If the correlation between the itemsets is positive, then it makes all possible combinations of the itemsets and the combination which satisfies the user specified minimum confidence is declared as a valid positive rule.
- If the correlation between the itemsets is negative, then it makes all possible combinations of the itemsets and the combination which satisfies the user specified minimum confidence is declared as a valid negative rule.

Jingrong Yang et. al algorithm is simple and finds PAR and NAR fast as compared to Xindong Wu et. al algorithm because it uses only three interestingness measures, i.e. support, confidence and correlation. But this algorithm is not effective as it does not find all possible negative association rules.

Third, Maria-Luiz Antonie et. al proposed an approach for PAR and NAR by extending support-confidence approach with a correlation coefficient given in the definition 15. In contrast to above two algorithms, it follows one step approach:

1. The algorithm starts with large one itemset. It gradually combines large L_{k-1} itemsets to find large two itemsets, large three itemsets, large four itemsets, and so on. After combining L_{k-1} itemsets it immediately finds the support of the itemsets. The itemsets which satisfy user-specified minimum support are included in large itemsets list. After adding it calculates the correlation coefficient for that large itemsets. The correlation coefficient for an itemsets an itemset may be either positive, negative or zero.
    - If the correlation coefficient is larger (positive) than the user specified minimum correlation coefficient then it generates all possible combinations of PAR. The confidence of these combinations is verified against user specified minimum confidence. Those satisfying it are declared as valid PAR with minimum support, minimum confidence and minimum correlation coefficient.
    - If the correlation coefficient is negative and the absolute value is larger than the user specified minimum correlation coefficient then it generates all possible combinations of NAR. The confidence of these combinations is verified against user specified minimum confidence. Those satisfying it are declared as valid NAR with minimum support, minimum confidence and minimum correlation coefficient.

Maria-Luiz Antonie et. al proposed a simple and fast approach to find PAR and NAR. Along with simplicity in the algorithm, it automatically adjusts the value of correlation coefficient if no rule is found with the user specified minimum correlation coefficient. But again, it does not explore the search space to find all NAR.

IV. CONCLUSIONS

We have surveyed some current algorithms proposed in the literature for mining of both PAR and NAR. Mining of PAR and NAR is very interesting and challenging because of the complexity and size of the search space. Still, very few authors have proposed algorithms to mine both PAR and NAR. To the best of our knowledge till date, only one or two algorithms [13] have explored all search space to find both PAR and NAR. Most of the other algorithms [14, 15] do not explore all search space.

Acknowledgements: This work is supported by Vice Chancellor Research Motivation Scheme (VCRMS) to college teachers through a university fund of North Maharashtra University, Jalgaon.

REFERENCES

[1] Agrawal, Rakesh, and Ramakrishnan Srikant. "Mining sequential patterns", *Data Engineering, 1995. Proceedings of the Eleventh International Conference on*. IEEE, 1995

[2] Sujatha V. "Improved user Navigation Pattern Prediction Technique from Web Log Data", *Procedia Engineering* 30 (2012): 92-99.

[3] Srivastava, Jaideep, et al. "Web usage mining: Discovery and applications of usage patterns from web data", *Acm Sigkdd Explorations Newsletter* 1.2 (2000), 12-23.

[4] Ramaraj, E., and N. Venkatesan. "Positive and Negative Association Rule Analysis in Health Care Database", *IJCSNS International Journal of Computer Science and Network Security* 8.10 (2008): 325-330.

[5] Kiran, R. Uday, and P. Krishna Re. "An improved multiple minimum support based approach to mine rare association rules", *Computational Intelligence and Data Mining, 2009. CIDM'09. IEEE Symposium on*. IEEE, 2009.

[6] T.P. Hong, C.S. Kuo, and S.C. Chi, "Trade-off between computation time and number of rules for fuzzy mining from quantitative data", *International Journal of Uncertainty, Fuzziness & Knowledge-Based Systems*, vol. 9, no 5, 2001, pp. 587–604.

[7] Hong, Tzung-Pei, and Yeong-Chyi Lee. "An overview of mining fuzzy association rules". *Fuzzy Sets and Their Extensions: Representation, Aggregation and Models*. Springer Berlin Heidelberg, 2008. 397-410.

[8] Brin, Sergey, et al. "Dynamic itemset counting and implication rules for market basket data", *ACM SIGMOD Record*. Vol. 26. No. 2. ACM, 1997.

[9] Srikant, Ramakrishnan, and Rakesh Agrawal, "Mining quantitative association rules in large relational tables", *Acm Sigmod Record*. Vol. 25. No. 2. ACM, 1996.

[10] Stephen G. Matthews, M.A. Gongora, A.A. Hopgood and S. Ahmadi, "Web usage mining with evolutionary extraction of temporal fuzzy association rules", *Knowledge - based Systems*, vol. 54, 2013, pp. 66-72.

[11] Cooley, Robert, Bamshad Mobasher, and Jaideep Srivastava. "Web mining: Information and pattern discovery on the world wide web", *Tools with Artificial Intelligence, 1997. Proceedings., Ninth IEEE International Conference on*. IEEE, 1997.

[12] Aggarwal, Charu C. and Jiawei Han, eds, " *Frequent pattern mining*", Springer, 2014.

[13] Wu, Xindong, Chengqi Zhang, and Shichao Zhang. "Efficient mining of both positive and negative association rules", *ACM Transactions on Information Systems (TOIS)* 22.3 (2004): 381-405.

[14] Yang, Jingrong, and Chunyu Zhao. "Study on the Data Mining Algorithm Based on Positive and Negative Association Rules", *Computer and Information Science* 2.2 (2009): 103-106.

[15] Antonie, Maria-Luiza, and Osmar R. Zaïane. "An associative classifier based on positive and negative rules". *Proceedings of the 9th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*. ACM, 2004.

[16] Zaiane, Maria-Luiza Antonie Osmar R., "Mining positive and negative association rules: An approach for confined rules", 2007: 12-34.

[17] Antonie, Luiza, Jundong Li, and Osmar Zaiane, "Negative association rules", *Frequent Pattern Mining*. Springer International Publishing, 2014. 135-145.

[18] Tan, Pang-Ning, Vipin Kumar, and Jaideep Srivastava. "Selecting the right interestingness measure for association patterns". *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2002.

[19] Ramasubbareddy, B., A. Govardhan, and A. Ramamohanreddy. "Mining Indirect Positive and Negative Association Rules", *International Conference on Advances in Computing and Communications*. Springer Berlin Heidelberg, 2011.

[20] Brin, Sergey, Rajeev Motwani, and Craig Silverstein. "Beyond market baskets: Generalising association rules to correlations", *ACM SIGMOD Record*. Vol. 26. No. 2. ACM, 1997.

[21] Silverstein, Craig, Sergey Brin, and Rajeev Motwani. "Beyond market baskets: Generalising association rules to dependence rules", *Data mining and knowledge discovery* 2.1 (1998): 39-68.