A Study on Positive and Negative Association rule mining

N.V.S.Pavan Kumar, L.JagaJeevan Rao, G.Vijay Kumar
Assistant Professor, School of Computing, KLUniversity
Assistant Professor, School of Computing, KLUniversity
Associate Professor School of Computing, KLUniversity

ABSTRACT:
In recent years negative association rule mining got much focus. In order to mine negative association rules, usually positive association rules are also found. Many of the algorithms developed for mining positive and negative association rules are based on support-confidence framework. Some of the algorithms use additional parameter(s) along with these two parameters and some of them are not. In this article, both kinds of algorithms are discussed. Though majority of the algorithms are meant for both positive and negative association rule mining, and have equal importance, our focus is mainly on developments in negative association rule mining. Though we were able to cover only few algorithms in this article, it is certainly helpful in literature survey of both positive and negative association rule mining.

Key words: Apriori Algorithm, Negative frequent item sets, Association rules, Chi-square

N.V.S.Pavan Kumar

INTRODUCTION:

Association is defined as the relation among different items. Both positive and negative association rule mining has their importance. Right from the Apriori algorithm, many algorithms and theses were proposed for the positive association rule mining. Where as very little effort or few algorithms were developed for negative association rule mining because finding negative item sets is expensive as the search space is exponentially large, compared to positive association rule mining. These association rules are developed from item sets. Consider the example, market-basket analysis, where the item sets which appear together are considered to be positive item sets and which do not appear together are considered as negative item sets.

Definitions:

A. Association Rules: Let I = {i1, i2…in} be a set of items. Let D be a set of transactions. Each transaction T is a set of items uniquely identified by TID such that T $\subseteq$ I. A transaction T is said to contain X (a set of items in I) if X $\subseteq$ T.

An association rule of the form "X => Y" is an implication, where X $\subseteq$I; Y $\subseteq$I, and X ∩ Y = Φ. X is called antecedent and Y is called consequent.

support of the rule X=> Y is s% of the transactions in D containing X U Y.

confidence of the rule X=>Y is c% of transactions in D that contain X also contain Y. strong rules are the association rules that have a support and confidence greater than given thresholds and this framework is known as the support-confidence framework.

A negative association rule is an implication of the form X => ┐Y (or ┐ X => Y or ┐ X => ┐ Y), where X$\subseteq$ I, Y $\subseteq$ I and X ∩ Y = Φ.

The support of X =>┐ Y is the frequency of occurrence of transactions with item set X in the absence of item set Y.

Confidence of X =>┐ Y can be calculated with P(X┐ Y)/P(X), where P (.) is the probability function.

Brin et al [1] named negative relationships for the first time in 1997. Their approach is based on chi-square. In his theses, independence between variables is checked by statistical test. The nature of the relation ship (positive or negative) is found by a correlation metric.
Savasere et al [2] in 1998 proposed an alternative approach, by combining positive frequent item sets based on their domain used to mine negative associations. Not only it is domain dependent but also it requires pre defined taxonomy, made it difficult to generalize.
Along with support-confidence frame work, Wu et. al [3] used another measure named *mininterest* in their algorithm to find both positive and negative association rules in the year 2002. According to them, the rule A→ B is interesting only if supp (AUB)-supp (A) supp (B)>=*mininterest*. But there is no formula suggested for calculating mininterest.
Maria-Luiza Antonie et al [4] proposed an algorithm to discover negative association rules with strong negative correlation between antecedents and consequents. In place of miniterest, they used another parameter called correlation coefficient (ρ) to find interestingness of association rules. This correlation coefficient (ρ) is added to the support-confidence framework to prune uninterested associations. This could be calculated as

$$\rho = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

Where Cov(X, Y) stands for covariance and σ_x ,σ_y stands for standard deviation of X and Y respectively.
If ρ=0 both the variables are independent
If ρ=1 both the variables positively correlated and
If ρ= -1 both the variables are negatively correlated.

Correlation coefficient ϕ is introduced by Pearson as

$$\phi = \frac{f_{11}f_{00} - f_{10}f_{01}}{\sqrt{f_{+0}f_{+1}f_{1+}f_{0+}}}$$

and could be transformed as

$$\phi = \frac{f_{11}(N - f_{10} - f_{01} - f_{11}) - f_{10}f_{01}}{\sqrt{f_{+0}f_{+1}f_{1+}f_{0+}}}$$

$$\phi = \frac{f_{11}N - f_{11}f_{10} - f_{11}f_{01} - f_{11}^2 - f_{10}f_{01}}{\sqrt{f_{+0}f_{+1}f_{1+}f_{0+}}}$$

$$\phi = \frac{f_{11}N - (f_{11} + f_{10})(f_{11} + f_{01})}{\sqrt{f_{+0}f_{+1}f_{1+}f_{0+}}}$$

$$\phi = \frac{Nf11 - f_{1+} * f_{f+1}}{\sqrt{f_{1+}(N - f_{1+})f_{+1}(N - f_{+1})}} \; .$$

Where X and Y are two binary variables and N is the size of dataset considered. The cells of 2x2 contingency table represent all the possible combinations of X and Y.

The strength of the correlation coefficient is discussed by Cohen [5]. According to him, the correlation value above 0.5 is large, 0.5 to 0.3 is moderate and 0.3 to 0.1 is small. Any thing less than 0.1 is not substantial and trivial. Hence they are not considered. Keeping this in mind, above algorithm is tested with correlation above 0.5 and slowly slides down till 0.1 and the results are observed.

Similar kind of approach is used in the algorithm PNAR, proposed by Honglei Zhu et al[6] in 2008 to mine both positive and negative association rules in transactional databases. In this they adopted metric called minimum support count for frequent negative item sets.
The rules of the form A=> not B and not A=> B are two contradictions, mined at a time and differentiated by using the correlation coefficient defined as
CorrAB= supp (AUB) / supp (A) supp (B)
They adopted a pruning strategy that do not consider the negative association rule of the form not A=>not B. this is for example, customers not to buy neither coffee not tea, is not much useful finding in market basket analysis. Different thresholds are used along with correlation and coefficient measure in this PNAR algorithm in order to find all valid association rules quickly.

e- NFIS is another kind of algorithm proposed by Xiangjun Dong et al [7] to find Negative frequent item sets based on positive frequent item sets in 2011. This paper mainly deals with NFIS (Negative Frequent Item Sets). They followed a method of extracting negative item sets only through positive item sets without using any additional metrics. This is a two step procedure. In step one, NFIS (negative frequent item sets) are mined (a not a, b not b) and in step two NAR (Negative Association Rules) are developed from these.
Frame work of e-NFIS

1. Positive Frequent Item Sets (PFIS) are find using any of the traditional methods like Apriory or FP-growth.
2. Negative Candidate Item Sets (NCIS) are generated from these PFIS.
   This is done by changing any m distinct Item Sets to their negative partners where M=1,2,3…, k-1 for a k-Item Set in PFIS.
3. Calculate the support of these NCIS using PFIS

For this calculation, they used an equation derived by X.Li et al [8]. This equation is developed based on inclusion- exclusion principles of set theory.

Ramasubbareddy et al [9] proposed an algorithm NAR (Negative Association Rules) to mine positive and negative association rules in 2010. This algorithm is developed based on Apriori Algorithm. Apriori algorithm has two steps namely *join* and *prune* steps. In step one (join), all frequent item sets of previous level are joined to obtain candidate item sets of present level. In step two (pruning), Apriori property is applied to find valid item sets denoted by PCk. NCk denotes negative candidate sets are obtained by replacing each literal by its negated item in PCk. Support of negative item sets can be obtained from positive item sets.

B. Kavitha Rani at al [10] recently (2011) published their work regarding the negative association rules i.e. for e.g. the products that conflict with each other or the products that complement each other in the market basket analysis. In this paper they had proposed an algorithm to find positive and negative association rules using an interesting measure called conviction along with support-confidence framework.

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

This is interpreted as the ratio of expected frequency of X occurs without Y divided by observed frequency of incorrect predictions. The range of conviction can be 0 to infinity.

CONCLUSION

In this article we presented our study on various developments in finding positive and negative association rule mining. Our study is mainly focused on negative association rules as it is difficult to find, compared to positive association rules. This is because the search space increases exponentially for negative Item Sets. So far many attempts were made in different ways to find negative association rules using support- confidence frame work. Importance of additional measures like correlation coefficient are also covered to some extent.

REFERENCES

[1].Brin, S., Motwani, R., Silverstein, C.: Beyond market basket: Generalizing association rules to correlations. In: Proc. of SIGMOD. Tucson, Arizona, 1997, 265–276.

[2]. 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

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

[4]. M.L. Antonie and O.R. Za¨ıane, ”*Mining Positive and* Negative Association Rules: an Approach for Confined *Rules”*, Proc. Intl.

[5]. Cohen, J.: Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum, New Jersey (1988)

[6]. Honglei Zhu, Zhigang Xu. :An Effective Algorithm for Mining Positive and Negative Association Rules. IEEE (2008)

[7]. Xiangjun Dong, Liang Ma, Xiqing Han.:e-NFIS: efficient Negative Frequent Item sets Mining only based on Positive Ones. IEEE(2011)

[8]. X. Li, Y. Liu, J. Peng, Z. Wu.: “the extended association rules and atom association rules”.Journal of computer research and development, vol.39, no.12,Dec2002.pp.1740-1750.

[9]. B.Ramasubbareddy, A.Govardhan, and A.Ramamohanreddy.: Mining Positive and Negative Association Rules, IEEE ICSE 2010,Hefeai, China, August 2010

[10]. B. Kavitha Rani, K.Srinivas, B.Ramasubba Reddy.:Mining Negative Association Rules,IJET vol.3(2),2011,100-105.

[11].Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. In: Proc. of SIGMOD. (1993) 207–216

[12].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/.

[13]. Yuan, X., Buckles, B., Yuan, Z., Zhang, J.: Mining negative association rules. In: Proc. of ISCC. (2002) 623–629

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

[15].B.Ramasubbareddy, A.Govardhan, and A.Ramamohanreddy. :Adaptive approaches in mining negative association rules. In intl.conference on ITFRWP-09, India Dec-2009.

[16]. R. Agrawal and R. Srikant.: Fast algorithms for mining associationrules. In VLDB, Chile, September 1994.

[17]. J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. In

SIGMOD, dallas, Texas, 2000.

[18]. D. Thiruvady and G. Webb. Mining negative association rules using grd. In PAKDD, Sydney, Australia, 2004

[19]. Gourab Kundu, Md. Monirul Islam, Sirajum Munir, Md. Faizul Bari ACN: An Associative Classifier with *Negative Rules* 11thIEEE International Conference on Computational Science and Engineering, 2008.

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