Asian Journal of Computer Science and Technology
ISSN: 2249-0701 (P) Vol.7 No.1, 2018, pp.46-49
© The Research Publication, www.trp.org.in
DOI: https://doi.org/10.51983/ajcst-2018.7.1.1831

Discovering Efficient Association Rule Mining via Correlation Analysis

C. Anuradha[1] and R. Anandavally[2]
[1&2]Assistant Professor, Department of Computer Science and Applications
Sreemath Sivagnana Balaya Swamigal Tamil, Arts & Science College, Mailam, Tamil Nadu, India
E-Mail: anumphil14@gmail.com, anandhi05@gmail.com

(Received 3 January 2018; Revised 25 January 2018; Accepted 5 February 2018; Available online 14 February 2018)

Abstract **- A Discovery of Association rule mining is an essential task in Data Mining. Traditional approaches employ a support confidence framework for finding association rule. This leads to the exploration of a number of uninteresting rules, such rules are not interesting to the users. To tackle this weakness, this paper examines the correlation measures to augment with support and confidence framework, which resulting in the mining of correlation rules. We then added an additional interesting measure based on statistical significance and correlation analysis. This paper reveals an overview of interesting measures and gives an insight into the discovery of more meaningful rules from large applications than traditional approach. Also it covers a theoretical issues associated with correlations that have yet to be explored.**

Keywords: Correlation, Cosine, null-invariant, support-confidence framework

I. INTRODUCTION

Data mining is used to extract the hidden information from large databases. It performs data analysis and may uncover important data patterns, contribute greatly to business strategies, knowledge bases, and scientific and medical research. The data patterns represent the knowledge embedded in the vast amount of data. One of the important techniques in finding frequent data pattern in mining is association rules [11]. Frequent item set mining leads to the discovery of associations and correlations among items in large transactional or relational or other data repositories.

The major role of the association rule mining is to find sets of binary variables that co-occur together frequently in a database [10]. The goal behind this is to identify groups of variables that are strongly correlated with each other or with a specific target variable. Moreover, even a strong association rules can be uninteresting and misleading. To examine this, we added an additional measure to find association to evaluate more meaning rules. Also this paper gives an overview of various correlation measures and their implementations to determine which would be good for mining rules from large data sets.

II. ASSOCIATION RULE ANALYSIS

Association rule mining can be identified based on the two objective measures.

1. The rule support X=>Y is taken to the probability of P (XUY).
2. The rule confidence X=>Y is taken to the probability of P (Y│X). Each of these measures is associated with a threshold. Rules below the threshold would likely reflect the noise or exceptions.

Although Objective measures are insufficient for identifying the interesting patterns. The rule generated by association rule mining may have redundancies that can be detected by correlation analysis. The support-confidence framework based on the statistics “behind” the data can be used as one step toward the goal of weeding out uninteresting rules from presentation to the user [1].

Example 1: Suppose we have data from Electronic stores with respect to the purchase of Apple I phone and Digital camera. Of the 10,000 transaction, data shows that 6,000 transactions include Apple I phone, while 7,500 included Digital camera, and 4000 included both. Let the data mining program is on the data using a minimum support 30% and a minimum confidence of 60%.The following association rule is:
buys (X,”Apple I Phone”)=>buys(X,”Digital camera”)
[Support=40%, Confidence=66%]

The above rule satisfies the threshold and thus is strong association rule but is misleading because the probability of purchasing camera is 75%, larger than 66%. Therefore the two items are negatively associated because purchase one item decreases the other. Also it does not measure the real strength of the correlation between the two items. To overcome this, correlation analysis measure can be used to mining association rules and thus resulting in a correlation rules.

A. Correlation Analysis

The word correlation is used in everyday life to denote some form of association. Correlation analysis measures the relationship between two quantitative variables and the degree of associations is also measured [2]. That is when correlation between two items, one item is called the “dependent” item and the other the “independent” item. The correlation coefficient can range between plus or minus one.

1. A coefficient of +1.0, a "perfect positive correlation," means that changes in the independent item will result in an identical change in the dependent item.

2. A coefficient of -1.0, a "perfect negative correlation," means that changes in the independent item will result in an identical change in the dependent item, but the change will be in the opposite direction.
3. A coefficient of zero means there is no relationship between the two items and has no effect on the dependent variable [2].

The goal is any change in the independent item will result in a change in the dependent item. This helps us to understand an indicator's predictive abilities [11]. Also correlation does not imply causality. That is, if X and Y are correlated, this does not imply that X causes Y or that Y causes X [2] [1] abilities [11]. Also correlation does not imply causality. That is, if X and Y are correlated, this does not imply that X causes Y or that Y causes X [2] [1]. We should mention the very interesting case where two related variables are separated by several steps in cause-effect chain of events Fig.1 illustrates this example.

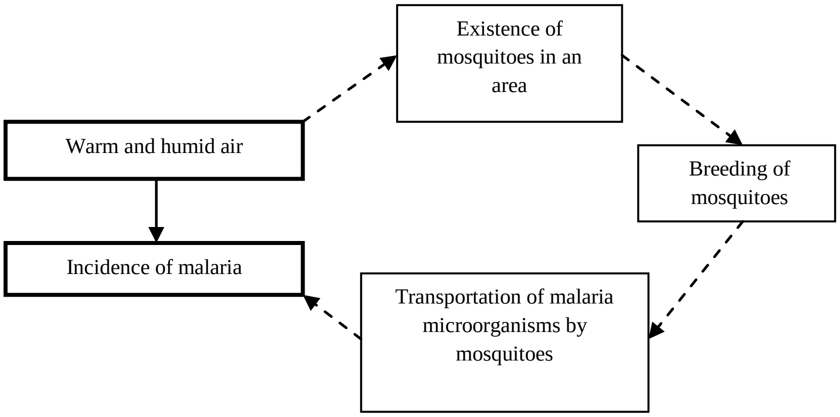
Existence of mosquitoes in an area
Warm and humid air
Breeding of mosquitoes
Incidence of malaria
Transportation of malaria microorganisms by mosquitoes

Fig. 1 Correlation does not imply Causality

The objective of association rule mining is to discover correlation relationship among a set of items. One difficulty is how to select a proper interestingness measures. An efficient method for generalizing associations to correlations is given in Brin,Motwani, and Silverstein [BMS97].There are also several correlation measures to determine which is good for mining large data sets. Since larger data sets typically have many null-transactions, it is important to consider the null-invariance property when selecting appropriate interesting measure in the correlation analysis. A measure is null-invariant if its value is free from the influence of null-transactions [9]. Therefore a good strategy is to understand their implications and limitation while interpreting correlation measures.

III. COMPARATIVE STUDY OF CORRELATION MEASURE

Once rules are extracted, the next step consists in picking out interesting rules and excludes the uninteresting [3]. To make the measures comparable all measures are defined using probabilities. We will first present some available measures, their meanings and then compare them on a series of datasets. The Table I shows the comparative studies of interesting measures done in [5]. The Key properties that a good measure M should satisfy:

P1: M=0 if A and B are statistically independent

P2: M monotonically increase with P (A, B) when P (A) and P (B) remain the same

P3: M monotonically decreases with P (A) (or P (B) when rest of the parameters P (A, B) and P (B) or P (A) remain unchanged.

It also explores that some of the measures have difficulty distinguishing correlation relationship because they are strongly influenced by null transactions [5]. A question arises which one is best at accessing correlation in all cases? Let's examine these measures in typical data sets in an illustration.

IV. CASE STUDY – TRANSACTIONAL DATA SETS

Here the correlation relationship between the purchase of two items coke and bottled water can be examined by in the form of Table II, a 2x2 contingency table, where an entry cw represents transaction containing both coke and bottled water.

TABLE I COMPARATIVE STUDIES OF INTERESTING MEASURES

Measures	Definition	Descriptions	Range	Null-invariant
χ^2	$\sum_{i,j=1}(o_{ij} - e_{ij})/ e_{ij}$	Need observed and expected value for computation. Less accurate	(0,∞)	No
Lift(X,Y)	P(XUY)/P(X)P(Y)	It Assesses the degree of lift. It is symmetric	(0,∞)	No
Coherence (X,Y)	Sup(XY)/Sup(X)+Sup(Y)-Sup(XY)	1. It needs order preserving transformation 2. Not applicable for unbalanced support	(0,1)	Yes
All_Conf(X)	Sup(X)/max_item_Sup(X)	1. It possesses the downward-closed closure property. 2. May not be appropriate for unbalanced support	(0,1)	Yes
Cosine(X,Y)	P(XUY)/Sqrt(P(X)P(Y))	1. It is harmonized lift measure. Consider only support of X and Y. 2. Most appropriate one for unbalanced cases.	(0,1)	Yes
Kulc(X,Y)	Sup(XUY)(1/Sup(X)+1/Sup(Y) 2	1. It is not sensitive with the number of transactions 2. Most appropriate one for unbalanced cases.	(0,1)	Yes
MaxConf (X,Y)	Max{Sup(XY)/Sup(X), Sup(XY)/Sup(Y)}	It is not sensitive to null transactions	(0,1)	Yes
Other Measures				
Collective Strength	C(Z) = (1-v(Z))/(1-E[v(Z)]) * E[v(Z)]/v(Z)	V(Z)-violation rate. It produces value 1 even if itemset appears more than expected	(0,∞)	No
Conviction (X,Y)	(1-P(Y))/1-Conf(X->Y)	It gives measure of implication and not just co-occurrence and Symmetric in nature	(0,∞)	No
Leverage (X, Y)	P(X and Y) - (P(X)P(Y))	Find out occurrences of XY are sold than expected from the independent sells. It suffers from rare item problem	(0,∞)	No
Corr(X,Y)	P(Y|X)/P(Y)	Determine if correlation is statistically significant.	(0,1)	No

TABLE II A 2X2 CONTINGENCY TABLE

	Coke	¬Coke	∑row
Water	cw	¬cw	w
¬ Water	c¬w	¬c¬w	¬w
∑col	c	¬c	∑

TABLE III COMPARISON OF MEASURES USING CONTINGENCY TABLE FROM DIFFERENT DATA SETS

Data Set	cw	¬cw	c¬w	¬c¬w	χ2	Lift	All Conf	Cosine	Coherence	Kulc	Max Conf
A1	10,000	1000	1000	100,000	9055.7	9.26	0.91	0.91	0.83	0.91	0.91
B1	1000	1000	1000	100,000	24730.7	25.75	0.5	0.5	0.33	0.5	0.5
C1	100	1000	1000	100,000	670.0	8.44	0.09	0.09	0.83	0.9	0.09
D1	1000	100	10,000	100,000	8,172.8	9.18	0.09	0.29	0.83	0.5	0.91
E1	10,000	1000	1000	100	0.0	1.0	0.91	0.91	0.83	0.91	0.91

From the Table III, we see that all the measures are good indicators for independent case, E1. Lift and χ2 are poor indicators of the other relationships, whereas all other measures are good. An interesting fact is that cosine is the better indicator when ¬cw and c¬w are unbalanced. Such a difference can be seen by Comparing C1 and D1.C1 should be more negatively correlated for c and w than D1 because cw is the smallest among three counts in C1 [1]. This can only be seen by checking the remaining measure values are identical in C1 and D1 except in Cosine. Also we found that Lift and χ2 have difficulty in finding correlation relationships because they are strongly affected by ¬c¬w. Besides this, the table explores Null-invariance is an important property for capturing meaningful correlations in large transactional databases was demonstrated.

In addition, a null-transaction is a transaction that does not contain any of the itemsets being examined. If we use correlation measures which are not null-invariant, the relationships between objects may appear or disappear simply by changing the number of transactions which do not contain items [9]. Analysis shows, a good strategy is to perform the all-confidence, kulc or cosine analysis first, and when the results shows that they are weakly Positively/negatively correlated, other analysis can be performed to assist in obtaining a more complete picture.

V. ISSUES IN CORRELATION ANALYSIS

1. Correlations are hard to interpret
2. Hidden variable

Consider this example: To infer that smoking Causes lung cancer, we would argue that people should stop smoking to lowering lung cancer rates. If smoking does not cause lung cancer, however, then stopping smoking would actually have no effect on lung cancer rates. This raises a question: How can a correlation not reflect causation? Hence the selection of best measures is still no universally accepted for judging interesting pattern

VI. CONCLUSION

In this article, we present an overview of some of the correlation measures and the importance of null-invariant measures. In addition, the use of only support and confidence measures to mine associations may leads to be uninteresting to the users. Thus, Correlation is a valuable type of scientific evidence in many applications. But first correlations must be confirmed as real, and then every possible causative relationship must be systematically explored. In general, the added measure substantially reduces the number of rules, and leads to discover more meaningful rules. However, there seems to be no single measure that works well for all cases. Although, in-depth research is still needed, so that field may have its long shine and the test results in not conclusive. For future, it would be interesting to see how this correlation analysis may influence real world problem, such as psychology, medicine and clustering.

REFERENCES

[1] Jiawei Han and Micheline Kamber, “Data Mining Concepts and Techniques”.

[2] [Online]. Available: http://www.bmj.com/about-bmj/resources-Readers/publications/statistics-square-one/11-Correlation-and-regression, *British Medical Journals*, published by BMJ Publishing Group.

[3] Tan, Pang-Ning, Vipin Kumar and Jaideep Srivastava, “Selecting the Right Interestingness Measure for Association Patterns”, in *Proc. of the 8th ACM SIGKDD, Int. Conf. on Knowledge discovery and data mining*, pp. 32-41, 2002.

[4] Merceron, Agathe and Kalina Yacef, “Revisiting Interestingness of Strong Symmetric Association Rules in Educational Data”, in *Proc. of Int. Workshop on Applying Data Mining in e-Learning, Greece*, pp. 3-12, 2007.

[5] Wu, Tianyi, Yuguo Chen and Jiawei Han, “Association Mining in Large Databases: A Re-Examination of its Measures”, in *European Conference on Principles of Data Mining and Knowledge Discovery, Springer*, pp. 621-628, 2007.

[6] Aggarwal, C. Charu and Philip S. Yu, “Mining Associations with the Collective Strength Approach”, *IEEE Transactions on Knowledge and Data Engineering,* Vol.13, No. 6, pp.863-873, 2001.

[7] Anita and Wasilewska, “Association Analysis”, Lecture Notes.

[8] Brin, Sergey, Rajeev Motwani and Craig Silverstein, “Beyond Market Baskets: Generalizing Association Rules to Correlations”, in *Acm Sigmod Record*, Vol. 26, No. 2, pp. 265-276, 1997.

[9] Kim, Sangkyum, Marina Barsky and Jiawei Han, “Efficient Mining of Top Correlated Patterns Based on Null- Invariant Measures”, in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer*, pp. 177-192, 2011.

[10] Achelis and Steven B, Technical Analysis from A to Z, New York, McGraw Hill, 2001.