Analysis and Design Interestingness Measure of Correlation: From Association Mining to Correlation Analysis

Parag M. Moteria, MCA, M.Phil.
MCA Department,
ISTAR
Vallabh Vidyanagar, India
paragmoteria@gmail.com

Dr. Y. R. Ghodasarad, MCA, Ph.D.
AIT Department
Anand Agricultural University
Anand, India
Yrghodasara77@yahoo.co.uk

***Abstract*— Association rules with low support thresholds and mining for long patterns can be uninteresting or misleading. Support and confidence measures are insufficient at filtering out uninteresting association rules. To tackle this weakness, support and confidence framework can be supplemented with additional interestingness measures based on statistical significance and correlation analysis or association attributes. Objective interestingness measures, based on the statistics can be used as one step toward the goal of weeding out uninteresting rules [1]. Important property null-(transaction) invariance, for measuring associations among events in large data sets, but many measures do not have this property [2]. In this study, we examine a set of null-invariant and not null-invariant interestingness measures and proposed new null-invariant interestingness measure.**

***Keywords*— Association rule, Correlation Analysis, Frequent pattern mining, Interesting measures, Null-invariance**

I. Introduction

Most association rule mining algorithms employ a support-confidence framework. Some time, many interesting rules can be found using low support threshold. Whether or not a rule is interesting can be assessed either subjectively or objectively. Ultimately, only the user can judge if a given rule is interesting and this judgment being subjective, may differ from one user to another. But objective interestingness measures based on the statistics can be used as one step toward the goal of weeding out uninteresting rules. This leads to correlation or association rule of the form A=>B [support , confidence, correlation/association]. Although minimum support and confidence thresholds help weed out or exclude the exploration of a good number of uninteresting rules, many rules so generated are still not interesting to the users [1]. In a typical transactional database, a particular item i (e.g. coffee) appearing in a transaction T (i.e. i ∈ T) is often a small probability event. Since most transaction do not contain i, they are null transactions with respect to i. If the association among a set of events being analyzed is affected by the transaction that contain none of them, such a measure is unlikely to be high quality [2]. We first look at how even strong association rules can be uninteresting and misleading. We then discuss how the support-confidence framework can be supplemented with additional interestingness measures based on statistical significance and correlation analysis or association attributes with and without null-invariance.

II. Problem Statement

Strong association rules can be uninteresting and misleading. Suppose we are interested in analyzing transactions with respect to the purchase of item1 and item2. Of the 10,000 transactions analyzed, the data show that 6,000 of the customer transactions included item1, while 8,000 included item2, and 4,000 included both item1 and item2. Suppose that a data mining program for discovering association rules is run on the data, using a minimum support of 30% and a minimum confidence of 60%.

The following association rule is discovered:

buys(X, “item1”)) => buys(X, “item2”) [support = 40%, confidence = 66%] (1)

Equation (1), is a strong association rule and would therefore be reported, since its support value of 4,000/10,000 =40% and confidence value of 4,000/6,000 =66% satisfy the minimum support and minimum confidence thresholds, respectively. However, Equation (1) is misleading because the probability of purchasing item2 is 80%, which is even larger than 66%. In fact, item1 and item2 are negatively associated because the purchase of one of these items actually decreases the likelihood of purchasing the other. Without fully understanding this phenomenon, we could easily make unwise business decisions based on Equation (1).

To measure the relationship between itemsets the support confidence framework can be supplemented with additional interestingness measures based on statistical significance and correlation analysis or association of attributes [1].

III. IMBALANCE RATIO (IR) MEASURE [2]

In many applications it is important to quantify to what extent a data set is controversial in order for a data analyst to correctly understand the data. Therefore, as a complement of the null-invariant association measures, we propose a new measure called Imbalance Ratio to gauge the degree of imbalance between two events. Denote by IR(A, B) the imbalance ratio of events A and B. IR(A, B) defined by

$$IR(A,B) = \frac{|\ sup(Ab) - sup(aB)\ |}{sup(AB) + sup(aB) + sup(Ab)} \quad (2)$$

In Equation (1),
a is complement of A
b is complement of B

Table I.

Range of IR is [0,1]	
IR measure	***Type of correlation***
0	Balance
>0	Imbalance

IV. VARIOUS CORRELATION MEASURES

A. lift [1] *measure*

Lift is a simple correlation measure that is given as follows. The occurrence of itemset A is independent of the occurrence of itemset B if P(A U B) = P(A)P(B); otherwise, itemsets A and B are dependent and correlated as events. This definition can easily be extended to more than two itemsets. The lift between the occurrence of A and B can be measured by computing

$$lift(A,B) = \frac{P(A\ U\ B)}{P(A)P(B)} \quad (3)$$

Table II.

lift is not null-invariant measure	
Range of lift is [0, ∞]	
lift(A,B)	***Type of correlation***
<1	Negative
1	Independent
>1	Positive

B. Yule's coefficient of association method [3]

This method is the most popular statistical method of studying association because here not only we can determine the nature of association, i.e. whether the attributes are positively associated, negatively associated or independent, but also the degree or extent to which the two attributes is associated. The Yule's coefficient is denoted by 'Q' and obtained by

$$Q = \frac{(AB)(\alpha\beta) - (A\beta).(\alpha B)}{(AB)(\alpha\beta) + (A\beta).(\alpha B)} \quad (4)$$

Table III.

Yule is not null-invariant measure	
Range of Yule is [-1, +1]	
Yule	***Type of Association***
<1	Negative
1	Independent
>1	Positive

C. all_confidence [1] measure

Given an itemset X = {i_1, i_2, …, i_k}, the all_confidence of X is defined as

$$all\_conf(A,B) = \frac{Sup(X)}{max\_item\_sup(X)} \quad (5)$$

In Equation (5), max{sup(i_j) | for all i_j Є X} is the maximum (single) item support of all the items in X and hence is called the max_item_sup of the itemset X.

Table IV.

all_conf is null-invariant measure	
Range of all_conf is [0,1]	
all_conf(A,B)	***Type of correlation***
0	Negative
0.5	Independent
1	Positive

D. cosine [1] *measure*

The cosine measure can be viewed as a harmonized lift measure. Given two itemsets A and B, the cosine measure of A and B is defined as

$$cosine(A,B) = \frac{P(A\ U\ B)}{\sqrt{P(A)\ P(B)}} \quad (6)$$

Two formulae, lift and cosine are similar except that for cosine the square root is taken on the product of the probabilities of A and B. This is an important difference, because by taking the square root, the cosine value is only influenced by the supports of A, B and A U B, and not by the total number of transactions.

Table V.

cosine is null-invariant measure	
Range of cosine is [0,1]	
cosine(A,B)	***Type of correlation***
0	Negative

0.5	Independent
1	Positive

E. coherence [4] *measure*

coherence is defined by

$$coherence(A,B) = \frac{sup(AB)}{sup(A) + sup(B) - sup(AB)} \quad (7)$$

Table VI.

coherence is null-invariant measure	
Range of coherence is [0,1]	
coherence(A,B)	***Type of correlation***
0	Negative
0.5	Independent
1	Positive

F. kulc [4][5] *measure*

Kulczynski measure is also known as kulc measure. kulc measure yields the average conditional probability that a characteristic is present in one item given that the characteristic is present in the other item. The measure is an average over both items acting as predictors. It is obtained by

$$kulc(A,B) = sup(AB)\left(\frac{1}{sup(A)} + \frac{1}{sup(B)}\right) \quad (8)$$

Table VII.

kulc is null-invariant measure	
Range of kulc is [0,1]	
kulc(A,B)	***Type of correlation***
0	Negative
0.5	Independent
1	Positive

V. PROPOSED WORK

Our proposed measure is denoted by YP. We use notations from 2X2 contingency table with following notations.

Table VIII.

	A	**A**	**Total**
B	(AB)	(aB)	(B)
b	(Ab)	(ab)	(b)
	(A)	(a)	N

In Table VIII, (A), (a), (B), (b), (AB), (Aa), (aB), (ab) and N represent class of attributes. The number of records assigned to class is called their count or frequencies or class frequencies. The number of records or units belonging to class is known as its frequency is denoted within bracket. Thus, (A) stands for the numbers of items of attribute A [6]. Proposed measure YP is given by,

$$YP(A,B) = \frac{2*sup(AB)}{sup(A) + Sup(B)} \quad (9)$$

Why range of YP is [0, 1] ?
From Table VIII,
$0 <= sup(AB) <= sup(A)$ (a)
$0 <= sup(AB) <= sup(B)$ (b)

Add (a) and (b),

$0 <= (2*sup(AB)) <= sup(A) + sup(B)$
$0 <= (2*sup(AB)) / (sup(A)+sup(B)) <= 1$

Hence, $0 <= YP <= 1$

Table IX

Proposed measure is null-invariant measure	
Range of YP is [0,1]	
Proposed measure	***Type of correlation***
0	Negative
0.5	Independent
1	Positive

VI. WHICH MEASURE INTUITIVELY REFLECTS THE TRUE RELATIONSHIP BETWEEN A (E.G. MILK) AND B (E.G. COFFEE) [2]

Given two arbitrary events A and B, we denote the support of A, B, and AB as sup(A), sup(B) and sup(AB), respectively and use A and sup(A) interchangeably when there is no ambiguity. Let μ be any of the five null-invariant measures. From the definitions as per above, we immediately have the following fundamental properties:

P1 $\mu \in [0, 1]$

P2 μ monotonically increases with sup(AB) when sup(A) and sup(B) remain unchanged; and it monotonically decreases with sup(A) (or sup(B)) when sup(AB) and sup(B) (or sup(A)) stay the same

P3 μ is symmetric under item permutations

P4 μ is invariant to scaling, i.e., multiplying a scaling factor to sup(AB), sup(A) and sup(B) will not affect the measure

A unified framework discloses the underlying philosophies of the measures and explains their inherent relationships, which in turn may help a user's decision-

making in selecting the right measure for different application domains. To begin with, we rewrite the definitions as per above cases into the form of conditional probabilities as shown in fig. 1. We convert the support counts into conditional probabilities using Equation (10):

$$P(a|b) = \frac{sup(ab)}{sup(b)} = \frac{sup(ab)}{sup(ab) + sup(\bar{a}b)}$$
$$P(b|a) = \frac{sup(ab)}{sup(a)} = \frac{sup(ab)}{sup(ab) + sup(a\bar{b})} \quad (10)$$

Measure	Definition
$AllConf(a,b)$	$\min\{P(a\|b), P(b\|a)\}$
$Coherence(a,b)$	$(P(a\|b)^{-1} + P(b\|a)^{-1} - 1)^{-1}$
$Cosine(a,b)$	$\sqrt{P(a\|b)P(b\|a)}$
$Kulc(a,b)$	$(P(a\|b) + P(b\|a))/2$
YP(a,b)	**$2 * (P(a\|b)^{-1} + P(b\|a)^{-1})^{-1}$**

Fig. 1

all_conf, Coherence and YP require sup(A) ≠ 0 or sup(B) ≠ 0, while Cosine and Kulc require sup(A) ≠ 0 and sup(b) ≠ 0. For the definitions in figure 1, we require sup(A) ≠ 0 and sup(B) ≠ 0, because otherwise P(a|b) and/or P(b|a) would be undefined. For simplicity, we hereafter assume that all five measures will be equal to 0 if sup(AB) = 0.

Following the rewritten definitions, we can generalize all five measures using the mathematical generalized mean (Kachigan 1991) [2]. Specifically, each of the null-invariant measures can be represented by the generalized mean of the two conditional probabilities *P(a|b)* and *P(b|a)* as

$$\mathbb{M}^k(P(a|b), P(b|a)) = \left(\frac{P(a|b)^k + P(b|a)^k}{2}\right)^{\frac{1}{k}} \quad (11)$$

In Equation (11), M denotes the mathematical generalized mean and k ∈ (−∞,+∞) is the exponent (k is a real number). We have the following lemma.

Lemma: Each null-invariant measure in fig. 1 can be expressed using Equation (1) with its corresponding exponent [2].

Here, we only express proposed null-invariant measure YP using Equation (11).

Put k = -1 in Equation (1), we get

$$\mathbf{M}^{-1}(P(a|b), P(b|a)) = YP(A,B)$$

VII. Result

In a typical transaction database, a product appearing in a transaction is called an event, and a set of products appearing in a transaction is called an event-set. Association analysis is to identify interesting (positive or negative) associations among a set of events. It is expected that a particular event happens with a very low probability.

Table X.

	Milk	**~Milk**	**Total**
Coffee	Mc	~mc	c
Not Coffee	m~c	~m~c	~c
	M	~m	Total

In table X, purchase history of two events milk and coffee. Following table enumerates six data sets in terms of a flattened contingency table and represent analysis of result of all_conf, cosine, coherence, kulc, YP, lift and Yule interesting measure of correlation or association of attributes with Imbalance Ratio (IR).

In the fig. 2, "SHADED (BLUE)" color represents similar result and "NON-SHADED (YELLOW)" color represents dissimilar result by various null-(transaction) variant measures say, add_conf, cosine, coherence, kulc and YP as well as not null-(transaction) variant measures say, lift and Yule. In data set D004, coherence is disagree with others null-(transaction) invariant measures even if data set is balanced. In data set D005 and D006, kulc is neutral as compare to others null-(transaction) invariant measures when data set unbalanced.

Data Set	AB	A~B	~AB	~A~B	ALL CONF	COSINE	COHERENCE	KULC	YP	LIFT	YULE	IR
					RESULT	RESULT	RESULT	RESULT	RESULT	RESULT	RESULT	RESULT
D01	10000	1000	1000	100000	0.90909	0.90909	0.83333	0.90909	0.90909	9.25620	0.99800	0.00000
D02	10000	1000	1000	100	0.90909	0.90909	0.83333	0.90909	0.90909	1.00000	0.00000	0.00000
D03	100	1000	1000	100000	0.09091	0.09091	0.04762	0.09091	0.09091	8.43802	0.81818	0.00000
D04	1000	1000	1000	100000	0.50000	0.50000	0.33333	0.50000	0.50000	25.75000	0.98020	0.00000
D05	1000	100	10000	100000	0.09091	0.28748	0.09009	0.50000	0.16529	9.18182	0.98020	0.89189
D06	1000	10	100000	100000	0.00990	0.09901	0.00990	0.50000	0.01961	1.97049	0.98020	0.98990

Fig. 2

VIII. Conclusion

We present a comprehensive study of null-invariant interestingness measures from association mining to correlation analysis. Study of Null-(transaction) invariant is very important for relationship in large data sets. We can analyze our proposed interestingness measure of null-(transaction) invariant measure YP is agree in all cases as like the result of cosine.

References

[1] Jiawei Hanand Micheline Kamber, "Data Mining Concepts and Techniques - Second Edition", ELSEVIER Morgan Kaufman Publisher pp. 67-70

[2] Tianyi Wu, Yuguo Chen and Jiawei Han,

Association Mining in Large Databases: A Re-Examination of Its Measures", Proc. 2007 Int. Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD'07), Warsaw, Poland, Sept. 2007, pp. 621–628 (Received: 12 February 2009 / Accepted: 10 December 2009 / Published online: 06 January 2010)

[3] http://www.assignmenthelp.net/assignment_help/yules-coefficient-of-association.php

[4] http://www.nyu.edu/classes/jcf/g22.3033-002/slides/session6/MiningFrequentPatternsAssociationAndCorrelations.pdf

[5] http://publib.boulder.ibm.com/infocenter/spsstat/v20r0m0/index.jsp?topic=%2Fcom.ibm.spss.statistics.help%2Falg_proximities_kulczynski2.htm

[6] S.P.Gupta, "Statistical Methods", Sultan Chand & Sons Educational Publishers, 39 Revised Edition – 2010 pp. 474 – 514

[7] Seyda Ertekin, Jian Huand, Leon Bottou, Lee Giles, "Learning on the Border: Active Learning in Imbalance Data Classification", CIKM'07, November 6–8, 2007, Lisboa, Portugal. Copyright 2007 ACM 978-1-59593-803-9/07/0011 URL: http://web.mit.edu/seyda/www/Papers/CIKM07_LearningontheBorder.pdf

[8] Camelia Lemnaru and Rodica Potolea, "Imbalanced Classification Problems: Systematic Study, Issues and Best Practices" URL: http://www.search.utcluj.ro/articole/Imbalanced%20Classification%20Problems.pdf

[9] David Jensen and Jennifer Neville, "Correlation and Sampling in Relational Data Mining" URL: http://www.cs.purdue.edu/homes/neville/papers/jensen-neville-interf2001.pdf

[10] Da Kuang, Charles X. Ling, and Jun Du, "Foundation of Mining Class-Imbalanced Data" URL: http://cling.csd.uwo.ca/papers/pakdd12_imb.pdf

[11] Sangkyum Kim, Marina Barsky and Jiawei Han, "Efficient Mining of Top Correlated Patterns Based on Null-Invariant Measures" URL: http://csci.viu.ca/~barskym/HOME_PAGE/publications/PKDDConference2011.pdf

[12] Nguyen, G. Hoang., Bouzerdoum, A. & Phung, S. (2009). "Learning pattern classification tasks with imbalanced data sets". In P. Yin (Eds.), Pattern recognition (pp. 193-208). Vukovar, Croatia: In-Teh. URL: http://ro.uow.edu.au/cgi/viewcontent.cgi?article=1806&context=infopapers&sei-redir=1&referer=http%3A%2F%2Fwww.google.co.in%2Furl%3Fsa%3Dt%26rct%3Dj%26q%3Dlearning%2520pattern%2520classification%2520tasks%2520with%2520imbalanced%2520data%2520sets%26source%3Dweb%26cd%3D1%26cad%3Drja%26ved%3D0CDMQFjAA%26url%3Dhttp%253A%252F%252Fro.uow.edu.au%252Fcgi%252Fviewcontent.cgi%253Farticle%253D1806%2526context%253Dinfopapers%26ei%3DI-0qUfHDOsL9rAeZ-oCwDA%26usg%3DAFQjCNHYv_kkhnv0tr1RS9cdVsrFxjl9Dg%26bvm%3Dbv.42768644%2Cd.bmk#search=%22learning%20pattern%20classification%20tasks%20imbalanced%20data%20sets%22

[13] Sofia Visa, Anca Ralescu, "Issues in Mining Imbalanced Data Sets - A Review Paper" URL: http://secs.ceas.uc.edu/~aralescu/PAPERS/VRMaics2005.pdf