ISSN (Print) : 2319-8613
ISSN (Online) : 0975-4024

Two Novel Pioneer Objectives of Association Rule Mining for high and low correlation of 2-varibales and 3-variables

Hemant Kumar Soni[#1, a], Sanjiv Sharma[*2] , A.K. Upadhyay[#3]
[1,3] Department of Computer Science and Engineering,
Amity School of Engineering and Technology, Amity University, Gwalior, Madhya Pradesh, India
[1a] soni_hemant@rediffmail.com
[2]Department of Computer science and Engineering,
Madhav Institute of Technology and Science, Madhya Pradesh , India

Abstract : **Association rule generation is a significant research area of data mining, which find out the relation between the set of items . Significant association rule mainly based on two objectives – support and confidence. Some other metrics are also available to evaluate the goodness, effectiveness and interestingness of an association rule. Therefore, the association rule mining problem can be treated as multi-objective optimization problem. In this paper, we discuss the various objectives and their limitation. It is found that, each and every objective are not suitable in every situation. Other than this , most of the objectives are defined for 2-variables only. Simultaneously, in certain situation correlation analysis does not show the positive and negative correlation between items. Authors proposed two novel objectives, high correlation and low correlation for 2-variables and 3-variables. Through numerical analysis it is found that proposed objective clearly indicate about the positive and negative correlation among items. These objectives also gives appropriate solution in those cases, where previously defined objectives have some limitations. Simultaneously it also works in Simpson's paradox situation successfully.**

Keyword : Association Rule Mining (ARM), Interest factor, Lift, Interestingness, Comprehensibility, Correlation analysis, correlation coefficient, high and low correlation.

I. INTRODUCTION

Association Rule Mining (ARM) is one of the technique used in Data Mining, is based on two objectives – support and confidence. Numerous algorithms has been discussed in [1,2] for frequent pattern generation and association rule mining. Application domain of ARM is very vast including time –series data analysis, e-commerce and recommendation systems[3-5]. The drawback of association rules generated by using support and confidence objectives are, if threshold value for support and confidence is low then it will generate too many rules. Whereas if the threshold value is high, it will generate few rules [6]. This resultant some important association rules may be missed. Sometimes association rules based on support and confidence, mislead us [7] . There are lack of interestingness and comprehensiveness are also found in this kind of association rules. Thus goodness, effectiveness and interestingness of association rules cannot be evaluated only in the basis of support and confidence [8].

Most of the data mining problem are multi-objective in nature [9]. There are many other parameters defined by the researchers to measure the goodness of association rules. Lift, J-measure, Interestingness, Comprehensibility, Interest Factor, Correlation Analysis are some of the objectives proposed in literature [7,10 - 15]. Almost every objectives have some drawback and every objective is also not applicable in every situation [10]. Only few objectives like Interest Factor, Cosine (IS), Piatetsky-Shapiro's (PS) and Jaccard Coefficient (ζ) support 3-variables [7]. In this paper authors proposed two novel objectives – High Correlation and Low Correlation for 2-variables and 3-variables. These objective indicate the positive correlation and negative correlation between items for 2-varibales and 3-variables.

The rest of the paper is organised as follows. In Section 2, we discuss the basic concepts of ARM and various objectives. In Section 3, we discuss the limitations of various objectives and their application domain. In section 4, we proposed two novel objectives- High Correlation and Low Correlation, for 2-variables and 3-variables. Analysis of proposed objectives and their comparison with the existing measures has been discussed in Section 5. Experimental evaluation and results has been discussed in section 6. Finally we summarize and discuss future work in Section 7.

II. BASIC CONCEPTS

Association rule mining can be defined formally as follows: Let $I = \{i_1, i_2, \ldots i_m\}$ be a set of literals, called items. Let D be a set of transactions, where each transaction T is a set of items such that $T \subseteq I$. associated with each transaction is a unique identifier, called its TID. In this concept say that a transaction T contains X, a set of some items (called itemset) in I, if $X \subseteq T$.

2.1 Association Rules

For a given transaction database T, An association rule is an implication of the form X ⇒ Y, where $X \subset I, Y \subset I, and\ X \cap Y = \Phi$, i.e. X and Y are two non-empty and non-intersecting itemsets. The rule $X \Rightarrow Y$ holds in the transaction set D with confidence c if c % of transactions in T that contain X also contain Y.

2.2 Support

A transaction T is said to support an item i_k, if i_k is present in T. T is said to support a subset of items $X \subseteq I$, if T support each item i_k in X. An itemset $X \subseteq I$ have a support s in D, denoted by s(X), if s% of transactions in D support X. There is a user-defined minimum support threshold, which is a fraction, i.e., a number in [0, 1].

$$Support\ (X \Rightarrow Y) = Support\ (X \cup Y) / |D| \quad \text{------ (1)}$$

2.3 Confidence

The confidence of rule $X \Rightarrow Y$ is the fraction of transactions in D containing X that also contain Y and indicates the strength of rule.

$$(X \Rightarrow Y) = Support\ (X \cup Y) / Support\ (X) \quad \text{------ (2)}$$

Other than Support and confidence, some other objectives such as Comprehensibility, Coverage, Cosine, lift, Laplace, Jaccard, J-measue, prevalence, surprise, recall, conviction, surprise and so on are also available in literature[10,16].

2.4 Comprehensibility

Comprehensibility of an association rule is quantified by the following expression:

$$Comprehensibility = \frac{\log(1+|Y|)}{\log(1+|X \cup Y|)} \quad \text{------ (3)}$$

where |itemset| means the number of attributes involved in the itemset. As a simple sentence, if the number of conditions in the antecedent part is less, the rule is more comprehensible.

2.5 Interestingness

Interestingness measure is used to quantify how much the rule is surprising for the user. As the most important purpose of association rule mining is to find some hidden information, it should extract rules that have comparatively less occurrence in the database. The following expression can be used to quantify the interestingness:

$$Interistngness = \left(\frac{Support\ (X \cup Y)}{Support\ (X)}\right) \times \left(\frac{Support\ (X \cup Y)}{Support\ (Y)}\right) \times \left(1 - \frac{Support\ (X \cup Y)}{Support\ (Z)}\right) \quad \text{------ (4)}$$

where Support(Z) is the number of records in the database [13,17].

However, most researchers have adopted Piatetsky-Shapiro's [18] argument that a rule cannot be interesting, if its antecedent and consequent are statistically independent.

2. 6 Lift

Lift compute the ratio between the rule's confidence and the support of the itemset in the rule consequent. Lift is equivalent to the ratio of the observed support to that expected if X and Y were statistically independent.

$$Lift = \frac{Confidence\ (x \rightarrow y)}{(Support\ (y))} \quad \text{------ (5)}$$

2.7 Interest Factor

For binary variables, lift is equivalent to another objective called interest factor which is defined as follows :

$$I(A,B) = \frac{S(A,B)}{S(A)\ X\ S(B)} = \frac{N\ f11}{f(1+)\ f(+1)} \quad \text{------ (6)}$$

2.8 Correlation Analysis

Correlation analysis is a statistical based technique for analyzing relationship between a pair of variables. For continuous variable, correlation is defined using Person correlation coefficient. For binary variables correlation can be measured as –

$$\phi = \frac{f(11)f(00) - f(10)f(01)}{\sqrt{(f(1+)f(+1)f(0+)f(+0)}} \quad \text{------ (7)}$$

The value of correlation range form -1(perfect negative correlation) to +1 (perfect positive correlation) [19]. If the variables are statistically independent than ϕ =0. The correlation between Tea and Coffee drinker given in Table 2 is -0.0625.

For assessing the worthiness of the association rule, objectives are the only measure. A contingency table consist the frequency count, which may used to measure objectives.

Table I. A 2-dimensional contingency table for 2-variables.

Variable	B	nB	Sum
A	f(1,1)	f(1,0)	f(1,+)
nA	f(0,1)	f(0,0)	f(0,+)
Sum	f(+,1)	f(+,0)	N

Table-I is an example of 2-dimensional contingency table for 2-variables, A and B. nA (Negative A) and nB (Negative B) indicate the absence from a transaction. The cells consist of f(i, j) represent the frequency count. Cell value f(1,1) represent the co-occurrence of A and B together whereas f(0,1) illustrate the absence of A but presence of B in transaction. Cell f(1,+) shows the support of A and cell f(+,1) demonstrate the support of B.

III. LIMITATION OF DIFFERENT OBJECTIVES

3.1 Limitations of the Support-Confidence Framework

Available technique to find out association rule mining is based on two objective- support and confidence. Many association rules can be identified if support is low and eliminated if support is high. Whereas the effect of confidence is more vital. This can be understand by the example given below. Table- II illustrate the choices of 1000 people who likes different beverage.

Table II. Preferences of 1000 people for beverage.

	COFFEE	nCOFFEE	
TEA	150	50	200
nTEA	650	150	800
	800	200	1000

With the help Table-II, we can identify the association rule Tea →Coffee. This rule has 15% support and 75% confidence, which are reasonable high. It means the people who like Tea also like Coffee. On the other hand, 80% of people drink Coffee, irrespective whether they drink Tea or not. Whereas the fraction of tea drinker who drink coffee is only 75%. Accordingly, a Tea drinker reduce the probability as a Coffee drinker from 80% to 75%. Hence the association rule Tea → Coffee is deceptive in spite of high confidence.

The snag of a confidence is that, it overlook the support of the item in the consequent part of the rule. In fact, if we check the support count of the Coffee drinker, then find out many of them who drink Tea also drink Coffee.

If we closely analyze the data given in Table-II, than some surprising facts are discover. We observe that proportion of Tea-Coffee drinker is quit less than the overall Coffee drinker. This indicate to an contrary relationship between Tea drinker and Coffee drinker.

3.2 Limitation of Interest factor

The occurrence of word pair {P, Q} and word pair {R,S} in same document, is given in below tables.

Table III. Contingency Table for the word Pairs {P,Q} and {R,S}.

	P	nP	
Q	880	50	930
nQ	50	20	70
	930	70	1000

	R	nR	
S	20	50	70
nS	50	880	930
	70	930	1000

As per the definition of interest factor and equation 6, the interest factor for word pairs {P,Q} and {R,S} is 1.02 and 4.08 respectively. The support count for word pairs {P,Q} is 88% and their interest factor is near to 1. It means P and Q are statistically independent. On the other hand support count of word pairs {R,S} is only 2% and their interest factor is 4.08. This is quite high in comparison of word pair {P,Q}. Consequently, the resultant value of interest factor of word pairs {P,Q} and {R.S} as given in Table-III is vey disquieting.

It is observe that confidence conceivably superior choice in this state. The calculated value of association between word pair {P,Q} is 94.6% is much higher than 28.6% for word pair {R,S}.

3.3 Limitation of Correlation Analysis

From the word association example given in Table-3, the shortcoming of correlation can be easily observe. In spite of co-occurrence of {P,Q} is more than {R,S}, the ϕ -coefficient for word pairs {P,Q} and {R,S} are identical, i.e. $\phi(P,Q) = \phi(R,S) = 0.232$.

Since ϕ-coefficient confer the same weight to both co-presence and co-absence of items. Hence ϕ-coefficient much appropriate for analyzing symmetric binary variables. If the sample size has been changed proportionately, the value of ϕ-coefficient will remain same. This is another drawback of this measure.

IV. HIGH AND LOW CORRELATION OBJECTIVES

In this section authors proposed 2 new objectives – High Correlation and Low Correlation, to calculate positive correlation and negative correlation between items for 2 variables and 3 variables.

4.1 High and Low Correlation objectives for 2-vaiables

ϕ- coefficient gives equal importance to both co-presence and co-absence of items in transactions. It is therefore more suitable for analyzing symmetric binary variable. To that overcome the drawback of correlation analysis, in this paper we proposed two new objective namely High Correlation ($h\phi_2$) and Low Correlation ($l\phi_2$) for 2-variables.

4.1.1 High Correlation for 2-variables : analyzing relationship between a pair of variables. It gives the importance to the co-presence of the variables. High Correlation compute as the ration of difference between Support(A,B) and Support(A, nB), with square root of Support(A, -) and Support(-,B)

$$h\phi_2 = \frac{f(1,1)-f(1,0)}{\sqrt{(f(1,+)f(+,1)}} \quad \text{------ (8)}$$

4.1.2. Low Correlation for 2-variables : analyzing relationship between a pair of variables. It gives the importance to the co-absence of the variables. Low Correlation compute as the ration of difference between Support(nA,nB) and Support(nA,B), with square root of Support(nA, -) and Support(-,nB)

$$l\phi_2 = \frac{f(0,0)-f(0,1)}{\sqrt{(f(0,+)f(+,0)}} \quad \text{------ (9)}$$

The High correlation between Tea and Coffee drinker given in Table 2 is **0.25** and the Low correlation between Tea and Coffee drinker is **-1.25**, where as the correlation between Tea and Coffee drinker given in Table 2 is -0.0625.

Simultaneously, the High correlation between P and Q given in Table 3 is **0.892473** and the Low correlation between P and Q is **-0.42857**and the High correlation between R and S is **-0.42857** and the Low correlation between P and Q is **0.892473,** which is vice versa. Whereas the correlation between P and Q is same as between R and S, which is **0.231951**.

4.2 High and Low Correlation objectives for 3-vaiables

In the literature, most of the objectives are defined for the relation between 2 variables. In this paper authors proposed 2 more new objectives which shows the relation among 3 variables. To illustrate the concept, we are using a three-dimensional contingency table for A,B and C as shown below.

Table IV. A 3-Dimensional Table for 3-variables.

C	B	n B	
A	f(1,1,1)	f(1,0,1)	f(1,+,1)
n A	f(0,1,1)	f(0,0,1)	f(0,+,1)
	f(+,1,1)	f(+,0,1)	f(+,+,1)

nC	B	n B	
A	f(1,1,0)	f(1,0,0)	f(1,+,0)
n A	f(010)	f(000)	f(0,+,0)
	f(+,1,0)	f(+,0,0)	f(+,+,0)

The cells of Table-IV, represent the presence and/or absence of the items in a transaction. For example f(I,J,K) shows the particular combination of items A,B and C. The cell value f(1,0,1) illustrate the presence of A and C but absence of B in number of transactions. Where as the cell f(1,+,1) represent the presence of A and C, irrespective of presence or absence of B in number of transaction.

4.2.1 High Correlation for 3-variables : analyzing relationship among 3 variables. It gives the importance to the co-presence of the variables.

$$h\phi_3 = \frac{f(1,1,1)-f(1,0,1)-f(1,1,0)-f(1,0,0)}{\sqrt{(f(1,+,1)f(1,+,0)f(+,1,1)f(+,1,0)f(+,+,1)}} \quad \text{------ (10)}$$

4.2.2 Low Correlation for 3-variables : analyzing relationship among 3 variables. It gives the importance to the co-absence of the variables.

$$l\phi_3 = \frac{f(0,0,0)-f(0,0,1)-f(0,1,0)-f(1,0,0)}{\sqrt{(f(0,+,1)f(+,0,1)f(+,0,0)f(0,+,0)f(+,+,0)}} \quad \text{------ (11)}$$

V. ANALYSIS OF PROPOSED OBJECTIVES

A relationship between sale of High definition Television (HDTV) and Exercise Machine(EM) is given in Table-V. A more classification of Table-5 data in the form of customer group is given in Table-6. The customer group comprises of College Students and working Adults .

Table V. A two-way contingency table between the sale of high-definition television and exercise machine.

Buy HDTV	Buy Exercise Machine		
	Yes	No	
Yes	99	81	180
No	54	66	120
	153	147	300

Table VI A three-way contingency table between the sale of high-definition television and exercise machine with customer group

Customer Group	Buy HDTV	Buy Exercise Machine		Total
		Yes	No	
College Students	Yes	1	9	10
	No	4	30	34
Working Adult	Yes	98	72	170
	No	50	36	86

Now Table 6 converted into another form of Three- dimensional contingency table as follows.

Table VII Contingency table for Working Adult.

C	B	nB	
A	98	50	148
nA	72	36	108
	170	86	256

Table VIII Contingency table for College Students

nC	B	nB	
A	1	4	5
nA	9	30	39
	10	34	44

Here A represent Buy-Exercise-Machine, nA represent Not-Buy-Exercise-Machine, B represent Buy-HDTV, nB represent Not-Buy-HDTV, C represent Working Adult, nC represent College Students respectively.

As revealed in Table-V, the relationship between buying of HDTV with EM has a 55% confidence whereas buying EM without HDTV has 45% confidence. At a glance, first rule looks more stronger than the second rule. However, a insightful analysis disclose that customer's category play a significant role in buying these items.

Table-VI include one more dimension i.e. customer group and show the buying patterns of HDTV and EM among college students and working adults. Table-VI also shows the breakup of the frequency according to the customer group. It also revealed that majority of the customers are working adults.

If we check the association between HDTV and EM for college students, the following are the output:

$$\text{confidence } \{\text{HDTV= Yes}\} \rightarrow \{\text{EM = Yes}\} = 1/10 = 10\%$$

$$\text{confidence } \{\text{HDTV= No}\} \rightarrow \{\text{EM = Yes}\} = 4/34 = 11.8\%$$

while the association between HDTV and EM for working adults are :

$$\text{confidence } \{\text{HDTV= Yes}\} \rightarrow \{\text{EM = Yes}\} = 98/170 = 57.7\%$$

$$\text{confidence } \{\text{HDTV= No}\} \rightarrow \{\text{EM = Yes}\} = 50/86 = 58.1\%$$

It is clear from the above result that irrespective of customer group, the customer who do not buying HDTV is more likely to buy EM. This result is simply contradict the previous result, when customers are not classified. With other metrics like Correlation, Odds Raito, or Interest, we still on the conclusion that buying of HDTV with EM is positively correlated in combined data and negatively correlated in separate data. This turnaround association is known as Simpson's paradox.

The large number of the customers who purchase HDTV and/or EM are working adults. Since total 85% of the customers are working adults, hence the observed relationship between HDTV and EM are much stringer in grouped data rather than ungrouped data.

The ϕ- coefficient for Table VII is, **-0.00471** and for Table VIII, it is **-0.0233**, which shows the negative correlation between items in both the cases. Whereas if we calculate the value of our proposed objectives, then we get **0.302612 for hϕ_2 and -0.37354 for lϕ_2** respectively for Table VII and **-0.42426 for hϕ_2 and 0.576697 for lϕ_2** respectively for Table VIII. Calculated value of our proposed objectives, High Correlation (hϕ_2)and Low Correlation (lϕ_2) for the same, is clearly indicated the positive correlation and negative correlation between items. Since the Table VIII has the values for college students (negative C), hence, in this case, the result for High Correlation (hϕ_2)and Low Correlation (lϕ_2) are reverse.

Table IX. A comparison of Interest factor, Correlation, low Correlation and high Correlation objectives.

Table No.	Support	Confidence	Interest Factor	ϕ	$l\phi_2$	$h\phi_2$	$l\phi_3$	$h\phi_3$
Table 2	Tea →Coffee = 15%	Tea → Coffee = 75%	0.9375	-0.063	-1.25	0.25	NA	NA
Table 3	P → Q = 88%	P → Q = 94.62%	1.017	0.232	-0.42857	0.892473	NA	NA
	R → S = 2%	P → Q = 28.57%	4.08	0.232	0.892473	-0.42857	NA	NA
Table 5	HDTV→EM = 33%	HDTV→EM = 55%	1.0784	0.0979	0.0903	0.1085	NA	NA
Table 7	A→B=38.28%	A->B=66.21%	0.9971	-0.005	-0.37354	0.302612	-0.0008	0.002396
Table 8	A→B= 2.27%	A->B = 20%	0.88	-0.023	0.576697	-0.42426		

Simultaneously, the value of High Correlation ($h\phi_3$) and Low Correlation ($l\phi_3$) is **0.002396 and -0.00082** respectively, commutatively for Table VII and Table VIII . Which is clearly shown the positive and negative correlation among items (in this case, for 3 items). The results shows that our proposed objectives give a solution in Simpson's paradox situation.

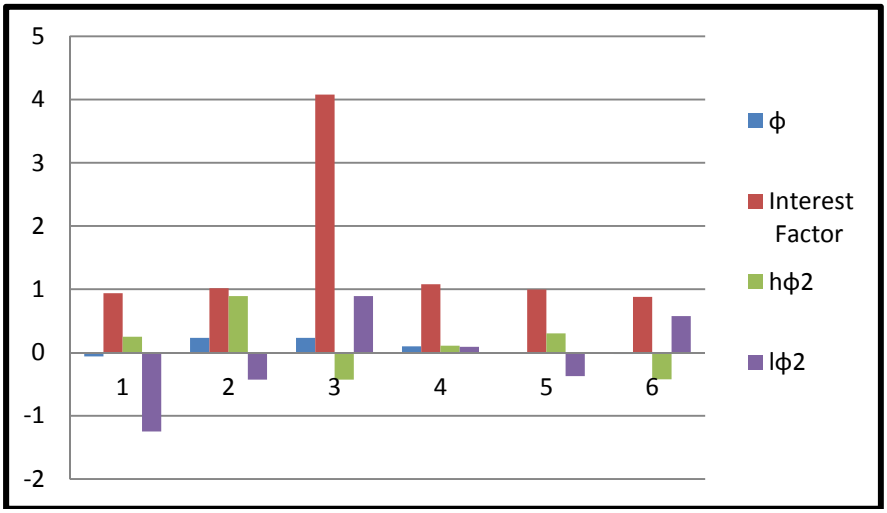

Fig. 1.- A comparison of ϕ, Interest Factor, $l\phi_2$, and $h\phi_2$ for Table -9.

As an example in [11], the contingency tables are given for E1-E10 in Table-X. In Table -XI, the ranking of these tables according to support, confidence, correlation coefficient (ϕ), interest factor(I), newly defined high-correlation ($h\phi_2$) and low-correlation($l\phi_2$) is given.

Table X. Ten examples of Contingency Table (CT)

CT	F11	F10	F01	F00	F1+	F0+	F+1	F+0	N
E1	8123	83	424	1370	8206	1794	8547	1453	20000
E2	8330	2	622	1046	8332	1668	8952	1048	20000
E3	9481	94	127	298	9575	425	9608	392	20000
E4	3954	3080	5	2961	7034	2966	3959	6041	20000
E5	2886	1363	1320	4431	4249	5751	4206	5794	20000
E6	1500	2000	500	6000	3500	6500	2000	8000	20000
E7	4000	2000	1000	3000	6000	4000	5000	5000	20000
E8	4000	2000	2000	2000	6000	4000	6000	4000	20000
E9	1720	7121	5	1154	8841	1159	1725	8275	20000
E10	61	2483	4	7452	2544	7456	65	9935	20000

Table XI. Ranking of Contingency Table using various objectives measure (1 is highest and 10 is the lowest rank).

CT	Support	Confidence	ϕ	IS	$h\phi_2$	$l\phi 2$
E1	3	3	1	6	3	4
E2	2	1	2	8	2	9
E3	1	2	3	10	1	7
E4	6	7	4	4	7	3
E5	7	4	5	3	5	5
E6	9	8	6	2	8	2
E7	4	5	7	5	4	6
E8	5	6	8	9	6	10
E9	8	9	9	7	9	8
E10	10	10	10	1	10	1

Each example is ranked according to its measures in decreasing order of magnitude. It is observe from Table - XI, that high correlation gives higher ranking to the contingency table that have high support value, for example E1,E2,E3 and E4. It is also found that low correlation gives high ranking to the contingency table that have high Interest factor value, for example E10, E6 and E4.

VI. EXPERIMENTAL EVALUATION

In this section we performed experimental analysis. A real-word data set, Retail Dataset is available at Frequent Itemset MIning (FIMI) repository (http://fimi.cs.helsinki.fi/data/). This Retail Dataset, is a large sparse data containing 16,470 distinct items in 88,162 transaction.

In [20], authors demonstrate an experiments to select rare association rule mining. Authors selected 15 contingency table from Retail Dataset. A group of users examined the contingency tables and gave the appropriate ranking. Table -XII show the contingency table chosen from Retail Dataset.

TABLE XII. The contingency table chosen from Retail Dataset.

	F11	F10	F01	F00	F1+	F0+	F+1	F+0
T1	130	74	71	87887	204	87958	201	87961
T2	124	127	56	87855	251	87911	180	87982
T3	106	41	120	87895	147	88015	226	87936
T4	99	175	201	87687	274	87888	300	87862
T5	106	90	138	87828	196	87966	244	87918
T6	5402	3053	36733	42974	8455	79707	42135	46027
T7	224	3673	3033	81232	3897	84265	3257	84905
T8	1740	19	13856	72547	88162	86403	15596	72566
T9	1206	83	40929	45174	87392	86103	42135	45257
T10	1416	40719	1520	44507	88162	46027	2936	85226

Table -XIII present the calculated value of low-correlation values ($l\phi_2$) of the contingency table chosen from Retail Dataset. The last column shows the ranking provided by the users [20]. Contingency Table1(T1) has been given the high rank because it represent an association between two rare variables. Rare association can be treated as lo-correlation ($l\phi_2$) between items. The lo-correlation ($l\phi_2$) rank and the user's rank are similar for top 5 contingency tables except T5. This similarity strength and proved our proposed low-correlation ($l\phi_2$) objectives for association rule mining.

Table XIII. A comparison between $l\phi_2$ rank and User's Rank of Rare Association or Low Correlation between items.

CT	$l\phi_2$ Value	$l\phi_2$ rank	User's Rank
T1	0.998369	1	1
T2	0.998323	2	2
T3	0.997721	3	3
T4	0.995573	5	4
T5	0.997135	4	5
T6	0.103039	9	6
T7	0.924509	6	7
T8	0.741209	7	8
T9	0.068003	10	9
T10	0.686349	8	10

The $l\phi_2$ rank and User's Rank show the rare association or low-correlation between items. It is clear from figure -1, the $l\phi_2$ and User's Rank are similar in top ranking cases, for example in T1,T2 and T3. A little reverse in case of T4 and T5 when the difference between the $l\phi_2$ values is only -0.00156. In case of T6 and T9, $l\phi_2$ rank is high when the value of $l\phi_2$ values is negligible. In case of T7,T8 and T10, the user's rank is high when $l\phi_2$ values decrease. This similarity strength and proved our proposed low-correlation ($l\phi_2$) objectives for association rule mining.

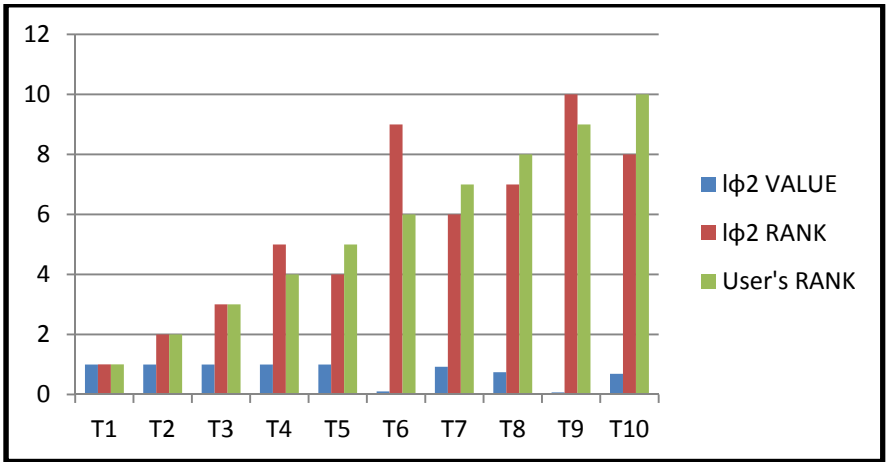

Fig. 2. Comparative Analysis between $l\phi_2$ and User's Rank on the basis of $l\phi_2$ value

VII. CONCLUSION

In this paper we describe the different objectives including support and confidence for association rule mining. We also discuss the limitation of the available objectives. We proposed two new objectives namely high correlation and low correlation for two and three variables and found that over proposed objectives gives better result and clear indication about the positive association and negative association between/among items. The proposed objectives also give the solution in Simpson's paradox situation. As given in experimental evaluation , our proposed objectives are near to the user's rank. This proved that our objectives gives near to accurate results. In future, these objectives can be used as a part of algorithm for generating effective association rules. Simultaneously these objectives can be tested for incremental data. Proposed objectives works for 2 and 3 variables only, that can be generalized for n-variables.

REFERENCES

[1] H. K. Soni, S. Sharma, M. Jain. "Frequent Pattern Generation Algorithms for Association Rule Mining : Strength and Challenges". In proceedings of IEEE International Conference on Electrical, Electronics and Optimization Techniques (ICEEOT), 2016, p. 3744-3747.
[2] M. Kumar, H. K. Soni. "A Comparative Study of Tree- based and Apriori-based Approaches for incremental Data Mining" . International Journal of Engineering Research in Africa, vol. ,23,pp 120-130, 2016.
[3] H. K. Soni, S. Sharma, P.K. Mishra. "Association Rule Mining : A data profiling and prospective approach". International Journal of Current Engineering and Scientific Research. Vol. 3, pp. 57-60, 2016.
[4] H. K. Soni, S. Sharma. "Plausible Characteristics of Association Rule Mining Algorithms for E-Commerce". Submitted to 3rd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB17).
[5] Neelam Mishra, Hemant Kumar Soni, Sanjiv Sharma. "A Comprehensive Survey of Data Mining Techniques on Time Series Data for Rainfall Prediction". Submitted to *Journal of ICT Research and Applications.*
[6] J. Han and M. Kamber. Data mining concepts and techniques. Morgan Kaufmann, 2001.

[7] P. Tan, M. Steinbach, and V. Kumar. Introduction to Data Mining. Addison-Wesley, 2005.
[8] A. Mukhopadhyay, U. Maulik, S. Bandyopadhyay, and C. A. Coello Coello, "Ssurvey of multi-objective evolutionary algorithms for data mining: Part-II", IEEE Transactions on Evolutionary Computation, vol. 18 (1), pp 20-35, 2014.
[9] A. Mukhopadhyay, U. Maulik, S. Bandyopadhyay, and C. A. Coello Coello, "A survey of multi-objective evolutionary algorithms for data mining: Part-I", IEEE Transactions on Evolutionary Computation, vol18 (1), pp1-16. 2014.
[10] P.N. Tan, V Kumar, J. Srivastava. "Selecting the right interestingness measure for association patterns". In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining ,2002, p 32-41.
[11] P.N. Tan, V Kumar, J. Srivastava. "Selecting the right objective measure for association analysis". Information System, vol. 29, pp 293-313, 2004.
[12] P.P. Wakabi-Waiswa, V. Baryamureeba. "Mining High Quality Association Rules Using Genetic Algorithms". In MAICS, 2011, p 73-78.
[13] H R Qodmanan, Mahdi Nasiri, Behrouz Minaei-Bidgoli. "Multi objective association rule mining with genetic algorithm without specifying minimum support and minimum confidence". Expert Systems with Applications , Vol. 38, pp 288–298, 2011.
[14] S. Dehuri, A. K. Jagadev, A. Ghosh and R. Mall. "Multi-objective Genetic Algorithm for Association Rule Mining Using a Homogeneous Dedicated Cluster of Workstations". American Journal of Applied Sciences, vol. 3 (11) pp. 2086-2095, 2006.
[15] Huynh, H. X., Guillet, F., Blanchard, J., Kuntz, P., Gras, R., & Briand, H. "A graphbased clustering approach to evaluate interestingness measures: A tool and a comparative study". Quality measures in data mining. Springer-Verlag. pp 25–50, 2007.
[16] Emna Ben Ayed, Mounir Ben Ayed. "A quality model for the evaluation of decision support systems based on a knowledge discovery from data process". Journal of Decision Systems, vol. 25(2), pp 95-117, 2016.
[17] Davis, L. (Ed.). Handbook of genetic algorithm. New York: Van Nostrand Reinhold. 1991.
[18] G Piatetsky-Shapiro, Discovery, analysis, and presentation of strong rules. In: G. Piatetsky-Shapiro, W. Frawley (eds.) Knowledge Discovery in Databases, AAAI/MIT Press ,1991, p 229–248.
[19] Bo Xiao. "The measures and mining method of credible association rule", IEEE International Conference on Network. 2009, p 322-325.
[20] A Surana, RU Kiran, PK Reddy. "Selecting a Right Interestingness Measure for Rare Association Rules". In Proceedings of the 15th International Conference on Management of Data,2010, p 115-124.

AUTHOR PROFILE

Hemant Kumar Soni received M.Sc. in Computer Science from Jiwaji University, Gwalior, Madhya Pradesh, India in the year 1996 and M. Tech (IT) from Bundelkhand University, Jhansi, Uttar Pradesh, India in the year 2006. He is pursuing Doctoral degree in Computer Science and Engineering from Amity University Madhya Pradesh, Gwalior, India. He has 21 years of teaching experience for UG and PG courses in Computer Science and presently working as offg. Head of the Department of Computer Science and Engineering at Amity University, Gwalior, Madhya Pradesh, India. His research interest in Data Mining and Soft Computing. He published many research papers in National, International Conferences and Scopus Indexed Journals. He is Reviewer of many referred journals. He received a Best Paper Award in an International Conference and organized number of National level events and conferences. He is a member of International Association of Engineers, Hongkong, Universal Association of Computer and Electronics Engineers (**UACEE**), The Institute of Research Engineers and Doctors, USA, Life Member of **ISTE** (Indian Society for Technical Education), India and Member of IAENG Society of Computer Science and Data Mining.