Chandraveer S.Deora, Sana Arora, Zaid Makani / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 3, Issue 2, March -April 2013, pp.208-215

Comparison of Interestingness Measures: Support-Confidence Framework versus Lift-Irule Framework

Chandraveer S.Deora, Sana Arora, Zaid Makani
*B.Tech Computers, 4th year Mukesh Patel School of Technology Management and Engineering JVPD Scheme Bhaktivedanta Swami Marg Vile Parle (W), Mumbai-400056
**B.Tech Computers, 4th year Mukesh Patel School of Technology Management and Engineering JVPD Scheme Bhaktivedanta Swami Marg Vile Parle (W), Mumbai-400056
***B.Tech Computers, 4th year Mukesh Patel School of Technology Management and Engineering JVPD Scheme Bhaktivedanta Swami Marg Vile Parle (W), Mumbai-400056

ABSTRACT

Data Mining is considered to be a step of paramount importance in the process of Knowledge Discovery in Databases. The term "Interestingness Measure" unequivocally forms a very essential aspect of extraction of "interesting" rules from databases. As there are a huge number of association rules or patterns that are generated by most Association Rule Mining Algorithms, there arises a need to prune away the unnecessary and unwanted rules. The rules that are crucial and indispensable can therefore be presented to the end user based on the application of these "Interestingness Measures". The reason this is done is so that the user gets a narrow focus on only those rules that will provide better business understanding and intelligence. However, there are a plethora of measures available today, and selecting the best amongst them requires a thorough research on each of them. This paper therefore provides a comparative study of certain important measures, thereby highlighting which measure is apt for application in which situation.

Keywords – Association Rule Mining, Confidence, Interestingness Measures, Irule, Lift, Support

1. INTRODUCTION

Discovering association rules between items in large databases is a frequent and important task in KDD. The main purpose is to discover hidden relations between the items of the various transactions of the database. This is known as Data Mining.

One of the most striking areas in the process of knowledge discovery is the analysis of the measures of interestingness of discovered patterns. With an increase in the number of discovered association rules, end-users, such as data analysts and decision makers, are recurrently confronted with a major task: Validation and selection the most interesting ones of those rules. This post-processing is done using the concept of Interestingness Measures.

The Apriori algorithm is a simple and well-known association rule algorithm which yields patterns that describe the relationship between the elements of a dataset. Its essentiality lies in the 'Large Item set' property which states that: 'Subsets of a large item set are also large'. The input to this algorithm is the dataset and a certain threshold: Support and Confidence percentage. This uses of the older frameworks of interestingness measures: Support-Confidence framework. It produces large or frequent item sets as its output. It is followed by Association Rule Mining which generates rules pertaining to the dataset. The unnecessary, unwanted rules however need to be pruned to yield more efficient, actionable and beneficial rules.

This is where the role of interestingness frameworks comes into play. It helps decide how interesting a certain rule is and how potent it is in business decision making.

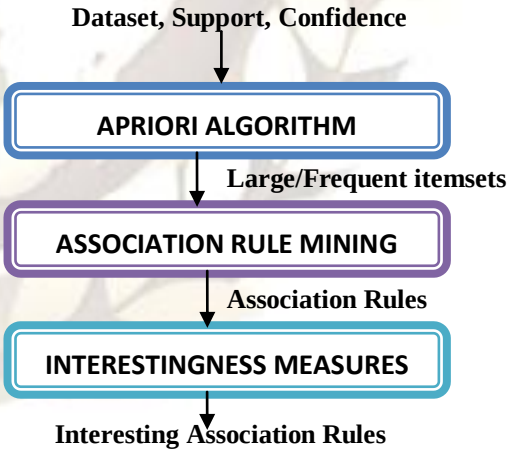

Figure 1: Flowchart of approach followed

Contribution of this paper:

i. Research on the twelve good interestingness measures with their corresponding IM values as per a sample dataset.
ii. Comparative study of all measures based on the % Reduction on a sample dataset.
iii. Application of certain good measures on five sample datasets to obtain results.

iv. Providing a solution as to which measure is preferred to be used in which scenario.

2. LITERATURE SURVEY

One important task for data mining algorithms is to search for patterns that are "surprising" in addition to being precise and coherent. The criterion on the basis of which the search is performed is called "Interestingness measures". It employs certain mathematical formulas using standard set operations. Each Interestingness Measure has its own significance with regards to what it depicts.

2.1. WHY INTERESTINGNESS MEASURES?

Interestingness measures play an important role in data mining. These measures are used to select and rank patterns according to the interest of the user. Consider any Association Rule Mining algorithm, Apriori or FP tree; most of them yield a large number of dependencies between the elements of the dataset. These dependencies or relationships are often termed as Association Rules. Many of these rules produced, provide results or outputs that do not cater well to the business needs. Also, the time taken to process these additional, unnecessary rules proves to be a hindrance to efficient processing. The space occupied by these rules also becomes enormous.

Therefore there arose a need for filtering away all those uninteresting rules that do not fit well into the business scenario. The question was: On what basis should the rules be filtered? Thus, the term "Interestingness Measures" was coined to denote a criterion that could help users identify patterns that were of interest to them.

2.2. NEED AND CRITERIA OF COMPARISON

It started with the 'Support-Confidence' framework. This dealt with two measures: "Support" and "Confidence". However it garnered certain criticisms as mentioned by Jianhua Liu, Xiaoping Fan, Zhihua Qu in their paper. Following this, several new and refined measures came into light, many of which were highly successful in overcoming the problems faced earlier.

But the question now becomes: With a large number of measures available, how do we select the best one? This demands a thorough research on all measures and a variety of datasets.

Another factor that needs to be understood here is: How does the user want to improve his output? Is it in terms of time complexity, space complexity, number of operations taken, number of steps? Here in this paper we have chosen an improvement in the '% Reduction'. % Reduction denotes the percentage of rules discarded. It is denoted by the formula:

% Reduction= (No. of rules rejected / No. of rules on which mining was applied) *100

Having conducted this survey on different sample datasets, a viable solution has been reached which depicts the most efficient measure that provides interesting rules. This is called the Irule.

2.3. RESEARCH SO FAR

Yuejin Zhang, Lingling Zhang, Guangli Nie, Yong Shi have presented a survey of Interestingness Measures for association rules This paper presents a review of the available literature on the various interestingness measures. It draws a comparison of twelve different measures on the basis of certain criteria. It therefore leaves it to the user to choose the best for his/her business application.

Jianhua Liu, Xiaoping Fan, Zhihua Qu have also proposed a new interestingness measure for mining association rules is proposed based on sufficiency measure of uncertain reasoning to improve the classical method of mining association rules.

3. SURVEY ON INTERESTINGNESS MEASURES

3.1. **Irule:** Irule indicates whether the contribution of U (or V) to the occurrence of T increases with V (or U) as a precondition. Therefore, Irule < 1 suggests that $U \cap V \rightarrow T$ is less interesting than $U \rightarrow T$ and $V \rightarrow T$. The value of Irule falls in $[0, +\infty)$. When Irule > 1, the higher Irule is, the more interesting the rule is. Therefore, our new measures are more useful than the traditional confidence and lift. [4]

Formula:

$$I = \frac{\text{Lift}(X \cap X1 \rightarrow Y)}{\text{Lift}(X \rightarrow Y) * \text{Lift}(X1 \rightarrow Y)} \quad (1)$$

Range: $[0, +\infty]$

If I > 1 then More Interesting Rules

If I ≤ 1 then Less Interesting Rules

3.2. **Correlation Coefficient:** Correlation Coefficient is defined as covariance divided by the standard deviation of the association rule. It is actually the measures the degree of linear dependency between a pair of random variable or association rule. It is also known as phi – Coefficient.

Formula:

$$I = \frac{P(X \cap Y) - P(X)P(Y)}{\sqrt{P(X)(1-P(X))} * \sqrt{P(Y)(1-P(Y))}} \quad (2)$$

***Range*:** [-1, +1]

If $I = 0$ then $X \rightarrow Y$ are independent

If $0 < I \leq 1$ then $X \rightarrow Y$ are Highly Dependent

If $0 > I \geq -1$ then $X \rightarrow Y$ are Loosely Dependent

3.3. **Lift:** It is a measure which predicts or classifies the performance of an association rule in order to enhance response. It helps to overcome the disadvantage of confidence by taking baseline frequency in account. [1] [3]

Formula:

$$I = \frac{P(X \cap Y)}{P(X) * P(Y)} \quad (3)$$

Range: [0, ∞]

If 0 < I < 1 then X→Y are negatively interdependent

If I=1 then interdependent

If ∞ > I > 1 then X→Y are positively interdependent

3.4. **Cosine:** Cosine measures the distance between antecedent and consequent of the association rule when they are considered as binary vectors.

Formula:

$$I = \frac{P(X \cap Y)}{\sqrt{P(X)} * \sqrt{P(Y)}} \quad (4)$$

***Range*:** [0, 1]

If I = 0 then No Overlap i.e. transaction contains item X without item Y

If I = 1 then vectors coincide i.e. transaction contains item X with item Y

3.5. **Pavillon:** Pavillon is a measure which takes negative examples (Contra-examples, Counter-example) in order to accept or reject the general trend to have Y when X is present.

Formula:

$$I = \frac{\overline{Y}}{N} - \frac{(X \cap \overline{Y})}{(X)} \quad (5)$$

Range: [0, 1]

If I = 1 then Most Interesting

If I = 0 then Least Interesting

3.6. **Laplace:** Laplace is a measure used in classification. It is a confidence estimator that takes support in order to calculate Laplace. As the value of support decreases value of Laplace also decreases which produce bad result.

Formula:

$$I = \frac{P(Y/X) * (N * P(X \cap Y) + 1)}{(N * P(X \cap Y)) + (2 * P(Y/X))} \quad (6)$$

Range: [0, 1]

If I = 1 then Most Interesting

If I = 0 then Least Interesting

3.7. **Conviction:** Conviction is interpreted as the ratio of the expected frequency that X occurs without Y (Incorrect prediction). It overcomes the weakness of confidence and lift. It attempts to measure the degree of implication of a rule.

Formula:

$$I = \frac{1 - P(X \cap Y)}{1 - P(Y/X)} \quad (7)$$

Range: [0.5, ∞]

If I = 1 then Rule is Independent

If I > 1 then Interesting Rules

If I = infinity then Logical Implication

3.8. **F-Measure:** **F_1 score** (also **F-score** or **F-measure**) is a measure of a test's accuracy. It considers both the precision *p* and the recall *r* of the test to compute the score: *p* is the number of correct results divided by the number of all returned results and *r* is the number of correct results divided by the number of results that should have been returned.

Formula:

Range: [0, 1]

$$I = \frac{2((X) - (X \cap \overline{Y})}{(X) + (Y)} \quad (8)$$

If I = 1 then Most Interesting

If I = 0 then Least Interesting

3.9. **Jaccard:** Jaccard measure the degree of overlap between the antecedent and consequent of the association rule. Jaccard coefficient assesses the distance between antecedent and consequent as the fraction of cases covered by both with respect to the fraction of cases covered by one of them.

Formula:

$$I = \frac{P(X \cap Y)}{P(X) + P(Y) - P(X \cap Y)} \quad (9)$$

Range: [0, 1]

If I = 1 then cover same cases (Most Interesting)

If I = 0 then cover different cases (Least Interesting)

3.10. **Leverage:** Leverage is a measure in which number of counting is obtained from the co-occurrence of the antecedent and consequent of the rule from the expected value.

Formula:

$$I = P(X \cap Y) - P(X) * P(Y) \quad (10)$$

Range: [-0.25,0.25]

If I = 0.25 then Most Interesting

If I = 0.25 then Least Interesting

3.11. **Support:** The Support of an itemset expresses how often the itemset appears in a single transaction in the database i.e. the support of an item is the percentage of transaction in which that items occurs.

Formula:

$$I = P(X \cap Y) = \frac{(X \cap Y)}{N} \quad (11)$$

Range: [0, 1]

If I = 1 then Most Interesting

If I = 0 then Least Interesting

3.12. **Confidence:** Confidence or strength for an association rule is the ratio of the number of transaction that contain both antecedent and

consequent to the number of transaction that contain only antecedent.

Formula:

$$I = P(Y/X) = \frac{P(X \cap Y)}{P(X)} \quad (12)$$

***Range*:** [0, 1]

If I = 1 then Most Interesting

If I = 0 then Least Interesting

Interestingness Measures	Formula
I Rule	$\frac{Lift(X \cap X1 \rightarrow Y)}{Lift(X \rightarrow Y) * Lift(X1 \rightarrow Y)}$
Correlation Coefficient	$\frac{P(X \cap Y) - P(X)P(Y)}{\sqrt{P(X)(1-P(X))} * \sqrt{P(Y)(1-P(Y))}}$
Lift	$\frac{P(X \cap Y)}{P(X) * P(Y)}$
Cosine	$\frac{P(X \cap Y)}{\sqrt{P(X)} * \sqrt{P(Y)}}$
Pavillon	$\frac{\overline{Y}}{N} - \frac{(X \cap \overline{Y})}{(X)}$
Laplace	$\frac{P(Y/X) * (N * P(X \cap Y) + 1)}{(N * P(X \cap Y)) + (2 * P(Y/X))}$
Conviction	$\frac{1 - P(X \cap Y)}{1 - P(Y/X)}$
F-Measure	$\frac{2((X) - (X \cap \overline{Y})}{(X) + (Y)}$
Jaccard	$\frac{P(X \cap Y)}{P(X) + P(Y) - P(X \cap Y)}$
Leverage	$P(X \cap Y) - P(X)*P(Y)$
Support	$I = P(X \cap Y) = \frac{(X \cap Y)}{N}$
Confidence	$I = P(Y/X) = \frac{P(X \cap Y)}{P(X)}$

4. Summary of Survey of Interestingness Measures

Table 1: A Summary of all Interestingness Measures compared

5. Experimentation Performed

5.1. Dataset Used:

The sample dataset is as shown above, a survey conducted in the western region to answer the following questions:

i. How does the color of one's hair affect his/her hobbies?

ii. How does the gender of a person affect his/her hobbies?

iii. Which hobby-store can be set up specifically for males/ females?

Table 2: Sample dataset 1

REGION	HAIR	GENDER	WORK	HEIGHT
West	Brown hair	Female	Stitching	Tall
West	Black hair	Female	Cooking	Tall
West	Black hair	Male	Painting	Medium

5.2. Association Rule Generation

An Apriori algorithm is applied on the above dataset.

Given:

Support= 66.67%

Confidence=50%

The large/frequent itemset obtained is:

{West, Female, Tall}

Association Rule Mining is carried out on the large itemsets generated and the following rules are obtained:

Table 3: Association Rules generated after applying Apriori algorithm on Sample dataset 1

Antecedent	→	Consequent
{West, Female}	→	{Tall}
{West, Tall}	→	{Female}
{ Female, Tall}	→	{West}
{Tall}	→	{West, Female}
{Female}	→	{West, Tall}
{West}	→	{ Female, Tall}

5.3. Application Table for Interestingness Measures

TABLE 4: The comparison of Interestingness values for all measures

Association Rule	I-Rule	Correlation Coefficient	Lift/ Interest	Cosine	Pavillon	Laplace	Conviction	F - Measure	Jaccard	Leverage	Support	Confidence
{West, Female} →{Tall}	1.002	1	1.5	1	0.666667	0.75	-	1	1	0.222222	0.666667	1
{West, Tall}→{Female}	1.002	1	1.5	1	0.666667	0.75	-	1	1	0.222222	0.666667	1
{Female, Tall}→{West}	1	-	1	0.816497	1	0.75	-	0.8	0.666667	0	0.666667	1
{Tall}→ {West, Female}	NA	1	1.5	1	0.666667	0.75	-	1	1	0.222222	0.666667	1
{Female} → {West, Tall}	NA	1	1.5	1	0.666667	0.75	-	1	1	0.222222	0.666667	1
{West} → {Female, Tall}	NA	-	1	0.816497	0	0.6	1	0.8	0.666667	0	0.666667	0.666667

5.4. % Reduction Table for Interestingness Measures

Table 5: % Reduction values for all Interestingness Measures

Interestingness Measures	% Reduction
I Rule	66.67
Correlation Coefficient	33.33
Lift	33.33
Cosine	0
Pavillon	0
Laplace	16.67
Conviction	16.67
F-Measure	0
Jaccard	0
Leverage	33.33
Support	0
Confidence	0

5.5. Support-Confidence Framework v/s Lift-Irule Framework

As observed, Irule gives a high % Reduction in the sample dataset above. The formula that is used to arrive at Irule is Lift, which in itself is a measure. The traditional framework used was the Support-Confidence framework which yields a very poor % reduction i.e. zero.

Therefore, a comparative study was drawn on the four measures of Support, Confidence, Lift and Irule on various sample datasets of different sizes. It was performed on five sample sets, three datasets that were small in size, and two that were large in size.

The standard values of Support and Confidence taken to carry out the comparison : Support=30%, Confidence=30%.

5.6. Sample Datasets

TABLE 2 was considered as the first sample. The other sample datasets were as follows:

Table 6: Sample dataset 2

TID	ITEMS BOUGHT
ID1	Scale, Pencil, Book
ID2	Pen. Pencil, Rubber
ID3	Scale, Pen, Pencil, Rubber
ID4	Pen, Rubber

Table 7: Sample dataset 3

TID	ITEMS BOUGHT
T100	Bread, Butter, Milk, Beer, Sandwich
T200	Bread, Butter, Milk, Curd
T300	Milk, Bread, Jam, Sandwich, Beer
T400	Beer, Jam, Curd, Sandwich

Table 8: Sample dataset 4

ATTRIBUTE	DESCRIPTION
Region	Region of respondent
Marital Status	Marital Status of respondent
Gender	Gender of respondent
Yearly Income	Income of respondent
Children	Number of children
Education	Education Achieved
Occupation	Preferred Occupation
Home Owner	Possess Home or not
Cars	Number of Cars Own
Commute Distance	Distance in Miles
Age	Given in years
Bike Buyer	Possess Bike or not

Table 9: Sample dataset 5

ATTRIBUTE	DESCRIPTION
Department Code	Departments of the Employee
Residential Area	Residential Address of the Employee
Office	Office Location of the Employee
Income	Annual Income
Allowance	Allowance amount
Marital Status	Marital Status of Employee
Children	Number of Children
Car	Number of Cars owned by Employee
Age	Age of the Employee
Project Manager	Assigned as Project Manager or not
Team Leader	Assigned as Team Leader or not

6. Results and Discussion

On comparing the four measures of Support, Confidence, Lift and Irule on the basis of % Reduction, the following results have been obtained.

Table 10: Final comparison of % Reduction between Support, Confidence, Lift and Irule

Data Set		Sample Data Set 1	Sample Data Set 2	Sample Data Set 3	Sample Data Set 4	Sample Data Set 5
Support	**0.2**	0	0	0	0	0
	0.4	100	0	0	85.71	100
	0.6	100	100	100	100	100
	0.8	100	100	100	100	100
Confidence	**0.2**	0	0	0	0	0
	0.4	3.33	0	0	18.36	0
	0.6	21.11	0	0	52.04	44.44
	0.8	21.11	42.85	66.667	76.53	88.88
Lift		6.667	14.28	33.33	41.83	11.11
I -Rule		96.67	57.14	50	93.87	66.67

The following points can be observed from the above table:

i. Users can select their measure of interest as per their business needs with different support-confidence threshold values given.

ii. For smaller datasets, the value of % Reduction obtained fluctuates hence any of the four measures can be used.

iii. For larger datasets, Irule and Support, both give a high % Reduction hence can be viable choices of selection.

iv. A % Reduction of 100 is not suitable, as it indicates an exclusion of all rules, leaving no rules considered, hence the purpose of selection of actionable rules is defeated.

v. Therefore, Irule undoubtedly proves to be the best measure in case of large datasets.

7. Graphical Analysis

The results obtained above are analyzed graphically. The following graphs have been plotted:

i. % Reduction comparison of all Interestingness Measures

ii. Comparison of % Reduction of four measures, Support, Confidence, Lift and Irule for different databases.

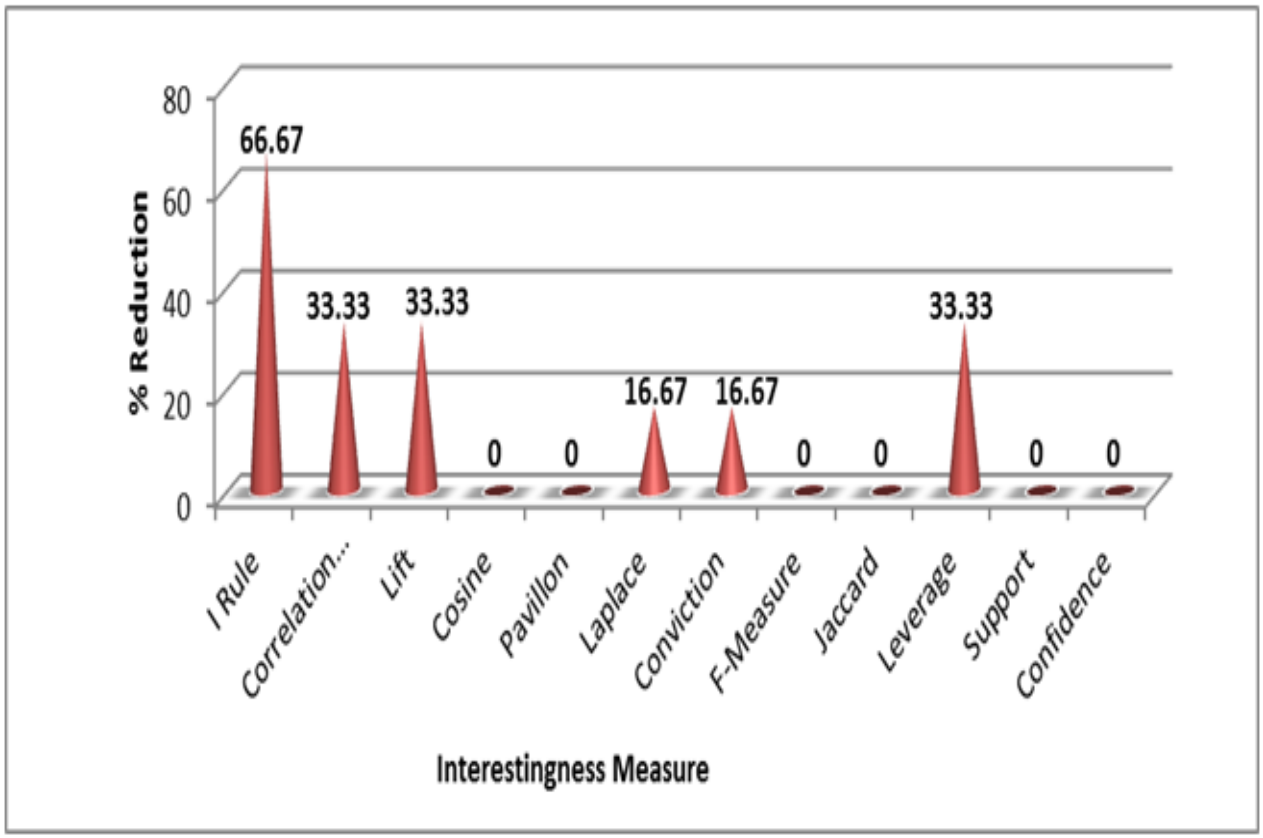

Figure 2: Graph Depicting Interestingness Measures V/S % Reduction for Sample Dataset 1

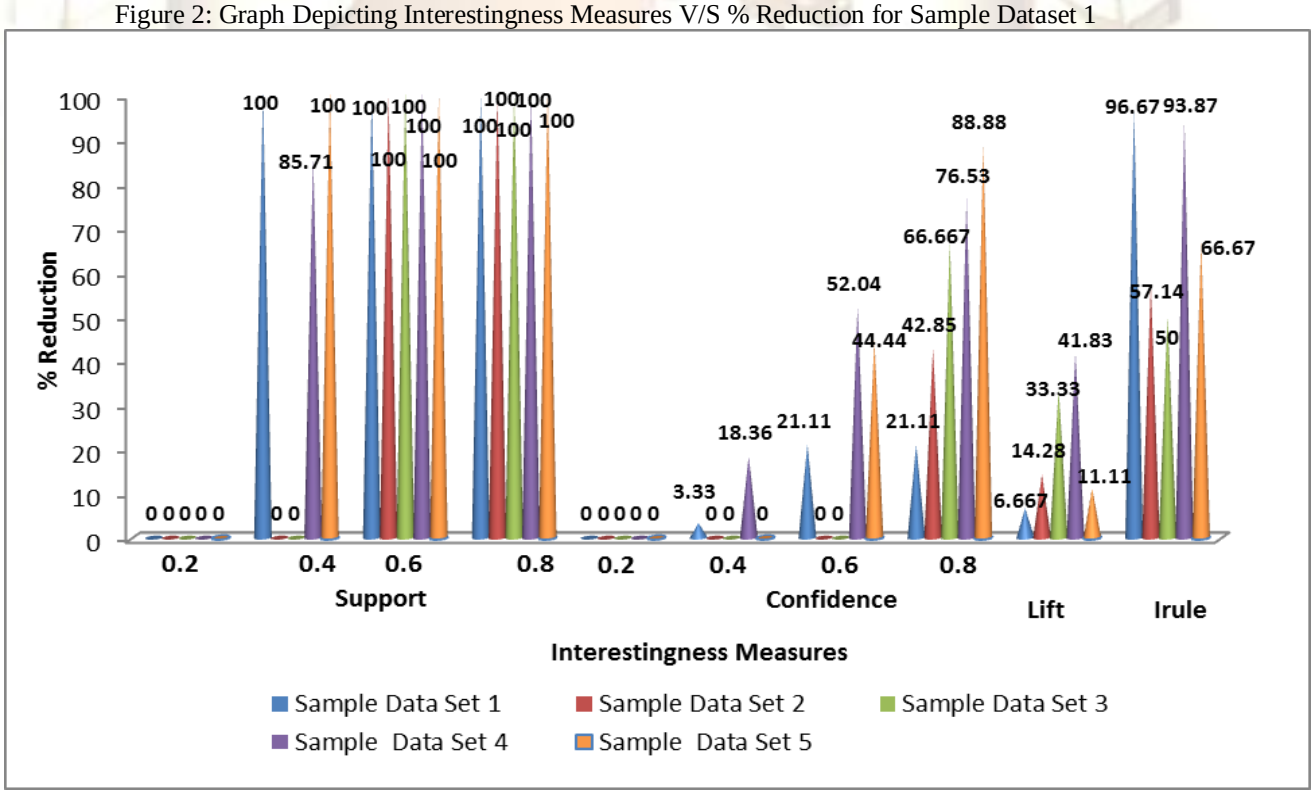

Figure 3: Graph Depicting Interestingness Measures(Support, Confidence, Lift, Irule) v/s % Reduction for all sample datasets

In Fig. 2 we observe that the % Reduction in case of Irule is the highest i.e. 66.66% which means that it gives us more efficient, actionable and interesting rules as compared to all other measures. Next, Lift can be a good evaluation measure as it is used in the formula for Irule. For support and confidence the % Reduction is 0 for the dataset considered. However it has been used for comparison further on because it is the traditional framework and has been used since a very long time.

In Fig. 3 we observe that for all various measures, Irule proves to be the best in yielding a high % Reduction and interesting rules. Support, however though it gives a % Reduction of 100 but it eliminates all rules so the entire purpose of yielding efficient rules is defeated.

CONCLUSION

From the experimentation conducted in this paper, the results obtained help us resolve the problem of choosing a certain Interestingness Measure. The new Lift-Irule framework has been proved to work better than the traditional Support-Confidence framework.

It has been noted that for small datasets as in Table 1, Table 2 and Table 3 we observe that Lift and Irule fluctuate in providing a good % Reduction thus leading to an unwavering selection. However, as the size of the dataset increases, as in more realistic situations, Irule and Support both give a high % Reduction however as Support gives a 100% reduction, it defeats the entire purpose of generating efficient rules. Therefore, Irule proves to be an immensely beneficial measure to obtain a high % Reduction yielding more interesting rules.

ACKNOWLEDGEMENT

We would like to thank our guide, Professor Prashasti Kanikar for all her guidelines and support provided by her while writing this paper. We would also like to thank the Head of the Department, Dr. Dhirendra Mishra whose constant support has led to the successful completion of this paper. We would also like to express our gratitude to the Dean of our college, Dr. S.Y.Mhaiskar for having provided us with the opportunity for writing this paper.

REFERENCES

Journal Papers:

[1] Zaid Makani, Sana Arora and Prashasti Kanikar. Article: A Parallel Approach to Combined Association Rule Mining. *International Journal of Computer Applications* 62(15), 2013, 7-13.

[2] T. Brijis, K. Vanhoof, G. Wets, Defining Interestingness for Association Rules, *International Journal "Information Theories & Applications", Vol.10*, 370 – 375.

[3] Prashasti Kanikar, Dr. Ketan Shah, Extracting Actionable Association Rules from Multiple Datasets, *International Journal of Engineering Research and Applications, Vol. 2, Issue 3*, May-Jun 2012, pp.1295-1300

[4] Prashasti Kanikar and Ketan Shah, An Efficient Approach for Extraction of Actionable Association Rules. *International Journal of Computer Applications* 54(11), 2012, 5-10.

[5] Yuejin Zhang, Lingling Zhang, Guangli Nie, Yong Shi, A Survey of Interestingness Measures for Association Rules, *International Conference on Business Intelligence and Financial Engineering*, 2009, 460 – 463.

[6] Jianhua Liu, Xiaoping Fan, Zhihua Qu, A New Interestingness Measure of Association Rules, *Second International Conference on Genetic and Evolutionary Computing*, 2008, 393 – 397.

[7] Philippe Lenca, Patrick Meyer, Benoit Vaillant, Stephane Lallich, On selecting interestingness measures for association rules: User oriented description and multiple criteria decision aid, *European Journal of Operational Research*, Volume 184, issue 2 (January 16, 2008), 610-626.

Thesis:

[8] Xuan – Hiep Huynh, *Interestingness Measure for Association Rules in a KDD process: Post processing of rules with ARQAT tool*, doctoral diss., University of Nantes, Nantes, 2010.

Proceedings Papers:

[9] Paulo J. Azevedo, Al´ıpio M. Jorge, Comparing Rule Measure for Predictive Association Rules, *Proceeding ECML of the 18th European conference on Machine Learning*, Springer–Verlag Berlin, Heidelberg, 2007, 510- 517.

[10] Pang– Ning Tan, Vipin Kumar, Jaideep Srivastava, Selecting the Right Interestingness Measure for Association Patterns, *Proceeding of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, USA, 2002, pp. 32 – 41.

[11] Merceron, A., and Yacef, K. Interestingness Measures for Association Rules in Educational Data. *Proceedings for the 1st International Conference on Educational Data Mining*, Montreal, Canada, 2008, 57 – 66.