Homogeneous Set of Association Rules

Selvarangam K[1], Ramesh Kumar K[2]

[1] Research scholar, Department of Computer Science and Engineering, Hindustan University, Chennai.

[2] Associate Professor, Department of Information Technology, Hindustan University, Chennai.

***Abstract*-An important and necessary task prior to knowledge mining from a set of association rules (ARs) is the determination of their interest. Since interesting rules only can lead to extract implicit knowledge. Interesting association rules may mined directly by data mining tools on applying interestingness measures while mining, or their interestingness determined by Interestingness measures after mining. However the consistent of the knowledge extracted from this rules with the actual knowledge present is a challenging question to data mining community. In this work, we developed a method to determine the level of knowledge present in a set of association rules by the value of homogeneity coefficient (HC). HC close to 1 leads to a homogeneous set of association rules, that is maximizing the interest, which will improve the consistency of mined knowledge with the actual knowledge present. Conversely HC close to zero indicates the set of rules may further divided into two or more homogeneous set of association rules.**

Keywords-Data mining,Association rule, Interestingness measures, Coefficient of variability, Homogeneity Coefficient.

I. INTRODUCTION

An approximate measure of the right thing is better than the exact measure of the wrong thing. Hence we may assume approximate measure on interesting rule will lead to better knowledge in the process knowledge discovery in data (KDD). Cluster analysis is a class of techniques used to classify objects or cases into relatively homogenous groups called clusters [2], [7], [8]. Objects in each cluster tend to be similar to each other and dissimilar to objects in other clusters. This is an approach to 'let the rules speak for themselves' by means of transactions. Application of clustering techniques might improve the understand ability of mined rules by bringing together 'similar' rules into the same cluster. It may be easier to infer item behaviour from rule clusters than from a rule list. This is because consecutive rules in a rule list may not have any relationship to each other. This can confound the user thus making the interpretation difficult. Clustering differs from grouping, in that there is no preconceived notion of the structure or the number of groups that may exist in the data [2]. The idea here is to look for a 'natural' structure in the data on the basis of which clusters are evolved. Researchers have used clustering and grouping as strategies to improve the understand ability of rules.

An association rule is an implication of the form A→B where $A \subset I$, $B \subset I$, $A \cap B = \emptyset$ and I is the item set. In this paper, we represent given Data set, in terms of association rule, that is the association rule A→B represented as a 2x2 contingency table as shown in the **Table. 1.** For the ease of use we will use the following notation thought in this paper such as number of transactions supporting A and B, by the alphabet 'a', number of transactions supporting A but not B, by the alphabet 'b', number of transactions not supporting A but supporting B by the alphabet 'c', and number of transactions not supporting both A and B by the alphabet'd'. Let N be the total number of transactions on the given data set, sum of a, b, c, and d always equals to N

Table 1.2x2 Contingency Table.

A→B	B	$\overline{B}$	
A	a	b	a+b
$\overline{A}$	c	d	c+d
	a+c	b+d	N

Statistically, variability is defined as the deviation from base point. Variability may calculate by range, mean, variance, deviations and coefficient of variation (CV). In our previous work [12] we ranked ARs by the value of CV. It is the fact that lower the CV leads, less deviation among the variables and higher the CV leads there will be more deviation among the variables. The CV predicts wrong deviation, when the variables having negative values or the mean of the variables become zero. And we know that if we measure temperature by Celsius and Fahrenheit units, the variation between Celsius and Fahrenheit units remains the same. While the Coefficient of Variation [3], defined as s/M, is often used to compare two standard deviations when their means differ substantially, it, too, is inadequate for present purposes: because s is not always smaller than the mean, it is possible for CV to be greater than 1-lack of a natural ceiling which, as in the case of s^2 and s, makes a definitive interpretation of the size of CV impossible. Because of this drawback of CV, we proposed a clustering technique using variability coefficient (VC) and homogeneity coefficient [11]. We assume that quality remains the same as interest in this study.

II. RELATED WORKS

Goktas and Isci [4] reviewed some common measures used to measure the association between two rules; the degree of association will determine the interestingness of ARs. Most of the measures used to determine the quality of association rules are build with mean and variance. Lent et al [10] have introduced the notion of a 'clustered' AR. A clustered AR is a rule that is formed by combining similar, 'adjacent' association rules to form a few general rules. Wang et al [14] maximizes certain interestingness criteria during the merging process. Toivonen et al [13] proposed another approach; Distance between two rules is defined as the number of transactions in which the two rules with the same consequents differ. Gupta et al [5] have proposed a normalized distance function called Conditional market-basket probability (CMPB) distance. This distance function tends to group all those rules that 'cover' the same set of transactions. Gupta et al [5] state "rules involving different items but serving equal purposes were found to be close good neighbors." [5] Thus, their approach is able to capture some amount of customer purchasing behavior. One of the limitations of both the schemes is the arbitrariness of the distance measures used for rule clustering [1]. Moreover, they do not develop any framework to concisely describe the generated rule clusters.

When it comes to quality of an association rule, how the quality of a rule is measured, to determine if it is useful, interesting, important etc. But there is no formal definition of quality and/or interestingness [7]. Currently there is a collection of different measures available which is partly due to the traditional methods of support and confidence being considered insufficient [10]. Most of the quality measures defined in terms of mean and variance are not able find the actual degree of association due to the impossibility of CV [12], when variance is greater than mean. The degree of homogeneity is a group exhibits on some measure and the difference in homogeneity is the group exhibits across two or more measures. These issues assume particular relevance when the interest lies in deciding whether to subdivide the set of ARs on the basis of the information at hand.

III. INTERESTINGNESS MEASURES

Hiep (2010) stated that patterns are transformed into value by the interestingness measures. Jeyachidra and Punithavalli (2014) developed their feature selection algorithm DWFS-CK by using the interesting measure Gini Index. The interestingness of a measure depends on both data structure and on the decision maker’s goal. Mcgarry (2005) classified these measures as objective and subjective in nature. Coverage, support, accuracy are criterias of objective and unexpectedness, actionable, novel are criteria under subjective nature. Liquing and Hamilton (2006) added semantic as additional nature. Also they extended criteria with conciseness, reliability, peculiarity, diversity and utility. Defining the Interestingness of a measure is complex, but we may define the interestingness of measure by the above stated criteria. Some measures may be relevant with some context but not with others. Hence the ranking may be different on different data sets.

A. Support

The rule A→B has support S in a transaction set S% of transaction contains $A \cup B$. That is fraction between number of transactions supporting A and B (a) and to the total number of transactions (N) will be the support and calculated by the **Eq. 1**. This is a basic measure used in the data mining literature to express the generality of an association rule

$$Support\ S = P(A \cup B) = \frac{a}{N} (1)$$

B. Confidence

The rule A→B has confidence C if C% transactions Contain A also contain B. That is, confidence value given by the conditional probability.

$$P\left({}^{B}/_{A}\right) = \frac{P(A \cup B)}{P(A)} = \frac{a}{a+b} \qquad (2)$$

C. Lift

Lift is also a basic measure in the data mining literature used to express the reliability of an association rule, which is defined by the Eq. 3. If there is no association between the variables A and B then the value of life become one.

$$Lift = \frac{P(A \cup B)}{P(A)P(B)} = \frac{Na}{(a+b)(a+c)} \qquad (3)$$

D. U Cost, S Cost and R Cost

Pecina and Schlesinger (2006) listed U Cost, S Cost and R Cost are heuristic association measures, which is used to find the association between bigrams (between two variables). These measures are defined by the following equations:

$$\text{U Cost} = \log\left(1 + \frac{\min(b,c) + a}{\max(b,c) + a}\right) (4)$$

$$\text{S Cost} = \log\left(1 + \frac{\min(b,c)}{a+1}\right)^{-\frac{1}{2}} (5)$$

$$\text{R Cost} = \log\left(1 + \frac{a}{a+b}\right) x \log\left(1 + \frac{a}{a+c}\right) \qquad (6)$$

E. T Combined Cost

This measure also a heuristic association measures used in many researches, listed by Pecina and Schlesinger [13], defined by **Eq. 7**.

$$T\ Combined\ Cost = \sqrt{UxSxR} \qquad (7)$$

Table 2. List of IS Measures

S.No	Measure	Formula
1	Support	$\frac{a}{N}$
2	Confidence	$\frac{a}{a+b}$
3	Lift	$\frac{Na}{(a+b)(a+c)}$
4	U Cost	$log\left(1 + \frac{min(b,c) + a}{max(b,c) + a}\right)$
5	S Cost	$log\left(1 + \frac{min(b,c)}{a+1}\right)^{-1/2}$
6	R Cost	$log\left(1 + \frac{a}{a+b}\right) xlog\left(1 + \frac{a}{a+c}\right)$
7	T Combined Cost	$\sqrt{UxSxR}$

IV. MATERIALS AND METHODS

Interestingness of set association rules will be calculated in this work by variability coefficient (VC) or by coefficient of homogeneity (HC). By fixing the threshold on VC or HC, we will cluster the association rule and make decision on the necessity of further division.

A. Variability coefficient

Variability consists of the differences in magnitude that exist in a set of occurrences of some measure. If at least one occurrence differs in magnitude from the others, the set of rules exhibits variability; if no difference occurs, then the set of rule does not exhibit variability. When only one occurrence differs in size from the others, the set exhibits minimum variability; and the greater the total difference in magnitude among the occurrences, the greater the variability exhibited by the set of rules. If variability is seen in this light, then its measure can be formulated as the sum of the observed differences among occurrences of a measure divided by the maximum possible sum of the differences. This is known as variability coefficient and express by the equation 8.

$$Variability\ Coefficient\ VC = \frac{OV}{MPV} \qquad (8)$$

Where: OV = Observed variation, MPV = Maximum possible variation

The value of VC always lies between 0 and 1. Since there is no variation in set of rule scores the OV become zero hence it is clear that VC become zero(by equation 8) In case of maximum variation among the rules scores, OV is equal to MPV hence VC become 1 in this case.

The observed variability (OV) is the sum of the absolute differences among occurrences of the measure at hand. A matrix arrangement of the differences among a group of scores is helpful in visualizing the calculations used to derive OV. Statistically it is the fact that, the maximum sum of differences in a set of scores will occur if half the scores have the lowest value contained in the set and the other half carry the highest value. For a comparison matrix of a data set half of which consists of one uniform value and half of which consists of a different uniform value, only comparisons of the two different values will yield nonzero remainders.

The derivation of MPV in (8) is based on the following reasoning: the maximum sum of differences in a set of scores will occur if half the scores have the lowest value contained in the set and the other half carry the highest value. For an even number of cases, the number of such comparisons is the number of scores in the group's lower half multiplied by the number of scores in the group's upper half, that is $\left(\frac{N}{2}\right)\left(\frac{N}{2}\right)$ and thus, the number of non-zero comparisons will equal the square of half the cases in the data set that is, $\left(\frac{N}{2}\right)^2$. The highest possible variability will consist of the product of this square and the sum of the comparisons of the two values. Thus, for a group of scores consisting of an even number of cases, MPV can be calculated as follows equation 9:

$$MPV = \left(\frac{N}{2}\right)^2 R \qquad (9)$$

Where, N = group size and R = the range, that is, the difference between the highest and lowest scores.

For a group of scores consisting of an odd number of cases, MPV can be calculated by equation 10:

$$MPV = \left(\frac{N-1}{2}\right)\left(\frac{N+1}{2}\right) R \qquad (10)$$

B. Homogeneity coefficient

A coefficient of homogeneity (HC) can be defined as the complement of VC; hence it is calculated by equation11:

$$HC = 1 - VC \qquad (11)$$

Since the VC value lies between 0 and 1, the HC value also lies between 0 and 1.

C. Calculation of HC and VC

Let us consider a relational data base R, and a set of association rules R$_1$, R$_2$, R$_3$,…, R$_n$ on R with rule score $x_1, x_2, x_3, \ldots, x_n$ $r$espectively. OV in equation 1 is the sum of absolute differences among the occurrence of the rules which is calculated by the equation 5. A matrix arrangement of the differences among a group of scores is helpful in visualizing the calculations used to derive OV. For the set of rules, the matrix is displayed in Table 3. The scores in Table 1appear vertically along the table's left as well as horizontally along its top. For each row, the cells represent the difference between the score on the left column and the other scores in the set. Each score on the horizontal list is subtracted from each of the scores on the vertical list and the remainder for each subtraction is recorded as an absolute value in the intersecting cell. If no difference emerges, a 0 is recorded.

$$OV = \sum \left|x_i - x_j\right| \ for\ all\ i,j \qquad (12)$$

The derivation of MPV in (8) is based on the following reasoning: the maximum sum of differences in a set of scores will occur if half the scores have the lowest value contained in the set and the other half carry the highest value. Let the least and highest score in table.3 be named $y_i\ and\ y_j$ respectively. The MPV calculation is represented in table.4.

D. Example calculation

Let us consider a relational data base R, and a set of association rules R$_1$, R$_2$, R$_3$, …, R$_{10}$ on R with scores 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, respectively. A matrix arrangement of the differences among rule scores is helpful in visualizing the calculations used to derive OV. For the above set of rules, the matrix is displayed in Table 5. The OV value is calculated by equation 12 and its value for the above set of rules is 820. The derivation of MPV for the above set of rules is displayed in table 5. The highest variation will occur if the data take the following values: 35, 35, 35, 35, 35, 75, 75, 75, 75, 75 and the MPV = 1000 (by equation. 2) by applying OV and MPV value in equation 8.

$$VC = \frac{820}{1000} = 0.82$$

And the HC value is given by equation 11.

$$HC = 1 - VC = 1 - 0.82 = 0.28$$

Table 3. Matrix arrangement of differences in rule scores

Score	$\boldsymbol{x_1}$	$\boldsymbol{x_2}$	$\boldsymbol{x_3}$	…	$\boldsymbol{x_i}$.	…	$\boldsymbol{x_n}$
$\boldsymbol{x_1}$	0	$\lvert x_1 - x_2\rvert$	$\lvert x_1 - x_3\rvert$		$\lvert x_1 - x_i\rvert$		$\lvert x_1 - x_n\rvert$
$\boldsymbol{x_2}$		0	$\lvert x_2 - x_3\rvert$		$\lvert x_2 - x_i\rvert$		$\lvert x_2 - x_n\rvert$
$\boldsymbol{x_3}$			0		$\lvert x_3 - x_i\rvert$		$\lvert x_3 - x_n\rvert$
.				.			.
$\boldsymbol{x_i}$					0		$\lvert x_i - x_n\rvert$
.						.	.
$\boldsymbol{x_n}$							0

Table 4. Matrix arrangement for MPV calculation

Scores	$\boldsymbol{y_i}$	$\boldsymbol{y_i}$	$\boldsymbol{y_i}$	…	$\boldsymbol{y_j}$.	…	$\boldsymbol{y_j}$
$\boldsymbol{y_i}$	0	0	0		$\lvert y_i - y_j\rvert$		$\lvert y_i - y_j\rvert$
$\boldsymbol{y_i}$		0	0		$\lvert y_i - y_j\rvert$		$\lvert y_i - y_j\rvert$
$\boldsymbol{y_i}$			0		$\lvert y_i - y_j\rvert$		$\lvert y_i - y_j\rvert$
.				.			.
$\boldsymbol{y_j}$					0		$\lvert y_i - y_j\rvert$
.						.	.
$\boldsymbol{y_j}$							0

According to the user knowledge expectation the set of rules generated from the relational data base R using data mining tools. For this set of rules VC and HC value calculated as above, based on the values of VC and HC we may concludethe interesting set of rules.

Table 5. Matrix arrangement of differences in a group scores

	30	35	40	45	50	55	60	65	70	75
30	0	5	10	15	20	25	30	35	40	45
35		0	5	10	15	20	25	30	35	40
40			0	5	10	15	20	25	30	35
45				0	5	10	15	20	25	30
50					0	5	10	15	20	25
55						0	5	10	15	20
60							0	5	10	15
65								0	5	10
70									0	5
75										0

V. IMPLEMENTAION

We generated a random set of association rule (shown in table .5) using IBM Quest data set generator. Coefficient variation (CV) [12] of the measures listed in table.6 is calculated for the set of association rules in table .5 and presented in Table.7.

Table 6. Matrix arrangement for MPV calculation

	30	30	30	30	30	75	75	75	75	75
30	0	0	0	0	0	45	45	45	45	45
30		0	0	0	0	45	45	45	45	45
30			0	0	0	45	45	45	45	45
30				0	0	45	45	45	45	45
30					0	45	45	45	45	45
75						0	0	0	0	0
75							0	0	0	0
75								0	0	0
75									0	0
75										0

The VC and HC values of the measures for the set of association measures in table .7 is calculated as shown in the example above and presented in table .8. The graph presented in figure.1 shows the deviation of the CV and VC values, and it is clear that VC and HC are complement to each other.

VI. CONCLUSION AND FUTURE WORK

In this work, we have presented a method by determining the variability coefficient, the value of VC is not depending on mean and variance, and hence the drawback on coefficient of variation will be eliminated. VC close to 1 means the set of rules exhibit more variations, and the rules produces more knowledge and they do not consistent with actual knowledge due to the over whelming. This supports the Geng and Hamilton [9] conclusion presented on their survey. Less homogeneity set of rules may divide further to make homogeneous set of rules. This work directs, when interest lies further subdividing of data in hand possibilities. Implementing on big data sets by the way of algorithm may be the extension of this work.

Table 7. Set of association rules

Rule	a	b	c	d
R_1	9624	86	64	226
R_2	8432	186	146	1236
R_3	7218	823	674	1285
R_4	6854	326	258	2562
R_5	543	318	9015	124
R_6	486	1281	7842	391
R_7	1343	2861	5236	560
R_8	883	3786	4631	700
R_9	3285	137	232	6346
R_{10}	3015	400	643	5942

Table 8.VC and HC values of association rules

	1	2	3	4	5	6	7
VC	81.39	84.75	69.9	66.97	78.6	91.74	88.58
HC	18.61	15.25	30.1	33.03	21.4	8.26	11.42

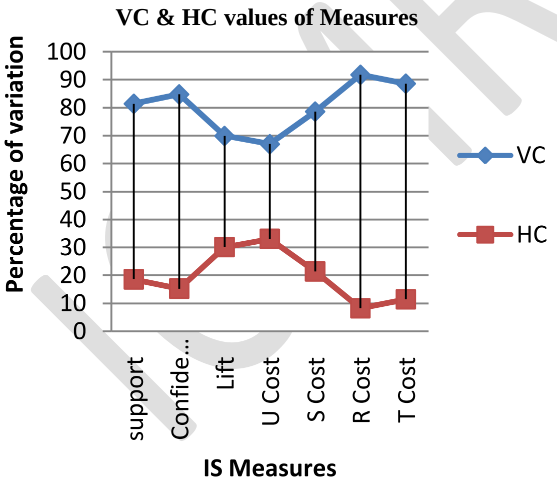

Figure 1. VC & HC values of IS Measure

REFERENCES

[1] AdomaviciusG,TuzhilinA,., (2001). Expert-driven validation of rule-based user models in personalization applications. Data Mining Knowledge Discovery 5(1/2): 33–58.

[2] Anderberg M R.,(1973).Cluster analysis for applications : New York: Academic Press (1973)

[3] Croxton, F.E., D.J. Crowden and S. Klein., (1967). Applied General Statistics. 3rd Edn., Prentice-Hall, New York, 754.

[4] Goktas, A. and Ö. Isci., (2011). A comparison of the most commonly used measures of association for doubly ordered square contingency tables via simulation. MetodološkiZvezki, 8: 17-37.

[5] Gupta G K, Strehl A, Ghosh J., (1999). Distance based clustering of association rules. Proc. Intelligent Engineering Systems through Artificial Neural Networks (ANNIE 1999) (St. Louis, MO: ASME Press) vol. 9, 759–76.

[6] Han, J. and M. Kamber, Data Mining., (2006). Concepts and Techniques. 2^{nd} Ed. Elsevier Inc. ISBN: 10: 1-55860-901-6, pp: 261-272.

[7] Jain A K, Murty M N, Flynn P J., (1999) Data clustering: A review.: ACM Comput. Surv. Vol. 31(3), 264–323.

[8] Kaufman L, Rousseeuw P J., (1990). Finding groups in data: An introduction to cluster analysis. New York: Wiley.

[9] Liquing, G. and H.J. Hamilton.,(2006). Interestingness measures for data mining: A survey.: ACM Comput. Surveys, vol. 38.

[10] Lent, B., Swami, A. N., and Widom, J., (1997). Clustering association rules:. In ICDE. 220-231.

[11] Martinez-Pons, M., (2013). Coefficient of variation. J. Mathem. Statistics, vol. 9, 62-64.

[12] Selvarangam K and Ramesh Kumar K., (2014). Selecting perfect interestingness measures by coefficient of variation based ranking algorithm,: J. Comput., vol. 10 , 1672 – 1679.

[13] Toivonen H, Klemettinen M, Ronkainen P, Hatonen K, Mannila H., (1999). Pruning and grouping discovered association rules.: Proc. Mlnet Workshop on Statistics, Machine Learning and Knowledge Discovery in Databases, Herakhion, Crete Greece.

[14] Wang K, Tay Soon H W, Liu B., (1998). Interestingness-based interval merger for numeric association rules. Proc. 4th Int. Conf. on Data Mining and Knowledge Discovery (KDD 98) New York: AAAI Press. 121–128.