VNU Journal of Science, Natural Sciences and Technology 24 (2008) 122-132

Ranking objective interestingness measures with sensitivity values

Hiep Xuan Huynh[1,*], Fabrice Guillet[2], Thang Quyet Le[1], Henri Briand[2]

[1] *College of Information and Communication Technology, Cantho University, Vietnam, No 1 Ly Tu Trong Street, An Phu Ward, Ninh Kieu District, Can Tho, Vietnam*

[2] *Polytechnic School of Nantes University, France.*

Received 31 October 2007

Abstract. In this paper, we propose a new approach to evaluate the behavior of objective interestingness measures on association rules. The objective interestingness measures are ranked according to the most significant interestingness interval calculated from an inversely cumulative distribution. The sensitivity values are determined by this interval in observing the rules having the highest interestingness values. The results will help the user (a data analyst) to have an insight view on the behaviors of objective interestingness measures and as a final purpose, to select the hidden knowledge in a rule set or a set of rule sets represented in the form of the most interesting rules.

Keywords: Knowledge Discovery from Databases (KDD), association rules, sensitivity value, objective interestingness measures, interestingness interval.

1. Introduction

Postprocessing of association rules is an important task in the Knowledge Discovery from Databases (KDD) process [1]. The enormous number of rules discovered in the mining task requires not only an efficient postprocessing task but also an adapted results with the user's preferences [2-7]. One of the most interesting and difficult approach to reduce the number of rules is to construct interestingness measures [8,7]. Based on the data distribution, the objective interestingness measures can evaluate a rule via its statistical factors. Depending on the user's point of view, each objective interestingness measures reflects his/her own interests on the data. Knowing that an interestingness measure has its own ranking on the discovered rules, the most important rules will have the highest ranks. As we known, it is difficult to have a common ranking on a set of association rules for all the objective interestingness measures.

In this paper we proposed a new approach for ranking objective interestingness measures using observations on the intervals of the distribution of interestingness values and the number of association rules having the highest interestingness values. We focused on the most significance interval in the inversely cumulative distribution calculated from each objective interestingness measures. The sensitivity evaluation is experimented on a rule set and on

* Corresponding author. E-mail: hxhiep@cit.ctu.edu.vn

a set of rule sets to rank the objective interestingness measures. The objective interestingness measures with the highest ranks will be chosen to find the most interesting rules from a rule set. The results will help the user to evaluate the quality of association rules and to select the most interesting rules as the useful knowledge. The results obtained are drawn from the ARQAT tool [9].

This paper is organized as follows. Section 2 introduces the post-processing stage in a KDD process with interestingness measures. Section 3 gives some evaluations based on the cardinalities of the rules as well as rule's interestingness distributions. Section 4 presents a new approach with sensitivity values calculated from the most interesting bins (a bin is considered as an interestingness interval) of an interestingness distribution in comparison with the number of best rules. Section 5 analyzes some results obtained from sensitivity evaluations. Finally, section 6 gives a summarization of the paper.

2. Postprocessintg of association rules

How to evaluate the quality of patterns (e.g., association rules, classification rules,...) issued from the mining task in the KDD process is often considered as a difficult and an important problem [6,7,10,1,3]. This work is lead to the validation of the discovered patterns to find the interesting patterns or hidden knowledge among the large amount of discovered patterns. So that, a postprocessing task is necessary to help the user to select a reduced number of interesting patterns [1].

2.1. Association rules

Association rule [2,4], taking an important role in KDD, is one of the discovered patterns issued from the mining task to represent the discovered knowledge. An association rule is modeled as $X_1 \wedge X_2 \wedge ... \wedge X_k \rightarrow Y_1 \wedge Y_2 \wedge ... \wedge Y_l$. Both of the two parts of an association rule (i.e., the antecedent and the consequence) are composed with many items (i.e., a set of items or *itemset*). An association rule can be described shortly as $X \rightarrow Y$ where $X \cap Y = \varnothing$.

2.2. Post-processing with interestingness measures

The notion of interestingness is introduced to evaluate the patterns discovered from the mining task [5,7,8,11-15]. The patterns are transformed in value by interestingness measures. The interestingness value of a pattern can be determined explicitly or implicitly in a knowledge discovery system. The patterns may have different ranks because their ranks depend strongly on the choice of interestingness measures. The interestingness measures are classified into two categories [7]: subjective measures and objective measures. Subjective measures explicitly depend on the user's goals and his/her knowledge or beliefs [7,16,17]. They are combined with specific supervised algorithms in order to compare the extracted rules with the user's expectations [7]. Consequently, subjective measures allow the capture of rule novelty and unexpectedness in relation to the user's knowledge or beliefs. Objective measures are numerical indexes that only rely on the data distribution [10,18-21,8]. Interestingness refers to the degree to which a discovered pattern is of interest to the user and is driven by factors such as novelty, utility, relevance and statistical significance [6,8]. Particularly, most of the interestingness measures proposed in the literature can be used for association rules [5,12,17-25]. To restrict the research area in this paper, we will working on objective interestingness measures only. So we can use the words objective interestingness measures, objective measures and interestingness measures interchangeably (see *Appendix* for a complete list of 40 objective interestingness measures).

3. Interestingness distribution

3.1. Interestingness calculation

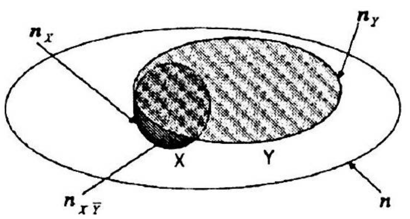


Fig. 1. Cardinalities of an association rule $X \rightarrow Y$.

Fig. 1 shows the 4 cardinalities of an association rule $X \rightarrow Y$ illustrated in a Venn diagram. Each rule set with its list of 4 cardinalities $n, n_X, n_Y, n_{X\bar{Y}}$ is then calculated by an objective measure respectively. The value obtained is called an interestingness value and stored in an *interestingness set*. The interestingness set is then sorted to have a *rank set*. The elements in the rank set is ranked due to its corresponding interestingness values. The higher the interestingness value the higher the rank obtained

For example, if the measure *Laplace* (see Appendix) has the formula $\frac{n_x + 1 - n_{x\bar{y}}}{n_x + 2}$ with $n_X = 120$ and $n_{X\bar{Y}} = 45$, so we can compute the interestingness value of this measure by:

$$v_i(Laplace) = \frac{n_X + 1 - n_{X\bar{Y}}}{n_X + 2} = \frac{120 + 1 - 45}{120 + 2} = \frac{76}{122} = 0.623$$

The other two necessary sets are also created. The first set is an *order set*. Each element of the order set is an order mapping *f*: 1 → 1 for each element in the corresponding interestingness set. The *value set* contains the list of interestingness values correspond to the position of the elements in the rank set (i.e. mapping *f*: 1 → 1).

For example, with 40 objective measures, one can obtain 40 interestingness sets, 40 order sets, 40 rank sets and 40 value sets respectively (see Fig. 2). Each data set type is saved in a corresponding folder. For instance, all the interestingness sets are stocked in an folder with the name INTERESTINGNESS. The other three folder names are ORDER, RANK and VALUE.

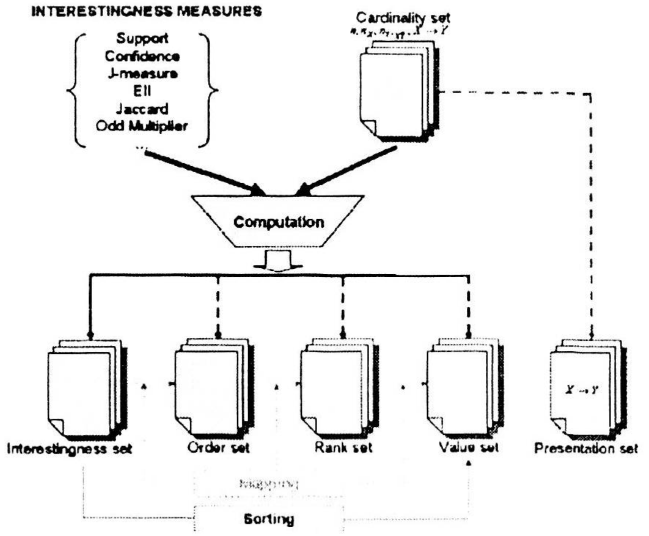


Fig. 2. The interestingness calculation module.

3.2. Distribution of interestingness values

The distribution of each measure can be very useful to the users. From this information the user can have a quick evaluation on the rule set. Some significant statistical characteristics about *minimum value, maximum value, average value, standard deviation value, skewness value, kurtosis value* are computed (see table 1). The shape information of the last two arguments are also determined. In addition, the histograms like frequency and inversely cumulative are also drawn (Fig. 3, Fig. 4 and table 2). The images are drawn with the support of the JFreeChart package [26]. We have added to this package the visualization of the inversely cumulative histogram. Table II illustrates an example of interestingness distribution from a rule set with 10 bins.

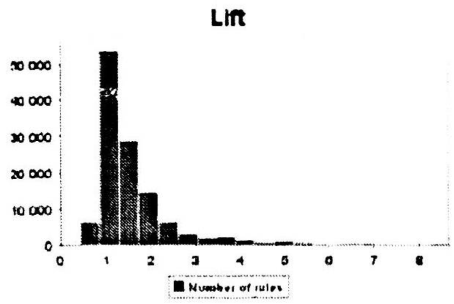


Fig. 3. Frequency histogram of the Lift measure from a rule set.

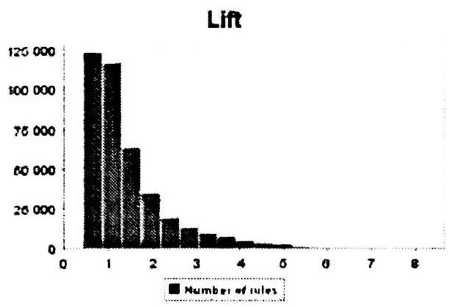


Fig. 4. Inversely cumulative histogram of the Lift measure from a rule set.

Assume that $\mathcal{R}$ is the set of p association rules, called a rule set. Each association rule r_i (i = 1..p) has an interestingness value v_i computed from a measure m.

Table 1. Some statistical indicators on a measure

Statistical significance	Symbol	Formula
Min	*min*	$\min(v_i)$
Max	*max*	$\max(v_i)$
Mean	*mean*	$\sum_{i=1}^{p} v_i$
Variance	*var*	$\frac{\sum_{i=1}^{p}(v_i - mean)^2}{p-1}$
Standard deviation	*std*	$\sqrt{var}$
Skewness	*skewness*	$\frac{\sum_{i=1}^{p}(v_i - mean)^3}{(p-1)\times std}$
Kurtosis	*kurtosis*	$\frac{\sum_{i=1}^{p}(v_i - mean)^4}{(p-1)\times var^2} - 3$

Table 2. Frequency and inversely cumulative bins

	Bins				
Histogram	1	2	3	4	5
Frequency	7	1	12	9	20
Relative frequency	0.031	0.004	0.053	0.040	0.880
Cumulative	7	8	20	29	49
Inversely cumulative	225	218	217	205	196

	Bins				
Histogram	6	7	8	9	10
Frequency	30	70	9	2	65
Relative frequency	0.133	0.311	0.040	0.008	0.288
Cumulative	79	149	158	160	225
Inversely cumulative	176	146	76	67	65

3.3. *Inversely cumulative histogram of interestingness values*

Interestingness histogram. An interestingness histogram is a histogram [27] in which the size of a category (i.e., a bin) is the number of rules having the same interval of interestingness values.

Suppose that the number of rules that fall into an interestingness interval i is h_i, the total number of bins are k, and the total number of rules is p. So the following constraint must be satisfied:

$$p = \sum_{i=1}^{k} h_i$$

Interestingness cumulative histogram. An interestingness cumulative histogram is a cumulative histogram [27] in which the size of a bin is the cumulative number of rules from the smaller bins up to the specified bin. The cumulative number of rules c_i in a bin i is determined as:

$$c_i = \sum_{j=1}^{i} h_j$$

For our purpose, we take the inversely cumulative distribution representation in order to show the number of rules that have been ranked higher than an eventually specified minimum threshold. Intuitively, the user can see exactly the number of rules that he will have to deal with in the case in which he/she will choose a particular value for the minimum threshold. The inversely cumulative number of rules ic_i can be computed as:

$$ic_i = \sum_{j=k}^{i} h_j$$

The number of bins k are directly dependent of the rule set size p. It is generated by the following Sturges formula [27]:

$$k = \frac{\max(v_i) - \min(v_i)}{Sturges's\ formula}$$

with:

i) $Sturges\ Formula = 1 + 3.3\log(p)$,

ii) $\max(v_i)$ and $\min(v_i)$ are the maximum interestingness value and minimum interestingness value respectively,

(iii) an interestingness value is represented by the symbol v_i.

4. Sensitivity values

4.1. Rule set characteristics

Before evaluating the sensitivity of the interestingness measures observed from interestingness distribution, we propose some arguments on rule set to give the user a quick observation on the characteristics of a rule set.

Each characteristic type is determined by a string representing its equation respectively. The purpose is to show the distributions underlying rule cardinalities, in order to detect "borderline cases". For instance, table 3 gives 16 necessary characteristic types in our study in which the first line gives the number of "logical" rules (i.e. rules without negative examples). The percentage of each characteristic type in the rule set is also computed.

Table 3. Characteristic types (remind that $n_{XY} = n_X - n_{X\bar{Y}}$)

N°	Type
1	$(n_{X\bar{Y}} = 0)$
2	$(n_X = n_{XY}) \wedge (n_Y \neq n_{XY}) \wedge (n \neq n_Y)$
3	$(n_Y = n_{XY}) \wedge (n_X = n_{XY}) \wedge (n \neq n_X)$
4	$(n_X = n_{XY}) \wedge (n_Y = n_{XY}) \wedge (n \neq n_X)$
5	$(n_X = n) \wedge (n_Y \neq n)$
6	$(n_Y = n) \wedge (n_X \neq n)$
7	$(n_X = n) \wedge (n_Y = n)$
8	$(n_X < n_Y)$
9	$(n_X < \frac{n_Y}{2})$
10	$(n_X < \frac{n_Y}{4})$
11	$(n_X < \frac{n_Y}{6})$
12	$(n_X < \frac{n_Y}{8})$
13	$(n_X < \frac{n_Y}{10})$
14	$(n_X = n_Y)$
15	$(n_{X\bar{Y}} = \frac{n_X}{2})$
16	$(n_{X\bar{Y}} = \frac{n_X \times n_Y}{2})$

Initially, the counter of each characteristic type is set to zero. Each rule in the rule set is then examined by its cardinalities to match the characteristic types. The complexity of the algorithm is linear $\mathcal{O}(p)$.

4.2. Sensitivity

The sensitivity of an interestingness measure is referred at the number of best rules (i.e., rules that have the highest interestingness values) that an interested user should have to analyze, and if these rules are still well distributed (have different assigned ranks), or all have ranks equal to the maximum assigned value for the specified data set. Table 4 shows a structure to be evaluated by the user. The sensitivity idea is inspired from [28].

Table 4. Sensitivity structure

rank	measure	inversely cumulative bins				histogram	Best rules
		1	2	3	...		

4.3. Average

Due to the fact that the number of bins is not the same when we have many rule sets to evaluate the sensitivity, so the number of rules that returned in the last interval also has not the same significance. Assume that the total number of measures to rank is fixed, the average ranks is used. The latter one is calculated according to the rank of each measure obtained from each rule set. A weight can be assigned to each rule set to favorite the level of importance, given by the user.

We use the average ranks to rank the measure over a set of rule sets based on the sensitivity values computed. The complement rule sets are benefited from this evaluation.

Table 5. Average structure to evaluate sensitivity on a set of rule set

rank	measure	rule set 1					rule set 2	...	avg. rank
		rank	first bin	last bin	image	best rule	...	...	

An average structure (see table 5) is constructed to have a quick evaluation on a set of rule sets. Each row represents a measure. The first two columns are represent the current rank of the measure. For each rule set, the rank, first bin, last bin, image and best rule assigned for each measure are represented. A remark is that the first and last bins are taken from the inversely cumulative distribution. The last column is the average rank of each measure calculated from all the rule sets studied.

5. Experiments

5.1. Rule sets

A set of four data sets [19] are collected, in which two data sets have opposite characteristics (i.e. correlated versus weakly correlated) and the others are two real-life data sets. Table 6 gives a quick description on these four data sets studied.

The categorical MUSHROOM data set ($\mathcal{D}_1$) from Irvine machine-learning database repository has 23 nominal attributes corresponding to the species of gilled mushrooms (i.e., edible or poisonous).

The synthetic T5I2D10k data set ($\mathcal{D}_2$) is obtained by simulating the transactions of customers in retailing businesses. The data set was generated using the IBM synthetic data generator [2]. $\mathcal{D}_2$ has the typical characteristic of the AGRAWAL data set T5I2D10k (T5: average size of the transactions is 5, I2: average size of the maximal potentially large itemsets is 2, D10k: number of items is 100).

The LBD data set ($\mathcal{D}_3$) is a set of lift breakdowns from the breakdown service of a lift manufacturer.

The EVAL data set ($\mathcal{D}_4$) is a data set of profiles of worker's performances which was used by the company *PerformanSe* to calibrate a decision support system in human resource management.

Table 6. Information on the data sets

Data set	Number of items	Transactions	
		Total	Average legnth
$\mathcal{D}_1$	128	8416	23
$\mathcal{D}_2$	81	9650	5
$\mathcal{D}_3$	92	2883	8.5
$\mathcal{D}_4$	30	2299	10

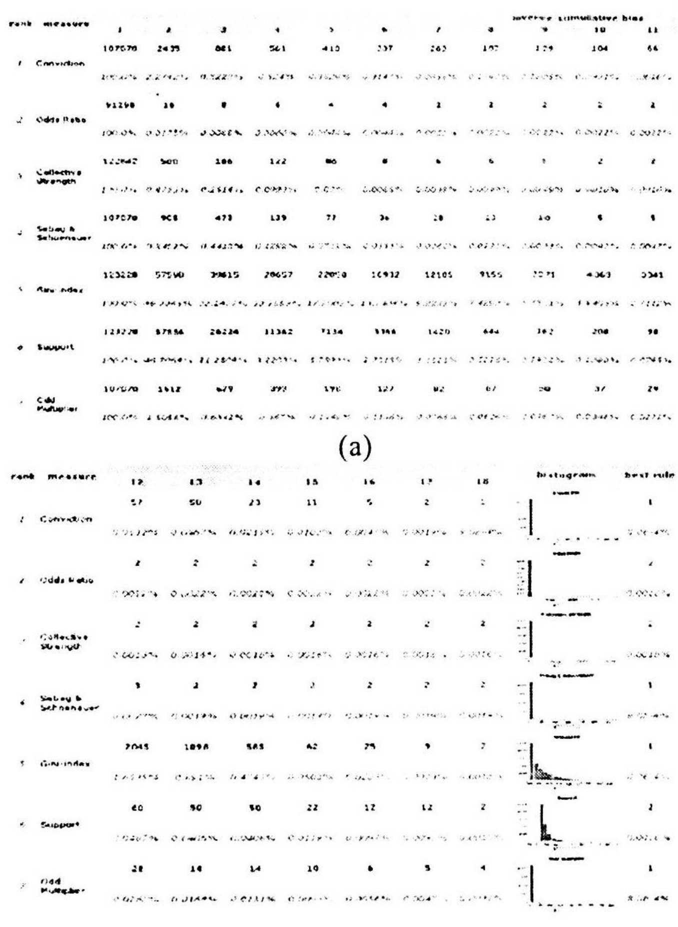
From the data sets discussed above, the corresponding rule sets (i.e., the set of association rules) are generated with the rule mining techniques [2].

Table 7. The rule sets generated

Data set	Rule set	Number of rules
$\mathcal{D}_1$	$\mathcal{R}_1$	123228
$\mathcal{D}_2$	$\mathcal{R}_2$	102808
$\mathcal{D}_3$	$\mathcal{R}_3$	43930
$\mathcal{D}_4$	$\mathcal{R}_4$	28938

5.2. Evaluation on a rule set

The sensitivity evaluation is based on the number of rules that falls in each interval is compared to rank the measures . For a measure on a rule set, the most significance interval will be the last bin (i.e., interval) of the inversely cumulative distribution. To have an approximation view on the sensitivity value, the number of rules has the maximum value is also retained. Fig. 5 (a) (b) shows the first seven measures that obtain the highest ranks. A remark is that the number of rules in the first interval is not always the same for all the measures because of the affectation of the number of NaN (not a number) values.



(a)

(b)

Fig. 5. Sensitivity rank on the $\mathcal{R}_1$ rule set.

An example of ranking two measures is given in Fig. 6 on the $\mathcal{R}_1$ rule set. The measure *Implication index* is ranked at the 13th place from a set of 40 measures while the measure *Rule Interest* is ranked at the 14th place. The meaning for this ranking is that the measure Implication index is more sensitive than the measure Rule Interest on $\mathcal{R}_1$ rule set even if the number of the most interesting rules returned with the maximum value is greater for the measure Rule Interest (3>2). The differences counted from each couple intervals, beginning from the last interval are quite important because the user will feel easier when looking at 11 rules in the last interval of the measure Implication index instead of looking at 64 rules from the same interval of the measure Rule Interest.

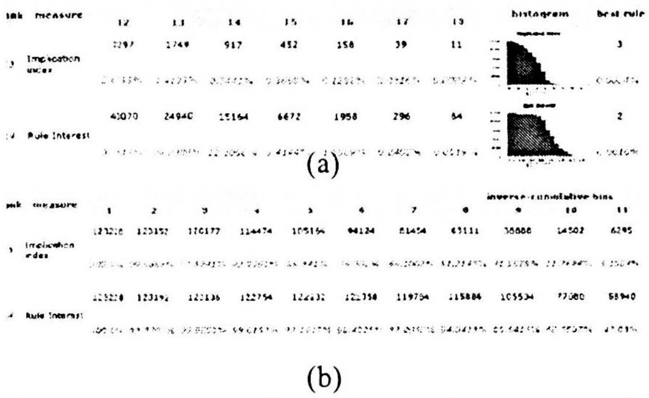


Fig. 6. Comparison of sensitivity values on a couple of measures of the $\mathcal{R}_1$ ruleset.

5.3. Evaluation on a set of rule sets

In Fig. 7 (a) (b), we can see the measure Implication Index goes strongly from place 13th in the $\mathcal{R}_1$ rule set to place 9th over all the set of the four rule sets while the measure Rule Interest goes lightly from place 14th to place 13th

(a)

rank	measure	MUSHROOM rank	first bin	last bin	image	best rule	T5I2D10k rank	first bin	last bin	image	best rule
1	Conviction	1	107070	1		1	9	99983	3		2
2	Odds Ratio	2	91298	2		2	4	97158	2		2
3	Gini-index	5	123228	2		1	1	102808	1		1
4	Odd Multiplier	7	107070	4		1	11	99983	7		7
5	Sebag & Schoenauer	4	107070	2		1	10	99983	3		2
6	J-measure	12	107070	8		1	8	99983	2		1
7	Support	6	123228	2		2	5	102808	2		2
8	Collective Strength	9	122842	2		2	7	102808	2		2
9	Implication index	13	123228	11		3	21	102808	191		1
10	Descriptive Confirm	8	123228	4		1	3	102808	1		1
11	TIC	18	123228	472		386	13	102808	14		14
12	Mutual Information	9	91220	4		2	15	97158	50		4
13	Rule Interest	14	123228	64		2	6	102808	2		2
14	Jaccard	19	123228	718		386	14	102808	42		14

(b)

rank	measure	LBD rank	first bin	last bin	image	best rule	EVAL rank	first bin	last bin	image	best rule	avg. rank
1	Conviction	3	43930	1		1	2	28938	1		1	3.75
2	Odds Ratio	6	43930	2		2	7	28938	2		2	4.75
3	Gini-index	7	43930	2		1	3	28938	2		1	5.25
4	Odd Multiplier	2	43930	1		1	2	28938	1		1	5.25
5	Sebag & Schoenauer	4	43930	1		1	3	28938	1		1	5.25
6	J-measure	5	43930	1		1	6	28938	1		1	7.75
7	Support	10	43930	4		2	11	28938	2		2	8.0
8	Collective Strength	11	43930	4		2	15	28938	4		2	8.5
9	Implication index	8	43930	2		1	4	28938	1		1	11.5
10	Descriptive Confirm	34	43930	320		1	5	28938	1		1	12.25
11	TIC	1	43930	0		43930	17	28938	8		2	12.25
12	Mutual Information	18	43930	6		2	20	28938	2		1	13.0
13	Rule Interest	13	43930	4		1	10	28938	10		1	13.0
14	Jaccard	12	43930	4		2	9	28938	2		2	13.5

Fig. 7. Sensitivity rank on all the set of rule sets (extracted).

6. Conclusion

Based on the sensitivity approach, we have ranked the 40 objective interestingness measures in order to find the most interesting rules in a rule set. By comparing the number of rules fallen in the most significant interestingness interval (i.e., the last bin in the inversely cumulative histogram) with the number of best rules (i.e., the number of rules having highest interestingness values), the sensitivity values have been determined. We have also proposed the sensitivity structure and the average structure to hold the sensitivity values on a single rule set as well as on a set of rule sets. The results obtained from the ARQAT tool [9] will provide some important aspects on the behaviors of the objective interestingness measures, as a supplementary view.

Together with the correlation graph approach [19], we will develop the dependant graph and the interaction graph by using the Choquet integral or the Sugeno integral [29,30]. These future results will provide a deeply insight view on the behaviors of interestingness measures on the knowledge represented in the form of association rules.

APPENDIX

N°	INTERESTINGNESS MEASURES	$f(n, n_X, n_Y, n_{X\bar{Y}})$
1	Causal Confidence	$1-\frac{1}{2}(\frac{1}{n_X}+\frac{1}{n_{\bar{Y}}})n_{X\bar{Y}}$
2	Causal Confirm	$\frac{n_X+n_{\bar{Y}}-4n_{X\bar{Y}}}{n}$
3	Causal Confirmed-Confidence	$1-\frac{1}{2}(\frac{3}{n_X}+\frac{1}{n_{\bar{Y}}})n_{X\bar{Y}}$
4	Causal Support	$\frac{n_X+n_{\bar{Y}}-2n_{X\bar{Y}}}{n}$
5	Collective Strength	$\frac{(n_X+n_{\bar{Y}}-2n_{X\bar{Y}})(n_X n_{\bar{Y}}+n_{\bar{X}} n_Y)}{(n_X n_Y+n_{\bar{X}} n_{\bar{Y}})(n_{\bar{X}Y}+n_{X\bar{Y}})}$
6	Confidence	$1-\frac{n_{X\bar{Y}}}{n_X}$
7	Conviction	$\frac{n_X n_{\bar{Y}}}{n n_{X\bar{Y}}}$
8	Coșine	$\frac{n_X-n_{X\bar{Y}}}{\sqrt{n_X n_Y}}$
9	Dependency	$\left\|\frac{n_{\bar{Y}}}{n}-\frac{n_{X\bar{Y}}}{n_X}\right\|$
10	Descriptive Confirm	$\frac{n_X-2n_{X\bar{Y}}}{n}$
11	Descriptive Confirmed-Confidence / Ganascia	$1-2\frac{n_{X\bar{Y}}}{n_X}$
12	EII (α=1)	$\sqrt{\varphi\times I^{\frac{1}{2\alpha}}}$
13	EII (α=2)	$\sqrt{\varphi\times I^{\frac{1}{2\alpha}}}$
14	Example & Contra-Example	$1-\frac{n_{X\bar{Y}}}{n_X-n_{X\bar{Y}}}$
15	F-measure	$\frac{2(n_X-n_{X\bar{Y}})}{n_X+n_Y}$
16	Gini-index	$\frac{(n_X-n_{X\bar{Y}})^2+n_{X\bar{Y}}^2}{n n_X}+\frac{n_{\bar{X}Y}^2+(n_{\bar{Y}}-n_{X\bar{Y}})^2}{n n_{\bar{X}}}-\frac{n_Y^2}{n^2}-\frac{n_{\bar{Y}}^2}{n^2}$
17	II	$1-\sum_{k=\max(0,n_X-n_Y)}^{n_{X\bar{Y}}}\frac{C_{n_Y}^{n_X-k}C_{n_{\bar{Y}}}^{k}}{C_n^{n_X}}$
18	Implication index	$\frac{n_{X\bar{Y}}-\frac{n_X n_{\bar{Y}}}{n}}{\sqrt{\frac{n_X n_{\bar{Y}}}{n}}}$
19	IPEE	$1-\frac{1}{2^{n_X}}\sum_{k=0}^{n_{X\bar{Y}}}C_{n_X}^{k}$
20	Jaccard	$\frac{n_X-n_{X\bar{Y}}}{n_Y+n_{X\bar{Y}}}$
21	J-measure	$\frac{n_X-n_{X\bar{Y}}}{n}\log_2\frac{n(n_X-n_{X\bar{Y}})}{n_X n_Y}+\frac{n_{X\bar{Y}}}{n}\log_2\frac{n n_{X\bar{Y}}}{n_X n_{\bar{Y}}}$
22	Kappa	$\frac{2(n_X n_{\bar{Y}}-n n_{X\bar{Y}})}{n_X n_{\bar{Y}}+n_{\bar{X}} n_Y}$

23	Klosgen	$\sqrt{\frac{n_X - n_{X\bar{Y}}}{n}}(\frac{n_Y}{n} - \frac{n_{X\bar{Y}}}{n_X})$
24	Laplace	$\frac{n_X + 1 - n_{X\bar{Y}}}{n_X + 2}$
25	Least Contradiction	$\frac{n_X - 2n_{X\bar{Y}}}{n_Y}$
26	Lerman	$\frac{n_X - n_{X\bar{Y}} - \frac{n_X n_Y}{n}}{\sqrt{\frac{n_X n_Y}{n}}}$
27	Lift / Interest factor	$\frac{n(n_X - n_{X\bar{Y}})}{n_X n_Y}$
28	Loevinger / Certainty factor	$1 - \frac{n n_{X\bar{Y}}}{n_X n_{\bar{Y}}}$
29	Mutual Information	$\frac{\frac{n_X - n_{X\bar{Y}}}{n}\log(\frac{n(n_X - n_{X\bar{Y}})}{n_X n_Y}) + \frac{n_{X\bar{Y}}}{n}\log(\frac{n n_{X\bar{Y}}}{n_X n_{\bar{Y}}}) + \frac{n_{\bar{X}Y}}{n}\log(\frac{n n_{\bar{X}Y}}{n_{\bar{X}} n_Y}) + \frac{n_{\bar{X}\bar{Y}}}{n}\log(\frac{n n_{\bar{X}\bar{Y}}}{n_{\bar{X}} n_{\bar{Y}}})}{\min(-(\frac{n_X}{n}\log(\frac{n_X}{n}) + \frac{n_{\bar{X}}}{n}\log(\frac{n_{\bar{X}}}{n})), -(\frac{n_Y}{n}\log(\frac{n_Y}{n}) + \frac{n_{\bar{Y}}}{n}\log(\frac{n_{\bar{Y}}}{n})))}$
30	Odd Multiplier	$\frac{(n_X - n_{X\bar{Y}})n_{\bar{Y}}}{n_Y n_{X\bar{Y}}}$
31	Odds Ratio	$\frac{(n_X - n_{X\bar{Y}})(n_{\bar{Y}} - n_{X\bar{Y}})}{n_{X\bar{Y}} n_{\bar{X}Y}}$
32	Pavillon / Added Value	$\frac{n_{\bar{Y}}}{n} - \frac{n_{X\bar{Y}}}{n_X}$
33	Phi-Coefficient	$\frac{n_X n_{\bar{Y}} - n n_{X\bar{Y}}}{\sqrt{n_X n_Y n_{\bar{X}} n_{\bar{Y}}}}$
34	Putative Causal Dependency	$\frac{3}{2} + \frac{4n_X - 3n_Y}{2n} - (\frac{3}{2n_X} + \frac{2}{n_{\bar{Y}}})n_{X\bar{Y}}$
35	Rule Interest	$\frac{n_X n_{\bar{Y}}}{n} - n_{X\bar{Y}}$
36	Sebag & Schoenauer	$\frac{n_X}{n_{X\bar{Y}}} - 1$
37	Support	$\frac{n_X - n_{X\bar{Y}}}{n}$
38	TIC	$\sqrt{TI(X \to Y) \times TI(\bar{Y} \to X)}$
39	Yule's Q	$\frac{n_X n_{\bar{Y}} - n n_{X\bar{Y}}}{n_X n_{\bar{Y}} + (n_Y - n_{\bar{Y}} - 2n_X)n_{X\bar{Y}} + 2n_{X\bar{Y}}^2}$
40	Yule's Y	$\frac{\sqrt{(n_X - n_{X\bar{Y}})(n_{\bar{Y}} - n_{X\bar{Y}})} - \sqrt{n_{X\bar{Y}} n_{\bar{X}Y}}}{\sqrt{(n_X - n_{X\bar{Y}})(n_{\bar{Y}} - n_{X\bar{Y}})} + \sqrt{n_{X\bar{Y}} n_{\bar{X}Y}}}$

References

[1] B. Baesens, S. Viaene, J. Vanthienen, "Post - processing of association rules," *SWP/KDD'00, Proceedings of The Special Workshop on Post-processing in conjunction with ACM KDD'00* (2000) 2.

[2] R. Agrawal, R. Srikant, "Fast algorithms fo r mining association rules," *VLDB'94, Proceedings of 20th International Conference on Very Large Data Bases* (1994) 487.

[3] R.J.Jr. Bayardo, R. Agrawal, "Mining the most interesting rules," KDD'99, *Proceedings of the 5th ACM SIGKDD International Confeference on Knowledge Discovery and Data Mining* (1999) 145.

[4] A. Ceglar, J.F. Roddick, "Association mining," *ACM Computing Surveys* 38(2) (2005) 1.

[5] S. Huang, G.I. Webb, "Efficiently identifying exploratory rules' significance," *Data Mining - Theory, Methodology, Techniques, and Applications – LNCS 3755* (2006) 64.

[6] G. Piatetsky-Shapiro, "Discovery, analysis, and presentation of strong rules," *Knowledge Discovery in Databases* (1991) 229.

[7] A. Silberschatz, A. Tuzhilin, "What makes patterns interesting in knowledge discovery systems," *IEEE Transactions on Knowledge and Data Engineering 5(6)* (1996) 970.

[8] G. Piatetsky-Shapiro, C.J. Matheus, "The interestingness of deviations," *AAAI'94, Knowledge Discovery in Databases Workshop* (1994) 25.

[9] H.X. Huynh, F. Guillet, H. Briand, "ARQAT: an exploratory analysis tool for interestingness measures," *ASMDA'05, Proceedings of the 11th International Symposium on Applied Stochastic Models and Data Analysis*, (2005) 334.

[10] N. Lavrac, P. Flach, B. Zupan, "Rule evaluation measures: a unifying view," *ILP'99, Proceedings of*

the 9th International Workshop on Inductive Logic Programming - LNAI 1634 (1999) 174.

[11] L. Geng, H.J. Hamilton, "Interestingness measures for data mining: A survey," *ACM Computing Surveys* 38(3) (2006) 1.

[12] C.C. Fabris, A.A. Freitas, "Discovering surprising instances of Simpson's paradox in hierarchical multidimensional data," *International Journal of Data Warehousing and Mining* 2(1) (2005) 27.

[13] R.J. Hilderman, "The Lorenz Dominance Order as a Measure of Interestingness in KDD," *PAKDD'02, Proceedings of the 6th Pacific- Asia Conference on Knowledge Discovery and Data Mining - LNCS 2336*, (2002) 177.

[14] R.J. Hilderman, H.J. Hamilton, "Measuring the interestingness of discovered knowledge: a Principled approach," *Intelligent Data Analysis* 7(4) (2003) 347.

[15] S. Bistarelli, F. Bonchi, "Interestingness is not a dichotomy: introducing softness in constrained pattern mining," *PKDD'05, 9th European Conference on Principles and Practice of Knowledge Discovery in Databases - LNCS 3721* (2005) 22.

[16] V. Bhatnagar, A. S. Al-Hegami and N. Kumar, "Novelty as a measure of interestingness in knowledge discovery," *International Journal of Information Technology* 2(1) (2005) 36.

[17] B. Padmanabhan, A. Tuzhilin, "On Characterization and Discovery of Minimal Unexpected Patterns in Rule Discovery," *IEEE Transactions on Knowledge and Data Engineering* 18(2) (2006) 202.

[18] P. Lenca, P. Meyer, B. Vaillant, S. Lallich, "On selecting interestingness measures for association rules: user oriented description and multiple criteria decision aid," *the European Journal of Operational Research* 184(2) (2008) 610. (In press).

[19] H.X. Huynh, F. Guillet, J.Blanchard, P. Kuntz, R. Gras, H. Briand, "A graph-based clustering approach to evaluate inte restingness measures: a tool and a comparative study. (Chapter 2)," *Quality Measures in Data Mining - Studies in Computational Intelligence*, Springer-Verlag Vol. 43 (2007) 25.

[20] J. Blanchard, F. Guillet, R. Gras, H. Briand, "Using information- theoretic measures to assess association rule interestingness," *ICDM'05 Proceedings of the 5th IEEE International Conference on Data Mining*, (2005) 66.

[21] D.R. Carvalho, A.A. Freitas, N.F.F. Ebecken, "Evaluating the correlation between objective rule interestingness measures and real human interest," *PKDD'05, the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases - LNAI 3731* (2005) 453.

[22] G. Adomavicius, A. Tuzhilin, "Expert-driven validation of rule-based user models in personalization applications," *Mining and Knowledge Discovery* 5(1-2) (2001) 33.

[23] R. Gras, P. Kuntz, "Discovering R-rules wit h a directed hierarchy," *Soft Computing - A Fusion of Foundations, Methodologies and Applications* 10(5) (2006) 453.

[24] G. Ritschard, D.A. Zighed, "Implication strength of classification rules," *ISMIS'06, Proceedings of the 16th International Symposium on Methodologies for Intelligent Systems - LNAI 4203* (2006) 463.

[25] Y. Jiang, K. Wang, A. Tuzhilin, A.W.C. Fu, "Mining Patterns That Respond to Actions," *ICDM'05, Proceedings of the Fifth IEEE International Conference on Data Mining* (2005) 669.

[26] http://www.jfree.org/jfreechart/index.php

[27] S. M. Ross, *Introduction to probability and statistics for engineers and scientists*, Wiley, 1987.

[28] H. Dalton, "The measurement of the inequality of incomes," *Economic Journal* 30 (1920) 348.

[29] J.L. Marichal, "An axiomatic approach of the discrete Sugeno integral as a tool to aggregate interacting criteria in a qualitative framework," *IEEE Transactions on Fuzzy Systems* 9 (1) (2001) 164.

[30] I. Kojadinovic, "Unsupervised aggregation of commensurate correlated attributes by means of the Choquet integral and entropy functionals," *International Journal of Intelligent Systems* 2007 (In press).