Semantic-based Relationship between Objective Interestingness Measures in Association Rules Mining

Rachasak Somyanonthanakul
School of ICT, SIIT, Thammasat University
Thammasat University
Pathum Thani, Thailand
d5922300164@g.siit.tu.ac.th

Monnapat Roonsamrarn
Faculty of Management Sciences
Panyapiwat Institute of Management
Nonthaburi, Thailand
monnapat.pim@gmail.com

Thanaruk Theeramunkong
School of ICT, SIIT, Thammasat University
Thammasat University
Pathum Thani, Thailand
thanaruk@siit.tu.ac.th

***Abstract*—This work investigates the semantic of 61 commonly used interestingness measures in order to explore their common and distinct characteristics, by means of a two-way contingency table of a pair of variables; A and B. As the first step, a synthetic data of six probability variables; $P(AB)$, $P(A\overline{B})$, $P(\overline{A}B)$, $P(\overline{A}\overline{B})$, $P(\overline{A})$ and $P(\overline{B})$ and profile of measurements are generated based on $P(A)$, $P(B)$, and $P(AB)$. The exploration will be done based on semantic relationship. Secondly, an extension is done to characterize among 61 interestingness measures. Thirdly, their similarity and dissimilarity among the measurments are investigated in terms of association and correlation points of view. Finally, the semantic hidden in the properties of each measure is revealed.**

***Index Terms*—Similarity measures, Association rules, Data mining**

I. Introduction

Association rules mining [1] [2] is widely used to find the cooccurrence of itemset in database using the interestingness measurements. Based on raw data, statistical methods play an important role to capture interesting patterns [3] [4]. Agrawal et. al. [2] proposed measurements to define the association patterns using two clssical measures, support and confidence which were later adapted to constraints. However, two measurements are not enough to capture the whole aspects of the interestingness rules.

Most of the measurements [5] have been described to discover patterns and solve this problem [6]. It is unclear whether the proposed measures had a strong relationship and truly effective measurement. Therefore, selecting an capable measure for suitable application becomes a crucial issue in data mining [7].

Nowadays, researchers focus on extract interestingness patterns from real world datasets [8]. However, it is unclear whether the real world datasets are appropriated and covered for the feature types of user interest. Currently, convolutional neural network (CNN) architectures demonstrated that synthesized data achieved an improvement on mean average precision when used as training data. The result shown improvement which spans from $\approx 5\%$ to $\approx 19\%$ across three widely used materials databases of real-world images, suggesting synthetic datasets can help to evaluate the performance of a data mining [9].

Besides the development of data mining, semantic datamining is described to apply for many applications. For example, [10] purposed the semantic similarity of the ontological distance metrics. These are later applied to explore the relationships between human phenotypes and bone dysplasia, suggesting semantic similarities can improve the efficiency of traditional interestingness measures in the association rule discovery process [11].

Therefore, the main objective is to improve and fulfill all above limit. Based on literature, a semantic is study of meanings that considers an explanations of the patterns [5]. For instance, semantic is involved with domain knowledge from the user, some researchers consider them a special type of subjective measure [12]. All possible interesting feature types of association pattern is investigated to explore a semantic based relationships between the objective interestingness measurements. The steps were listed as follows. First, all possible interesting feature types of co-occurrence cases generation were synthesized in term of association pattern. Second, the association rules were generated by using standard Apriori algorithm. the distance analysis. The last, semantic relation between interestingness measures were explored by associative rule mining techniques and correlation analysis.

The paper were organized as follows. Section 2 described related work. Section 3 is methods and section 4 show all of results. Finally, Section 5 concludes the paper and future work.

II. Related work

Based on our study focuses, the literature reviews can be categorized into two parts; 1) objective interestingness measurements and 2) semantic association rules mining.

A. Objective interestingness measurements

Since, a number of interestingness measures have been proposed to capture knowlegge form numerrous data [5]. Researchers intended to find association patterns that are interesting and useful for user applications.

Tan et al. [3] analyzed 21 measures using 10 synthesized datasets (E1-E10) from contingency table [13]. Their [14] results suggest that each measure has different properties which make them useful for some application domains but not for others. They also purposed an algorithm for helping users to select a suitable measure from the small dataset.

Ohsaki et al. [11] evaluated 40 interestingness measures using clinical datasets of meningitis and hepatitis. The results showed that the interestingness measures, accuracy, chi-square measure for one quadrant, relative risk, uncovered negative, and peculiarity, have a stable, reasonable performance in estimating real human interest in the medical domain.

Lenca et al. [15] described that the selection of interestingness measures is based on two-steps. The fiest step defined interestingness measures by their classical properties and then, multi-criteria decision was applied to aid for user who not a data mining expert.

Recently, Tew C. et al. [8] analyzed rule-ranking behavior of 61 interestingness measures which were conducted on the rules generated 110 different datasets. They concluded that domain knowledge is essential to selecte an appropriate interestingness measure for a particular task and business objective.

B. Semantic association rules mining

Semantic association rules mining is a technique to combine data mining and semantic techniques for post-mining and selection of association rules [16]. The exponential development of semantic web, numerous linked data from social community, companies or end-users are produced which connect data were share hidden relationship or "semantic association" [17]. Blanchard J et al. [18] studied in semantics-based classification of rule interestingness measures. The experiments showed that according to three criteria such as the subject, the scope, and the nature of the measure are novel and useful for classification of interestingness measures. They also showed that these criteria seem to us essential to grasp the meaning of the measures which aid user to select appropriate measures. Moreover, the classification allows one to compare the rules to closely related concepts such as similarities, implications, and equivalences.

Liu et al [19] proposed a general post-filtering framework to enhance robustness and accuracy of semantic concept detection using association and temporal analysis for concept knowledge discovery. They also described that co-occurrence of several semantic concepts could imply the presence of other concepts. The association mining techniques used to investigate inter-concept association relationships from annotations. The experiments from public dataset called TRECVID 2005 showed that post-filtering framework is both efficient and effective in improving the accuracy of semantic concept detection in video [20].

Paul et al. [10] explored the semantic similarity of the ontological distance metrics. The experiments showed that confidence appears to be the best interestingness measure regardless of way in which is computed, traditional or semantic. The use of semantics provides a marginal, but consistent, improvement in accuracy over traditional measures, suggesting use semantic association mining can improved performance of traditional data mining.

Table I
A TWO-WAY CONTINGENCY TABLE REPRESENTING THE FREQUENCY COUNTS FOR VARIABLE A AND B

	B	$\overline{B}$	$B+\overline{B}$
A	f_{11}	$f_{10}=f_{1+}-f_{11}$	f_{1+}
$\overline{A}$	$f_{01}=f_{+1}-f_{11}$	$f_{00}=N-f_{11}-f_{1+}-f_{+1}$	$f_{0+}=f_{01}-f_{00}$
$A+\overline{A}$	f_{+1}	$f_{+0}=f_{10}-f_{00}$	N

III. METHODS

In this section, we describe our approaches to find the semantic of interestingness measures.

A. Generate Dataset

We start with a case generated to cover all possible on contigency patterns. Then, all co-occurance $(A \rightarrow B)$ is computed using the selected 21 interstingness measures [8].

The characteristics are defined to investigate the semantic relationships. Two systematic methods,the similarity of association rule and distance analysis, are developed to explore the final results.

We generate a large number of synthesis association patterns $(A \rightarrow B)$ to cover all possible pattern. Each pattern describes association patterns which is defined using a two-way (2×2) contigency table. Each pattern is 9-variables (f_{11}, f_{+1}, $f_{1+}, f_{10}, f_{01}, f_{00}, f_{0+}, f_{+0}$ and N) as shown in Table 1.

B. Case Generation

Let f_{11} is the number of A and B collocated with each other. Given f_{1+} (= f_{11} + f_{10}) is the row summary representing maginal frequency of A. We define f_{+1} (= f_{11} + f_{01}) is column summary, showing maginal frequency of B. N is the total number of transaction. The dependent variables are following; $f_{10} = (f_{1+} - f_{11})$, $f_{01} = (f_{+1} - f_{11})$, and $(f_{00} = N - f_{01} - f_{10} - f_{11})$. The association pattern $(A \rightarrow B)$ is conducted the synthetic dataset with conditionnal probability using contingency table of six probability; $P(A,B)$, $P(A,\overline{B})$, $P(\overline{A},B)$, $P(\overline{A},\overline{B})$, $P(A)$, and $P(B)$.

The Algorithm 1 is defined to generated association pattern $(A \rightarrow B)$ that works on two input varibales. We define $S = \emptyset$ to be an empty dataset. Denote by the $Init$ is a lower bound and End is an upper bound in the range, respectively. Three independent variables f_{1+}, f_{+1}, and f_{11} are assigned a lower (i) and upper bound (j) to sequence of each iteration (Line 2-4). Five independent variables f_{10}, f_{11}, f_{01}, f_{00}, f_{0+}, and f_{+0} are computed on the two-way contingency constrains following Table1 (Line 5-9). Association pattern is generated under conditions to satisfy the constraints (Line 10).

C. Case Comparison

The Algorithm 2 generates case comparison. All associaton patterns are computed with all interestingness measurements (Line 2-4). A synthesis pattern that consists of each association rule and interestingness values obtained via the matrix A

Algorithm 1 Case Generation

Input: Init, End
Output: **S** Association patern
Initialisation : $i : Init, j : End$
1: $S = \emptyset$
2: **for** f_{1+} = i to j **do**
3: **for** f_{+1} = i to j **do**
4: **for** f_{11} = i to j **do**
5: $f_{10} = f_{1+} - f_{11}$
6: $f_{01} = f_{+1} - f_{11}$
7: $f_{00} = j - f_{11} - f_{1+} - f_{+1}$
8: $f_{0+} = f_{01} + f_{00}$
9: $f_{+0} = f_{10} + f_{00}$
10: **if** $(f_{11} \leq j) \wedge (f_{01} \geq 0) \wedge (f_{10} \geq 0)$ **then**
11: $S = \{f_{1+}, f_{+1}, f_{11}, f_{10}, f_{01}, f_{00}, f_{0+}, f_{+0}\}$
12: **end if**
13: **end for**
14: **end for**
15: **end for**

(Line 7). The comparison method is propoesd to generate a synthesis pattern. Every patterns are compared head-to-head with each of the other candidates pattern. A comparison result is represented in term of the compared cases. A positive tendency (P) is a incressing of interestingness value (Line 18-20). A negative tendency (N) is a decressing of interestingness value (Line 21-23). A equal tendency (E) is an identical interestingness value (Line 24-26). A dynamic pattern is generated from all possible association patterns. A number of dynamic pattern depents on three input varibales.

D. Semantic of Association Rule

Algorithm 3 provides steps for semantic association of interestingness measurements. After process association mining of comparison patterns with interestingness measures, association rules were generated by using Apriori algorithm (Line 2). The symmetric rule is a criteria for association rule extracted, one-antecedent to one-consequent (Line 3-4).

Any symmetric rule is compared rule-by-rule with each of the other candidate symmetric rules (Line 7-8). A bilateral symmetry rules is selected to the sematic rule (Line 9-11). The semantic rule is ranked using the confidence value to identify a strong level of interestingness measure.

E. Distance Analysis

We used Pearson's correlation coefficient to analyze the distance of similarity between interestingness measurement. A comparison pattern is converted to compatable with Pearson's correlation.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y}^2)}} \quad (1)$$

Three tendency patterns Positive(P), Negative(N), and Equal(E) are convert to (P = 1), (N = -1), and (E = 0). Equation (1) is Pearson's correlation coefficient formula.

Algorithm 2 Case Comparison

Input: S : Association patern, I : of interesting measurement
Output: D : Comparison dataset
Initialisation : $|S| \times |I|$ matrix A
1: **for** $s \in S$ **do**
2: **for** $i \in I$ **do**
3: $v_s(i)$ = value of s on i
4: $A_{si} = v_s(i)$
5: **end for**
6: **end for**
7: l = a number of record on A
8: **for** $m = 1$ to l **do**
9: **for** $n = m + 1$ to $l - 1$ **do**
10: **for** $i \in I$ **do**
11: $k++$
12: D_{jk} = *Tendency*(A, m, n, i)
13: $j++$
14: D_{jk} = *Tendency*(A, n, m, i)
15: **end for**
16: **end for**
17: **end for***Tendency*(A, m, n, i)
18: **if** $A_{ni} < A_{mi}$ **then**
19: $tend$ = P
20: **end if**
21: **if** $A_{ni} > A_{mi}$ **then**
22: $tend$ = N
23: **end if**
24: **if** $A_{ni}A_{mi}$ **then**
25: $tend$ = E
26: **end if**
27: **return** $tend$
28: **return** S

Algorithm 3 Semantic Rule Mining

Input: D : Comparison dataset
Output: R : Semantic Rule
1: **for** $d \in D$ **do**
2: F_i = association rules i generated by Apriori on d
3: LHS_i = left hand side of symmetry rule i on F
4: RHS_i = right hand side of symmetry rule i on F
5: **end for**
6: i = a number of symmetric association rules on D
7: **for** $m = l$ to i **do**
8: **for** $n = l$ to i **do**
9: **if** $(LHS_m = RHS_n) \wedge (RHS_m = LHS_n)$ **then**
10: $R \leftarrow LHS_m = RHS_m$
11: $R \leftarrow LHS_n = RHS_n$
12: **end if**
13: **end for**
14: **end for**
15: **return** R

Table II
THE SAMPLE OF VARIABLE PROBABILITY

No.	f_{1+}	f_{+1}	f_{11}	f_{00}	f_{01}	f_{10}	f_{0+}	f_{+0}	N
1	500	500	**300**	300	200	200	500	500	1000
2	500	500	**400**	400	100	100	500	500	1000
1 vs.2	E	E	P	P	N	N	E	E	E
2 vs.1	E	E	N	N	P	P	E	E	E

Denote, the x and y are interestiness measurement in dynamic pattern. The sample mean ($\bar{x}$, $\bar{y}$) are the corrected. Pearson's correlation coefficient r ranges from -1 to 1. We defined r of 1 indicates that comparing interestingness measures is perfect relationship, on a line for which y increases as x. Given r is -1 implies that all interestingness values lie on a line for which y decreases as x increases. The last, r of 0 implies that there is no linear correlation between interestingness measurements.

IV. EXPERIMENTS

In this section, we describe the experiment to explore the semantic interestingness measure. We described the comparative association patterns generated on their relationship type of categorization. The first, we generate dataset in Algorithm 1 to create the result of association pattern. Second, we analyzed the semantics of associative rules by conducting the synthesized comparative association patterns on 61 interestingness measures. Algorithm 2 used dataset to generate comparative pattern and wrote to another csv file. Then, Aogrithm 3 is the correlation of association rules that were tested to explore strong realtionship or semantic association rules. The result used Weka 3.8.2 to use associa function for Apiori algorithm. The distance analysis was performed standard Python library to generate the resule. We implemented in the Anaconda Python, and ran on 32 CPUs 2.10GHz, 503 GB of RAM at Sirindhorn International Institute of Technology, Thailand.

A. Case Comparison Dataset

Following algorithm Algorithm II, we generated 286 feasible two-way contingency tables to conduct our experiment. Table II shows an example, synthetic pattern is contained 2×2 contingency tables. The synthetic pattern is composed of occuring number that described by variables. We set all integer numbers satisfying $f_{11} < f_{1+}$, $f_{11} < f_{+1}$ and 1,000 that increase 100 in all variables for every step increasing. The variable f_{1+} and f_{+1} are fixed 500 and f_{11} increases from 300 to 400. The comparison results show E (equal) in variable f_{1+} and f_{+1}. The variable f_{11} show P (positive) of pair 1 vs.2 and N (negative). In this step, we apply the pairwise comparison method for twenty-one interestingness measures in every pattern (81,550 patterns). The interestingness measurement was compared to show tendency in each pair. The pattern of all records was measured using twenty-one interestingness measures. The result of comparison give positive trend (P) for a pair record least than, negative trend (N) for great than and equal trend (E) for a tie.

B. Semantic Association Rule Mining

Comparison patterns was computed the distance analysis using the Pearsons correlation. The matrix (61×61) was analyzed using Perasons correlation coefficient between interestingness measurement. We selected the highest correlation of paires compare with semnatic association. TableII showed experiment of semantic interestingness measures by confidence and support. The fourteen paire of semantic interestingness measures were explored. The result shown 40 association rules that the confidence is 1.000. All paires of semantic interestingness measures pairs are high support ranging (0.4671 - 0.4643), suggesting that the confidence is promising for user judgments. In correlation coefficient [8], the semantics rules can be sperated in ten groups. TableIII presents 40 association rules ordering by support. The summary of semantic rules are presentd in TableIII

C. Similarity of Interestingness measure

Figure1 presents a cluster of interestingness measurement. We perform Algorithm 2 on interestingness measures that produces a distance matrix which, after clustering, present in Figure1. A structure of the dendrogram, with a number of positively groups correlated measures adding in larger cluster, and a few relatively independent measures (e.g., Logical Necessity, Implication Index, CCS). The other groups, sizable cluster of generally positively correlated measures (i.e., distance less than 20, or correlation greater than 20), one from k-measure Prevalence (8 measures), and the other from EIC II DIR (21 measures), and a group of 28 measures (from Rick Relative Information Gain) that tend be negatively correlated with these.

V. DISCUSSION AND CONCLUSION

In this paper, we have proposed a systematic methods find the semantic relationship between interestingness measurement. A comparison dataset is generated cover all possible asoociation patterns. [1] [2] [6] [8] [11] [13] used the real-world dataset investigate interestingness measurements. All dataset is static dataset and not cover all possible association patterns. Semantic Association is our method that proposed investigate semantic relationship. Our method used confidence measure sematic relationship that give the strong rule [10]. Pearson Correlation [8] and [13] is computed describe degree of association that was proposed previous work. Our study, we introduced a method generate semantic relation of association rules by using all possible relation types synthetic co-concurrence patterns. The methods performed association data mining and compared each interestingness measures that of another measure in order characterize their similarity. The high confident and correlation is represented their semantic relation. Finally, we remind that our research has focused on interestingness measures in the context of association rule mining. Given the recent interest in synthesis pattern sets. Furture work, we could be applied the valuable result conduct our methodology and extend our study include interestingness measures for systematically.

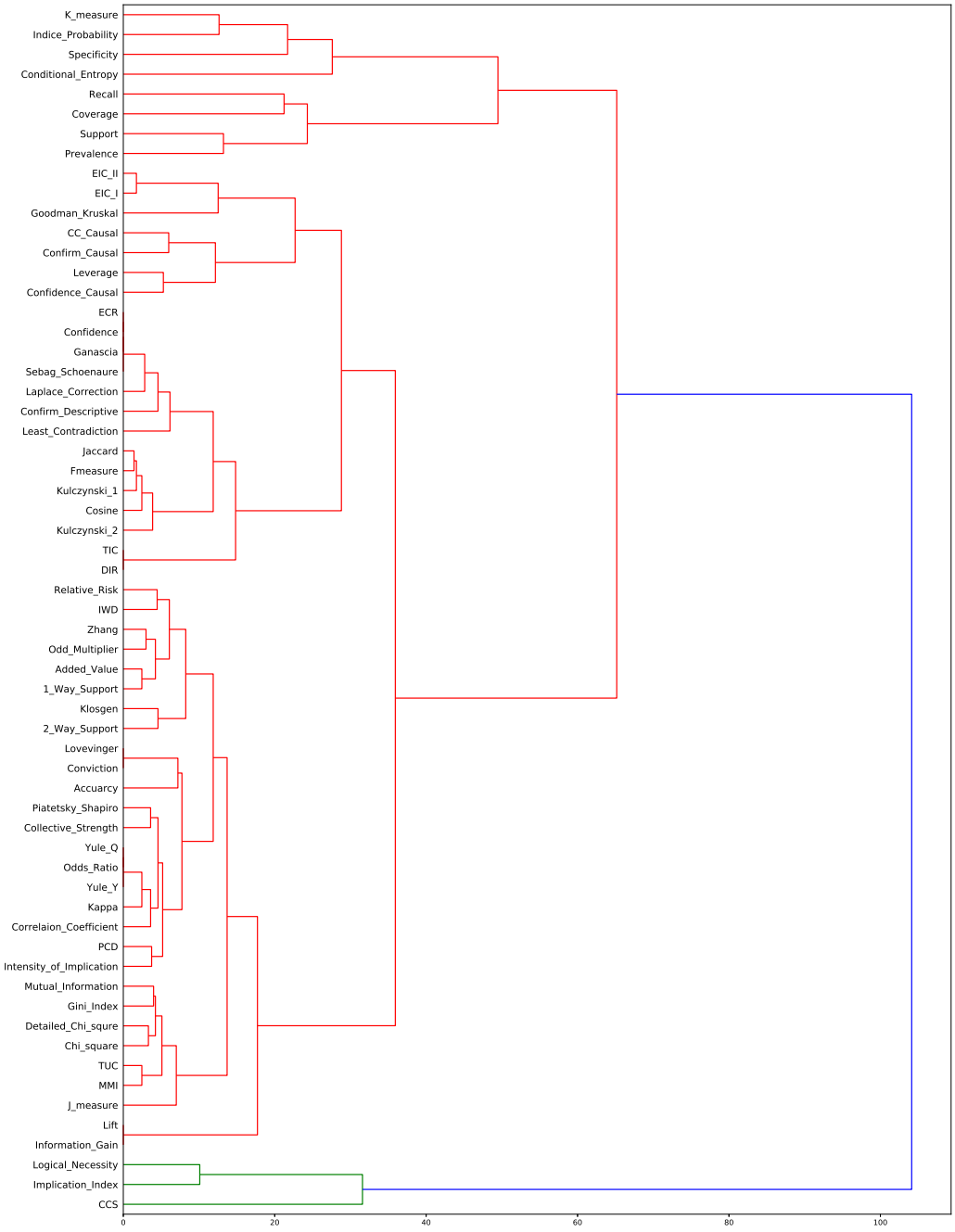

Figure 1. Hierarchical cluster analysis is visualized the correlation result of 61 interestingness measurement by turning the results a dendrogram.

Table III
SEMANTIC ASSOCIATION RULE MINING

No.	Measure 1		Measure 2	Supp	Conf
1	Information Gain = Negative	→	Lift = Negative	0.4761	1.0000
2	Lift = Negative	→	Information Gain = Negative	0.4761	1.0000
3	Information Gain = Positive	→	Lift = Positive	0.4761	1.0000
4	Lift = Positive	→	Information Gain = Positive	0.4761	1.0000
5	Odd Ratio = Negative	→	Yule Q = Negative	0.4722	1.0000
6	Yule Q = Negative	→	Odd Ratio = Negative	0.4722	1.0000
7	Odd Ratio = Positive	→	Yule Q = Positive	0.4722	1.0000
8	Yule Q = Positive	→	Odd Ratio = Positive	0.4722	1.0000
9	Odd Ratio = Negative	→	Yule Y = Negative	0.4722	1.0000
10	Yule Y = Negative	→	Odd Ratio = Negative	0.4722	1.0000
11	Odd Ratio = Positive	→	Yule Y = Positive	0.4722	1.0000
12	Yule Y = Positive	→	Odd Ratio = Positive	0.4722	1.0000
13	Yule Q = Negative	→	Yule Y = Negative	0.4722	1.0000
14	Yule Y = Negative	→	Yule Q = Negative	0.4722	1.0000
15	Yule Q = Positive	→	Yule Y = Positive	0.4722	1.0000
16	Yule Y = Positive	→	Yule Q = Positive	0.4722	1.0000
17	Confidence = Negative	→	Example and Counterexample Rate = Negative	0.4643	1.0000
18	Example and Counterexample Rate = Negative	→	Confidence = Negative	0.4643	1.0000
19	Confidence = Positive	→	Example and Counterexample Rate = Positive	0.4643	1.0000
20	Example and Counterexample Rate = Positive	→	Confidence = Positive	0.4643	1.0000
21	Confidence = Negative	→	Sebag Schoenaure = Negative	0.4643	1.0000
22	Sebag Schoenaure = Negative	→	Confidence = Negative	0.4643	1.0000
23	Confidence = Positive	→	Sebag Schoenaure = Positive	0.4643	1.0000
24	Sebag Schoenaure = Positive	→	Confidence = Positive	0.4643	1.0000
25	Confidence = Positive	→	Ganascia = Positive	0.4643	1.0000
26	Ganascia = Positive	→	Confidence = Positive	0.4643	1.0000
27	Confidence = Negative	→	Ganascia = Negative	0.4643	1.0000
28	Ganascia = Negative	→	Confidence = Negative	0.4643	1.0000
29	Example and Counterexample Rate = Negative	→	Ganascia = Negative	0.4643	1.0000
30	Ganascia = Negative	→	Example and Counterexample Rate = Negative	0.4643	1.0000
31	Example and Counterexample Rate = Positive	→	Ganascia = Positive	0.4643	1.0000
32	Ganascia = Positive	→	Example and Counterexample Rate = Positive	0.4643	1.0000
33	Example and Counterexample Rate = Negative	→	Sebag Schoenaure = Negative	0.4643	1.0000
34	Sebag Schoenaure = Negative	→	Example and Counterexample Rate = Negative	0.4643	1.0000
35	Example and Counterexample Rate = Positive	→	Sebag Schoenaure = Positive	0.4643	1.0000
36	Sebag Schoenaure = Positive	→	Example and Counterexample Rate = Positive	0.4643	1.0000
37	Ganascia = Negative	→	Sebag Schoenaure = Negative	0.4643	1.0000
38	Sebag Schoenaure = Negative	→	Ganascia = Negative	0.4643	1.0000
39	Ganascia = Positive	→	Sebag Schoenaure = Positive	0.4643	1.0000
40	Sebag Schoenaure = Positive	→	Ganascia = Positive	0.4643	1.0000

ACKNOWLEDGMENT

This research is financially supported under the Thammasat Universitys research fund, Center of Excellence in Intelligent Informatics, Speech and Language Technology and Service Innovation (CILS), and Intelligent Informatics and Service Innovation (IISI) Research Center, the Thailand Research Fund under grant number RTA6080013, as well as the STEM workforce Fund by National Science and Technology Development Agency (NSTDA).

REFERENCES

[1] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '93. New York, NY, USA: ACM, 1993, pp. 207–216. [Online]. Available: http://doi.acm.org/10.1145/170035.170072

[2] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in *Proceedings of the 20th International Conference on Very Large Data Bases*, ser. VLDB '94. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1994, pp. 487–499. [Online]. Available: http://dl.acm.org/citation.cfm?id=645920.672836

[3] P.-N. Tan, V. Kumar, and 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*. ACM, 2002, pp. 32–41.

[4] P.-N. Tan and V. Kumar, "Interestingness measures for association patterns: A perspective," in *Proc. of Workshop on Postprocessing in Machine Learning and Data Mining*, 2000.

[5] L. Geng and H. J. Hamilton, "Interestingness measures for data mining: A survey," *ACM Computing Surveys (CSUR)*, vol. 38, no. 3, p. 9, 2006.

[6] R. J. Hilderman and H. J. Hamilton, *Knowledge discovery and interestingness measures: A survey*. Citeseer, 1999.

[7] X. Li, H. Zhou, K. Shimada, K. Hirasawa *et al.*, "Analysis of various interestingness measures in class association rule mining," *SICE Journal of Control, Measurement, and System Integration*, vol. 4, no. 4, pp. 295–304, 2011.

[8] C. Tew, C. Giraud-Carrier, K. Tanner, and S. Burton, "Behavior-based clustering and analysis of interestingness measures for association rule mining," *Data Mining and Knowledge Discovery*, vol. 28, no. 4, pp. 1004–1045, 2014.

[9] G. Kalliatakis, A. Sticlaru, G. Stamatiadis, S. Ehsan, A. Leonardis, J. Gall, and K. D. McDonald-Maier, "Material Classification in the Wild: Do Synthesized Training Data Generalise Better than Real-World Training Data?" *ArXiv e-prints*, Nov. 2017.

[10] R. Paul, T. Groza, J. Hunter, and A. Zankl, "Semantic interestingness measures for discovering association rules in the skeletal dysplasia domain," *Journal of biomedical semantics*, vol. 5, no. 1, p. 8, 2014.

[11] M. Ohsaki, H. Abe, S. Tsumoto, H. Yokoi, and T. Yamaguchi, "Evaluation of rule interestingness measures in medical knowledge discovery in databases," *Artificial Intelligence in Medicine*, vol. 41, no. 3, pp. 177–196, 2007.

[12] H. Yao and H. J. Hamilton, "Mining itemset utilities from transaction databases," *Data & Knowledge Engineering*, vol. 59, no. 3, pp. 603–626, 2006.

[13] P.-N. Tan, V. Kumar, and J. Srivastava, "Selecting the right objective measure for association analysis," *Inf. Syst.*, vol. 29, no. 4, pp. 293–313, Jun. 2004. [Online]. Available: http://dx.doi.org/10.1016/S0306-4379(03)00072-3

[14] K. McGarry, "A survey of interestingness measures for knowledge discovery," *The knowledge engineering review*, vol. 20, no. 1, pp. 39–61, 2005.

[15] P. Lenca, P. Meyer, B. Vaillant, and S. Lallich, "On selecting interestingness measures for association rules: User oriented description and multiple criteria decision aid," *European journal of operational research*, vol. 184, no. 2, pp. 610–626, 2008.

[16] C. Kruengkrai, V. Sornlert, L. Anich, W. Buranasing, and T. C. Orn, "Semantic relation extraction from a cultural database," 2012.

[17] X. Jiang, X. Zhang, W. Gui, F. Gao, P. Wang, and F. Zhou, "Summarizing semantic associations based on focused association graph," in *International Conference on Advanced Data Mining and Applications*. Springer, 2012, pp. 564–576.

[18] J. Blanchard, F. Guillet, and P. Kuntz, "Semantics-based classification of rule interestingness measures," 2009.

[19] K.-H. Liu, M.-F. Weng, C.-Y. Tseng, Y.-Y. Chuang, and M.-S. Chen, "Association and temporal rule mining for post-filtering of semantic concept detection in video," *IEEE Transactions on Multimedia*, vol. 10, no. 2, pp. 240–251, 2008.

[20] T. M. Khoshgoftaar, M. Golawala, and J. Van Hulse, "An empirical study of learning from imbalanced data using random forest," in *Tools with Artificial Intelligence, 2007. ICTAI 2007. 19th IEEE international conference on*, vol. 2. IEEE, 2007, pp. 310–317.

Table IV
INTERESTINGNESS MEASUREMENT

No	Measurement	Formular
1	1-way Support (1SW)	$P(B\|A)log_2\frac{P(B\|A)}{P(B)}$
2	2-way Support (2SW)	$P(AB)log_2\frac{P(B\|A)}{P(B)}$
3	Accuracy (ACC)	$P(AB)+P(\overline{A}\,\overline{B})$
4	Added Value (AV)	$P(B\|A)+P(B)$
5	Chi-Square χ^2	$\frac{(P(AB)-P(A)P(B))^2N}{P(A)P(\overline{A})P(B)P(\overline{B})}$
6	Collective Strength (CS)	$\frac{P(AB)-P(A\overline{B})}{P(A)P(B)+P(\overline{A})P(\overline{B})}\frac{1-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(AB)-P(\overline{A}\,\overline{B})}$
7	Complement Class Support (CCS)	$\frac{P(A\overline{B})}{P(\overline{B})}$
8	Conditional Entropy (CE)	$-P(B\|A)\log_2P(B\|A)-P(\overline{B}\|A)\log_2P(\overline{B}\|A)$
9	Confidence (CON)	$P(B\|A)$
10	Confidence Causal (CDC)	$\frac{1}{2}(P(B\|A)+P(\overline{A}\|\overline{B}))$
11	Confirm Causal (CRC)	$P(AB)+P(\overline{A}\,\overline{B})-2P(A\overline{B})$
12	Confirm Descriptive (CRD)	$P(AB)-P(A\overline{B})$
13	Confirmed Confidence Causal (CCC)	$\frac{1}{2}(P(B\|A)+P(\overline{A}\|\overline{B}))-P(\overline{B}\|A)$
14	Conviction (CVC)	$\frac{P(A)P(\overline{B})}{P(A\overline{B})}$
15	Correlation Coefficient (CCO)	$\frac{P(AB)-P(A)P(B)}{\sqrt{P(A)P(B)P(\overline{A})P(\overline{B})}}$
16	Cosine (COS)	$\frac{P(AB)}{\sqrt{P(A)P(B)}}$
17	Coverage (COV)	$P(A)$
18	Dilated Chi-square (D(2))	$(\frac{P(A)P(\overline{A})P(B)P(\overline{B})}{(min(min(P(A),P(\overline{A})),min(P(B),P(\overline{B})))min(max(P(A),P(\overline{A})),max(P(B),P(\overline{B})))^2})^\alpha\chi^2$
19	Directed Information Ratio (DIR)	$\begin{cases}-\infty & \text{if } P(B)=1\\ 0 & \text{if } P(B)\le\frac{1}{2}\ and\ P(B\|A)\le\frac{1}{2}\\ 1+P(B\|A)\log_2P(B\|A)+P(\overline{B}\|A)\log_2P(\overline{B}\|A) & \text{if } P(B)\le\frac{1}{2}\ and\ P(B\|A)>\frac{1}{2}\\ 1+\frac{1}{P(B)\log_2P(B)+P(\overline{B})\log_2P(\overline{B}))} & \text{if } P(B)>\frac{1}{2}\ and\ P(B\|A)\le\frac{1}{2}\\ 1-\frac{P(B\|A)\log_2P(B\|A)+P(\overline{B}\|A)\log_2P(\overline{B}\|A)}{P(B)\log_2P(B)+P(\overline{B})\log_2P(\overline{B})} & \text{if } P(B)>\frac{1}{2}\ and\ P(B\|A)>\frac{1}{2}\end{cases}$
20	Entropic Implication Intensity 1 (EII1)	non
21	Entropic Implication Intensity 2 (EII2)	$\sqrt{IIM((1-H^\alpha_{B\|A})(1-H^\alpha_{A\|\overline{B}})^{\frac{1}{2\alpha}}}$ where, $H_{X\|Y}=-P(X\|Y)\log_2P(X\|Y)-P(\overline{X}\|Y)\log_2P(\overline{X}\|Y)$
22	Example and Counterexample Rate (ECR)	$1-\frac{P(A\overline{B})}{P(AB)}$
23	F-Measure (FM)	$\frac{2P(A\|B)P(B\|A)}{P(A\|B)+P(B\|A)}$
24	Ganascia (GAN)	$2P(B\|A)-1$
25	Gini Index (GI)	$P(A)(P(B\|A)^2+P(\overline{B}\|A)^2)+P(\overline{A})(P(B\|\overline{A})^2+P(\overline{B}\|\overline{A})^2)-P(B)^2-P(\overline{B})^2$
26	Goodman-Kruskal' s (GK)	$\frac{max(P_1,P_2)+max(P_3,P_4)+max(P_1,P_3)+max(P_2,P_4)-max(P(A),P(\overline{A}))-max(P(B),P(\overline{B}))}{2-max(P(A),P(\overline{A}))-max(P(B),P(\overline{B}))}$
27	Implication Index (IIN)	$\frac{\sqrt{N}P(A\overline{B})-P(A)P(\overline{B})}{\sqrt{P(A)P(\overline{B})}}$
28	Indice Probabiliste d'Ecart d'Equilibre	$1-\frac{1}{2^{N_A}}\sum_{k=0}^{N_{A\overline{B}}}\binom{N_A}{k}$
29	Information Gain (IG)	$\log_2\frac{P(AB)}{P(A)P(B)}$
30	Intensive of Implication (IIM)	$\frac{1}{2}-\frac{1}{2}\text{sgn}(\frac{IIN}{\sqrt{2}})\sqrt{1-e^{-(\frac{IIN}{\sqrt{2}})^2\frac{\frac{4}{\pi}+0.147(\frac{IIN}{\sqrt{2}})^2}{1+0.147(\frac{IIN}{\sqrt{2}})^2}}}$
31	Interestingness Weighting Dependency	$((\frac{P(AB)}{P(A)P(B)})^l-1)P(AB)^m$
32	Jaccard (JAC)	$\frac{P(AB)}{P(A)+P(B)-P(AB)}$
33	J-measure (JM)	$P(AB)\log_2\frac{P(B\|A)}{P(B)}+P(A\overline{B})\log_2\frac{P(\overline{B}\|A)}{P(\overline{B})}$
34	Kappa (κ)	$\frac{P(B\|A)P(A)+P(\overline{B}\|\overline{A})P(\overline{A})-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A)P(B)-P(\overline{A})P(\overline{B})}$
35	Klösgen (KLO)	$\sqrt{P(A)}(P(B\|A)-P(B)$
36	K-measure (KM)	$P(B\|A)\log_2\frac{P(B\|A)}{P(B)}+P(\overline{B}\|\overline{A})\log_2\frac{P(\overline{B}\|\overline{A})}{P(\overline{B})}-P(B\|A)\log_2\frac{P(B\|A)}{P(\overline{B})}-P(\overline{B}\|\overline{A})\log_2\frac{P(\overline{B}\|\overline{A})}{P(\overline{B})}$
37	Kulczynski 1 (KU1)	$\frac{P(AB)}{P(A\overline{B})+P(\overline{A}B)}$
38	Kulczynski 2 (KU2)	$\frac{1}{2}(\frac{P(AB)}{P(A)}+\frac{P(AB)}{P(B)})$
39	Laplace Correction (LAC)	$\frac{NP(AB)+1}{NP(A)+k}$
40	Least Contradiction (LEC)	$\frac{P(AB)-P(A\overline{B})}{P(B)}$
41	Leverage (LEV)	$P(B\|A)-P(A)P(B)$
42	Lift (LIF)	$\frac{P(B\|A)}{P(B)}$
43	Loevinger (LOE)	$1-\frac{P(A\overline{B})}{P(A)P(\overline{B})}$
44	Logical Necessity (LON)	$\frac{P(\overline{A}\|B)}{P(\overline{A}\|\overline{B})}$
45	Mutual Information (MI)	$P(AB)\log_2\frac{P(AB)}{P(A)P(B)}+P(A\overline{B})\log_2\frac{P(A\overline{B})}{P(A)P(\overline{B})}+P(\overline{A}B)\log_2\frac{P(\overline{A}B)}{P(\overline{A})P(B)}+P(\overline{A}\,\overline{B})\log_2\frac{P(\overline{A}\,\overline{B})}{P(\overline{A})P(\overline{B})}$
46	Normalized Mutual Information (NMI)	$\frac{MI}{-P(A)\log_2P(A)-P(\overline{A})\log_2P(\overline{A})}$
47	Odd Multiplier (OM)	$\frac{P(AB)P(\overline{B})}{P(B)P(A\overline{B})}$
48	Odd Ratio (OR)	$\frac{P(AB)P(\overline{A}\,\overline{B})}{P(A\overline{B})P(\overline{A}B)}$
49	Piatetsky-Shapiro (PS)	$N(P(AB)-P(A)P(B))$
50	Prevalence (PRE)	$P(B)$
51	Putative Causal Dependency (PCD)	$\frac{1}{2}(P(B\|A)-P(B))+(P(\overline{A}\|\overline{B})-P(\overline{A}))-(P(\overline{B}\|A)-P(\overline{B}))-(P(A\|\overline{B})-P(A))$
52	Recall (REC)	$P(A\|B)$
53	Relative Risk (REL)	$\frac{P(B\|A)}{P(B\|\overline{A})}$
54	Sebag-Schoenaure (SS)	$\frac{P(AB)}{P(A\overline{B})}$
55	Specificity (SPE)	$P(\overline{B}\|\overline{A})$
56	Support (SUP)	$P(AB)$
57	Theil Uncertainty Coefficient (TUC)	$\frac{MI}{-P(B)\log_2P(B)-P(\overline{B})\log_2P(\overline{B})}$
58	TIC	$\sqrt{DIR(A\Rightarrow B)DIR(\overline{B}\Rightarrow\overline{A})}$
59	Yule's Q (YQ)	$\frac{P(AB)P(\overline{A}\,\overline{B})-P(A\overline{B})P(\overline{A}B)}{P(AB)P(\overline{A}\,\overline{B})+P(A\overline{B})P(\overline{A}B)}$
60	Yule's Y (YY)	$\frac{\sqrt{P(AB)P(\overline{A}\,\overline{B})}-\sqrt{P(A\overline{B})P(\overline{A}B)}}{\sqrt{P(AB)P(\overline{A}\,\overline{B})}+\sqrt{P(A\overline{B})P(\overline{A}B)}}$
61	Zhang (ZHA)	$\frac{P(AB)-P(A)P(B)}{max(P(AB)(1-P(B),P(B)(P(A)-P(AB))}$