Interestingness Measures for Classification Rule Mining: Model Selection Ability

Pannapa Changpetch[1], Dennis K. J. Lin[2]

[1]Department of Mathematical Sciences, Bentley University,
Waltham, MA 02452, USA

[2]Department of Statistics, Pennsylvania State University,
University Park, PA 16802, USA

Abstract

This study analyzes the ability of interestingness measures to capture the correct predictive model through a new simulation design. The simulation is designed to be general enough to allow fair conclusions to be drawn without depending on subjective opinions. We found that the relative success of interestingness measures in capturing the correct model depends on two major factors: (1) the characteristics of the model and (2) the weight of each of three components—confidence, support, and the probability of result or class—in the measure's formula. No measure was found to perform best in all scenarios. However, we found that two groups of measures work well with different scenarios and that these groups of measures complement each other. That is, in any given scenario, the measures in one or the other of these two groups would perform the best. Therefore, in actual practice, when the characteristics of data are unknown, we propose using representatives of these two groups—confidence and lift—to capture the predictive models as either of these will capture the predictive model that best fits the given scenario.

Keywords: *Classification rule mining, Interestingness measures, Predictive model*

1. Introduction

A methodology for exploring relationships among items or variables in the form of rules, association rules analysis is a popular data mining technique introduced in the early 1990s [1]. A large number of studies have applied association rules analysis in a wide variety of research areas, including biology ([2], [3]), business and marketing ([4], [5], [6]), geography ([7], [8]), agriculture ([9], [10]), education ([11], [12], [13]), photography ([14], [15], [16]), and economics ([17], [18]).

The set of rules can be used for other purposes, including classification through a technique called classification rule mining (CRM), which is a subset of association rules analysis. The purpose of classification rule mining is straightforward: it is to find the rules—selected via interestingness measures—that are important to a given dataset. The selected rules will form the predictive model which is used to classify the response.

The classification rule mining technique has improved over time. For example, alternative interestingness measures have been developed and/or applied ([19], [20], [21]). And, researchers have compared measures in order to establish relationships between them and to identify the properties of each. In this regard, Ohsaki et al. (2004) [22] and Tan et al. (2004) [19] compared rankings established by experts with those established by interestingness measures for specific datasets. Lenca et al. (2004) [23] used a multicriteria decision strategy to produce a measures selection method whereby the properties of each measure are provided and the measures are selected based on the properties desired by users. In this method, it falls to the practitioner to weigh the significance of each property of each interestingness measure. In Vaillant et al.'s (2004) [24] work, based on their properties and their rankings of rules in a specified rule set, interestingness measures are clustered into groups. Geng and Hamilton (2006) [20] surveyed interestingness measures and analyzed them in several ways including by summarizing their properties and objectives. Kannan and Bhaskaran (2009) [25] studied interestingness measures that can be used for pruning with clustering. The literature covers interestingness measures for specific kinds of data. For example, Merceron and Yacef (2008) [26] and Pandey and Pal (2011) [27] studied interestingness measures for educational data, whereas Anandhavalli et al. (2010) [28]

studied interestingness measures for spatial gene expression databases.

In the present study, we analyze the behavior of respective interestingness measures—in terms of the ability of each to find the correct predictive model—through a new simulation design. Our objective is to identify the factors that affect the performance of each measure. We compare and analyze interestingness measures based on how they perform in a simulation designed to be general enough to allow fair conclusions to be drawn without depending on expert opinions. Based on our research in the literature, we regard our study as the most thorough endeavor to date to compare measures via a simulation protocol. We found that the success of interestingness measures in terms of capturing the correct model depends on two major factors: (1) the characteristics of the model and (2) the weight of each of three components—confidence, support, and the probability of class—in the measure's formula. Measures are also grouped based on their components and the recommended measures are given for use in practice.

This paper is organized as follows. Section 2 presents a review of classification rule mining and of the interestingness measures. Section 3 presents the simulation process and comparison procedure. Section 4 demonstrates the simulation results and provides an analysis. Section 5 offers a discussion and concluding remarks.

2. Classification Rule Mining and Interestingness Measures

Association rules analysis is a methodology designed to explore relationships among items in the form of rules. Each rule has two parts: the first comprises left-hand side item(s), or condition(s), and the second is a right-hand side item, or a result. The rule is always represented as a statement: If *condition*, then *result* [29]. When association rules analysis was introduced, the two measurements used were support (*s*), *computed* by s = Prob (*condition* and *result*), and confidence (*c*), computed by c = Prob (*condition* and *result*)/Prob (*condition*). Association rules analysis finds all the rules that meet two key thresholds: minimum support and minimum confidence [1].

A set of rules that meets these two thresholds can be used for other purposes, including classification. In fact, a technique called classification rule mining (CRM)—a subset of association rules analysis—was developed to find a set of rules in a database that would constitute an accurate classifier ([30], [31]). This technique uses an item to represent a pair consisting of a main effect and its corresponding integer value. More specific than association rules analysis, CRM has only one target, which must be specified in advance. The target is generally the response, which means the result of the rule (the right-hand-side item) can only be the response and its class. Therefore, the left-hand-side item (the condition) consists of the explanatory variable and its level. For example, assume there are k binary factors, X_1, X_2, …, X_k, and a binary response, Y. All variables have two levels, one denoted by 0 and the other by 1. Many rules can be generated by CRM, including If $X_1 = 1$, then $Y = 1$ where $X_1 = 1$ is the condition and $Y = 1$ is the result.

The literature on association rules analysis specifies numerous interestingness measures. And, based on a thorough search of the data mining papers and journals in the recent literature, we identified 27 measures (summarized with detailed notations in Table 1).

The 27 measures in Table 1 can be rewritten using three components: confidence (*c*), support (*s*), and the probability of class or result P(B). Some measures have only confidence (*c*) in their formula, e.g., confidence, example, and counter-example. Some measures have both confidence (*c*) and P(B), e.g., lift and added value. Some measures have all three components, e.g., cosine and implication index. Therefore, we categorize measures based on the component(s) in the formula of each, as shown in Table 2. These three groups appear to have the same components in their formulae.

Note that a limitation of our study is the lack of noise. We traded this limitation with the fair comparison and analysis among measures without the factor of noise involved. As a consequence, several measures yielded invalid values (from a zero denominator) or fixed values without noised added. These measures, which we have omitted, include the certainty factor, conviction, Loevinger, the odds multiplier, the odds ratio, Sebag-Schoenauer, Yule's Q, Yule's Y, Goodman and Kruskal and Zhang. In the following sections, we analyze the performance of each of these 27 measures via our proposed simulation framework.

Table 1: Interestingness measures (in alphabetical order)

For the rule If A, then B, *P(A)* is the probability of condition (A); *P(B)* is the probability of result (B) $P(AB) = P(A \cap B)$, $P(A') = 1 - P(A)$, $P(B') = 1 - P(B)$; support (*s*) = *P*(*AB*); and confidence (*c*) = *P*(*AB*)/*P*(*A*).

No.	Measure	Formula	Rewritten formula
1	Accuracy	$P(AB) + P(A'B')$	$1 - \frac{s}{c} - P(B) + 2s$
2	Added value	$P(B\|A) - P(B)$	$c - P(B)$
3	Confidence	$\frac{P(AB)}{P(A)}$	C
4	Collective strength	$\left(\frac{P(AB)+P(A'B')}{P(A)P(B)+P(A')P(B')}\right)\times \left(\frac{1-P(A)P(B)-P(A')P(B')}{1-P(AB)-P(A'B')}\right)$	$\left(\frac{1+2s-\frac{s}{c}-P(B)}{1+\frac{2sP(B)}{c}-\frac{s}{c}-P(B)}\right)\times \left(\frac{\frac{s}{c}+P(B)-\frac{2sP(B)}{c}}{\frac{s}{c}+P(B)-2s}\right)$
5	Cosine	$\frac{P(AB)}{\sqrt{P(A)P(B)}}$	$\sqrt{\frac{sc}{P(B)}}$
6	Dice index	$\frac{P(AB)}{P(AB)+\frac{1}{2}(P(A'B)+P(AB'))}$	$\frac{cs}{cs+\frac{1}{2}cP(B)+\frac{1}{2}s}$
7	Directed contribution to chi-squared	$\frac{\sqrt{N}(P(AB)-P(A)P(B))}{\sqrt{P(A)P(B)}}$	$\sqrt{\frac{ncs}{P(B)}}\left(1-\frac{P(B)}{c}\right)$
8	Example and counterexample rate	$1-\frac{P(AB')}{P(AB)}$	$2-\frac{1}{c}$
9	Ganascia index	$\frac{P(AB)-P(AB')}{P(A)}$	$2c-1$
10	Gini index	$P(A)\left(P(B\|A)^2+P(B'\|A)^2\right)+ P(A')\left(P(B\|A')^2+P(B'\|A')^2\right)- P(B)^2-P(B')^2$	$sc+\left(\frac{s}{c}-s\right)^2\frac{c}{s}+\frac{(P(B)-s)^2}{\left(1-\frac{s}{c}\right)}+ \frac{\left(1-\frac{s}{c}-P(B)+s\right)^2}{\left(1-\frac{s}{c}\right)}-P(B)^2-(1-P(B))^2$
11	Implication index	$-\frac{\sqrt{N}(P(AB)-P(A)P(B'))}{\sqrt{P(A)P(B')}}$	$-\sqrt{\frac{ns}{c}}\left(\frac{1+c-P(B)}{\sqrt{1-P(B)}}\right)$
12	Information gain	$\log\left(\frac{P(AB)}{P(A)P(B)}\right)$	$\log\left(\frac{c}{P(B)}\right)$

13	Jaccard	$\frac{P(AB)}{(P(A)+P(B)-P(AB))}$	$\frac{s}{\left(\frac{s}{c}-s+P(B)\right)}$
14	J-measure	$P(AB)\log\left\{\frac{P(B\|A)}{P(B)}\right\}+$ $P(AB')\log\left\{\frac{P(B'\|A)}{P(B')}\right\}+$	$s\log\left(\frac{c}{P(B)}\right)+$ $\left(\frac{s}{c}-s\right)\log\left(\frac{1-c}{1-P(B)}\right)$
15	Kappa	$\frac{P(AB)+P(A'B')-P(A)P(B)-P(A')P(B')}{1-P(A)P(B)-P(A')P(B')}$	$\frac{2s(c-P(B))}{cP(B)+s-2sP(B)}$
16	Klosgen	$\sqrt{P(AB)}\left(P(B\|A)-P(B)\right)$	$\sqrt{s}\left(c-P(B)\right)$
17	Kulczynski index	$\frac{1}{2}\left(\frac{P(AB)}{P(A)}+\frac{P(AB)}{P(B)}\right)$	$\frac{1}{2}\left(c+\frac{s}{P(B)}\right)$
18	Laplace	$\frac{N(AB)+1}{N(A)+2}$	$\frac{N(s)+1}{N\left(\frac{s}{c}\right)+2}$
19	Least contradiction	$\frac{P(AB)-P(AB')}{P(B)}$	$\frac{s\left(2-\frac{1}{c}\right)}{P(B)}$
20	Lift	$\frac{P(AB)}{P(A)P(B)}$	$\frac{c}{P(B)}$
21	One-way support	$P(B\|A)\log_2\left\{\frac{P(AB)}{P(A)P(B)}\right\}$	$c\log_2\left\{\frac{c}{P(B)}\right\}$
22	Piatetsky-Shapiro	$P(AB)-P(A)P(B)$	$s\left(1-\frac{P(B)}{c}\right)$
23	Relative risk	$\frac{P(B\|A)}{P(B\|A')}$	$\frac{(c-s)}{(P(B)-s)}$
24	Rogers-Tanimoto index	$\frac{1-P(A'B)-P(AB')}{1+P(A'B)+P(AB')}$	$\frac{\left(1-P(B)-\frac{s}{c}+2s\right)}{\left(1+P(B)+\frac{s}{c}-2s\right)}$
25	Tau-b	$\frac{P(AB)-P(A)P(B)}{\sqrt{P(A)P(B)P(A')P(B')}}$	$\sqrt{\frac{s(c-P(B))^2}{(c-s)P(B)(1-P(B))}}$
26	Two-way support	$P(AB)\log_2\left\{\frac{P(AB)}{P(A)P(B)}\right\}$	$s\log_2\left\{\frac{c}{P(B)}\right\}$
27	Two-way support variation	$P(AB)\log_2\left\{\frac{P(AB)}{P(A)P(B)}\right\}+$ $P(AB')\log_2\left\{\frac{P(AB')}{P(A)P(B')}\right\}+$ $P(A'B)\log_2\left\{\frac{P(A'B)}{P(A')P(B)}\right\}+$ $P(A'B')\log_2\left\{\frac{P(A'B')}{P(A')P(B')}\right\}$	$J-measure+$ $(P(B)-s)\log_2\left(\frac{c(P(B)-s)}{(c-s)P(B)}\right)+$ $(1+s-\frac{s}{c}-P(B))\log_2\left(\frac{1+s-\frac{s}{c}-P(B)}{\left\{1-\frac{s}{c}\right\}(1-P(B))}\right)$

Note: The Formula column states the original formula. The Rewritten formula states the rewritten formula as the form of support (*s*), confidence (*c*), and the probability of result or class (P(B)).

Table 2 Groups of interestingness measures based on components in their formulae.

Group	Component	Measures
1	*c*	Confidence, example and counter-example, Ganascia index, and Laplace
2	*c* and P(B)	Added value, information gain, lift, one-way support
3	*c*, *s*, and P(B)	Accuracy, collective strength, cosine, dice index, directed contribution to chi-squared, Gini index, implication index, Jaccard, J-measure, kappa, Klosgen, Kulczynski index, least contradiction, Piatetsky-Shapiro, relative risk, Roger-Tanimoto index, specificity, tau-b, two-way support, and two-way support variation

3. Simulation Process and Comparison Procedure

We used simulations to compare the interestingness measures in terms of performance. Our objective was to analyze the properties of the measures and the factors that affect the ability of each to find the correct predictive model or the correct response for four scenarios. We also observed the performance of each when the number of variables was larger. Therefore, we compared the performance of each measure when 6 binary variables were used with the performance of each when 10 binary variables were used. By using simulations, we were able to observe the ability of each measure to find the correct predictive model, which is the same as finding the correct response. The simulation was designed to be general enough to allow fair conclusions to be drawn. Section 3.1 presents detailed information about the design matrix, and Section 3.2 describes the simulation procedures.

3.1 Design Matrix

Two design matrices were used for this study: (1) a full factorial design for 6 binary variables, X_1–X_6, referred to as design matrix 2(6) with a total of 64 observations, and (2) a full factorial design for 10 binary variables, X_1–X_{10}, referred to as design matrix 2(10) with a total of 1,024 observations. We used these two design matrices mainly to study changes in each measure's performance as the number of variables increased. We used the full factorial design for our study, as this design covers all possible combinations among the independent variables and is fair in terms of sampling. This design does not produce any bias that would affect comparisons among the measures. Moreover, the design can easily be modified by adding and/or eliminating variables.

3.2 Simulation Process

For each design matrix, the simulation is separated into two parts: Part I generates the datasets, and Part II tests the performance of the interestingness measures with the generated datasets. An overview of the entire simulation process is shown in Figure 1.

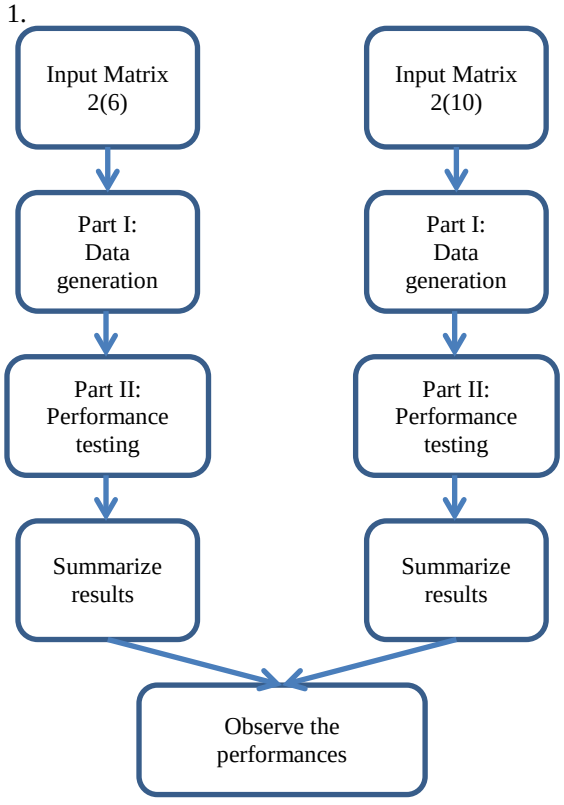

Fig. 1 Flowchart of overall simulation process.

The simulation process can be briefly summarized as follows. For each design matrix, a dataset is generated for each of the four scenarios of interest. Then, the generated datasets are used to test the ability of each of 27 measures to find the correct predictive model, which in our simulation is the same as finding the correct response. Next, the results from the performance testing are summarized. (Part I: data generation and Part II: performance testing are explained in more detail next.) Then, for the two design matrices, the measures are compared in regard to the respective ability of each to find the correct response. The Part I data-generating procedure is shown in Figure 2.

Part I: Data-generation Procedure

This is the first part of the simulation procedure. For our study, there are four scenarios of interest. To represent each scenario, we developed a data set in order to compare the 27 measures in terms of performance. The four datasets representing the four scenarios are shown in Table 3.

Given the input matrix, we generate response (y) for each scenario in Table 3. By combining the response with the input matrix, we produce the dataset. As there are four scenarios, we generate four datasets to use in Part II: performance testing.

However, before giving the details of the procedure for Part II, we will give detailed descriptions of the scenarios of interest, as explained next.

There are two models for each scenario, as shown in Table 3. The first model has class 1 as a result (model class 1), whereas the second model has class 0 as a result (model class 0). These two models are equivalent, which means that they generate the same response; i.e., Model 1 and Model 1^c are equivalent, Model 2 and Model 2^c are equivalent, Model 3 and Model 3^c are equivalent, and Model 4 and Model 4^c are equivalent. In other words, both Model 1 and Model 1^c are correct predictive models for scenario 1, both Model 2 and Model 2^c are correct predictive models for scenario 2, both Model 3 and Model 3^c are correct predictive models for scenario 3, and both Model 4 and Model 4^c are correct predictive models for scenario 4.

We noted earlier that the simulation is designed to be general enough to produce results from which reasonable conclusions can be drawn. This generalizability arises from our design in which all the models used cover cases of balanced and unbalanced models between model class 1 and model class 0, which is the key to analyzing each measure's performance.

Definition: A model is considered balanced when all its rules are the same size. In other words, the number of variables in the condition is the same for all the rules in the model.

The scenarios in Table 3 show all the cases when (i) both model class 1 and model class 0 are balanced, (ii) only one of the model classes is balanced, and (iii) neither model class is balanced.

For scenario 1, Model 1 is unbalanced (the rule size for Rule 1a is one and the rule size for Rule 1b is two), whereas Model 1^c is balanced (the rule size for both Rule 1′a and 1′b is two). For scenario 2, Model 2 is unbalanced, whereas Model 2^c is balanced. However, for scenario 1, the probability of class 1 (the class for the unbalanced Model 1) is higher than the probability of class 0 (the class for the balanced Model 1^c). For scenario 2, the probability of class 0 (the class for the balanced Model 2^c) is higher than the probability of class 1 (the class for the unbalanced Model 2). For scenario 3, Model 3 and Model 3^c are both balanced. For scenario 4, neither Model 4 nor Model 4^c is balanced.

Part II: Performance-testing Procedure

The second part of the simulation process is designed to test the performance of each rules selection measure. The main goal is to observe each measure's ability to find the correct predictive model. As noted earlier, there is more than one correct predictive model in each scenario; therefore, we designed the procedure to capture the correct response, which covers all the correct predictive models for each dataset. The procedure for Part II is shown in Figure 3.

The performance-testing procedure for each measure is described in detail next.

Step 1: Generate all rules with the main effect, two-way interactions, three-way interactions, and four-way interactions for each dataset.

Note that the rule with the main effect contains only one variable in the condition. The rule with the two-way interactions has two variables in the condition. The rule with the three-way interactions has three variables in the condition. The rule with the four-way interactions has four variables in the condition.

Step 2: Rank the rules in the following order:

i) Value of the measure
ii) Value of support (*s*)
iii) Rule size (the number of variables in the condition)—the smaller the rule size, the lower-ranked (better) the rule is.

Step 3: Search for the correct response from rule 1 to the higher-ranked (worse) rules. The rules that are unnecessary to or contradict the dataset are discarded.

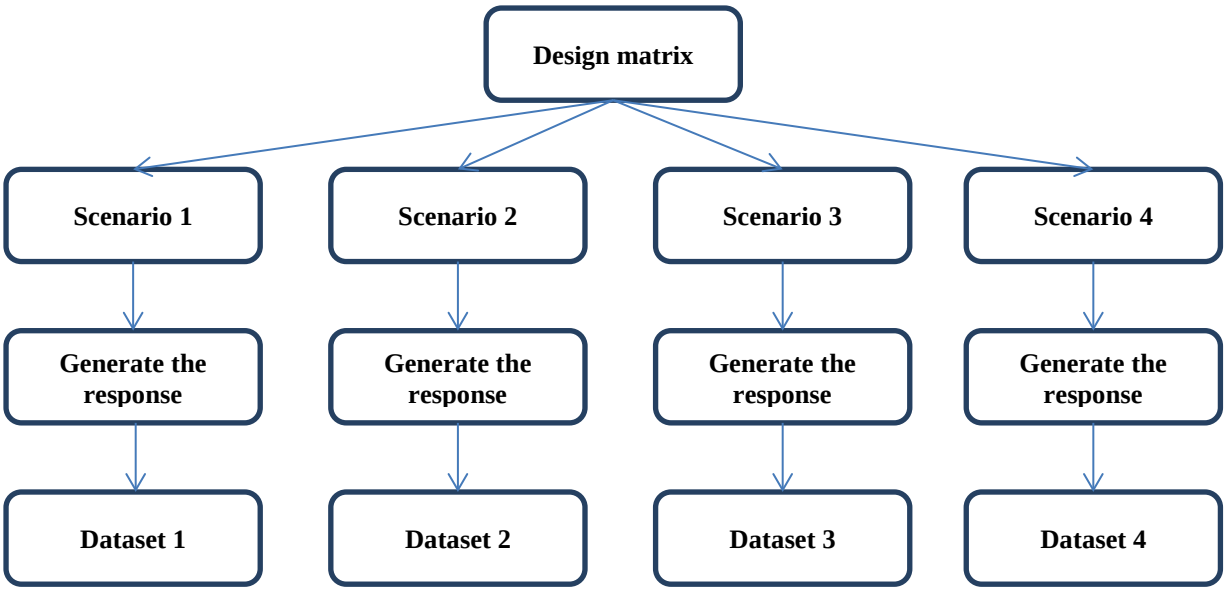

Fig. 2 Data-generation procedure for each design matrix.

Table 3: Scenarios of interest

Scenario	Model class 1	Rules in model class 1	Model class 0	Rules in model class 0
1	Model 1	1a: X1 = 0 → Y = 1 1b: X2 = 0, X3 = 0 → Y = 1	Model 1^c	1′a: X1 = 1, X2 = 1 → Y = 0 1′b: X1 = 1, X3 = 1 → Y = 0
2	Model 2	2a: X1 = 0, X2 = 0 → Y = 1 2b: X3 = 0, X4 = 0, X5 = 0 → Y = 1	Model 2^c	2′a: X1 = 1, X3 = 1 → Y = 0 2′b: X1 = 1, X4 = 1 → Y = 0 2′c: X1 = 1, X5 = 1 → Y = 0 2′d: X2 = 1, X3 = 1 → Y = 0 2′e: X2 = 1, X4 = 1 → Y = 0 2′f: X2 = 1, X5 = 1 → Y = 0
3	Model 3	3a: X1 = 0, X2 = 0 → Y = 1 3b: X3 = 0, X4 = 0 → Y = 1	Model 3^c	3′a: X1 = 1, X3 = 1 → Y = 0 3′b: X1 = 1, X4 = 1 → Y = 0 3′c: X2 = 1, X3 = 1 → Y = 0 3′d: X2 = 1, X4 = 1 → Y = 0
4	Model 4	4a: X1 = 0, X2 = 0 → Y = 1 4b: X1 = 0, X3 = 0, X4 = 0 → Y = 1	Model 4^c	4′a: X1 = 1 → Y = 0 4′b: X2 = 1, X3 = 1 → Y = 0 4′c: X2 = 1, X4 = 1 → Y = 0

Step 4: Record the number of rules needed to obtain the correct response. For example, we search for the correct response from rule 1 to rule *k*. (This means that rule *k* combined with some other lower-ranked (better) rule(s) generates the correct response.) The number *k* is recorded as the number of rules needed. If there are ties, the highest number of rules with all ties combined will be recorded.

Step 5: Summarize the number of rules needed to obtain the correct responses for each of the four datasets.

Note that we repeat this process for all the interestingness measures.

In summary, the datasets for the four scenarios in Table 3 are generated. Each measure is then tested with all four datasets, and the number of rules needed to find the correct response for each dataset is recorded.

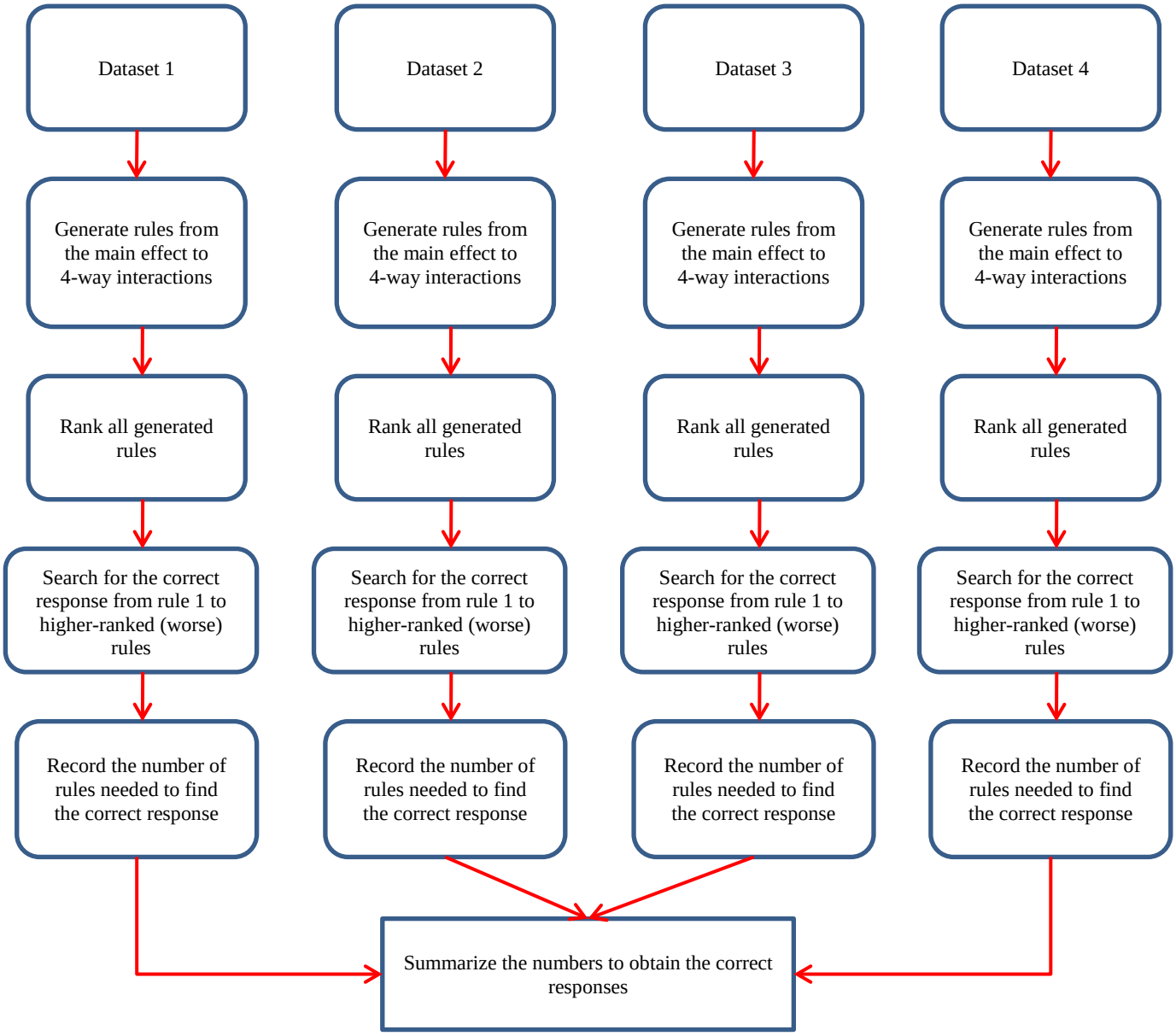

Fig. 3 Performance-testing procedure for each measure.

4. Simulation results and analysis

In this section, we analyze the factors that affect the ability of the measures to find the correct response. Note that to capture the correct response is the same as to capture all possible correct predictive models, as there is more than one correct model for each dataset. The performance of each of the measures for both design matrices is shown in Table 4.

From Table 4, it can be seen that the number of rules needed to capture the correct response for each of the four measures in group 1 is the same for all scenarios. Similarly, the number of rules needed to capture the correct response for each of the four measures in group 2 is the same for all scenarios.
In this section, we give some definitions and properties in order to provide the basis of our analysis for this comparison.

Table 4: Number of rules needed to find the correct response for each measure

Measure	Design matrix 2(6)				Design matrix 2(10)			
	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 1	Scenario 2	Scenario 3	Scenario 4
Conf	14	7	6	14	22	7	6	22
Examp	14	7	6	14	22	7	6	22
Ganas	14	7	6	14	22	7	6	22
Lapla	14	7	6	14	22	7	6	22
Added	2	10	2	10	2	18	2	18
Infor	2	10	2	10	2	18	2	18
Lift	2	10	2	10	2	18	2	18
Onesu	2	10	2	10	2	18	2	18
Accur	4	10	2	14	4	18	2	22
Coole	4	14	2	14	4	22	2	22
Cosin	4	14	2	18	4	14	2	26
Dice	4	16	2	22	4	24	2	38
Direc	2	10	2	10	2	18	2	18
Ginii	5	10	2	13	5	18	2	21
Impli	3	7	2	11	3	7	2	19
Jacca	4	16	2	22	4	24	2	38
Jmeas	2	10	2	11	2	18	2	19
Kappa	4	14	2	14	4	22	2	22
Klosq	2	10	2	10	2	18	2	18
Kulcz	4	9	2	13	4	9	2	21
Least	4	12	2	15	4	12	2	23
PS	4	11	2	16	4	11	2	24
Relat	3	12	2	11	3	20	2	19
Roger	4	10	2	14	4	18	2	22
Taub	4	10	2	12	4	18	2	20
Twosu	3	12	2	14	3	20	2	22
Twova	5	10	2	13	5	18	2	21

Note: Refer to Table 2, interestingness measures in group 1 are the first four measures (Conf–Lapla). Interestingness measures in group 2 are the next four measures (Added–Onesu). The rest are measures from group 3.

Definition: A child rule is a rule whose condition is the subset of its parent rule when the result is the same as its parent rule. As an example, If X1 = 0 and X2 = 0, then Y = 1 is the child rule of the rule If X1 = 0, then Y = 1.

For scenario 1, the measures in group 2 perform the best whereas the measures in scenario 1 perform the worst. Note that most measures in group 3 perform worse than group 2's measures do (except the directed contribution to chi-square, J-measure, and Klosgen) but perform better than group 1's measures do.

This is the scenario in which the probability of result P(B) for the class of the unbalanced model is higher than for the balanced model (Model 1 and Model 1^c from scenario 1). Note that for Rule 1a, support (*s*) is higher than for Rule $1a^c$ and that this also holds for Rule 1b and Rule $1b^c$. This means that support (*s*) for rules on the unbalanced side is higher than support (*s*) for rules on the balanced side. The affected measures contain only confidence (*c*) in their respective formulae. Measures in group 1 prefer the unbalanced model over the balanced model; e.g., these measures prefer Model 1 over Model 1^c. Note that these measures do not contain either support (*s*) or P(B) in their formulae. However, as we used support (*s*) as the second-ordered criterion to rank rules, these measures prefer the unbalanced side with higher support (*s*) over the balanced side with less support (*s*). Consider that when the number of variables increases, the additional child rule of the smaller-sized rule is added to the number of rules needed to find the correct model for the affected measures, i.e. measures in group 1. On the other hand, the other measures are not affected.

For scenario 2, the measures in group 1 perform the best whereas the performance of the measures in group 2 worsens when the number of variables increases, i.e., from design matrix 2(6) to design matrix 2(10). Some measures in group 3, such as the implication index and the Kulczynski index, perform well, but some, such as Jaccard and the dice index, perform poorly.

This is the scenario in which the probability of result P(B) for the class of the balanced model is higher than for the class of the unbalanced model (Model 2^c and Model 2 from scenario 2). Moreover, the unbalanced model has a larger rule size than does the balanced model. As the larger rule size implies smaller support (*s*), the unbalanced model has the rule with the least support (*s*) among all the rules for scenario 2. That is, rule 2a has less support (*s*) than either rule 2′a, 2′b, 2′c, 2′d, 2′e, or 2′f.

Note that this scenario does not affect the measures in group 1, as these measures prefer the balanced side with higher support (*s*) over the unbalanced side with smaller support (*s*). The measures in group 2 are clearly negatively affected by this scenario, as they prefer the unbalanced model over the balanced model based on the smaller P(B).

This scenario causes difficulty for some of the measures in group 3, as the lower P(B) and the smallest support (*s*) are both on the unbalanced model. Therefore, some of the measures in this category prefer the unbalanced model over the balanced model. The measures in group 3 that prefer the balanced model over the unbalanced model include cosine, the implication index, least contradiction, the Kulczynski index, and PS. The key factor is the difference in the weight of support (*s*) and P(B) in those measures and the difference in the values of support (*s*) and P(B) between the balanced and the unbalanced side.

For scenario 3, the performance of the measures in group 1 is the worst whereas all the other measures can capture the correct response within two rules. For this scenario 3, both model class 1 and model class 0 are balanced and no measures are affected when the number of variables increases.

For scenario 4, the measures in group 2 perform the best. Some measures in group 3 such as directed contribution to chi-square, and Klosgen, perform better than the measures in group 1. However, some measures in group 3, such as Jaccard and the Dice index, perform worse than measures in group 1. Consider that when the number of variables increases, all the measures are affected as both model class 1 and model class 0 are unbalanced, as with Model 4 and Model 4^c in scenario 4. The number of child rules of the lower-sized rules (Rule 4a and Rule 4^ca) increases. This higher number of child rules is included in the number of rules needed to find the correct response. Therefore, the number of rules needed to find the correct response increases if the number of variables increases.

5. Conclusions

This study analyzes the behavior of interestingness measures through a new framework designed for the study of the behavior of interesting measures with different scenarios. The results show that interestingness measures perform differently depending on the scenario. According to our findings, two factors that actually explain how each measure reacts to the data are (1) the weight of each of three components for each measure, i.e., confidence (*c*), support (*s*), and the probability of result P(B); and (2) the characteristics of the data, which we explained by the designed models, balanced vs unbalanced.

Even though we did not find one measure that always performs the best, we learned that measures in group 1 and measures in group 2 perform well with different scenarios. Measures in group 1 perform the best with scenario 2, whereas measures in group 2 perform the best with scenarios 1, 3, and 4. In practice, when the characteristics of the

datasets are unknown, we recommend using two measures: confidence (as the representative of measures from group 1) and lift (as the representative of measures from group 2) to find the prediction models, as either one will capture the predictive model the best in any scenario.

References

[1] R. Agrawal, S. Srikant, "Fast algorithms for Mining Association Rules", In Proceedings of the ACM SIGMOD conference on management of data, 1994, 487–499.

[2] C. Becquet, S. Blachon, B. Jeudy, J. F. Boulicaut, and O. Gandrillon, "Strong association rule mining for large gene expression data analysis: A case study on human SAGE data", Genome Biology, 3, 2002, 1–16.

[3] R. J. Kuo, S. Y. Lin, and C. W. Shih, "Mining association rules through integration of clustering analysis and ant colony system for health insurance database in Taiwan", Expert Systems with Applications, 33(3), 2007, 794–808.

[4] S. W. Changchien, and T. Lu, "Mining association rules procedure to support on-line recommendation by customers and products fragmentation", Expert Systems with Applications, 20, 2001, 325–335.

[5] K. Wang, S. Zhou, Q. Yang, and J. M. S. Yeung, "Mining customer value: From association rules to direct marketing", Data Mining and Knowledge Discovery, 11(1), 2005, 57–79.

[6] M. J. Shih, D. R. Liu, and M. L. Hsu, "Discovering competitive intelligence by mining changes in patent trends", Expert Systems with Applications, 37(4), 2010, 2882–2890.

[7] A. Appice, M. Ceci, A. Lanza, F. A. Lisi, and D. Malerba, "Discovery of spatial association rules in georeferenced census data: A relational mining approach", Intelligent Data Analysis, 7(6), 2003, 541–566.

[8] A. J. T. Lee, R. W. Hong, W. M. Ko, W. K. Tsao, and H. H. Lin, "Mining spatial association rules in image databases", Information Sciences, 177(7), 2007, 1593–1608.

[9] K. Matsumoto, "An experimental agricultural data mining system", In Proceedings of the First International Conference on Discovery Science (DS'98), Japan, 1998, 439–440.

[10] D. K. McIver, and M. A. Friedl, "Using prior probabilities in decision tree classification of remotely sensed data", Remote Sensing of Environment, 81, 2002, 253–261.

[11] J. Garcia, C. Romero, S. Ventura, and T. Calders, "Drawbacks and solutions of applying association rules mining in learning management systems", In Proceedings of the International Workshop on Applying Data Mining in e-learning (ADM'07), Crete, Greece, 2007, 13–22.

[12] P. Garcia, A. Amandi, S. Schiaffino, and M. Campo, "Evaluating Bayesian networks' precision for detecting students' learning styles", Computer and Education Journal, 49, 2007, 794–808.

[13] A. Merceron, and K. Yacef, "Interestingness measures for association rules in educational data", In Proceedings of Educational Data Mining Conference (EDM'08), 2008, 57–66.

[14] Q. Ding, Q. Ding, and W. Perrizo, "Association rule mining on remotely sensed images using p-trees", In Proceedings of Pacific Asia Conference on Knowledge Discovery and Data Mining (PAKDD'02), 2002.

[15] J. Tesic, S. Newsam, and B. S. Manjunath, "Mining image datasets using perceptual association rules", In SIAM'03 Workshop on Mining Scientific and Engineering Datasets, 2003.

[16] K. H. Liu, M. F. Weng, C. Y. Tseng, Y. Y. Chuang and M. S. Chen, "Association and temporal rule mining for post-processing of semantic concept detection in video", IEEE TMM, 10(2), 2008, 240–251.

[17] K. Dopfer, and J. Potts, "Evolutionary realism: A new ontology for economics", Journal of Economic Methodology, 11, 2004, 195–212.

[18] J. Mossong, N. Hens, M. Jit, et al., "Social contacts and mixing patterns relevant to the spread of infectious diseases", PLoS Medicine, 5, 2008, 381–391.

[19] P. N. Tan, V. Kumar, and J. Srivastava, "Selecting the right objective measure for association analysis. Information Systems", 29(4), 2004, 293–313.

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

[21] J. Blanchard, F. Guillet, and P. Kuntz, "Post-mining of association rules: Techniques for effective knowledge extraction, Yanchang Zhao, Chengqi Zhang, Longbing Cao (Ed.)", 2009, 56–79.

[22] M. Ohsaki, S. Kitaguchi, K. Okamoto, H. Yokoi, and T. Yamaguchi, "Evaluation of rule interestingness measures with a clinical dataset on hepatitis", In Proceedings of the 8th European Conference on Principles of Data Mining and Knowledge Discovery (Pkdd 2004), Pisa, Italy, 2004, 362–373.

[23] P. Lenca, P. Meyer, B. Vaillant, and S. Lallich, "A multicriteria decision aid for interestingness measure selection" Tech. Rep. Lussi-Tr-2004-01-En, Lussi Department, Get/Enst, Bretagne, France, 2004.

[24] B. Vaillant, P. Lenca, and S. Lallich, "A clustering of interestingness measures", In Proceedings of the 7th International Conference on Discovery Science (DS 2004), Padova, Italy, 2004, 290–297.
[25] S. Kannan and R. Bhaskaran, "Association rule pruning based on interestingness measures with clustering", International Journal of Computer Science Issues, 6(1), 2009, 35–43.
[26] A. Merceron, and K. Yacef, "Interestingness Measures for Association Rules in Educational Data" In International Conference on Educational Data Mining, Montreal, Canada, 2008, 57–66.
[27] U. K. Pandey, and S. Pal, "A data mining view on class room teaching language", International Journal of Computer Science Issues, 8(2), 2011, 277–282.
[28] M. Anandhavalli, M. K. Ghose, and K. Gauthaman, "Interestingness measure for mining spatial gene expression data using association rule", Journal of Computing, 2(1), 2010, 110–114.
[29] M. J. A. Berry, and G. Linoff, Data Mining Techniques: For Marketing, Sales, and Customer Support. John Wiley & Sons, 1997.
[30] B. Liu, W. Hsu, and Y. Ma, "Integrating classification and association rule mining", In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD'98), New York, 1998.
[31] J. R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, 1992.

Dr. Pannapa Changpetch is an assistant professor from the Department of Mathematical Sciences, Bentley University, MA, USA. She received her B.Eng. Industrial Engineering from Chulalongkorn University, Thailand, her M.Eng. Industrial Enineering from the University of Wisconsin-Madison and her Ph.D. in Statistics from the Pennsylvania State University.

Dr. Dennis K. J. Lin is a University Distinguish Professor from the Department of Statistics, the Pennsylvania State University. He received his doctoral degree in Statistics from the University of Wisconsin-Madison. Dr. Lin is an elected fellow of ASA (*American Statistical Association*), IMS (*Institute of Mathematical Statistics*) and ASQ (*American Society for Quality*), an elected member of ISI (*International Statistical Institute*), a lifetime member of ICSA (*International Chinese Statistical Association*), and a fellow of RSS (*Royal Statistical Society*). He is also the recipient of the 2015 Shewhart Medal (ASQ), the 2014 Hunter Award (ASQ/ASA), the 2010 Don Owen Award (ASA) and the 2004 Faculty Scholar Medal Award at Penn State University.