
Machine Learning, 62, 175–198, 2006

DOI: 10.1007/s10994-005-5066-8

On Mining Summaries by Objective Measures of Interestingness

NAIM ZBIDI zbidinaim@yahoo.fr
Institut Supérieur de Gestion de Tunis, 41, Rue de la Liberté-Cité, Bouchoucha- 2000 Tunis —Le Bardo, Tunisia

SAMI FAIZ sami.faiz@insat.rnu.tn
Institut National des Sciences Appliquées et de Technologie, Boulevard de la Terre, BP. 676- 1080 Tunis, Cedex Tunisia

MOHAMED LIMAM mohamed.limam@isg.rnu.tn
Institut Supérieur de Gestion de Tunis, 41, Rue de la Liberté-Cité, Bouchoucha- 2000 Tunis —Le Bardo, Tunisia

Editor: Stan Matwin

Published online: 29 January 2006

Abstract. Knowledge discovery in databases is used to discover useful and understandable knowledge from large databases. A process of knowledge discovery consists of two steps, the data mining step and the evaluation step. In this paper, evaluating and ranking the interestingness of summaries generated from databases, which is a part of the second step, is studied using diversity measures. Sixteen previously analyzed diversity measures of interestingness are used along with three not previously considered ones, brought from different well-known areas. The latter three measures are evaluated theoretically according to five principles that a measure must satisfy to be qualified acceptable for ranking summaries. A theoretical correlation study between the eight measures that satisfy all five principles is presented based on mathematical proofs. An empirical evaluation is conducted using three real databases. Then, a classification of the eight measures is deduced. The resulting classification is used to reduce the number of measures to only two, which are the best over all criteria, and that produce non-similar results. This helps the user interpret the most important discovered knowledge in his decision making process.

Keywords: data mining, diversity measures, association rules

1. Introduction

One of the most important problems in the field of knowledge discovery is the development of effective interestingness measures. These measures are divided into objective measures, those based upon the structure of discovered patterns (Han & Kamber, 2001), and subjective measures, those based upon the belief and the class of users who examine the patterns (Silberschatz & Tuzhilin, 1995). The latter measures are also divided by Silberschatz and Tuzhilin (1996) into two classes: actionability measures where a pattern is interesting if the user can do something to his advantage with it, and unexpectedness measures which rate a pattern as interesting if it is surprising to the user.

One approach to giving good measures of interestingness is the use of diversity ones as heuristic measures of interestingness. In the context of ranking the interestingness

Table 1. A sales database.

Store	Qty	Amount
1	2	20
2	3	30
3	1	10
4	4	40
5	1	10
6	3	30

of summaries, generated from databases, Hilderman and Hamilton (1999) proposed the use of sixteen measures. In this paper, besides those proposed, three measures not used previously are presented and evaluated. The goal is to select particular measures that are representative of distinct classes of interestingness measures. This will make the decision making easier and more objective: easier by reducing the number of measures and more objective by using only the ones which are the best over all criteria instead of applying all members of the set of measures or arbitrarily choosing one of them.

The use of valid interestingness measures, satisfying all required criteria, is very important in this context since the number of generated summaries from the same database may reach many thousands. The assessment of such number of summaries is a very hard task.

The remainder of this paper is organized as follows. Section 2 gives some preliminaries by defining a *summary* on which our methodology of treating the databases is based. In Section 3, diversity measures are discussed. In Section 4, we show that the new measures satisfy the five principles that a useful measure must satisfy. Then, a theoretical correlation study is made. In Section 5, we propose an empirical evaluation of the measures by describing their distribution characteristics. Hence, the eight theoretically acceptable measures are applied to three real databases and their classification is deduced.

2. The summary concept

Let S be a set of tables summarizing a given database using generalization techniques such as Concept Hierarchy (CH) or Domain Generalization Graph (DGG) that replace attribute values with more general concepts according to user specifications. Hence, S is a set of *summaries*. To explain this concept, we propose the following example: Table 1 gives a sales database and Figure 1[1] presents a CH and a DGG for this database.

A possible representation deduced from the DGG is the following set $S = S_1, S_2, S_3$, where S_i is summary i such that summary S_1, given in Table 2, gives a representation according to the first rule. Summaries S_2 and S_3, presented in Tables 3 and 4 respectively, give representations according to the second and the third rules.

The summaries correspond to the following three hypothetical rules:

– Sales are influenced by the fact of being in different regions.
– Sales are influenced by the fact of being in different cities.
– Sales are influenced by the fact of being in different stores.

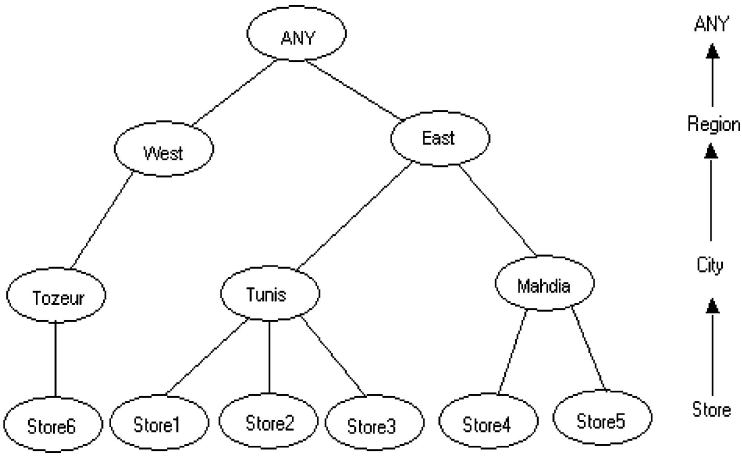

Figure 1. A concept hierarchy and a domain generalization graph for the store attribute.

These three rules might be chosen a priori according to the initial database by taking into consideration existing generalization relations in the DGG.

The last columns in Tables 2– 4, or *Count*, represent the number of occurrences of the values in the store column and they are the important ones for our study.

3. Heuristic measures of interestingness

The process of generating a summary from a database produces a set of tuples which are unique. Hence, they can be considered as a population with a probability distribution. Hilderman and Hamilton (1999) discussed sixteen heuristic measures of interestingness which are based on diversity measures taken from different well-known areas such as statistics, ecology, information theory, and management. Hilderman and Hamilton (2001) mentioned that these measures evaluate the distribution within a tuple and assign

Table 2. Summary S_1.

Store	Qty	Amount	Count
East	11	110	5
West	3	30	1

Table 3. Summary S_2.

Store	Qty	Amount	Count
Tunis	6	60	3
Mahdia	5	50	2
Tozeur	3	30	1

Table 4. Summary S_3.

Store	Qty	Amount	Count
Store1	2	20	1
Store2	3	30	1
Store 3	1	10	1
Store 4	4	40	1
Store 5	1	10	1
Store 6	3	30	1

Table 5. Results of the hypothetical example.

Summary	Interestingness value	Ranks assigned
S_1	x_1	1
S_2	x_2	2
S_3	x_3	3

Table 6. Notation.

m	The total number of tuples in a summary
n_i	The value of the derived count attribute for tuple t_i
$N = \sum_{i=1}^{m} n_i$	The total count (It is assumed that $N \geq 2$)
$p_i = \frac{n_i}{N}$	The actual probability for tuple t_i
$\overline{q} = \frac{1}{m}$	The uniform probability for tuple t_i (for all i=1,2, . . . ,m
$\overline{u} = \frac{N}{m}$	The count for tuple t_i (i=1,2,. . .,m) according to a uniform distribution
$r_i = \frac{n_i + \overline{u}}{2N}$	The probability for tuple t_i

a single real-valued index that represents its interestingness. Thus, they allow the assignment of a rank to the current summary according to the set of summaries generated from the original database.

In our example in Section 2, we assume that, by using a given measure of interestingness, the values generated are x_1, x_2 and x_3 for summaries S_1, S_2 and S_3, respectively. Also we assume that $x_1 > x_2 > x_3$[2]. Results are presented in Table 5 which says that summary S_1 is the most interesting one. This allows us to conclude that sales are more influenced by the fact of being in different regions than of being in different cities or different stores which means that the first rule is the most interesting.

To introduce these measures, the used notation is presented in Table 6.[3] Table 7 presents the measures previously studied by Hilderman and Hamilton (1999, 2000, 2001).

In the following, three not previously considered measures of interestingness are presented. Some modifications are introduced to the original formulas to produce valid interestingness measures. Note that the measures are independent of measurement units.

Table 7. Hilderman and Hamilton's set of interestigness measures.

$I_{\text{Variance}} = \frac{\sum_{i=1}^{m}(p_i-\overline{q})^2}{m-1}$
$I_{\text{Simpson}} = \sum_{i=1}^{m} p_i^2$
$I_{\text{Shannon}} = -\sum_{i=1}^{m} p_i \log_2 p_i$
$I_{Total} = \text{m} \times I_{Shannon}$
$I_{\text{Max}} = \log_2 m$
$I_{\text{McIntosh}} = \frac{N-\sqrt{\sum_{i=1}^{m} n_i^2}}{N-\sqrt{N}}$
$I_{\text{lorentz}} = \overline{q}\sum_{i=1}^{m}(m-i+1)p_i$
$I_{\text{Gini}} = \frac{\overline{q}\sum_{i=1}^{m}\sum_{j=1}^{m}\lvert p_i-p_j\rvert}{2}$
$I_{Berger} = \max(p_i)$
$I_{\text{Shutz}} = \frac{\sum_{i=1}^{m}\lvert p_i-\overline{q}\rvert}{2m\overline{q}}$
$I_{\text{Bray}} = \frac{\sum_{i=1}^{m} min(n_i,\overline{u})}{N}$
$I_{\text{Whittaker}} = 1-\left(0.5\sum_{i=1}^{m}\lvert p_i-\overline{q}\rvert\right)$
$I_{\text{Kullback}} = \log_2 m-\left(\sum_{i=1}^{m} p_i \log_2 \frac{p_i}{\overline{q}}\right)$
$I_{\text{MacArthur}} = \left(-\sum_{i=1}^{m} r_i \log_2 r_i\right)-\left(\frac{\left(-\sum_{i=1}^{m} p_i \log_2 p_i\right)+\log_2 m}{2}\right)$
$I_{\text{Theil}} = \frac{\sum_{i=1}^{m}\lvert p_i \log_2 p_i-\overline{q}\log_2\overline{q}\rvert}{m\overline{q}}$
$I_{\text{Atkinson}} = 1-\left(\prod_{i=1}^{m}\frac{p_i}{\overline{q}}\right)^{\overline{q}}$

3.1. The I_{Rae} measure

The I_{Rae} measure, introduced by Rae and Taylor (1970), is based on an index of ethnic fractionalization. It measures the degree of ethnic diversity and it is given by

$$I_{\text{Rae}} = \frac{\sum_{i=1}^{m} n_i(n_i-1)}{N(N-1)}. \tag{1}$$

When applied to Table 2, I_{Rae} yields 0.667.

3.2. The I_{CON} measure

The I_{CON} measure, introduced by Egghe and Rousseau (1991), is based on concentration measuring. It is given by

$$I_{\text{CON}} = \sqrt{\frac{\left(\sum_{i=1}^{m} p_i^2\right)-\overline{q}}{1-\overline{q}}}. \tag{2}$$

When applied to Table 2, I_{CON} yields 0.667.

3.3. *The I_{Hill} measure*

The I_{Hill} measure, introduced by Hill (1973), is based on a compound diversity measure which depends on the species proportional abundance. It is given by

$$I_{\text{Hill}} = 1 - \frac{1}{\sqrt{\sum_{i=1}^{m} p_i^3}}. \tag{3}$$

When applied to Table 2, I_{Hill} yields -0.310.

4. Theoretical evaluation

4.1. *Interestingness principles*

Hilderman and Hamilton (2000, 2001) proposed five principles that a measure should satisfy to be acceptable for ranking the interestingness of discovered summaries. Among the five principles, Egghe and Rousseau (1991) used four to classify concentration measures. The following notation is used:

- $(n_1, \ldots, n_m)$ is a vector of values derived from the database (e.g. the count values in the example in Table 2) such that $n_1 \geq n_2 \cdots \geq n_m$.
- $f(n_1, \ldots, n_m)$ is a function of m variables which is a general measure of interestingness.

The principles are used below to evaluate measures of interestingness for summaries deduced from a single dataset, so N is fixed (n_i's and N have the same meaning as in Section 3).

Zero-valued n_i's are eliminated because we suppose that a zero in the *count* column is without importance for our study.

4.1.1. Minimum value principle (P1). *Given a vector $(n_1, n_2, \ldots, n_m)$ if $n_i = n$ for all i; $f(n_1, n_2, \ldots, n_m) = f(n, n, \ldots, n)$ attains its minimum.*

P1 means that the interestingness is at its minimum level when the tuple counts are all equal. For example (3, 3), (17, 17, 17), etc.

4.1.2. Maximum value principle (P2). *Given a vector $(n_1, n_2, \ldots, n_m)$ if $n_1 = N - m + 1$, $n_i = 1$ for $i = 2, \ldots, m$ and $N > m$, $f(n_1, n_2, \ldots, n_m) = f(N - m + 1, 1, \ldots, 1)$ attains its maximum.*

This means that the interestingness must have its maximum value in the case of perfect concentration between the tuple counts. For example for $m = 2$ and 3 respectively, $N = 5$ and 45 respectively, we have (4, 1) and (43, 1, 1).

4.1.3. Skewness principle (P3). *Given a vector $(n_1, \ldots, n_m)$ where $n_1 = N - m + 1$, $n_i = 1$, $i = 2, \ldots, m$, $N > m$, and a vector $(n_1 - c, n_2, \ldots, n_m, n_{m+1}, \ldots, n_{m+c})$,* where $n_1 - c > 1$ *and $n_i = 1$, $i = 2, \ldots, m + c$; then $f(n_1, n_2 \ldots, n_m) > f(n_1 - c, n_2, \ldots n_m, n_{m+1}, \ldots n_{m+c})$.*

In the case of perfect concentration, P3 means that a summary containing m tuples will be more interesting than a summary containing $(m + c)$ tuples. For example $f(44, 1) > f(43, 1, 1)$.

4.1.4. Permutation invariance principle (P4). *Given a vector* $(n_1, \ldots, n_m)$ *and any permutation* $(i1, \ldots, im)$ *of* $(1, \ldots, m)$; $f(n_1, \ldots, n_m) = f(n_{i1}, \ldots, n_{im})$.

This principle states that interestingness is not a labeled property, meaning that interestingness does not change when the order of tuple counts changes. For example $f(1, 2, 3) = f(1, 3, 2) = f(2, 3, 1) = f(2, 1, 3) = f(3, 2, 1) = f(3, 1, 2)$.

4.1.5. Transfer principle (P5). *Given a vector* $(n_1, \ldots, n_m)$; $0 < c < n_j$ *and* $n_i \geq n_j$; *then* $f(n_1, \ldots, n_i + c, \ldots, n_j - c, \ldots, n_m) > f(n_1, \ldots, n_i, \ldots, n_j, \ldots, n_m)$.

If we make a strictly positive transfer from a tuple count to another whose count is greater, P5 states that interestingness increases. For example $f(6, 5, 2, 1) > f(6, 4, 3, 1)$.

4.2. Proofs

In this section, it is shown that the three proposed measures are acceptable to evaluate and rank the interestingness of discovered summaries according to the five principles described in Section 4.1. Mathematical proofs are given in the following.

4.2.1. Principle P1. *Given a vector* $(n_1, n_2, \ldots, n_m)$ *if* $n_i = n$ *for all* i; $f(n_1, n_2, \ldots, n_m) = f(n, n, \ldots, n)$ *attains its minimum.*

I_{Rae}: It needs to be shown that for any vector $(n + c, n - d_2, \ldots, n - d_m)$ whose values are not uniformly distributed where at least one $d_i > 0$ and $\sum_{i=2}^{m} d_i = c$;

$$f(n + c, n - d_2, \ldots, n - d_m) > f(n, n, \ldots, n),$$

we need to determine the sign of

$$\frac{\sum_{i=1}^{m} n(n-1)}{N(N-1)} - \frac{(n+c)(n+c-1) + \sum_{i=2}^{m}(n-d_i)(n-d_i-1)}{N(N-1)},$$

which has the same sign as

$$\begin{aligned} A &= mn^2 - mn - \left[n^2 + c^2 + 2nc - n - c + \sum_{i=2}^{m} n^2 + \sum_{i=2}^{m} d_i^2 \right. \\ &\quad \left. -2n \sum_{i=2}^{m} d_i - \sum_{i=2}^{m} n + \sum_{i=2}^{m} d_i \right] \\ &= mn^2 - mn - \left(mn^2 + c^2 - mn + \sum_{i=2}^{m} d_i^2 \right) \end{aligned}$$

$$= c^2 - \sum_{i=1}^{m} d_i^2 < 0.$$
$$\Rightarrow I_{\text{Rae}} \text{ satisfies } P_1.$$

I_{CON}: It is needed only to show that

$$\sum_{i=1}^{m} p_i^2 < \left(\frac{n+c}{N}\right)^2 + \sum_{i=2}^{m} \left(\frac{n-d_i}{N}\right)^2.$$
$$\sum_{i=1}^{m} p_i^2 = \sum_{i=1}^{m} \left(\frac{n_i}{N}\right)^2 = \sum_{i=1}^{m} \left(\frac{n}{N}\right)^2 = \sum_{i=1}^{m} \left(\frac{1}{m}\right)^2 = m\left(\frac{1}{m^2}\right) = \frac{1}{m}.$$

Hence, we need to find the sign of

$$\frac{N^2}{mN^2} - \frac{(n+c)^2}{N^2} - \frac{1}{N^2} \sum_{i=2}^{m} (n-d_i)^2,$$

which has the same sign as

$$\begin{aligned} B &= \frac{N^2}{m} - (n+c)^2 - \sum_{i=2}^{m} (n-d_i)^2 \\ &= \frac{N^2}{m} - n^2 - 2nc - c^2 - \sum_{i=2}^{m} n^2 + 2n \sum_{i=2}^{m} d_i - \sum_{i=2}^{m} d_i^2 \\ &= \frac{N^2}{m} - mn^2 - c^2 - \sum_{i=2}^{m} d_i^2 \\ &= \frac{m^2 n^2}{m} - mn^2 - c^2 - \sum_{i=2}^{m} d_i^2 \\ &= -c^2 - \sum_{i=2}^{m} d_i^2 < 0 \\ &\Rightarrow I_{\text{CON}} \text{ satisfies } P_1. \end{aligned}$$

I_{Hill}: We need to show that

$$\sum_{i=1}^{m} p_i^3 < \left(\frac{n+c}{N}\right)^3 + \sum_{i=2}^{m} \left(\frac{n-d_i}{N}\right)^3,$$

where

$$\sum_{i=1}^{m} p_i^3 = \sum_{i=1}^{m} \left(\frac{n_i}{N}\right)^3 = \sum_{i=1}^{m} \left(\frac{n}{N}\right)^3$$

which is equivalent to determining the sign of

$$\begin{aligned} C &= \sum_{i=1}^{m} n^3 - (n+c)^3 - \sum_{i=2}^{m} (n-d_i)^3 \\ &= mn^3 - (n^3 + 3n^2c + 3nc^2 + c^3) - \left(\sum_{i=2}^{m} n^3 - 3n^2 \sum_{i=2}^{m} d_i + 3 \sum_{i=2}^{m} n d_i^2 - \sum_{i=2}^{m} d_i^3 \right) \\ &= -3nc^2 - c^3 - 3n \sum_{i=2}^{m} d_i^2 + \sum_{i=2}^{m} d_i^3 < 0, \text{ since } c^3 = \left(\sum_{i=2}^{m} d_i \right)^3 > \sum_{i=2}^{m} d_i^3. \\ &\Rightarrow I_{\text{Hill}} \text{ satisfies } P_1. \end{aligned}$$

4.2.2. Principle P2. *Given a vector* $(n_1, n_2, \ldots, n_m)$ *if* $n_1 = N - m + 1$, $n_i = 1$ *for* $i = 2, \ldots, m$ *and* $N > m$, $f(n_1, n_2, \ldots, n_m) = f(N\text{-}m + 1, 1, \ldots, 1)$ *attains its maximum.*
I_{Rae}: Let $n_1 = n_1' + c$. We need to show that

$$\begin{aligned} & f(n_1' + c, n_2' - d_2, \ldots, n_m' - d_m) f(n_1', n_2', \ldots, n_m') \\ D &= (n_1' + c)(n_1' + c - 1) + \sum_{i=2}^{m} (n_i' - d_i)(n_i' - d_i - 1) - \sum_{i=1}^{m} n_i'(n_i' - 1) \\ &= n_1'^2 + 2n_1'c + c^2 - n_1' - c + \sum_{i=2}^{m} n_i'^2 + \sum_{i=2}^{m} d_i^2 - 2 \sum_{i=2}^{m} n_i' d_i \\ &\quad - \sum_{i=2}^{m} n_i' + \sum_{i=2}^{m} d_i - \sum_{i=1}^{m} n_i'^2 + \sum_{i=1}^{m} n_i' \\ &= \sum_{i=1}^{m} n_i'^2 + 2n_i'c + c^2 - \sum_{i=1}^{m} n_i' - c + \sum_{i=2}^{m} d_i^2 - 2 \sum_{i=2}^{m} n_i' d_i \\ &\quad + \sum_{i=2}^{m} d_i - \sum_{i=1}^{m} n_i'^2 + \sum_{i=1}^{m} n_i' \\ &= c^2 + \sum_{i=2}^{m} d_i^2 + 2n_1'c - 2 \sum_{i=2}^{m} n_i' d_i \\ &= c^2 + \sum_{i=2}^{m} d_i^2 + 2n_1' \sum_{i=2}^{m} d_i - 2 \sum_{i=2}^{m} n_i' d_i > 0. \\ &\Rightarrow I_{\text{Rae}} \text{ satisfies } P_2. \end{aligned}$$

I_{CON}: It is sufficient to show that

$$\left(\frac{n_1' + c}{N} \right)^2 + \sum_{i=2}^{m} \left(\frac{n_i' - d_i}{N} \right)^2 > \sum_{i=1}^{m} \left(\frac{n_i}{N} \right)^2$$

or that

$$(n_1' + c)^2 + \sum_{i=2}^{m}(n_i' - d_i)^2 > \sum_{i=1}^{m} n_i^2.$$

$$\begin{aligned} E &= n_1'^2 + 2n_1'c + c^2 + \sum_{i=2}^{m} n_i'^2 - 2\sum_{i=2}^{m} n_i'd_i + \sum_{i=2}^{m} d_i^2 - \sum_{i=1}^{m} n_i^2 \\ &= \sum_{i=1}^{m} n_i'^2 + 2n_1'c + c^2 - 2\sum_{i=1}^{m} n_i'd_i + \sum_{i=1}^{m} d_i^2 - \sum_{i=1}^{m} n_i'^2 \\ &= 2n_1'c + c^2 - 2\sum_{i=1}^{m} n_i'd_i + \sum_{i=1}^{m} d_i^2 > 0. \\ &\Rightarrow I_{\mathrm{CON}} \text{ satisfies } P_2. \end{aligned}$$

I_{Hill}: We need only to show that

$$\left(\frac{n_i' + c}{N}\right)^3 + \sum_{i=2}^{m}\left(\frac{n_i' + d_i}{N}\right)^3 > \sum_{i=1}^{m}\left(\frac{n_i}{N}\right)^3$$

or

$$\begin{aligned} (n_1' + c)^3 + \sum_{i=2}^{m}(n_i' + d_i)^3 &> \sum_{i=1}^{m} n_i^3 \\ F &= n_1'^3 + 3n_1'^2c + 3n_1'c^2 + c^3 + \sum_{i=2}^{m} n_i'^3 - 3\sum_{i=2}^{m} n_i'^2d_i \\ &\quad + 3\sum_{i=2}^{m} n_i'd_i^2 - \sum_{i=2}^{m} d_i^3 - \sum_{i=1}^{m} n_i'^3 \\ &= 3n_1'^2c + 3n_i'c^2 + c^3 - 3\sum_{i=2}^{m} n_i'^2d_i + 3\sum_{i=2}^{m} n_i'd_i^2 - \sum_{i=2}^{m} d_i^3 > 0. \\ &\Rightarrow I_{\mathrm{Hill}} \text{ satisfies} P_2. \end{aligned}$$

4.2.3. Principle P3. *Given a vector* $(n_1, \ldots, n_m)$ *where* $n_1 = N - m + 1$, $n_i = 1$, $i = 2, \ldots, m$, $N > m$, *and a vector* $(n_1 - c, n_2, \ldots, n_m, n_{m+1}, \ldots, n_{m+c})$, *where* $n_1 - c > 1$ *and* $n_i = 1$, $i = 2, \ldots, m + c$; *then* $f(n_1, n_2 \ldots, n_m) > f(n_1 - c, n_2, \ldots n_m, n_{m+1}, \ldots n_{m+c})$.

I_{Rae}: We need to show that

$$\sum_{i=1}^{m} n_i(n_i - 1) > (n_1 - c)(n_1 - c - 1) + \sum_{i=2}^{m+c}(n_i)(n_i - 1)$$

we have

$$\sum_{i=2}^{m} n_i(n_i - 1) = 0$$

because

$$n_i = 1; \quad i = 2, \ldots, m + c.$$

Therefore, it is obvious that

$$\sum_{i=1}^{m} n_i(n_i - 1) > \sum_{i=1}^{m} (n_i c)(n_i c - 1).$$
$$\Rightarrow I_{\text{Rae}} \text{ satisfies } P_3.$$

I_{CON}: We need only to show that

$$\sum_{i=1}^{m} \left(\frac{n_i}{N}\right)^2 > \left(\frac{n_1 c}{N}\right)^2 + \sum_{i=2}^{m+c} \frac{1}{N}$$

or

$$\sum_{i=1}^{m} n_i^2 > (n_1 - c)^2 + \sum_{i=2}^{m+c} 1$$
$$G = n_1^2 + m - 1 - n_1^2 + 2n_1 c - c^2 - (m + c - 1)$$
$$= 2n_1 c - c^2 - c,$$

which has the same sign as

$$2n_1 - c - 1 > 0.$$
$$\Rightarrow I_{\text{CON}} \text{ satisfies } P_3.$$

I_{Hill}: We need only to show that

$$\sum_{i=1}^{m} \left(\frac{n_i}{N}\right)^3 > \left(\frac{n_1 - c}{N}\right)^3 + \sum_{i=2}^{m+c} \left(\frac{1}{N}\right)^3$$

or

$$n_1^3 + \sum_{i=2}^{m} 1 > (n_1 - c)^3 + \sum_{i=2}^{m+c} 1$$
$$H \quad = n_1^3 + m + 1 - (n_1^3 - 3n_1^2 c + 3n_1 c^2 - c^3 + m + c - 1)$$
$$= 2 + 3n_1^2 c - 3n_1 c^2 + c^3 - c$$

$$= 2 + c^3 + 3n_1c(n_1 - c) - c > 0.$$
$$\Rightarrow I_{\text{Hill}} \text{ satisfies } P_3.$$

4.2.4. Principle P4. *Given a vector* $(n_1, \ldots, n_m)$ *and any permutation* $(i1, \ldots, im)$ *of* $(1, \ldots, m)$; $f(n_1, \ldots, n_m) = f(n_{i1}, \ldots, n_{im})$

All measures satisfy P4 since ranking order is not a factor.

4.2.5. Principle P5. *Given a vector* $(n_1, \ldots, n_m)$; $0 < c < n_j$ *and* $n_i \geq n_j$; *then* $f(n_1, \ldots, n_{i+c}, \ldots, n_j - c, \ldots, n_m) > f(n_1, \ldots, n_i, \ldots, n_j, \ldots, n_m)$.

I_{Rae}: We have to show that

$$\sum_{i=1}^{j-1} \frac{n_i(n_i-1)}{N(N-1)} + \frac{(n_j+c)(n_j+c-1)}{N(N-1)} + \sum_{i=j+1}^{k-1} \frac{n_i(n_i-1)}{N(N-1)} + \frac{(n_kc)(n_kc-1)}{N(N-1)} + \sum_{i=k+1}^{m} \frac{n_i(n_i-1)}{N(N-1)} > \sum_{i=1}^{j-1} \frac{n_i(n_i-1)}{N(N-1)} + \frac{n_j(n_j-1)}{N(N-1)} + \sum_{i=j+1}^{k-1} \frac{n_i(n_i-1)}{N(N-1)} + \frac{n_k(n_k-1)}{N(N-1)} + \sum_{i=k+1}^{m} \frac{n_i(n_i-1)}{N(N-1)},$$

which is equivalent to show the following:

$$(n_j + c)(n_j + c - 1) + (n_k - c)(n_k - c - 1) - n_j(n_j - 1) - n_k(n_k - 1) > 0$$
$$I = 2cn_j + 2c^2 - 2cn_k + n_j + n_k$$
$$= n_j + n_k + 2c(c + n_j - n_k) > 0.$$

Assuming that

$$n_1 \geq n_2 \geq \cdots \geq n_m,$$
$$\Rightarrow I_{\text{Rae}} \text{ satisfies } P_5.$$

I_{CON}: We need only to show that

$$\sum_{i=1}^{j-1} \left(\frac{n_i}{N}\right)^2 + \left(\frac{n_j+c}{N}\right)^2 + \sum_{i=j+1}^{k-1} \left(\frac{n_i}{N}\right)^2 + \left(\frac{n_k-c}{N}\right)^2 + \sum_{i=k+1}^{m} \left(\frac{n_i}{N}\right)^2 > \sum_{i=1}^{j-1} \left(\frac{n_i}{N}\right)^2 + \left(\frac{n_j}{N}\right)^2 + \sum_{i=j+1}^{k-1} \left(\frac{n_i}{N}\right)^2 + \left(\frac{n_k}{N}\right)^2 + \sum_{i=k+1}^{m} \left(\frac{n_i}{N}\right)^2$$
$$J = (n_j + c)^2 + (n_kc)^2 - n_j^2 - n_k^2$$
$$= 2c^2 + 2cn_j - 2cn_k > 0.$$
$$\Rightarrow I_{\text{CON}} \text{ satisfies } P_5.$$

I_{Hill}: We need only to show that

$$\sum_{i=1}^{j-1}\left(\frac{n_i}{N}\right)^3 + \left(\frac{n_j+c}{N}\right)^3 + \sum_{i=j+1}^{k-1}\left(\frac{n_i}{N}\right)^3 + \left(\frac{n_k-c}{N}\right)^3 + \sum_{i=k+1}^{m}\left(\frac{n_i}{N}\right)^3 > \sum_{i=1}^{j-1}\left(\frac{n_i}{N}\right)^3 + \left(\frac{n_j}{N}\right)^3 + \sum_{i=j+1}^{k-1}\left(\frac{n_i}{N}\right)^3 + \left(\frac{n_k}{N}\right)^3 + \sum_{i=k+1}^{m}\left(\frac{n_i}{N}\right)^3$$

$$\begin{aligned} K &= (n_j+c)^3 + (n_k-c)^3 - n_j^3 - n_k^3 \\ &= n_j^3 + 3n_j^2c + 3n_jc^2 + c^3 + n_k^3 - 3n_k^2c + 3n_kc^2 - c^3 - n_j^3 - n_k^3 \\ &= 3n_j^2c + 3n_jc^2 - 3n_k^2c + 3n_kc^2 > 0. \\ &\Rightarrow I_{\text{Hill}} \text{ satisfies } P_5. \end{aligned}$$

Given the results presented in this section, we conclude that our three measures satisfy the five principles and hence, they are acceptable for evaluating and ranking the interestingness of discovered knowledge. Adding these three measures to the five ones presented by Hilderman and Hamilton (2001), we end up with a set of 8 measures that are theoretically acceptable for such a task.

Table 7 summarizes the status of all interestingness measures mentioned with respect to the five principles, where a "×" means that a measure satisfies the respective principle. In Table 7, we added the properties of the three new measures as compared to the table presented by Hilderman and Hamilton (2001).

4.3. Theoretical correlation study

To check if there are possible correlations between future results generated by theoretically acceptable measures, we proceed by mathematical proofs. When there is not a linear transformation, we opt to study the tendency or the behaviour of the measure's functions.

First, it is noticed that the measure I_{Variance} is a linear transformation of I_{Simpson}:

$$\begin{aligned} I_{\text{Variance}} &= \frac{\sum_{i=1}^{m}(p_i-\overline{q})^2}{m-1} = \frac{\sum_{i=1}^{m}(p_i^2 - 2p_i\overline{q} + \overline{q}^2)}{m-1} \\ &= \frac{\sum_{i=1}^{m} p_i^2 - 2\overline{q}\sum_{i=1}^{m} p_i + n\overline{q}^2}{m-1} \\ &= \frac{\sum_{i=1}^{m} p_i^2 - 2\overline{q} + n\overline{q}^2}{m-1}, \end{aligned}$$

while

$$\sum_{i=1}^{m} p_i = 1$$

$$I_{\text{Variance}} = \frac{I_{\text{Simpson}} - 2\overline{q} + n\overline{q}^2}{m-1}.$$

Second, $I_{\text{Total}} = m \times I_{\text{Shannon}}$ which is a linear transformation of I_{Shannon}.

The following shows that there is a possible relation between the results generated by I_{Rae} and I_{McIntosh}:

$$\begin{aligned} I_{\text{Rae}} &= \frac{\sum_{i=1}^{m} n_i(n_i-1)}{N(N-1)} = \frac{\sum_{i=1}^{m} n_i^2 - N}{N^2-N} = -\frac{N-\sum_{i=1}^{m} n_i^2}{N^2-N} \\ &= -\frac{N-\sqrt{\sum_{i=1}^{m} n_i^2}+\sqrt{\sum_{i=1}^{m} n_i^2}-\sum_{i=1}^{m} n_i^2}{(N-\sqrt{N})(N+\sqrt{N})} \\ &= -\left(\frac{I_{\text{McIntosh}}}{N+\sqrt{N}} + \frac{\sqrt{\sum_{i=1}^{m} n_i^2}-\sum_{i=1}^{m} n_i^2}{N^2-N}\right). \end{aligned}$$

The measure I_{Rae} is a linear transformation of I_{Simpson} as demonstrated in the following:

$$\begin{aligned} I_{\text{Rae}} &= \frac{\sum_{i=1}^{m} n_i(n_i-1)}{N(N-1)} = \sum_{i=1}^{m} \frac{n_i}{N}\frac{n_i-1}{N-1} = \sum_{i=1}^{m} p_i \frac{n_i-1}{N-1} \\ &= \frac{N}{N-1}\sum_{i=1}^{m} p_i\left(\frac{n_i}{N}-\frac{1}{N}\right) = \frac{N}{N-1}\sum_{i=1}^{m} p_i\left(p_i-\frac{1}{N}\right) \\ &= \frac{N}{N-1}\sum_{i=1}^{m}\left(p_i^2-\frac{p_i}{N}\right) = \frac{N}{N-1}\sum_{i=1}^{m} p_i^2 - \frac{1}{N-1}\sum_{i=1}^{m} p_i \\ &= \frac{N}{N-1} I_{\text{Simpson}} - \frac{1}{N-1}. \end{aligned}$$

A possible correlation between results generated by I_{Shannon} and I_{Simpson} is established by the following:

$$I_{\text{Shannon}} = -\sum_{i=1}^{m} p_i \log_2 p_i = -\sum_{i=1}^{m} p_i \frac{\log p_i}{\log 2} = -\frac{1}{\log 2}\sum_{i=1}^{m} p_i \log p_i,$$

knowing that $p_i \log p_i$ has the same variation and behaviour of p_i^2.

The relation between I_{CON} and I_{Simpson} is given by the following:

$$I_{\text{CON}} = \sqrt{\frac{\left(\sum_{i=1}^{m} p_i^2\right)-\overline{q}}{1-\overline{q}}} = \sqrt{\frac{\left(I_{\text{Simpson}}\right)-\overline{q}}{1-\overline{q}}}.$$

From this section we conclude that theoretically, results generated by I_{Variance}, I_{Simpson}, I_{Shannon}, I_{McIntosh}, I_{Rae} and I_{Total} could be correlated and that I_{CON}, without considering the square root, is a linear transformation of I_{Simpson}. Finally, we were not able to relate I_{Hill} to any other measure.

In the following, we will check if these theoretical dependencies are apparent when the measures are applied to actual databases.

5. Experimental evaluation

The experimental evaluation of the interestingness measures is divided into three parts. The first one consists of their evaluation using a fictitious database. The second part consists of an application to three real databases and the last one consists of a study of similarities in the ranks assigned to summaries generated from the databases. Finally a classification of the eight measures is conducted.

5.1. Evaluation of the interestingness measures

By using the input data of Table 8 which is used previously by Hilderman and Hamilton (2001), we analyzed the statistical distribution of the indices values produced by the 19 measures. The input data used here consists of 16928 vectors which are the list of all possible ordered arrangements of a set of 50 objects among 10 classes. Vectors are ordered as follows: the first one presents the case of perfect concentration and the last, the case of uniform distribution.

Our purpose is to study the attitude of every measure and to check if its distribution is close to the Standard Normal Distribution (SND). The SND is taken as a point of reference in this step. This distribution plays a crucial role in a large body of statistics because it is very tractable analytically and its symmetry makes it an attractive choice for many population models. In addition to the following, the normal distribution can be used as an approximation of a large variety of distributions in large samples via the

Table 8. Summary of interestingness properties regarding the five principles.

Measure	P1	P2	P3	P4	P5
I_{Variance}	×	×	×	×	×
I_{Simpson}	×	×	×	×	×
I_{Shannon}	×	×	×	×	×
I_{McIntosh}	×	×	×	×	×
I_{Total}	×	×	×	×	×
I_{Rae}	×	×	×	×	×
I_{CON}	×	×	×	×	×
I_{Hill}	×	×	×	×	×
I_{Lorentz}	×	×			×
I_{Gini}	×	×		×	×
I_{Berger}	×	×	×	×	
I_{Shutz}	×	×		×	
I_{Bray}	×	×		×	
$I_{\text{Whittaker}}$	×	×		×	
$I_{\text{MacArthur}}$	×	×		×	×
I_{Theil}	×	×		×	
I_{Atkinson}	×	×		×	×
I_{Kullback}				×	

Table 9. Ordered arrangements of 50 objects among 10 classes.

50 Objects/ 10 classes
(41, 1, 1, 1, 1, 1, 1, 1, 1, 1)
(40, 2, 1, 1, 1, 1, 1, 1, 1, 1)
(39, 3, 1, 1, 1, 1, 1, 1, 1, 1)
. . .
. . .
. . .
(6, 6, 5, 5, 5, 5, 5, 5, 4, 4)
(6, 5, 5, 5, 5, 5, 5, 5, 5, 4)
(5, 5, 5, 5, 5, 5, 5, 5, 5, 5)

Central Limit Theorem. Hence, the minimum value, the maximum value, the number of tuples less than, or greater than, the middle and the skewness and kurtosis coefficients are determined.

The skewness coefficient measures the symmetry of a distribution. A zero value means that the distribution is symmetric and a positive (negative) value means that it is clustered more to the left (right). The kurtosis measures the peakedness or flatness of a distribution. A zero value means that the distribution has a SND peak and a positive (negative) value means that the distribution has a sharper (flatter) peak than the SND.

For example, the minimum value of I_{Rae} is 0.08163, the maximum value is 0.66938, there are 16761 tuples less than middle, and 167 tuples greater than middle. The skewness coefficient is equal to 1.844 meaning that the distribution is asymmetric and clustered more to the left. The kurtosis coefficient is equal to 5.570, meaning that the distribution has a sharper peak than the SND. For I_{CON}, the minimum value is 0, the maximum value is 0.8, there are 14784 tuples less than middle, and 2144 tuples greater than middle. The skewness coefficient is equal to 0.716 which means that the distribution is asymmetric and it is clustered more to the left, but it is less than I_{Rae}. The kurtosis coefficient is equal to 0.883 meaning that the distribution has a sharper peak than the SND and a flatter peak than I_{Rae}.

The summary of the first step of the experimental evaluation is given in Table 9 (where I_{Max} is not calculated because it is a constant in this case) and in Table 10. These two tables include a calculation of the values concerning the three not previously considered measures and a re–calculation of the existing ones. The top eight measures in these two tables are the ones which satisfy the five principles.

Histograms of the distributions of index values generated by our three measures for the vectors described in Table 8 are presented in figures 2, 3 and 4, where the horizontal axes describe the intervals of the values generated and the vertical axes describe the number of vectors in each interval.

5.2. Application to real databases

The second step of this experimental evaluation consists of generating summaries from three databases and evaluating their interestingness by the eight selected measures. The

Table 10. Distribution characteristics of the interestingness measures.

Measure	Min	Max	Middle	<Middle	>Middle
I_{Variance}	0	0.06400	0.03200	16760	168
I_{Simpson}	0.10000	0.67600	0.38800	16760	168
I_{Shannon}	1.25066	3.32193	2.28630	613	16315
I_{McIntosh}	0.20710	0.79640	0.50175	509	16419
I_{Total}	12.50660	33.21930	22.86295	613	16315
I_{Rae}	0.08163	0.66938	0.37551	16761	167
I_{CON}	0	0.80000	0.40000	14784	2144
I_{Hill}	0.99920	0.99999	0.99959	1720	15208
I_{Lorentz}	0.55000	0.91000	0.73000	4703	12225
I_{Gini}	0	0.72000	0.36000	4671	12256
I_{Berger}	0.10000	0.82000	0.46000	15836	1092
I_{Schutz}	0	0.72000	0.36000	9996	6932
I_{Bray}	0.52000	1.00000	0.76000	4209	12719
$I_{\text{Whittaker}}$	0.28000	1.00000	0.64000	7441	9487
$I_{\text{MacArthur}}$	0	0.42084	0.21042	15683	1245
I_{Theil}	0	2.14143	1.07072	5549	11379
I_{Atkinson}	0	0.71006	0.35503	11432	5496
I_{Kullback}	1.25066	3.32193	2.28630	613	16315
I_{Max}	–	–	–	–	–

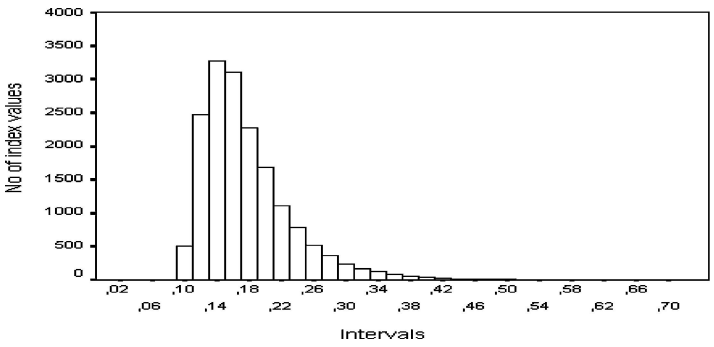

Figure 2. Histogram of I_{Rae}.

first one is a research awards database (D1), which is available in the public domain from the Natural Sciences and Engineering Research Council of Canada (NSERC). It consists of a set of tables giving information about the distribution and the amounts of the awards granted to researchers in 67 universities in Canada. This database is used in previous data mining research such as the work presented by Carter and Hamilton (1995a, 1995b). The second one is a heart disease diagnosis database (D2) from the UCI ML. It consists of 270 observations and gives information about the attributes that can influence the fact of being affected. The third is Hayes Roth database (D3) from the

Table 11. Kurtosis and skewness of the interestingness measures.

Measure	Kurtosis	Skewness
I_{Variance}	5.570	1.844
I_{Simpson}	5.570	1.844
I_{Shannon}	1.357	–0.958
I_{McIntosh}	2.316	–1.243
I_{Total}	1.357	–0.957
I_{Rae}	5.570	1.844
I_{CON}	0.883	0.716
I_{Hill}	0.360	–0.791
I_{Lorentz}	–0.233	–0.144
I_{Gini}	–0.233	–0.144
I_{Berger}	1.139	0.976
I_{Schutz}	–0.131	0.132
I_{Bray}	0.145	–0.345
$I_{\text{Whittaker}}$	–0.131	–0.132
$I_{\text{MacArthur}}$	0.485	0.684
I_{Theil}	–0.236	–0.056
I_{Atkinson}	–0.422	0.166
I_{Kullback}	1.357	–0.957
I_{Max}	–	–

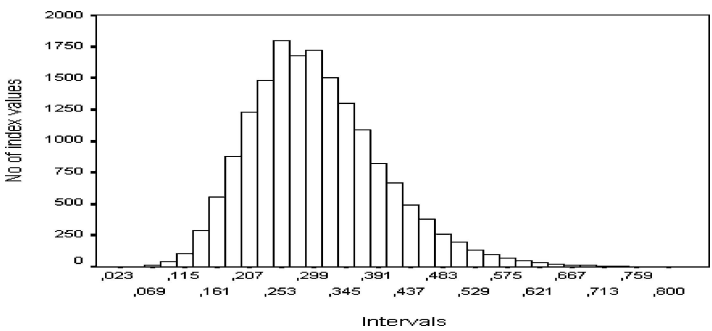

Figure 3. Histogram of I_{CON}.

UCI ML which consists of 132 persons and gives information about the attributes that influence the classification of persons into classes.

Results of the eight measures are presented in Tables 12 , 13 and 14 , where real values are assigned to the summaries generated from the initial database according to the set of indices used as heuristic measures of interestingness. The purpose is to rank these summaries by interestingness and then, select the most interesting ones.

The *S* column contains the summary numbers (summaries are ordered increasingly from the one which has the fewest number of tuples to the summary with the greatest

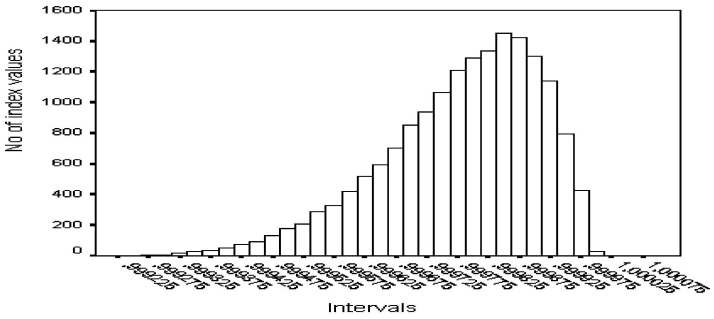

Figure 4. Histogram of I_{Hill}.

number of tuples). The *No of T* column describes the number of tuples in each summary, the remainder of the columns give the interestingness of each summary according to the set of measures described above.

For example, for the I_{Rae} measure and with D1, summaries are numbered from 1 to 12. The number of tuples in summary 1 is 4, summary 2 contains 5 tuples, and so on until summary 12 which has 67 tuples. For this measure, summary 2 which has 5 tuples, has an I_{Rae} value equal to 0.5807, which is the highest value, so it is the most interesting, summary 4 with an I_{Rae} value equal to 0.3989 is the second highest, summary 12 with an I_{Rae} equal to 0 is the least interesting one. According to these values, ranks are assigned to the 12 summaries. Therefore, summary 2 is ranked first, summary 4 is second, etc. When the same analysis is conducted for the I_{CON} measure, summary 2 is ranked first (most interesting) and summary 12 is the twelfth (least interesting). Tables 15–17 give the ranks assigned to each summary. In the case of equal interestingness, the same rank is assigned. According to the ranks presented, we notice that there are measures that give the same order of summaries, which means, there are similarities in the results and in the ranks deduced. This result caused us to study these similarities.

Table 12. Empirical results for D1.

S	No of T	I_{Variance}	I_{Simpson}	I_{Shannon}	I_{McIntosh}	I_{Total}	I_{Rae}	I_{CON}	I_{Hill}
1	4	0.0015	0.2546	1.9861	0.5643	7.9443	0.2433	0.0785	–2.8973
2	5	0.0967	0.5870	1.1855	0.2664	5.9274	0.5807	0.6955	–0.5427
3	8	0.0107	0.1998	2.5445	0.6299	20.3556	0.1877	0.2924	–3.5796
4	9	0.0371	0.4079	1.9272	0.4116	17.3444	0.3989	0.5778	–1.0728
5	10	0.0099	0.1891	7373	0.6438	27.3728	0.1768	0.3147	–3.6764
6	12	0.0063	0.1522	3.0055	0.6948	36.0658	0.1393	0.2740	–5.0049
7	18	0.0035	0.1147	3.5467	0.7533	63.8414	0.1013	0.2503	–6.3450
8	21	0.0029	0.1054	3.7154	0.7694	78.0231	0.0918	0.2462	–7.1205
9	21	0.0031	0.1094	3.7073	0.7624	77.8526	0.0959	0.2547	–6.4707
10	26	0.0017	0.0806	4.0570	0.8157	105.4820	0.0669	0.2094	–8.9532
11	29	0.0015	0.0777	4.2499	0.8215	123.2460	0.0638	0.2117	–9.1126
12	67	0	0.0149	6.0661	1.0000	406.4280	0	0	–66.0000

Table 13. Empirical results for D2.

S	No of T	$I_{Simpson}$	$I_{Variance}$	$I_{Shannon}$	$I_{McIntosh}$	I_{Total}	I_{Rae}	I_{CON}	I_{Hill}
1	2	0,0632	0.5632	0.9068	0.2657	1.8136	0.5616	0.5616	–0.7030
2	2	0.3951	0.8951	0.3095	0.0574	0.6191	0.8947	0.5616	–0.7030
3	3	0.0751	0.4835	1.1548	0.3244	3.4643	–1.0087	0.5616	–0.7030
4	3	0.1001	0.5336	1.1205	0.2870	3.3614	0.5318	0.5616	–0.7030
5	4	0.0841	0.5022	1.2157	0.3102	4.8628	0.5003	0.5616	–0.8646
6	5	0.0288	0.3150	1.7931	0.4672	8.9657	0.3125	0.5616	–2.0336
7	5	0.0573	0.4290	1.4577	0.3674	7.2886	0.4269	0.5616	–1.2130
8	6	0.0225	0.2792	2.0502	0.5021s	12.3010	0.2766	0.3675	–2.2876
9	6	0.0292	0.3126	2.0098	0.4695	12.0590	0.3100	0.4184	–1.7888
10	8	0.0197	0.2628	2.2573	0.5190	18.0581	0.2600	0.3968	–2.3868
11	9	0.0090	0.1832	2.6793	0.6090	24.1135	0.1802	0.2848	–3.8925
12	10	0.0163	0.2469	2.3524	0.5357	23.5244	0.2441	0.4041	–2.6383
13	11	0.0065	0.1554	2.8905	0.6450	31.7957	0.1523	0.2664	–5.0293
14	16	0.0063	0.1568	3.1339	0.6432	50.1430	0.1536	0.3171	–4.3116
15	20	0.0023	0.0920	3.6887	0.7418	73.7729	0.0887	0.2103	–8.5338

Table 14. Empirical results for D3.

S	No of T	$I_{Simpson}$	$I_{Simpson}$	$I_{Shannon}$	$I_{McIntosh}$	I_{Total}	I_{Rae}	I_{CON}	I_{Hill}
1	2	0.1327	0.6327	0.7990	0.2241	1.5981	0.6299	0.5152	–0.4923
2	4	0.0227	0.3182	1.7880	0.4775	7.1520	0.3130	0.3015	–1.9758
3	6	0.0091	0.2124	2.3753	0.5906	14.2517	0.2063	0.2341	–3.4447
4	8	0.0119	0.2082	2.5525	0.5955	20.4198	0.2022	0.3084	–3.2820
5	12	0.0022	0.1076	3.3577	0.7361	40.2924	0.1007	0.1625	–7.7885
6	16	0.0040	0.1220	3.4264	0.7127	54.8219	0.1153	0.2520	–5.9773
7	23	0.0013	0.0711	4.0804	0.8034	93.8486	0.0640	0.1698	–11.6671
8	28	0.0017	0.0825	4.0938	0.7807	114.6270	0.0755	0.2203	–9.1923
9	39	0.0004	0.0421	4.8755	0.8705	190.1440	0.0348	0.1301	–19.5354
10	51	0.0007	0.0541	4.7794	0.8407	243.7520	0.0468	0.1875	–14.9488
11	57	0.0002	0.0289	5.4314	0.9090	309.5920	0.0215	0.1076	–28.9620
12	69	5.84E-07	0.0200	0.8506	0.9405	403.6880	0.0125	0.0746	–44.6432

5.3. Similarity study

The last step in the experimental evaluation is to study similarities deduced from ranks assigned to generated summaries. Ranking similarities between measures was conducted by Hilderman, Hamilton and Barber (1999a) using the Gamma correlation coefficient. In this paper, the Spearman's rank correlation coefficient is used. Computed correlations are given in Tables 18–20 and for databases D1, D2 and D3, respectively.

Table 15. Ranks assigned for D1.

I_{Variance}	I_{Simpson}	I_{Shannon}	I_{McIntosh}	I_{Total}	I_{Rae}	I_{CON}	I_{Hill}
11	3	3	3	2	3	11	3
1	1	1	1	1	1	1	1
3	4	4	4	4	4	4	4
2	2	2	2	3	2	2	2
4	5	5	5	5	5	3	5
5	6	6	6	6	6	5	6
6	7	7	7	7	7	7	7
8	9	9	9	9	9	8	9
7	8	8	8	8	8	6	8
9	10	10	10	10	10	10	10
10	11	11	11	11	11	9	11
12	12	12	12	12	12	12	12

Table 16. Ranks assigned for D2.

I_{Variance}	I_{Simpson}	I_{Shannon}	I_{McIntosh}	I_{Total}	I_{Rae}	I_{CON}	I_{Hill}
5	2	2	2	2	2	11	2
1	1	1	1	1	1	1	1
4	5	4	5	4	5	5	5
2	3	3	3	3	3	3	
3	4	5	4	5	4	2	4
8	7	7	7	7	7	9	9
6	6	6	6	6	6	4	6
9	9	9	9	9	9	10	8
7	8	8	8	8	8	6	7
10	10	10	10	10	10	8	10
12	12	12	12	12	12	13	12
11	11	11	11	11	11	7	11
13	14	13	14	13	14	14	14
14	13	14	13	14	13	12	13
15	15	15	15	15	15	15	15

A correlation equal to 1 means a total similarity in the ranks assigned. A threshold of 95% is used to declare two ranks as similar or highly correlated. Correlations are used to classify the measures into homogeneous classes. In a homogeneous class of measures, all measures give the same order or are highly correlated with minimum intra–classes differences and maximum inter–classes differences.

According to our empirical study, summary 2 for D1 and D2 and summary 1 for D3 are preferred by all measures. For the three considered databases, I_{Hill} produced highly correlated results with many other measures which is in contradiction with the

Table 17. Ranks assigned for D3.

I_{Variance}	I_{Simpson}	I_{Shannon}	I_{McIntosh}	I_{Total}	I_{Rae}	I_{CON}	I_{Hill}
1	1	1	1	1	1	1	1
2	2	2	2	2	2	3	2
4	3	3	3	3	3	5	4
3	4	4	4	4	4	2	3
6	6	5	6	5	6	9	6
5	5	6	5	6	5	4	5
8	8	7	8	7	8	8	8
7	7	8	7	8	7	6	7
10	10	10	10	9	10	10	10
9	9	9	9	10	9	7	9
11	11	11	11	11	11	11	11
12	12	12	12	12	12	12	12

Table 18. Spearman's correlations for D1.

Measure	I_{Variance}	I_{Simpson}	I_{Shannon}	I_{McIntosh}	I_{Total}	I_{Rae}	I_{CON}	I_{Hill}
I_{Variance}	1.000	0.748	0.748	0.748	0.685	0.748	0.979	0.748
I_{Simpson}	0.748	1.000	1.000	1.000	0.993	1.000	0.727	1.000
I_{Shannon}	0.748	1.000	1.000	1.000	0.993	1.000	0.727	1.000
I_{McIntosh}	0.748	1.000	1.000	1.000	0.993	1.000	0.727	1.000
I_{Total}	0.685	0.993	0.993	0.993	1.000	0.993	0.664	0.993
I_{Rae}	0.748	1.000	1.000	1.000	0.993	1.000	0.727	1.000
I_{CON}	0.979	0.727	0.727	0.727	0.664	0.727	1.000	0.727
I_{Hill}	0.748	1.000	1.000	1.000	0.993	1.000	0.727	1.000

theoretical assumptions. This contradiction can be due to the nature of data used in this study.

D1 produces the following two classes:

- **Class 1:** $\{I_{\text{Variance}}, I_{\text{CON}}\}$,
- **Class 2:** $\{I_{\text{Simpson}}, I_{\text{Shannon}}, I_{\text{McIntosh}}, I_{\text{Rae}}, I_{\text{Hill}}, I_{\text{Total}}\}$.

D2 and D3 produce the following classes:

- **Class 1:** $\{I_{\text{CON}}\}$,
- **Class 2:** $\{I_{\text{Variance}}, I_{\text{Simpson}}, I_{\text{Shannon}}, I_{\text{McIntosh}}, I_{\text{Rae}}, I_{\text{Hill}}, I_{\text{Total}}\}$.

The purpose of this classification is to select one representative measure from each class. The criteria used are the skewness and kurtosis coefficients, as determined in Section 5.1. Hence, the best measure in a class is the one which is less skewed and

Table 19. Spearman's correlations for D2.

Measure	I_{Variance}	I_{Simpson}	I_{Shannon}	I_{McIntosh}	I_{Total}	I_{Rae}	I_{CON}	I_{Hill}
I_{Variance}	1.000	0.971	0.971	0.971	0.971	0.971	0.871	0.971
I_{Simpson}	0.971	1.000	0.993	1.000	0.993	1.000	0.786	0.989
I_{Shannon}	0.971	0.993	1.000	0.993	1.000	0.993	0.768	0.982
I_{McIntosh}	0.971	1.000	0.993	1.000	0.993	1.000	0.786	0.989
I_{Total}	0.971	0.993	1.000	0.993	1.000	0.993	0.768	0.982
I_{Rae}	0.971	1.000	0.993	1.000	0.993	1.000	0.786	0.989
I_{CON}	0.871	0.786	0.768	0.786	0.768	0.786	1.000	0.793
I_{Hill}	0.971	0.989	0.982	0.989	0.982	0.989	0.793	1.000

Table 20. Spearman's correlations for D3.

Measure	I_{Variance}	I_{Simpson}	I_{Shannon}	I_{McIntosh}	I_{Total}	I_{Rae}	I_{CON}	I_{Hill}
I_{Variance}	1.000	0.993	0.979	0.993	0.972	0.993	0.937	1.000
I_{Simpson}	0.993	1.000	0.986	1.000	0.979	1.000	0.916	0.993
I_{Shannon}	0.979	0.986	1.000	0.986	0.993	0.986	0.867	0.979
I_{McIntosh}	0.993	1.000	0.986	1.000	0.979	1.000	0.916	0.993
I_{Total}	0.972	0.979	0.993	0.979	1.000	0.979	0.846	0.972
I_{Rae}	0.993	1.000	0.986	1.000	0.979	1.000	0.916	0.993
I_{CON}	0.937	0.916	0.867	0.916	0.846	0.916	1.000	0.937
I_{Hill}	1.000	0.993	0.979	0.993	0.972	0.993	0.937	1.000

whose peak is the closest to the peak of the SND. From class 1, I_{CON} is chosen because it has skewness and kurtosis coefficient values close to 0 in the three databases. From class 2, I_{Hill} is chosen for the same reason.

6. Conclusion

Based on three real databases and an experimental evaluation, a classification of interestingness measures into two classes was deduced. This classification reduces the number of measures needed to evaluate databases such as D1, D2 and D3, to only two measures by taking a representative one from each class. This is in spite of the contradictions encountered between the theoretical and the empirical checking of correlations. The choice of the representative measure from each class is based on the evaluation of skewness and kurtosis coefficients. The advantage of this reduction is to allow the user to take under consideration only the ranks deduced from the two representative measures. Using only two measures eases the decision making and the interpretation of generated results.

Our empirical study, based on only three databases, which are of relatively small size, could be extended by using large databases. Moreover, future research is needed to suggest new enhanced interestingness measures.

Acknowledgments

The authors would like to acknowledge the editor and three anonymous referees whose suggestions and comments improved the content of this paper.

Notes

1. Adapted from the figure given by Hilderman, Hamilton and Cercone (1999b)
2. For this example, we assume that a higher value produced by the heuristic measure means higher interestingness which is not always the case.
3. The same notation is used by Hilderman and Hamilton (1999).

References

Carter, C. L., & Hamilton, H. J. (1995a). Fast, incremental generalization and regeneralization for knowledge discovery from large databases. In *Proceedings of the Eighth Florida Artificial Intelligence Symposium*. (pp. 319–323), Melbourne, Florida.

Carter, C. L., & Hamilton, H. J. (1995b). Performance evaluation of attribute-oriented algorithms for knowledge discovery from databases. In *Proceedings of the Seventh IEEE International Conference on Tools with Artificial Intelligence (ICTAI'95)*. (pp. 486–489), Washington, D.C.

Egghe, L., & Rousseau, R. (1991). Transfer principles and a classification of concentration measures. *Journal of the American Society for Information Science (JASIS), 42:7*, 479–489.

Han, J., & Kamber, M. (2001). *data mining: Concepts and techniques*. Morgan Kaufmann Publishers.

Hilderman, R. J., & Hamilton, H. J. (1999). Heuristic measures of interestingness. In *Proceedings of the Third European Conference on the Principles of Data Mining and Knowledge Discovery (PKDD'99)*. (pp. 232–241), Prague, Czech Republic.

Hilderman, R. J., & Hamilton, H. J. (2000). Principles for mining summaries using objective measures of interestingness. In *Proceedings of the Twelfth IEEE International Conference on Tools with Artificial Intelligence (ICTAI'00)*. (pp. 72–81), Vancouver, Canada.

Hilderman, R. J., & Hamilton, H. J. (2001). Evaluation of interestingness measures for ranking discovered knowledge. *Lecture Notes in Computer Sciences, 2035*, 247–259.

Hilderman, R. J., Hamilton, H. J., & Barber, B. (1999a). Ranking the interestingness of summaries from data mining systems. In *Proceedings of the 12th International Florida Artificial Intelligence Research Symposium (FLAIRS'99)*. (pp. 100–106), Orlando, U.S.A..

Hilderman, R. J., Hamilton, H. J., & Cercone, N. (1999b). Data mining in large databases using domain generalization graphs. *Journal of Intelligent Information Systems, 13:3*, 195–234.

Hill, M. O. (1973). Diversity and evenness: A unifying notation and its consequences. *Ecology, 54*, 427–432.

Rae, D. W., & Taylor, M. (1970). *The Analysis of Political Cleavages*. New Haven: Yale University Press.

Silberschatz, A., & Tuzhilin, A. (1995). On objective measures of interestingness in knowledge discovery. In *Proceedings of The First International Conference on Knowledge Discovery and Data Mining (KDD'95)*. (pp. 275–281), Montreal, Canada.

Silberschatz, A., & Tuzhilin, A. (1996). What makes patterns interesting in knowledge discovery. *IEEE Transactions on Knowledge and Data Engineering, Special Issue on Data Mining, 5:6*, 970–974.

Received August 11, 2003
Final Revision June 25, 2005
Accepted August 30, 2005