Evaluation of Interestingness Measures for Ranking Discovered Knowledge

Robert J. Hilderman[1] and Howard J. Hamilton[2]

[1]Saskatchewan Population Health and Evaluation Research Unit
[1,2]Department of Computer Science
University of Regina
Regina, Saskatchewan, Canada S4S 0A2
{Robert.Hilderman,Howard.Hamilton}@uregina.ca

Abstract. When mining a large database, the number of patterns discovered can easily exceed the capabilities of a human user to identify interesting results. To address this problem, various techniques have been suggested to reduce and/or order the patterns prior to presenting them to the user. In this paper, our focus is on ranking summaries generated from a single dataset, where attributes can be generalized in many different ways and to many levels of granularity according to taxonomic hierarchies. We theoretically and empirically evaluate thirteen diversity measures used as heuristic measures of interestingness for ranking summaries generated from databases. The thirteen diversity measures have previously been utilized in various disciplines, such as information theory, statistics, ecology, and economics. We describe five principles that any measure must satisfy to be considered useful for ranking summaries. Theoretical results show that only four of the thirteen diversity measures satisfy all of the principles. We then analyze the distribution of the index values generated by each of the thirteen diversity measures. Empirical results, obtained using synthetic data, show that the distribution of index values generated tend to be highly skewed about the mean, median, and middle index values. The objective of this work is to gain some insight into the behaviour that can be expected from each of the measures in practice.

1 Introduction

When mining a large database, the number of patterns discovered can easily exceed the capabilities of a human user to identify interesting results. To address this problem, various techniques have been suggested to reduce and/or order the patterns prior to presenting them to the user. For example, in [3], it is shown that the most interesting rules may reside along a support/confidence border. A technique is described in [20] that discovers interesting rules via an interactive process that seeks to classify rules that are not interesting. In [8], a measure is described that determines the interestingness (called surprise there) of discovered knowledge via the explicit detection of Simpson's Paradox. An approach is described in [7] that utilizes a distance metric to evaluate the importance of a rule by considering its unexpectedness in terms of other rules in its neighborhood.

Our focus is on the use of diversity measures for ranking summaries generated from a single dataset, where attributes can be generalized in many different ways and to many levels of granularity according to taxonomic hierarchies. We introduced this use of diversity measures in [10] and [11]. An empirical analysis found that highly ranked, concise summaries provided a reasonable starting point for further analysis of discovered knowledge. It was also shown that for selected sample datasets, the order in which some of the measures rank summaries is highly correlated, but the rank ordering can vary substantially when different measures are used. In [12], the notion of a summary was extended to include other forms of knowledge representation, and we showed that these other forms are also amenable to ranking using diversity measures. And significant progress has been made into more theoretical issues regarding formal principles for diversity measures used as measures of interestingness in data mining applications [14].

In this paper, we evaluate thirteen diversity measures as heuristic measures of interestingness for ranking summaries in data mining applications. We describe five principles that any measure must satisfy to be considered useful for ranking summaries. Our theoretical results show that only four of the thirteen diversity measures satisfy all of the principles. We then analyze the distribution of the index values generated by each of the thirteen diversity measures. Empirical results, obtained using synthetic data, show that the distribution of index values generated tend to be highly skewed about the mean, median, and middle index values. The objective of this work is to gain some insight into the behaviour that can be expected from each of the measures in practice so that when choosing a candidate interestingness measure, we can determine which of the five principles are satisfied, and then knowing the behavioural characteristics of each measure, judge the suitability of the candidate interestingness measure for the intended application.

The remainder of the paper is organized as follows. In Section 2, we describe several forms of knowledge representation, which we collectively refer to as summaries, and motivate the need for ranking discovered knowledge. In Section 3, we provide a brief overview of thirteen diversity measures introduced and evaluated as heuristic measures of interestingness in previous work. In Section 4, we describe five principles that useful diversity measures must satisfy, and identify those diversity measures satisfying the five principles. In Section 5, we present experimental results describing the distribution of index values generated by each of the thirteen measures. We conclude in Section 6 with a summary of our work and suggestions for future research.

2 Background and Motivation

Let a *summary* S be a relation defined on the columns $\{(A_1, D_1), (A_2, D_2), \ldots, (A_n, D_n)\}$, where each (A_i, D_i) is an attribute-domain pair. Also, let $\{(A_1, v_{i1}), (A_2, v_{i2}), \ldots, (A_n, v_{in})\}$, $i = 1, 2, \ldots, m$, be a set of m unique tuples, where each (A_j, v_{ij}) is an attribute-value pair and each v_{ij} is a value from the domain D_j associated with attribute A_j. Let attribute A_k be a derived attribute whose values v_{ik}, from the domain D_k, for each attribute-value pair (A_k, v_{ik}) is an aggregation of values from the the unconditioned data present in the original database. For example, a sample summary is shown in Table 1. Table 1 is a generalized relation in which retail sales transactions have been aggregated to show the derived attributes *Quantity*, *Amount*, and *Count* (i.e., number of transactions) by *Region*.

Table 1. A generalized relation

Region	*Quantity*	*Amount*	*Count*
North	12	$150.00	7
South	5	$325.00	2
West	8	$200.00	4
East	11	$275.00	3

The summary definition given above can also be naturally extended to include summaries that are multi-dimensional. For example, another sample summary, is shown in Figure 1. Figure 1 shows a data cube in which retail sales transactions have been aggregated in three dimensions, where the *Item* attribute is on the vertical dimension, *Transact.Loc* is on the horizontal, and *Cust.Loc* is on the diagonal. *Transact.Loc* is the city where the sales transaction was processed, and *Cust.Loc* is the city where the sales transaction was initiated. Here we show each cell containing two values (due to space limitations); the top value is the quantity of items aggregated from sales transactions (i.e., *Quantity*), and the bottom value is the number of transactions aggregated (i.e., *Count*).

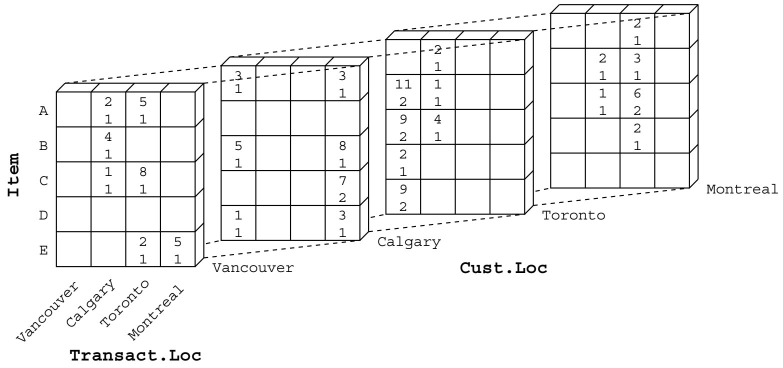

Figure 1. A data cube

Of course, numerous methods could be used to guide the generation of summaries, such as concept hierarchies [5], domain generalization graphs [15], Galois lattices [9], conceptual graphs [4], and formal concept analysis [22]. Also, summaries could more generally include many other forms of knowledge representation, such as database views, association rules, itemsets, and web search results.

However, when given hundreds, or even thousands of summaries (possibly multi-dimensional), it is simply not feasible to determine the most interesting summaries or dimensions using a manual technique. What is needed are effective measures of interestingness to assist in the interpretation and evaluation of the discovered knowledge. The development of such measures is currently an active research area in KDD. Such measures are broadly classified as either objective or subjective. *Objective measures* are based upon the structure of discovered patterns, such as the frequency with which combinations of items appear in sales transactions [1]. *Subjective measures* are based upon user beliefs or biases regarding relationships in the data, such as an approach utilizing Bayes Rule to revise prior beliefs [18]. Here we focus on objective measures of interestingness.

3 Objective Interestingness Measures

The tuples in a summary or dimension generated from a database are unique, and therefore, can be considered to be a population with a structure that can be described by some frequency or probability distribution. Here, we review thirteen diversity measures, described in detail in [10], and shown in Figure 2, that evaluate the frequency or probability distribution of the values in a derived attribute to assign a single real-valued index that represents its interestingness relative to other summaries or dimensions generated from the same database.

In Figure 2, let m be the total number of tuples in a summary. Let n_i be the value contained in the derived attribute for tuple t_i. Let $N = \sum_{i=1}^{m} n_i$ be the total of the derived attribute. Let p be the actual probability distribution of the tuples based upon the values n_i. Let $p_i = n_i/N$ be the actual probability for tuple t_i. Let q be a uniform probability distribution of the tuples. Let $\bar{u} = N/m$ be the value for tuple t_i, $i = 1, 2, \ldots, m$ according to the uniform distribution q. Let $\bar{q} = 1/m$ be the probability for tuple t_i, for all $i = 1, 2, \ldots, m$ according to the uniform distribution q. Let r be the probability distribution obtained by combining the values n_i and $\bar{u}$. Let $r_i = (n_i + \bar{u})/2N$, be the probability for tuples t_i, for all $i = 1, 2, \ldots, m$ according to the distribution r.

The measures shown in Figure 2 are well-known measures of dispersion, dominance, inequality, and concentration that have previously been successfully applied in several areas of the social, ecological, information, and computer sciences. Although the terminology varies depending upon the application, the concept of diversity has been considered a useful one for analyzing many phenomena. For example, in ecology, various measures of diversity have been proposed and studied to aid in understanding the variability of populations of organisms within different types of habitat [17]. Diversity measures have also been used by economists and social scientists to study the distribution of income between different socioeconomic groups and geographical regions [2]. In information theory, diversity

$$I_{Variance} = \frac{\sum_{i=1}^{m} (p_i - \bar{q})^2}{m-1}$$

$$I_{Simpson} = \sum_{i=1}^{m} p_i^2$$

$$I_{Shannon} = -\sum_{i=1}^{m} p_i \log_2 p_i$$

$$I_{McIntosh} = \frac{N - \sqrt{\sum_{i=1}^{m} n_i^2}}{N - \sqrt{N}}$$

$$I_{Lorenz} = \bar{q} \sum_{i=1}^{m} (m - i + 1) p_i$$

$$I_{Gini} = 0.5 \left(\sum_{i=1}^{m} \sum_{j=1}^{m} |p_i \bar{q} - p_j \bar{q}| \right)$$

$$I_{Berger} = \max(p_i)$$

$$I_{Schutz} = \frac{\sum_{i=1}^{m} |p_i - \bar{q}|}{2m\bar{q}}$$

$$I_{Bray} = \frac{\sum_{i=1}^{m} \min(n_i, \bar{u})}{N}$$

$$I_{Whittaker} = 1 - \left(0.5 \sum_{i=1}^{m} |p_i - \bar{q}| \right)$$

$$I_{MacArthur} = \left(-\sum_{i=1}^{m} r_i \log_2 r_i \right) - \left(0.5 \left(-\sum_{i=1}^{m} p_i \log_2 p_i \right) + \log_2 m \right)$$

$$I_{Theil} = \frac{\sum_{i=1}^{m} |p_i \log_2 p_i - \bar{q} \log_2 \bar{q}|}{m\bar{q}}$$

$$I_{Atkinson} = 1 - \left(\prod_{i=1}^{m} \frac{p_i}{\bar{q}} \right)^{\bar{q}}$$

Figure 2. Thirteen diversity measures

measures are used to measure the information content in messages [21]. Diversity measures have been used to describe the linguistic differences between the inhabitants of neighboring geographic regions [16]. More general treatments attempt to define the concept of diversity and develop a related theory of diversity measurement [19, 23].

4 Theoretical Results

We now describe principles of interestingness against which the utility of candidate interestingness measures can be assessed. We do this through the mathematical formulation of five principles that must be satisfied by any acceptable diversity measure for ranking the interestingness of discovered knowledge using our, or a similar, technique. Proofs are omitted due to space considerations, so refer to [13] and [14] for complete details. We study functions f of m variables, $f(n_1, \ldots, n_m)$, where f denotes a general measure of diversity, m and each n_i (n_i assumed to be non-zero) are as defined in the previous section, and $(n_1, \ldots, n_m)$ is a vector corresponding to the values in a derived numeric measure attribute (e.g., the *Count* values from the examples in Section 2)for some arbitrary summary whose values are arranged in descending order such that $n_1 \geq \ldots \geq n_m$ (except for discussions regarding I_{Lorenz}, which requires that the values be arranged in ascending order). The principles presented here are for ranking the interestingness of summaries generated from a single dataset, so we assume that N is fixed. We justify the non-zero assumption for the n_i's, as follows. If the value of the *Count* attribute for a particular tuple is zero, there are two possible reasons. Either the combination of domain values being counted in the tuple can occur in practice, but no occurrences have been encountered during the mining process, or else the combination of domain values being summarized cannot occur in practice, and no occurrences will ever be encountered (i.e., such an entity does not exist). So, to preserve and simplify the general applicability of our technique, we make no assumptions regarding the possibility of occurrence of particular combinations of domain values. We now begin by specifying two fundamental principles.

Minimum Value Principle (P1). Given a vector $(n_1, \ldots, n_m)$, where $n_i = n_j$, $i \neq j$, for all i, j, $f(n_1, \ldots, n_m)$ attains its minimum value.

P1 specifies that the minimum interestingness should be attained when the tuple counts are all equal (i.e., uniformly distributed). For example, given the vectors $(2, 2)$, $(50, 50, 50)$, and $(1000, 1000, 1000, 1000)$, we require that the index value generated by f be the minimum possible for the respective values of m and N.

Maximum Value Principle (P2). Given a vector $(n_1, \ldots, n_m)$, where $n_1 = N-m+1$, $n_i = 1$, $i = 2, \ldots, m$, and $N > m$, $f(n_1, \ldots, n_m)$ attains its maximum value.

P2 specifies that the maximum interestingness should be attained when the tuple counts are distributed as unevenly as possible. For example, given the vectors $(3, 1)$, $(148, 1, 1)$, and $(3997, 1, 1, 1)$, where $m = 2, 3,$ and 5, respectively, and $N = 4, 150,$ and 4000, respectively, we require that the index value generated by f be the maximum possible for the respective values of m and N.

The behaviour of a measure relative to satisfying both *P1* and *P2* is significant because it reveals an important characteristic about its fundamental nature as a measure of diversity. A measure of diversity can generally be considered either a *measure of concentration* or a *measure of dispersion*. A measure of concentration can be viewed as the opposite of a measure of dispersion, and we can convert one to the other via simple transformations. For example, if g corresponds to a measure of dispersion, then we can convert it to a measure of concentration f, where $f = \max(g) - g$. Here we only consider measures of concentration. A measure was considered to be a measure of concentration if it satisfied P1 and P2 without transformation. A measure was considered to be a measure of dispersion if it satisfied P1 and P2 following transformation. All measures of dispersion were transformed into measures of concentration prior to our analysis.

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

P3 specifies that a summary containing m tuples, whose counts are distributed as unevenly as possible, will be more interesting than a summary containing $m + c$ tuples, whose counts are also distributed as unevenly as possible. For example, given the vectors $(999, 1)$ and $(997, 1, 1, 1)$ (i.e., $c = 2$), we require that $f(999, 1) > f(997, 1, 1, 1)$.

Permutation Invariance Principle (P4). Given a vector $(n_1, \ldots, n_m)$ and any permutation $(i_1, \ldots, i_m)$ of $(1, \ldots, m)$, $f(n_1, \ldots, n_m) = f(n_{i_1}, \ldots, n_{i_m})$.

P4 specifies that every permutation of a given distribution of tuple counts should be equally interesting. That is, interestingness is not a labeled property, it is only determined by the distribution of the counts. For example, given the vector $(2, 4, 6)$, we require that $f(2, 4, 6) = f(4, 2, 6) = f(4, 6, 2) = f(2, 6, 4) = f(6, 2, 4) = f(6, 4, 2)$.

Transfer Principle (P5). Given a vector $(n_1, \ldots, n_m)$ and $0 < c < n_j$, $f(n_1, \ldots, n_i + c, \ldots, n_j - c, \ldots, n_m) > f(n_1, \ldots, n_i, \ldots, n_j, \ldots, n_m)$.

P5, adapted from [6], specifies that when a strictly positive transfer is made from the count of one tuple to another tuple whose count is greater, then interestingness increases. For example, given the vectors $(10, 7, 5, 4)$ and $(10, 9, 5, 2)$, we require that $f(10, 9, 5, 2) > f(10, 7, 5, 4)$.

Those measures satisfying the above principles of interestingness are shown in Table 2. In Table 2, the *P1* to *P5* columns describe the five principles, and a measure that satisfies a principle is indicated by the *bullet* symbol (i.e., •).

Table 2. Measures satisfying the five principles

Measure	*P1*	*P2*	*P3*	*P4*	*P5*
$I_{Variance}$	•	•	•	•	•
$I_{Simpson}$	•	•	•	•	•
$I_{Shannon}$	•	•	•	•	•
$I_{McIntosh}$	•	•	•	•	•
I_{Lorenz}	•	•			•
I_{Gini}	•	•		•	•
I_{Berger}	•	•	•	•	
I_{Schutz}	•	•		•	
I_{Bray}	•	•		•	
$I_{Whittaker}$	•	•		•	
$I_{MacArthur}$	•	•		•	•
I_{Theil}	•			•	
$I_{Atkinson}$	•	•		•	•

5 Experimental Results

We now analyze the distribution of the index values generated by each of the thirteen measures. Input data consists of two populations of vectors shown in

Table 3, where index values for 16,928 vectors (i.e., all possible ordered arrangements of a population of 50 objects among 10 classes) and 2,611 vectors (i.e., all possible ordered arrangements of a population of 50 objects among 5 classes) were generated. The choice of vectors to evaluate here was made somewhat arbitrarily, but it does provide a large, controlled population of index values in which a gradual change in evenness occurs from the most highly skewed distribution in the first vector, to the uniform distribution in the last vector.

Table 3. Ordered arrangements of two populations

50 objects / 10 classes	*50 objects / 5 classes*
(41, 1, 1, 1, 1, 1, 1, 1, 1, 1)	(46, 1, 1, 1, 1)
(40, 2, 1, 1, 1, 1, 1, 1, 1, 1)	(45, 2, 1, 1, 1)
(39, 3, 1, 1, 1, 1, 1, 1, 1, 1)	(44, 3, 1, 1, 1)
⋮	⋮
(6, 6, 5, 5, 5, 5, 5, 5, 4, 4)	(11, 11, 10, 10, 8)
(6, 5, 5, 5, 5, 5, 5, 5, 5, 4)	(11, 10, 10, 10, 9)
(5, 5, 5, 5, 5, 5, 5, 5, 5, 5)	(10, 10, 10, 10, 10)

Histograms of the absolute frequencies of the index values for the vectors in Table 3 were generated for each measure. Again, due to space limitations, we cannot show all of these histograms. However, sample histograms of the index values generated for the population of 50 objects among 10 classes by $I_{Variance}$ and I_{Schutz} are shown in Figures 3 and 4, respectively. In Figures 3 and 4, the horizontal and vertical axes describe intervals for the index values generated and the number of index values that fall in each interval, respectively. For example, the histogram for $I_{Variance}$ shows that 68 index values were generated on the interval $(0.000, 0.0009]$, 1,106 on $(0.0009, 0.003]$, 2,464 on $(0.003, 0.005]$, 3,006 on $(0.005, 0.007]$, 2,581 on $(0.007, 0.008]$, 2,055 on $(0.008, 0.010]$, 1,549 on $(0.010, 0.012]$, and 4,099 on the remaining intervals in $(0.012, 0.065]$. A curve describing the standard normal distribution (SND) of the index values is superimposed over the observed frequencies.

To provide a summary description of each histogram, we can use the skewness and kurtosis for the distribution of index values. *Skewness* is a measure of the symmetry of a distribution. It has a value of zero when the distribution is a symmetrical curve (i.e., as in a SND). If the skewness is different from zero, then the distribution is asymmetrical. A positive (negative) value indicates the index values are clustered more to the left (right) of the mean, with most of the extreme index values to the right (left) of the mean. In general, for positive (negative) skewness, we have mode $\leq$ median $\leq$ mean (mean $\leq$ median $\leq$ mode). *Kurtosis* is a measure of the relative peakedness of a distribution and indicates the extent to which outliers cause the distribution to differ from the SND. When a distribution follows the SND, it has value of zero. When the value is greater

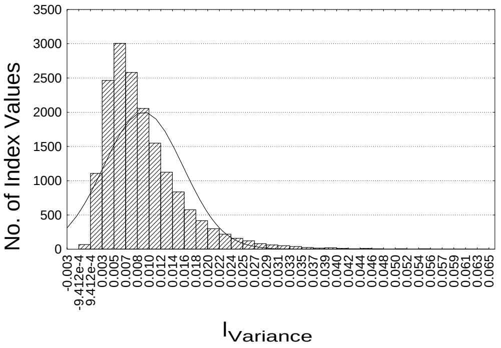

Figure 3. Histogram for $I_{Variance}$

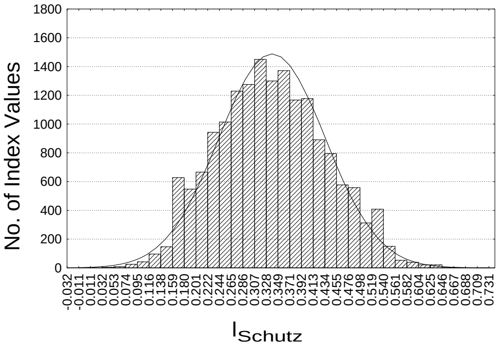

Figure 4. Histogram for I_{Schutz}

than (less than) zero, the distribution has a sharper (flatter) peak than the SND and is more (less) prone to containing outliers.

The skewness and kurtosis for all measures are shown in Table 4. In Table 4, mnemonics are provided as an aid to interpreting the curves described by the values. The skewness mnemonics describe the symmetry of the frequency distribution in relation to the mean (i.e., AL = asymmetrical left, AR = asymmetrical right, NS = near symmetrical, and S = symmetrical) and the kurtosis mnemonics describe the relative peakedness of the frequency distribution in relation to the SND (i.e., SP = sharp peaked, NSN = near standard normal, MP = more peaked, and LP = less peaked). For example, the histogram for $I_{Variance}$, shown in Figure 3, has a skewness and kurtosis of approximately 1.8 and 5.6, respectively. This means that the distribution of index values is asymmetrical to the left of the mean (i.e., AL) and more sharply peaked than the SND (i.e., SP). Similarly, in the histogram for I_{Schutz}, shown in Figure 5, the distribution of index values is near symmetrical (i.e., NS) and less peaked than the SND (i.e., LP). The other measures in Table 4 can also be interpreted similarly.

Table 4. Skewness and kurtosis of the index values for the two populations

	50 objects / 10 classes				*50 objects / 5 classes*			
Measure	*Skewness*		*Kurtosis*		*Skewness*		*Kurtosis*	
$I_{Variance}$	1.84421	AL	5.571732	SP	1.55959	AL	3.273237	SP
$I_{Simpson}$	1.84421	AL	5.571732	SP	1.55959	AL	3.273237	SP
$I_{Shannon}$	-0.95761	AR	1.357844	MP	-1.03452	AR	1.391038	MP
$I_{McIntosh}$	-1.24351	AR	2.317341	SP	-1.13072	AR	1.496420	SP
I_{Lorenz}	0.14435	S	-0.232495	NSN	0.02128	S	-0.317871	NSN
I_{Gini}	-0.14435	S	-0.232495	NSN	-0.02128	S	-0.317871	NSN
I_{Berger}	0.97607	AL	1.139526	SP	0.75039	AL	0.264196	SP
I_{Schutz}	0.13192	NS	-0.130277	LP	0.27521	NS	-0.076436	LP
I_{Bray}	-0.13192	NS	-0.130277	LP	-0.27521	NS	-0.076436	LP
$I_{Whittaker}$	-0.13192	NS	-0.130277	LP	-0.27521	NS	-0.076436	LP
$I_{MacArthur}$	0.68369	AL	0.485805	MP	0.86586	AL	0.883313	MP
I_{Theil}	-0.05563	S	-0.236451	NSN	0.68371	AL	1.112360	MP
$I_{Atkinson}$	0.16650	NS	-0.422023	LP	0.30949	AL	-0.476633	LP

We now determine the number of index values generated by each measure that are less than and greater than the middle index value (i.e., $(minimum + maximum)/2$), and less than and greater than the median (i.e., the value for which 50% of the generated index values lie below and 50% lie above). Our belief is that a useful measure of interestingness should generate index values that are reasonably distributed throughout the range of possible values (such as in a SND). Again, we analyze the index values generated from the two populations shown in Table 3, with the results shown in Tables 5 and 6. In Tables 5 and 6, the *Minimum* and *Maximum* columns describe the minimum and maximum index values generated by each measure, respectively, the *Middle* column describes the middle index value, the < *Middle* and > *Middle* columns describe the number of index values less than and greater than the middle index value, respectively, and the *Median* column describes the median index value. For example, for the $I_{Variance}$ measure, the minimum and maximum index values are 0.0 and 0.064, respectively, the middle index value is 0.032, 16,761 (167) index values lie below (above) the middle index value, and the median index value is 0.00791. The distribution of index values in Tables 5 and 6 is highly skewed about the middle and median values for most of the measures. Isolated exceptions include I_{Bray} and $I_{Whittaker}$ in Table 5, and I_{Lorenz} and I_{Gini} in Table 6.

6 Conclusion and Future Research

The use of diversity measures for ranking the interestingness of summaries generated from databases is a new application area. Here we theoretically and experimentally analyzed thirteen diversity measures. Five principles of interestingness for useful diversity measures were described. Theoretical results showed that only four of the thirteen diversity measures satisfied all five principles. Experimental results showed that the distribution of index values, in relation to the mean, is least skewed for I_{Lorenz}, I_{Gini}, I_{Schutz}, I_{Bray}, and $I_{Whittaker}$, but these

Table 5. Distribution of index values for 50 objects among 10 classes

Measure	*Minimum*	*Maximum*	*Middle*	< *Middle*	> *Middle*	*Median*
$I_{Variance}$	0.0	0.064	0.032	16761	167	0.007911
$I_{Simpson}$	0.1	0.676	0.388	16761	167	0.1712
$I_{Shannon}$	1.250664	3.321928	2.286295	613	16315	2.860161
$I_{McIntosh}$	0.207096	0.7964	0.50175	509	16419	0.682799
I_{Lorenz}	0.214	0.55	0.37	12353	4575	0.338
I_{Gini}	0.107	0.275	0.185	4786	12142	0.169
I_{Berger}	0.14	0.82	0.46	15836	1092	0.28
I_{Schutz}	0.0	0.72	0.36	10751	6177	0.34
I_{Bray}	0.28	1.0	0.64	7549	9379	0.66
$I_{Whittaker}$	0.28	1.0	0.64	7549	9379	0.66
$I_{MacArthur}$	0.0	0.420842	0.21042	15683	1245	0.114606
I_{Theil}	0.0	2.141432	1.07072	5550	11378	1.21593
$I_{Atkinson}$	0.0	0.71	0.35503	11432	5496	0.296977

Table 6. Distribution of index values for 50 objects among 5 classes

Measure	*Minimum*	*Maximum*	*Middle*	< *Middle*	> *Middle*	*Median*
$I_{Variance}$	0.0	0.162	0.081	2507	104	0.0258
$I_{Simpson}$	0.2	0.848	0.524	2507	104	0.3032
$I_{Shannon}$	0.562179	2.321928	1.44205	164	2447	1.940238
$I_{McIntosh}$	0.092165	0.643839	0.36800	200	2411	0.523381
I_{Lorenz}	0.24	0.6	0.42	1496	1115	0.412
I_{Gini}	0.12	0.300	0.21	1183	1428	0.0.206
I_{Berger}	0.2	0.92	0.56	2180	431	0.42
I_{Schutz}	0.0	0.72	0.36	1850	761	0.3
I_{Bray}	0.28	1.0	0.64	939	1672	0.7
$I_{Whittaker}$	0.28	1.0	0.64	939	1672	0.7
$I_{MacArthur}$	0.0	0.427524	0.213765	2425	186	0.099571
I_{Theil}	0.0	1.759749	0.879875	2357	254	0.566115
$I_{Atkinson}$	0.0	0.784944	0.39247	1964	647	0.283374

measures are poorly behaved, containing a sharp peak, or multiple sharp peaks, in the frequency distribution of the index values. The remaining eight measures were skewed asymmetrically in relation to the mean, and more or less peaked than the SND. The experimental results also show that the distribution of the index values is highly skewed, in relation to the middle and median values, for most of the measures.

Future research will focus on extending the theory of interestingness for diversity measures used to rank summaries. New principles will be developed for ranking the interestingness of summaries generated from different sources (i.e., related, but physically, logically, or temporally independent databases).

References

1. R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proceedings of the 20th International Conference on Very Large Databases (VLDB'94)*, pages 487–499, Santiago, Chile, September 1994.

2. A.B. Atkinson. On the measurement of inequality. *Journal of Economic Theory*, 2:244–263, 1970.
3. R.J. Bayardo and R. Agrawal. Mining the most interesting rules. In *Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining (KDD'99)*, pages 145–154, San Diego, California, August 1999.
4. I. Bournaud and J.-G. Ganascia. Accounting for domain knowledge in the construction of a generalization space. In *Proceedings of the Third International Conference on Conceptual Structures*, pages 446–459. Springer-Verlag, August 1997.
5. C.L. Carter and H.J. Hamilton. Efficient attribute-oriented algorithms for knowledge discovery from large databases. *IEEE Transactions on Knowledge and Data Engineering*, 10(2):193–208, March/April 1998.
6. H. Dalton. The measurement of the inequality of incomes. *Economic Journal*, 30:348–361, 1920.
7. G. Dong and J. Li. Interestingness of discovered association rules in terms of neighborhood-based unexpectedness. In X. Wu, R. Kotagiri, and K. Korb, editors, *Proceedings of the Second Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'98)*, pages 72–86, Melbourne, Australia, April 1998.
8. A.A. Freitas. On objective measures of rule surprisingness. In J. Zytkow and M. Quafafou, editors, *Proceedings of the Second European Conference on the Principles of Data Mining and Knowledge Discovery (PKDD'98)*, pages 1–9, Nantes, France, September 1998.
9. R. Godin, R. Missaoui, and H. Alaoui. Incremental concept formation algorithms based on galois (concept) lattices. *Computational Intelligence*, 11(2):246–267, 1995.
10. R.J. Hilderman and H.J. Hamilton. Heuristic measures of interestingness. In J. Zytkow and J. Rauch, editors, *Proceedings of the Third European Conference on the Principles of Data Mining and Knowledge Discovery (PKDD'99)*, pages 232–241, Prague, Czech Republic, September 1999.
11. R.J. Hilderman and H.J. Hamilton. Heuristics for ranking the interestingness of discovered knowledge. In N. Zhong and L. Zhou, editors, *Proceedings of the Third Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'99)*, pages 204–209, Beijing, China, April 1999.
12. R.J. Hilderman and H.J. Hamilton. Applying objective interestingness measures in data mining systems. In *Proceedings of the 4th European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD'00)*, pages 432–439, Lyon, France, September 2000.
13. R.J. Hilderman and H.J. Hamilton. Principles for mining summaries: Theorems and proofs. Technical Report CS 00-01, Department of Computer Science, University of Regina, February 2000. Online at http://www.cs.uregina.ca/research/Techreport/0001.ps.
14. R.J. Hilderman and H.J. Hamilton. Principles for mining summaries using objective measures of interestingness. In *Proceedings of the Twelfth IEEE International Conference on Tools with Artificial Intelligence (ICTAI'00)*, pages 72–81, Vancouver, Canada, November 2000.
15. R.J. Hilderman, H.J. Hamilton, and N. Cercone. Data mining in large databases using domain generalization graphs. *Journal of Intelligent Information Systems*, 13(3):195–234, November 1999.
16. S. Lieberson. An extension of Greenberg's linguistic diversity measures. *Language*, 40:526–531, 1964.
17. A.E. Magurran. *Ecological diversity and its measurement*. Princeton University Press, 1988.
18. B. Padmanabhan and A. Tuzhilin. A belief-driven method for discovering unexpected patterns. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD'98)*, pages 94–100, New York, New York, August 1998.
19. G.P. Patil and C. Taillie. Diversity as a concept and its measurement. *Journal of the American Statistical Association*, 77(379):548–567, 1982.
20. S. Sahar. Interestingness via what is not interesting. In *Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining (KDD'99)*, pages 332–336, San Diego, California, August 1999.
21. C.E. Shannon and W. Weaver. *The mathematical theory of communication*. University of Illinois Press, 1949.
22. G. Stumme, R. Wille, and U. Wille. Conceptual knowledge discovery in databases using formal concept analysis methods. In J. Zytkow and M. Quafafou, editors, *Proceedings of the Second European Conference on the Principles of Data Mining and Knowledge Discovery (PKDD'98)*, pages 450–458, Nantes, France, September 1998.
23. M.L. Weitzman. On diversity. *The Quarterly Journal of Economics*, pages 363–405, May 1992.