Mining Interesting Correlated Contrast Sets

Mondelle Simeon, Robert J. Hilderman, and Howard J. Hamilton

Department of Computer Science
University of Regina
Regina, Saskatchewan, Canada S4S 0A2
{simeon2m, hilder, hamilton}@cs.uregina.ca

Abstract. Contrast set mining has been developed as a data mining task which aims at discerning differences across groups. These groups can be patients, organizations, molecules, and even time-lines. A valid correlated contrast set is a conjunction of attribute-value pairs that are highly correlated with each other and differ significantly in their distribution across groups. Although the search for valid correlated contrast sets produces a comparatively smaller set of results than the search for valid contrast sets, these results must still be further filtered in order to be examined by a domain expert and have decisions enacted from them. In this paper, we apply the minimum support ratio threshold which measures the ratio of maximum to minimum support across groups. We propose a contrast set mining technique which utilizes the minimum support ratio threshold to discover maximal valid correlated contrast sets. We also demonstrate how four probability-based objective measures developed for association rules can be used to rank contrast sets. Our experiments on real datasets demonstrate the efficiency and effectiveness of our approach.

1 Introduction

Discovering the differences between groups is a fundamental problem in many disciplines. Groups are defined by a selected property that distinguish one group from the other. The search for group differences can be applied to a wide variety of objects such as patients, organizations, molecules, and even time-lines. The group differences sought are novel, implying that they are not obvious or intuitive, potentially useful, implying that they can aid in decision-making, and understandable, implying that they are presented in a format easily understood by human beings. It has previously been demonstrated that contrast set mining is an effective method for mining group differences from observational multivariate data [1] [2] [3] [4] [5] [6].

Existing contrast set mining techniques can produce a prohibitively large set of differences across groups with varying levels of interestingness [2] [5]. For example, if we have a dataset with five attributes, each with two possible values, our search for differences would cause us to examine all 226 possible combinations of attribute values. On larger datasets, the search space becomes inordinately large, and the search becomes prohibitively expensive producing a large number

of results which ultimately must be analyzed and evaluated by a domain expert. In our previous work we demonstrated that utilizing mutual information and all confidence to select only the attributes and attribute-values that are highly correlated creates a smaller search space and a smaller number of "interesting" contrast sets [5]. Additionally, we demonstrated that using the minimum support ratio threshold, which measures the ratio of maximum to minimum support across groups, also discovers "more interesting" contrast sets during the search process [6]. In this paper, we combine our findings and propose a contrast set mining technique which utilizes the minimum support ratio threshold to discover contrast sets whose attributes and attribute-values are highly correlated.

The remainder of this paper is organized as follows. In Section 2, we briefly review related work. In Section 3, we describe the correlated contrast set mining problem. In Section 4, we examine four probability-based measures that have been developed for association rules and demonstrate their use for ranking contrast sets. In Section 5, we introduce our algorithm for mining correlated contrast sets that meet our minimum support ratio threshold. In Section 6, we present a summary of experimental results from a series of mining tasks. In Section 7, we conclude and suggest areas for future work.

2 Related Work

The STUCCO (Search and Testing for Understandable Consistent Contrasts) algorithm [1] is the original technique for mining contrast sets. The objective of STUCCO is to find statistically significant contrast sets from grouped categorical data. It employs a modified Bonferroni statistic to limit Type I errors resulting from multiple hypothesis tests. In other work, STUCCO forms the basis for a method proposed to discover negative contrast sets [7] that can include negation of terms in the contrast set. The main difference is their use of Holm's sequential rejective method [8] for the independence test.

The CIGAR (Contrasting Grouped Association Rules) algorithm [2] is a contrast set mining technique that specifically identifies which pairs of groups are significantly different and whether the attributes in a contrast set are correlated. CIGAR utilizes the same general approach as STUCCO, however it focuses on controlling Type II errors through increasing the significance level for the significance tests and by not correcting for multiple comparisons. Like STUCCO, CIGAR is only applicable to discrete-valued data.

Contrast set mining has also been applied to continuous data. Early work focussed on the formal notion of a time series contrast set and an efficient algorithm was proposed to discover contrast sets in time series and multimedia data [9]. Another approach utilized a modified equal-width binning interval, where the approximate width of the intervals is provided as a parameter to the model [3]. The methodology used is similar to STUCCO except that the discretization step is added before enumerating the search space.

The COSINE (Contrast Set Exploration using Diffsets) [4], GENCCS (Generate Correlated Contrast Sets) [5], and GIVE (Generate Interesting Valid con-

trast sEts) [6] algorithms are contrast set mining techniques that incorporate a vertical data format, a back tracking search algorithm, and simple discretization in mining maximal contrast sets from both discrete and continuous-valued attributes. GENCCS extends COSINE by utilizing mutual information and all-confidence to select the attribute-interval pairs that are most highly correlated. GIVE extends COSINE to discover maximal valid contrast sets which have a high ratio of maximum to minimum support across groups. The results demonstrate that both COSINE and GENCCS are more efficient than STUCCO and CIGAR, even at very low minimum support difference thresholds, and GIVE produces more interesting contrast sets than COSINE and GENCCS.

3 Problem Definition

Let $\mathcal{A} = \{a_1, a_2, \ldots, a_n\}$ be a set of n distinct attributes. We use $\mathcal{Q}$ and $\mathcal{C}$ to denote the set of *quantitative* attributes and the set of *categorical* attributes respectively. Let $\mathcal{V}(a_k)$ be the domain of values for a_k. An *attribute-interval pair*, denoted as $a_k : [v_{kl}, v_{kr}]$, is an attribute a_k associated with an interval $[v_{kl}, v_{kr}]$, where $a_k \in \mathcal{A}$ and $v_{kl}, v_{kr} \in \mathcal{V}(a_k)$. Further, if $a_k \in \mathcal{C}$ then $v_{kl} = v_{kr}$. Similarly, if $a_k \in \mathcal{Q}$, then $v_{kl} \leq v_{kr}$. Let $\mathcal{T} = \{x_1, x_2, \ldots, x_p\}$ where $x_k \in \mathcal{V}(a_k)$, $1 \leq k \leq p$, be a *transaction*. Let $\mathcal{D}$ be a set of transactions, called the *database*. Let $\{i_1, i_2, \ldots, i_m\}$ be a set of m distinct values from the set $\{1, 2, \ldots, n\}$, $m < n$. Let $\mathcal{F} = \{a_{i_1}, a_{i_2}, \ldots, a_{i_m}\}$, $a_{i_k} \in \mathcal{A}$, be a set of m distinct class attributes.

Let $G = \{a_1 : [v_{1l}, v_{1r}], \ldots, a_m : [v_{ml}, v_{mr}]\},\ a_k \in \mathcal{F}, 1 \leq k \leq m,\ a_i \neq a_j,\ \forall i, j$, be a set of m distinct class attribute-interval pairs, called a *group*. Let $X = \{a_1 : [v_{1l}, v_{1r}], \ldots, a_q : [v_{ql}, v_{qr}]\},\ a_i, a_j \in \mathcal{A} - \mathcal{F},\ 1 \leq k \leq q \leq n - m,\ a_i \neq a_j,\ \forall i, j$, be a set of distinct attribute-interval pairs, called a *quantitative contrast set*. Henceforth, we refer to a quantitative contrast set as simply a *contrast set*. A contrast set, X, is called k-specific, if $|X| = k$. The *support* of a contrast set, X, *in a database,* $\mathcal{D}$, denoted as $supp(X)$, is the percentage of transactions in $\mathcal{D}$ containing X. The support of a contrast set, X, in a group, G, denoted as $supp(X, G)$, is the percentage of transactions in $\mathcal{D}$ containing $X \cup G$.

A contrast set X associated with b mutually exclusive groups. $G_1, G_2, \ldots, G_b$ is called a *valid contrast set* if, and only if, the following four criteria are satisfied:

$$\exists ij\, supp(X, G_i) \neq supp(X, G_j), \tag{1}$$

$$\max_{ij} |supp(X, G_i) - supp(X, G_j)| \geq \epsilon, \tag{2}$$

$$supp(X) \geq \sigma, \tag{3}$$

and

$$\max_{i}^{n} \left\{ \frac{supp(Y, G_i)}{supp(X, G_i)} \right\} \geq \kappa, \tag{4}$$

where ϵ is called the *minimum support difference threshold*, σ is the *minimum frequency threshold*, κ called the *minimum subset support ratio threshold*, and

$Y \subset X$ with $|Y| = |X| + 1$. Criterion 1 ensures that the contrast set represents a true difference between the groups. Contrast sets that meet this criterion are called *significant*. Criterion 2 ensures the effect size. Contrast sets that meet this criterion are called *large*. Criterion 3 ensures that the contrast set occurs in a large enough number of transactions. Contrast sets that meet this criterion are called *frequent*. Criterion 4 ensures that the support of the contrast set in each group is different from that of its superset. Contrast sets that meet this criterion are called *specific*.

A valid contrast set is called *maximal* if it is not a subset of any other valid contrast set. A valid contrast set is called *interesting* if the ratio of maximum to minimum support across the groups is sufficiently large. Formally, for a valid contrast set, X, the ratio is given by

$$\lambda(X) = \begin{cases} \infty, & \text{if } \min_i^n \{supp(X, G_i)\} = 0 \\ \dfrac{\max_i^n \{supp(X, G_i)\}}{\min_i^n \{supp(X, G_i)\}}, & \text{otherwise.} \end{cases}$$

A large value for $\lambda(X)$ implies that X occurs in significantly fewer transactions in one group G_i than in some other group G_j. A value of ∞ indicates that X is absent from at least one group G_i and present in at least one other group G_j.

A valid contrast set is called a λ *contrast set* if, and only if,

$$\lambda(X) \geq \omega, \tag{5}$$

where ω is a user-defined *minimum support ratio threshold*. Criterion 5 ensures that the ratio of maximum and minimum support across all groups is sufficiently large. A λ contrast set is called a ∞ *contrast set* if, and only if,

$$\lambda(X) = \infty.$$

A valid contrast set is called *correlated*, if every attribute-interval pair that it contains is correlated with every other attribute-interval pair in the contrast set, where correlation is determined by mutual information [10] and all-confidence [11]. For example, given a 3-specific correlated contrast set $\{A_1 : [0,0], A_2 : [1,1], A_3 : [0,0]\}$, $A_1 : [0,0]$ is correlated with $A_3 : [1,1]$, $A_1 : [0,0]$ is correlated with $A_3 : [0,0]$, and $A_2 : [1,1]$ is correlated with $A_3 : [0,0]$.

If we have two 1-specific contrast sets $P : [v_{il}, v_{ir}]$ and $Q : [v_{jl}, v_{jr}]$, we can represent our knowledge of $P : [v_{il}, v_{ir}]$ and $Q : [v_{jl}, v_{jr}]$ in the contingency table shown in Table 1.

Then we can define the mutual information of $P : [v_{il}, v_{ir}]$ and $Q : [v_{jl}, v_{jr}]$, $I(P : [v_{il}, v_{ir}]; Q : [v_{jl}, v_{jr}])$, as follows:

$$I(P : [v_{il}, v_{ir}]; Q : [v_{jl}, v_{jr}]) = log_2 \left(\frac{A * N}{(A + B) \times (A + C)} \right).$$

Mutual information is a good measure of dependency between attributes, thus we must measure $I(P : [v_{il}, v_{ir}]; Q : [v_{jl}, v_{jr}])$, $\forall v_i \in \mathcal{V}(P)$ and $\forall v_j \in \mathcal{V}(Q)$.

Table 1: Contingency table for $P : [v_{il}, v_{ir}]$ and $Q : [v_{jl}, v_{jr}]$

	$Q : [v_{jl}, v_{jr}]$	$\neg Q : [v_{jl}, v_{jr}]$	
$P : [v_{il}, v_{ir}]$	A	B	$n(P : [v_{il}, v_{ir}])$
$\neg P : [v_{il}, v_{ir}]$	C	D	$n(\neg P : [v_{il}, v_{ir}])$
	$n(Q : [v_{jl}, v_{jr}])$	$n(\neg Q : [v_{jl}, v_{jr}])$	N

Thus, we can represent the mutual information of P and Q as follows:

$$I(P;Q) = \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{A}{N} \times I(P : [v_{il}, v_{ir}]; Q : [v_{jl}, v_{jr}])$$

where $n = |\mathcal{V}(P)|$ and $m = |\mathcal{V}(Q)|$.

For the two 1-specific contrast sets $P : [v_{il}, v_{ir}]$ and $Q : [v_{jl}, v_{jr}]$, we can define the all-confidence of $\{P : [v_{il}, v_{ir}] | \forall v_i \in \mathcal{V}(P)\}$ and $\{Q : [v_{jl}, v_{jr}] | \forall v_j \in \mathcal{V}(Q)\}$, $A_c(P : [v_{il}, v_{ir}]; Q : [v_{jl}, v_{jr}])$, as follows:

$$A_c(P : [v_{il}, v_{ir}]; Q : [v_{jl}, v_{jr}]) = \frac{A}{\max\{(A+B), (A+C)\}}.$$

Formally, a valid contrast set, $X = \{a_1 = [v_{1l}, v_{1r}], a_2 = [v_{2l}, v_{2r}], \ldots, a_n = [v_{nl}, v_{nr}]\}$, is a correlated contrast set if, and only if, the following two conditions are satisfied:

$$\forall a_i, a_j \in X, I(a_i; a_j) \geq \psi \tag{6}$$

$$\forall a_i, a_j \in X,\ \forall v_i \in \mathcal{V}(a_i),\ \forall v_j \in \mathcal{V}(a_j),\ A_c(a_i : [v_{il}, v_{ir}], a_j : [v_{jl}, v_{jr}]) \geq \xi \tag{7}$$

where ψ is a minimum mutual information threshold, and ξ is a minimum all-confidence threshold. Criterion 6 ensures that each attribute in a correlated contrast set tells a great amount of information about every other attribute. Criterion 7 ensures that each attribute-interval pair is highly correlated with every other attribute-interval pair.

Given a database $\mathcal{D}$, a minimum support difference threshold ϵ, a minimum frequency threshold σ, a minimum subset support ratio threshold κ, and a minimum support ratio threshold ω, a minimum mutual information threshold, ψ, and a minimum all-confidence threshold, ξ, our goal is to find all maximal λ correlated contrast sets (i.e., all maximal valid correlated contrast sets that satisfy Equations 5 and 3).

4 Ranking Methods

A contrast set mining process typically returns many valid contrast sets. Consequently, measures are needed to rank the relative interestingness of the valid

Table 2: Contingency table for $X \rightarrow G_i$

	G_i	$\neg G_i$	Σ *Row*
X	$n(X, G_i)$	$N(X, \neg G_i)$	$n(X)$
$\neg X$	$n(\neg X, G_i)$	$n(\neg X, \neg G_i)$	$n(\neg X)$
Σ *Column*	$n(G_i)$	$n(\neg G_i)$	N

contrast sets prior to presenting them to the end-user. In our previous work we demonstrated the suitability of five measures for ranking contrast sets: distribution difference, unusualness, coverage, lift, and interestingness factor for ranking contrast sets [5] [6].

Probability-based objective measures have been studied extensively as a means for evaluating the generality and reliability of association rules [12] [13] [14]. Contrast set mining is closely related to association rule mining and thus, we propose that these probability-based objective measures can be adapted for the task of ranking contrast sets. In this section, we propose five measures and demonstrate their use in ranking contrast sets.

Here we define the variables used in the ranking methods described in this section. A contrast set, X, is represented by a set of association rules, $X \rightarrow G_1, X \rightarrow G_2, \ldots, X \rightarrow G_n$, where $G_1, G_2, \ldots, G_n$ are unique groups. Let $n(X, G_i)$ be the number of instances of X in G_i. Let $n(X, \neg G_i)$ be the number of instances of X in groups other than G_i (that is, the number of times X occurs in $G_1, \ldots, G_{i-1}$, G_{i+1}, G_n). Let $(\neg X, G_i)$ be the number of instances of contrast sets other than X in G_i. Let $n(\neg X, \neg G_i)$ be the number of instances of contrast sets other than X in groups other than G_i. Let N be the total number of instances.

The values $n(X, G_i), n(X, \neg G_i), n(\neg X, G_i)$, and $n(\neg X, \neg G_i)$ actually correspond to the observed frequencies at the intersection of the rows and columns in a 2×2 contingency table for the association rule $X \rightarrow G_i$, such as the one shown in Table 2. Rows represent the occurrence of the contrast set and the columns represent occurrence of the groups.

Table 3a shows the number of instances for five valid contrasts sets represented by the rows V, W, X, Y, and Z in each of the groups represented by the columns labelled A, B, C, and D. For example, the number of instances of contrast set V in groups A, B, C, and D corresponding to association rules $V \rightarrow A$, $V \rightarrow B$, $V \rightarrow C$, and $V \rightarrow D$ is 312, 198, 36, and 30, respectively. A sample contingency table for the rule $V \rightarrow A$ is shown in Table 3b.

The values in Tables 3a and 3b are used throughout this section to illustrate how the various interestingness measures are calculated and how the contrast sets are ranked.

4.1 Leverage

The *leverage* (LEV) of an association rule $X \rightarrow G_i$ is the difference between the actual proportion of instances in which both X and G_i occur and the propor-

Table 3: Summary Data

(a) Results Summary for V, W, X, Y, and Z

Contrast Sets	A	B	C	D
V	312	198	36	30
W	62	10	23	13
X	104	66	12	10
Y	12	34	3	15
Z	36	108	18	30
Total	1210	384	69	65

(b) Contingency table for $V \rightarrow A$

	A	$\neg A$	Σ *Row*
V	312	264	576
$\neg V$	898	254	1152
Σ *Column*	1210	518	1728

tion that would be expected if X and G_i were statistically independent of each other [15] and is given by

$$\text{LEV}(X \rightarrow G_i) = p(G_i|X) - P(X)P(G_i) = \frac{n(X, G_i)}{n(X)} - \frac{n(X)}{N} \times \frac{n(G_i)}{N}.$$

The values for LEV occur on the interval [-1, 1]. Values near one indicate strong independence between X and G_i.

The leverage for a contrast set, X, is determined by the group G_i, for which the leverage is largest. Thus, the leverage of X is given by

$$\text{LEV}(X) = \max_i \text{LEV}(X \rightarrow G_i).$$

For example, the leverage of V in A is

$$\text{LEV}(V \rightarrow A) = \frac{312}{576} - \frac{576}{1728} \times \frac{1210}{1728} = 0.308.$$

4.2 Change of support

The *change of support* (CS) of of an association rule $X \rightarrow G_i$ measures the correlation between X and G_i [13] and is given by

$$\text{CS}(X \rightarrow G_i) = p(G_i|X) - p(G_i) = \frac{n(X, G_i)}{n(X)} - \frac{n(G_i)}{N}.$$

The values for change of support occur in the interval [-0.5, 1]. Values near 1 indicate strong correlation between X and G_i.

The CS for a contrast set, X, is determined by the group, G_i, for which change of support is largest, and is given by

$$\text{CS}(X) = \max_i \text{CS}(X \rightarrow G_i).$$

For example, the Change of Support of V in A is

$$\text{CS}(V \rightarrow A) = \frac{312}{576} - \frac{1210}{1728} = -0.159.$$

4.3 Yule's Q coefficient

The *Yule's Q coefficient* (YQ) of an association rule $X \rightarrow G_i$ measures the degree to which X and G_i are associated with each other [16] and is given by

$$\begin{aligned}\mathrm{YQ}(X \rightarrow G_i) &= \frac{p(X,G_i)p(\neg X \neg G_i) - p(X,\neg G_i)p(\neg X, G_i)}{p(X,G_i)p(\neg X \neg G_i) + p(X,\neg G_i)p(\neg X, G_i)} \\ &= \frac{n(X,G_i)n(\neg X \neg G_i) - n(X,\neg G_i)n(\neg X, G_i)}{n(X,G_i)n(\neg X \neg G_i) + n(X,\neg G_i)p(\neg X, G_i)}.\end{aligned}$$

The values for Yule's Q coefficient occur on the interval [-1,1]. Values near 1 indicate strong association between X and G_i. Values near -1 indicate complete dissociation between X and G_i. 0 indicates complete independence between X and G_i.

The Yule's Q coefficient for a contrast set, X, is determined by the group, G_i, for which Yule's Q coefficient is largest. Thus, the Yule's Q coefficient of X is given by

$$\mathrm{YQ}(X) = \max_i \mathrm{YQ}(X \rightarrow G_i).$$

For example, the Yule's Q coefficient of V in A is

$$\mathrm{YQ}\ (V \rightarrow A) = \frac{312 \times 254 - 264 \times 898}{312 \times 254 + 264 \times 898} = -0.498.$$

4.4 Conviction

The *conviction* (CONV) of an association rule $X \rightarrow G_i$ measures true implication between X and G_i [17] and is given by

$$\mathrm{CONV}(X \rightarrow G_i) = \frac{p(X)p(\neg G_i)}{p(X,\neg G_i)} = \frac{n(X) \times n(\neg G_i)}{N \times n(X,\neg G_i)}$$

The values for conviction occur in the interval $[1, \infty]$. A value equal to 1 indicates X and G_i are completely unrelated. Values further from 1 indicate that X and G_i are more strongly related.

The conviction of a contrast set for all groups is the sum of the individual conviction values for the contrast set in each group. Thus the conviction of a contrast set X is given by

$$\mathrm{CONV}(X) = \sum_i \mathrm{CONV}(X \rightarrow G_i).$$

For example, the conviction of V in A is

$$\mathrm{CONV}(V \rightarrow A) = \frac{576 \times 518}{1728 \times 264} = 0.654.$$

Table 4 shows how each of the contrast sets, V, W, X, Y, and Z were ranked by each of the six measures. Though Z was ranked highest most frequently, and V was ranked the lowest, most frequently, the order of the five contrast sets, was unique for each measure.

Table 4: Rankings

Contrast Set	LEV	COS	YQ	CONV
V	5	4	4	3
W	2	3	2	5
X	4	4	5	3
Y	3	2	1	2
Z	1	1	3	1

5 Mining Interesting Valid Correlated Contrast Sets

Given a database $\mathcal{D}$, a minimum support difference threshold ϵ, a minimum frequency threshold σ, a minimum subset support ratio threshold κ, and a minimum support ratio threshold ω, a minimum mutual information threshold, ψ, a minimum all-confidence threshold, ξ, and a ranking measure, m, the GENICCS(Generate Interesting Correlated Contrast Sets), algorithm presented in Algorithm 1 finds all the maximal λ correlated contrast sets in a given dataset (i.e, all the contrast sets that satisfy Equations 1, 2, 3, 4, 5, 6, and 7). It modifies the backtracking search technique proposed in [5] for mining correlated contrast sets.

Algorithm 1 GENICCS($\mathcal{D}, \mathcal{F}, \epsilon, \sigma, \kappa, \omega, \psi, \xi, m$)

Input: A dataset, $\mathcal{D}$, and parameters, $\epsilon, \sigma, \kappa, \omega, \psi, \xi$, and m
Output: The set of ranked valid lambda correlated contrast sets W

for each $a_i \in \mathcal{A},\ a_i \notin \mathcal{F}$ **do**
  if $a_i \in \mathcal{Q}$ **then**
    $\mathcal{V}(a_i)$ =DISCRETIZE(a_i)
  end if
  $B = B \cup \{a_i : [v_{il}, v_{ir}]\},\ \ \forall v_i \in \mathcal{V}(a_i)$
end for
for each $a_i : [v_{il}, v_{ir}] \in B$ **do**
  for each $a_j : [v_{jl}, v_{jr}] \in B\ a_j > a_i$ **do**
    if $I(a_i; a_j) \geq \psi$ **then**
      if $A_c(a_i : [v_{il}, v_{ir}], a_j : [v_{jl}, v_{jr}] \geq \xi$ **then**
        $R = R \cup \{a_i : [v_{il}, v_{ir}], a_j : [v_{jl}, v_{jr}]\}$
      end if
    end if
  end for
end for
$B = B \cup R$
C_0 =COMBINE($\{\}, B, \epsilon, \sigma, 0, \omega, W$)
TRAVERSE($\{\}, C_0, W, \epsilon, \sigma, \kappa, \omega, t$)
RANK(W, m)
return W

First, GENICCS determines all the 1-specific contrast sets from the domain $\mathcal{V}$ of each attribute in the dataset not occurring in $\mathcal{F}$ (lines 1 to 6). Quantitative attributes are discretized (line 3) to determine a $\mathcal{V}$ set from which 1-specific quantitative contrast sets can be generated. We use the discretization algorithm previously described in [4]. Each of the 1-specific contrast sets are added to B (line 5). Second, GENICCS combines the 1-specific contrast sets with each other in lexicographic order to form 2-specific contrast sets. Each 2-specific contrast set whose mutual information and all-confidence is at least ψ and ξ, respectively, are added to a list R (lines 9 to 13). The contrast sets in R are merged with those in B, and stored in B (line 16). Third, GENICCS calls the COMBINE subroutine to determine which of the contrast sets in B are valid (line 17). Fourth, GENICCS calls the MINE subroutine to combine the contrast sets in C_0 with each other to produce more specific contrast sets, determine which are valid, and store them in W (line 18). Fifth, GENICCS calls the RANK subroutine to rank the contrast sets in W using a ranking measure m (line 19). Finally, GENICCS returns the set of ranked correlated contrast sets in W (line 20).

5.1 COMBINE

Given a prefix P, a combine set H, a set of valid contrast sets W, a minimum support difference threshold ϵ, a minimum frequency threshold σ, a minimum subset support ratio threshold κ, and a minimum support ratio threshold ω, the COMBINE Algorithm, shown in Algorithm 2, builds new correlated contrast sets from P and H that satisfy Equations 1, 2, 3, 4 and 5.

COMBINE begins by combining the prefix P with each member y of the possible set of combine elements, H, to create a new contrast set z (line 3). It then determines the diffset D_z, frequency F_z, and potential combine set C_z, for z (line 4). COMBINE determines whether z satisfies Equations 1, 2, 3, and 4 (line 5). COMBINE also checks whether Equation 3 is satisfied (line 6). Any z which satisfies Equation 3 is potentially maximal and is added to W if it has no superset already in W (lines 6 to 9). Otherwise, COMBINE adds any z which satisfies Equation 5 to C (lines 10 to 12). COMBINE re-orders the contrast sets in C, in increasing cardinality of their potential combine sets, then in increasing value of their frequencies (line 15). Finally, it returns C (line 16).

5.2 TRAVERSE

Given a prefix P_l, a combine set C_l, a minimum support difference threshold ϵ, a minimum frequency threshold σ, a minimum subset support ratio threshold κ, and a minimum support ratio threshold ω, the TRAVERSE Algorithm, shown in Algorithm 3, traverses the search space for maximal λ correlated contrast sets that satisfy Equations 1, 2, 3, 4, and 5.

TRAVERSE begins by determining the next prefix, P_{l+1} and a new possible set of combine elements, H_{l+1}, for that prefix (lines 1 to 12). TRAVERSE prunes away the subtree of P_{l+1} if $P_{l+1} \cup H_{l+1}$ is subsumed by an existing contrast set (lines 13 to 15). Otherwise, P_{l+1} is extended. TRAVERSE then calls COMBINE

Algorithm 2 COMBINE$(P, H, W, \epsilon, \sigma, \kappa, \omega)$

$C = \emptyset$
for each $y \in H$ **do**
  $z = P \cup \{y\}$
  Determine D_z, F_z, C_z
  if significant(z) & large(z, ϵ) & frequent(z, σ) & specific(z, κ) **then**
    if $\lambda(z) == \infty$ **then**
      if $Z \not\supseteq z : Z \in W$ **then**
        $W = W \cup \{z\}$
      end if
    else if $\lambda(z) \geq \omega$ **then**
      $C = C \cup \{z\}$
    end if
  end if
end for
Sort C in increasing $|C_z|$ then in increasing F_z
return C

Algorithm 3 TRAVERSE$(P_l, C_l, W_l, \epsilon, \sigma, \kappa, \omega)$

for each $x \in C_l$ **do**
  $P_{l+1} = \{x\}$
  $H_{l+1} = \emptyset$
  Let $P'_{l+1} = P_{l+1} - P_l$
  for each $y \in C_l$ **do**
    if $y > P_{l+1}$ **then**
      Let $y' = y - P_l$
      if $y' \neq P'_{l+1}$ & $y' \in C_{P_l}$ **then**
        $H_{l+1} = H_{l+1} \cup \{y\}$
      end if
    end if
  end for
  if $Z \supseteq P_{l+1} \cup H_{l+1} : Z \in W_l$ **then**
    return
  end if
  $C_{l+1} =$ COMBINE$(P_{l+1}, H_{l+1}, W_l, \epsilon, \sigma, \kappa, \omega)$
  if $C_{l+1} \neq \emptyset$ **then**
    if $Z \not\supseteq P_{l+1} : Z \in W_l$ **then**
      $W_l = W_l \cup P_{l+1}$
    end if
  else
    $W_{l+1} = \{W \in W_l : x \in W\}$
  end if
  if $C_{l+1} \neq \emptyset$ **then**
    TRAVERSE$(P_{l+1}, C_{l+1}, W_{l+1}, \epsilon, \sigma, \kappa, \omega)$
  end if
  $W_l = W_l \cup W_{l+1}$
end for

to create a new combine set for the next level of the search (line 16). P_{l+1} is added to W_l, if C_{l+1} is empty and W_l does not contain any supersets of P_{l+1} (lines 17 to 20). Otherwise, TRAVERSE is called again with P_{l+1}, C_{l+1}, and the set of new local maximal contrast sets, W_{l+1}, created such that it contains that only the contrast sets in W_l that subsume contrast sets in P_l (lines 22 to 26). After the recursion completes, W_l is updated with the elements from W_{l+1} (line 27). For a more detailed description of TRAVERSE, see [6].

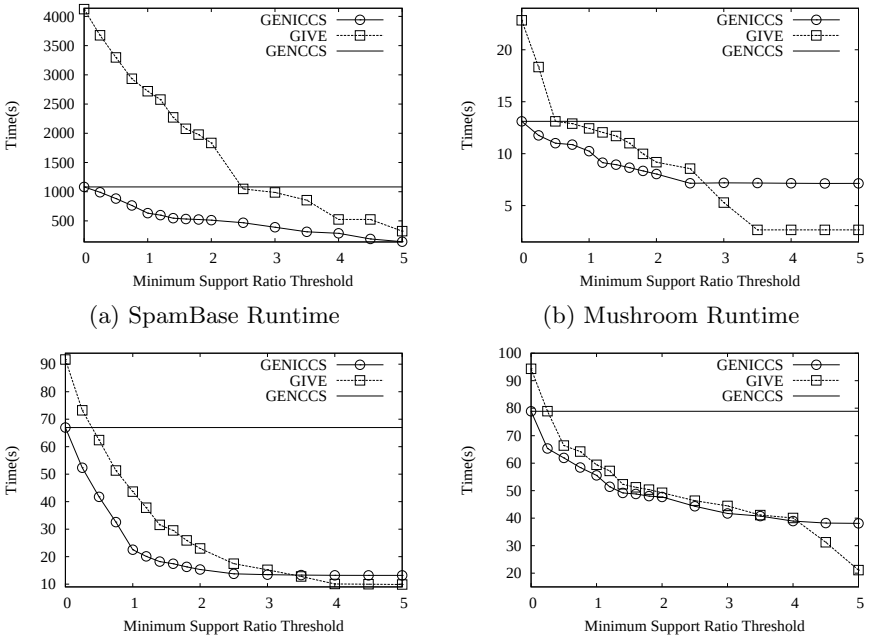
6 Experimental Results

In this section, we present the results of an experimental evaluation of our approach. Our experiments were conducted on an Intel dual core 2.40GHz processor with 4GB of memory, running Windows 7 64-bit. Discovery tasks were performed on four real datasets obtained from the UCI Machine Learning Repository [18]. Table 5 lists the name, the number of transactions, the number of attributes, the number of groups for each dataset, and the cardinality of the longest combine set before and after mutual information and all confidence were applied. These datasets were chosen because of their use with previous contrast set mining techniques and the ability to mine valid contrast sets with high specificity.

Our experiments evaluate whether the use of mutual information and all confidence in limiting the size of the search space, and the use of the minimum support ratio threshold (MSRT) during the search process produces a smaller number of "more interesting" contrast sets in less time, than using either method separately. Thus we compare GENICCS with GENCCS, and GIVE. We do not use STUCCO and CIGAR for comparison, as they do not incorporate either mutual information, all confidence, or the MSRT. Additionally, we have previously demonstrated the efficiency and effectiveness of GENCCS over STUCCO and CIGAR [5].

Table 5: Dataset Description

Data Set	**# Transactions**	**# Attributes**	**# Groups**	**Cardinality of Longest Combine Set**	
				Before MI&AC	**After MI& AC**
Census	32561	14	5	145	16
Mushroom	8124	22	2	198	35
Spambase	4601	58	2	560	75
Waveform	5000	41	3	390	80



(a) SpamBase Runtime

(b) Mushroom Runtime

(c) Census Runtime

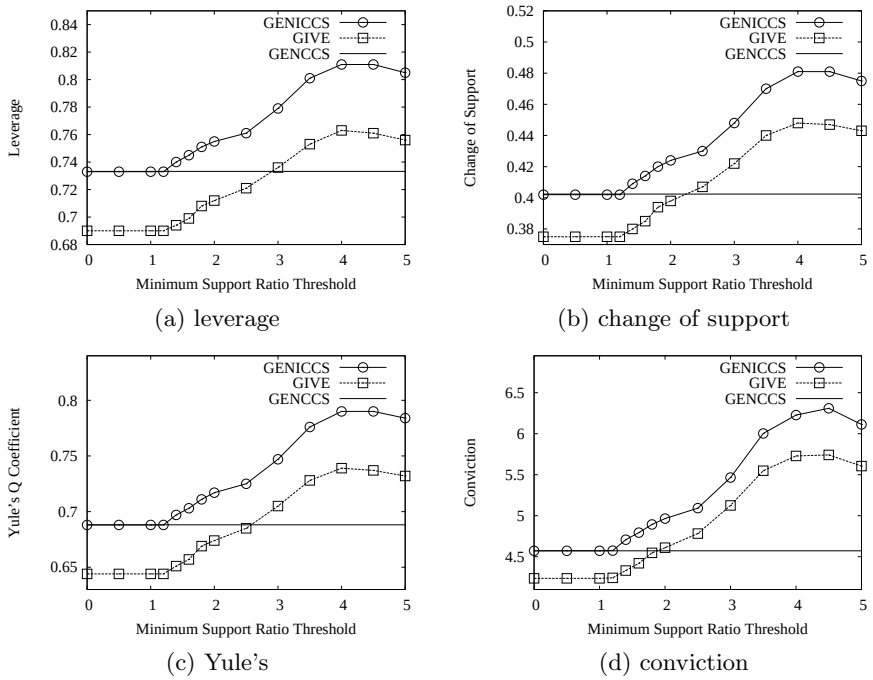
(d) Waveform Runtime

Fig. 1: Summary of runtime results

6.1 Performance of GENICCS

We first examine the efficiency of GENICCS by measuring the time taken to complete the contrast set mining task as the MSRT varies. For each of GENICCS, GENCCS, and, GIVE we set the significance level to 0.95, the minimum support difference threshold and minimum subset ratio threshold to 0, respectively, and averaged the results over 10 runs. Figure 1 shows the number of valid contrast sets discovered and the run time for each of the datasets as the MSRT is varied. For both GENICCS and GENCCS, we set ψ and ξ to be the mean mutual information, and mean all confidence values, respectively. For GENCCS, these mutual information and all confidence values were shown previously to be optimal [5].

Figures 1a, 1b, 1c, and 1d, show that GENICCS has the same run time as GENCCS when the MSRT is 0, but decreases significantly as the MSRT is increased. Since the MSRT serves as a constraint, as we increase its value, fewer correlated contrast sets satisfy this constraint and GENICCS becomes more efficient than GENCCS. As the MSRT approaches 3, 3.5, and 4 for the Mushroom, Census, and Waveform datasets, respectively, the cost of using the mutual information and all confidence begins to outweigh the benefits of using the MSRT as GIVE begins to outperform GENICCS. GIVE never outperforms GENICCS on the spambase dataset.



(a) leverage (b) change of support

(c) Yule's (d) conviction

Fig. 2: Summary of interestingness results

6.2 Interestingness of the Contrast Sets

We examine the effectiveness of GENICCS by measuring the average leverage, change of support, yule's Q, and conviction for the valid contrast sets discovered as the MSRT varies, as shown in Figure 2. We focus on the Waveform dataset as it had not only the smallest run time difference between GENICCS and GIVE, 17%, but also the largest combine set, 80, after mutual information and all confidence were applied. The average leverage, change of support, yule's Q coefficient, and conviction of the valid contrast sets discovered by GIVE and GENCCS for each dataset is also provided for comparison.

Figure 2a, 2b, 2c, and 2d shows that the maximal contrast sets discovered by GENICCS are more interesting, when measured by the average leverage, change of support, yule's Q coefficient, and conviction, than those discovered by either GIVE or GENCCS. The magnitude of the difference is significant even at higher MSRT values where GIVE outperforms GENICCS as shown in Figure 1d, which implies that even though GIVE is less expensive, GENICCS produces better quality contrast sets.

6.3 Number of Contrast Sets

We examine the effect of the MSRT on the number of contrast sets discovered by GENICCS. Here again, we focus on the Waveform dataset. Table 6 shows

Table 6: Number of Maximal Contrast Sets

MSRT	# of Maximal Contrast Sets	
	GENICCS	GIVE
0	17861	32842
0.5	17861	32842
1	17861	32842
1.2	16919	30678
1.4	8475	15064
1.6	6586	10424
1.8	5041	7097
2	4134	5780
2.5	3039	3691
3	1439	1789
3.5	712	881
4	494	618
4.5	351	456
5	285	384
GENCCS		17861

the number of contrast sets discovered by GENICCS and GIVE as the MSRT is varied. The number of contrast sets discovered by GENCCS is also provided for comparison.

Table 6 shows that at an MSRT of 5, GENICCS discovers 285 maximal correlated contrast sets, while GIVE discovers 384 maximal contrast sets. This represents a 98.4% and a 97.9% reduction, respectively, from the 17861 discovered by GENCCS, and a 99.1% and 98.8% reduction, respectively, from the 32842 discovered by GIVE when the MSRT is 0. The number of contrast sets produced by GIVE when the MSRT is 0 is the largest superset of maximal valid contrast sets. Having already established the interestingness of these contrast sets, and the efficiency to generate them, we also postulate that they would be a more manageable set for a domain expert to examine.

7 Conclusion

In this paper, we proposed a contrast set mining technique, GENICCS, for mining maximal valid correlated contrast sets that meet a minimum support ratio threshold. We compared our approach with two previous contrast set mining approaches, GIVE and GENCCS, and found our approach to be comparable in terms of efficiency but more effective in generating interesting correlated contrast sets. We also introduced four probability-based interestingness measures and demonstrated how they can be used to rank correlated contrast sets. In addition, the results showed that the use of the MSRT allowed for an efficient discovery of a smaller more interesting subset of contrast sets that would be more manageable for a domain expert to examine. Future work will further incorporate space reduction techniques with additional interestingness measures.

References

1. Bay, S.D., Pazzani, M.J.: Detecting group differences: Mining contrast sets. Data Min. Knowl. Discov. **5**(3) (2001) 213–246
2. Hilderman, R., Peckham, T.: A statistically sound alternative approach to mining contrast sets. Proceedings of the 4th Australasian Data Mining Conference (AusDM'05) (Dec. 2005) 157–172
3. Simeon, M., Hilderman, R.J.: Exploratory quantitative contrast set mining: A discretization approach. In: ICTAI '07: Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence - Vol.2 (ICTAI 2007), Washington, DC, USA, IEEE Computer Society (2007) 124–131
4. Simeon, M., Hilderman, R.J.: Cosine: A vertical group difference approach to contrast set mining. In: Advances in Artificial Intelligence - 24th Canadian Conference on Artificial Intelligence, Canadian AI 2011, St. John's, Canada, May 25-27, 2011. Proceedings. (2011) 359–371
5. Simeon, M., Hilderman, R.J.: Genccs: A correlated group difference approach to contrast set mining. In: Machine Learning and Data Mining in Pattern Recognition - 7th International Conference, MLDM 2011, New York, NY, USA, August 30 - September 3, 2011. Proceedings. (2011) 140–154
6. Simeon, M., Hilderman, R.J., Hamilton, H.J.: Mining interesting contrast sets. In Lorenz, P., Dini, P., eds.: INTENSIVE 2012, The Fourth International Conference on Resource Intensive Applications and Services. (2012)
7. Wong, T.T., Tseng, K.L.: Mining negative contrast sets from data with discrete attributes. Expert Syst. Appl. **29**(2) (2005) 401–407
8. Holm, S.: A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics **6** (1979) 65–70
9. Lin, J., Keogh, E.J.: Group sax: Extending the notion of contrast sets to time series and multimedia data. In: PKDD. (2006) 284–296
10. Cover, T.M., Thomas, J.A.: Elements of information theory. Wiley-Interscience, New York, NY, USA (2006)
11. Han, J.: Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2005)
12. Lavrac, N., Flach, P.A., Zupan, B.: Rule evaluation measures: A unifying view. In: Proceedings of the 9th International Workshop on Inductive Logic Programming. ILP '99, London, UK, UK, Springer-Verlag (1999) 174–185
13. Tan, P.N., Kumar, V., Srivastava, J.: Selecting the right objective measure for association analysis. Inf. Syst. **29**(4) (2004) 293–313
14. Geng, L., Hamilton, H.J.: Interestingness measures for data mining: A survey. ACM Comput. Surv. **38**(3) (2006) 9
15. Piatetsky-Shapiro, G.: Discovery, analysis, and presentation of strong rules. In Piatetsky-Shapiro, G., Frawley, W., eds.: Knowledge Discovery in Databases. AAAI/MIT Press, Cambridge, MA (1991)
16. Yule, G.U.: On the association of attributes in statistics: With illustrations from the material of the childhood society, &c. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character **194**(252-261) (1900) 257–319
17. Brin, S., Motwani, R., Ullman, J.D., Tsur, S.: Dynamic itemset counting and implication rules for market basket data. SIGMOD Rec. **26**(2) (1997) 255–264
18. Asuncion, A., Newman, D.: UCI machine learning repository (2007)