Mining Interesting Contrast Sets

Mondelle Simeon Robert J. Hilderman Howard J. Hamilton
Department of Computer Science
University of Regina
Regina, SK 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 contrast set is a conjunction of attribute-value pairs that differ significantly in their distribution across groups. The search for valid contrast sets can produce a prohibitively large number of results which must be further filtered in order to be examined by a domain expert and have decisions enacted from them. In this paper, we introduce the notion of the minimum support ratio threshold to measure the ratio of maximum and minimum support across groups. We propose a contrast set mining technique to discover maximal valid contrast sets which meet a minimum support ratio threshold. We also introduce five interestingness measures and demonstrate how they can be used to rank contrast sets. Our experiments on real datasets demonstrate the efficiency and effectiveness of our approach, and the interestingness of the contrast sets discovered.**

***Keywords*-contrast set mining; group differences; data mining.**

I. 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].

Existing contrast set mining techniques can produce a prohibitively large set of differences across groups with varying levels of interestingness [2] [5]. For example, suppose we want to find out which demographic and socio-economic characteristics differentiate between women who use short-term, long-term, or no contraceptive methods. We could use data, as shown in Table I, with five such characteristics: wife currently working, husband currently working, has children, has high standard of living, and has high media exposure, where 1 indicates the characteristic is true, and 0 that it is false. There are 30 possible combinations of characteristics that differentiate between the women, however, they are not all equally interesting. For instance, assume we found that all the women who are working and have high media exposure use either short-term or long-term methods whereas those who are not working and do not have high media exposure, are equally likely to use either a short-term, long-term or no contraceptive method. Perhaps then, we could use the former result in a media marketing campaign targeted to that specific group of women, while the latter result which is less conclusive, is considered “uninteresting”. We propose using a measure, called the minimum support ratio threshold, to discover “interesting” group differences during the search process.

Table I
SAMPLE DATASET

	wife currently working	husband currently working	has children	has high standard of living	has high media exposure
TID	A	B	C	D	E
1	1	0	1	1	1
2	0	1	1	0	1
3	1	1	0	0	1
4	0	0	1	1	1
5	1	1	1	1	1
. . .	. . .	. . .	. . .	. . .	. . .
768	0	0	0	1	1

The remainder of this paper is organized as follows. In Section II, we briefly review related work. In Section III, we describe the correlated contrast set mining problem. In Section IV, we provide an overview of the search framework for contrast set mining. In Section V, we introduce our algorithm for mining contrast sets that meet our minimum support ratio threshold. In Section VI, we present a summary of experimental results from a series of mining tasks. In Section VII, we conclude and suggest areas for future work.

II. 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 [6] that can include negation of terms in the contrast set. The main difference is their use of Holm's sequential rejective method [7] 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 [8]. 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) algorithm [4] is a contrast set mining technique that uses a vertical data format, diffsets, a back tracking search algorithm, and simple discretization in mining maximal contrast sets from both discrete and continuous-valued attributes. The results demonstrate that COSINE is more efficient than STUCCO and CIGAR, even at very low minimum support difference thresholds. The GENCCS (Generate Correlated Contrast Sets) algorithm [5] is a contrast set mining technique that extends COSINE by utilizing mutual information and all-confidence to select the attribute-value pairs that are most highly correlated. The results show that GENCCS is more efficient and produced more interesting contrast sets than STUCCO and CIGAR.

III. Problem Definition

Let $\mathcal{A} = \{a_1, a_2, \dots, a_n\}$ 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 $\mathcal{T} = \{x_1, x_2, \dots, x_p\}$ 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 $\{i_1, i_2, \dots, i_m\}$ be a set of m distinct values from the set $\{1, 2, \dots, n\}$, $m < n$. Let $\mathcal{F} = \{a_{i_1}, a_{i_2}, \dots, a_{i_m}\}$, $a_{i_k} \in \mathcal{A}$, be a set of m distinct class attributes.

Let $G = \{a_1 : [v_{1l}, v_{1r}], \dots, a_m : [v_{ml}, v_{mr}]\}$, $a_k \in \mathcal{F}, 1 \leq k \leq m$, $a_i \neq a_j$, $\forall i, j$, be a set of m distinct class attribute-interval pairs, called a *group*. Let $X = \{a_1 : [v_{1l}, v_{1r}], \dots, a_q : [v_{ql}, v_{qr}]\}$, $a_i, a_j \in \mathcal{A} - \mathcal{F}$, $1 \leq k \leq q \leq n - m$, $a_i \neq a_j$, $\forall i, j$, 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. $G_1, G_2, \dots, G_b$ is called a *valid contrast set* (CS) 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 and minimum support across the groups is sufficiently large. Formally, for a valid contrast set, X, the ratio is given by

$$\lambda(X) = \begin{cases} \infty, & \text{if } \min_i^n \{supp(X, G_i)\} \\ & = 0, \\ \dfrac{\max_i^n \{supp(X, G_i)\}}{\min_i^n \{supp(X, G_i)\}}, & \text{otherwise.} \end{cases} \tag{5}$$

A large value for $\lambda(X)$ implies that X occurs in significantly fewer transactions in one group G_i than in some other group

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-188-5

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* (λ-CS) if, and only if,

$$\lambda(X) \geq \omega, \tag{6}$$

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

$$\lambda(X) = \infty. \tag{7}$$

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, our goal is to find all the maximal λ contrast sets (i.e., all maximal valid contrast sets that satisfy Equations 6 and 7).

IV. Background

A. Search for Contrast Sets

Our algorithm uses a backtracking search paradigm in order to enumerate all maximal group differences. Backtracking algorithms are useful because they allow us to iterate through all the possible configurations of the search space. Consider the partial search space tree shown in Figure 1. The root of the tree corresponds to the combine set $\{A : [0, 0], A : [1, 1], B : [0, 0], B : [1, 1], C : [0, 0], C : [1, 1], D : [0, 0], D : [1, 1], E : 0, 0], E : [1, 1]\}$, which is composed of the 1-specific contrast sets from the attributes shown in Table I. Each attribute can take a value of 0 or 1. All these contrast sets share the empty prefix in common.

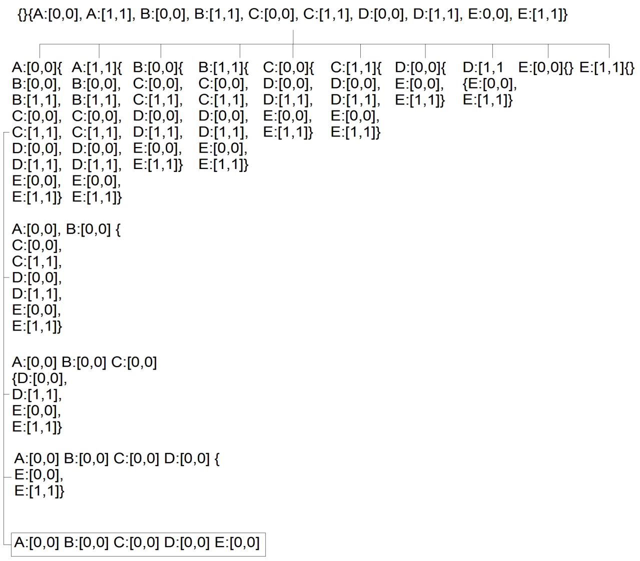

{}{A:[0,0], A:[1,1], B:[0,0], B:[1,1], C:[0,0], C:[1,1], D:[0,0], D:[1,1], E:0,0], E:[1,1]}
A:[0,0]{ B:[0,0], B:[1,1], C:[0,0], C:[1,1], D:[0,0], D:[1,1], E:[0,0], E:[1,1]}
A:[1,1]{ B:[0,0], B:[1,1], C:[0,0], C:[1,1], D:[0,0], D:[1,1], E:[0,0], E:[1,1]}
B:[0,0]{ C:[0,0], C:[1,1], D:[0,0], D:[1,1], E:[0,0], E:[1,1]}
B:[1,1]{ C:[0,0], C:[1,1], D:[0,0], D:[1,1], E:[0,0], E:[1,1]}
C:[0,0]{ D:[0,0], D:[1,1], E:[0,0], E:[1,1]}
C:[1,1]{ D:[0,0], D:[1,1], E:[0,0], E:[1,1]}
D:[0,0]{ E:[0,0], E:[1,1]}
D:[1,1 {E:[0,0], E:[1,1]}
E:[0,0]{}
E:[1,1]{}
A:[0,0], B:[0,0] { C:[0,0], C:[1,1], D:[0,0], D:[1,1], E:[0,0], E:[1,1]}
A:[0,0] B:[0,0] C:[0,0] {D:[0,0], D:[1,1], E:[0,0], E:[1,1]}
A:[0,0] B:[0,0] C:[0,0] D:[0,0] { E:[0,0], E:[1,1]}
A:[0,0] B:[0,0] C:[0,0] D:[0,0] E:[0,0]

Figure 1. Search Tree: Box indicates a maximal contrast set

Formally, for a set of contrast sets with prefix P, $[P] = \{X_1, X_2, \cdots, X_n\}$, the intersection of PX_i with all of PX_j with $j > i$ is performed to obtain a new combine set $[PX_i]$ where the contrast set PX_iX_j satisfies Equations 1, 2, 3, 4, and 6. For example, from $[A : [0, 0]] = \{B : [0, 0], B : [1, 1], C : [0, 0], C : [1, 1], D : [0, 0], D : [1, 1], E : [0, 0], E : [1, 1]\}$, we obtain $[A : [0, 0]B : [0, 0]] = \{C : [0, 0], C : [1, 1], D : [0, 0], D : [1, 1], E : [0, 0], E : [1, 1]\}$ for the next level of the search tree. A node with an empty combine set such as $[E : [0, 0]$ need not be explored further.

B. Data Format

Contrast set mining algorithms using the vertical format have been shown to be very effective and usually outperform horizontal approaches [5] [4]. Our algorithm also uses a vertical data format in representing the data.

C. Ranking Methods

A contrast set mining task has the potential to return many contrast sets. Consequently, measures are needed to rank the relative interestingness of the contrast sets prior to presenting them to the end-user. Much work has been done on various measures of interestingness. For more on this, see [9] [10]. Ideally, a measure would be used to rank the contrast sets as well as describe them, akin to the support, confidence, leverage and lift measures used in association rule mining. In this section, we propose four 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, such as the one shown in Table II for the association rule $X \rightarrow G_i$. Rows represent the occurrence of the contrast set and the columns represent occurrence of the groups.

Table II
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

Distribution Difference: The *distribution difference* (DD) of a contrast set measures how different the support for a group in the contrast set is from the support for the group in the entire dataset [3]. Formally, the distribution difference for X, in G_i is given by

$$\mathit{DD}(X \rightarrow G_i) = \left| \frac{n(X, G_i)}{n(X)} \times \frac{N}{n(G_i)} - 1 \right|.$$

For example, assume that in the entire dataset 40% of individuals are male and 60% are female. Now assume that we have two contrast sets where 65% are male and 35% are female in the first, and 42% are male and 58% are female in the second. In comparing these contrast sets, the first is more interesting because it deviates more from the distribution in the entire dataset. The distribution difference captures that. The distribution difference can have a minimum value of zero, which indicates that the instances in the contrast set occur in the same distribution across the groups in comparison to the distribution in the entire dataset. A large distribution difference indicates significant variance in the distribution across the groups.

The *aggregate distribution difference* of a contrast set is the sum of the distribution difference values over all the groups. Formally, the aggregate distribution difference for X across $G_1, G_2, \ldots, G_n$ is given by

$$\mathit{Aggregate\ DD}(X) = \sum_{i}^{n} \mathit{DD}(X \rightarrow G_i).$$

Unusualness: Unusualness is a measure of interestingness used in subgroup discovery [11]. Given a set of instances possessing some property of interest, a *subgroup* is a subset of instances in which the statistical characteristics of the property of interest are ”unusual”. Instances in the subgroup can be described by an association rule, $X \rightarrow Y$, where the property of interest is represented by the consequent, Y, and the antecedent, X, an itemset. Weighted relative accuracy (WRAcc) is used to evaluate the quality (i.e., unusualness) of the induced association rules.

Contrast set mining has been shown to be an extension of subgroup discovery, where each group represents a different property of interest [12]. Thus, we can use the weighted relative accuracy to measure the *unusualness* (UN) of X in G_i. Formally, unusualness is given by

$$\begin{aligned} \mathit{UN}(X, G_i) &= p(X) \times (p(G_i|X) - p(G_i)) \\ &\approx \frac{n(X)}{N} \times \left(\frac{n(X, G_i)}{n(X)} - \frac{n(G_i)}{N} \right). \end{aligned}$$

Possible values for unusualness range from -1 to 1. The unusualness of X, is determined by the group for which the unusualness is largest. Thus, the unusualness of X, across $G_1, G_2, \ldots, G_n$ is given by

$$\mathit{Maximum\ UN}(X) = \max_{i} \mathit{UN}(X, G_i).$$

Coverage: The *coverage* of an association rule, $X \rightarrow G_i$ is the proportion of instances in the dataset where X is true in G_i [10], and is given by

$$\mathit{Coverage}(X \rightarrow G_i) = p(X) = \frac{n(X)}{N}.$$

Possible values for coverage range from 0 to 1, inclusive, where contrast sets that occur more frequently have a higher coverage.

The coverage of a contrast set for all groups, called the *aggregate coverage*, is the sum of the individual coverage values for the contrast set in each group. The aggregate coverage of a contrast set X in $G_1, G_2, \ldots, G_n$ is given by

$$\mathit{Aggregate\ Coverage}(X) = \sum_{i=1}^{n} \mathit{Coverage}(X \rightarrow G_i)$$

Lift: The *lift* of an association rule, $X \rightarrow G_i$, measures how many times X and G_i actually occur together compared to the number of times X and G_i would be expected to occur together if they where statistically independent [13] and is given by

$$\mathit{Lift}(X \rightarrow G_i) = \frac{p(X, G_i)}{P(X)P(G_i)} = \frac{N \times n(X, G_i)}{n(X) \times n(G_i)}$$

Possible values for lift range from 0 to infinity, inclusive.

The lift for a contrast set across all groups called the *aggregate lift*, is the sum of the lift for the contrast set in each group. Formally,

$$\mathit{Aggregate\ Lift}(X) = \sum_{i=1}^{n} \mathit{Lift}(X \rightarrow G_i).$$

Interestingness Factor: The four interestingness measures described above can be used individually to rank discovered contrast sets. However, they can can also be used in combination to determine the most interesting contrast sets based on multiple measures. The *Interestingness Factor* (IF) of a contrast set is the average of it's rank over a set of the selected interestingness measures, and is given by

$$\mathit{IF}(X) = \frac{\sum_{i=1}^{n} r_i \times w_i}{\sum_{i=1}^{n} w_i \times m_i}$$

where n is the number of interestingness measures, r_i is the rank of the contrast set by a measure i, w_i is the weight of the ranking measure i (i.e., a user-defined parameter indicating the relative importance of measure i), and m_i is the maximum rank of measure i. Possible values for the interestingness factor can range from 0 to 1, where values close to 0 indicate contrast sets which are more interesting, while values close to 1 indicate contrast sets which are less interesting. An interestingness factor of 1 indicates that the contrast set was ranked the lowest by each method.

V. Mining Interesting Valid Contrast Sets

GIVE (Generate Interesting Valid contrast sEts) presented in Algorithm 1, finds all the maximal valid contrast sets in a given dataset (i.e, all the contrast sets that satisfy Equations 1, 2, 3, 4, and 6). It adapts several tenets of the backtracking search technique proposed in [4] for contrast set mining.

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

Input: A dataset, $\mathcal{D}$, and ranking method, m

Output: The set of ranked interesting valid contrast sets W

```
 1: for each $i \in \mathcal{A}, \mathcal{A} \in \mathcal{D}, i \notin \mathcal{F}$ do
 2:     if $i \in \mathcal{Q}$ then
 3:         $\mathcal{V}(i) =$ DISCRETIZE($i$)
 4:     end if
 5:     $B = B \cup \mathcal{V}(i)$
 6: end for
 7: $C_0$ =COMBINE($\{\}, B, \epsilon, \sigma, 0, \omega, W$)
 8: Sort each $C_0$ in increasing $|C_x|$ then in increasing $F_x$
 9: TRAVERSE($\{\}, C_0, W, \epsilon, \sigma, \kappa, \omega$)
10: RANK($W, m$)
11: return W
```

GIVE begins by first determining all the 1-specific contrast sets from the domain $\mathcal{V}$ of each attribute in the dataset not occurring in $\mathcal{F}$, and storing them in B (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]. GIVE determines valid 1-specific contrast sets by calling the subroutine COMBINE, with the empty prefix $\{\}, B, \epsilon, \sigma, 0, \omega,$ and W which will hold all the valid contrast sets at the end (line 7). The set of valid 1-specific contrast sets is re-ordered in increasing cardinality of the combine set and frequency (line 8). Thus contrast sets with a lower frequency at one level are less likely to produce contrast sets that meet our frequency threshold on the next level. GIVE then calls the subroutine, TRAVERSE, with the empty prefix, $\{\}, C_0, W, \epsilon, \sigma, \kappa,$ and ω (line 9). The valid contrast sets are ranked by a method m (line 10).

A. 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 ω, the COMBINE Algorithm, shown in Algorithm 3, builds new contrast sets from P and H that satisfy Equations 1, 2, 3, 4 and 6.

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). For each z, it calculates its diffset, D_z, its potential combine set, C_z, and its frequency, F_z (line 4). It then determines whether z satisfies Equations 1, 2, 3, and 4 (line 5). COMBINE also checks whether Equation 7 is satisfied (line 6). Any z which meets this criteria is potentially maximal and no further processing of the subset tree has to be done. z is added to W if it has no superset already in W (lines 7 to 9). Otherwise, COMBINE checks whether Equation 6 is satisfied. Any z which meets this criteria is added to the combine set, C (lines 10 to 11). Finally C is returned.

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

```
 1: $C = \emptyset$
 2: for each $y \in H$ do
 3:     $z = P \cup \{y\}$
 4:     Determine $D_z$, $F_z$, $C_z, \alpha_L$
 5:     if significant($z, \alpha_L$) & large($z, \epsilon$) & frequent($z, \sigma$) &
        specific($z, \kappa$) then
 6:         if $\lambda(z) == \infty$ then
 7:             if $Z \not\supseteq z \cup H : Z \in W$ then
 8:                 $W = W \cup \{z\}$
 9:             end if
10:         else if $\lambda(z) \geq \omega$ then
11:             $C = C \cup \{z\}$
12:         end if
13:     end if
14: end for
15: return $C$
```

B. 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 all, maximal or λ contrast sets that satisfy Equations 1, 2, 3, 4, and 6.

TRAVERSE begins by determining the next prefix, P_{l+1} (line 2). It then determines a new possible set of combine elements, H_{l+1}, by first stripping the prefix P_{l+1} of the previous prefix P_l, creating $P^{'}_{l+1}$ (line 4). It then determines from the list of elements in C_l, those which are greater than (appear after) P_{l+1} (recall from above, that P_{l+1} was also chosen from C_l) (line 6). For any such element, y, TRAVERSE strips it of the prefix P_l, creating $y^{'}$ (line 7). It then checks whether $P^{'}_{l+1}$ is not equal to $y^{'}$ and whether it is in the combine set of P_l (line 8). $P^{'}_{l+1}$ and $y^{'}$ are 1-specific contrast sets and if they originate from the same attribute, they cannot be part of a new contrast set, as we require contrast sets to have unique attributes. If $y^{'}$ is in the combine set of P_l then it will be in the combine set of P_{l+1}. If both conditions are met, y is added to H_{l+1} (line 9). TRAVERSE repeats this for every member of C_l.

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

```
1:  for each $x \in C_l$ do
2:      $P_{l+1} = \{x\}$
3:      $H_{l+1} = \emptyset$
4:      Let $P'_{l+1} = P_{l+1} - P_l$
5:      for each $y \in C_l$ do
6:          if $y > P_{l+1}$ then
7:              Let $y' = y - P_l$
8:              if $y' \neq P'_{l+1}$ & $y' \in C_{P_l}$ then
9:                  $H_{l+1} = H_{l+1} \cup \{y\}$
10:             end if
11:         end if
12:     end for
13:     if $|W_l| > 0$ then
14:         if $Z \supseteq P_{l+1} \cup H_{l+1} : Z \in W_l$ then
15:             return
16:         end if
17:     end if
18:     $C_{l+1}$ = COMBINE$(P_{l+1}, H_{l+1}, W_l, \epsilon, \sigma, \kappa, \omega)$
19:     Sort $C_{l+1}$ by increasing $F_z$, $\forall z \in C_{l+1}$
20:     if $C_{l+1} \neq \emptyset$ then
21:         if $Z \not\supseteq P_{l+1} : Z \in W_l$ then
22:             $W_l = W_l \cup P_{l+1}$
23:         end if
24:     else
25:         $W_{l+1} = \{W \in W_l : x \in W\}$
26:     end if
27:     if $C_{l+1} \neq \emptyset$ then
28:         TRAVERSE$(P_{l+1}, C_{l+1}, W_{l+1}, \epsilon, \sigma, \kappa, \omega)$
29:     end if
30:     $W_l = W_l \cup W_{l+1}$
31: end for
```

The cardinality of the current set of contrast sets, is determined and if it is greater than zero, TRAVERSE checks if $P_{l+1} \cup H_{l+1}$ is subsumed by an existing contrast set. If yes, the current and subsequent contrast sets in C_l can be pruned away (lines 13 to 17). If not, an extension is necessary. TRAVERSE then creates a new combine set for the next level of the search by using the subroutine COMBINE (line 18). The combine set, C_{l+1}, is sorted in increasing order of the frequency of its members (line 19). Any contrast set not in the combine set refers to a node pruned from the search tree. TRAVERSE checks if C_{l+1} is empty and if P_{l+1} is not a subset of any contrast set in W_l, P_{l+1} is added to W_l (lines 20 to 23). Otherwise, a new set of local contrast sets, W_{l+1}, is created based such that only the contrast sets in W_l that contain all the contrast sets in P_l are added to W_{l+1} (line 25). This allows the number of contrast sets of interest to be narrowed down as recursive calls are made. If C_{l+1} is not empty, TRAVERSE is called again with P_{l+1}, C_{l+1}, and the set of new local maximal contrast sets, W_{l+1} (lines 27 to 28). After the recursion completes, the set of maximal valid contrast sets or λ valid contrast sets, W_l, is updated with the elements from W_{l+1} (lines 30).

Table III
DATASET DESCRIPTION

Data Set	# Transactions	# Attributes	# Groups
Census	32561	14	5
Mushroom	8124	22	2
Spambase	4601	58	2
Waveform	5000	41	3

VI. EXPERIMENTAL RESULTS

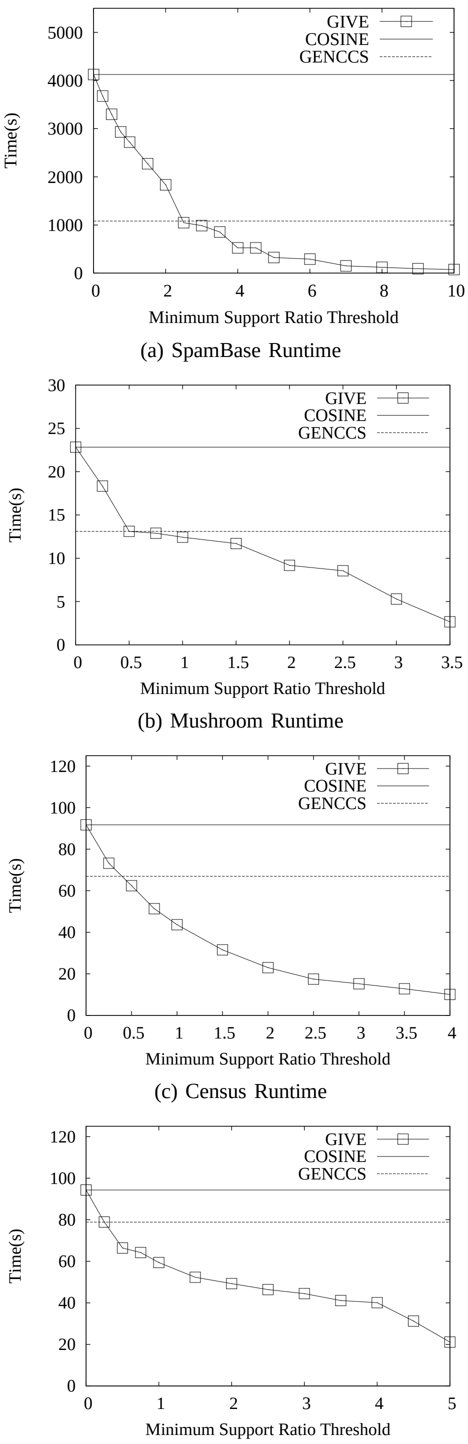
In this section, we present the results of an experimental evaluation of our approach. Our experiments were conducted on an Intel dual core 2.40GHz processor with 4GB of memory, running Windows 7 64-bit. Discovery tasks were performed on four real datasets obtained from the UCI Machine Learning Repository [14]. Table III lists the name, number of transactions, number of attributes, and the number of groups for each dataset. 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.

A. Performance of GIVE

We first examine the efficiency of GIVE by measuring the time taken to complete the contrast set mining task as the minimum support ratio threshold (MSRT) varies. We set the significance level to 0.95, the minimum support difference and minimum subset ratio to 0, respectively, and average the results over 10 consecutive runs. Figure 2 shows the number of valid contrast sets discovered and the run time for each of the datasets as the MSRT is varied. Results are only shown for MSRT values which produce substantial changes in the time. The time taken by COSINE and GENCCS for each dataset is also provided for comparison. For GENCCS, we set use the mean mutual information, and mean all confidence value, as the mutual information threshold, and all confidence threshold, respectively, as these were shown previously to be optimal [5]. Figures 2a, 2b, 2c, and 2d, show that GIVE is as efficient as COSINE when the minimum support ratio threshold is 0 but less than that of GENCCS. GIVE becomes more efficient than GENCCS when the MSRT is greater than 2.5, 0.75, 0.25, and 0.75 for the Spambase, Mushroom, Waveform and Census datasets, respectively. Since the MSRT serves as a constraint, as we increase its value, fewer contrast sets satisfy this constraint and GIVE becomes more efficient.

B. Effectiveness of GIVE

We examine the effectiveness of GIVE by measuring the average unusualness of the valid contrast sets discovered



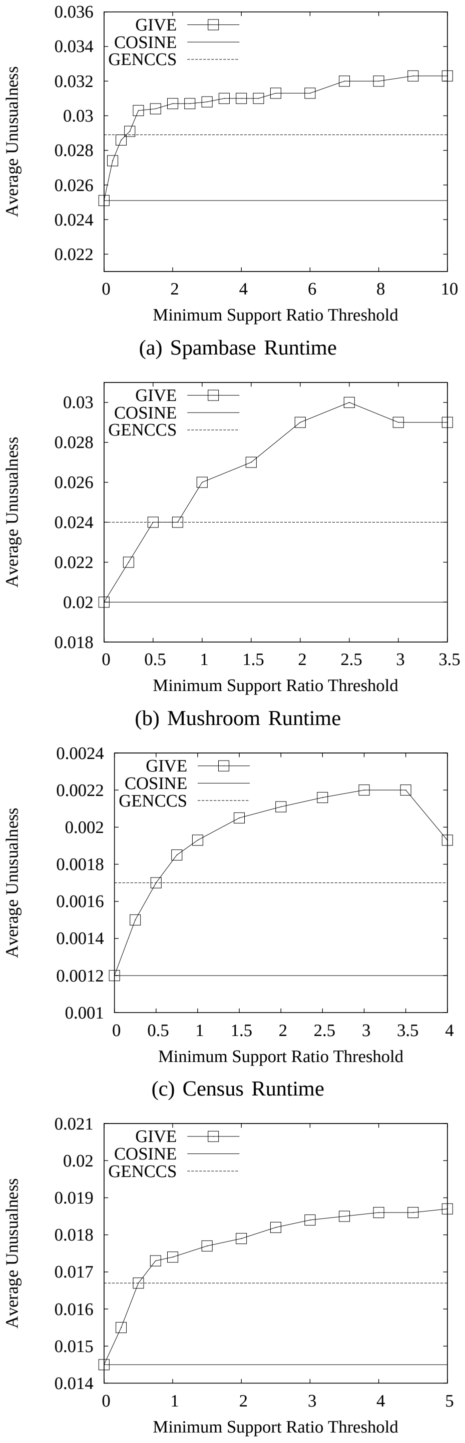
(a) SpamBase Runtime

(b) Mushroom Runtime

(c) Census Runtime

(d) Waveform Runtime

Figure 2. Summary of runtime results



(a) Spambase Runtime

(b) Mushroom Runtime

(c) Census Runtime

(d) Waveform Runtime

Figure 3. Summary of interestingness results

as the MSRT varies, as shown in Figure 3. The average unusualness of the valid contrast sets discovered by COSINE and GENCCS for each dataset is also provided for comparison. Figure 3a, 3b, 3c, and 3d shows that the maximal contrast sets discovered by GIVE are more interesting, when measured by the average unusualness, than those discovered by either GENCCS or COSINE. The magnitude of the difference is significant even at lower

MSRT values where GENCCS outperforms GIVE as shown in Figures 2a, 2c, 2b, and 2d, which implies that even though GENCCS is less expensive, GIVE produces better quality contrast sets. Similar results were observed for the average distribution difference and lift, respectively, and are not shown due to space considerations.

C. Effect of λ on the Search Process

We further explored how the quality of the contrast sets discovered is affected by using λ in the search process by comparing the average interestingness factor for contrast sets that are found by COSINE and GENCCS that are also found by GIVE with those that are not found by GIVE. Table IV shows the average IF using all four measures equally weighted for each of the four datasets. In Table IV, each row shows the average IF of the contrast sets found by COSINE and GIVE, COSINE and ¬GIVE, GENCCS and GIVE, and, GENCCS and ¬GIVE, respectively for each dataset. For example, for the Census dataset, the average interestingness factor of the contrast sets found by COSINE that are also found by GIVE is 0.45. For the Census, Mushroom, and Spambase datasets, the average IF of the contrast sets found by both COSINE and GIVE is significantly lower than those found by COSINE only. This shows that using λ in the search process does not compromise the quality of the contrast sets discovered. With the Waveform dataset, this difference is smaller and not as significant. For GENCCS, a similar observation can be made, though the difference in the average IF is significantly different for all four datasets.

Table IV
EFFECTIVENESS OF λ

Data Set	COSINE & GIVE	COSINE & ¬GIVE	GENCCS & GIVE	GENCCS & ¬GIVE
Census	0.45	0.54	0.34	0.42
Mushroom	0.38	0.49	0.32	0.39
Spambase	0.45	0.58	0.40	0.49
Waveform	0.66	0.69	0.57	0.65

VII. CONCLUSION

In this paper, we introduced the notion of the minimum support ratio threshold and proposed a contrast set mining technique, GIVE, for mining maximal valid contrast sets that meet a minimum support ratio threshold. We compared our approach with two previous contrast set mining approaches, COSINE and GENCCS, and found our approach to be comparable in terms of efficiency but more effective in generating interesting contrast sets. We also introduce five interestingness measures and demonstrated how they can be used to rank contrast sets. In addition, the results showed that the contrast sets discovered by GIVE had an average interestingness factor that was significantly higher than those produced only by COSINE or GENCCS. Future work will incorporate space reduction techniques with additional interestingness measures.

REFERENCES

[1] S. D. Bay and M. J. Pazzani, “Detecting group differences: Mining contrast sets,” *Data Min. Knowl. Discov.*, vol. 5, no. 3, pp. 213–246, 2001.

[2] R. Hilderman and T. Peckham, “A statistically sound alternative approach to mining contrast sets,” *Proceedings of the 4th Australasian Data Mining Conference (AusDM’05)*, pp. 157–172, Dec. 2005.

[3] M. Simeon and R. J. Hilderman, “Exploratory Quantitative Contrast Set Mining: A Discretization Approach,” in *ICTAI (2)*, 2007, pp. 124–131.

[4] ——, “COSINE: A Vertical Group Difference Approach to Contrast Set Mining,” in *Canadian Conference on AI*, 2011, pp. 359–371.

[5] ——, “GENCCS: A Correlated Group Difference Approach to Contrast Set Mining,” in *MLDM*, 2011, pp. 140–154.

[6] T.-T. Wong and K.-L. Tseng, “Mining negative contrast sets from data with discrete attributes,” *Expert Syst. Appl.*, vol. 29, no. 2, pp. 401–407, 2005.

[7] S. Holm, “A simple sequentially rejective multiple test procedure,” *Scandinavian Journal of Statistics*, vol. 6, pp. 65–70, 1979.

[8] J. Lin and E. J. Keogh, “Group sax: Extending the notion of contrast sets to time series and multimedia data,” in *PKDD*, 2006, pp. 284–296.

[9] R. J. Hilderman and H. J. Hamilton, *Knowledge Discovery and Measures of Interest*. Norwell, MA, USA: Kluwer Academic Publishers, 2001.

[10] L. Geng and H. J. Hamilton, “Interestingness measures for data mining: A survey,” *ACM Comput. Surv.*, vol. 38, no. 3, p. 9, 2006.

[11] B. Kavsek and N. Lavrac, “Apriori-sd: Adapting association rule learning to subgroup discovery,” *Applied Artificial Intelligence*, vol. 20, no. 7, pp. 543–583, 2006.

[12] P. Kralj, N. Lavrac, D. Gamberger, and A. Krstacic, “Contrast set mining for distinguishing between similar diseases,” in *AIME*, 2007, pp. 109–118.

[13] S. Brin, R. Motwani, J. D. Ullman, and S. Tsur, “Dynamic itemset counting and implication rules for market basket data,” *SIGMOD Rec.*, vol. 26, no. 2, pp. 255–264, 1997.

[14] A. Asuncion and D. Newman, “UCI machine learning repository,” 2007. [Online]. Available: http://www.ics.uci.edu/∼mlearn/MLRepository.html