Rules for Contrast Sets

Paulo J. Azevedo *
CCTC, Departamento de Informática
Campus de Gualtar
Universidade do Minho
4710-057 Braga
Portugal
pja@di.uminho.pt

Abstract

In this paper we present a technique to derive rules describing contrast sets. Contrast sets are a formalism to represent groups differences. We propose a novel approach to describe directional contrasts using rules where the contrasting effect is partitioned into pairs of groups. Our approach makes use of a directional Fisher Exact Test to find significant differences across groups. We used a Bonferroni within-search adjustment to control type I errors and a pruning technique to prevent derivation of non significant contrast set specializations.

Keywords: Contrast Sets, Association Rules, Fisher exact Test, Bonferroni Adjustment

*This work was Supported by Fundação Ciência e Tecnologia, Project PFound, Project ProtUnf, FEDER and Programa de Financiamento Plurianual de Unidades de I & D.

1 Introduction

As pointed by several authors [3, 19, 9], Contrast Sets mining is a fundamental task in data mining. The aim is to understand the differences among contrasting groups. Learning about groups differences can be important in several domains. For instance, in a census data set to find that there is a fundamental difference between salesmen with high (PhDs) and salesmen with average education (Bachelors). A contrast set is conjunction of characteristics describing a sub-population that occurs with different occurrence along different groups. Take as an example to contrast individuals: at different times (*papers accepted at a conference in 1998 against 2008*), for different spatial locations (*find the distinguishing features of location x for human DNA, versus location x for mouse DNA*), across different classes (*find the differences between people with brown hair, versus those with blond hair*).

Several applications of contrast sets mining in different domains can be found reflecting several properties that can be contrasted. For instance, [13] reformulate the notion of contrast sets for time series data. They redefine it to be the set of key patterns that are maximally different across time series. In [11] the problem of distinguishing between thrombolic brain stroke and embolic brain stroke is addressed using a particular approach to contrast set mining through subgroup discovery.

STUCCO [3, 4], is the standard approach to this mining task. Its search procedure is based on a breadth-first search framework to derive itemsets representing contrasts along different datasets (groups). It uses a two-sided chi-squared test to determine significant contrast sets. Type I errors are controlled through a levelwise cutoff adjustment based on the number of performed tests. STUCCO also uses a particular definition of interest to discard contrast set specializations that represent no new information.

We propose a novel approach (Rules for Contrast Sets) to contrast set mining that is more robust in preventing derivations of false discoveries and non relevant specializations. Our approach is association rule based where a frequent itemset mining engine is redesigned to derived rules expressing pairwise contrasts. Each rule is derived as long as the expressed contrast is significant according to a Fisher exact test. We readjust the techniques introduced in [17] and [18] to determine significance in a contrasts sets setting. Specializations of a contrast set are considered whenever they pass a pruning filter also based on the same statistical test. Since multiple tests are performed, the Bonferroni-like adjustment proposed in [18] is used to

obtained a suitable cutoff for this task. We also show evidence that the two-sided chi-squared test and the associated cutoff adjustment used in [4] are unappropriated for this task. In particular, depending on the number of contrasting groups, we show that the two-sided test tends to yield overestimated or underestimated p-values.

The rest of the paper is organized as follows. In the next section we briefly survey the STUCCO approach to contrast set mining. In section 3 we present our approach. A novel form of rules for representing contrast sets is presented as well as a strategy for detecting contrasts which includes a method to prevent type I errors. A pruning method to avoid the derivation of non relevant contrast sets is also described. Section 4 presents evaluation where specific patterns derived by STUCCO and our approach are analyzed. Finally, related work and conclusions are put forward.

2 Contrast Sets

In [4] the search algorithm STUCCO (Search and Testing for Understandable Consistent Contrasts) is proposed to find all contrast sets whose support differs meaningfully across groups. A contrast set is represented as an itemset. A group is a user supplied dataset where contrast sets occur. Groups are represented by items e.g. G_i, G_j, typically attribute values. Formally, the goal is to find those contrast sets (cs) where:

$$\exists ij \; P(cs|G_i) \neq P(cs|G_j) \tag{1}$$

and

$$max_{i,j}|sup(cs, G_i) - sup(cs, G_j)| \geq \delta \tag{2}$$

where δ is a user defined threshold called the minimum support difference. Support in this setting measures the proportional incidence of the contrast set within a group (G_i and G_j). Contrast sets where eq. (1) is statistically valid are called *significant*, and contrast sets where eq. (2) is met is referred as *large*. If both requirements are achieved, then the contrast set is considered a deviation. The statistical significance criterion ensures that the contrast set represents a true difference between the groups. The second criterion measures the effect size and ensures that everything that is reported to the user provides a large enough effect to be important.

2.1 Significant Differences

One can assess if a contrast set is *significant* by testing the null hypothesis that contrast set support is equal across all groups or, equivalently, contrast set support is independent of group membership. The standard test for independence of variables in contingency tables is the chi-square test. It works by computing the statistic χ^2:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{3}$$

where O_{ij} is the observed frequency count for the cell in row i and column j . E_{ij} is the expected frequency count in cell ij given independence of the row and column variables and is calculated as follows: $E_{ij} = \sum_j O_{ij} \sum_i O_{ij}/N$ with N being the total number of observations. The result is then compared to the distribution of χ^2 when the null hypothesis is true. If the observed frequencies follow a multinomial distribution and the expected values are not too small, then the χ^2 statistic has an approximately chi-square distribution. Equation (1) can be read as the alternative hypothesis for the χ^2 test.

To determine if the differences in proportions are significant, a test α level is chosen. The choice of α sets the maximum probability of rejecting the null hypothesis when it is true. For a single test, α is commonly set to 0.05. Considering table 3 as our example, we calculate $\chi^2 = 5.09$ with 1 degrees of freedom which has a p-value of 0.024. Since the p-value is less than the 0.05 cutoff, it can be inferred that the null hypothesis is likely false and that contrast set support and group membership are not independent.

2.2 Controlling search error

With a single test, α sets the maximum probability of falsely rejecting the null hypothesis. It is well known that when performing multiple tests, the probability of false rejection can be highly inflated. This is particularly true in data mining, where a large number of hypotheses (in a scale of millions) are tested. For example, if the null hypothesis is always true and we made 1000 tests each at $\alpha = 0.05$, we would obtain on average 50 "significant" differences. Falsely rejecting the null hypothesis, i.e., concluding that there is a difference when none exists, is known as a Type I error or false positive. Type I error can be controlled for a family of tests by using a more restricted α cutoff for the individual tests. The α_i levels used for each individual test can be related to a global α (the expected error rate) by using the Bonferroni inequality: given any set of events $e_1, e_2, ..., e_n$, the probability

of their union ($e_i \vee e_2 \vee ... \vee e_n$) is less than or equal to the sum of the individual probabilities. Applied to hypothesis testing, we let e_i be the rejection of the ith hypothesis h_i. Consequently, h_i is rejected if $p_i \leq \alpha_i$ where $\sum \alpha_i \leq \alpha$. Usually $\alpha_i = \alpha/n$, where n is the total number of tests.

To solve the problem of incrementally reporting patterns after a level is mined and to discriminate patterns by their size (which is related to test power), STUCCO uses a specific Bonferroni adjustment. Since the Bonferroni method holds as long as $\alpha \geq \sum_i \alpha_i$, a different α for tests at different levels of the search tree can be defined as:

$$\alpha_l = min(\frac{\alpha}{2^l}/|C_l|, \alpha_{l-1}) \tag{4}$$

where α_l is the cutoff at level l and $|C_l|$ is the number of candidates (number of hypothesis evaluated) at level l. Notice that we have at most $\alpha_l \leq \alpha/(2^l \times |C_l|)$. Since $|C_l|$ does not include the entire search space of contrast sets (only accounts for candidates) from which those to be evaluated were selected, this method does not enforce the desired upper bound of α on this risk of false discoveries.

2.3 Pruning

STUCCO prunes contrasts sets that are deviations but are clearly not interesting by using two constraints:

- Allow only specialization of a contrast set cs if the support is different from the support of cs. Specializations with the same support often represent trivial findings.
- Prune specializations of cs that yield group support distribution similar to cs. For instance, assume a group with a much higher support for cs then all the others. Assume also that, independently of the items added to cs, this group preserves the large support difference to the others. Thus, the contrast set specialization should be discarded since the relation between groups appears to be fixed.

STUCCO implements these two ideas using the following equations:

$$\exists i \; P(csG|G_i) \neq P(csS|G_i) \tag{5}$$

$$max_i|sup(csG, G_i) - sup(csS, G_i)| \geq \delta_s \tag{6}$$

being $csG \subseteq csS$. Notice yet another user provided parameter δ_s. Equation (5) requires an extra χ^2 test. STUCCO reports contrast set in a level by level mode. Thus, a more specific cs is reported if it is surprising in relation to previously shown patterns. STUCCO estimates the probability of a conjunction based on its subsets and from this the expected frequency counts is obtained. A more specific cs is reported if the expected count are different (following equations (5) and (6)) from the observed counts. STUCCO makes use of sophisticated (and computationally expensive) techniques like *Iterative Proportional Fitting* (IPF) [7] to obtain the maximum likelihood estimates.

3 The RCS approach

Our proposal redesigns an association rules engine to derive rules describing contrast sets. In spite of making use of a specific implementation [1], any frequent itemset mining algorithm capable of deriving rules along the search process could be adopted to derive rules describing contrasts. Efficient algorithms exhibiting these features are described in [8]. To determine when a contrast set is significant, we adapt Webb's approach for significant association rules derivation (described in [16], [17] and [18]) to the problem of finding meaningful differences across groups. We also use the notion of support within a group, as in STUCCO.

To describe meaningful differences we use rules composed of an itemset in the antecedent and a list of pairs of groups in contrast in the consequent. The antecedent represents the contrast set. Each pair in the consequent exhibits the directional difference between two groups.

Our algorithm does not perform a STUCCO like level by level reporting. Rather it explores the search space using a depth-first search framework, likewise other frequent itemset mining algorithms [8]. As soon as a frequent itemset is found it goes through a set of procedures to evaluate whether it can be reported as a contrast set. Instead of considering the set of groups against the contrast set candidate (itemset), as in STUCCO, each pair of groups jointly with the newly derived itemset is evaluated. This translates to a Fisher exact test to determine significance and a pruning procedure to determine whether a specialization of a contrast set provides an added value on the differences between two groups. Our algorithm also verifies preservation of support. That is, when a specialization of an itemset preserves support it is discarded and not considered for contrast set evaluation. This pruning

is obtained by embedding the *Parent Equivalence Pruning* technique (PEP) [5] into our contrast sets mining algorithm.

A depth-first approach exhibits certain advantages. Although, this type of algorithms does not fully explore the downward closure property of support, it leads to an efficient rule based algorithm. Using such an approach is fundamental to achieve efficient solutions to contrast set mining. Actually, the reference to [8] might wrongly suggest that our approach is itemset based. In fact, our implementation is towards proposals like OPUS search [15] and [20] or even a more recent approach to discrimination rules like [12]. Thus, finding a frequent itemset is here to build the antecedent of a contrast rule i.e. the contrast set. Within the same algorithm the sufficient statistics describing the subpopulation of the rule are computed and the necessary statistical test of significance is performed.

Deriving rules yield important advantages in relation to itemset based algorithms. First, pruning along the same branch of the depth search tree can be performed considering the results of the statistical tests. When deriving a rule that satisfies all statistical demands one can determine whether any specialization of that rule will also pass these tests. In the negative case, the branch of the depth-first tree can be pruned and a considerable amount of redundant computation that would occur in an itemset based algorithm is avoided. Further, a rule based algorithm, like the one in [12], is almost insensitive to the number of groups (databases) where contrast sets are to be found.

Since our approach is association rules based, the data is concentrated in a single repositorium, contrasting with STUCCO which expects different data sources for different groups. It also can process data in attribute/value format: tabular data where rows are records and columns represent attributes values, and in basket format: a set of transactions where each transaction contains a variable number of items. In this way, groups are represented by attribute values or by items, depending on the data being on attribute/value or basket format. Thus, a group is a set of records (transactions) where a certain attribute value (item) occurs. Typically, the algorithm derives contrast sets rules considering the attribute (or set of items) supplied by the user.

3.1 Describing Contrast Sets

Rules for describing contrast sets across groups are formally defined as:

$$G_1 >> G_2, ..., G_i >> G_j \leftarrow cs \tag{7}$$

where cs is the itemset representing the contrast set and $G_1, G2, ..., G_i, G_j$ the list of groups. The list of pairs of the form $G_x >> G_y$ in the consequent represent the several directional differences between groups.

Consider the following example from the *Adult* dataset with 3 groups - BSc, MSc, PhD:

```
Gsup = 0.17191 | 0.04121  p = 1.1110878451E-017 education=Doctorate >> education=Masters
Gsup = 0.17191 | 0.01681  p = 3.0718399575E-040 education=Doctorate >> education=Bachelors
Sup(CS) =  0.03097                                          <---   workclass=State-gov  &
                                                                   class > 50K.
```

The rule is to be read as: *the occurrence of the contrast set "working for the state government and making an income of more than 50K" is significantly larger within people holding an PhD than a MSc. A significant difference in the same direction also occurs between PhD and BSc holders.* This type of rules present an important advantage in relation to STUCCO reported interest patterns.

For each contrast set cs, the pairs of groups were cs represents a significant difference are individually reported. For each contrast, the support within groups and the Fisher test p-value is displayed. A measure of association for cs/groups and the absolute occurrence of the cs within groups are also reported, but not shown in this figure. The support of the cs (3.09%) within the dataset is also displayed. Although the above pattern is displayed as a single rule, all pruning processes and statistical tests will consider it as two separate rules (a rule for each contrast).

For the following pattern reported by STUCCO[1] (describing absolute and relative group support, degrees of freedom, the χ^2 statistics and p-value):

```
hours_per_week = ]20.6:40.2]
2880 857 161 | 0.537815 0.497388 0.389831
============================
d.f.    chi^2    pvalue
2       38.37    4.65e-09
============================
```

[1]Along the paper, we present tables that are the actual output derived by STUCCO

one could be tempted to conclude that there is a contrast between BSc, MSc and PhD holders. However, the equivalent rule is

```
Gsup  = 0.53782 | 0.38983 p = 4.2124973374E-009 education=Bachelors >> education=Doctorate
Sup(CS) =  0.52036                                        <--- hours_per_week= ]20.6:40.2]
```

which is more informative since it specifies that the only *relevant* contrast is between BSc and PhDs.

3.2 Detecting differences across pairs of groups

To determine differences among contrasting groups i.e. to enforce eq (1), we apply a Fisher test to itemset occurrence across groups. Contrary to STUCCO, we analyze each pair of groups individually instead of the entire groups set. This enables the user to spot differences from one group to another instead of just reporting that there is a significant difference in proportions across n groups (being $n > 2$). These observations can be represented by 2×2 contingency tables like table 1. Fisher Exact test is a directional (one-sided) test to determine on a contingency table whether observed proportions are significant. The p-value is computed as follows:

$$p = \sum_{i=0}^{min(b,c)} \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{(a+b+c+d)!(a+i)!(b-i)!(c-i)!(d+i)!} \tag{8}$$

which is basically the computation of the sum of the probabilities of more extreme (or as extreme) contingency tables than the observed one. More extreme tables are obtained by increasing the cells along the diagonal (a and d) in table 1 and decreasing the cells off the diagonal (b and c). There are exactly $min(b, c)+1$ of such tables. Fisher test is exact, one-sided and suitable for small samples which is the appropriate for this application, whereas χ^2 is an approximate, two-sided test and unreliable for small samples. In our case, the cells in table 1 are $a = sup(cs, G_i)$, $b = sup(cs, G_j)$, $c = sup(\neg cs, G_i) \equiv sup(G_i) - sup(cs, G_i)$ and $d = sup(\neg cs, G_j) \equiv sup(G_j) - sup(cs, G_j)$, where $sup(G_i)$ is the number of examples (support) representing group G_i. Instead of using the traditional Stirling's approximation, our Fisher test implementation derives a more accurate p-value by using a table with pre-computed results for the log factorial function.
For contrast set mining, the null hypothesis is

$$H_0 : P(cs|G_i) \leq P(cs|G_j) \tag{9}$$

The alternative hypothesis expresses that the proportions observed in our contingency table are not a chance artifact of the sample data. In this way, instead of eq. (1) we use $\exists i, j\ P(cs|G_i) > P(cs|G_j)$.

3.3 Controlling the report of false discoveries

We adapt the proposal for layered critical values proposed in [18] to control the extraction of false contrast set rules. This is achieved by deriving a more stringent cutoff value to prevent type I errors. Webb reuses the adjustment described in equation (4) to derive a new α_i cutoff for each level (length in number of items for the antecedent of an association rule). The idea is to create a Bonferroni-like adjustment when the number of tests is not known in advance. Notice that the adjustment used in [4] considers the number of hypothesis evaluated rather than the size of the search space. That is, only accounts for candidate patterns that satisfy a set of constraints where eq. (2) is included. However, candidates are the patterns most likely to pass the statistical test. Thus, the critical value should be adjusted by the number of patterns from which those to be tested are selected instead of the number of times the statistical test is applied.

In [17] an upper limit is required on the length of the antecedent of an association rule, in our case contrast set length, to derive the search space size (the number of potential rules). Since a minimal support constraint is used, the size of the search can be calculated in advance either because the user provides an upper limit or the number of frequent items is used as the maximal contrast set length. The search space size is usually much smaller at the lower contrast set lengths. However, we do not want to disproportionately weight the available critical value mass toward the smaller lengths. In consequence we use

$$\alpha'_L = \alpha/(L_{max} \times S_L) \tag{10}$$

where L is the level of the search space, S_L is the number of rules at level L and L_{max} is the maximum value of L. The latter can either be a user provided value or the number of frequent items. The aim of this formula is to ensure that $\alpha \geq \sum_{L=1}^{L_{max}} \alpha'_L \times S_L$

To calculate S_L (the number of contrast set rules) for basket-format data, it is enough to compute:

$$S_L = cons \times \sum_{i=1}^{maxx} \left(\begin{array}{c} m - cons \\ i \end{array} \right) \tag{11}$$

where m is the number of items in the dataset, G the number of items describing groups and $maxx$ is the upper bound (in number of items) on the contrast set size. A contrast set has at least one item. $cons = \begin{pmatrix} G \\ 2 \end{pmatrix}$ is the number of combinations containing 2 out of G group items. In attribute/value-format data, the number of attributes (excluding the group attribute) is used for L_{max}. In this data format S_L can be computed as:

$$S_L = cons \times \sum_{j=1}^{maxx} C_{att \setminus \{g\},j,m} \tag{12}$$

being g the group attribute and att the set of attributes present in the dataset. Assuming that att_k is the number of values the attribute k contains and that $C_{att,j,k}$ represents the number of combinations of up to j items where items contain only values for attributes $att_1, ..., att_k$, the number of itemsets that potentially define a contrast set is then:

$$C_{att,j,k} = \begin{cases} \#att_k, & j = 1, k = 1 \\ 0, & j > 1, k = 1 \\ C_{att,1,k-1} + \#att_k, & j = 1, k > 1 \\ C_{att,j,k-1} + (\#att_k \times C_{att,j-1,k-1}), & otherwise \end{cases} \tag{13}$$

3.4 Pruning Non Relevant Rules

An important sub-task for this mining process is to determine which specializations of a contrast set are relevant. STUCCO relies on expected frequencies to report specifications of a contrast set. Other proposals like [9] only consider specializations that yield a ϕ-coefficient improvement, where ϕ is a measure of association between the contrast sets and the pair of groups. We argue that a specialization of a contrast set cs should only be reported (and considered relevant) when it yields a significant improvement on the differences across the pair of groups. This can occur when the specialization is a contrast in the same direction as the general cs and improves that contrast, or when it contradicts the direction of the general cs. Thus, a contrast set must be checked against all its generalizations, including the contrast set $\emptyset$.

For this task we will consider rules for contrast sets as association rules and perform a test of significance improvement between a rule and all its generalizations, as proposed in [17][2]. In this way, eq. (5) is replaced by the

[2]Actually, Webb only compares rules against direct generalizations i.e. $y \leftarrow \emptyset$ and $y \leftarrow x - z$ where z is a single item, are direct generalizations of $y \leftarrow x$.

alternative hypothesis $\exists i\ P(cs|G_i) > P(cs - z|G_i)$.

We test a rule $G_i >> G_j \leftarrow cs$ with level 0 generalization i.e. a rule with empty body, being $sup(G_i) = n_i$ and $sup(G_j) = n_j$ (assuming $n_i >> n_j$). A contrast set cs is relevant if it contradicts $G_i >> G_j$ for $\emptyset$ or observations in table 2 are significant. Both conditions will be captured by a Fisher exact test.

Notice, however, that this table contains the same observations as in table 1. Consequently, checking for improvement between a contrast set cs and $\emptyset$ was already obtained (as described in section 3.2) when the validity of the contrast set cs was verified.

For level $j > 0$, cs is relevant if $\forall cs - z \subset cs$, (where z is an itemset) either $G_i >> G_j$ contradicts direction in $cs - z$ or observations in table 4 are significant. Again, the Fisher exact test will capture both conditions. Both tables 2 and 4 contain independent (disjoint) observations in the four cells.

As in section 3.3, these statistical tests use an adjusted cutoff value to prevent type I errors. Although for each candidate rule n tests are performed (being n the number of generalizations for that rule), it is the number of rules in the search space to be considered for the adjustment instead of the number of hypothesis used. The reason for this is because, for one rule, we are only checking whether any of the hypothesis (from the disjunction $H_{0,1} \vee H_{0,2} \vee ... \vee H_{0,n}$) is violated. Actually, in this case, when multiple hypotheses are tested to assess only whether any does not hold, it is the risk of type II rather than type I error that is increased by the multiple hypothesis testing.

An efficient implementation of this pruning process is achieved by making use of a specific data structure to represent rules [2]. Our contrast rules are derived along the search process. Consequently, due to the use of a depth-first search, when deriving a contrast set not all its subsets are available. Thus, our pruning process implementation is not guaranteed to be complete. However, our approach guarantees the enforcement of:

$$\begin{array}{l} \forall cons \leftarrow cs \in ResultSet,\ \forall x \subset cs\ \wedge\ cons \leftarrow x \in ResultSet, \\ Fisherpvalue(cons \leftarrow x, cons \leftarrow cs) \leq \alpha'_L \end{array} \tag{14}$$

where $ResultSet$ is the set of contrast rules returned at the end of the mining process and $L = length(cs)$. Hence, completeness can only be ensured

by a post-pruning process. The same problem was identified in [19] using Magnum Opus. It was reported when using a Binomial sign test to derive association rules representing contrast sets.

Algorithm 1 summarizes our approach. In line 3, each node of the depth-first tree is considered i.e. each potential antecedent. The algorithm discards any antecedent preserving parental support (PEP technique). Then, in lines 4-5, each pair of groups together with the candidate contrast (antecedent i) is processed. A Fisher exact test is performed to determine significance of the contrast within the pair. In the positive case, lines 5-8, relevance pruning is applied by checking whether the candidate rule is significant in relation to all its generalizations. Finally, in line 7, relevant rules are added to the result set.

input : dataset D, list of groups G, minsup ms, cutoff α
output: ResultSet of contrast rules RS
1 Compute $\alpha'_1, \alpha'_2, ..., \alpha'_n$, for a supplied n or using number of frequent_items/attributes;
2 $RS := \emptyset$;
3 **foreach** *node* $i \in depth\_first\_search(D, ms) : sup(i) \geq ms$ **do**
4 **foreach** *pair* $g_a, g_b \in G$ *where* $sup(i, g_a) >> sup(i, g_b)$ **do**
5 **if** $Fisherpvalue(g_a, g_b, i, \emptyset) \leq \alpha'_{length(i)}$ **then**
6 **if** $\forall g_a >> g_b \leftarrow cs \in RS$: $cs \subset i$ & $Fisherpvalue(g_a, g_b, i, cs) \leq \alpha'_{length(i)}$ **then**
7 $RS := RS \cup \{g_a >> g_b \leftarrow i\}$;
8 **end**
9 **end**
10 **end**
11 **end**

Algorithm 1: (RCS) Rules for Contrast Sets.

This approach avoids the problem of processing multiple pairwise combinations of datasets [12] i.e. to analyze each pair of groups which requires to process pairs of datasets.

The fact the algorithm is rule based enables a set of interesting computational savings. These are obtained by new opportunities of pruning. For instance, consider the rule $p_i >> p_j \leftarrow body$ with supports $sup(body, p_i) = n_i$ and $sup(body, p_j) = n_j$. The node in the search space corresponding to the antecedent of the rule is worth expanding (specialize) if the fisher test be-

tween this rule and a specialization $body \cup a$ where $sup(body \cup a, p_i) = n_i$ and $sup(body \cup a, p_j) = 0$ yields success. Since the latter artificial rule represents the greatest possible contrast, failing to pass the test enables to prune all expansions in the search space for this rule in this node.

4 Evaluation

All experiments were run on an Intel based PC with a Dual-core processor and 4 Gb of main memory. STUCCO was run using the default parameters setting. In all datasets, *Caren* (version 2.5.2., which implements the RCS algorithm) was run with 1% minimal support. Our implementation deals with both formats (attribute/value and basket data format) whereas STUCCO only deals with attribute data.

Datasets described in table 5 were obtained from the UCI Machine Learning repository. Datasets *adult*, *mushroom* and *ipums* were also used for evaluation in [4] and [9]. In the *adult* dataset, numeric attributes were preprocessed using equi-width discretization with 5 bins. Groups are defined by a subset of 3 values from the attribute `education`. In the *ipums* equi-depth was used with 10 bins as the maximal number of intervals. Groups represent collection of federal census for the years 1970, 1980 and 1990. This data set contains a 1 in 1000 sample of the Los Angeles and Long Beach area from the original data. A constraint of maximal contrast sets length of 5 items was imposed for this dataset. For length larger than 5 items STUCCO runs out of memory.

According to the definition of deviation, interest and relevance used by STUCCO and RCS, one would expect a small set of common patterns derived by both algorithms (see table 6). There are at least four reasons for this expectation:

- RCS cutoff adjustment is more stringent than STUCCO adjustment,
- Fisher test is directional whereas χ^2 is two-sided test. This tends to yield, in the majority of situations, a bigger p-value for χ^2 in datasets with 2 groups. For more than 2 groups the inverse tends to occur,
- STUCCO uses $2 \times G$ tables where RCS always uses a set of 2×2 contingency tables,
- Relevance pruning is more robust than STUCCO interest pruning.

For the *adult* dataset, the set of contrast sets reported by RCS only covers 23 patterns discovered by STUCCO. All of these patterns are length 1. Since RCS uses a more restricted adjusted cutoff ($\alpha_1' = 1.10229E^{-5}$), patterns `marital.status` = `Divorced`, `occupation` = `Tech_support`, `sex` = `Male` and `native_country` = `United_States` are not derived. For length larger than 1, no patterns coincides. However, STUCCO strangely derives the following redundant pattern (which contradicts example 4 in [4]):

```
relationship = Husband AND sex = Male
2433 886 265 | 0.454342 0.514219 0.641646
============================
d.f.    chi^2    pvalue
2       65.39    6.33e-15
============================
```

It also derives `relationship` = `Husband`, which occurs exactly with the same proportions as the pattern above. Our algorithm discards the former because it occurs with the same support as the general. STUCCO discards the redundant pattern if δ_s of eq. (6) is raised to 0.06 (reporting only 30 interesting patterns). However it still derives patterns like

`relationship` = `Husband` `AND` `race` = `White` `AND` `sex` = `Male`

and also

`age` =]46.2 : 60.8] `AND` `relationship` = `Husband` `AND` `sex` = `Male`.

The largest contrast set derived by RCS is:

```
Gsup  = 0.14286 | 0.00598 p = 2.3517318499E-046 education=Doctorate >> education=Bachelors
Sup(CS) =  0.01882                                <---  workclass=State-gov  &  class=>50K
                                                        &  occupation=Prof-specialty
```

which is not derived by STUCCO whose largest pattern is:

```
occupation = Adm-clerical AND workclass = Private AND class = <=50K
AND native_country = United-States
372 51 6 | 0.0694678 0.0295995 0.0145278
============================
d.f.    chi^2    pvalue
2       53.17    2.85e-12
============================
```

This pattern is not derived by RCS because, for instance, the p-value for the contrast `BSc` >> `PhD` is $2.5360E^{-9} > \alpha_4' = 6.03433E^{-10}$.

For datasets with more than 2 groups, RCS tends to derive less patterns than STUCCO. The first reason is because RCS uses a more stringent cut-off. This is a consequence of the adjustment described in eq. (10) which

uses the rule's space instead of number of statistical tests performed. For instance, in *ipums*, STUCCO uses $\alpha'_1 = 5.93824E^{-05}$, $\alpha'_2 = 4.682E^{-07}$, $\alpha'_3 = 1.28348E^{-08}$, $\alpha'_4 = 8.37733E^{-10}$ and $\alpha'_5 = 8.80848E^{-11}$. RCS uses $\alpha'_1 = 8.13008E^{-6}$, $\alpha'_2 = 4.04668E^{-8}$, $\alpha'_3 = 3.08378E^{-10}$, $\alpha'_4 = 3.19915E^{-12}$ and $\alpha'_5 = 4.23705E^{-14}$. The second reason is because the pruning applied by RCS is more effective than STUCCO notion of interest. e.g. for the *adult* dataset STUCCO derives 5486 deviations (before interest pruning) where RCS derives 72400 patterns when relevance pruning is not used.

Exceptions occur in the *lympho* and *zoo* datasets. Zero interest patterns found by STUCCO at these datasets are explained by the very small size of some groups, yielding expected cell counts less than 3 in a 2×4 and 2×7 contingency tables, which disables the use of the χ^2 statistical test.

For the *mushroom*, since it is a 2 groups dataset (implies 1 degree of freedom), the obtained p-values by STUCCO and RCS should be approximately similar. Consider the following two examples of STUCCO and equivalent RCS patterns:

```
Gsup  = 0.02860 | 0.00000  p = 1.3866570410E-036  CLASS=p >> CLASS=e
Sup(CS) =  0.01379                                   <---    HABIT=u &
                                                             CCOLOR=w
                    HABIT = u AND CCOLOR = w
                    0 112 | 0 0.0286006
                    ============================
                    d.f.    chi^2    pvalue
                    1       122.03    2.27e-28
                    ============================

Gsup  = 0.17110 | 0.08172  p = 2.1817481460E-034  CLASS=e >> CLASS=p
Sup(CS) =  0.12802                                   <---    CCOLOR=w
                    CCOLOR = w
                    720 320 | 0.171103 0.081716
                    ============================
                    d.f.    chi^2    pvalue
                    1       145.18    1.96e-33
                    ============================
```

As these examples illustrate, the χ^2 test tends to yield bigger p-values than the Fisher exact test. This is because χ^2 is a two-sided test. For instance, Yates' correction for the χ^2 test correlates well with the two-sided p-value Fisher test [7] e.g. for $\texttt{HABIT} = \texttt{u}$, STUCCO p-value is $5.41E^{-24}$, RCS p-value is $\texttt{8.75651E}^{-025}$ and the two-sided Fisher test p-value is $\texttt{1.15713E}^{-24}$. Analogous phenomena are observed in datasets *flare*, *tic-tac-toe* and *house-votes*.

This is also a good example to illustrate the information gain obtained with RCS rules. The first rule is a specialization of the second but where

direction of contrast is inverted. This detail is much more difficult to grasp when analyzing STUCCO patterns, mainly when more than 2 groups are assessed.

For dataset with more than 2 groups, STUCCO patterns tend to have smaller p-value since the χ^2 test is performed on $2 \times G$ tables rather than sets of 2×2 tables e.g. for $\texttt{school} = \texttt{0} \,\&\, \texttt{inctot} =]-\texttt{2.0} : \texttt{0.0}]$ in *ipums* STUCCO p-value is $5.44E^{-156}$ and RCS only contrast is $\texttt{year} = \texttt{1990} >> \texttt{year} = \texttt{1970}$ with $\texttt{p} = \texttt{2.31170E}^{-111}$. This discrepancy combined with a more conservative cutoff value explains the difference in number of derived patterns by RCS and STUCCO in this dataset.

Interestingly, although *Caren* is a JAVA based implementation, it tends to be faster for datasets with long patterns than STUCCO e.g. *ipums*. Part of this performance is explained by the fact that STUCCO makes use of the expensive IPF technique and a proper p-value calculation for the χ^2 (not just table lookup). The experiments with *ipums* also suggest that STUCCO requires more memory than RCS due to its breath-first search approach.

Figure 1 shows running time of *Caren* along a range of different minimal supports using the *adult* dataset with 3 groups. The algorithm scales well with minimal support variation.

Table 7 describe results for an experiment using different variants of the UCI *adult* dataset. The purpose is to show group scalability and the effect of dataset size and number of rules derived on runtime performance. The first dataset contains groups *PhD* and *BSc*, the second these two groups plus group *MSc*, and finally the last dataset contains all seven groups. RCS algorithm scales well with number of groups. However it is clearly bounded by dataset size and number of derived rules.

Table 7 exhibits an interesting phenomenon on number of derived rules reduction along number of groups. For instance, from 2 to 3 groups the number of derived rules reduces from 43 to 41. There are several reasons that explain this effect: First, minsup is given as a relative value (1% in this case). When the number of groups is increased, the relative support will drop and it can be the case that the contrast set support no longer satisfy the minimal support threshold. Secondly, the adjusted cutoff value tends to drop as the number of groups rises. This is because the number of rules associated to the search space (S_L in eq. (10)) tends to increase. For instance, in *adult* with 2 groups $\alpha_2' = 7.720E^{-07}$ whereas with 3 groups $\alpha_2' = 2.480E^{-07}$ and with 7 groups $\alpha_2' = 3.328E^{-08}$ Finally, when pruning non relevant rules these adjusted cutoff values are also used. Thus, a rule

can be relevant in relation to all its generalization in one dataset but non relevant in the next dataset where an extra group is added.

An alternative experiment was performed using different variants of the previous datasets. All datasets are now size 1000 records and are obtained by equal size sampling from each group. In this setting, group distribution is uniform which dissipates the previous effect of different groups size increments. Table 8 describes these results in number of derived rules. Figure 3 displays runtime performance for this experiment on number of groups variation. Our algorithm scales well on this factor. However, it is the number of rules derived that bounds runtime performance as shown in figure 2 (linear interpolation yields a higher slope on data from figure 2 than on figure 3).

5 Related Work

Several proposals are described in the literature. CIGAR [9], reuses STUCCO algorithm embedding new constraints and new concepts for deviation and interest. It reuses equations (1) and (2) but also adds constraints on minimal group support for a contrast set and minimal correlation. Correlation is measured using the phi-coefficient. Using table 1, ϕ is defined as:

$$\phi = \frac{(a \times c) - (b \times d)}{\sqrt{(a+c)(b+d)(a+b)(c+d)}} \tag{15}$$

Having $\phi = 0.0$ indicates independence. Values for $\phi \leq 0.29$ suggest weak or no association. The authors in [9] propose the minimal value for correlation to be fixed as 0.3. We have implemented this constraint in our algorithm using the suggested threshold. However, it turned out to be highly conservative yielding as few as 3 patterns for the *adult* dataset. Specialization of a contrast set is derived if it yields an improvement on the ϕ value. A minimal correlation threshold is required. This feature introduces an additional difficulty for the user since the value for ϕ largely restrains the number of derived patterns. With these new constraints an extra number of 3 new parameters are required.

CIGAR replaces the chi-squared test by the Yates' correction. However, it does not implement any cutoff adjustment to cope with the derivation of false discoveries. Further, Yates' correction still does not specify direction of departure from H_0. In a similar way to our approach, CIGAR decomposes the $2 \times G$ contingency tables into 2×2 tables so that a specific contrast

between two groups can be reported. CIGAR tends to derive more contrast sets than STUCCO.

Emerging patterns is an association rules based approach to report differences among datasets. Also proposed in the same conference edition as STUCCO, Emerging Patterns [6] are patterns that capture differences between two databases. The pattern ep is an Emerging Pattern if the $growth\_rate(ep) = \frac{sup_1(ep)}{sup_2(ep)} \geq \rho$, where ρ is a user provided threshold. The original proposal was designed to derive patterns for two groups (databases). In [12] an extension to handle emerging pattern in multiple databases (groups) is proposed. This proposal directly generates δ-discriminative equivalence classes which avoids redundancy that occurs when deriving jumping emerging patterns (itemsets that occurs in one group but never in the others).

In [14], a survey relating contrast set mining, subgroup mining and emerging pattern mining is presented. This paper contributes towards a novel understanding of these subareas of data mining by presenting a unified terminology and by describing the three tasks as variants of an unique supervised descriptive rule discovery task.

Group difference is also studied in [19]. Being an association rule approach, it tackles the problem by restricting the consequent of rules to be the attribute values that define group membership. A Binomial sign test is used to determine groups differences. However, according to [9], this proposal only performs a within-groups comparison rather than a between-groups comparison.

Subgroup mining can be also seen as a form of contrast finding. Example is [10] where numeric properties of interest are analyzed instead of categorical groups.

6 Conclusion

The proposal presented in this paper seems to corroborate the claim in [19] that a redesigned association rules engine is capable of efficiently tackle the problem of deriving contrast sets. Our approach requires less parameters from the user when compared to STUCCO. This is a desirable feature since it leads towards a parameter-free data mining. A large number of parameters tends to confuse any non-expert user. We make use of the Fisher Exact test instead of the traditional χ^2 approach to check independence between contrast sets and groups. First, for testing significance on differences across

groups one requires a one-tail test, whereas the χ^2 test is a two-tail test. Fisher test yields results indicating departure from the null hypothesis in a specific direction, whereas the χ^2 assesses departures in both directions. This χ^2 characteristic can lead to type I or type II errors. Also, the χ^2 test is an approximated test and known to be unreliable for small counts.

The rules reported by our algorithm describe the specific contrasts where differences actually occur. For instance, STUCCO relies on the differences among a set of groups, not specifying which groups contrast occur. The RCS can be redesigned to report the top-k patterns where, for instance, only the top-k lowest p-value contrast sets rules are derived.

7 Repeatability Assessment

Caren (version 2.5.2) with the RCS implementation is available at [1]. Each dataset version, used for RCS and STUCCO, are also available at this site.

8 Acknowledgments

Thanks to Prof. M. Pazzani for kindly providing the code for the STUCCO algorithm.

References

[1] P. J. Azevedo. Caren - class project association rule engine. http://www.di.uminho.pt/∼pja/class/caren.html.

[2] P. J. Azevedo. A data structure to represent association rules based classifiers. Technical report, DI, Universidade do Minho, 2005.

[3] S. D. Bay and M. J. Pazzani. Detecting change in categorical data: Mining contrast sets. In *KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 302–306, 1999.

[4] S. D. Bay and M. J. Pazzani. Detecting group differences: Mining contrast sets. *Data Min. Knowl. Discov.*, 5(3):213–246, 2001.

[5] D. Burdick, M. Calimlim, J. Flannick, J. Gehrke, and T. Yiu. Mafia: A maximal frequent itemset algorithm. *IEEE Trans. Knowl. Data Eng.*, 17(11):1490–1504, 2005.

[6] G. Dong and J. Li. Efficient mining of emerging patterns: Discovering trends and differences. In *KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 43–52, 1999.

[7] B. Everitt. *The Analysis of Contingency Tables*, volume 45 of *Monographs on Statistics and Applied Probability*. Chaoman & Hall-CRC, 1997.

[8] B. Goethals and M. J. Zaki, editors. *FIMI '03, Frequent Itemset Mining Implementations, Proceedings of the ICDM 2003 Workshop on Frequent Itemset Mining Implementations, 19 December 2003, Melbourne, Florida, USA*, volume 90 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2003.

[9] R. J. Hilderman and T. Peckham. Statistical methodologies for mining potentially interesting contrast sets. In F. Guillet and H. J. Hamilton, editors, *Quality Measures in Data Mining*, volume 43 of *Studies in Computational Intelligence*, pages 153–177. Springer, 2007.

[10] A. M. Jorge, P. J. Azevedo, and F. Pereira. Distribution rules with numeric attributes of interest. In *Knowledge Discovery in Databases: PKDD 2006, 10th European Conference on Principles and Practice of*

Knowledge Discovery in Databases, Berlin, Germany, September 18-22, 2006, Proceedings, pages 247–258, 2006.

[11] P. Kralj, N. Lavrac, D. Gamberger, and A. Krstacic. Contrast set mining for distinguishing between similar diseases. In R. Bellazzi, A. Abu-Hanna, and J. Hunter, editors, *AIME*, volume 4594 of *Lecture Notes in Computer Science*, pages 109–118. Springer, 2007.

[12] J. Li, G. Liu, and L. Wong. Mining statistically important equivalence classes and delta-discriminative emerging patterns. In P. Berkhin, R. Caruana, and X. Wu, editors, *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, California, USA, August 12-15, 2007*, pages 430–439. ACM, 2007.

[13] J. Lin and E. J. Keogh. Group sax: Extending the notion of contrast sets to time series and multimedia data. In *Knowledge Discovery in Databases: PKDD 2006, 10th European Conference on Principles and Practice of Knowledge Discovery in Databases, Berlin, Germany, September 18-22, 2006, Proceedings*, pages 284–296, 2006.

[14] P. Novak, N. Lavrac, and G. Webb. Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. *Journal of Machine Learning Research*, 10:377–403, 2009.

[15] G. I. Webb. Efficient search for association rules. In *KDD*, pages 99–107, 2000.

[16] G. I. Webb. Discovering significant rules. In T. Eliassi-Rad, L. H. Ungar, M. Craven, and D. Gunopulos, editors, *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, August 20-23, 2006*, pages 434–443. ACM, 2006.

[17] G. I. Webb. Discovering significant patterns. *Machine Learning*, 71(1):131, 2008.

[18] G. I. Webb. Layered critical values: a powerful direct-adjustment approach to discovering significant patterns. *Machine Learning*, 71(2-3):307–323, 2008.

[19] G. I. Webb, S. M. Butler, and D. A. Newlands. On detecting differences between groups. In L. Getoor, T. E. Senator, P. Domingos, and

C. Faloutsos, editors, *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 24 - 27, 2003*, pages 256–265. ACM, 2003.

[20] G. I. Webb and S. Zhang. K-optimal rule discovery. *Data Min. Knowl. Discov.*, 10(1):39–79, 2005.

Table 1: A generic Contingency table for two groups and a contrast set.

	G_i	G_j	$\sum row$
cs	a	b	a + b
$\neg cs$	c	d	c + d
$\sum column$	a + c	b + d	n

Table 2: Contingency table for cs versus $\emptyset$

	G_i	G_j	$\sum row$
cs	$sup(cs, G_i)$	$sup(cs, G_j)$	
$\emptyset$	$n_i - sup(cs, G_i)$	$n_j - sup(cs, G_j)$	
$\sum column$	n_i	n_j	$n_i + n_j$

Table 3: Contingency table for example $Stress \times Location$.

	Location=urban	Location=rural	$\sum row$
$Stress = high$	194	355	549
$\neg(Stress = high)$	360	511	871
$\sum column$	554	866	1420

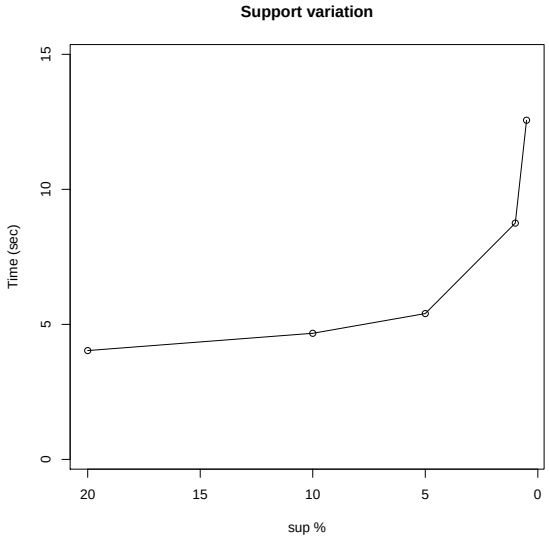

Figure 1: Scalability on support

Table 4: Contingency table for cs versus csg

	G_i	G_j	$\sum row$
cs	$sup(cs, G_i)$	$sup(cs, G_j)$	
$cs - z$	$sup(cs - z, G_i) - sup(cs, G_i)$	$sup(cs - z, G_j) - sup(cs, G_j)$	
$\sum column$	$sup(cs - z, G_i)$	$sup(cs - z, G_j)$	$sup(cs - z, G_i) + sup(cs - z, G_j)$

Table 5: Datasets

Name	Tuples	Attributes	Items	Groups
mushroom	8142	23	118	2
flare	1066	11	33	2
house-votes	435	17	34	2
tic-tac-toe	958	10	29	2
adult	7491	15	111	3
lympho	148	19	63	4
zoo	101	17	43	7
ipums	23348	61	413	3

Table 6: Summary of experiments with RCS and STUCCO (number of derived patterns and runtime)

Dataset	#RCS	#STUCCO	Time(RCS)	Time(STUCCO)
mushroom	1819	1588	12s29	9m8s02
flare	21	20	2s60	0s02
house-votes	74	44	3s54	0s05
tic-tac-toe	34	9	2s910	0s02
adult	41	47	8s75	2s76
lympho	11	0	2s20	0s02
zoo	28	0	1s80	0s01
ipums	11604	25241	477m0s	1022m7s

Table 7: Experiments showing scalability with number of group using the *adult* dataset.

#Groups	Time	#Rules	#Dataset	*education*
2	5s	43	5768	PhD+Bsc
3	9s	41	7491	+Msc
4	11s	68	7942	+12th
5	25s	136	18425	+HS-grad
6	34s	143	19001	+Prof-School
7	44s	135	20068	+Assoc-acdm

Table 8: Number of group versus time versus number of rules using the *adult* dataset with fixed size 1000 records.

#Groups	Time	#Rules
2	2.850	6
3	3.082	16
4	3.747	18
5	3.972	24
6	4.476	27
7	5.207	28

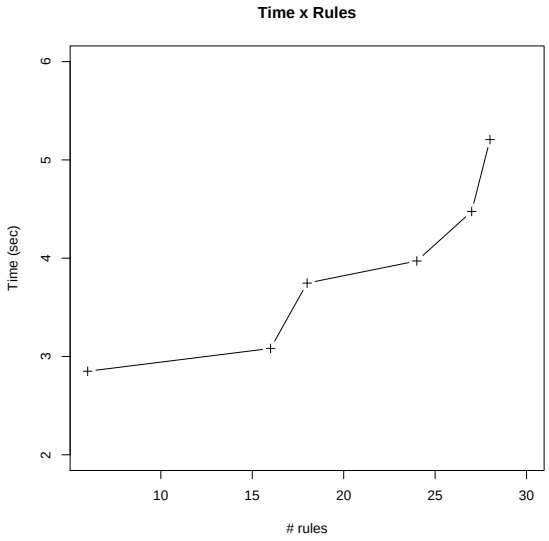

Figure 2: Scalability on number of rules derived

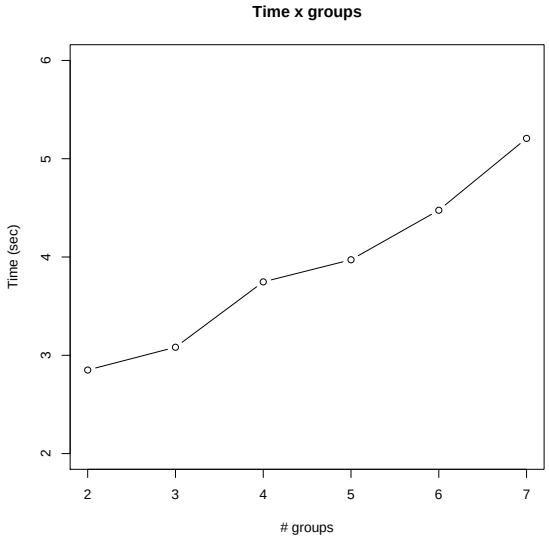

Figure 3: Scalability on number of groups