Contrasting the Contrast Sets: An Alternative Approach

Amit Satsangi
Department of Computing Science
University of Alberta, Canada
amit@cs.ualberta.ca

Osmar R. Zaïane
Department of Computing Science
University of Alberta, Canada
zaiane@cs.ualberta.ca

Abstract

The need to identify significant differences between contrasting groups or classes is ubiquitous and thus was the focus of many statisticians and data miners. Contrast sets, conjunctions of attribute-value pairs significantly more frequent in one group than another, were proposed to describe such differences, which lead to the introduction of a new data mining technique - contrast-set mining. A number of attempts have been made in this regard by various authors; however, no clear picture seems to have emerged. In this paper, we try to address the problem of finding meaningful contrast sets by using Association Rule based analysis. We present the results for our experiments for interesting contrast sets and compare these results with those obtained from the well-known algorithm for contrast sets-STUCCO.

1. Introduction

A commonly asked question for data analysis in any discipline is: "How can several contrasting groups be compared against each other?" Depending on the context this leads to specific questions like-which categories of students are more likely to accept an admission offer from a University? What are the specific characteristics that best differentiate between patients with a specific disease and normal patients? What distinguishes between the customers that buy more than some value and those that buy less than another threshold? What is the difference between male and female managers, all other things being equal? Do postgraduate degree holders fare better in their career, than those who hold only a bachelors degree?

The differences between the contrasting groups can be described in terms of conditional probabilities (i.e. the probability of a group given some conjunctions of attribute value pairs), such as:

P(Degree=Bachelors | Income=high ^ Position=Manager)= 34%,
and
P (Degree=Doctorate | Income=high ^ Position=Manager) = 43%

These conditional probabilities are actually equivalent to the two association rules given by:

Degree=Bachelors → Income=high ^ Position=Manager (34%),
and
Degree=Doctorate → Income=high ^ Position=Manager (43%)

where the percentages represent the support for the association rule within each group and the consequent is called a contrast[1] set.

Contrast set mining was introduced as emerging pattern mining [5] by Dong et al. using the framework of Association rule-based technique introduced in [3] by Bayardo et al. in their Max-Miner algorithm. The problem was independently researched by Bay et al. [1] in the context of statistical significance of contrast sets, by employing a Max-Miner like approach in the search space; a more detailed description and evaluation can be found in [2]. The authors propose an algorithm called STUCCO (**S**earching and **T**esting for **U**nderstandable **C**onsistent **CO**ntrasts) for determining the statistical significance of contrast-sets; they use a canonical ordering of nodes in the search space by using set-enumeration trees, and employ χ^2 testing of two-dimensional contingency tables, along with modified Bonferroni method for controlling Type I error (or false positives, i.e. finding only but not necessarily all significant contrast sets).

Later, Webb et al. [8] used rule-discovery techniques, in their algorithm called Magnum Opus, by

[1] The original definition of contrast set, as given in [1], has been modified here for reasons discussed later on.

employing a heuristic approach to cut down on the number of contrast sets in the search space. They conclude that contrast set mining is a special case of the more general rule-discovery task. Finally, Hilderman et al. [6] consider a different approach, whereby they employ three additional constraints to the STUCCO framework, and seek to control Type II error (or false negatives i.e. attempting to reduce the missed significant contrast sets). Using their algorithm called CIGAR (**C**ontrast**I**ng **G**rouped **A**ssociation **R**ules), the authors find a different lot of contrast-sets from that of STUCCO; they conclude that both STUCCO and CIGAR represent valid alternative solutions to the problem of identifying contrast sets.

2. Problem Definition

Any relation with observations defined on attributes can be translated into a set of transactions *D* such that each example E in *D* is described by a vector of m attribute-value pairs A1 = V1, A2 = V2, …Am = Vm; each Vi is selected from a finite set $\{Vi_1, Vi_2, \ldots Vi_n\}$ such that the elements of this set take only discrete values. One attribute in D is such that its value V_{jk} in E is used to assign E into one of n mutually exclusive groups G1, G2, …Gn. In [1] and [2] a contrast set X is defined as a conjunction of attribute-value pairs on G_1, G_2, G_n, such that no A_i occurs more than once.

Thus we get rules of the form $(A_j = V_{jk}) \rightarrow X$, where the antecedent determines the group membership, while the consequent is called a contrast set. Contrast set mining aims to identify all contrast sets for which the support is significantly different across groups. STUCCO achieves this end by imposing two constraints

$$\exists_{ij} P(X|G_i) \neq P(X|G_j) \qquad (1)$$

$$\max_{ij} \left| \text{support}(X, G_i) - \text{support}(X, G_j) \right| \geq \text{min\_dev} \qquad (2)$$

The support of a contrast set X for a group G_k, given by support(X,G_k) is the fraction of the examples in G_k where the contrast set is true. The first constraint (equation 1) is called the significance condition; it checks for the statistical significance of the contrast-set, the second condition (equation 2) is called the largeness condition; when both the conditions are met it is called a deviation. *min_dev* is a user defined threshold called the minimum support difference.

3. Problems with Related Work

In [1], [2] and [6] the contrast sets are reported as belonging to the association rules such as *Group* ➔ *contrast set* (for brevity we will, hence forth, refer to these kinds of contrast sets as the “first kind”). The authors do not consider other kind of contrast sets (henceforth referred to as the “second kind”) that come from the rules of the type *contrast set* ➔ *Group*. In [8], the authors consider only the second kind of contrast sets. Later, we show that only the second kinds of contrast sets exist. The second issue regarding the methodology involved in previous works is that there seems to be no consensus on the kind of filter to be used to prune the search space. In their concluding remarks in [8], the authors mention that neither STUCCO nor Magnum Opus applies a perfect filter, and that while STUCCO seemed to discard some contrasts of potential value, Magnum Opus appears to include contrast sets that were probably spurious, thus highlighting the inadequacy of the two approaches.

In [7], the authors prove that Magnum Opus actually does a within-groups comparison rather than a between-groups comparison and thus generates only a subset of the contrast sets generated by STUCCO. The claim made by Peckham *et al.* is true, at the same time it seems to be adding more confusion to the field because in [8], Webb *et al.* had claimed and reported results showing that Magnum Opus produced all the contrast sets generated by STUCCO and a few more interesting ones that STUCCO failed to produce. Thus it seems that all the approaches so far have been unable to tackle the root of the problem and there seems to be no agreement on this issue.

4. An Alternative Approach

Bay *et al.* state that employing association rules has three problems. First, there are too many rules to compare. Second, in some cases, there are rules in one group that have no match amongst any of the other groups. Third, even with matched rules, proper statistical comparison has to be made to see if the differences in the support and confidence are significant; if the contrast sets are mined separately and these comparisons are employed on them afterwards, one loses opportunities to prune the sets.

We concede that their first and third objections are true, however, that only affects the total time required and not the accuracy; also regarding their third objection we argue, later, that it can actually be a significant advantage if contrast set mining is used as an intermediate step to study certain problems. Their second argument has implications on the accuracy of the results, and rightly so, however, having found a way to overcome this issue we decided to use association rules to investigate the problem of finding contrast sets because of the inadequacies of the earlier mentioned techniques, and their conflicting conclusions.

Association rules form the backbone of all the previously mentioned approaches, and hence the accuracy of the results obtained by this approach cannot be questioned, even if this approach might be slower. Our hypothesis was that association rules, being the foundation of this problem, will generate all the “interesting” and “useful” contrast sets that were generated by STUCCO and potentially many more. While our approach still aims at identifying the contrast sets that satisfy the deviation conditions of STUCCO (i.e. to find the significant and large contrast sets), it does so using association rules, and in the process does not suffer from the same shortcomings as Magnum Opus (regarding the within-group comparison as opposed to an inter-group comparison as identified by Hilderman *et al.*)

4.1. Finding Deviations

The solution that we employed to overcome the problem regarding the presence of rules in one group that have no match amongst any of the other groups (the objection raised by Bay et al.), was to modify the largeness condition, for such cases. Consider a contrast set such as *(Income=high Λ Position=Manager Λ Sex=male)*, in the group *Degree=Doctorate* with a support of 43%. For the sake of argument let us assume that the above contrast set does **not** exist in the group *Degree=Bachelors*; we call such contrast sets α-contrast sets. The normal contrast sets, called β-contrast sets, are those for which the contrast set exists in at least two groups. In the case of α-contrast sets the largeness condition cannot *normally* be applied owing to absence of sufficient support for the second group; we propose that the minimum support used for generating the association rules for that group should instead be used, in the largeness condition. Thus the modified largeness condition is given by:

$$\left| \text{support}(X|G_i) - \text{min\_supp\_apriori} \right| \geq \text{min\_dev} \qquad (3)$$

The justification for this comes from the fact that if the association rule corresponding to the contrast set *(Income=high Λ Position=Manager Λ Sex=male)* is not found in the group *Degree=Bachelors*, then it must either be the case that the support for the contrast set was either much less than the minimum support used for generating the association rules or it could be, or just a shade less than the minimum support. In either case we have no way of knowing which condition was true because of the very fact that the association rule was absent, however, the minimum support forms an upper-bound in this case, and hence represents the worst case analysis in the largeness condition above. If the potential α-contrast set satisfies (3) then it should be considered as satisfying the largeness condition. By employing this condition we were able to keep a significant number of contrast sets that would have been wrongly pruned. Note that the assumption that the support for the contrast set is zero because it does not appear in the set of Association Rules ($\tilde{A}$) would be wrong; consider the case that the actual support in the dataset for a particular association rule was 1.9% (for e.g.) while the min_support used in the Apriori code was 2.0%, and hence that association rule did not appear in $\tilde{A}$. This does not imply a support of 0% for that Association Rule; had we used 1.9% as the value min_support we would have found that particular Association Rule in $\tilde{A}$ that were extracted from the dataset.

4.2. Contrast Sets: First and Second Kind

We ran an association rule program[2] on our data sets and discovered that the number of association rules generated for the first kind of *potential* contrast sets was far too less (always less than 1%) than the number of association rules corresponding to the second kind of contrast sets. In all of these cases the contrast set was composed of only single item-sets. Initially puzzling, this result was easy to interpret; consider the case where we have two groups in the data set, and assuming that they occur approximately equally in the data-thus 50% of the records belong to group 1 while the other 50% belong to group 2. The support for the rule *Group1*➔ *A, B, C* will be: $P(A \cap B \cap C \cap Group1)/P(Group1)$.

Given that *P(Group1)* is very high (~0.5) the support for it will be very low implying that the minimum support inputted for generating association rules must be very low. We used a value of 1% for the

[2] Christian Borgelt’s implementation of Apriori version 4.28 [4]

minimum support and found that only single item-sets on the right hand side are able to meet these conditions. On the other hand for the rules of the second kind: *A,B,C* ➔ *Group*, the denominator is the $P(A \cap B \cap C)$, which is small and hence the value of minimum support that goes into Apriori code can be relatively high. Having laid this issue to rest, we decided to consider only contrast sets (and hence association rules) of the second kind.

5. Experimental Results

In this section we present the results of our experimental evaluation and comparison of the contrast sets obtained from STUCCO and Association Rules. STUCCO was supplied by the original author-Dr. Stephen Bay. STUCCO is implemented in C++ and was compiled using g++ (version 3.4.4) run on Linux (2.6.9-42.0.3Elsmp). We implemented our code in Java 1.4.1 and ran it on a Linux (kernel version 2.6.9-42.0.3Elsmp) PC with a 2.4 GHz. AMD 64 bit Processor (4000+) and a 2 GB of memory. The Apriori code [4] is written in C. Our Java code encapsulates Apriori to extract the relevant contrast sets from the association rules results of the Apriori program.

5.1. The Datasets

We ran STUCCO and our code on different datasets and we report here three of those: Mushroom, Breast Cancer and Adult Census. The Mushroom dataset describes characteristics of gilled mushrooms; it available from the UCI Machine Learning Repository (www.ics.uci.edu/~mlearn/MLRepository.html). The Adult Census dataset is a small subset of the Adult Census Data: Census Income (1994/1995) dataset-a survey dataset from the U.S. Census Bureau. The Breast Cancer dataset, again obtained from the UCI Machine Learning Repository (as above), was collected by physicians and the data belongs to two groups: recurring and non-recurring.

The characteristics of the datasets are shown in Table 1 where the Tuples column describes the number of tuples in the dataset, the Attributes column describes the number of attributes, the Values column describes the number of unique values contained in the attributes, and the Groups column describes the number of distinct groups-as defined by the number of unique values in the grouping attribute.

Table 1. Properties of the datasets

Dataset	**Tuples**	**Attributes**	**Values**	**Groups**
Mushroom	8,142	23	130	2
Adult Census	826	13	129	2
Breast Cancer	286	9	53	2

5.2. The Algorithm

The cleaned data, for each group, is stored in a separate file and the Apriori program is run for all of these separately. Apriori generates association rules and we "grep" all those association rules of the kind *Group* 🡐 *contrast-set*, and store them in separate files again. At this point we have as many files as the number of groups; we then sort these association rules and feed these files with sorted association rules to a java program. The program reads the first lines of these files initially and starts comparing the right hand side of these association rules to check if they form a contrast set. A lexicographical comparison of these "Strings" is carried out and the one that is lexicographically smaller than others is marked as zero while the rest are all ones. If more than one strings match than they are marked zero. If there is only one zero then it is a potential α-contrast sets and hence the modified largeness condition is checked for that string, while if there are more than one zeros those strings are checked for the deviation condition.

After that one more line is read only from the files(s) that has/have a zero corresponding to them; once again a lexicographical minimum is found and so on until all the lines from all the files are read. If the end of a file(s) is reached while other files still have lines to be read, then no more lines are read from that file(s).

5.3. The effect of maximal number of item-sets

Christian Borgelt's code for generating association rules using Apriori analysis requires a maximal number of items per set (**n**) where by the default value is 5. Potentially this corresponds to the maximum number of attribute-value pairs in a discovered contrast set. As we did not know of an optimal value for **n,** in advance, we decided to vary this number all the way from 2 to

13 (the maximum number of attributes) for the Adult Census dataset, and similarly for the Mushroom dataset.

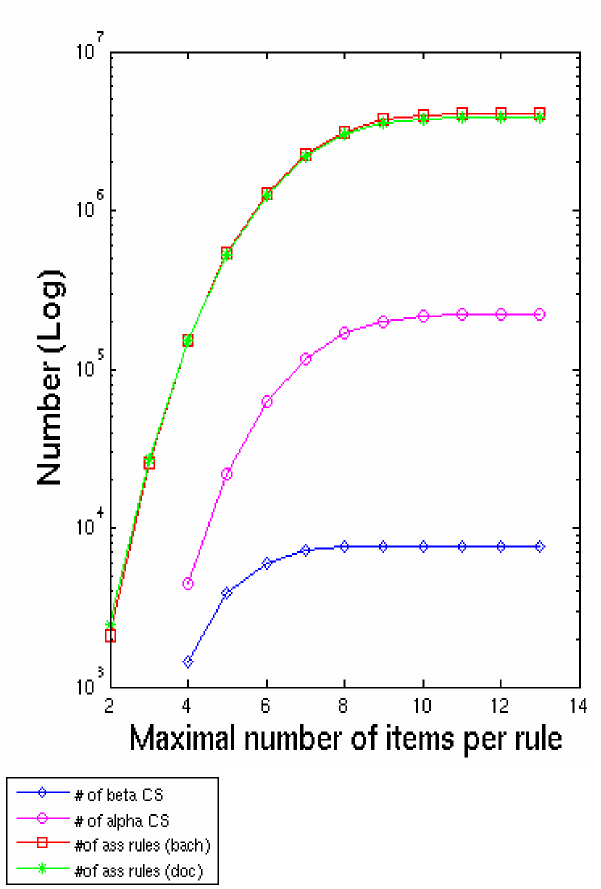

Figure 1 Behavior of association rules with contrast sets for the Adult dataset **(**“bach” stands for bachelor; “doc” stands for doctorate- the two groups).

We expected the number of contrast sets to increase, initially, with the increase in the value of **n**, and then we expected it to start decreasing above a certain value of **n,** however to our surprise the results were different, and at the same time interesting. They are presented in Figures 1 and 2 for the two data sets. The plot for the Mushroom dataset does not go beyond 7 maximal items per rule because we hit the limit of the maximum possible size of a file at that point (because we have more than 10 million association rules, for each group). It is clear both from Figure 1 and 2 that as the maximal number of items per rule increase, the number of association rules increase for both the groups. In Figure 1, the maximal number of items per rule at which the curve for the number of association rules become *almost* flat is approximately 9.

While the curve for the number of β-contrast sets becomes flat at 7 maximal items per association rule, corresponding number for the α-contrast sets is 9, showing that it follows the association rules. The number of α-contrast sets that are found is more than an order of magnitude higher than the number of β-contrast sets. Also the plot shows the initial rate of increase of all the curves, with the maximal number of items per rule, is much higher (e.g. from 4 to 6) than the rate of increase later (e.g. 6 to 8), thus signifying that this rate decreases with the number of maximal items per rule.

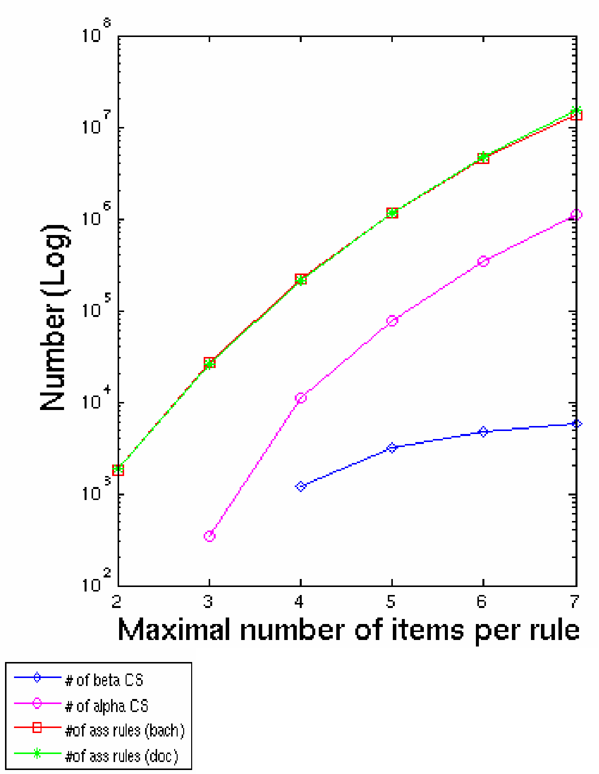

Figure 2 Behavior of association rules with contrast sets for the Mushroom dataset

For the case of Mushroom data set it is clear that the curve for the number of α-contrast sets seems to follow the curve for the number of association rules for both the groups, and also the fact that the curve for β-contrast sets seems to be close to flattening out while the other curves still have a rising trend. Figures 1 and 2 clearly show that there is a marked distinction between the α-contrast sets and the β-contrast sets.

5.4. A Comparison of the contrast sets

For the Adult data set STUCCO found 24 interesting contrast sets out a total number of 919 identified deviations. All the contrast sets found by STUCCO were also present in the interesting contrast sets that we generated with our approach. We ranked the contrast sets from our code in terms of their interestingness (i.e. confidence differential). There seemed to be many interesting contrast sets in our list that STUCCO missed. For the Breast Cancer data set STUCCO found only 18 deviations and 5 interesting contrast sets, again all of these belonged to our list of discovered interesting contrast sets and very few of STUCCO's contrast sets lie in our list of top 50. A comparative analysis for the Mushroom data set could not be performed because STUCCO's output for that dataset were garbage values, and no contrast set was discovered by it while our approach pinpointed many relevant contrast sets.

6. Conclusion and Future Work

Our analysis on the Adult Census dataset and the Breast Cancer dataset shows that Association Rule based analysis is more correct and finds all the contrast sets that are found by STUCCO, and some more potentially interesting ones that STUCCO fails to discover. We have also shown that only one kind of Association Rules make sense-the second kind. We have provided a new method for treating the α-contrast sets, which in turn finds a large number of contrast sets that would otherwise have been pruned. We found that while the number of contrast sets increases almost exponentially with the maximal number of allowed items per set (initially), and then it tapers off, this behaviour is different for α- and β-contrast sets.

We believe that our work has implications both for clustering–using the contrast sets obtained from the data, and analyzing the quality of clustering that is carried out by any of the known methods. Contrast sets can discriminate among clusters and thus help describe and label clustering results; we plan to carry out further exploration in this regard. Contrast sets can also be used to improve the accuracy of a classifier. A more systematic way to evaluate results from mining contrast sets is to study their impact on classifiers in terms of accuracy improvement.

Acknowledgements

The code for STUCCO was provided by the original author, Dr. Stephen Bay, and we would like to acknowledge his help in this regard.

References

[1] S.D. Bay and M.J. Pazzani. Detecting change in categorical data: Mining contrast sets. In Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'99), pages 302-306, San Diego, U.S.A., August 1999.

[2] S.D. Bay and M.J. Pazzani. Detecting group differences: Mining contrast sets. Data Mining and Knowledge Discovery, 5(3): 213-246, 2001.

[3] R. J. Bayardo, Efficiently mining long patterns from databases, In proceedings of the ACM SIGMOD Conference on Management of Data, 1998.

[4] C. Borgelt, Fast Implementation of the Apriori Algorithm available at:
http://fuzzy.cs.uni-magdeburg.de/~borgelt/apriori.html

[5] G. Dong and J. Li. Efficient mining of emerging patterns: Discovering trends and differences.

[6] R.J. Hilderman and T. Peckham. A Statistically Sound Alternative Approach to Mining Contrast Sets. In Proceedings of the 4th Australasian Data Mining Conference (AusDM), pages 157-172, Sydney, Australia, December, 2005.

[7] Terry Peckham. Contrasting interesting grouped association rules. Master's thesis, University of Regina, 2005.

[8] G.I. Webb, S. Butler, and D. Newlands. On detecting differences between groups. In Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'03), pages 256-265, Washington, D.C., U.S.A., August 2003.