Learning Statistically Significant Contrast Sets

M. Shazan Mohomed Jabbar and Osmar R. Zaïane

Department of Computing Science, University of Alberta,
Edmonton, Canada
{mohomedj,zaiane}@ualberta.ca

Abstract. Contrast set learning is important to discover control variables that can distinguish different groups in a dataset. Association rule mining has an inherent connection to the contrast set learning problem and has also been used to address it. All of the association rule based contrast set learning techniques use support-confidence based methods and inherit their limitations. In recent years statistically significant rule mining has become a viable alternative to address those limitations. We propose a novel contrast set learning approach based on statistically significant rule mining that eliminates the limitations in using traditional rule mining approaches and identifies statistically significant contrast sets. We evaluated our method by building a classifier using the discovered contrast sets. The performance of our classifier, while our method is not for classification per se, reveals the effectiveness of our approach in distinguishing the groups.

Keywords: Contrast sets, Association rules, Characterizing groups

1 Introduction

Finding meaningful groups in a dataset is one of the main tasks in data mining known as clustering. Characterizing discovered clusters remains a challenge. For instance, when given a clinical dataset with two clusters, healthy and non-healthy, an intriguing question researchers would like to ask is: "what are the key variables that can differentiate between healthy and non-healthy people?". The difference between contrasting groups in such situations can be described using conditional probabilities [5]. As an example consider, $P(Smoking \wedge Exposure-to-carcinogens|Healthy)$ and $P(Smoking \wedge Exposure-to-carcinogens|Non-healthy)$. These conditional probabilities can be interpreted as association rules as follows: $Healthy \implies (Smoking \wedge Exposure - to - carcinogens)$ and $Non-healthy \implies (Smoking \wedge Exposure-to-carcinogens)$. The consequent of these rules could be representing a contrast set [5].

Contrast set mining was introduced as emerging pattern discovery [2]. A well known contrast set learning algorithm is STUCCO [1]. The contrast set learning problem is intrinsic to the association rule mining problem where several of the previous methods directly exploit this connection. Magnum Opus [4] and CIGAR [3] are two such approaches. Some of these previous techniques are based on the *first-kind* of association rules: $Group \implies Contrast - set$ [1], [3] whereas the others have primarily based on the *second-kind* of rules: $Contrast - set \implies Group$ [4, 5]. Recent works in the literature have empirically proved that only the *second-kind* of contrast sets are possible [5]. Most of the existing techniques are based on the traditional Apriori [7] like rule discovery techniques, which depend on two threshold values called support and confidence and inherits the limitations imposed by them.

In recent years, statistically significant association rule mining has become a viable alternative to address such limitations in traditional approaches. Kingfisher [8] is one such proposed pioneering technique. Kingfisher uses Fisher's exact test to measure the statistical significance of a rule against the null hypothesis. Null hypothesis outlines that the antecedents and the consequent of a given rule are independent. When developing a contrast set learning method, statistically significant association rule mining techniques could be of great use not just because it can eliminate the limitations posed by traditional methods, but also the detected contrast sets would be statistically significant.

2 Problem Definition

In association rule analysis and contrast set learning, we deal with a transaction database D such that each sample transaction E in D can be defined as a vector of size m. Let $A = \{A_1, A_2, ..., A_m\}$ be a set of attributes called *items*. Then a transaction E can be defined as a vector consisting of attribute-value pairs $A_1 = V_1, A_2 = V_2, ..., A_m = V_m$ where each $V_j \in \{V_{j1}, V_{j2}, ..., V_{jn}\}$ and n is a finite number. Given these definitions an **association rule** is an implication of the form $X \implies Y$ where $X \subset A$, $Y \subset A$ and $X \cap Y = \emptyset$. Confidence c in $X \implies Y$ is the percentage of data instances in D containing X also contains Y (i.e. $P(Y|X)$). Support s for $X \implies Y$ is the percentage of data instances in D containing $X \cup Y$. Traditional algorithms discover strong association rules by verifying that their s and c exceed some user defined threshold. **Classification Association Rules** (**CARs**) are an important class of association rules. Given a set of class labels $C = \{c_1, c_2, ..., c_q\}$ where each instant E in D is associated with a class label c_i and q is the number of classes, a CAR can be defined as an association rule of the form $X \implies c_i$.

A **contrast set** is a conjunction of attribute-value pairs defined on mutually exclusive classes from C such that no A_i occurs more than once. Some of the early research works identify contrast sets as sets belonging to rules of the form $Class \implies Contrast - set$. However according to recent literature we have only considered contrast sets that are identified as sets belonging to rules of the form $Contrast - set \implies Class$ [5]. The STUCCO algorithm detects significant contrast sets by imposing two constraints, known as **deviation conditions**, on the candidate patterns. When both the conditions are met it is defined as a deviation and the candidate pattern is identified as a valid contrast set [1]. These conditions are as follows:

$$\exists_{i,j} P(X|c_i) \neq P(X|c_j) \tag{1}$$

$$\max_{i,j} |support(X, c_i) - support(X, c_j)| \geq min\_dev \tag{2}$$

3 Proposed Approach

We propose a technique, which uses statistically significant CARs, to discover valid contrast sets from a given transaction dataset. Our approach consists of two phases: 1) Generate statistically significant CARs from the given dataset; 2) Mine statistically significant contrast sets from those rules.

3.1 Classification Association Rule Generation

In phase 1 we mine statistically significant CARs. The statistical significance of CARs can be determined by testing for its dependency. Hence, the CAR, $X \implies c_i$ is considered statistically significant at level α, if the probability p of observing an equal or stronger dependency in a dataset complying with the null hypothesis is not greater than α. In the null hypothesis it is assumed that X and c_i are independent. The probability p (i.e. p-value) can be calculated from a cumulative hypergeometric distribution using Fisher's exact test [8].

The significance level α is set to be 0.05. Calculated p-values can be used to prune non significant rules using this α. Kingfisher algorithm [8] accomplishes this task by using an enumeration tree with branch and bound search. In our specific case to detect CARs, we define the shape of the rules to be outputted by the Kingfisher algorithm. This constrained version of the Kingfisher we used obtains only rules with the type $X \implies c_i$ [6].

3.2 Discovering Contrast Sets

In phase 2 we discover statistically significant contrast sets using the CARs obtained in phase 1. To achieve this, we propose a new set of deviation conditions, based on the original conditions from STUCCO, as follows:

$$\exists_{i,j} p_F(X \implies c_i) \neq p_F(X \implies c_j) \tag{3}$$

$$\min_{i,j} |p_F(X \implies c_i) - p_F(X \implies c_j)| \leq max\_dev \tag{4}$$

where p_F refers to the p-value of the particular rule. Equation 3 attempts to capture the statistical significance of the contrast set while Equation 4 attempts to capture whether the statistical significance of the candidate contrast set across different classes is sufficiently large given the maximum deviation threshold. Note that in contrast to the original conditions introduced in STUCCO, we use a maximum deviation threshold because lower the p_F value better the statistical significance is.

Original deviation conditions from STUCCO was unable to capture the contrast sets which only exist in one group. Based on some of the recent works in the literature [5] we addressed this issue by introducing the condition in Equation 5 which uses the maximum threshold value or level of significance of the

Kingfisher algorithm (i.e. $Kingfisher(p_F)$). Whenever a candidate contrast set is absent in a group, this threshold value can provide a lower bound to the p_F value of that candidate set. This new condition is as follows:

$$\min_{i,j} |p_F(X \implies c_i) - Kingfisher(p_F)| \leq max\_dev \tag{5}$$

For the maximum deviation threshold we experimented with using the $Avg.(p_F)$ value and $Avg.(p_F) + \sigma(p_F)$ of the rules generated in Phase 1. $\sigma(p_F)$ is the standard deviation of the p-value distribution.

4 Experiments

We conducted experiments on 18 public datasets from the UCI ML Repository [1] to evaluate our method in identifying valid contrast sets. As previously explained, in our experiments, we used a constrained version of the Kingfisher algorithm in the first phase of our method with a 5% level of significance. In the second phase of our method we discovered contrast sets using the three deviation conditions (Equation 3, 4 and 5) we introduced earlier. All the experiments we conducted are 10 fold cross validated and results are averaged in Table 1.

We implemented a baseline method which uses support-confidence based rules to learn contrast sets. To be fair we have obtained the minimum confidence and support from the statistically significant rule set and used them as threshold values to obtain association rules from Apriori like algorithms. Then using Equation 1 and 2 we identified contrast sets. However, when we compared the contrast sets detected by our method and this baseline approach, from UCI datasets, it is revealed that there is a little to no overlap between them.

To further investigate our approach, we used the contrast sets obtained by our method to build an associative classifier, CS^2 (**C**lassification based on **S**tatistically Significant **C**ontrast **S**ets), following similar works [10]. Having a better classification accuracy would mean that we have identified contrast sets which can meaningfully differentiate classes. As shown in Algorithm 1 (line 7-12), CS^2 first recognizes the subset of contrast sets which can contribute to classify a given data instance. Then, based on class, it categorizes this subset of rules. Aggregate function *sum* can be applied to each of these categories to obtain a representative measure for each class. Next a class can be assigned to the data instance by using another aggregate function *min* on the representative measures. We have compared the classification accuracy of CS^2 with several other standard classifiers on UCI datasets even though our target is not to build a classifier but to discover accurate contrast sets. Their classification accuracies are reported in Table 1. We chose classifiers, C4.5 [9], CBA [10] and CPAR [11], to compare. Average accuracy in all 18 datasets indicates that CS^2 has a classification accuracy very close to that of other standard classifiers. It even outperforms CBA algorithm, proving that statistically significant rules can provide quality contrast sets. Results with contrast sets used as baseline were very low and not reported in Table 1.

[1] `http://archive.ics.uci.edu/ml/`

Algorithm 1 CS^2 Algorithm

```
INPUT: Database D, Object O, Attributes A, Classes C, Level-of-Significance $\alpha$
CAR = Kingfisher(D, A, C, $\alpha$)
CSet = $\emptyset$
for all rule $r$ in $CAR$ do
    if $r$ suffice Eq. 3, 4 or 5
        CSet = CSet $\cup$ r
end for
CSet$_{new}$ = $\emptyset$
for all c-set $c$ in $CSet$ do
    if $c.antecedent \subseteq O.antecedent$
        CSet$_{new}$ = CSet$_{new}$ $\cup$ c
end for
Divide CSet$_{new}$ to subsets based on class labels: $S_1, S_2, ...S_n$
for all $S_i$ in $S_1, S_2, ...S_n$ do
    sum all the $ln$ p$_F$ values in each subset
end for
Assign the class with lowest some of p$_F$ to O
RETURN $O.label$
```

Table 1. Comparison of classification results: C4.5, CBA, CPAR and CS^2.

Dataset	#cls	#rec	C4.5	CBA	CPAR	CS^2	
						$Avg.(p_F)$	$Avg.(p_F) + \sigma(p_F)$
adult	2	48842	78.8	**84.2**	77.3	83.4	83.7
anneal	6	898	76.7	94.5	**95.1**	84.1	76.8
breast	2	699	91.5	**94.1**	93.0	82.1	79.1
cylBands	2	540	69.1	**76.1**	70.0	63.8	63.8
flare	9	1389	82.1	**84.2**	63.9	73	64.7
glass	7	214	65.9	**68.4**	64.9	61.7	63.5
heart	5	303	**61.5**	57.8	53.8	55.5	57.8
hepatitis	2	155	84.1	42.2	75.5	81.5	**85.3**
horseColic	2	368	70.9	78.8	**81.2**	77.1	76
ionosphere	2	351	84.6	32.5	**88.9**	78.6	74.3
iris	3	150	91.3	93.3	**94.7**	93.3	93.3
led7	10	3200	**73.8**	73.1	71.3	71.6	73.1
mushroom	2	8124	92.8	46.7	**98.5**	94.7	95.8
pageBlocks	5	5473	92.0	90.9	**92.5**	90	89.7
penDigits	10	10992	70.5	**92.3**	80.5	68.4	70.3
pima	2	768	71.7	**74.6**	74.0	67.5	65.6
wine	3	178	75.8	49.6	88.2	91.4	**92**
zoo	7	101	91.0	40.7	**94.1**	81.2	92
Average			79.1	70.7	79.9	77.7	77.6

5 Conclusion and Future Work

We proposed to mine statistically significant contrast sets by using statistically significant association rules. Our results indicated that the contrast sets learned using our method and the contrast sets learned using traditional association rules have little to no similarity. However, we proved that, on average, our CS^2 classifier based on the proposed approach had close or better performance with other standard classifiers, providing evidence to the quality of the contrast sets we learned. Note again that our purpose is not to build a new classifier but classification was used as a means for validation. In our current work, we only explored the possibility of using association rules of the *second-kind* to mine contrast sets. However, since we introduced the use of statistically significant rules, the possibility of finding and using *first-kind* of association rules should be explored. Ideally, a combination of both type of rules may provide a better set of candidate contrast sets.

References

1. Bay, S. D., Pazzani, M. J.: Detecting Group Differences: Mining Contrast Sets. Data Mining and Knowledge Discovery, vol. 5, no. 3, pp. 213–246. Springer (2001)
2. Dong, G., Li, J.: Efficient Mining of Emerging Patterns: Discovering Trends and Differences. In: 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 43–52. ACM (1999)
3. Hilderman, R. J., Peckham, T: A Statistically Sound Alternative Approach to Mining Contrast Sets. In: 4th Australia Data Mining Conference (AusDM-05), pp. 157–172. (2005)
4. Webb, G. I., Butler, S., Newlands, D.: On Detecting Differences Between Groups. In: 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 256-265. ACM (2003)
5. Satsangi, A., Zaïane, O. R.: Contrasting the contrast sets: An alternative approach. In: 11th International Database Engineering and Applications Symposium (IDEAS 2007), pp. 114–119. IEEE (2007)
6. Li, J., Zaïane, O. R.: Associative Classification with Statistically Significant Positive and Negative Rules. In: 24th ACM International Conference on Information and Knowledge Management, pp. 633–642. ACM (2015)
7. Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules. In: 20th International Conference in Very Large Data Bases (VLDB), pp. 487-499. (1994)
8. Hämäläinen, W.: Kingfisher: An Efficient Algorithm for Searching for Both Positive and Negative Dependency Rules with Statistical Significance Measures. Knowledge and Information Systems, vol. 32, no. 2, pp. 383–414. (2012)
9. Quinlan, J. R.: C4. 5: Programs for Machine Learning. Morgan Kaufmann Publishers, Inc., (1993)
10. Liu, B., Hsu, W., Ma, Y.: Integrating Classification and Association Rule Mining. In: 4th International Conference on Knowledge Discovery and Data Mining, pp. 80–86. (1998)
11. Yin, X., Han, J.: CPAR: Classification based on Predictive Association Rules. In: 3rd SIAM International Conference on Data Mining, pp. 369–376. (2003)