On detecting differences between groups

Yi Yang
Department of Computing Science
University of Alberta

Contrast-Set Mining

- Understanding the differences between contrasting groups is a fundamental task in data analysis

- "Contrast-set Mining"

  S. D. Bay and M. J. Pazzani
  Detecting change in categorical data: Mining contrast sets. 1999

- A new technique in data mining ?

If yes, is it somehow related to previous data mining techniques such as association rule mining, classification, etc?

On detecting differences between groups

Geoffrey I. Webb, Shane M. Butler, Douglas Newlands
2003 ACM SIGKDD

- A study is undertaken to compare contrast-set mining with existing rule-discovery techniques.
- Collaboration with a retail store
- Surprise...?

Outline

- Introduction
- The three techniques
  - STUCCO
  - Magnum Opus
  - C4.5rules
- Comparison
- Rule Quality Assessment
- Conclusion

Introduction

- Based on a project to evaluate how contrast-set mining differs from pre-existing forms of rule-discovery in an applied context:
  - One of Australia's largest discount department store companies
  - Retail activities of two different days
  - 6 stores; several departments
  - Task:
    to highlight how the "baskets" of departments differed between 2 days

Three Techniques

- STUCCO
  - Search and Testing for Understandable Consistent Contrasts
  - Specialized for mining contrast-sets.
  - Proposed by Bay and Pazzani
- Magma Opus
  - A commercial implementation of OPUS_AR rule-discovery algorithm.
  - Rules: antecedent --> consequent
- C4.5rules
  - Classification-rule discovery
  - Treat groups as classes

STUCCO

- Find contrasts "significant" and "large"
  - Significant:

$$\exists ij\; P(cset|G_i) \neq P(cset|G_i)$$

  - Large:

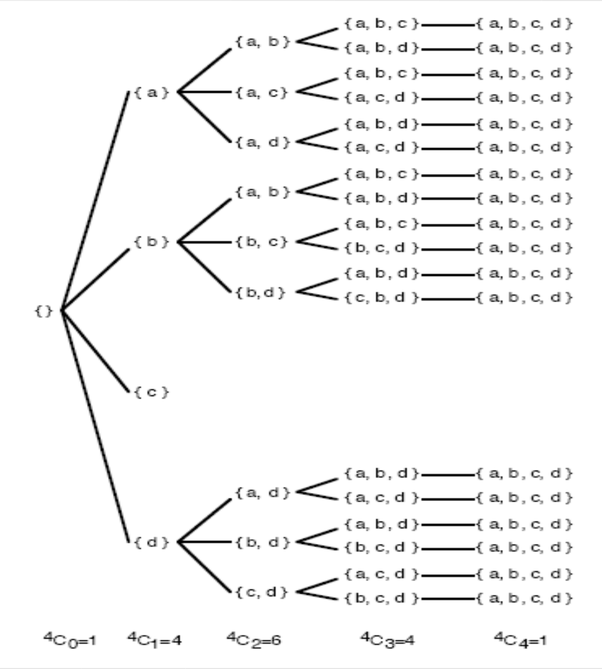
$$\max_{ij} |support(cset, G_i) - support(cset, G_j)| \geqslant \delta$$

  where δ is a user-defined threshold called the *minimum support-difference*

  - Rule filter: chi-square test

Magnum Opus

- OPUS algorithm (Optimized Pruning for Unordered Search):
  - search tree;
  - identifies excluded operators;
  - prunes descendent trees;
  - ...



- Magnum Opus
  - performs association-rule-like search
  - does NOT find frequent-itemsets
  - no requirement for minimum support, but requires rule value & maximum number of rules

Magnum Opus (cont.)

- Rule: antecedent --> consequent

  antecedent = cond1∧ cond2∧ ...}

- Measures of rule value:
  - Support
  - Confidence (called strength)
  - Lift
  - Coverage

    support of antecedent
  - Leverage (default measure)

    degree to which the observed joint frequency of the antecedent and consequent differ from their joint frequency

$$leverage(a \rightarrow c) = support(a \cup c) - support(a) \times support(c)$$

C4.5rules

- Discovers classification rules
  1. discovers a decision tree
  2. converts tree to a set of rules
  3. simplifies those rules

- Different from contrast-set/association-rule discovery
  - CS/AR find all rules that satisfies some constraint
  - CR find rules that are sufficient to predict classes

- Adaption to contrast-set mining:
  - Groups are encoded as a class variable
  - Learn rules to distinguish the groups

Application

- Data
  - 2 days of transactions
  - 6 stores, aggregated to the department level
  - To contrast the purchasing behavior of customers on the two days

- Configuration and parameters
  - STUCCO
    - *Significance level = 0.05*
    - *Minimum support-difference = 0.01*
  - C4.5rules
    - *Default settings*
  - Magnum Opus
    - *Rule value: leverage*
    - *Maximum number of rules: 1000*

Table 1: Descriptive statistics

Statistic	**Day 1 (August-14th)**	**Day 2 (August-21st)**
No. transactions on each day	6296	6906
Average no. depts. per transaction	1.55	1.93
Top department	1100 items from dept 929	1349 items from dept 929
Second top department	845 items from dept 805	1213 items from dept 805
Third top department	708 items from dept 220	849 items from dept 851
Fourth top department	653 items from dept 60	841 items from dept 340
Fifth top department	483 items from dept 845	796 items from dept 60
Sixth top department	449 items from dept 340	666 items from dept 855
Seventh top department	442 items from dept 901	638 items from dept 845
Eighth top department	415 items from dept 905	608 items from dept 901
Ninth top department	414 items from dept 685	556 items from dept 355
Tenth top department	407 items from dept 170	507 items from dept 270

Comparison

	STUCCO	Magnum Opus	C4.5rules
Total # of rules	19	83	24
# of single-value rules	19	56	5
# of two-value rules	0	23	2
# of three-value rules	0	4	3
# of multi(>3)-value rules	0	0	14

- Rules discovered by STUCCO are all single-value rules;
- Magnum Opus discovered all rules found by STUCCO;
- C4.5 discovered rules up to 51 conditions (51-value rules).

Example of rules: STUCCO

Table 2: A contrast set as output by STUCCO

```
220 = 1
434 257 | 0.0689327 0.037214
==============================
d.f.    chi^2    pvalue
1       66.80    3.00e-16
==============================
```

Contrast Set

Proportion of transactions

Number of transactions on each day that contained dept 220

chi-square test of significance

Example of rules: Magnum Opus

Table 3: Six rules as output by Magnum Opus

```
851 -> August-21st [Coverage=0.049 (649);
Support=0.038 (500); Strength=0.770; Lift=1.47;
Leverage=0.0122 (160)]

855 -> August-21st [Coverage=0.043 (574);
Support=0.033 (432); Strength=0.753; Lift=1.44;
Leverage=0.0100 (131)]

855 & 851 -> August-21st [Coverage=0.009 (119);
Support=0.008 (104); Strength=0.874; Lift=1.67;
Leverage=0.0032 (41)]

220 -> August-14th [Coverage=0.052 (691);
Support=0.033 (434); Strength=0.628; Lift=1.32;
Leverage=0.0079 (104)]

335 -> August-14th [Coverage=0.007 (98);
Support=0.006 (74); Strength=0.755; Lift=1.58;
Leverage=0.0021 (27)]

220 & 355 -> August-21st [Coverage=0.001 (15);
Support=0.001 (13); Strength=0.867; Lift=1.66;
Leverage=0.0004 (5)]
```

- Rules 1-2: the proportion of customers buying from each of dept. 851 and 855 on the 2nd day was higher than the 1st.
- Rule 3: this effect was heightened when customers that bought from both departments in a single transaction were considered.
- Rules 4-6: Whereas items for dept. 220 and 355 were each purchased more frequently on day 1 than day 2, a greater proportion of customers bought items from both departments on the day 2 than day 1.

Example of rules: c4.5rules

```
 Rule 645:
        261 = 1
    ->  class August-21st  [86.8%]

Rule 628:
        405 = 0
        60 = 0
        901 = 0
        957 = 0
        200 = 0
        920 = 0
        903 = 0
        345 = 1
        999 = 0
    ->  class August-21st  [84.2%]

Rule 472:
        370 = 0
        870 = 0
        957 = 1
        855 = 0
        640 = 0
        830 = 0
        851 = 0
        285 = 0
        620 = 0
        250 = 0
        335 = 0
        440 = 0
        235 = 0
    ->  class August-14th  [55.6%]
```

- Value in brackets is the confidence of the rule
- Most rules contain many “negative” conditions where dept=0
- Are negative conditions useful? Will be assessed by domain experts

Table 5: Comparison of rules discovered

Dept.	Magnum Opus Rule Num. (Single condition)	Rule Num. (Multiple conditions)	STUCCO Rule Num.	C4.5rules Rule Num.	p
851	1	19	5	7	0.00000
855	2	19, 51	6	9	0.00000
490	10		9	11	0.00000
520	12		8	14	0.00000
405	16		12		0.00000
335	27				0.00000
870	17	51	11	13	0.00000
875	20	61	10		0.00000
261	36			2	0.00000
620	24	59		10	0.00000
410	21	69	13		0.00000
355	14	52, 60, 63, 66	17	17	0.00001
500	22	78	15		0.00002
685	4	62	7	12	0.00002
170	18	62, 67	18	22	0.00005
440	47			4	0.00007
270	15	39, 60	19		0.00007
80	26				0.00019
980	40				0.00022
360	23				0.00027
265	35				0.00049
465	57			6	0.00071
830	25				0.00073

Relationship between STUCCO and Magnum Opus

- STUCCO

  $$\exists ij\; P(cset|G_i) \neq P(cset|G_i)$$

- Magnum Opus
  - Rule filter:

    $$For\; rule\; a \rightarrow c,\; P(c|a) > P(c)$$

  - If the antecedents are treated as contrast sets and the consequents as groups:

    $$\exists i\; P(G_i|cset) > P(G_i)$$

- THEOREM. If all *csets* belong to a group ($\sum_{i=1}^{l} P(G_i) = 1.0$) and no group is empty ($\forall i : 1 \leq i \leq l, 0.0 < P(G_i) \leq 1.0$) then

  $$\exists i\; P(G_i \,|\, cset) > P(G_i) \equiv \exists ij\; P(cset \,|\, G_i) \neq P(cset \,|\, G_j) \qquad (9)$$

Relationship between STUCCO and Magnum Opus

This led to the realization that contrast-set mining is a special case of the more general rule-discovery task.

Rule Quality Assessment

- Domain experts from the retail collaborators: retail marketing managers.
- Rules expressed in natural language:

  On August 21st customers were 7.6 times more likely to purchase items from department 445 (MENSWEAR; Mens Nightwear) than they were on August 14th. They were bought in 2.2% of transactions on August 21st and 0.3% of transactions on August 14th.
- Two questions were asked:
  1. Is this rule surprising?
  2. Is this rule potentially useful to the organization?

Rule Quality Assessment (cont.)

Table 7: Summary of assessments

System	Total no rules	Surprising	Potentially Useful
Magnum Opus (1 Dept.)	56	12	15
Magnum Opus (2 Depts.)	23	10	5
Magnum Opus (3 Depts.)	4	1	1
Magnum Opus (All)	83	23	21
STUCCO	19	2	5

- Only a lower proportion of rules discovered by STUCCO are “surprising”, and that proportion for Magnum Opus is much higher
- The proportion of contrasts being “potentially useful” is similar between STUCCO and Magnum Opus.

Rule Quality Assessment (cont.)

- Assessment of negative conditions (dept = 0)
  - *On October 22nd customers were 5.0 times more likely to purchase items from department 123 (INFANTS; Diapers) and nothing from department 345 (BEVERAGES; Beer) than they were on July 5th. This occurred in 2.5% of transactions on October 22nd and 0.5% of transactions on July 5th.*
- Response from industry collaborators:
  - *While negative conditions of these form were of potential value, these specific rules did not appear to be of interest and were more difficult to interpret than the Magnum Opus and STUCCO rules.*
- Classification rule discovery is not an appropriate approach to contrast discovery
- Negative conditions may be of value (at least in this application)

Conclusion

- We discovered that the core contrast-set discovery task is strictly equivalent to a special case of the more general rule-discovery task (though contrast discovery is still a valuable data mining task).
  -->
- Existing rule-discovery techniques can be applied to perform the core contrast-discovery task

- There issues for further investigation:
  - Selection of a rule filter: chi-square test or binomial sign test (Magnum Opus)?
  - Tuning of parameters: better performance?
  - Contrast description to help user better understand

References

[1] Geoffrey I. Webb, Shane M. Butler, Douglas Newlands. On Detecting Differences Between Groups. *In Proc. 2003 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*

[2] Stephen D. Bay, Michael J. Pazzani. Detecting Change in Categorical Data: Mining Contrast Sets. *In Proc. 1999 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*

[3] Geoffrey. I. Webb. OPUS: An efficient admissible algorithm for unordered search. *Journal of Artificial Intelligence Research*

Thanks for your attention!

Questions?