Clustering Rules Using Empirical Similarity of Support Sets

B. Golden
S. Lele
K. Ozga
University of Maryland
E. Wasil
American University

Fourth International Conference on Discovery Science
November 25-28, 2001
Washington, DC

Introduction

- Often rule generating algorithms produce too many rules.
- We develop a procedure for pruning a given set of rules such that there is minimal loss of support.
- Use cluster analysis to remove redundant rules.

Procedure

- Step 1. Partition initial set of rules into groups based on the predicted consequent.
- Step 2. Partition each rule group into distinct clusters such that the antecedents of rules within a given cluster point to similar record sets.
- Step 3. Rules within a given cluster are ranked on the basis of their support values and only the top rule is retained.

Methodology

- The similarity between two rules R_i and R_j is given by:

$$\text{sim}(i, j) \equiv \#(S_i \cap S_j)/min\{\#(S_i)\ ,\ \#(S_j)\},$$

where S_i denotes the support set of the i^{th} rule, i.e., the set of records of the data set where the rule applies and is true and $\#(S_i)$ denotes the cardinality of S_i.

Data Set

- Large marketing data set.
- Determine whether a given customer will re-use a particular product or not.
- 438,808 records, about 22% were re-users.
- Training set of 60,000 records and test set of 10,000 records were randomly selected.

Experiment

- Ninety rules were extracted from the training set using C4.5.
- All rules indicated re-use by the customer.
- Similarity matrix constructed for all 90 rules.
- Cluster analysis performed using complete linkage with a threshold level of 0.5.

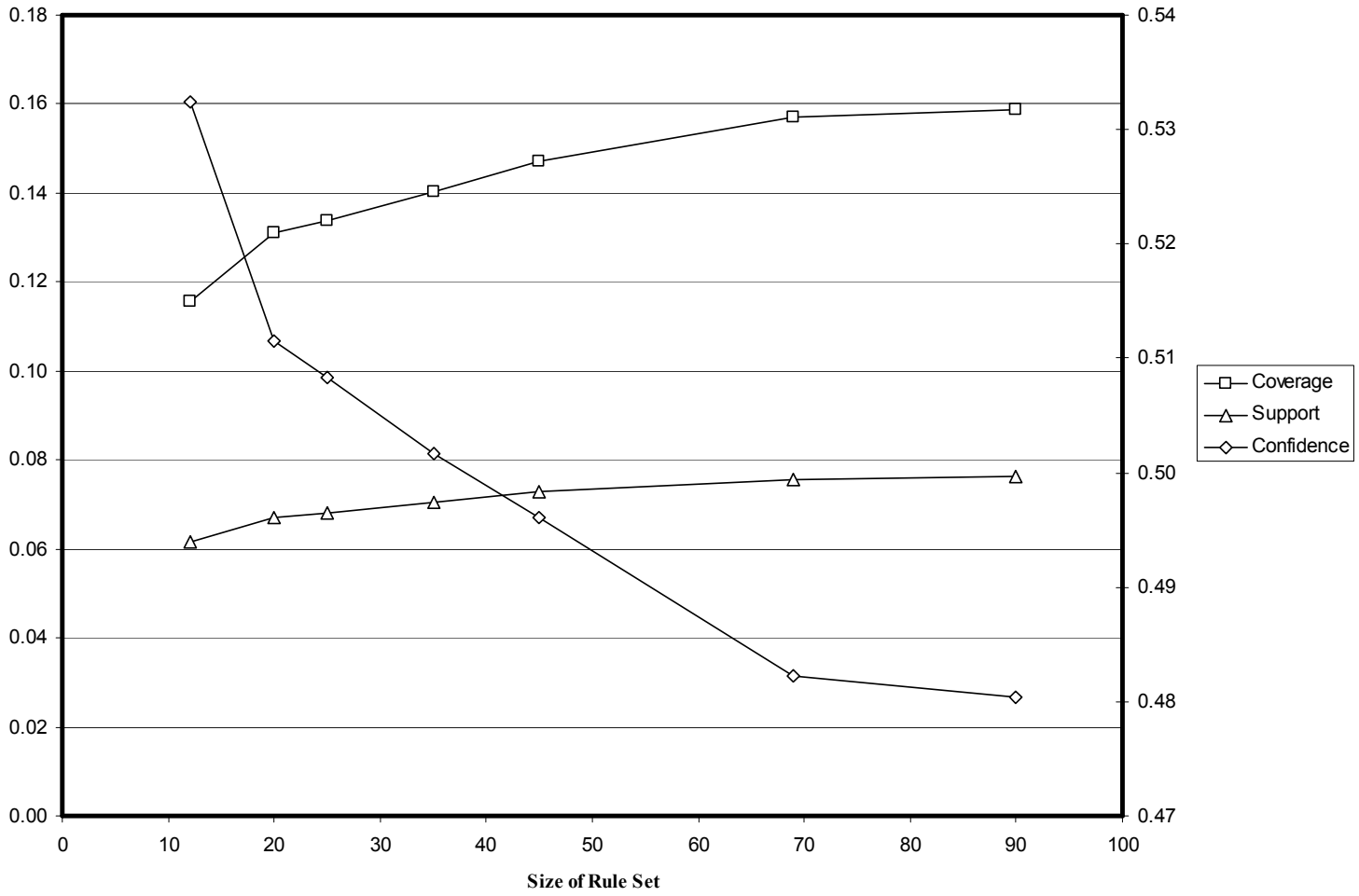
Cluster Analysis

- Cluster analysis resulted in 69 clusters, of which 57 clusters contained a single rule, while 12 contained more than one rule.
- Rules in each cluster were sorted on the basis of their individual support values, and the rule with the maximum support was retained from each cluster.
- Rules were further pruned by dropping rules with low values of support.
- We examined nested rule sets of 45, 35, 25, 20, and 12 rules.

Table 1. Performance measures of successively pruned rule sets.

Size of Rule Set	Support		Coverage		Confidence	
	Training	Test	Training	Test	Training	Test
90	0.0775	0.0762	0.1586	0.1587	0.4878	0.4804
69	0.0769	0.0757	0.1571	0.1569	0.4896	0.4823
45	0.0736	0.0729	0.1471	0.1470	0.5004	0.4960
35	0.0711	0.0706	0.1403	0.1402	0.5071	0.5017
25	0.0687	0.0680	0.1337	0.1339	0.5137	0.5083
20	0.0678	0.0671	0.1309	0.1311	0.5180	0.5116
12	0.0621	0.0616	0.1155	0.1157	0.5378	0.5324

Figure 1. Performance measures on test data for different sizes of rule sets. Support and coverage follow scale on left, confidence follows scale on right.



Results

- The numbers for the test set show a similar pattern to those for the training set.
- In comparing the initial set of 90 rules with the post-clustering rule set of 69 rules, we see that the pruning of 21 rules (23% of the initial rule set) is accompanied by only a small drop in the value of support.

Conclusions

- Presented a method for rule pruning based on cluster analysis.
- Instead of looking at the logical construction of various rules, we looked at the support sets of rules to determine similarity.
- We reduced the size of a rule set without a significant decrease in support.