Generating Compact Rough Cluster Descriptions Using an Evolutionary Algorithm

Kevin Vogts and Nigel Pope

School of Management, University of Canterbury, Christchurch, New Zealand
Kevin.Voges@canterbury.ac.nz
School of Marketing, Griffith University, Nathan Campus, Brisbane, Australia
N.Pope@griffith.edu.au

1 Cluster Analysis

Cluster analysis is a technique used to group objects into clusters such that similar objects are grouped together in the same cluster. Early methods were derived from multivariate statistics. Some newer methods are based on rough sets, introduced by Pawlak [3], [4]. An extension of rough sets to rough clusters was introduced in [5]. The lower approximation (LA) of a rough cluster contains objects that only belong to that cluster, and the upper approximation (UA) contains objects that may belong to more than one cluster. An EA can be used to find a set of lower approximations of rough clusters that provide the most comprehensive coverage of the data set with the minimum number of clusters.

2 Rough Clustering Algorithm

The building block of the data structure is the *template* [2]. Let $S = (U, A)$ be an information system, where U is the set of data objects, and A is the set of attributes of the data objects. Any clause of the form $D = (a \in Va)$ is called a *descriptor*, and the value set Va is the *range* of D. A *template* (T) is a conjunction of unique descriptors defined over attributes from $B \subseteq A$. That is, $T = \Lambda_{a \in B} (a \in Va)$ is a *template* of S. The data structure acted on by the EA is a cluster solution, C, which is defined as any conjunction of k unique templates. This data structure was encoded as a simple two-dimensional array with a variable length equal to the number of unique templates in the cluster solution and a fixed width equal to the number of attributes being considered.

A number of considerations determine the fitness measure. Firstly, the algorithm maximizes the data set coverage c, defined as the fraction of the universe of objects that matches the set of templates in C. Secondly, the algorithm minimizes k, the number of templates in C. Finally the accuracy a, of each template needs to be maximized [4]. This is the sum of the cardinal value of the LA divided by the cardinal value of the UA defined by each template in C. The *fitness* value, f, of each cluster solution is defined as the coverage multiplied by the accuracy divided by the number of templates in C.

K. Deb et al. (Eds.): GECCO 2004, LNCS 3103, pp. 1332–1333, 2004.

A multi-point recombination operator was used to generate valid rough cluster solutions. Templates are randomly selected from each parent, and added to the offspring after checking that they are unique to the cluster solution. Two mutation operators were used. One randomly sampled a unique template from the list of valid templates and added it to the cluster solution, and the other randomly removed a template from the cluster solution. Repair operators were not required.

3 Application

The technique was used to analyze data from a study of beer preferences in young adults, who were asked to evaluate possible attributes to be considered when making a purchasing decision. Five attributes were used: image, packaging, price, alcohol content, and place sold.[1] A rough cluster analysis was conducted, partitioning the study participants into distinct clusters based on which attributes were considered important. The best cluster solution obtained achieved coverage of 94.8% of the data set using 13 templates. The accuracy of each template ranged from 0.50 to 1.00. This EA-based technique was able to find a rough cluster solution that covered a large percentage of the data set with a small number of templates. The technique also overcame some limitations of previous techniques, such as those that require the number of clusters be specified in advance [1], or those that generate too many clusters to be easily interpretable [5].

References

1 Lingras, P.: Rough Set Clustering for Web Mining. In: *Proceedings of 2002 IEEE International Conference on Fuzzy Systems*. (2002)

2 Nguyen, S. H.: Regularity Analysis and its Applications in Data Mining. In: Polkowski, L., Tsumoto, S., Lin, T. Y. (eds.): *Rough Set Methods and Applications: New Developments in Knowledge Discovery in Information Systems*. Physica-Verlag, Heidelberg New York (2000) 289 - 378

3 Pawlak, Z.: Rough Sets. *International Journal of Information and Computer Sciences*. 11:5 (1982) 341 - 356

4 Pawlak, Z.: *Rough Sets: Theoretical Aspects of Reasoning About Data*. Kluwer, Boston (1991)

5 Voges, K. E., Pope, N. K. Ll., Brown, M. R.: Cluster Analysis of Marketing Data Examining On-line Shopping Orientation: A Comparison of *K*-means and Rough Clustering Approaches. In: Abbass, H.A. Sarker, R.A., Newton, C.S. (eds.): *Heuristics and Optimization for Knowledge Discovery*. Idea Group Publishing, Hershey, PA (2002) 207-224

[1] Our thanks to Michael Veitch and Kevin Nicholson for providing this data