

Analysing Destination Image Data Using Rough Clustering

*Kevin E. Voges, Department of Management, University of Canterbury**
kevin.voges@canterbury.ac.nz

Nigel K. Ll. Pope, Department of Marketing, Griffith University
n.pope@griffith.edu.au

Abstract

Cluster analysis is a fundamental data analysis technique, but many clustering methods have limitations, such as requiring initial starting points and requiring that the number of clusters be specified in advance. This paper describes an evolutionary algorithm based rough clustering algorithm, which is able to overcome these limitations. Rough clusters use sub-clusters called lower and upper approximations. The lower approximation of a rough cluster contains objects that only belong to that cluster, while the upper approximation contains objects that can belong to more than one cluster. The approach therefore allows for multiple cluster membership for data objects. This rough clustering algorithm was tested on a large data set of perceptions of city destination image attributes, and some preliminary results are presented.

Introduction

Tourism is an important economic activity in many countries, and its importance becomes particularly noticeable after natural disasters such as a tsunami (Stateman, 2005), or when there is a sudden drop in tourism numbers (Fernandes, 2002). Tourism is a form of consumer behaviour that is intimately connected to the image of the country being visited, so country image is an important construct in any study of tourism behaviour.

There is quite an extensive literature addressing the country image concept, dating back to the 1970s (Gunn, 1972; Hunt, 1975). Konecnik (2005) gives an excellent overview of the literature, including a comprehensive list of country image attributes. Studies have looked at individual countries, such as Australia (Hanlan and Kelly, 2005; Son, 2005; Son and Pearce, 2005), New Zealand (Morgan, Pritchard and Piggott, 2003), Portugal (Silvestre and Correia, 2005), Slovenia (Konecnik, 2005), and the United States (Bonn, Joseph and Dai, 2005; Obenour, Lengfelder and Groves, 2005). Specific tourist markets have also been studied, such as outbound Chinese tourists (Kim, Guo and Agrusa, 2005), nature-based destinations (Obenour, Lengfelder and Groves, 2005), and business tourism (Hankinson, 2005). Important destination characteristics identified include “sun and sand” (Silvestre and Correia, 2005), safety, and beautiful scenery (Kim, Guo and Agrusa, 2005). Research has also reported on the influence of visitor characteristics (Boo and Busser, 2005) and information (Hanlan and Kelly, 2005) on destination image. Studies have also considered the strategic implications of perceptions of country image (Bonn, Joseph and Dai, 2005; Konecnik, 2005).

This paper presents the results of a cluster analysis of a large data set of over 6,000 records, containing individual perceptions of city destination image attributes. The purpose of the paper is to demonstrate the rough clustering analysis technique, and show its scalability to large data sets. As an outcome of this cluster analysis, the paper reports clusters of linked consumer perceptions that could be of benefit to marketers in promoting their tourism destination.

Methodology

Cluster Analysis

Cluster analysis is a fundamental technique in both traditional data analysis and in data mining. The technique is defined as grouping ‘individuals or objects into clusters so that objects in the same cluster are more similar to one another than they are to objects in other clusters’ (Hair, Anderson, Tatham and Black, 1998, p. 470). Many clustering methods have been identified, with one of the most commonly used non-hierarchical methods being the k-means approach (MacQueen, 1967). More recently, techniques have been developed using approaches derived from the field of computation intelligence, for example the use of rough sets and evolutionary algorithms.

Rough Sets

The theory of rough sets was developed by Pawlak (1982, 1991), and is based on the assumption that a certain amount of information is associated with an object, expressed by means of attributes that describe the object. The basic concepts of rough sets have their equivalents in traditional statistical analysis, with an *information system* equivalent to a data matrix (that is, a table of values of attributes), an *object* equivalent to a data record, and *attributes* equivalent to variables. For example, if objects are customers in a database, the available information consists of various measures such as demographics and purchase history, which are attributes used to describe each customer. One advantage of rough sets theory over traditional data analysis is that the information (data) is analysed using the assumptions of set theory, and none of the traditional assumptions of multivariate analysis, such as normality of distributions and homogeneity of variance/covariance matrices, are relevant. For an introduction to rough sets, also known as approximation sets, see Pawlak (1991), Munakata (1998), or Pal and Skowron (1999).

A rough set is defined by two sets, called the lower and upper approximations. These are formally defined as:

Let $S = (U, A)$ be an information system, and let $B \subseteq A$ and $X \subseteq U$. We can describe the subset X using only the information contained in the attribute values from the subset B by constructing two subsets, referred to as the B -lower and B -upper approximations of X , and denoted as $B_*(X)$ and $B^*(X)$ respectively, where:

$$B_*(X) = \{ x \mid [x]_B \subseteq X \} \text{ and } B^*(X) = \{ x \mid [x]_B \cap X \neq \emptyset \} \quad (1, 2)$$

The lower approximation, defined in (1), contains objects that are definitely in the subset X and the upper approximation, defined in (2), contains objects that may or may not be in X . In practical data analysis a third subset is also useful, the boundary region, which is the difference between the upper and lower approximations.

Most of the early literature in rough sets concentrated on a specific type of information system, known as a decision system, where at least one of the attributes is a decision attribute (equivalent to a dependent variable in traditional statistical analysis). These applications of rough sets theory use the technique for classification problems, where prior group membership is known (Greco, Matarazzo and Slowinski, 2000; Pawlak, 2000, 2001). More recent rough sets literature has used the technique for clustering problems, where prior group membership is not known. This approach has become known as rough clustering (do Prado, Engel and Filho, 2002; Lingras, 2001, 2002; Voges, Pope and Brown, 2002).

Rough Clustering

The concept of a rough cluster was introduced in Voges et al. (2002), as a simple extension of the notion of rough sets. A rough cluster was defined in a similar manner to a rough set – that is with a lower and upper approximation. The lower approximation of a rough cluster contains objects that only belong to that cluster, and the upper approximation contains objects in the cluster that are also members of other clusters. In rough clustering an object can belong to more than one cluster, whereas in traditional clustering objects belong to one cluster.

Most attempts to apply rough sets theory to clustering have had limitations. The approach developed by Voges et al. generated a large number of clusters, and it was not certain whether the lower approximations of each cluster provided the best coverage of the data set. The approach developed by Lingras (2001, 2002) required that the number of clusters be specified in advance, information that is not always available for large data sets. Voges and Pope (2004) and Voges (2006) presented an extension to their original rough clustering approach that attempted to overcome the limitations of these previous attempts. This approach used an evolutionary algorithm to maximize the coverage of the data set by the clusters, without pre-specifying the number of clusters required.

An Evolutionary Algorithm Based Rough Clustering Approach

Two important considerations in the use of evolutionary algorithms are the data structure that the algorithm operates on, and the level of fitness of each data structure, which helps determine the effectiveness of the solution generated.

Data Structure

In this hybrid technique, the basic building block of the data structure used for describing a rough cluster is the *template*, as described in Nguyen (2000). Formally, let $\mathbf{S} = (U, A)$ be an information system. Any clause of the form $D = (a \in V_a)$ is called a *descriptor*, with the value set V_a called the *range* of D . A template is a conjunction of unique descriptors defined over attributes from $B \subseteq A$. Any propositional formula $T = \bigwedge_{a \in B} (a \in V_a)$ is called a template of \mathbf{S} . To create a viable description of a cluster using a template, at least two attributes from B need to be chosen. This results in compact, but non-trivial, descriptions of the rough cluster being produced. Template T is simple if any descriptor of T has a range of one element. Templates with descriptors having a range of more than one element are called generalized. In the example presented below, only simple templates are used.

The data structure acted on by the evolutionary algorithm is a cluster solution, \mathbf{C} , which is defined as any conjunction of k unique templates,

$$\mathbf{C} = T_1 \wedge T_2 \wedge T_3 \dots \wedge T_k \quad (3)$$

This data structure can be 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. Possible values in the template were the same as the values in the data set (0, 1, 2 or 3), with 0 being used as a “don’t care” value. A template describes a partition of U , and the conjunction of templates contained in a cluster solution results in some templates having both lower approximations (that is, objects satisfying one template only) and upper approximations (that is, objects satisfying more than one template). Consequently \mathbf{C} is a rough cluster solution

Fitness Measure

A number of criteria need to be considered when developing a fitness measure for rough clustering. Firstly, the rough clustering algorithm aims to maximize the data set coverage C , which is defined as the fraction of the universe of objects that matches the set of templates in the cluster solution, C . Secondly, the algorithm aims to minimize k , the number of templates in the cluster solution, C . Finally the accuracy a , of each template needs to be maximized (Pawlak, 1991).

More formally, for any $X \subseteq U$, the set of objects $\{x \in X : \forall a \in B a(x) \in V_a\}$ from X satisfying any template T_i is denoted by $[T_i]X$. $[T_i]^*X$ is a lower approximation if x is unique to that set. $[T_i]^*X$ is an upper approximation if x is contained in $[T_i]X$ and at least one other set $[T_j]X$. We therefore define the following values:

$$c = (\sum | [T_j]^*X |) / | U |, \text{ where } \{1 \leq j \leq k\} \quad (4)$$

That is, the coverage c , is the sum of the cardinal values of the lower approximations of each template in the cluster solution, C , divided by the cardinal value of U , the full data set.

$$a = \sum (| [T_i]^*X | / | [T_i]X |), \text{ where } \{1 \leq j \leq k\} \quad (5)$$

That is, the accuracy a , is the sum of the cardinal value of the lower approximation divided by the cardinal value of the upper approximation for each template in the cluster solution, C .

The fitness value, f , of each cluster solution, C , is defined as the coverage multiplied by accuracy divided by the number of templates in C .

$$f = c \times (a / k) \quad (6)$$

Rough Clustering of City Image Data

The data used in the analyses was collected as part of a wider study, where 6,240 participants were asked to subjectively rank attributes of eleven cities on a three-point scale (High, Moderate, and Low). The eleven city destinations in the Asia Pacific region were Adelaide, Brisbane, Darwin, Melbourne, Perth, Sydney, Auckland, Christchurch, Hong Kong, Singapore, and Tokyo. Initially the questionnaire had a total of 22 items. To simplify the data for analysis, the number of variables was reduced using factor analysis. This factor analysis revealed a stable seven-factor structure, with two items per factor. The seven factors were Language, Safety, Adventure, Standard of Living, Sun and Sand, Information, and Prices.

An evolutionary algorithm based rough clustering algorithm was applied to this data. The rough cluster analysis partitioned the data set of individual perceptions into distinct clusters, based on combinations of attributes considered relevant by the participants.

The “best” cluster solution obtained is shown in Table 1. This cluster achieved coverage of 91.4% of the data set. Table 1 shows the seven templates that comprise this cluster solution, with two of the templates fully enclosed within other templates. Template 1 describes the largest cluster in the solution, comprising 33.6% of the sample. It describes perceptions of cities with a high or medium standard of living and with a high level of interest and adventure. A sub-cluster (Template 1*) completely contained within this cluster (comprising 32.7% of the cluster and 11.0% of the whole sample), describes perceptions of cities with a high standard of living, a high level of interest and adventure, and for which a high level of tourist information is available. Template 2 (28.8% of the sample), describes perceptions of cities with medium levels of adventure, medium beaches and weather, and medium price.

Table 1: “Best” Cluster Solution for Rough Cluster Analysis of City Image Data

T	Variables							Size	%
	Language	Safety	Adventure	Standard of Living	Sun and Sand	Information	Prices		
1	-	-	High	High/Medium	-	-	-	2096	33.6%
1*	-	-	High	High	-	High	-	685	11.0%
2	-	-	Medium	-	Medium	-	Medium	1799	28.8%
3	Medium	-	Low	-	-	-	-	653	10.5%
3*	Medium	-	Low	-	-	-	Medium	575	9.2%
4	-	-	Medium	-	Low	-	Medium	609	9.8%
5	-	-	Medium	-	High	-	Medium	546	8.8%

Template 3 (10.5% of the sample) describes perceptions of cities with medium perceptions of (same) language and low adventure levels – these could be considered to be describing “safe” cities to visit. A sub-cluster (Template 3*) completely contained within this cluster (comprising 88.1% of the cluster and 9.2% of the whole sample), describes perceptions of cities with medium perceptions of (same) language and low adventure levels, coupled to medium price levels. Template 4 (9.8% of the sample) comprises perceptions of cities with medium adventure, low concern with beaches and weather, and medium concern with price. Template 5 (8.8% of the sample) comprises perceptions of cities with medium adventure, high concern with beaches and weather, and medium concern with price.

Conclusion

The use of rough sets theory in data analysis provides a useful alternative to quantitative techniques based on “traditional” statistical approaches. Due to space limitations, this paper has not presented comparisons between rough set approaches and traditional methods, such as discriminant analysis and k-means clustering. However, such comparisons are available in the literature. See, for example, Beynon, Curry and Morgan (2001) for a comparison between rough classification and discriminant analysis, and Voges, Pope and Brown (2002) for a comparison between rough clustering and k-means clustering.

Rough clusters allow an object to belong to multiple clusters. The research presented in this paper uses templates (conjunctions of attribute-value descriptors) to describe the cluster solution. An evolutionary algorithm was used to find a rough cluster solution that covers the largest percentage of the data set with the smallest number of accurate lower approximations. Previous studies using evolutionary algorithms have required that the number of clusters be specified in advance, a major limitation with large or complex data sets, which is overcome in this paper. The paper has also demonstrated the value of the technique with a large (>6,000) data set. Further work will involve expanding the template descriptions to include generalized templates and applying the technique to more data sets.

Rough clustering can be conceptualised as extracting concepts from the data, rather than strictly delineated sub-groupings as in more traditional clustering techniques. The concepts provide interpretations of different tourist preferences present in the data. Such concepts can be an aid to marketers attempting to uncover different segments of consumers. As such, it is a promising technique deserving further investigation.

References

- Beynon, M. J., Curry, B., and Morgan, P. 2001. Knowledge discovery in marketing: An approach through rough set theory. *European Journal of Marketing* 35 (7/8), 915-935.
- Bonn, M. A., Joseph, S. M., and Dai, M. 2005. International versus domestic visitors: An examination of destination image perceptions. *Journal of Travel Research* 43 (3), 294-301.
- Boo, S., and Busser, J. A. 2005. The hierarchical influence of visitor characteristics on tourism destination images. *Journal of Travel and Tourism Marketing* 19 (4), 55-68.
- do Prado, H. A., Engel, P. M., and Filho, H. C. 2002. Rough clustering: An alternative to find meaningful clusters by using the reducts from a dataset. In J. J. Alpigini, J. F. Peters, A. Skowron, and N. Zhong (Eds.). *Rough Sets and Current Trends in Computing*, Third International Conference RSCTC 2002 LNCS 2475. Berlin: Springer-Verlag, pp. 234-238.
- Fernandes, E. 2002. Polishing the country's image: The economic effect of a 15% decline in the number of visitors prompts a campaign to relaunch India as a holiday destination. *Financial Times*, September 23 2002, p.2.
- Greco, S., Matarazzo, B., and Slowinski, R. 2000. Extension of the rough set approach to multicriteria decision support. *INFOR* 38 (3), 161-195.
- Gunn, C. 1972. *Vacationscape*. Austin, TX: Bureau of Business Research, University of Texas.
- Hair, J. E., Anderson, R. E., Tatham, R. L., and Black, W. C. 1998. *Multivariate Data Analysis* (5th ed.). London: Prentice-Hall International.
- Hankinson, G. 2005. Destination brand images: a business tourism perspective. *Journal of Services Marketing* 19 (1), 24-32.
- Hanlan, J., and Kelly, S. 2005. Image formation, information sources and an iconic Australian tourist destination. *Journal of Vacation Marketing* 11 (2), 163-177.
- Hunt, J. D. 1975. Image as a factor in tourism development. *Journal of Travel Research* 13 (4), 1-7.
- Kim, S. S., Guo, Y., and Agrusa, J. 2005. Preference and positioning analyses of overseas destinations by mainland Chinese outbound pleasure tourists. *Journal of Travel Research* 44 (2), 212-220.
- Konecnik, M. 2005. Slovenia as a tourism destination: Differences in image evaluations perceived by tourism representatives from closer and more distant markets. *Economic and Business Review for Central and South-Eastern Europe* 7 (3), 261-282.
- Lingras, P. 2001. Unsupervised rough set classification using GAs. *Journal of Intelligent Information Systems* 16 (3), 215-228.
- Lingras, P. 2002. Rough set clustering for web mining. In *Proceedings of 2002 IEEE International Conference on Fuzzy Systems*.
- MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. In L. M. Le Cam, and J. Neyman (Eds.) *Proceedings of the Fifth Berkeley Symposium on Mathematics, Statistics and Probability*, Volume 1. Berkeley, CA: University of California Press, pp. 281-298.
- Munakata, T. 1998. *Fundamentals of the New Artificial Intelligence: Beyond Traditional Paradigms*. New York: Springer-Verlag.
- Morgan, N. J., Pritchard, A., and Piggott, R. 2003. Destination branding and the role of the stakeholders: The case of New Zealand. *Journal of Vacation Marketing* 9 (3), 285-299.
- Nguyen, S. H. 2000. Regularity analysis and its applications in data mining. In L. Polkowski, S. Tsumoto and T. Y. Lin (Eds.) *Rough Set Methods and Applications: New Developments in Knowledge Discovery in Information Systems*. Heidelberg: Physica-Verlag, pp. 289-378.

- Obenour, W., Lengfelder, J., and Groves, D. 2005. The development of a destination through the image assessment of six geographic markets. *Journal of Vacation Marketing* 11 (2), 107-113.
- Pal, S. K., and Skowron A. (Eds.) 1999. *Rough-Fuzzy Hybridization: A New Trend in Decision Making*. Singapore: Springer.
- Pawlak, Z. 1982. Rough sets. *International Journal of Information and Computer Sciences* 11, 341-356
- Pawlak, Z. 1991. *Rough Sets: Theoretical Aspects of Reasoning About Data*. Boston: Kluwer.
- Pawlak, Z. 2000. Rough sets and decision analysis. *INFOR* 38 (3), 132-144.
- Pawlak, Z. 2001. Rough sets and decision algorithms. In W. Ziarko and Y. Yao (Eds.) *Rough Sets and Current Trends in Computing (Second International Conference, RSCTC2000)*. Berlin: Springer, pp. 30-45.
- Silvestre, A. L., and Correia, A. 2005. A second-order factor analysis model for measuring tourists' overall image of Algarve, Portugal. *Tourism Economics* 11 (4), 539-554.
- Son, A. 2005. The measurement of tourist destination image: applying a sketch map technique. *International Journal of Tourism Research* 7 (4/5), 279-294.
- Son, A., and Pearce, P. 2005. Multi-faceted image assessment: International students' views of Australia as a tourist destination. *Journal of Travel and Tourism Marketing* 18 (4), 21-35.
- Stateman, A. 2005. After the tsunami: The complexities of promoting areas damaged by disasters. *Public Relations Tactics* 12 (2), 10.
- Voges, K. E. 2006. Rough clustering of destination image data using an evolutionary algorithm. *Journal of Travel and Tourism Marketing (Special Issue on "New Quantitative Models in Travel and Tourism Research")* 21 (4), 121-137
- Voges, K. E., and Pope, N. K. Ll. 2004. Generating compact rough cluster descriptions using an evolutionary algorithm. In K. Deb et al. (Eds.) *GECCO2004: Genetic and Evolutionary Algorithm Conference - LNCS 3103*. Berlin: Springer-Verlag, pp. 1332-1333.
- Voges, K. E., Pope, N. K. Ll., and Brown, M. R. 2002. Cluster analysis of marketing data examining on-line shopping orientation: A comparison of k-means and rough clustering approaches. In H. A. Abbass, R. A. Sarker, and C. S. Newton (Eds.) *Heuristics and Optimization for Knowledge Discovery*. Hershey, PA: Idea Group Publishing pp. 207-224.