An Information-Theoretic Approach to Non-Parametric Clustering

Emmanuel M. Pothos* and **Nick Chater****

*Department of Experimental Psychology
University of Oxford
Oxford OX1 3UD, UK.
pothos@psy.ox.ac.uk

**Department of Psychology
University of Warwick
Coventry CV4 7AL, UK.
n.chater@warwick.ac.uk

Introduction

A fundamental problem for any statistician faced with a noisy data set or for any living creature trying to survive is to identify how much (if any) useful structure exists in a noisy input. In other words, in a given data set one wishes to identify as much redundancy as possible so as to reduce the amount of information needed to code for the data.

In both cases clustering techniques aim at identifying groups of individuals within an input space such that the ones within clusters are more similar to the ones outside clusters, so that the domain is partitioned into disjoint or properly inclusive subsets. We interpret the clustering problem as one of data compression within an information theory framework and use the minimum description length principle (henceforth MDL; Rissanen, 1978) to derive a non-parametric algorithm which identifies the clustering configuration maximally compressing a given data set. Encouraging human experimental results are presented, highly suggestive of a low-level categorization model based on information gain.

Clustering by MDL

The problem of identifying the set of clusters that best capture the statistical structure of the domain can be redescribed as follows. Suppose we are interested in transmitting a set of ordered relations among n objects (so that we have $n*(n-1)/2$ relations), of the form $a<b$, $a<d$, $b<c$ etc.). If we were to transmit each relation individually then we would require $n*(n-1)/2$ bits.

Specifying a cluster is equivalent to saying that all the distances *within* clusters are less than all the distances *between* clusters so that the total number of relations that needs to be transmitted is reduced. In particular, if s constraints are introduced by a set of clusters, then the information gain is s bits. However, in general no cluster will be perfect, so that some of these constraints will be wrong, and also the particular way clusters divide up the data set needs to be described. According to the MDL principle, a cluster solution will be advantageous to the extent that the above costs are less than the information gain associated with the constraints imposed by clusters. Summing up, the information gain associated with transmitting a cluster solution instead of all relations individually would be:

Information Gain = Constraints - Costs, where Costs are given by $\left(\log(s+1)+\log({}^{r}C_{s})\right)+\sum_{v=0}^{n}(-1)^{v}\frac{(n-v)^{r}}{(n-v)!v!}$.

(See Chater & Pothos, in preparation.)

Implementation of the algorithm

The algorithm proceeds by grouping cases into bigger and bigger clusters. At the beginning, all items are considered individual clusters. At each step, all possible cluster combinations are formed and the total (domain) transmission costs are calculated. The cluster that leads to the greatest information gain over the previous configuration is the one to be formed. The algorithm stops when no more information gain is possible.

Performance of the algorithm

Four simple data sets (10 points embedded in a 2D Euclidean space) were used to illustrate performance of the algorithm. The first three data sets were highly structured and the algorithm indeed partitioned the data points in the way that seemed to reflect most faithfully the structure of the domains. The fourth data set was constructed so that there would be little structure and this led to a very low information gain associated with the final configuration.

Comparisons with human results

The data points in each of the data sets used in validating the performance of the algorithm were then printed on A4 sheets of paper and we asked human subjects to partition these domains in a way that seemed "natural and intuitive", by drawing a curve round the points. Chi-squared tests revealed that the algorithm cluster solutions were the only ones that were produced significantly more frequently against chance for the first three data sets, while for the last one (the little structure data set) all solutions were at chance frequencies

References

Chater, N., & Pothos, E. M. (in preparation). Non-parametric clustering by minimum description length.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14, 465-471.