

A Likelihood Approach for Determining Cluster Number

William D. Shannon, Tsvika Klein, Robert Culverhouse

**Washington University in St. Louis School of Medicine
660 S. Euclid Ave, Campus Box 8005
St. Louis, MO 63119**

Phone/Fax: 314-454-8356/314-454-5113

Email: shannon@lva.wustl.edu, tklein@im.wustl.edu, rob@frodo.wustl.edu

Abstract

Deciding where to cut the dendrogram produced by a hierarchical cluster analysis is known as a stopping rule. Heuristic approaches proposed for solving this problem have been based on statistics such as the proportion of variance accounted for by the clusters. The statistic is calculated on each of the sets of clusters produced by cutting the dendrogram at successive heights. The number of clusters in the set that optimizes the statistic estimates the true number of clusters. This is a reasonable ad hoc measure, but is not based on a probability model of cluster distributions. Therefore, the properties of statistical decision theory (e.g., Type I and II error) do not hold, and the associated P-values are meaningless. In this presentation we propose a novel stopping rule based on a probability model for graphical objects. The application of probability models to hierarchical trees is highly speculative, but is based on prior published work (Banks and Constantine 1998; Shannon and Banks 1999). We extend this prior work to derive a likelihood and a likelihood-ratio test (LRT) for determining the number of clusters in a dataset. We are aware that the criteria for the LRT (Lehmann 1999) are not fully met so P-values will be approximations at best, though bootstrap P-values might easily be estimated. We are beginning to contrast the likelihood-ratio test stopping rule with other existing ad hoc approaches. In our paper we present this method for the first time and show some very preliminary results.

Introduction

Cluster analysis can be viewed as a data reduction method in that all the objects in a group can be represented by an 'average' of that group. In cluster analysis, the goal is to find groups such that objects within a group are more similar to each other than they are to objects in different groups. A large number of approaches are available for clustering, including hierarchical clustering (Everitt and Rabe-Hesketh 1997; Eisen, Spellman et al. 1998; Shannon, Culverhouse et al. 2003) and k-means clustering (Hartigan and Wong 1979) from the statistical literature, and self-organizing maps (Kohonen 1990) and artificial neural networks from the machine learning literature. These algorithms are equivalent in terms of performance (i.e., one method does not dominate all others). Our research focuses on hierarchical clustering, though the methods we develop are applicable to other clustering methods. For an overview of hierarchical clustering the reader is referred to our recent review article (Shannon, Culverhouse et al. 2003). There are numerous references to the mathematics and multivariate statistics used in cluster analysis, clustering methods from both statistics and machine learning, and the application of clustering methods to biological data (Legendre and Legendre 1998; Hastie, Tibshirani et al. 2001; Timm 2002).

Unlike standard statistical methods such as the t-test and analysis-of-variance, clustering does not have a probabilistic foundation. Because of this, there are no statistical tests to determine how many subgroups exist naturally in the data. While it is possible to compute a formal test statistic, such as an F-test statistic, the assumptions of the statistical test are not met and the P-value has no interpretation. Pseudo-statistical methods to determine the correct number of clusters in a data set have recently been developed. However, these methods are ad hoc, have not generally been accepted by the statistical community, and have not been routinely incorporated into data analyses. One such method is to use the number of clusters that optimizes the Gap statistic comparing within-cluster dispersion to dispersion under the null hypothesis (Hastie, Tibshirani et al. 2000; Hastie, Tibshirani et al. 2001; Timm 2002). Another approach uses a perturbation, or sensitivity analysis, method. This method clusters the data, then, after introducing a small amount of random noise to the data, reclusters the data. The clusters from the perturbed data are then compared to the results to the original clustering (Bittner, Meltzer et al. 2000). This approach is based on consensus methods for combining cluster results (McMorris and Neumann 1983). Many other methods developed over the years have been studied extensively and none shown to have optimal statistical properties for a wide range of data structures (Milligan and Cooper 1985).

In the absence of statistical tests, external criteria are used to choose the number of clusters. One such criterion is that if splitting a tree at a particular height produces clusters of patients that are nearly homogeneous with regard to an important property, the split is deemed appropriate. For example, if splitting a tree at a particular height results in mostly tumor samples in one cluster and mostly normal samples in the other, the split is considered interesting. The variables used to create that hierarchical tree would then be studied to see if and how they are involved in tumor biology, specifically in distinguishing tumor from normal tissue. The obvious problem with this approach is the subjective nature of deciding which external criteria to use and what 'mostly' means when saying that the clusters contain mostly one type of sample or another.

A second difficulty with cluster analysis is that the algorithms are guaranteed to produce clusters from any data. There is currently no accepted way to test the null hypothesis that there are no clusters in the data (e.g., data are distributed uniformly). For this reason, caution is required in interpreting the results of a cluster analysis. The results always need

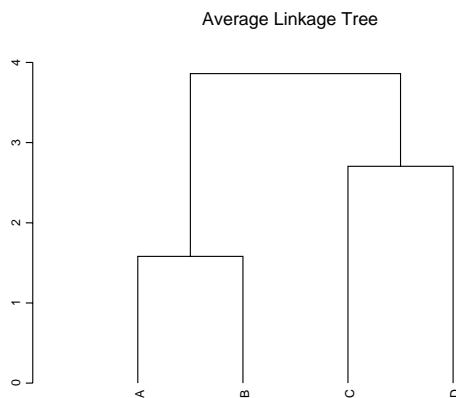
to be examined to see if it is plausible that the resulting clusters are natural groups and not just artifacts of the algorithm.

In spite of these two shortcomings, cluster analysis has proven itself to be a powerful exploratory data analysis tool. In fact, clustering forms the foundation for many applied areas including evolutionary biology, microarray data analysis, and market segmentation in business. It also is cited as the predominant approach for data mining.

This paper describes an approach based on probability and graph theory to formally decide the optimal number of clusters in a dataset. Our hope is to provide a new tool based on formal statistical and probabilistic thinking to assist the applied data analyst in using clustering methods.

Stopping Rules

We consider the special case of a hierarchical clustering where the cluster model can be represented by a dendrogram. Consider the adjacent dendrogram where four objects, A, B, C, and D, have been clustered using an average clustering algorithm. Once the dendrogram has been fit the analyst must decide where to cut the tree to induce a partition of the objects. For this simple case, the possible solutions are not to cut the tree (one cluster $\{A, B, C, D\}$), to cut the tree at level 3 (two clusters $\{A, B\}$, $\{C, D\}$), to cut the tree at level 2 (three clusters $\{A, B\}$, $\{C\}$, $\{D\}$), or to cut the tree at level 1 (four clusters $\{A\}$, $\{B\}$, $\{C\}$, $\{D\}$).



Many rules have been proposed for deciding the number of clusters and are often referred to as 'stopping rules' since they determine where to stop cutting the dendrogram. For an overview and discussion of their relative accuracies in a large-scale simulation study, the reader is referred to Milligan (1985).

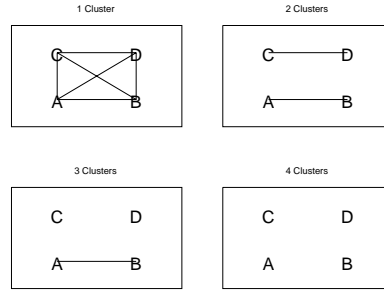
Current stopping rules generally are heuristic approaches based on optimization criteria. For

example, a likelihood ratio test compares k to $k-1$ clusters sequentially applied to each new cut of a tree. (Note: each new cut results in one cluster being split into two clusters.) Although this method was modeled after Wilks' likelihood ratio criterion assuming multivariate normality, subsequent work showed this criterion is not asymptotically distributed as chi-square. Thus, P-values obtained from this method are approximations. Another stopping rule is the cubic clustering criterion available in SAS. This method selects the number of clusters based on the proportion of variance explained by the clusters. This criterion makes major assumptions concerning the distribution and size of the clusters that are generally not realistic. When these assumptions are violated, the performance of this criterion is significantly degraded.

These and other stopping rules are not rigorous decision criteria since no test of statistical significance is performed. The method we propose in this paper is an attempt to develop a stopping rule providing a valid P-value based on a firm theoretical foundation with few assumptions made regarding the structure of the data or clusters.

Representing Clusters as Graphs

Our method requires representing the clusters as complete graphs. Consider the adjacent figure of graphs for each level of cutting the above dendrogram. Four possible graphs can be induced from this tree depending on the number of clusters: 1 cluster where each object A, B, C, D are connected by an edge, 2 clusters where objects A and B are connected and objects C and D are connected by an edge, etc.



In other words, each data point is represented by a node and, for any two points in the same cluster, their corresponding nodes have an undirected edge between them. Using this representation, cluster results can be modeled using a probability distribution defined on sets of graphs. We outline this model in the next section.

Probability Model

In this section we define a probability model applicable to graphs, and show how it can be used to measure the variability of cluster analysis and to estimate the number of clusters in a data set. This model allows the use of the standard tools of statistical inference such as likelihood ratio tests, confidence interval estimation, and goodness-of-fit tests. It has previously been derived and applied to several types of graphs such as cluster trees, digraphs, and recursively partitioned trees (Banks and Constantine 1998; Shannon and Banks 1999).

Let G be a finite set of graphs, denote the elements of G by g , and let $d : G \times G \rightarrow \mathfrak{R}$ be an arbitrary metric on G . Consider the distribution on G defined by

$$p(g) = C e^{-\tau d(g, g^*)}, \quad \forall g \in G$$

where g^* is the central graph, $\tau > 0$ is a concentration parameter, C is the normalizing constant, and d is Hamming distance (defined as the number of edge discrepancies between two graphs (Hamming 1950)). Intuitively we can view g^* as the ‘mean’ and τ as a measure of precision (the inverse of variance) of the distribution. As $\tau \rightarrow 0$, every $g \in G$ becomes equiprobable; thus all g are modal and there is no central tendency.

When τ is large, sampled graphs will strongly concentrate around g^* . If this is the case, observed data will contain much useful information for the estimation of g^* .

For a random sample of n observed graphs, $g_1, g_2, \dots, g_n \in G$, the log-likelihood of this probability model is

$$\ln L(g^*, \tau) = -n \ln \left[\sum_{g \in G} e^{-\tau d(g, g^*)} \right] - \tau \sum_{i=1}^n d(g_i, g^*).$$

In our application we assume g^* and τ are unknown parameters to be estimated from the data.

Maximum Likelihood Estimates

The MLE's \hat{g}^* and $\hat{\tau}$ can be used to summarize a set of graphs within the framework of a probability model. We describe below how we would obtain these MLE's from a set of graphs. Banks and Constantine (Banks and Constantine 1998) derived the following closed form solutions for the ML estimates $(\hat{g}^*, \hat{\tau})$ for connected graphs. One finds that \hat{g}^* contains an edge between two nodes if and only if the edge is in more than half of the sample graphs. Given \hat{g}^* we calculate

$$\hat{\tau} = -\ln \left(\frac{W}{1-W} \right) \quad \text{where} \quad W = m \sum_{i=1}^n d(g_i, \hat{g}^*),$$

n is the number of graphs in the sample, and m is a constant determined by the number of nodes. We used a bootstrap algorithm to generate the sample of graphs, though other approaches, such as using various distance metrics and clustering algorithms, could be used. The bootstrap algorithm is described in more detail below.

Likelihood Ratio Test

Given our bootstrap sample of graphs, the likelihood function can be estimated and a likelihood-ratio test can be calculated to test various value of k , the number of clusters, by

$$\lambda_k = -2 \ln \frac{L(\hat{\mathbf{g}}_0^*, \hat{\tau}_0)}{L(\hat{\mathbf{g}}_k^*, \hat{\tau}_k)}$$

where $(\hat{\mathbf{g}}_0^*, \hat{\tau}_0)$ are the MLEs of central graph and estimate of τ for the null. The null might be specified as no clusters ($k = 1$) or in a sequential manner where there are one fewer cluster ($k - 1$) than the number currently being tested by the MLEs $(\hat{\mathbf{g}}_k^*, \hat{\tau}_k)$ for the population of graphs induced by cutting the bootstrap dendrograms at level k .

We will assess the performance of λ by observing how often its maximum value occurs at the true k . In our examples this should be at $k = 3$ clusters. (In the future we have planned more extensive simulation studies to more accurately test its performance, but for now only report on this simple metric of performance.)

Sampling Graphs

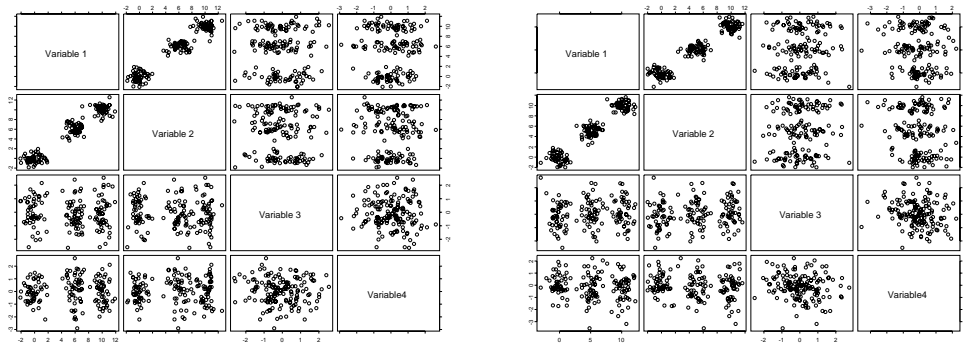
Our approach requires that a sample of graphs representing the population of graphs that could be obtained by cluster analysis be available for calculating the MLEs. We use bootstrapping to generate this sample of graphs. If we are interested in testing for k possible clusters, we proceed as follows:

1. Generate a bootstrap sample of the objects to be clustered;
2. Fit a dendrogram to the bootstrap sample;
3. Cut the bootstrap dendrogram at the level to generate k clusters;
4. For each of the k clusters, calculate the cluster's mean vector over the variables used in the analysis;
5. Assign each object in the original dataset to the closest mean vector – objects assigned to the same mean vector are in the same cluster and so are connected by an edge;
6. Repeat Steps 1-5 to generate the population of graphs.

We assume the variability in the sample of graphs is related to the stability of the clusters obtained by cutting each bootstrap tree at level k . Suppose a subset of the objects being clustered have very small pairwise distances so are always clustered together. The objects in this subset will always have an edge connecting them in each graph induced by cutting the dendrograms at level k . This subset will then appear as a connected group (i.e., will lie within a cluster) in the MLE $\hat{\mathbf{g}}^*$. Furthermore, suppose other objects are sometimes connected to this subset and other times are not connected to this subset. Depending on how often these other objects are connected, which is a function of the proximity of these objects to the subset, they may or may not be connected in the MLE $\hat{\mathbf{g}}^*$.

Examples

We simulated two examples to test the feasibility of our method. Each simulation consists of three multivariate normal groups. Each variable had variance 1 and all covariance's were 0. In the first example shown in the first scatterplot matrix, each group is defined by the first two variables whose means are (0, 0), (6, 6), and (10, 10), and two other uninformative variables with means (0, 0). In the second example shown in the other scatterplot matrix, the second group is shifted so the first two variables have means (5, 5). All other parameters are the same.



In example 1 the second group is shifted towards group 3 and away from group 1. In the second example the three groups are equally separated.

We ran the algorithm on 100 simulations of each example, where each simulation consisted of 50 data points from each group. For each simulation we generated a population of 100 bootstrapped graphs to fit the MLEs $(\hat{g}^*, \hat{\tau})$. The values of the likelihood ratio and τ are averaged over the 100 simulations. The results for k ranging from 2 to 10 are as follows:

Clusters (k)	Example 1		Example 2	
	Likelihood Ratio (λ)	Tau (τ)	Likelihood Ratio (λ)	Tau (τ)
2	Undefined	Infinite	-15047	2.49
3	Undefined	Infinite	Undefined	Infinite
4	-10161	3.06	-10634	3.00
5	-11765	2.86	-7707	3.44
6	-12481	2.77	-12453	2.77
7	-8458	3.31	-14576	2.54
8	-12328	2.79	-13966	2.60
9	-12597	2.76	-11965	2.83
10	-12055	2.82	-11254	2.92

When τ is 0, the sample graphs exhibited no variability. This caused the likelihood ratio to be undefined. In the first example, this result occurred because every bootstrap dendrogram at $k = 2$ separated group 1 into one cluster and groups 2 and 3 into another cluster, and at $k = 3$ split groups 2 and 3 into their own clusters. This occurred each time because of the large separation among all the groups, and because group 2 was shifted towards group 3.

In example 2 when setting $k = 2$ some bootstrap dendrograms clustered group 2 with group 1, and other dendrograms clustered group 2 with group 3. This resulted in variability among the graphs and τ was estimated at 2.49. When setting $k = 3$ each bootstrap dendrogram then perfectly separated groups 1, 2, and 3 and there was again no variability in the graphs making τ infinite and the likelihood ratio undefined.

Discussion

We have presented a stopping rule method that is more statistically rigorous using a likelihood ratio test, and have presented some very preliminary results. Several questions remain open that we are currently investigating.

First, can we use the undefined likelihood ratio test (LRT) when τ is infinite as the decision rule for determining the number of clusters when there is little variability among the sample of graphs? In these simple examples the lack of variability among the graphs was due to large separation of groups in the data, and the undefined likelihood ratio occurred at the correct number of clusters. Will this be a serious problem in realistic settings? Second, what are the properties of the LRT over a range of k ? The probability model defined by Banks and Constantine provides a new approach to the analysis of graphical objects, and the performance of this measure is still under investigation. Third, how well does our decision rule using the LRT perform with respect to other measures of clustering performance, such as the adjusted Rand statistic, which are more related to the field of cluster analysis? Fourth, since the criteria guaranteeing that the LRT will be distributed as a chi-square statistic (Lehmann 1999) are not fully met, will the standard assumption of the LRT as a chi-squared statistic be sufficiently close to serve as a good approximation, or will we need to develop a permutation approach for calculating the P-value?

g. Literature Cited

- Banks, D. and G. Constantine (1998). "Metric models for random graphs." Journal of Classification **15**(199-223).
- Bittner, M., P. Meltzer, et al. (2000). "Molecular classification of cutaneous malignant melanoma by gene expression profiling." Nature **406**(6795): 536-40.
- Eisen, M., P. Spellman, et al. (1998). "Cluster analysis and display of genome-wide expression patterns." Proceedings of the National Academy of Sciences **95**: 14863-14868.
- Everitt, B. and S. Rabe-Hesketh (1997). The analysis of proximity data. New York City, John Wiley.
- Hamming, R. (1950). "Error detecting and error correcting codes." Bell Systems Technical Journal **29**: 147-160.
- Hartigan, J. and M. Wong (1979). "A k-means clustering algorithm." Applied Statistics **28**: 100-108.
- Hastie, T., R. Tibshirani, et al. (2000). "Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns." Genome Biol **1**(2): RESEARCH0003.
- Hastie, T., R. Tibshirani, et al. (2001). The elements of statistical learning. New York City, Springer.
- Kohonen, T. (1990). "The self-organizing map." Proceedings of the IEEE **78**: 1464-1479.
- Legendre, P. and L. Legendre (1998). Numerical Ecology. New York City, Elsevier.
- Lehmann, E. (1999). Elements of Large Sample Theory. New York City, Springer.
- McMorris, F. and D. Neumann (1983). "Consensus functions defined on trees." Mathematical Social Sciences **4**: 131-136.
- Milligan, G. and M. Cooper (1985). "An examination of procedures for determining the number of clusters in a data set." Psychometrika **50**: 159-179.
- Shannon, W. and D. Banks (1999). "Combining classification trees using maximum likelihood estimation." Statistics In Medicine **18**(6): 727-740.
- Shannon, W., R. Culverhouse, et al. (2003). "Analyzing microarray data using cluster analysis." Pharmacogenomics **4**(1): 41-51.
- Timm, N. (2002). Applied Multivariate Analysis. New York City, Springer.