Constructing Optimal Ultrametrics

N. Sriram
National University of Singapore

Scott Lewis
Bell Laboratories

Abstract: Clique optimization (CLOPT) is a family of graph clustering procedures that construct parsimonious ultrametrics by executing a sequence of divisive and agglomerative operations. Every CLOPT procedure is associated with a distinct graph-partitioning heuristic. Seven HCS methods, a mathematical programming algorithm, and two CLOPT heuristics were evaluated on simulated data. These data were obtained by distorting ultrametric partitions and hierarchies. In general, internally optimal models yielded externally optimal models. By recovering near-optimal solutions more consistently, CLOPT2 emerged as the most robust technique.

Keywords: Clustering; Graph; Heuristic; Hierarchical; Simulation.

1. Introduction

Ultrametrics are extensively used to model proximity data by researchers in many disciplines (Anderberg 1973). A variety of clustering techniques have been developed to construct ultrametric trees (Gordon 1987). If the data satisfy the ultrametric inequality, then many methods (e.g., single link and

This paper is based on portions of the first author's dissertation. Research and manuscript preparation for this article were supported by a National University of Singapore Grant RP 900042 to the first author. We thank the editor and referees for many valuable suggestions.

Author's Address: Correspondence regarding this paper should be directed to: N. Sriram, 10 Kent Ridge Crescent, Department of Social work and Psychology, National University of Singapore, Singapore 0511. e-mail: swknasri@nusvm.nus.sg.
complete link) construct identical hierarchies (Johnson 1967). In general, however, empirical data are not perfectly ultrametric, and the various methods approximate the data by different ultrametric models. Here, the researcher is confronted with the dilemma of selecting from diverse ultrametrics the particular solution that is, in some sense, "optimal."

To compare the quality of competing models, the notion of optimality must be made precise. In this paper we will consider two conceptions of optimality. A model is *internally optimal* if it optimizes (maximizes or minimizes) some function that indexes the degree of fit of the model to the data. Many functions can measure the degree of fit between the model and the data. One common index is the sum of the squares of the differences between the model and the proximity data. The construction of internally optimal ultrametrics, for both the absolute difference and the least squares functions, is an NP-hard problem (Krivánek 1986; Krivánek and Moravek 1986). NP-hardness implies that any algorithm guaranteeing optimal ultrametrics will have exponential time complexity; i.e., in general, the time required to construct the globally optimal ultrametric increases exponentially with the number of objects. An example is the globally convergent least squares ultrametric tree fitting procedure outlined by Chandon and De Soete (1984). Consequently, most optimization procedures seek local maxima (minima) in restricted solution spaces.

An alternative conception of optimality is commonplace in the literature devoted to the validation of clustering methods (Dubes and Jain 1979; Milligan 1981a). In these studies artificial data, distortions of known structures, are used to compare clustering procedures. A model is *externally optimal* if it optimizes some index of the fit between the model and the (known) structure underlying the data. Here, the emphasis is on how well the method recovers the underlying structure and less on how closely it approximates the data. In practical applications, there is usually no external criterion with which to evaluate the fit of the model. Thus, optimization procedures use internal criteria to converge to their results. In many instances, the model that is internally optimal is also externally optimal; however, as we shall see later, this is not generally the case.

The popular agglomerative hierarchical clustering schemes (HCS) used to construct ultrametrics are, in general, not explicit optimization procedures.\(^1\) Instead, they carry out a sequence of transformations that do not

---

1. But see Hansen, Jaumard, and Frank (1989) for the characterization of the single link algorithm as an optimization procedure.