Chapter 4

Determination of the Number of Clusters in a Data Set:
A Stopping Rule × Clustering Algorithm Comparison

Derrick S. Boone Sr.
Wake Forest University - Schools of Business, USA

ABSTRACT

The accuracy of “stopping rules” for determining the number of clusters in a data set is examined as a function of the underlying clustering algorithm being used. Using a Monte Carlo study, various stopping rules, used in conjunction with six clustering algorithms, are compared to determine which rule/algorithm combinations best recover the true number of clusters. The rules and algorithms are tested using disparately sized, artificially generated data sets that contained multiple numbers and levels of clusters, variables, noise, outliers, and elongated and unequally sized clusters. The results indicate that stopping rule accuracy depends on the underlying clustering algorithm being used. The cubic clustering criterion (CCC), when used in conjunction with mixture models or Ward’s method, recovers the true number of clusters more accurately than other rules and algorithms. However, the CCC was more likely than other stopping rules to report more clusters than are actually present. Implications are discussed.
INTRODUCTION

Although market segmentation is a key element of a firm’s marketing strategy, its implementation can be problematic (Yankelovich & Meer, 2006). For example, firms must determine not only how to assign consumers to specific market segments, but also the number of segments, if any, that exist. Cluster analysis, the analytical technique most often used in market segmentation studies, addresses the challenges associated with assigning consumers to specific segments (Dickinson, 1990; Viren, Wedel, & Wilms, 1996; Chaturvedi et al., 1997; Wedel & Kamakura, 2000). “Stopping rules,” analytical techniques for determining the number of clusters in a data set, address the challenges associated with determining the number of segments that exist in a market.

Cluster analysis, given its importance not only in business but also in other fields (e.g., psychometrics, taxonomy, and signal processing), is a fairly well researched topic. Stopping rules should be similarly well researched because determining the number of clusters in a data set is often an integral part of cluster analysis applications. However, much of the research investigating stopping rules has tended to focus on the development of new rules rather than a comparative assessment of the rules currently in existence (Arabie, Hubert, & De Soete, 1996; Li et al., 2008; Liao & Ng, 2009; Milligan & Cooper, 1985; Wedel & Kamakura, 2000).

This research oversight is problematic because previous findings regarding the accuracy of stopping rules may be misleading (Milligan & Hirtle, 2003). Accordingly, the goals of this research are threefold. First, this research seeks to determine if the accuracy of a given stopping rule varies as a function of the underlying clustering algorithm. Second, this research seeks to compare the relative accuracies of various rules/algorithms under data conditions likely to be encountered by researchers. However, given the complexities of real-world data and the suitability of different algorithms in different data environments, this research does not attempt to identify a single rule/algorithm combination that is universally superior. Instead, and third, this research seeks to identify a set of rules/algorithms that may serve as a guide for researchers and practitioners when trying determine the number of clusters in a data set.

The accuracy rates of six stopping rules, used in conjunction with six clustering algorithms, are examined using 648 artificially generated data sets that contain multiple numbers and levels of variables, observations, clusters, noise, outliers, and elongated and unequally sized clusters. The results indicate that certain rule/algorithm combinations are better than others for determining the number of clusters in a data set.

The remainder of the paper is organized as follows. In the next section, the literature on stopping rules is briefly reviewed. The comparative stopping rules and clustering algorithms used in the study are then presented, followed by a discussion of how the data sets were generated and the experimental procedures established. The accuracy rates of the rule/algorithm combinations are reported, and finally conclusions and limitations and directions for future research are offered.

LITERATURE REVIEW

The problem of determining the number of clusters in a data set is relevant across a wide variety of disciplines such as business, psychology, sociology, statistics, biology, engineering, and computer science (Li et al., 2008; Liao & Ng, 2009; Milligan & Cooper, 1985; Wedel & Kamakura, 2000). The following review briefly highlights many of these techniques and while not exhaustive, is designed to provide a sense of the various approaches that have been proposed (Carmone, Kara, & Maxwell, 1999; Dolnicar & Leisch, 2010; Li et al., 2008; Milligan & Cooper, 1985; Milligan & Hirtle, 2003; Maulik & Bandyopadhyay, 2002; Salem & Nandi, 2009).
13 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the product's webpage:  
www.igi-global.com/chapter/determination-number-clusters-data-set/70951?camid=4v1

www.igi-global.com/e-resources/library-recommendation/?id=1

Related Content

Explaining Involuntary Spinoffs from Teams  
www.igi-global.com/chapter/explaining-involuntary-spinoffs-teams/70959?camid=4v1a

The Information Sector in the Economy and its Strategic Value  
www.igi-global.com/chapter/information-sector-economy-its-strategic/38425?camid=4v1a

Decision Support System for Sanitary Teams Activities  
www.igi-global.com/article/decision-support-system-for-sanitary-teams-activities/124322?camid=4v1a

Hybrid Multi-Criteria Models: Joint Health and Safety Unit Selection on Hybrid Multi-Criteria Decision Making  
Ömer Faruk Efe (2020). Multi-Criteria Decision Analysis in Management (pp. 62-84).  
www.igi-global.com/chapter/hybrid-multi-criteria-models/249265?camid=4v1a