Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2006) 526–544
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

A toolbox for K-centroids cluster analysis

Friedrich Leisch*

Department of Statistics and Probability Theory, Vienna University of Technology, 1040 Vienna, Austria

Received 3 June 2005; received in revised form 11 October 2005; accepted 13 October 2005
Available online 4 November 2005

Abstract

A methodological and computational framework for centroid-based partitioning cluster analysis using arbitrary distance or similarity measures is presented. The power of high-level statistical computing environments like R enables data analysts to easily try out various distance measures with only minimal programming effort. A new variant of centroid neighborhood graphs is introduced which gives insight into the relationships between adjacent clusters. Artificial examples and a case study from marketing research are used to demonstrate the influence of distances measures on partitions and usage of neighborhood graphs.

Keywords: Cluster analysis; Distance measures; R

1. Introduction

New developments for partitioning cluster analysis have been dominated by algorithmic innovations over the last decades, with the same small set of distance measures used in most of them. Especially the machine learning literature is full of "new" cluster algorithms for Euclidean or—to much lesser extent—L^1 distance (also called Manhattan distance). For hierarchical cluster analysis it has always been natural to treat distance (or similarity measures) and details of the cluster algorithm like the linkage method at par, and most software implementations reflect this fact, offering the user a wide range of different distance measures. In programmatic environments like S (Becker et al., 1988) hierarchical clustering often works off a distance matrix, allowing for arbitrary distance measures to be used.

Most monographs on cluster analysis do the same for partitioning cluster analysis and state the corresponding algorithms in terms of arbitrary distance measures, see e.g., Anderberg (1973) for an early overview. However, partitioning cluster analysis has often been the method of choice for segmenting "large" data sets, with the notion of what exactly a "large" data set is changing considerably over time. So computational efficiency always was an important consideration for partitioning cluster algorithms, narrowing the choice of distance methods to those where closed-form solutions for cluster centroids exist: Euclidean and Manhattan distance. Other distance measures have been used for the task, of course, but more often than not "new" cluster algorithms got invented to deal with them.

The goal of this paper is to shift some of the focus in partitioning cluster analysis from algorithms to the distance measure used. Algorithms have an important influence on clustering (convergence in local optima, dependency on starting values, etc.), but their efficiency should no longer be the main consideration given modern computing power. As an example for an "unusual" distance measure, Heyer et al. (1999) use the so-called jacknife correlation for clustering

* Tel.: +43 1 58801 10715; fax: +43 1 58801 10798.
E-mail address: friedrich.leisch@tuwien.ac.at.

0167-9473/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2005.10.006

gene expression profiles, which is the average correlation after removing each time point once. Many other distance measures could be useful in applications if their usage does not put too much computational burden like programming effort onto the practitioner.

This paper is organized as follows: in Section 2, we define the notation used throughout the paper, give a general formulation of the K-centroids cluster analysis (KCCA) problem and write several popular cluster algorithms for metric spaces as special cases within this unifying framework. Section 3 shows how one can adapt the framework to arbitrary distance measures using examples for clustering nominal spaces, and especially binary data. Section 4 describes a flexible implementation of the framework that can be extended by users with only minimal programming effort to try out new distance measures. Section 5 introduces a new version of cluster neighborhood graphs which allows for visual assessment of the cluster structure. Finally, Section 6 demonstrates the methods on a real data set from tourism marketing.

2. The K-centroids cluster problem

2.1. Distances, similarities and partitions

Let $\mathcal{X}$ denote a space of deterministic or random variables, μ a (probability) measure on the Borel sets of $\mathcal{X}$, $\mathcal{C}$ a set or space of admissible centroids, and let

$$d(\mathbf{x}, \mathbf{c}): \mathcal{X} \times \mathcal{C} \to \mathbb{R}^+$$

denote a distance measure on $\mathcal{X} \times \mathcal{C}$. A set of K centroids

$$C_K = \{\mathbf{c}_1, \ldots, \mathbf{c}_k\}, \quad \mathbf{c}_k \in \mathcal{C}$$

induces a *partition* $\mathcal{P} = \{\mathcal{X}_1, \ldots, \mathcal{X}_k\}$ of $\mathcal{X}$ into K disjoint clusters by assigning each point to the segment of the closest centroid

$$\mathcal{X}_k = \{\mathbf{x} \in \mathcal{X} | c(\mathbf{x}) = \mathbf{c}_k\},$$
$$c(\mathbf{x}) = \operatorname*{argmin}_{\mathbf{c} \in C_K} d(\mathbf{x}, \mathbf{c}).$$

If we use a similarity measure $s(\mathbf{x}, \mathbf{c}) \in [0, 1]$ instead of a distance measure we get a partition by defining

$$c(\mathbf{x}) = \operatorname*{argmax}_{\mathbf{c} \in C_K} s(\mathbf{x}, \mathbf{c}).$$

Note that a similarity measure can easily be turned into a distance measure (and vice versa), e.g., the transformation

$$d(\mathbf{x}, \mathbf{c}) = \sqrt{1 - s(\mathbf{x}, \mathbf{c})}$$

turns any non-negative definite matrix of similarities into a matrix of so-called Euclidean distances (Everitt et al., 2001). A matrix of pairwise (arbitrary) distances of objects is called Euclidean if the objects can be represented by points in a Euclidean space which have the same pairwise distances as the original objects, which is an appealing property, especially for visualization. For notational simplicity we will use only distances in the following, all algorithms can also be written for similarities (basically all minimization problems are replaced by maximizations).

Definition. The set of centroids C_K is called *canonical* for $(\mathcal{X}, \mathcal{C}, d)$, if

$$\int_{\mathcal{X}_k} d(\mathbf{x}, \mathbf{c}_k) \, \mathrm{d}\mu(\mathbf{x}) \leqslant \int_{\mathcal{X}_k} d(\mathbf{x}, \mathbf{c}) \, \mathrm{d}\mu(\mathbf{x}) \quad \forall \mathbf{c} \in \mathcal{C} \; \forall \mathcal{X}_k \in \mathcal{P},$$

where $\mathcal{P}$ is the partition induced by C_K.