Clustering in the Identification of Space Models

Maribel Yasmina Santos
University of Minho, Portugal

Adriano Moreira
University of Minho, Portugal

Sofia Carneiro
University of Minho, Portugal

INTRODUCTION

Clustering is the process of grouping a set of objects into clusters so that objects within a cluster have high similarity with each other, but are as dissimilar as possible to objects in other clusters. Dissimilarities are measured based on the attribute values describing the objects (Han & Kamber, 2001).

Clustering, as a data mining technique (Cios, Pedrycz, & Swiniarski, 1998; Groth, 2000), has been widely used to find groups of customers with similar behavior or groups of items that are bought together, allowing the identification of the clients' profile (Berry & Linoff, 2000). This article presents another use of clustering, namely in the creation of Space Models.

Space Models represent divisions of the geographic space in which the several geographic regions are grouped accordingly to their similarities with respect to a specific indicator (values of an attribute). Space Models represent natural divisions of the geographic space based on some geo-referenced data.

This article addresses the development of a clustering algorithm for the creation of Space Models – STICH (*Space Models Identification Through Hierarchical Clustering*). The Space Models identified, integrating several clusters, point out particularities of the analyzed data, namely the exhibition of clusters with outliers, regions which behavior is strongly different from the other analyzed regions.

In the work described in this article, some assumptions were adopted to the creation of Space Models through a clustering process, namely:

- Space Models must be created by looking at the data values available and no constraints must be imposed for their identification.
- Space Models can include clusters of different shapes and sizes.
- Space Models are independent of specific domain knowledge, like the specification of the final number of clusters.

The following sections, in outline, include: (i) an overview of clustering, its methods and techniques; (ii) the STICH algorithm, its assumptions, its implementation and the results for a sample dataset; (iii) future trends; and (iv) a conclusion with some comments about the proposed algorithm.

BACKGROUND

Clustering (Grabmeier, 2002; Jain, Murty, & Flynn, 1999; Zaït & Messatfa, 1997) is a discovering process (Fayyad & Stolorz, 1997) that identifies homogeneous groups of segments in a dataset.

Han & Kamber (2001) state that clustering is a challenging field of research integrating a set of special requirements. Some of the typical requirements of clustering in data mining are:

- *Scalability* in order to allow the analysis of large datasets, since clustering on a sample of a database may lead to biased results.
- *Discovery of clusters with arbitrary shapes* since clusters can be of any shape. Some existing clustering algorithms identify clusters that tend to be spherical, with similar size and density.
- *Minimal domain knowledge* since the clustering results can be quite sensitive to the input parameters, like the number of clusters required by many clustering algorithms.
- *Ability to deal with noisy data* avoiding the identification of clusters that were negatively influenced by outliers or erroneous data.
- *Insensitivity to the order of input records* since there are clustering algorithms that are influenced

by the order in which the available records are analyzed.

Two of the well-known types of clustering algorithms are based on partitioning and hierarchical methods. These methods and some of the corresponding algorithms are presented in the following subsections.

Clustering Methods

Partitioning Methods

A partitioning method constructs a partition of n objects into k clusters where $k \leq n$. Given k, the partitioning method creates an initial partitioning and then, using an iterative relocation technique, it attempts to improve the partitioning by moving objects from one cluster to another. The clusters are formed to optimize an objective-partitioning criterion, such as the distance between the objects. A good partitioning must aggregate objects such that objects in the same cluster are similar to each other, whereas objects in different clusters are very different (Han & Kamber, 2001).

Two of the well-known partitioning clustering algorithms are the *k-means* algorithm, where each cluster is represented by the mean value of the objects in the cluster, and the *k-medoid* algorithm, where each cluster is represented by one of the objects located near the centre of the cluster.

Hierarchical Methods

Hierarchical methods perform a hierarchical composition of a given set of data objects. It can be done bottom-up (Agglomerative Hierarchical methods) or top-down (Divisive Hierarchical methods). Agglomerative methods start with each object forming a separate cluster. Then they perform repeated merges of clusters or groups close to one another until all the groups are merged into one cluster or until some pre-defined threshold is reached (Han & Kamber, 2001). These algorithms are based on the inter-object distances and on finding the nearest neighbors objects. Divisive methods start with all the objects in a single cluster and, in each successive iteration, the clusters are divided into smaller clusters until each cluster contains only one object or a termination condition is verified.

The next subsection presents two examples of clustering algorithms associated with partitioning and hierarchical methods respectively.

Clustering Techniques

The K-Means Algorithm

Partitioning-based clustering algorithms such as *k-means* attempt to break data into a set of k clusters (Karypis, Han, & Kumar, 1999).

The *k-means* algorithm takes as input a parameter k that represents the number of clusters in which the n objects of a dataset will be partitioned. The obtained division tries to maximize the *Intracluster* similarity (a measurement of the similarity between the objects inside a cluster) and minimize the *Intercluster* similarity (a measurement of the similarity between different clusters). That is to say a high similarity between the objects inside a cluster and a low similarity between objects in different clusters. This similarity is measured looking at the centers of gravity (centroids) of the clusters, which are calculated as the mean value of the objects inside them.

Given the input parameter k, the *k-means* algorithm works as follows (MacQueen, 1967):

1. Randomly selects k objects, each of which initially represents the cluster centre or the cluster mean.
2. Assign each of the remaining objects to the cluster to which it is the most similar, based on the distance between the object and the cluster mean.
3. Compute the new mean (centroid) of each cluster.

After the first iteration, each cluster is represented by the mean calculated in step 3. This process is repeated until the criterion function converges. The squared-error criterion is often used, which is defined as:

$$E = \sum_{i=1}^{k} \sum_{j=1}^{l} o_j \in C_i \left| o_j - m_i \right|^2 \quad (1)$$

where E is the sum of the square-error for the objects in the dada set, l is the number of objects in a given cluster, o_j represents an object, and m_i is the mean value of the cluster C_i. This criterion intends to make the resulting k clusters as compact and as separate as possible (Han & Kamber, 2001).

The *k-means* algorithm is applied when the mean of a cluster can be obtained, which makes it not suitable for the analysis of categorical attributes[1]. One of the disadvantages that can be pointed out to this method is the necessity for users to specify k in advance. This method is also not suitable for discovering clusters with nonconvex shapes or clusters of very different sizes (Han & Kamber,