Effects of including spatial information in clustering multivariate image data

Patrick W.T. Krooshof, Thanh N. Tran, Geert J. Postma, Willem J. Melssen, Lutgarde M.C. Buydens

Multivariate image data provide detailed information in variable and image space. Most traditional clustering methods are based on variable information only and ignore spatial information. A method based on both variable and spatial information could improve the results substantially.

In this review, we study the benefits and the pitfalls of including spatial information in chemometric clustering techniques. Spatial information is taken into account in initialization of clustering parameters, during cluster iterations by adjusting the similarity measure or at a post-processing step. We illustrate the effect of taking spatial information into account by a univariate synthetic data set and two real-world multivariate data sets. We show that methods that include neighboring pixel information in the clustering procedure improve the performance accuracy of the clustering in most cases. Homogeneous regions in the image are better recognized and the amount of noise is reduced by these methods.

Keywords: Image clustering; Initialization; Pattern recognition; Post-processing; Spatial information

Patrick W.T. Krooshof, Geert J. Postma, Willem J. Melssen, Lutgarde M.C. Buydens*
Institute for Molecules and Materials, Analytical Chemistry, Radboud University Nijmegen, Toernooiveld 1, NL-6525 ED Nijmegen, The Netherlands

Thanh N. Tran
Research and Technology Chemicals Department, Akzo Nobel Chemicals bv, Velperweg 76, NL-6800 SB Arnhem, The Netherlands

*Corresponding author.
Tel.: +31 24 3653180;
Fax: +31 24 3652653;
E-mail:
L.Buydens@science.ru.nl

1. Introduction

Pattern recognition is a popular chemometric technique used in many applications. It aims to divide a set of objects (a series of measurements) into several categories, based on a similarity measure [1–3]. The resulting classes contain objects that are more similar to each other, compared to objects in the other classes. New, unidentified objects can then be assigned to the class that contains objects that are most similar to these new objects. Classification based on the available patterns that have already been classified is called supervised pattern recognition [1,4]. Pattern recognition can also be unsupervised, in which case no predefined categories are available: the classes are obtained by the data itself. Unsupervised pattern recognition is also called clustering [1–3].

Although clustering is used in many sciences for explorative research, we will focus in this article on the area of multivariate image analysis [5,6]. In general, a multivariate image can be considered as a stack of images, which contains multiple variables per image pixel. An example is data obtained from remote-sensing measurements of the Earth's surface. The data comprise multivariate images, each image recorded at a different wavelength, to identify or to study surface materials. Because of the large amount of data, it is difficult and time consuming to examine such images in detail. Clustering is used to automate data processing and to facilitate the analysis of large images [5,6].

Besides the variable information of each pixel, region information is also available in image data [4,7]. As most images are expected to contain homogeneous regions, neighboring pixels have a high probability of being of the same class. Thus, pixels that are spatially close to each other are more likely to be similar, compared to pixels from other regions in the image. The inclusion of this spatial information in the clustering of image data is expected to improve the final clustering result in most cases [7–9]. Spatial information can be included in the similarity measure for clustering data, before clustering in an initialization step or after partitioning of the data in a post-processing step. However, spatial information was, until recently, most often neglected: standard clustering techniques do not take it into account [5].

0165-9936/$ - see front matter © 2006 Elsevier Ltd. All rights reserved. doi:10.1016/j.trac.2006.09.002

Nomenclature

Symbols: Simple indices are explained in the text.

β	Spatial dependency parameter
$d_{euc}(\mathbf{x}_i,\mathbf{x}_j)$	Euclidean distance between object $\mathbf{x}_i$ and $\mathbf{x}_j$
E	Criterion function to determine the optimal partitioning of the data set
$f(\mathbf{x}_i)$	Distribution function: density of objects $\mathbf{x}_i$
K	Total number of (chosen) clusters in the data set
L	Likelihood criterion function
$\boldsymbol{\mu}$	Mean of the data set
$\boldsymbol{\mu}_k$	Mean of cluster k
N	Total number of objects in the data set
P	Total number of variables
γ	Fuzziness parameter
s_i	Neighborhood of object $\mathbf{x}_i$
$\mathbf{C}$	Covariance of the data set
τ_k	Mixture proportion of cluster k
u_{ik}	Probability of object $\mathbf{x}_i$ belonging to cluster k
$w(\mathbf{x}_i,s_i,k)$	Weighted function, dependent on the neighborhood of object $\mathbf{x}_i$
$\mathbf{x}_i$	i-th object (pixel) of the data set
$\mathbf{x}_j$	j-th object (pixel) of the data set

In this article, we discuss the benefits and the pitfalls of using spatial information. Some basic clustering techniques are used for clustering multivariate image data. We discuss and clarify differences between methods by applying them to three data sets [10]:

- a univariate synthetic data set;
- a multispectral image for minced meat; and,
- a remote-sensing Synthetic Aperture Radar (SAR) image, taken over the Flevoland area in The Netherlands.

To show the impact of spatial information on clustering multivariate image data, we use comparable methods that do and do not include spatial information for clustering the three data sets.

2. Clustering techniques

The heart of the clustering machine is formed by a similarity measure between a set of objects. In clustering, it is usual to calculate the (dis)similarity between two objects, $\mathbf{x}_i$ and $\mathbf{x}_j$, using a distance measure [1,2,4,5]. The most popular distance measure is the Euclidean distance:

$$d_{euc}(\mathbf{x}_i,\mathbf{x}_j) = \sqrt{\sum_{l=1}^{P}(\mathbf{x}_{il}-\mathbf{x}_{jl})^2} \tag{1}$$

where $\mathbf{x}_i = \{\mathbf{x}_{i1},\ldots,\mathbf{x}_{iP}\}$, in which P denotes the number of variables.

Another distance measure that is widely used is the Mahalanobis distance [2,4,11], which calculates the distance between an object, $\mathbf{x}_i$, and the centroid (mean) of a group of objects, named a cluster. The major difference with the standard Euclidean distance is that the Mahalanobis distance takes the covariance of a cluster into account (i.e. the size and the shape of the clusters).

Besides the many possible dissimilarity functions, there is an even larger variety of clustering methods [1,2,4,12]. The different algorithms can be categorized into three main types of clustering: partitional [9,13]; hierarchical [14,15]; and, density-based [16,17], but we do not discuss the last of these in this article.

2.1. Partitional clustering

Partitional clustering methods try to partition the data into a certain number of clusters in an optimal way, according to a certain criterion or cost function. For example, in the K-means algorithm [9,13,18], the sum of squares of the within-cluster distances, E, is minimized by iteratively transferring objects between clusters (Equation (2)). The minimum value for the criterion function is obtained if the data is partitioned into well-separated, compact clusters. A compact cluster contains objects, for which the distance to the mean of the cluster, $\boldsymbol{\mu}_k$, is relatively small.

$$E = \sum_{k=1}^{K}\sum_{i\in k} d^2(\mathbf{x}_i,\boldsymbol{\mu}_k) \tag{2}$$

The K-means algorithm starts with K randomly selected, cluster centers. Next, the objects of the data set are assigned to the cluster with the smallest distance measure, $d(\mathbf{x}_i,\boldsymbol{\mu}_k)$, and the cluster centers are updated. This process continues until a stop criterion is met, such as a threshold for criterion E, the number of iterations or stabilization of the clustering result.

K-means results in a ‘‘hard’’ clustering of the data, where each object is assigned to only one cluster. Fuzzy clustering extends this assignment to associate each object to every cluster, using a membership function. The higher the degree of membership for a particular cluster, the more probable it is that the object belongs to that cluster. Fuzzy c-means [13,19,20], which is a variant of K-means, includes such a membership function in the criterion function, E. The fuzzy clustering result can easily be converted to a ‘‘hard’’ clustering by assigning an object to the cluster with the highest degree of membership.