Available online at www.sciencedirect.com
SCIENCE DIRECT
Pattern Recognition 39 (2006) 1315–1324
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/patcog

Scalable visual assessment of cluster tendency for large data sets

Richard J. Hathaway[a], James C. Bezdek[b], Jacalyn M. Huband[b,*]

[a]*Department of Mathematical Sciences, Georgia Southern University, Statesboro, GA 30460, USA*
[b]*Computer Science Department, University of West Florida, Pensacola, FL 32514, USA*

Received 12 August 2005; received in revised form 6 February 2006; accepted 10 February 2006

Abstract

The problem of determining whether clusters are present in a data set (i.e., assessment of cluster tendency) is an important first step in cluster analysis. The visual assessment of cluster tendency (VAT) tool has been successful in determining potential cluster structure of various data sets, but it can be computationally expensive for large data sets. In this article, we present a new scalable, sample-based version of VAT, which is feasible for large data sets. We include analysis and numerical examples that demonstrate the new scalable VAT algorithm.

Keywords: Clustering; Similarity measures; Cluster validity; Data visualization; Scalability

1. Introduction

Recently, extended versions of fuzzy c-means clustering algorithms for very large image [1], general object [2] and relational [3] data sets have been developed. These extended c-means algorithms have two types of useful application: (1) to provide faster clustering results when the large data set is still small enough so that application of a conventional form of c-means clustering is possible; and (2) to simply provide (any) clustering results when the data set is so large that application of a conventional version of c-means is not practical (either because of the time or space required). A requirement to run any of the extended (or conventional) forms of c-means clustering is a good choice for c, the number of clusters. The purpose of this paper is to describe, analyze and demonstrate a visual method for determining the number of clusters that can be applied to very large data sets in a computationally efficient manner. The new method is a sample-based version of the visual assessment of cluster tendency procedure from Ref. [4]. We begin with some necessary background.

Our focus is a type of preliminary data analysis related to the pattern recognition problem of clustering. Clustering or cluster analysis is the problem of partitioning a set of objects $O = \{o_1, \ldots, o_N\}$ into c self-similar subsets based on available data and some well-defined measure of (cluster) similarity. In some cases, a geometric description of the clusters (e.g. by "cluster centers" in data space) is also desired and some clustering methods produce such geometric descriptors. The type of clusters found is strongly related to the properties of the mathematical model that underlies the clustering method. All clustering algorithms will find an arbitrary (up to $1 \leqslant c \leqslant N$) number of clusters, even if no "actual" clusters exist. Therefore, a fundamentally important question to ask before applying any particular (and potentially biasing) clustering algorithm is: Are clusters present at all?

The problem of determining whether clusters are present as a step prior to actual clustering is called the *assessment of clustering tendency*. Various formal (statistically based) and informal techniques for tendency assessment are discussed in Refs. [5,6]. The technique proposed here is visual, and visual approaches for various data analysis problems have been widely studied in the last 30 years; [7,8] are standard sources for many visual techniques. The basis for the new method for large data sets developed in this article is the

* Corresponding author. Tel.: +1 850 474 2304; fax: +1 850 857 6056.
E-mail address: jhuband@cs.uwf.edu (J.M. Huband).

0031-3203/$30.00
doi:10.1016/j.patcog.2006.02.011

visual assessment of tendency (VAT) procedure from Ref. [4]. The VAT approach presents pairwise dissimilarity information about the set of objects $O=\{o_1,\ldots,o_N\}$ as a square digital image with N^2 pixels, after the objects are suitably reordered so that the image is better able to highlight potential cluster structure. To go further into the VAT approach requires some additional background on the types of data typically available to describe the set $O=\{o_1,\ldots,o_N\}$.

There are two common data representations of O upon which clustering can be based. When each object in O is represented by a (column) vector $\mathbf{x}$ in $\mathscr{R}^s$, the set $X=\{\mathbf{x}_1,\ldots,\mathbf{x}_N\}\subset\mathscr{R}^s$ is called an *object data* representation of O. The kth component of the ith feature vector (x_{ki}) is the value of the kth feature (e.g., height, weight, length, etc.) of the ith object. It is in this data space that practitioners sometimes seek geometrical descriptors (often called prototypes) of the clusters. Alternatively, when each *pair* of objects in O is represented by a relationship between them, then we have *relational data*. The most common case of relational data is when we have (a matrix of) dissimilarity data, say $R=[R_{ij}]$, where R_{ij} is the pairwise dissimilarity measure (usually a distance) $d(o_i,o_j)$ between objects o_i and o_j, for $1\leqslant i,\ j\leqslant N$. More generally, R can be a matrix of similarities based on a variety of measures [9,10].

The VAT tool is widely applicable because it displays a reordered form of dissimilarity data, which itself can *always* be obtained from the original data for O. If the original data consists of a matrix of pairwise (symmetric) similarities $S=[S_{ij}]$, then dissimilarities can be obtained through several simple transformations. For example, we can take

$$R_{ij}=S_{\max}-S_{ij}, \tag{1}$$

where $S_{\max}$ denotes the largest similarity value. If the original data set consists of object data $X=\{\mathbf{x}_1,\ldots,\mathbf{x}_N\}\subset\mathscr{R}^s$, then R_{ij} can be computed as $R_{ij}=\|\mathbf{x}_i-\mathbf{x}_j\|$, using any convenient norm on $\mathscr{R}^s$. If the original data has missing components (is incomplete), then any existing data imputation scheme can be used to "fill in" the missing part of the data prior to processing. A discussion of various options for inexpensively handling the missing data is given in Ref. [4]. The main point here is that the dissimilarity data needed to apply VAT is available in virtually *all* numerical data sets.

The original VAT procedure is stated next. We assume that R is symmetric, and has nonnegative off-diagonal entries and zero diagonal entries. In general, the functions, arg max and arg min, in Steps 2 and 3 are set valued, so that the procedure selects any of the optimal arguments. The reordering found by VAT is stored in array $P=(P(1),\ldots,P(N))$.

VAT: Visual Assessment of (Cluster-) Tendency

Input: The user supplies the full $N\times N$ matrix of pairwise dissimilarities R.

Step 1. Set $K=\{1,2,\ldots,N\}$.
Select $(i,j)\in\arg\min_{p\in\mathrm{K},q\in\mathrm{K}}\{R_{pq}\}$.
Set $P(1)=i$; $I=\{i\}$; and $J=K-\{i\}$.

Step 2. For $t=2,\ldots,N$:
Select $(i,j)\in\arg\min_{p\in\mathrm{I},q\in\mathrm{J}}\{R_{pq}\}$.
Set $P(t)=j$; Replace $I\leftarrow I\cup\{j\}$ and $J\leftarrow J-\{j\}$.
Next t.

Step 3. Form the ordered dissimilarity matrix $\tilde{R}=[\tilde{R}_{ij}]=[R_{P(i)P(j)}]$, for $1\leqslant i,\ j\leqslant N$.

Step 4. Display $\tilde{R}$ as an intensity image, scaled so that $\max\{\tilde{R}_{ij}\}$ corresponds to white and 0 corresponds to black.

The VAT ordering algorithm can be implemented in $\mathscr{O}(N^2)$ time complexity and is similar to Prim's algorithm for finding a minimal spanning tree (MST) of a weighted graph (see, for example [11] for a description of Prim's algorithm). The main differences between VAT and Prim's algorithm are that: (i) we are not interested in representing the MST, but only in finding the order in which the vertices are added as it is grown; and (ii), we specify a method for choosing the initial vertex that depends on the maximum edge weight in the underlying complete graph. (This choice of initial vertex gives nicer images, by avoiding a phenomenon known as "zigzagging", which is discussed in Ref. [4].) The permuted indices of the N objects are stored in the array P. Note that distances in $\tilde{R}$ are *not* recomputed; instead, we simply rearrange the rows (and columns) of R to construct $\tilde{R}$.

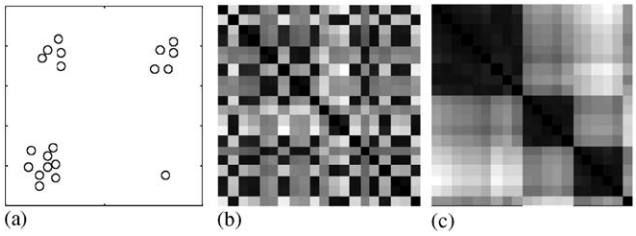

Fig. 1. (a) Object data. (b) Image for original R. (c) Image for VAT-ordered $\tilde{R}$.

From Ref. [4], we repeat a small example in Fig. 1 to show the reader how well-separated cluster structure is indicated as dark diagonal blocks in the intensity image display of the VAT-ordered $\tilde{R}$. Fig. 1(a) gives a scatter plot of a small data set in $\mathscr{R}^2$. A display of the relational data matrix $R=[r_{ij}]=[\|\mathbf{x}_i-\mathbf{x}_j\|]$ using the Euclidean norm to convert X to R in Fig. 1(b) does not indicate the structure of the data set. After the relational matrix R is reordered by VAT and displayed as $\tilde{R}$ in Fig. 1(c), the structure is apparent. We see $c=4$ clusters in view 1(c), indicated by the four dark blocks along the main diagonal. Moreover, the size of each block corresponds directly to the number of points in each cluster. Notice the singleton! Certainly, VAT is not needed when a scatter plot such as Fig. 1(a) is possible, but we use this simple example of object data in $\mathscr{R}^2$ to help the reader correlate (visual) spatial clusters with VAT images.