A Brief Overview of Relational Clustering Methods

Olfa Nasraoui
Department of Computer Engineering & Computer Science
University of Louisville,
olfa.nasraoui_AT_louisville.edu

Two types of data are used in pattern recognition, object and relational data. Object data is the most common type of data and is in the form of the usual data set of feature vectors. Relational data is less common than object data and consists of the pairwise relations (similarities or dissimilarities) between each pair of implicit objects. Such a relation is usually stored in a relation matrix and no other knowledge is available about the objects being clustered. Because relational data is less common than object data, relational pattern recognition methods are not as well developed as their object counterparts, particularly in the area of robust clustering. However, relational methods are becoming a necessity as relational data becomes more and more common. For instance, information retrieval and data mining are all applications which could greatly benefit from pattern recognition methods that can deal with relational data.

There are two types of clustering algorithms for relational data. Hierarchical algorithms include local or graph-theoretic methods, while partitional algorithms are global objective function driven. We focus on partitional methods because they are the most commonly used methods for object data, and they have a lower computational complexity. Most objective-function based relational clustering methods assume that the relation matrix, $\mathbf{R}$, is of the dissimilarity type. The earliest models have been proposed by Ruspini [Rus70], Roubens [Rou78], Diday [Did75], and Windham [Win85]. Hathaway and Bezdek [HDB89] reformulated the FCM objective function to be able to work on relational data by eliminating the prototypes from the FCM objective function. The Relational Fuzzy C Means (RFCM) has the following objective

$$\min_{\mathbf{U}} \quad \sum_{i=1}^{C} \frac{\sum_{j=1}^{N}\sum_{k=1}^{N} u_{ij}^{m} u_{ik}^{m} r_{jk}}{2\sum_{t=1}^{N} u_{it}^{m}}. \tag{1}$$

where

$$r_{jk} = \|\mathbf{x}_j - \mathbf{x}_k\|^2. \tag{2}$$

They showed that minimization of the FCM and RFCM objective functions are equivalent provided $\mathbf{R}$ satisfies (2), i. e., there exists a set of N object data in $\mathcal{R}^p$ such that the pairwise distances define $\mathbf{R}$, for some integer $p < N-1$. In this case, RFCM can be considered as the relational dual of FCM. In order to derive the necessary update equations for the RFCM, Hathaway and Bezdek [HDB89] proved that the squared Euclidean distance, $d_{ik}^2 = \|\mathbf{x}_j - \mathbf{c}_i\|^2$, from feature vector $\mathbf{x}_j$ to the center of the i^{th} cluster, $\mathbf{c}_i$, can be written in

terms of the relation matrix $\mathbf{R}$ as follows

$$d_{ik}^2 = (\mathbf{R}\mathbf{v}_i)_k - \mathbf{v}_i^t \mathbf{R}\mathbf{v}_i/2. \tag{3}$$

where $\mathbf{v}_i$ is the membership vector defined by

$$\mathbf{v}_i = \frac{(u_{i1}^m, \ldots, u_{iN}^m)^t}{\sum_{j=1}^N u_{ij}^m}. \tag{4}$$

Equation (3) allows the computation of the distances between the data points and cluster prototypes in each iteration when only the relational data, $\mathbf{R}$, are given. Therefore, a relational dual of FCM exists for the special case where the object data and relational data satisfy (2). This means that even when only relational data is available in the form of an $N \times N$ relation matrix, the relational dual of FCM is expected to perform in an equivalent way to FCM provided that the relation matrix, $\mathbf{R}$, is Euclidean, i.e., there exists a set of N points in $\mathcal{R}^{N-1}$, called a realization of $\mathbf{R}$, satisfying (2).

When a realization does not exist for the relation matrix, $\mathbf{R}$, the relational dual of the FCM may fail mainly because some of the distances computed using (3) may be negative. To overcome this problem, we can use the β-spread transform [HB94] to convert a non-Euclidean matrix $\mathbf{R}$ into an Euclidean Matrix $\mathbf{R}_\beta$ as follows

$$\mathbf{R}_\beta = \mathbf{R} + \beta(\mathbf{M} - \mathbf{I}) \tag{5}$$

where β is a suitably chosen scalar, $\mathbf{I} \in \mathcal{R}^{N \times N}$ is the identity matrix and $\mathbf{M} \in \mathcal{R}^{N \times N}$ satisfies $M_{jj} = 1$ for $1 \leq j \leq N$. It was suggested in [HB94] that the distances d_{ik}^2 be checked in every iteration for negativity, which indicates a non-Euclidean relation matrix. In that case, the β-spread transform should be applied with a suitable value of β to make the d_{ik}^2 positive again. An underestimate for the lower bound on β was derived [HB94] and related to the necessary shift, $\Delta\beta$, that is needed to make the distances positive. This result can be summarized as

$$\Delta\beta = \max_{i,k}\{-2d_{ik}^2 / \|\mathbf{v}_i - \mathbf{e}_k\|^2\}, \tag{6}$$

where $\mathbf{e}_k$ denotes the k^{th} column of the identity matrix.

Acknowledgment

This work is supported by the National Science Foundation (CAREER Award IIS-0133948 to O. Nasraoui). Partial support of earlier stages of this work by the Office of Naval Research grant (N000014-96-1-0439) and by the National Science Foundation Grant IIS 9800899 is also gratefully acknowledged.

Bibliography

Did75
E. Diday.
Classification automatique sequentielle pour grands tableaux.
Revue Francaise d'Automatique Informatique et Recherche Operationelle, pages 29-61, 1975.

HB94
R. J. Hathaway and J. C. Bezdek.
Nerf c-means: Non-euclidean relational fuzzy clustering.
Pattern Recognition, 27(3):429-437, 1994.

HDB89
R. J. Hathaway, J. W. Davenport, and J. C. Bezdek.
Relational duals of the c-means algorithms.
Pattern Recognition, 22:205-212, 1989.

Rou78
M. Roubens.
Pattern classification problems and fuzzy sets.
Fuzzy Sets and Systems, 1:239-253, 1978.

Rus70
E. Ruspini.
Numerical methods for fuzzy clustering.
Information Science, 12:319-350, 1970.

Win85
M. P. Windham.
Numerical classification of proximity data with assignment measures.
J. Classification, 2:157-172, 1985.

About this document ...

A Brief Overview of Relational Clustering Methods

This document was generated using the **LaTeX**2HTML translator Version 2002 (1.62)

The command line arguments were:

latex2html `-split 0 -image_type gif RelationalClustering.tex`
