Generating multidimensional embeddings based on fuzzy memberships

ID - 101

Rovetta Stefano[1], Francesco Masulli[1], Maurizio Filippone[1]
[1]Department of Computer and Information Sciences, University of Genova

Motivation

Exploratory analysis of genomic data sets using unsupervised clustering techniques is often affected by problems due to the small cardinality and high dimensionality of the data set. These problems may be eased by performing clustering in an embedding space. This brings about the problem of selecting an appropriate transformation to perform the required multidimensional embedding, which should be able to keep the necessary information while reducing data dimensionality.

If the cardinality of the data set is small compared to the input space dimensionality, then the matrix of mutual distances or other pairwise pattern evaluation methods such as kernels may be used to represent data sets in a more compact way. Following this approach, the data matrix is replaced by a pairwise dissimilarity matrix D.

Methods

We have proposed an embedding technique based on the concept of fuzzy membership, which is related but different from dissimilarity-based representations. A set of vectors are selected from the data set. These are termed probes and are used as reference points for the rest of the data set. Probes are interpreted as fuzzy points; for each of the remaining points in the data set, the fuzzy membership to a probe can be evaluated. Therefore, for each point an ordered set of membership values is defined, one for each probe, and this ordered set can be used as a new feature vector to represent the point itself, embededd in a space induced by the probes. We will call this representation space the Membership Embedding Space (MES).

We may observe that a point in the embedding space will be represented by a vector containing only few non-null components (depending on the support of the membership function), in correspondence of the closest probes in the original feature space. In our experiments, the memberships of fuzzy sets centered on the probes were modeled as Gaussians normalized over all probes. Here we propose a generative technique based on Simulated Annealing to select sets of probes of small cardinality. An appropriate generalized energy is defined to represent clustering quality and clustering complexity for the probes.

Results

When applied to clustering, the approach has been demonstrated to lead to significant improvements with respect to the application of clustering algorithms in the original space and in the distance embedding space. We present results based on standard data commonly available on line, to make them readily comparable with other approaches. These results indicate that the method supports high quality clustering solutions using compact sets of probes.

Availability: http://mlsc.disi.unige.it/

Email: ste@disi.unige.it