BITS 2005 meeting

Randomized maps to estimate cluster reliability using high dimensional genomic data

A. Bertoni, R. Folgieri, F. Ruffino, G. Valentini
DSI, Dipartimento di Scienze dell' Informazione, Università degli Studi di Milano

Abstract

Motivation:
Discovering new subclasses of pathologies and expression signatures related to specific phenotypes are challenging problems in the context of gene expression data analysis.
To pursue these objectives, we need to estimate the natural number and the stability of the discovered clusters.
To this end, new approaches based on random subspaces and bootstrap methods have been recently proposed.

Methods:
We present a method based on randomized embedding between euclidean subspaces to assess the stability of clusters characterized by low cardinality and very high dimensionality.
In particular we propose a cluster stability measure based on similarity between randomly projected data obeying the Johnson Lindenstrauss lemma, in order to control the distortion induced by randomized maps.
As a by-product of our approach we may also assess the stability of the overall clustering (thus estimating the number of "natural clusters" in a data set), and the confidence of the assignments of each example to each cluster.
The proposed approach may be applied to any clustering algorithm, comprising classical hierarchical and fuzzy clustering.

Results:
At first we evaluated the distortion induced by the random mappings from very high to lower dimensional euclidean spaces using high dimensional synthetic data, showing that we may obtain distortions lower than that predicted by the Johnson Lindenstrauss lemma.
Then we applied the proposed stability indices, based on embeddings into lower dimensional spaces with limited distortion, to both synthetic and gene expression data,.
In particular we computed the s-index (stability index) specific for each cluster, the overall validity index S that estimates the reliability of the overall clustering, and the AC (Assignment-Confidence) index that estimates the reliability of the membership of a specific example to a specific cluster.
Results with synthetic and gene expression data clustered with classical hierarchical clustering algorithms show the effectiveness of the proposed approach.