

Model-Based Linear Manifold Clustering

Abstract

A new paradigm of clustering called “Linear Manifold Clustering” which is based on linear manifolds is designed, analyzed, and evaluated throughout this thesis. A Linear manifold is a translated subspace. Linear manifold clustering seeks to identify groups of points that are embedded in lower dimensional linear manifolds. The “birth” of this paradigm of clustering is a consequence of what we believe is a need for an important yet overlooked cluster model, and as a result of what we identify as an acute need to sufficiently address certain clustering requirements that current state of art methods are unable to address.

In many problem domains it is assumed that linear models are sufficient enough to describe and capture the data’s inherent structure. Yet very few remote attempts have been made to devise clustering methods able to identify or learn mixtures of linear manifolds. None of these attempts posed the problem in a model-based statistical setting intended to model and understand the underlying “process” responsible for generating sets of points that lie on lower dimensional linear manifolds. In this thesis we introduce a formal stochastic *linear manifold cluster model*. Based on this model we present a series of results and techniques demonstrating the applicability of the linear manifold clustering paradigm to a wide range of applications. We show that this model is a generalization of other more common and somewhat limited cluster models. This generalization allows for less assumptions to be imposed on the data, which typically yield biased results, and more freedom for the data to “speak for itself”. An emphasis is put on the paradigms of *pattern* and *correlation clustering* and on the application of DNA microarray analysis, where we show that pattern clusters or correlations manifest themselves as linear manifolds in the data space. Based on the linear manifold cluster model we present two clustering algorithms: one for clustering or learning mixtures of linear manifolds, and the other tailored to the application of correlation clustering and DNA micro array analysis. The efficacy of these techniques is demonstrated by a series of experiments on synthetic and real data sets.

Most clustering methods focus only on the grouping aspects of clustering and lack the ability to provide any scientific content. In this thesis, we also present two linear manifold based modeling techniques that deliver scientific content and with which data can be described. One based on a *probabilistic density estimation* model with which statistical inference such as predictions can be based upon. The other, a model which describes the linear dependencies in the data in the form of a set of linear equations.