Sparse Nonparametric Feature Extraction for High-Dimensional Data Classification

黃孝雲
輔仁大學統計資訊學系

Abstract

High-dimensional data sets, such as hyperspectral image, financial tick-by-tick data, and microarrays, are getting more and more accessible due to the progress of the new technologies. The classification task in high-dimensional data is of great interesting. For example, how to classify the observed pixels to the known different landcover types is one of the main purposes for collecting hyperspectral images.

For classification purpose, high-dimensional data might offer more information for each individual observation so that the classification performance is expecting to be better. However, the fact is that we might not have enough observations (frequently, the number of observations is much less than the number of dimensionality) in our classification task, so that the blessing of the dimensionality turns to the curse of the dimensionality. One common solution of this problem is to reduce the dimensionality via the Fisher's linear discriminant analysis (LDA).

LDA is also called the discriminant analysis feature extraction (DAFE) is a rather robustness and theoretically sound dimension reduction method. But, it has four major drawbacks. One is it works well only if the distributions of classes are normal-like distributions. The second drawback is that only min(c-1, d) features can be extracted, where c is the number of classes and d is the dimension of the observation. That is, when d is much greater than c, only c-1 features can be extracted. The c-1 features are suboptimal in a Bayes sense, although they are optimal based on the specified criterion. The third drawback is that if the within-class scatter matrix is singular, which is often occurs in high-dimensional problems, the performance will be poor. The forth is that the extracted features are lack of interpretation, that is, lack of sparseness.

In this talk, a new feature extraction named sparse nonparametric feature extraction (SNFE) is introduced. The SNFE is proposed to avoid the mentioned drawbacks of the LDA and thus with better performance for the high-dimensional classification problems. Some evaluations with real-world data sets, including hyperspectral images and microarray data sets, are performed and the results will be demonstrated in this talk as well.