F

Feature Extraction/Selection in High-Dimensional Spectral Data

Seoung Bum Kim
The University of Texas at Arlington, USA

INTRODUCTION

Development of advanced sensing technology has multiplied the volume of spectral data, which is one of the most common types of data encountered in many research fields that require advanced mathematical methods with highly efficient computation. Examples of the fields in which spectral data abound include near-infrared, mass spectroscopy, magnetic resonance imaging, and nuclear magnetic resonance spectroscopy.

The introduction of a variety of spectroscopic techniques makes it possible to investigate changes in composition in a spectrum and to quantify them without complex preparation of samples. However, a major limitation in the analysis of spectral data lies in the complexity of the signals generated by the presence of a large number of correlated features. Figure 1 displays a high-level diagram of the overall process of modeling and analyzing spectral data.

The collected spectra should be first preprocessed to ensure high quality data. Preprocessing steps generally include denoising, baseline correction, alignment, and normalization. Feature extraction/selection identifies the important features for prediction, and relevant models are constructed through the learning processes. The feedback path from the results of the validation step enables control and optimization of all previous steps. Explanatory analysis and visualization can provide initial guidelines that make the subsequent steps more efficient.

This chapter focuses on the feature extraction/selection step in the modeling and analysis of spectral data. Particularly, throughout the chapter, the properties of feature extraction/selection procedures are demonstrated with spectral data from high-resolution nuclear magnetic resonance spectroscopy, one of the widely used techniques for studying metabolomics.

BACKGROUND

Metabolomics is global analysis for the detection and recognition of metabolic changes in biological systems in response to pathophysiological stimuli and to the intake of toxins or nutrition (Nicholson et al., 2002). A variety of techniques, including electrophoresis, chromatography, mass spectroscopy, and nuclear magnetic resonance, are available for studying metabolomics. Among these techniques, proton nuclear magnetic resonance (^{1}H-NMR) has the advantages of high-resolution, minimal cost, and little sample preparation (Dunn & Ellis, 2005). Moreover, the tech-

Figure 1. Overall process for the modeling and analysis of high-dimensional spectra data

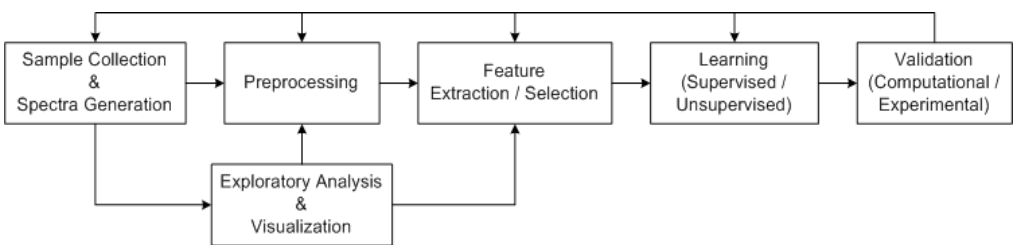
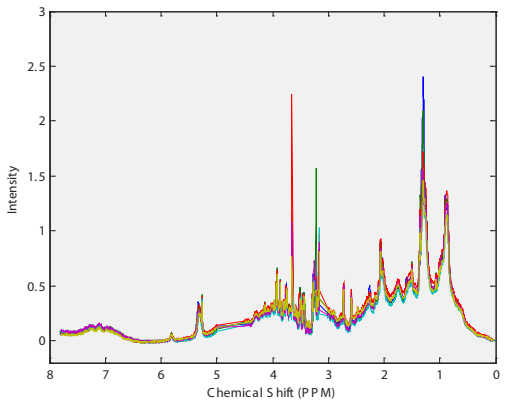


Figure 2. Multiple spectra generated by a 600MHz ^{1}H-NMR spectroscopy



nique generates high-throughput data, which permits simultaneous investigation of hundreds of metabolite features. Figure 2 shows a set of spectra generated by a 600MHz ^{1}H-NMR spectroscopy. The *x*-axis indicates the chemical shift within units in parts per million (ppm), and the *y*-axis indicates the intensity values corresponding to each chemical shift. Traditionally, chemical shifts in the *x*-axis are listed from largest to smallest. Analysis of high-resolution NMR spectra usually involves combinations of multiple samples, each with tens of thousands of correlated metabolite features with different scales.

This leads to a huge number of data points and a situation that challenges analytical and computational capabilities. A variety of multivariate statistical methods have been introduced to reduce the complexity of metabolic spectra and thus help identify meaningful patterns in high-resolution NMR spectra (Holmes & Antti, 2002). Principal components analysis (PCA) and clustering analysis are examples of unsupervised methods that have been widely used to facilitate the extraction of implicit patterns and elicit the natural groupings of the spectral dataset without prior information about the sample class (e.g., Beckonert et al., 2003). Supervised methods have been applied to classify metabolic profiles according to their various conditions (e.g., Holmes et al., 2001). The widely used supervised methods in metabolomics include Partial Least Squares (PLS) methods, *k*-nearest neighbors, and neural networks (Lindon et al., 2001).

Although supervised and unsupervised methods have been successfully used for descriptive and predictive analyses in metabolomics, relatively few attempts have been made to identify the metabolite features that play an important role in discriminating between spectra among experimental conditions. Identifying important features in NMR spectra is challenging and poses the following problems that restrict the applicability of conventional methods. First, the number of features present usually greatly exceeds the number of samples (i.e., tall and wide data), which leads to ill-posed problems. Second, the features in a spectrum are correlated with each other, while many conventional multivariate statistical approaches assume the features are independent. Third, spectra comprise a number of local bumps and peaks with different scales.

MAIN FOCUS OF CHAPTER

Feature Extraction/Selection Process

It is important to distinguish between feature extraction and feature selection, although much of the literature

Figure 3. Overview of feature extraction, feature selection, and a combination of feature extraction and selection