INDEPENDENT COMPONENT ANALYSIS FOR DATA MINING

By

Guangyin Zeng

An Abstract of a Thesis Submitted to the Graduate

Faculty of Rensselaer Polytechnic Institute

in Partial Fulfillment of the

Requirements for the Degree of

DOCTOR OF PHILOSOPHY

Major Subject: Decision Sciences & Engineering Systems

The original of the complete thesis is on file
in the Rensselaer Polytechnic Institute Library

Examining Committee:

Mark J. Embrechts, Thesis Adviser

Thomas R. Willemain, Member

Wai Kin Chan, Member

Curt M. Breneman, Member

Rensselaer Polytechnic Institute
Troy, New York

November 2007
(For Graduation December 2007)

ABSTRACT

Data mining is the science of *extracting novel and potentially useful information from large data sets* [68]. An important problem in data mining is to represent multivariate data and extract useful features from the data. For simplicity, the observed data is usually assumed to be a linear mixture of some latent variables. Well-known linear transformation methods include Principal Component Analysis (PCA), Partial Least Squares (PLS), Canonical Correlation Analysis (CCA), and Factor Analysis (FA). These methods find uncorrelated components from the data. Independent Component Analysis (ICA) is a recently developed method for finding a linear transformation in which the extracted components are mutually independent, which is a stronger condition than uncorrelated. This thesis focuses on algorithms and applications of ICA for data mining.

Several different ICA algorithms have been published to solve the Blind Source Separation (BSS) problem based on different principles, which measure statistical independence in different ways. In this research, an extended ICA algorithm is proposed within the framework of information-theoretic learning. The performance of this algorithm and existing ICA algorithms is studied. The close connection between PCA and ICA is investigated within the information-theoretic framework learning as well. It is shown that ICA becomes PCA when the higher-order terms in the Gram-Charlier series are dropped. ICA is introduced as a general data preprocessing tool. And a pseudo-inverse ICA filtering algorithm is introduced for data cleansing. The ICA filtering algorithm is able to filter out any components including noise from the data. This helps improve the quality of the data, leading to improved results when applying data mining techniques to the cleaned data. In addition to being used in the derivation of the ICA algorithms, the information theory is also introduced for outlier detection problems.

ICA is applied to three applications: cleaning the terahertz spectra of explosives, denoising the terahertz images of explosives, and separating the terahertz spectrum of water vapor.