Chapter 9

Improving Sparsity in Kernelized Nonlinear Feature Extraction Algorithms by Polynomial Kernel Higher Order Neural Networks

Zhao Lu
Tuskegee University, USA

Gangbing Song
University of Houston, USA

Leang-San Shieh
University of Houston, USA

ABSTRACT

As a general framework to represent data, the kernel method can be used if the interactions between elements of the domain occur only through inner product. As a major stride towards the nonlinear feature extraction and dimension reduction, two important kernel-based feature extraction algorithms, kernel principal component analysis and kernel Fisher discriminant, have been proposed. They are both used to create a projection of multivariate data onto a space of lower dimensionality, while attempting to preserve as much of the structural nature of the data as possible. However, both methods suffer from the complete loss of sparsity and redundancy in the nonlinear feature representation. In an attempt to mitigate these drawbacks, this article focuses on the application of the newly developed polynomial kernel higher order neural networks in improving the sparsity and thereby obtaining a succinct representation for kernel-based nonlinear feature extraction algorithms. Particularly, the learning algorithm is based on linear programming support vector regression, which outperforms the conventional quadratic programming support vector regression in model sparsity and computational efficiency.

DOI: 10.4018/978-1-61520-711-4.ch009
1. INTRODUCTION

For the purpose of performing pattern recognition in high-dimensional spaces, the pre-processing procedures of mapping the data into a space of lower dimensionality is usually necessary and vital to achieve higher recognition accuracy. In general, a reduction in the dimensionality of the input space will be accompanied by a loss of some of the information which discriminates between different classes. Hence, the objective in dimensionality reduction is to preserve as much of the relevant information as possible. The well-known Principal component analysis (PCA) and Fisher discriminant analysis (FDA) are two important methods in this direction, and they have been used widely in the field of pattern recognition for their great practical significance.

By calculating the eigenvectors of the covariance matrix of the original inputs, principal component analysis (PCA) linearly transforms a high-dimensional input vector into a low-dimensional one whose components are uncorrelated (Diamantaras and Kung, 1996). The new coordinate values by which we represent the data are called principal components. It is often the case that a small number of principal components are sufficient to account for most of the structure in the data. PCA has been successfully used in the realms of face recognition (Kirby and Sirovich, 1990), radar target recognition (Lee and Ou, 2007) and faulty diagnosis (Yoon and MacGregor, 2004), and so on. Essentially PCA is an orthogonal linear transformation into a lower-dimensional coordinate system that preserves maximum variance in the data, thus minimizing mean-square error, computed as a subset of the Karhunen-Loève rotation.

Although PCA is useful in obtaining a lower dimensional representation of the data, as an unsupervised feature extraction algorithm, the directions that are discarded by PCA might be exactly the directions that are needed for distinguishing between classes. Contrary to the unsupervised PCA for feature extraction, by taking the label information of the data into account, Fisher discriminant analysis is a technique to find a direction \( w \) that separate the class means well (when projected onto the found direction) while achieving a small variance around these means. The quantity measuring the difference between the means is called \textit{between class variance} and the quantity measuring the variance around these class means is called \textit{within class variance}, respectively. Hence, the objective is to find a direction that maximizes the between class variance while minimizing the within class variance at the same time. The Fisher discriminant analysis also bears strong connections to least squares and Bayes classification.

On the other hand, PCA and FDA both seek to find a linear transformation to represent the data. The features extracted by them are therefore limited to linear combinations of input features, which results in their incapability in capturing more complex nonlinear-correlations and overestimate of the true dimensionality of the data (Bishop, 2006); for instance, PCA’s linear model is suboptimal for image data (Mekuz, Bauckhage & Tsotsos, 2009). In an attempt to surmount these restrictions, the kernel principal component analysis (KPCA) and kernel Fisher discriminant (KFD) were proposed as the nonlinear generalizations of PCA and FDA (Schölkopf, Burge & Smola, 1998; Mika et al, 1999). In KPCA, the original inputs were nonlinearly mapped into a high-dimensional feature space where the linear PCA was calculated, which is equivalent to nonlinear PCA in the original input space. Incorporating the kernel method into the conventional PCA, KPCA is very powerful in extracting nonlinear features by taking into account the high order information of the original inputs implicitly, and eventually leading to the excellent generalization performance for learning (Rosipal et al., 2001; Cao et al., 2003). Similar to that in KPCA, the idea of the KFD is to solve the problem of Fisher linear discriminant in a kernel feature space, thereby yielding a nonlinear discriminant in the input space. Particularly, for classification problems, a crucial advantage of the Fisher discriminant algorithm over standard support vector learning...