

A data mining-based subset selection for enhanced discrimination using iterative elimination of redundancy

Hyun-Woo Cho *

Department of Industrial and Information Engineering, University of Tennessee, Knoxville, TN 37996, USA

Abstract

The presence of redundant or irrelevant features in data mining may result in a mask of underlying patterns. Thus one often reduces the number of features by applying a feature selection technique. The objective of feature selection is to get a feature subset that has the best performance. This work proposes a new feature selection method using orthogonal filtering and nonlinear representation of data for an enhanced discrimination performance. An orthogonal filtering is implemented to remove unwanted variation of data. The proposed method adopts kernel principal component analysis, one of nonlinear kernel methods, to extract nonlinear characteristics of data and to reduce the dimensionality of data. The proposed feature selection method is based on the selection criterion of linear discriminant analysis in an environment of iterative backward feature elimination. The performance of the proposed method is compared with those of three different methods. The results showed that it outperforms the three methods. The use of filtering and a kernel method was shown to be a promising tool for an efficient feature selection.

© 2007 Elsevier Ltd. All rights reserved.

Keywords: Variable selection; Support vector machine; Backward elimination; Kernel principal component analysis; Discriminant analysis

1. Introduction

A common problem in data analysis occurs when a large number of features or variables hinder our investigating some patterns present in data. They often are not all equally informative, and many possible features are considered in analyzing data because the relevant features are unknown a priori. The presence of redundant or irrelevant features may result in a mask of underlying patterns of data. Thus one often tries to reduce the number of features to be considered by applying some feature selection schemes (Akbarian & Bishnoi, 2001; Guyon and Elisseeff, 2003).

The objective of feature or variable selection is to get a feature subset that has the most discriminative power in a set of features available. Knowledge about the most important features can help analyze and interpret data of interest.

The feature selection has many potential benefits such as data visualization and interpretation, dimension and storage reduction, and training time and performance improvement (Cortes & Vapnik, 1995). Fig. 1 shows the three-dimensional illustration as an example of feature selection. The use of only two features (right) may significantly improve the classification performance whilst three classes of data are not discriminated well in original three features.

Nonlinear kernel-based methods are fast becoming standard tools for solving various problems. Successful applications of these methods have been reported in various fields such as classification, regression, unsupervised learning, pattern recognition, prediction, etc. (Bach & Jordan, 2002; Müller, Mika, Rätsch, Tsuda, & Schölkopf, 2001; Schölkopf, Smola, & Müller, 1998; Schölkopf et al., 1999). The choice of linear or nonlinear techniques in data analysis are determined by the characteristics of data. The data with linear patterns can be analyzed using linear techniques, in which both linear and nonlinear techniques are expected to have a good performance (Norvilas, Negiz,

* Tel.: +1 865 974 7655; fax: +1 865 974 0588.

E-mail address: hwcho@postech.ac.kr

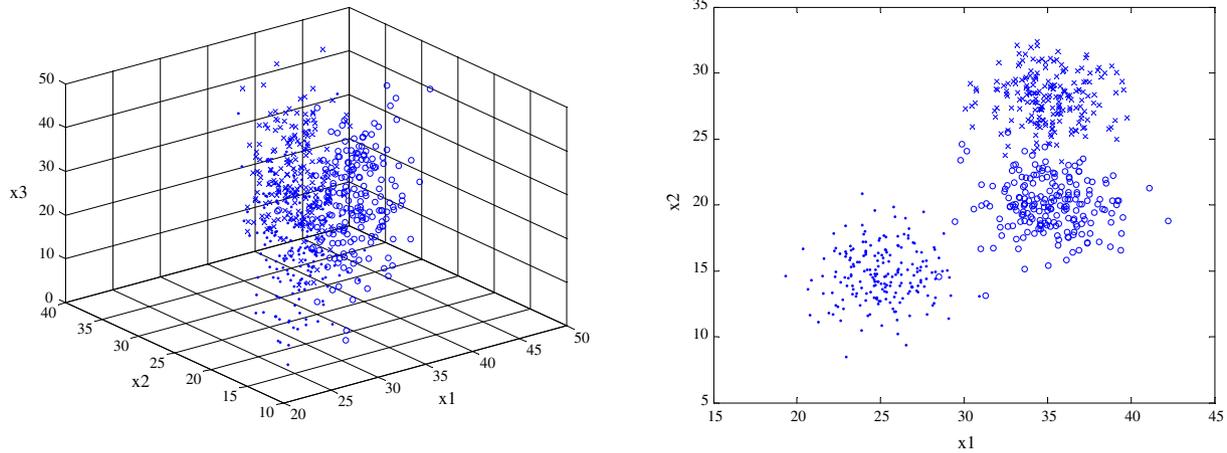


Fig. 1. Plots of simple data for classification of three groups.

Decicco, & Cinar, 2000; Qian, Li, Jiang, & Wen, 2004). However, using linear techniques for nonlinear data will deteriorate analysis results.

In this respect, many powerful kernel methods have been developed and used successfully in many application areas: support vector machine (SVM), kernel principal component analysis (KPCA), kernel Fisher discriminant analysis (KFDA), kernel partial least squares (KPLS), and kernel independent component analysis (KICA) (Bach & Jordan, 2002; Baudat & Anouar, 2000; Cortes & Vapnik, 1995; Qin, 2003; Rosipal & Trejo, 2001; Schölkopf et al., 1998). The basic idea of kernel methods is that input data are mapped into a kernel feature space by a nonlinear mapping function and then these mapped data are analyzed. Several researchers developed the feature selection methods based on support vector machines (SVM) (Guyon, Weston, Barnhill, & Vapnik, 2002; Mao, 2004; Rakotomamonjy, 2003; Weston, Elisseeff, Schölkopf, & Tipping, 2003). Guyon et al. (2002) proposed the SVM-recursive feature elimination (RFE) for selection of genes in micro-array data. The objective of SVM-RFE is to identify a subset of predetermined size of all features available for inclusion in the support vector classifier. Weston et al. (2001) developed a feature selection method that utilizes leave-one-out error rate through gradient descent methods. Rakotomamonjy (2003) also proposed various selection criteria (e.g., generalization error bound) in a SVM-RFE context.

This work proposes a new feature selection method for a classification. It employs orthogonal filtering and nonlinear representation of data as pre-treatment steps for an enhanced discrimination performance. An orthogonal filtering is needed to remove unwanted variation of data. The proposed method focuses on KPCA to extract nonlinear characteristics of data and to decrease high-dimensionality of data. The proposed feature selection scheme is executed with the selection criterion of linear discriminant analysis in an environment of iterative backward feature elimination. This paper is organized as follows. First, a review of SVM and kernel version of principal component

analysis is given. Then the proposed method is presented, which is followed by a case study to demonstrate the proposed method. Finally, concluding remarks are given.

2. Kernel methods: SVM and KPCA

The goal of SVM is to determine hyperplane by minimizing generalization error, which corresponds to maximizing the margin between the separating hyperplane and data (Vapnik, 1995). In SVM input data are first mapped into high-dimensional feature space where optimal decision function can be obtained. As shown in Fig. 2, an optimal separating hyperplane is found which maximizes the margin. It should be noted that we don't need actual computation in high-dimensional space because of kernel trick.

This decision function satisfies inequality constraints

$$y_i(\mathbf{w}\Phi(\mathbf{x}_i) + b) - 1 \geq 0 \forall i. \tag{1}$$

The optimal decision function is obtained by minimizing $1/2\|\mathbf{w}\|^2$ with constraints (1). Non-separable problems are solved by introducing ξ_i and Lagrangian

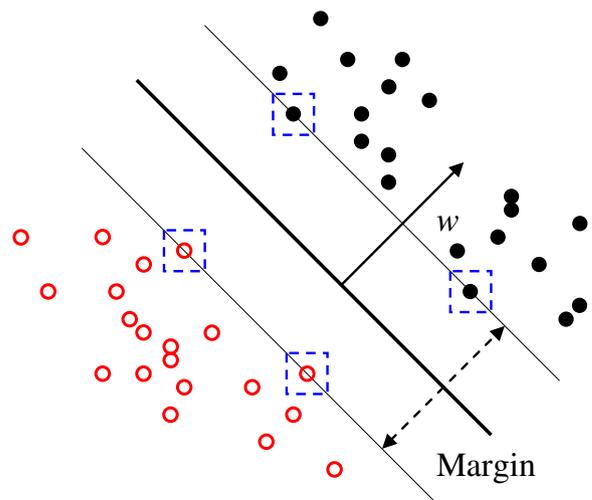


Fig. 2. An illustration of SVM for two-class classification problem.

متن کامل مقاله

دریافت فوری ←

ISIArticles

مرجع مقالات تخصصی ایران

- ✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
- ✓ امکان دانلود نسخه ترجمه شده مقالات
- ✓ پذیرش سفارش ترجمه تخصصی
- ✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
- ✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
- ✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
- ✓ دانلود فوری مقاله پس از پرداخت آنلاین
- ✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات