Feature Reduction for Support Vector Machines

Shouxian Cheng
Planet Associates, Inc., USA

Frank Y. Shih
New Jersey Institute of Technology, USA

INTRODUCTION

The *Support Vector Machine* (SVM) (Cortes and Vapnik,1995; Vapnik, 1995; Burges, 1998) is intended to generate an optimal separating hyperplane by minimizing the generalization error without the assumption of class probabilities such as Bayesian classifier. The decision hyperplane of SVM is determined by the most informative data instances, called *Support Vectors* (SVs). In practice, these SVMs are a subset of the entire training data. By now, SVMs have been successfully applied in many applications, such as face detection, handwritten digit recognition, text classification, and data mining. Osuna et al. (1997) applied SVMs for face detection. Heisele *et al.* (2004) achieved high face detection rate by using 2nd degree SVM. They applied hierarchical classification and feature reduction methods to speed up face detection using SVMs.

Feature extraction and reduction are two primary issues in feature selection that is essential in pattern classification. Whether it is for storage, searching, or classification, the way the data are represented can significantly influence performances. Feature extraction is a process of extracting more effective representation of objects from raw data to achieve high classification rates. For image data, many kinds of features have been used, such as raw pixel values, Principle Component Analysis (PCA), Independent Component Analysis (ICA), wavelet features, Gabor features, and gradient values. Feature reduction is a process of selecting a subset of features with preservation or improvement of classification rates. In general, it intends to speed up the classification process by keeping the most important class-relevant features.

BACKGROUND

Principal Components Analysis (PCA) is a multivariate procedure which rotates the data such that the maximum variabilities are projected onto the axes. Essentially, a set of correlated variables are transformed into a set of uncorrelated variables which are ordered by reducing the variability. The uncorrelated variables are linear combinations of the original variables, and the last of these variables can be removed with a minimum loss of real data. PCA has been widely used in image representation for dimensionality reduction. To obtain m principal components, a transformation matrix of $m \times N$ is multiplied by an input pattern of $N \times 1$. The computation is costly for high dimensional data.

Another well-known method of feature reduction uses Fisher's criterion to choose a subset of features that possess a large between-class variance and a small within-class variance. For two-class classification problem, the within-class variance for i-th dimension is defined as

$$\sigma_i^2 = \frac{\sum_{j=1}^{l} (g_{j,i} - m_i)^2}{l-1}, \tag{1}$$

where l is the total number of samples, $g_{j,i}$ is the i-th dimensional attribute value of sample j, and m_i is the mean value of the i-th dimension for all samples. The Fisher's score for between-class measurement can be calculated as

$$S_i = \left| \frac{m_{i,class1} - m_{i,class2}}{\sigma_{i,class1}^2 + \sigma_{i,class2}^2} \right| . \tag{2}$$

Copyright © 2009, IGI Global, distributing in print or electronic forms without written permission of IGI Global is prohibited.

By selecting the features with the highest Fisher's scores, the most discriminative features between class 1 and class 2 are retained.

Weston et al. (2000) developed a feature reduction method for SVMs by minimizing the bounds on the leave-one-out error. Evgenious et al. (2003) introduced a method for feature reduction for SVMs based on the observation that the most important features are the ones that separate the hyperplane the most. Shih and Cheng (2005) proposed an improved feature reduction method in input and feature space for the 2nd degree polynomial SVMs.

MAIN FOCUS

In this section, we present an improved feature reduction method for the 2nd degree polynomial SVMs. In the input space, a subset of input features is selected by ranking their contributions to the decision function. In the feature space, features are ranked according to the weighted support vector in each dimension. By applying feature reduction in both input and feature space, a fast non-linear SVM is designed without a significant loss in performance. Here, the face detection experiment is used to illustrate this method.

Introduction to Support Vector Machines

Consider a set of l labeled training patterns $(\mathbf{x}_1, y_1)$, $(\mathbf{x}_2, y_2)$, ..., $(\mathbf{x}_i, y_i)$, ..., $(\mathbf{x}_l, y_l)$ where $\mathbf{x}_i$ denotes the i-th training sample and $y_i \in \{1, -1\}$ denotes the class label. For two-class classification, SVMs use a hyperplane that maximizes the margin (i.e., the distance between the hyperplane and the nearest sample of each class). This hyperplane is viewed as the *Optimal Separating Hyperplane* (OSH).

If the data are not linearly separable in the input space, a non-linear transformation function $\Phi(\cdot)$ is used to project $\mathbf{x}_i$ from the input space to a higher dimensional feature space. An OSH is constructed in the feature space by maximizing the margin between the closest points $\Phi(\mathbf{x}_i)$ of two classes. The inner-product between two projections is defined by a kernel function $K(\mathbf{x}, \mathbf{y}) = \Phi(\mathbf{x}) \cdot \Phi(\mathbf{y})$. The commonly-used kernels include polynomial, Gaussian RBF, and Sigmoid kernels.

The decision function of the SVM is defined as

$$f(\mathbf{x}) = \mathbf{w} \cdot \Phi(\mathbf{x}) + b = \sum_{i=1}^{l} \alpha_i y_i K(\mathbf{x}, \mathbf{x}_i) + b \,, \tag{3}$$

where $\mathbf{w}$ is the support vector, α_i is the Lagrange multiplifier, and b is a constant. The optimal hyperplane can be obtained by maximizing

$$W(\boldsymbol{\alpha}) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \, \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \,, \tag{4}$$

subject to

$$\sum_{i=1}^{l} \alpha_i y_i = 0 \;\text{ and } 0 \le \alpha_i \le C,$$

where C is regularization constant that manages the tradeoff between the minimization of the number of errors and the necessity to control the capacity of the classifier.

Feature Reduction in Input Space

A feature reduction method is proposed for the 2nd-degree polynomial SVM with kernel $K(\mathbf{x}, \mathbf{y}) = (1 + \mathbf{x} \cdot \mathbf{y})^2$ After training, the decision function for a pattern $\mathbf{x}$ is defined in Box 1, where s is the total number of support vectors, $\mathbf{x}_i$ is the i-th support vector, and $x_{i,k}$ and x_k are respectively the k-th dimension for the support vector $\mathbf{x}_i$ and the pattern $\mathbf{x}$. The component in the k-th dimension (where $k = 1, 2, \ldots, N$) is shown in Box 2.

The m features with the largest contributions to the decision function are selected from the original N features. The contribution can be obtained by

$$F(k) = \int_V f(\mathbf{x}, k) dP(\mathbf{x}), \tag{7}$$

where V denotes the input space and $P(\mathbf{x})$ denotes the probability distribution function. Since $P(\mathbf{x})$ is unknown, we approximate $F(k)$ using a summation over the support vectors as shown in Box 3.

6 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/chapter/feature-reduction-support-vector-machines/10922

Related Content

Intelligent Query Answering

Zbigniew W. Rasand Agnieszka Dardzinska (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition (pp. 1073-1078).*

www.irma-international.org/chapter/intelligent-query-answering/10954

Data Mining in Genome Wide Association Studies

Tom Burr (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition (pp. 465-471).*

www.irma-international.org/chapter/data-mining-genome-wide-association/10861

Topic Maps Generation by Text Mining

Hsin-Chang Yangand Chung-Hong Lee (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition (pp. 1979-1984).*

www.irma-international.org/chapter/topic-maps-generation-text-mining/11090

Histograms for OLAP and Data-Stream Queries

Francesco Buccafurri (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition (pp. 976-981).*

www.irma-international.org/chapter/histograms-olap-data-stream-queries/10939

Database Sampling for Data Mining

Patricia E.N. Lutu (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition (pp. 604-609).*

www.irma-international.org/chapter/database-sampling-data-mining/10883