CCIT Report #533

Journal of Machine Learning Research X (2005) X-X Submitted 1/2005; Published X/XX

May 2005

Feature Selection by Global Minimization of a Generalization Bound

Dori Peleg DORIP@TX.TECHNION.AC.IL
Department of Electrical Engineering
Technion, Haifa 32000, Israel

Ron Meir RMEIR@EE.TECHNION.AC.IL
Department of Electrical Engineering
Technion, Haifa 32000, Israel

Editor: Unknown

Abstract

A feature selection algorithm is presented based on the global minimization of a data-dependent generalization error bound. Feature selection and scaling algorithms often lead to non-convex optimization problems, which in many previous approaches were addressed through gradient descent procedures, which can only guarantee convergence to a local minimum. We propose an alternative approach, whereby the global solution of the non-convex optimization problem is derived by an equivalent convex conic optimization problem. Highly competitive numerical results on both artificial and real-world data sets are reported. The relation of the algorithm to the support vector machine algorithm is also discussed.

Keywords: Feature Selection, Dimensionality Reduction, Classification, Generalization Error Bounds, Statistical Learning Theory.

1. Introduction

This paper presents a new approach to feature selection for classification where the goal is to learn a decision rule from a training set of pairs $S_n = \left\{x^{(i)}, y^{(i)}\right\}_{i=1}^{n}$, where $x^{(i)} \in \mathbb{R}^d$ are input patterns and $y^{(i)} \in \{-1, 1\}$ are the corresponding labels. The goal of a classification algorithm is to find a separating function $f(\cdot)$, based on the training set, which will generalize well, i.e. classify new patterns with as few errors as possible. Feature selection schemes often utilize, either explicitly or implicitly, scaling variables, $\{\sigma_j\}_{j=1}^{d}$, which multiply each feature. The aim of such schemes is to optimize an objective function over $\sigma \in \mathbb{R}^d$.

Feature selection can be viewed as the special case $\sigma_j \in \{0, 1\}$, $j = 1, \ldots, d$, where a feature j is removed if $\sigma_j = 0$. The more general case of feature *scaling* is considered here, namely $\sigma_j \geq 0$, $j = 1, \ldots, d$. Clearly feature selection is a special case of feature scaling.

The overwhelming majority of feature selection algorithms in the literature, separate the feature selection and classification tasks, while solving either a combinatorial or a non-

©2005 Peleg and Meir.