Neurocomputing 73 (2010) 1760–1773

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Weighted feature extraction with a functional data extension ☆

Luis Gonzalo Sánchez Giraldo [a,*,1], Germán Castellanos Domínguez [b]

[a] Electrical and Computer Engineering Department, University of Florida, Gainesville, USA
[b] Electrical, Electronics and Computing Engineering Department, Universidad Nacional de Colombia, Manizales, Caldas, Colombia

ARTICLE INFO

Available online 19 March 2010

Keywords:
Dimensionality reduction
Feature selection
Feature extraction
Principal component analysis
Regularized discriminant analysis

ABSTRACT

Dimensionality reduction has proved to be a beneficial tool in learning problems. Two of the main advantages provided by dimensionality reduction are interpretation and generalization. Typically, dimensionality reduction is addressed in two separate ways: variable selection and feature extraction. However, in the recent years there has been a growing interest in developing combined schemes such as feature extraction with built-in feature selection. In this paper, we look at dimensionality reduction as a rank-deficient problem that embraces variable selection and feature extraction, simultaneously. From our analysis, we derive a weighting algorithm that is able to select and linearly transform variables by fixing the dimensionality of the space where a relevance criterion is evaluated. This step enforces sparseness on the resulting weights. Our main goal is dimensionality reduction for classification problems. Namely, we introduce modified versions of principal component analysis (PCA) by expectation maximization (EM) and linear regularized discriminant analysis (RDA). Finally, we propose a simple extension of WRDA that deals with functional features. In this case, observations are described by a set of functions defined over the same domain. Methods were put to test on artificial and real data sets showing high levels of generalization even for small sized training samples.

1. Introduction

Identifying relevant features has been discussed in the past by several authors such as [1,2], It has been shown theoretically and experimentally that dimensionality reduction can improve the generalization ability of a learning algorithm. One of the most common issues that encourages the use of dimensionality reduction is the problem of over-fitting to the training data when the number of observations (size of the sample) is small compared to the number of variables that represent each instance. A common paradigm in pattern recognition systems is to take advantage of prior knowledge that can be used to tailor features to effectively describe objects. However, very often this prior knowledge is not available or cannot be easily incorporated; in such cases it is desirable to employ features that are rather generic. The key issue is how to use these sets of features, wisely [3]. Therefore, reducing the size of data either by encoding or removing the redundant and irrelevant information becomes necessary if one wants to achieve good performance in the prediction. Functional regularization as well as Bayesian methods have attempted to attack this problem by restricting the set of hypothesis that can be implemented by a learning machine, theoretical as well as empirical evidence justify their widespread use. Among the algorithms that belong to this context, we can find the support vector machines, relevance vector machines, Gaussian processes for classification. Although, these methods have succeeded on various tasks, they do not solve the problem of identifying irrelevant information, explicitly. As a consequence, performance tends to degrade in the presence of many variables that do not reflect relevant information for the problem. Dimensionality reduction comes into play as a very important stage to overcome the above limitation. It is then not surprising that research on this field has remained active during the last years [4,5].

Dimensionality reduction techniques are mainly divided into two groups: feature selection and feature extraction methods; both attempt to reduce dimensionality, nonetheless,they are based on different objectives. Feature selection can be understood as a explicit selection of a subset from input set of variables; whereas feature extraction comprises transformations of the input set to obtain a new set of variables that can represent the problem under some optimality criterion. In short, the problem consists on finding a subset of features that can be efficiently encoded preserving the relevant information related to the task.

☆This research was carried out under grant funded by COLCIENCIAS; Centro de Investigacion e Innovacion de Excelencia ARTICA.

* Corresponding author. Computational Neuro-Engineering Laboratory, NEB 486, University of Florida, Gainesville, USA. Tel.: +1 646 895 4860.
E-mail addresses: luisitobarcito@ufl.edu (L.G.S. Giraldo), cgcastellanosd@unal.edu.co (G.C. Domínguez).
[1] This work was done while the author was a student at Universidad Nacional de Colombia, Manizales.

0925-2312/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2009.12.030

Feature selection can be addressed either as a binary search or as a weighting procedure. Binary search involves explicit enumeration of the subsets of features by assigning to each subset an indicator variable (e.g. binary vector). Feature weighting relax this constraint by letting the indicator variables take continuous values to weight each feature. The reason for using continuous values is the possibility to include differentiable penalties in the target function that allows the use of mathematical programming tools to optimize weights. Recent work in this area has been directed towards methods that incorporate sparseness; clear examples of this trend are L^1 regularization [6], commonly known as the Lasso approach and combined L^1 and L^2 [7], which is called elastic-net regularization, and related methods such as the ones presented in [8]. In this work, we address the problem of dimensionality reduction by finding a relevant subset of projected features (joint feature selection and extraction); this formulation corresponds to a weighted feature extraction. We relate the problem of finding relevant information to a rank deficient formulation for dimensionality reduction that suggests some desirable properties for weighting schemes. We focus our discussion around formulations for supervised weighted principal component analysis (WPCA) and weighted regularized discriminant analysis (WRDA) as suitable methods for feature extraction with built-in feature selection. Weighted versions of PCA have already been employed for image processing applications [9]. Within the statistics community this formulation is understood as a generalized form of PCA [10]. The generalized PCA consider weighting operations on both instances and variables. For example, in [28] a linear projection that takes into account local structure and class information is presented. The derived solution can be understood as a form of PCA that performs weighting on the instances. It is clear then that the choice of the weights is rather an open problem. In our case the ratio between two matrix traces serves as the objective function that guides the variable weighting. This scheme turns out to be a constrained optimization problem where constraints are optimization problems themselves.

One interesting question is whether these built in feature selection methods can be extended to other problems where explicit enumeration seem to be the only reasonable approach. Functional data analysis appears to be one of such cases. The extension of the existing multivariate methods to more involved representations such as infinite dimensional objects is not quite obvious. Our work explores a first attempt to extend our feature selection method to the functional data analysis framework.

The paper is organized as follows: a review and reformulation of the concept of relevance in terms of relevant mappings is presented; this alternative view agrees with our approach to the problem as a rank deficient problem, for which we provide a simplified treatment in terms of linear operators in Hilbert spaces; then we derive algorithms for WPCA and WRDA. We also propose a simple extension to more general type of representation were each object is represented by a set of stochastic processes with the same index set. Finally, we provide some results on artificial as well as real data. For the functional adaptation methodology, we test on artificially generated Gaussian processes with very interesting results.

2. Relevance

Roughly speaking, the purpose of dimensionality reduction is to find a transformation of the original data (initial representation) that preserves the relations with some target variable while maintaining the set of descriptors as small as possible. Notice, this definition also considers the cases of non-linear transformations that can be thought as mappings to high dimensional spaces, on which we want to keep the dimensionality of the mapped data as low as possible (a subspace of the high dimensional space).

Definition 2.1 (*Relevance [11,12]*). On the basis of prior evidence E, a hypothesis H is considered, and the change in the likelihood of H due to additional evidence I is examined. If the likelihood of H is changed by the addition of I to E, I is said to be relevant to H on evidence E; otherwise it is irrelevant. In particular, if the likelihood of H is increased due to the addition of I to E, I is said to be positively relevant to H; if the likelihood of H is decreased, I is said to be negatively relevant.

The above definition of relevance agrees with the common sense notion of the word. The evidence is provided by the features and their relevance to the hypothesis is measured in terms of an objective function which increases or decreases according to the employed set of descriptors. Now, we need a way to quantify the relevance in order to translate the above statement into a concrete implementation. In machine learning, the most widespread definition of relevance compares conditional probability measures of predicted variables (hypothesis) given different sets of predictors (prior and additional evidence) [1]. Let $\boldsymbol{\xi}=\{\xi_1,\xi_2,\ldots,\xi_p\}$ be the full set of descriptors, $\boldsymbol{\xi}_i^-=\boldsymbol{\xi}\backslash\xi_i$ (the complement of ξ_i in $\boldsymbol{\xi}$), and Y the target variable.

Definition 2.2 (*Strong relevance*). A descriptor ξ_i is strongly relevant iff

$$P(Y|\boldsymbol{\xi}_i^-,\xi_i)\neq P(Y|\boldsymbol{\xi}_i^-).$$

Definition 2.3 (*Weak relevance*). A descriptor ξ_i is weakly relevant iff

$$P(Y|\boldsymbol{\xi}_i^-,\xi_i)=P(Y|\boldsymbol{\xi}_i^-),$$

and $\exists\boldsymbol{\xi}_i^*\subset\boldsymbol{\xi}_i^-$, such that

$$P(Y|\boldsymbol{\xi}_i^*,\xi_i)\neq P(Y|\boldsymbol{\xi}_i^*).$$

Corollary 2.1 (*Irrelevance*). *A descriptor ξ_i is irrelevant iff*

$$\forall\boldsymbol{\xi}_i^*\subset\boldsymbol{\xi}_i^-,P(Y|\boldsymbol{\xi}_i^*,\xi_i)=P(Y|\boldsymbol{\xi}_i^*).$$

Consider the set of objects $\mathcal{X}$, associated with each $x\in\mathcal{X}$ there is an element $y\in\mathcal{Y}$. Let $\mathcal{B}_{\mathcal{X}\times\mathcal{Y}}$ be the $\sigma-$algebra of $\mathcal{X}\times\mathcal{Y}$, and a probability measure $P_{\mathcal{X}\times\mathcal{Y}}$, so we have a measure space. We will consider $\boldsymbol{\xi}$ as a set of measurable functions each one from $\mathcal{X}$ to their range space $\mathcal{R}(\xi_i)\subseteq\Omega_i$; we refer to the space containing the range of this function as the initial representation space $(\Omega_1\times\cdots\times\Omega_p=\Omega,\mathcal{B}_\Omega)$. Now consider a set of measurable functions $T=\{T_j:\Omega\mapsto\Omega_j^*\}$, where $\Omega^*=\Omega_1\times\cdots\times\Omega_c$ is an alternative representation space provided with a Borel $\sigma-$algebra $\mathcal{B}_\Omega$. We say that T is relevant iff

$$P(Y|\boldsymbol{\xi})=P(Y|T)\quad\text{and}\quad P(Y|\boldsymbol{\xi})\neq P(Y|T\backslash T_i)\quad\text{for all } i.$$

In the supervised setting, where correspondences between $\mathcal{X}$ and $\mathcal{Y}$ are available, the above definition fits either for feature selection or feature extraction. The first case can be understood as a permutation and clipping map to some of set of relevant descriptors, by this we do not really mean the map implies the smallest achievable dimension, but in terms of strong and weak relevance an optimal subset must contain only strongly relevant elements. Regarding feature extraction, the simplest example would be a two-class linear discriminant analysis with idealized conditions i.e. equal, isotropic within class covariance matrices and different mean vectors. In this case the conditional probability of the class by projecting the data onto the line with the maximum Fisher score does not eliminate the necessary information to obtain the same labels. The unsupervised problem is a