Pattern Recognition 44 (2011) 650–661

Contents lists available at ScienceDirect

Pattern Recognition

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

Signature extraction using mutual interdependencies

Heiko Claussen [a,b,*], Justinian Rosca [a], Robert Damper [b]

[a] Siemens Corporation, Corporate Research, 755 College Road East, Princeton, NJ 08540, USA
[b] University of Southampton, School of Electronics and Computer Science, Highfield, Southampton SO17 1BJ, UK

ARTICLE INFO

Article history:
Received 20 March 2009
Received in revised form
25 August 2010
Accepted 18 September 2010

Keywords:
Algorithms
Signal processing
Pattern classification
Signal analysis
Speaker recognition
Face recognition

ABSTRACT

Recently, mutual interdependence analysis (MIA) has been successfully used to extract representations, or "mutual features", accounting for samples in the class. For example, a mutual feature is a face signature under varying illumination conditions or a speaker signature under varying channel conditions. A mutual feature is a linear regression that is equally correlated with all samples of the input class. Previous work discussed two equivalent definitions of this problem and a generalization of its solution called generalized MIA (GMIA). Moreover, it showed how mutual features can be computed and employed. This paper uses a parametrized version GMIA(λ) to pursue a deeper understanding of what GMIA features really represent. It defines a generative signal model that is used to interpret GMIA(λ) and visualize its difference to MIA, principal and independent component analysis. Finally, we analyze the effect of λ on the feature extraction performance of GMIA(λ) in two standard pattern recognition problems: illumination-independent face recognition and text-independent speaker verification.

1. Introduction

Statistical pattern recognition methods such as Fisher's linear discriminant analysis (FLDA) [9], canonical correlation analysis (CCA) [16] or ridge regression [25] aim to model or extract the essence of a dataset. The goal is to find a simplified data representation that retains the information that is necessary for subsequent tasks such as classification or prediction. Each method uses a different viewpoint and criteria to find this "optimal" representation. Furthermore, pattern recognition problems implicitly assume that the number of observations is usually much higher than the dimensionality of each observation. This allows one to study characteristics of the distributional observations and design proper discriminant functions for classification. For instance, FLDA is used to reduce the dimensionality of a dataset by projecting data points on a space that maximizes the ratio of the between- and within-class scatter of the training data. In this way, FLDA aims to find a simplified data representation that retains the discriminant characteristics for classification. On the other hand, CCA assumes one common source in two datasets. The dimensionality of the data is reduced by retaining the space that is spanned by pairs of projecting directions in which the datasets are maximally correlated. In contrast, ridge regression finds a linear combination of the inputs that best fits a desired response.

In this paper, we present alternative criteria to find an "optimal" dataset representation. We aim to extract an invariant representation of high-dimensional instances of a single class, where the number of input instances N is smaller than their dimensionality D. An invariant is a property or feature of the input data that does not change within its class. Approaches that have been designed for this purpose are mutual interdependence analysis (MIA) and generalized MIA (GMIA) [4–6]. We revisit both methods in Sections 2 and 3, respectively, and parametrize GMIA with λ, which subsumes MIA for $\lambda = 0$. In Section 4, we introduce a generative model for GMIA(λ). On synthetic data, we demonstrate that GMIA(λ) extracts features unlike approaches such as PCA and ICA. Also we show how these features differ from the sample mean. Section 5 evaluates the discriminative quality of GMIA(λ) features for illumination-invariant face recognition on synthetic data. Section 6 analyses the effect of λ on real data for illumination-invariant face recognition and text-independent speaker verification. The document concludes with a summary and directions for future work.

2. Mutual interdependence analysis (MIA)

MIA was first introduced by the authors in Claussen et al. [4] to uniquely represent high-dimensional samples of a single class. The understanding of how this problem can be succinctly and

* Corresponding author at: Siemens Corporation, Corporate Research, 755 College Road East, Princeton, NJ 08540, USA.
E-mail addresses: Heiko.claussen@siemens.com (H. Claussen), Justinian.rosca@siemens.com (J. Rosca), rid@ecs.soton.ac.uk (R. Damper).

0031-3203/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2010.09.012

elegantly stated has been evolved and generalized [6]. In this section we present an up to date statement of MIA.

2.1. Scatter-based definition of MIA

Throughout this paper, $\mathbf{x}_i^{(p)} \in \mathbb{R}^D$ denotes the ith input vector, $i=1\ldots N^{(p)}$ in class p. Furthermore, we use $\mathbf{X}^{(p)} \subseteq \mathbf{X}$ to represent a matrix with columns $\mathbf{x}_i^{(p)}$ and $\mathbf{X}$ to denote the matrix with columns $\mathbf{x}_i$ of all K classes. Moreover, $\boldsymbol{\mu}=(1/N)\sum_{i=1}^{N}\mathbf{x}_i$, $\underline{\mathbf{1}}$ is a vector of ones and $\mathbf{I}$ represents the identity matrix.

Assume that we wish to find a class representation $\mathbf{w}^{(p)}$ of high-dimensional data vectors $\mathbf{x}_i^{(p)}$ ($D \geq N^{(p)}$). A common first step is to select features so as to reduce the dimensionality of the data. However, because of possible loss of information, this preprocessing is not always desirable. Therefore, we aim to find a class representation of similar or same dimensionality as the inputs.

The quality of such a representation can be evaluated by its correlation with the class instances. Our intuition is that a superior class representation is highly correlated and also has a small variance of the correlations over all instances in the class. The former condition ensures that most of the signal energy in the samples is captured. The latter condition is indicative of membership in a single class. Note that only vectors in the span of the class instances contribute to the cross-correlation value. Therefore, in the absence of prior knowledge, it is reasonable to constrain the search for a class representation $\mathbf{w}$ to the span of the training vectors $\mathbf{w}=\mathbf{X}^{(p)}\cdot\mathbf{c}$, where $\mathbf{c}\in\mathbb{R}^{N^{(p)}}$. This problem definition is the motivation for the MIA criterion proposed in Claussen et al. [4].

The MIA representation for class p is defined as a direction $\mathbf{w}_{\text{MIA}}^{(p)} \in \mathbb{R}^D$ that minimizes the projection scatter of the class p inputs, under the linearity constraint to be in the span of $\mathbf{X}^{(p)}$:

$$\mathbf{w}_{\text{MIA}}^{(p)} = \underset{\mathbf{w},\mathbf{w}=\mathbf{X}^{(p)}\cdot\mathbf{c}}{\operatorname{argmin}} (\mathbf{w}^T\cdot(\mathbf{X}^{(p)}-\boldsymbol{\mu}^{(p)}\cdot\underline{\mathbf{1}}^T)\cdot(\mathbf{X}^{(p)}-\boldsymbol{\mu}^{(p)}\cdot\underline{\mathbf{1}}^T)^T\cdot\mathbf{w}) \tag{1}$$

Note that the original space of the inputs spans the space of the mean subtracted inputs plus possibly one additional dimension. Indeed, the mean subtracted inputs, which are linear combinations of the original inputs, sum to zero. Mean subtraction cancels linear independence resulting in a 1D span reduction. The following two theorems describe the MIA solution.

Theorem 2.1. *The minimum of the criterion in Eq.* (1) *is zero if the inputs* $\mathbf{x}_i$ *are linearly independent.*

If inputs are linearly independent and span a space of dimensionality $N \leq D$, then the subspace of the mean subtracted inputs in Eq. (1) has dimensionality $N-1$. There exists an additional dimension in $\mathbb{R}^N$, orthogonal to this subspace. Thus, the scatter of the mean subtracted inputs can be made zero. The existence of a solution where the criterion in Eq. (1) becomes zero is indicative of an invariance property of the data.

Theorem 2.2. *The solution of Eq.* (1) *is unique (up to scaling) if the inputs* $\mathbf{x}_i$ *are linearly independent.*

By solving in the span of the original rather than mean subtracted inputs, a closed form solution of Eq. (1) can be found [4]:

$$\mathbf{w}_{\text{MIA}}^{(p)} = \zeta\mathbf{X}^{(p)}\cdot(\mathbf{X}^{(p)^T}\cdot\mathbf{X}^{(p)})^{-1}\cdot\underline{\mathbf{1}} \quad \text{where } \zeta \text{ is a constant} \tag{2}$$

Consider that $(\mathbf{X}^{(p)^T}\cdot\mathbf{X}^{(p)})^{-1}\cdot\underline{\mathbf{1}}$ is a column vector. The structure of the solution shows that $\mathbf{w}$ is a data-dependent transformation representing a linear combination of the input observations.

The mathematical structure of this MIA solution has a striking similarity with linear regression. Indeed this result can be obtained as follows. Let us assume the regression problem $\mathbf{y}=\mathbf{X}\cdot\boldsymbol{\beta}$. We are looking for $\boldsymbol{\beta}$ such that the unknown regression $\mathbf{y}$ is equally correlated with all inputs $\mathbf{X}^T\cdot\mathbf{y}=\underline{\mathbf{1}}$. It can be shown that the solution to this problem is given by Eq. (2) with $\zeta=1$ and $\mathbf{y}=\mathbf{w}$. In Section 3, we return to the discussion of similarities between the two problems. Eq. (2) computes a unique representation we call MIA with the property of invariant correlation with all samples in the input. This uniqueness indicates that MIA captures an inherent property of the input data.

2.2. CCA-based definition of MIA

The minimum variance criterion is also used in other data analysis approaches such as FLDA. However, this theory does not apply when analyzing data from one class. This motivated the comparison with CCA as a generalization of FLDA, and the discovery of an equivalent, CCA-based formulation of the MIA problem. We revisit this new definition following and extending Claussen et al. [6]. First, we review CCA and its FLDA equivalent formulation. Thereafter, we extend this formulation to address the MIA problem.

If a common source $\mathbf{s}\in\mathbb{R}^N$ influences two datasets $\mathbf{X}\in\mathbb{R}^{D\times N}$ and $\mathbf{Z}\in\mathbb{R}^{K\times N}$, of possibly different dimensionality, CCA is used to extract this inherent similarity. The goal of CCA is to find two vectors to project the datasets such that their projection lengths are maximally correlated. Let $\mathbf{C_{XZ}}$ denote the cross covariance matrix between the datasets $\mathbf{X}$ and $\mathbf{Z}$. Then the CCA problem is given by maximization of the objective function

$$J(\mathbf{a},\mathbf{b}) = \frac{\mathbf{a}^T\cdot\mathbf{C_{XZ}}\cdot\mathbf{b}}{\sqrt{\mathbf{a}^T\cdot\mathbf{C_{XX}}\cdot\mathbf{a}}\cdot\sqrt{\mathbf{b}^T\cdot\mathbf{C_{ZZ}}\cdot\mathbf{b}}} \tag{3}$$

over the vectors $\mathbf{a}$ and $\mathbf{b}$. The CCA problem can be solved by a singular value decomposition (SVD) of $\mathbf{C_{XX}}^{-1/2}\cdot\mathbf{C_{XZ}}\cdot\mathbf{C_{ZZ}}^{-1/2}$ [19]. The solution is obtained by solving the two eigenvector problems:

$$\left(\mathbf{C_{XX}}^{-1/2}\cdot\mathbf{C_{XZ}}\cdot\mathbf{C_{ZZ}}^{-1}\cdot\mathbf{C_{ZX}}\cdot\mathbf{C_{XX}}^{-1/2}\right)\cdot\mathbf{a}=\lambda\mathbf{a} \tag{4}$$

and

$$\left(\mathbf{C_{ZZ}}^{-1/2}\cdot\mathbf{C_{ZX}}\cdot\mathbf{C_{XX}}^{-1}\cdot\mathbf{C_{XZ}}\cdot\mathbf{C_{ZZ}}^{-1/2}\right)\cdot\mathbf{b}=\lambda\mathbf{b} \tag{5}$$

We hypothesize that the maximally correlated projections $\mathbf{X}^T\cdot\mathbf{a}$ and $\mathbf{Z}^T\cdot\mathbf{b}$ represent an estimate of the common source.

Canonical correlation analysis can be used to extract classification relevant information from a set of inputs. Indeed, let $\mathbf{X}$ be the union of all data points and $\mathbf{Z}$ the table of corresponding class memberships, $k=1,\ldots,K$ and $i=1,\ldots,N$:

$$\mathbf{Z}_{ki} = \begin{cases} 1 & \text{if } \mathbf{x}_i \in \mathbf{X}^{(k)} \\ 0 & \text{otherwise} \end{cases}$$

All classification relevant information is represented by this classification table. Therefore, this information is retained in those input components of $\mathbf{X}$ that originate from a common virtual source with the classification table. It has been shown [2,19,14,1] that this special CCA approach is equivalent to FLDA.

CCA with $\mathbf{Z}$ given by the class membership can be modified to extract a representation of inputs from a single class, similar to MIA. One possible interpretation of CCA is from the point of view of the cosine angle between the (non-mean-subtracted) vectors $\mathbf{a}^T\cdot\mathbf{X}$ and $\mathbf{Z}^T\cdot\mathbf{b}$. The aim is to find a vector pair that results in a minimum angle. We will use a modified CCA criterion (MCCA) as follows. First, consider the original inputs rather than the mean subtracted covariance matrices; second, the class membership table for data from a single class collapses to a vector and $\mathbf{b}$ to a scalar, therefore $\mathbf{Z}^T\cdot\mathbf{b}=\underline{\mathbf{1}}\cdot\mathbf{b}$. Thus, criterion Eq. (3) becomes