Neurocomputing ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at ScienceDirect

Neurocomputing

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

Application of Linear Regression Classification to low-dimensional datasets

Mehmet Koç [a,*], Atalay Barkana [b]

[a] Department of Electrical and Electronics Engineering, Bilecik Seyh Edebali University, Bilecik 11210, Turkey
[b] Department of Electrical and Electronics Engineering, Anadolu University, Eskisehir 26555, Turkey

ARTICLE INFO

Article history:
Received 23 April 2013
Received in revised form
26 August 2013
Accepted 1 October 2013
Communicated by B. Apolloni

Keywords:
Correlation matrix
Subspace methods
Linear Regression Classification
Class-featuring information compression

ABSTRACT

The Traditional Linear Regression Classification (LRC) method fails when the number of data in the training set is greater than their dimensions. In this work, we proposed a new implementation of LRC to overcome this problem in the pattern recognition. The new form of LRC works even in the case of having low-dimensional excessive number of data. In order to explain the new form of LRC, the relation between the predictor and the correlation matrix of a class is shown first. Then for the derivation of LRC, the null space of the correlation matrix is generated by using the eigenvectors corresponding to the smallest eigenvalues. These eigenvectors are used to calculate the projection matrix in LRC. Also the equivalence of LRC and the method called Class-Featuring Information Compression (CLAFIC) is shown theoretically. TI Digit database and Multiple Feature dataset are used to illustrate the use of proposed improvement on LRC and CLAFIC.

1. Introduction

Subspace methods are widely used in several pattern recognition (PR) areas [1–3]. Also Linear Regression Classification (LRC) [4] has become a popular subspace method in face recognition area. Several methods inspired from LRC are proposed by the researchers [5–7]. In appearance-based methods a $w \times h$ sized image is represented with a point in wh-dimensional space, i.e., for a 30×40 dimensional image, where feature space turns out to be 1200 dimensional. Since the number of feature vectors in the training set is very small with respect to the dimension of the feature vectors in face recognition problems, LRC can be easily applied as a subspace method. But LRC cannot be directly applied to the PR problems where the number of samples is greater than the dimension of the samples because the projection of the data onto the subspace spanned by training set samples will cover the whole space, that is, the predictor of a class will span the whole space.

CLAFIC is one of the earliest and well-known subspace methods [8,9]. It is used in many pattern recognition areas [10–12]. In this method, Principal Component Analysis (PCA) is used to compute the basis vectors of the class specific subspace. A test sample is assigned to the class where it has the largest length in that subspace.

In this paper, we proposed a method that makes the LRC method applicable to PR problems where the number of samples (N) in a class is greater than the dimension of the samples (n). In this proposal, we take the advantage of the relation between the predictor of a class and the eigenvectors of the correlation matrix. Firstly, it is shown that the column space of the predictor of a class is the same subspace spanned by the correlation matrix's eigenvectors that correspond to the nonzero eigenvalues. In the case of having larger number of samples than their dimensions, $n < N$, all the eigenvalues of the correlation matrix are nonzero and its null space does not exist. However the correlation matrix's eigenvectors corresponding to the smallest eigenvalues (since all are positive) can be used to build a similar subspace idea with the null space of the correlation matrix. In all these cases we also show that LRC and CLAFIC methods yield identical results in classification.

It is better to start with the definition of a correlation matrix for a set of feature vectors not to cause any misunderstanding.

Definition. Let $\boldsymbol{w}_i,\ \ i = 1, \ldots, N$ be the feature vectors of a class or they are the data in the training set of a class. Then the correlation matrix can be calculated as $\boldsymbol{R} = \boldsymbol{W}\boldsymbol{W}^T$ [13], where $\boldsymbol{W}$ is a matrix of the form

$$\boldsymbol{W} = [\boldsymbol{w}_1 \vdots \boldsymbol{w}_2 \vdots \cdots \vdots \boldsymbol{w}_N] \quad (1)$$

It is known that the correlation matrix $\boldsymbol{R}$ is positive semidefinite and symmetric, therefore all the eigenvalues are nonnegative [13]. This definition will be used in the following sections.

A review of LRC is given in Section 2. Also relations between hat matrix, predictor and the correlation matrix of a class are given in

* Corresponding author. Tel.: +90 228 214 1552
E-mail addresses: mehmet.koc@bilecik.edu.tr, mehmetkoc12@gmail.com (M. Koç), atalaybarkan@anadolu.edu.tr (A. Barkana).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.10.009

the same section if $n > N$. Implementation of LRC, when having $n < N$ case, is explained in Section 3. The equivalence of LRC and CLAFIC is shown in Section 4. Experimental work is given is Section 5 and the conclusion is given in Section 6.

2. Review of Linear Regression Classification

In this section a review of LRC is given first. Also the relations between the hat matrix, the predictor, and the correlation matrix are given later. It is proved that the hat matrix and the projection matrix obtained by using the eigenvectors corresponding to the nonzero eigenvalues are the same projection matrices. Therefore the estimations of vector $\boldsymbol{a}$ obtained by using these two projection matrices will be the same.

2.1. Linear Regression Classification

Assume that we have C classes and each class has N samples in the n-dimensional feature space. Let $\boldsymbol{w}_k^i$, $k = 1, \ldots, N$ be the n-dimensional feature vectors in the training set of the ith class. It is also assumed that n is larger than N. Then the predictor of the ith class is

$$\boldsymbol{W}_i = [\boldsymbol{w}_1^i \vdots \boldsymbol{w}_2^i \vdots \cdots \vdots \boldsymbol{w}_N^i]. \tag{2}$$

If a feature vector $\boldsymbol{a}$ belongs to the ith class, then it can be represented by the linear combinations of these feature vectors with an error $\boldsymbol{\varepsilon}$ according to LRC. Hence

$$\boldsymbol{a} = \boldsymbol{W}_i \boldsymbol{\beta}_i + \boldsymbol{\varepsilon} \tag{3}$$

where $\boldsymbol{\beta}_i$ is $N \times 1$ dimensional parameter vector. The sum of error squares is

$$S = \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} = (\boldsymbol{a} - \boldsymbol{W}_i \boldsymbol{\beta}_i)^T (\boldsymbol{a} - \boldsymbol{W}_i \boldsymbol{\beta}_i)$$

After the minimization with respect to $\boldsymbol{\beta}_i$, the estimation of the vector parameters becomes

$$\tilde{\boldsymbol{\beta}}_i = (\boldsymbol{W}_i^T \boldsymbol{W}_i)^{-1} \boldsymbol{W}_i^T \boldsymbol{a} \tag{4}$$

Then the estimation of the vector $\boldsymbol{a}$ is

$$\tilde{\boldsymbol{a}}_i = \boldsymbol{W}_i (\boldsymbol{W}_i^T \boldsymbol{W}_i)^{-1} \boldsymbol{W}_i^T \boldsymbol{a} \tag{5}$$

Here the projection matrix

$$\boldsymbol{H}_i = \boldsymbol{W}_i (\boldsymbol{W}_i^T \boldsymbol{W}_i)^{-1} \boldsymbol{W}_i^T \tag{6}$$

is called the *hat matrix* for class $\boldsymbol{i}$. The classification is done according to the following distance criteria, i.e.,

$$C^* = \underset{i}{\operatorname{argmin}} \{\|\boldsymbol{a} - \tilde{\boldsymbol{a}}_i\|\}, \quad i = 1, 2, \ldots, C \tag{7}$$

2.2. Relations between hat matrix, predictor, and correlation matrix of a class

Let the subspace spanned by the predictor of the class i be V_i, then

$$V_i = span\,\{\boldsymbol{w}_1^i, \boldsymbol{w}_2^i, \ldots, \boldsymbol{w}_N^i\} = span\{\boldsymbol{W}_i\} \tag{8}$$

Let T_i be the orthonormal vector set $\{\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots \boldsymbol{q}_N\}$ that can be obtained by applying the Gram-Schmidt orthogonalization to the predictor $\boldsymbol{W}_i$ of the ith class, and let $\boldsymbol{Q}_i$ be a matrix whose columns are the vectors of the set T_i, that is,

$$\boldsymbol{Q}_i = [\boldsymbol{q}_1^i \vdots \boldsymbol{q}_2^i \vdots \cdots \vdots \boldsymbol{q}_N^i] \tag{9}$$

Then the columns of $\boldsymbol{Q}_i$ also span V_i, thus $V_i = span(\boldsymbol{Q}_i)$. The following theorem is given to prove this claim.

Theorem 1. *The hat matrix of the ith class* $\boldsymbol{H}_i = \boldsymbol{W}_i (\boldsymbol{W}_i^T \boldsymbol{W}_i)^{-1} \boldsymbol{W}_i^T$ *and the projection matrix onto the column space of* $\boldsymbol{Q}_i$ *are the same, that is,* $\boldsymbol{H}_i = \boldsymbol{Q}_i \boldsymbol{Q}_i^T$.

Proof. $\boldsymbol{W}_i$ can be factorized using QR decomposition as $\boldsymbol{W}_i = \boldsymbol{Q}_i \boldsymbol{M}_i$. Additionally, $\boldsymbol{W}_i$ and $\boldsymbol{Q}_i$ with orthogonal columns span the same subspace. $\boldsymbol{M}_i$ is an upper triangular matrix. The $(j,k)th$ entry of $\boldsymbol{M}_i$ is given as.

$$m_{jk} = \begin{cases} \left(\boldsymbol{w}_j^i\right)^T \boldsymbol{q}_k^i, & k \le j \\ 0, & k > j \end{cases}$$

$$\begin{aligned} \boldsymbol{W}_i (\boldsymbol{W}_i^T \boldsymbol{W}_i)^{-1} \boldsymbol{W}_i^T &= (\boldsymbol{Q}_i \boldsymbol{M}_i)((\boldsymbol{Q}_i \boldsymbol{M}_i)^T \boldsymbol{Q}_i \boldsymbol{M}_i)^{-1} (\boldsymbol{Q}_i \boldsymbol{M}_i)^T \\ &= \boldsymbol{Q}_i \boldsymbol{M}_i (\boldsymbol{M}_i^T \boldsymbol{Q}_i^T \boldsymbol{Q}_i \boldsymbol{M}_i)^{-1} \boldsymbol{M}_i^T \boldsymbol{Q}_i^T \\ &= \boldsymbol{Q}_i \boldsymbol{M}_i (\boldsymbol{M}_i^T \boldsymbol{I} \boldsymbol{M}_i)^{-1} \boldsymbol{M}_i^T \boldsymbol{Q}_i^T \\ &= \boldsymbol{Q}_i \boldsymbol{M}_i \boldsymbol{M}_i^{-1} \boldsymbol{M}_i^{-T} \boldsymbol{M}_i^T \boldsymbol{Q}_i^T \\ &= \boldsymbol{Q}_i \boldsymbol{Q}_i^T \end{aligned}$$

This completes the proof. □

In the following two theorems it will be shown that the subspace spanned by the predictor of a class in Eq. (2) and the range space of the correlation matrix are the same subspaces.

Theorem 2. *Let the null space of the ith class correlation matrix* $\boldsymbol{R}_i$ *be* $Null\,(\boldsymbol{R}_i) = \{\boldsymbol{x} \in \mathbb{R}^n | \boldsymbol{R}_i \boldsymbol{x} = 0\}$ *and let the complementary subspace of* V_i *be* $V_i^\perp = \{\boldsymbol{x} \in \mathbb{R}^n | \boldsymbol{x}^T \boldsymbol{w}^i = 0, \;\; \forall \boldsymbol{w}^i \in V_i\}$. *Then* $Null\,(\boldsymbol{R}_i) = V_i^\perp$ *holds.*

Proof. (i) Let $\boldsymbol{x}$ be a vector lying in $V_i^\perp$ then $\boldsymbol{x}^T \boldsymbol{w}^i = 0$ for all $\boldsymbol{w}^i \in V_i$.

$$\boldsymbol{R}_i \boldsymbol{x} = \boldsymbol{W}_i \boldsymbol{W}_i^T \boldsymbol{x} = \left[\boldsymbol{w}_1^i \vdots \boldsymbol{w}_2^i \vdots \cdots \vdots \boldsymbol{w}_N^i\right] \begin{bmatrix} (\boldsymbol{w}_1^i)^T \\ (\boldsymbol{w}_2^i)^T \\ \vdots \\ (\boldsymbol{w}_N^i)^T \end{bmatrix} x = \left[\boldsymbol{w}_1^i \vdots \boldsymbol{w}_2^i \vdots \cdots \vdots \boldsymbol{w}_N^i\right] \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \boldsymbol{0}.$$

Thus $\boldsymbol{x} \in Null(\boldsymbol{R}_i)$, so

$$V_i^\perp \subset Null\,(\boldsymbol{R}_i). \tag{10}$$

(ii) Let $\boldsymbol{x}$ be a feature vector in the null space of the correlation matrix of the ith class, $\boldsymbol{R}_i$, then $\boldsymbol{R}_i \boldsymbol{x} = \boldsymbol{0}$. Multiply the equation from the left side by $\boldsymbol{x}^T$,

$$\boldsymbol{x}^T \boldsymbol{R}_i \boldsymbol{x} = 0$$

$$\boldsymbol{x}^T \boldsymbol{W}_i \boldsymbol{W}_i^T \boldsymbol{x} = 0$$

$$(\boldsymbol{W}_i^T \boldsymbol{x})^T \boldsymbol{W}_i^T \boldsymbol{x} = 0$$

$$||\boldsymbol{W}_i^T x|| = 0$$

$$(\boldsymbol{w}_k^i)^T \boldsymbol{x} = 0, \quad i = 1, 2, \ldots, N \tag{11}$$

We know that $\boldsymbol{w}_k^i$'s in Eq. (11) form a basis for V_i, then $\boldsymbol{x} \in V_i^\perp$. Therefore the following holds:

$$Null\,(\boldsymbol{R}_i) \subset V_i^\perp \tag{12}$$

By combining Eqs. (10) and (12), we will end up with $Null\,(\boldsymbol{R}_i) = V_i^\perp$. □

Theorem 3. *If* $\boldsymbol{u}_k^i$, $k = 1, 2, \ldots, N$ *are the eigenvectors corresponding to the nonzero eigenvalues of the correlation matrix* $\boldsymbol{R}_i$, *then* $V_i = span\{\boldsymbol{u}_1^i, \boldsymbol{u}_2^i, \ldots, \boldsymbol{u}_N^i\}$.

Proof. It is known that $Null\,(\boldsymbol{R}_i) = span\{\boldsymbol{u}_{N+1}^i, \boldsymbol{u}_{N+2}^i, \ldots, \boldsymbol{u}_n^i\}$ where $\boldsymbol{u}_k^i$, $k = N+1, \ldots, n$ are the eigenvectors corresponding to the zero eigenvalues. From Theorem 2, $Null\,(\boldsymbol{R}_i) = V_i^\perp$, then $V_i^\perp = span$

Please cite this article as: M. Koç, A. Barkana, Application of Linear Regression Classification to low-dimensional datasets, Neurocomputing (2013), http://dx.doi.org/10.1016/j.neucom.2013.10.009