Pattern Recognition Letters 89 (2017) 46–52

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Nonparametric kernel sparse representation-based classifier

Alireza Esmaeilzehi[a], Hamid Abrishami Moghaddam[a,b,*]

[a] Faculty of Electrical Engineering, K.N. Toosi University of Technology, Tehran, Iran
[b] GRAMFC Unité de Génie Biophysique et Médical, Faculté de Médecine, Université de Picardie Jules Verne, AMIENS, France

ARTICLE INFO

Article history:
Received 2 June 2016
Available online 16 February 2017

Keywords:
Sparse representation-based classifier
Compressive sensing
Kernel method
Nonparametric discriminant analysis
Face recognition

ABSTRACT

Sparse representation-based classifier (SRC) and kernel sparse representation-based classifier (KSRC) are founded on combining pattern recognition and compressive sensing methods and provide acceptable results in many machine learning problems. Nevertheless, these classifiers suffer from some shortcomings. For instance, SRC's accuracy drops against samples from same directional classes or KSRC's output declines when data is not normally distributed in kernel space. This paper introduces nonparametric kernel sparse representation-based classifier (NKSRC) as a generalized framework for SRC and KSRC. First, it applies kernel on samples to overcome data directionality and then employs nonparametric discriminant analysis (NDA) to reduce data dimensionality in kernel space alleviating concern about data distribution type. The experimental results of NKSRC demonstrate its superiority over SRC and KSRC–LDA and its equal or superior performance with respect to KSRC–PCA on some synthetic, four well-known face recognition and several UCI datasets.

1. Introduction

One of the most recent classifiers which provides appropriate performance in various fields of pattern recognition problems is sparse representation-based classifier (SRC) [23]. SRC was founded on compressive sensing (CS) theory. CS tries to find the sparsest solution of an underdetermined linear equations system which has infinite answers. CS procedure utilizes L_0-norm minimization to find the unique sparse solution of equations system [5,1]. Since L_0-norm function is discontinuous; the minimization process confronts some obstacles and often may be infeasible. In this situation, the approximations like L_1-norm function which results also in sparse solution and is appropriate for signal recovery could be used as an alternative for L_0-norm function. Some cost functions like basis pursuit (BP) [8] or least absolute shrinkage and selection operator (LASSO) [22] use this approximation to reach the sparsest solution. SRC has found various applications in which non-stationary signals were investigated. In EEG-based brain computer interfaces, SRC-based methods provided better performance with respect to the other approaches from both classification and robustness points of view [20,21].

Moreover, in different disease assessment problems like tongue geometric feature analysis [24], SRC was utilized for distinguishing between healthy and disease patterns.

In spite of remarkable performance of SRC or its extended versions like extended SRC [9] in some face recognition databases [23], its shortcomings in some pattern recognition problems, in particular those in which samples come from same directional classes are noticeable [25]. When distributions of classes are in the same direction, some linear relationships can be found between samples from different classes. In other words, a test sample may not only be represented as a linear combination of some samples from its own class, but also it can be sparsely represented by linear combination of some samples from other classes. In the latter case, the residuals of other classes become significant and SRC misclassifies the corresponding test sample. One of the inventive manners for enhancing SRC's performance in this situation is utilization of kernel space [16]. Accordingly, samples are carried to kernel space and due to kernel's nonlinearity, the classes lose their directionality. Samples in kernel space are often of high or even infinite dimensionality and L_1-minimization in this space is impractical. One of the most attractive solutions is adding a dimensionality reduction (DR) stage in kernel space before optimizing the corresponding cost function. This procedure was introduced as kernel sparse representation-based classifier (KSRC) by Zhang et al. [25]. This approach uses reputed DR methods like linear discriminant analysis (LDA) [13], principal component analysis (PCA) [14] and random projection (RP) [4].

When kernel is applied on samples, there is no guarantee that the distributions of primary dataset are preserved and if the data has not Gaussian distribution in kernel space, parametric DR approaches like LDA or PCA would be compromised theoretically. On

* Corresponding author.
E-mail addresses: a.esmailzehi@ee.kntu.ac.ir (A. Esmaeilzehi), moghaddam@kntu.ac.ir (H. Abrishami Moghaddam).

http://dx.doi.org/10.1016/j.patrec.2017.02.007
0167-8655/

the other hand, RP is not the most accurate way for DR in every situation. Thus, other DR methods that do not rely on data distributions are preferable. One of the remarkable procedures that is less dependent on distribution of samples, is nonparametric discriminant analysis (NDA) [3,12].

This paper proposes the nonparametric kernel sparse representation-based classifier (NKSRC) approach to alleviate kernel space distribution dependency of KSRC. The new approach is demonstrated to outperform its predecessors in terms of classification accuracy in several pattern recognition datasets.

In Section 2, some related works are reviewed. In Section 3, NKSRC method is introduced. Section 4 includes experimental results of NKSRC method on some artificial datasets, four face recognition databases and also some datasets from UCI machine learning repository [11]. Finally, Section 5 concludes from the results of the paper.

2. Related works

2.1. Sparse representation-based classifier

SRC tries to represent a test sample as a linear combination of the training samples from one class. Let the training set matrix be $\mathbf{X} \in R^{m \times n}$ and include samples from c classes with m dimensions. For a test vector $\mathbf{x} \in R^m$, the SRC problem is defined as follows:

$$\min_{\boldsymbol{\alpha}} \|\boldsymbol{\alpha}\|_1 \text{ s.t. } \mathbf{x} = \mathbf{X}\boldsymbol{\alpha} \tag{1}$$

where $\boldsymbol{\alpha} \in R^n$ is a sparse vector that contains only nonzero elements corresponding to the test sample's class. By computing class' residuals for the test sample and finding minimum of them, the label of the test data is determined

$$\hat{y} = \arg \min_{i=1,\ldots,c} r_i(\mathbf{x}) = \left\|\mathbf{x} - \mathbf{X}\boldsymbol{\delta}_i\right\|_2 \tag{2}$$

where $\boldsymbol{\delta}_i = [\boldsymbol{\delta}_i(\boldsymbol{\alpha}_1), \boldsymbol{\delta}_i(\boldsymbol{\alpha}_2), \ldots, \boldsymbol{\delta}_i(\boldsymbol{\alpha}_n)]^t$ is the selection function for the i-th class in which

$$\boldsymbol{\delta}_i(\boldsymbol{\alpha}_j) = \begin{cases} \boldsymbol{\alpha}_j & y_j = i \\ 0 & o.w. \end{cases} \tag{3}$$

In the above equation, $y_j \in \{1, 2, \ldots, c\}$ is the class label corresponding to the j-th element in the sparse vector. However, in many practical problems the test sample is polluted with bounded energy noise and in real world situation, the SRC constraint alters as

$$\min_{\boldsymbol{\alpha}} \|\boldsymbol{\alpha}\|_1 \text{ s.t.} \|\mathbf{x} - \mathbf{X}\boldsymbol{\alpha}\|_2 \leq \varepsilon \tag{4}$$

where ε is a tiny positive number representing the noise energy. SRC exhibited phenomenal performance particularly in some face recognition problems [23]. Nevertheless, its accuracy declines when the training samples come from same directional classes. Indeed, when the means of more than two classes are laid on a straight line, they are considered as directional. In such circumstance, SRC may represent a test sample not only as a linear combination of constituents from its own class, but also from other directional classes. Obviously, this representation lacks the crucial sparsity property and consequently degrades the classification accuracy. Therefore, other tricks must be added to SRC to improve its yield.

2.2. Kernel sparse representation-based classifier

In KSRC, at first we apply a kernel function to the dataset and as a result samples are taken to a new space in which they are not in the same direction. Consequently, SRC could be utilized in the kernel space. The dimensionality of training and test samples in the new space relies on the type of kernel function. In addition, the kernel function must satisfy Mercer's conditions like continuity, symmetry and being positive semidefinite.

Some popular kernels that meet Mercer's condition are sigmoid kernel, polynomial kernel and Gaussian radial basis function (RBF). Each of them transfers sample to a specific higher dimensional space which is determined by the form of the corresponding kernel function. For instance, RBF kernel is defined as

$$K(\mathbf{x}, \mathbf{y}) = e^{-\gamma(\|\mathbf{x}-\mathbf{y}\|_2^2)} \tag{5}$$

where γ is a positive constant called RBF kernel parameter. Moreover, it can be proved that RBF kernel takes every sample to an infinite dimensional space. This issue is not practical in investigations of different classifiers. Now problem with the bounded energy noise on test sample can be expressed as

$$\min_{\boldsymbol{\alpha}} \|\boldsymbol{\alpha}_1\| \text{ s.t.} \left\|\boldsymbol{\phi}(\mathbf{x}) - \boldsymbol{\Phi}\boldsymbol{\alpha}\right\|_2 \leq \varepsilon \tag{6}$$

where $\boldsymbol{\phi}(\mathbf{x}) \in R^D$ is the test sample in kernel space and $\boldsymbol{\Phi} \in R^{D \times n}$ is a training dictionary which is obtained by applying kernel on $\mathbf{X}$. However, in many cases after applying kernel, the dimensionality of kernel space (D) is much greater than sample's original dimensionality (m) which leads to time consuming procedure or even infeasible optimization. To overcome this inconvenient, a dimensionality reduction step must be added which converts the problem to a feasible optimization. The ultimate form of KSRC can be formulated as

$$\min_{\boldsymbol{\alpha}} \|\boldsymbol{\alpha}\|_1 \text{ s.t.} \left\|\mathbf{P}^t\boldsymbol{\phi}(\mathbf{x}) - \mathbf{P}^t\boldsymbol{\Phi}\boldsymbol{\alpha}\right\|_2 \leq \varepsilon \tag{7}$$

where t denotes the transpose of matrix and $\mathbf{P} \in R^{D \times d}$ ($d \ll D$) is one of LDA or PCA transforms. However, in general after applying kernel, the distribution of samples in kernel space differs from their distribution in the initial space and this matter may require employing other DR approaches that are less dependent on samples' distribution.

3. Nonparametric kernel sparse representation–based classifier

In NKSRC algorithm, at first the data is transferred to a higher dimensional space using a kernel function which satisfies the Mercer's theorem. If $\boldsymbol{\phi}(\mathbf{x})$ and $\boldsymbol{\Phi}$ are the test sample and training dictionary in kernel space respectively, then the optimization problem in (4) will be rewritten as:

$$\min_{\boldsymbol{\alpha}} \|\boldsymbol{\alpha}\|_1 \text{ s.t.} \left\|\boldsymbol{\phi}(\mathbf{x}) - \boldsymbol{\Phi}\alpha\right\|_2 \leq \varepsilon \tag{8}$$

$$\boldsymbol{\phi}^t(\mathbf{x})\boldsymbol{\phi}(\mathbf{x}) = K(\mathbf{x}, \mathbf{x}) \tag{9}$$

where K is the kernel function. In general, kernel feature space is high (possibly infinite) dimensional. Now, NDA is applied to reduce the number of equations for optimization. Let $\mathbf{T} \in R^{d \times D}$ be the NDA transform for DR in which d and D represent the reduced and kernel space dimensionality, respectively. Consequently, the NKSRC problem could be shown as:

$$\min_{\boldsymbol{\alpha}} \|\boldsymbol{\alpha}\|_1 \text{ s.t.} \left\|\mathbf{T}\boldsymbol{\phi}(\mathbf{x}) - \mathbf{T}\boldsymbol{\Phi}\boldsymbol{\alpha}\right\|_2 \leq \varepsilon \tag{10}$$

In NDA, the transformation matrix $\mathbf{T}$ is obtained from linear combination of training samples in kernel space. Zhang et al. [25] showed that when projection matrix incorporates linear combination of training samples and $D \gg n$, (10) can be solved in n dimensional space using kernel function. Here, we are inclined to show how $\mathbf{T}$ is obtained from training samples and also how it is not dependent on sample's distributions. In contrast to parametric DR transforms, $\mathbf{T}$ which is obtained from NDA is not founded on Gaussian assumption of data. Transformation matrix $\mathbf{T}$ uses nearest neighbor (NN) [15] to define within and between class scatter matrices and then optimizes the cost function for maximizing the