Pattern Recognition 45 (2012) 1104–1118

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

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

Beyond sparsity: The role of L_1-optimizer in pattern classification

Jian Yang [a,b,*], Lei Zhang [c], Yong Xu [d], Jing-yu Yang [a]

[a] Department of Computer Science, Nanjing University of Science and Technology, Nanjing 210094, PR China
[b] Computation and Neural Systems, California Institute of Technology, Pasadena, CA 91125, USA
[c] Biometric Research Centre, Department of Computing, Hong Kong Polytechnic University, Kowloon, Hong Kong
[d] Bio-Computing Research Centre, Shenzhen Graduate School of Harbin Institute of Technology, Shenzhen, China

ARTICLE INFO

Article history:
Received 28 September 2010
Received in revised form
23 July 2011
Accepted 22 August 2011
Available online 30 August 2011

Keywords:
Sparse representation
Pattern classification
Classifier
Feature extraction

ABSTRACT

The newly-emerging sparse representation-based classifier (SRC) shows great potential for pattern classification but lacks theoretical justification. This paper gives an insight into SRC and seeks reasonable supports for its effectiveness. SRC uses L_1-optimizer instead of L_0-optimizer on account of computational convenience and efficiency. We re-examine the role of L_1-optimizer and find that for pattern recognition tasks, L_1-optimizer provides more classification meaningful information than L_0-optimizer does. L_0-optimizer can achieve sparsity only, whereas L_1-optimizer can achieve closeness as well as sparsity. Sparsity determines a small number of nonzero representation coefficients, while closeness makes the nonzero representation coefficients concentrate on the training samples with the same class label as the given test sample. Thus, it is closeness that guarantees the effectiveness of the L_1-optimizer based SRC. Based on the closeness prior, we further propose two kinds of class L_1-optimizer classifiers (CL_1C), the closeness rule based CL_1C (C-CL_1C) and its improved version: the Lasso rule based CL_1C (L-CL_1C). The proposed classifiers are evaluated on five databases and the experimental results demonstrate advantages of the proposed classifiers over SRC in classification performance and computational efficiency for large sample size problems.

1. Introduction

"Sparse (or sparsity)" becomes a popular term in neuroscience, information theory and signal processing and related areas in the past decade [1–10]. Vinje and Gallant's studies suggested that primary visual cortex (area V1) uses a sparse code to efficiently represent natural scenes. The receptive fields function forms a sparse representation of the visual world during natural vision [1]. Olshausen and Field [2] and Serre [3] revealed that the firing of the neurons with respect to a given input image is typically highly sparse if these neurons are viewed as an overcomplete dictionary of base signal elements at each visual stage. All of these findings form a physiological basis for sparse coding and sparse representation.

Sparse coding and sparse representation has recently aroused intensive interest pattern recognition and computer vision area. Labusch et al. [11] presented a simple sparse-coding strategy for digit recognition and achieved state-of-the-art results on the MNIST benchmark. Zhou et al. [12] presented a sparse principal component analysis (SPCA), which uses the Lasso (elastic net) to produce modified principal components with sparse loadings and yields encouraging results for regular multivariate data and gene expression arrays. Subsequently, different formulations of SPCA and sparse linear discriminant analysis have been developed [13–15]. Cai et al. [16] suggested a sparse projection over graph and showed its power for document classification. Qiao et al. [17] put forward a sparse preserving projection technique and demonstrated its effectiveness for face recognition. Actually, Qiao et al.'s sparse preserving projection can be viewed as a special case of L_1-graph under a general dimensionality reduction framework [18–20]. Recently, Wright et al. presented a sparse representation based classification method and successfully applied it to recognize human faces with varying lighting condition, occlusion and disguise [21]. In addition, Wright et al. [20] reviewed other sparse representation methods that were applied to different vision tasks such as image super-resolution [22], image denoising and inpainting [23], signal and image classification [24–27], etc. In most of these applications, using sparsity as a prior leads to state-of-the-art results.

This paper focuses on sparse representation based classification. The basic idea of Wright et al.'s sparse representation based classification (SRC) method is to represent a given test sample as a sparse linear combination of all training samples; the sparse nonzero representation coefficients are supposed to concentrate on the training samples with the same class label as the test sample. The sparsest solution can be sought by solving the L_0-optimization problem. However, solving L_0-optimization problem is NP hard

* Corresponding author at: Department of Computer Science, Nanjing University of Science and Technology, Nanjing 210094, PR China.
E-mail addresses: csjyang@mail.njust.edu.cn, jianyang@caltech.edu (J. Yang), cslzhang@comp.polyu.edu.hk (L. Zhang), laterfall2@yahoo.com.cn (Y. Xu), yangjy@mail.njust.edu.cn (J.-y. Yang).

0031-3203/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2011.08.022

and even difficult to approximate [28]. Recent development in the emerging theory of sparse representation and compressed sensing [29,30,5] reveals that finding the solution of the L_0 optimization problem is equivalent to finding the solution of the L_1 optimization problem for certain dictionaries. The L_1-optimizer is therefore used instead of the L_0-optimizer in SRC.

Regarding SRC, a fundamental problem is: when one uses all classes of training samples to represent a given test sample, why does the small number of nonzero representation coefficients concentrate on the homo-class training samples? Wright and Ma [31] and Wright et al. [20] addressed the extended L_1-minimization model based error correction problem and interpreted why accurate recovery of sparse signals is possible even if the corruption error is almost dense. But the fundamental problem mentioned remains open, just as said in [20] "—the striking discriminative power of the sparse representation still lacks rigorous mathematical justification". In this paper, our intention is to seek some reasonable supports for SRC.

We begin with an example of the two-class handwritten numerical recognition problem in which the L_0-solution fails while the L_1-Solution succeeds for classification. This fact indicates that the sparest representation gained by the L_0-optimizer is not sufficient for classification. Conversely, the L_1-optimizer may not achieve the sparest solution, but achieves the meaningful solution for correct classification. We then introduce the closeness theory to reveal the connection of the L_1-solution to classification. The L_1-norm of nonzero weights can provide a metric to measure the degree of closeness between the testing sample and its support training samples, while the L_0-norm cannot. The effectiveness of SRC is due to the closeness prior: the homo-class representation leads to the minimal L_1-norm of nonzero weights. The physical meaning of minimizing L_1-norm of weights becomes clearer if a weight-sum-to-one constraint is imposed onto the L_1-optimizer, i.e., searching for the support training samples such that their centroid is closest to the given test sample in the sense of L_1-norm.

We further introduce the theory of (global) neighborliness and local neighborliness of quotient polytope associated with a dictionary, and use it to in-depth analyze the role of L_1-optimizer in pattern recognition. In global neighborliness cases where the quotient polytope associated with the dictionary formed by all training samples is t-neighborly, L_1-optimizer achieves both sparsity and closeness globally. In such cases, L_1-solution equals to L_0-solution, i.e., the globally sparsest solution. This sparsest solution determines the set of support training samples that is closest to the given testing sample. In local neighborliness cases where the quotient polytope associated with the dictionary formed by class training samples is t-neighborly, L_1-optimizer achieves sparsity locally and closeness globally. In such cases, L_1-solution is a locally sparse solution, possibly not the globally sparsest solution, but it is the solution which is most meaningful for classification. Beyond neighborliness, the degree of sparsity of L_1-solution cannot be guaranteed, but its effectiveness for classification can still be guaranteed, i.e., the L_1-solution determines the set of support training samples that is closest to the given testing sample.

Based on the closeness analysis, we present two class L_1-optimizer classifiers (CL_1C). To this end, we first provide theoretical, geometrical and computational justifications for supporting the class training samples based representation. We then present the closeness rule based CL_1C (C-CL_1C), which uses the *closeness* (i.e., the L_1-norm of the representation coefficients) as a criterion to make a decision. A normalized version of C-CL_1C is obtained based on geometrical meaning of the solution of the constrained L_1-optimizer. To overcome the limitation of C-CL_1C, which restricts the testing sample to lie on faces of the class polytopes and only suits for large sample size problems, we further present the Lasso rule based CL_1C (L-CL_1C) and its normalized version. To test the proposed classifiers, we finally use four databases which involve different recognition tasks: the AR database for gender recognition, the CENPARMI database for handwritten numeral Recognition, the NUST603 database for handwritten Chinese character recognition, the Extended Yale B database for face recognition. The experimental results demonstrate the effectiveness of the proposed classifiers.

2. Outline of sparse representation-based classifier

Suppose there are c known pattern classes. Let $\mathbf{A}_i$ be the matrix formed by the training samples of Class i, i.e., $\mathbf{A}_i=[\mathbf{y}_{i1},\mathbf{y}_{i2},\ldots,\mathbf{y}_{iM_i}]\in R^{N\times M}{}_i$, where M_i is the number of training samples of Class i. Let us define a matrix $\mathbf{A}=[\mathbf{A}_1,\mathbf{A}_2,\ldots,\mathbf{A}_c]\in R^{N\times M}$, where $M=\sum_{i=1}^{c}M_i$. The matrix $\mathbf{A}$ is obviously composed of entire training samples.

Given a test sample $\mathbf{y}$, we represent $\mathbf{y}$ in a overcomplete dictionary whose basis vectors are training sample themselves, i.e., $\mathbf{y}=\mathbf{A}\mathbf{w}$. This system of linear equation is underdetermined if $N<M$. The idea of sparse representation based classification is motivated by the following observation: a valid test sample $\mathbf{y}$ can be sufficiently represented using only the training samples from the same class. The representation is naturally sparse if training sample size is large enough. The sparser the recovered representation coefficient vector $\mathbf{w}$ is, the easier it will be to accurately determine the identity of the test sample $\mathbf{y}$ [21].

The sparsest solution to $\mathbf{y}=\mathbf{A}\mathbf{w}$ can be sought by solving the following optimization problem:

$$(L_0)\ \hat{\mathbf{w}}_0=\arg\min\|\mathbf{w}\|_0,\ \text{subject to}\ \mathbf{A}\mathbf{w}=\mathbf{y}, \tag{1}$$

where $\|\cdot\|_0$ denotes the L_0-norm, which counts the number of nonzero entries in a vector.

Solving L_0 optimization problem in Eq. (1), however, is NP hard and extremely time-consuming. Fortunately, recent research efforts reveal that for certain dictionaries, if the solution $\hat{\mathbf{w}}_0$ is spare enough, finding the solution of the L_0 optimization problem is equivalent to finding the solution to the following L_1 optimization problem [5,29,30]:

$$(L_1)\ \hat{\mathbf{w}}_1=\arg\min\|\mathbf{w}\|_1,\ \text{subject to}\ \mathbf{A}\mathbf{w}=\mathbf{y}. \tag{2}$$

This problem can be solved in polynomial time by standard linear programming algorithms [33]. A more efficient algorithm, e.g., the homotopy algorithm which has a computational complexity that is linear to the size of the training set, is available recently [34].

After obtaining the sparsest solution $\hat{\mathbf{w}}_1$, we can design a sparse representation based classifier (SRC) in terms of the class reconstruction residual. Specifically, for Class i, let $\delta_i: R^N\to R^N$ be the characteristic function that selects the coefficients associated with the ith class. For $\mathbf{w}\in R^N$, $\delta_i(\mathbf{w})$ is a vector whose only nonzero entries are the entries in $\mathbf{w}$ that are associated with Class i. Using only the coefficients associated with the ith class, one can reconstruct a given test sample $\mathbf{y}$ as $\hat{\mathbf{y}}_i=\mathbf{A}\delta_i(\hat{\mathbf{w}}_1)$. The corresponding class reconstruction residual is defined by

$$r_i(\mathbf{y})=\|\mathbf{y}-\hat{\mathbf{y}}_i\|_2=\|\mathbf{y}-\mathbf{A}\delta_i(\hat{\mathbf{w}}_1)\|_2. \tag{3}$$

The SRC decision rule is: if $r_l(\mathbf{y})=\min_i r_i(\mathbf{y})$, $\mathbf{y}$ is assigned to Class l.

For convenience, the training samples (or basis vectors) associated with nonzero representation coefficients are called the *support training samples* (or support basis vectors) in the remainder of the paper, which is in spirit consistent with the concept of support vectors in support vector machine (SVM) literature [32].