A Similar K-SVD Optimization Algorithm

Generalizing the K-Means and the Bayesian tracking

Renjie WU
Graduate Schol of Information, Produvtion and systems
Waseda Univercity
Kitakyushu-shi, Japan
Wurj-sjtu-waseda@ruri.waseda.jp

Prof S.Kamata
Graduate Schol of Information, Produvtion and systems
Waseda Univercity
Kitakyushu-shi, Japan
kam@waseda.jp

***Abstract*—The last decade have seen tremendous improvement in the development of new image information processing and computational tools based on sparse representation. Today, in the information sciences, computer vision and image processing, the development of sparse representation algorithms led to convenient tools to transient compressed image (data) rapidly, to remove noise from image, and to get the super-resolution image. In the study of sparse representation of images, overcomplete dictionary is used. It contains prototype image-atoms. In this way, the images are described by sparse linear combinations of theses atoms. In this field has concentrated mainly on the design of a better dictionary. The generalized K-Means algorithm (K-SVD) [1] taught us a very good case. This paper has proposed an optimization algorithm adopting the Bayesian tracking and K-SVD analysis method. We analyze this algorithm and demonstrate its results on image data.**

Keywords-Sparese Repressentation; Bayesina Prior; K-SVD; Atom decomposition, Dictionaty.

I. Introduction

A. Sparse Representation Modeling

Consider a matrix $D \in R^{n \times K}$ with $n < K$, and define the underdetermined linear system of equations $Dx = y$. This system has more unknowns than equations, and thus it has either no solution, if b is not in the span of the matrix A, or infinitely many solutions.

In order to solve a problem that has an infinite number of solutions, we shall hereafter assume that D is a full-rank matrix, implying that its columns span the entire R^n.

Defining the general optimization problem (P_J):

$$(P_J): \min_x J(x) \quad s.t. \quad y = Dx. \tag{1}$$

A signal $y \in R^n$ can be represented as a sparse linear combination of atoms $d_j \in D, j = 1, 2, ..., K$. The vector $x \in R^K$ contains the representation coefficients of the signal y.

In engineering, this sparsest representation is the solution of either [2]:

$$(P_0): \min_x \|x\|_0 \quad s.t. \quad y = Dx. \tag{2}$$

Or, we can rewrite the relaxed equality:

$$(P_{0,\varepsilon}): \min_x \|x\|_0 \quad s.t. \quad \|y - Dx\|_2^2 \le \varepsilon. \tag{3}$$

From a linear combination angle, the optimization model:

$$(P_{0,\tau_0}): \min_x \|y - Dx\|_2^2 \quad s.t. \quad \|x\|_0 \le \tau_0. \tag{4}$$

In this paper, we use (4) to analyze optimization problem.

B. Overcomplete Dictionary

A traditional overcomplete dictionary (choosing pre-constructed dictionaries, such as DCT, Wavelets [3], contourlets, curvelets [4], and more) are typically limited in their ability to handle the signals. Furthermore, most of those dictionaries are restricted to signals of a certain type (particular to stylized 'cartoon-like' image content, clear edge and texture structure). Therefore we need an approach to optimize a dictionary that overcomes these limitations. Normally, we use sparse solution to update the overcomplete dictionary.

In this paper, we consider a overcomplete dictionary that each atom is updated individually based on Bayesian prior. The proposed algorithm can converge faster than the K-SVD algorithm.

II. Generalizing the K-Means (K-SVD) Alogonithm

A. The K-Means Algorithm

K-Means is an algorithm of cluster analysis. We defined N numbers of signals $Y = \{y_i\}_{i=1}^N, N >> K$, a code-book $D = \{d_j\}_{j=1}^K$.

In vector quantization (VQ) [5], if D is preset, then, for each signal

© the authors

$$\forall k \neq j, \left\| y_i - De_j \right\|_2^2 \leq \left\| y_i - De_K \right\|_2^2 . \quad (5)$$

If only one atom is allowed in the signal decomposition and furthermore, the coefficient multiplying it must be one. There is a variant of the VQ coding method.

In this case, the overall MSE of the Y can be written as

$$E^2 = \left\| Y - DX \right\|_2^2 . \quad (6)$$

The objective function:

$$\min_{D,X} \left\{ \left\| Y - DX \right\|_2^2 \right\} \quad s.t. \quad \forall i, x_i = e_k . \quad (7)$$

Where, vector $e_i = (\underbrace{0,0,...,0,1}_{i},\underbrace{0,...,0,0}_{K-i})$.

The K-means applies two steps at the process of each iteration:

- Assignment step: calculate the X
- Update step: calculate the D

B. The K-SVD Algorithm

We are liable to come to K-SVD modeling via K-Means Algorithm [1].

Let overcomplete $D = \left\{ d_j \right\}_{j=1}^{K} \in R^{n \times K}$,

signals $Y = \left\{ y_i \right\}_{i=1}^{N}$ and sparse solutions $X = \left\{ X_i \right\}_{i=1}^{N}$

From (4), we can get an optimization problem:

$$\min_{D,X} \left\{ \left\| Y - DX \right\|_2^2 \right\} \quad s.t. \quad \forall i, \left\| x_i \right\|_1 \leq \tau_0 . \quad (8)$$

In K-SVD algorithm, first it fixes and aims to find the best coefficient matrix. In this case, an approximation pursuit method is used. As long as we predetermined number of nonzero entries τ_0, the approximate solution can be found.

Second, it needs to search for a better dictionary. This process updates one column at a time, fixing all columns in D except one d_k, and finding a new column d_k and new values for its coefficients that best reduce the MSE.

From (8), we isolate the dependency on d_k, the penalty term can be rewritten as

$$\left\| Y - DX \right\|_2^2 = \left\| Y - \sum_{j=1}^{K} d_j x_j^T \right\|_2^2 = \left\| \left(Y - \sum_{j \neq k} d_j x_j^T \right) - d_k x_k^T \right\|_2^2 . \quad (9)$$

Where, x_j^T stands for the j-th row of X. the update targets both d_k and x_k^T, and refers to the term in parentheses,

$$E_k = Y - \sum_{j \neq k} d_j x_j^T . \quad (10)$$

as a known pre-computed error matrix.

The optimal d_k and x_k^T minimizing equation (9) are the rank-1 approximation of E_k, and can be obtained via an SVD, but this typically would yield a dense vector x_k^T, implying that we increase the number of non-zeros in X.

For this problem, define ω_k as the group of indices pointing to examples $\left\{ y_i \right\}$ that use the atom d_k, those where $x_k^T(i)$ is nonzero.

$$\omega_k = \left\{ i \,|\, 1 \leq i \leq K, x_k^T(i) \neq 0 \right\} . \quad (11)$$

Define a restriction operator Ω_k as a matrix of size $N \times |\omega_k|$, with ones on the $\left(\omega_k(i), i \right)$ th entries and zeros elsewhere.

Then, we may return to (9).

$$\left\| E_k \Omega_k - d_k x_k^T \Omega_k \right\|_2^2 . \quad (12)$$

Minimizing equation (9) is convergent.

Of course, also we can use the Least-Squares method to solving

$$\min_{x_k^T \Omega_k} \left\| E_k \Omega_k - d_k x_k^T \Omega_k \right\|_2^2 \Rightarrow \left(x_k^T \Omega_k \right)^T = \frac{\Omega_k^T E_k^T d_k}{\left\| d_k \right\|_2^2} . \quad (13)$$

Once updated, it is kept fixed, and we update d_k by

$$\min_{d_k} \left\| E_k \Omega_k - d_k x_k^T \Omega_k \right\|_2^2 \Rightarrow d_k = \frac{E_k \Omega_k \left(x_k^T \Omega_k \right)^T}{\left\| \left(x_k^T \Omega_k \right)^T \right\|_2^2} . \quad (14)$$

© the authors

III. OPTIMIZATION ALOGONITHM BASED ON BAYESIAN TRACKING

A. Detailed Description

We shall discuss the new optimization algorithm in detail. First, Let us revert to the core questions in dictionary-learning.

Let us consider the following objective function:

$$\min_{D,\{x_i\}_{i=1}^{N}} \sum_{i=1}^{N} \|y_i - Dx_i\|_2^2 \quad s.t. \quad \|x_i\|_0 \le \tau_0, 1 \le i \le N. \tag{15}$$

In terms of well posedness of the dictionary-learning problem presented above, a fundamental question is whether there is a uniqueness property underlying this problem.

In this paper, we will use a Bayesian approach to calculate the approximate solution.

Assuming the signal is defined as $E_k = Dx_k + \varepsilon$.

From (10) and (12), the relationship between the source signals and the dictionary is defined as

$$R_k = \langle E_k, d_k \rangle = x_k + \sum_{j \ne k} x_j b_{jk} + v_k. \tag{16}$$

There, $b_{jk} = \langle d_j, d_k \rangle$, $v_k = \langle \varepsilon, d_k \rangle$.

We know the estimated value of other coefficients (previous iteration). Then, (16) can be rewritten as

$$R_k - \sum_{j \ne k} \hat{x}_j b_{jk} = x_k + \sum_{j \ne k} \left(x_j - \hat{x}_j \right) b_{jk} + v_k. \tag{17}$$

Here, $\hat{x}_j$ is the estimate of j-th coefficient.

For convenience, definition of as follows:

$$m_k = \sum_{j \ne k} \hat{x}_j b_{jk}, \gamma_k = \sum_{j \ne k} \left(x_j - \hat{x}_j \right) b_{jk} + v_k. \tag{18}$$

Then, we can get the assumption about H_1 and H_2.

$$\begin{aligned} H_1 &: R_k - m_k = x_k + \gamma_k. \\ H_2 &: R_k - m_k = \gamma_k. \end{aligned} \tag{19}$$

$P(H_1 \mid R)$ and $P(H_2 \mid R)$ are effective posterior probability and invalid posterior probability of the k-th atom in the dictionary [6].

In this modeling, the value of spare coefficient effective probability 1-p and invalid probability p is defined. If the sparse coefficient errors $x_j - \hat{x}_j$ to meet the Gaussian distribution (variance is defined as $\delta_{j,\varepsilon_x}^2$), then γ_k is also a Gaussian distribution (variance is defined as $\delta_{\gamma_k}^2$). So, we can get the following formula:

$$\frac{1-p}{\sqrt{2\pi(\delta_{\gamma_k}^2 + \delta_{x_k}^2)}} \exp\left(\frac{-(R_k - m_k)}{2(\delta_{\gamma_k}^2 + \delta_{x_k}^2)} \right) > \frac{p}{\sqrt{2\pi\delta_{\gamma_k}^2}} \exp\left(\frac{-(C_k - m_k)}{2\delta_{\gamma_k}^2} \right). \tag{20}$$

There, sparse coefficient to meet the Gaussian distribution (variance is defined as $\delta_{x_k}^2$). p, $\delta_{\gamma_k}^2$ and $\delta_{x_k}^2$ is an unbiased estimate of the source signal.

By the (20), we can get the judgment rule of the hypothesis testing:

$$P(x_j \mid H_1) = |R_k - m_k| > Th_k, Th_k = \sqrt{2(\delta_{\gamma_k}^2 + \delta_{x_k}^2) \frac{\delta_{\gamma_k}^2}{\delta_{x_k}^2} \ln\left(\frac{p}{1-p} \cdot \sqrt{\frac{\delta_{\gamma_k}^2 + \delta_{x_k}^2}{\delta_{\gamma_k}^2}} \right)}. \tag{21}$$

We can get initial threshold and the final threshold as follows:

$$Th_k^{(0)} = Th\,|_{\delta_{\gamma_k}^{(0)}}, Th_k^{(\infty)} = Th\,|_{\delta_\gamma = \delta_e} \approx \delta_e \sqrt{2\ln\left(\frac{p}{1-p} \cdot \frac{\delta_\gamma}{\delta_e} \right)}. \tag{22}$$

From the (22), analysis shows that the threshold converges to $Th_k^{(\infty)}$.

B. Optimization Algorithm

Task: Train a dictionary D to sparsely represent the data $\{y_i\}_{i=1}^{N}$, by approximating the solution to the problem posed in (4).

Initialization: Initialize mainloop=0 and

Initialize Dictionary: Build $D_{(0)} \in R^{n \times K}$, either by using random entries, or using K randomly chosen examples.

Normalization: Normalize the columns of $D_{(0)}$

Main Iteration: Increment mainloop by 1, and apply

Sparse Coding Stage: Use a pursuit algorithm to approximate the solution of

$$\hat{x}_i = \arg\min_x \|y_i - D_{(mainloop-1)} x\|_2^2 \quad s.t. \quad \|x\|_0 \le \tau_0$$

Obtaining sparse representations $\hat{x}_i$ for $1 \le i \le N$.

These form the matrix $X_{(mainloop)}$.

Dictionary-Update stage: Use the following procedure to update the columns of the dictionary and

© the authors

obtain $D_{(mainloop)}$: Repeat for $k = 1, 2, ..., K$

1) Define the group of examples that use the atom d_k ,

$$\Omega_k = \{ i \mid 1 \le i \le K, X_{(mainloop)}[k,i] \ne 0 \}$$

2) Compute the residual matrix

$$E_k = Y - \sum_{j \ne k} d_j x_j^T$$

3) Restrict E_k by choosing only the columns corresponding to Ω_k

4) Under the judgment rule of the hypothesis testing, decrease $\delta^2_{j,\varepsilon_x}$, iterative calculation until the threshold to the minimum $Th_k^{(\infty)}$

Stopping Rule: if the change in $\left\| Y - D_{(mainloop)} X_{(mainloop)} \right\|_2^2$ is small enough, stop.

Otherwise, apply another iteration.

Output: the desired result is D

Figure 1. The Optimization Algorithm.

IV. EXAMPLES

We turn to present an elementary experiment performed on natural image data. We train a dictionary for sparsely representing patches of size 8×8 extracted from the image Barbara, shown in Fig. 2(a). We extract these patches to train on. The number of iterations of the K-SVD, proposed optimization algorithm was 50 and 20.

The truncated Bayesian prior process infers the subset of dictionary elements employed to represent the data.

A. Filling In Missing Pixels

35% pixels missing image and reconstructions are shown in Fig. 2. In this case, the algorithm gets stuck on a saddle-point steady-state solution.

(a) (b) (c)

Figure 2. Filling in missing pixels, (a) source image, (b) 35% pixels missing image, the PSNR is 31.23, (c) reconstructed results, the PSNR is 32.57

B. Reconstruction Error

Fig. 3 displays a comparison of compressive sensing (CS) Measurement when applying BP, OMP, K-SVD and proposed optimization algorithm on the Barbara image. The relative reconstruction errors were computed and are displayed.

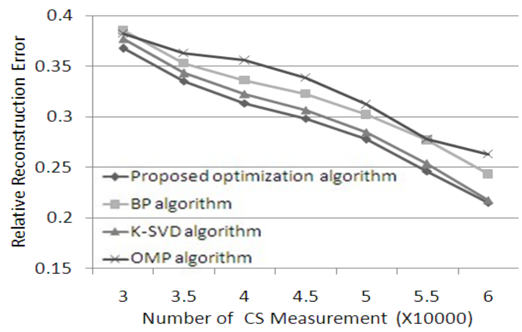

Figure 3. Reconstruction results.

V. CONCLUSION

In this paper, we presented an optimization algorithm based on Bayesian. It can significantly speed up convergence on the dictionary update stage. And it is concluded that more accurate atoms can be obtained. But, there are two problems cannot be solved. First we cannot guarantee that these algorithms obtain the global minimum of the penalty function posed in (4). Second, we cannot guarantee a monotonic nonincreasing prenalty value as a function of the iterations. The multi-scale analysis may be an effective way to solve these problems.

REFERENCES

[1] M. Aharon, M. Elad, and A. Bruckstein. “K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation,” IEEE Transactions on signal processing, 54(11):4311, 2006.

[2] Jean-Luc Starck, Michael Elad, and David L. Donoho,“Image decomposition via the combination of sparse representations and a variational approach,” IEEE Trans. Image Process., Feb. 2004

[3] R. Gastaud and J. L. Starck, “Dynamic range compression : A new method based on wavelet transform,” in Astron. Data Anal. Software Systems Conf., Strasbourg, 2003.

[4] J. L. Starck, E. J. Candes, and D. L. Donoho, “The curvelet transform for image denoising,” IEEE Trans. Image Process., vol. 11, pp. 670–684, 2002.

[5] A. Gersho and R. M. Gray, “Vector Quantization and Signal Compression”. Norwell, MA: Kluwer Academic, 1991.

[6] Z. Zhang, B.D. Rao. “Sparse signal recovery with temporally correlated source vectors using sparse Bayesian learning” [J]. IEEE Journal of Selected Topics in Signal Processing, 2011, 5(5): 912-926.

© the authors