A greedy pursuit approach to classification using multi-task multivariate sparse representations

Charles Jeon⋆, Umamahesh Srinivas†, Vishal Monga†
⋆ Department of Electrical and Systems Engineering, University of Pennsylvania, USA
† Department of Electrical Engineering, Pennsylvania State University, USA

July 2012

Abstract

In this report, we propose an extension of the well-known Simultaneous Orthogonal Matching Pursuit (SOMP) algorithm to solve a multi-task multivariate classification problem using sparse representations. This corresponds to the case where an event is described by multiple representations, and separate training dictionaries are designed for each such representation.

I. Notation

Let $\boldsymbol{y}_i \in \mathbb{R}^m, i = 1, \ldots, T$ be T different representations[1] of the same physical event, which is to be classified into one of C different classes. Let $\boldsymbol{Y} := [\boldsymbol{y}_1 \ \ldots \ \boldsymbol{y}_T] \in \mathbb{R}^{m \times T}$. Assuming n training samples/events in total, we design T dictionaries $\boldsymbol{D}_i \in \mathbb{R}^{m \times n}, i = 1, \ldots, T$, corresponding to the T representations. We define a new composite dictionary $\boldsymbol{D} := [\boldsymbol{D}_1 \ \ldots \ \boldsymbol{D}_T] \in \mathbb{R}^{m \times nT}$. Further, each dictionary $\boldsymbol{D}_i$ is represented as the concatenation of the sub-dictionaries from all classes corresponding to the i-th representation of the event:

$$\boldsymbol{D}_i := [\boldsymbol{D}_i^1 \ \boldsymbol{D}_i^2 \ \ldots \ \boldsymbol{D}_i^C], \tag{1}$$

where $\boldsymbol{D}_i^j$ represents the collection of training samples for representation i that belong to the j-th class. So, we have:

$$\boldsymbol{D} := [\boldsymbol{D}_1 \ \ldots \ \boldsymbol{D}_T] = [\boldsymbol{D}_1^1 \ \boldsymbol{D}_1^2 \ \ldots \ \boldsymbol{D}_1^C \quad \ldots \quad \boldsymbol{D}_T^1 \ \boldsymbol{D}_T^2 \ \ldots \ \boldsymbol{D}_T^C]. \tag{2}$$

An important assumption in designing $\boldsymbol{D}$ is that the k-th column from each of the dictionaries $\boldsymbol{D}_i, i = 1, \ldots, T$, taken together offer multiple representations of the k-th training sample/event.

II. Multi-task multivariate sparse representations

A test event $\boldsymbol{Y}$ can now be represented as a linear combination of training samples as follows:

$$\boldsymbol{Y} = [\boldsymbol{y}_1 \ \ldots \ \boldsymbol{y}_T] = \boldsymbol{D}\boldsymbol{S} = \left[\boldsymbol{D}_1^1 \ \boldsymbol{D}_1^2 \ \ldots \ \boldsymbol{D}_1^C \quad \ldots \quad \boldsymbol{D}_T^1 \ \boldsymbol{D}_T^2 \ \ldots \ \boldsymbol{D}_T^C\right] [\boldsymbol{\alpha}_1 \ \ldots \ \boldsymbol{\alpha}_T], \tag{3}$$

where the coefficient vectors $\boldsymbol{\alpha}_i \in \mathbb{R}^{nT}, i = 1, \ldots, T$, and $\boldsymbol{S} = [\boldsymbol{\alpha}_1 \ \ldots \ \boldsymbol{\alpha}_T] \in \mathbb{R}^{nT \times T}$.

We examine the structure of the coefficient matrix $\boldsymbol{S}$ and make some crucial observations.

- It is reasonable to assume that the i-th representation of the test event (i.e. $\boldsymbol{y}_i$) can be approximately represented by the linear span of the training samples belonging to the i-th representation alone (i.e. only those training samples in $\boldsymbol{D}_i$). So the columns of $\boldsymbol{S}$ have the following structure: each vector $\boldsymbol{\alpha}_i$ has non-zero coefficients only in the locations corresponding to the columns of $\boldsymbol{D}_i$ and has zeros elsewhere. As a result, $\boldsymbol{S}$ exhibits block-diagonal structure.
- Each representation $\boldsymbol{y}_i$ of the test event is a *sparse* linear combination of the training samples in $\boldsymbol{D}_i$. Suppose the event belongs to class $c \in \{1, \ldots, C\}$; then only those coefficients in $\boldsymbol{\alpha}_i$ that correspond to $\boldsymbol{D}_i^c$ are expected to be non-zero.
- Furthermore, the non-zero weights of training samples in the linear combination exhibit one-to-one correspondence across representations. If the j-th training sample from the c-th class in $\boldsymbol{D}_1$ has a non-zero contribution to $\boldsymbol{y}_1$, then for all $i = 2, \ldots, T$, $\boldsymbol{y}_i$ has non-zero contributions from the j-th training sample of the c-th class in $\boldsymbol{D}_i$.

This suggests a joint sparsity model similar to the model introduced in [1]. However, the multi-task nature of the problem with different dictionaries $\boldsymbol{D}_i$ does not permit us to apply the SOMP algorithm from [1] directly. Since $\boldsymbol{S}$ obeys column correspondence, we introduce a new matrix $\boldsymbol{S}' \in \mathbb{R}^{n \times T}$ as the transformation of $\boldsymbol{S}$ with the zero coefficients removed,

$$\boldsymbol{S}' = \begin{bmatrix} \boldsymbol{\alpha}_1^1 & \ldots & \boldsymbol{\alpha}_i^1 & & \boldsymbol{\alpha}_T^1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{\alpha}_1^C & \ldots & \boldsymbol{\alpha}_i^C & \ldots & \boldsymbol{\alpha}_T^C \end{bmatrix},$$

where $\boldsymbol{\alpha}_i^j$ refers to the sub-vector extracted from $\boldsymbol{\alpha}_i$ that corresponds to coefficients from the j-th class. Note that, in the i-th column of $\boldsymbol{S}'$, only the coefficients corresponding to $\boldsymbol{D}_i$ are retained (for $i = 1, \ldots, T$).

[1] The term *multi-task* is used to refer to these multiple representations in some application domains such as heterogeneous sensor fusion.

We can now apply row-sparsity constraints similar to the approach in [1]. Our modified optimization problem becomes:

$$\hat{\boldsymbol{S}}' = \arg\min_{\boldsymbol{S}'} \|\boldsymbol{S}'\|_{\text{row}-0} \quad \text{subject to} \quad \|\boldsymbol{Y} - \boldsymbol{D}\boldsymbol{S}\|_F \leq \epsilon, \tag{4}$$

for some tolerance $\epsilon > 0$. We minimize the number of non-zero rows, while the constraint guarantees a good approximation.

The matrix $\boldsymbol{S}$ can be transformed into $\boldsymbol{S}'$ by introducing matrices $\boldsymbol{H} \in \mathbb{R}^{nT \times T}$ and $\boldsymbol{J} \in \mathbb{R}^{n \times nT}$,

$$\boldsymbol{H} = \text{diag}\left[\mathbf{1}\ \mathbf{1}\ \ldots\ \mathbf{1}\right], \boldsymbol{J} = \left[\boldsymbol{I}_n\ \boldsymbol{I}_n\ \ldots\ \boldsymbol{I}_n\right],$$

where $\mathbf{1} \in \mathbb{R}^n$ is the vector of all ones, and $\boldsymbol{I}_n$ denotes the n-dimensional identity matrix. Finally, we obtain $\boldsymbol{S}' = \boldsymbol{J}\left(\boldsymbol{H} \circ \boldsymbol{S}\right)$, where $\circ$ denotes the Hadamard product, $\left(\boldsymbol{H} \circ \boldsymbol{S}\right)_{ij} \triangleq h_{ij}s_{ij}$ for all i, j.

III. Extension of SOMP for multi-task multivariate sparse representations

Eq. (4) represents a hard optimization problem due to presence of the non-invertible transformation from $\boldsymbol{S}$ to $\boldsymbol{S}'$. We bypass this difficulty by proposing a modified version of the SOMP algorithm for the multi-task multivariate case.

Recall that the original SOMP algorithm gives K distinct atoms (assuming K iterations) from a dictionary $\boldsymbol{D}$ that best represent the data matrix $\boldsymbol{Y}$. In every iteration k, SOMP measures the residual for each atom in $\boldsymbol{D}$ and creates an orthogonal projection with maximal correlation. Extending this to the multi-task setting, for every representation $i, i = 1, \ldots, T$, we can identify the index set that gives the highest correlation with the residual at the k-th iteration as follows:

$$\lambda_{i,k} = \arg\max_{j=1,\ldots,n} \sum_{q=1}^{T} w_q \left\|\boldsymbol{R}_{k-1}^t \boldsymbol{d}_{q,j}\right\|_p, p \geq 1,$$

where w_q denotes the weight (confidence) assigned to the q-th representation, $\boldsymbol{d}_{q,j}$ represents the j-th column of $\boldsymbol{D}_q, q = 1, \ldots, T$, and the superscript $(\cdot)^t$ indicates the matrix transcript operator. After finding $\lambda_{i,k}$, we modify the index set to:

$$\Lambda_{i,k} = \Lambda_{i,k-1} \bigcup \lambda_{i,k}, i = 1, \ldots, T.$$

Thus, by finding the index set for the T distinct representations, we can create an orthogonal projection with each of the atoms in their corresponding representations. The algorithm is summarized below in Algorithm 2.

Algorithm 1 SOMP for multi-task multivariate sparse representation-based classification

Input: Dictionary $\boldsymbol{D}$ as defined in Section I, signal matrix $\boldsymbol{Y}$, number of iterations K

Initialization: residual $\boldsymbol{R}_0 = \boldsymbol{Y}$, index set $\Lambda_0 = \phi$, iteration counter $k = 1$

while $k \leq K$ **do**

(1) Find the index of the atom that best approximates all residuals:
$\lambda_{i,k} = \arg\max_{j=1,\ldots,n} \sum_{q=1}^{T} w_q \left\|\boldsymbol{R}_{k-1}^t \boldsymbol{d}_{q,j}\right\|_p, p \geq 1$

(2) Update the index set $\Lambda_{i,k} = \Lambda_{i,k-1} \bigcup \{\lambda_{i,k}\}, i = 1, \ldots, T$

(3) Compute the orthogonal projector $\boldsymbol{p}_{i,k} = \left(\boldsymbol{D}_{\Lambda_{i,k}}^t \boldsymbol{D}_{\Lambda_{i,k}}\right)^{-1} \boldsymbol{D}_{\Lambda_{i,k}}^t \boldsymbol{y}_i$, for $i = 1, \ldots, T$, where $\boldsymbol{D}_{\Lambda_{i,k}} \in \mathbb{R}^{n \times k}$ consists of the k atoms in $\boldsymbol{D}_i$ indexed in $\Lambda_{i,k}$

(4) Update the Residual Matrix $\boldsymbol{R}_k = \boldsymbol{Y} - \left[\boldsymbol{D}_{\Lambda_{1,k}}\boldsymbol{p}_{1,k}\ \ \ldots\ \ \boldsymbol{D}_{\Lambda_{T,k}}\boldsymbol{p}_{T,k}\right]$

(5) Increment k: $k \leftarrow k + 1$

end while

Output: Index set $\Lambda_i = \Lambda_{i,K}, i = 1, \ldots, T$; sparse representation $\hat{\boldsymbol{S}}'$ whose non-zero rows indexed for each representation by Λ_i, $i, i = 1, \ldots, T$, are the K rows of the matrix $\left(\boldsymbol{D}_{\Lambda_{i,K}}^t \boldsymbol{D}_{\Lambda_{i,K}}\right)^{-1} \boldsymbol{D}_{\Lambda_{i,K}}^t \boldsymbol{Y}$.

References

[1] J. A. Tropp, A. C. Gilbert, and M. J. Strauss, "Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit," *Signal Processing*, vol. 86, pp. 572–588, Apr. 2006.