Neurocomputing 238 (2017) 340–350

Contents lists available at ScienceDirect

Neurocomputing

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

# Maximum-margin sparse coding

Chien-Liang Liu[a], Wen-Hoar Hsaio[b,*], Bin Xiao[c], Chun-Yu Chen[c], Wei-Liang Wu[c]

[a] Department of Industrial Engineering and Management, National Chiao Tung University, Taiwan
[b] Information Management Center, National Chung-Shan Institute of Science and Technology, Taiwan
[c] Department of Computer Science, National Chiao Tung University, Taiwan

ARTICLE INFO

*Article history:*
Received 19 June 2016
Revised 9 December 2016
Accepted 30 January 2017
Available online 8 February 2017

Communicated by Jiayu Zhou

*Keywords:*
Maximum-margin
Sparse coding
Block coordinate descent

ABSTRACT

This work devises a maximum-margin sparse coding algorithm, jointly considering reconstruction loss and hinge loss in the model. The sparse representation along with maximum-margin constraint is analogous to kernel trick and maximum-margin properties of support vector machine (SVM), giving a base for the proposed algorithm to perform well in classification tasks. The key idea behind the proposed method is to use labeled and unlabeled data to learn discriminative representations and model parameters simultaneously, making it easier to classify data in the new space. We propose to use block coordinate descent to learn all the components of the proposed model and give detailed derivation for the update rules of the model variables. Theoretical analysis on the convergence of the proposed MMSC algorithm is provided based on Zangwill's global convergence theorem. Additionally, most previous research studies on dictionary learning suggest to use an overcomplete dictionary to improve classification performance, but it is computationally intensive when the dimension of the input data is huge. We conduct experiments on several real data sets, including Extended YaleB, AR face, and Caltech101 data sets. The experimental results indicate that the proposed algorithm outperforms other comparison algorithms without an overcomplete dictionary, providing flexibility to deal with high-dimensional data sets.



## 1. Introduction

The last decade has witnessed the great success of sparse coding, which has become a widely used framework in machine learning, signal processing, and neuroscience. The aim of sparse coding is to learn sets of overcomplete bases or dictionaries, namely the number of bases is greater than the dimensionality of the training examples [25], such that one can present data as a linear combination of the basis vectors. The advantage of over-complete bases is that these basis vectors are more likely to capture structures and patterns inherent in the input data. However, over-complete bases result in high-dimensional coding results when the dimension of the input data is huge, and dealing with high-dimensional datasets is always a challenging task in machine learning.

Sparse coding can model inhibition between the bases by sparsifying their activations, and the learned bases resemble the receptive fields of neurons in the visual cortex [37]. One advantage of using sparse vectors is that sparse representation allows to compute similarities very fast [5]. Furthermore, learning the dictionary from the training examples has been shown to produce state-of-the-art results compared to using off-the-shelf bases such as Fourier or wavelet bases. As a result, this work proposes to use sparse representation in the proposed model.

Existing dictionary learning can generally be divided into two categories, unsupervised dictionary learning and supervised dictionary learning [16]. The unsupervised learning approaches do not consider labeled data in learning dictionary, and focus on minimizing reconstruction errors with an $\ell_1$ regularization term to model data vectors as sparse linear combinations of basis elements. Unlike some other unsupervised learning techniques such as principal component analysis (PCA), sparse coding can be applied to learning overcomplete basis and does not impose that the basis vectors to be orthogonal, allowing more flexibility to adapt the representation to the data. It has been successfully applied to compressive sensing [38,55], computer vision [33,52] and image classification [28,29,49]. Many unsupervised dictionary learning algorithms have been devised in the last decade [1,9,24]. Compared to unsupervised dictionary learnin, supervised dictionary learning uses labeled data to learn classification-oriented dictionary, and recent research indicates that dictionaries constructed via supervised learning yield better classification results [16,31,36,50]. Dictionary learning can be viewed as a matrix factorization with sparsity constraint problem, and non-negative has shown to be an important property of matrix factorization, since it tends to learn parts-based decomposition of

* Corresponding author.
*E-mail address:* bass28.cs96g@g2.nctu.edu.tw (W.-H. Hsaio).

http://dx.doi.org/10.1016/j.neucom.2017.01.071
0925-2312/

images [23]. Thus, several research studies [8,47] have considered to impose non-negative constraint into dictionary learning recently.

Among supervised dictionary learning algorithms, some approaches incorporate discriminative terms into the objective function, such that the learning algorithms can learn a discriminative dictionary. Zang and Li [62] devised a discriminative K-SVD method called D-KSVD, which incorporates classification loss into the objective function to simultaneously learn a linear classifier and dictionary. Jiang et al. [16] proposed a label consistent K-SVD algorithm called LC-KSVD to learn a discriminative dictionary for sparse coding. To learn a compact and discriminative dictionary for sparse coding, a new label consistent constraint called discriminative sparse-code error is additionally introduced for LC-KSVD. Furthermore, LC-KSVD integrates general reconstruction error, classification error and discriminative sparse-code error to form a unified objective function. As a result, LC-KSVD can be viewed as an extension of K-SVD algorithm [1]. Yang et al. [58] proposed to use Fisher discrimination criterion to simultaneously learn class-specific sub-dictionaries and to make the coefficients more discriminative. Mairal et al. [32] proposed a generative/discriminative model for sparse signal representation and classification from learned dictionary and classification model.

Intuitively, the feature representation that is compatible with the underlying classifier can greatly enhance the performance of a learning system, so we propose to learn the feature representations and parameters of classifier simultaneously. Additionally, support vector machine (SVM) [46] is famous for its strong generalization guarantees derived from the maximum-margin property, inspiring us to consider maximum-margin in the model to propose a maximum-margin sparse coding (MMSC) framework, which jointly considers reconstruction loss and hinge loss in the model. We optimize the proposed framework with block coordinate descent, in which the block variables include sparse coefficients, dictionaries, and classifier parameters. The proposed framework uses maximum-margin constraint to obtain theoretical generalization guarantee, and uses sparse coding to learn a discriminative feature presentation. Once the training phrase is completed, the representations for training and testing data, dictionaries, and the parameters of the classifier are obtained. Then, classification of the testing data is achieved by using the obtained representations of testing data and the parameters of the classifier.

The SVM has been widely recognized as a state-of-the-art classification algorithm and has been applied to many practical applications [6,19,27]. The role of sparse coding in the proposed framework resembles kernel trick used in SVM, which maps data points into a high-dimensional space to presumably make the separation easier in that space. Therefore, using sparse coding and maximum-margin in the proposed framework resemble the two important properties of SVM, namely kernel trick and maximum-margin. The previous work that is related to the proposed framework is maximum-margin dictionary learning (MMDL) [26], which considers the learning of bag of visual words (BOV) model and the training of classifier with maximum margin criteria simultaneously. Several differences exist between the two frameworks. First, the proposed framework considers sparse coefficients, dictionaries and linear SVM parameter; while MMDL considers dictionaries and SVM parameter. Second, the proposed framework considers dictionary and hinge loss when optimizing sparse coefficients; while MMDL only considers dictionary to find feature representation. Finally, we use both labeled data and unlabeled data in the proposed framework to learn a discriminative and generalized dictionary; while MMDL only uses labeled data. We conduct experiments on several data sets, and the experimental results indicate that the proposed algorithm outperforms other algorithms. Besides the proposed framework and experiments, a proof of convergence is presented in the paper.

The following notations are used throughout the paper. We denote matrices by uppercase letters and vectors by bold-faced lowercase letters. Given a $n \times K$ matrix $\mathbf{D}$, the $(i, j)$ entry of the matrix $\mathbf{D}$ is $\boldsymbol{d}_{ij}$, the $i$th row of $\mathbf{D}$ is $\tilde{\mathbf{d}}_i^T \in \mathbb{R}^K$, and the $j$th column of $\mathbf{D}$ is $\mathbf{d}_j \in \mathbb{R}^n$. The $\ell_1$ norm of $\boldsymbol{\alpha}$ is $\| \boldsymbol{\alpha} \|_1 = \sum_{i,j} |\alpha_{ij}|$. Finally, we introduce an indicator variable $\mathbf{e}_j = [0, 0, \ldots, 1, \ldots, 0]^T$, denoting the $j$th element is 1, and the other elements are 0.

## 2. Maximum-margin sparse coding

Feature learning transforms raw data into a representation that can be effectively exploited in a supervised learning task such as classification. Among the feature learning algorithms, dictionary learning has shown its success in applications such as image processing, audio processing, and document analysis. Given data points $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_m]$, where $\mathbf{x}_i \in \mathbb{R}^n$, the problem can be formulated as learning a dictionary $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_K]$ as listed in Eq. (1), where $\mathbf{d}_j \in \mathbb{R}^n$ and $\lambda$ is a constant controlling the sparsity. The task is to minimize the reconstruction error, and uses $\ell_1$ regularization to enforce that the feature representation $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_m]$ is sparse.

$$\min_{\mathbf{D}} \sum_{i=1}^{m} \| \mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i \|_2^2 + \lambda \| \boldsymbol{\alpha}_i \|_1 \tag{1}$$

Once the dictionary learning process is completed, one can use the dictionary $\mathbf{D}$ to transform each data point into a sparse representation $\boldsymbol{\alpha}$, namely sparse linear combinations of basis elements. In image classification task, one can use SVM with sparse representation $\boldsymbol{\alpha}$ along with their corresponding labels to train a classification model, and has recently led to state-of-the-art results. In other words, the above approaches includes two phases. The first phase is an unsupervised feature learning phase, using available examples to learn a code book $\mathbf{D}$. The second phase is a supervised learning phase to learn model parameters with new feature representations.

### 2.1. Problem definition of supervised sparse coding

In supervised sparse coding for classification task, the labeled data and the underlying classifier are both considered in the objective function. The goal is to impose additional constraints on dictionary learning such that the learned feature representation can improve the underlying classifier. Given a training data set $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$, where $\mathbf{x}_i \in \mathbb{R}^n$ and $y_i \in \{1, \ldots, p\}$, the goal is to learn a feature representation $\boldsymbol{\alpha}_i \in \mathbb{R}^K$, such that the classification can benefit from the new representation. For the data $\mathbf{x}_i$ belonging to $t$th class, we use one-against-all scheme to represent the label vector, namely, $\mathbf{y}_i = [-1, \ldots, -1, 1, -1, \ldots, -1]^T$, denoting the $t$th element is 1, and the other elements are $-1$. We formulate the supervised sparse coding as a joint problem of reconstruction loss $\mathcal{L}_r$ and classification loss $\mathcal{L}_c$ as listed in Eq. (2), where $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_m]$ represents labels and $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_p]$ denotes classifier parameter.

$$\min \mathcal{L}_r(\mathbf{X}, \mathbf{D}, \boldsymbol{\alpha}) + \mathcal{L}_c(\boldsymbol{\alpha}, \mathbf{Y}, \mathbf{W}) \tag{2}$$

Several previous research studies have used the supervised sparse coding framework listed in Eq. (2) to devise algorithms. For example, the supervised dictionary learning [32] uses logistic loss to denote $\mathcal{L}_c$, and LC-KSVD uses a linear classifier in $\mathcal{L}_c$. The NMFSVM algorithm proposed by Gupta and Xiao [13] uses the above framework to identify the decomposition and classification parameters. The idea behind NMFSVM is to combine non-negative matrix factorization (NMF) objective with SVM, but several differences exist between the proposed method and NMFSVM. First, NMF in NMFSVM is related to matrix decomposition, while matrix