STRUCTURE-AWARE CLASSIFICATION USING SUPERVISED DICTIONARY LEARNING

Yael Yankelevsky and Michael Elad

Computer Science Department
Technion - Israel Institute of Technology
Haifa 32000, Israel
{yaelyan, elad}@cs.technion.ac.il

ABSTRACT

In this paper, we propose a supervised dictionary learning algorithm that aims to preserve the local geometry in both dimensions of the data. A graph-based regularization explicitly takes into account the local manifold structure of the observations. A second graph regularization gives similar treatment to the feature domain and helps in learning a more robust dictionary. Both graphs can be constructed from the training data or learned and adapted along the dictionary learning process. The combination of these two terms promotes the discriminative power of the learned sparse representations and leads to improved classification accuracy. The proposed method was evaluated on several different datasets, representing both single-label and multi-label classification problems, and demonstrated better performance compared with other dictionary based approaches.

Index Terms— supervised dictionary learning, sparse coding, graph Laplacian, classification

1. INTRODUCTION

Dictionary learning (DL) aims to learn a set of atoms such that a given signal can be well approximated by a sparse linear combination of these atoms. The standard DL problem is formulated as

$$\arg \min_{D,X} \|Y - DX\|_F^2 \quad \text{s.t.} \quad \|x_i\|_0 \leq T \quad \forall i \qquad (1)$$

where $Y \in \mathbb{R}^{n \times N}$ is the data matrix, $X \in \mathbb{R}^{K \times N}$ contains the sparse representations and $D \in \mathbb{R}^{n \times K}$ is an overcomplete dictionary with normalized columns (atoms).

Despite the popularity of standard DL methods in many domains, their performance in classification tasks is suboptimal, since an accurate reconstruction is not as important for classification as the discrimination capability of the dictionary. This motivates the emergence of supervised DL

techniques, exploiting both the existing label information and the underlying structure of the data.

Some of these methods attempt to learn a separate sub-dictionary for each class (e.g. [1–5]). Consequently, the sparse codes over the learned dictionary are used as features on which a classifier is trained. More sophisticated approaches (e.g. [6–11]) learn a discriminative dictionary by introducing a classification-error term into the objective function, and enforcing some discriminative criteria on the optimized sparse coefficients. By doing so, these methods form a unified learning problem and learn the dictionary and classifier jointly.

Of the latter category, we focus on the Label Consistent K-SVD (LC-KSVD) method [9] for joint learning of an over-complete dictionary D and an optimal linear classifier W:

$$\arg \min_{D,W,A,X} \|Y - DX\|_F^2 + \alpha \|Q - AX\|_F^2 + \beta \|H - WX\|_F^2$$
$$\text{s.t.} \quad \|x_i\|_0 \leq T \quad \forall i, \qquad (2)$$

where $H \in \mathbb{R}^{q \times N}$ is a binary matrix containing the labels of the training data (out of q possible classes), and $Q \in \mathbb{R}^{K \times N}$ associates label information with each dictionary atom, thus forcing signals from the same class to have similar sparse representations. The minimized objective hence balances between the reconstruction error $\|Y - DX\|_F^2$, the label consistency $\|Q - AX\|_F^2$ and the classification error $\|H - WX\|_F^2$. These terms can be fused together, leading to a standard formulation:

$$\arg \min_{\tilde{D},X} \left\| \tilde{Y} - \tilde{D}X \right\|_F^2 \quad \text{s.t.} \quad \|x_i\|_0 \leq T \quad \forall i, \quad (3)$$

where $\tilde{Y} = \begin{bmatrix} Y \\ \sqrt{\alpha}Q \\ \sqrt{\beta}H \end{bmatrix}$ and $\tilde{D} = \begin{bmatrix} D \\ \sqrt{\alpha}A \\ \sqrt{\beta}W \end{bmatrix}$.

Equation (3) can be efficiently solved using the K-SVD algorithm [12], which iteratively alternates between a sparse coding step (optimization over X) and a dictionary update step (that updates each dictionary atom in $\tilde{D}$ along with its related coefficients from X). Having completed the training process, the individual components D and W can be recovered from $\tilde{D}$. Consequently, classification of a new signal

is simply performed by sparse coding over the dictionary D and applying the learned classifier W on the resulting sparse coefficient vector, choosing the class that yields the highest score.

In [13], we have presented an unsupervised DL algorithm for graph signals. The algorithm takes into account the underlying structure of the data in both the feature and the manifold domains using graph smoothness constraints. Furthermore, the underlying structure, encapsulated by a graph Laplacian matrix, can be learned within the dictionary learning process to promote the desired smoothness.

In this paper, we propose an extension of our dual graph regularized DL algorithm to a supervised setting by applying the same ideas to the LC-KSVD approach [9]. The novelty of the LC-KSVD algorithm lies in the requirement that objects from the same class have similar sparse codes over some dictionary. While this method was shown to yield satisfactory classification results, outperforming other dictionary based methods (e.g. [12], [8], [1]), we argue that optimizing a separate sub-dictionary for each class or directly relating the dictionary atoms with specific classes is overly restrictive and highly sensitive to the initialization of the algorithm. We therefore replace the label consistency constraint with a graph-based smoothness regularization that leverages the label information and promotes the discriminative nature of the sparse codes. Additionally, we propose to simultaneously take into account the underlying structure of the training data in the feature domain such that the feature dependencies are preserved in the learned dictionary atoms.

In the sequel, we describe the proposed algorithm and demonstrate its efficiency in simulations.

2. GRAPH-CONSTRAINED SUPERVISED DICTIONARY LEARNING

2.1. Introducing the Data Manifold Structure

Our proposed algorithm is based on the LC-KSVD approach [9]. In order to take into account the local geometrical structure of the data manifold, we shall model the relationships between different data samples using a graph and require smoothness of the sparse codes over the graph topology.

Given a set of training samples $\{y_1, ..., y_m\} \in \mathbb{R}^n$, let us construct a weighted graph $\mathcal{M}$ with m vertices, where each node represents a training data point. The weight w_{ij} assigned to the edge connecting the i-th and j-th nodes is designed to be inversely proportional to the distance between them. A common choice uses a Gaussian kernel function

$$w_{ij} = \exp\left(-\frac{\|y_i - y_j\|_2^2}{\varepsilon_{\mathcal{M}}}\right). \tag{4}$$

where $\varepsilon_{\mathcal{M}}$ is a properly chosen scale parameter. The graph adjacency matrix $W^{\mathcal{M}}$ consists of the edge weights w_{ij}. The

graph Laplacian $L_{\mathcal{M}}$ is then defined as $L_{\mathcal{M}} = D^{\mathcal{M}} - W^{\mathcal{M}}$, where the degree matrix $D^{\mathcal{M}}$ is a diagonal matrix whose entries are $D_{ii}^{\mathcal{M}} = \sum_j w_{ij}$.

Following ideas from manifold learning and spectral graph theory, the graph Laplacian $L_{\mathcal{M}}$ can be used as a smoothness operator to preserve the local manifold structure. Similarly to the methods proposed in [14, 15], we incorporate $L_{\mathcal{M}}$ into the objective function as a regularizer of the form $Tr(XL_{\mathcal{M}}X^T)$. Denoting the i-th column of X by x_i, we observe that

$$Tr(XL_{\mathcal{M}}X^T) = \frac{1}{2}\sum_{i,j} w_{ij}\|x_i - x_j\|_2^2. \tag{5}$$

This term therefore encourages similar signals, having a large proximity measure w_{ij}, to have similar sparse codes, thus satisfying the commonly known manifold assumption [16]. In other words, it promotes smoothness of the obtained sparse representations X along the geodesics of the underlying data manifold, described by the graph Laplacian $L_{\mathcal{M}}$.

Replacing the original label consistency term with the graph regularization, the new DL problem formulation reads

$$\arg\min_{D,W,X} \|Y - DX\|_F^2 + \beta\|H - WX\|_F^2 + \gamma Tr(XL_{\mathcal{M}}X^T)$$

$$\text{s.t.} \quad \|x_i\|_0 \leq T \quad \forall i. \tag{6}$$

Fusing the first two components together similarly to the form presented in Equation (3), we obtain the graph-regularized supervised DL problem

$$\arg\min_{\tilde{D},X} \left\|\tilde{Y} - \tilde{D}X\right\|_F^2 + \gamma Tr(XL_{\mathcal{M}}X^T)$$

$$\text{s.t.} \quad \|x_i\|_0 \leq T \quad \forall i \tag{7}$$

where $\tilde{Y} = \begin{bmatrix} Y \\ \sqrt{\beta}H \end{bmatrix}$ and $\tilde{D} = \begin{bmatrix} D \\ \sqrt{\beta}W \end{bmatrix}$.

2.2. Introducing Feature Interdependencies

In order to take into account the interdependencies in the feature domain as well, we shall construct a second graph to model the relationships between different features, and require smoothness of the dictionary atoms over this new graph. That is, if two features behave similarly across the training signals, this behavior should be reflected in the structure of the learned atoms. Explicitly, given the set of training samples $\{y_1, ..., y_m\} \in \mathbb{R}^n$, let us construct a weighted graph $\mathcal{G}$ with n vertices, where each node represents a feature (corresponding to a row in the data matrix). The weight assigned to the edge connecting the i-th and j-th nodes is again designed to be inversely proportional to the distance between them using a Gaussian kernel. The graph adjacency matrix $W^{\mathcal{G}}$ consists of the edge weights, $D^{\mathcal{G}}$ is the corresponding degree matrix, and the graph Laplacian is defined as $L_{\mathcal{G}} = D^{\mathcal{G}} - W^{\mathcal{G}}$.

In a symmetric view to the manifold graph regularization, the feature graph should be integrated through a regularization term of the form $Tr(D^T L_{\mathcal{G}} D)$ applied to the dictionary matrix D. Since in the current formulation (7) the dictionary D only occupies a subset of the matrix $\tilde{D}$, we shall zero-pad $L_{\mathcal{G}}$ to the dimension of $\tilde{D}$:

$$\tilde{L}_{\mathcal{G}} = \begin{bmatrix} L_{\mathcal{G}} & 0 \\ 0 & 0 \end{bmatrix} \tag{8}$$

As the extended matrix $\tilde{L}_{\mathcal{G}}$ satisfies $\tilde{D}^T \tilde{L}_{\mathcal{G}} \tilde{D} = D^T L_{\mathcal{G}} D$, the regularization may equivalently be imposed on $\tilde{D}$.

The result is the dual graph constrained DL problem:

$$\arg \min_{\tilde{D},X} \left\| \tilde{Y} - \tilde{D}X \right\|_F^2 + \gamma Tr(X L_{\mathcal{M}} X^T) + \rho Tr(\tilde{D}^T \tilde{L}_{\mathcal{G}} \tilde{D})$$
$$\text{s.t.} \quad \|x_i\|_0 \leq T \quad \forall i \tag{9}$$

Solving this problem requires significant modifications of the K-SVD algorithm. This can be done using the ***graph^{2}DL*** algorithm we proposed in [13] that reflects the added restrictions. Upon completion of the learning process, D and W can be recovered from $\tilde{D}$, and classification is again performed by sparse coding the test signals using D and applying the classifier W on the resulting coefficients.

2.3. Learning Dependencies

The feature dependencies were so far inferred from the patterns detected in the training set. For this purpose, an initial weight matrix $W^{\mathcal{G}}$ for the features graph can be constructed by computing the pairwise distances between rows in the data matrix Y, which correspond to the different features:

$$w_{ij} = \exp \left(-\frac{\|Y(i,:) - Y(j,:)\|_2^2}{\varepsilon_{\mathcal{G}}} \right) \tag{10}$$

To better handle partial correlations, these dependencies can be learned and adapted along with the dictionary learning process, as previously suggested in [13]. Having obtained an intermediate dictionary D, the graph can be re-estimated based on the dictionary instead of the original input data Y, according to the following optimization problem:

$$\arg \min_{L_{\mathcal{G}} \in \Omega_L^n} \rho Tr(D^T L_{\mathcal{G}} D) + \mu \|L_{\mathcal{G}}\|_F^2, \tag{11}$$

where

$$\Omega_L^N = \{ L \in \mathbb{R}^{N \times N} \mid L = L^T, L_{ij} \leq 0 \ (i \neq j),$$
$$L\underline{1} = \underline{0}, \ Tr(L) = N \} \tag{12}$$

defines the set of trace-normalized valid graph Laplacian matrices of size $N \times N$. The term $\|L_{\mathcal{G}}\|_F^2$ was added to the objective function to control the sparsity of the resulting Laplacian

matrix. In a similar manner, the manifold graph Laplacian $L_{\mathcal{M}}$ may also be adapted by solving

$$\arg \min_{L_{\mathcal{M}} \in \Omega_L^m} \gamma Tr(X L_{\mathcal{M}} X^T) + \eta \|L_{\mathcal{M}}\|_F^2. \tag{13}$$

Both (11) and (13) can be cast as quadratic problems with linear constraints which could be solved using existing convex optimization tools, as elaborated in [17].

The complete algorithm, as summarized in Algorithm 1, is assembled by alternating between the graph constrained supervised DL problem posed in Equation (9), and adapting the Laplacian matrices $L_{\mathcal{G}}$ and $L_{\mathcal{M}}$.

Algorithm 1 Graph Constrained Supervised Dictionary Learning

- Initialize $\tilde{D}^{(0)}$
- Iterate for $k = 1, 2, ...$
  - Run a few iterations of **graph^{2}DL** (see [13], Algorithm 3), initialized with $\tilde{D}^{(k-1)}$, to solve Equation (9) and obtain $\tilde{D}^{(k)}, X^{(k)}$.

  - Update the graph Laplacians (Equations (11),(13)).
- Return $\tilde{D}^{(k)}, X^{(k)}$

---

3. EXPERIMENTAL RESULTS

3.1. Single-Label Classification

First, we evaluate the performance of the proposed algorithm for single-label image classification on the AR Face database [18] and on the Extended YaleB database [19].

The AR Face database consists of over 4000 images: 26 images per person for 126 different people. The images feature frontal view faces with different facial expressions, illumination conditions, and occlusions (sun-glasses and scarf). Following the standard evaluation procedure, we use a subset of the database consisting of 2600 images from 50 male subjects and 50 female subjects. Randomly selected 20 images per person constitute the training set, and the rest are used for testing. Each image is represented by a 540-dimensional feature vector using the procedure described in [9].

The Extended YaleB database contains 2414 frontal-face images of 38 individuals, captured under various illumination conditions. We randomly selected half of the images (about 32 images per person) as the training set and the rest are used for testing. Each image is represented by a 504-dimensional feature vector.

We compare the proposed algorithm to the two variants proposed in [9]: LC-KSVD1, which refers to (2) for $\beta = 0$, and LC-KSVD2, which refers to (2) for $\alpha, \beta \neq 0$. Two variants of our method are evaluated: **SupGraphDL**, standing for Algorithm 1 without learning the graph Laplacians $L_{\mathcal{G}}$ and $L_{\mathcal{M}}$, and **SupGraphDL-L**, in which the two Laplacians

are adapted throughout the process. The obtained results are summarized in Table 1. Note that we used a random partition of the data into training and testing sets, therefore the results obtained for LC-KSVD represent average performance and are different from the best case results presented in [9]. Nevertheless, when the incorporated graph is not arbitrarily constructed but optimized to properly convey the local data structure, our algorithm (SupGraphDL-L) achieves a higher classification accuracy for both evaluated datasets.

Method	AR	YaleB
LC-KSVD1	84.17	93.12
LC-KSVD2	85.00	93.29
SupGraphDL	84.93	92.89
SupGraphDL-L	**85.33**	**93.44**

Table 1: Classification accuracy (%) for the AR Face dataset and the Extended YaleB dataset.

3.2. Multi-Label Classification

Next, we evaluate the proposed algorithm for the more challenging task of multi-label classification. In this setting, each instance may be associated with multiple classes simultaneously. Thus, exploiting the interdependency between labels can significantly affect the success of the classification algorithm.

As an initial step, we extended our method to support multi-label classification, by altering the binary label matrix H to allow multiple non-zeros per column. The classification procedure was also extended to support multiple labels: instead of choosing the class yielding the maximal score, the relevant labels were selected as those reaching a result above a threshold, i.e. $\Omega_i = \{\ell : [Wx_i](\ell) \geq 0.5\}$, where x_i is the sparse representation of a test sample y_i over the dictionary D, and W is the learned classifier.

Furthermore, the previous results indicate that the graph constructions significantly impact the performance. For the multi-label classification settings, we propose a bilateral proximity metric for $L_{\mathcal{M}}$ consisting of both signal values and label data. That is, we multiply w_{ij} in (4) by $\exp\left(-\frac{d_h(h_i,h_j)^2}{\epsilon_h}\right)$ where h_i, h_j are the label vectors corresponding to y_i, y_j and $d_h(\cdot, \cdot)$ denotes the Hamming distance.

The algorithm was evaluated for two datasets: natural scene images and yeast gene functionality.

The natural scene dataset consists of 2000 natural scene images, each belonging to one or more out of 5 possible semantic classes: *desert*, *mountains*, *sea*, *sunset* and *trees*. Half of the images were used for training and the rest constitute the test set. Each image is represented by a 294-dimensional feature vector using the procedure described in [20]. The extracted features are spatial color moments in the LUV space,

which are commonly used in the scene classification literature.

The yeast dataset [21] is formed by micro-array expression data and phylogenetic profiles, and includes 2417 genes, 1500 of which are used for training and the rest constitute the test set. Each gene is represented by a 103-dimensional feature vector, and associated with a set of functional groups out of 14 possible classes (such as *metabolism*, *transcription* and *protein synthesis*).

Similarly to the single-label classification experiment, we compare two variants of our method (with and without adapting the graphs) to LC-KSVD1 and LC-KSVD2 from [9].

To assess the accuracy of the algorithms in the multi-label experiments, we use the average precision measure as defined in [20]. The classification results summarized in Table 2, indicate that our algorithm clearly outperforms the other methods.

	Scene	Yeast
LC-KSVD1	81.48	61.17
LC-KSVD2	82.57	61.21
SupGraphDL	82.78	64.69
SupGraphDL-L	**83.80**	**67.91**

Table 2: Classification accuracy (%) for the multi-label Scene and Yeast datasets.

In the multi-modal scenario, the potential overlap between different combinations of classes implies that a more complex underlying structure could be learned and exploited. As expected, the impact of the structural constraints in the multi-modal scenario is very significant, and more pronounced compared with the single-label examples. For the yeast dataset, which is known to be difficult, the achieved improvement in classification accuracy is almost 7%. Moreover, though increasing the computational complexity, it can be observed that adapting the Laplacian matrices further improves the classification results for both datasets.

4. CONCLUSIONS

In this paper, we extended our previously proposed dual graph regularized DL algorithm to a supervised setting. In the new algorithm, the dictionary atoms are encouraged to preserve the feature similarities as detected in the training data and encapsulated by the graph Laplacian $L_{\mathcal{G}}$, thus leading to a better representative and more robust dictionary. By adhering to the intrinsic geometrical structure of the data manifold, as captured by the graph Laplacian $L_{\mathcal{M}}$, the resulting sparse codes are more discriminative and can significantly enhance classification performance.

Experiments performed on different datasets demonstrate that the proposed algorithm yields very good classification results, outperforming other supervised DL algorithms for both single-label and multi-label classification tasks.

5. REFERENCES

[1] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Yi Ma, "Robust face recognition via sparse representation," *IEEE TPAMI*, vol. 31, no. 2, pp. 210–227, Feb 2009.

[2] M. Yang, L. Zhang, J. Yang, and D. Zhang, "Metaface learning for sparse representation based face recognition," in *ICIP*, 2010, pp. 1601–1604.

[3] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Fisher discrimination dictionary learning for sparse representation," in *ICCV*, 2011, pp. 543–550.

[4] I. Ramirez, P. Sprechmann, and G. Sapiro, "Classification and clustering via dictionary learning with structured incoherence and shared features," in *CVPR*, 2010, pp. 3501–3508.

[5] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Sparse representation based fisher discrimination dictionary learning for image classification," *Int. J. Comput. Vis.*, vol. 109, no. 3, pp. 209–232, 2014.

[6] J. Mairal, F. R. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Supervised dictionary learning," in *NIPS*, pp. 1033–1040. 2009.

[7] D.-S. Pham and S. Venkatesh, "Joint learning and dictionary construction for pattern recognition," in *CVPR*, June 2008, pp. 1–8.

[8] Q. Zhang and B. Li, "Discriminative K-SVD for Dictionary Learning in Face Recognition," in *CVPR*, June 2010, pp. 2691–2698.

[9] Z. Jiang, Z. Lin, and L. S. Davis, "Learning a Discriminative Dictionary for Sparse Coding via Label Consistent K-SVD," in *CVPR*, 2011, pp. 1697–1704.

[10] J. Mairal, F. Bach, and J. Ponce, "Task-Driven Dictionary Learning," *IEEE TPAMI*, vol. 34, no. 4, pp. 791–804, April 2012.

[11] M. J. Gangeh, A. Ghodsi, and M. S. Kamel, "Kernelized Supervised Dictionary Learning," *IEEE Trans. Sig. Proc.*, vol. 61, no. 19, pp. 4753–4767, Oct. 2013.

[12] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation," *IEEE Trans. Sig. Proc.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[13] Y. Yankelevsky and M. Elad, "Dual Graph Regularized Dictionary Learning," *IEEE Trans. Sig. Inf. Proc. over Net.*, vol. 2, no. 4, pp. 611–624, Dec. 2016.

[14] M. Zheng, J. Bu, C. Chen, C. Wang, L. Zhang, G. Qiu, and D. Cai, "Graph Regularized Sparse Coding for Image Representation," *IEEE Trans. Img. Proc.*, vol. 20, no. 5, pp. 1327–1336, May 2011.

[15] K. N. Ramamurthy, J. J. Thiagarajan, P. Sattigeri, and A. Spanias, "Learning Dictionaries with Graph Embedding Constraints," in *ASILOMAR*, Nov 2012, pp. 1974–1978.

[16] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Comput.*, vol. 15, no. 6, pp. 1373–1396, June 2003.

[17] X. Dong, D. Thanou, P. Frossard, and P. Vandergheynst, "Learning Graphs from Signal Observations under Smoothness Prior," in *arXiv:1406.7842*, 2014.

[18] A. Martinez and R. Benavente, "The AR Face Database," Tech. Rep. 24, CVC, June 1998.

[19] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From Few to Many: Illumination Cone Models for Face Recognition Under Variable Lighting and Pose," *IEEE TPAMI*, vol. 23, no. 6, pp. 643–660, June 2001.

[20] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern Recognition*, vol. 37, no. 9, pp. 1757–1771, 2004.

[21] A. Elisseeff and J. Weston, "A Kernel Method for Multi-Labelled Classification," in *NIPS*, 2001, pp. 681–687.