Graph-Constrained Supervised Dictionary Learning for Multi-Label Classification

Yael Yankelevsky and Michael Elad
Computer Science Department
Technion – Israel Institute of Technology

Abstract—**In this work, we tackle the problem of multi-label classification using a sparsity-based approach. Multi-label classification problems, in which each instance is associated with a set of multiple labels, have received significant attention over the past few years due to the ongoing growth of data dimensions and availability. However, the dependency between labels poses new challenges to existing classification techniques. We propose a supervised dictionary learning algorithm suited for the multi-label setting. The suggested scheme introduces a novel graph Laplacian regularization that encapsulates the training set labels. This regularization explicitly takes into account the local manifold structure of the data, thus promoting the discriminative power of the learned sparse representations. Experiments on two different real-world multi-label learning problems, i.e. natural scene classification and yeast gene functional analysis, demonstrate that our proposed algorithm achieves superior performance to other dictionary based approaches as well as some established multi-label learning algorithms.**

I. Introduction

Sparse and redundant representations have been successfully applied to various computer vision and image processing problems, such as image denoising [1], super resolution [2], etc. The underlying idea is approximating a given signal as a sparse linear combination of items from an over-complete dictionary. Since the choice of a representative dictionary is crucial to the success of sparse coding, improved performance can be achieved by adapting the dictionary to a set of training signals rather than using a predefined basis. Several methods have therefore been proposed to efficiently learn an over-complete dictionary from the data, such as the Method of Optimal Directions (MOD) [3] and K-SVD [4].

The dictionary learning problem is given by

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

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 over-complete dictionary with normalized columns (atoms).

Despite the popularity of general dictionary learning methods, they only take into account the representation accuracy and do not consider the discrimination capability of the dictionary. Therefore, their performance in classification tasks may be sub-optimal. To overcome this limitation, many supervised dictionary learning algorithms have recently been proposed that exploit the training data label information in various ways.

In most previous methods, the dictionary and classifier are learned separately. The straightforward approach for utilizing the label information is learning a separate sub-dictionary for each class (e.g. [5], [6], [7], [8]). Consequently, each test signal can be classified according to its reconstruction error using the class sub-dictionaries. Alternatively, the sparse codes can be used as features based on which a classifier is trained.

More sophisticated approaches (e.g. [9], [10], [11], [12]) introduce a classification-error term into the objective function, and enforce some discriminative criteria on the optimized sparse coefficients. By doing so, these methods form a unified problem and learn the dictionary and classifier jointly. Within this category we elaborate on the Label Consistent K-SVD (LC-KSVD) method [12]. The LC-KSVD algorithm jointly learns an over-complete dictionary D and an optimal linear classifier W by solving:

$$\begin{aligned}\arg\min_{D,W,A,X} & \|Y - DX\|_F^2 + \alpha\|Q - AX\|_F^2 + \beta\|H - WX\|_F^2 \\ \text{s.t.} \quad & \|x_i\|_0 \le T \quad \forall i.\end{aligned} \tag{2}$$

In this formulation, $H \in \mathbb{R}^{m\times N}$ is a binary matrix indicating the labels of the training data, such that $H_{ij} = 1$ if the signal y_j belongs to the i-th class (out of m possible classes). The binary matrix $Q \in \mathbb{R}^{K\times N}$ associates a label with each atom, such that $Q_{ij} = 1$ if the signal y_j and the atom d_i share the same label. The term $\|Q - AX\|_F^2$ thus enforces that X approximate the discriminative sparse codes Q, encouraging 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:

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

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

The research leading to these results has received funding from the European Research Council under European Union's Seventh Framework Program, ERC Grant agreement no. 320649.

978-1-5090-2152-9/16/$31.00 ©2016 IEEE

Equation (3) can be efficiently solved using the K-SVD algorithm [4], or any other fast algorithm recently developed for this purpose. Having completed the training process, a new signal is classified by sparse coding over the dictionary D and applying the learned classifier W on the sparse coefficient vector, choosing the class that yields the highest score.

While achieving impressive results, the previously proposed dictionary-based classification methods are limited to the single-label scenario. A more complicated problem is multi-label classification, which has become prevalent in recent years due to the increase of data volumes and availability of online labeling services. Such problems exist in several domains such as text mining, where a document may be associated with multiple topics; gene functional analysis, where each gene can belong to multiple functional classes; and natural scene classification, where each natural scene image may contain several objects and belong to multiple categories. Since each instance may be associated with multiple classes simultaneously, exploiting the interdependency between labels can significantly affect the success of a multi-label classification algorithm.

Some well-known approaches to multi-label classification include decision trees [13] and AdaBoost [14]. Several works have attempted to extend single-label classification approaches to handle the multi-label scenario. Prominent among them are the Multi-Label K-Nearest Neighbors algorithm (ML-KNN) [15] and the Instance-Based Logistic Regression (IBLR) [16].

In [17], we have proposed an unsupervised dictionary learning algorithm that takes into account the underlying structure of the data in both the feature and the manifold domains using graph smoothness constraints. In [18] we extended this algorithm to a supervised setting by applying similar ideas to the LC-KSVD approach. In this paper, we further generalize the supervised dictionary learning method to the more challenging multi-label setting. Thereafter, we add an adaptive threshold class to our proposed scheme, which will improve the ability to distinguish the relevant categories from the irrelevant ones. Additionally, we suggest replacing the label consistency term with a less restrictive graph Laplacian regularization, that promotes the discriminative nature of the sparse codes without explicitly learning a separate dictionary per class.

II. Supervised Dictionary Learning for Multi-Label Classification

Our proposed algorithm is based on the LC-KSVD approach [12]. Initially, we extended this method to support multi-label classification, by altering the binary label matrix H to allow multiple non-zeros per column. The classification procedure should also be extended to support multiple labels. Similarly to the single label problem, classification of a new test signal y_i is performed by sparse coding over the dictionary D and applying the optimized classifier W to the resulting coefficient vector x_i. Then, instead of choosing the class yielding the maximal score (the largest entry of Wx_i), the relevant labels are selected as those reaching a result above a threshold, i.e. $\Omega_i = \{\ell : [Wx_i](\ell) \geq 0.5\}$. Subsequently, we suggest two extensions to the algorithm: optimizing the classification threshold, and replacing the label consistency constraint with a graph-based smoothness regularization.

A. Adaptive classification threshold

A common practice in multi-label learning is to optimize a thresholding function which dichotomizes the label space into relevant and irrelevant label sets. In order to apply a similar concept to the dictionary-based approach, we introduce an additional threshold category, which is now optimized in the combined dictionary learning process. In practice, this is achieved by expanding the $m \times K$ classifier matrix W (m being the number of classes and K the number of dictionary atoms) to include an additional row, accounting for the new reference class. The label set for each signal is then determined by considering the result of each classifier with respect to the reference class. Put formally, we replace the term $\|H-WX\|_F^2$ with $\|H - MWX\|_F^2$, where $M \in \mathbb{R}^{m\times(m+1)}$ is defined as

$$M = \begin{bmatrix} & -1 \\ I_m & \vdots \\ & -1 \end{bmatrix} \qquad (4)$$

with I_m denoting the $m \times m$ identity matrix. Consequently, Equation (3) becomes

$$\arg\min_{\tilde{D},X} \left\|\tilde{Y} - \tilde{M}\tilde{D}X\right\|_F^2 \qquad \text{s.t.} \qquad \|x_i\|_0 \leq T \quad \forall i, \qquad (5)$$

where we defined the extended matrix $\tilde{M} = \begin{bmatrix} I_n & & \\ & I_K & \\ & & M \end{bmatrix}$.

In order to solve Equation (5), we suggest a modification to the K-SVD algorithm [4]. Adopting the K-SVD formulation, we perform sequential update of each atom along with its related coefficients. Let v_j denote the j-th column of X^T, so that v_j^T is the j-th row of X. For the j-th atom update, the error term could thus be reformulated as $\|\tilde{Y} - \tilde{M}\tilde{D}X\|_F^2 = \|E_j - \tilde{M}\tilde{d}_j v_j^T\|_F^2$ where $E_j \triangleq \tilde{Y} - \sum_{i\neq j} \tilde{M}\tilde{d}_i v_i^T$.

To preserve the representation sparsity, the update support is restricted to samples using the j-th atom . Let $E_j^R, (v_j^T)^R$ denote the restricted versions of E_j, v_j^T respectively. The optimization problem for the j-th atom is:

$$\arg\min_{\tilde{d}_j,(v_j^T)^R} \|E_j^R - \tilde{M}\tilde{d}_j(v_j^T)^R\|_F^2. \qquad (6)$$

This problem can be solved by alternating between updates of $\tilde{d}_j$ and $(v_j^T)^R$, leading to the following closed-form update rules:

$$(v_j^T)^R = \frac{\tilde{d}_j^T \tilde{M}^T E_j^R}{\tilde{d}_j^T \tilde{M}^T \tilde{M} \tilde{d}_j}, \qquad (7)$$

$$d_j = \frac{(\tilde{M}^T\tilde{M})^{-1}\tilde{M}^T E_j^R v_j^R}{\|v_j^R\|_2^2}. \tag{8}$$

For $\tilde{M} = I$, (7),(8) coincide with the K-SVD solution.

B. Graph Laplacian regularization

The main contribution of LC-KSVD is introducing the requirement that objects from the same class have similar sparse codes over the dictionary. While this requirement is reasonable, we find the current formulation, which directly associates each dictionary atom with a specific class, highly restrictive. Instead, inspired by our previous work [17], we propose learning a single dictionary and encouraging similar signals to have similar sparse codes using a graph Laplacian regularization. The proposed regularization leverages the label information and promotes the discriminative nature of the sparse codes. Explicitly, we suggest to 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 $\{y_1, ..., y_N\} \in \mathbb{R}^n$, we construct a weighted graph $\mathcal{G}$ with N vertices, where each node represents a 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. The graph adjacency matrix $W^{\mathcal{G}}$ consists of the edge weights w_{ij}. The graph Laplacian L is then defined as $L = D^{\mathcal{G}} - W^{\mathcal{G}}$, where the degree matrix $D^{\mathcal{G}}$ is a diagonal matrix whose entries are $D_{ii}^{\mathcal{G}} = \sum_j w_{ij}$.

Similarly to the methods proposed in [19], [20], [17], we incorporate the graph Laplacian L into the objective function as a regularizer of the form $Tr(XLX^T)$. Since $Tr(XLX^T) = \frac{1}{2}\sum_{i,j} w_{ij}\|x_i - x_j\|_2^2$, where x_i is the i-th column of X, this term encourages similar signals, having a large proximity measure w_{ij}, to have similar sparse codes.

The new formulation therefore explicitly considers the local geometrical structure of the data, such that the obtained sparse representations X vary smoothly along the geodesics of the underlying data manifold, as described by the Laplacian L. By preserving locality, the resulting sparse codes can have more discriminating power and hence better facilitate classification tasks.

For the multi-label classification settings, we propose a bilateral proximity metric consisting of both signal values and label data. That is, denote the signals by y_i, y_j and the corresponding labels by h_i, h_j, then

$$w_{ij} = \exp\left(-\frac{d_h(h_i,h_j)^2}{\epsilon_1}\right)\exp\left(-\frac{d_y(y_i,y_j)^2}{\epsilon_2}\right) \tag{9}$$

where $d_h(\cdot,\cdot)$ denotes the Hamming distance between the binary label vectors and $d_y(\cdot,\cdot)$ denotes the Euclidean distance between the data samples. The constructed manifold graph is therefore not only data driven but also integrates the auxiliary features given by the training set labels.

Combining the graph regularization into the dictionary learning task, instead of the original label consistency term, the new formulation reads

$$\begin{aligned} &\arg\min_{\tilde{D},X} \left\|\tilde{Y} - \tilde{M}\tilde{D}X\right\|_F^2 + \gamma Tr(XLX^T) \\ &\text{s.t.} \ \ \|x_i\|_0 \le T \quad \forall i, \end{aligned} \tag{10}$$

where $\tilde{Y} = \begin{pmatrix} Y \\ \sqrt{\beta}H \end{pmatrix}$, $\tilde{D} = \begin{pmatrix} D \\ \sqrt{\beta}W \end{pmatrix}$ and $\tilde{M} = \begin{bmatrix} I_n & \\ & M \end{bmatrix}$.

Similarly to the original framework, after the training process completes, the individual components can be recovered from $\tilde{D}$. Consequently, classification is performed by sparse coding the test signals using D and applying the generalized classifier MW on the resulting coefficients.

Solving Equation (10) requires significant modifications of the K-SVD algorithm, beyond those mentioned in the previous subsection. Having introduced the graph constraint, the update rule for the sparse coefficients related to each atom should be altered to reflect the added restriction. Explicitly, the coefficients update rule of the dictionary update step, previously given by Equation (7), is replaced by

$$v_j^R = \left(\gamma L + \tilde{d}_j^T\tilde{M}^T\tilde{M}\tilde{d}_j I\right)^{-1}(E_j^R)^T\tilde{M}\tilde{d}_j. \tag{11}$$

More importantly, the sparse coding step will now diverge from the standard form. This calls for a new pursuit technique, as described in detail in the sequel.

III. Manifold Regularized Sparse Coding

Denote the effective dictionary $\hat{D} = \tilde{M}\tilde{D}$. The manifold regularized sparse coding task is thus formulated as follows:

$$\begin{aligned} &\arg\min_{X} \left\|\tilde{Y} - \hat{D}X\right\|_F^2 + \gamma Tr(XLX^T) \\ &\text{s.t.} \ \ \|x_i\|_0 \le T \quad \forall i \end{aligned} \tag{12}$$

Due to the imposed graph constraint, the problem is no longer separable, and the sparse representations of different dataset signals are now dependent on each other. Previous work [19] proposed to solve Equation (12) by replacing the ℓ_0 norm with ℓ_1 and using a coordinate descent approach and subgradient methods.

In [17], we have proposed a different solution based on the Alternating Direction Method of Multipliers (ADMM) [21], which enables simultaneous update of all columns of X. In this approach, the non-convex sparsity constraint is separated from the rest and Equation (12) is reformulated as

$$\begin{aligned} &\arg\min_{X} \|\tilde{Y} - \hat{D}X\|_F^2 + \gamma Tr(XLX^T) \\ &\text{s.t.} \quad X = Z, \qquad \|z_i\|_0 \le T \quad \forall i. \end{aligned} \tag{13}$$

The augmented Lagrangian is then given by

$$\mathcal{L}_\rho(X,Z,U) = f(X) + g(Z) + \rho\|X - Z + U\|_2^2, \tag{14}$$

where $f(X) = \|\tilde{Y} - \hat{D}X\|_F^2 + \gamma Tr(XLX^T)$, $g(Z) = \mathcal{I}(\|z_i\|_0 \le T \ \forall i)$ for an indicator function $\mathcal{I}()$, and U is the scaled dual form variable.

The ADMM iterative solution consists of sequential optimizations of $\mathcal{L}_\rho$ over each of the variables X,Z, and U.

For the sub-problem of updating X, omitting the sparsity requirement has led to a quadratic objective. By simple derivation, this problem reduces to solving a Sylvester equation [22]:

$$(\hat{D}^T\hat{D}+\rho I)X+\gamma XL=\hat{D}^T\tilde{Y}+\rho(Z-U). \tag{15}$$

Since the eigenvalues of $(\hat{D}^T\hat{D}+\rho I)$ and $(-\gamma L)$ are distinct, a unique solution X is guaranteed [23].

The sub-problem of updating Z reduces to a shrinkage problem, requiring merely a sparse projection of $X+U$. To obtain it, hard thresholding is applied to $X+U$ such that only the T largest entries of each column are kept. We denote this projection operator by $\mathcal{P}_T$.

The graph regularized sparse coding algorithm is summarized in Algorithm 1.

Algorithm 1 Graph Regularized Sparse Coding

```
Initialize:
    $X^{(0)} = \arg\min_X \|\tilde{Y}-\hat{D}X\|_F^2$  s.t.  $\|x_i\|_0 \le T \quad \forall i,$
    $Z^{(0)} = X^{(0)}$ , $U^{(0)} = 0.$
Iterate: for $k = 1, 2, ...$
    • Update $X^{(k)}$ as the solution of
      $(\hat{D}^T\hat{D}+\rho I)X+\gamma XL = \hat{D}^T\tilde{Y}+\rho\left(Z^{(k-1)}-U^{(k-1)}\right)$
    • Update $Z^{(k)} = \mathcal{P}_T\left(X^{(k)}+U^{(k-1)}\right)$
    • Update $U^{(k)} = U^{(k-1)}+X^{(k)}-Z^{(k)}$
Output: The desired result is $Z^{(k)}$.
```

Note that since the problem is non-convex, ADMM is not guaranteed to converge. Nevertheless, for initialization with the standard sparse coding (e.g. using OMP), convergence is empirically observed within a few iterations.

IV. Simulation and Results

In the previous sections, we proposed a supervised dictionary learning algorithm for multi-label classification, based on the LC-KSVD [12]. The algorithm, denoted *graphDL-ML*, incorporates the data manifold regularization and includes an adaptive classification threshold.

We shall compare the proposed algorithm to three other methods: LC-KSVD1 and LC-KSVD2, which refer to the two variants proposed in [12], and ML-KNN [15]. The ML-KNN approach is, to our best knowledge, the state-of-the-art for multi-label classification, and was shown to give superior results to some well-established multi-label learning methods. For comparison with ML-KNN, we use the parameter value $K = 10$, in accordance with the simulations presented in [15].

One of the challenges in multi-label prediction is the additional notion of being partially correct. To account for partial prediction, we use five different multi-label evaluation metrics to evaluate the performance of the compared algorithms: the Hamming loss, One-error, Coverage, Ranking loss and Average Precision. Details of these evaluation metrics can be found in [15]. We note that for the average precision, a good classification should lead to large values, while for the first four measures - small values are desired. For both evaluated datasets, ten-fold cross-validation was performed and the mean results were used for comparison.

A. Natural scene classification

The algorithms were first evaluated for multi-label natural scene classification. In this task, each natural scene image may belong to several semantic classes simultaneously. Given a set of manually labeled training samples, the goal is to output a label set whose size is unknown a-priori for each unseen sample. The experimental 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. A few examples of this dataset are depicted in Figure 1.

Each image is represented by a 294-dimensional feature vector using the procedure described in [24]. The extracted features are spatial color moments in the LUV space, which are commonly used in the scene classification literature.

The obtained results are summarized in table I. Our algorithm clearly outperforms both the other dictionary based methods and the ML-KNN approach. The observed improvement in classification accuracy is between $1.7\% - 3.9\%$.

B. Yeast micro-array dataset

Next, we evaluate the algorithms for yeast gene functional classification. The yeast dataset 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*).

The obtained classification results are summarized in Table II. For this dataset, which is known to be difficult [25], our

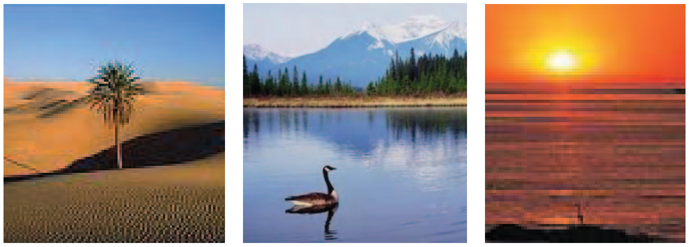
(a) (b) (c)

Fig. 1: Examples of multi-labeled images: (a) desert+trees, (b) mountains+sea+trees, (c) sea+sunset.

	Hamming loss	One-error	Coverage	Ranking loss	Average precision
LC-KSVD1	0.2404	0.2281	2.0132	0.2193	0.8148
LC-KSVD2	0.2456	0.2192	1.9298	0.2047	0.8257
ML-KNN	0.2825	0.2456	2.0132	0.2295	0.8038
graphDL-ML	**0.2254**	**0.1711**	**1.8904**	**0.1879**	**0.8427**

TABLE I: Experimental results for the natural scene image dataset

	Hamming loss	One-error	Coverage	Ranking loss	Average precision
LC-KSVD1	0.2774	0.4209	8.5060	0.3353	0.6117
LC-KSVD2	0.2764	0.4231	8.4591	0.3340	0.6121
ML-KNN	0.1980	0.2345	6.4144	0.1715	0.7585
graphDL-ML	0.2485	0.3609	7.8931	0.2862	0.6606

TABLE II: Experimental results for the yeast dataset

algorithm was unable to outperform the ML-KNN method. However, in the category of dictionary based methods, it performs significantly better, with almost 5% improvement in classification accuracy.

V. Conclusion

In this paper, we have proposed a multi-label extension to dictionary based classification methods, that includes an adaptation of the classification threshold and an introduction of a novel graph-based regularization that promotes the discriminative nature of sparse codes.

By adhering to the intrinsic geometrical structure of the data manifold, as captured by the graph Laplacian, the resulting sparse codes have better discriminating power and can significantly enhance classification performance. This is especially meaningful in multi-label classification problems, where exploiting the interdependency between labels poses an additional challenge to the classification algorithms. This was addresses in our proposed algorithm by the graph Laplacian regularization, which was used for incorporating this dependency into the learning process.

Experiments performed on two different datasets demonstrate that the proposed method yields very good classification results, and outperforms other supervised dictionary learning algorithms even for very challenging multi-label classification tasks.

References

[1] M. Elad and M. Aharon, “Image denoising via sparse and redundant representations over learned dictionaries,” *IEEE Trans. Img. Proc.*, vol. 15, no. 12, pp. 3736–3745, 2006.

[2] J. Yang, J. Wright, T. S. Huang, and Y. Ma, “Image super-resolution via sparse representation,” *IEEE Trans. Img. Proc.*, vol. 19, no. 11, pp. 2861–2873, 2010.

[3] K. Engan, S. O. Aase, and J. Hakon Husoy, “Method of optimal directions for frame design,” in *ICASSP*, vol. 5, 1999, pp. 2443–2446.

[4] 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.

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

[6] 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.

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

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

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

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

[11] Q. Zhang and B. Li, “Discriminative k-svd for dictionary learning in face recognition,” in *CVPR*, June 2010, pp. 2691–2698.

[12] 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.

[13] C. Vens, J. Struyf, L. Schietgat, S. Džeroski, and H. Blockeel, “Decision trees for hierarchical multi-label classification,” *Mach. Learn.*, vol. 73, no. 2, pp. 185–214, Nov. 2008.

[14] R. E. Schapire and Y. Singer, “Boostexter: A boosting-based system for text categorization,” *Mach. Learn.*, vol. 39, no. 2-3, pp. 135–168, 2000.

[15] M.-L. Zhang and Z.-H. Zhou, “Ml-knn: A lazy learning approach to multi-label learning,” *Pattern Recognition*, vol. 40, no. 7, pp. 2038–2048, Jul. 2007.

[16] W. Cheng and E. Hüllermeier, “Combining instance-based learning and logistic regression for multilabel classification,” *Mach. Learn.*, vol. 76, no. 2-3, pp. 211–225, Sep. 2009.

[17] Y. Yankelevsky and M. Elad, “Dual graph regularized dictionary learning,” *IEEE Transactions on Signal and Information Processing over Networks*, 2016. [Online]. Available: http://dx.doi.org/10.1109/TSIPN.2016.2605763

[18] ——, “Structure-aware classification using supervised dictionary learning,” 2016. [Online]. Available: http://arxiv.org/abs/1609.09199

[19] 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.

[20] K. N. Ramamurthy, J. J. Thiagarajan, P. Sattigeri, and A. Spanias, “Learning dictionaries with graph embedding constraints,” in *ASILOMAR*, Nov 2012, pp. 1974–1978.

[21] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, “Distributed optimization and statistical learning via the alternating direction method of multipliers,” *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, Jan. 2011.

[22] J. Sylvester, “Sur l’equations en matrices $px = xq$,” *Comptes Rendus Acad. Sci. Paris*, vol. 99, no. 2, pp. 67–71,115–116, 1884.

[23] R. Bhatia, *Matrix Analysis*. Springer-Verlag, New York, 1997.

[24] 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.

[25] A. Elisseeff and J. Weston, “A kernel method for multi-labelled classification,” in *NIPS*, 2001, pp. 681–687.