Semi-Supervised Dimension Reduction for Multi-label Classification

Buyue Qian
Dept. of Computer Science
University of California, Davis
byqian@ucdavis.edu

Ian Davidson
Dept. of Computer Science
University of California, Davis
davidson@cs.ucdavis.edu

Abstract

A significant challenge to make learning techniques more suitable for general purpose use in AI is to move beyond i) complete supervision, ii) low dimensional data and iii) a single label per instance. Solving this challenge would allow making predictions for high dimensional large dataset with multiple (but possibly incomplete) labelings. While other work has addressed each of these problems separately, in this paper we show how to address them together, namely the problem of semi-supervised dimension reduction for multi-labeled classification, SSDR-MC. To our knowledge this is the first paper that attempts to address all challenges together. In this work, we study a novel joint learning framework which performs optimization for dimension reduction and multi-label inference in semi-supervised setting. The experimental results validate the performance of our approach, and demonstrate the effectiveness of connecting dimension reduction and learning.

Introduction

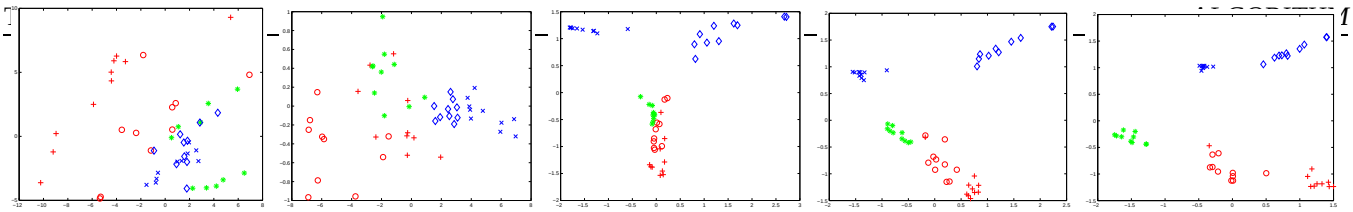
Motivation. Typical learning algorithm assumes each instance has exactly one label, unlabeled instances can be ignored for training and the data is in low-dimensional space. However, much data available today violates these assumptions. This is particularly true for complex objects such as images, video and music which can have multiple labels that are only partially filled in. The relaxation of each of these assumptions gives rise to the fields of multi-label learning, semi-supervised learning and dimension reduction respectively. While each of these fields has been well studied producing much excellent work, little work has looked at multiple relaxations simultaneously. The purpose of this work is to address all three problems (semi-supervision, multi-label and dimension reduction) simultaneously which to our knowledge is the first paper to attempt this. We believe (and will experimentally show) that this is advantageous as each problem is best not solved independently of the others.

Consider this simple experiment to illustrate the weakness of existing approaches. We collect 50 frontal and well aligned face images of five people in ten different expressions, each of which associated with four attributes, i.e. name, sex, bearded, glasses (see Figure 1, the face images are projected into a 2D space by different dimension reduction techniques, where the five symbols denote five people, and the three colors indicate the attributes associated to the face images. "Red" stands for female, unbearded, and non-glasses; "green" denotes male, unbearded, and non-glasses; and "blue" indicates male, bearded, glasses. For Principal Component Analysis (PCA) (Jolliffe 2002), an unsupervised dimension reduction technique, we see that it finds a mapping of the images into a 2D space (Figure 2(a)) where the people are not well separated. Though supervised dimension reduction is useful, it requires that each image has been completely labeled which is often not the case if labeling is expensive or not readily available. Thus we only label 30% images for supervision, we see that PCA+LDA (Belhumeur et al. 1997) performs only marginally better in Figure 2(b) because the missing labels can not be inferred. Since dimension reduction and learning algorithm could benefit each other, we establish the connection between them by our SSDR-MC approach. As shown in Figure 2(c) to 2(e), our algorithm allows the interaction between dimension reduction and label inference. This leads to accurate predictions and improvement in the dimension reduction result, iteration by iteration, until convergence. During the iterative process, images belonging to the same person or associated with similar attributes aggregate gradually while dissimilar images move far apart. Note that green marks are nearer to red marks than to blue marks, since they share more labels.

Figure 1: Sample face images

Related Works. Various dimension reduction methods have been proposed to simplify learning problems, and can be categorized as unsupervised, supervised, and semi-supervised. In contrast to traditional classification tasks where classes are mutually exclusive, the classes in multi-label learning are actually overlapped and correlated. Thus, two specialized multi-label dimension reduction algorithms have been proposed in (Zhang and Zhou 2008) and (Yu, Yu, and Tresp 2005), both of which try to capture the correlations between

Copyright © 2010, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.



(a) PCA (b) PCA+LDA (c) SSDR-MC Iteration 1 (d) SSDR-MC Iteration 2 (e) SSDR-MC Iteration 3

Figure 2: 2D projected faces from different methods. Symbols denote different people, and colors denote different attributes. Best viewed in color.

multiple labels. However, the usefulness of these methods is weakened by the lack of complete label knowledge, which is very expensive to obtain and even impossible for those extremely large dataset, e.g. web images annotation. In order to utilize unlabeled data, there are many semi-supervised multi-label learning algorithms (Chen et al. 2008) (Sun, Ji, and Ye 2008) been proposed, which solve learning problem by optimizing the objective function over graph or hypergraph. However, the performance of such approaches is limited by the lack of connection between dimension reduction and learning. To the best of our knowledge, (Ji and Ye 2009) is the first attempt to establish the connection between dimension reduction and multi-label learning, but it suffers from the inability of utilizing unlabeled data.

Overview, Contribution and Claims. In this work, we propose a general-purpose joint learning framework called SSDR-MC. We exploit reconstruction error as the criterion to measure how well a data point is represented by its nearest neighbors, which will lead us to locate the intrinsic geometric relations between data points, and provide helpful information for both dimension reduction and label inference. We establish the connection between dimension reduction and multi-label learning by an alternating optimization procedure. First we learn an optimal weight matrix (which decides the resultant dimension reduction) from both feature description and the available associated labels; then we infer the missing labels based on the weight matrix and the initial labels; repeat this procedure until the predicted labels stabilize. The alternating optimization can be viewed as a process during which the labeled data points **gradually propagate their labels** to those unlabeled data points along the weighted edges in the neighborhood. The main contribution of our work is that we tightly connect the dimension reduction and multi-label learning, and also successfully introduce semi-supervision. We show the SSDR-MC approach has following benefits:1) It is a general-purpose multi-label learning algorithm, especially for high dimensional data; 2) It incorporates the correlation between multiple labels; 3) Simultaneously solving dimension reduction and multi-label learning is beneficial; 4) Alternating optimization can avoid the decay of label influence during the label propagation process; 5) The iterative optimization procedure can converge in a small number of steps; 6) The parameters in our approach are easily tunable. In the following sections, we validate these claims, and demonstrate the effectiveness and efficiency of our algorithm.

Algorithm

In this section, we outline our algorithm starting from describing notations and the objective function. We then propose an alternating optimization approach where one unknown is held constant and the other is optimally solved. We summarize our algorithm in Table 1 and provide an approach for spectral embedding.

Notation

To deal with multi-label problem, we define a finite label set $\mathcal{C} = (1, \cdots, c)$, thus there are at most c labels associated with each data point. Given a data point set $\mathcal{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_l, \mathbf{x}_{l+1}, \cdots, \mathbf{x}_n) \subset \mathbb{R}^m (l \ll n)$, without loss of generality, we let the first l points be labeled and set a prior label matrix $Y \in \mathbb{B}^{c \times l}$, where $Y_{ij} = 1$ if $\mathbf{x}_j \in \mathcal{C}_i$ and $Y_{ij} = 0$ otherwise. An asymmetric *k-NN* graph $G(V, E)$ can be constructed over the n data points, in which an edge e_{ij} is established only if node v_j is among the k nearest neighbors of node v_i. We also define the weight matrix for dimension reduction $W \in \mathbb{R}^{n \times n}$ (we set $W_{ii} = 0$ to avoid self-reinforcement), and the binary classification matrix $F = [\mathbf{f}_1, \cdots \mathbf{f}_n] \in \mathbb{B}^{c \times n}$ of which $\mathbf{f}_i \in \mathbb{B}^c$ is the predicted label vector corresponding to instance $\mathbf{x}_i$. We let $\mathcal{N}_i$ denote the index set composed of k nearest neighbors of $\mathbf{x}_i$ ($\mathbf{x}_{\mathcal{N}_{i,j}}$ is the j-th nearest neighbor of $\mathbf{x}_i$). The instance $\mathbf{x}_i$ with missing label will have its labels inferred from the set $\mathcal{N}_i$.

Objective Function

Our aim is to solve dimension reduction and transductive inference for multi-label learning simultaneously. For this task, a reasonable choice of cost function is reconstruction error (Roweis and Saul 2000), which attempts to discover nonlinear structure in high dimensional data by exploiting the local symmetries of linear reconstructions. In this way, both data vector and label vector can be represented by a weighted linear combination of the corresponding k nearest neighbor vectors, and the problem is to optimize the weight matrix W and classification matrix F simultaneously. Thus we formulate the objective function as:

$$\mathcal{Q}(W, F) = (1-\alpha) \sum_{i=1}^{n} \|\mathbf{x}_i - \sum_{j \in \mathcal{N}_i} W_{ij} \mathbf{x}_j\|^2 + \alpha \sum_{i=1}^{n} \|\mathbf{f}_i - \sum_{j \in \mathcal{N}_i} W_{ij} \mathbf{f}_j\|^2 \tag{1}$$

where the first term is the data vector reconstruction error, which measures the error between the position of a point written as a linear combination of its nearest neighbors. The

second term is the label vector reconstruction error, which measures the error of writing the labels of a point as a linear combination of the labels of its nearest neighbors. The tuning parameter $0 \leq \alpha \leq 1$ determines how much the weights should be biased by the labels. It is worth to point out that the weight matrix W is invariant to **rotations** and **rescalings**, which follows immediately from the form of eq. (1). In order to guarantee the invariance to **translations**, we enforce the sum-to-one constraint upon each row of W. We use the same weight matrix W for both data vectors and label vectors, and constrain the prior labels to be unchangeable ($F_l = Y$) so as to capture the label information. The optimal W and F matrix can be obtained by minimizing the reconstruction error function, and thus our problem can be expressed as a constrained optimization:

$$\begin{aligned} \min \ & \mathcal{Q}(W, F) \\ s.t. \quad & F_l = Y \\ & \sum_{j \in \mathcal{N}_i} W_{ij} = 1, \, i = 1, \cdots, n. \end{aligned} \tag{2}$$

Alternating Optimization

In our proposed joint learning framework, the cost function involves two variables to be optimized. While simultaneously recovering both unknowns (the binary F and the continuous W) is intractable (the reduction is from the mixed integer programming problem) instead we solve eq. (2) for each unknown optimally (assuming the other unknown is a constant) in closed form and create an iterative approach based on these two steps. The minimization of $\mathcal{Q}(W, F)$ iterates between dimension reduction weights learning step and transductive inference step until F stabilized, which will asymptotically lead to a reliable local optimal weighting and labeling.

Learning Weights for Dimension Reduction In the weight learning step, we assume F is fixed, and we then solve for the weight matrix W as a constrained least squares problem in closed form. Since the optimal weights used to reconstruct a particular point is computed only from its own neighbor set, each row of the W can be obtained independently. We let the column vector $\mathbf{w}_i$ be composed of the non-zero entries in the i-th row of W in the order of k nearest neighbors, the problem turns to minimizing the following function:

$$\begin{aligned} \min_{\mathbf{w}_i} \quad & (1-\alpha)\|\mathbf{x}_i - X_i^{\mathcal{N}}\mathbf{w}_i\|^2 + \alpha\|\mathbf{f}_i - F_i^{\mathcal{N}}\mathbf{w}_i\|^2 \\ s.t. \quad & \mathbf{w}_i^T\mathbf{1} = 1 \end{aligned} \tag{3}$$

where $X_i^{\mathcal{N}} = [\mathbf{x}_{\mathcal{N}_{i,1}}, \cdots, \mathbf{x}_{\mathcal{N}_{i,k}}]$ is the neighbor set of data vector $\mathbf{x}_i$, and $F_i^{\mathcal{N}} = [\mathbf{f}_{\mathcal{N}_{i,1}}, \cdots \mathbf{f}_{\mathcal{N}_{i,k}}]$ is the neighbor set of label vector $\mathbf{f}_i$. Note that we use the same index set $\mathcal{N}_i$ for both of them, which is determined only by the data vectors. $\mathbf{1}$ denotes a $k \times 1$ all-one-vector. Using a Lagrange multiplier to enforce the constraint $\mathbf{w}_i^T\mathbf{1} = 1$, the optimal solution can be written in terms of the inverse local covariance matrix

$$\mathbf{w}_i = \frac{\left[(1-\alpha)P_i + \alpha Q_i\right]^{-1}\mathbf{1}}{\mathbf{1}^T\left[(1-\alpha)P_i + \alpha Q_i\right]^{-1}\mathbf{1}} \tag{4}$$

where $P_i = (\mathbf{x}_i\mathbf{1}^T - X_i^{\mathcal{N}})^T(\mathbf{x}_i\mathbf{1}^T - X_i^{\mathcal{N}}) \in \mathbb{R}^{k\times k}$ denotes the local covariance matrix of data vector $\mathbf{x}_i$, and $Q_i = (\mathbf{f}_i\mathbf{1}^T - F_i^{\mathcal{N}})^T(\mathbf{f}_i\mathbf{1}^T - F_i^{\mathcal{N}}) \in \mathbb{R}^{k\times k}$ indicates the local covariance matrix of label vector $\mathbf{f}_i$. As long as we obtain $\mathbf{w}_i$ of all the instances ($i = 1, \cdots, n$), the optimal weight matrix W can be constructed by simply placing each weight to its corresponding coordinates in the matrix.

Since the weight matrix W is obtained by optimization over both features and labels rather than calculating from a certain distance metrics, the label matrix F can be viewed as a supervisor, which guarantees the W partially (biased by α) fit to the label information. Notice that eq. (4) can not guarantee all the weights are non-negative. After observing the experiments, we found that negative weights are sparse and relatively small (generally $\ll 0.1$). Therefore, a straightforward solution is to discard the negative weights which we found does not affect the learning accuracy. However, in our presented work we allow the weights to be negative. The explanation is, if a positive weight means two points are similar, then on the contrary, a negative weight indicates they are dissimilar. Moreover, if a positive weight means the corresponding neighbor constructively contributes to the label prediction, then a negative weight indicates destructive contribution.

Transductive Inference In our multi-label inference step, the goal is to fill in those missing labels based on the weight matrix W. To recover the optimal F in closed form, we relax the binary classification matrix F to be real-valued. Since only the second term (label reconstruction error) in eq. (1) are accessed in minimization, we rewrite it in matrix format:

$$\begin{aligned} \min_{F} \quad & \mathcal{Q}(F) = \frac{1}{2} tr\left(F(I-W)^T(I-W)F^T\right) \\ s.t. \quad & F_l = Y \end{aligned} \tag{5}$$

The cost function above is convex allowing us to recover the optimal F by setting the partial derivative $\frac{\partial \mathcal{Q}}{\partial F} = 0$

$$\begin{cases} (I-W)F^T = 0 \\ F_l = Y \end{cases} \tag{6}$$

The optimization problem above yields a large, sparse system of linear equations, which could be solved by a number of standard methods. The most straightforward one is the close-form solution via matrix inversion. To compute the solution explicitly in terms of matrix operations, we split the weight matrix W into 4 blocks after the l-th row and column $W = \begin{bmatrix} W_{ll} & W_{lu} \\ W_{ul} & W_{uu} \end{bmatrix}$. Letting $F = \begin{bmatrix} F_l \\ F_u \end{bmatrix}$ where F_u denotes the missing label vectors, the optimal F_u can be recovered in closed form via matrix inversion:

$$F_u^T = (I - W_{uu})^{-1}W_{ul}F_l^T = (I - W_{uu})^{-1}W_{ul}Y^T \tag{7}$$

To complete the prediction at each iteration, we set up a threshold H where $0 \leq H \leq 1$. Then, ceil all the entries in $F_u \geq H$ to 1, and floor all the entries in $F_u < H$ to 0.

As we manipulate each multi-label vector as an entirety throughout the computation, the correlation and overlap between labels have been incorporated into our approach indeed. The transductive inference step can be viewed as process during which the labeled data points propagate labels along the weighted edges to their neighbors. In other word, each predicted label vector of a data point is actually the weighted linear combination of its neighbors' label vectors. In the same sense, the alternating optimization can be interpreted as a progressive label propagation process, i.e. we only make confident predictions at each iteration, and the size of predicted labels gradually grows. In this way, the label influence decay effect could be dramatically reduced.

Algorithm Summary

We summarize the SSDR-MC approach in Table 1.

Table 1: SSDR-MC Algorithm

Input	$\mathcal{X} \in \mathbb{R}^{m\times n}$, $Y \in \mathbb{B}^{c\times l}$, $F \in \mathbb{B}^{c\times n}$, $F_l = Y$, $0 \le \alpha \le 1, 0 \le H \le 1, k \ge 1$
Steps	(i) given F, solve weight matrix W by eq. (4) (ii) given W, recover matrix F_u by eq. (7) (iii) update predictions to F using threshold H repeat (i) (ii) (iii) until F converges, $\triangle\|F\| = 0$
Output	weight matrix W, predicted label matrix F_u

Spectral Embedding

Since weight matrix W completely captured the the intrinsic geometric relations between data points, we can use it to perform dimension reduction. Now, we show how to compute the dimension reduced data vector explicitly. Note we do not need do this for learning only dimension reduction. Let d denote the desired number of dimensions of the feature vector, the dimension reduced instance $\mathbf{x'}_i$ minimizes the embedding cost function:

$$\mathcal{Q}(X') = \sum_{i=1}^{n} \|\mathbf{x'}_i - \sum_{j=1}^{n} W_{ij}\mathbf{x'}_j\|^2 \tag{8}$$

where $X' \in \mathbb{R}^{d\times n}$ is the dimension reduced data matrix. The embedding cost in eq. (8) defines a quadratic form in the vector $\mathbf{x'}_i$. Subject to constraints that make the problem well-posed, the minimization can be solved as a sparse eigen decomposition problem

$$\min \ \mathcal{Q}(X') = tr\left(X'MX'^T\right) \tag{9}$$

where $M = (I - W)^T(I - W)$. We can find the optimal embedding by computing the bottom $d + 1$ eigenvectors of the matrix M, then discard the smallest eigenvector which it is an unit vector with all equal components (represents a free translation mode of eigenvalue zeros). The remaining d eigenvectors are the optimal embedding coordinates.

Implementation Issues and Discussion

Solution for Weak Prior Knowledge

In order to extend the generality of our approach, we provide another efficient solution to utilize the incomplete and noisy labeled data. Since we enforced the initial labeled data to be unchangeable previously, the solution provided in eq. (7) suffers from two problems: i) in multi-label learning task, the knowledge of labels for a certain labeled instance may not be complete ii) there may be considerable noise scattered in labeled data. A reasonable solution for the two problems is to relax the constraint by adding a new term to the inference cost function, namely local fitting penalty (Zhou et al. 2004) allowing slight changes of the fixed/prior labels. We extend the prior label matrix Y to be a $c \times n$ matrix (fill the missing labels with 0 for the first iteration), and the new inference cost function can be written as

$$\min_F \quad \mathcal{Q}(F) = \frac{1}{2}tr\left\{F(I - W)^T(I - W)F^T + \beta(F - Y)(F - Y)^T\right\} \tag{10}$$

where coefficient $\beta > 0$ balances the reconstruction error and local fitting penalty. If we set $\beta = \infty$, the cost function will reduce to eq. (5). The minimization problem is straightforward since the cost function is convex and unconstrained

$$\frac{\partial \mathcal{Q}}{\partial F^*} = 0 \quad \Longrightarrow \quad (I - W)F^{*T} + \beta(F^* - Y)^T = 0 \tag{11}$$

Thus, the optimal F^* can be recovered as:

$$F^{*T} = \left(\frac{I - W}{\beta} + I\right)^{-1} Y^T \tag{12}$$

After each iteration, we update confident predictions in F to Y. When the prior label knowledge is weak, the optimal F can be recovered by this relaxed solution instead of eq. (7).

Efficiency Improvement

Observing eq. (4), we see that the denominator of the fraction is a constant which rescales the sum of $\mathbf{w}_i$ to 1. Therefore, in practice, a more efficient way to recover the optimal $\mathbf{w}_i$ is simply to solve the linear system of equations, and then rescale the sum of weights to 1. Let L_i denote the mixed local covariance matrix $(1 - \alpha)P_i + \alpha Q_i$. The linear system of equations corresponding to instance $\mathbf{x}_i$ can be written as: $L_i\mathbf{w}_i = \mathbf{1}$. The optimal $\mathbf{w}_i$ can be recovered efficiently by solving this linear system, and then rescale the sum of $\mathbf{w}_i$ to 1. When the local covariance matrices is singular ($k > m$ or $k > c$), the linear system of equations can be conditioned by adding a small multiple of the identity matrix

$$L_i \leftarrow L_i + \frac{\xi tr(L_i)}{k} I \tag{13}$$

where ξ is a very small value ($\xi \ll 0.1$).

Convergence

For the solution of F provided in eq. (12), there is a guaranteed convergence since we update the confident predictions to Y after each iteration (adding the most confident predictions to labeled set). However, the solution of F provided in eq. (7) cannot guarantee a convergence. Therefore, it is possible that the predictions of current iteration oscillate and backtrack from predicted labelings in previous iterations. A straightforward method to remove backtracking, inconsistency and unstable oscillation is to set up a small tolerance T. If the number of different entries between the current prediction F_c the previous prediction F_p is smaller than the tolerance T, the iteration will be terminated and the last prediction matrix F will be output as the final classification result.

But the tolerance T is rarely useful, since our alternating optimization converges in a small number of iterations for most cases in practice. Additionally, the co-occurrence of quick convergence and high classification accuracy in experiments implies that we can achieve both of them (will experimentally show) by selecting appropriate parameters (α and H).

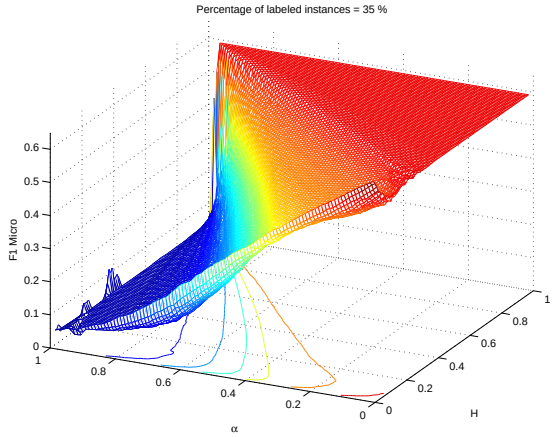
Experiment Evaluation

Dataset and Experiment Settings

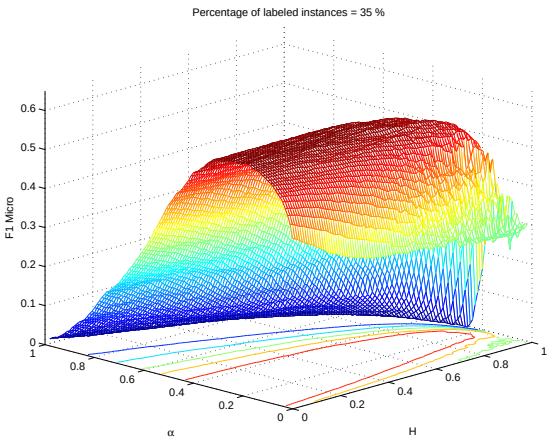
To show the generality of our approach, we carry out experiments on three different types of real world data (Chang and Lin 2001). **Yeast**: gene dataset consists of 2,417 samples, each of which belongs to one or more of 14 distinct functional categories. The feature vector is in 103-dimensional space, and associated with 4.24 labels averagely. **Scene**: image dataset consists of 2,407 natural scene images, each of which is represented as a 294-dimensional feature vector and belongs to one or more (1.07 in average) of 6 categories: "sunset","urban","fall foliage","field","mountain","beach". **SIAM TMC 2007**: text dataset for SIAM Text Mining Competition 2007 consisting of 28,596 text samples, each of which belongs to one or more (2.21 in average) of 22 categories. In experiment, we only take a randomly selected subset containing 3,000 samples from the original dataset, then use binary term frequencies and normalize each instance to unit length (30,438-dimensional feature vector). We compare our algorithm with three baseline models: 1) **RankSVM** (Elisseeff and Weston 2001), a state-of-the-art supervised multi-label classification algorithm; 2) **PCA+RankSVM**, perform PCA dimension reduction as a separate step before RankSVM; 3) **ML-GFHF**, the multi-label version (two-dimensional optimization) of harmonic function (Chen et al. 2008). For fairness, we use the same parameter setting throughout the experiment. For RankSVM, we choose RBF kernel function (σ = the average of Euclidean distances between all pairs of data points), and fix the penalty coefficient $C = 1000$. For ML-GFHF, we construct a *k-NN* ($k = 15$) graph and also use the RBF kernel ($\sigma = \sum_{i=1}^{n} \parallel \mathbf{x}_i - \mathbf{x}_{ik} \parallel /n$, where $\mathbf{x}_{ik}$ is the k-th nearest neighbor of $\mathbf{x}_i$) to recover the edge weights. For our approach, we choose eq. (7) as the inference function, set the number of neighbors $k = 15$, the tuning parameter $\alpha = 0.1$, the threshold $H = 0.3$, and the tolerance $T = 5$. For the performance evaluation purpose, we exploit the standard metrics: micro-average F1 score (F1 Micro) (Yang 1999).

Parameter Selection

To explore the parametrical stability of our approach, we evaluate the performance of SSDR-MC on the two representative datasets (Yeast, most samples have several labels; Scene, most samples have only one label) under a series of varying parameter settings. We randomly select 35% data points from the dataset as the labeled data, and then increase α and H gradually from 0.01 to 0.99 with a step size of 0.01. By observing the result shown in Figure 3, we see that there is a large continuous region of parameter settings (marked by the red boundary at the bottom of each figure), in which the performance of our approach is excellent and stable. Since the reliable region of parameter settings takes a relatively large area (55.4% and 35.6% of total area for Yeast and Scene respectively), we can conclude that the two parameters in our framework are easily tunable. Based on the experimental result, SSDR-MC can achieve an accurate multi-label prediction by choosing $0.05 \leqslant \alpha \leqslant 0.25$ and $0.2 \leqslant H \leqslant 0.6$. Another interesting phenomenon we observed is the co-occurrence of reliable region and quick convergence. In the reliable parameter region, the alternating optimization we proposed always converges in a small number (4 to 8) of steps, which means the convergent problem could also be solved by choosing appropriate parameters.



(a) Yeast

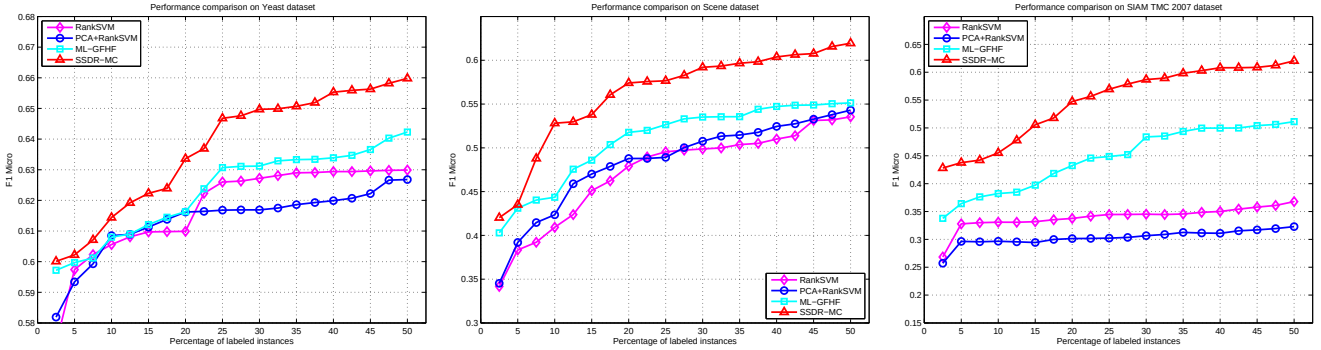


(b) Scene

Figure 3: F1 Micro of SSDR-MC with respect to α and H. Best viewed in color.

Performance Comparison

To comprehensively compare our algorithm with the three base models, we apply the four algorithms on the three datasets with a series of varying sized labeled data. In each trial, we randomly select a portion of instances from the dataset as the labeled set, and the rest of the data points will be used as the testing set. The portion of labeled data gradually increases from 2.5% to 50% with a step size of 2.5%,



(a) Yeast (b) Scene (c) SIAM TMC 2007

Figure 4: Performance comparison by F1 Micro score

and the result shown in Figure 4 is based on **the average over 50 trials**. For the three real-world datasets we explored in our experiments, our approach performs **statistically significantly better than the competitors at the 98.77% level** when all experimental results (regardless of step size) are pooled together. We have shown that by choosing appropriate parameters, our framework is expected to achieve a quick convergence. We evaluate the efficiency of the proposed alternating optimization by comparing the time consumptions with the three competitors. Although the optimization is iterative, statistically, SSDR-MC is comparably efficient as the others, and even more efficient if the data is in high-dimensional space. The reason is that the weights in our approach are recovered by solving a linear system of equations (eq. (13)) while others need to explicitly calculate all pairwise distances. We observe from Figure 4(c) that PCA+RankSVM does not outperform RankSVM on the high-dimensional data SIAM TMC 2007, which indicates the lack of connection between dimension reduction and learning algorithm will limit the usefulness of dimension reduction dramatically. As the superior performance of our algorithm, we demonstrated the the effectiveness of connecting them together, especially when the data is high-dimensional. Also, we see from the result that the performance of the proposed algorithm improves monotonically as the size of the labeled data increase.

Conclusion

As applications in data mining and machine learning move towards demanding domains, they must move beyond the restrictions of complete supervision, single-labels and low-dimensional data. The SSDR-MC algorithm is to our knowledge the only work that attempts to address all three limitations simultaneously. SSDR-MC can be viewed as simultaneously solving for two sets of unknowns: filling in the missing labels and identifying the projection vectors that make points with similar labels close together and points with different labels far apart. The superior experimental performance of our approach demonstrates the usefulness of connecting dimension reduction and learning, as well as the effective use of both labeled and unlabeled data.

Acknowledgements

The authors are appreciative to ONR for funding this research via grant ONR (N000140910712 P00001).

References

Belhumeur, P. N.; Hespanha, J. P.; Hespanha, P.; and Kriegman, D. J. 1997. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *PAMI* 19:711–720.

Chang, C.-C., and Lin, C.-J. 2001. *LIBSVM: a library for support vector machines*.

Chen, G.; Song, Y.; Wang, F.; and Zhang, C. 2008. Semi-supervised multi-label learning by solving a sylvester equation. *SIAM Conference on Data Mining*, 410–419.

Elisseeff, A. E., and Weston, J. 2001. A kernel method for multi-labelled classification. *NIPS 14*, 681–687.

Ji, S., and Ye, J. 2009. Linear dimensionality reduction for multi-label classification. In *Proceedings of the 21st international jont conference on Artifical intelligence*, 1077–1082.

Jolliffe, I. T. 2002. *Principal Component Analysis*. New York: Springer-Verlag, 2nd edition.

Roweis, S. T., and Saul, L. K. 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science* 290:2323–2326.

Sun, L.; Ji, S.; and Ye, J. 2008. Hypergraph spectral learning for multi-label classification. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 668–676.

Yang, Y. 1999. An evaluation of statistical approaches to text categorization. *Journal of Information Retrieval* 1:67–88.

Yu, K.; Yu, S.; and Tresp, V. 2005. Multi-label informed latent semantic indexing. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, 258–265.

Zhang, Y., and Zhou, Z.-H. 2008. Multi-label dimensionality reduction via dependence maximization. In *AAAI'08: Proceedings of the 23rd National Conference on Artificial Intelligence*, 1503–1505.

Zhou, D.; Bousquet, O.; Lal, T. N.; Weston, J.; and Schlkopf, B. 2004. Learning with local and global consistency. In *Advances in Neural Information Processing Systems*, 321–328.