Semi-Supervised Dimension Reduction for Multi-label Classification

Buyue Qian and Ian Davidson
Dept. of Computer Science,
University of California, Davis

AAAI-10: Twenty-Fourth Conference on Artificial Intelligence
Atlanta July 13 ,2010

Outline

1. Motivation
2. A Simple Example
3. Algorithm
   - 3.1 Objective Function
   - 3.2 Alternating Optimization
   - 3.3 Algorithm Summary
   - 3.4 Spectral Embedding
4. Discussion
   - 4.1 Weak Prior Knowledge
   - 4.2 Convergence
5. Experiment
   - 5.1 Experiment Settings
   - 5.2 Parameter Selection
   - 5.3 Result
6. Conclusion

1. Motivation

Typical learning algorithm assumes

1) Single label per instance
2) Complete supervision
3) Low dimensional data

The relaxation of each of these assumptions gives rise to the fields of

1) Multi-label learning
2) Semi-supervised learning
3) Dimension reduction

However, we want solve them together, and believe that this is advantageous as each problem is best not solved independently of the others.

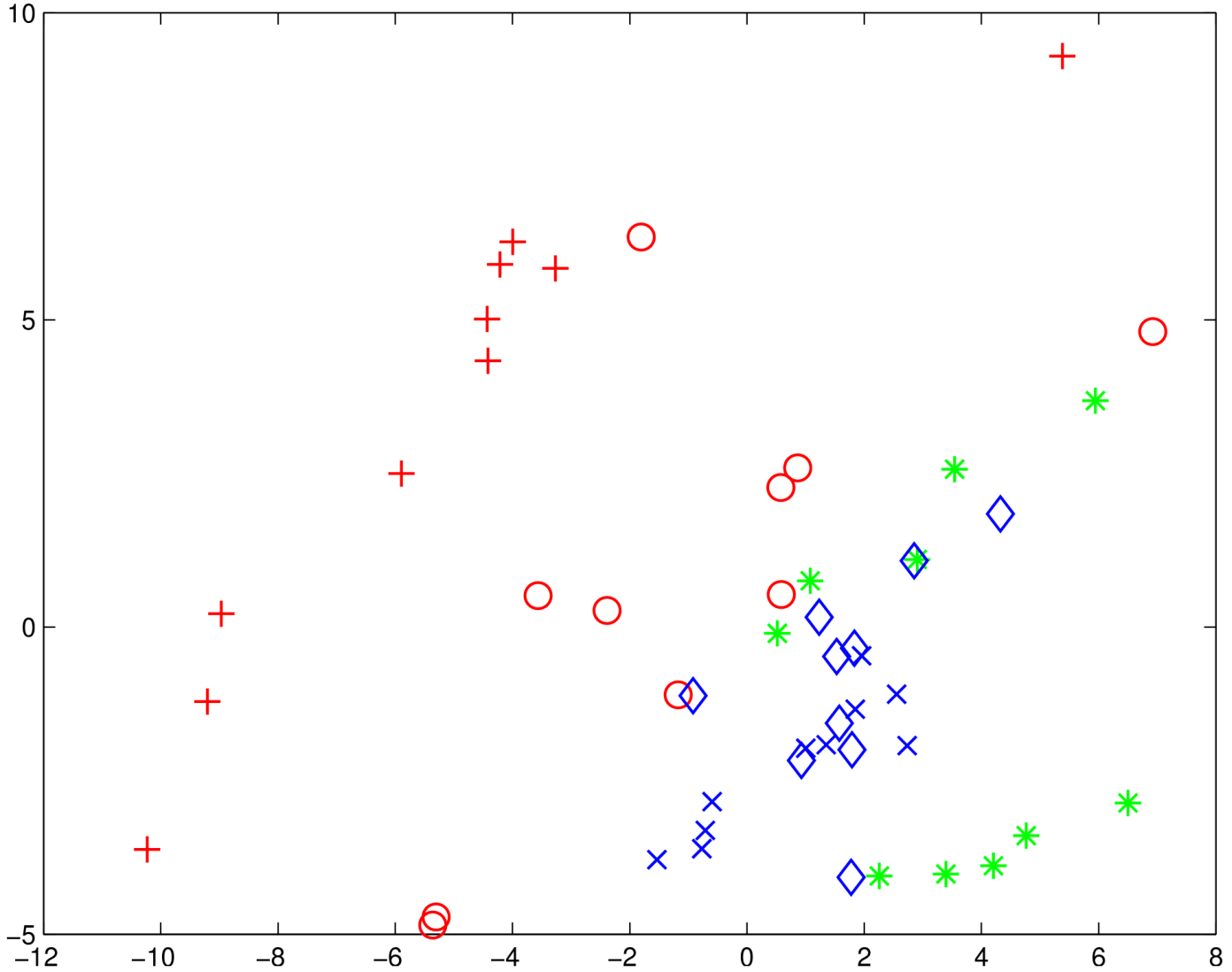
2. A Simple Example

50 face images of 5 people in 10 different expressions, each image associated with 4 attributes, name, gender, bearded or not, glasses or not

These face images are projected into 2D space by using different dimension reduction techniques.

In the following figures, different symbols denote different people, the three colors indicate the attributes:
“red”: female, not bearded, non-glasses
“green”: male, not bearded, non-glasses
“blue”: male, bearded, glasses

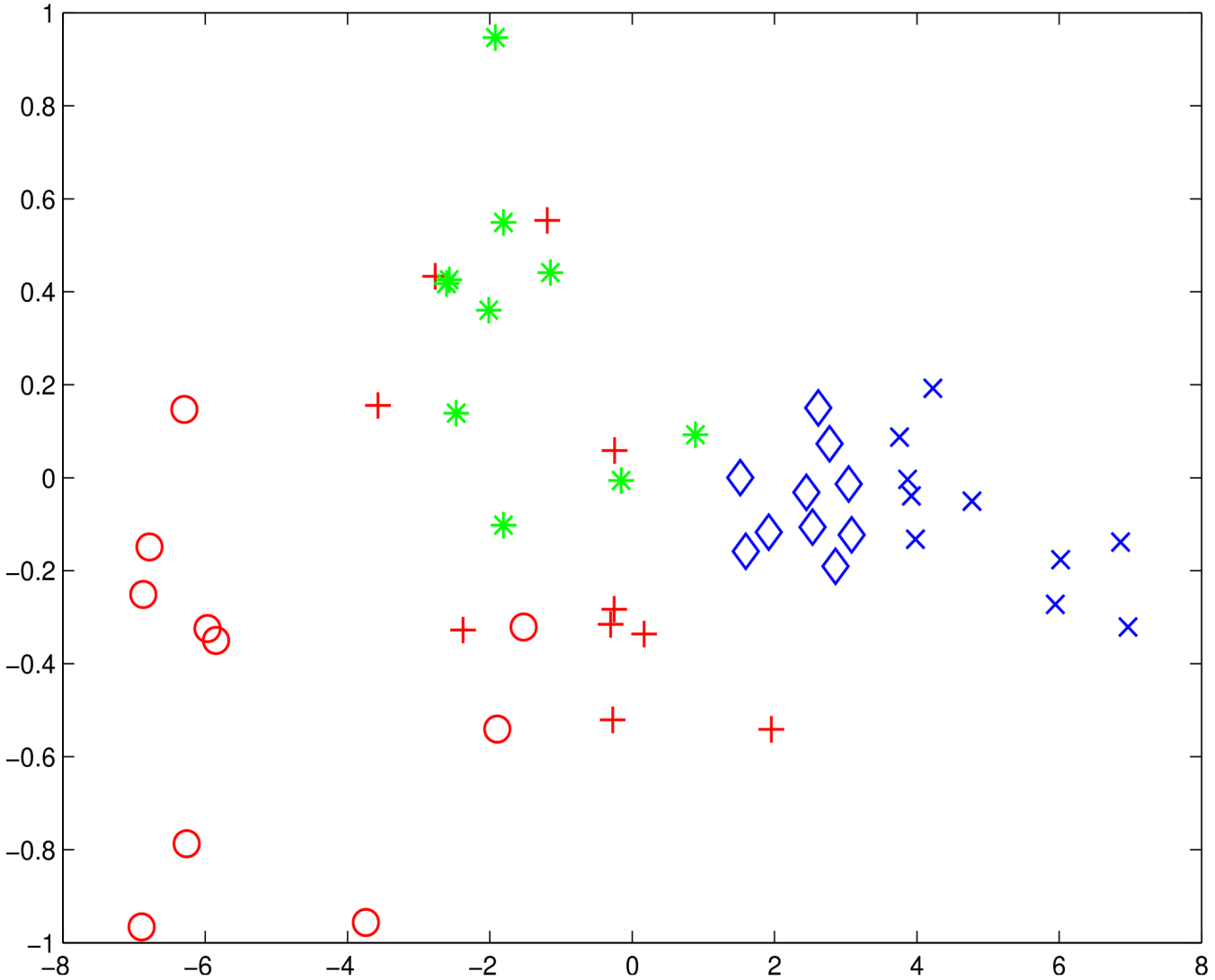
PCA



Not well separated

“red”: female, not bearded, non-glasses
“green”: male, not bearded, non-glasses
“blue”: male, bearded, glasses

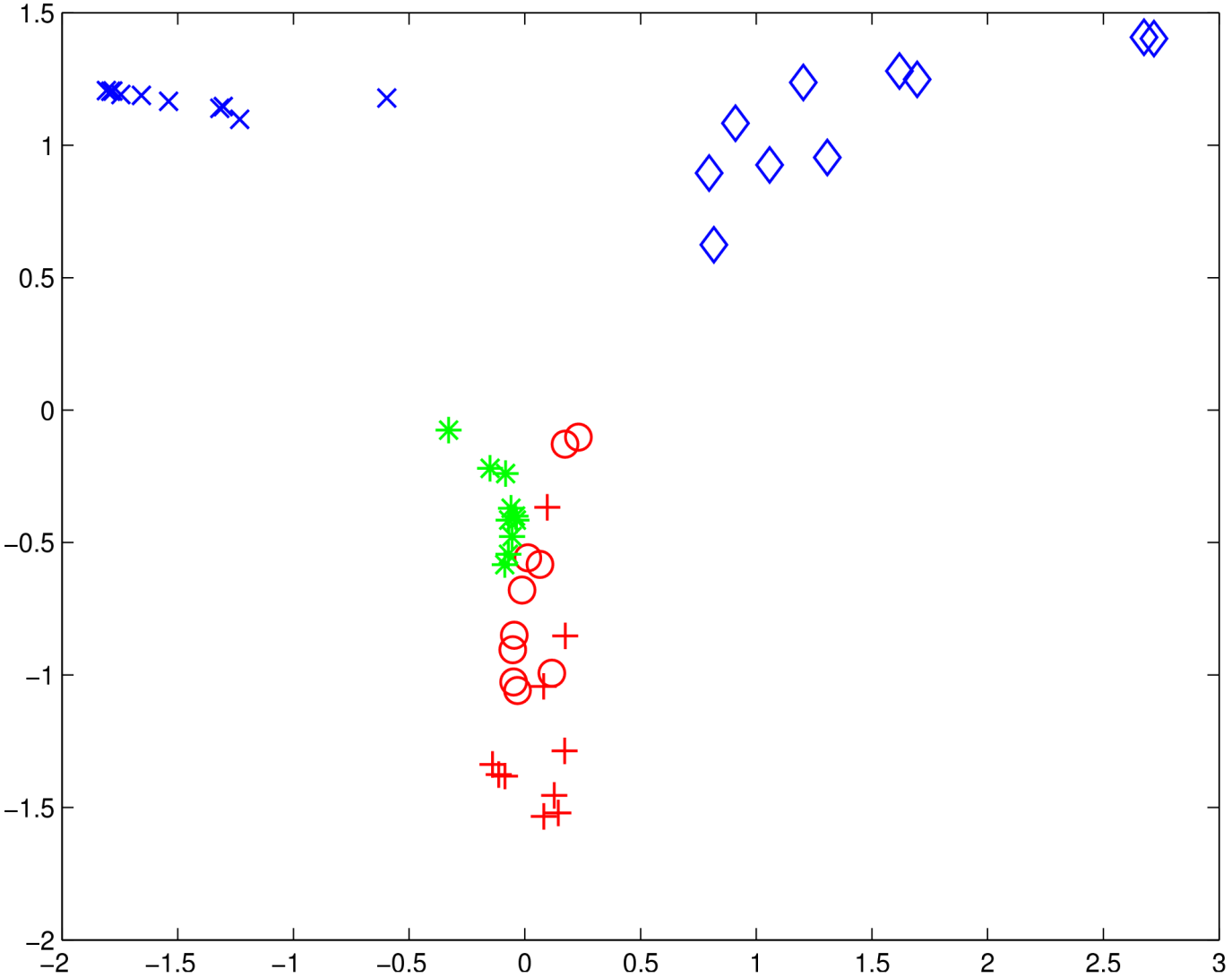
PCA+LDA



Marginally better

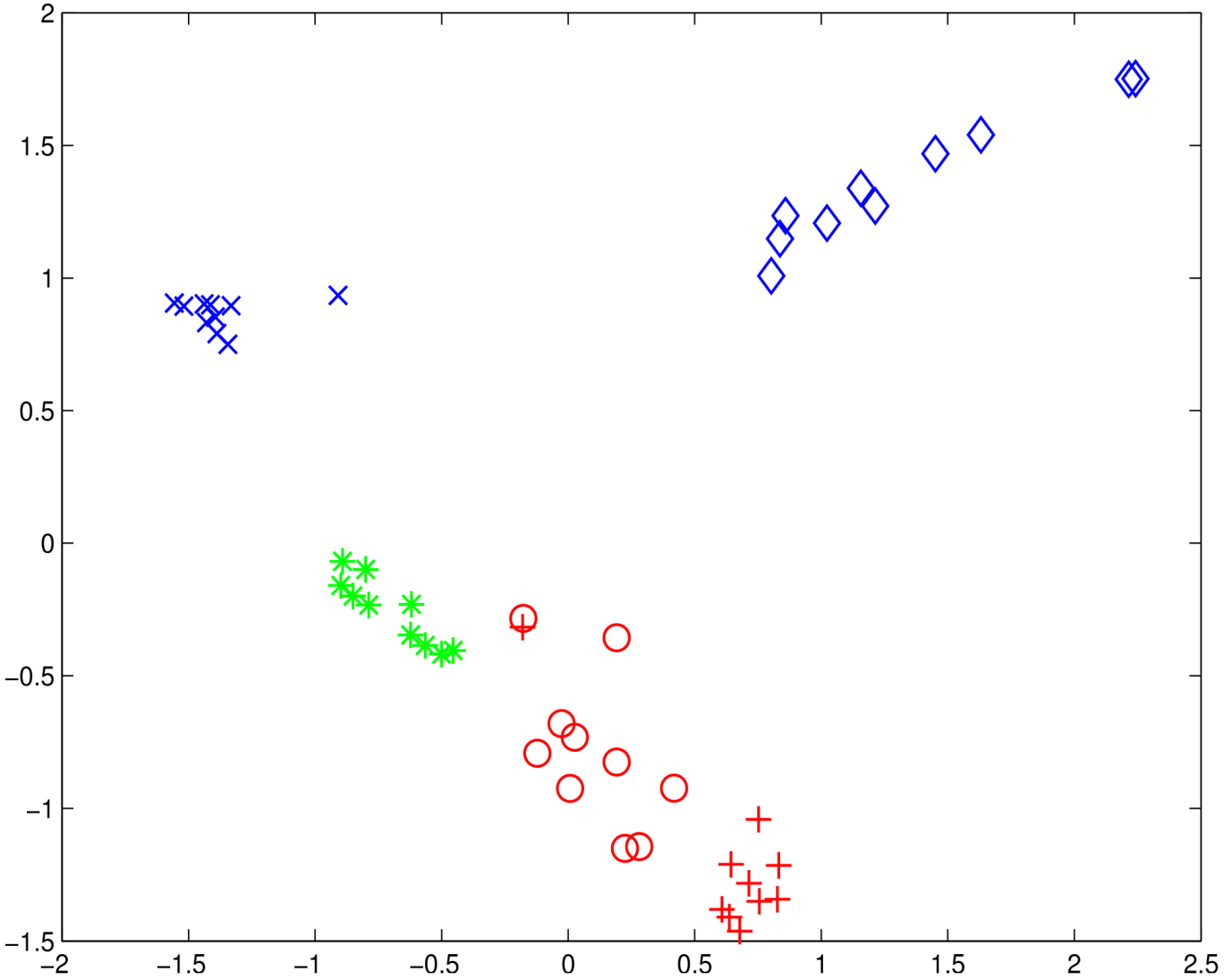
“red”: female, not bearded, non-glasses
“green”: male, not bearded, non-glasses
“blue”: male, bearded, glasses

Result of our approach: Iteration #1



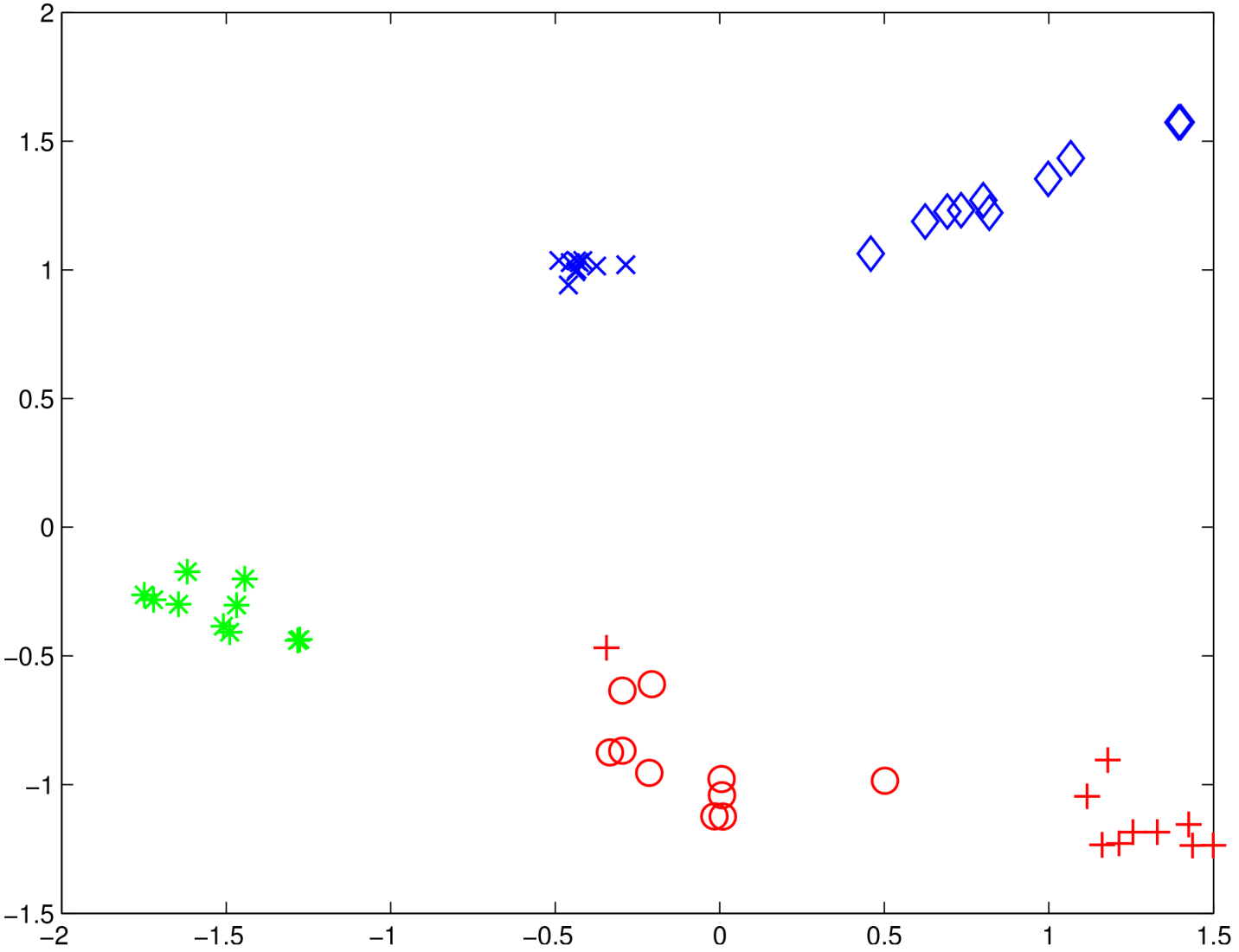
“red”: female, not bearded, non-glasses
“green”: male, not bearded, non-glasses
“blue”: male, bearded, glasses

Result of our approach: Iteration #2



“red”: female, not bearded, non-glasses
“green”: male, not bearded, non-glasses
“blue”: male, bearded, glasses

Result of our approach: Iteration #3



“red”: female, not bearded, non-glasses
“green”: male, not bearded, non-glasses
“blue”: male, bearded, glasses

3. Algorithm

3.1 Objective function

Assuming both data point and label vector can be represented by a weighted linear combination of the corresponding nearest neighbors, we adopt reconstruction error for the cost function

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

Weight matrix W is invariant to rotation and rescaling, enforce the sum-to-one constraint to guarantee the invariance to translation.

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

3.2 Alternating Optimization

While simultaneously recovering both unknowns is intractable, instead, we solve each unknown optimally (assuming the other unknown is a constant) in closed form, and create a two-step iterative approach.

i) Learning weight matrix

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

ii) Label inference

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

3.3 Algorithm Summary

Input:

data points, initial labels, parameters (k, α, H)

Training Stage:

Repeat

i) learning weight matrix

ii) label inference

Until F stabilized

Output:

weight matrix, predicted labels

3.4 Spectral Embedding

We exploit reconstruction error as the cost function for embedding

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

The minimization problem can be solved as a sparse eigen decomposition problem

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

where $M = (I-W)^T(I-W)$

Let d denote the desired dimension, the optimal embedding can be recovered by computing the bottom d+1 eigenvectors of matrix M, then discard the smallest eigenvector which is an unit vector with all equal components (free translation mode of eigenvalue zero).

4. Discussion

4.1 Weak Prior Knowledge

By weak prior knowledge, we mean:

i) an instance could be only partially labeled.

ii) there may be considerable noise scattered in labeled data.

A reasonable solution is to relax the constraint by adding a local fitting penalty term to the cost function, which allows slight changes to the initial labels.

4.2 Convergence

There is no theoretical guarantee that our approach will converge.

Two straightforward solutions:

i) fix the confident predictions

ii) set up a small tolerance

In experiment, we observed that both quick convergence and high learning accuracy can be reached by choosing appropriate parameters.

5.1 Experimental Settings

Dataset

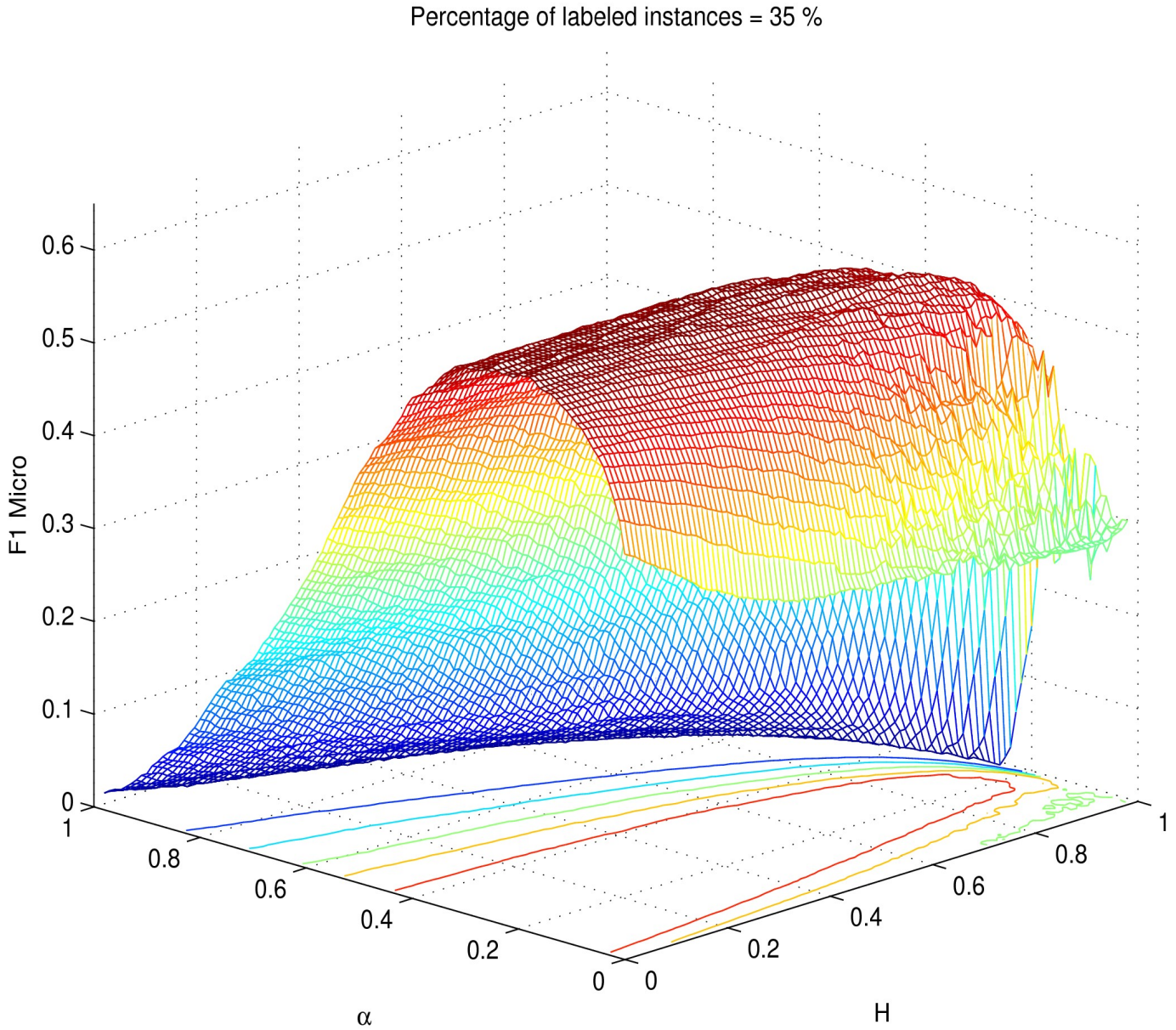
Yeast: gene dataset, 2417 samples, dimension 103, average # labels: 4.24
Scene: image dataset, 2407 images, dimension 294, average # labels: 1.07
SIAM TMC 2007: text dataset, 28,596 text samples, dimension 30,438, average # of labels: 2.21

Competitors

RankSVM: a state-of-the-art supervised multi-label classifier
PCA+RankSVM: perform PCA as a separate step before learning
ML-GFHF: an multi-label version of harmonic function (two dimensional optimization)

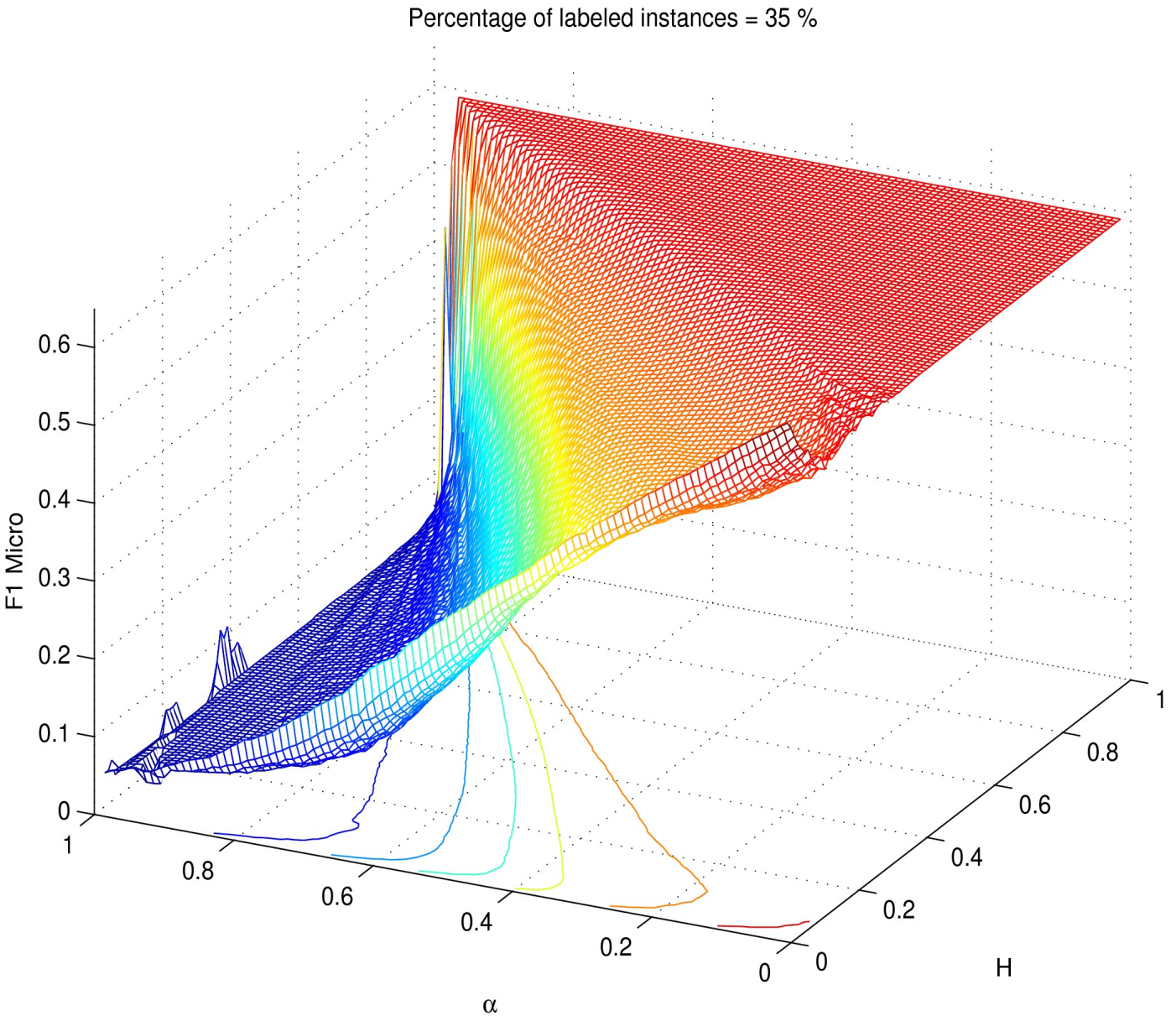
5.2 Parameter Selection

Scene: F1 Micro with respect to α and H



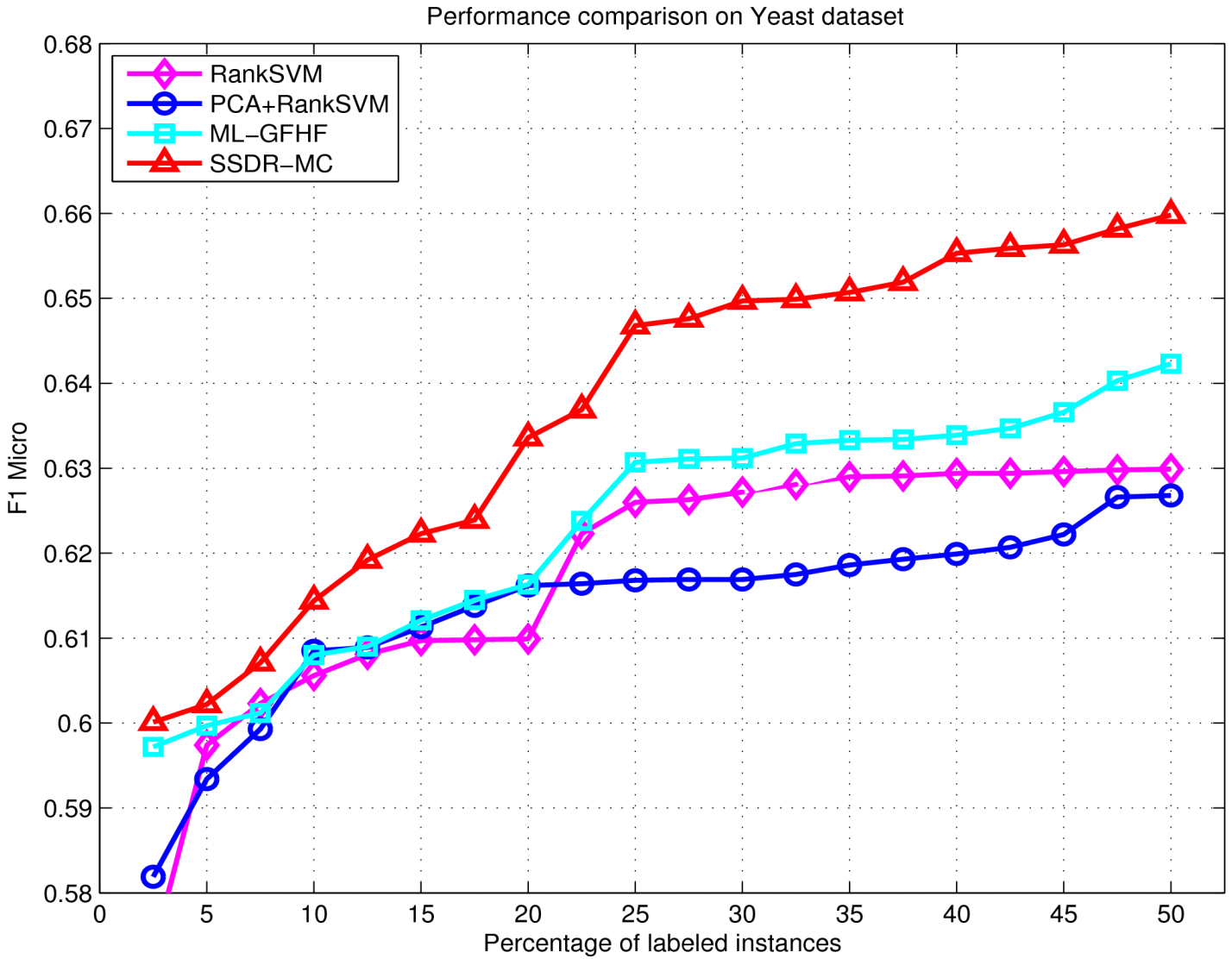
5.2 Parameter Selection Continued

Yeast: F1 Micro with respect to α and H



5.3 Experimental Result

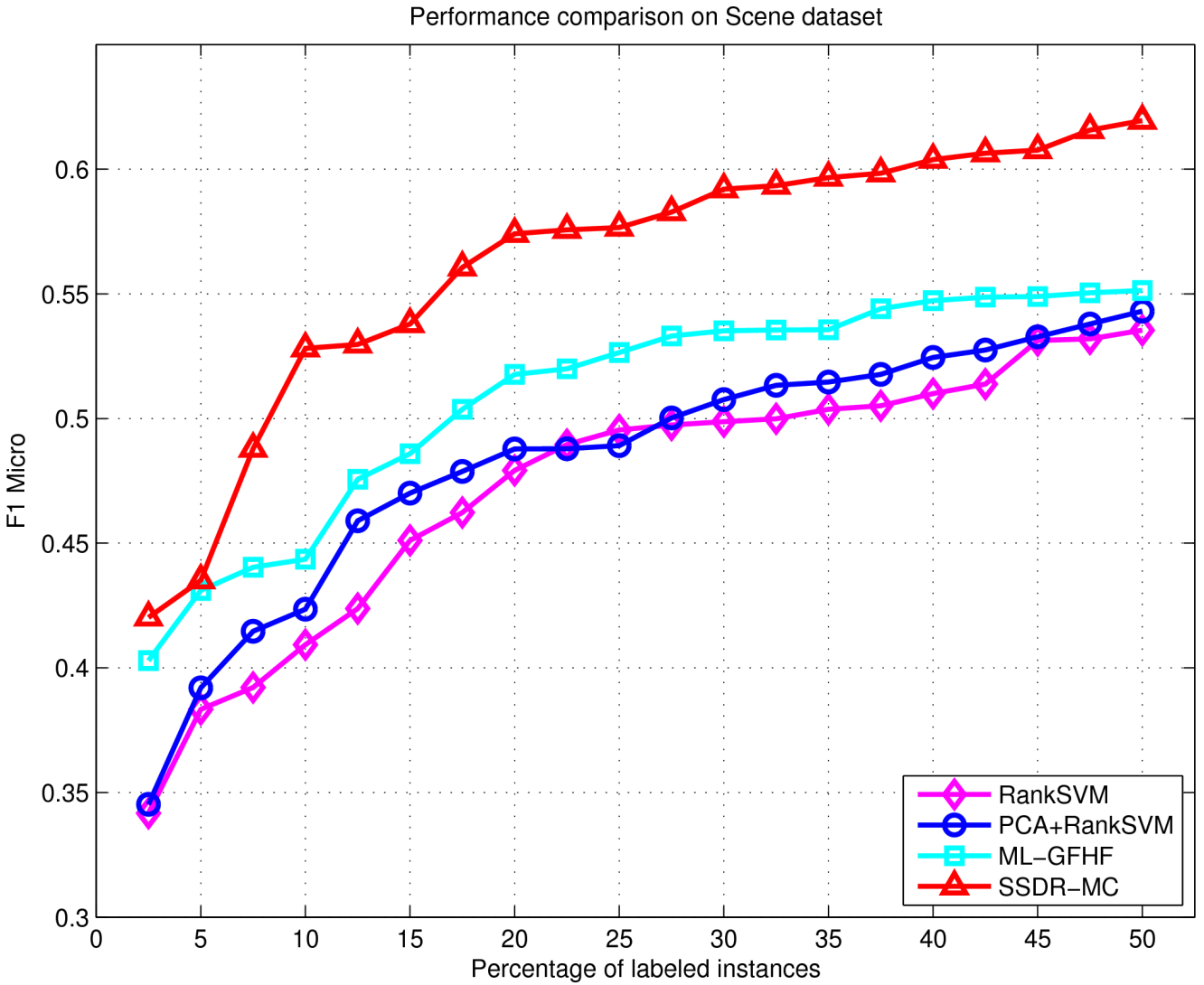
The portion of labeled instances gradually increases from 2.5% to 50%



Performance Comparison on **Yeast**

5.3 Experimental Result Continued

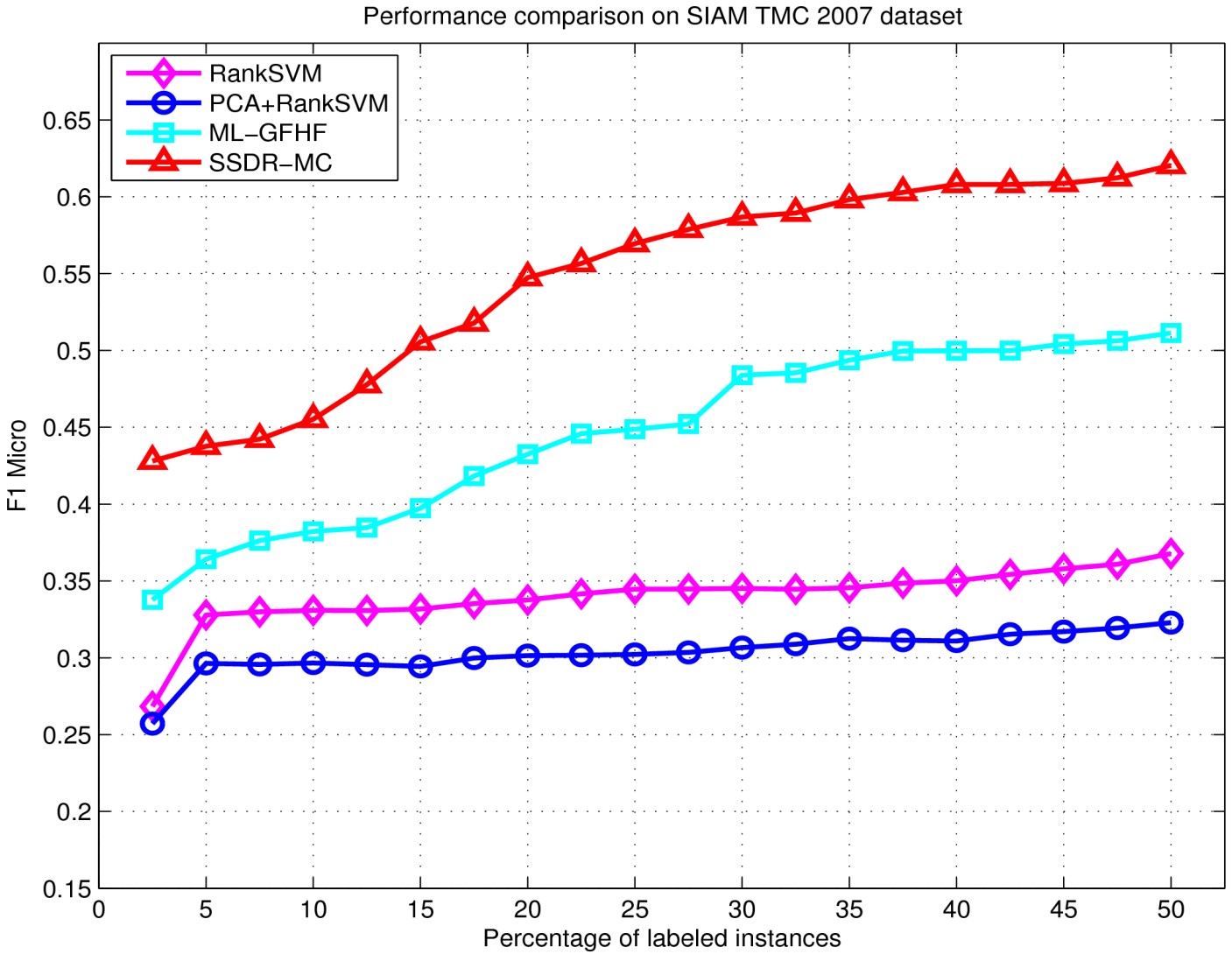
The portion of labeled instances gradually increases from 2.5% to 50%



Performance Comparison on **Scene**

5.3 Experimental Result Continued

The portion of labeled instances gradually increases from 2.5% to 50%



Performance Comparison on **SIAM TMC 2007**

6. Conclusion

As applications in data mining and machine learning move towards demanding domains, they must move beyond the restriction of complete supervision, single-label and low-dimensional data.

In this work, we establish the connection between dimension reduction and learning by an alternating optimization procedure:

1) learn a weight matrix from both feature description and available labels.
2) Infer the missing labels based on the weight matrix.

It can be viewed as simultaneously solving for two sets of unknowns: filling in the missing labels and identifying the projection that makes data points with similar labels close together while move points with different labels far apart.

Thank you for listening!
Question?