Published in final edited form as:

IEEE EMBS Int Conf Biomed Health Inform. 2017 February ; 2017: . doi:10.1109/bhi.2017.7897191.

Improving Multi-class Classification for Endomicroscopic Images by Semi-supervised Learning

Hang Wu[1] **[Student Member, IEEE]**, **Li Tong**[1] **[Student Member, IEEE]**, **May D. Wang**[1] **[Senior Member, IEEE]**

[1.]Dept. of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA 30332, USA

Abstract

Optical Endomicroscopy (OE) is a newly-emerged biomedical imaging modality that can help physicians make real-time clinical decisions about patients' grade of dysplasia. However, the performance of applying medical imaging classification for computer-aided diagnosis is primarily limited by the lack of labeled images. To improve the classification performance, we propose a semi-supervised learning algorithm that can incorporate large sets of unlabeled images. Our real-world endo-microscopic imaging datasets consist of 425 labeled images and 2,826 unlabeled ones. With semi-supervised learning algorithms, we improved multi-class classification performance over supervised learning algorithms by around 10% in all evaluation metrics, namely precision, recall, F1 score and Cohen-Kappa statistics.

Introduction

Optical endomicroscopy (OE) is a newly-emerged biomedical imaging modality that can help physicians make real-time clinical decisions about patients' grade of dysplasia. Applying medical imaging processing, specifically image classification techniques, have shown the potential to assist physicians in their decision-making process. Such classification works by first obtaining a feature representation for each labeled image and then applying supervised learning to learn a proper mapping from the features of images to their most likely labels.

Despite the simplicity of the pipeline, the performance of such pipeline is limited by the nature of medical imaging applications and supervised learning algorithms. As suggested by statistical learning theory (Vapnik 2013), the performance of supervised learning algorithms decreases as the number of training data decreases. In medical imaging applications, however, it is very costly and practically impossible for physicians to manually annotate each image collected for patients, resulting in a limited size dataset of labeled images and a large collection of unlabeled images.

To overcome this obstacle, a natural question to ask is: *Can we incorporate unlabeled images to improve existing classification performance?* Semi-supervised learning, which proposes a

Corresponding author: May D. Wang maywang@gatech.edu.

framework to augment supervised learning algorithms with unlabeled data, is perfect for the task. In this paper, we proposed a fast version of a kernel-based semi-supervised learning algorithm, called label spreading (D. Zhou et al. 2003), by approximating the kernel matrix with a lower rank matrix. The proposed modification is shown to speed up the algorithm by a large magnitude while obtaining almost identical performance.

Method

Problem Formulation

Given a set of images represented their feature vectors in $\mathbb{R}^D$, $\mathcal{X} = \{x_1, \ldots, x_L, x_{L+1}, \ldots, x_N\}$, the first L images $\mathcal{X}_L = \{x_1, \ldots, x_L\}$ are labeled as $Y_L = \{y_1, \ldots, y_L\}$ from a multi-class label set $\mathcal{C} = \{1, \ldots, C\}$, and the rest $N - L$ images $\mathcal{X}_U = \{x_{L+1}, \ldots, x_N\}$ are unlabeled. Our goal is to improve prediction results incorporating unlabeled images.

Algorithm: Label Spreading

We will base our algorithm on one successful semi-supervised learning algorithm called Label Spreading (D. Zhou et al. 2003). Let $\mathcal{F}$ denote the label matrix of $\mathbb{R}^{N \times C}$, where each entry $\mathcal{F}_{ij} = p(y_i = j)$ denotes the probability of data point i is from class j. With $\mathcal{F}$, we can classify each data point i to a label $\hat{y}_i = \operatorname{argmax}_j \mathcal{F}_{ij}$, and we can further define the prediction matrix $\hat{Y} = \left[\hat{Y}_{ij}\right]$ where $\hat{Y}_{ij} = 1$ if $\hat{y}_i = j$ and 0 otherwise.

The algorithm works as follows:

Form the affinity matrix W defined by $W_{ij} = \kappa(x_i, x_j)$, where κ is some kernel function measuring the similarity between two data points

Construct the matrix $S = D^{-1/2} W D^{-1/2}$, where D is the diagonal matrix with $D_{ii} = \sum_j W_{ij}$

Iterate

$$F(t+1) = \alpha S F(t) + (1 - \alpha) Y$$

until convergence

Let F^* denote the limit of the sequence $\{F(t)\}$.

Label each point x_i as $y_i = \operatorname{argmax}_j F^*_{ij}$

The paper then proceeds to develop a closed-form solution for the problem, which writes as $F^* = (I - \alpha S)^{-1} Y$. In practice, when the scale of the problem is not too large, we could directly invoke the matrix inversion formula. With a large number of samples, however, it will be beneficial to use the iterative version of the algorithm, because matrix inversion requires $O(N^3)$ time complexity, and the iterative version has less time complexity.

The algorithm has connections to the well-known spectral clustering (Luxburg 2007), and actually, the first steps are the same as spectral clustering. Unlike distance-based clustering, spectral clustering focuses on the connectivity of data points as in a graph, and can divide data into disconnected subgraphs. For example, as shown in Figure 1, we have two clusters of data points that are disconnected, and spectral clustering can perfectly discover such patterns, while distance-based clustering, for example, K-means algorithms, fails to correctly classify data points that match human intuition.

For the case of semi-supervised learning, unlabeled data points and labeled data points can also form a graph structure together, whose edges indicate similarities between each pair of them. Known label information is propagated in the graph so that each unlabeled data point can be classified based on the combined information from its neighbors.

Nyström Approximation of the Affinity Matrix

One of the major limitation of the algorithm above lies in the construction of the affinity matrix W, as the calculation of W requires $N(N-1)/2$ pairwise comparisons between each pair of the N samples.

We propose to accelerate the algorithm by introducing a low-rank approximation to the affinity matrix W, using the well-established technique called Nyström approximation (Williams and Seeger 2000).

To approximate $\hat{W}$, we first randomly sample m columns of W without replacement, and then compute two sub-matrices, namely $W_{m,m}$, W_{n_m}, which are the corresponding blocks of the original matrix W. Afterward, we set $\hat{W}$ to $W_{n,m} W_{m,m}^{-1} W_{m,n}$, which helps reduce the time complexity from $O(N^3)$ to $O(m^2 N)$. We can illustrate the process as in Figure 2. In practice, the method has shown to improve speed while not significantly decreasing the accuracy (Williams and Seeger 2000).

In summary, our modified label spreading algorithm uses the low-rank approximation of W in the first step of the original algorithm, and our following experiments show that this seemingly simple modification improve the prediction accuracy significantly.

Experiments

Dataset Description

Our endo-microscopic dataset contains images collected from patients undergoing endoscope-based Confocal Laser Endomicroscopy (eCLE) procedures for BE. In total, 425 images were labeled as one of the eight classes by an expert gastrointestinal endomicroscopists and another 2,826 unlabeled were used in the semi-supervised learning algorithms. Although binary classification of patients into high- and low-risk groups are easier and more common in practice, understanding the exact subtypes of OE images can help physicians detect BE associated neoplasia and suggest better treatment plans (Sharma, Meining, and others 2011).

Features are extracted using SIFT (Lowe 1999), which is verified by our previous work to work better for endomicroscopic images (Kothari et al. 2016). After the extraction, each image is represented as a vector in $\mathbb{R}^{500}$.

Experiment Configuration

We use common multi-class classification metrics, namely precision, recall, F1 score, and Cohen-Kappa Score (Smeeton 1985). The first three are all computed by a weighted average of the original metrics of all classes, and range in [0,1], with 1 denoting a perfect classification. Cohen-Kappa Score measures the inter-rater agreement for categorical items, and it ranges between −1 and 1 and the higher the better.

For our algorithm, we use an approximation rank of 50, which is much smaller than the full size 2000, and the maximum number of iterations to be 100.

In our experiments, we compare our algorithm to widely adopted supervised classification algorithms, namely support vector machine (SVM), random forests(RF), and logistic regression(LR), which are implemented by the Scikit-Learn Package (Pedregosa et al. 2011). We also compare our algorithms to the original version of label spreading (denoted as ‘LP’).

Results: Classification Performance

Figure 3 shows the multi-class classification performance of our algorithms against standard supervised learning algorithms. Among four metrics, the semi-supervised algorithms both outperform common baselines by a large margin and increase the performance by around 11%, which shows that when labeled data is limited, introducing unlabeled dataset indeed improves the classification performance.

Between our proposed modification and the original label spreading algorithm, the performance is almost identical, which indicates that a rank = 50 approximation is close enough for the algorithm to perform almost the same as the full matrix version. We will further study the influence of the rank on the results in the next section.

Results: On the Rank of approximation

We vary the rank of approximation from 50 to 1000, then measure the prediction performance and running time. The results presented here are averaged across ten runs, and all parameters except rank are selected the same.

We plotted four performance measures against rank in Figure 4 and also fit a trend line using linear regression. As the rank increases, the prediction performance slights decreases. And the best performing rank is around 50, comparable to the results obtained without low-rank approximation.

As for the running time, since based on basic time complexity analysis, we know the time required for such approximation is $O(m^2N)$, where m is the rank, we fit a quadratic regression trend line on the figure, and the resulting Figure 5 verifies our theoretical analysis. The y-axis of Figure 5 is $T(approximation)/T(full\ rank)$, which showed the ratio of using low-rank approximation to using full affinity matrix. If we use the full rank matrix instead of pairwise approximation, the running time was recorded as around 120s, and the speed-up we obtained from the approximation with rank = 50, is about 40 times.

Related Work

Multi-class classification, where each sample comes from one of the C categories ($C > 2$), is an important yet challenging tasks. One of the most popular approach to convert binary classification algorithms, including support vector machine and logistic regression, to multi-class case, where several binary-class estimators are constructed. Typical conversion schemes are one-versus-rest and one-versus-one, and we refer readers to (Bishop 2006) for detailed introduction. Several algorithms naturally support multi-class classification, such as k-nearest neighbor, Naïve Bayes, decision trees, and random forests. Recently researchers have researched into how to explore the hierarchy on the output class space. For example, Choromanska et al. (Choromanska and Langford 2015) developed a tree construction approaches on the class hierarchies and achieves a logarithmic time complexity in online prediction. Bengio et al. (Bengio, Weston, and Grangier 2010) improved the classification by learning embedding for each of the class and constructing a tree-like classifier.

Semi-supervised learning aims to improve traditional classification algorithms by introducing unlabeled data, since labeled instances are often expensive to obtain (Zhu 2005). Conventional methods mainly fall into three categories: wrapper methods including self-training or co-training (Z. H. Zhou and Li 2005), transductive discriminative methods such as transductive SVM (Joachims 1999), and perhaps, the most well studied, graph-based algorithms (Liu and Chang 2009) including label spreading. With the vast popularity of deep learning, deep generative models have also been applied to semi-supervised learning (Kingma et al. 2014) and achieved great success in large-scale image classification tasks.

Kernel matrix approximation falls into the recently developed randomized algorithms for matrices and data (Boyd 2011). As the size of data grows even larger, it will be difficult to perform conventional analysis on a single machine, and randomized algorithms tackle the challenges by sampling a subset of the data and extrapolate to approximate the original matrices.

Conclusion & Future Work

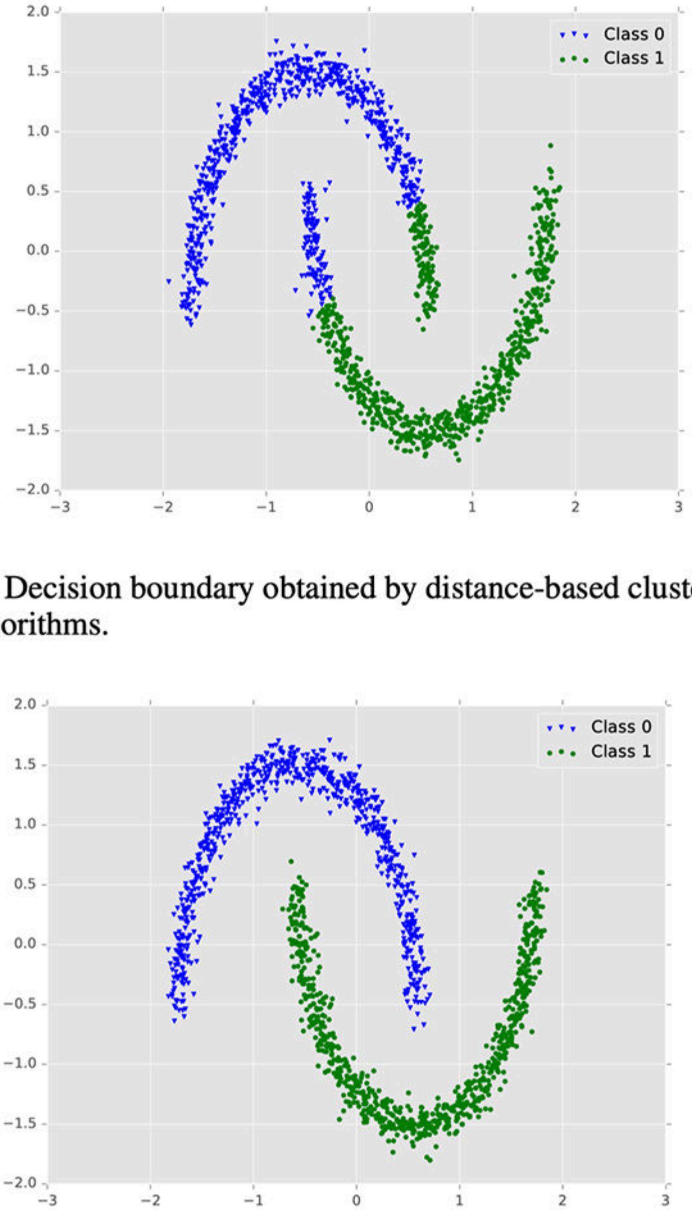
This work demonstrates that by adding unlabeled data into our prediction model, we can improve the prediction performance over traditional supervised learning algorithms significantly. The proposed method relies on computing the kernel matrix between all pairs of data and we speeded up the original method by introducing a fast low-rank approximation to the kernel matrix. Extensive experiments show that our solution is both fast and effective. The current work focuses on conventional kernel-based methods, and in future, it will be beneficial to introduce deep generative models, such as variational auto-encoders, to understand the generating mechanisms of the images we are analyzing.

Acknowledgement

We thank Dr. Kevin E. Woods for the help throughout the project and the dataset he provided. We also thank Hamid Hassanzadeh and Ryan Hoffman for their kind comments and suggestions on the manuscript.

References:

1. Bengio Samy, Weston Jason, and Grangier David. 2010 “Label Embedding Trees for Large Multi-Class Tasks,” 163–71.
2. Bishop CM. 2006 Machine learning and pattern recognition Information Science; Statistics. Springer.
3. Boyd Michael W Mahoney. 2011 “Randomized Algorithms for Matrices and Data.” Foundations and Trends in Machine Learning 3 (2): 123–224.
4. Choromanska Anna E, and John Langford. 2015 “Logarithmic Time Online Multiclass prediction,” 55–63.
5. Joachims T 1999 “Transductive inference for text classification using support vector machines.” ICML.
6. Kingma Diederik P, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. 2014 “Semi-supervised Learning with Deep Generative Models,” 3581–9.
7. Kothari S, Wu H, Tong L, and Woods KE. 2016 “Automated risk prediction for esophageal optical endomicroscopic images.” 2016 IEEE-EMBS ….
8. Liu Wei, and Chang Shih-Fu. 2009 “Robust multi-class transductive learning with graphs.” In 2009 Ieee Computer Society Conference on Computer Vision and Pattern Recognition Workshops (Cvpr Workshops), 381–88. IEEE.
9. Lowe DG . 1999 “Object recognition from local scale-invariant features.” Computer Vision.
10. Luxburg Ulrike von. 2007 “A tutorial on spectral clustering.” Statistics and Computing 17 (4): 395–416.
11. Pedregosa Fabian, Varoquaux Gaël, Gramfort Alexandre, Michel Vincent, Thirion Bertrand, Grisel Olivier, Blondel Mathieu, et al. 2011 “Scikit-learn: Machine Learning in Python.” Journal of Machine Learning Research () 12 (Oct): 2825–30.
12. Sharma Prateek, Meining Alexander R, and others. 2011 “Real-Time Increased Detection of Neoplastic Tissue in Barrett’s Esophagus with Probe-Based Confocal Laser Endomicroscopy.” Gastrointestinal Endoscopy 74 (3). Elsevier: 465–72. [PubMed: 21741642]
13. Smeeton NC. 1985 “Early history of the kappa statistic.”
14. Vapnik V 2013 “The nature of statistical learning theory.”
15. Williams Christopher K I, and Seeger Matthias W. 2000 “Using the Nyström Method to Speed Up Kernel Machines.” NIPS.
16. Zhou Dengyong, Bousquet Olivier, Navin Lal Thomas, Weston Jason, and Schölkopf Bernhard. 2003 “Learning with Local and Global Consistency.” NIPS.
17. Zhou ZH, and Li M. 2005 “Semi-Supervised Regression with Co-Training.” IJCAI.
18. Zhu X 2005 “Semi-supervised learning literature survey.”



(a) Decision boundary obtained by distance-based clustering algorithms.

(b) Decision boundary obtained by spectral clustering algorithms.

Figure. 1:

Different decision boundaries obtained two types of clustering algorithms show that spectral cluster.

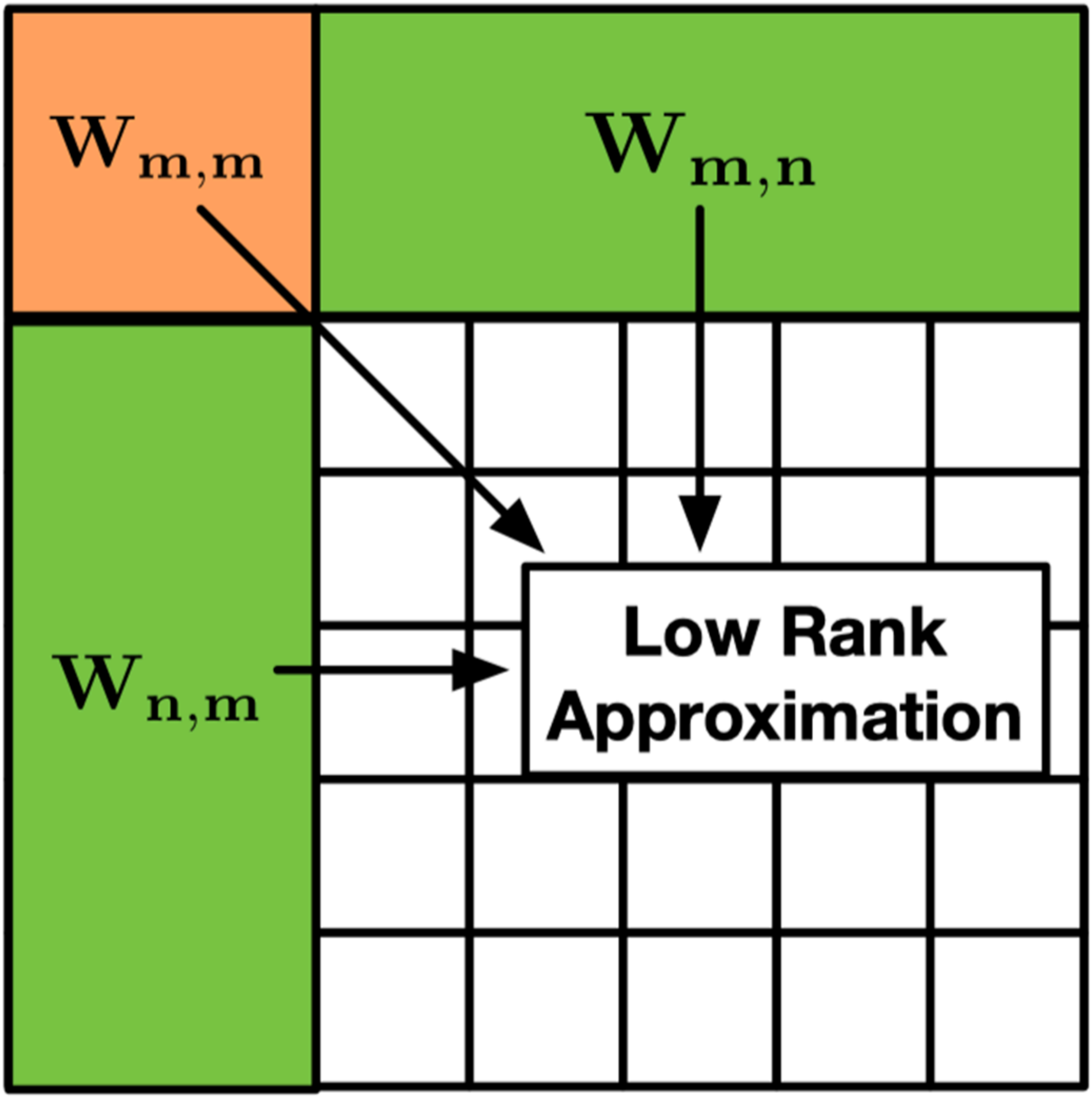


Figure. 2:

To perform Nystrom approximation of a large matrix, we only need to compute two relative small sub-matrices (colored in the figure), and then approximate the rest entries (white boxes in the figure) with basic matrix multiplication.

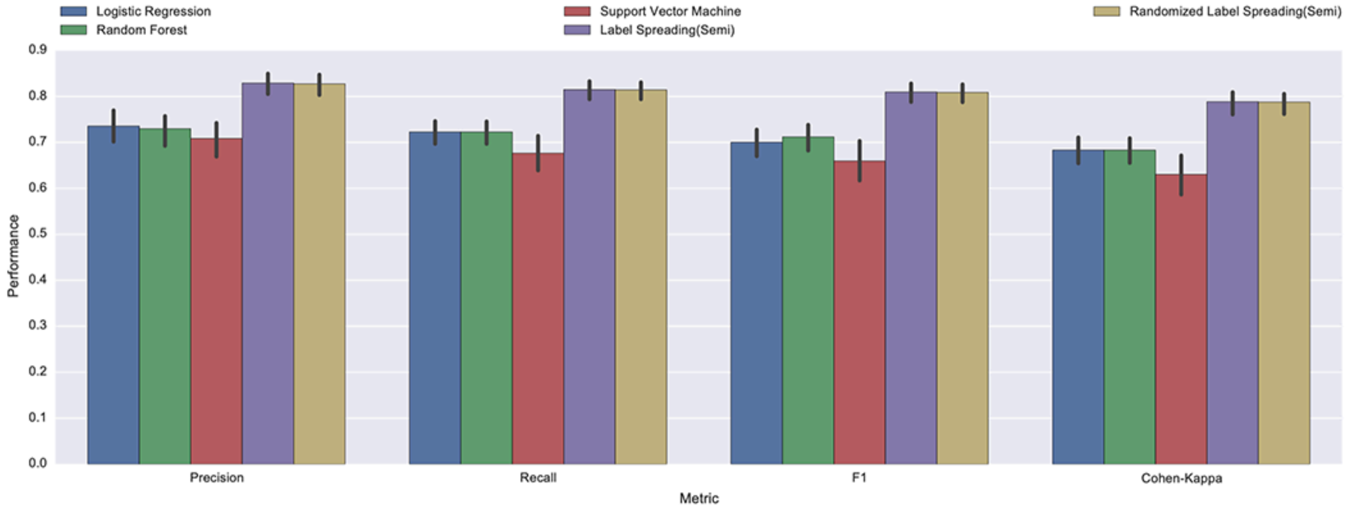


Figure. 3:

Semi-supervised learning algorithms (Label Spreading and our proposed Randomized Label Spreading) outperform supervised learning algorithms in terms of precision, recall, F1 and Cohen-Kappa statistics by a large margin.

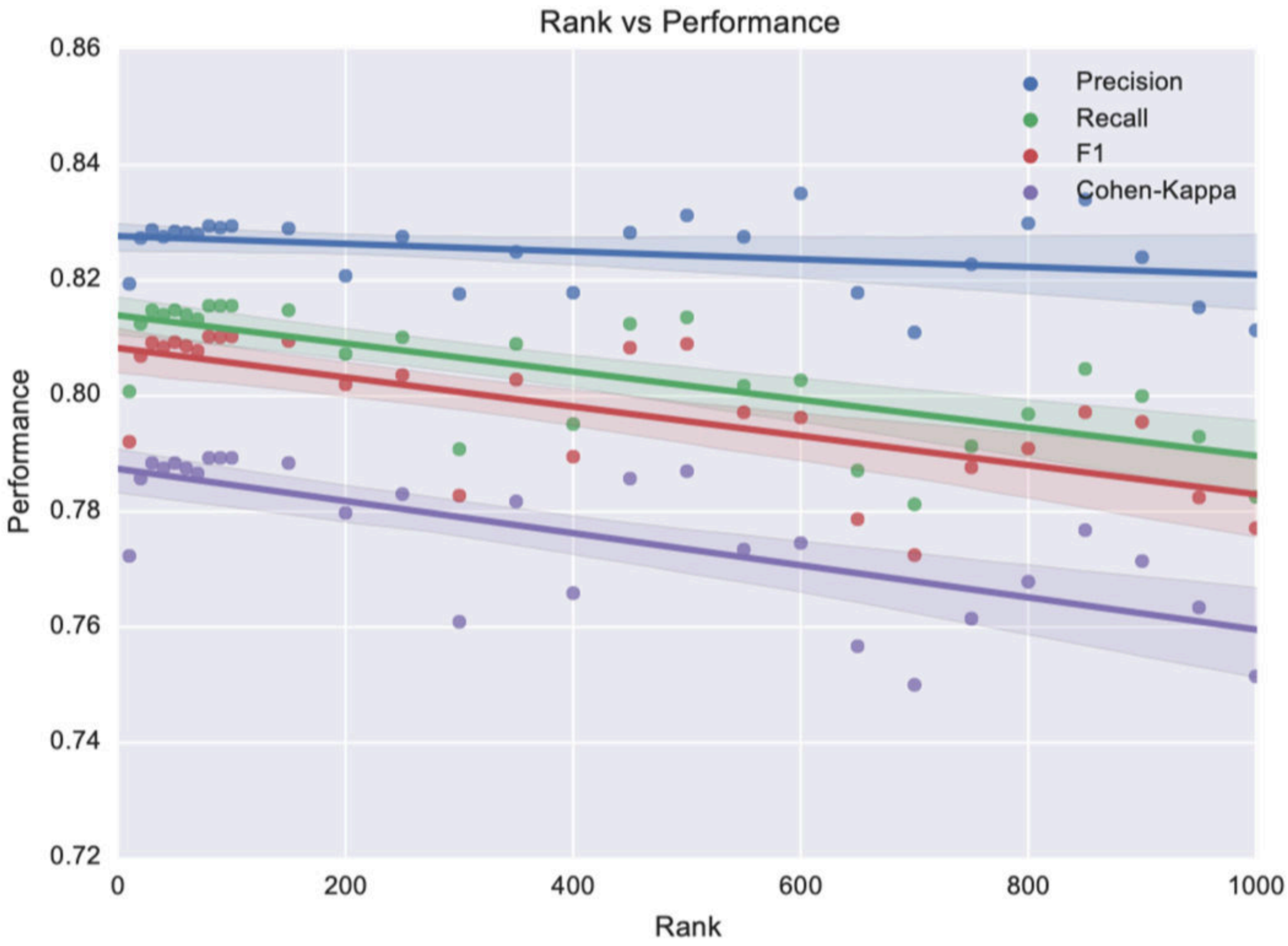


Figure. 4:
As the rank increases, classification performance decreases, yet still higher than purely supervised algorithms.

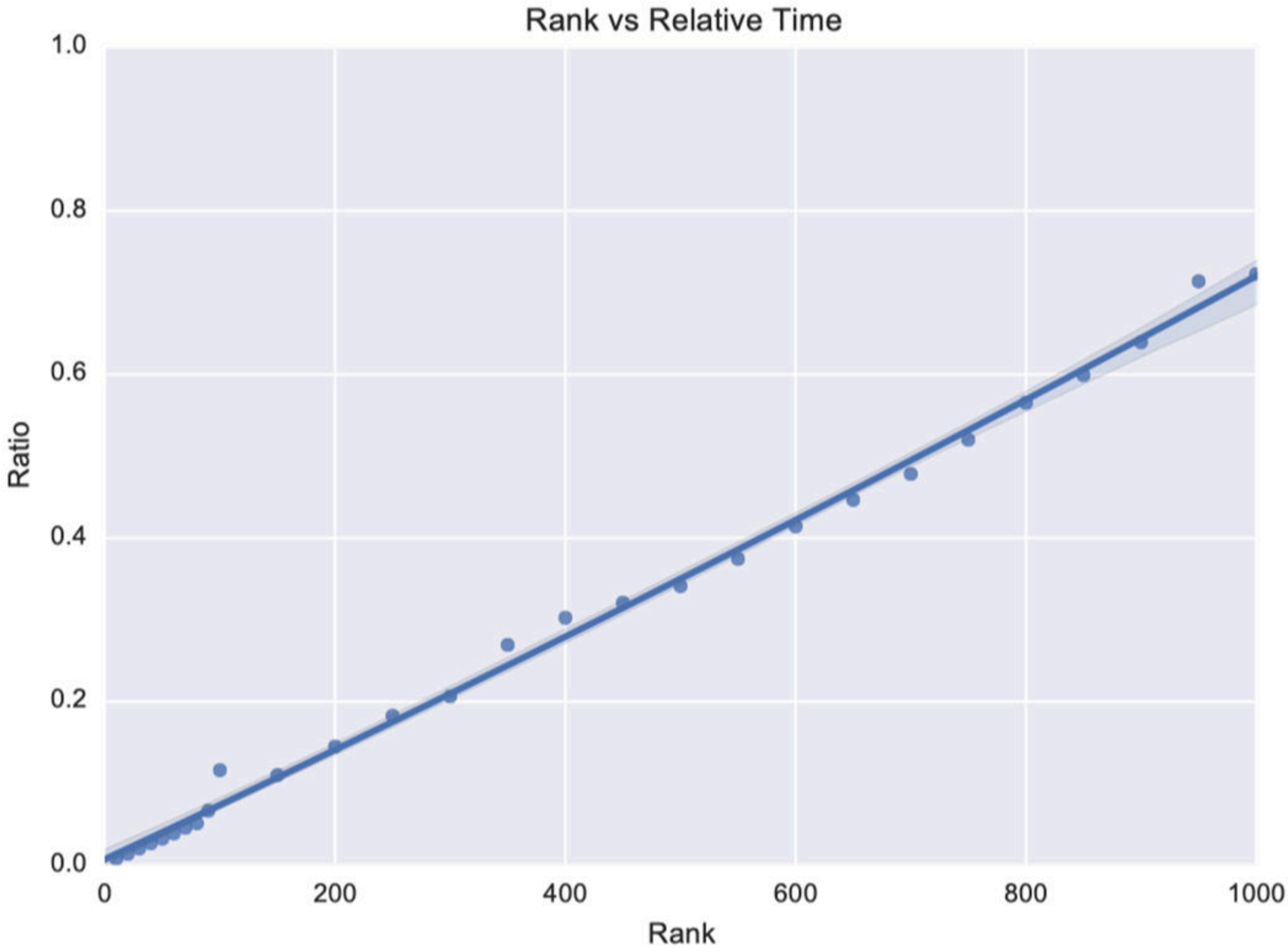


Figure. 5:

As the rank increases, running time increases quadratically accordingly. The regular full rank affinity matrix computation is averaged to be around 120s.