Regularized Semi-supervised Classification on Manifold

Lianwei Zhao[1], Siwei Luo[1], Yanchang Zhao[2], Zhihai Wang[1]

[1] School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China
lw_zhao@126.com
[2] Faculty of Information Technology, University of Technology, Sydney, Australia

Abstract. Semi-supervised learning gets estimated marginal distribution P_X with a large number of unlabeled examples and then constrains the conditional probability $p(y \mid x)$ with a few labeled examples. In this paper, we focus on a regularization approach for semi-supervised classification. The label information graph is first defined to keep the pairwise label relationship and can be incorporated with neighborhood graph which reflects the intrinsic geometry structure of P_X . Then we propose a novel regularized semi-supervised classification algorithm, in which the regularization term is based on the modified Graph Laplacian. By redefining the Graph Laplacian, we can adjust and optimize the decision boundary using the labeled examples. The new algorithm combines the benefits of both unsupervised and supervised learning and can use unlabeled and labeled examples effectively. Encouraging experimental results are presented on both synthetic and real world datasets.

1 Introduction

The problem of learning from labeled and unlabeled examples has attracted considerable attention in recent years. It can be described as follows: with l labeled examples $M = \{x_i, y_i\}_{i=1}^{l}$ drawn from an unknown probability distribution $P_{X\times Y}$ and u unlabeled examples $\{x_j\}_{j=l+1}^{l+u}$ drawn from the marginal distribution P_X of $P_{X\times Y}$, how to learn $P_{X\times Y}$ by exploiting the marginal distribution P_X ? It is also known as semi-supervised learning, and a number of algorithms have been proposed for it, including Co-training [6], random field models [7,8] and graph based approaches [9, 10].

However, learning from examples has been seen as an ill-posed inverse problem [11] and regularizing the inverse problem means finding a meaning stable solution, so in this paper we focus on regularization approaches. Measure based regularization [12] assumes that two points connected by a line going through high density region should have the same label. Based on this assumption, the regularizer is weighted with data density. The idea of information regularization [13] is that labels should not

change too much in regions where marginal is high, so regularization penalty that links marginal to the conditional distribution is introduced and it is expressed in terms of mutual information $I(x;y)$ as a measure of label complexity. Both of the above two methods take density into consideration, and can get the decision boundary that lies in the region of low density in 2D example. However, it is difficult to apply them in high-dimensional real world data sets.

Manifold regularization [1-4] assumes that two points close in the input space should have the same label, and exploits the geometry of the marginal distribution to incorporate unlabeled examples within a geometrically motivated regularization term. However, after incorporating an additional regularization term, there are two regularization parameters. It not only makes it difficult to find a solution, but needs improvement in theory. In addition, how to choose appropriate values for regularization parameters is a new problem.

In this paper, we first define the label information graph, and then incorporate it with neighborhood graph. Based on modified Graph Laplacian regularizier, we propose a novel regularized semi-supervised classification algorithm. There is only one regularization parameter reflecting the tradeoff between the Graph Laplacian and the complexity of solution. The labeled examples can be used to redefine the Graph Laplacian and further to adjust and optimize the decision boundary. Experimental results show that our algorithm can use unlabeled and labeled examples effectively and is more robust than Transductive SVM and LapSVM.

This paper is organized as follows. Section 2 briefly reviews Graph Laplacian and semi-supervised learning assumption. In section 3, we define label information graph with labeled examples and propose the regularized semi-supervised classification algorithm. Experimental results on synthetic and real world data are shown in section 4, followed by conclusions in section 5.

2 Related Works

2.1 Graph Laplacian

Graph Laplacian [5] has played a crucial role in several recently developed algorithms [14,15], because it approximates the natural topology of data and is simple to compute for enumerable based classifiers. Let's consider a neighborhood graph $G = (V, E)$ whose vertices are labeled or unlabeled example points $V = \{x_1, x_2, \cdots, x_{l+u}\}$ and whose edge weights $\{W_{ij}\}_{i,j=1}^{l+u}$ represent appropriate pairwise similarity relationship between examples. The neighborhood of x_j can be defined as those examples which are closer than ε or the k nearest neighbors of x_j. To ensure that the embedding function f is smooth, a natural choice is to get empirical estimate $I(G)$, which measures how much f varies across the graph:

$$I(G) = \frac{1}{2\sum_{i,j} W_{ij}} \sum_{i,j=1}^{l+u} (f(x_i) - f(x_j))^2 W_{ij} \quad \textbf{(1)}$$

where $2\sum_{i,j} W_{ij}$ is normalizing factor, so that $0 \le I(G) \le 1$.

Defining $\hat{f} = [f(x_1, \cdots, f(x_{l+u})]^T$, and $L = D - W$ as Graph Laplacian matrix, where D is diagonal matrix given by $D_{ii} = \sum_{j=1}^{l+u} W_{ij}$, $I(G)$ can be rewritten as:

$$I(G) = \frac{1}{2\sum_{i,j} W_{ij}} \hat{f}^T L \hat{f} \quad \textbf{(2)}$$

2.2 Semi-supervised Learning Assumptions

In the semi-supervised learning framework, the marginal distribution P_X is unknown, so we must get empirical estimates of P_X using a large number of unlabeled examples and then constrain the conditional $p(y \mid x)$ with a few labeled examples. However, there is no identifiable relation between the P_X and the conditional $p(y \mid x)$, so the relationship between them must be assumed. Manifold regularization[1,2] assumes that two points that are close in the input space should have the same label. In other words, the conditional probability distribution $p(y \mid x)$ varies smoothly along the geodesics in the intrinsic geometry of P_X.

3 ReguSCoM: Regularized Semi-supervised Classification on Manifold

3.1 Our Motivation

We have noticed that the knowledge of the joint probability distribution $p(x, y)$ is enough to achieve perfect classification in supervised learning. We divide the process of semi-supervised learning into two steps. Firstly we get the empirical estimates of the marginal distribution P_X using both labeled and unlabeled examples and estimate $\hat{p}(y \mid x)$ according to the information carried about the distribution of labels. Secondly, we adjust $\hat{p}(y \mid x)$ to $p(y \mid x)$ using a few labeled examples and then get $p(x, y) = p(y \mid x) p(x)$. The first step can be considered as semi-supervised classification, while the second step is supervised learning.

We have assumed that if two points $x_1, x_2 \in X$ are close in the input space, then the conditional $p(y \mid x_1)$ and $p(y \mid x_2)$ are near in intrinsic geometry of P_X. In manifold regularization [1] this assumption is represented by adjacency matrix, i.e., edge weights $\{W_{ij}\}_{i,j=1}^{l+u}$. However, this adjacency matrix doesn't take into consideration the information carried by labeled examples. The regularization term $I(G)$, especially for binary case classifiers, is proportional to the number of separated neighbors, that is, the number of connected pairs that are classified differently by decision boundary. Therefore for labeled examples x_i and x_j, if they are of the same label, they should not be separated by the decision boundary, so we can redefine the relationship between x_i and x_j by strengthening it. If x_i and x_j have the different labels, we can weaken it.

3.2 Definition of Label Information Graph

In the manifold learning, one of the key assumptions is that the data lie on a low dimensional manifold M and this manifold can be approximated by a weighted graph constructed with all the labeled and unlabeled examples. So the performance of the learning algorithm significantly depends on how the graph is constructed.

We consider all the sample points $\{x_1, x_2, \cdots, x_{l+u}\}$, including both the labeled and unlabeled examples. When the support of P_X is a compact submanifold M, the geometry structure can be approximated using the Graph Laplacian with both labeled and unlabeled examples. The Least Squares algorithm solves the problem with the squared loss function $\sum_{i=1}^{l} V(x_i, y_i, f) = \sum_{i=1}^{l} (y_i - f(x_i))^2$, which is based on the minimizing the error on the labeled examples. It is important to observe that

$$\begin{aligned} 2(l-1)\sum_{i=1}^{l}(y_i - f(x_i))^2 &\geq \sum_{i,j=1}^{l}((y_i - f(x_i)) - (y_j - f(x_j)))^2 \\ &= \sum_{i,j=1}^{l}((f(x_i) - f(x_j)) - (y_i - y_j))^2 \end{aligned} \tag{3}$$

If $\sum_{i=1}^{l}(y_i - f(x_i))^2 \to 0$, then

$$\sum_{i,j=i}^{l}((f(x_i) - f(x_j)) - (y_i - y_j))^2 \to 0 \tag{4}$$

So if $|y_i - y_j| < \delta$, then $|f(x_i) - f(x_j)| < \varepsilon$, where $\delta, \varepsilon \to 0$, *and* $\delta, \varepsilon > 0$.

We define $(l+u) \times (l+u)$ matrix J as follows.

$$J_{ij} = \begin{cases} 1 \text{ or } \quad W_{ij}, & \text{if } \quad i, j \le l \text{ and } \left| y_i - y_j \right| < \delta \\ 0 \text{ or } -W_{ij}, & \text{if } \quad i, j \le l \text{ and } \left| y_i - y_j \right| \ge \delta \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

This can be seen as a label information graph $G' = (V, E')$, whose vertices are the labeled or unlabeled example points $V = \{x_1, x_2, \cdots, x_{l+u}\}$ and whose edge weights J_{ij} represent appropriate pairwise label relationship between labeled examples i and j.

According to the label information graph, the right of the equation 3 can be rewritten as follows:

$$\sum_{i,j=1}^{l} ((f(x_i) - f(x_j)) - (y_i - y_j))^2 = \sum_{i,j=1}^{l+u} (f(x_i) - f(x_j))^2 J_{ij} \tag{6}$$

This term can be seen as label information carried by labeled examples and penalizes classifiers that separate the examples having the same labels.

Remark: In graph G', weight J_{ij} just represents appropriate pairwise label relationship between i and j. If labeled example i has the same label as j, they should not be separated by decision boundary. This relationship must not be represented only by element J_{ij}. For example, for large scale problems, this relationship J_{ij} can be represented by a geodesic path $J_{ik_1}, J_{k_1k_2}, \cdots J_{k_nj}$, which can be computed by finding a shortest path $(i, k_1, k_2, \cdots k_n, j)$ from i to j in graph G'.

3.3 Classifier Based on the Modified Graph Laplacian

In this section, we consider the problem of using the manifold structure to improve the performance of the classifier f, where $f : X \in M \to Y$. In most situations, the manifold is approximated by a graph constructed with all examples and f is defined on the vertices of the graph, so a stabilizer is necessary. An important class of stabilizers is squares of norms on reproducing kernel Hilbert spaces (RKHS). The squared norm $\|f\|_K^2$ is used as stabilizer to penalize high oscillation of various types. The geometry structure of the marginal distribution P_X is incorporated as a regularization term based on the neighborhood graph [1,2]. In order to exploit the label information, equation 6 is also introduced as a penalty term based on the label information graph.

The neighborhood graph and the label information graph have the same vertices and can be incorporated together. So the optimization problem has the following objective function:

$$\min_{f \in H} H[f] = \gamma \|f\|_K^2 + \sum_{i,j=1}^{l+u} (f(x_i) - f(x_j))^2 (W_{ij} + J_{ij}) \quad (7)$$
$$= \gamma \|f\|_K^2 + \hat{f}^T L_a \hat{f}$$

where $L_a = D - (W + J)$, D is diagonal matrix given by $D_{ii} = \sum_{j=1}^{l+u} (W_{ij} + J_{ij})$ and γ is a regularization parameter that controls the complexity of the clustering function. It has the same form as unsupervised regularization spectral clustering [1] The existence, uniqueness and an explicit formula describing the solution of this minimizing problem are given by the Representer theorem. Then the solution of the problem has the unique solution:

$$f(x) = \sum_{i=1}^{l+u} \alpha_i K(x, x_i) \quad (8)$$

where α can be solved by an eigenvalue problem and the regularization parameter γ can be selected by the approach of L-curve. For binary classification problem, classifier function f is constant within the region of input space associated with a particular class, that is $Y = \{-1, 1\}$.

3.4 Learning Algorithm

The crux of the proposed learning algorithm is to redefine the Graph Laplacian based on the clustering hypothesis and then adjust the semi-supervised classification with the labeled examples.

The complete semi-supervised learning algorithm (ReguSCoM) consists of the following five steps.

Step 1. Construct adjacency graph $G = (V, E)$ with $(l + u)$ nodes using k nearest neighbors. Choose edge weights W_{ij} with binary or heat kernel weights, construct label information graph $G' = (V, E')$, and then compute the Graph Laplacian L_a.

Step 2. Regularized semi-supervised classification. At this step, we use the objective function given by equation 7.

Step 3. Label the unlabeled examples. Firstly, we select one labeled example from $M = \{x_i, y_i\}_{i=1}^{l}$. Without loss of generality, we select $\{x_1, y_1\}$, so all the examples clustering with $\{x_1, y_1\}$ will have the same label y_1 as $\{x_1, y_1\}$, while the others will have the label different from y_1. So for every $\{x_i, y_i\} \in M$, we get a label $\hat{y}_i$.

Step 4. Compute $\sum_{i=1}^{l} |y_i - \hat{y}_i|^2$. Stop if $\sum_{i=1}^{l} |y_i - \hat{y}_i|^2 \le threshold$, otherwise, select the i th labeled example where $i = \arg\max_i |y_i - \hat{y}_i|$.

Step 5. Adjust the weights J_{ij}. For the selected i th example, we can find the labeled examples j satisfying $\{|y_j - y_i| \le \delta, |y_j - \hat{y}_j| \le \varepsilon, 1 \le j \le l\}$, and then adjust the weight J_{ij} and re-compute the matrix L_a. Goto step 2.

4 Experimental Results

4.1 Synthetic Data

We first conducted experiments on two moons dataset. The dataset contains 200 unlabeled sample points, and all the labeled points are sampled from the unlabeled points randomly. Figure 1 (left) shows the results of unsupervised manifold regularization clustering without labeled points, where the curves represent the decision boundary. After adjusted by one labeled point for each class using Regularized Semi-supervised Classification on manifold (ReguSCoM) proposed in this paper, the decision boundary has little change as shown in Figure 1 (right). The reason lies in that this dataset has regular geometry structure and the manifold regularization clustering can find this structure. The Graph Laplacian based algorithm can implement perfectly the cluster assumption that the decision boundary does not separate the neighbors.

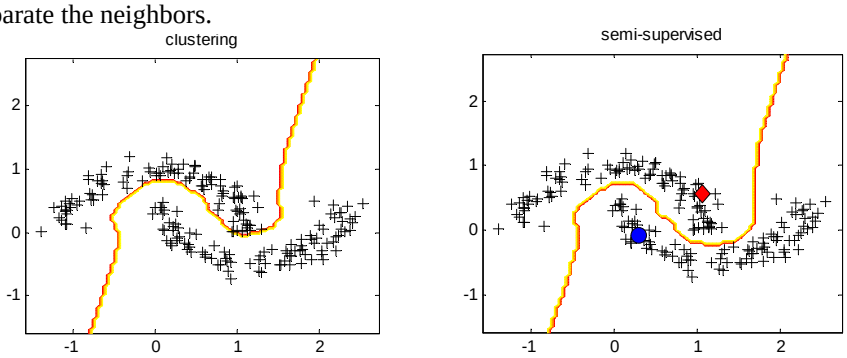

Fig. 1. The result of unsupervised regularization clustering and Regularized Semi-supervised Classification with only one labeled points for each class on two moons dataset

Figure 2 shows the results of semi-supervised classification using ReguSCoM algorithm on two moons dataset with Guassian noise and 0, 1, 3, and 5 labeled points added respectively. With 0 labeled points it can be regarded as unsupervised manifold regularization clustering. From the figure, it is clear that unsupervised classification failed to find the optimal decision boundary. The reason is that the dataset loses the regular geometry structure when noise added. With more labeled examples added, the decision boundary can be adjusted appropriately. With only 5 labeled points for each class, the proposed algorithm can find the optimal solution shown in Figure 2.

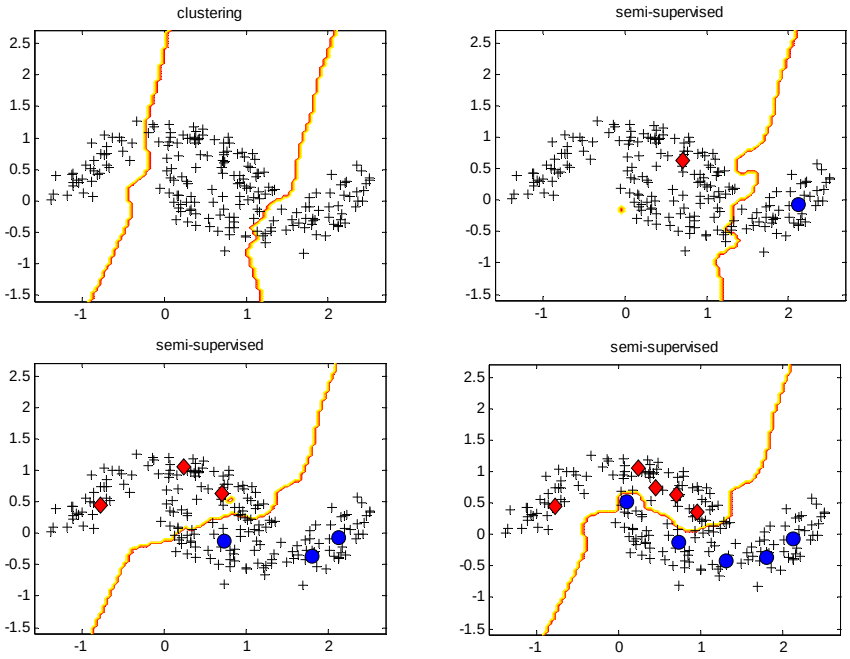

Fig. 2. Regularized Semi-supervised classification on two moons dataset added with Guassian noise and 0, 1, 3, and 5 labeled points respectively.

4.2 Real World Datasets

In this section, we will show the experimental results on two real world datasets, USPS dataset and Isdolet dataset from UCI machine learning repository. We constructed the graph with 6 nearest-neighbors and used the binary weight of the edge of the neighborhood graph, that is $W_{ij} = 0$ *or* 1.

We first used Isolet database of letters of the English alphabet spoken in isolation. We chose isolet1+2+3+4 dataset of 6238 examples and considered the task of binary classifying one of spoken letter from another. Figure 3 (left) shows the mean error rates with the increasing of number of labeled examples using ReguSCoM.

We also show the results of 45 binary classification problems using USPS dataset. We used the first 400 images for each handwritten digit, and processed using PCA to 100 dimensions as in [1]. Figure 3 (right) shows that the mean error rates decrease with the increase of number of labeled examples. We compare the error rate of ReguSCoM with Transductive SVM and LapSVM at the precision-recall breakeven points in the ROC curves, as shown in Figure 4. We choose Polynomial kernel of degree 3, as in [1]. Experimental results show clearly that ReguSCoM is of higher accuracy than Transductive SVM and LapSVM.

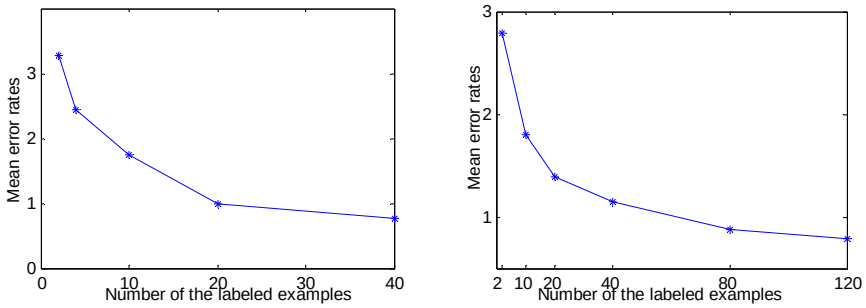

Fig. 3. Mean error rates with the number of labeled examples at the precision-recall breakeven points on Isolet (left) and USPS (right) dataset

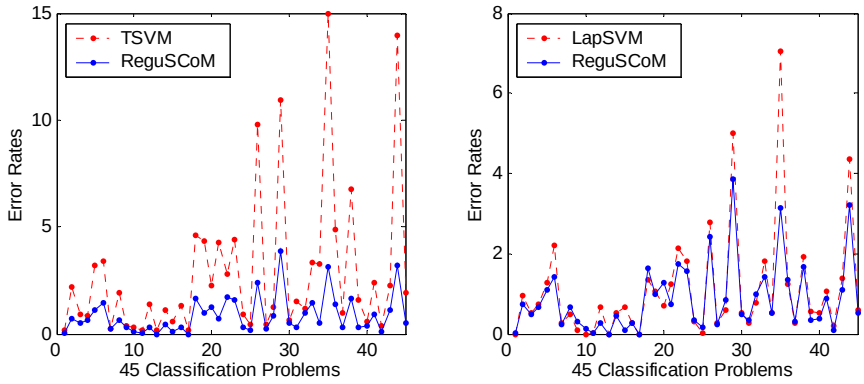

Fig. 4. Comparing the error rate of ReguSCoM, Transductive SVM, and LapSVM at the precision-recall breakeven points

5 Conclusions

Learning from examples has been seen as an ill-posed inverse problem and semi-supervised learning is to benefit from a large number of unlabeled examples and a few labeled examples. We propose a novel regularized semi-supervised classification algorithm on manifold (ReguSCoM) in this paper. The regularization term not only represents the intrinsic geometry structure of P_X that implies the information of classification, but reflects the label information carried by labeled examples. Our method yields encouraging experimental results on both synthetic data and real world datasets and the results demonstrate effective use of both unlabeled and labeled data. In future work, we will explore the link to other semi-supervised leaning algorithms in theory and will investigate other alternative training approaches based on manifold learning to improve performance of semi-supervised learning algorithm. To attack nonlinear ill-posed inverse problem will also be part of our future work.

Acknowledgements

We would like to thank M. Belkin for useful suggestion. The research is supported by the National Natural Science Foundations of China (60373029) and the National Research Foundation for the Doctoral Program of Higher Education of China (20050004001).

Reference:

1. Belkin M., Niyogi P., Sindhwani V. Manifold Regularization: A Geometric Framework for Learning from Examples. Department of Computer Science, University of Chicago, TR-2004-06.
2. Belkin M., Niyogi P., Sindhwani V. On Manifold Regularization. Department of Computer Science, University of Chicago, TR-2004-05.
3. Belkin M., Matveeva I., Niyogi P. Regression and Regularization on Large Graphs. In Proceedings of the Conference on Computational Learning Theory, 2004.
4. Belkin M., Niyogi P. Using Manifold Structure for Partially Labeled Classification, NIPS 2002,Vol. 15.
5. Belkin M., Niyogi P. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation, Neural Computation, June 2003
6. Blum A., Mitchell T. Combining labeled and unlabeled data with co-training. In Proceedings of the Conference on Computational Learning Theory, 1998.
7. Szummer M., Jaakkola T. Partially labeled classification with markov random walks. NIPS 2001,Vol. 14.
8. Zhu X., Ghahramani Z., Lafferty J. Semi-supervised learning using Gaussian fields and harmonic functions. ICML 2003.
9. Blum A., Chawla S. Learning from Labeled and Unlabeled Data using Graph Mincuts, ICML 2001.
10. Zhou D., Bousquet O, Lal TN, Weston J., Schoelkopf B., Learning with Local and Global Consistency, NIPS 2003, Vol. 16.
11. Ernesto De Vito,Lorenzo Rosasco, Andrea Caponnetto, Umberto De Giovannini, Francesca Odone.Learning from Examples as an Inverse Problem. Journal of Machine Learning Research, 6 (2005) 883–904.
12. Bousquet O., Chapelle O, Hein M. Measure Based Regularization, NIPS 2003, Vol. 16.
13. Szummer, M., Jaakkola T. Information regularization with partially labeled data. NIPS 2002,Vol. 15.
14. Krishnapuram B., Williams D., Xue Ya, Hartemink A., Carin L., Figueiredo M. A. T. On Semi-Supervised Classification. NIPS 2004,Vol. 17.
15. K´egl B., Wang Ligen. Boosting on manifolds: adaptive regularization of base classifiers. NIPS 2004,Vol. 17.