TIKHONOV REGULARIZATION AND SEMI-SUPERVISED LEARNING ON LARGE GRAPHS

Mikhail Belkin, Irina Matveeva and Partha Niyogi

University of Chicago, Department of Computer Science

ABSTRACT

We consider the problem of labeling a partially labeled graph. This setting may arise in a number of situations from survey sampling to information retrieval to pattern recognition in manifold settings. It is also of potential practical importance, especially when data is abundant, but labeling is expensive or requires human assistance. Our approach develops a framework for regularization on such graphs parallel to Tikhonov regularization on continuous spaces. The algorithms are very simple and involve solving a single, usually sparse, system of linear equations. Using the notion of algorithmic stability, we derive bounds on the generalization error and relate it to the structural invariants of the graph.

1. INTRODUCTION

We consider the problem of predicting the labels on vertices of a partially labeled graph. Consider a weighted graph $G = (V, E)$ where $V = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ is the vertex set and E is the edge set. Associated with each edge $e_{ij} \in E$ is a weight W_{ij}. If there is no edge present between $\mathbf{x}_i$ and $\mathbf{x}_j$, $W_{ij} = 0$. Imagine a situation where a subset of these vertices are labeled with values $y_i \in \mathbb{R}$. We wish to predict the values of the rest of the vertices. In doing so, we would like to exploit the structure of the graph. In particular, in our approach we will assume that the weights are indications of the affinity of nodes with respect to each other and consequently are related to the potential similarity of the y values, these nodes are likely to have. We will propose an algorithm for regularization on graphs.

This general problem arises in a number of different settings. In survey sampling, one has a database of individuals along with their preference profiles that determines a graph structure based on similarity of preferences. One wishes to estimate a survey variable (e.g. hours of TV watched, amount of cheese consumed, etc.). Rather than survey the entire set of individuals every time, which might be impractical, one may sample a subset of the individuals and then attempt to infer the survey variable for the rest of the individuals. In Internet and information retrieval applications, one is often in possession of a database of objects that have a natural graph structure (or more generally affinity matrix). One may wish to categorize the objects into various classes but only a few (object, class) pairs may be obtained by access to a supervised oracle. In the Finite Element Method for solving PDEs, one sometimes evaluates the solution at some of the points of the finite element mesh and one needs to estimate the value of the solution at all other points. A final example arises when data is obtained by sampling an underlying manifold embedded in a high dimensional space. In recent approaches to dimensionality reduction, clustering and classification in this setting, a graph approximation to the underlying manifold is computed. Semi-supervised learning in this manifold setting reduces to a partially labeled classification problem of the graph. This last example is an instantiation of transductive learning where other approaches include the Naive Bayes for text classification in [11], transductive SVM [14, 9], the graph mincut approach in [3], and the random walk on the adjacency graph in [13]. We also note the closely related work [10], which uses kernels and in particular diffusion kernels on graphs for classification.

We consider this problem in some generality and introduce a framework for regularization on graphs. Two algorithms are derived within this framework. The resulting optima have simple analytical expressions. If the graph is sparse, the algorithms are fast and, in particular, do not require the computation of multiple eigenvectors as is common in many spectral methods (including our previous approach [1]). Another advantage of the current framework is that we are able to provide theoretical guarantees for the generalization error. Using techniques from algorithmic stability we show that generalization error is bounded in terms of the smallest nontrivial eigenvalue (Fiedler number) of the graph. Finally some experimental evaluation is conducted suggesting that this approach to partially labeled classification is competitive.

Several groups of researchers have been recently investigating related ideas. In [16] the authors propose the Label Propagation algorithm for semi-supervised learning, which is similar to our Interpolated Regularization when $S = L$. In [15] a somewhat different regularizer together with the

e-mail:misha@math.uchicago.edu, matveeva@cs.uchicago.edu, niyogi@cs.uchicago.edu

normalized Laplacian is used for semi-supervised learning. The ideas of spectral clustering motivated the authors of [5] to introduce Cluster Kernels for semi-supervised learning. Another related work is [12].

2. TIKHONOV REGULARIZATION ON GRAPHS

2.1. Smoothness Functionals on Graphs

To approximate a function on a graph G, with the weight matrix W_{ij} we need a notion of a "good" function. One way to think about such a function is that is that it does not make too many "jumps". We formalize that notion (also see [1]), by the smoothness functional $\mathcal{S}(f) = \sum_{i \sim j} W_{ij}(f_i - f_j)^2$ where the sum is taken over the adjacent vertices of G. For "good" functions f the functional $\mathcal{S}$ takes small values.

It is important to observe that

$$\sum_{i \sim j} W_{ij}(f_i - f_j)^2 = \mathbf{f}^T L \mathbf{f}$$

where L is the Laplacian $L = D - W$,

$$D = \mathrm{diag}(\sum_i W_{1i}, \ldots, \sum_i W_{ni})$$

This is a basic identity in the spectral graph theory and provides some intuition for the remarkable properties of the graph Laplacian L.

Other smoothness matrices, such as $L^p, p \in \mathbb{N}$, $\exp(tL)$, $t \in \mathbb{R}$ are also possible. In particular, L^2 often seems to work well in practice.

2.2. Algorithms

Let $G = (V, E)$ be a graph with n vertices and the weight matrix W_{ij}. For the purposes of this paper we will assume that G is connected and that the vertices of the graph are numbered. We would like to regress a function $f : V \to \mathbb{R}$. f is defined on vertices of G, however we have only partial information, say, for the first k vertices. That is $f(\mathbf{x}_i) = y_i$, $1 \leq i \leq k$. The labels can potentially be noisy. We also allow data points to have multiplicities, i.e. each vertex of the graph may appear more than once with the same or different y value.

We precondition the data by mean subtracting first. We take $\tilde{\mathbf{y}} = (y_1 - \bar{y}, \ldots, y_k - \bar{y})$ where $\bar{y} = \frac{1}{k}\sum y_i$. This is needed for stability of the algorithms as will be seen in the theoretical discussion.

Algorithm 1: Tikhonov regularization (parameter $\gamma \in \mathbb{R}$). The objective is to minimize the square loss function plus the smoothness penalty.

$$\tilde{\mathbf{f}} = \underset{\substack{\mathbf{f}=(f_1,\ldots,f_n) \\ \sum f_i = 0}}{\mathrm{argmin}} \frac{1}{k}\sum_i (f_i - \tilde{y}_i)^2 + \gamma \mathbf{f}^t S \mathbf{f}^t$$

S here is a smoothness matrix, e.g. $S = L$ or $S = L^p$, $p \in \mathbb{N}$. The condition $\sum f_i = 0$ is needed to make the algorithm stable. It can be seen by following the proof of Theorem 1 that necessary stability and the corresponding generalization bound cannot be obtained unless the regularization problem is constrained to functions with mean 0.

Without the loss of generality we can assume that the first l points on the graph are labeled. l might be different from the number of sample points k, since we allow vertices to have different labels (or the same label several times).

The solution to the quadratic problem above is not hard to obtain by standard linear algebra considerations. If we denote by $\mathbf{1} = (1, 1, \ldots, 1)$ the vector of all ones, the solution can be given in the form

$$\tilde{\mathbf{f}} = (k\gamma S + I_k)^{-1}(\tilde{\mathbf{y}} + \mu \mathbf{1})$$

Here $\tilde{\mathbf{y}}$ is the n-vector

$$\mathbf{y} = (\sum_i y_{1i}, \sum_i y_{2i}, \ldots, \sum_i y_{mi}, 0, \ldots, 0)$$

where we sum the labels corresponding to the same vertex on the graph.

I_k is a diagonal matrix of multiplicities

$$I_k = \mathrm{d}iag\,(n_1, n_2, \ldots, n_l, 0, \ldots, 0)$$

where n_i is the number of occurrences of vertex i among the labeled point in the sample. μ is chosen so that the resulting vector $\mathbf{f}$ is orthogonal to $\mathbf{1}$. Denote by $s(\mathbf{f})$ the functional

$$s : \mathbf{f} \to \sum_i f_i$$

Since s is linear, we obtain $0 = s(\tilde{\mathbf{f}}) = s\left((k\gamma S + I_k)^{-1}\tilde{\mathbf{y}}\right) + s\left((k\gamma S + I_k)^{-1}\mathbf{1}\right)$. Therefore we can write

$$\mu = -\frac{s\left((k\gamma S + I_k)^{-1}\tilde{\mathbf{y}}\right)}{s\left((k\gamma S + I_k)^{-1}\mathbf{1}\right)}$$

Note that dropping the condition $\mathbf{f} \perp \mathbf{1}$ is equivalent to putting $\mu = 0$.

Algorithm 2: Interpolated Regularization (no parameters).

Here we assume that the values $y_1, \ldots, y_k$ have no noise. Thus the optimization problem is to find a function of maximum smoothness satisfying $f(\mathbf{x}_i) = \tilde{y}_i$, $1 \leq i \leq k$:

$$\tilde{\mathbf{f}} = \underset{\substack{\mathbf{f}=(\tilde{y}_1,\ldots,\tilde{y}_k,f_{k+1},\ldots,f_n) \\ \sum f_i = 0}}{\mathrm{argmin}} \mathbf{f}^t S \mathbf{f}$$

As before S is a smoothness matrix, e.g. L or L^2. However, here we are not allowing multiple vertices in the sample. We partition S as

$$S = \begin{pmatrix} S_1 & S_2 \\ S_2^T & S_3 \end{pmatrix}$$

where S_1 is a $k \times k$ matrix, S_2 is $k \times n-k$ and S_3 is $(n-k) \times (n-k)$. Let $\tilde{f}$ be the values of f, where the function is unknown, $\tilde{f} = (f_{k+1}, \ldots, f_n)$.

Straightforward linear algebra yields the solution:

$$\tilde{f} = S_3^{-1} S_2^T ((\tilde{y_1}, \ldots, \tilde{y_k})^T + \mu \mathbf{1})$$

$$\mu = -\frac{s\,(S_3^{-1}\, S_2^T\, \tilde{\mathbf{y}})}{s\,(S_3^{-1}\, S_2^T\, \mathbf{1})}$$

The regression formula is very simple and has no free parameters. However, the quality of the results depends on whether S_3 is well conditioned.

It can be shown that Interpolated Regularization is the limit case of Tikhonov regularization when γ tends to 0.

3. THEORETICAL ANALYSIS

In this section we investigate some theoretical guarantees for the generalization error of regularization on graphs. We use the notion of algorithmic stability, first introduced by Devroye and Wagner in [7]. We used a theorem of Bousquet and Elisseeff ([4]) to obtain the bounds.

For the lack of space we omit all proofs. See our Technical Report ([2]) for the details.

The goal of a learning algorithm is to learn a function on some space V from examples. Given a set of examples T the learning algorithm produces a function $f_T : V \to \mathbb{R}$. Therefore a learning rule is a map from data sets into functions on V. We will be interested in the case where V is a graph. The empirical risk $R_k(f)$ (with the square loss function) is a measure of how well we do on the training set:

$$R_k(f) = \frac{1}{k} \sum_1^k (f(\mathbf{x}_i) - y_i)^2$$

The generalization error $R(f)$ is the expectation of how well we do on all points, labeled or unlabeled.

$$R(f) = E_\mu \left(f(\mathbf{x}) - y(\mathbf{x})\right)^2$$

where the expectation is taken over an underlying distribution μ on $V \times \mathbb{R}$ according to which the labeled examples are drawn.

As before denote the smallest nontrivial eigenvalue of the smoothness matrix S by λ_1. If S is the Laplacian of the graph, this value, first introduced by Fiedler in [8] as algebraic connectivity and is sometimes known as the Fiedler constant, plays a key role in spectral graph theory.

Theorem 1 (Generalization Performance of Graph Regularization). *Let γ be the regularization parameter, T be a set of $k \geq 4$ vertices $\mathbf{x}_1, \ldots, \mathbf{x}_k$, where each vertex occurs no more than t times, together with values $y_1, \ldots, y_k$, $|y_i| \leq M$. Let f_T be the regularization solution using the smoothness functional S with the second smallest eigenvalue λ_1. Assuming that $\forall \mathbf{x} |f_T(\mathbf{x})| \leq K$ we have with probability $1-\delta$ (conditional on the multiplicity being no greater than t):*

$$|R_k(f_T) - R(f_T)| \leq \beta + \sqrt{\frac{2\log(2/\delta)}{k}} \left(k\beta + (K+M)^2\right)$$

where

$$\beta = \frac{3M\sqrt{tk}}{(k\gamma\lambda_1 - t)^2} + \frac{4M}{k\gamma\lambda_1 - t}$$

We do not discuss the issue of multiplicity here but note that in many situations t can be taken to be one. We see that the generalization error it decreases at a rate $\frac{1}{\sqrt{k}}$. It is important to note that the estimate is nearly independent of the total number of vertices n in the graph. We say "nearly" since the probability of having multiple points increases as k becomes close to n and since the value of λ_1 may (or may not) implicitly depend on the number of vertices.

The only thing that is missing is an estimate for K. Below we give two such estimates, one for the case of general S and the second when the smoothness matrix is the Laplacian L.

Proposition 2. *With λ_1, K, M and γ as above we have the following inequality:*

$$K \leq \frac{M}{\sqrt{\lambda_1 \gamma}}$$

A different, possibly sharper inequality can be obtained when $S = L$. Note the the diameter of the graph is typically far smaller than the number of vertices. For example, when G is a n-cube, the number of vertices is 2^n, while the diameter is n.

Proposition 3. *Let $W = \min_{i \sim j} w_{ij}$ be the smallest nonzero weight of the graph G. Assume G is connected. Let D be the unweighted diameter of the graph, i.e. the maximum length of the shortest path between two points on the graph. Then the maximum entry K of the solution to the γ-regularization problem with y's bounded by M satisfies the following inequality:*

$$K \leq M \sqrt{\frac{D}{\gamma W}}$$

A useful special case is

Corollary 4. *If all weights of G are either 0 or 1, then*

$$K \leq M \sqrt{\frac{D}{\gamma}}$$

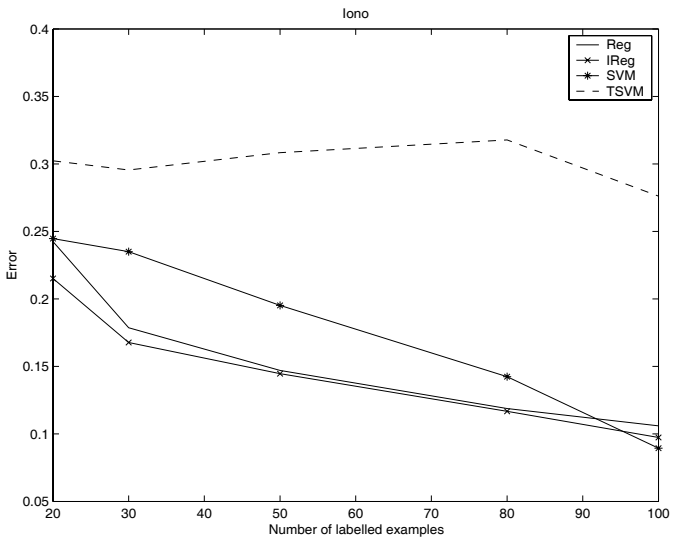

Fig. 1. Classification error for regularization, interpolated regularization compared to the Support Vector Machine (SVM) and the Transductive Support Vector Machine (TSVM) for different numbers of labeled points.

4. EXPERIMENTS

As an example, we compare the performance of regularization and interpolated regularization on the Ionosphere dataset from the UC Irvine Machine Learning Repository. The parameter γ for the regularization algorithm is determined using the leave-one-out cross-validation. We construct the graphs using 7 nearest neighbors and binary weights. Because of the space constraints, we do not show results of other experiments but note that the performance is generally competitive.

5. CONCLUSIONS

In a number of different settings, there is a need to fill in the labels (values) of a partially labeled graph. We have provided a principled framework for Tikhonov regularization on such graphs. The algorithms proposed are simple and require solving a single, usually sparse, system of linear equations.

It is important to note that if the graph arises from the local connectivity of data obtained from sampling an underlying manifold, then our approach has natural connection to regularization on that manifold.

6. REFERENCES

[1] M. Belkin, P. Niyogi, *Semi-supervised Learning on Riemannian Manifolds*, Machine Learning Journal, Special Issue on Clustering, to appear.

[2] M. Belkin, I. Matveeva, P. Niyogi, *Regression and Regularization on Large Graphs*, University of Chicago CS Technical Report, 2003.

[3] A. Blum, S. Chawla, *Learning from Labeled and Unlabeled Data using Graph Mincuts*, ICML, 2001,

[4] Bousquet, O., A. Elisseeff, *Algorithmic Stability and Generalization Performance*. Advances in Neural Information Processing Systems 13, 196-202, MIT Press, 2001,

[5] Chapelle, O., J. Weston and B. Scholkopf,*Cluster Kernels for Semi-Supervised Learning*, Advances in Neural Information Processing Systems 15. (Eds.) S. Becker, S. Thrun and K. Obermayer,

[6] Fan R. K. Chung, *Spectral Graph Theory,* Regional Conference Series in Mathematics, number 92, 1997

[7] L.P. Devroye, T. J. Wagner, *Distribution-free Performance Bounds for Potential Function Rules*, IEEE Trans. on Information Theory, 25(5): 202-207, 1979.

[8] M. Fiedler, *Algebraic connectibity of graphs*, Czechoslovak Mathematical Journal, 23(98):298–305, 1973.

[9] T. Joachims, *Transductive Inference for Text Classification using Support Vector Machines*,Proceedings of ICML-99, pps 200–209, 1999.

[10] I.R. Kondor, J. Lafferty, *Diffusion Kernels on Graphs and Other Discrete Input Spaces*, Proceedings of ICML, 2002.

[11] K. Nigam, A.K. McCallum, S. Thrun, T. Mitchell, *Text Classification from Labeled in Unlabeled Data*, Machine Learning 39(2/3), 2000,

[12] A. Smola and R. Kondor, *Kernels and Regularization on Graphs*, COLT/KW 2003,

[13] Martin Szummer, Tommi Jaakkola, *Partially labeled classification with Markov random walks*, Neural Information Processing Systems (NIPS) 2001, vol 14.,

[14] V. Vapnik, *Statistical Learning Theory*, Wiley, 1998,

[15] D. Zhou, O. Bousquet, T.N. Lal, J. Weston and B. Schoelkopf, *Learning with Local and Global Consistency*, Max Planck Institute for Biological Cybernetics Technical Report, June 2003,

[16] X. Zhu, J. Lafferty and Z. Ghahramani, *Semi-supervised learning using Gaussian fields and harmonic functions*, Machine Learning: Proceedings of the Twentieth International Conference, 2003.