SOME NEW DIRECTIONS IN GRAPH-BASED SEMI-SUPERVISED LEARNING

Xiaojin Zhu, Andrew B. Goldberg, Tushar Khot

Department of Computer Sciences
University of Wisconsin-Madison
Madison, WI, USA 53706
{jerryzhu, goldberg, tushar}@cs.wisc.edu

ABSTRACT

In this position paper, we first review the state-of-the-art in graph-based semi-supervised learning, and point out three limitations that are particularly relevant to multimedia analysis: (1) rich data is restricted to live on a single manifold; (2) learning must happen in batch mode; and (3) the target label is assumed smooth on the manifold. We then discuss new directions in semi-supervised learning research that can potentially overcome these limitations: (i) modeling data as a mixture of multiple manifolds that may intersect or overlap; (ii) online semi-supervised learning that learns incrementally with low computation and memory needs; and (iii) learning spectrally sparse but non-smooth labels with compressive sensing. We give concrete examples in each new direction. We hope this article will inspire new research that makes semi-supervised learning an even more valuable tool for multimedia analysis.

Index Terms— semi-supervised learning, multi-manifold, online learning, compressive sensing, graph

1. STATE-OF-THE-ART IN GRAPH-BASED SEMI-SUPERVISED LEARNING AND LIMITATIONS

Semi-supervised learning encompasses many different model assumptions [1, 2]. Graph-based semi-supervised learning is an important family of methods that make the following common assumption. Let $x_i, x_j \in \mathcal{X}$ be two input items, and $y_i, y_j \in \mathcal{Y}$ be their labels. Usually, $\mathcal{X} \subseteq \mathbb{R}^D$ and $\mathcal{Y} = \{-1, 1\}$ for binary classification, but multiclass classification and regression are common, too. Let $d(\cdot, \cdot)$ be an appropriate distance measure on $\mathcal{X}$. The graph-based assumption states that if $d(x_i, x_j)$ is small, then $y_i \approx y_j$. This assumption applies regardless of whether x_i, x_j are labeled. If x_1 is labeled, and there is a sequence of unlabeled items $x_2, \ldots, x_k$ such that $d(x_i, x_{i+1})$ is small for $i = 1 \ldots k-1$, then the label y_1 will propagate along the sequence.

Formally, a graph is formed with nodes being labeled data $\{(x_i, y_i)\}_{i=1}^n$ and unlabeled data $\{x_i\}_{i=n+1}^{n+m}$. The undirected edges reflect similarity between nodes: the edge weight w_{ij} between nodes x_i, x_j is large if $d(x_i, x_j)$ is small. A common choice of edges is to connect each node to its kNNs with weight 1, and disconnect it from all other nodes (weight 0). Let W be the $(n+m) \times (n+m)$ weight matrix, and D the diagonal degree matrix with $D_{ii} = \sum_j w_{ij}$. Let $L = D - W$ be the unnormalized graph Laplacian matrix (the normalized Laplacian $D^{-1/2} L D^{-1/2}$ can be used, too). Let f be a function on the graph. Then the graph assumption is equivalent to having a small *energy* $f^\top L f = 1/2 \sum_{i,j} (f(x_i) - f(x_j))^2 w_{ij}$.

This assumption is behind graph-based semi-supervised methods such as Mincut [3], graph random walk [4], Gaussian Random Fields [5], local and global consistency [6], spectral graph transducer [7], manifold regularization [8, 9], and many other variants. In particular, Belkin et al. generalize graph-based learning to the manifold setting, where $\mathcal{X}$ is assumed to be a low dimensional manifold in $\mathbb{R}^D$, the labels y change smoothly on the manifold, and the graph constructed on labeled and unlabeled training data is a random realization of the manifold. This provides an elegant conceptual model. The graph-based assumption has been extended to directed edges like links between Web pages [10] and dissimilarity edges [11, 12]. Applications of graph-based semi-supervised learning abound.

Despite their success, we point out three major limitations of graph-based methods:

(1) Current methods assume that $\mathcal{X}$ is a single manifold, or multiple well-separated manifolds. Therefore, it makes sense to create the graph W using kNN edges, or Gaussian weighted edges $w_{ij} = \exp\left(-\lambda d(x_i, x_j)^2\right)$, where $d(\cdot, \cdot)$ is based on Euclidean distance. In both cases, nearby nodes are strongly connected and are assumed to have similar labels. However, in multimedia data, the distribution of objects might form multiple manifolds that intersect or partially overlap with each other. For example, in motion segmentation from video images, the tracked feature points on different objects form multiple intersecting and overlapping manifolds [13]. Even though each individual manifold obeys the label smoothness assumption, nearby items on different man-

We would like to thank the Wisconsin Alumni Research Foundation. AG is supported in part by a Yahoo! Key Technical Challenges Grant.

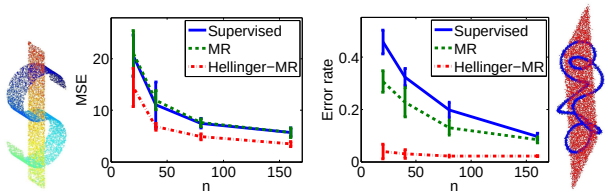
ifolds may not satisfy this assumption. Straightforward application of existing graph-based semi-supervised learning will not achieve optimal performance.

(2) Current methods learn in batch mode. That is, they require the training set to be available all at once. However, consider a robot with a camera that continuously takes video of its surroundings, and learns the names of various objects. A human annotator provides object names (labels) only occasionally on selected video frames. This is therefore semi-supervised learning. But the robot cannot afford to store the massive amount of mostly unlabeled video before learning. It requires an "anytime classifier" that is ready at all times, while continuously improving itself. And training must be cheap and quick. What we need is semi-supervised learning that operates in online mode.

(3) Current methods assume label smoothness on the graph. As dissimilarity edges show, this may not always be the case [11, 12]. In general, the relation between the label and the underlying graph can be studied from the perspective of harmonic analysis. It is well-known that the traditional smoothness assumption is equivalent to favoring low frequency components of the graph spectrum [14]. Recent advances in compressive sensing (see, e.g., [15]) allow learning from an arbitrary combination of low and high frequency components, as long as the number of components is small. We present, to our knowledge, the first connection between compressive sensing and graph-based transduction.

2. MULTI-MANIFOLD LEARNING

We recently introduced a novel graph as a first step in addressing data containing a mixture of manifolds [16]. The idea is to assign edge weights based on differences in local geometry around each item x. Our intuition is that items on different manifolds, or in regions with different density, should be considered dissimilar and lead to low edge weights. Computationally, we compare local regions using Hellinger distance, which is sensitive to local manifold structures. We start by estimating the local sample covariance matrix Σ_x around a randomly selected set of anchor items x. Then, the squared Hellinger distance between two anchor points x_i, x_j is $H^2(p,q) = \frac{1}{2}\int\left(\sqrt{p(x)}-\sqrt{q(x)}\right)^2 \mathrm{d}x$, where $p = \mathcal{N}(x;0,\Sigma_{x_i})$ and $q = \mathcal{N}(x;0,\Sigma_{x_j})$ are zero mean Gaussians with those local sample covariance matrices. The Hellinger distance H is symmetric, in $[0,1]$, small when the local geometry is similar, and large when there is significant difference in density, manifold dimensionality or orientation (see Figure 2 in [16]). Finally, we build a sparse kNN graph over the labeled and anchor unlabeled items as follows: Each such x is connected by a weighted, undirected edge to its k nearest Mahalanobis neighbors. Note that, since Σ_x captures the local geometry around x, we "follow the manifold" by using the Mahalanobis distance as the local distance metric at x: $d_M^2(x,x') = (x-x')^\top \Sigma_x^{-1}(x-x')$. The neighborhood size k is set to grow with dataset size. The graph edges are weighted using the standard RBF scheme, but with Hellinger distance: $w_{ij} = \exp\left(-\lambda H^2(p,q)\right)$. See Figure 3 in [16] for an example graph using this weighting scheme. In short, the graph combines locality and geometry: an edge has large weight when the two nodes are close in Mahalanobis distance, and have similar covariance structure. Importantly, it effectively separates intersecting and overlapping manifolds into individual pieces.



(a) Dollar sign (b) Surface-helix

Fig. 1. Multi-manifold learning with Hellinger-graphs.

We demonstrate the effectiveness of this Hellinger graph with manifold regularization [8] on two synthetic datasets. **Dollar sign** is a regression dataset containing two intersecting manifolds with target values varying greatly across intersection points (Figure 1(a), color indicates y). **Surface-helix** is a classification dataset with a 1D toroidal helix intersecting a surface—each manifold is a separate class (Figure 1(b)). Figure 1 compares three learners on these datasets: **[Supervised]**: supervised learner (kernel regression or SVM) trained on labeled data only, ignoring unlabeled data. **[MR]**: standard manifold regularization (LapRLS or LapSVM) using a Euclidean-based 3NN graph [8]. **[Hellinger-MR]**: manifold regularization (LapRLS or LapSVM) using this novel Hellinger graph. See [16] for details about the parameters governing the Hellinger graph. All other parameters were tuned using 5-fold cross validation. All datasets start with $M = 20,000$ unlabeled items, from which we select $m \sim O(M/log(M))$ anchor items. Figure 1 shows performance on a separate test set of 20,000 items, averaged over 10 trials. For the dollar sign data, standard MR performs only as well as supervised learning, while Hellinger-MR achieves statistically significantly better MSE in all four n conditions (based on paired t-tests). For the surface-helix data, the three methods are all statistically significantly different for all n, with Hellinger-MR making the best use of unlabeled data.

Open issues surrounding our Hellinger graph remain, including how to select anchors, and how to optimize parameters. Furthermore, some other metrics over matrices or probability distributions may be more appropriate than Hellinger distance for this purpose. Finally, it could be useful to exploit the labeled data for detecting and validating the presence of multiple manifolds with differing target values.

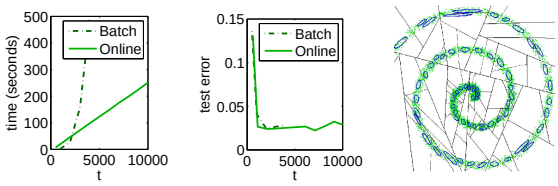


Fig. 2. Online semi-supervised learning

3. ONLINE SEMI-SUPERVISED LEARNING

We argue that the following is an important and practical setting, especially for real-time multimedia applications:

1. At time t an adversary picks (x_t, y_t) and shows x_t.
2. The learner predicts $f_t(x_t)$.
3. With (small) probability p the adversary reveals y_t. Otherwise x_t remains unlabeled.
4. The learner updates its predictor to f_{t+1}, even when y_t is not given. Repeat with $t = t + 1$.

This is clearly online learning. It differs from the standard online setting in that learning happens even on unlabeled data. The goal is to update f_t in such a way that there is no regret, i.e., the wrong predictions the online procedure makes over time are comparable to a batch learner, which has access to the same input *simultaneously* but has to use the single best fixed predictor. A good online semi-supervised learning algorithm should achieve zero regret with sublinear space and time complexity.

As a concrete example, our online semi-supervised algorithm in [17] employs online convex programming with an asymptotic zero-regret guarantee. At the heart of the algorithm is a gradient step in kernel space $f_{t+1} = f_t - \eta_t \left.\frac{\partial J_t(f)}{\partial f}\right|_{f_t}$, where η_t is a stepsize that decays as $O(1/\sqrt{t})$. The term $J_t(f)$ is the *instantaneous risk* functional. When summing over time, we recover the standard batch manifold regularization risk $J(f) = \frac{1}{n+m}\sum_{t=1}^{n+m} J_t(f)$. The definition of $J_t(f)$ can be found in [17]; suffice it to say that the instantaneous graph energy is $\sum_{i=1}^{t-1}(f(x_i) - f(x_t))^2 w_{it}$. That is, it involves the edges from x_t to all previous nodes in the graph. However, its complexity grows linearly with t. One approximation is to use a buffer of fixed size τ: $\frac{t}{\tau}\sum_{i=t-\tau}^{t-1}(f(x_i) - f(x_t))^2 w_{it}$. That is, old nodes from τ steps ago are discarded. Figure 2 compares batch vs. online semi-supervised learning (manifold regularization)'s running time and test error on MNIST digit recognition 1 vs. 2. The online algorithm achieves a desirable constant learning complexity at each step, and has comparable accuracy as batch mode.

Keeping a fixed buffer of recent input is not optimal ultimately. The dynamic graph constructed on items in the buffer only reflects a random and noisy snapshot of the underlying manifold structure. Given the same space constraint, it is better to form a summary of all the input so far. One possibility is some form of online clustering that forms a mixture model (e.g., Gaussian mixtures) on the input, using a fixed number of mixing components, as in Figure 2(right). A "hyper-graph" can then be maintained on the mixture model, with nodes being the components. Graph-based learning proceeds on the hyper-graph, which is a fixed-size summary of the manifold seen so far. This is the idea behind the use of Random Projection Trees in [17]; see also [18].

Looking forward, we need more efficient online semi-supervised algorithms with theoretical guarantees. The combination of online semi-supervised learning and online active learning also deserves attention.

4. LEARNING NON-SMOOTH LABELS ON GRAPHS

The spectrum of the graph is the set of eigenvalue, eigenvector pairs $\{(\lambda_i, \psi_i)\}_{i=1}^{n+m}$, where the Laplacian $L = \sum_i \lambda_i \psi_i \psi_i^\top$. If we sort λ from small to large, then $\lambda_1 = \ldots = \lambda_k = 0$ if and only if the graph has k disconnected components. The eigenvectors $\Psi = \{\psi_i\}$ form an orthonormal basis. Any target label function on the graph can be decomposed into $f = \sum_i \alpha_i \psi_i$, where ψ_1 corresponds to the lowest frequency component, and ψ_{n+m} the highest frequency component. The function's energy can be shown to be $f^\top L f = \sum_i \lambda_i \alpha_i^2$.

Existing semi-supervised learning algorithms assume that f is smooth with respect to the graph. This is equivalent to assuming large (non-zero) α's for small i, and small (zero) α's for large i. In the future, one may wish to allow non-smooth labels f to model richer data, which must still be learnable from a small number of labeled points. Compressive sensing offers such guarantee if f is spectrally sparse, i.e., if only $S \ll n+m$ of the α's are non-zero. These S non-zero components can occupy any frequency, thus allowing non-smooth f and generalizing the graph smoothness assumption.

Our key insight is that *transductive learning on graphs corresponds to compressive sensing using the* $(n+m) \times (n+m)$ *canonical basis* $\Phi = I$. The $n \times (n+m)$ sensing matrix consists of n random rows selected from Φ. The sensing matrix simply reads out the label values at n nodes; these correspond to the n labeled data items. Importantly, when the graph is "nice," i.e., without small (nearly) disconnected components, the graph spectrum basis Ψ is incoherent with the canonical basis Φ. This allows the exact recovery of the whole f from the n observations when $n \geq C\mu^2(\Phi, \Psi) S \log(n+m)$, where $\mu(\Phi, \Psi) = \sqrt{n+m}\max_{i,j} |\phi_i^\top \psi_j|$ is the coherence (the lower the better).

We now give a concrete example. Consider a closed chain graph (i.e., a ring) with $n+m = 1024$ nodes and edge weights 1. The graph spectrum Ψ is the discrete Fourier basis, whose coherence with the canonical basis is $\sqrt{2}$. Let $S = 3$, and $f = -\psi_5 - 1.3\psi_8 + \psi_{63}$, which is spectrally sparse yet non-smooth as shown in Figure 3(left). We take n measurements

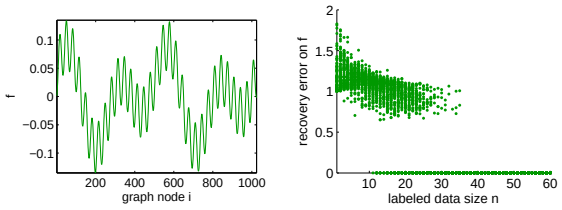


Fig. 3. Compressive sensing on a closed-chain graph

from the canonical basis (i.e., select labeled items); thus we have labels $y = f$ on those n random nodes. We then solve the standard ℓ_1 minimization problem to recover $\hat{\alpha}$ (thus $\hat{f}$ on the whole graph). We vary n from 1 to 60. For each n we run 100 trials; each takes n random rows from Φ to form the sensing matrix. For each trial, we compute the recovery error $\|f - \hat{f}\|_{\ell_2}/\|f\|_{\ell_2}$. Each trial is a dot in Figure 3(right). It seems exact recovery happens when $n > 35$ for this graph.

Much work remains in improving this novel way of performing transduction on a graph. Potential research directions include: identifying real-world problems with spectrally sparse labels on graphs, finding bases that are more localized than the Laplacian spectrum yet still incoherent with the canonical sensing basis, and studying label acquisition mechanisms when the sensing basis is not canonical (e.g., random matrices).

5. CONCLUSIONS

We have presented three new research directions for graph-based semi-supervised learning and our initial approaches at solving these novel problems. We hope this article will inspire new research, making semi-supervised learning an even more valuable tool for multimedia analysis.

6. REFERENCES

[1] Olivier Chapelle, Alexander Zien, and Bernhard Schölkopf, Eds., *Semi-supervised learning*, MIT Press, 2006.

[2] Xiaojin Zhu, "Semi-supervised learning literature survey," Tech. Rep. 1530, Department of Computer Sciences, University of Wisconsin, Madison, 2005.

[3] Avrim Blum and Shuchi Chawla, "Learning from labeled and unlabeled data using graph mincuts," in *Proc. 18th International Conf. on Machine Learning*, 2001.

[4] Martin Szummer and Tommi Jaakkola, "Partially labeled classification with Markov random walks," in *Advances in Neural Information Processing Systems, 14*, 2001, vol. 14.

[5] Xiaojin Zhu, Zoubin Ghahramani, and John Lafferty, "Semi-supervised learning using Gaussian fields and harmonic functions," in *The 20th International Conference on Machine Learning (ICML)*, 2003.

[6] Dengyong Zhou, Olivier Bousquet, Thomas Lal, Jason Weston, and Bernhard Schl̈kopf, "Learning with local and global consistency," in *Advances in Neural Information Processing System 16*, 2004.

[7] Thorsten Joachims, "Transductive learning via spectral graph partitioning," in *Proceedings of ICML-03, 20th International Conference on Machine Learning*, 2003.

[8] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *Journal of Machine Learning Research*, vol. 7, pp. 2399–2434, November 2006.

[9] Vikas Sindhwani, Partha Niyogi, and Mikhail Belkin, "Beyond the point cloud: from transductive to semi-supervised learning," in *ICML05, 22nd International Conference on Machine Learning*, 2005.

[10] Dengyong Zhou, Jiayuan Huang, and Bernhard Schölkopf, "Learning from labeled and unlabeled data on a directed graph," in *ICML*, 2005.

[11] Andrew Goldberg, Xiaojin Zhu, and Stephen Wright, "Dissimilarity in graph-based semi-supervised classification," in *AISTATS*, 2007.

[12] Wei Tong and Rong Jin, "Semi-supervised learning by mixed label propagation," in *AAAI*, 2007.

[13] R. Tron and R. Vidal, "A benchmark for the comparison of 3-D motion segmentation algorithms," in *CVPR*, 2007.

[14] F. R. K. Chung, *Spectral graph theory, Regional Conference Series in Mathematics, No. 92*, American Mathematical Society, 1997.

[15] Emmanuel Candés and Michael Wakin, "An introduction to compressive sampling," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 21–30, 2008.

[16] Andrew Goldberg, Xiaojin Zhu, Aarti Singh, Zhiting Xu, and Robert Nowak, "Multi-manifold semi-supervised learning," in *AISTATS*, 2009.

[17] Andrew B. Goldberg, Ming Li, and Xiaojin Zhu, "Online manifold regularization: A new learning setting and empirical study," in *ECML PKDD*, 2008.

[18] Xiaojin Zhu and John Lafferty, "Harmonic mixtures: combining mixture models and graph-based methods for inductive and scalable semi-supervised learning," in *ICML*, 2005.