ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Graph-Based Semi-Supervised Learning for Large Scale Data Processing Using SVM

N. Deshai, M.Tech
Assistant Professor,
Department Information Technology,
S.R.K.R Engineering College,
Bhimavaram.

Dr. I. Hemalatha, M.Tech, Ph.D
Professor,
Department of Information Technology,
S.R.K.R Engineering College,
Bhimavaram.

Dr. G. P. Saradhi Varma, M.Tech, Ph.D
Department Computer science and Engineering'
S.R.K.R Engineering College,
Bhimavaram.

Abstract:
Here a new approach is proposed to study the performance of graph-based semi-supervised learning methods, under the assumptions that the dimension of data p and their number n grow large at the same rate and that the data arise from a Gaussian mixture model. Unlike small dimensional systems, the large dimensions allow for a Taylor expansion to linearize the weight (or kernel) matrix W, thereby providing in closed form the limiting performance of semi-supervised learning algorithms. This notably allows to predict the classification error rate as a function of the normalization parameters and of the choice of the kernel function. Despite the Gaussian assumption for the data, the theoretical findings match closely the performance achieved with real datasets, particularly here on the popular MNIST database.

Keywords:
Semi-supervised learning, graphs, performance analysis, random matrix theory

Introduction:
Semi-supervised learning (SSL) methods use small amounts of labeled data along with large amounts of unlabeled data to train prediction systems. Such approaches have gained widespread usage in recent years and have been rapidly supplanting supervised systems in many scenarios owing to the abundant amounts of unlabeled data available on the Web and other domains. Annotating and creating labeled training data for many predictions tasks is quite challenging because it is often an expensive and labor-intensive process.

On the other hand, unlabeled data is readily available and can be leveraged by SSL approaches to improve the performance of supervised prediction systems. There are several surveys that cover various SSL methods in the literature. The majority of SSL algorithms are computationally expensive; for example, transductive SVM. Graph-based SSL algorithms are a subclass of SSL techniques that have received a lot of attention recently, as they scale much better to large problems and data sizes. These methods exploit the idea of constructing and smoothing a graph in which data (both labeled and unlabeled) is represented by nodes and edges link vertices that are related to each other. Edge weights are defined using a similarity function on node pairs and govern how strongly the labels of the nodes connected by the edge should agree.

Graph-based methods based on label propagation work by using class label information associated with each labeled "seed" node, and propagating these labels over the graph in a principled, iterative manner. These methods often converge quickly and their time and space complexity scales linearly with the number of edges |E| and number of labels m. Successful applications include a wide range of tasks in computer vision information retrieval (IR) and social networks and natural language processing (NLP); for example, class instance acquisition and relation prediction, to name a few. Several classification and knowledge expansion type of problems involve a large number of labels in real world scenarios. For instance, entity-relation classification over the widely used Freebase taxonomy requires learning over thousands of labels which can grow further by orders when extending to

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

open-domain ex- traction from the Web or social media; scenarios involving complex overlapping classes or fine-grained classification at large scale for natural language and computer vision applications . Unfortunately, existing graph-based SSL methods cannot deal with large m and |E| sizes. Typically individual nodes are initialized with sparse label distributions, but they become dense in later iterations as they propagate through the graph. Talukdar and Cohen recently proposed a method that seeks to overcome the label scale problem by using a Count-Min Sketch to approximate labels and their scores for each node. This reduces the memory complexity to O(logm) from O(m). They also report improved running times when using the sketch-based approach. However, in real world applications, the number of actual labels k associated with each node is typically sparse even though the overall label space may be huge; i.e., $k < m$. Cleverly leveraging sparsity in such scenarios can yield huge benefits in terms of efficiency and scalability. While the sketching technique from approximates the label space succinctly, it does not utilize the sparsity (a naturally occurring phenomenon in real data) to full benefit during learning.

Related Work:

In many real world classification tasks, the number of labeled instances is very few due to the prohibitive cost of manually labeling every single data point, while the number of unlabeled data can be very large since they are easy to obtain. Traditional classification algorithms, known as supervised learning, only make use of the labeled data, therefore prove insufficient in these situations. To address this problem, semi-supervised learning has been developed, which makes use of unlabeled data to boost the performance of supervised learning. In particular, graph-based semi-supervised learning algorithms have proved to be effective in many applications, such as hand-written digit classification [Zhu et al., 2003; Zhu et al., 2005], medical image segmentation [Grady and Funka-Lea, 2004], word sense disambiguation [Niu, Ji and Tan, 2005], image retrieval [He et al., 2004], etc.

Compared with other semi-supervised learning methods, such as TSVM [Joachims, 1999], which finds the hyperplane that separates both the labeled and unlabeled data with the maximum margin, graph-based semi-supervised learning methods make better use of the data distribution revealed by unlabeled data. In graph-based semi-supervised learning, a weighted graph is first constructed in which both the labeled and unlabeled data are represented as vertices. Then many of these methods can be viewed as estimating a function on the graph [Zhu, 2005]. Based on the assumptions that nearby points in the feature space are likely to have the same label, the function is defined to be locally smooth and consistent with the labeled data. Finally, the classification labels are obtained by comparing the function value and a pre-specified threshold. For example, in the Gaussian random fields and harmonic function method, the learning problem is formulated in terms of a Gaussian random field on the graph, and the mean of the field serves as the function [Zhu et al., 2003].

Another example is the local and global consistency method, in which the function at each point is iteratively determined by both the information propagated from its neighbors and its initial label [Zhou et al., 2004]. Yet another example is the graph mincut method whose function corresponds to partitioning the graph in a way that roughly minimizes the number of similar pairs of examples that are given different labels [Blum and Chawla, 2001]. In the mincut method, the function can only take binary values. Up till now, graph-based semi-supervised learning methods are generally approached from the discriminative perspective [Zhu, 2005] in that the function on the graph corresponds to posterior probabilities in one way or another. In the discriminative setting, however, the use of unlabeled data does not necessarily guarantee better decision boundaries. In addition, there is no clear explanation why the function on the graph should correspond to posterior probabilities from statistics point of view. In this paper, we propose a new graph-based semi-

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

supervised learning method from the generative model perspective. Specifically, the class conditional probabilities and the class priors are estimated from the weighted graph. The potential advantages involve several aspects: first, it can be theoretically justified that in the ideal cases where the two classes are separable, the output functions in terms of certain eigenvectors of the graph converge to the class conditional probabilities as the number of training data goes to infinity. In non-ideal cases, our functions still provide a good estimate of the class conditional probabilities. Finally, the estimated class priors make use of both the labeled and unlabeled data, which compensate for the lack of label information in many practical situations. Experimental results show that our approach leads to better performance than other existing graph-based methods on a variety of datasets. Hence we can claim both stronger theoretical justification and better empirical results.

System Architecture

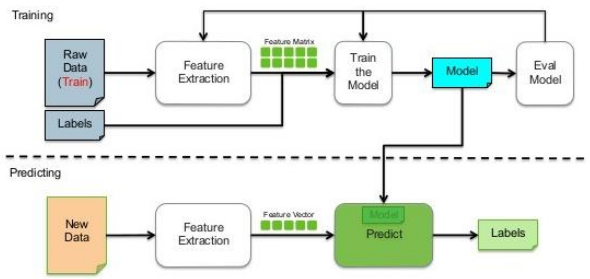

Fig: Semi-Supervised Learning

Implementation

Graph SSL Optimization

We learn a label distribution ˆ Y by minimizing the convex objective function:

$$C(\hat{Y}) = \mu_1 \sum_{v \epsilon V_1} 8vv \left|\left|\hat{Y}v - Yv\right|\right|_2^2 + \mu_2 \sum_{v \epsilon V, u \epsilon N(v)} wvu \left|\left|\hat{Y}v - \hat{Y}u\right|\right|^2 + \mu_3 \sum_{v \epsilon v} \left|\left|\hat{Y}v - U\right|\right|_2^2 \quad (1)$$

$$\text{s.t.} \sum_{l=1}^{L} \hat{Y}_{vl} = 1, Vv$$

where N(v) is the (incoming) neighbor node set of the node v, and U is the (uniform) prior distribution over all labels. The above objective function models that: 1) the label distribution should be close to the gold label assignment for all the seeds; 2) the label distribution of a pair of neighbors should be similar measured by their affinity score in the edge weight matrix; 3) the label distribution should be close to the prior U, which is a uniform distribution. The setting of the hyper parameters μ_i will be discussed. The optimization criterion is inspired from and similar to some existing approaches such as Adsorption and MAD but uses a slightly different objective function, notably the matrices have different constructions. In Section 5, we also compare our vanilla version against some of these baselines for completeness. The objective function in Equation 1 permits an efficient iterative optimization technique that is repeated The objective function in Equation 1 permits an efficient iterative optimization technique that is repeated 2The graph G can be directed or undirected depending on the task. Following most existing works in the literature, we use undirected edges for E in our experiments. until convergence. We utilize the Jacobi iterative algorithm which defines the approximate solution at the (i+1)th iteration, given the solution of the (i)th iteration as follows:

$$\hat{Y}_{vl}^{(i)} = \frac{1}{M_{vl}} \left(\mu_1 s_{vv} Y_{vl} + \mu_2 \sum_{u \epsilon N(v)} w_{vu} \hat{Y}_{ul}^{(i=1)} + \mu_3 U_l\right)$$

$$M_{vl} = \mu_1 s_{vv} + \mu 2 \sum_{u \epsilon N(u)} w_{vu} + \mu_3 \quad (2)$$

where i is the iteration index and $U_l = \frac{1}{m}$ which is the uniform distribution on label l. The iterative procedure starts with $\hat{Y}_{vl}^{(0)}$ v_l which is initialized with seed label weight Y_{vl} if v ϵ V_l, else with uniform distribution $\frac{1}{m}$. In each iteration I, $\hat{Y}_{vl}^{(i)}$ v_l aggregates the label distribution $\hat{Y}_{v}^{(i-1)}$ at iteration i−1 from all its neighbors uϵN(v). More details for deriving the update equation can be found in.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

We use the name EXPANDER to refer to this vanilla method that optimizes Equation 1. DIST-EXPANDER: Scaling To Large Data In many applications, semi-supervised learning becomes challenging when the graphs become huge. To scale to really large data sizes, we propose DISTEXPANDER, a distributed version of the algorithm that is directly suited towards parallelization across many machines. We turn to Pregel and its open source version. Graph as the underlying framework for our distributed algorithm. These systems follow a Bulk Synchronous Parallel (BSP) model of computation that proceeds in rounds. In every round, every machine does some local processing and then sends arbitrary messages to other machines. Semantically, we think of the communication graph as fixed, and in each round each node performs some local computation and then sends messages to its neighbors.

The specific systems like Pregel and Giraph build infrastructure that ensures that the overall system is fault tolerant, efficient, and fast. The programmer's job is simply to specify the code that each vertex will run at every round. Previously, some works have explored using MapReduce framework to scale to large graphs . But unlike these methods, the Pregelbased model is far more efficient and better suited for graph algorithms that fit the iterative optimization scheme for SSL algorithms. Pregel keeps vertices and edges on the machine that performs computation, and uses network transfers only for messages. MapReduce, however, is essentially functional, so expressing a graph algorithm as a chained MapReduce requires passing the entire state of the graph from one stage to the next—in general requiring much more communication and associated serialization overhead which results in significant network cost.

Algorithm 1 DIST-EXPANDER Algorithm

1: Input: A graph G = (V,E,W), where V = $V_l \cup V_u$ V_l = seed/labeled nodes, Vu = unlabeled nodes

2: Output: A label distribution $\hat{Y}v = \hat{Y}v1\ \hat{Y}v2 \ldots\ \hat{Y}vm$ for every node v ϵ V minimizing the overall objective function (1). Here, $\hat{Y}v_l$ represents the weight of label l assigned to the node v.

3: Let L be the set of all possible labels, |L| = m.

4: Initialize $\hat{Y}^0_{vl}$ with seed label weights if vϵVl, else$\frac{1}{m}$.

5: (Graph Creation) Initialize each node v with its neighbors N(v) = {u : (v, u) ϵ E}.

6: Partition the graph into p disjoint partitions V_1, ……..,Vp, where $\cup_i$Vi = V

7: for i = 1 to max_ iter do

8: Process individual partitions Vp in parallel.

9: for every node v ϵ Vp do

10: (Message Passing) Send previous label distribution $\hat{Y}^{i=1}_v$ to all neighbors u ϵ N(v).

11: (Label Update) Receive a message Mu from its neighbor u with corresponding label weights $\hat{Y}^{i=1}_v u$ .Process each message M1...M|N(v)| and update current label distribution $\hat{Y}^1_v$ iteratively using Equation (2).

12: end for

13: end for

Results:

In this section, we present the comparative experimental results on two datasets: Cedar Buffalo binary digits database [Hull, 1994], and a document genre-classification dataset [Liu et al., 2003]. Our algorithm is compared with two other graph-based semi-supervised learning methods: Gaussian random fields and the local and global consistency method. We did not compare with supervised learning methods, such as one nearest neighbor, since they have been proved to be less effective than Gaussian random fields based on experimental results. Here designed two kinds of experiments: balanced and unbalanced. In the balanced case, the ratio of labeled points from each class is always the same as the class priors; in the unbalanced case, if not explained otherwise, we fix the total number n_1 of labeled points, and perturb the number of positive labeled points around $n_1/2$ with a Gaussian distribution of mean 0 and standard deviation $n_1/10$.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

In each experiment, we gradually increase the number of labeled data, perform 20 trials for each labeled data volume, and average the accuracy at each volume point.

6.1 Cedar Buffalo Binary Digits Database:

We first perform experiments on Cedar Buffalo binary digits database [Hull, 1994] including two classification tasks: classifying digits "1" vs "2", with 1100 images in each class; and odd vs even digits, with 2000 images in each class (400 images for each digit). The data we use are the same as those used in [Zhu et al., 2003]. Here $Ø(x_i, x_j) = (2\pi\sigma^2)^{-d/2} \exp(-\|x_i - x_j\| / 2\sigma^2)$ where is the average distance between each data point and its 10 nearest neighbors.

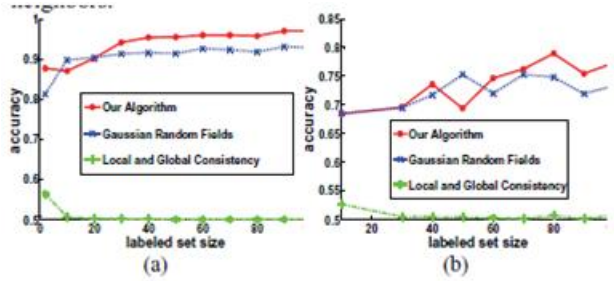

Figure 3. Balanced Classification. (a): 1 vs 2; (b) odd vs even

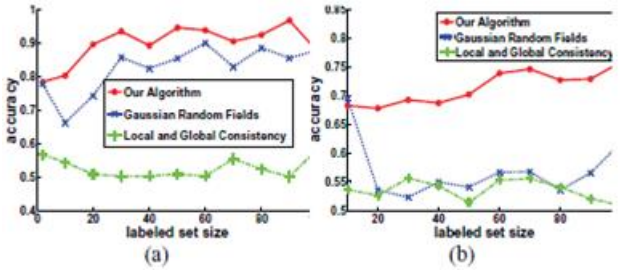

Figure 4. Unbalanced Classification. (a): 1 vs 2; (b) odd vs even

Figure 3(a) and 3(b) show the results of the two classification tasks in the balanced case. The performance of our algorithm is comparable with Gaussian random fields, and both of them are much better than the local and global consistency method. Figure 4(a) and 4(b) show the results in the unbalanced case. In this situation, the performance of Gaussian random fields is much worse than in the balanced case, while the performance of our algorithm is comparable to the balanced case. This is because the class mass normalization procedure adopted in Gaussian random fields depends on the labeled set only to estimate the class priors; while our algorithm makes use of both the labeled and the unlabeled set to estimate the class priors. Therefore, it is more robust against the perturbation in the proportion of the positive and negative data in the labeled set.

6.2 Genre Dataset:

Genre classification is to classify the documents based on its writing styles, such as political articles and movie reviews. The genre dataset that we use consists of documents from 10 genres, including biographies (b), interview scripts (is), movie reviews (mr), product reviews (pr), product press releases (ppr), product descriptions on store websites (pd), political articles on newspapers (pa), editorial papers on politics (ep), news (n), and search results from multiple search engines using 10 queries (sr). We randomly select 380 documents from each category to compose the whole dataset of 3800 documents. Each document is processed into a "tf.idf" vector, which is generated based on the top 10,000 most frequent words in this dataset after stemming, with the header and stop words removed. Here $Ø(x_i, x_j) = \exp(-(1-(x_i.x_j/(\|x_i - x_j\| /0.03)$,which is borrowed from [Zhu et al., 2003] and roughly measures the similarity between documents. The only difference is that we keep all the edges instead of keeping edges for only 10 nearest neighbors. Next we perform experiments to compare the three algorithms. The results are provided in Figure 5 and Figure 6 respectively.

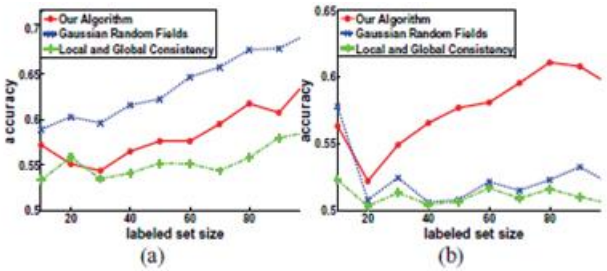

Figure 5. Classification between Random Partitions. (a): balanced; (b): unbalanced

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

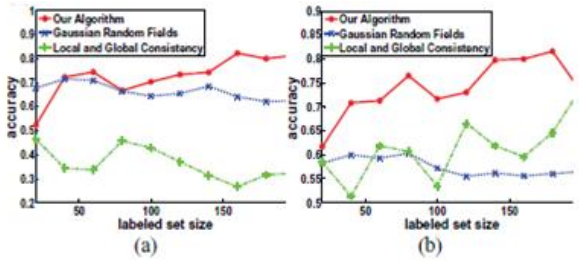

Figure 6. Unbalanced Classification. (a): pa vs other; (b) b vs other

For Figure 5, we randomly partition the 10 categories into two classes, i.e. pa, pr, sr, b, and is, vs mr, ppr, pd, ep and n. Figure 5(a) and 5(b) correspond to the balanced and unbalanced cases respectively. In the balanced case, Gaussian random fields are better than our algorithm and the local and global consistency method. This might be because the function $Ø(x_i, x_j)$does not have some of the nice properties required by Theorem 2. However, in the unbalanced case, Gaussian random fields tend to suffer a lot. On the contrary, our algorithm is quite robust despite of the perturbation. In Figure 6, we try to classify pa and b against all the other categories. In these experiments, the class priors are 0.1 for the positive class and 0.9 for the negative class. However, here we provide equal numbers of positive and negative points in the labeled set. From the figures, we can see that the performance of our algorithm is rather stable, while the performance of both Gaussian random fields and the local and global consistency method is largely affected by the misleading labeled set, since they only depend on the labeled set to estimate the class priors, either explicitly or implicitly.

Conclusion and Future Work:

In this paper, we propose a novel graph-based semi-supervised learning method to estimate both the class conditional probabilities and the class priors. It is a generative model, in contrast to existing graph-based methods, which are essentially discriminative. In the ideal case, the estimated class conditional probabilities have been proved to converge to the true value.

In the general case, our algorithm can still output reasonable estimates of the class conditional probabilities. For data points outside the training set, the class conditional probabilities are estimated via kernel regression. When estimating the class priors, we effectively use the unlabeled data to make up for the labeled data with unrepresentative class prior distributions. Experimental results on two datasets demonstrate the superiority of our algorithm over recent existing graph-based semi-supervised learning methods, especially when the proportion in the labeled set is not the same as the class priors. In our experiments, we notice that in some cases, adding even a single labeled point into the labeled set brings about significant improvement in classification accuracy; while in other cases, adding many labeled points into the labeled set does not help improve the performance. Currently we are incorporating active learning into our framework. Particularly, we are interested in determining when to invoke active learning (not just which instances to label) in order to achieve the biggest gain while minimizing incremental labeling cost.

References:

[1][Blum and Chawla, 2001] Blum, A., & Chawla, S. (2001). Learning from labeled and unlabeled data using graph
mincuts. ICML.

[2][Chung, 1997] Chung, F. R. K. (1997). Spectral graph theory, regional conference series in mathematics, no. 92. American Mathematical Society.

[3] [Grady and Funka-Lea, 2004] Grady, L., & Funka-Lea, G. (2004). Multi-label image segmentation for medical applications based on graph-theoretic electrical potentials. ECCV04, workshop on Computer Vision Approaches to Medical Image Analysis and Mathematical Methods in Biomedical Image Analysis.

[4][He et al., 2004] He, J., Li, M., Zhang, H. J., Tong, H., & Zhang, C. (2004). Manifold-ranking based

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

image retrieval. Proc. 12th ACM International Conf. on Multimedia. [Hull, 1994] Hull, J. J. (1994). A database for handwritten text recognition research. IEEE Transactions on Pattern Analysis and Machine Intelligence, 16.

[5][Joachims, 1999] Joachims, T. (1999). Tranductive Inference for text classification using Support Vector Machines. ICML.

[6][Liu et al., 2003] Liu, Y., Carbonell, J., & Jin, R. (2003). A New Pairwise Ensemble Approach for Text Classification. ECML.

[7][Niu, Ji and Tan, 2005] Niu, Z. Y., Ji, D. H., & Tan, C. L. (2005). Word sense disambiguation using label propagation based semi-supervised learning. Proc. 43rd Meeting of the Association for Computational Linguistics.

[8][Zhou et al., 2004] Zhou, D., bousquet, O., Lal, T., Weston, J., & Schlkopf, B. (2004). Leaning with local and global consistency. NIPS.

[9][Zhu et al., 2003] Zhu, X., Ghahramani, Z., & Lafferty, J. (2003). Semi-supervised learning using Gaussian fields and harmonic functions. ICML.

[10][Zhu et al., 2005] Zhu, X., & Lafferty, J. (2005). Harmonic mixtures: combining mixture models and graph-based methods for inductive and scalable semi-supervised learning. ICML.

[11][Zhu, 2005] Zhu, X. (2005). Semi-supervised learning with graphs. Doctoral dissertation, School of Computer Science, Carnegie Mellon University.ŶŶ.