Graph Quality Judgement: A Large Margin Expedition*

Yu-Feng Li Shao-Bo Wang Zhi-Hua Zhou
National Key Laboratory for Novel Software Technology, Nanjing University
Collaborative Innovation Center of Novel Software Technology and Industrialization
Nanjing 210023, China
{liyf,wangsb,zhouzh}@lamda.nju.edu.cn

Abstract

Graph as a common structure of machine learning, has played an important role in many learning tasks such as graph-based semi-supervised learning (GSSL). The quality of graph, however, seriously affects the performance of GSSL; moreover, an inappropriate graph may even cause *deteriorated* performance, that is, GSSL using unlabeled data may be outperformed by direct supervised learning with only labeled data. To this end, it is desired to judge the quality of graph and develop performance-safe GSSL methods. In this paper we propose a large margin separation method LEAD for safe GSSL. Our basic idea is that, if a certain graph owns a high quality, its predictive results on unlabeled data may have a large margin separation. We should exploit the large margin graphs while keeping the small margin graphs (which might be risky) to be rarely exploited. Based on this recognition, we formulate safe GSSL as Semi-Supervised SVM (S3VM) optimization and present an efficient algorithm. Extensive experimental results demonstrate that our proposed method can effectively improve the safeness of GSSL, in addition achieve highly competitive accuracy with many state-of-the-art GSSL methods.

1 Introduction

As a convenient way to describe data relationship, graph is a kind of important structure in machine learning and has played a significant role in many learning tasks [Ng *et al.*, 2001; Zhu, 2007] such as graph-based semi-supervised learning (GSSL) [Blum and Chawla, 2001; Zhu *et al.*, 2003; Zhou *et al.*, 2004]. With the help of graph, GSSL obtains many advantages such as closed-form solution, promising learning performance, and thus GSSL attracts significant attention since it was proposed and has been widely applied in a large amount of applications [Liu *et al.*, 2012].

However, it is widely known that the graph quality, rather than the learning/optimization algorithm, seriously affects the performance of GSSL methods [Zhu, 2007; Belkin and Niyogi, 2008; Wang and Zhang, 2008; Jebara *et al.*, 2009]. Moreover, it has been found in many empirical results [Zhou *et al.*, 2004; Belkin and Niyogi, 2004; Wang and Zhang, 2008; Karlen *et al.*, 2008] that an inappropriate graph may cause GSSL to deteriorate the performance. That is, GSSL with the use of unlabeled data not only does not improve performance, but sometimes it may even be outperformed by direct supervised learning with only a small amount of labeled data. Such a deficiency hinders GSSL to play an important role in more applications. To this end, studying the quality of the graph and developing performance-*safe* GSSL method are desired. Specifically, it is desirable to judge the quality of graph such that GSSL could often improve performance, while in the worst case it will not be outperformed by direct supervised learning with only labeled data. This task, to our best knowledge, has not been studied in literature.

In this paper, we present a large margin approach named LEAD (LargE margin grAph quality juDgement) inspired by the success of large margin criterion [Vapnik, 1998]. Our basic idea is simple and intuitive. That is, when a certain graph owns a high quality, its predictive results on the unlabeled data may have a large margin separation. Table 1 illustrates the idea through four representative data sets. It can be seen that when one graph owns a better accuracy on the unlabeled data, it may also suffer a smaller hinge loss (a loss w.r.t. large margin separation) [Vapnik, 1998]. In other words, large margin separation may help judge the quality of graph. Based on this recognition, we should exploit the large margin graphs while keeping the small margin graphs (which might be risky) to be rarely exploited. This motivates our proposed LEAD method. LEAD views the predictive results of multiple graphs as features, and then constructs a large margin classifier based on these features. These procedures consequently formulate safe GSSL as Semi-Supervised SVM (S3VM) [Vapnik, 1998; Joachims, 1999] optimization and an efficient algorithm is presented. Extensive experimental results demonstrate that LEAD can effectively improve the safeness of GSSL, in addition achieve highly competitive accuracy with many state-of-the-art GSSL methods.

The paper is organized as follows. Section 2 briefly reviews related work. Section 3 presents the proposed method. Empirical results are reported in Section 4. Finally, Section 5 gives a conclusive remark of this paper.

*This research was supported by NSFC (61333014, 61403186), JiangsuSF (BK20140613), 863 Program (2015AA015406).

Table 1: Illustration on large margin separation assumption for graph quality. Two 5-nearest neighbor (5NN) graphs with different distances are used as examples. 'Accuracy' refers to the accuracy (%) on unlabeled data. 'Hinge Loss' refers to the average hinge loss (a loss w.r.t. large margin separation) on unlabeled data when applying S3VM on the predictive result of the graph. 10 labeled instances are used and average results (mean±std) over 20 random splits are reported.

Dataset	Domain	5NN Graph with Euclidean Distance		5NN Graph with Manhattan Distance	
		Accuracy	Hinge Loss	Accuracy	Hinge Loss
breast-cancer	life	91.6±3.1	0.529±0.110	92.6±2.7	0.370±0.106
coil	image	60.9±6.2	0.341±0.109	62.4±6.9	0.276±0.109
musk-1	physical	57.7±4.9	0.632±0.139	56.4±4.5	0.671±0.145
text	text	52.3±3.3	0.964±0.006	50.3±0.0	0.994±0.009

2 Related Work

GSSL attracts significant attention since it was proposed. A large number of works have been presented on optimizing a label assignment of the unlabeled data, e.g. [Blum and Chawla, 2001; Zhu *et al.*, 2003; Joachims, 2003; Zhou *et al.*, 2004; Belkin *et al.*, 2006; Jebara *et al.*, 2009] and constructing the graph, e.g. [Carreira-Perpiñán and Zemel, 2005; Jebara *et al.*, 2009; Wang and Zhang, 2008]. For optimizing the label assignment, just name a few, Zhu et al. [2003] proposed to use the belief propagation technique and derived a closed-form solution; Zhou et al.[2004] formulated GSSL as a convex regularization problem with a simple iterative algorithm. For the graph construction, many kinds of graphs have been considered. For example, k-nearest neighbor graph, ϵ-neighborhood graph, minimal spanning tree [Carreira-Perpiñán and Zemel, 2005], b-matching graph [Jebara *et al.*, 2009], locally linear reconstruction graph [Wang and Zhang, 2008], etc. It is also reported that rather than the type of the graph, the selection of distance metric used in the graph will also significantly affect the final GSSL performance [Zhu, 2007]. There are also approaches, e.g., [Zhu *et al.*, 2005; Argyriou *et al.*, 2005] proposed to optimize the graph construction as well as the label assignment simultaneously.

It is now widely accepted that, with the deepening of research, the quality of the graph construction plays a key role to the learning performance of GSSL method [Zhu, 2007; Belkin and Niyogi, 2008; Wang and Zhang, 2008; Jebara *et al.*, 2009]. What is more serious, as stated in many empirical studies [Zhou *et al.*, 2004; Belkin and Niyogi, 2004; Wang and Zhang, 2008; Karlen *et al.*, 2008], an inappropriate graph can even cause a degenerated performance. How to identify the quality of graph so as to avoid the hurt of using unlabeled data in GSSL has not been deeply studied yet.

There are some discussions in literatures, e.g., [Cozman *et al.*, 2002; Ben-David *et al.*, 2008; Balcan and Blum, 2010] on the reasons about the decreased performance of general semi-supervised learning. For example, Cozman et al. [2002] conjectured that the performance degradation on semi-supervised learning is caused by incorrect model assumptions. However, without large amount of domain knowledge, it is difficult to make correct model assumptions. Ben-David et al. [2008], from a theoretical perspective, pointed out that semi-supervised learning is not necessarily able to achieve better generalization performance when there is not sufficient domain knowledge. Yet they did not give a feasible solution. Balcan and Blum [2010] showed that when unlabeled data is able to provide a good regularizer, a purely inductive supervised SVM on labeled data using such a regularizer guarantees a good generalization. However, deriving such a good regularizer is quite difficult.

Recently there are some efforts devoted to develop safe semi-supervised learning approaches, e.g., [Li and Zhou, 2005; 2015; Li *et al.*, 2016]. However, developing safe GSSL approaches remains challenging. To our best knowledge, this paper is the first proposal on this aspect.

3 The LEAD Method

To judge the quality of graph so as to alleviate the performance degradation of GSSL methods, we present the LEAD method with the use of large margin principle. As mentioned, when a certain graph owns a high quality, its predictive results on the unlabeled data may derive a large margin separation. In other words, if one predictive result derived by a graph has a small margin, then it may cause a higher risk of performance degradation when using the graph. Therefore, given multiple graphs with unknown quality, one should encourage to use the graphs with large margin, rather than the graphs with small margin, and consequently reduce the chances of performance degradation. Based on the above assumption, by treating the predictive results of multiple graphs as new input training features, the proposed method constructs a large margin classifier for both the labeled and unlabeled data to derive a relatively safe predictive result for GSSL.

Specifically, in GSSL, given a few labeled instances $\{\mathbf{x}_i, y_i\}_{i=1}^{l}$ and a large amount of unlabeled instances $\{\mathbf{x}_j\}_{j=l+1}^{l+u}$ ($l \ll u$) where $y \in \{+1, -1\}$ is the output label for input instance $\mathbf{x}$. Let $\{G_t = (\mathcal{V}, \mathcal{E}_t, \mathbf{W}_t)\}_{t=1}^{T}$ denote multiple graphs where T refers to the number of graphs. Here $\mathcal{V}$ is a set of $l+u$ nodes each corresponds to one instance. $\mathcal{E}$ is a set of undirected edges between node pairs. $\mathbf{W} \in \mathbb{R}^{(l+u)\times(l+u)}$ is a nonnegative and symmetric adjacency weighted matrix associating with $\mathcal{E}$ in G, i.e., the weight w_{ij} on the edge $e_{ij} \in \mathcal{E}$ reflects the similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$.

For each graph, GSSL aims to infer an optimal label assignment $\mathbf{z} = [z_1, \dots, z_{l+u}]$ such that the label inconsistency with respect to the graph as well as the existing labeled instances is minimized. This procedure is formulated as,

$$\begin{aligned} \min_{\mathbf{z}} \quad & \sum_{e_{ij}\in\mathcal{E}} w_{ij}\|z_i - z_j\|^2 \qquad (1)\\ \text{s.t.} \quad & z_i = y_i,\ i = 1,\ldots,l;\\ & z_j \in [-1,1],\ j = l+1,\ldots,l+u; \end{aligned}$$

Eq.(1) is a convex quadratic optimization and could be solved via many efficient GSSL algorithms, for example, the Harmonic method by [Zhu *et al.*, 2003]. Let $\mathbf{z}^{(t)} = [z_1^{(t)},\ldots,z_{l+u}^{(t)}]$ denote the optimal solution of Eq.(1) with respect to the graph G_t, $t = 1,\ldots,T$. Without sufficient domain knowledge, it is difficult to distinguish the quality of these solutions. If one chooses an inappropriate solution, GSSL may cause a performance degradation. To this end, we propose to use large margin principle to help distinguish the quality of the solutions.

Specifically, let $\mathbf{u}_i$ denote a vector where each entry corresponds to the predictive result of each graph on instance $\mathbf{x}_i$, i.e., $\mathbf{u}_i = [z_i^{(1)},\ldots,z_i^{(T)}]$. We regenerate a new semi-supervised training set where $\{\mathbf{u}_i, y_i\}_{i=1}^{l}$ denotes the new labeled examples and $\{\mathbf{u}_j\}_{j=l+1}^{l+u}$ the new unlabeled instances. A large margin linear classifier is then build to separate both the new labeled and unlabeled data. Intuitively, the large margin classifier is lateral to the use of large margin graphs, and avoids the direct utilization of small margin graph, and therefore the chance of performance degradation can be reduced.

Formally, we seek to find a linear classifier $f(\mathbf{u}) = \mathbf{w}'\mathbf{u} + b$ and a label assignment of unlabeled data $\hat{\mathbf{y}} = [\hat{y}_{l+1},\ldots,\hat{y}_{l+u}]$ that minimize the following optimization,

$$\begin{aligned} \min_{\mathbf{w},\hat{\mathbf{y}}} \quad & \frac{1}{2}\|\mathbf{w}\|_2^2 + C_1\sum_{i=1}^{l}\ell(y_i f(\mathbf{u}_i)) + C_2\sum_{j=l+1}^{l+u}\ell(\hat{y}_j f(\mathbf{u}_j))\\ \text{s.t.} \quad & \hat{y}_{l+j} \in \{+1,-1\},\ \ j = 1,\ldots,u;\\ & |\frac{\sum_{j=l+1}^{l+u}\hat{y}_j}{u} - \frac{\sum_{i=1}^{l}y_i}{l}| \le \beta \qquad (2) \end{aligned}$$

where $\ell(v) = \max\{0, 1-v\}$ is the hinge loss in large margin separation and the last constraint is *balanced constraint* [Joachims, 1999] that enforces the class ratio on unlabeled data to be closely related to that of labeled data (β is a small constant). C_1 and C_2 are parameters trading off the losses on the labeled and unlabeled data, respectively.

Eq.(2) is no more than the classical semi-supervised SVM (S3VM) [Vapnik, 1998; Joachims, 1999]. That is, safe GSSL is posed as S3VM optimization. To our best knowledge this is the first time to connect safe GSSL to S3VM.

To solve the optimization in Eq.(2), one direct solution is the use of traditional S3VM algorithms [Joachims, 1999; Chapelle *et al.*, 2008; Li *et al.*, 2013]. However, these algorithms do not fully exploit the structure in our setting, and may be not efficient. For example, here linear kernel is used while traditional S3VMs often consider the use of non-linear kernels. To this end, we present an efficient alternating optimization algorithm for the solving of Eq.(2).

It alternatively optimizes $\mathbf{w}$ (or $\hat{\mathbf{y}}$) by fixing the $\hat{\mathbf{y}}$ (or $\mathbf{w}$) as constants. Specifically, when $\hat{\mathbf{y}}$ is fixed, Eq.(2) is a standard linear SVM form and $\mathbf{w}$ can be solved by a linear SVM package like LIBLINEAR [Fan *et al.*, 2008] efficiently. When $\mathbf{w}$ is fixed, it has been proved that the rank of the elements of the optimal $\hat{\mathbf{y}}$ will be consistent with that of the prediction $\mathbf{w}'\mathbf{u} + b$ on unlabeled data [Zhang *et al.*, 2007]. Therefore, the optimal $\hat{\mathbf{y}}$ can be derived by the following closed-form solution.

Algorithm 1 The LEAD Method

```
Input: labeled instances $\{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$, unlabeled instances $\{\mathbf{x}_j\}_{j=l+1}^{l+u}$, regularization coefficients $C_1$ and $C_2$, a set of candidate graphs $\{G_t = (\mathcal{V}, \mathcal{E}_t, \mathbf{W}_t)\}_{t=1}^{T}$, the label assignment $\hat{\mathbf{y}}^0 = [\hat{y}_{l+1}^0,\ldots,\hat{y}_{l+u}^0]$ of a direct supervised learning method using only labeled instances $\{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$
Output: a label assignment on the unlabeled instances $\hat{\mathbf{y}} = [\hat{y}_{l+1},\ldots,\hat{y}_{l+u}]$
Perform a GSSL algorithm (in this paper the methods in [Zhu et al., 2003] and [Zhou et al., 2004] are both implemented for comparison) on a set of graphs $\{G_t\}_{t=1}^{T}$; record $\mathbf{z}^{(t)} = [z_1^{(t)},\ldots,z_{l+u}^{(t)}]$ as the optimal solution with respect to the graph $G_t$, $t = 1,\ldots,T$
Regenerate a new training data set. $\{\mathbf{u}_i, y_i\}_{i=1}^{l}$ denotes the new labeled instances and $\{\mathbf{u}_j\}_{j=l+1}^{l+u}$ the unlabeled instances, where $\mathbf{u}_i = [z_i^{(1)},\ldots,z_i^{(T)}]$, $i = 1,\ldots,l+u$
Initialize $\hat{\mathbf{y}} = \mathrm{sign}(\frac{1}{T}\sum_{t=1}^{T}\mathbf{z}^{(t)})$ and $\hat{C}_2 = 10^{-6}C_1$
repeat
    repeat
        Fix $\hat{\mathbf{y}}$ and update the solution of $\mathbf{w}$ via a linear SVM solver like LIBLINEAR [Fan et al., 2008]
        Fix $\mathbf{w}$ and update the solution of $\hat{\mathbf{y}}$ via Eq.(3)
    until the objective of Eq.(2) does not decrease
    $\hat{C}_2 = 2\hat{C}_2$
until $\hat{C}_2 \ge C_2$
If $\hat{y}_{l+j}(\mathbf{w}'\mathbf{u}_{l+j} + b) \le 1$, then $\hat{y}_{l+j} = \hat{y}_{l+j}^0$, $j = 1,\ldots,u$
return $\hat{\mathbf{y}} = [\hat{y}_{l+1},\ldots,\hat{y}_{l+u}]$
```

$$\hat{y}_{l+j} = \begin{cases} +1 & \text{if } r_j \le \big(\frac{2\sum_{i=1}^{l}y_i}{l} - \beta\big)u\\ -1 & \text{if } r_j \ge \big(\frac{2\sum_{i=1}^{l}y_i}{l} + \beta\big)u\\ \mathrm{sign}(\mathbf{w}'\mathbf{u}_{l+j} + b) & \text{otherwise} \end{cases} \qquad (3)$$

where $\{r_1,\ldots,r_u\}$ are the ranks of the predictions on the unlabeled data $\{\mathbf{u}_{l+1},\ldots,\mathbf{u}_{l+u}\}$ (sorted in a descending order). The larger prediction, the smaller rank. To further improve the quality of the solution, inspired by [Joachims, 1999], we first assign a small importance for the unlabeled data (i.e., C_2 is initialized as a very small constant), and then gradually increase the importance of the unlabeled data (i.e., the value of C_2) until it reaches an upper bound. Finally, the unlabeled instances lying within the margin remain risky to use and thus their labels are assigned with the direct supervised learning method. The pseudocode of the proposed LEAD method is given in Algorithm 1.

4 Experiment

4.1 Setting

To evaluate the effectiveness of our proposal, we conduct experimental comparison on a number of binary data sets[1] that cover a wide range of properties (Table 2). The sample size ranges from 294 to more than 1,5000. The feature dimensionality ranges from 5 to more than 10,000. The proportion of classes (i.e., ratio of the number of positive samples to that of negative samples) ranges from less than 0.2 to more than 1.5.

LEAD is compared with the following methods.

- **1NN**: The baseline supervised learning method. Each unlabeled instance is assigned with the label of its nearest labeled instance.
- **Harmonic**[2]: The classical harmonic function method proposed in [Zhu *et al.*, 2003]. Without sufficient domain knowledge of graph construction, k-nearest neighbor graph is recognized as a good candidate graph [Zhu, 2007]. Therefore, three types of nearest neighbor graphs (namely 3, 5 and 7 nearest neighbor graphs) are considered for the Harmonic method.
- **LLGC**: The classical LLGC (local and global consistency) method proposed in [Zhou *et al.*, 2004]. Similar to the Harmonic method, 3, 5 and 7 nearest neighbor graphs are employed.
- **CGL**[3]: This is a GSSL method [Argyriou *et al.*, 2005] that learns a graph from multiple candidate graphs through worst-case.
- **Majority Voting**: We also compare with the majority voting method which is known as a popular approach in dealing with multiple predictive results and has been found useful in many situations [Zhou, 2012].
- **GSSL-CV**: We further compare with the cross-validation method which is widely accepted as a popular method to perform model selection. Here the cross-validation method is performed to select a graph from a set of candidate graphs via the cross-validation result.

Table 2: Experimental Data Sets

Data	# Dim	# Pos	# Neg	# Total
heart-hungarian	12	106	188	294
vertebral	6	100	210	310
ionosphere	33	225	126	351
horse	25	136	232	368
musk-1	166	207	269	476
credit	15	383	307	690
breast-cancer	9	241	458	699
mammographic	5	445	516	961
coil	241	750	750	1,500
digit1	241	734	766	1,500
text	11,960	750	750	1,500
usps	241	300	1,200	1,500
spambase	57	1,813	2,788	4,601
musk-2	166	1,017	5,581	6,598
twonorm	20	3,697	3,703	7,400
mushroom	21	3,916	4,208	8,124
mnist4vs9	629	6,824	6,958	13,782
mnist3vs8	631	7,141	6,825	13,966
mnist7vs9	600	7,293	6,958	14,251
mnist1vs7	652	7,877	7,293	15,170

The distance metric used to determine nearest neighbors about the above compared methods is set as the Euclidean distance. For the Harmonic and CGL method, the parameters are set to the recommended ones in the package. For the LLGC method, since the authors do not share the code, it is implemented by ourselves and the parameter α is set to 0.99 recommended in the paper. For the Majority Voting method, the discrete label assignments from a set of candidate graphs are integrated and the label ratio of unlabeled data is enforced to be similar to that of labeled data. For the GSSL-CV method, 5-fold cross-validation is conducted (we have conducted other types of cross-validation method, like 2-fold and 10-fold cross-validation, and 5-fold cross-validation performs the best). For the LEAD method, the parameters C_1, C_2 and β are set to 1, 0.01 and 0.02 for all the experimental settings in this paper. 9 candidate graphs from 3, 5 and 7 nearest neighbor graphs based on 3 distance metrics (i.e., Euclidean, Manhattan and Cosine distance) [Zhu, 2007] are exploited, and correspondingly the GSSL predictive results of LEAD are from the output of classical GSSL methods (i.e., the Harmonic and LLGC methods) on the above graphs. For each data set, 10 instances are labeled and the rest are unlabeled. The class ratio is maintained on both sets. Each experiment is repeated 20 times, and the average accuracy (mean ± std) on the unlabeled data is reported.

4.2 Performance Results

Table 3 shows the comparison result. The proposed LEAD method obtains highly competitive accuracy with compared GSSL methods in the ability of performance improvement. Firstly, in terms of the average accuracy, the proposed LEAD method obtains competitive performance when exploiting both the Harmonic and the LLGC method. Secondly, in terms of the frequency of the performance improvement, LEAD obtains quite good performance (It achieves significant improvement in 9 and 12 cases, which are the most among all the compared methods using the Harmonic and the LLGC method, respectively).

More importantly, unlike the compared GSSL methods that will significantly cause a decreased performance in many cases (regardless of the use of the Harmonic method or the LLGC method), the proposed LEAD method does not decrease the performance significantly. Such a advantage, will help GSSL to play a role in more applications, especially those which require a high reliability for the exploitation of unlabeled data.

Both the Majority Voting method and the GSSL-CV method are to some extent capable of reducing the chance of the performance degradation and improving the accuracy, however they still decrease the performance in some cases,

[1] Downloaded from http://olivier.chapelle.cc/ssl-book/benchmarks.html and http://archive.ics.uci.edu/ml/datasets.html

[2] http://pages.cs.wisc.edu/~jerryzhu/pub/harmonic_function.m

[3] http://cvn.ecp.fr/personnel/andreas/code/graph/index.html

Table 3: Accuracy (mean ± std) for the compared methods and our LEAD method with 10 labeled instances. If the performance is significantly better/worse than 1NN (paired t-tests at 95% significance level), the corresponding entries are bolded/underlined. The average accuracy over 20 data sets is listed for comparison. The win/tie/loss counts against 1NN are summarized and the method with the smallest number of losses is bolded. '-' means there are some data sets where CGL method could not be terminated within 24 hours and thus the average accuracy is not available (denoted by 'N/A') for comparison.

Data sets	1NN		Harmonic		CGL	Majority Voting	GSSL-CV	LEAD
		3NN Graph	5NN Graph	7NN Graph				
breast-cancer	94.0 ± 2.6	**95.7 ± 2.1**	**95.6 ± 1.1**	**95.3 ± 1.4**	94.7 ± 2.7	92.2 ± 0.8	**95.8 ± 2.1**	**94.1 ± 2.5**
coil	60.4 ± 5.2	**67.7 ± 8.0**	**64.2 ± 5.9**	62.8 ± 6.9	**65.6 ± 6.3**	**65.3 ± 6.9**	**67.6 ± 7.1**	**63.5 ± 6.9**
credit	72.0 ± 6.1	70.2 ± 6.0	70.3 ± 8.4	67.9 ± 9.6	68.0 ± 6.7	71.5 ± 6.9	70.1 ± 6.8	72.2 ± 5.8
digit1	73.3 ± 3.9	**87.4 ± 4.9**	**87.4 ± 5.6**	**86.1 ± 6.7**	**93.6 ± 2.3**	**92.4 ± 2.7**	**90.8 ± 4.5**	**79.1 ± 3.8**
heart-hungarian	76.2 ± 7.3	73.9 ± 5.8	74.9 ± 4.5	76.2 ± 5.5	75.7 ± 10.0	76.2 ± 6.4	75.4 ± 5.7	76.3 ± 7.2
horse	62.9 ± 5.0	64.5 ± 7.6	63.6 ± 8.4	64.3 ± 6.7	60.6 ± 8.1	64.6 ± 5.0	65.3 ± 7.0	62.9 ± 5.1
ionosphere	73.4 ± 6.8	72.0 ± 6.3	72.7 ± 8.6	72.8 ± 8.9	69.9 ± 2.9	75.1 ± 5.7	75.1 ± 7.0	74.0 ± 6.9
mammographic	73.5 ± 5.9	66.6 ± 5.8	67.8 ± 5.1	69.3 ± 5.2	71.7 ± 6.2	70.5 ± 6.1	66.0 ± 5.6	74.2 ± 4.9
mnist1vs7	92.3 ± 3.3	**97.7 ± 4.7**	**97.6 ± 4.3**	95.4 ± 9.5	-	**97.4 ± 0.3**	**97.8 ± 4.7**	**96.8 ± 1.1**
mnist3vs8	79.1 ± 3.8	63.5 ± 18.2	67.7 ± 21.0	62.1 ± 18.2	-	84.3 ± 19.5	77.8 ± 21.0	**80.5 ± 3.8**
mnist4vs9	67.0 ± 6.2	67.8 ± 16.5	62.7 ± 14.5	58.6 ± 11.8	-	**76.6 ± 13.7**	72.4 ± 15.8	**67.4 ± 6.3**
mnist7vs9	76.0 ± 3.7	70.2 ± 20.0	63.7 ± 18.5	59.4 ± 14.1	-	79.3 ± 15.3	80.9 ± 15.2	**77.9 ± 4.3**
mushroom	79.3 ± 7.9	79.3 ± 7.9	79.3 ± 7.9	79.3 ± 7.9	-	79.1 ± 7.9	**80.4 ± 7.0**	80.1 ± 7.3
musk-1	60.5 ± 4.0	61.1 ± 5.8	60.3 ± 6.0	60.6 ± 5.4	60.3 ± 5.3	61.8 ± 5.0	61.0 ± 6.6	60.5 ± 4.1
musk-2	73.9 ± 6.7	**76.2 ± 5.9**	**78.5 ± 5.9**	**78.7 ± 6.4**	57.3 ± 5.7	**77.2 ± 4.6**	**80.0 ± 5.4**	75.2 ± 5.6
spambase	71.9 ± 5.5	62.4 ± 6.3	62.5 ± 5.6	61.8 ± 7.6	-	62.7 ± 10.3	62.8 ± 6.5	**72.5 ± 5.0**
text	59.2 ± 5.6	54.9 ± 4.0	53.2 ± 3.0	53.9 ± 5.4	**63.1 ± 6.3**	56.0 ± 4.6	54.3 ± 4.0	59.2 ± 5.6
twonorm	89.2 ± 3.6	60.3 ± 18.2	57.5 ± 15.3	62.6 ± 19.0	**96.8 ± 0.2**	90.5 ± 13.8	88.3 ± 14.8	**89.2 ± 3.6**
usps	81.3 ± 2.8	80.8 ± 0.9	80.5 ± 0.8	80.3 ± 0.7	62.6 ± 6.8	70.7 ± 3.3	80.8 ± 0.9	81.3 ± 2.8
vertebral	70.3 ± 7.5	71.0 ± 3.4	72.1 ± 2.6	71.9 ± 3.1	68.0 ± 9.0	71.0 ± 4.9	72.4 ± 5.0	70.4 ± 7.5
Ave. Accuracy	74.3 ± 9.9	72.2 ± 11.4	71.6 ± 12.0	71.0 ± 11.9	N/A	75.7 ± 11.3	75.7 ± 11.6	75.4 ± 10.2
Win/Tie/Loss		5/10/5	5/9/6	3/9/8	4/6/4	5/10/5	6/11/3	**9/11/0**

Data sets	1NN		LLGC		CGL	Majority Voting	GSSL-CV	LEAD
		3NN Graph	5NN Graph	7NN Graph				
breast-cancer	94.0 ± 2.6	**95.9 ± 0.7**	**95.4 ± 1.1**	94.7 ± 1.5	94.7 ± 2.7	92.7 ± 0.6	**96.1 ± 0.7**	**94.1 ± 2.5**
coil	60.4 ± 5.2	**66.8 ± 6.5**	**63.6 ± 6.2**	61.7 ± 6.4	**65.6 ± 6.3**	**64.0 ± 6.1**	**67.4 ± 6.7**	**64.3 ± 6.6**
credit	72.0 ± 6.1	72.3 ± 5.3	69.7 ± 7.6	66.1 ± 8.2	68.0 ± 6.7	73.3 ± 6.2	72.6 ± 7.1	72.5 ± 5.5
digit1	73.3 ± 3.9	**90.4 ± 3.5**	**90.4 ± 3.3**	**90.2 ± 3.7**	**93.6 ± 2.3**	**91.5 ± 3.1**	**91.2 ± 3.7**	**84.2 ± 3.7**
heart-hungarian	76.2 ± 7.3	75.7 ± 4.9	73.8 ± 4.2	72.7 ± 6.0	75.7 ± 10.0	74.9 ± 7.2	75.7 ± 5.4	**76.4 ± 7.2**
horse	62.9 ± 5.0	65.2 ± 6.1	64.4 ± 5.7	62.9 ± 5.3	60.6 ± 8.1	63.3 ± 5.1	65.3 ± 6.2	62.9 ± 5.1
ionosphere	73.4 ± 6.8	70.6 ± 7.7	68.2 ± 7.0	67.2 ± 6.4	69.9 ± 2.9	71.5 ± 7.6	71.1 ± 7.2	73.3 ± 6.9
mammographic	73.5 ± 5.9	67.7 ± 5.9	69.8 ± 4.3	71.5 ± 4.8	71.7 ± 6.2	71.7 ± 4.8	67.3 ± 5.7	74.1 ± 4.7
mnist1vs7	92.3 ± 3.3	**98.6 ± 1.5**	**98.8 ± 0.8**	**98.8 ± 0.7**	-	**97.4 ± 0.4**	**98.4 ± 1.7**	**98.7 ± 1.1**
mnist3vs8	79.1 ± 3.8	**95.5 ± 1.7**	**95.5 ± 2.0**	**95.4 ± 2.1**	-	**95.8 ± 1.8**	**95.8 ± 2.0**	**93.6 ± 2.1**
mnist4vs9	67.0 ± 6.2	**88.8 ± 7.7**	**87.8 ± 7.4**	**87.1 ± 7.7**	-	**88.3 ± 7.7**	**88.8 ± 7.6**	**84.6 ± 7.4**
mnist7vs9	76.0 ± 3.7	**94.7 ± 3.9**	**93.7 ± 4.5**	**93.5 ± 4.4**	-	**94.1 ± 4.1**	**93.8 ± 4.2**	**91.5 ± 4.3**
mushroom	79.3 ± 7.9	79.3 ± 7.9	79.3 ± 7.9	79.3 ± 7.9	-	79.1 ± 7.9	**80.4 ± 7.0**	80.2 ± 7.1
musk-1	60.5 ± 4.0	61.4 ± 5.2	59.4 ± 5.2	60.0 ± 4.6	60.3 ± 5.3	60.7 ± 5.1	61.4 ± 5.1	60.2 ± 4.1
musk-2	73.9 ± 6.7	74.7 ± 6.3	**77.3 ± 4.6**	**78.6 ± 4.9**	57.3 ± 5.7	**77.7 ± 3.8**	**77.9 ± 5.6**	75.9 ± 5.1
spambase	71.9 ± 5.5	72.1 ± 5.3	73.2 ± 4.8	72.7 ± 6.3	-	**79.4 ± 5.1**	71.3 ± 5.7	**78.1 ± 2.9**
text	59.2 ± 5.6	58.7 ± 4.0	58.0 ± 4.6	57.8 ± 5.8	**63.1 ± 6.3**	60.1 ± 5.3	57.9 ± 5.2	59.2 ± 5.6
twonorm	89.2 ± 3.6	**96.0 ± 0.5**	**96.3 ± 0.6**	**96.5 ± 0.7**	**96.8 ± 0.2**	**97.1 ± 0.3**	**96.6 ± 0.5**	**94.6 ± 1.0**
usps	81.3 ± 2.8	**82.9 ± 2.2**	81.8 ± 1.9	81.0 ± 1.4	62.6 ± 6.8	76.1 ± 4.6	**83.0 ± 2.2**	**81.5 ± 2.8**
vertebral	70.3 ± 7.5	69.6 ± 2.2	68.4 ± 2.2	68.0 ± 1.3	68.0 ± 9.0	64.9 ± 7.1	71.3 ± 4.0	**70.7 ± 7.2**
Ave. Accuracy	74.3 ± 9.9	78.8 ± 13.0	78.2 ± 13.4	77.8 ± 13.6	N/A	78.7 ± 12.8	79.2 ± 13.0	78.5 ± 11.9
Win/Tie/Loss		9/9/2	9/9/2	7/10/3	4/6/4	9/9/2	11/8/1	**12/8/0**

especially when the accuracies of 3, 5 and 7 nearest neighbor graphs are worse than that of the baseline 1NN method.

Table 4 further gives the results of the compared methods and our proposed method on 30 labeled instances. As can be seen, similar to the situations in Table 3, our proposal method effectively improves the safeness of GSSL methods, in addition achieves a highly competitive accuracy in comparison to the state-of-the-art GSSL methods.

Previous discussions are all based on the use of Euclidean distance.To further study the influence of the distance metric, Figure 1 shows the accuracy improvement results of the compared methods against baseline 1NN method with the use of the Manhattan and Cosine distance, respectively. Similarly, unlike the compared methods that may seriously decrease the accuracy in many cases, our LEAD method can often robustly improve the accuracy (the paired t-tests at 95% significance level shows that LEAD will not significantly decrease the accuracy in all the cases).

4.3 Ability in Judging the Graph Quality

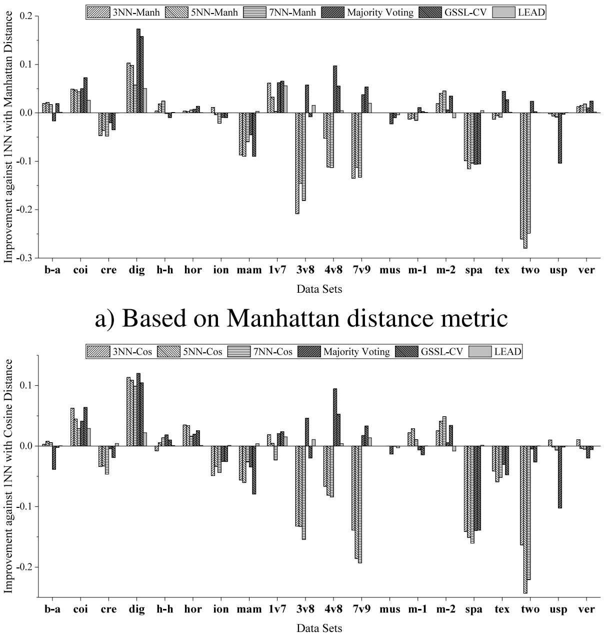
We study the ability of our proposed method in judging the quality of the graphs. Specifically, let w_1^l, w_2^l, w_3^l and w_1^s, w_2^s, w_3^s denote the largest 3 and the smallest 3 coefficients of the learned $\mathbf{w}$ in our proposed method. Suppose acc^l, acc^s and acc^{ave} are the average accuracies of the graphs corresponding to $\{w_1^l, w_2^l, w_3^l\}$, $\{w_1^s, w_2^s, w_3^s\}$ and all the coefficients in $\mathbf{w}$, respectively. Figure 2 shows the average result of $acc^l - acc^{ave}$ and $acc^s - acc^{ave}$ on 20 random splits for all

Table 4: Accuracy (mean ± std) for the compared methods and our LEAD method with 30 labeled instances.

Method	1NN	Harmonic			Majority Voting	GSSL-CV	LEAD
		3NN Graph	5NN Graph	7NN Graph			
Ave. Accuracy	80.0 ± 8.7	82.0 ± 10.9	81.5 ± 10.4	80.9 ± 10.6	83.6 ± 10.9	83.9 ± 11.1	82.2 ± 9.4
Win/Tie/Loss		10/5/5	10/5/5	10/6/4	13/5/2	14/4/2	**18/2/0**

Method	1NN	LLGC			Majority Voting	GSSL-CV	LEAD
		3NN Graph	5NN Graph	7NN Graph			
Ave. Accuracy	80.0 ± 8.7	82.3 ± 12.2	81.4 ± 13.1	80.9 ± 13.4	82.6 ± 12.1	82.5 ± 12.3	83.2 ± 10.2
Win/Tie/Loss		9/5/6	9/5/6	9/5/6	10/5/5	10/5/5	**16/4/0**

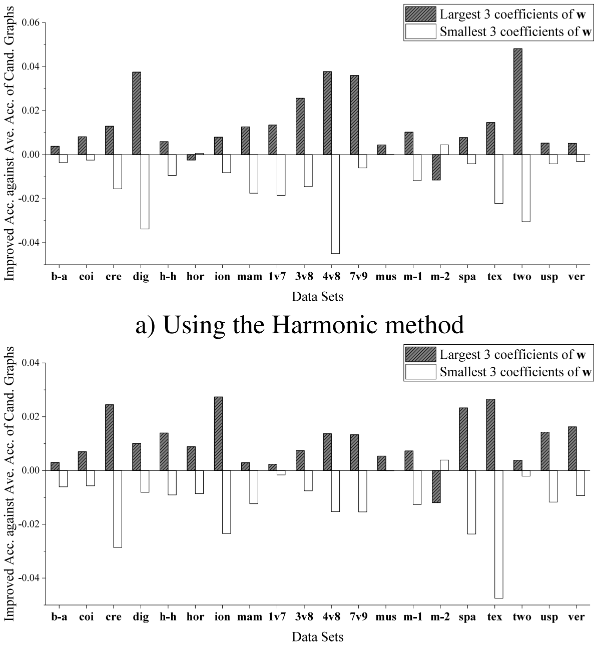
Figure 1: The performance improvement against 1NN on 10 labeled instances with different distance metrics, i.e., Manhattan distance (a) and Cosine distance (b), respectively.



a) Based on Manhattan distance metric

b) Based on Cosine distance metric

Figure 2: The improved accuracy against the average accuracy of the candidate graphs, for the graphs with the largest 3 and the smallest 3 coefficients of the learned coefficients in the large margin classifier **w**, respetively.



a) Using the Harmonic method

b) Using the LLGC method

the experimental data sets. From the figure, it can be observed that in most cases the performance of the largest 3 coefficients is much better than that of the smallest ones. This result indicates that large margin principle can indeed be helpful to judge the quality of the graphs.

5 Conclusion

In this paper we study to judge the graph quality in GSSL, a key factor for GSSL performance, and develop a safe GSSL method LEAD that does not deteriorate the performance when using unlabeled data. Our main contribution is to propose a large margin assumption for the graph quality. Specifically, when a graph owns a high quality, its prediction on unlabeled data may have a large margin separation. Intuitively, one should exploit the large margin graphs and rarely use the small margin graphs (which might be risky), and therefore reduce the chance of performance degeneration when using unlabeled data. We consequently formulate safe GSSL as the classical Semi-Supervised SVM (S3VM) optimization. Extensive experimental results demonstrate that large margin principle is helpful in judging the graph quality and improving the safeness of GSSL.

References

[Argyriou *et al.*, 2005] A. Argyriou, M. Herbster, and M. Pontil. Combining graph laplacians for semi-supervised learning. In *Advances in Neural Information Processing Systems 18*, pages 67–74. MIT Press, Cambridge, MA, 2005.

[Balcan and Blum, 2010] M. F. Balcan and A. Blum. A discriminative model for semi-supervised learning. *Journal of the ACM*, 57(3), 2010.

[Belkin and Niyogi, 2004] M. Belkin and P. Niyogi. Semi-Supervised Learning on Riemannian Manifolds. *Machine Learning*, 56:209–239, 2004.

[Belkin and Niyogi, 2008] M. Belkin and P. Niyogi. Towards a theoretical foundation for laplacian-based manifold methods. *Journal of Computer and System Sciences*, 74(8):1289–1308, 2008.

[Belkin *et al.*, 2006] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7:2399–2434, 2006.

[Ben-David *et al.*, 2008] S. Ben-David, T. Lu, and D. Pál. Does unlabeled data provably help? worst-case analysis of the sample complexity of semi-supervised learning. In *Proceedings of the 21st Annual Conference on Learning Theory*, pages 33–44, Helsinki, Finland, 2008.

[Blum and Chawla, 2001] A. Blum and S. Chawla. Learning from labeled and unlabeled data using graph mincuts. In *Proceedings of the 8th International Conference on Machine Learning*, pages 19–26, Williamstown, MA, 2001.

[Carreira-Perpiñán and Zemel, 2005] M. Á. Carreira-Perpiñán and R. S. Zemel. Proximity graphs for clustering and manifold learning. In *Advances in Neural Information Processing Systems 17*, pages 225–232. MIT Press, Cambridge, MA, 2005.

[Chapelle *et al.*, 2008] O. Chapelle, V. Sindhwani, and S. S. Keerthi. Optimization techniques for semi-supervised support vector machines. *Journal of Machine Learning Research*, 9:203–233, 2008.

[Cozman *et al.*, 2002] F. G. Cozman, I. Cohen, and M. Cirelo. Unlabeled data can degrade classification performance of generative classifiers. In *Proceedings of the 15th International Florida Artificial Intelligence Research Society Conference*, pages 327–331, Pensacola Beach, FL, 2002.

[Fan *et al.*, 2008] R.E. Fan, K.W. Chang, C.J. Hsieh, X.R. Wang, and C.J. Lin. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

[Jebara *et al.*, 2009] T. Jebara, J. Wang, and S. F. Chang. Graph construction and b-matching for semi-supervised learning. In *Proceedings of the 26th International Conference on Machine Learning*, pages 441–448, Montreal, Canada, 2009.

[Joachims, 1999] T. Joachims. Transductive inference for text classification using support vector machines. In *Proceedings of the 16th International Conference on Machine Learning*, pages 200–209, Bled, Slovenia, 1999.

[Joachims, 2003] T. Joachims. Transductive learning via spectral graph partitioning. In *Proceedings of the 20th International Conference on Machine Learning*, pages 290–297, Washington, DC, 2003.

[Karlen *et al.*, 2008] M. Karlen, J. Weston, A. Erkan, and R. Collobert. Large Scale Manifold Transduction. In *Proceedings of the 25th International Conference on Machine Learning*, pages 775–782, Helsinki, Finland, 2008.

[Li and Zhou, 2005] M. Li and Z.-H. Zhou. SETRED: Self-training with editing. In *Proceedings of the 9th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 611–621, Hanoi, Vietnam, 2005.

[Li and Zhou, 2015] Y.-F. Li and Z.-H. Zhou. Towards making unlabeled data never hurt. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(1):175–188, 2015.

[Li *et al.*, 2013] Y.-F. Li, J. T. Kwok, I. Tsang, and Z.-H. Zhou. Convex and scalable weakly label SVMs. *Journal of Machine Learning Research*, 14:2151–2188, 2013.

[Li *et al.*, 2016] Y.-F. Li, J. T. Kwok, and Z.-H. Zhou. Towards safe semi-supervised learning for multivariate performance measures. In *Proceedings of 30th AAAI Conference on Artificial Intelligence*, Phoenix, AZ, 2016.

[Liu *et al.*, 2012] W. Liu, J. Wang, and S.F. Chang. Robust and scalable graph-based semisupervised learning. *Proceedings of the IEEE*, 100(9):2624–2638, 2012.

[Ng *et al.*, 2001] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems 14*, pages 849–856. MIT Press, Cambridge, MA, 2001.

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

[Wang and Zhang, 2008] F. Wang and C. Zhang. Label propagation through linear neighborhoods. *IEEE Transactions on Knowledge and Data Engineering*, 20(1):55–67, 2008.

[Zhang *et al.*, 2007] K. Zhang, I. W. Tsang, and J. T. Kwok. Maximum margin clustering made practical. In *Proceedings of the 24th International Conference on Machine Learning*, pages 1119–1126, Corvallis, OR, 2007.

[Zhou *et al.*, 2004] D. Zhou, O. Bousquet, T. Navin Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems 16*, pages 595–602. MIT Press, Cambridge, MA, 2004.

[Zhou, 2012] Z.-H. Zhou. *Ensemble Methods: Foundations and Algorithms*. Chapman & Hall, Boca Raton: FL, 2012.

[Zhu *et al.*, 2003] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In *Proceedings of the 20th International Conference on Machine learning*, pages 912–919, Washington, DC, 2003.

[Zhu *et al.*, 2005] X. Zhu, J. Kandola, Z. Ghahramani, and J. Lafferty. Nonparametric transforms of graph kernels for semi-supervised learning. In *Advances in Neural Information Processing Systems 17*, pages 1641–1648. MIT Press, Cambridge, MA, 2005.

[Zhu, 2007] X. Zhu. Semi-supervised learning literature survey. Technical report, University of Wisconsin-Madison, 2007.