Semi-supervised vs. Cross-domain Graphs for Sentiment Analysis

Natalia Ponomareva
University of Wolverhampton, UK
nata.ponomareva@wlv.ac.uk

Mike Thelwall
University of Wolverhampton, UK
m.thelwall@wlv.ac.uk

Abstract

The lack of labeled data always poses challenges for tasks where machine learning is involved. Semi-supervised and cross-domain approaches represent the most common ways to overcome this difficulty. Graph-based algorithms have been widely studied during the last decade and have proved to be very effective at solving the data limitation problem. This paper explores one of the most popular state-of-the-art graph-based algorithms - label propagation, together with its modifications previously applied to sentiment classification. We study the impact of modified graph structure and parameter variations and compare the performance of graph-based algorithms in cross-domain and semi-supervised settings. The results provide a strategy for selecting the most favourable algorithm and learning paradigm on the basis of the available labeled and unlabeled data.

1 Introduction

Sentiment classification is an active area of research concerned with the automatic identification of sentiment strength or valence in texts. Being a special case of topic classification, it can benefit from all well-known classification algorithms. However, as sentiment classification relies on sentiment markers rather than frequent topic words, it potentially needs more data for satisfactory performance. When a limited amount of labeled data is available, cross-domain learning (CDL) or semi-supervised learning (SSL) approaches are commonly used. CDL techniques endeavour to exploit existing annotated data from a different domain (i.e. different topic and/or genre) but their success largely depends on how similar the source and target domains are. In contrast, SSL relies on a small amount of labeled data from the same domain which requires additional annotations.

Graph-based (GB) learning has been intensively studied in the last ten years (Zhu et al., 2003; Joachims, 2003; Talukdar and Crammer, 2009; Subramanya and Bilmes, 2011) and applied to many NLP tasks. In particular, in the field of sentiment analysis GB models have been employed for sentiment classification (Pang and Lee, 2004; Goldberg and Zhu, 2006; Wu et al., 2009), automatic building of sentiment lexicons (Hassan and Radev, 2010; Xu et al., 2010), cross-lingual sentiment analysis (Scheible et al., 2010) and social media analysis (Speriosu et al., 2011). The popularity of GB algorithms is not accidental: they not only represent a competitive alternative to other SSL techniques (co-training, transductive SVM, etc.) but also feature a number of remarkable properties, including scalability (Bilmes and Subramanya, 2011) and easy extension to multi-class classification (Zhu et al., 2003). GB algorithms exploit the ability of the data to be represented as a weighted graph where instances are vertices and edges reflect similarities between instances. Higher edge weights correspond to more similar instances and vice versa. GB approaches assume smoothness of the label function on the graph so that strongly connected nodes belong to the same class. In this paper we focus on the adaptation of a widely used Label Propagation (LP) algorithm (Zhu and Ghahramani, 2002) to semi-supervised and cross-domain sentiment classification.

The goal of our research is two-fold. First, we attempt to formalise and unify the research on GB approaches in the field of sentiment analysis. In particular, we conduct a comparison between LP and its variants and study the impact of different graph structures and parameter values on algorithm performance. We also demonstrate that GB-SSL and GB-CDL accuracies are competitive or superior to the accuracies shown by other SSL and CDL techniques.

Second, most research on sentiment classification which deals with limited or no in-domain labeled data usually favours one learning paradigm

Proceedings of Recent Advances in Natural Language Processing, pages 571–578,
Hissar, Bulgaria, 7-13 September 2013.

- SSL or CDL. However, in real life situations out-of-domain labeled data is often available, and therefore focusing only on SSL means overlooking the potential of already existing resources. At the same time, relying only on out-of-domain data might be risky as CDL accuracy largely depends on the properties of in-domain and out-of-domain data sets, e.g., domain similarity and complexity (Ponomareva and Thelwall, 2012a; Ponomareva and Thelwall, 2012b). Thus, it is important to investigate what data properties determine the choice of either CDL or SSL and what amount of in-domain labeled data is needed to outperform CDL accuracy. In light of this, the second objective of the paper is to develop a strategy for selecting the most appropriate learning paradigm under limited data conditions.

The paper is organised as follows. Section 2 presents the LP algorithm and its variants, some of which have been recently proposed for the sentiment classification task. Section 3 describes our approach to building the sentiment graph. Section 4 contains an extensive comparative analysis of LP and its variants in CDL and SSL settings. Section 5 lists some works on sentiment classification and GB learning related to our research. Finally, Section 6 defines the strategy suggesting the best algorithm and learning paradigm under limited data conditions and gives directions for further research.

2 Graph-based Approaches

2.1 Label Propagation

LP was one of the first GB algorithms to be developed, introduced by Zhu and Ghahramani (2002). It represents an iterative process that at each step propagates information from labeled to unlabeled nodes until convergence, i.e. when node labels do not change from one iteration to another. LP can be seen as weighted averaging of labels in a node neighbourhood where the influence of neighbours is defined by edge weights. In case of sentiment classification, the nodes are documents and the edge weights indicate the closeness of document ratings.

Let us introduce a formalism for a description of the algorithm. Let $G = (V, E)$ be an undirected graph with n vertices $V = \{x_1, ..., x_n\}$ connected through edges $E = \{(x_i, x_j)\}$. Assume that the first l nodes are labeled with $Y_l = \{y_1, ..., y_l\}$ while the remaining u nodes are unlabeled. Clearly $l + u = n$. We consider a binary classification problem, i.e. $y_i \in \{0, 1\}$, although the algorithm can be easily extended to multi-class cases. The task is to assign labels $\hat{Y}_u = \{\hat{y}_{l+1}, ...\hat{y}_n\}$ to unlabeled nodes. Let $W = (w_{ij})$ be a weight matrix on E with elements corresponding to the similarity between x_i and x_j, and let $\bar{W} = (\bar{w}_{ij})$ be its normalised version:

$$\bar{w}_{ij} = \frac{w_{ij}}{\sum_j w_{ij}} \qquad (1)$$

LP is formally presented in Algorithm 1.

Algorithm 1. LP

```
1. Initialise $\hat{Y} = (y_1, ..., y_l, 0, ..., 0)$
2. Propagate $\hat{Y} \leftarrow \bar{W}\hat{Y}$
3. Clamp the labeled data: $\hat{Y}_l \leftarrow Y_l$
4. Repeat from 2 until convergence
```

Bengio et al. (2006) demonstrated that LP is equivalent to minimising a quadratic cost function:

$$C(\hat{Y}) = \sum_{ij} w_{ij}(\hat{y}_i - \hat{y}_j)^2 \rightarrow min \qquad (2)$$

Zhu et al. (2003) showed that if we consider a continuous label space $\hat{y} \in \mathbf{R}$ instead of the discrete there exists a harmonic function delivering a closed form solution to the optimisation problem: Let us split the normalised weight matrix $\bar{W}$ into four sub-matrices:

$$\bar{W} = \begin{pmatrix} \bar{W}_{ll} & \bar{W}_{lu} \\ \bar{W}_{ul} & \bar{W}_{uu} \end{pmatrix} \qquad (3)$$

The harmonic solution of (2) can be given by:

$$\hat{Y}_u = (I - \bar{W}_{uu})^{-1}\bar{W}_{ul}Y_l \qquad (4)$$

Zhu et al. (2003) pointed out that if classes are not well-separated then the final distribution of classes can be highly skewed. To avoid unbalanced classification they adopt the class mass normalisation (CMN) procedure which scales the output values on the basis of the class priors. Let q be the desirable proportion for the classes and let $\sum_i \hat{y}_i$ and $\sum_i (1 - \hat{y}_i)$ be the masses of classes 1 and 0 respectively. The decision rule for $\hat{y}_i$ to belong to the class 1 can then be represented as:

$$q\frac{\hat{y}_i}{\sum_i \hat{y}_i} > (1 - q)\frac{1 - \hat{y}_i}{\sum_i (1 - \hat{y}_i)} \qquad (5)$$

2.2 Modifications to the LP algorithm

The graph structure used in LP does not differentiate between labeled and unlabeled neighbours. However, in some cases it might be beneficial to give them different impacts. For example, in SSL

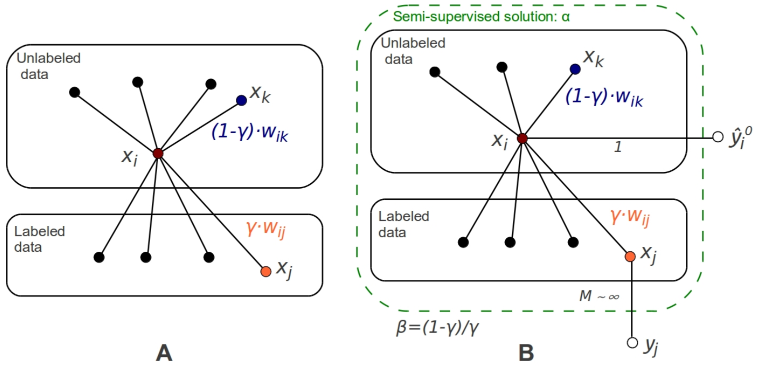

Figure 1: Modified graph structures for the LP algorithm.
A Different impact of labeled and unlabeled nodes; **B** incorporation of predictions by external classifiers

it is natural to rely more on labeled data whose labels are identified with a high level of confidence. In contrast, for CDL highly reliable labels do not help much when source and target data are very different and it might be better to prioritise unlabeled examples. Let us introduce a coefficient γ with $\gamma \in (0, 1)$ responsible for the proportion of influence between labeled and unlabeled data, so that $\gamma < 0.5$ gives preference to unlabeled and $\gamma > 0.5$ to labeled examples. This modification (LP_γ) leads to the redistribution of the weight function on graph edges (Figure 1A).

An approach very similar to LP_γ has been proposed by Wu et al. (2009) for cross-domain sentiment classification. The suggested method has two main differences from LP_γ. First, the weight matrices W_{uu} and W_{ul} are normalised separately instead of using the same scaling factor for labeled and unlabeled data. This difference has no effect as long as the scaling factors for both matrices are similar. However, this might not be the case for cross-domain graphs. Indeed, if source and target domains are very different so that out-of-domain neighbours are much farther away than in-domain neighbours, the scaling factors can have different orders of magnitude. Second, the updated values of unlabeled nodes are normalised after each iteration using the CMN procedure which fixes data skewing. As we will see in Section 4, these differences lead to a large performance increase in the results of GB-CDL. We formalise the method of Wu et al. (2009) (further called LP_γ^n, where "n" states for normalisation) in Algorithm 2.

Algorithm 2. LP_γ^n

```
1. Normalise separately $W_{uu}$ and $W_{ul}$
2. Initialise $Y_l$ and $Y_u$
3. Propagate $\hat{Y}_u \leftarrow (1-\gamma)\bar{W}_{uu}\hat{Y}_u + \gamma\bar{W}_{ul}\hat{Y}_l$
4. Normalise $\hat{Y}_u$ with $CMN$
5. Repeat from 3 until convergence
```

We can further improve the graph structure in Figure 1A by incorporating external classifiers for unlabeled examples. This was implemented by Goldberg and Zhu (2006) in an application for semi-supervised multi-class sentiment classification (Figure 1B). In this modification, each labeled and unlabeled vertex is connected to a dongle node which is a labeled node with either the true value y_i or prediction $\hat{y}_i^0$ given by an external classifier. This LP variant (called $LP_{\alpha\beta}$) is able to take advantage of different sources of information. It relies on two main parameters, α and β. β is an analogue of γ in LP_γ, $\beta = \frac{1-\gamma}{\gamma}$. Parameter α controls the weight of the GB solution compared to the initial predictions. Specifically, α close to 0 gives more importance to the initial predictions whilst high values of α prioritise the GB solution. For further details about the implementation of $LP_{\alpha\beta}$ the reader is invited to refer to Goldberg and Zhu (2006).

3 Sentiment Graph Construction

Construction of a good graph with an adequate approximation to similarity between data instances is key for the successful performance of GB algorithms (Zhu, 2005). Sentiment classification requires a similarity metric which assigns values to a pair of documents on the basis of their sentiments, so that documents with the same sentiment obtain high similarity scores and documents of opposite sentiments obtain low scores. This implies that vector representation of the data must contain sentiment markers rather than topic words. Previous research suggests several possible vector representations for documents. Pang and Lee

(2005) proposed PSP-based similarity and document representation as (PSP, 1-PSP), where PSP is the percentage of positive sentences in a document. They used an additional classifier for learning sentence polarity that was trained on external data with user-provided scores. As a result, the PSP values gave a high correlation with document ratings. Goldberg and Zhu (2006) also used in-domain labeled data to approximate sentiment similarity for semi-supervised sentiment classification. In particular, they constructed a vector representation based on document words. The weight of words was calculated using their mutual information with positive and negative classes from the external data set. The main disadvantage of both of the above approaches is that they require labeled in-domain data. The principal purpose of our research is to develop a learning strategy when a limited amount of labeled data is available.

Research on sentiment analysis suggests that certain parts of speech, e.g., adjectives, verbs and adverbs, are good sentiment markers (Pang and Lee, 2008). Thus, we represent a document as a vector of unigrams and bigrams and filter out those that do not contain above parts of speech. As nouns can also convey sentiments, we extend our feature space by the nouns listed in the SO-CAL-dictionaries (Taboada et al., 2010). The similarity between two documents is measured by the cosine similarity between their vector representations.

Another issue that needs to be tackled when constructing a graph is connectivity. Graphs can be fully connected or sparse. The former representation, besides its high computational cost, usually performs worse than sparse models (Zhu, 2005). The most common way to construct sparse graphs is to introduce either a threshold for the number of nearest neighbours k (kNN graphs) or a maximum proximity radius ϵ which removes edges with weights less than ϵ (ϵNN graphs). According to Zhu (2005) all kNN graphs tend to perform well empirically. Following this observation as well as our own experiments with ϵNN graphs, which showed no significant difference in the performance, we choose the kNN graph structure for all our models. Moreover, unlike much previous work we distinguish labeled and unlabeled nodes in a way that we connect each unlabeled node with k_l labeled and k_u unlabeled neighbours, where k_l and k_u can be different. This modification is justified empirically (see Section 4).

4 Experiments

4.1 Data and Experimental Objectives

In our experiments we use the popular multi-domain data set (Blitzer et al., 2007) comprising Amazon product reviews on 4 topics: books (BO), electronics (EL), kitchen appliances (KI) and DVDs (DV). Reviews are rated using a binary scale, 1-2 star reviews are considered as negative and 4-5 star reviews as positive. The data within each domain are balanced: they contain 1000 positive and 1000 negative reviews.

We experiment with these data in two different settings: CDL and SSL. In CDL settings we assume that there are 2 data sets: one labeled (source) and the other unlabeled (target). The task is to label the target data on the basis of the information given by the source data. In SSL settings we assume that we have a limited amount of labeled data and vast amount of unlabeled data and we aim to classify some test data belonging to the same domain. As both settings use some labeled data all algorithms described in Section 2 can be easily applied to these tasks. In our experiments we examine the performance of LP and its 3 variants: LP_γ, LP_γ^n and $LP_{\alpha\beta}$. We also compute normalise values of the obtained results: $LP_\gamma + CMN$ and $LP_{\alpha\beta} + CMN$.

Our experiments aim to answer four questions:

1. Which modifications of graph structure improve the algorithm performance and which algorithm delivers the best results?
2. Can GB-CDL approach fully-supervised in-domain accuracy levels?
3. How much labeled data does GB-SSL approach need to achieve the performance of fully-supervised classification?
4. Do GB algorithms provide results comparable with other state-of-the-art CDL and SSL techniques?

4.2 Cross-domain Learning

Previous studies on CDL agreed that properties of source and target data determine the results given by CDL algorithms. Asch and Daelemans (2010) and Plank and van Noord (2011) focused on the similarity between source and target data sets as the main factor influencing the CDL accuracy loss. Our previous research (Ponomareva and Thelwall, 2012a) brought forward another data

source-	baseline	LP	LP_{γ}	LP_{γ}	$LP_{\alpha\beta}$	$LP_{\alpha\beta}$	LP_{γ}^{n}	SCL	SFA	in-domain
target				$+CMN$		$+CMN$				accuracy
EL-BO	65.5	68.5	69.0	70.3	69.2	70.5	72.3	75.4	75.7	78.6
KI-BO	64.7	68.8	69.2	69.9	69.2	71.5	73.9	68.6	74.8	78.6
DV-BO	74.4	**78.5**	**79.9**	**80.4**	**80.3**	**81.1**	**80.9**	**79.7**	**77.5**	78.6
BO-EL	70.0	69.8	70.0	73.8	73.2	74.1	77.4	77.5	72.5	81.2
KI-EL	79.7	**83.3**	**83.0**	**83.8**	**83.4**	**83.7**	**82.3**	**86.8**	**85.1**	81.2
DV-EL	67.2	74.1	74.3	74.9	74.1	76.2	78.9	74.1	76.7	81.2
BO-KI	69.5	73.0	74.8	76.3	76.1	77.0	81.4	78.9	78.8	82.9
EL-KI	81.6	**82.3**	**83.8**	**84.7**	**85.0**	**86.1**	**84.1**	**85.9**	**86.8**	82.9
DV-KI	70.2	75.3	75.5	76.2	77.3	77.6	80.9	81.4	80.8	82.9
BO-DV	76.5	78.0	77.0	**79.5**	**78.8**	**80.8**	**78.6**	75.8	**81.4**	79.6
EL-DV	71.3	71.3	72.3	73.0	74.7	74.6	74.6	76.2	77.2	79.6
KI-DV	70.1	71.0	72.5	72.8	72.8	75.2	76.3	76.9	77.0	79.6
average	71.7	74.5	75.1	76.3	76.2	77.3	78.4	78.1	78.7	80.6

Table 1: Accuracies (%) of GB algorithms in CDL settings (accuracies within the 95% confidence interval of the in-domain accuracies are highlighted).

property called domain complexity which we defined as vocabulary richness and approximated by the percentage of rare words. We showed a non-symmetry of the accuracy drop, specifically, that it tends to be higher when source data are more complex. We also demonstrated:

a) similarity between BO and DV on the one hand, and between EL and KI on the other hand;

b) a higher level of complexity of BO and DV with respect to EL and KI.

We exploit these findings to analyse the GB-CDL results. The four data sets give 12 combinations of source-target pairs and, therefore, 12 series of experiments. Our experimental setup includes 2 stages: parameter tuning and algorithm testing. We randomly extract 400 examples from the target data and use them as the development data set for tuning the parameters α, $\beta(\gamma)$, k_u and k_l. The parameter search is run over the following ranges: $k_u \in \{5, 10, 20, 50, 100\}$, $k_l \in \{5, 20, 50, 100, 200, 400\}$, $\beta \in \{0.2, 0.5, 1, 2, 5\}$, $\alpha \in \{1, 2, 5, 10, 50, 100, 200\}$. $LP_{\alpha\beta}$ also requires initial approximations for the labels which we obtain by applying a linear-kernel SVM[1] classifier trained on the source data. The best set of parameter values is established on the basis of the highest average accuracy over all source-target pairs.

Analysing the optimal set of parameter values we observe an overall agreement between the algorithms on the choice of $\beta(\gamma)$ with a preference for high values of $\beta = 5$ and correspondingly low values of $\gamma = 0.2$. This implies that GB algorithms in CDL settings heavily rely on labels provided by in-domain neighbours. Optimal value of α is obtained to be 200 as low values of α ($\alpha < 10$) keep output labels very close to the supervised solution. In most cases, the best results are achieved for low $k_u \leq 10$ and relatively high $k_l = 100$, which confirms the importance of separate parameters for the number of labeled and unlabeled neighbours. The obtained optimal parameter values are used in algorithms' testing conducted over the remaining 1600 examples from the target data.

GB-CDL accuracies are presented in Table 1. The baseline stands for the performance of a linear-kernel SVM classifier trained on the source data. The in-domain accuracies computed on the target data with 5-fold cross-validation give an estimation of the CDL performance upper bound. All LP variants improve the LP results, although the effect of some parameters is rather modest, e.g. γ. Incorporating external classifiers leads to an accuracy gain of more than 1% on average which is consistent over the domain pairs. The CMN procedure also brings a considerable contribution with overall accuracy increase of 1%. The highest results are achieved by LP_{γ}^{n} which outperforms $LP_{\gamma} + CMN$ by 2.5%.

All GB algorithms show a significant improvement over the baseline. Moreover, the accuracy gain given by the best two methods LP_{γ}^{n} and $LP_{\alpha\beta} + CMN$ reaches 5-6% on all domain pairs.

[1] We used the LIBSVM library (Chang and Lin, 2011).

GB-CDL demonstrates excellent results for pairs with similar source and target (DV-BO, BO-DV, KI-EL and EL-KI) outperforming in-domain supervised classification. At the same time, GB accuracies are rather low for pairs with large discrepancies between source and target data. In this respect, LP_γ^n is promising as it can "fix" the domain discrepancies for some source-target pairs: BO-EL, DV-EL, BO-KI and DV-KI. Keeping in mind that EL and KI have lower values of lexical richness than BO and DV, we can presume that LP_γ^n works better when the target domain is simple. This could be due to the fact that for simple domains the weight function better approximates the actual similarities between documents, but further research is necessary before such a conclusion can be drawn with high confidence.

GB algorithms demonstrate competitive performance with respect to other state-of-the-art approaches, namely SCL (Blitzer et al., 2007) and SFA (Pan et al., 2010). Indeed, Table 1 shows that the difference between average accuracies of SCL, SFA and the two best GB algorithms are not statistically significant. However, the GB approach is more beneficial for multi-class classification as its adaptation to this task is straightforward.

4.3 Semi-supervised Learning

SSL experiments are carried out separately for each domain. We randomly divide our data into 5 folds where one is used for parameter tuning and 4 for testing the algorithms in the cross-validation setup. Thus, in every experiment, 400 examples are used for testing/tuning and the remaining 1600 instances are split into labeled and unlabeled sets. We gradually increase the amount of labeled data from 50 to 800 to analyse the impact of the labeled data size on the algorithms' performance.

In contrast to the CDL experiments, we substitute k_l by the proportion of labeled neighbours Δ_l with respect to the labeled data size. We find this parameter more natural for variable sizes of labeled data. The best value for Δ_l is searched for in the range $\{0.05, 0.1, 0.2, 0.3, 0.4, 0.5\}$. The search for remaining parameters is run in the same ranges as for CDL and the optimal set is established on the basis of the highest average accuracy over all domains and labeled data sizes. Optimal value for β is found to be quite low: $\beta = 0.5$ (γ = 0.7) which is consistent with our expectations of the algorithms' preference for more reliable labeled data from the same domain. All algorithms agree on low values of k_u and Δ_l, showing best results for $k_u = 5$ and $\Delta_l = 0.1$ or 0.2.

GB-SSL accuracies are presented in Table 2[2]. The baseline corresponds to the accuracy given by a linear-kernel SVM classifier trained on the same portion of labeled data. We observe that GB-SSL algorithms outperform the in-domain results with 600-700 labeled examples. Moreover, relatively high accuracies (within the 95% confidence interval of the in-domain accuracies) can be achieved with only 500 labeled examples.

We also compare GB-SSL with two state-of-the-art SSL approaches tested on the same data (Dasgupta and Ng, 2009; Li et al., 2010) (Table 2). The method of Dasgupta and Ng (2009) combines spectral clustering with active learning. The authors report the accuracy for 100 and 500 labeled examples selected by active learning. The accuracies shown by $LP_{\alpha\beta} + CMN$ are significantly higher than the accuracies obtained by their method with an average difference of approximately 4% for both sizes of labeled data. Li et al. (2010) adopt a co-training approach which deploys classifiers trained on personal and impersonal view data sets. Although the co-training achieves very high accuracies for the KI domain it gives considerably worse results for the domains of BO and DV. Averaging accuracies across domains gives 71.4% for $LP_{\alpha\beta} + CMN$ vs. 64.5% for the co-training when 100 labeled examples are used and 77.2% vs. 74.7% for 300 examples. Moreover, unlike the proposed co-training approach the GB algorithms are much more robust delivering equally good results across all data sets.

5 Related Work

There are several fields related to our research. GB-SSL has received extensive attention from the research community (Zhu et al., 2003; Joachims, 2003; Talukdar and Crammer, 2009; Subramanya and Bilmes, 2011). Two of the most recent methods proposed in this field are Modified Adsorption (MAD) and Measure Propagation (MP), which present some advantages over LP. However, preliminary experiments we performed using MAD did not lead to very promising results and more experiments are necessary. Our paper is also re-

[2] We deliberately reduced the number of algorithms reported in this paper due to space constraints and similar behaviour of some LP variants.

No. labeled data	50	100	200	300	400	500	600	700	800	in-domain accuracy
books										
SVM	60.3	65.2	71.8	71.8	73.2	74.9	76.1	76.8	76.3	78.6
$LP_\gamma + CMN$	68.0	71.1	72.7	75.5	**77.6**	**78.5**	**79.3**	**80.2**	**81.1**	
LP_γ^n	65.5	69.9	73.1	76.6	**78.0**	**78.7**	**80.0**	**80.1**	**79.7**	
$LP_{\alpha\beta} + CMN$	66.5	70.8	73.1	75.5	75.4	**78.2**	**79.3**	**79.9**	**80.1**	
Dasgupta and Ng (2009)	–	62.1	–	–	–	73.5	–	–	–	
Li et al. (2010)	–	60.1	73.0	71.6	–	–	–	–	–	
electronics										
SVM	57.4	66.6	72.3	73.9	75.1	76.7	77.5	78.2	79.0	81.2
$LP_\gamma + CMN$	70.6	74.2	76.7	77.9	79.2	**80.6**	**80.1**	**80.6**	**81.5**	
LP_γ^n	66.7	72.8	77.4	79.4	**79.9**	**81.0**	81.0	**81.3**	**82.0**	
$LP_{\alpha\beta} + CMN$	69.9	74.1	77.8	78.4	78.9	**80.6**	**81.6**	**81.8**	**82.8**	
Dasgupta and Ng (2009)	–	70.6	–	–	–	77.5	–	–	–	
Li et al. (2010)	–	70.0	77.0	78.2	–	–	–	–	–	
kitchen										
SVM	60.0	69.2	74.1	75.8	76.8	78.1	77.5	79.9	80.1	82.9
$LP_\gamma + CMN$	70.7	73.2	76.8	79.1	80.6	80.8	**81.8**	**82.5**	**82.2**	
LP_γ^n	68.3	71.4	76.7	80.1	81.0	**81.9**	**82.4**	**82.7**	**83.5**	
$LP_{\alpha\beta} + CMN$	71.4	74.2	76.5	79.5	80.3	**82.0**	**81.8**	**83.2**	**83.5**	
Dasgupta and Ng (2009)	–	74.1	–	–	–	78.4	–	–	–	
Li et al. (2010)	–	78.6	79.0	**83.3**	–	–	–	–	–	
DVDs										
SVM	53.8	63.4	70.6	73.9	75.0	75.9	76.0	77.8	77.1	79.6
$LP_\gamma + CMN$	65.8	67.1	71.7	74.2	76.5	78.0	**80.0**	**80.8**	**81.4**	
LP_γ^n	65.2	66.3	72.3	75.1	**78.3**	**79.2**	**80.3**	**80.6**	**80.9**	
$LP_{\alpha\beta} + CMN$	65.2	66.3	72.1	75.3	77.3	**78.4**	**80.0**	**80.4**	**80.2**	
Dasgupta and Ng (2009)	–	62.7	–	–	–	73.4	–	–	–	
Li et al. (2010)	–	49.5	63.0	65.5	–	–	–	–	–	

Table 2: Accuracies (%) of GB algorithms in SSL settings (accuracies within the 95% confidence interval are highlighted; accuracies outperforming the in-domain accuracies are underlined).

lated to work in cross-domain sentiment classification and the results we obtain are comparable to those reported by (Blitzer et al., 2007; Pan et al., 2010). The SSL methods discussed in Section 4.3 (Dasgupta and Ng, 2009; Li et al., 2010) offer an interesting alternative to GB algorithms, but their results are substantially lower.

6 Conclusions and Future Work

This paper has explored GB algorithms in CDL and SSL settings. The evaluation of the GB-CDL algorithms has shown that the best methods, $LP_{\alpha\beta} + CMN$ and LP_γ^n, consistently improve the baseline by 5-6% for all domain pairs. Therefore, if source and target domains are similar (i.e. the baseline classifier loses less than 5% accuracy when adapted from the source to target domain) GB-CDL algorithms are a competitive alternative to the fully supervised techniques. Moreover, we have shown that if the target domain has low complexity, the LP_γ^n algorithm can deliver good performance even for quite different domain pairs.

For large discrepancies between source and target data GB-SSL can help to achieve good results with a reasonably small amount of labeled data. Specifically, even 500 labeled examples are enough to ensure performance within a 95% confidence interval of the in-domain accuracy.

In the future, we plan to compare GB-SSL and GB-CDL for multi-class sentiment classification. This extension should be straightforward as GB algorithms can be easily adapted to multi-class cases. In addition, we will include in our experiments other algorithms such as MAD and MP.

References

Vincent Van Asch and Walter Daelemans. 2010. Using domain similarity for performance estimation. In *Proceedings of the Workshop on Domain Adaptation for Natural Language Processing, ACL'10*, pages 31–36.

Yoshua Bengio, Olivier Delalleau, and Nicolas Le Roux, 2006. *Semi-Supervised Learning*, chapter 11. Label Propagation and Quadratic Criterion, pages 193–216. The MIT Press.

Jeff Bilmes and Amarnag Subramanya, 2011. *Scaling up Machine Learning: Parallel and Distributed Approaches*, chapter 15. Parallel Graph-Based Semi-Supervised Learning, pages 307–330. Cambridge University Press.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of ACL '07*, pages 440–447.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Sajib Dasgupta and Vincent Ng. 2009. Mine the easy, classify the hard: A semi-supervised approach to automatic sentiment classification. In *Proceedings of ACL-AFNLP'09*, pages 701–709.

Andrew B. Goldberg and Xiaojin Zhu. 2006. Seeing stars when there aren't many stars: graph-based semi-supervised learning for sentiment categorization. In *Proceedings of TextGraphs '06*, pages 45–52.

Ahmed Hassan and Dragomir R. Radev. 2010. Identifying text polarity using random walks. In *Proceedings of ACL '10*, pages 395–403.

T. Joachims. 2003. Transductive learning via spectral graph partitioning. In *Proceedings of ICML'03*.

Shoushan Li, Chu-Ren Huang, Guodong Zhou, and Sophia Yat Mei Lee. 2010. Employing personal/impersonal views in supervised and semi-supervised sentiment classification. In *Proceedings of ACL '10*, pages 414–423.

Sinno Jialin Pan, Xiaochuan Niz, Jian-Tao Sunz, Qiang Yangy, and Zheng Chen. 2010. Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of WWW '10*.

Bo Pang and Lillian Lee. 2004. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of ACL '04*, pages 271–278.

Bo Pang and Lillian Lee. 2005. Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of ACL '05*, pages 115–124.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Barbara Plank and Gertjan van Noord. 2011. Effective measures of domain similarity for parsing. In *Proceedings of ACL '11*, pages 1566–1576.

Natalia Ponomareva and Mike Thelwall. 2012a. Bibliographies or blenders: Which resource is best for cross-domain sentiment analysis? In *Proceedings of CICLing '12*.

Natalia Ponomareva and Mike Thelwall. 2012b. Do neighbours help? an exploration of graph-based algorithms for cross-domain sentiment classification. In *Proceedings of EMNLP '12*.

Christian Scheible, Florian Laws, Lukas Michelbacher, and Hinrich Schütze. 2010. Sentiment translation through multi-edge graphs. In *Coling 2010: Posters*, pages 1104–1112.

Michael Speriosu, Nikita Sudan, Sid Upadhyay, and Jason Baldridge. 2011. Twitter polarity classication with label propagation over lexical links and the follower graph. In *Proceedings of EMNLP '11*, pages 53–63.

A. Subramanya and J. Bilmes. 2011. Semi-supervised learning with measure propagation. *Journal of Machine Learning Research*, 12:3311–3370.

M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. 2010. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

Partha Pratim Talukdar and Koby Crammer. 2009. New regularized algorithms for transductive learning. In *Proceedings of ECML PKDD 2009*.

Qiong Wu, Songbo Tan, and Xueqi Cheng. 2009. Graph ranking for sentiment transfer. In *Proceedings of ACL-IJCNLP '09 (Short Papers)*, pages 317–320.

Ge Xu, Xinfan Meng, and Houfeng Wang. 2010. Build chinese emotion lexicons using a graph-based algorithm and multiple resources. In *Proceedings of Coling '10*, pages 1209–1217.

Xiaojin Zhu and Zoubin Ghahramani. 2002. Learning from labeled and unlabeled data with label propagation. Technical report, Carnegie Mellon University.

Xiaojin Zhu, Zoubin Ghahramani, and John Lafferty. 2003. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International Conference on Machine Learning (ICML '03)*, pages 912–919.

Xiaojin Zhu. 2005. *Semi-Supervised Learning with Graphs*. Ph.D. thesis, Carnegie Mellon University.