A Multi-view Approach for Semi-Supervised Scientific Paper Classification

Víctor A. Laguna [1], **Alneu de Andrade Lopes**[1]

[1]Instituto de Ciências Matemáticas e de Computação - ICMC
Universidade de São Paulo/Campus de São Carlos – Brazil

{vlaguna,alneu}@icmc.usp.br

Abstract. In this paper we show that combining information from a citation-based network with the traditional bag of words representation in a semi-supervised framework like Cotraining significantly improves scientific paper classification accuracy. We carried out experiments showing that the examples labeled by the classifier based on the citation network representation and used to increment the training set are mostly correctly labeled. This fact contributes to improve the overall accuracy of cotraining classifiers, even when the citation-based classifier, separately, is not as accurate as the classifier based on the bag of words. The results suggest that citation network information significantly improves the performance of the classifiers, mainly when labeled instances are scarce.

1. Introduction

The increasing storage capacity of electronic devices, associated with the growth and popularization of the Web, promoted the proliferation of large and diverse collections of data in electronic format [Lyman et al. 2003]. However, in most cases, the quality of this data is not appropriate for automatic knowledge discovery tasks. For such tasks, considerable effort is required from experts in preprocessing tasks such as cleansing, structuring, and labeling the data. Although unlabeled data is usually easy to collect, this preprocessing effort is difficult, expensive, time consuming, and often impractical. In this context, semi-supervised algorithms have been proposed since they address this problem of using a small quantity of labeled data with a large quantity of unlabeled one for classification purposes.

Cotraining [Blum and Mitchell 1998] is one of the most effective semi-supervised algorithms to provide a framework for using unlabeled examples to improve classification accuracy. Its basic idea is to boost accuracy by using a small number of labeled examples in a training set iteratively augmented with the most confident predictions from two classifiers. This approach differs from the Selftraining algorithm [Nigam and Ghani 2000] by the number of representations and classifiers. While Selftraining uses only one representation and trains just one classifier, Cotraining uses two representations and trains two classifiers (one for each representation). Nevertheless, the main difficulty is to obtain two representations which satisfy the Cotraining assumptions, particularly, the conditional independence between representations.

Most document classification methods employ text representations based in a set of features formed by terms deemed relevant in document contents, known as “bag of words”. A feature-value vector is used where those features are weighted by functions

of the frequency of the corresponding terms in the document (or collection). Though many textual domains present a richer and more informative structure based on relations between documents [Bockhorst and Craven 2002], traditional classification approaches extract features from document contents only.

Particularly for scientific papers, documents can be linked together by a citation network that may be useful to infer some kind of similarity between papers that have common citations patterns. However, the citation network may not be sufficient to train a good classifier because obtaining an accurate network model may be difficult. In a real scenario, where a given scientific paper collection should be classified, is common few or none of the references cited in a document also be part of the collection. That is, if the citation network were represented in a table where the rows represent the collection of papers and the columns represent the references, a fairly sparse matrix will be produced. Hence, when it is used as training data will likely not succeed in achieving a good performance classifier.

Although a classifier trained based only on the citation network features usually is not as accurate as a bag of words based classifier, we show that combining both a network-based classifier with a content-based classifier in a Cotraining schema improves classification performance. We also provide empirical results indicating that the Cotraining framework overcomes the Selftraining algorithm in this scenario.

The remainder of the paper is organized as follows. In the next Section, we briefly introduce the background and related work. In Section 3, we present the concepts of Cotraining and Selftraining, as well as the generation process of the representation based on a citation network. In Section 4, we present experimental results on a collection of scientific papers obtained from a computer science dataset. Finally, in Section 6 we present the conclusions and discuss future work.

2. Background

Link mining is an emerging research area that focuses on the study of links between examples in mining tasks [Lu and Getoor 2003]. A lot of effort has been addressed to Link Based Classification, specially in domains when relations among their constituents can be explicitly represented. Particularly for textual applications several relational domains exist, like Web pages, connected by hyperlinks, and scientific papers, connected by citations networks.

Chakrabarti [Chakrabarti et al. 1998] proposed a Web pages categorization algorithm which combines features from linked pages, in addition with features extracted from contents of the pages to be classified. Such approach generates an extended feature vector for each page which attempts to take advantage of the information obtained from the neighbors to improve classification accuracy. Oh and colleagues [Oh et al. 2000] extended this work studying which features of the linked pages should be used for augmenting the feature vector. As a result, they have shown that neighbor's information is useful but it has to be carefully used to avoid noisy representations. Chikhi and colleagues [Chikhi et al. 2008] used a similar approach to find out groups of topics in scientific collections. They combined the features of a content-based vector with the features obtained from the neighbors in a citation network. Their approach outperforms similar techniques.

Specifically, in classifying scientific papers some researchers have successfully

employed the underlying citation network to improve classification performance. Cao and Gao proposed to refine the labels obtained by traditional statistical classifiers using the information present in the citation networks [Cao and Gao 2005]. In their approach, they update the class probability distribution, given as output by a statistical content-based classifier, using a simple linear function that measures the influence of the neighbor's classes in the document. In a later work, Zhang proposed two more complex functions (linear and probabilistic) to update labels in the citations, showing that relations contain information useful to achieve a better performance [Zhang et al. 2006]. Unlike the technique proposed in this paper, those investigations studied the use of the citation networks when sufficient labeled data is available.

Different semi-supervised algorithms were used when the number of initial labeled examples is small (see [Bennett and Demiriz 1998, Brefeld and Scheffer 2004] as examples). In this context, Cotraining was successfully employed in text classification problems [Nigam and Ghani 2000, Balcan et al. 2005], where it is common to use a feature vector splitting to obtain the two views. Furthermore, those authors studied text classification with representations that are not totally independent. However better performance is expected when two independent views are combined in the Cotraining framework.

We propose a semi-supervised approach similar to the on proposed by Loo and coleagues [Lu and Getoor 2003], who developed a logistic regression model that aims at predicting unavailable labels of Web pages by computing the posterior probability of each label for each page. Nevertheless, unlike the approach proposed here the use of the two representations is integrated in the probability calculation and not for training different classifiers.

3. Semi-supervised text classification using citation networks

In this Section, the Cotraining and Selftraining algorithms are briefly introduced, as well as the assumptions for the Cotraining framework. We also present our approach for the citation-based network representation.

3.1. The Cotraining algorithm

Cotraining is a traditional algorithm employed when multiple views of the training data are available. Its purpose is to train classifiers (one for each view) boosted by using unlabeled data. In this paper, information on the citation structure and the contents of documents are integrated in the Cotraining framework. This integration of representations aims at training classifiers that overcome the lack of labeled instances.

After generating two representations of the data, the Cotraining algorithm is applied as follows. Given a set L of labeled examples and a set U of unlabeled examples, each one represented by two views, two classifiers trained with L are induced. These classifiers label the set u which is a subset of U. Then each classifier chooses the $n + p$ most confident labeled examples from u to augment the set L. With $n + p$ examples selected for each classifier, there will be $2n + 2p$ less examples in u that should be replenished by drawing $2n + 2p$ random examples from U. This procedure iterates k times, finally obtaining two semi-supervised trained classifiers.

According to Blum e Mitchell, the values for n and p must consider the underlying data distribution and should be defined by the user. Blum e Mitchell justify the existence of the set u to force the classifiers select examples respecting the underlying class distribution (i.e. u always has the same number of n negative examples and p positive examples). They also define a third "combined" classifier, which computes the probability of an example to belong to a given class by multiplying the probabilities calculated by each trained classifier.

The Cotraining algorithm considers the following assumptions about the representations: they should be redundant for prediction (i.e. they should agree in the predicted label in most cases); and they should be conditionally independent given the class (i.e. the information represented by each view should not be correlated). In scientific corpus classification, the representation based on the citation network can be seen as independent of the representation based on the content of the papers. For example, if the texts were translated, the representation based in the contents would change at all whereas the network representation would remain without alterations.

3.2. Modeling a feature-vector representation based on citation networks

Obtaining a representation of a textual data set independent from that given by its bag of words does not seem to be a trivial task. In this work, a citation-based network was employed to represent citation links among documents in addition to a bag of words representation. Specifically, such network was considered because it is an explicit network in this domain.

The wide variety of bibliographical reference formats, common typos, abbreviations, translation or transliteration of contents in the references turn the citation-based network construction into a complex task. Furthermore, for the citation-based network to be considered informative enough for mining tasks, it should present recurrent patterns between similar objects. In other words, for text classification, a citation network will be useful to improve classification performance if scientific papers of a same class have some citations in common.

The citation-based network is represented by a feature vector based in the adjacency matrix of the unweighted directed graph that maps the citations among instances. This feature vector was generated by modifying the adjacency matrix so that each row represents an instance in the training corpus, and each feature represents all the papers in the citation network. In this way, let p_i be the i^{th} paper in the corpus, and let $cout_i$ be the number of cited references in p_i and cin_i the number of citations that p_i receives from other papers. Let t be the total number of features (both papers present in corpus and just cited) in this model. The feature vector of pi will have t features, where $cin_i + cout_i$ features will have a value of 1 and $t - (cin_i + cout_i)$ will have a value of 0. In other words, each feature f_j will have value of 1 if p_i cites or is cited by f_j. An additional column was inserted at the end of this matrix with the class value. This approach is illustrated in Figure 1, which shows two groups of citations with different class values.

3.3. The Selftraining algorithm

Selftraining is an incremental algorithm that uses only one view of the training data. Its goal is to boost a weak classifier by initially classifying the unlabeled instances,

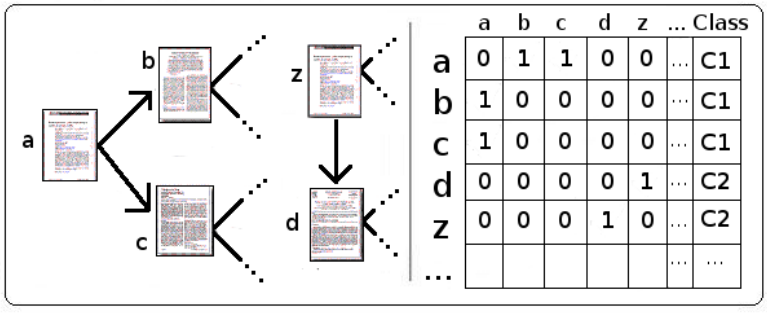

Figure 1. Feature vector representation of the citation network

and, iteratively, increment the labeled set with the most confident classified examples [Nigam and Ghani 2000].

In the experiments, Selftraining was employed just with the representation based in the contents of the documents (bag of words), since, isolated, it achieves better results than using the citation network representation. This approach was compared to the proposed in this work (using Cotraining), to evaluate our hypothesis that the network vision helps to improve classification accuracy when combined in a Cotraining model.

4. Experiments

We consider a realistic scenario where a scientific paper collection must be classified and just a few examples labeled by experts are available. A Cotraining framework was evaluated with one vision based on the citation network and the other based on the bag of words.

4.1. Combining representations with Cotraining

In this section, results for the semi-supervised scientific paper classification are presented. In order to test the multi-view semi-supervised approach, we have used a corpus referred to as CBR-ILP-IR-SON, comprised by collected scientific papers in four different subjects: Case-Based Reasoning (CBR), Inductive Logic Programming (ILP), Information Retrieval (IR), and Sonification (SON). It consists of 675 scientific papers obtained from LNAI publications (CBR and ILP datasets)[1] and retrieved from the Web (IR and SON datasets) .

Due to the binary nature of the Cotraining framework, for each class classifiers were trained considering examples from a given class as positive and the rest as negative examples. As one observes in Table 1, data sets with different class distributions were generated from the CBR-ILP-IR-SON corpus.

The Cotraining algorithm needs a ranking of probabilities of each example in order to choose the most confident labeled examples. Classifiers were thus trained separately for each view using the Naïve Bayes algorithm. The Naïve Bayes algorithm has been successfully used in Text Classification tasks [Domingos and Pazzani 1997, Mccallum and Nigam 1998] and Cotraining semi-

[1] http://www.springer.com/series/1244

supervised tasks [Blum and Mitchell 1998, Nigam and Ghani 2000]. Hence, it seems suitable to analyze and compare the contribution of the network-based classifier.

Table 1. Datasets detail

Name	%possitive class	%negative class
CBR-NotCBR	48%	52%
IR-NotIR	25%	75%
ILP-NotILP	19%	81%
SON-NotSON	14%	86%

To compare the multi-view approach (Cotraining) with the one-view approach (Selftraining), we trained a Naïve Bayes classifier with 3% labeled trained examples, and evaluated its performance using 10-fold cross validation. The average on 10 runs was obtained on 15 iteration for each run and parameters p and n proportional to the distribution of positive and negative examples, respectively. Hence, for the CBR-NotCBR dataset these parameters were set to $p = 1$ and $n = 5$; $p = 2$ and $n = 7$ for the IR-NotIR; $p = 2$ and $n = 8$ for the ILP-NotILP; and $p = 1$ and $n = 7$ for the SON-NotSON. Finally, the Selftraining algorithm with a bag of words based classifier was run with the same parameter settings, in order to ensure a fair comparison.

Results are summarized in Table 2. Classification accuracies in two different moments for each dataset are presented: before and after applying the Cotraining algorithm (without and with Cotraining, respectively). The classifiers based on the bag of words and on the citation network are referred to as BOW and Network, respectively. The result for the $combined$ classifier is also presented. Note (see Table 3) that for the first three datasets the Cotraining classifiers outperform the classifier obtained by the Selftraining algorithm. In the last dataset the Selftraining classifier achieves a better performance. This is because the number of positives exmples labeled initially is very small in this data set (SON-notSON), representing a quite poor part of the network (a subgraph with a few or no edges at all). This fact, added to the sparse network representation, decreases the classifier's capability of expressing an accurate network-based model. We empirically demonstrate this hypotheses by incrementing the number of initial labeled examples from 3 to 10%. With this change the Cotraining classifiers outperform the Selftraining Classifier also in this data set.

Table 2. Error rates of the classifiers trained without and with Cotraining

Name	without Cotraining		with Cotraining		
(3% training)	**BOW**	**Network**	**BOW**	**Network**	**Combined**
CBR-notCBR	9,48%(5,5)	40,64%(5,9)	1,06%(0,3)	20,25%(3,4)	**1,04%(0,2)**
IR-notIR	21,06%(2,4)	28,28%(11,65)	**15,47%(2,0)**	19,74%(0,8)	16,07%(2,1)
ILP-notILP	16,21%(1,9)	16,36%(2,5)	**1,02%(0,5)**	4,69%(0,4)	1,05%(0,5)
SON-notSON	14,24%(1,2)	18,18%(9,58)	**11,18%(2,06)**	13,69%(0,1)	11,98%(1,24)

4.2. Analyzing the contribution of the citation network representation

Although the network-based classifier may not be as accurate as the bag of words based classifier (indeed, its error rate is much higher in most cases), one can notice that performance of both classifiers improve when using the Cotraining framework. In order to

Table 3. Comparing error rates for Cotraining and Selftraining

Name	*Cotraining (Combined)*	*Selftraining (BOW)*
CBR-notCBR (3% training)	**1,04%(0,2)**	8,47%(3,0)
IR-notIR (3% training)	**16,07%(2,1)**	18,00%(1,92)
ILP-notILP (3% training)	**1,05%(0,5)**	6,12%(1,8)
SON-notSON (3% training)	11,98%(1,24)	**9,18%(3,9)**
SON-notSON (10% training)	**4,80%(0,4)**	5,77%(0,5)

evaluate the network based contribution, we computed in each loop the error rate of the network-based classifier, only for the $n + p$ examples used to augment the set of labeled examples L. Figure 2 shows the error rate of the network-based classifier in each iteration step for the CBR-NotCBr and IR-NotIR datasets (the values shown represented the average of 10 runs with 15 iterations). The other data sets show a similar behavior. Notice that the error rates in each iteration step, for these instances, are much smaller than the overall error rate. Consequently, the contribution of the network-based classifier is significant to improve the global classification accuracy.

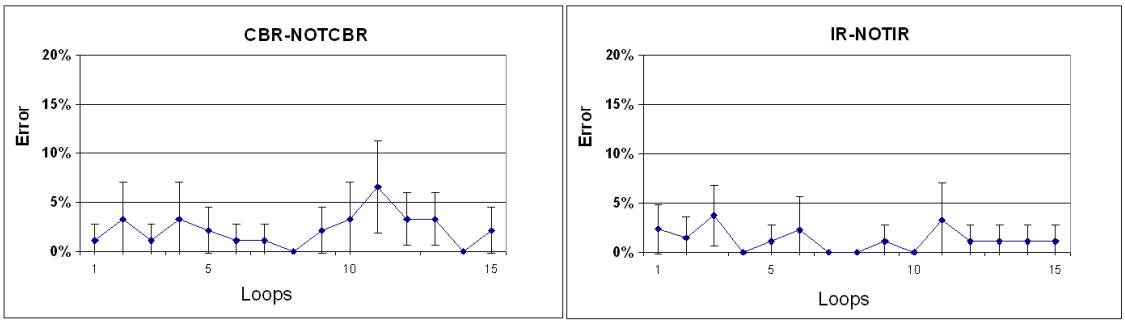

Figure 2. Error rates of the network based classifier in each iteration step.

5. Conclusions and future work

We have demonstrated empirically that using a Cotraining framework with one vision based in a bag of words and another in a citation-based network, significantly improves the classification accuracy of a scientific paper collection. The experiments suggest that the number of labeled examples used to train an initial weak classifiers should be defined considering the quality of the patterns described by the citation network (which, in this case, is the weakest classifier).

In future work, new approaches to reduce the impact of sparse network representation will be studied, as well as different kinds of networks (like co-authorship and topic networks).

6. Acknowledgment

This research was supported by CAPES.

References

Balcan, M., Blum, A., and Yang, K. (2005). Co-training and expansion: Towards bridging theory and practice. In L.K., S., Weiss, Y., and Bottou, L., editors, *in Neural Information Processing Systems*, volume 17, pages 89–96. MIT Press, Cambridge, MA.

Bennett, K. P. and Demiriz, A. (1998). Semi-supervised support vector machines. In *Advances in Neural Information Processing Systems*, pages 368–374. MIT Press.

Blum, A. and Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *COLT' 98: Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100, New York, USA. ACM.

Bockhorst, J. and Craven, M. (2002). Exploiting relations among concepts to acquire weakly labeled training data. In *ICML '02: Proceedings of the Nineteenth International Conference on Machine Learning*, pages 43–50, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Brefeld, U. and Scheffer, T. (2004). Co-em support vector learning. In *In Proceedings of the International Conference on Machine Learning*, pages 121–128.

Cao, M. D. and Gao, X. (2005). Combining contents and citations for scientific document classification. In *AI 2005: proceedings of 18th Australian Joint Conference on Artificial Intelligence*, pages 143–152, Sydney, Australia. Springer-Verlag New York.

Chakrabarti, S., Dom, B., and Indyk, P. (1998). Enhanced hypertext categorization using hyperlinks. In *SIGMOD '98: Proceedings of the 1998 ACM SIGMOD international conference on Management of data*, pages 307–318, New York, NY, USA. ACM.

Chikhi, N. F., Rothenburger, B., and Aussenac-Gilles, N. (2008). Combining link and content information for scientific topics discovery. In *ICTAI '08: Proceedings of the 2008 20th IEEE International Conference on Tools with Artificial Intelligence*, pages 211–214, Washington, DC, USA. IEEE Computer Society.

Domingos, P. and Pazzani, M. (1997). On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning*, 29:103–130.

Lu, Q. and Getoor, L. (2003). Link-based classification using labeled and unlabeled data. In *Proceedings of the ICML-2003 Workshop on The Continuum from Labeled to Unlabeled Data*, pages 496–503, Washington DC, USA.

Lyman, P., Varian, H. R., Charles, P., Good, N., Jordan, L. L., and Pal, J. (2003). How much information? Technical report, Regents of the University of California.

Mccallum, A. and Nigam, K. (1998). A comparison of event models for naive bayes text classification. In *AAAI Workshop on "Learning for Text Categorization"*, pages 41–48. AAAI Press.

Nigam, K. and Ghani, R. (2000). Understanding the behavior of co-training. In *Proceedings of KDD-2000 Workshop on Text Mining*, pages 15–17, Boston, MA, USA.

Oh, H.-J., Myaeng, S. H., and Lee, M.-H. (2000). A practical hypertext catergorization method using links and incrementally available class information. In *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 264–271, New York, NY, USA. ACM.

Zhang, M., Gao, X., Cao, M. D., and Ma, Y. (2006). Modelling citation networks for improving scientific paper classification performance. In *PRICAI 2006: proceedings of 9th Pacific Rim International Conference on Artificial Intelligence*, pages 413–422, Guilin, China. Springer-Verlag New York, Inc.