PE-PUC: A Graph Based PU-Learning Approach for Text Classification

Shuang Yu and Chunping Li

School of Software, Tsinghua University, Beijing 100084, China
yushuang@mails.tsinghua.edu.cn,
cli@tsinghua.edu.cn

Abstract. This paper presents a novel solution for the problem of building text classifier using positive documents (P) and unlabeled documents (U). Here, the unlabeled documents are mixed with positive and negative documents. This problem is also called PU-Learning. The key feature of PU-Learning is that there is no negative document for training. Recently, several approaches have been proposed for solving this problem. Most of them are based on the same idea, which builds a classifier in two steps. Each existing technique uses a different method for each step. Generally speaking, these existing approaches do not perform well when the size of P is small. In this paper, we propose a new approach aiming at improving the system when the size of P is small. This approach combines the graph-based semi-supervised learning method with the two-step method. Experiments indicate that our proposed method performs well especially when the size of P is small.

1 Introduction

Text classification is the technique of automatically assigning categories or classes to unlabeled documents. With the ever-increasing volume of text documents from various online sources, an automatic text classifier can save considerable time and human labor. Recently, a new direction of text classification problem becomes recognized, which is called PU-Learning [6] [7] [5] [8]. P represents the given labeled positive set; U represents the given unlabeled set, which is mixed with positive and negative documents. Usually, the positive set contains documents from a special topic and the negative set contains documents from diverse topics. PU-Learning is a special problem of text classification, where classifiers are built using labeled positive documents and unlabeled documents.

PU-Learning is of great use in the task of accurately labeling documents as positive and negative with respect to a special class. It is particularly useful when the user wants to find positive documents from many text collections or sources. For example, a student is interested in the field of text classification and has collected some papers of this field from ICML, now he wants to find papers of text classification from ICDM. At this time, PU-Learning is helpful. The papers collected from ICML are positive documents (P), all the papers in the ICDM are unlabeled documents (U).

P. Perner (Ed.): MLDM 2007, LNAI 4571, pp. 574–584, 2007.

With the help of a PU-Learning system, the user can get the papers he wants from ICDM automatically.

Recently several approaches have been proposed for solving the PU-Learning problem, such as typically the two-step methods such as PNB and PNCT [8] [3]. However, current existing methods can not perform well in some cases. Experiments show that especially when the labeled positive set P for training is relative small, the classification result is not satisfactory. The main reason is due to the uniqueness of PU-Learning: 1) a large portion of training documents is unlabeled and no labeled negative documents are given; 2) the positive class contains documents from a special topic while the negative class contains documents from diverse topics. When the size of P is small, it can hardly reflect the true feature distribution of the positive class. Small P and the high diversity of the negative class will make building a good classifier extremely difficult [8].

Graph based semi-supervised learning [11] is usually effective for the classification task in the case of small size of labeled training. This kind of methods assumes label smoothness over the graph. In other words, they are smooth with respect to the intrinsic structure revealed by the given labeled and unlabeled data. Current graph-based methods mainly include spectral methods [4], random walks [9], graph mincuts [1], Gaussian random field and harmonic functions [12], etc. The characteristics of the graph based semi-supervised motivate us think of using this kind of methods to solve the PU-Learning problem with small positive dataset P. In this paper, we propose a novel method aiming at solving the PU-Learning problem when the given positive dataset P is lacking. To overcome the difficulty caused by small size of positive dataset, we combine the graph-based method with classical two-step methods of PU-Learning in an effective way, and further present our approach called PE-PUC for constructing the *positive document enlarging PU classifier*.

The organization of the paper is as follows. In Section 2 we introduce the background of the two-step method for PU-Learning and the graph-based semi supervised learning. In Section 3, we present our PE-PUC approach by combining the graph based method to solve the PU-Learning problem with respect to text classification. In Section 4, we give the evaluation of our PE-PUC approach with experimental results. In Section 5, we have the concluding remarks and the future work.

2 PU-Learning and Graph Based Semi-supervised Learning

2.1 Two-Step Method of PU-Learning

The given training data for PU-Learning is the labeled positive dataset P and the unlabeled dataset U. The key feature of PU-Learning is that there is no labeled negative data for training, which makes the task of building classifier challenging. One class of algorithm for solving this problem is based on a two-step strategy.

Step 1: Identify a set of reliable negative documents RN, from the given unlabeled dataset U. In this step, several techniques can be used, such as naïve Bayesian approach, spy technique, 1-DNF and Rocchio algorithm, etc.

Step 2: Build a set of classifiers by applying a classification algorithm iteratively using the given labeled positive documents P, the extracted negative documents RN and the remaining unlabeled documents U-RN; at last, select a good classifier from the set. In this step, Expectation Maximization (EM) algorithm and Support Vector Machine (SVM) usually are used.

2.2 Graph-Based Semi-supervised Learning

Graph-based semi-supervised learning [11] [12] considers the problem of learning with labeled and unlabeled data. The problem can be described as follows.

Given a point set $\chi = \{x_1,...,x_l,x_{l+1},...,x_{l+n}\} \subset \mathbb{R}^m$ and a label set $C = \{1,...,c\}$, the first l points of χ are labeled as $y_i \in C$, here, each class of C at least has one point. The remaining n points are unlabeled. The task is to predict the label of unlabeled points. The graph-based method using the concept and characteristic of graph, compute the similarities between nodes and propagate according to a given rule until reach a global stable state. The points with high similarity are considered to have the same label.

3 PE-PUC Approach: Positive Document Enlarging PU Classifier

As indicated in Section 1, current two-step methods cannot work well when P is small. In order to solve the PU-Learning problem, the two-step methods first extract a set of reliable negative documents from U in Step 1. The key requirement for this step is that the identified negative documents from the unlabeled set must be reliable and relative pure, that is to say, with no or very few positive documents in RN. If not so, too many noisy documents will damage the performance of classifier, which is built in Step 2. When the size of P is small, P is too small to reflect the true feature distribution of the positive class. In the two-step methods, whatever technique we use in Step 1, it is difficult to get reliable RN. In other words, after Step 1, many positive documents may be extracted from U as negative ones and put into RN. In Step 2, the noisy RN and the small P will make it impossible to build good classifiers.

Our PE-PUC method proposes a solution of the PU-Learning problem with small P. Intuitively, if we can extract some positive documents from U to enlarge P, we will possibly extract RN with high precision in Step 1. However, it is difficult to extract positive documents from U because: 1) U is large in size and high in diversity; 2) only a small portion of U is positive. It is difficult to avoid importing some negative documents into P when enlarging P. Those noisy documents can not improve the system but make it even poorer. To enlarge P with high precision, we present the PE-PUC algorithm using the graph-based semi-supervised techniques with main steps in Figure 1.

PE-PUC (P,U)

Input: the given labeled positive documents, P, the given unlabeled documents, U;

Output: PU Classifier

1. Based on P, extract a set of negative documents, RN, from U;
2. Enlarge P: Extract a set of reliable positive documents, RP, from U-RN;
3. P'= P∪RP, U'=U-RP, extract a set of negative documents, RN', from U';
4. Build the final classifier using P', RN' and U'-RN'.

Fig. 1. PE-PUC algorithm

3.1 Extracting Negative Documents from U

We use the naïve Bayesian method to extract negative documents RN from U and get the remaining unlabeled dataset U-RN. The detail of the procedure is shown in Figure 2.

The reason for labeling each document in U with the class label "-1" is that the proportion of positive documents in U is usually very small. In order to build the naïve Bayesian classifier, we firstly assume U is negative. Since naïve Bayesian method can tolerate some noise, this assumption is feasible.

Extract RN (P, U)

Input: the given labeled positive documents, P, the given unlabeled documents, U;

Output: a set of reliable negative documents, RN, a set of remaining unlabeled documents, U-RN.

1. Label each document in P with the class label 1;
2. Label each document in U with the class label -1;
3. Build a naïve Bayesian classifier, NB-C, using P and U;
4. Classify U using NB-C;
5. RN ← documents which are classified as negative; U-RN ← documents which are classified as positive;

Fig. 2. Algorithm for extracting reliable negative documents

3.2 Enlarge P: Extracting RP from U-RN

In order to solve the PU-Learning problem with small P, we try to enlarge P by extracting some reliable positive documents from U-RN. Now we give the detail of this procedure.

Given a point set $\chi = \{x_1,...,x_l,x_{l+1},...,x_{l+n}\} \subset \mathbb{R}^m$, the first l points of χ are labeled positive documents; the remaining points of χ are unlabeled documents which are to be ranked according to their relevance to the labeled positive documents. Let $d: \chi \times \chi \to \mathbb{R}$ denotes a matrix on χ, this matrix assigns each pair x_i, x_j the distance $d(x_i, x_j)$, and $f: \chi \to \mathbb{R}$ denotes a ranking function which assigns each data

point of χ a ranking score. Finally, we define a vector $y=[y_1,...y_{l+n}]^T$, in which $y_1,...y_l=1$, referring to the labeled positive documents, and $y_{l+1},...y_{l+n}=0$, referring to the unlabeled documents.

Graph based semi-supervised learning is an effective approach to deal with small size of labeled training for the purpose of classification. But for PU-Learning based classification, the problem is that we don't have any negative documents for propagation. Thus, an improved graph-based algorithm for extracting RP from U-RN is proposed as shown in Figure 3. An intuitive description of the algorithm is to randomly select a set of positive documents from P and put them into PL, which is used as the seeds for propagation, and then a weighted graph is formed which takes each point in PL∪ (U-RN) as a vertex. A positive ranking score to each point in PL is further assigned while zero to the remaining ones, and all the points then spread their scores to the nearby points via the weighted graph. This spread process is repeated until a global stable state is reached, and all the points except the seed points will have their own scores according to which they will be ranked. The resultant ranking score of an unlabeled document in U-RN is in proportion to the probability that it is relevant to the positive class, with large ranking score indicating high probability. So, at last, we can choose a number of the top ranked documents as reliable positive documents and use them to enlarge P.

Enlarge P (P, U-RN, λ)

Input: a set of positive documents, P, a set of unlabeled documents, U-RN, the percentage of U-RN which will be extracted as positive documents, λ, $\lambda\in(0,1)$;

Output: a set of positive documents, RP;

1. RP $\leftarrow \varnothing$, $n \leftarrow$ the number of documents in U-RN;
2. Randomly select l documents from P and put them in PL;
3. Form the affinity matrix W, $W_{ij}=\exp\left[-\|x_i-x_j\|^2/2\sigma^2\right]$ if $i\neq j$, $W_{ii}=0$;
4. Symmetrically normalize W by $S=D^{-1/2}WD^{-1/2}$. D is the diagonal matrix with (i,i)-element equal to the sum of the ith row of W;
5. $f^*=(1-\alpha)(1-\alpha S)^{-1}y$, $\alpha\in(0,1)$. Rank each document $x_i, i\in[l+1,l+n]$ according to the ranking score in f^* (largest ranked first);
6. RP $\leftarrow$ the top ranked documents in U-RN ($|\mathrm{RP}|=\lceil\lambda\times|U-RN|\rceil$)

Fig. 3. Algorithm for enlarging P

But when P is extremely small, only several labeled positive documents are known for training. In this case, just extracting RP from U-RN to enlarge P may not improve the performance distinctly. Here we propose a repeated extraction approach to take place of the second step in PE-PUC, namely, enlarging P repeatedly. The procedure of repeated extraction is shown in Figure 4.

Repeated Extraction

Input: the given labeled positive documents, P, the given unlabeled documents, U, the number of iteration, m, $\Lambda = \{\lambda_1, ..., \lambda_i, ..., \lambda_m\}$;

Output: a set of reliable positive documents, RP;

1: **for** $i = 1 : m$ **do**
2: get U-RN from Extract RN (U, P);
3: get RP from Enlarge P (P, U-RN, λ_i);
4: P←P∪RP, U←U-RP, $i = i + 1$;
5: **end for**
6: **return** RP

Fig. 4. Algorithm for repeated extraction

In PU-Learning problem, the negative class consists of diverse topics. It is the diversity that makes it difficult to extract RP from U. Thus, the main issue is to find a way to deal with the problem of diversity. The key to semi-supervised learning problem is the prior assumption of consistency: 1) nearby points are likely to have the same label; 2) points on the same structure (such as a cluster) are likely to have the same label. The classifying function, which is constructed by the graph-based method, is sufficiently smooth with respect to the intrinsic structure revealed by the given labeled and unlabeled data. Using this method to extract RP, the propagation of ranking score reflects the relationship of all the data points (each document in PL and U-RN now is looking as a point in the graph), since in the feature space, distant points will not have similar ranking scores unless they belong to the same cluster consisting of many points that help to link the distant points, and nearby points will have similar ranking scores unless they belong to different clusters. As we use positive documents as seeds for propagation, so after convergence, the documents with higher-ranking scores are more likely positive documents.

Another reason that we adopt the graph-based method is that it needs few labeled documents for propagation. This characteristic accords with our situation when the size of P is small. No matter how small |P| is, this method is relative feasible.

In this step, we extract positive documents from U-RN but not from U. This is reasonable. The given unlabeled set U is mixed with positive and negative documents. Usually the proportion of positive documents in U is quite small, and the negative documents are of high diversity, so it is difficult to extract positive documents from U with high precision. Moreover, the number of documents in U is quite large, which will make the computation complicated and time consuming. According to our experiments, when P is small, most of the negative documents in U are extracted into RN, and a lot of positive documents in U are also selected into RN. In other words, RN is of high recall but low precision. Under this circumstance, the number of documents in U-RN is much smaller than the number of documents in U, so the computation of the graph-based method is easy. In addition, the proportion of positive documents in U-RN is much larger than the proportion of positive documents in U, which makes the extraction with high precision possible.

3.3 Build the Final Classifier

In the process of enlarging P, a reliable positive set RP is extracted from U-RN and added to P. Then, we use P' and U' as the new input, and get the newly extracted negative documents, which is defined as set RN'. The final classifier is built based on P', RN' and U'-RN'. In our work, we use two techniques to build the final classifier, one is based on the naïve Bayesian method and the other is based on the Expectation-Maximization (EM) algorithm.

For the naïve Bayesian method, we directly build the final classifier with P' and RN'. For the EM algorithm, we build a set of classifiers using P', RN' and U'-RN'. EM iteratively runs naïve Bayesian algorithm to revise the probabilistic label of each document in set U'-RN'. The iteration of EM at each time generates a naïve Bayesian classifier. After convergence, we can get several classifiers. Since it is not easy to catch the best classifier, we choose the better one between the first classifier and the classifier at convergence as the final result.

4 Experiment and Evaluation

4.1 Experiment Setup

In the experiment, we use the 20 Newsgroups[1] dataset. For the 20newsgroup, there are totally 20 different classes, where each class contains about 1,000 documents. We use each newsgroup as the positive set and the rest of the 19 groups as the negative set, which creates 20 datasets. For each dataset, 30% of the documents are randomly selected as test documents, the rest (70%) are used as training documents. The training datasets are selected as follows. γ% of the documents from the positive class is first selected as the positive set P. The rest of the positive documents and negative documents are mixed to form the unlabeled set U. Our work focuses on the situation when |P| is small, so γ is ranged from 1% to10% for evaluating our method.

In our experiment, we use NB-NB and NB-EM as the baseline systems which are adopted in [8]. In the process of enlarging P, 10 documents are randomly selected from P as the seeds for propagation. We compute the affinity matrix W with $\sigma = 1.0$ and iteration with $\alpha = 0.99$. The number of positive documents selected from U-RN is set according to |U-RN| and the parameter λ. We test different λ settings to get a better result. In the last step, naïve Bayesian and EM algorithms are used to build the final classifier, which are represented as PE-PUC-NB and PE-PUC-EM, respectively. We use the popular F-score on the positive class as the evaluation measure.

4.2 Result Evaluation

The PE-PUC Method Can Give Better Results When P Is Small

Table 1 is the average of F-scores of the 20 datasets for each γ setting. Columns 2 and 3 show the results of the baseline systems. Columns 4 and 5 show the results of our PE-PUC approach. The comparative result of the experiment is shown in Figure 5.

[1] http://www.cs.cmu.edu/afs/cs/project/theo-11/www/naive-bayes/20_newsgroups.tar.gz

Table 1. The results of Newsgroup20

γ%	NB-NB	NB-EM	PE-PUC	
			NB	EM
1	0.063	0.429	0.272	0.451
2	0.155	0.499	0.474	0.512
3	0.192	0.511	0.538	0.538
4	0.253	0.524	0.597	0.597
5	0.321	0.530	0.648	0.648
6	0.370	0.531	0.625	0.625
7	0.421	0.568	0.611	0.627
8	0.464	0.590	0.630	0.666
9	0.497	0.599	0.642	0.679
10	0.530	0.625	0.657	0.690

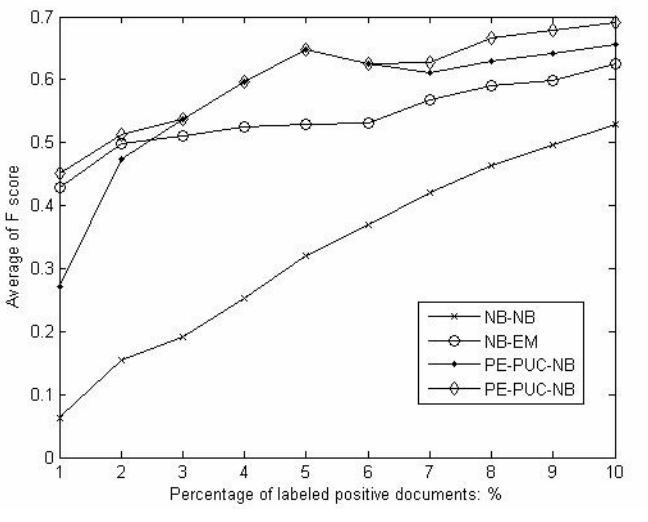

Fig. 5. Experiment results for the 20Newsgroup

The results indicate that our PE-PUC method performs better than the baseline systems significantly when P is small. For some cases, using the EM algorithm to build the final classifier can boost the systems. However, for the other cases, EM gives the same result as NB. As we use a classifier selection mechanism with the EM algorithm, which is able to select the first classifier if it is better than the one at

convergence, so we can see, for some instances, the iteration of EM algorithm cannot boost the systems but degraded them.

Analysis of the Enlarging P Procedure

In our work, the graph-based method is used to extract positive documents from U-RN. Now we further analyze the function of this procedure according to the experimental results.

(1) The number of positive documents extracted from U-RN affects the performance of the system.

We set the size of RP according to the number of documents in U-RN and the parameter λ, where $|\mathrm{RP}| = \lceil \lambda \times |U - RN| \rceil$. Table 2 gives the results of different λ settings for $\gamma = 9\%$ in term of F-score. Due to the space limitation, here we just list out the results of the four classes. The results of the other classes behave in the similar way. From the results, we can observe that different λ produces different results.

Table 2. $\gamma = 9\%$, Results of different λ settings

	PE-PUC					
	NB-NB			**NB-EM**		
P / λ	0.6	0.7	0.8	0.6	0.7	0.8
Crypt	0.921	0.918	0.909	0.921	0.918	0.909
Electronics	0.437	0.449	0.461	0.576	0.578	0.580
Med	0.396	0.417	0.388	0.396	0.417	0.388
Space	0.601	0.641	0.673	0.781	0.785	0.790

(2) Enlarging P can help to extract negative documents with higher precision.

As indicated in Section 3, the key requirement for the extraction of RN is high precision, which is a main problem when P is small. Table 3 gives the results of the precision of RN, which is the average of 20 datasets for $\gamma = 10\%$. As we can see from Table 3, the enlarging P procedure can help to extract negative documents with higher precision.

Table 3. $\gamma = 10\%$, Precision of RN

Method	**PE-PUC**			**Two-step**
λ	$\lambda = 0.6$	$\lambda = 0.7$	$\lambda = 0.8$	
$\overline{PrecisionOfRN}$	0.9771	0.9786	0.9756	0.9703

(3) Effectiveness of the repeated extraction approach

Another phenomenon shown in our experiment is that when the number of positive documents is extremely small, e.g. $\gamma \le 5\%$, the number of documents in U-RN will be very small. The reason is that when P is extremely small, P is too small to represent the distribution of the positive class, so most of the documents in U will be extracted into RN as negative ones. In this case, the number of positive documents can be extracted from U-RN is small, which limits the performance of our PE-PUC approach. To solve this problem, we conceive the repeated extraction approach to gradually enlarge P. From Table 4 shows that our approach is effective when P is extremely small. The value in the form is the average of F -scores of the 20 datasets for each γ setting.

Table 4. The results of PE-PUC with Repeated Extraction Approach

$\gamma\%$	$m=1$		$m=2$		m=3	
	NB	EM	NB	EM	NB	EM
1	0.187	0.437	0.213	0.425	0.272	0.451
2	0.321	0.510	0.389	0.501	0.474	0.512
3	0.392	0.518	0.522	0.522	0.538	0.538
4	0.474	0.533	0.597	0.597	0.535	0.535
5	0.563	0.563	0.648	0.648	0.600	0.600

5 Conclusion and Future Work

In this paper, we present a novel approach called PE-PUC to solve the PU-Learning problem when the positive dataset P is small. PU-Learning refers to the problem of learning a classifier from positive and unlabeled data. A typical kind of method for solving this problem is a so called two-step method. However, the two-step method cannot perform well when the positive dataset P is small. In our PE-PUC approach, the graph-based method is combined with the two-step method, which is used to extract some reliable positive documents from the unlabeled dataset to enlarge P. A comprehensive evaluation shows that our PE-PUC approach outperforms current existing PU-Learning algorithms especially when positive dataset is small.

In the future work, our research will further focus on the parameter selection, namely, to effectively determine the most suitable λ and m settings in a pure mechanical way with respect to different datasets.

Acknowledgments. This work was finished with the supports of China 973 Research Project under Grant No. 2002CB312006.

References

[1] Blum, A., Chawla, S.: Learning from labeled and unlabeled data using graph minicuts. In: Proceedings of the 18th International Conference on Machine Learning, pp. 19–26 (2001)
[2] Denis, F., Gilleron, R., Tommasi, M.: Text classification and co-training from positive and unlabeled examples. In: Proceedings of the ICML-03 Workshop on Continuum from Labeled to Unlabeled Data, pp. 80–87 (2003)
[3] Denis, F., et al.: Learning from positive and unlabeled examples. Journal of Theoretical Computer Science 1(248), 70–83 (2005)
[4] Joachims, T.: Transductive learning via spectral graph partitioning. In: Proceedings of the 20th International Conference on Machine Learning, pp. 290–297 (2003)
[5] Lee, W.S., Liu, B.: Learning with positive and unlabeled examples using weighted logistic regression. In: Proceedings of the Twentieth International Conference on Machine Learning, 448–455 (2003)
[6] Li, X., Liu, B.: Learning to classify text using positive and unlabeled data. In: Proceedings of the 18th International Joint Conference on Artificial Intelligence, pp. 587–594 (2003)
[7] Liu, B., Lee, W.S., Yu, P., Li, X.: Partially supervised classification of text documents. In: Proceedings of the 19th International Conference on Machine Learning, pp. 387–394 (2002)
[8] Liu, B., et al.: Building text classifiers using positive and unlabeled examples. In: Proceedings of the Third IEEE International Conference on Data Mining, pp. 179–188 (2003)
[9] Szummer, M., Jaakkola, T.: Partially labeled classification with Markov random walks. Advances in Neural Information Processing Systems, 945–952 (2002)
[10] Yu, H., Han, J., Chang, K.: PEBL: Positive example based learning for Web page classification using SVM. In: Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery in Databases, pp. 239–248 (2002)
[11] Zhou, D., et al.: Learning with local and global consistency. Advances in Neural Information Processing Systems, 321–328 (2003)
[12] Zhu, X., Ghahramani, Z., Lafferty, J.: Semi-supervised learning using Gaussian fields and harmonic functions. In: Proceedings of the 20th International Conference on Machine Learning, pp. 912–919 (2003)