

An Evaluation of Two-Step Techniques for Positive-Unlabeled Learning in Text Classification

Azam Kaboutari
Computer Department
Islamic Azad University,
Shabestar Branch
Shabestar, Iran

Jamshid Bagherzadeh
Computer Department
Urmia University
Urmia, Iran

Fatemeh Kheradmand
Biochemistry Department
Urmia University of Medical
Sciences
Urmia, Iran

Abstract: Positive-unlabeled (PU) learning is a learning problem which uses a semi-supervised method for learning. In PU learning problem, the aim is to build an accurate binary classifier without the need to collect negative examples for training. Two-step approach is a solution for PU learning problem that consists of two steps: (1) Identifying a set of reliable negative documents. (2) Building a classifier iteratively. In this paper we evaluate five combinations of techniques for two-step strategy. We found that using Rocchio method in step 1 and Expectation-Maximization method in step 2 is most effective combination in our experiments.

Keywords: PU Learning; positive-unlabeled learning; one-class classification; text classification; partially supervised learning

1. INTRODUCTION

In recent years, the traditional machine learning task division into supervised and unsupervised categories is blurred and a new type of learning problems has been raised due to the emergence of real-world problems. One of these partially supervised learning problems is the problem of learning from positive and unlabeled examples and called Positive-Unlabeled learning or PU learning [2]. PU learning assumes two-class classification, but there are no labeled negative examples for training. The training data is only a small set of labeled positive examples and a large set of unlabeled examples. In this paper the problem is supposed in the context of text classification and Web page classification.

The PU learning problem occurs frequently in Web and text retrieval applications, because Oftentimes the user is looking for documents related to a special subject. In this application collecting some positive documents from the Web or any other source is relatively easy. But Collecting negative training documents is especially requiring strenuous effort because (1) negative training examples must uniformly represent the universal set, excluding the positive class and (2) manually collected negative training documents could be biased because of human's unintentional prejudice, which could be detrimental to classification accuracy [6]. PU learning resolves need for manually collecting negative training examples.

In PU learning problem, learning is done from a set of positive examples and a collection of unlabeled examples. Unlabeled set indicates random samples of the universal set for which the class of each sample is arbitrary and may be positive or negative. Random sampling in Web can be done directly from the Internet or it can be done in most databases, warehouses, and search engine databases (e.g., DMOZ¹).

Two kinds of solutions have been proposed to build PU classifiers: the two-step approach and the direct approach. In this paper, we review some techniques that are proposed for step 1 and step 2 in the two-step approach and evaluate their performance on our dataset that is collected for identifying diabetes and non-diabetes WebPages. We find that using

Rocchio method in step 1 and Expectation-Maximization method in step 2 seems particularly promising for PU Learning.

The next section provides an overview of PU learning and describes the PU learning techniques considered in the evaluation - the evaluation is presented in section 3. The paper concludes with a summary and some proposals for further research in section 4.

2. POSITIVE-UNLABELED LEARNING

PU learning includes a collection of techniques for training a binary classifier on positive and unlabeled examples only. Traditional binary classifiers for text or Web pages require laborious preprocessing to collect and labeling positive and negative training examples. In text classification, the labeling is typically performed manually by reading the documents, which is a time consuming task and can be very labor intensive. PU learning does not need full supervision, and therefore is able to reduce the labeling effort.

Two sets of examples are available for training in PU learning: the positive set P and an unlabeled set U . The set U contains both positive and negative examples, but label of these examples not specified. The aim is to build an accurate binary classifier without the need to collect negative examples. [2]

To build PU classifier, two kinds of approaches have been proposed: the two-step approach that is illustrated in Figure 1 and the direct approach. In The two-step approach as its name indicates there are two steps for learning: (1) Extracting a subset of documents from the unlabeled set, as reliable negative (RN), (2) Applying a classification algorithm iteratively, building some classifiers and then selecting a good classifier. [2]

Two-step approaches include S-EM [3], PEBL [6], Roc-SVM [7] and CR-SVM [8]. Direct approaches such as biased-SVM [4] and Probability Estimation [5] also are offered to solve the problem. In this paper, we suppose some two-step approaches for review and evaluation.

¹ <http://www.dmoz.org/>

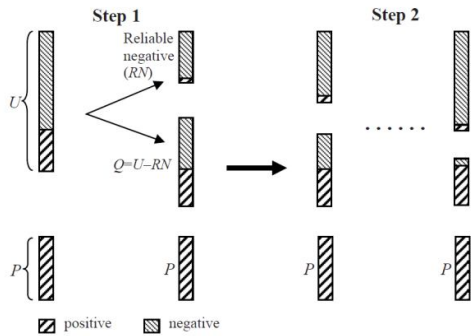


Figure 1. Two-step approach in PU learning [2].

2.1 Techniques for Step 1

For extracting a subset of documents from the unlabeled set, as reliable negative five techniques are proposed:

2.1.1 Spy

In this technique small percentage of positive documents from P are sampled randomly and put in U to act as “spies” and new sets P_s and U_s are made respectively. Then the naïve Bayesian (NB) algorithm runs using the set P_s as positive and the set U_s as negative. The NB classifier is then applied to assign a probabilistic class label $\Pr(+|d)$ to each document d in U_s . The probabilistic labels of the spies are used to decide which documents are most likely to be negative. S-EM [3] uses Spy technique.

2.1.2 Cosine-Rocchio

It first computes similarities of the unlabeled documents in U with the positive documents in P using the cosine measure and extracts a set of potential negatives PN from U . Then the algorithm applies the Rocchio classification method to build a classifier f using P and PN . Those documents in U that are classified as negatives by f are regarded as the final reliable negatives and stored in set RN . This method is used in [8].

2.1.3 IDNF

It first finds the set of words W as positive words that occur in the positive documents more frequently than in the unlabeled set, then those documents from the unlabeled set that do not contain any positive words in W extracted as reliable negative and used for building set RN . This method is employed in PEBL [6].

2.1.4 Naïve Bayesian

It builds a NB classifier using the set P as positive and the set U as negative. The NB classifier is then applied to classify each document in U . Those documents that are classified as negative denoted by RN . [4]

2.1.5 Rocchio

This technique is the same as that in the previous technique except that NB is replaced with Rocchio. Roc-SVM [7] uses Rocchio technique.

2.2 Techniques for Step 2

If the set RN contains mostly negative documents and is sufficiently large, a learning algorithm such as SVM using P and RN applied in this step and it works very well and will be able to build a good classifier. But often a very small set of negative documents identified in step 1 especially with IDNF technique, then a learning algorithm iteratively runs till it converges or some stopping criterion is met. [2]

For iteratively learning approach two techniques proposed:

2.2.1 EM-NB

This method is the combination of naïve Bayesian classification (NB) and the EM algorithm. The Expectation-Maximization (EM) algorithm is an iterative algorithm for maximum likelihood estimation in problems with missing data [1].

The EM algorithm consists of two steps, the Expectation step that fills in the missing data, and the Maximization step that estimates parameters. Estimating parameters leads to the next iteration of the algorithm. EM converges when its parameters stabilize.

In this case the documents in Q ($= U - RN$) regarded as having missing class. First, a NB classifier f is constructed from set P as positive and set RN as negative. Then EM iteratively runs and in Expectation step, uses f to assign a probabilistic class labels to each document in Q . In the Maximization step a new NB classifier f is learned from P , RN and Q . The classifier f from the last iteration is the result. This method is used in [3].

2.2.2 SVM Based

In this method, SVM is run iteratively using P , RN and Q . In each iteration, a new SVM classifier f is constructed from set P as positive and set RN as negative, and then f is applied to classify the documents in Q . The set of documents in Q that are classified as negative is removed from Q and added to RN . The iteration stops when no document in Q is classified as negative. The final classifier is the result. This method, called I-SVM is used in [6].

In the other similar method that is used in [7] and [4], after iterative SVM converges, either the first or the last classifier selected as the final classifier. The method, called SVM-IS.

3. EVALUATION

3.1 Data Set

We suppose the Internet as the universal set in our experiments. To collect random samples of Web pages as unlabeled set U we used DMOZ, a free open Web directory containing millions of Web pages. To construct an unbiased sample of the Internet, a random sampling of a search engine database such as DMOZ is sufficient [6].

We randomly selected 5,700 pages from DMOZ to collect unbiased unlabeled data. We also manually collected 539 Web pages about diabetes as positive set P to construct a classifier for classifying diabetes and non-diabetes Web pages. For evaluating the classifier, we manually collected 2500 non-diabetes pages and 600 diabetes page. (We collected negative data just for evaluating the classifier.)

3.2 Performance Measure

Since the F-score is a good performance measure for binary classification, we report the result of our experiments with this measure. F-score is the harmonic mean of precision and recall. Precision is defined as the number of correct positive predictions divided by number of positive predictions. Recall is defined as the number of correct positive predictions divided by number of positive data.

3.3 Experimental Results

We present the experimental results in this subsection. We extracted features from normal text of the content of Web pages, and then we perform stopwording, lowercasing and stemming. Finally, we get a set of about 176,000 words. We used document frequency (DF), one of the simple unsupervised feature selection methods for vocabulary and vector dimensionality reduction [9].

The document frequency of a word is the number of documents containing the word in the training set, in our case in P_{UU}. Then we create a ranked list of features, and returns the *i* highest ranked features as selected features, which *i* is in {200, 400, 600, 1000, 2000, 3000, 5000, 10000}.

As discussed in Section 2, we studied 5 techniques for Step 1 and 3 techniques for Step 2 (EM-NB, I-SVM and SVM-IS). Clearly, each technique for first step can be combined with each technique for the second step. In this paper, we will empirically evaluate only the 5 possible combinations of methods of Step 1 and Step 2 that available in the LPU², a text learning or classification system, which learns from a set of positive documents and a set of unlabeled documents.

These combinations are S-SVM which is Spy combined with SVM-IS, Roc-SVM is Rocchio combined with SVM-IS, Roc-EM is Rocchio+EM-NB, NB-SVM is Naïve Bayesian+ SVM-IS and NB-EM is Naïve Bayesian+ EM-NB.

In our experiments, each document is represented by a vector of selected features, using a bag-of-words representation and term frequency (TF) weighting method which the value of each feature in each document is the number of times (frequency count) that the feature (word) appeared in the document. When running SVM in Step 2, the feature counts are automatically converted to normalized tf-idf values by LPU. The F-score is shown in Figure 2.

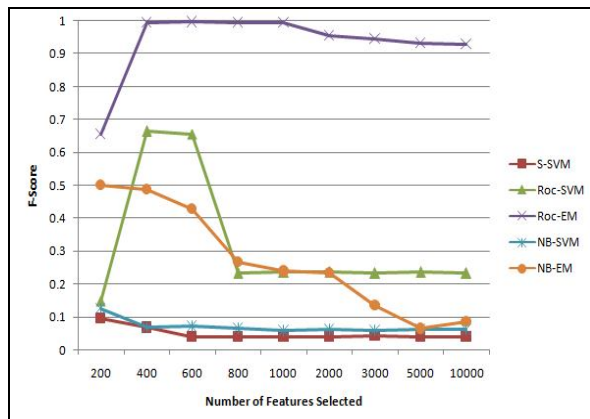


Figure 2. Results of LPU using DF feature selection method.

As Figure 2 shows, very poor results are obtained in S-SVM which Spy is used in Step 1 and SVM-IS is used in Step 2. Since we obtain better results in other combinations that SVM-IS is used in Step 2, we conduct that Spy is not a good technique for Step 1 in our experiments. By using NB in step 2, results are improved and best results we have obtained in our experiments when using Rocchio technique in Step 1. Figure 2 also shows that how using EM-NB instead of SVM-IS in Step 2 can improve results significantly.

The average of all F-score in each combination of techniques of Step 1 and Step 2 are shown in Table 1. As seen in Table 1 and Figure 2 Roc-EM is the best combination in our experiments which Rocchio technique is used in Step 1 and EM-NB is used in Step 2.

Table 1. Comparison of two-step approaches results.

	S-SVM	Roc-SVM	Roc-EM	NB-SVM	NB-EM
Average F-score	0.0489	0.3191	0.9332	0.0698	0.2713

4. CONCLUSIONS

In this paper, we discussed some methods for learning a classifier from positive and unlabeled documents using the two-step strategy. An evaluation of 5 combinations of techniques of Step 1 and Step 2 that available in the LPU system was conducted to compare the performance of each combination, which enables us to draw some important conclusions. Our results show that in the general Rocchio technique in step 1 outperforms other techniques. Also, we found that using EM for the second step performs better than SVM. Finally, we observed best combination for LPU in our experiments is R-EM, which is Rocchio, combined with EM-NB.

In our future studies, we plan to evaluate other combinations for Step 1 and Step 2 for Positive-Unlabeled Learning.

5. REFERENCES

- [1] Dempster, N. Laird and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, 1977, 39(1): pp. 1-38.
- [2] Liu and W. Lee, "Partially supervised learning", In "Web data mining", 2nd ed., Springer Berlin Heidelberg, 2011, pp. 171-208.
- [3] Liu, W. Lee, P. Yu and X. Li, "Partially supervised classification of text documents," In *Proceedings of International Conference on Machine Learning(ICML-2002)*, 2002.
- [4] B. Liu, Y. Dai, X. Li, W. Lee and Ph. Yu, "Building text classifiers using positive and unlabeled examples," In *Proceedings of IEEE International Conference on Data Mining (ICDM-2003)*, 2003.
- [5] Elkan and K. Noto, "Learning classifiers from only positive and unlabeled data," In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2008)*, 2008.
- [6] H. Yu, J. Han and K. Chang, "PEBL: Web page classification without negative examples", *Knowledge and Data Engineering, IEEE Transactions on*, vol.16, no.1, pp. 70- 81, Jan. 2004.
- [7] X. Li and B. Liu. "Learning to classify texts using positive and unlabeled data". In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI-2003)*, 2003.
- [8] X. Li, B. Liu and S. Ng, "Negative Training Data can be Harmful to Text Classification," In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-2010)*, 2010.
- [9] X. Qi and B. Davison, "Web page classification: Features and algorithms," *ACM Comput. Surv.*, 41(2): pp 1-31, 2009.

² <http://www.cs.uic.edu/~liub/LPU/LPU-download.html>