

A Simple Probabilistic Approach to Learning from Positive and Unlabeled Examples

Dell Zhang

School of Computer Science & Information Systems
Birkbeck, University of London
dell.z@ieee.org

Wee Sun Lee

Department of Computer Science
National University of Singapore
leews@comp.nus.edu.sg

Abstract

We propose a simple probabilistic approach to learning from positive and unlabeled examples, and show experimentally that it can approximate or outperform other state-of-the-art approaches to this problem in spite of its simplicity. By employing a linear-time learning algorithm such as PrTFIDF, our approach can be highly efficient and scalable.

1 Introduction

The classic problem of *supervised learning* is basically to train a classifier $y = f(\mathbf{x}) \in \{+1, -1\}$ based on a set of labeled examples L and then use it to predict the labels of test examples (Mitchell 1997). Since it is often hard or expensive to get labeled data but unlabeled data are widely available, recently there is a considerable interest on the problem of *semi-supervised learning*, i.e., learning from a small set of labeled examples L together with a large set of unlabeled examples U , such as (Blum & Mitchell 1998; Nigam et al. 2000).

We address a special kind of semi-supervised learning in this paper. Suppose we are interested in one specific class of data and seek to distinguish the interesting instances (i.e., positive examples) with others (i.e., negative examples), we often lack labeled negative examples on hand, because it is not so common to collect or label uninteresting instances. In such situations, L contains labeled positive examples only so we also denote it by P . This problem of *learning from positive and unlabeled examples* is prevalent in real-world applications. To automatically filter web pages according to a user's preference, the browsed or bookmarked pages can be used as positive examples while unlabeled examples can be easily collected from the web. To automatically find machine learning literature, the ICML papers can be used as positive examples while unlabeled examples can be easily collected from the ACM or IEEE digital library. To automatically

identify cancer patients, the patients who are known to have cancers can be used as positive examples while unlabeled examples can be easily collected from the hospital's patient database. To automatically discover future customers for direct marketing, the current customers of the company can be used as positive examples while unlabeled examples can be purchased at a low cost compared with obtaining negative examples.

The key feature of this problem is that there is no labeled negative example, which makes conventional supervised or semi-supervised learning techniques inapplicable. It is possible to discard the unlabeled examples and learn only from the positive examples, e.g., using One-Class SVM (Scholkopf et al. 2001). However, in general it should be beneficial to take advantage of the unlabeled examples rather than throwing them away. Some theoretical studies of this problem can be found in (Muggleton 1996; Denis 1998; Liu et al. 2002).

We would like to first review the existing approaches to this problem in section 2, then propose a simple probabilistic approach in section 3 and show its effectiveness experimentally in section 4, finally make conclusion in section 5.

2 Existing Approaches

A few approaches to learning from positive and unlabeled examples have been proposed in recent years.

The PNB method (Denis et al. 2002) modifies the Naïve Bayes algorithm (Mitchell 1997) to solve this problem. Its basic idea is to construct the model of negative examples by statistically removing the effect of positive examples from the model of unlabeled examples. The PNCT method (Denis et al. 2003) extends this idea to the Co-Training setting (Blum & Mitchell 1998) where the data have two redundant views (feature sets) that are compatible but conditionally independent, so as to make it work even when the number of positive examples is small. Their major shortcoming is that they require the

knowledge of the prior probability of the positive examples which is usually not available in practice.

One family of methods to solve this problem takes a two-step strategy:

- (1) identifying a set of reliable negative examples from the unlabeled set U ;
- (2) building a series of classifiers by iteratively applying a classification algorithm and then selecting a good classifier from them.

The S-EM method (Liu et al. 2002) uses the so-called “spy” technique in the first step, and then uses the Naïve Bayes algorithm (Mitchell 1997) together with the Expectation Maximization (EM) technique (Dempster et al. 1977) in the second step. The RC-SVM method (Li & Liu 2003) uses the Rocchio algorithm (Rocchio 1971) together with the K-Means clustering technique (MacQueen 1967) in the first step, and then uses the Support Vector Machine (SVM) algorithm (Cristianini & Shawe-Taylor 2000) in the second step. The PEBL method (Yu et al. 2004) and its refined faster version SVMC (Yu 2005) exploit the margin maximization property of the Support Vector Machine (SVM) algorithm (Cristianini & Shawe-Taylor 2000). They use the 1-DNF algorithm or the Rocchio algorithm (Rocchio 1971) or the One-Class SVM algorithm (Scholkopf et al. 2001) in the first step, and then iteratively trains a binary SVM classifier until it converges in the second step. The underlying idea of these two step methods is to iteratively increase the number of unlabeled examples that are classified as negative while maintaining the positive examples correctly classified. This idea has been justified to be effective for this problem in (Liu et al. 2002). For a comprehensive survey of this type of methods, please refer to (Liu et al. 2003).

The PN-SVM method (Fung et al. 2005) goes further along this way to challenge the tough situation where the number of positive examples is extremely small and the negative class is very diverse. It extracts not only negative examples but also positive examples from the unlabeled set U using the K-Means clustering technique (MacQueen 1967), and then simply applies the Support Vector Machine (SVM) algorithm (Cristianini & Shawe-Taylor 2000).

Another family of methods to solve this problem reduces this problem to the problem of learning with high one-sided noise by treating the unlabeled set U as a set of noisy negative examples. The W-LR method (Lee & Liu 2003) performs logistic regression after weighting the examples to handle noise rates of greater than a half. The Biased SVM (B-SVM) method (Liu et al. 2003) directly uses the asymmetric cost formulation (Morik et al. 1999) of the Support Vector Machine (SVM) algorithm (Cristianini & Shawe-Taylor 2000).

3 Our Approach

Let C_+ and C_- denote the positive and negative classes respectively. Given an instance \mathbf{x} , it is natural to do classification by comparing $\Pr[C_+|\mathbf{x}]$ and $\Pr[C_-|\mathbf{x}]$, i.e., a reasonable classification function is

$$f(x) = \text{sgn}(\Pr[C_+|\mathbf{x}] - \Pr[C_-|\mathbf{x}]).$$

If the positive and unlabeled examples are representative, we can reasonably assume that for an example in C_+ , it is randomly left unlabeled with an unknown probability $p \in [0,1]$ (Denis 1998). In other words, any example in C_+ has probability $1-p$ to be in the set P , and probability p to be in the set U . Remember that all example in C_- are unlabeled (in the set U). Therefore

$$\Pr[P|\mathbf{x}] = \Pr[C_+|\mathbf{x}](1-p),$$

$$\Pr[U|\mathbf{x}] = \Pr[C_+|\mathbf{x}]p + \Pr[C_-|\mathbf{x}].$$

A straightforward calculation would show that

$$\begin{aligned} & (1+p)\Pr[P|\mathbf{x}] - (1-p)\Pr[U|\mathbf{x}] \\ &= (1+p)\Pr[C_+|\mathbf{x}](1-p) \\ &= -(1-p)(\Pr[C_+|\mathbf{x}]p + \Pr[C_-|\mathbf{x}]) \\ &= (1-p)\Pr[C_+|\mathbf{x}] - (1-p)\Pr[C_-|\mathbf{x}] \\ &= (1-p)(\Pr[C_+|\mathbf{x}] - \Pr[C_-|\mathbf{x}]). \end{aligned}$$

Consequently

$$\begin{aligned} & \Pr[C_+|\mathbf{x}] - \Pr[C_-|\mathbf{x}] \\ &= \left(\frac{1+p}{1-p} \right) \Pr[P|\mathbf{x}] - \Pr[U|\mathbf{x}], \end{aligned}$$

Let $b = (1+p)/(1-p)$. Now we see that the classification function can be equivalently written as

$$f(\mathbf{x}) = \text{sgn}(b\Pr[P|\mathbf{x}] - \Pr[U|\mathbf{x}]).$$

Following the above insight directly, we propose a new method to learn $f(\mathbf{x})$ from P and U in two stages:

- (1) estimating $\Pr[P|\mathbf{x}]$ and $\Pr[U|\mathbf{x}]$;
- (2) estimating b .

3.1 Estimating $\Pr[P|\mathbf{x}]$ and $\Pr[U|\mathbf{x}]$

A lot of learning algorithms, e.g., Naïve Bayes (Mitchell 1997), could be used to estimate $\Pr[P|\mathbf{x}]$ and $\Pr[U|\mathbf{x}]$ from P and U . In our experiments on text data (section 4), we chose the PrTFIDF algorithm (Joachims 1997). It runs very fast because it only needs a single sequential scan over the dataset.

Given a document collection D (in our case $P \cup U$) and one of its subsets S (in our case P or U), PrTFIDF calculates $\Pr[S|\mathbf{x}]$ to be:

$$\Pr[S|\mathbf{x}] = \sum_{w \in V} \Pr[S|w] \Pr[w|\mathbf{x}],$$

where V means the vocabulary (the set of all distinct words in D). Let $TF(w, \mathbf{x})$ denote the term frequency of the word w in the document \mathbf{x} . Then the probabilities in the above formula are estimated as:

$$\Pr[w|\mathbf{x}] = \frac{TF(w, \mathbf{x})}{\sum_{w' \in \mathbf{x}} TF(w', \mathbf{x})},$$

$$\Pr[S|w] = \frac{\sum_{\mathbf{x} \in S} \Pr[w|\mathbf{x}]}{\sum_{\mathbf{x} \in D} \Pr[w|\mathbf{x}]}.$$

3.2 Estimating b

We are not able to compute b straightforwardly by $(1+p)/(1-p)$ because the probability p is unknown. To overcome this obstacle, we employ a separate validation set (that also consists of labeled positive and unlabeled examples), and select the value of b that can make the resulting classifier achieve optimal performance on the validation set.

The need to learn from positive and unlabeled examples often arise in information retrieval (Baeza-Yates & Ribeiro-Neto 1999) situations, where we have a collection of positive examples and would like to retrieve more positive examples from a source of unlabeled examples. A commonly used performance measure in information retrieval is the F -score, $F = 2pr/(p+r)$, where p is the precision $\Pr[C_+ | f(\mathbf{x})=1]$ and r is the recall $\Pr[f(\mathbf{x})=1 | C_+]$. Unfortunately it is unclear how to estimate the F score without labeled negative examples, so we turn to a similar performance measure proposed in (Lee & Liu 2003):

$$pr/\Pr[C_+] = r^2/\Pr[f(\mathbf{x})=1].$$

This performance measure behaves similarly to the F score in the sense that it is large only when both p and r are large. Since r can be estimated using the labeled positive examples in the validation set and $\Pr[f(\mathbf{x})=1]$ can be estimated using the entire validation set, we are able to compute this performance measure without labeled negative examples. In this way, a good value of b can be found even though p is not known in advance.

3.3 The Novelty

Our approach is different with PNB (Denis et al. 2002) and PNCT (Denis et al. 2003), because it does not require to know $\Pr[C_+]$ beforehand.

Our approach is unlike the existing two-step methods including S-EM (Liu et al. 2002), RC-SVM method (Li & Liu 2003), PEBL (Yu et al. 2004) and SVMC (Yu 2005), and unlike PN-SVM (Fung et al. 2005) either, because it does not attempt to identify a set of reliable negative examples first.

Our approach is somewhat similar to W-LR (Lee & Liu 2003) and B-SVM (Liu et al. 2003), and it can be considered as the generative (as opposed to discriminative) modeling version of their approach. Furthermore, our approach is able to employ any probabilistic learning algorithm as long as it is not too sensitive to noise. By employing a linear-time learning algorithm such as PrTFIDF (Joachims

1997), our approach can be highly efficient and scalable.

We call our approach using the PrTFIDF algorithm (Joachims 1997) Biased-PrTFIDF (or B-Pr for short).

4 Experiments

We have empirically evaluated B-Pr on text data and compared it with other state-of-the-art approaches in terms of the macro-averaged F -score (Yang & Liu 1999).

4.1 Experiments on the Reuters Dataset

The Reuters-21578 dataset¹ consists of 21,578 newswire articles that are distributed in 135 categories. Only the ten largest categories were used, namely (0)acq; (1)corn; (2)crude; (3)earn; (4)grain; (5)interest; (6)money-fx; (7)ship; (8)trade; (9)wheat.

Table 1: Experimental results on the Reuters dataset.

C_+	$p = 0.55$			$p = 0.85$		
	PEBL	RC-SVM	B-Pr	PEBL	RC-SVM	B-Pr
(0)	0.891	0.909	0.920	0.001	0.867	0.907
(1)	0.663	0.645	0.610	0.000	0.822	0.601
(2)	0.798	0.811	0.830	0.000	0.801	0.760
(3)	0.956	0.923	0.957	0.000	0.891	0.940
(4)	0.900	0.903	0.863	0.020	0.869	0.753
(5)	0.770	0.616	0.748	0.000	0.724	0.744
(6)	0.714	0.764	0.830	0.000	0.785	0.785
(7)	0.672	0.843	0.742	0.008	0.596	0.547
(8)	0.728	0.728	0.800	0.000	0.778	0.772
(9)	0.783	0.792	0.650	0.000	0.854	0.597
Avg.	0.788	0.793	0.795	0.003	0.799	0.741

Our experiments on this dataset is with the same setting as (Li & Liu 2003) in order to allow comparison. All documents were pre-processed by removal of stop-words and removal of words that occurred no more than 5 times in the dataset. For each category (e.g., ‘trade’), a binary text classification task (‘trade’ vs. ‘non-trade’) is formulated. For each task, true positive documents were randomly left unlabeled with probability p , while true negative documents were first randomly discarded with probability $1-p$ and the rest were included in the set of unlabeled documents. Two different values of p (0.55 and 0.85) were tried. The dataset was randomly split into two sets: a training set containing 70% of the documents; a validation set containing 30% of the documents. The test set was just the set of unlabeled documents.

¹ <http://www.daviddlewis.com/resources/testcollections/reuters21578/>

Table 1 shows the F -scores on the Reuters-21578 dataset using our proposed B-Pr method and two other methods: PEBL (Yu et al. 2004) and RC-SVM (Li & Liu 2003). The results for PEBL and RC-SVM are taken directly from (Li & Liu 2003). In summary, B-Pr > RC-SVM > PEBL (when $p = 0.55$); RC-SVM > B-Pr > PEBL (when $p = 0.85$). Our approach worked better than PEBL and RC-SVM when $p = 0.55$. Its performance was much better than PEBL but slightly worse than RC-SVM when $p = 0.85$.

4.2 Experiments on the 20NG Dataset

The 20NG database² consists of posts from 20 newsgroups, namely (0)atheism; (1)autos; (2)space; (3)graphics; (4)motorcycles; (5)christian; (6)ms-win; (7)baseball; (8)guns; (9)pc; (10)hockey; (11)mideast; (12)mac; (13)crypt; (14)politics; (15)x-win; (16)electronics; (17)religion; (18)for-sale; (19)med.

Table 2: Experimental results on the 20NG dataset.

C_+	$p = 0.3$			$p = 0.7$		
	S-EM	W-LR	B-Pr	S-EM	W-LR	B-Pr
(0)	0.546	0.655	0.703	0.577	0.552	0.689
(1)	0.644	0.804	0.811	0.542	0.710	0.767
(2)	0.844	0.887	0.875	0.631	0.738	0.824
(3)	0.513	0.643	0.731	0.480	0.498	0.644
(4)	0.784	0.870	0.907	0.698	0.799	0.851
(5)	0.679	0.741	0.819	0.644	0.636	0.789
(6)	0.572	0.717	0.694	0.471	0.633	0.571
(7)	0.705	0.830	0.897	0.705	0.727	0.868
(8)	0.673	0.730	0.769	0.580	0.618	0.706
(9)	0.505	0.619	0.641	0.470	0.541	0.592
(10)	0.830	0.897	0.933	0.803	0.826	0.884
(11)	0.855	0.873	0.907	0.840	0.796	0.877
(12)	0.515	0.767	0.692	0.450	0.677	0.658
(13)	0.849	0.868	0.861	0.721	0.843	0.828
(14)	0.567	0.620	0.677	0.488	0.564	0.629
(15)	0.628	0.760	0.797	0.618	0.636	0.733
(16)	0.527	0.666	0.727	0.349	0.527	0.652
(17)	0.440	0.540	0.588	0.421	0.495	0.510
(18)	0.709	0.732	0.760	0.573	0.591	0.675
(19)	0.838	0.867	0.848	0.745	0.782	0.800
Avg.	0.661	0.754	0.782	0.590	0.659	0.727

Our experiments on this dataset is with the same setting as (Li & Liu 2003) in order to allow comparison. All documents were pre-processed by removal of stop-words and removal of words that occurred no more than 5 times in the dataset. For each newsgroup (e.g., ‘politics’), a binary text classification task (‘politics’ vs. ‘non-politics’) is formulated. For each task, true positive documents

were randomly left unlabeled with probability p , while true negative documents were always left unlabeled. Two different values of p (0.3 and 0.7) were tried. The dataset was randomly split into three sets: a training set containing 50% of the documents; a validation set containing 20% of the documents; and a test set containing 30% of the examples.

Table 2 shows F -scores on the 20NG dataset using our proposed B-Pr method and two other methods: S-EM (Liu et al. 2002) and W-LR (Lee & Liu 2003). The results for S-EM and W-LR are taken directly from (Lee & Liu 2003). In summary, B-Pr > W-LR > S-EM (when $p = 0.3$); B-Pr > W-LR > S-EM (when $p = 0.7$). Our approach outperformed S-EM and W-LR in both cases.

5 Conclusion

In this paper we propose a simple probabilistic approach to learning from positive and unlabeled examples. It is easy to understand and implement due to its simplicity, yet it has been shown to be very effective and efficient in practice.

References

- Baeza-Yates, R. & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*, Addison-Wesley.
- Blum, A. & Mitchell, T. (1998). Combining Labeled and Unlabeled Data with Co-Training. *Proceedings of the 11th Annual Conference on Computational Learning Theory (COLT)*. pp. 92-100. Madison, WI.
- Cristianini, N. & Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines*. Cambridge, UK, Cambridge University Press.
- Dempster, A. P., Laird, N. M., et al. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B* 39(1): 1-38.
- Denis, F. (1998). PAC Learning from Positive Statistical Queries. *Proceedings of the 9th International Conference on Algorithmic Learning Theory (ALT)*. pp. 112-126. Otzenhausen, Germany.
- Denis, F., Gilleron, R., et al. (2002). Text Classification from Positive and Unlabeled Examples. *Proceedings of the 9th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU)*. pp. 1927-1934. Annecy, France.
- Denis, F., Laurent, A., et al. (2003). Text Classification and Co-Training from Positive and Unlabeled Examples. *Proceedings of the ICML 2003 Workshop: The Continuum from Labeled to*

² <http://www-2.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/news20.html>

- Unlabeled Data*. pp. 80-87, 2003. Washington, DC.
- Fung, G. P. C., Yu, J. X., et al. (2005). Text Classification without Labeled Negative Documents. *Proceedings of the 21st International Conference on Data Engineering (ICDE)*. pp. 594-605. Tokyo, Japan.
- Joachims, T. (1997). A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. *Proceedings of the 14th International Conference on Machine Learning (ICML)*. pp. 143-151. Nashville, Tennessee, USA.
- Lee, W. S. & Liu, B. (2003). Learning with Positive and Unlabeled Examples using Weighted Logistic Regression. *Proceedings of the 20th International Conference on Machine Learning (ICML)*. pp. 448-455. Washington DC, USA.
- Li, X. & Liu, B. (2003). Learning to Classify Texts Using Positive and Unlabeled Data. *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI)*. pp. 587-594. Acapulco, Mexico.
- Liu, B., Dai, Y., et al. (2003). Building Text Classifiers Using Positive and Unlabeled Examples. *Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM), Melbourne, FL*. pp. 179-188.
- Liu, B., Lee, W. S., et al. (2002). Partially Supervised Classification of Text Documents. *Proceedings of the 19th International Conference (ICML)*. pp. 387-394. Sydney, Australia.
- MacQueen, J. (1967). Some Methods for Classification and Analysis of Multivariate Observations. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*. 1, pp. 281-297. L. M. L. a. N. Neyman: University of California Press.
- Mitchell, T. (1997). *Machine Learning*, McGraw Hill.
- Morik, K., Brockhausen, P., et al. (1999). Combining Statistical Learning with a Knowledge-Based Approach - A Case Study in Intensive Care Monitoring. *Proceedings of the 16th International Conference on Machine Learning (ICML)*. pp. 268-277. Bled, Slovenia.
- Muggleton, S. (1996). Learning from Positive Data. *Proceedings of the 6th International Workshop on Inductive Logic Programming (ILP)*. pp. 358-376. Stockholm, Sweden.
- Nigam, K., McCallum, A., et al. (2000). Text Classification from Labeled and Unlabeled Documents using EM. *Machine Learning* 39(2-3): 103-134.
- Rocchio, J. J. (1971). Relevance Feedback in Information Retrieval. *The SMART Retrieval System: Experiments in Automatic Document Processing*. pp. 313-323. G. Salton: Prentice-Hall.
- Scholkopf, B., Platt, J., et al. (2001). Estimating the Support of a High-Dimensional Distribution. *Neural Computation* 13(7): 1443-1471.
- Yang, Y. & Liu, X. (1999). A Re-examination of Text Categorization Methods. *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. pp. 42-49. Berkeley, CA.
- Yu, H. (2005). Single-Class Classification with Mapping Convergence. *Machine Learning*.
- Yu, H., Han, J., et al. (2004). PEBL: Web Page Classification without Negative Examples. *IEEE Transactions on Knowledge and Data Engineering (TKDE)* 16(1): 70-81.