Dealing with Different Distributions in Learning from Positive and Unlabeled Web Data

Xiaoli Li
School of Computing
National University of Singapore/
Singapore-MIT Alliance
Singapore 117543
lixl@comp.nus.edu.sg

Bing Liu
Department of Computer Science
University of Illinois at Chicago
851 South Morgan Street
Chicago, IL 60607-7053
liub@cs.uic.edu

ABSTRACT

In the problem of learning with positive and unlabeled examples, existing research all assumes that positive examples *P* and the hidden positive examples in the unlabeled set *U* are generated from the same distribution. This assumption may be violated in practice. In such cases, existing methods perform poorly. This paper proposes a novel technique A-EM to deal with the problem. Experimental results with product page classification demonstrate the effectiveness of the proposed technique.

Categories & Subject Descriptors: H.2.8 [**Database Applications**]: Data mining

General Terms: Algorithms, Experimentation

Keywords: Classification, positive and unlabeled learning.

1. INTRODUCTION

In traditional classification, a classifier is built using labeled training data of every class. In the past few years, a partially supervised classification problem is also studied. In this problem, one has a set *P* of positive examples of a particular class and a set *U* of unlabeled examples that contains examples from class *P* and also other types of examples (called negative examples). One wants to build a classifier to classify the examples in *U* into cases from *P* and cases not from *P*. As there is no labeled negative example, traditional classification techniques are not applicable. In the past two years, several techniques [5, 7, 2, 3, 4] were proposed to solve the problem. These techniques mainly use a two-step strategy. The first step tries to identify a set of reliable negative documents from *U*. The second step builds a classifier by iteratively applying a classification algorithm, i.e. EM [1] or SVM.

All the existing techniques assume that positive examples in *P* and the hidden positive examples in *U* are generated from the same distribution. In the context of the Web or text documents, this means that the word features in these positive documents in both *P* and *U* are similar and with similar frequencies. Existing techniques also assume that the proportion of positive examples in *U* is small. These assumptions may be violated in practice. For example, one wants to collect all printer pages from the Web. One can use the printer pages from one site (e.g., amazon.com) as the set *P* of positive pages and use product pages from another Web site (e.g., cnet.com) as *U*. He/She wants to classify all the pages in *U* into printer pages and non-printer pages. Although printer pages from the two sites have many similarities, they can also be quite different. Additionally, *U* (e.g., cnet.com) may also contain a large number of printer pages, which make the proportion of positive examples in *U* quite large. In such cases, directly apply existing methods give very poor results. The main reason is that the first step is unable to give reliable negative pages. Consequently, the second step builds poor classifiers.

This paper proposes a novel technique to deal with this problem. The proposed method (called A-EM for Augmented EM) is in the framework of EM [5, 6]. The proposed technique has two novelties for dealing with the above problems:

- We add a number of irrelevant documents (which are definitely negative documents) in *U*. This reduces the proportion of positive documents in *U*, which enables us to compute the parameters of the classifier more accurately.
- The EM algorithm generates a sequence of classifiers. However, the performances of this sequence of classifiers may not be necessarily improving. This is a well-known phenomenon that has been documented in a number of papers [5, 6]. We then propose a classifier selection criterion to select a good classifier from the set of classifiers produced by EM. Although there are existing classifier selection methods given in [5, 3], they perform poorly also due to the different data distributions identified above.

We have performed a large number of experiments using product pages. Our experimental results show that the new method outperforms existing methods dramatically.

2. The PROPOSED TECHNIQUE

The A-EM algorithm is given in Figure 1. Initially, we assign each positive document d_i in *P* the class label "+" (line 2), and each document d_j in unlabeled set *U* the class label "-" (line 3). Let us ignore *O* in line 1 for the time being. Using this initial labeling a naïve Bayesian (NB) classifier can be built (line 4). This classifier is then applied to classify documents in *U* to obtain the posterior probability ($P(+|d_j)$ and $P(-|d_j)$) for each document in *U*. We can then iteratively employ the revised posterior probability to build a new NB classifier. The process goes on until the parameters converge.

In Figure 1, the key piece of information needed for classification is $P(w_t|c_j)$, where w_t is a word and c_j is a class. If there are a large number of positive examples in *U* or there are many keywords that are indicative of positive documents also occurring in *U* very often, then the NB classifier will not be able to separate positive and negative classes well because for these features NB is not sure whether they are representative of positive or negative class.

```
Algorithm A-EM(P, U, O)
1.   Let $N = U \cup O$;
```

WWW 2004, May 17–22, 2004, New York, New York, USA.
ACM 1-58113-912-8/04/0005.

```
2.  For each $d_i \in P$ let $P(+|d_i) = 1$, $P(-|d_i) = 0$;
3.  For each $d_i \in N$ let $P(+|d_i) = 0$, $P(-|d_i) = 1$;
4.  Build the initial naïve Bayesian classifier NB-C;
5.  Loop while classifier parameters change
6.      For each document $d_i \in N$
7.          Compute $P(+|d_i)$ and $P(-|d_i)$ using NB-C;
8.          Update $P(c_j)$ and $P(w_t|c_j)$ with the new probabilities in
                step 7 (a new NB-C is being built in the process)
9.  Select a good classifier from the series of classifiers
    produced by EM. // each iteration of EM produces a NB classifier.
```

Figure 1 A-EM algorithm with Naïve Bayes classifier

To deal with this problem, we introduce additional irrelevant (negative) documents O into the original unlabeled set U (line 1 in Figure 1). This changes the probability $P(w_t|-)$. Obviously, the proportion of positive documents in $O+U$ is reduced and consequently $P(w_t|-)$ is reduced for a positive keyword w_t. Note that $P(w_t|+)$ does not change because we do not add anything in the positive set P. In effect, we amplify or boost the positive features. In classifying documents in U, those positive documents are likely to get much higher values of $P(+|d_i)$, and lower values of $P(-|d_i)$. This means that we have boosted the similarity of positive documents in P and U, which allows us to build more accurate classifiers.

EM generates a sequence of classifiers. A classifier selection criterion is needed in order to select a good classifier from the set of classifiers produced by EM. Since the distribution of the documents in positive training set P are not the same as that of the positives in unlabeled set U, the two existing techniques [3, 5] do not work because they both depend on P. Our proposed technique depends primarily on the unlabeled set U. So the distribution difference will not cause a major problem (line 9).

Here we use the F value to evaluate the performance of the classifier in each iteration of EM. Suppose TP, FN, FP, TN are the number of true positive, false negative, false positive and true negative respectively, we have (p is precision and r is recall)

$$F = \frac{2 * p * r}{p + r} = \frac{2TP}{(TP + FP) + (TP + FN)} \quad (1)$$

Note that $TP+FP$ is the number of documents that are classified as positive (we denote the document set as CP) and $TP+FN$ is the actual number of positive documents in U (we denote it as PD, and it is a constant). So the F value can be expressed as:

$$F = \frac{2TP}{|CP| + PD} \quad (2)$$

Here we choose to use an estimate of change in F value to decide which iteration of EM to select as the final classifier. From equation (3), the change in F value from iteration i-1 to i is

$$\Delta_i = \frac{F_i}{F_{i-1}} = \frac{TP_i}{TP_{i-1}} * \frac{|CP_{i-1}| + PD}{|CP_i| + PD} \quad (3)$$

In the EM algorithm, we select iteration i as our final classifier if Δ_i is the last iteration with value greater than 1. Note that in equation (3), $|CP_{i-1}|$ and $|CP_i|$ are the number of documents classified as positive in iteration i and i+1 respectively. We estimate PD by using the number of documents classified as positive when EM converges. Then the question is how to estimate TP_i / TP_{i-1}. Our idea here is that first we get a set K of representative keywords for the positive class. For a document, the more positive keywords it contains, the more likely it belongs to the positive class. Hence, we use

$$\sum_t^{|K|} N(w_t, d_i), d_i \in CP_i \Big/ \sum_t^{|K|} N(w_t, d_i), d_i \in CP_{i-1} \quad (4)$$

to estimate TP_i / TP_{i-1}, where $\sum_t^{|k|} N(w_t, d_i), d_i \in CP_i$ is the total number of keywords in the document set CP_i. Intuitively, for a set CP_i (documents classified as positive) in an EM iteration, the larger the total number of positive keywords are in CP_i, the more true positive documents it contains. For instance, if CP_i contains more printer keywords, then it is likely that CP_i contains more true printer pages.

3. EMPIRICAL EVALUATION

Our empirical evaluation is done using Web pages from 5 commercial Web sites, Amazon, CNet, PCMag, J&R and ZDnet. We choose Web pages that focus on the following categories of products: Notebook, Digital Camera, Mobile Phone, Printer and TV. The construction of positive set P and unlabeled set U is done as follows: we use Web pages of a particular type of product from a single site ($Site_i$) as positive pages P, e.g., camera pages from Amazon. The unlabeled set U is the set of all product pages from another site ($Site_j$) ($i \neq j$), e.g., CNet. We also use U as the test set in our experiments because our objective is to extract those positive pages in U, e.g., camera pages in CNet. The irrelevant document set O is from two large document corpora: 20 Newsgroup and Reuters. Due to space limitations, Table 1 only shows the average classification results of various techniques by adding Reuters and 20newsgroup as irrelevant data. A-EM outperforms other methods dramatically and adding what kind of data is not very important as long as they are negative.

Table1 Comparison of various techniques

Adding	Roc	RocSVM	PEBL	**A-EM**
Reuters	0.645	0.734	0.723	**0.872**
20newsgroup	0.667	0.726	0.721	**0.891**

4. ACKNOWLEDGMENTS

Bing Liu's work is supported by the National Science Foundation under the NSF grant IIS-0307239.

5. REFERENCES

[1] A. Dempster, N. Laird & D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 1977.

[2] F. Denis, R. Gilleron and M. Tommasi. "Text classification from positive and unlabeled examples." *IPMU-02*. 2002.

[3] W. Lee, & Bing Liu. "Learning with positive and unlabeled examples using weighted logistic regression. *ICML-2003*.

[4] X. Li, & B. Liu. Learning to classify text using positive and unlabeled data. *IJCAI-2003*.

[5] B. Liu, W. Lee, P. Yu, & X. Li. Partially supervised classification of text documents. *ICML-2002*.

[6] K. Nigam, A. McCallum, S. Thrun, & T. Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 2000.

[7] H. Yu, J. Han, & K. Chang. PEBL: Positive example based learning for Web page classification using SVM. *KDD-2002*.