# Positive Unlabelled Learning for Document Classification

**Xiao-Li Li**
*Institute for Infocomm Research, Singapore*

**See-Kiong Ng**
*Institute for Infocomm Research, Singapore*

## INTRODUCTION

In traditional supervised learning, a large number of labeled positive and negative examples are typically required to learn an accurate classifier. However, in practice, it is costly to obtain the class labels for large sets of training examples, and oftentimes the negative examples are lacking. Such practical considerations motivate the development of a new set of classification algorithms that can learn from a set of labeled positive examples $P$ augmented with a set of unlabeled examples $U$ (which contains both hidden positive and hidden negative examples). That is, we want to build a classifier using $P$ and $U$ in the absence of negative examples to classify the data in $U$ as well as future test data. This problem is called the Positive Unlabelled learning problem or PU learning problem.

For instance, a computer scientist may want to build an up-to-date repository of machine learning (ML) papers. Ideally, one can start with an initial set of ML papers (e.g., a personal collection) and then use it to find other ML papers from related online journals or conference series, e.g., Artificial Intelligence journal, AAAI (National Conference on Artificial Intelligence), IJCAI (International Joint Conferences on Artificial Intelligence), SIGIR (ACM Conference on Research & Development on Information Retrieval), and KDD (ACM International Conference on Knowledge Discovery and Data Mining) etc. With the enormous volume of text documents on the Web, Internet news feeds, and digital libraries, finding those documents that are related to one's interest can be a real challenge.

In the application above, the class of documents that one is interested in is called the **positive documents** (i.e. the ML papers in the online sources). The set of known positive documents are represented as $P$ (namely, the initial personal collection of ML papers). The unlabelled set $U$ (papers from AAAI Proceedings etc) contains two groups of documents. One group contains documents of class $P$, which are the hidden positive documents in $U$ (e.g., the ML papers in an AAAI Proceedings). The other group, which comprises the rest of the documents in $U$, are the **negative documents** (e.g., the non-ML papers in an AAAI Proceedings) since they do not belong to positive class. Given a positive set $P$, **PU learning** aims to identify a particular class $P$ of documents from $U$ or classify the future test set into positive and negative classes. Note that collecting unlabeled documents is normally easy and inexpensive, especially those involving online sources.

## BACKGROUND

A theoretical study of PAC learning from positive and unlabeled examples under the statistical query model was first reported in (Denis, 1998). It assumes that the proportion of positive instances in the unlabeled set is known. Letouzey *et al.* (Letouzey, Denis, & Gilleron, 2000; Denis, Gilleron, & Letouzey, 2005) presented a learning algorithm based on a modified decision tree algorithm in this model. Muggleton (Muggleton, 1997) followed by studying the problem in a Bayesian framework where the distribution of functions and examples are assumed known. (Liu, Lee, Yu, & Li, 2002) reported sample complexity results and provided theoretical elaborations on how the problem may be solved.

A number of practical algorithms, S-EM, PEBL and Roc-SVM (Liu et al., 2002; Yu, Han, & Chang, 2002; Li & Liu, 2003) have also been proposed. They all conformed to the theoretical results in (Liu et al., 2002), following a two-step strategy: (1) identifying a set of reliable negative documents $RN$ from the unlabeled set (called *strong negative documents* in PEBL), and (2) building a classifier using $P$ (positive set), $RN$ (negative set) and $U$-$RN$

(unlabelled set) through applying an existing learning algorithm (such as EM or SVM) once or iteratively. Their specific differences are described in the next subsection "Existing Techniques S-EM, PEBL and Roc-SVM".

Other related works include: Lee and Liu's weighted logistic regression technique (Lee & Liu, 2003) and Liu et al.'s biased SVM technique (Liu, Dai, Li, Lee, & Yu, 2003). Both required a performance criterion to determine the quality of the classifier. In (Fung, Yu, Lu, & Yu, 2006), a method called PN-SVM was proposed to deal with the case when the positive set is small where it assumes that the positive examples in $P$ and the hidden positives in $U$ were all generated from the same distribution. More recently, PU learning was used to identify unexpected instances in the test set (Li, Liu, & Ng, 2007b). PU learning was also useful for extracting relations, identifying user preferences and filtering junk email, etc (Agichtein, 2006; Deng, Chai, Tan, Ng, & Lee., 2004; Schneider, 2004; Zhang & Lee., 2005).

## MAIN FOCUS

We will first introduce state-of-art techniques in PU learning for document classification, namely, S-EM (Liu et al., 2002), PEBL (Yu et al., 2002) and Roc-SVM (Li & Liu, 2003). Then, we will describe a document classification application which requires PU learning with only a small positive training set.

### Existing Techniques S-EM, PEBL and Roc-SVM

As mentioned earlier, the existing techniques (S-EM, PEBL and Roc-SVM) all use a two-step strategy. We will focus on discussing the first step (negative example extraction) of the three methods (Spy, 1DNF, Rocchio respectively), since the second step of the three methods is essentially similar.

The main idea of S-EM is to use a spy technique to identify some *reliable negative documents* from the unlabeled set $U$. S-EM works by sending some "spy" documents from the positive set $P$ to the unlabeled set $U$. The technique makes the following assumption: since the spy documents from $P$ and the hidden positive documents in $U$ are positive documents, the spy documents should behave identically to the hidden positive documents in $U$ and can thus be used to reliably infer the behavior of the unknown positive documents.

S-EM randomly samples a set $S$ of positive documents from $P$ and puts them in $U$. Next, it runs the naïve Bayesian (NB) technique using the set $P - S$ as positive and the set $U \cup S$ as negative. The NB classifier is then applied to classify each document $d$ in $U \cup S$, i.e., to assign it a probabilistic class label $\Pr(+|d)$, where "+" represents the positive class. Finally, it uses the probabilistic labels of the spies to decide which documents are most likely to be positive (the remaining documents are thus the reliable negative). S-EM sets a threshold using the probabilistic labels of spies which controls to extract most (85%) of spies from $U$, indicating that it can also extract out most of the other hidden positives from $U$ since they behave identically.

In comparison, PEBL (Positive Example Based Learning) tries to extract reliable (strong) negative documents by using 1DNF method. Those documents in $U$ that do not contain any positive features are regarded as reliable negative documents. 1DNF method first builds a positive feature set $PF$ containing words that occur in the positive set $P$ more frequently than in the unlabeled set $U$. Then it tries to filter out possible positive documents from $U$. A document in $U$ that does not contain any positive feature in $PF$ is regarded as a reliable negative document.

Unlike S-EM and PEBL, Roc-SVM performs negative data extraction using the Rocchio method. Roc-SVM uses the positive set $P$ as positive training data and $U$ as negative training data to build a Rocchio classifier where positive and negative prototype vectors $\vec{c}^{+}$ and $\vec{c}^{-}$ are constructed. In classification, for each document $\vec{d}'$ in unlabelled set $U$ , it simply uses the cosine measure (Salton & McGill, 1986) to compute the similarity of $\vec{d}'$ with each prototype vector. The class whose prototype vector is more similar to $\vec{d}'$ is assigned to the document. Those documents classified as negative form the reliable negative set $RN$.

Let us compare the above three methods for extracting reliable negative documents. Although S-EM is not sensitive to noise, it can be problematic when the EM's assumptions (the data is generated by a mixture model, and there is a one-to-one correspondence between mixture components and classes) do not hold (Liu et al., 2002). PEBL is sensitive to the number of positive documents. When the positive data