Chapter V
Partially Supervised Text Categorization

Xiao-Li Li
Institute for Infocomm Research, Singapore

ABSTRACT

In traditional text categorization, a classifier is built using labeled training documents from a set of predefined classes. This chapter studies a different problem: partially supervised text categorization. Given a set $P$ of positive documents of a particular class and a set $U$ of unlabeled documents (which contains both hidden positive and hidden negative documents), we build a classifier using $P$ and $U$ to classify the data in $U$ as well as future test data. The key feature of this problem is that there is no labeled negative document, which makes traditional text classification techniques inapplicable. In this chapter, we introduce the main techniques S-EM, PEBL, Roc-SVM and A-EM, to solve the partially supervised problem. In many application domains, partially supervised text categorization is preferred since it saves on the labor-intensive effort of manual labeling of negative documents.

INTRODUCTION

Text categorization is an important problem and has been studied extensively in machine learning, information retrieval and natural language processing. To build a text classifier, the user first collects a set of training documents, which are labeled with predefined or known classes (labeling is often done manually). A classification algorithm is then applied to the training data to build a classifier which is subsequently employed to assign the predefined classes to documents in a test set (for evaluation) or future instances (in practice). This approach to building classifiers is called supervised learning/classification because the training documents all have pre-labeled classes.
Over the past few years, a special form of text categorization, partially supervised text categorization, is proposed. This problem can be regarded as a two-class (positive and negative) classification problem, where there are only labeled positive training data, but no labeled negative training data. Due to the lack of negative training data, the classifier building is thus partially supervised. Since traditional classification techniques require both labeled positive and negative documents to build a classifier, they are thus not suitable for this problem. Although it is possible to manually label some negative documents, it is labor-intensive and very time consuming. Partially supervised text categorization studies to build a classifier using only a set of positive documents and a set of unlabeled documents. Note that partially supervised text categorization is also called PU learning where $P$ and $U$ represent “positive set” and “unlabeled set” respectively. Additionally, although PU learning is commonly studied in the context of text and Web page domain, the general concepts and the algorithms are also applicable to the classification tasks in other domains. Next, we discuss the problem definition and some possible applications of PU learning.

With the growing volume of text documents on the Web, Internet news feeds, and digital libraries, one often wants to find those documents that are related to one’s interest. For instance, one may want to build a repository of machine learning (ML) papers. First, one can start with an initial set of ML papers (e.g., an ICML Proceedings). One can then find those ML papers from related online journals or conference series, e.g., AI journal, AAAI, IJCAI, SIGIR, and KDD etc. Note that collecting unlabeled documents is normally easy and inexpensive in many text or Web page domains, especially those involving online sources.

The ability to build classifiers without negative training data is particularly useful if one needs to find positive documents from many text collections or sources. Given a new collection, the algorithm can be run to find those positive documents. Following the above example, given a collection of AAAI papers (unlabeled set), one can run the algorithm to identify those ML papers. Given a set of SIGIR papers, one can run the algorithm again to find those ML papers. If traditional classification techniques are used to deal with the application above, a user has to label the negative training set for every online sources, such as AAAI, SIGIR and KDD proceedings, etc (it is thus time consuming). In other words, one cannot use the classifier built using ML papers as positive set ($P$) and manually labeled non-ML papers in AAAI as negative set ($N$) to classify the SIGIR papers (as a test set) because they are from different domains — the classifier will suffer since the distribution of negative test documents in SIGIR (Information Retrieval papers) is largely different from that in the negative training documents in $N$ (Artificial Intelligence papers). A user would obviously prefer techniques that can provide accurate classification without manual labeling any negative document.

In the application above, the class of documents that one is interested in is called the positive class documents, or simply positive documents (e.g., the ML papers in an ICML Proceedings). The set of positive documents are represented as $P$. The unlabelled set $U$ (e.g., AAAI Proceedings) contains two parts of documents. One is also the documents of class $P$, which are the hidden positive documents in $U$. The rest of the documents in $U$ are called the negative class documents or simply 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 $U$. The diagram of PU learning is shown in the Figure 1.

Now, we can state the problem more formally as follow:

**Problem Statement:** Given a set $P$ of positive documents that we are interested in, and a set $U$ of unlabeled documents (the mixed set), which contains both positive documents and negative documents,