Leveraging Unlabeled Data for Classification

Yinghui Yang
University of California, Davis, USA

Balaji Padmanabhan
University of South Florida, USA

INTRODUCTION

Classification is a form of data analysis that can be used to extract models to predict categorical class labels (Han & Kamber, 2001). Data classification has proven to be very useful in a wide variety of applications. For example, a classification model can be built to categorize bank loan applications as either safe or risky. In order to build a classification model, training data containing multiple independent variables and a dependant variable (class label) is needed. If a data record has a known value for its class label, this data record is termed "labeled". If the value for its class is unknown, it is "unlabeled". There are situations with a large amount of unlabeled data and a small amount of labeled data. Using only labeled data to build classification models can potentially ignore useful information contained in the unlabeled data. Furthermore, unlabeled data can often be much cheaper and more plentiful than labeled data, and so if useful information can be extracted from it that reduces the need for labeled examples, this can be a significant benefit (Balcan & Blum 2005). The default practice is to use only the labeled data to build a classification model and then assign class labels to the unlabeled data. However, when the amount of labeled data is not enough, the classification model built only using the labeled data can be biased and far from accurate. The class labels assigned to the unlabeled data can then be inaccurate.

How to leverage the information contained in the unlabeled data to help improve the accuracy of the classification model is an important research question. There are two streams of research that addresses the challenging issue of how to appropriately use unlabeled data for building classification models. The details are discussed below.

BACKGROUND

Research on handling unlabeled data can be approximately grouped into two streams. These two streams are motivated by two different scenarios.

The first scenario covers applications where the modeler can acquire, but at a cost, the labels corresponding to the unlabeled data. For example, consider the problem of predicting if some video clip has suspicious activity (such as the presence of a "most wanted" fugitive). Vast amounts of video streams exist through surveillance cameras, and at the same time labeling experts exist (in law enforcement and the intelligence agencies). Hence labeling any video stream is possible, but is an expensive task in that it requires human time and interpretation (Yan et al 2003). A similar example is in the "speech-to-text" task of generating automatic transcriptions of speech fragments (Hakkani-Tur et al 2004, Raina et al 2007). It is possible to have people listen to the speech fragments and generate text transcriptions which can be used to label the speech fragments, but it is an expensive task. The fields of active learning (e.g. MacKay (1992), Saar-Tsechansky & Provost (2001)) and optimal experimental design (Atkinson 1996) addresses how modelers can selectively acquire the labels for the problems in this scenario. Active learning acquires labeled data incrementally, using the model learned so far to select particularly helpful additional training examples for labeling. When successful, active learning methods reduce the number of instances that must be labeled to achieve a particular level of accuracy (Saar-Tsechansky & Provost (2001)). Optimal experimental design studies the problem of deciding which subjects to experiment on (e.g. in medical trials) given limited resources (Atkinson 1996).

The second scenario, the focus in this chapter, covers applications where it is not possible to acquire the

unknown labels or such acquisition is not an option. The extreme cases of the previous scenario where the costs are prohibitively high can also be considered in this set. For example, consider the problem of predicting the academic performance (i.e. the graduating GPA) of thousands of current applicants to an undergraduate program. Ample data exists from the performance of ex-students in the program, but it is impossible to "acquire" the graduating GPA of current applicants. In this case is the unlabeled data (i.e. the independent variables of the current applicants) of any use in the process of building a model? A stream of recent research (Blum & Mitchell (1998), Joachims (1999), Chapelle (2003)) addresses this problem and presents various methods for making use of the unlabeled data for this context.

To some extent, approaches used to learning with missing values can be applied to learning the labels of the unlabeled data. One standard approach to learning with missing values is the EM algorithm (Dempster et al. 1977). The biggest drawback of such approaches is that they need to assume the class label follows a certain distribution.

A second approach for this (Blum & Mitchell, 1998) is co-training (and variants (Yarowsky, 1995)) which was initially applied to Web page classification, since labeling Web pages involves human intervention and is expensive. The idea is to first learn multiple classifiers from different sets of features. Each classifier is then used to make predictions on the unlabeled data and these predictions are then treated as part of the training set for the other classifiers. This approach works well for Web page categorization since one classifier can be trained based on words within pages while another (using different features) can be trained on words in hyperlinks to the page. This approach is in contrast with self-training where a classifier uses its own (selected) predictions on the unlabeled data to retrain itself.

Another approach is to use clustering and density estimation to first generate a data model from both the labeled and unlabeled data (e.g. Chapelle, 2003). The labels are then used for labeling entire clusters of data, or estimating class conditional densities which involves labeling of the unlabeled data dependent on their relative placement in the data space with respect to the original labeled data. A popular approach for implementing this idea is using generative mixture models and the EM algorithm. The mixture models are identified using unlabeled data, and then the labeled data is used to determine the classes to assign to the (soft) clusters generated from the combined data.

There is also work on integrating these ideas into a specific classifier, such as the development of Transductive Support Vector Machines (Joachims 1999). Extending the concept of finding optimal boundaries in a traditional SVM, this work develops methods to learn boundaries that avoid going through dense regions of points both in labeled as well as unlabeled data.

To summarize, the prior research on learning with unlabeled data focuses either on selecting unlabeled data to acquire labels, or use models built on labeled data to assign labels to unlabeled data.

MAIN FOCUS OF THE CHAPTER

In this chapter, we focus on an approach for using the unlabeled data in the case where labels cannot be acquired. This approach is different from the ones discussed above in that it does not involve assigning labels to the unlabeled data. Instead, this approach augments the features (independent variables) of the labeled data to capture information in the unlabeled data. This approach is based on the intuition that the combined labeled and unlabeled data can be used to estimate the joint distribution of attributes among the independent variables better, than if this was estimated from the labeled data alone. Specifically, if interactions (or patterns) among variables turn out to be useful features for modeling, such patterns may be better estimated using all available data.

The Approach

The approach and alternatives for comparison are pictorially illustrated in Figure 1. The approach presented is the path on the right (#3). First the column represented by the target attribute (Y) is removed, and the labeled and unlabeled data are combined into one large dataset. Then a pattern discovery procedure (e.g. a procedure to discover frequent itemsets) is applied to learn a set of patterns from this data. Let the number of patterns learned from just the independent variables of both the labeled and unlabeled data be Q_2. Each of these patterns then is used to create binary variables P_1, P_2, … , P_{Q2} indicating whether each given pattern is present in each record of the dataset. For example, for pattern number Q_2, we check whether a data record contains