读书报告

09118119 黄一凡

# 自己提出的问题

#### 1.Self-Training是不是会导致泛化能力不够强，因为它是用自己训练的结果来教自己，感觉会产生很大的误差

的确会出现这个问题。查阅资料后了解到，Self-Training是一个较早提出的半监督算法，所以相对较为落后。

Self-Training首先根据少量labeled数据训练出分类器，由于可用数据较少，分类器很有可能效果并不理想，这就容易造成 “一步错，步步错”的情况。

#### 2.在Co-Training中提到，用来训练分类器的两个特征的子集is sufficient for learning the target classification function。但是在训练之前我们是怎么得到这个结论的？还是说这只是一个assumption?

经过讨论我们认为这确实是一个assumption，或者可以人为手工进行相应的判断。

# 别人提出的问题

#### 1.Co-Training可否将属性分为三个集合训练三个分类器进行学习呢（或者更多）？

理论上可以这样进行，准确度也许还会提高，但是在实际中，往往不能这样。因为Co-Training需要满足的一个假设是两个特征子集 是相互独立的。如果有三个子集，那么这三个子集也需要相互独立，在现实操作中很难满足。

#### 2.5.1.1中通过leave-one-out在已标注数据集中交叉验证精度的方法选取μ是什么意思？

作为一个超参数，我们需要提前确定它的值。这时，我们采用leave-one-out的方法，首先在labeled data中拿出一个 instance，将剩余labeled data和unlabeled data进行训练，对取不同的值，得到多个classifier，再用这些classifier 对一开始拿出的labeled data instance进行验证，从而判断哪一个为最优。

#### 3.Mincut的代价函数是如何得到的，有何意义？

Mincut的代价函数如下：

其中，为所属的类别，我们用0或1表示。文章前面提到，当两个instance的相似度越高，其也就相应越大。 由此，为了最小化这个代价函数，对于相似度高的我们需要尽可能使，对于相似度低的我们需要尽可能使。

#### 4.5.11中mixture model究竟指什么？意义何在？为什么要有这前提假设？

朴素贝叶斯法是一种生成模型，它的目标是学习到生成数据的机制。如果我们的数据不符合一个模型分布，那么朴素贝叶斯 就不能学到其不存在的模型。

# 读书计划

#### 本周所读：

5.1

#### 下周计划：

《统计学习方法》第一到二章，另外继续学习高斯混合分布和朴素贝叶斯法

# 读书摘要

下面是我读书时做的一些笔记整理：

# 5 Partially Supervised Learning

## 5.1 Learning from Labeled and Unlabeled Examples

This section learns from a small set of labeled examples and a large set of unlabeled examples.

One key point to note is that although the number may be small, every class must have some labeled examples.

LU Learning: Labeled and Unlabeled Learning

### 5.1.1 EM Algorithm with Naïve Bayesian Classification

The EM algorithm consists of two steps, the Expectation step (or E-step), and the Maximization step (or M-step). - The E-step basically fills in the missing data based on the current estimation of the parameters - The M-step, which maximizes the likelihood, re-estimates the parameters. - EM converges to a local minimum when the model parameters stabilize - EM algorithm is not really a specific “algorithm”, but is a framework or strategy

### 5.1.2 Co-Training

Using the same training data to build two classifiers using two subsets of features.

The first assumption is that the example distribution is compatible with the target functions.

The second assumption is that the features in one set of an example are conditionally independent of the features in the other set.

The key idea of co-training is that classifier f1 adds examples to the labeled set that are used for learning f2 based on the X2 view, and vice versa.

**The second assumption, also called the conditional indenpence assumption is very important.**

### 5.1.3 Self-Training

The basic idea of this method is that the classifier uses its own predictions to teach itself.

### 5.1.4 Transductive Support Vector Machine

One way to use unlabeled data in training SVM is by selecting the labels of the unlabeled data in such a way that the resulting margin of the classifier is maximized.

Training for the purpose of labeling known (unlabeled) test instances is referred to as transduction, giving rise to the name transductive SVM.

### 5.1.5 Graph-Based Methods

Graph-based LU learning methods can be viewed as extensions of nearest neighbor supervised learning algorithms that work with both labeled and unlabeled instances.

We discuss three types of graph-based LU learning methods below: mincut, Gaussian fields and spectral graph transducer.

### 5.1.6 Discussion

#### Does the Unlabeled Set Always Help?

The answer is no: - When the assumptions are not true, the unlabeled data may harm learning - Researchers have not shown that when the labeled data set is sufficiently large, the unlabeled data still help. - Manual labeling more text documents may not be as difficult as it seems in some applications

#### How to evaluation LU learning?

1. If we use the evaluation same as traditional classification problems, which needs quantities of labeled data to act as test set, this contradicts the LU learning problem statement for the short in labeled data
2. We first use the classifier generated by LU learning to classify the unlabeled set or a new test set and then sample some classified documents to be checked manually in order to estimate the classification accuracy.