方骏-2020年3月1日-读书报告

1. 自己提出的问题的理解：
2. 提出的问题1：EM算法的两个假设都是什么意思？

讨论后的理解：数据是由一个混合模型产生的；混合成分和类别之间有一一对应关系。数据如果不是由混合模型产生，那么贝叶斯模型便不能应用到数据中，必须是有混合的分类，否则无法计算概率。那么混合成分必须是只有一个类别对应，这个类别不能再根据这个数据进一步细分成类别子集去对应，否则就无法应用该方法。

1. 提出的问题2：Co-training的第二个假设怎么影响算法的

讨论后的理解：从同一个数据提取的不同特征之间应该是条件独立的。如果不是独立的，而是相关的，那么之和用一个数据集进行分类就没有区别了，既然要起到用多个分类器的作用，就必须让这几个属性是相互独立的，才能产生不同的分类器，在进行无标签数据标签后加入对方数据集时，才可以起到随机的效果。

1. 别人提出的问题的理解：
2. 问题3：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。

自己的理解：这个是可能会出现的。因为这个方法是比较早提出的半监督方法，所以很简单，没有假设条件，所以可以应用的范围比较广，但是效果可能会比较差，因为在选择去标签数据时会比较困难，要选择可信度比较高的，但是这个可信度标准很难选择，而且需要逐渐严格，否则很容易会让训练的结果跑偏，因此误差很大是有可能的。但是选择好可信度就可以减少这种误差。

1. 问题4： 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?

自己的理解：这是一个假设，也就是要用这个方法，就必须要满足独立的几个属性这一假设。选择数据的时候找的也必须是这种可以将一个特征从多个属性来看待的，否则这个联合训练就和self-training没有区别了。

1. 问题5： 种子的类别标识表示了聚类结果中各个簇的类别标识是什么意思？

自己的理解：先对有标签的几个数据进行聚类得到几个簇，然后再利用这些簇对无标签的聚类，这些有标签的数据就是种子，这些种子就是表明了这几个簇的类别到底是什么，因为这些聚类就是根据有标签数据的类别计算出来的，所以完全可以这么代表。

1. 读书计划

1、本周完成的内容章节：5.1

2、下周计划：《统计机器学习》第一章

四、读书摘要及理解

1、首先是半监督学习，和监督学习以及无监督学习不同，算是一种折衷且针对具体问题的方法。在很多的学习任务中，存在着标签数据集较少的情况，这时运用监督学习的效果不够好，因为标签集太少。直接用无监督学习又不值得，所以半监督学习是一种可以很好运用的方法，它先学习较少的标签数据，联合无标签学习提高学习分类的准确性，这就是半监督学习，将有标签和无标签的数据都运用到方法中。

2、EM方法是其中一种比较经典的方法。EM算法包括两步，是E-step和M-step，E-step基本上是根据参数的当前估计来填充缺失的数据。最大化似然的M-step重新估计参数。缺失的数据就是缺失的数据标签，用有标签数据计算出的贝叶斯分类模型预测无标签数据的标签概率，这样所有的数据就有了标签，再去学习一个新的贝叶斯模型。下一个循坏再去修改响应的参数，知道参数不再改变，就完成了EM模型。但是这个模型应用的基础是满足两个假设，分别是：数据是由一个混合模型产生的；混合成分和类别之间有一一对应关系。如果这两个假设不满足，会影响算法的准确性，解决方法有二：减少无标签数据的权重，在计算公式概率时将无标签数据的权重减少；精准的找到混合成分，然后把这些成分的类别一一对应找到，进行类别细分。

3、Co-training方法是充分利用数据的特征进行计算的模型。根据数据属性的一个子集的数据构建一个分类器，然后用另一个子集构建另一个分类器，这两个子集是针对数据提供了两个不同的角度来看待同一个属性的，这样就可以得到两组数据，共同学习。学习了两个分类器f1和f2，用f1和f2再去预测无标签的数据，对每一个标签ci选择ni个可信度最高的数据，赋予标签ci，加入对方的有标签数据，再开始下一个循环，直到无标签数据为空。这个方法也有两个假设：特征集上的目标分类函数预测相同的标签；从同一个数据提取的不同特征之间应该是条件独立的。这个方法要求高的地方就是怎么样才能做到将一个数据特征从两个不同的角度去看，且这两个角度必须独立且得到的分类必须一样。这样在进行计算的时候，将f1得到的加标签的数据加入另一个角度的标签数据集时，这几个数据才能呈现出随机的样子，否则就无法保证准确性。

4、Self-training方法就比较简单了，就是直接用标记数据训练一个分类器，用到无标签数据上，选择前几个可信度最高的无标签数据加上标签，加入有标签数据，在进行下一次循环。这个方法的关键就是怎么选择可信度高的无标签数据，这个方法没有假设，应用很广，但是就是比较难选择数据，很容易很引起分类的偏离。

5、Transductive Support Vector Machines方法就是选择让分类器边距最大的类别赋值给无标签数据的。

6、Graph-Based Methods方法是基于图形学的各种方法。这些方法的基本思想是将有标签和无标签的实例视为图中的顶点，在图中使用相似函数定义实例之间的边权重。然后，使用具有由较大权重连接的相似实例的图来帮助标记未标记实例，使得由具有较大权重的边连接的顶点的标签趋于一致。方法有三种：mincut，Gaussian和fields and spectral graph transducer。

mincut：该方法将与有标签数据关联的顶点赋值为0或1，无标签数据的顶点通过如下的损失函数来赋值，让该损失函数最小：



该公式的优点是即使是一个组合优化问题，也能在多项式时间内求解。将其转换为最大流问题是一种很好的解决方法。

Gaussian Fields：最小化的损失函数是：

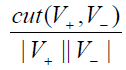


解决问题的公式可以写做：



其中vU是向量v的一部分，包含分配给未标记实例的值，vL是向量v的一部分，包含分配给已标记实例的值，DUU是D的子矩阵，由W中与未标记实例相关联的行的项之和组成。

Spectral Graph Transducer：mincut公式的一个潜在问题是mincut cost函数倾向于选择不平衡削减，即正类或负类中的实例数量远远超过另一类中的实例数量。该方法提出的损失函数是这样的：



最优化问题精简成：

