Web Data Mining第五次读书报告

报告人：71117415-喻泽弘

读书时间:2.24-3-1

读书进度：web Data Mining第五章

问题与解答

- 个人提出的问题
  - Co-Training可否将属性分为多个集合并且运用多个分类器进行学习?

    讨论认为，Co-Training可以将属性分为多个集合并且运用多个集合进行学习，但是这个集合需要满足Co-Training的预定条件，即：

    - 多个集合的分布需要和目标函数兼容，也就是说对于同一条元组，同个分类器在不同集合内预测的结果需要保持绝大部分的相同，即$f_1(x) = f_2(x) = f_3(x) = c$
    - 多个集合需要满足条件独立，即集合与集合的元素之间不会相互影响，对我们分布的预测不会造成影响

    现实情况中，对于两个集合要满足上述条件都十分困难，对于多个集合来讲，这些条件就变得不切实际了，并且运用多个分类器从一定程序上来讲，会增加模型的复杂度，但不一定会有更好的模型泛化性能以及预测结果准确率。
  - 如何理解EM算法的两个假设？Mix-model以及one-one correspondence between mixture components and classes.

    对EM算法的两个假设：

    - the data is generated by a mixture model:我认为这句话想要表达的意思就是，数据是通过混合模型生成出来的。混合模型，我的理解就是例如正态分布这种方式的分布，通过这种分布可以使得生成出来的数据更加均衡，从而算法能够能够考虑各个范围的数据，从而算法的泛化性能更好
    - one-one correspondence between mixture components and classes:第二个假设，我个人的理解是类别和子类别需要一一对应，比如：运动这个大类别中可能包含篮球、足球、羽毛球、等等。但是运动这个大类别不能包含数学，如果包含了数学，那么数学明显就是运动这个大类的脏数据，从而会导致训练结果变差。因此，对于第二个假设而言就是，类别和它的子类别需要一一对应的关系。
- 别人提出的问题
  - Self-Training是不是会导致泛化性能不够，因为它是用自己训练的结果来训练自己，感觉会造成很大的误差?

    对于Self-Training，肯定会有可能算法的误差很大，关键点在于初始算法的选择，如果初始算法本身就不适合模型的话，那么Self-Training显然不会有很好的结果，同时算法最重要的一步就是初始数据的选择，初始数据必须选择十分有把握的数据，这样也有利于后续的运算。
  - 为什么Co-Training 中的 confidential independent is a somewhat unrealistic assumption in practice?

    对于这个问题，首先Co-Training的第二个假设是两个子集合是条件独立的，但是在一些现实情况在，并不能找出这个集合，比如，某服装厂的数据有两个属性，身高以及年龄，如果对该服装厂的数据使用Co-Training的方法显然是不合适的，因为身高和年龄这两个数据并不是条件独立的，也就满足不了Co-Training的第二个假设。

下周计划

进入李航老师的《统计学习》一书进行学习，并尽快完成第一章的阅读

读书收获

- 认识到了半监督的学习方法，这是一种介于无监督学习以及监督学习之间的一种方法，该方法适用于数据集中既有有标签的数据，也有没有标签的数据。从而我们便需要根据有标签的数据生成一种算法，并运用这种算法去给无标签数据添加标签，从而再进行学习
- 认识到了半监督学习中的Co-Training方法，该方法根据数据集的性质，将数据集划分为两个条件独立的子集和，再对两个子集合分别使用分类器进行学习，并且通过对两个分类器的结果进行加权，从而得出一个较为精准的预测值。注意两点：
    - 子集合条件独立：只有满足子集合条件独立这个条件，我们才能将分类器用于两个子集合，不然的话，两个子集合的属性可能互相影响，从而两个分类器就可能存在一定程度的相似性，从而，会对算法的精确程度造成一定程序的影响
    - 两个分类器对于同一类别的数据进行预测，需要满足大部分数据预测值都相同，如果，两个分类器对于已知类别的数据预测不同的话，这说明两个分类器的训练效果并不是很理想，需要对数据集进行调整