读书报告

09017244

郑健雄

1. 自己提出的问题的理解：
2. 提出的问题1：3.2.4节中，处理决策树额外问题中的Handling Skewed Class Distribution的第二种方法Another solution is to rank the new cases according to how likely they may be intrusions.要如何理解？如何对新的样例进行排序？排序后又对调整模型有什么好处？

讨论后的理解：此处给出的方法实际上并不是使用决策树模型，而是说使用别的模型直接给出一些最可能是入侵的样本点，然后进行人工选择的方法。因为即使由于入侵样本点占比很小导致概率很低，这些样本点本身还是存在概率，也可以进行排序，这样的话，人工只需要从很少的点中进行筛选，所以也是一种可行的方法。

1. 提出的问题2：比起单纯看模型的accuracy来说，计算recall和precision可以提供哪些额外的信息来进一步评估模型？

讨论后的理解：accuracy实际上并不是一个很全面的评估标准，其并不能完全反应模型的完整性和准确性，比如说网络入侵模型，其准确度很高，也有很高的precision，但是其recall很低，可能接近于0，说明对于入侵样本，其分类全部误分，正确分类数非常小，分类不完整。新的指标可以表现出模型更具体的不足之处，帮助人们进行改进。

1. 别人提出的问题的理解：
2. 问题3：对于决策树生成算法而言，如何计算sorting of a continuous attribute takes |D|log|D|的时间复杂度。

自己的理解：对于一个连续属性来说，其要进行二分操作，而对于每个样本点，二分操作的前提是这个连续属性值有序，所以在进行分类之前，对于每一个连续attribute，每一个样本点都需要进行排序，而对于D个样本点，最快的排序复杂度也是DlogD。

1. 问题4：以cross-validation为例，按理来说k次交叉验证后应得到的k个模型分别有不同的accuracy，那取这k个accuracy的平均值有何意义？它并不能代表k个模型中的任意一个。

自己的理解：这个均值accuaracy并不是用来寻找模型的，它的价值在于对学习方法或者模型进行一个评估，通过建立多个模型来进行评估，可以看出这个方法对于训练该数据集是否有效，是不是一个合适的方法。

1. 读书计划

1、本周完成的内容章节：第五章

2、下周计划：整理回顾web data mining，继续阅读统计学习方法

四、读书摘要及理解

1、读书摘要及理解

本周讨论内容:

第三章：

1. 监督学习基本概念：第三章的主要内容是讲监督学习，监督学习的目标是根据已有的数据产生一个分类器，使其可以对未来的数据进行分类。监督学习的一般流程是将数据集分为训练集和测试集，使用机器学习算法通过训练集构筑模型，再通过测试集对其准确性进行评估。

可以认为对于给定的数据集D,给定的工作T以及给定的测试标准M,如果计算机系统通过学习数据集D,其工作T在测试标准M下有所改善，则称之为进行了学习。

2. 决策树模型：决策树模型一般会寻找能够让分类结果更为纯净或者说错误情况最小的分类属性进行分类，使得子节点的不确定程度尽可能下降，在决策树模型中使用信息增益的概念来理解。决策树模型的构建是一个递归的过程。

一般我们使用C4.5算法进行决策树构建，其引入了信息论中的熵概念来表征不确定度或者说不纯粹度。具体寻找分类属性使用了信息增益或者信息增益比来计算分类前后的信息增加量。可以说Impurity Function在决策树模型中是最重要的部分，它直接决定了决策树模型的效率和准确度。

决策树模型一般处理的是属性值分散的属性，对于连续数字属性，其可以使用二分或者区间的方法将其作为一种离散属性来操作。

同时，决策树模型本身也存在过拟合问题以及数据残缺问题，书中也给出了处理的方法。

3 分类器的评价方法：在拥有一个分类器模型之后，我们需要对其进行评价，此处给出了一系列评价准则和评价参数。

评价方法给出了三种，分别是：Holdout Set，Multiple Random Sampling，Cross-Validation。Holdout Set是常用的方法，将数据集分为训练集，测试集并且有的时候再添加一个验证集，它主要用在大数据集上。而对于数据量不充足的情况，一种就是Multiple Random Sampling，多次随机采样，这样的好处是可以挖掘出更多的信息，减少随机性。另外一种方法就是Cross-Validation，交叉验证和多次随机采样也很类似，也是减少结果的随机性，尽可能地挖掘出数据背后隐藏的信息。后两种方法虽然有效，但是也是不得已而为之的一种思路。

分类参数给出了Precision, Recall, F-score and Breakeven Point四种，precision表达了在所有分类为正的数据中正确分类数目的比例，而recall在实际正类数据中正确分类数目的比例，一般而言，p和r之间有一种负相关关系，而一个好的模型，p和r应该都比较高。F-score and Breakeven Point是对上面两种参数的延申。这两个参数可以帮助我们找到均衡点。

阅读内容:

第五章:

1. 半监督学习本质上还是一种监督学习，其未标注数据的意义在于帮助形成一个比较好的监督学习模型。其目标是使用未标注数据来改善监督学习。半监督学习的提出考虑到人工标注一些数据或者文本非常耗时间，所以能否提出一种方法来减少人力，而直接利用原始数据。半监督学习的一个要求是即使标注数据很少，但是每个类的数据至少有一个，这个方法类似于K-means方法中的种子的概念。

2. EM算法:EM算法的思路分为两步，E步为根据现有数据，填充缺失的数据，M步则根据加入新数据的数据集，再一次更新模型的参数，使得其概率最大，其运用了贝叶斯概念的思想，和朴素贝叶斯实际上十分类似，其不同在于需要填充数据，否则，如果数据集完整，其实一次就可以得出模型参数，而不是局部最优解。

3. Co-training: Co-traning代表同时训练两个分类器，其假设是数据集中的属性是存在冗余信息的，对于一个数据集来说，其属性集合可能拥有两个可以进行分类预测的子集，所以可以分别训练模型。那么一般情况下，就会根据两个模型，而且理想情况下，它们之间的数据是不相关的，并且预测效果是差不多的，这样就可以通过两个模型来给无标签数据贴标签，加入训练集，再训练从而改善模型的效果。其和EM算法类似，也是不断加入数据，不断重复训练，但是其效果很好。

此外，还有自学习算法，转移向量机算法和图算法也可以用来进行监督学习。