各种分类算法的优缺点 - 学习笔记1.0 - 经管之家(原人大经济论坛)

各种分类算法比较

最近在学习分类算法，顺便整理了各种分类算法的优缺点。

1决策树（Decision Trees）的优缺点

决策树的优点：

一、 决策树易于理解和解释.人们在通过解释后都有能力去理解决策树所表达的意义。

二、 对于决策树，数据的准备往往是简单或者是不必要的.其他的技术往往要求先把数据一般化，比如去掉多余的或者空白的属性。

三、 能够同时处理数据型和常规型属性。其他的技术往往要求数据属性的单一。

四、 决策树是一个白盒模型。如果给定一个观察的模型，那么根据所产生的决策树很容易推出相应的逻辑表达式。

五、 易于通过静态测试来对模型进行评测。表示有可能测量该模型的可信度。

六、 在相对短的时间内能够对大型数据源做出可行且效果良好的结果。

七、 可以对有许多属性的数据集构造决策树。

八、 决策树可很好地扩展到大型数据库中，同时它的大小独立于数据库的大小。

决策树的缺点：

一、 对于那些各类别样本数量不一致的数据，在决策树当中,信息增益的结果偏向于那些具有更多数值的特征。

二、 决策树处理缺失数据时的困难。

三、 过度拟合问题的出现。

四、 忽略数据集中属性之间的相关性。

2 人工神经网络的优缺点

人工神经网络的优点：分类的准确度高,并行分布处理能力强,分布存储及学习能力强，对噪声神经有较强的鲁棒性和容错能力，能充分逼近复杂的非线性关系，具备联想记忆的功能等。

人工神经网络的缺点：神经网络需要大量的参数，如网络拓扑结构、权值和阈值的初始值；不能观察之间的学习过程，输出结果难以解释，会影响到结果的可信度和可接受程度；学习时间过长,甚至可能达不到学习的目的。

3 遗传算法的优缺点

遗传算法的优点：

一、 与问题领域无关切快速随机的搜索能力。

二、 搜索从群体出发，具有潜在的并行性，可以进行多个个体的同时比较，鲁棒性好。

三、 搜索使用评价函数启发，过程简单。

四、 使用概率机制进行迭代，具有随机性。

五、 具有可扩展性，容易与其他算法结合。

遗传算法的缺点：

一、 遗传算法的编程实现比较复杂,首先需要对问题进行编码,找到最优解之后还需要对问题进行解码,

二、 另外三个算子的实现也有许多参数,如交叉率和变异率,并且这些参数的选择严重影响解的品质,而目前这些参数的选择大部分是依靠经验.没有能够及时利用网络的反馈信息,故算法的搜索速度比较慢，要得要较精确的解需要较多的训练时间。

三、 算法对初始种群的选择有一定的依赖性，能够结合一些启发算法进行改进。

4 KNN算法(K-Nearest Neighbour) 的优缺点

KNN算法的优点：

一、 简单、有效。

二、 重新训练的代价较低（类别体系的变化和训练集的变化，在Web环境和电子商务应用中是很常见的）。

三、 计算时间和空间线性于训练集的规模（在一些场合不算太大）。

四、 由于KNN方法主要靠周围有限的邻近的样本，而不是靠判别类域的方法来确定所属类别的，因此对于类域的交叉或重叠较多的待分样本集来说，KNN方法较其他方法更为适合。

五、 该算法比较适用于样本容量比较大的类域的自动分类，而那些样本容量较小的类域采用这种算法比较容易产生误分。

KNN算法缺点：

一、 KNN算法是懒散学习方法（lazy learning,基本上不学习），一些积极学习的算法要快很多。

二、 类别评分不是规格化的（不像概率评分）。

三、 输出的可解释性不强，例如决策树的可解释性较强。

四、 该算法在分类时有个主要的不足是，当样本不平衡时，如一个类的样本容量很大，而其他类样本容量很小时，有可能导致当输入一个新样本时，该样本的K个邻居中大容量类的样本占多数。该算法只计算“最近的”邻居样本，某一类的样本数量很大，那么或者这类样本并不接近目标样本，或者这类样本很靠近目标样本。无论怎样，数量并不能影响运行结果。可以采用权值的方法（和该样本距离小的邻居权值大）来改进。

五、 计算量较大。目前常用的解决方法是事先对已知样本点进行剪辑，事先去除对分类作用不大的样本。

5 支持向量机（SVM）的优缺点

SVM的优点：

一、 可以解决小样本情况下的机器学习问题。

二、 可以提高泛化性能。

三、 可以解决高维问题。

四、 可以解决非线性问题。

五、 可以避免神经网络结构选择和局部极小点问题。

SVM的缺点：

一、 对缺失数据敏感。

二、 对非线性问题没有通用解决方案，必须谨慎选择Kernelfunction来处理。

6 朴素贝叶斯的优缺点

优点：

一、 朴素贝叶斯模型发源于古典数学理论，有着坚实的数学基础，以及稳定的分类效率。

二、 NBC模型所需估计的参数很少，对缺失数据不太敏感，算法也比较简单。

缺点：

一、 理论上，NBC模型与其他分类方法相比具有最小的误差率。但是实际上并非总是如此，这是因为NBC模型假设属性之间相互独立，这个假设在实际应用中往往是不成立的（可以考虑用聚类算法先将相关性较大的属性聚类），这给NBC模型的正确分类带来了一定影响。在属性个数比较多或者属性之间相关性较大时，NBC模型的分类效率比不上决策树模型。而在属性相关性较小时，NBC模型的性能最为良好。

二、 需要知道先验概率。

三、 分类决策存在错误率

7 Adaboosting方法的优点

一、 adaboost是一种有很高精度的分类器。

二、 可以使用各种方法构建子分类器，Adaboost算法提供的是框架。

三、 当使用简单分类器时，计算出的结果是可以理解的。而且弱分类器构造极其简单。

四、 简单，不用做特征筛选。

五、 不用担心overfitting。

8 Rocchio的优点

Rocchio算法的突出优点是容易实现，计算（训练和分类）特别简单，它通常用来实现衡量分类系统性能的基准系统，而实用的分类系统很少采用这种算法解决具体的分类问题。

9各种分类算法比较

根据这篇论文所得出的结论,

Calibrated boosted trees的性能最好，随机森林第二，uncalibrated bagged trees第三,calibratedSVMs第四，uncalibrated neural nets第五。

性能较差的是朴素贝叶斯，决策树。

有些算法在特定的数据集下表现较好。

参考文献：

[1] 罗森林, 马俊, 潘丽敏.数据挖掘理论与技术[M].电子工业出版社.2013.126-126

[2] 杨晓帆,陈廷槐.人工神经网络固有的优点和缺点[J].计算机科学.1994(vol.21).23-26

[3] Steve.遗传算法的优缺点.http://blog.sina.com.cn/s/blog_6377a3100100h1mj.html

[4] 杨建武.文本自动分类技术.

www.icst.pku.edu.cn/course/mining/12-13spring/TextMining04-%E5%88%86%E7%B1%BB.pdf

[5] 白云球工作室. SVM(支持向量机)综述.http://blog.sina.com.cn/s/blog_52574bc10100cnov.html

[6] 张夏天. 统计学习理论和SVM的不足（1）.http://blog.sciencenet.cn/blog-230547-248821.html

[7] RichCaruana，AlexandruNiculescu-Mizil.An Empirical Comparison of Supervised LearningAlgorithms.2006