决策树等机器学习算法研究

摘　要：无

Research on Machine Learning Algorithms such as Decision Tree

Abstract：None

决策论中，决策树（Decision tree）由一个决策图和可能的结果组成，用来创建到达目标的规划。决策树建立并用来辅助决策，是一种特殊的树结构。决策树是一个利用像树一样的图形或决策模型的决策支持工具，包括随机事件结果，资源代价和实用性。它是一个算法显示的方法。决策树经常在运筹学中使用，特别是在决策分析中，它帮助确定一个能最可能达到目标的策略。如果在实际中，决策不得不在没有完备知识的情况下被在线采用，一个决策树应该平行概率模型作为最佳的选择模型或在线选择模型算法。决策树的另一个使用是作为计算条件概率的描述性手段。

机器学习中，决策树是一个预测模型；他代表的是对象属性与对象值之间的一种映射关系。树中每个节点表示某个对象，而每个分叉路径则代表某个可能的属性值，而每个叶节点则对应从根节点到该叶节点所经历的路径所表示的对象的值。决策树仅有单一输出，若欲有复数输出，可以建立独立的决策树以处理不同输出。数据挖掘中决策树是一种经常要用到的技术，可以用于分析数据，同样也可以用来作预测。

从数据产生决策树的机器学习技术叫做决策树学习,通俗说就是决策树。

一个决策树包含三种类型的节点：

1. 决策节点：通常用矩形框来表示
2. 机会节点：通常用圆圈来表示
3. 终结点：通常用三角形来表示

决策树学习也是数据挖掘中一个普通的方法。在这里，每个决策树都表述了一种树型结构，它由它的分支来对该类型的对象依靠属性进行分类。每个决策树可以依靠对源数据库的分割进行数据测试。这个过程可以递归式的对树进行修剪。 当不能再进行分割或一个单独的类可以被应用于某一分支时，递归过程就完成了。另外，随机森林分类器将许多决策树结合起来以提升分类的正确率。

决策树同时也可以依靠计算条件概率来构造。

决策树如果依靠数学的计算方法可以取得更加理想的效果。 数据库已如下所示：

(x, y) = (x1, x2, x3…, xk, y)

相关的变量 Y 表示我们尝试去理解，分类或者更一般化的结果。其他的变量 x1, x2, x3 等则是帮助我们达到目的的变量。决策树的有几种常用的生成方式：

1. 分类树分析是当预计结果可能为离散类型使用的概念。
2. 回归树分析是当局域结果可能为实数（例如房价，患者住院时间等）使用的概念。
3. CART 分析是结合了上述二者的一个概念。CART 是 Classification And Regression Trees 的缩写.
4. CHAID （Chi-Square Automatic Interaction Detector）

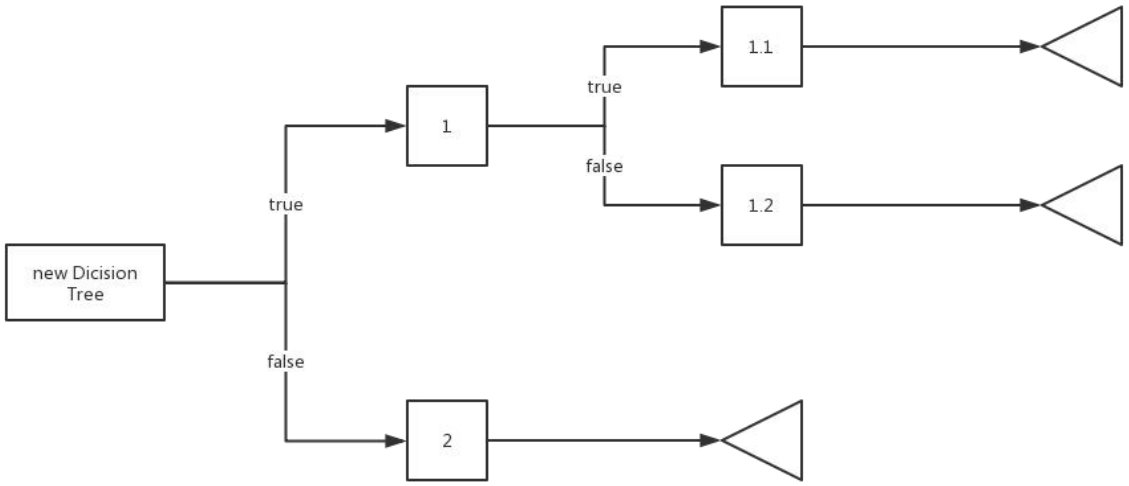

图 1 决策树示意图

在问题一中，我们用的是 CART 的方法生成的决策树，三个数据集（BCW，PID，VLC）中各自挑选前 10%的数据作为测试集，剩下的作为训练集进行训练，然后用训练结束的模型对预测集进行预测，然后我们可以把预测信息通过以下几个指标数值化：

指标	含义
True	预测正确
False	预测错误
TP	真正例
FN	假反例
FP	假正例
TN	真反例
P	查准率
R	查全率
AUC	Area under the Curve of ROC

表 1 问题一名词解释

训练结果如下：

指标	数值
True	92.7536%
False	7.2464%
TP	50.7246%
FN	4.3478%
FP	2.8986%
TN	42.0290%
P	94.5946%
R	92.1052%
AUC	0.8787

表 2 BCW

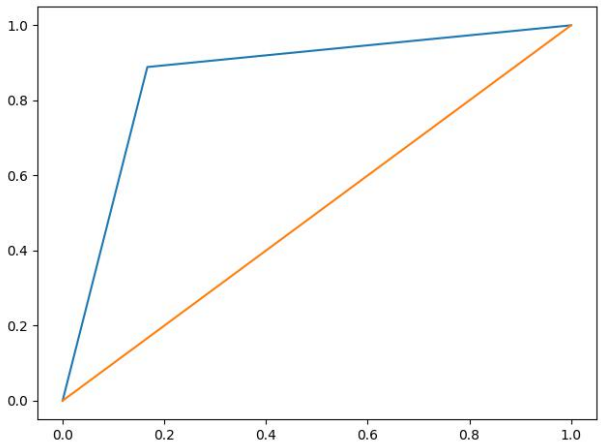

图 5 BCW 的 ROC 曲线

指标	数值
True	69.2308%
False	30.7692%
TP	53.8462%
FN	7.6923%
FP	23.0769%
TN	15.3846%
P	70.0000%
R	87.5000%
AUC	0.6833

表 3 VLC

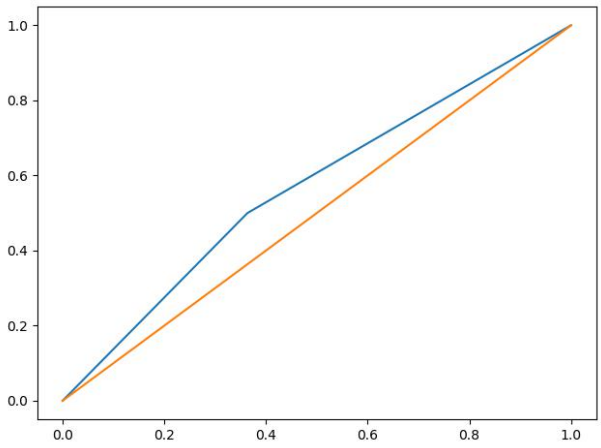

图 6 VLC 的 ROC 曲线

指标	数值
True	68.4211%
False	31.5789%
TP	23.6842%
FN	14.4737%
FP	17.1053%
TN	44.7368%
P	58.0645%
R	62.0690%
AUC	0.6870

表 4 PID

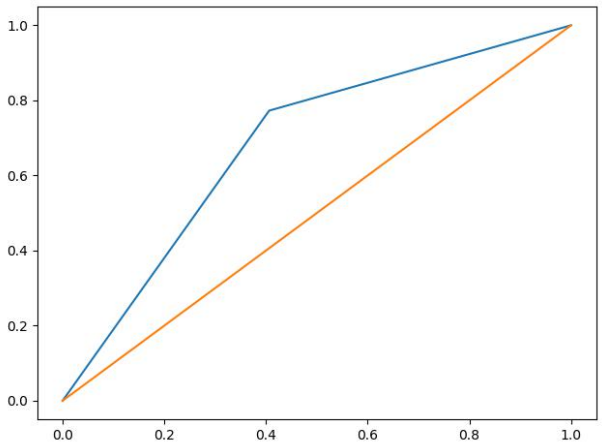

图 7 PID 的 ROC 曲线

通过训练我们发现，预测结果的正确率会随着训练次数的增加而小幅度增加，最终达到一个稳定值。

然后我们开始解决问题二，对于问题二，我们采用决策树，神经网络和支持向量机队数据集 wine 进行分类预测。

数据集中的 Label 有 3 种，也就是这与问题一中的二分类问题不同，是一个三分类问题，因为没有数据说明，所以我们默认 Label 为 1 的葡萄酒为最优，Label 为 3 的葡萄酒为最差，Label 为 2 的则为中等。

因为是三分类问题，所以没办法继续使用问题一中的二分类数值指标，所以我们为了数值化分类的结果，需要把二分类问题，拓展成三分类问题。

为了方便理解，我们对以下指标命名仅采用最简单的方式。

指标	含义
True	预测正确
False	预测错误
11	实际为 1 预测结果为 1
12	实际为 1 预测结果为 2
13	实际为 1 预测结果为 3
21	实际为 2 预测结果为 1
22	实际为 2 预测结果为 2
23	实际为 2 预测结果为 3
31	实际为 3 预测结果为 1
32	实际为 3 预测结果为 2
33	实际为 3 预测结果为 2

表 4 问题二名词解释

然后，我们分别采用三种模型对该数据集进行训练，我们从数据集中选取前 10%的数据作为测试集，剩下的数据作为训练集进行训练。

通过训练我们发现，若选用前 10%的数据作为测试集的话，三种模型的预测正确率均达到了惊人的 100%，我们逐步增加测试集，减少训练集，可以得到以下结果：

训练集比例	DT	SVM	CNN
10%	100%	100%	100%
15%	91.67%	100%	100%
20%	96.86%	100%	100%
25%	100%	100%	100%
30%	97.92%	95.83%	100%
35%	98.21%	94.64%	98.21%
40%	96.88%	95.31%	96.88%
45%	95.83%	95.83%	98.61%
50%	97.50%	96.25%	97.50%
55%	90.91%	92.05%	97.72%

表 5 调参结果

综上，我们可以发现，CNN 的训练情况最好 ，拟合度最高。

通过分析数据我们发现，在只训练最后几组数据的情况下，该模型就能够基本拟合，从而导致预测概率接近 100%。