
数据挖掘-决策树C5.0（ Decision Tree ） ， 示例

作者：writer 来源：新浪博客

本文原地址：<https://www.iikx.com/news/statistics/818.html>

本文仅供学习交流之用，版权归原作者所有，请勿用于商业用途！

数据挖掘-决策树C5.0(Decision Tree)

是否参与

文件

生成

预览

?

模型

查看器

汇总

设置

注解

1

2

所有

%

i

家长不鼓励节点中有30个样本，其中不参与的28个，占92.3%（此规则的置信度）

家长是否鼓励 = Not Encouraged [模式: No] ⇒ No (30; 0.933)

家长是否鼓励 = Encouraged [模式: Yes] (39)

在校综合评价指数 ≤ 48.052 [模式: No] ⇒ No (15; 0.8)

在校综合评价指数 > 48.052 [模式: Yes] ⇒ Yes (24; 0.833)

根据MDLP的熵分组结果，及分组的值为48.052；

家庭人均年收入、是否参加过无偿献血没有进入决策树，对是否参与社会公益活动的影响很小。

确定

取消

应用

重置

2 / 8

比较 \$C-是否参与 与 是否参与

'Partition'	1_Training		2_Testing	
正确	57	82.61%	24	77.42%
错误	12	17.39%	7	22.58%
总计	69		31	

1_Training: 训练样本, 共69个, 正确率为 57/69=82.61%

2_Testing: 测试样本, 共31个, 正确率为 24/31=77.42%

\$C-是否参与 的重合矩阵 (行表示实际值)

'Partition' = 1_Training	No	Yes
No	37	7
Yes	5	20
'Partition' = 2_Testing	No	Yes
No	19	4
Yes	3	5

\$CC-是否参与 的置信度值报告

'Partition' = 1_Training	
范围	0.765 - 0.906
平均正确性	0.849
平均不正确性	0.813
正确性始终高于	0.906 (观测值的 0%)
不正确性始终低于	0.765 (观测值的 0%)
93.33% 以上的准确性	0.808
2.0 以上的折叠正确性	0.933 (观测值的 80.76%)
'Partition' = 2_Testing	
范围	0.765 - 0.906
平均正确性	0.874
平均不正确性	0.83
正确性始终高于	0.906 (观测值的 0%)
不正确性始终低于	0.765 (观测值的 0%)
90% 以上的准确性	从未达到需求等级
2.0 以上的折叠正确性	0.895 (观测值的 80.76%)

范围: 全部样本100个, 置信度的最低值与最高值之间;

预测正确的样本, 其置信度的 平均值 0.849; 错误预测的为0.813

置信度高于0.906且预测正确的样本为0;

置信度低于0.765且预测错误的样本为0;

置信度大于0.808的样本中, 有93.33%样本预测正确。

训练样本中大于0.808的样本总数为30

个, 其中有两个预测错误, 那么预测

正确率为 28/30=93.33%

置信度大于0.895的样本, 预测正确率是全部样本的2倍。 80.76%为: 置信度>0.895的样本占总样本的比率。

是否参与



 文件
  生成
  视图
  预览
 

模型

查看器

汇总

设置

注解

1

2

所有





家长是否鼓励 = Not Encouraged [模式: No] ⇒ No (30; 0.933)

家长是否鼓励 = Encouraged [模式: Yes] (39)

在校综合评价指数 ≤ 48.052 [模式: No] ⇒ No (15; 0.8)

在校综合评价指数 > 48.052 [模式: Yes] ⇒ Yes (24; 0.833)

是否参与

模型 汇总 设置 注解

1 2 3 所有

规则用于 Yes - 包含 1 个规则

- 规则 1 用于 Yes (24; 0.808)
 - 如果 在校综合评价指数 > 48.052
 - 并且 家长是否鼓励 = Encouraged
 - 则 Yes

规则用于 No - 包含 2 个规则

- 规则 1 用于 No (30; 0.906)
 - 如果 家长是否鼓励 = Not Encouraged
 - 则 No
- 规则 2 用于 No (29; 0.839)
 - 如果 在校综合评价指数 <= 48.052
 - 则 No

默认: No

计算器

标准

0.83333333333333333333

1、预测结果是 "Yes" 的规则:

规则1为: 在校综合评价指数 > 48.052 and 家长是否鼓励 = "Encouraged" and Partition = "1_Training"

共24个样本, 其中正确的样本为20个; 正确率为83.33%; 置信度的平均值为0.808。

2、预测结果为 "No" 的规则:

规则1为: 家长是否鼓励 = "Not Encouraged" and Partition = "1_Training"

共30个样本, 其中正确的样本为28个; 正确率为93.33%; 置信度的平均值为0.906

Boosting技术规则1

文件 生成 视图

模型 查看器 汇总 设置 注解

1 2 所有

家长是否鼓励 = Not Encouraged [模式: No] => No (30; 0.933)

家长是否鼓励 = Encouraged [模式: Yes] (39)

在校综合评价指数 <= 48.052 [模式: No] => No (15; 0.8)

在校综合评价指数 > 48.052 [模式: Yes] => Yes (24; 0.833)

使用Boosting技术后, 生成了5个规则, 这是规则1

从决策树上直接获得的推理规则

是否参与

模型 查看器 汇总 设置 注解

1 2 3 所有

规则用于 Yes - 包含 1 个规则

- 规则 1 用于 Yes (24; 0.808)
 - 如果 在校综合评价指数 > 48.052
 - 并且 家长是否鼓励 = Encouraged
 - 则 Yes

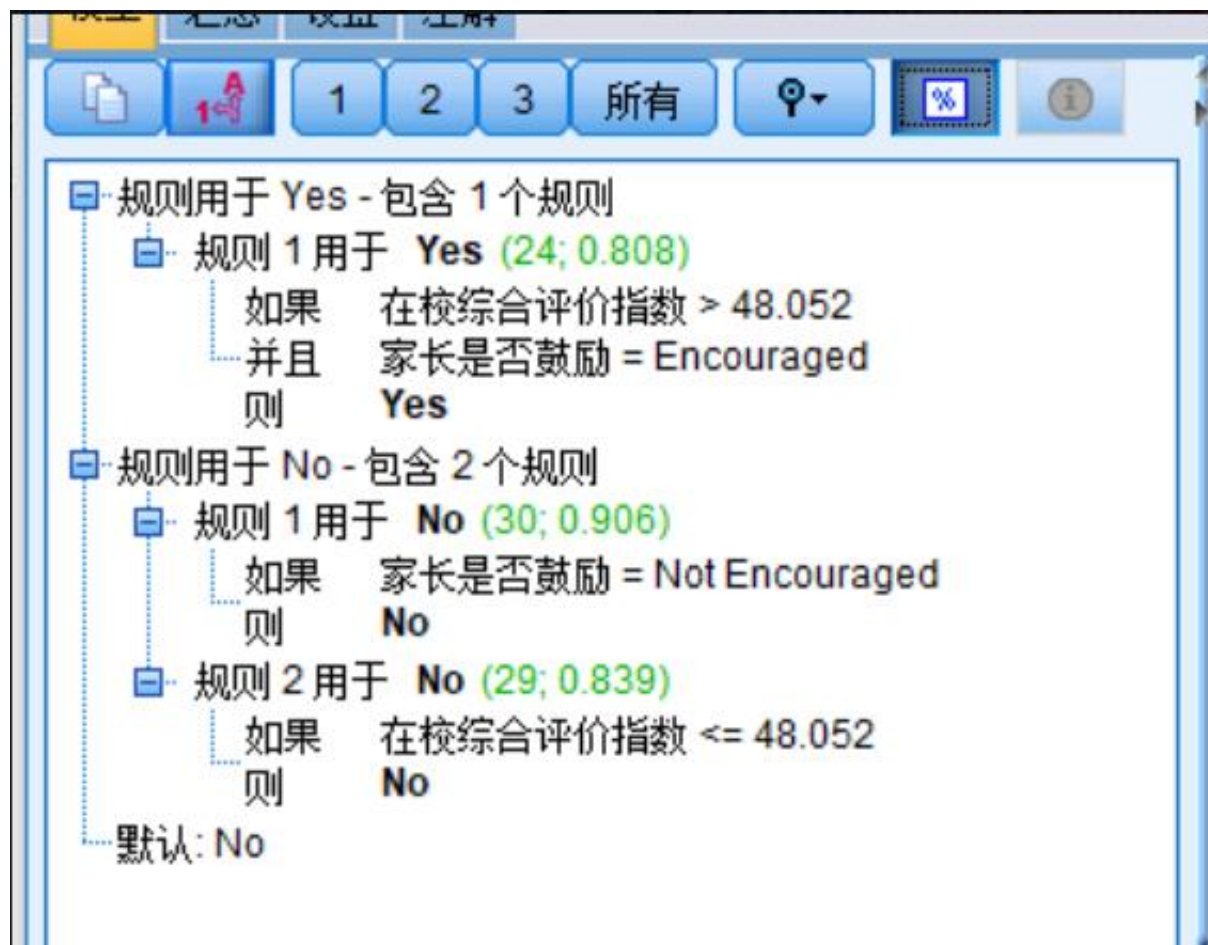
规则用于 No - 包含 2 个规则

- 规则 1 用于 No (30; 0.906)
 - 如果 家长是否鼓励 = Not Encouraged
 - 则 No
- 规则 2 用于 No (29; 0.839)
 - 如果 在校综合评价指数 <= 48.052
 - 则 No

默认: No

C5.0直接生成规则

根据PRISM算法自动生成的推理规则



是否参与

文件

生成

预览

?

模型

查看器

汇总

设置

注解

1

2

所有

%

92.3% (此规则的置信度)

家长是否鼓励 = Not Encouraged [模式: No] $\Rightarrow$ No (30; 0.933)

家长是否鼓励 = Encouraged [模式: Yes] (39)

在校综合评价指数 ≤ 48.052 [模式: No] $\Rightarrow$ No (15; 0.8)

在校综合评价指数 > 48.052 [模式: Yes] $\Rightarrow$ Yes (24; 0.833)

根据MDLP的熵分组结果，及分组的值为48.052；

家庭人均年收入、是否参加过无偿献血没有进入决策树，对是否参与社会公益活动的影响很小。

确定

取消

应用

重置

家长不鼓励节点中有30个样本，其中不参与的28个，占92.3% (此规则的置信度)

6/8

Table (11 个字段, 100 条记录)

实际分类										
预测值										
预测置信度 (此规则置信度)										
表	注解	编号	是否无偿献血	家庭人均年收入	在校综合评价指数	家长是否鼓励	是否参与	Partition	SCC 是否参与	SCC 是否参与
1		1.0...	Yes	46580.000	50.649	Not Encouraged	No	1_Training	No	0.906
2		2.0...	Yes	40896.500	77.922	Not Encouraged	No	1_Training	No	0.906
3		3.0...	Yes	63482.000	53.247	Encouraged	Yes	1_Training	Yes	0.808
4		4.0...	No	40454.000	88.312	Not Encouraged	No	2_Testing	No	0.906
5		5.0...	Yes	7333.000	32.468	Not Encouraged	No	1_Training	No	0.906
6		6.0...	No	17617.000	57.143	Not Encouraged	No	1_Training	No	0.906
7		7.0...	Yes	33540.000	63.636	Not Encouraged	No	1_Training	No	0.906
8		8.0...	Yes	48171.000	53.247	Not Encouraged	Yes	1_Training	No	0.906
9		9.0...	Yes	40896.500	23.377	Not Encouraged	No	1_Training	No	0.906
10		10...	Yes	73325.000	76.623	Encouraged	Yes	1_Training	Yes	0.808
11		11...	Yes	33153.000	66.234	Not Encouraged	No	1_Training	No	0.906
12		12...	Yes	10331.000	42.857	Not Encouraged	No	1_Training	No	0.906
13		13...	No	33505.000	58.442	Not Encouraged	Yes	2_Testing	No	0.906
14		14...	No	30052.000	19.481	Encouraged	Yes	1_Training	No	0.765
15		15...	Yes	24579.000	57.143	Not Encouraged	No	2_Testing	No	0.906
16		16...	Yes	37497.000	14.286	Not Encouraged	No	2_Testing	No	0.906
17		17...	Yes	31572.000	48.052	Encouraged	No	1_Training	Yes	0.808
18		18...	No	41070.000	400.000	Not Encouraged	No	1_Training	No	0.906

推理规则的置信度

置信度 = $28/30 = 0.93333$; 检验样本中30个28个预测正确, 所以预测置信度为0.93333; 表中显示的置信度为“普拉斯调整后的预测置信度”: $(28+1) / (30+2) = 0.9625$;

下图就是推理规则, 是直接从决策树获得的, 用文字表述!

家长是否鼓励 = Not Encouraged [模式: No] $\Rightarrow$ No (30; 0.933)
 家长是否鼓励 = Encouraged [模式: Yes] (39)
 在校综合评价指数 ≤ 48.052 [模式: No] $\Rightarrow$ No (15; 0.8)
 在校综合评价指数 > 48.052 [模式: Yes] $\Rightarrow$ Yes (24; 0.833)

家长不鼓励的节点中有30个样本, 其中不参与的28个, 占93.3% (推理规则的置信度)

比较 \$C\$ 是否参与与 是否参与

'Partition'	1_Training		2_Testing	
正确	57	82.61%	24	77.42%
错误	12	17.39%	7	22.58%
总计	69		31	

1_Training: 训练样本, 共69个, 正确率为 57/69=82.61%

2_Testing: 测试样本, 共31个, 正确率为 24/31=77.42%

\$C\$ 是否参与的重合矩阵 (行表示实际值)

'Partition' = 1_Training	No	Yes
No	37	7
Yes	5	20
'Partition' = 2_Testing	No	Yes
No	19	4
Yes	3	5

\$C\$ 是否参与的置信度值报告

'Partition' = 1_Training	
范围	0.765 - 0.906
平均正确性	0.849
平均不正确性	0.813
正确性始终高于	0.906 (观测值的 0%)
不正确性始终低于	0.765 (观测值的 0%)
93.33% 以上的准确性	0.808
2.0 以上的折叠正确性	0.933 (观测值的 80.76%)
'Partition' = 2_Testing	
范围	0.765 - 0.906
平均正确性	0.874
平均不正确性	0.83
正确性始终高于	0.906 (观测值的 0%)
不正确性始终低于	0.765 (观测值的 0%)
90% 以上的准确性	从未达到需求等级
2.0 以上的折叠正确性	0.895 (观测值的 80.76%)

更多 统计方法 请访问 <https://www.iikx.com/news/statistics/>

本文版权归原作者所有，请勿用于商业用途，[爱科学iikx.com](https://www.iikx.com)转发