选课推荐建模文档

1. 问题重述

项目最终的目标是开发一个集推荐选课、查看课表、评价课程等功能的应用。其中推荐选课功能需要建立一个选课模型，以对公共课程进行量化考虑，找到当前条件下，学生最值得选的课程。

1. 名词定义

推荐度：RD（Recommend Degree），最终权衡课程是否值得选的指标；

校区匹配度：MDC（Matching Degree of Campus），开课的校区与学生所在校区是否相同,为布尔值；

时段匹配度：MDS（Matching Degree of Schooltime），上课的时间段（共划分为四个时间段：早课、上午、下午和晚上）与学生预期的时间段的匹配程度, 最期望的值为2，最不期望的值为0，其他为1；

类型匹配度：MDT（Matching Degree of Type），课程类型与学生预期的类型的匹配程度，最期望的值为2，最不期望的值为0，其他为1；

选中率：SR（Selection Rate），某门课程当前的选中率, 50~100%为 **高**选中率；10~49%为**一般**选中率； 0~9%为**低**选中率；

学分：CR（CRedit），某门课程的学分（1~6）；

课容量：TC（Teaching Content），某门课程总的学生人数上限，小于50人为**小**容量，值为1；大于50人小于100人为**中**容量，值为2，大于100人为**大**容量，值为1；

1. 模型建立

3.1决策树模型介绍

决策树模型就是为了求出一系列规则，按照规则划分数据，得到预测结果。可以把决策树看做有决策块和终止块组成；

3.2决策树的优点：

1）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；

2）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；

3）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。

4）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；

5）易于通过静态测试对模型进行评测。表示有可能测量该模型的可信度；

6）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；

7）可以对有许多属性的数据集构造决策树；

8）决策树可很好地扩展到大型数据库，且它的大小独立于数据库的大小。

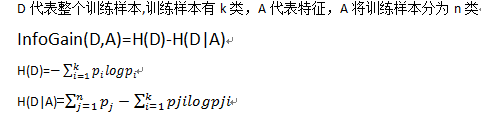
3.3 决策树建模过程：

1）特征选择；

2）决策树生成

3）剪枝

根据信息增益进行特征选择，这就涉及到信息熵的概念，信息熵描述的是数据的混乱程度，信息增益描述的加入特征后数据混乱程度的减小程度，所以信息增益的值等于信息熵减去条件熵的值：



每一步选择信息增益最大的特征作为决策块，最终生成决策树。

3.4 选课推荐建模

通过发放调查问卷，我们统计公选课程各特征与是否选课填入下表：

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ID | MDC  (T、F) | MDS  (0, 1, 2) | MDT  (0, 1, 2) | SR  (高,一般,低) | CR  (1~6) | TC  (小、大、中) | 是否  选择 |
| 1 | T | 2 | 2 | 一般 | 3 | 中 | 是 |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |

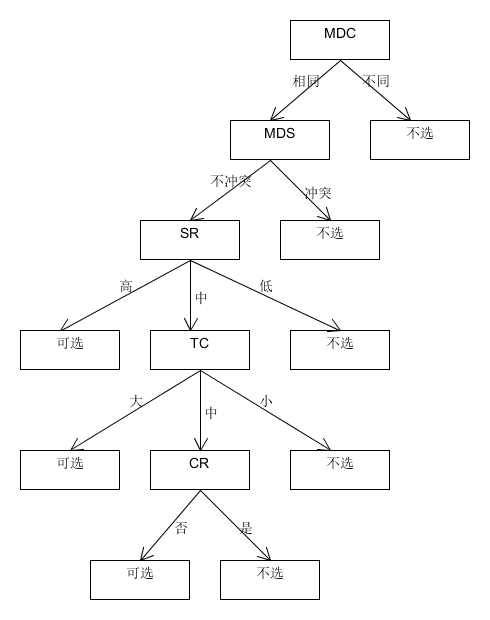
表3-1

由表3-1我们可以计算出各个指标的信息增益值如下：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| MDC | MDS | MDT | SR | CR | TC |
| X1 | X2 | X3 | X4 | X5 | X6 |

表3-2

由此我们可以得到选课推荐的决策树模型：



对于决策树模型判断为不可选的课程，将其排在所有可选课程的后面，推荐度置为0；而对于决策树模型判断为可选的课程，需要进一步衡量推荐度进行排序，我们以归一化后的信息增益值为权重(依次记为λ1，λ2,..)进行线性模拟，得到推荐度的模型：



最后，假设可选课程数为X，则按照推荐度的值由大到小排名，选取前X门课程；