基于加权朴素贝叶斯分类法的成绩预测模型

文 王小丽　远俊红

摘要　目前许多高校已经开始使用数据挖掘技术对教育数据进行深入研究，从而提取有价值的规律和信息来改进和调整教学策略。本文采用加权朴素贝叶斯分类法对学生的考试成绩进行预测和分析，以一种基于互信息量的方法赋予各属性权值，最后通过实验证明本文采用的分类法对于预测成绩具有较好的准确度。

【关键词】朴素贝叶斯 成绩预测 信息熵

1 引言

目前，各大高校由于在校生规模较以往扩大许多，在教学过程中积累了大量的有关学生学习成绩的资源，但传统的方法难以在这些海量的数据中寻找到有价值的信息，教学管理者和决策者们都迫切需要通过更高层次的数据分析来揭示其教学中的规律，从而更好的开展教学工作。于是许多学者开始使用数据挖掘技术去研究这些教育数据中潜藏的知识和信息，为高校教学的提高和改革提供更科学有效地支持。本文建议使用朴素贝叶斯模型对学生成绩进行预测和分析，找出对学生成绩影响较大的因素，将这些处理的结果有效地用于完善教学系统的设计、控制和评价中，从而及时改进和调整教学策略，进而提高高校的教学质量。

2 加权朴素贝叶斯模型

朴素贝叶斯分类（NBC）是数据挖掘中最基础有效的分类算法之一，它有着完整的数学基础和稳定的分类效率。如果有一个数据样本 X 和描述它的 n 个属性，即 $X=\{x_1,x_2,\ldots,x_n\}$，而类别变量 Y 有 m 个属性，即 $Y=\{y_1,y_2,\ldots,y_m\}$，在朴素贝叶斯分类时假设各个样本属性相对于类别条件是独立的，满足公式：

$$P(X\mid y_m)=\prod_{i=1}^{n}P(x_i\mid y_m) \quad (1)$$

其中概率 $P(x_k\mid y_m)$ 可以由收集得来的训练数据集中直接计算得到。

根据贝叶斯定理，数据样本 相对于类别变量的后验概率值为：

$$P(y_m\mid X)=\frac{P(X\mid y_m)P(y_m)}{P(X)} \quad (2)$$

朴素贝叶斯分类法预测 X 应当属于具有最高后验概率的类，也就是说，朴素贝叶斯分类将未知样本分配给类别 y_i，当且仅当：

$$P(y_i\mid X)>P(y_j\mid X),1<j<m,i\neq j \quad (3)$$

朴素贝叶斯分类法与其他算法相比有着最小的误差率，但其前提条件限制比较严格，只有当对象的各个属性之间都相互独立时，使用朴素贝叶斯模型可以得到最佳分类效果，然而在学生成绩预测的几个研究属性之间很难满足这个条件，例如：任课教师的资历很可能对学生的上课兴趣和教师评价有着一定的关联，而学生平时所学习的专业也有可能影响学生对电脑的熟悉度等。为了弥补这个缺陷，有学者提出了加权朴素贝叶斯分类模型（WNBC），其主要思想是为每个属性赋予不同的权值，从而使得朴素贝叶斯方法得以扩展，降低算法对属性独立性的要求，同时也有利于提高分类的效率。加权朴素贝叶斯分类模型被定义为：

$$V_{wnb}(X)=\arg\max_{y}P(y)\prod_{i=1}^{n}p(x_i\mid y)^{w_i} \quad (4)$$

其中 w_i 代表属性 x_i 的权值。显然如果权值越大，该属性对分类的影响就越大，因此加权朴素贝叶斯分类的关键问题就在于如何确定不同属性的权值。

3 确定属性权值

本文使用的是一种基于互信息量的分类

<< 上接 86 页

%>

综上所述，本系统本着操作简便、界面简单而友好。系统整体主要从整个教育系统的教师专业发展规划为立足点，通过从教师本人的个人发展规划需求为基本单元，进一步整合，然后根据教育系统领导、教研员、镇街、片区等各个方面的需求为线，把整个教师的发展规划与整个教育区的对人才的需求、对继续教育的管理等系统的发展规划有机地结合起来，形成一套较完整的教师专业发展规划的管理系统。系统采用更为灵活的 B/S 结构，采用网络操作形式，使得用户可以从不同的浏览平台、不同的操作平台来进行操作。

参考文献

[1] 刘宇．管理信息系统 [M]．北京：北京大学出版社．2009

[2] 求是科技． ASP 信息管理系统开发实例导航 [M]．北京：人民邮电出版社．2005

[3] 杨伟民． 发展规划的理论和实践 [M]．北京：清华大学出版社． 2010

[4] 戴维·乔纳斯．学会用技术解决问题－一个构建主义的视角 [M]．第二版．北京：教育科学出版社．2007、2: 28

作者单位

广东煤炭地质局　广东省广州市　510170

模型，它能根据信息论中的平均互信息量的值来计算条件属性的权重，充分考虑各个条件属性对类别的影响，从而大大提高分类算法的效率。

对于一个数据样本 $X=\{x_1,x_2,...,x_n\}$ 和类别变量 $Y=\{y_1,y_2,...,y_m\}$，x_i 与 y_i 之间的互信息量，也称为事件信息的概率计算公式为：

$$I(x_i;y_i)=\log\frac{p(x_i\mid y_i)}{p(x_i)}=\log\frac{p(x_i,y_i)}{p(x_i)p(y_i)} \quad (5)$$

如果 $p(x_i)=p(x_i\mid y_i)$ 时，说明事件 x_i 和 y_i 相互独立，互信息量 $I(x_i;y_i)$ 为 0；如果 $p(x_i\mid y_i)>p(x_i)$，那么说明当 y_j 出现时，x_i 出现的几率随之增加，此时互信息量 $I(x_i;y_i)$ 为正数；相反的，如果 $p(x_i\mid y_i)<p(x_i)$，那么意味着当 y_i 出现，x_i 出现的概率也变小，此时互信息量 $I(x_i;y_i)$ 为负数。通过计算互信息量可以用来衡量 x_i 和 y_i 之间的相关性指标。

因此可以计算每个属性 x_i 相对于类别变量的权重：

$$w_i=I(x_i;Y)=\sum_{i=1}^{m}p(x_iy_i)\log\frac{p(x_i\mid y_i)}{p(x_i)} \quad (6)$$

得到权重向量 $(w_1,w_2,...,w_n)$。

在设计学生成绩分析预测模型时，可以先计算这些属性的权重，主要目的是能区分不同属性对决策的不同影响，一般对分类有较大影响的属性，我们赋予它较大的权重；相反地，如果该属性对分类影响较小，则权重值也应当越小，从而可以增强对分类影响大的属性的学习，也有利于提高分类学习的效率。

4 构建学生成绩预测模型

以目前各高校都必须要求学生参加的计算机等级考试为例，我们采集了某高校大一新生的数据，整理出可能影响学生计算机等级考试成绩的因素，包括学生的高中的文理科计算机会考成绩、生源地，入校后就读专业、任课教师的条件、学生对计算机的兴趣和了解等，制成数据调查表，要求学生如实填写信息。收集调查表后进行数据录入和预处理，并根据算法需要对文本数据进行必要的转换，例如将学生学习兴趣度归纳为三个等级，分别是很感兴趣，一般，和没兴趣，形成可适用于算法的数据表如下所示：

为了验证预测模型的效果，我们将数据库中 70% 的数据做为训练集，用于构建分类预测模型，而将剩下的 30% 做为测试集，用于测试模型的准确度。首先可以根据公式（6）计算出每个属性的权值如下：

为了验证方法的准确度和效率，我们将本文所采用的加权朴素贝叶斯分类法和传统贝叶斯分类法做了比较试验，分别在取数据测试集样本量为 500、1000、1500、2500、3000 时得到的准确度如下表所示：

通过两种分类法测试结果的比较，可以看出加权朴素贝叶斯分类法比传统朴素贝叶斯分类法的准确度要略高些，而如果训练集数据量越大，那么加权朴素贝叶斯分类法的优势就更加明显。

5 总结

本文使用加权朴素贝叶斯分类法来对高校中学生的考试成绩做了预测分析，各个属性的权值是由根据信息论观点下的平均互信息量计算得来的，实验表明，本文采用的加权朴素贝叶斯分类法在准确度上比传统朴素贝叶斯方法要更好，特别是在取样的数据集量比较大时，分类的效果越好，可以做为预测和分析学生成绩的一种有效方法。

表 1: 学生基本信息表

学号	文理科	生源地	专业	熟练度	教师评价	教师职称	兴趣度
010101	文科	昆明	会计	高	好	助教	高
010102	文科	楚雄	广告	中	一般	讲师	低
010103	理科	弥勒	财务	中	差	讲师	中

表 2: 属性权值表

文理科	生源地	专业	熟练度	教师评价	教师职称	兴趣度
0.3541	0.5678	0.8415	0.8749	0.7543	0.8124	0.8069

表 3: 准确度比较

数据量	500	1000	1500	2000	2500	3000
WNBC	58.74	62.27	71.25	85.62	90.56	92.14
NBC	61.45	65.47	70.48	83.06	86.15	89.62

参考文献

[1] 孔丽英，数据挖掘在计算机等级考试中的应用 [J]，计算机教育，2010,1,38-41

[2] 数据挖掘导论

[3] （Harry Z, Sheng S. Learning Weighted Naïve Bayes with Accurate Ranking[J]. The Fourth IEEE International Conference on Data Minging, 2004, 341-356）

[4] 李立萍，张明友，信息论导引 [M]，成都：电子科技大学出版社，2005,32-70

[5] 张龙飞，基于互信息的朴素贝叶斯改进模型研究 [D]，长春：吉林大学，2010,27-38

[6] 张震，胡学钢，基于互信息量的分类模型 [J]，计算机应用，2011,6.

作者单位

云南经济管理职业学院　云南省昆明市五华区 650106