贝叶斯决策树方法在招生数据挖掘中的应用

黄春华[1,2],陈忠伟[2],李石君[1]
(1. 武汉大学 计算机学院,湖北 武汉 430072;
2. 广西英华国际职业学院 工信学院,广西 钦州 535000)

摘 要:文中首先简单介绍了贝叶斯决策树方法的基本思想,该方法结合了贝叶斯分类的先验信息方法和决策树分类的信息增益方法的优点,加入贝叶斯节点弥补了决策树不能处理具有二义性或存在缺失值数据的缺点。在此基础上,文中设计了一种基于朴素贝叶斯方法和 ID3 算法的贝叶斯决策树算法——NBDT-ID3 算法,并给出了该算法的设计及分析过程。然后将该算法应用到高职招生数据挖掘中,对新生报到情况进行分析与预测,并在 Matlab 环境下进行了实验验证。实验结果表明,NBDT-ID3 算法在付出一定时间代价的情况下,不仅可以获得更高的分类精度,而且在处理二义性、不完整或不一致数据方面具有更好的效果。

关键词:数据挖掘;贝叶斯决策树;分类;招生数据;报到预测
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2016)04-0114-05
doi:10.3969/j.issn.1673-629X.2016.04.025

Application of Bayesian Decision Tree Method in Admission Data Mining

HUANG Chun-hua[1,2], CHEN Zhong-wei[2], LI Shi-jun[1]
(1. School of Computer, Wuhan University, Wuhan 430072, China;
2. Dept. of Industry and Information, Guangxi Talent International College, Qinzhou 535000, China)

Abstract: It simply introduces the basic thought of Bayesian decision tree method in this paper, which takes advantage of the prior information method for Bayesian classification and the information gain method of decision tree, and makes up for the decision tree cannot handle the ambiguity data and the missing value by adding Bayesian node. On this basis, a Bayesian decision tree algorithm based on Naïve Bayesian method and ID3 algorithm is presented named NBDT-ID3 algorithm. The algorithm process of the design and analysis is introduced. Then the algorithm is applied to higher vocational admission data mining, which analyzes and forecasts the new student registration. It is tested and verified under the Matlab environment. The experimental results show that NBDT-ID3 algorithm not only can get higher classification accuracy but also behave well in handling the ambiguity, incomplete or incongruous data in the case of paying certain of time.

Key words: data mining; Bayesian decision tree; classification; admission data; registration forecasting

0 引 言

招生工作一直是民办高职院校工作的重中之重,因为生源是其生存之本。如何有针对性地开展招生工作,既能提高新生的报到率又能节省招生成本,一直是民办高职院校非常关心的问题之一。数据挖掘技术是通过分析大量不完整的、模糊的、随机的数据来发现隐藏的、潜在有用的知识和规则的过程[1]。学校可以通过结合数据挖掘技术和招生工作经验,对历年招生数据进行分析,从中寻找到有价值的信息,以此指导学校制定合理的招生计划,将有限的人力物力用在能“产出”大量生源的地方,提高新生报到率,达到招生效益最大化。

目前用于招生数据挖掘的方法有关联规则、决策树分类、支持向量机等[2-3],但是每一类方法都有一定的应用局限性。决策树分类算法是以实例为基础的归纳学习算法,通过信息增益来构建决策树,只需要在训练和测试这两个阶段进行简单的比较,对数据类别的要求不高,计算过程简单,主要着眼于从一组给定的无次序、无规则样本数据中推理出以决策树表示的分类规则,结果表现直观[4]。但是该类算法的主要缺点是

收稿日期:2015-07-15 **修回日期**:2015-10-21 **网络出版时间**:2016-03-22
基金项目:中央高校基本科研业务费专项基金项目(20420140057);湖北省自然科学基金项目(2014CFB289)
作者简介:黄春华(1985-),女,硕士,讲师,研究方向为数据挖掘、SQL 数据库技术及应用。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20160322.1521.072.html

对缺失或二义性数据难以产生正确的分支,以致影响整个决策树的生成,从而降低了分类的准确性[4]。针对这个不足之处,可以将贝叶斯分类方法引入决策树学习模型中,前者具有坚实的数学基础且算法具有简单直观、易实现、时空开销小、健壮性小等优点[5]。这样不仅可以更好地处理包含不一致性或不完整等非规律性数据的集合,还可以将先验知识与概率背景融入决策树分类模型中[6]。

目前基于贝叶斯决策树的数据挖掘算法已经得到许多学者的研究并被应用到不同的领域中。尹婷等[7]将基于贝叶斯决策树的方法应用到电信企业客户流失分析与预测中;徐哲等[8]将贝叶斯决策树方法应用到识别英文现在分词的词性中;王琦[9]构建了一种基于贝叶斯决策树算法的垃圾邮件识别机制。

在简单介绍了贝叶斯决策树方法基本思想的基础之上,文中详细给出了一种基于朴素贝叶斯方法和 ID3 算法的贝叶斯决策树分类算法,并根据民办高职院校招生工作及其数据特点,将该算法应用到高职招生数据挖掘中,主要对新生报到情况的分析与预测进行了初步研究。

1　贝叶斯决策树方法

1.1　贝叶斯分类方法

贝叶斯分类方法基于贝叶斯定理,其关键在于使用概率表示各种形式的不确定性,即通过变换事件的先验概率及后验概率,配合决定分类特性的各属性彼此间是相互独立的假设来预测分类的结果[10]。下面以朴素贝叶斯(Naïve Bayesian)分类方法为例,给出一个贝叶斯分类方法的工作过程[11-12]。

(1)设 D 是训练元组和它们相关联的类标号的集合,通常每个元组用一个 k 维属性向量 $X=(x_1,x_2,\cdots,x_k)$ 表示,描述由 k 个属性 $A_1,A_2,\cdots,A_k$ 对元组的 k 个测量。

(2)假定有 l 个类别 $C_1,C_2,\cdots,C_l$,给定元组 X,分类法将预测 X 属于具有最高后验概率的类别(在条件 X 下)。根据贝叶斯定理的公式可得:

$$p(C_i \mid X)=\frac{p(X \mid C_i)p(C_i)}{p(X)} \tag{1}$$

其中:$p(C_i)$ 是先验概率;$p(C_i \mid X)$ 是后验概率。

由此可知,朴素贝叶斯分类法预测 X 属于类别 C_i 当且仅当 $p(C_i \mid X) > p(C_j \mid X)$,其中 $1 \leqslant j \leqslant l$,且 $i \neq j$。

(3)因为 $p(X)$ 对所有类别均为常数,所以只需保证 $p(X \mid C_i)p(C_i)$ 取值最大即可。如果类的先验概率都是未知的,则可以假定这些类是等概率的,即 $p(C_1)=p(C_2)=\cdots=p(C_l)$,此时就只需对 $p(X \mid C_i)$ 最大化,否则对 $p(X \mid C_i)p(C_i)$ 最大化。另外需注意的是,类的先验概率 $p(C_i)$ 可以由 $|C_{i,D}|/|D|$ 估计,其中 $|C_{i,D}|$ 表示 D 中类别 C_i 的训练元组数。

(4)当给定的数据集中具有许多属性时,计算 $p(X \mid C_i)$ 的开销可能会很大,可以通过做类条件独立的朴素假定来降低计算开销。因此有:

$$p(X \mid C_i)=p(x_1 \mid C_i)p(x_2 \mid C_i)\cdots p(x_k \mid C_i)=\prod_{t=1}^{k}p(x_t \mid C_i) \tag{2}$$

x_j 表示元组 X 在属性 A_j 上的值,$p(x_j \mid C_i)$ 可以很容易地由训练元组进行估计。对于每个属性,主要考察该属性是分类的还是连续值的。如果 A_j 是分类属性,则 $p(x_j \mid C_i)$ 是 D 中属性 A_j 的值为 x_j 的类的元组数除以 D 中类别 C_i 的元组数 $|C_{i,D}|$;如果 A_j 是连续值属性,则可假定连续值属性服从高斯分布,即 $p(x_j \mid C_i)=g(x_j,\mu_{C_i},\sigma_{C_i})$,其中 μ_{C_i} 和 σ_{C_i} 分别表示类别 C_i 训练元组属性 A_j 的均值和标准差。

(5)为了预测 X 的类别标号,对每个类别 C_i,计算 $p(X \mid C_i)p(C_i)$。则朴素贝叶斯分类法预测 X 属于类别 C_i 可最终表述为当且仅当 $p(X \mid C_i)p(C_i) > p(X \mid C_j)p(C_j)$,其中 $1 \leqslant j \leqslant l, i \neq j$。根据式(2)可进一步得到:

$$p(C_i)\prod_{t=1}^{k}p(x_t \mid C_i) > p(C_j)\prod_{t=1}^{k}p(x_t \mid C_j) \tag{3}$$

即被预测的类别标号是使 $p(X \mid C_i)p(C_i)$ 最大的类 C_i。

1.2　决策树

决策树(Decision Tree)又称为判定树,是一种以树状结构形式来表达的预测分析模型,是数据挖掘技术中一种重要的分类方法。根据给定的一个类标号未知的实例,可以在决策树上测试该实例的属性值,并跟踪一条由根到叶子节点的路径,则该叶子节点就存放着该实例的类预测。决策树的主要优点是描述简单,分类速度快,特别适合大规模的数据处理[4]。图 1 是一棵决策树。

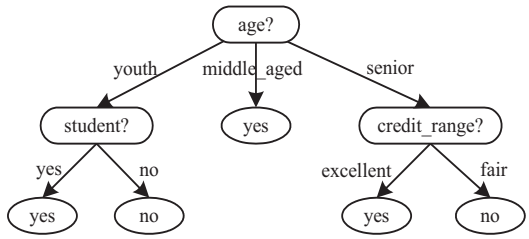

图 1　决策树举例

1.3　贝叶斯决策树方法简介

定义:在原有决策树的两个属性测试节点之间加入一个能够根据贝叶斯原理进行函数计算[13]的新节点,该节点即是贝叶斯节点(Bayesian Node,BN)。相

应地将具有贝叶斯节点的决策树称为贝叶斯决策树(Bayesian Decision Tree,BDT),其结构如图 2 所示。

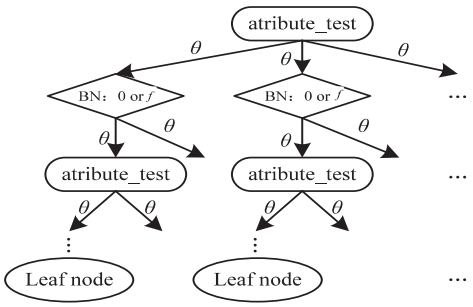

图 2 BDT 的结构

由图 2 可知,BN 包含两个值:0 和 f。当 BN 取值为 0 时,该节点只需根据属性测试条件 θ 直接转向下一个属性测试节点,不必进行任何计算;当 BN 取值为 f 时,该节点需要计算函数 f 的值,并根据属性测试条件 θ 转向下一个属性测试节点,即当 BN 取值为 f 时,下一个属性节点的选择依赖于两点:函数 f 的值和属性测试条件 θ。这里的函数 f 根据具体情况可以是朴素贝叶斯公式也可以是其他贝叶斯公式。

需要说明的一点是,当根据函数 f 和属性测试条件 θ 进行下一属性节点的选择时,都采用 IF……THEN ……的表达形式进行描述[6]。

2 算法的设计及分析

2.1 算法设计思路

根据贝叶斯决策树分类算法的基本思想,以下给出一种基于朴素贝叶斯方法和 ID3 算法的贝叶斯决策树分类算法(NBDT-ID3)的设计思路:

(1)当使用决策树的信息增益方法就可确定选择某个属性的分支时,BN 的取值为 0。其中 ID3 算法信息增益的计算方法[11]如下所述:

设 D 是一个标记类元组的训练集,假设类标号属性具有 m 个不同值,分别定义了 m 个不同的类别 $C_i(i=1,2,\cdots,m)$。设 $C_{i,D}$ 是 D 中类别 C_i 元组的集合,$|D|$ 和 $|C_{i,D}|$ 分别表示 D 和 $C_{i,D}$ 中元组的个数。则对于 D 中的元组分类所需要的期望信息为:

$$\mathrm{Info}(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \tag{4}$$

其中,p_i 为 D 中任意元组属于类别 C_i 的非零概率,用 $|C_{i,D}|/|D|$ 估计。

假设要按某个属性 A 划分 D 中的元组,其中属性 A 根据训练数据的观测值具有 v 个不同值 $\{a_1,a_2,\cdots,a_v\}$。可以用属性 A 将 D 划分为 v 个子集 $\{D_1,D_2,\cdots,D_v\}$,其中 $D_j(j=1,2,\cdots,v)$ 包含 D 中的元组,它们对应于属性 A 的值为 a_j。如果 A 作为测试属性,那么这些子集对应于由 D 的节点生长出来的分枝。基于按属性 A 划分对 D 的元组分类所需要的期望信息为:

$$\mathrm{Info}_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \mathrm{Info}(D_j) \tag{5}$$

其中,$|D_j|/|D|$ 表示第 j 个分区子集的权重。

信息增益定义为原来的信息需求(仅基于类比例)与新的信息需求(对 A 划分后)之间的差值,即:

$$\mathrm{Gain}(A) = \mathrm{Info}(D) - \mathrm{Info}_A(D) \tag{6}$$

(2)当数据分类具有二义性,即数据对象的分类类别无法确定或属性值丢失时,BN 的取值为 f。这里的 f 选择为朴素贝叶斯公式,即根据以前的经验知识或实验结果得出该数据对象的先验概率值,再以此值来判断可以先将其分到某些类中,然后运用贝叶斯分类方法确定这些类的后验概率值,最后选择后验概率值最大的那一类作为该数据对象的所属类别[6]。

2.2 算法流程

根据以上设计思路,给出 NBDT-ID3 算法流程:

输入:数据集 $\{X_1,X_2,\cdots,X_n\}$,其中每个数据 X_i 具有 m 个属性 $x_{ij}(i=1,2,\cdots,n;j=1,2,\cdots,m)$;

输出:显示或打印出对数据集 $\{X_1,X_2,\cdots,X_n\}$ 已划分到各个相关类别 $C_k(k=1,2,\cdots)$ 中的数据。

(1)根据事先给定的类别特征或属性确定要生成的类别集合 $\{C_1,C_2,\cdots,C_l\}$,并确定类别数目 l。

(2)运用 2.1 节中信息增益的计算方法先确定优先判断的属性,然后确定要进行分类的数据 $X_i(i=1,2,\cdots)$ 的某个或某些属性,属性值与相应的类别相关。

(3)当属性选择和数据分类都无二义性时,BN 的取值为 0,直接根据属性测试条件转向下一个属性测试,转到(2),否则转到(4)。

(4)对 X_i 进行分类。若 X_i 确定对应某一类别 C_k,则将 X_i 划分到该类别中;若 X_i 不能确定划分到哪一个类别中,而是与某些类别都可能相关,则根据 1.1 中所述的朴素贝叶斯分类方法计算出最大的 $p(X_i|C_k)p(C_k)$ 值,并将 X_i 划分到相应类别中。

(5)BN 的取值为 f,且 $f=\max(p(X_i|C_k)p(C_k))$,转到(3)。

2.3 算法分析

NBDT-ID3 算法仍然具有与决策树分类算法的产生规则易于理解、分类速度相对较快等相似的优点[6]。该算法主要包括两项工作:判断是否要计算 f 值和判断是否要计算属性的后验概率值。根据上述的算法流程,最坏的情况就是需要计算所有数据的后验概率值。假设共有 n 个数据待分类,且每个数据有 m 个属性,需要把它们划分到 k 个类别中,计算一个数据的后验概率值需要时间 t_1,计算信息增益值需要时间 t_2,此时算法的计算时间为:

$$(t_1 + mt_2) \cdot n \cdot k = nkt_1 + nmt_2 \tag{7}$$

当 $m = n = k$ 时,计算时间为 $n^2t_1 + n^3t_2$,则此时算法的时间复杂度为 $O(n^3)$ 。

NBDT-ID3 算法自身具有的优点如下:

(1)具有更高的分类精度和准确率。分类一般按照数据的某个或某些属性进行,假如根据数据集计算出来的两个不同属性的信息增益值相等,则属性的选择出现了二义性。大量的数据二义性必然会对数据集的分类精度和准确率产生不良影响。而 NBDT-ID3 算法通过引入朴素贝叶斯方法,可很好地利用先验信息去处理这些数据二义性,提高分类的精度和准确率。

(2)具有更强的分类鲁棒性。数据挖掘一般处理的都是海量数据,这些数据由于主客观原因难免会存在大量不完整、不一致和噪声等干扰数据。可以通过预处理的方法[11]对这些干扰数据进行处理,但该解决方法一般较为耗时耗力。NBDT-ID3 算法通过运用朴素贝叶斯方法,可以根据历史数据的先验信息或经验来消除不一致的数据,平滑不完整的数据,排除噪声数据等[6],相对而言省时省力,且具有更好的处理效果,从而增强了数据分类的鲁棒性。

3　NBDT-ID3 算法的应用

3.1　数据准备及预处理

因为该学院的新生来源主要分为高考统招生和三校生两类,其中三校生通过中职对口的招生方式进行录取,招生来源一般是定向的,因此只对高考统招生的数据进行挖掘分析。实验数据来源于该学院 2012-2014 年实际的高考统招生信息。

因为不同年份招生数据表的格式有所差异,存在着相同含义的属性用不同字段名称表示的情况。比如在 2012 年数据表中用“入学成绩”表示高考成绩,在 2013 年数据表中则用“总分”表示高考成绩。为了保证数据挖掘的有效性,必须先将这些属性名称统一表示。经过初步分析,首先删除掉数据集中那些明显与数据挖掘不相关的字段,比如年份、考生姓名、身份证号、联系地址等,初步保留那些可能与招生数据挖掘相关的字段:考生号、性别、考生类别、高考成绩、报考科类、录取专业、录取专业代码和报到情况。

根据高职招生业务及其数据的特点,可以对招生数据做进一步的处理以更有利于数据挖掘工作的进行。依据全国高职高专专业目录中专业代码的含义,可以将录取专业进行泛化处理[11];依据考生号的组成含义,可以得到每位新生的生源地区信息;采用合适的数学方法[3]对高考成绩进行离散化处理,划分出每个考生的成绩等级。最终处理得到的数据如表 1 所示。

表 1　最终处理得到的数据示例

序号	性别	生源地区	考生类别	高考成绩	录取专业分类	报到情况
10001	女	桂南	农村应届	中等	财经大类	是
10002	男	桂南	农村应届	良好	土建大类	是
10003	女	桂东	城镇应届	良好	交通运输大类	是
10004	男	桂东	城镇应届	中等	艺术设计与传媒大类	是
10005	男	桂东	城镇应届	一般	制造大类	否
10006	女	桂北	农村应届	良好	电子信息大类	否
10007	女	桂北	农村应届	一般	旅游大类	是
10008	男	桂中	农村往届	良好	农林渔牧大类	是
10009	男	桂西	城镇应届	良好	财经大类	否
10010	男	区外	农村应届	中等	文化教育大类	否
…	…	…	…	…	…	…

3.2　算法的检验与性能评价

为了验证 NBDT-ID3 算法在高职新生报到预测中的应用性能,在 Matlab 环境下分别运用 ID3 决策树算法和 NBDT-ID3 算法对预处理后的招生数据集进行训练和测试,并对实验结果进行对比说明。预处理后的招生数据集共有 2 625 条新生信息记录,其中报到新生人数 1 782 人,未报到新生人数 843 人。随机抽取其中 2/3 的数据作为训练集建立基于贝叶斯决策树预测模型得到预测结果,再运用该模型对剩余的 1/3 数据进行新生报到情况的预测,然后从覆盖率和命中率两个方面对预测结果和实际结果进行对比分析。

覆盖率:实际报到预测也是报到的新生人数 X 占所有实际报到的新生人数的比例,它是描述模型普适性的指标[7],用 α 表示,其计算公式为:

$$\alpha = \frac{X}{X + Y} \times 100\% \tag{8}$$

其中, Y 为实际报到但预测是未报到的新生人数。

命中率:实际报到预测也是报到的新生人数 X 占所有预测为报到的新生人数的比例,它是描述模型精确度的指标[7],用 β 表示,其计算公式为:

$$\beta = \frac{X}{X + Z} \times 100\% \tag{9}$$

其中, Z 为预测报到但实际并未报到新生人数。

最后得到仅应用 ID3 决策树算法模型与运用基于 NBDT-ID3 算法的贝叶斯决策树模型得到的训练结果和检验结果对比情况,见表 2。

表 2　两种决策树模型训练结果和检验结果对比情况　　%

决策树模型	训练结果		检验结果	
	覆盖率	命中率	覆盖率	命中率
NBDT-ID3 算法决策树	91.68	90.21	83.74	81.33
ID3 算法决策树	87.92	86.08	78.46	73.15

从表 2 的对比结果可以看出,两种决策树模型的训练结果在覆盖率和命中率上都比检验结果的好,但基于 NBDT-ID3 算法的决策树模型比 ID3 决策树算法

模型无论是在训练结果还是检验结果上覆盖率和命中率都高一些,说明前者能获得较好的预测效果。

另外,建模规则和实施分类的时间也会对系统效率和性能产生影响[14],所以有必要对算法的训练时间和分类时间进行验证和比较,以进一步评价算法的性能。同样在 Matlab 环境下,对 NBDT-ID3 算法与 ID3 算法在数据集训练执行过程中所需的训练时间之比和分类时间之比进行验证和比较,结果如图 3 所示。

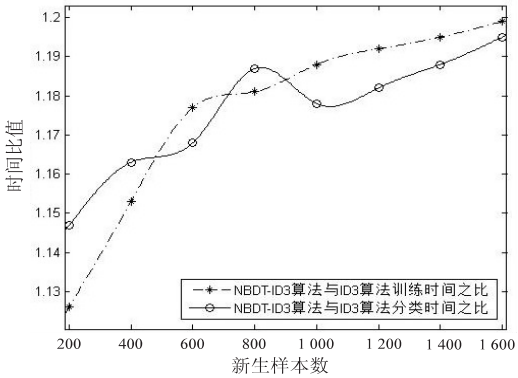

图 3 两种算法训练时间和分类时间对比结果

从图中可以看出,NBDT-ID3 算法的训练时间和分类时间都比 ID3 算法的长。这是因为在构建决策树时 NBDT-ID3 算法需额外插入 BN,在分类时 NBDT-ID3 算法需对选择 BN 值为 f 的节点进行后验概率计算,从而造成了额外的时间开销,但从整体上看,两者的训练时间和分类时间相差不大,时间比值保持在 1.12 ~ 1.2,基本符合理想增长的趋势。

为了验证 NBDT-ID3 算法数据分类的鲁棒性,分别从 UCI 机器学习数据库 Anneal、Balance-scale、Vowel 中随机抽取 3 个数据集进行分类测试,同样在 Matlab 环境下运用 ID3 决策树和 NBDT-ID3 算法对数据集进行分类,比较这两种算法在建树时间之比和分类精度上的情况,结果如表 3 所示。

表 3 两种算法数据分类的鲁棒性检验结果对比情况

数据库名称	样本总数	属性个数	样本缺失率/%	NBDT-ID3 算法与 ID3 算法建树时间之比	分类精度/%	
					NBDT-ID3 算法	ID3 算法
Anneal	1 082	36	56.84	2.58	86.04	84.83
Balance-scale	894	25	22.71	2.26	71.33	71.15
Vowel	661	12	75.04	2.91	66.84	64.76

从表 3 中可以看出,在样本缺失率较高的情况下,NBDT-ID3 算法因为要计算更多选择 BN 值为 f 的节点的后验概率值,所以比 ID3 算法需要更长的建树时间,但在付出时间代价的情况下,NBDT-ID3 算法能较好地提高分类精度。由此说明,在付出一定时间代价的情况下,NBDT-ID3 算法不仅能提高分类精度,而且在处理数据不完整、不一致等缺失样本时具有更强的分类鲁棒性。

4 结束语

根据贝叶斯决策树方法的基本思想,设计了一种基于朴素贝叶斯方法和 ID3 算法的贝叶斯决策树分类算法——NBDT-ID3 算法,并详细给出了该算法的设计及分析过程。然后将该算法应用到高职招生数据挖掘中,对新生报到情况进行预测分析。实验结果表明,NBDT-ID3 算法在付出一定时间代价的情况下,可以获得更好的分类效果,并且对具有二义性、不完整或不一致的数据具有更好的处理效果。如何更加有效地将这种基于贝叶斯决策树的分类方法运用到民办高职院校招生数据的挖掘分析中,更好地为学校招生工作提供科学而直观的决策支持,是接下来需要进一步研究的工作。

参考文献:

[1] 朱志勇,徐长梅,刘志兵,等. 基于贝叶斯网络的客户流失分析研究[J]. 计算机工程与科学,2013,35(3):155-158.

[2] 孙晓莹,郭飞燕. 数据挖掘在高校招生预测中的应用研究[J]. 计算机仿真,2012,29(4):387-391.

[3] 詹柳春. 数据挖掘技术在高校招生录取数据中的应用研究[D]. 广州:华南理工大学,2012.

[4] Quilan J R. Induction of decision tree[J]. Machine Learning,1986,1(1):81-106.

[5] Palacios-Alonso M A,Brizuela C A,Sucar L E. Evolutionary learning of dynamic Naïve Bayesian classifiers[J]. Journal of Automated Reasoning,2010,45(1):21-37.

[6] 樊建聪,张问银,梁永全. 基于贝叶斯方法的决策树分类算法[J]. 计算机应用,2005,25(12):2882-2884.

[7] 尹 婷,马 军,覃锡忠,等. 贝叶斯决策树在客户流失预测中的应用[J]. 计算机工程与应用,2014,50(7):125-128.

[8] 徐 哲,刘 循. 贝叶斯决策树在英文现在分词词性识别中的应用[J]. 计算机应用,2009,29(9):2571-2574.

[9] 王 琦. 基于贝叶斯决策树算法的垃圾邮件识别机制[C]//"智慧城市和绿色 IT"2011 年通信与信息技术新进展——第八届中国通信学会学术年会. 湖北,武汉:出版者不详,2011.

[10] 张依杨,向 阳,蒋锐权,等. 朴素贝叶斯算法的 MapReduce 并行化分析与实现[J]. 计算机技术与发展,2013,23(3):23-26.

[11] Han Jiawei,Kamber M,Pei Jian. 数据挖掘概念与技术[M]. 北京:机械工业出版社,2014:217-218.

[12] 黄宇达,王迤冉. 基于朴素贝叶斯与 ID3 算法的决策树分类[J]. 计算机工程,2012,38(14):41-43.

[13] Friedman N,Geiger D,Goldszmidt M. Bayesian network classifiers[J]. Machine Learning,1997,29(2-3):131-163.

[14] Jing Y,Pavlovi V,Rehg J M. Boosted Bayesian network classifiers[J]. Machine Learning,2008,73(2):155-184.