文章编号：2096-1618(2018)03-0274-07

基于决策树-LMBP神经网络的学生成绩分析及预测模型的研究

吴 强，方 睿，韩 斌，贾 川，浦 东
(成都信息工程大学网络空间安全学院，四川 成都 610225)

摘要：在大数据技术背景和建设智慧校园新阶段下，教育数据挖掘已成为一个新的潮流趋势。结合决策树和LMBP神经网路算法的优点，构建基于这两种算法的分析预测模型，并应用于教育数据挖掘中，实现了应用创新。实验证明通过决策树分析影响学生成绩的主要因子，并在LMBP神经网络模型进行分类预测拥有较单个模型更小的均方误差和更高的分类准确率。

关 键 词：教育数据挖掘；LMBP神经网络；决策树C4.5
中图分类号：TP301.6 **文献标志码**：A
doi：10.16836/j.cnki.jcuit.2018.03.009

0 引言

近些年，随着学生、科研、教学等数据在高校的不断积累，以及数据挖掘技术在各行各业的不断成功应用，教育数据挖掘(educational data mining，EDM)这一研究方向受到大量专家学者关注。2017年，第十届国际教育数据挖掘大会(EDM2017)在武汉开幕，将教育数据挖掘研究热潮推向了一个新高度[1]。

如今中国科教事业不断发展，大学生数量比例逐年增加，同时中国各高校毕业大学生能力不足和综合素质过低的问题却日趋严峻。而建立符合素质教育要求，促进学生成长、教师发展、学校教学质量提高的教育评价体系，已成为新课程改革中的一项重要任务[2]。教育部《基础教育课程改革纲要(试行)》提出建立促进学生全面发展的评价体系。评价不仅要关注学生的学业成绩，而且要发现和发展学生多方面的潜能，了解学生发展中的需求，帮助学生认识自我，建立自信，发挥评价的教育功能，促进学生在原有水平上的发展[3]。

因此，在科技发展的今天，需要借助先进的数据挖掘技术，从学生成绩分析预测入手，建立有效合理的数据挖掘预测模型，科学地分析影响学生成绩的相关因素，预测学生成绩及发展趋势，协助教育管理者发现学生的优缺点，从而正确地评价学生、引导学生，使学生全面发展。

收稿日期：2017-12-06
基金项目：四川省科技计划资助项目(2017GZ0331)

1 理论基础及研究方法

现阶段中国学者对学生成绩预测模型做了一些研究，胡帅等[4]提出了基于PCA-RBF网络的学生写作成绩预测模型，主张使用主成分分析(PCA)进行数据降维，利用RBF网络进行学生成绩预测，虽然有效地提高了模型的收敛速度和准确率，但是未能得出学生成绩的影响因子，使该模型的可理解性大大降低，不利于教育分析者针对学生各方面情况做出优化教学工作、改善管理的措施。王黎黎等[5]提出了基于决策树C4.5算法的学生成绩预测模型，分析并找出了影响学生成绩的主要因素和规则，为学生制定学习计划和预测成绩提供了理论依据。王华等[6]利用了改进的关联规则挖掘算法分析了学业课程中的联系和不及格课程情况，得出学生课程成绩之间的内在关联关系，但是未能分析影响学生成绩其他因素，无法做出合理的决策正确地引导学生发展。何楚等[7]提出的基于频繁模式谱聚类课程关联分析模型和学生成绩预测模型能够比较准确地对课程进行分类及对学生未来可能不及格科目进行预测，具有一定的参考价值。陈子健等[8]计算相关系数和信息增益来确定学业成绩的影响因素，然后通过Bagging、Boosting和RandomForest的集成分类器对学生成绩进行预测。虽然该模型拥有较单一分类器较高的分类准确率，但是效率问题却没有得到详细验证。综上可知，目前已有的学生成绩预测模型大多都基于一种白盒模型、黑盒模型或者一种关联规则算法，模型构建单一，没有利用各个模型的优缺点进行相互弥补。

1.1　决策树 C4.5 算法简述

在诸多的分类方法中,决策树(decision tree)是一种常用的、直观的快速分类方法[9],以内部节点、分支和叶子节点的组合形式表示。通过构造决策树可直观表示对象属性和对象类之间的映射关系,其中决策树的内部节点代表每个属性上的测试,每个分支代表一个测试输出,每个叶子节点代表对象类别。决策树的生成和剪枝都有对应的算法,根据不同的算法就有不同的决策树。其中使用最普遍的就是 C4.5 决策树算法。

设 T 为一个包含 $|T|$ 个数据样本的集合,类别属性有 m 个不同的值,对应 m 个不同的类别集合 $Ci, i \in \{1,2,3,\cdots,m\}$,则通过数据样本 T 构造一个决策树需要以下步骤:

(1)计算 $\mathrm{Info}(T) = -\sum_{i=1}^{m} P_i \log_2 P_i$, 式中 P_i 是 T 中任意元组输入类 Ci 的概率,$\mathrm{Info}(T)$ 为样本 T 的信息熵(entropy),其大小代表识别 T 中元组的类别所需的平均信息量;

(2)计算 $\mathrm{Info}_A(T) = \sum_{j=1}^{v} \frac{|T_j|}{|T|} \times \mathrm{Info}(T_j)$, 式中 v 表示样本按 A 属性划分的子集个数,T_j 表示样本中按 A 属性划分的属于 j 子集的样本,$\mathrm{Info}(T_j)$ 表示要识别该子集类别所需的平均信息量,$\mathrm{Info}_A(T)$ 表示按 A 划分 T 的元组分类所需要的平均信息量;

(3)计算 $\mathrm{Gain}(A) = \mathrm{Info}(T) - \mathrm{Info}_A(T)$;$\mathrm{Gain}(A)$ 表示通过属性 A 划分样本 T 的信息增益;

(4)计算 $\mathrm{SplitInfo}_A(T) = -\sum_{j=1}^{v} \frac{|T_j|}{|T|} \times \log_2 \frac{|T_j|}{|T|}$; 表示 A 的分裂信息大小;

(5)计算 $\mathrm{GainRatio}(A) = \frac{\mathrm{Gain}(A)}{\mathrm{SplitInfo}(A)}$;表示 A 的信息增益率;

(6)选择具有最大增益率属性作为分裂属性,构造决策树内部节点,然后该分裂属性处输出划分尽可能“纯”的分裂子集,构造决策树分支;

(7)循环(1)~(7),直到样本子集为空或者满足其他算法停止条件,则算法停止。

1.2　LMBP 神经网络算法简述

Levenberg 和 Marquradt 分别于 1944 年和 1963 年对非线性最小方差优化问题进行了研究,极大地改进了以搜索方向为核心的迭代优化算法[10],使该算法同时具有高斯-牛顿法极高的收敛速度,又具有梯度下降法全局收敛的特性。

LMBP 与其他 BP 算法不同之处在于误差计算上,设第 k 次迭代的误差函数为

$$E(W^k) = \frac{1}{2}\sum_{i=1}^{T} | Y_i - Y'_i |^2 = \frac{1}{2}\sum_{i=1}^{T} e_i^2(W^k) \quad (1)$$

式中 Y_i 是算法期望输出值,Y'_i 是算法实际输出值,T 为样本数,k 表示迭代次数,$\boldsymbol{W}$ 向量为网络权值 w 和阈值 $basis$ 的组合向量,则可得权重增量。

$$\Delta W = W^{k+1} - W^k = [J^{\mathrm{T}}(W^k)J(W^k) + \eta I]^{-1} J^{\mathrm{T}}(W^k)e(W^k) \quad (2)$$

式中 $\boldsymbol{J}(\boldsymbol{W^k})$ 为 Jacobian 矩阵,$\boldsymbol{J^{\mathrm{T}}}(\boldsymbol{W^k})$ 为 $\boldsymbol{J}(\boldsymbol{W^k})$ 的转置矩阵,$\boldsymbol{I}$ 为单位向量,η 为用户自定义的学习率,Jacobian 定义如下:

$$\boldsymbol{J}(\boldsymbol{W^k}) = \begin{vmatrix} \frac{\partial e_1(\boldsymbol{W^k})}{\partial \boldsymbol{W}_1} & \frac{\partial e_1(\boldsymbol{W^k})}{\partial \boldsymbol{W}_2} & \cdots & \frac{\partial e_1(\boldsymbol{W^k})}{\partial \boldsymbol{W}_n} \\ \frac{\partial e_2(\boldsymbol{W^k})}{\partial W_1} & \frac{\partial e_2(\boldsymbol{W^k})}{\partial \boldsymbol{W}_2} & \cdots & \frac{\partial e_2(\boldsymbol{W^k})}{\partial \boldsymbol{W}_n} \\ .\cdots & \cdots & \cdots & \cdots \\ \frac{\partial e_T(\boldsymbol{W^k})}{\partial \boldsymbol{W}_1} & \frac{\partial e_T(\boldsymbol{W^k})}{\partial \boldsymbol{W}_2} & \cdots & \frac{\partial e_T(\boldsymbol{W^k})}{\partial \boldsymbol{W}_n} \end{vmatrix}$$

LMBP 算法步骤简述如下:

(1)设定误差允许值 ε,初始化迭代次数 $k=0$,常数 η,β $(0<\beta<1)$,并初始化权值 $W(0)$

(2)计算算法输出以及误差 $E(W^k)$;

(3)计算 *Jacobian* 矩阵 $\boldsymbol{J}(\boldsymbol{W^k})$;

(4)计算 $\Delta \boldsymbol{W}$;

(5)若 $E(\boldsymbol{W^k}) \leqslant \varepsilon$,则算法结束;

(6)若 $E(\boldsymbol{W^k}) > \varepsilon$,计算误差指标 $E(\boldsymbol{W^{k+1}})$;

(7)若 $E(\boldsymbol{W^{k+1}}) < E(\boldsymbol{W^k})$,则令 $k=k+1$,$\eta = \eta\beta$,转到第(2)步继续执行循环;

(8)若 $E(\boldsymbol{W^{k+1}}) > E(\boldsymbol{W^k})$,计算 $\eta = \eta/\beta$,转到第(4)步继续执行循环。

基于上述理论,各取所长,利用决策树计算信息增益率筛选因子的能力、可理解性高的优点,和 LMBP 神经网络收敛速度快、分类准确率高、鲁棒性高、可线性预测的优点,提出了基于决策树C4.5-LMBP神经网络学生成绩预测模型。

2　学生数据收集和预处理

从商业角度看,数据挖掘过程的初始阶段(商业目标、数据获取、数据理解和处理)非常重要,要求理解项目目标和商业需求,再转化为一个数据挖掘问题的定义,并做出一个达到该目标的初步计划[11]。在构建学生成绩预测模型之前,首先需要收集学生成绩数据以及

可能影响学生成绩的各方面数据,在分析者充分熟悉数据集类型和特征的基础上,凭借专业经验对数据进行预处理操作,然后选择最适合学生成绩预测及建立模型的算法。从文献[12]的数据集中抽取 1044 条结构化数据用于研究,其中包括了学生个人信息、生活数据、学生成绩数据等,数据集及其特征集如表 1 ~2 所示。

表 1 学生综合数据信息表

School	Sex	Age	Address	Famsize		Absences	G1	G2	G3
GP	F	18	U	GT3		4	0	11	11
GP	F	17	U	GT3	…	2	9	11	11
MS	F	16	R	GT3		0	12	12	12
		…					…		

表 2 学生数据特征说明表

编号	属性	特征说明
1	School	Binary:"GP" or "MS"
2	Sex	Binary:"F" or "M"
3	Age	Numeric: from 15 to 22
4	Address	Binary: "U" - urban or "R" - rural
5	Famsize	Binary: "LE3" - less or equal to 3 or "GT3" - greater than 3
6	Pstatus	Parent´s cohabitation status (binary: "T" - living together or "A" - apart)
7	Medu	Mother´s education (numeric: 0-none,1-primary , 2 - 5th to 9th grade, 3 - secondary or 4 - higher)
8	Fedu	Father´s education (numeric: 0-none, 1-primary , 2 - 5th to 9th grade, 3 - secondary or 4 - higher)
9	Mjob	Mother´s job (nominal: "teacher", "health" , "services" , "at_home" or "other")
10	Fjob	Father´s job (nominal: "teacher", "health", "services", "at_home" or "other")
11	Reason	Reason to choose this school (nominal: close to "home", school "reputation", "course" preference or "other")
12	Guardian	Student´s guardian (nominal: "mother", "father" or "other")
13	Traveltime	Home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
14	Studytime	Weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15	Failures	Number of past class failures (numeric: n if 1<=n<3, else 4)
16	Schoolsup	Extra educational support (binary: yes or no)(yes:1,no:0)
17	Famsup	Family educational support (binary: yes or no))(yes:1,no:0)
18	Paid	Extra paid classes within the course subject (Math or Portuguese) (binary: yes or no))(yes:1,no:0)
19	Activities	Extra-curricular activities (binary: yes or no))(yes:1,no:0)
20	Nursery	Attended nursery school (binary: yes or no))(yes:1,no:0)
21	Higher	Wants to take higher education (binary: yes or no))(yes:1,no:0)
22	Internet	Internet access at home (binary: yes or no))(yes:1,no:0)
23	Romantic	With a romantic relationship (binary: yes or no))(yes:1,no:0)
24	Famrel	Quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25	Freetime	Free time after school (numeric: from 1 - very low to 5 - very high)
26	Goout	Going out with friends (numeric: from 1 - very low to 5 - very high)
27	Dalc	Workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28	Walc	Weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29	Health	Current health status (numeric: from 1 - very bad to 5 - very good)
30	Absences	Number of school absences (numeric: from 0 to 93)
31	G1	First period grade (numeric: from 0 to 20)
32	G2	Second period grade (numeric: from 0 to 20)
33	G3	Final grade (numeric: from 0 to 20)(output target)

通过观察表 2 可得该学生数据是带有混合数据类型的数据集,其中含有 33 个维度信息,13 种二元数据,4 种标称数据以及 16 种数值数据。由于数据维度和数据类型较为复杂,为简化数据集,提高数据挖掘的可操作性,对学生数据集进行数据预处理操作。

属性子集选择。一般对于初次收集的学生数据集可能含有较多的属性,其中有些属性与数据挖掘任务是不相关、弱相关或冗余的,应该删除,使数据类的概率分布尽可能接近使用属性所得到的原始分布[13]。

通过 SPSS 分析学生各属性与预测目标 G3 之间的χ^2系数,设置显著水平 α 为0.05,各个属性相关卡方系数表和冗余属性相关系数表见表 3、表 4。从表 3 可得与 G3 存在明显弱相关的属性为 Sex、Activities、Pstaus、Age、Nursery、Health、Guardian、Famrel,将这些属性删除以减少维度,其他属性留给后续实验部分。

表 3　各个属性卡方系数表

	Sex	Activities	Pstatus	Age	Nursery	Health	Guardian	Famrel	…
χ^2	1.66	1.70	2.01	2.70	2.75	7.47	11.80	14.26	…
Sig(>0.05)	0.64	0.63	0.57	0.43	0.43	0.82	0.06	0.28	…
	Edusup	Absences	Alc	Higher	Studyrate	School	Failures	G1	G2
χ^2	15.65	17.30	22.00	7.47	33.40	34.36	190.89	1014.28	1724.05
Sig(<0.05)	0.016	0.044	0.001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001

根据表 4 可得 Mjob 、Fjob 和 Medu、Fedu 之间存在冗余,为简化数据集,提高描述精度,删除 Mjob 、Fjob 两个属性,Medu、Fedu 留给后续实验部分。

表 4　冗余属性相关系数表

χ^2	Mjob	Fjob
Medu	604.613	153.712
Fedu	188.228	304.863

数据重复项删除和缺失项检查。降维后的数据可能包含有重复项,重复记录增加了数据冗余度和数据分析的工作量,应删除。对于越大的数据集丢失的数据可能性越大,数据丢失最好办法就是用数据降维和特征选取来缩小数据集[14],通过检查未发现数据重复及丢失情况。

数据离群点或异常数据检查(和去除)。离群点和异常数据是和大多数数据的理论取值不一样的数据。一般来讲,异常数据可能由测量错误、记录错误等人为疏忽造成[15]。对于异常点的处理一般较好的办法就是检查并删除含有异常点的记录。通过检查发现 3 条学生记录的缺席数据存在异常,具体如图 1 所示,找到这 3 条异常数据,并直接删除。

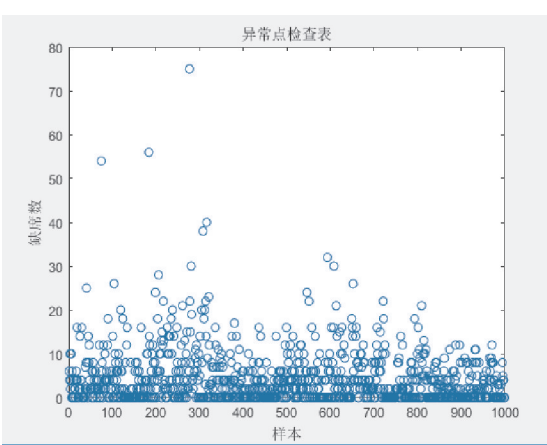

图 1　学生缺席次数分布表

属性聚集和构造。数据聚集是对数据进行汇总或者聚集,而属性构造是通过两个或多个原属性构造出新属性并添加到属性集中,新属性有时候可以更好地描述数据集的特征并减少数据维度。具体属性构造方法需依据分析者经验和对数据熟悉的情况来定,不存在普遍适用的构造方法。具体操作如下:

(1) 将 Dacl 和 Wacl 2 个属性通过取均值的方式聚集成 Acl 新属性,表示每周平均酒精消费水平;

(2) 将 Medu、Fedu 2 个属性通过取均值的方式聚集成 famedu,表示家庭教育水平;

(3) 将 schoolsup 和 famsup 属性合成一个新标称属性 edusup (nominal: "none", "mid" or "much")表示教育支持综合水平,"none" = {schoolsup:"no",famsup:"no"},"much" = {schoolsup:"yes", famsup:"yes"},"mid" = {schoolsup:"no", famsup:"yes"}/{schoolsup:"yes",famsup:"no"};

(4) 将 romantic 和 goout 属性合成新标称属性 friendrel (nominal: "bad", mid" or "good")表示社交关系的好坏,"good" = {romantic:"yes", goout:5},"bad":{romantic="no", goout<=2},"mid":{其他情况};

(5) 将 address 和 famsize 属性合成一个新属性 wealthy(nominal: "not", "mid" or "much"),表示家庭富有程度."much": {address:"u", famsize:"GT3"},"not":{address="r",famsize="LE3"},"mid"={其他情况};

(6) 将 traveltime、studytime 和 freetime 3 个属性构造出新属性 $\text{studyrate}=\frac{\text{studytime}}{\text{traveltime}+\text{freetime}}$,表示学生相对学习时间比率。

数值概念分层化。即属性的原始数据值用区间或者较高层的概念替换,通过数值概念分层化可以节约数据挖掘时间,提高算法效率。现将 failures、absences、G1、G2、G3 进行如下数值归约:

（1）Famedu 的值集为{0,1,2,3,4,5}，归约为{pri:0-1,mid:2-3,higher:4-5}；

（2）Studyrate 的值集为{0.11,0.13,0.140.17,0.2,0.25,0.29,0.33,0.4,0.45,0.5,0.57,0.6,0.67,0.7,0.75,0.8,1,1.33,1.5,2}，归约为{low:0.11-0.5,mid:0.5-1,high:>1}；

（3）Alc 的值集为{1,2,3,4,5}，归约为{low:1-2,mid:3-4,high:5}；

（4）failures 的值集为{0,1,2,3}，归约为{none:0,exist:>0}；

（5）absences 的值集是{ from 0 to 93}，归约为{bit:0-9,few:10-19,many:20-29,more:>=30}；

（6）G1 的值集为{from 0 to 20}，归约为{bad:0-5,mid:6-10,good:11-15,excellent:16-20}；

（7）G2 的值集为{from 0 to 20}，归约为{bad:0-5,mid:6-10,good:11-15,excellent:16-20}；

（8）G3 的值集为{from 0 to 20}，归约为{bad:0-5,mid:6-10,good:11-15,excellent:16-20}。

现将得到的学生数据通过数据处理得到表 5。

表 5 学生综合数据信息表(预处理后)

School	Wealthy	Famedu	Reason		Absences	G1	G3	G3
GP	much	pri	course		bit	mid	mid	excellent
GP	mid	mid	course	…	bit	good	good	good
GP	mid	mid	course		bit	mid	mid	mid
				…				

从表 5 可得预处理后的学生数据全为标称类型数据，拥有 16 个维度，较原始维度减少了一半多，数据冗余度和弱相关性大大降低，整体质量较高，可理解性较强，可以用于下一步实验处理。

3 学生成绩影响因子分析及预测过程

为提高学生成绩预测模型的精度和可靠度，先用决策方法对表 5 的学生数据进行指标筛选，再利用 LMBP 神经网络模型对测试样本的学生成绩进行预测分析。

3.1 决策树构建

决策树算法运算是通过 WEKA3.8 软件中 J48 算法实现的。从数据集中抽取 1000 多条数据作为训练数据集导入到 weka 软件中，选择 J48 算法作为决策树生成算法，通过调整决策树参数剪枝参数(confidencefactor)和最小叶子实例数(minNumObj)可得到最优分类率决策树如图 2 所示。

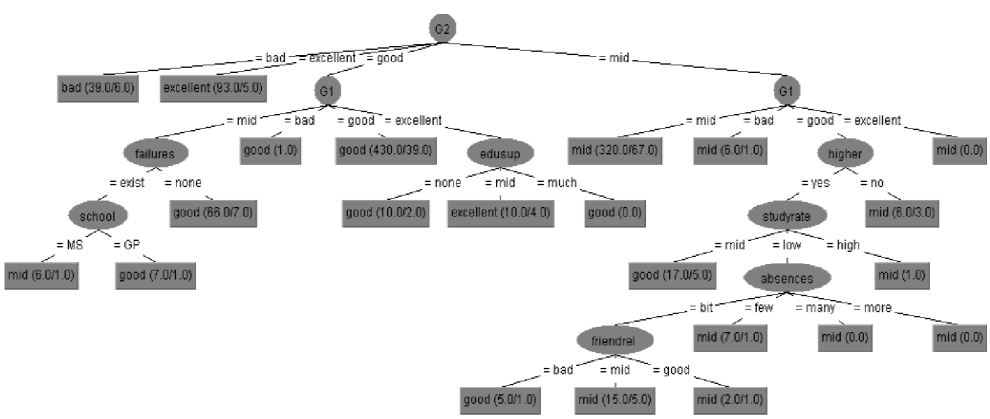

图 2 C4.5 最优决策树(confidenceFactor:0.4,minNumObj:5)

从图 2 中可得决策树通过学生数据训练筛选得出 G2、G1、failures、edusup、higher、school、studyrate、absences、friendrel 9 个指标为影响学生 G3 成绩的主要因子，其中 G2、G1 对 G3 成绩影响最大。

```
=== Confusion Matrix ===

  a   b   c   d   <-- classified as
 33  30   0   0 |   a = bad
  6 274  38   0 |   b = mid
  0  54 478   6 |   c = good
  0   1  33  88 |   d = excellent
```

图 3 误分类矩阵

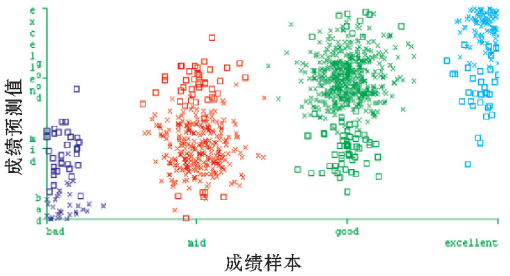

图 4　误分类可视化图

图 3 为误分类矩阵图,矩阵对角线{33,274,478,88}为成绩分类等级{bad,mid,good,excellent}正确分类实例数,其余表示误分类实例数,计算可得误分类率为0.17。

图 4 为误分类可视化图。其中十字表示正确分类,方框表示错误分类,深蓝色代表 bad 分类,红色代表 mid 分类,绿色代表 good 分类,浅蓝色代表 excellent 分类,显然,从图中可得被误分类为 mid 的实例数最多。

综上 9 个指标对学生成绩 G3 进行预测时,准确率最高。也就是说 9 个指标对 G3 成绩影响系数最大,同时也在表 3 卡方相关系数表的最后几个强相关属性中得到验证。这样不仅表明信息增益率与卡方系数具有强弱相关属性的筛选能力,也表明该决策树模型筛选出影响学生成绩的主要因子是完全有效和合理的。

3.2　LMBP 预测学生成绩

通过第一步筛选出来的预测指标按照神经网络输入格式标准化处理后得到表 6 数据。

表 6　神经网络输入数据表

序号	Studyrate	G1	G2	G3
1	2	1	1	2
2	2	2	1	1
3	2	2	1	1
…		…		…

使用 Matlab 神经网络工具箱将上表数据导入到 LMBP 神经网络中,通过调整 LMBP 神经网络参数,当神经网络层数为 2,第一层隐藏层节点数为 6,第二层隐藏层节点数为 5 时,神经网络达到最优结构如图 5 所示。

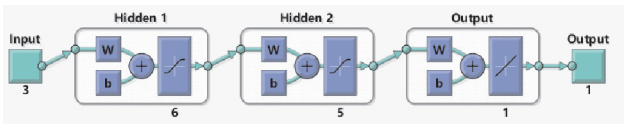

图 5　LMBP 神经网络结构模型

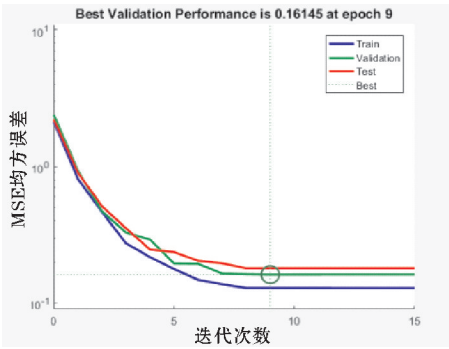

图 6　迭代次数误差均方图

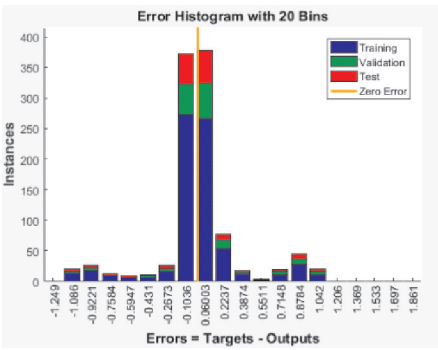

图 7　实例误差均方图

由图 6 可知当迭代次数为第 8 次时,验证集的误差分类率达到最低(0.1527),迭代次数达到第 15 次时该算法终止,最终准确率达84.73%。图 7 中 X 轴表示误差(样本值-实际值)大小,Y 轴表示实例数,可知实验数据整体误差在-0.1 ~0.06,该模型预测的结果误差范围较小,说明用该模型预测学生成绩较为理想。

3.3　模型评估

将表 5 中数据分别用于 RandomTree、RandomForest、Adaboost、BP 神经网络这 4 个分类模型进行分类预测,分类结果如表 7 所示。

表 7　各模型预测结果

模型 / 参数	RandomTree	RandomForest	Adaboost	BP 神经网络	决策树-LMBP
准确率/%	73.29	80.2	73.1	75.6	84.73
平均绝对误差	0.1424	0.1436	0.284	0.1305	0.194

由表 7 可得构建的决策树-LMBP 模型较其他模型拥有较高的准确率和较小的平均绝对误差,表明该模型具有一定优势。综上所有实验结果表明,通过决策树模型筛选出的学生成绩影响因子应用于 LMBP 神

经网络预测模型不仅再次缩减数据维度(缩减 7 个维度,缩减比率43.75%),而且可以得出较高的分类准确率,同时借助决策树白盒模型的特点提高了整个模型的可理解性。

4 结束语

利用决策树算法白盒模型可理解性高的特点,结合 LMBP 神经网络黑盒模型分类准确率高、收敛速度快、能无限逼近任意函数等的特点,构建了基于决策树-LMBP 神经网络分析预测模型,并首次应用在教育数据挖掘中。通过实验虽然获得较满意的结果,但是还存在不足:通过单个卡方相关系数来排除弱相关和冗余属性可能不够完善,除了卡方相关系数外应通过综合协方差系数、皮尔森系数等来比较相关性的大小,可更加可靠地排除弱相关、不相关或者冗余因子;数据预处理中的属性构造虽然降低了数据的维度,但由于缺乏学生数据分析经验,可能导致构造的特征质量不高,所以在实验前应通过专家拜访等形式来提高数据预处理的质量;实验数据集数量不足导致实验模型线性预测的优点没有很好地体现。

参考文献:

[1] 连迅. 中外教育数据挖掘专家齐汇聚武汉分享研究成果[EB/OL]. http://www.hb.xinhuanet.com/2017-06/26/c_1121206912.htm,2017.

[2] 国家基础教育课程改革"促进教师发展与学生成长的评级研究"项目组. 教育性评价[M]. 北京:中国轻工业出版社,2005.

[3] 胡中锋. 教育评价学[M]. 北京:中国人民大学出版社,2009.

[4] 胡帅,顾艳,姜华. 基于 PCA-RBF 网络的学生写作成绩预测模型[J]. 计算机与现代化,2016(1):69-72.

[5] 王黎黎,刘学军. 决策树 C4.5 算法在成绩分析中的应用[J]. 河南工程学院学报(自然科学版),2014(4):69-73.

[6] 王华,刘萍. 改进的关联算法在学生成绩预警中的应用[J]. 计算机工程与设计,2015(3):679-682.

[7] 何楚,宋健,卓桐. 基于频繁模式谱聚类的课程关联分析模型和学生成绩预测算法研究[J]. 计算机应用研究,2015(10):2930-2933.

[8] 陈子健,朱晓亮. 基于数据挖掘的在线学习者学业成绩预测建模研究[J]. 中国电化教育,2017(12):75-81.

[9] 陈安,陈宁,周龙骧,等. 数据挖掘技术及应用[M]. 北京:科学出版社,2006.

[10] 史忠值. 神经网络[M]. 北京:高等教育出版社,2009.

[11] MehmedKantardzic. 数据挖掘:概念、模型、方法和算法(第二版)[M]. 王晓海,吴志刚,译. 北京:清华大学出版社,2013.

[12] P Cortez, A Silva. Using Data Mining to Predict Secondary School Student Performance[C]. In A. Brito and J. Teixeira Eds. ,Proceedings of 5th FUture BUsiness TEChnology Conference(FUBUTEC 2008),2008:5-12.

[13] JiaweiHan MicelineKamber. 数据挖掘概念与技术(第三版)[M]. 范明,孟小峰,译. 北京:机械工业出版社,2012.

[14] 梁循. 数据挖掘算法与应用[M]. 北京:北京大学出版,2006.

[15] Pang-Ning Tan, MichaelSteinbach, Vipin Kumar. 数据挖掘导论[M]. 北京:人民邮电出版社,2006.

Research and Analysison Students′ Achievement and Prediction Model based on Decision Tree and LMBP Neural Network

WU Qiang, FANG Rui, HAN Bin, JIA Chuan, PU Dong

(College of Cybersecurity, Chengdu University of Information Technology, Chengdu 610225, China)

Abstract: Education data mining has become a new trend in the new stage of large data technology background and the construction of intelligent campus. Based on the advantages of decision tree and LMBP neural network, the analysis and prediction models were constructed and theywere applied in educational data mining for the first time to achieve application innovation. The experiment results show that the main factors which affectstudents′achievement are analyzed through decision tree, and the LMBP neural network model used forclassifying and predicating has smaller mean square error and higher classification accuracy than single model.

Keywords: decision tree C4.5; LMBP; EDM