DOI:10.13878/j.cnki.jnuist.2019.03.010

谢娟英[1]　张宜[1,2]　陈恩红[3]

学生成绩关键因素挖掘与成绩预测

摘要

为了探索影响学习成绩的关键因素,为学生学习、教师教学和学校管理提供帮助,采用密度全局 K-means 算法对 UCI 机器学习数据库的葡萄牙学生数据、陕西蒲城县第三高级中学的学生数据进行聚类分析,挖掘影响学生成绩的相关因素,并对学生成绩进行预测分析.葡萄牙学生数据挖掘发现:学生成绩与其所在学校、家庭住址、母亲学历、家庭有无网络有极大相关性,与父亲受教育程度、上学路上花费时间、想上大学、在谈恋爱也有一定相关性.蒲城县第三高级中学学生数据分析发现:学生成绩与其监护人、父母年龄、父母学历、学习态度、课后学习量之间有极大相关性.成绩预测聚类结果显示:预测成绩与实际成绩一致.中外学生数据挖掘揭示:学生成绩与父母受教育程度,特别是母亲受教育程度密切相关,母亲受教育程度越高,孩子学习成绩越好;孩子成长过程中,父母作为监护人的陪伴作用不容忽视;激励和引导学生树立远大理想,调动学生学习的主动性,对学习成绩和成长至关重要;缩小城乡教育差距势在必行.

关键词

教育数据挖掘;学生成绩分析;密度全局 K-means 算法;关联分析;预测分析

中图分类号 TP181

文献标志码 A

收稿日期 2019-05-10

资助项目 国家自然科学基金(61673251);中央高校基本科研业务费项目(GK201701006,GK201806013)

作者简介

谢娟英,女,博士,教授,主要研究方向为机器学习、数据挖掘、生物医学分析.xiejuany@snnu.edu.cn

陈恩红(通信作者),男,博士,教授,主要研究方向为机器学习、数据挖掘、社会网络、个性化推荐系统.cheneh@ustc.edu.cn

1　陕西师范大学　计算机科学学院,西安,710062

2　陕西省蒲城县第三高级中学,蒲城,715500

3　中国科学技术大学　计算机科学与技术学院,合肥,230026

0　引言

学生成绩是评估学校教育质量的重要依据[1],也是评价学生是否掌握所学知识的重要方式,因此寻找影响学生成绩的关键因素非常重要.目前国内外还鲜见关于影响学生成绩的关键因素,以及根据学生现有成绩预测学生未来成绩的研究[2].尤其在应试教育环境下,这方面的研究一直被忽视,原因之一是该研究需要专业的数据挖掘技术.另外,教育数据挖掘在教育中的重要性和有用性还未被充分认识.

教育数据挖掘能够从现有数据中发现隐藏在其中的内在联系与规律,为学生学习、教师教学和学校管理提供帮助[3].教育数据挖掘已经得到我们国家的重视,2018 年起,国家自然科学基金委专门为此设立了相应的学科方向.教育数据挖掘的目的是利用数据挖掘技术对教育数据进行分析,挖掘其内在规律,发现影响学生成绩的关键因素,准确预测学生成绩[4],为提高学生学习成绩、提高学校教学质量提供帮助与支持[5].

Khan[6]从印度一所高中选取了 400 名学生,男、女生各一半,通过对学生在校表现进行聚类分析,挖掘影响学生社会成就的因素,研究发现:对女生而言,家庭经济条件好,则其社会成就高,而对男生则刚好相反,家庭经济条件差的反而社会成就高.Al-Radaideh 等[7]分别运用 ID3、C4.5 和朴素贝叶斯方法,对一所大学的学生 C++课程期末考试成绩进行预测,评估不同方法的分类规则和生成规则,发现决策树模型预测效果更好.Hijazi 等[8]从巴基斯坦一所学校选取了 225 名男生和 75 名女生,对影响大学生在校成绩的因素进行研究,通过回归分析得出:家庭收入和母亲受教育程度对学生在校成绩影响较大.Chapman[9]对几个国家学生参加家庭补课情况进行研究得出:印度有更多学生会接受家庭补课,且该举措可以提高学生成绩,但学生的家庭经济条件对补课强度影响很大.Ayesha 等[10]用 K-means 算法对学生学习行为进行聚类分析并预测其学习成绩,通过对学生课堂测验、期中考试、期末考试等进行研究,将得到的相关信息及时反馈给班主任,帮助教师更好地管理学生,提高学生学习成绩.Bhardwaj 等[11]从印度一所大学选取了 300 名学生,通过建立预测模型对这些学生的成绩进行研究,结果发现:学生母亲受教育程度、家庭年收入、家庭住址、以往学习成绩、生活习惯和家庭因素等对学生成绩有很大影响.舒

忠梅等[12]采用回归分析和神经网络方法对影响大学生学习成绩的因素进行分析得出:学生自身努力程度与学习时间对学习成绩有很大影响,学校资源和校园风气也很大程度影响学生成绩.

以上分析显示教育数据挖掘意义非凡,但该类研究在我国还未得到充分重视,面对科技强国趋势,教育数据挖掘对中国教育意义重大.对教育数据进行聚类、关联及预测分析,找出影响学生成绩的关键因素,对学生学习、教师教学和学校管理都非常有意义,特别是在我们着力培养创新型人才的情况下,教育数据挖掘可为培养创新型人才提供理论支撑与借鉴.

为此,本文提出采用密度全局 K-means 算法[13]对 UCI 机器学习库的 student performance 数据集[14-15]和蒲城县第三高级中学学生数据集进行聚类分析,通过中、外两组教育数据,挖掘与学生成绩高度相关的因素,提出学生成绩预测模型,为因材施教、提高学生学习成绩、提高学校教育质量提供支持.

1 成绩影响因素聚类分析方法

聚类无需任何先验知识,是数据挖掘的一项主要技术,其形式化描述为:设 $\boldsymbol{X}=\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n\}\subset\mathbf{R}^p$ 是包含 n 个样本的数据集,每个样本含有 p 维属性,聚类就是将 $\boldsymbol{X}$ 划分为 k 个类簇 $\{C_1,\cdots,C_k\}$,使 $\boldsymbol{X}=\bigcup_{j=1}^{k}C_j$,$\bigcap_{j=1}^{k}C_j=\Phi$,且使聚类误差平方和 $E=\sum_{i=1}^{n}\sum_{j=1}^{k}\chi(\boldsymbol{x}_i)\|\boldsymbol{x}_i-\boldsymbol{m}_j\|$ 达到最小,其中 $\boldsymbol{m}_j(j=1,\cdots,k)$ 是第 j 类簇 C_j 的质心,$\chi(\boldsymbol{x}_i)=\begin{cases}1, & \boldsymbol{x}_i\in C_j,\\ 0, & \boldsymbol{x}_i\notin C_j.\end{cases}$

基于划分的 K-means 算法[16]是经典聚类算法,其思想简单且收敛速度快,但其聚类结果会受离群点影响,且常收敛于局部最优解.全局 K-means[17]、快速 K-medoids[18]、邻域 K-medoids[19]等是 K-means 的改进算法,但这些改进算法未能克服 K-means 算法只能发现球状簇的缺陷.基于图论的最小生成树算法[20]可以发现任意形状的簇,但效率较低.快速搜索密度峰值点算法[21]可以发现任意形状的类簇,收敛速度快,但其一步样本分配策略会带来“多米诺骨牌”效应,一旦一个样本分配错误,会带来一系列的样本分配错误,致使其不能发现严重重叠数据的真实分布[22].

密度全局 K-means 算法[13]作为全局 K-means 算法[17]的改进,通过引入样本分布密度选取初始聚类中心,使初始聚类中心位于样本分布密集区域,且相互之间距离较远,避免了噪音点对聚类结果的影响,使算法尽可能收敛到全局最优解.密度全局 K-means 算法大大改进了全局 K-means 算法的聚类效果[13].实验证明密度全局 K-means 算法对有挑战性的实际问题具有很好的聚类效果[23],因此,本文选择密度全局 K-means 算法对学生成绩进行聚类分析,比较聚类所得结果的各类簇中心在不同属性的分布,发现各类簇中心分布差异比较大的属性,作为影响学生成绩的关键因素.

2 成绩预测方法

学生成绩预测是指以学生现有成绩对学生的未来成绩进行预测,本文实验的两个数据集的成绩预测方法分别介绍如下.

2.1 葡萄牙学生成绩预测

对葡萄牙学生成绩预测,以前两次成绩预测第 3 次成绩.具体思想是:以前两次成绩均值作为学生成绩,采用密度全局 K-means 对该数据集的学生样本进行聚类,分别计算前两次成绩到样本聚类结果所在簇的距离,以两个距离均值作为第 3 次成绩到相应簇的距离,从而估算出第 3 次考试成绩.

2.2 蒲城县第三高级中学学生成绩预测

对蒲城县第三高级中学学生的成绩预测,以 2019 届学生成绩预测 2020 届学生的成绩.采用距离最近原则,对每个 2020 届学生寻找距离其最近的 2019 届学生,以距离最近的 2019 届学生的成绩预估 2020 届相应的学生的成绩.具体方法是:计算 2020 届各学生样本到 2019 届每个学生样本的距离,找出距离最近的样本,若这样的样本不止一个,则以距离最近的多个学生的成绩均值作为 2020 届相应学生的成绩预测值,否则以距离最近的那个学生的成绩作为 2020 届相应学生的预测成绩.

3 预测成绩评价方法

预测成绩准确与否,采用聚类评价指标进行评价.聚类效果评价方法包括:内部评价指标和外部评价指标[24].内部评价指标不依赖于任何先验知识,外部评价指标需要先验知识.聚类误差平方和是常用的内部评价指标.外部评价指标包括 Rand 指数[25]、Jaccard 系数[26]、Adjusted Rand Index(ARI)参数[27]、聚类准确率[28]等.其中,ARI 参数是目前最好的聚类

有效性评价准则[29].

本文采用密度全局 K-means 对预测成绩与原始成绩分别进行聚类,计算预测成绩聚类结果的聚类误差平方和,以原始成绩的聚类结果为准,计算预测成绩聚类结果的 Rand 指数、Jaccard 系数、ARI 参数、聚类准确率等指标,分析预测成绩与实际成绩的一致性.设密度全局 K-means 对预测成绩的聚类结果为 $P=\{P_1,\cdots,P_s\}$,对实际成绩的聚类结果为 $S=\{S_1,\cdots,S_m\}$. Rand 指数、Jaccard 系数、ARI 参数、聚类准确率的定义分别如式(1)—(4)所示.Rand 指数与 Jaccard 系数的取值范围为[0,1],ARI 参数的取值范围为[-1,1].各参数取值越接近 1,说明预测成绩与实际成绩越一致.

1)Rand 指数:

$$R=(a+d)/(a+b+c+d). \tag{1}$$

2)Jaccard 系数:

$$J=a/(a+b+c). \tag{2}$$

3)ARI 参数:

$$I_{\mathrm{AR}}=\frac{2(ad-bc)}{(a+b)(b+d)+(a+c)(c+d)}. \tag{3}$$

4)聚类准确率(accuracy):

$$R_{\mathrm{acc}}=\frac{r}{n}. \tag{4}$$

式(1)—(4)中:a 表示在 S 和 P 中都属于同一类簇的样本对数目;b 表示在 S 中属于同一类簇,但在 P 中不属于同一类簇的样本对数目;c 表示在 S 中不属于同一类簇,但在 P 中属于同一类簇的样本对数目;d 表示在 S 和 P 中都不属于同一类簇的样本对数目;r 为正确聚类的样本数,n 为总样本数.本研究中 $m=s$,即对预测成绩和实际成绩进行密度全局 K-means 聚类的类簇数 K 相同.

4 实验数据与数据预处理

4.1 实验数据

第 1 个实验数据是来自 UCI 机器学习数据库[15]的学生成绩数据集 student performance[14].该数据集有 649 个样本,包含葡萄牙 2 所中等学校学生的成绩、人数、学校功能等 33 个属性.数据集详细信息描述如表 1 所示.其中 3 次考试成绩均采用 20 分制.

第 2 个实验数据是本文作者张宜通过对蒲城县第三高级中学的家访报告、问卷调查收集获得的 681 个 2019 届和 596 个 2020 届学生的相关数据,包括学生成绩、个人表现、家庭情况等 10 个属性.数据集详细信息如表 2 描述.其中的考试成绩为学生在高一第一学期的期末成绩,总成绩满分为 960 分,包括语文 120 分、数学 120 分、英语 120 分、政治 100 分、历史 100 分、地理 100 分、物理 100 分、化学 100 分、生物 100 分.

4.2 缺失数据处理

常用的缺失数据处理方法有删除法、填补法等[30-32].本文第 1 个数据集不存在缺失数据,不用进行缺失数据处理.第 2 个数据集有部分缺失数据,对成绩缺失的样本,鉴于填补缺失成绩会偏离实际情况,而且含有缺失成绩的样本很少,采取删除法处理.对父母信息缺失的单亲家庭样本,分别建立单亲父亲、单亲母亲模型进行分析,避免填补法偏离实际情况,以及删除法带来的单亲家庭样本缺失.

4.3 离散属性预处理

为采用密度全局 K-means 进行聚类分析,对两个学生数据集的离散型属性进行实值化预处理.处理方法是:假设 x_{if} 是第 i 个样本在第 f 维离散型属性的取值,M_f 是第 f 维属性的取值状态数,z_{if} 为第 i 个样本的第 f 维属性实值化后的值,则 $z_{if}=\dfrac{x_{if}}{M_f-1}$.

4.4 年龄与成绩属性标准化处理

年龄属性和成绩属性相比于其他属性取值过大,若不进行标准化处理,则相当于对年龄和成绩属性赋予很大权重,导致最终聚类结果受年龄和成绩属性影响很大.因此,采用式(5)的最大最小化方法将样本年龄及成绩标准化到[0,1]区间.

$$\mathrm{New}M=\frac{M-\min}{\max-\min}(\mathrm{Newmax}-\mathrm{Newmin})+\mathrm{Newmin}, \tag{5}$$

式(5)中,M 为年龄或成绩原始值,max 和 min 分别为该属性的最大、最小值,Newmax 和 Newmin 分别为新区间的最大值 1 和最小值 0,New$M$ 为 M 标准化后的值.

5 实验结果与分析

本文实验环境为:Win7 32 bit 操作系统,Matlab R2012a 软件,4 GB 内存,AMD Turion(tm) II P520 Dual-Core Processor 2.30 GHz.

5.1 Student performance 数据集

5.1.1 关键因素分析

数据集最后 3 个属性为 3 次考试成绩,采用密

表 1　student performance 数据集描述

Table 1　Student performance data set description

序号	属性描述	属性取值
1	学生学校	GP(Gabriel Pereira)=0;MS(Mousinho da Silveira)=1
2	学生性别	女=0;男=1
3	学生年龄	[15,22]
4	家庭住址	城市=0;农村=1
5	家庭人数	不超过 3 人=0;超过 3 人=1
6	父母生活状态	同居=0;分离=1
7	母亲受教育程度	未上学=0;小学=1;5~9 年级=2;中等教育=3;高等教育=4
8	父亲受教育程度	未上学=0;小学=1;5~9 年级=2;中等教育=3;高等教育=4
9	母亲工作	教师=1;医护=2;服务业=3;在家=4;其他=0
10	父亲工作	教师=1;医护=2;服务业=3;在家=4;其他=0
11	选择学校原因	离家近=1;学校声誉=2;喜欢其课程=3;其他=0
12	监护人	母亲=1;父亲=2;其他=0
13	去学校路上花费时间	小于 15 min=1;15~30 min=2;30~60 min=3;大于 60 min=4
14	每周学习时间	小于 2 h=1;2~5 h=2;5~10 h=3;大于 10 h=4
15	过去失败次数	1 次=1;2 次=2;3 次=3;大于等于 4 次=4
16	学校对教育的支持	是=1;否=0
17	家庭对教育的支持	是=1;否=0
18	参加补课	是=1;否=0
19	课外活动	是=1;否=0
20	上过幼儿园	是=1;否=0
21	想上大学	是=1;否=0
22	家里有网络	是=1;否=0
23	在谈恋爱	是=1;否=0
24	家庭关系好坏	从非常差到非常好,依次取值 1~5
25	放学后的自由时间	从非常少到非常多,依次取值 1~5
26	和朋友外出次数	从非常少到非常多,依次取值 1~5
27	周内饮酒量	从非常少到非常多,依次取值 1~5
28	周末饮酒量	从非常少到非常多,依次取值 1~5
29	当前健康状况	从非常差到非常好,依次取值 1~5
30	缺勤次数	[0,93]
31	第 1 阶段成绩	[0,20]
32	第 2 阶段成绩	[0,20]
33	期末成绩	[0,20]

表 2　蒲城县第三高级中学学生数据集描述

Table 2　Student data set description of the third high school in Pucheng county

序号	属性描述	属性取值
1	监护人	母亲=3;父亲=2;其他=1
2	母亲年龄	[35,50]
3	母亲学历	大学及以上=5;高中=4;初中=3;小学=2;未上学=1
4	父亲年龄	[35,55]
5	父亲学历	大学及以上=5;高中=4;初中=3;小学=2;未上学=1
6	是否想读大学	是=1;否=0
7	上课听讲状态	完全集中=3;大多数时间集中=2;大多数时间不集中=1
8	能否按时完成作业	全部完成=3;部分完成=2;大部分不能完成=1
9	课后学习量	大量=3;适量=1;少量或无=1
10	考试成绩	[0,960]

度全局 K-means 算法进行聚类分析,挖掘影响学生成绩的关键因素时,以前两次成绩均值作为样本成绩.聚类结果的簇质心曲线如图 1 所示.

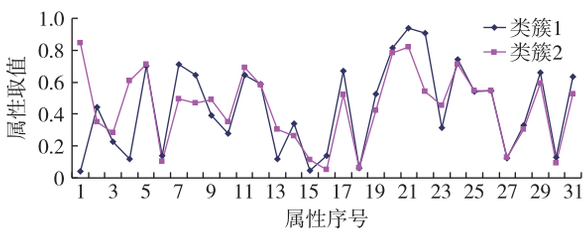

图 1　密度全局 K-means 算法对 UCI 的 student performance 数据集聚类结果簇质心曲线

Fig. 1　The centroids of clusters by density based global K-means on student performance dataset from UCI

图 1 实验结果揭示:聚类结果在第 1、4、7 和 22 属性的差异非常大,该 4 个属性分别是学生所在学校、家庭地址、母亲受教育程度和家里是否有网络;聚类结果在第 8、13、21、23 属性也有一定差异.因此,除了学生所在学校、家庭住址、母亲学历及家庭接入网络情况对学生成绩影响很大外,父亲受教育程度、去学校路上花费时间、想上大学、在谈恋爱这 4 个因素对学生学习成绩也有一定影响.由此可见,孩子成绩不仅受到家庭环境等外在因素影响,还受到主观意愿等内在因素影响.影响孩子成绩的关键因素是:孩子就读学校、家庭地址、母亲受教育程度和家庭经济状况.

5.1.2　成绩关联分析

根据影响孩子成绩的 4 个关键属性分别对学生采用密度全局 K-means 算法进行聚类分析,得到该 4 个属性与学习成绩的关联关系,如图 2 所示.

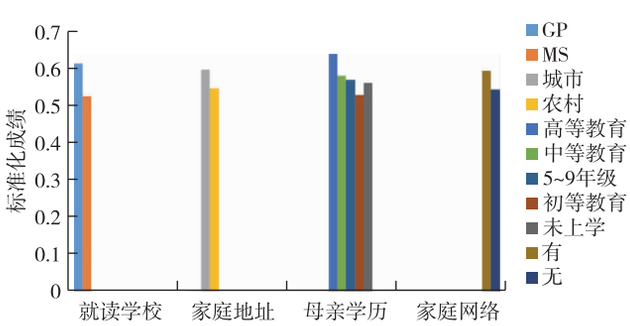

图 2　学生成绩与就读学校、家庭住址、母亲受教育程度、家庭有无网络的关联分析

Fig. 2　Correlation analysis between student performance and his school,home address,his mother's education, and his home's network exists or not

图 2 显示:GP 学校的学生成绩高于 MS 学校的学生;城市学生的成绩高于农村学生的成绩;母亲受过高等教育的学生成绩最好,然后依次是母亲受过中等教育、5~9 年级教育和母亲未上过学的学生,成绩最差的是母亲受过初等教育的学生;家里有网络的学生比家里没有网络的学生成绩更优秀.此处需要注意:在受过教育的母亲中,受教育程度越高的母亲,孩子学习成绩越好,但母亲未上过学的孩子,成绩优于母亲受过初等教育的孩子,说明不能排除母亲因家庭经济状况而读不起书,但会投入精力去引导孩子读好书的情况.

综上分析可得:学生就读学校、家庭住址、母亲受教育程度和家庭经济状况(家里有无网络)与学生成绩密切相关.这符合一般规律:家住城市并且家里有网络的学生,通常经济条件较好,拥有更多学习资源,学习成绩普遍较好,而家住农村又没有网络的学生,往往经济条件较差,可使用的资源较少,学习成绩相对较差;受教育程度越高的母亲有能力对孩子在学习方面给予指导和帮助,学生成绩自然就会越好,而受教育程度较低的母亲,往往对孩子的引导能力较差.

5.1.3　成绩预测分析

以学生的前两次考试成绩预测其第 3 次考试成绩.根据 2.1 节描述的学生成绩预测方法,预测出学生第 3 次考试成绩,然后按学生实际成绩递减顺序对样本进行编号,将预测成绩和实际成绩进行对比.实际成绩与预测成绩的对比曲线如图 3 所示.

从图 3 实验结果可以看出:预测成绩与实际成绩基本一致,围绕实际成绩上下有小幅度浮动.因此,2.1 节提出的成绩预测方法能够较准确地预测学生成绩.

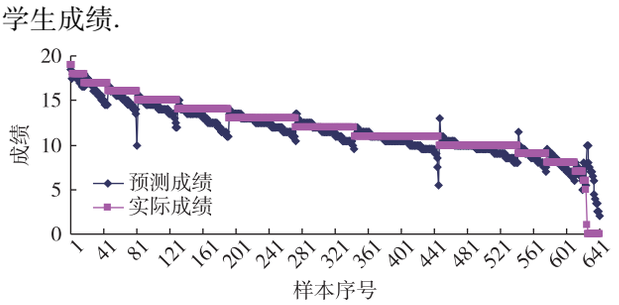

图 3　第 3 次考试预测成绩与实际成绩对比

Fig. 3　Comparison between the predictive scores and the true ones of the third test

对样本前 30 个属性+第 3 次考试实际成绩构成的数据集、前 30 个属性+第 3 次考试预测成绩构成

的数据集,分别采用密度全局 K-means 进行聚类,得到两个聚类结果.以第 3 节描述的评价指标评价预测成绩与实际成绩的相符度.各评价指标如表 3 所示.

表 3　预测成绩聚类结果指标值

Table 3　The evaluation criterion values of the clustering of the predictive scores

	评价指标				
	E	Rand 指数	Jaccard 系数	ARI 参数	聚类准确率
取值	2 213.6	0.993 8	0.988 4	0.987 7	99.69%

表 3 实验结果显示,样本预测成绩和实际成绩分别进行密度全局 K-means 聚类分析,有 99.69%的样本的聚类结果相同,预测成绩聚类结果的 Rand 指数、Jaccard 系数和 ARI 参数非常接近 1,说明提出的成绩预测方法所得模型对不同层次学生的成绩预测得非常准确.

5.2　蒲城县第三高级中学学生成绩分析

5.2.1　成绩影响因素分析

为准确分析影响学生成绩的因素,将蒲城县第三高级中学的学生数据分为单亲、非单亲数据分别进行聚类分析.使用密度全局 K-means 对 2019 届非单亲家庭的 626 名学生聚成 4 个类簇.图 4 展示了聚成 4 个簇的质心在各属性的取值曲线.单亲与非单亲家庭学生成绩的聚类分析结果对比如图 5 所示.

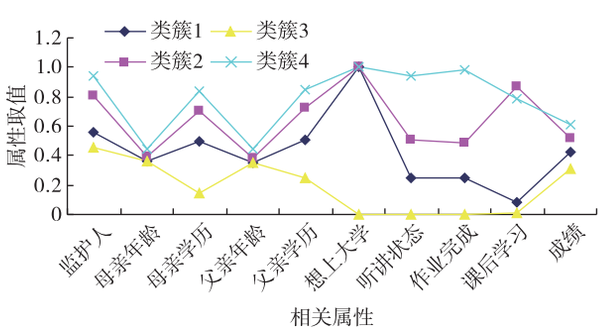

图 4　2019 届学生聚成结果质心在各属性的取值曲线

Fig. 4　The centroid curve of the clustering results of the grade 2019

图 4 结果显示:簇 4 学生的成绩最好,簇 2 学生的成绩次之,簇 1 学生的成绩位居第三,簇 3 学生的成绩最差.聚类结果在监护人、母亲学历、父亲学历、听讲状态、作业完成情况 5 个属性的取值差异较大.是否想读大学的主观意愿对学生成绩有一些影响,但不是主要因素,因为除了学习成绩最差的类簇,其他学生均有想读大学的主观意愿.就课后学习量属性来看,学习成绩较好的第 4、第 2 类簇,课后学习量属性的取值远高于学习成绩较差的第 1、第 3 类簇,学习成绩最差的第 3 类簇在该属性的取值最低,但学习成绩最好的簇 4 在该属性的取值低于学习成绩次之的簇 2 在该属性的取值.因此,课后学习量对学生成绩会产生影响,但并不是完全正相关,课后学习量不是越多越好,当学习量达到一定程度时,如果继续增加,并不能提升学生成绩.学生父母亲年龄基本相当,学习成绩最好的簇 4 的学生,父母亲相对年轻.由此可见:学生成绩与监护人、父母亲学历、听讲状态、作业完成情况密切相关.

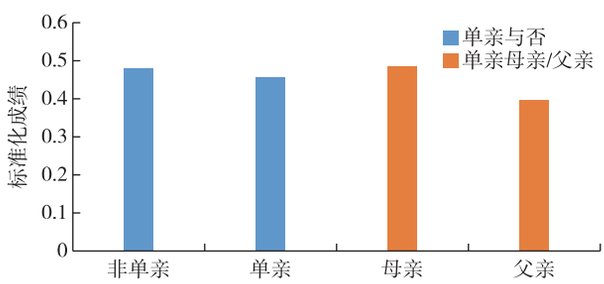

图 5　单亲与非单亲家庭孩子成绩对比

Fig. 5　The performance comparison of children from single or non-single parent families

图 4 聚类结果显示父母亲年龄取值非常接近,为了验证父母亲年龄对孩子学习是否有影响,对聚类结果的父母亲年龄进行方差分析,表 4 给出了聚类结果的学生父母年龄方差及学生成绩.

表 4　聚类结果的学生父母年龄方差与学生成绩

Table 4　The variance of parent ages and student performance of the clustering results

类簇	母亲年龄方差	父亲年龄方差	学生成绩
1	0.088 5	0.073 1	0.421 1
2	0.075 7	0.060 6	0.518 3
3	0.096 8	0.070 5	0.307 9
4	0.014 6	0.017 5	0.611 2

表 4 结果显示:学生成绩最好的簇 4,父母年龄的方差最小;成绩较好的簇 2,父母年龄方差较小;簇 1 学生成绩较差,母亲年龄方差较大,父亲年龄方差最大;成绩最差的簇 3,母亲年龄方差最大,父亲年龄方差较大.由此可见,若父母年龄处于平均水平,即方差越小,则孩子成绩较好,而父母过于年轻或年长,即方差越大,孩子成绩则相对较差.这与适龄生

育理论相一致,特别是母亲的年龄不宜过大.

从图 5 可见,非单亲家庭学生成绩高于单亲家庭学生成绩,说明单亲家庭对孩子的成长有负面影响.单亲家庭中,单亲母亲家庭的孩子成绩优于单亲父亲家庭的孩子成绩,可见母亲对孩子的教育比父亲更重要,在孩子的成长过程中起着关键性作用.

5.2.2 成绩关联结果分析

根据学生的监护人、父母亲学历、听讲状态、作业完成情况、课后学习量各属性对学生进行聚类分析,得到各属性与成绩的关联关系,如图 6 所示.

图 6 显示,监护人为母亲的学生成绩最好,监护人为父亲的次之,监护人为其他(一般多为祖父母)的学生成绩最差.母亲学历为大学及以上的学生成绩最好,母亲学历为高中、初中、小学的学生,成绩依次变差.父亲学历为大学及以上的学生成绩最好,然后依次是父亲学历为高中、初中和小学的学生.上课注意力完全集中听讲的学生成绩最好,上课大多数时间能集中注意力听讲的学生次之,上课大多数时间不能集中注意力听讲的学生成绩最差.能够按时完成全部作业的学生成绩最好,按时完成部分作业的学生成绩次之,大部分作业不能按时完成的学生成绩最差.课后学习量适量的学生成绩最优,课后学习量很大的学生次之,课后学习量较少或没有的学生成绩最差.

由此可见,学生的监护人对其成绩有很大影响,父母亲尤其母亲在孩子成长过程中起着非常关键的作用,祖父母等其他人对孩子的教育作用远不及父母.父母学历对学生成绩影响很大,父母学历越高,学生成绩越好,父母学历越低,学生成绩越差.上课听讲状态对学生成绩影响较大,上课越注意力集中、认真听讲的学生成绩越优秀,上课听讲状态不认真、注意力不集中的学生成绩较差.学生作业完成情况对其成绩影响较大,能够按时完成作业的学生成绩优秀,而不能按时完成作业的学生成绩较差.课后学习量对学生的学习成绩影响较大,但在课后学习量达到一定程度时,一味增加学习量并不能提高学生学习成绩.

5.2.3 成绩预测结果分析

以 2.2 节介绍的学生成绩预估方法,根据 2019 届学生成绩预测 2020 届学生成绩.然后按照成绩递减顺序对学生进行编号,比较 2020 届各学生的预测成绩和实际成绩.各学生预测成绩与实际成绩对比如图 7 所示.

从图 7 可见:预测成绩与实际成绩几乎完全一致,以实际成绩为主线,上下仅有很小幅波动,由此可见,2.2 节提出的成绩预测方法所得预测模型能够较为准确地预测学生成绩.

对 2020 届学生的实际考试成绩和预测成绩分别进行密度全局 K-means 聚类,得到原始成绩聚类

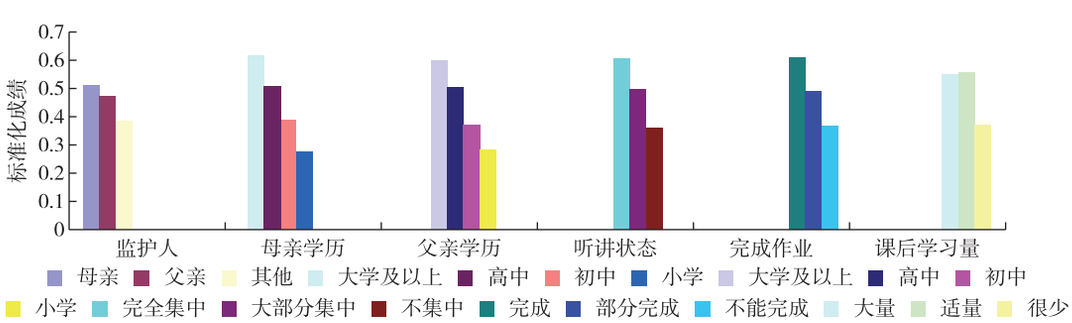

图 6 各属性与学生成绩的关联关系

Fig. 6 The correlation analysis between each attribute and student performance

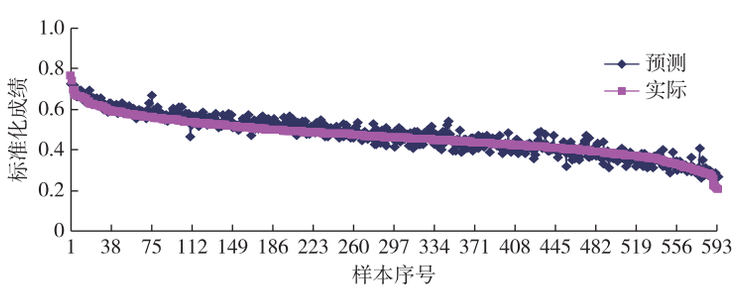

图 7 2020 届学生预测成绩与实际成绩对比

Fig. 7 Comparisons between predictive and true scores of grade 2020

结果和预测成绩聚类结果.对比两个聚类结果,计算预测成绩对应聚类结果的各评价指标值,如表5所示.

表5 2020届学生预测成绩聚类结果评价指标

Table 5 Evaluation criterion values of clustering results of predictive scores for grade 2020

	评价指标				
	E	Rand指数	Jaccard系数	ARI参数	聚类准确率
取值	233.642 0	1	1	1	100%

表5聚类结果指标显示:2020届学生的预测成绩和实际成绩分布完全一致,说明2.2节提出的预测方法能得到非常好的预测模型,对不同层次学生的成绩预测非常准确.

6 结论与建议

对葡萄牙两所学校的学生数据集和中国陕西蒲城县第三高级中学的学生数据集分别进行聚类、关联以及成绩预测分析.葡萄牙学生数据分析发现:学生成绩与其所在学校、家庭住址、母亲受教育程度、家庭有无网络情况密切相关.城市孩子比农村孩子成绩优秀;母亲受教育程度越高,孩子成绩越优秀;家里有网络的孩子比家里没有网络的孩子成绩更优.成绩预测揭示:预测成绩与实际成绩整体相符,99.69%的样本预测结果与实际一致,预测模型能对不同层次学生的成绩进行准确预测.蒲城县第三高级中学学生数据分析得出:学生成绩与监护人、父母年龄、父母学历、学习态度(听讲状态、作业完成情况)、课后学习量密切相关.监护人为母亲的学生成绩最优,监护人为父亲的学生成绩次之,监护人为其他人的学生成绩最差;父母学历越高,学生成绩越好,尤其母亲学历对孩子成绩影响极大;父母年龄处于平均水平的孩子,成绩优于父母年龄较大或较小的孩子;上课认真听讲、作业完成情况好的学生,成绩优秀;课后学习量适量的学生,成绩优于课后学习量很大的学生.单亲与非单亲家庭学生数据分析得出:非单亲家庭学生的成绩优于单亲家庭学生成绩;单亲母亲孩子的成绩优于单亲父亲孩子的成绩.成绩预测发现:由上届成绩能准确预测下届学生成绩.

中、外学生数据分析发现:

1)父母尤其母亲在孩子成长过程中的陪伴作用不容忽视;

2)父母受教育程度,尤其是母亲受教育程度,与孩子学习成绩绝对正相关;

3)学习态度会对学习成绩产生极大影响,培养孩子的积极主动学习态度至关重要;

4)课后作业量适度至关重要,适量作业有助于提高学习成绩,繁重的课后作业则适得其反;

5)缩小城乡教育差距是亟待解决的问题;

6)适龄生育、优生优育作为国家政策依然需要倡导.

参考文献

References

[1] Jin H J, Wang X R, Wang Y L, et al. Study and application of genetic algorithm in computer test construction[C]// IEEE International Symposium on Communications and Information Technology, 2005. ISCIT2005, 12-14 Oct. 2005, Beijing, China, 2005: 424-427

[2] 谭庆.基于K-means聚类算法的试卷成绩分析研究[J].河南大学学报(自然科学版),2009,39(4):412-415
TAN Qing. Analysis and research of grades of examination paper based on K-means clustering algorithm[J]. Journal of Henan University(Natural Science),2009,39(4):412-415

[3] 王盛.教育数据挖掘促进高校学生个性化学习途径分析[J].考试周刊,2014(34):176
WANG Sheng. Analysis of educational data mining in promoting individualized learning of university students[J]. Exam Weekly,2014(34):176

[4] 彭涛,丁凌云.基于教育数据挖掘学生表现预测模型构建研究[J].黑龙江高教研究,2015,33(11):55-58
PENG Tao, DING Lingyun. Performance prediction model based ondata mining based education students[J]. Heilongjiang Researches on Higher Education, 2015, 33(11):55-58

[5] 洪雪峰.教育数据挖掘下的学习效果探析[J].长沙铁道学院学报(社会科学版),2014(2):196-198
HONG Xuefeng. The exploring analysis of learning effect under educational data mining[J]. Journal of Changsha Railway University(Social Sciences),2014(2):196-198

[6] Khan Z N. Scholastic achievement of higher secondary students in science stream [J]. Journal of Social Sciences,2005,1(2):84-87

[7] Al-Radaideh Q A, Al-Shawakfa E M, Al-Najjar M I. Mining student data using decision trees[C]// The 2006 International Arab Conference on Information Technology (ACIT'2006), Yarmouk University, Jordan.2006

[8] Hijazi S T, Naqvi S M M R. Factors affecting students performance acase of private colleges[J]. Bangladesh e-Journal of Sociology,2006,3(1):90-100

[9] Chapman D W. The shadow education system: private tutoring and its implications for planners[J]. Economicsof Education Review,2001,20(6):608-609

[10] Ayesha S, Mustafa T, Sattar A R, et al. Data mining model

for higher education system[J].European Journal of Scientific Research,2010,43(1):24-29

[11] Bhardwaj B K, Pal S. Data mining: aprediction for performance improvement using classification [J]. International Journal of Computer Science and Information Security,2011,9(4):136-140

[12] 舒忠梅,屈琼斐.基于教育数据挖掘的大学生学习成果分析[J].东北大学学报(社会科学版),2014,16(3):309-314
SHU Zhongmei, QU Qiongfei. An analysis of university students learning outcome based on educational data[J]. Journal of Northeastern University (Social Science), 2014,16(3):309-314

[13] 谢娟英,蒋帅,王春霞,等.一种改进的全局K-均值聚类算法[J].陕西师范大学学报(自然科学版),2010,38(2):18-22
XIE Juanying, JIANG Shuai, WANG Chunxia, et al. An improved global K-means clustering algorithm[J].Journal of Shaanxi Normal University(Natural Science Edition), 2010,38(2):18-22

[14] Cortez P,Silva A.Using data mining to predict secondary school student performance[C]//Proceedings of 5th Future Business Technology Conference(FUBUTEC 2008), 2008:5-12

[15] Dua D,Karra T E.UCI machine learning repository[EB/OL].[2019-05-18].http://archive.ics.uci.edu/ml

[16] Huang Z X.Extensions to the k-means algorithm for clustering large data sets with categorical values[J]. Data Mining and Knowledge Discovery,1998,2(3):283-304

[17] Likas A,Vlassis N,Verbeek J J.The global k-means clustering algorithm[J].Pattern Recognition,2003,36(2): 451-461

[18] Park H S,Jun C H.A simple and fast algorithm for K-medoidsclustering[J]. Expert Systems with Applications, 2009,36(2):3336-3341

[19] 谢娟英,郭文娟,谢维信.基于邻域的K中心点聚类算法[J].陕西师范大学学报(自然科学版),2012,40(4):16-22
XIE Juanying, GUO Wenjuan, XIE Weixin. A neighborhood-based K-medoids clustering algorithm[J]. Journal of Shaanxi Normal University (Natural Science Edition),2012,40(4):16-22

[20] 王小乐,刘青宝,陆昌辉,等.一种最小生成树聚类算法[J].小型微型计算机系统,2009,30(5):877-882
WANG Xiaole, LIU Qingbao, LU Changhui, et al. Minimum spanning tree clustering algorithm[J]. Journal of Chinese Computer Systems,2009,30(5):877-882

[21] Rodriguez A,Laio A.Clustering by fast search and find of density peaks[J].Science,2014,344(6191):1492-1496

[22] 谢娟英,高红超,谢维信.K近邻优化的密度峰值快速搜索聚类算法[J].中国科学:信息科学,2016,46(2): 258-280
XIE Juanying, GAO Hongchao, XIE Weixin. K-nearest neighbors optimized clustering algorithm by fastsearch and finding the density peaks of adataset[J]. Science China:Information Science,2016,46(2):258-280

[23] 张宜,谢娟英,李静,等.红斑鳞状皮肤病的聚类分析[J].济南大学学报(自然科学版),2017,31(3): 181-187
ZHANG Yi, XIE Juanying, LI Jing, et al. Clustering analysis for erythemato-squamous diseases[J].Journal of University of Jinan(Science and Technology),2017,31(3):181-187

[24] 谢娟英,周颖,王明钊,等.聚类有效性评价新指标[J].智能系统学报,2017,12(6):873-882
XIE Juanying,ZHOU Ying,WANG Mingzhao,et al.New criteria for evaluating the validity of clustering[J].CAAI Transactions on Intelligent Systems, 2017, 12(6): 873-882

[25] Hubert L,Arabie P.Comparing partitions[J].Journal of Classification,1985,2(1):193-218

[26] 杨燕,靳蕃,Kamel M.聚类有效性评价综述[J].计算机应用研究,2008,41(6):1631-1632
YANG Yan, JIN Fan, Kamel M. Survey of clustering validity evaluation [J]. Application Research of Computer,2008,41(6):1631-1632

[27] 于剑,程乾生.模糊聚类方法中的最佳聚类数的搜索范围[J].中国科学:E辑,2002,32(2):274-280
YU Jian,CHENG Qiansheng.The search range of optimal cluster number in fuzzy clustering methods[J].Science in China:Series E,2002,32(2):274-280

[28] 谢娟英,郭文娟,谢维信,等.基于样本空间分布密度的初始聚类中心优化K-均值算法[J].计算机应用研究,2012,29(3):888-892
XIE Juanying,GUO Wenjuan,XIE Weixin,et al.K-means clustering algorithm based on optimal initial centers related to pattern distribution of samples in space[J].Application Research of Computers,2012,29(3):888-892

[29] Vinh N X, Epps J, Nailey J. Information theoretic measures for clustering comparison: is acorrection for chance necessary[M]. New York: ACM Press, 2009: 1073-1080

[30] Han J W,Kamber M.数据挖掘概念与技术[M].范明,孟小峰.译.北京:机械工业出版社,2001
Han J W, Kamber M. Data mining: concepts and techniques [M]. Translated by FAN Ming, MENG Xiaofeng.Beijing:Machinery Industry Press,2001

[31] Grzymala-BusseJ W,Hu M.A Comparison of several approaches to missing attribute values in data mining[M]// Grzymala-Busse J W, Hu M. eds. Rough Sets and Current Trends in Computing.Berlin,Heidelberg:Springer Berlin Heidelberg, 2001: 378-385. DOI: 10.1007/3-540-45554-x_46

[32] 乔珠峰,田凤占,黄厚宽,等.缺失数据处理方法的比较研究[J].计算机研究与发展,2006(增刊1): 171-175
QIAO Zhufeng,TIAN Fengzhan,HUANG Houkuan,et al. A comparasionstudy of missing value datasets processing methods[J].Journalof Computer Research and Development,2006(supl):171-175

[33] 方洪鹰.数据挖掘中数据预处理的方法研究[D].重庆:西南大学,2009
FANG Hongying. Data processing method of dimensionless [D]. Chongqing: Southwest University,2009

南京信息工程大学学报(自然科学版),2019,11(3):316-325
Journal of Nanjing University of Information Science and Technology(Natural Science Edition),2019,11(3):316-325

Mining the key factors behind student performance and predicting students' examination scores

XIE Juanying[1] ZHANG Yi[1,2] CHEN Enhong[3]
1 School of Computer Science, Shaanxi Normal University, Xi'an 710062
2 The Third Senior Middle School of Pucheng County, Pucheng County 715500
3 School of Computer Science and Technology, University of Science and Technology of China, Hefei 230026

Abstract Understanding the key factors that influence student performances will help students, teachers, and administrators to improve the performance of the students. To this end, the density-based global K-means algorithm is adopted to perform cluster analysis of the student performance data from the UCI machine learning repository for two secondary education Portuguese schools and for a senior middle school of the Pucheng county in the Shaanxi province. The results for the two Portuguese schools reveal that student performance is strongly related to the specific school where the student is enrolled, and location of residence, mother's education level, and if the network is available or not in the family. Education level of the father, the time the student takes on the way to school, the willingness of the student to go to college, and whether the student is in love are factors affecting the student performance to some extent. The results of the third senior middle school demonstrate that student performance is strong related to their guardians, parents' age, parents' education level, learning attitude of the student, and the time the student devotes to courses after classes. In addition, the results indicate that scores of a student for the upcoming examination can be predicted with the available ones and that the predicted scores coincide with the actual ones. The studies in this paper demonstrate that student performance is strong related to parents' education level, especially to mother's education level. The higher the level of education of the mother, the better the student performance. Parents cannot ignore their role in the individual growth of children. It is important to teach students to study actively to improve their achievements. Finally, it is imperative that the education gap between the urban and rural areas is narrowed.

Key words educational data mining; student performance analysis; density-based global K-means; association analysis; prediction analysis