基于 K-means 算法的研究生入学成绩分析

李春生,刘 涛,于 澍,张可佳
(东北石油大学 计算机与信息技术学院,黑龙江 大庆 163318)

摘 要: 研究生入学成绩是导师初步了解学生学习能力、学习风格、制定研究生培养方案的重要参考指标。随着学校招生规模的扩大,学生人数的增加,研究生入学成绩的日趋复杂,传统的分析方法已经不能满足当前对于研究生入学成绩分析的需要。通过应用 K-means 聚类算法对研究生入学成绩进行分析,将研究生入学成绩进行分类,发现学生成绩分布的特点,找出成绩之间的关系,了解学生各科的学习状况,找到适合学生发展的方向,以实现个性化的研究生教育和培养,所得结果为研究生培养方案的制定与研究生进行研究方向的选择提供了借鉴意义。首先,分析了几种主要聚类算法应用于研究生入学成绩的适用性;其次,介绍了 K-means 聚类算法;最后,对研究生入学成绩进行数据分析、预处理。通过实验证明了 K-means 聚类算法在研究生入学成绩分析中的实用性。

关键词: 研究生入学成绩;聚类分析;K-means 算法;成绩分析
中图分类号: G424.7 **文献标识码:** A **文章编号:** 1673-629X(2019)02-0162-04
doi: 10.3969/j.issn.1673-629X.2019.02.034

Analysis of Enrollment of Graduate Students Based on K-means Algorithm

LI Chun-sheng, LIU Tao, YU Shu, ZHANG Ke-jia
(School of Computer and Information Technology, Northeast Petroleum University, Daqing 163318, China)

Abstract: Postgraduate enrollment is an important reference for instructors to understand students' learning ability, learning style and development of postgraduate training programs. With the expansion of enrollment scale, the increase of students and the growing complexity of the postgraduate enrollment, the traditional analysis methods can no longer meet the current needs of graduate enrollment analysis. Through the application of K-means clustering algorithm to analyze the results of graduate enrollment, the graduate enrollment score is classified to find out the characteristics of student achievement distribution and the relationship between achievements, and the learning situation of students in each subject is understood to find a direction suitable for student development, achieving personalized graduate education and training. The results provide a reference for the formulation of postgraduate training programs and the selection of research direction for graduate students. Firstly, the applicability of several major clustering algorithms for graduate student enrollment performance is analyzed. Secondly, the K-means clustering algorithm is introduced. Finally, the data of graduate enrollment grades are analyzed and preprocessed. The practicability of K-means clustering algorithm in the analysis of postgraduate enrollment score is demonstrated by experiments.

Key words: graduate enrollment scores; clustering analysis; K-means algorithm; grade analysis

0 引 言

研究生入学成绩作为学生在毕业学校期间表现好坏的重要评价标准之一,是学生检测之前学习状态、学习态度的依据,是学校之间相互了解的重要参考标准,然而大多数学校对于学生的入学成绩仅仅停留在登记、分类、储存等表面的工作,缺乏对学生入学成绩背后潜在信息的深入分析,从而造成了教学资源的浪费。通过对研究生入学成绩的分析,可以帮助导师了解学生的学习能力、学习风格,帮助学生制定个性化培养方案。同时,学生的入学成绩反映出学生毕业学校的办

收稿日期: 2018-03-19 修回日期: 2018-07-13 网络出版时间: 2018-11-15
基金项目: 国家自然科学基金(51774090);黑龙江省自然科学基金(F2015020);黑龙江省教育科研专项引导性创新基金项目(2017YDL-12);黑龙江省教育规划重大课题(GJ20170006)
作者简介: 李春生(1960-),男,博士,教授,博导,研究方向为数据挖掘与智能系统、软件集成技术、图像处理与模式识别、智能仪器与计算机控制系统;刘 涛(1994-),男,研究生,研究方向为数字媒体技术。
网络出版地址: http://kns.cnki.net/kcms/detail/61.1450.TP.20181115.1048.048.html

学条件,教学水平,人才培养的质量。因此,对研究生入学成绩进行合理有效的深度挖掘,可以更好地帮助学生进行研究方向的选择,帮助其找到适合自己发展的方向,实现研究生的个性化教育和培养。

应用聚类分析算法对研究生入学成绩进行分析,能够发现学生成绩分布的特点,找出成绩之间的关系,弥补传统分析方法对研究生入学成绩分析的不足,为教学管理者提供了决策指导。

1　聚类分析

聚类分析属于探索性、无监督的数据分析方法,即将给定的数据元素进行划分,使高相似性的数据元素归为一类,低相似性的数据元素归为不同类。相似度的高低是通过计算两个数据元素之间的距离来判断的[1]。欧几里得公式是常用的距离计算公式,表示如下:

$$\mathrm{dist}(X,Y)=\sqrt{\sum_{i=0}^{n}(x_i-y_i)^2} \tag{1}$$

主要的聚类分析算法包括层次聚类算法、基于密度的聚类算法、基于网格的聚类算法、基于模型的聚类算法和划分聚类算法[2]。

1.1　层次聚类算法

层次聚类算法是运用层次分解的方法将给定的数据集合形成满足某种条件的聚类树。层次聚类算法具有很高的聚类精度,但其时间复杂度和空间复杂度较高的缺点依旧无法避免[3]。随着研究生人数的迅速增长,入学成绩数据量的增大,层次聚类算法因其自身的缺点,无法高效地实现对研究生入学成绩的分析。

1.2　基于密度的聚类算法

基于密度的聚类算法给定密度阈值,将密度超过阈值的区域连接,形成相同密度区[4]。使用该算法的基础是数据分布的密度差距较大,由于研究生入学成绩首先要超过国家分数线,研究生入学成绩的分布很难出现密度差距较大的区域,因此基于密度的聚类算法不适用于研究生入学成绩的分析。

1.3　基于网格的聚类算法

基于网格的聚类算法将给定的数据元素划分成若干网格单元,然后进行网格单元的凝聚和分裂[5]。基于网格的聚类算法优缺点明显,优点是聚类速度快,缺点是无法处理分布不规则的数据。研究生入学成绩在空间上分为多个维度,数据的分布具有随机性易形成不规则分布,因此基于网格的聚类算法对于研究生入学成绩的分析缺乏精确性。

1.4　基于模型的聚类算法

基于模型的聚类算法将每个聚类构想成数学模型,然后通过聚类使数据元素找到自己的对应模型,用统计法得到类个数的过程。常见的算法有统计学方法和神经网络方法。该算法复杂度较高,执行效率缓慢,因此也不适用于对研究生入学成绩进行分析。

1.5　划分聚类算法

划分聚类算法运用分裂的方法将数据集进行分类,分组的结果使同类元素之间的相似性高,异类元素之间相似性低,并且保证每一类中都包含一个或多个数据元素,每个数据元素只属于一个类。划分聚类算法具有算法简单、复杂性低、收敛速度快、聚类效果较好等特点,加之研究生入学成绩有复试分数线作为"基础分数",没有明显的噪声点的影响,因此适用于研究生入学成绩的分析。常见的算法有 K-means 算法、K-medoids 算法及 CLARANS 算法等[6],其中典型的代表算法就是 K-means 聚类算法。

2　K-means 聚类算法

K-means 算法是划分聚类算法中一种经典的聚类方法,在生物应用、图像分析、市场调查等领域中应用广泛[7-9]。以 K-means 算法为关键字在中国知网上进行搜索,发现对于 K-means 算法的研究虽然在某个年份中有略微的下降趋势,但其总体上是持续上升的,如图 1 所示(2018 年为预测数值)。

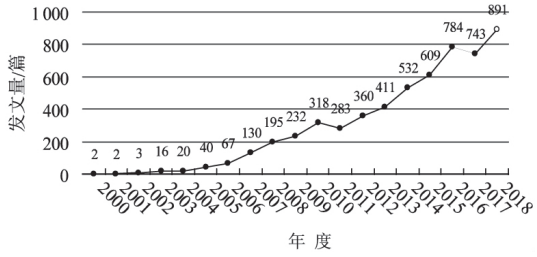

图 1　K-means 聚类算法研究趋势

K-means 聚类的算法思想为,给定包含 X 个 d 维数据的数据集 $M=\{m_1,m_2,\cdots,m_n\}(m_i\in R^d)$,及将要生成数据的子集数目 K,K-means 算法将给定的数据集分为 K 组[10-11]。每个分组为一个类 $C=\{C_k,i=1,2,\cdots,K\}$,每个类 C_k 都有一个中心 O_i。以欧氏距离作为数据之间相似性的判断标准,计算类中数据点到聚类中心 O_i 的距离平方和,计算公式为:

$$J(C_k)=\sum_{m_i\in C_k}\|m_i-o_k\|^2 \tag{2}$$

聚类的最终的目的是使同类中所有数据元素到其聚类中心距离的平方和 $J(C)$ 值最小[12]。

$$J(C)=\sum_{k=1}^{K}J(C_k)=\sum_{i=1}^{n}d_{ki}\|m_i-o_k\|^2 \tag{3}$$

$$d_{ki}=\begin{cases}1,\text{若 } m_i\in c_i\\0,\text{若 } m_i\notin c_i\end{cases} \tag{4}$$

算法流程如图 2 所示。

K-means 聚类算法可以对大型的数据集进行高

效的分类、聚类，其复杂度是 $O(nkt)$，其中 n 为数据元素的个数，k 为聚类个数，t 为迭代次数[13]。

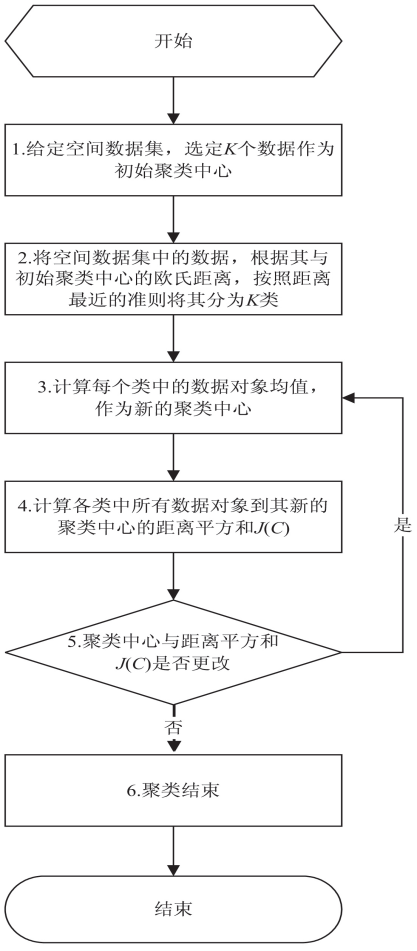

图 2 K-means 聚类算法流程

3 K-means 聚类算法在学生入学成绩分析中的应用

本次实验用统计分析软件 SPSS 对数据进行分析。

3.1 数据来源

实验数据来源某高校 2013 年、2014 年马克思主义理论专业研究生入学成绩。马克思主义理论专业学生共 115 人，其中 2013 年学生 66 人，2014 年 49 人。

3.2 数据预处理

实验中对于数据的处理采用忽略元组的方法，删除外语考试为非英语一、免试入学、专业课考试为中国近代史的学生共 31 人，得到最终实验数据情况为：马克思主义理论学生总人数为 104 人，其中 2013 年学生 62 人，2014 年 42 人。预处理结果（部分）如表 1 所示。

表 1 马克思主义理论专业研究生入学成绩

学号	[201] 英语一	[101] 政治理论	[705] 马克思主义原理	[826] 中国化马克思主义
138019020494	43	62	121	128
138019020495	53	75	129	124

续表 1

学号	[201] 英语一	[101] 政治理论	[705] 马克思主义原理	[826] 中国化马克思主义
138019020496	64	65	131	121
138019020497	42	67	128	125
138019020498	55	70	136	137
138019020499	64	72	135	134
138019020500	52	67	134	134
138019020501	58	74	138	137
138019020502	68	73	139	134
138019020503	65	70	103	127
138019020504	62	67	117	107
138019020505	48	78	114	103
138019020506	51	55	125	121
……	……	……	……	……

3.3 聚类数的处理

将处理过后的数据导入 SPSS 软件，利用 K-均值聚类。首先，将 2013 年马克思主义原理专业研究生的入学的成绩分别进行分析，初始的聚类中心随机产生，聚类数目定为 5。当迭代次数为 6 时，任何中心的最大绝对坐标更改为 0，初始中心间的最小距离则为 29.563。其次，对 2014 年马克思主义原理专业研究生入学成绩进行分析，$K=5$ 得到的初始中心间的最小距离为 41.061。

3.4 聚类结果分析

最终聚类中心与每个聚类中的案例数如表 2 和表 3 所示。

表 2 2013 年最终聚类中心及案例数

项目	聚类				
	1	2	3	4	5
[201] 英语一	44	62	57	65	54
[101] 政治理论	60	62	69	73	66
[705] 马克思主义原理	121	106	119	136	130
[805] 中国化马克思主义	125	129	113	134	130
聚类中的案例数	13	6	11	9	23

由表 2 分析可知，第一类学生共 13 人，占总学生人数的 21%，总体成绩状况较差，低分出现概率最高区域，专业课能力相对较好，英语成绩水平有待提高，培养方案中应加强对英语能力的培养；第二类学生共 6 人，占总学生人数的 9.7%，其中马克思主义原理成绩较低，其他成绩相对较好，表明该类学生马克思主义原理的知识结构相对匮乏，应加强对于专业课能力的培养，研究方向可考虑中国化马克思主义相关领域；第三类学生共 11 人，占总学生人数的 17.8%，这类学生成绩处于中间水平，专业课能力有待提高；第四类学生

共 9 人,占总学生人数的 14.5%,这类学生的总体情况稳定,成绩较为优秀,没有弱科情况; 第 5 类学生共 23 人,占总学生人数的 37%,这类学生成绩相对较好,但相比较其他学科来说英语较弱。

表 3　2014 年最终聚类中心及案例数

项目	聚类				
	1	2	3	4	5
[201] 英语一	83	48	50	59	55
[101] 政治理论	59	66	63	66	69
[705] 马克思主义原理	128	105	130	124	132
[805] 中国化马克思主义	123	126	115	96	134
聚类中的案例数	1	14	11	4	12

由表 3 分析可知,第一类学生共 1 人,占总学生人数的 2.4%,总体成绩状况较好,政治理论成绩较低,英语成绩突出,专业课能力相对较好,培养方案中应加强对政治理论能力的培养; 第二类学生共 14 人,占总学生人数的 33.3%,其中英语和马克思主义原理成绩较低,其他成绩相对较好,应加强英语以及马克思主义理论方面的研究培养,研究方向应趋向于中国化马克思主义相关领域; 第三类学生共 11 人,占总学生人数的 26.2%,这类学生英语和中国化马克思主义成绩相对较弱,应加强这两方面能力的培养,但这类马克思主义原理成绩突出,研究方向应趋向于此研究领域; 第四类学生共 4 人,占总学生人数的 9.5%,这类学生中国化马克思主义成绩较低,其他成绩相对较好应加强关于马克思主义中国化的研究; 第 5 类学生共 12 人,占总学生人数的 28.6%,这类学生成绩相对较好,英语成绩相比较低,专业课水平较高,基础扎实。

将 2013 年和 2014 年聚类情况表进行纵向分析,如图 3 所示,其中纯色填充表示 2013 年情况,宽上对角线填充表示 2014 年情况。

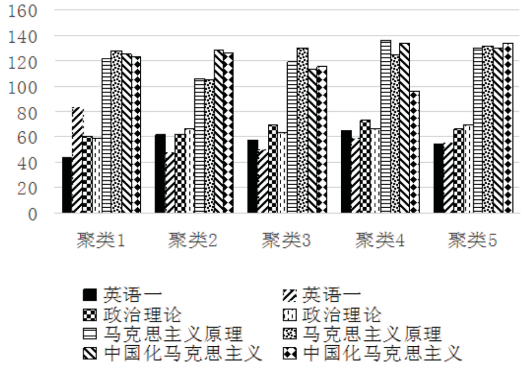

图 3　2013 年与 2014 年入学成绩聚类对比

由图 3 可得,除英语最低聚类中心 2013 年为 44 分,而 2014 年为 48 分,2013 年略低以外,其他科目成绩的最高值均在 2013 年,成绩最低值均在 2014 年,说明了 2013 年学生的学习能力,专业课功底相对高于 2014 年的学生。

4　结束语

聚类分析作为数据挖掘中的一种重要的技术手段和方法,已经广泛应用于各个领域。在教育信息化[14]的发展趋势下,数据挖掘技术应用于教育领域已经成为必然。文中通过列举法对几种常见的聚类分析算法在研究生入学成绩领域的适用性进行说明,得出划分聚类算法应用于研究生入学成绩中效果最优的结论,并运用划分聚类算法中的典型算法 K-means 进行论证,以实例证明了 K-means 算法对研究生入学成绩进行分析的可行性。分析结果科学合理地反映了学生的学习状态及学习能力,为研究生的培养,专业培养方案的制定提供了有利的依据。

参考文献:

[1] 曾　旭,司马宇.K-Means 算法在计算机等级考试成绩分析中的应用[J].软件导刊,2012,11(11):19-21.
[2] 吴夙慧,成　颖,郑彦宁,等.K-means 算法研究综述[J].现代图书情报技术,2011(5):28-35.
[3] 孙　海.层次聚类算法的改进[D].哈尔滨: 哈尔滨工程大学,2014.
[4] 杨佳润.数据挖掘之聚类分析算法综述[J].通讯世界,2017(16):291.
[5] 陈向东.数据挖掘常用聚类算法分析与研究[J].数字技术与应用,2017(4):151-152.
[6] 叶福兰.基于 K-means 均值算法的学生成绩分析—以福州外语外贸学院信息管理与信息系统专业为例[J].贵阳学院学报: 自然科学版,2017,12(3):17-20.
[7] 王　勇,刘建平,蔡长霞.一种改进的 K-means 聚类算法[J].工业控制计算机,2010,23(8):91-93.
[8] JAIN A K.Data clustering: 50 years beyond K-means[J].Pattern Recognition Letters,2010,31(8):651-666.
[9] 华婷婷.K-means 聚类算法研究[J].黄山学院学报,2013,15(5):17-19.
[10] 孙　菲,张健沛,董　野,等.基于标准偏移量的学生成绩 K-means 聚类分析算法研究[J].齐齐哈尔大学学报: 自然科学版,2015,31(2):57-64.
[11] 王　千,王　成,冯振元,等.K-means 聚类算法研究综述[J].电子设计工程,2012,20(7):21-24.
[12] 周丽华,黄成泉,王　林.一种自动模糊聚类的算法[J].统计与决策,2014(20):16-19.
[13] HAR-PELED S,MAZUMDAR S.On coresets for k-means and k-median clustering[C]//Thirty-sixth ACM symposium on theory of computing.Chicago,IL,USA: ACM,2004: 291-300.
[14] 单玉双.聚类算法在学生成绩分析中的应用研究[D].阜新: 辽宁工程技术大学,2009.