基于改进的 K-means 聚类算法的学生成绩分析

张　云

（安徽开放大学 教务处，合肥　230022）

摘要：从开放大学教务系统中选取 2021 年春季学期计算机科学与技术本科期末考试相关成绩作为数据样本，使用改进的 K-means 聚类算法对开放教育学生成绩进行聚类分析，从而获取学生成绩与年龄之间的潜在关系。研究结果表明，改进的 K-means 聚类算法可以更好地分析不同年龄段学生的相关学习数据，为教师提供更多的学生学习数据信息。教师可以根据结果采取相应的措施，从而提升学生的学习效果。

关键词：开放大学；K-means 算法；k 值选择；成绩分析

中图分类号：TP3-0　　**文献标识码**：A　　**文章编号**：2097-0625(2022)03-0092-05

一、引言

安徽开放大学的主要教育形式是现代远程教育。近年来随着互联网技术的发展，线上教学模式在开放教育中的应用日益广泛。与普通高校的全日制学生不同，开放教育学生来自各行各业，需要平衡工学矛盾；学生年龄跨度比较大，学习目的和学习需求呈现多元化。线上学习可以随时随地进行，学生可以根据自己学习情况重复观看课程视频，也可以利用碎片时间进行学习。线上学习为开放教育的学生提供了非常便利的学习方式[1]，因此学生的线上学习需求旺盛。但是，线上学习也存在一些缺点，如课堂上师生间缺乏交流互动，无法及时了解学生对知识点的掌握情况，无法直观了解学生处于哪个年龄层次等。这些信息的缺乏影响了教师对课程教学效果的判断以及制定合适的教学计划。

学生的考试成绩是检查学生学习情况以及检验教师教学成果的重要途径。通过开放大学的教务系统可以查询学生的学习成绩，包含形考成绩、卷面成绩以及综合成绩。为了解线上教学模式的效果，了解不同年龄阶段学生线上课堂知识点的掌握情况，深层次挖掘成绩数据里面所隐藏的有价值信息十分必要。本文对不同年龄段学生期末考试相关成绩进行聚类分析，挖掘学生线上学习成效与学生年龄之间的关系，为教师针对不同年龄段的学生制定个性化教学方案提供理论依据，从而进一步提升开放教育的教学质量。

二、K-means 算法

K-means 算法是一种有效的聚类划分方法 ，其主要思想是：在给定 K 值和初始簇中心的情况下，把数据对象划分到距离其最近的簇中心所代表的类簇中，所有数据对象分配完成之后，根据一个簇内的所有数据对象重新计算该类簇的中心，然后再迭代进行分配和更新簇内中心的步骤，直至簇内中心点的变化很小，或者达到指定的迭代次数[2]。

假定 $D=\{X_1,X_2,X_3,\cdots,X_n\}$ 是具有 n 个数据对象集合，其中每个数据对象都具有 m 个维度的属性。K-means 算法就是以欧式距离作为衡量数据对象间相似度的指标，将数据样本 D 中 n 个数据对象依据数据对象之间相似性划分到 k 个类簇 $C_1,C_2,C_3,\cdots,C_k$ 中，其中 $1\leqslant k\leqslant n$ 。集合中每个对象都划分到与其距离最近的簇内中心所在的类簇中[3-5]。

K-means 具体算法流程如下[6-8]：

(1)从数据对象 D 中随机选取 K 个对象作为初

收稿日期：2021-12-31

基金项目：安徽省高等学校省级质量工程教学研究项目“开放教育省管课程线上考试创新研究与实践”（项目编号：2020jyxm0268）、“基于用户体验的安徽开放大学教学教务一体化管理平台设计研究”（项目编号：2020jyxm0267）；安徽开放大学青年项目“基于物联网技术的室内消防定位技术研究”（项目编号：QN202110）

作者简介：张　云（1990－），女，安徽宣城人，助教，硕士。研究方向：数据挖掘。

始类簇中心。

(2)根据公式(1)计算当前每个簇内数据对象到簇中心的欧式距离：

$$d(X_i,C_j)=\sqrt{\sum_{t=1}^{m}(X_{it}-C_{jt})^2} \quad (1)$$

式(1)中 X_i 表示第 i 个数据对象，$1\leqslant i\leqslant n$，C_j 表示第 j 个类簇中心，$1\leqslant j\leqslant k$，X_{it} 表示第 i 个对象的第 t 个属性，C_{jt} 表示第 j 个类簇中心的第 t 个属性，$1\leqslant t\leqslant m$ 。

(3)依次比较每个数据对象到簇中心的距离，将数据对象划分到距离最近的簇内中心的类簇中，得到 k 个类簇 $\{S_1,S_2,S_3,\cdots,S_k\}$，根据公式(2)重新计算新的类簇中心，

$$C_l=\frac{\sum\limits_{X_i\in S_l}X_i}{|S_l|} \quad (2)$$

式(2)中 C_l 表示第 l 个类簇中心，$1\leqslant l\leqslant k$，$|S_l|$ 表示第 l 个类簇中数据对象的数量。

(4)重复步骤(2)和(3)直至类簇中心不变或达到指定的迭代次数。

三、基于加权的 K-means 聚类成绩数据分析

(一)数据选取

数据的质量决定了聚类模型预测结果，其涉及很多因素，比如数据的真实性、时效性、完整性等[9]。现实中，我们拿到的真实数据可能由于数据收集过程中的主客观因素，而包含了部分无效信息、失真信息或异常数据，这些数据都是造成模拟训练结果与实际情况不相符的主要因素。

从开放大学教务系统中选取了安徽开放大学计算机科学与技术本科班软件工程课程 2021 年春季学期期末考试相关成绩进行分析，所选成绩样本数据来自全省所有选择该课程的学生，共计样本成绩数据 240 条。通过 K-means 算法与加权 K-means 算法分析结果进行比较，选择更适用于进行成绩分析的方法，分析不同年龄层次的开放教育学生的考试成绩与学生学习效果之间的关系。为了分析结果更加真实有效，我们选取含有成绩数据的学生成绩进行模拟训练。

(二)K 值的选取

初始聚类数 K 值的选取，若偏离真实值，则会直接影响数据样本聚类的效果。本文通过“手肘法”来确定初始聚类数 K 值。“手肘法”是通过聚类数 K 与误差平方和(SSE)之间对应的关系，在 K 与 SSE 的关系中，我们可以通过观察拐点来确定 K 值。随着 K 值的增加，并且 K 值未达到真实值之前，SSE 值下降的幅度会很大，数据对象的聚类会更加明显，数据对象之间的聚合程度会逐渐增大；当 K 值达到聚类数的真实值时，SSE 值的下降幅度会迅速减缓；而当 K 值大于真实值并且继续增加时，SSE 值会逐渐趋于平缓[9-10]。

K 值确定步骤：

(1)通过式(1)确定每个类簇中的对象到簇内中心的欧式距离；

(2)计算每个类簇内的误差平方和；

(3)计算当前数据对象的总误差和；

(4)K 为 $K+1$ 时，重复计算当前数据样本对象的 SSE 值，确定 K 值。

误差平方和公式：

$$SSE=\sum_{i=1}^{K}\sum_{X\in C_i}|X-C_i|^2 \quad (3)$$

式(3)中 K 为聚类的数量，X 为当前簇 C_i 中的数据对象，c_i 为当前簇 C_i 的中心。

使用 SPSS 软件的 K 聚类分析功能对课程成绩进行聚类分析，聚类数 K 从 2 增加到 11 的聚类结果如图 1 所示，可以看出曲线在 $K=7$ 时有明显的拐点，即“手肘”，根据“手肘法”可以确定该数据的真实聚类数应为 7。

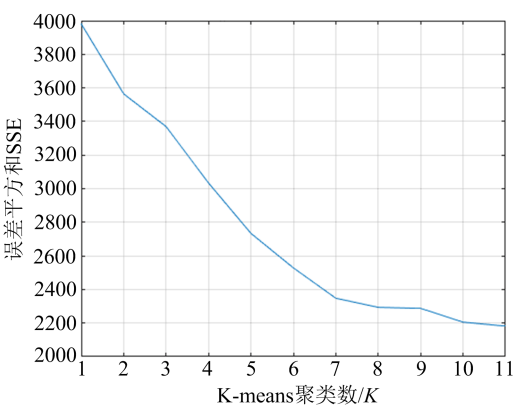

图 1　手肘图像结果

(三) K-means 权重选取

从公式(1)可知，如果数据对象的某一属性变化范围大，那么在计算聚类成员到聚类中心的距离时，它的权重就会相对较高，最终会导致聚类中心更偏向该属性。为更直观说明这一问题，本文实验使用 K-

means 算法对学生的期末考试卷面成绩和年龄进行聚类分析，结果如图 2 所示。

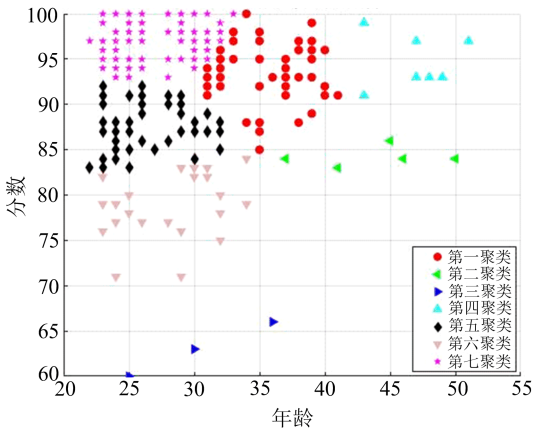

图 2　K-means 聚类算法成绩分析结果

从图 2 可以看出聚类有明显的不合理之处，样本数据的聚类中心倾向于成绩属性，在年龄属性上跨度较大。以“★”所代表聚类为例，该聚类左右两侧存在明显的界限，该聚类中右侧的样本数据更应该与“·”所代表的聚类归为一类。这是因为在成绩与年龄二维坐标中，成绩的取值范围是 0～100，而年龄的取值范围为 20～51。成绩的取值范围约为年龄取值范围的 3 倍，由式(1)可知，使用普通 K-means 算法进行聚类分析时，成绩维度在欧氏距离中占比更重，聚类中心会倾向于成绩属性。

通过式(1)计算欧氏距离时，若某一个属性的取值范围明显大于其他属性，那么该属性在欧氏距离中比重就会偏重。为了消除这种现象，可以将样本数据的每个属性调整到相同的取值范围，即在进行聚类分析前，给每一个属性乘上不同的系数 a_t 得到 X'_i，其中 $X'_i=\{a_1X_{i1},a_2X_{i2},\cdots,a_{t-1}X_{i(t-1)},a_tX_{it}\}$。

$$a_t=\frac{\max(s_1,s_2,s_3\cdots,s_{t-1},s_t)}{s_t} \tag{4}$$

公式(4)中 s_t 为第 t 个属性的取值范围。

故本实验加权后采用的欧氏距离计算公式为：

$$\begin{aligned} d(X'_i,C'_j)&=\sqrt{\sum_{t=1}^{m}((X'_{it}-C'_{jt})^2)}\\ &=\sqrt{\sum_{t=1}^{m}a_t^2(X_{it}-C_{jt})^2}\\ &=\sqrt{\sum_{t=1}^{m}k_t(X_{it}-C_{jt})^2} \end{aligned} \tag{5}$$

公式(5)中 $k_t=a_t^2$，k_t 为每个维度的权值，X'_i 为处理后数据对象，X'_{it} 为处理后数据对象 X'_i 的第 t 个属性，C'_j 为处理后数据对象类簇中心，C'_{jt} 为类簇中心 C'_j 的第 t 个属性，X_i 为原始数据对象，X_{it} 为原始数据对象 X_i 的第 t 个属性，C_j 为原始类簇中心，C_{jt} 为原始类簇中心 C_j 的第 t 个属性，m 表示数据对象有 m 个属性值。

使用公式(5)作为欧氏距离公式相当于对每个属性赋以不同的权重，使用加权后的算法重新对包含学生成绩与年龄的样本数据进行聚类分析。结果如图 3 所示。

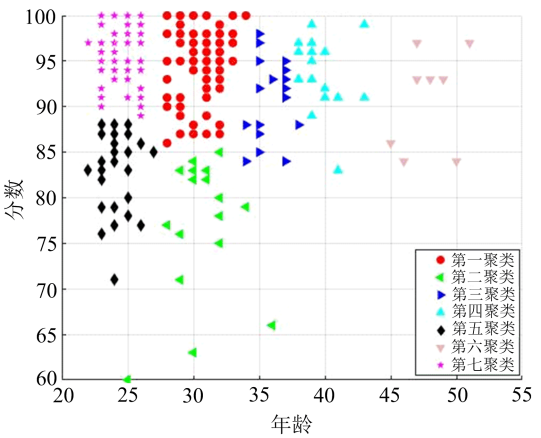

图 3　加权的 K-means 聚类算法成绩分析结果

从图 3 可以看出，样本数据不再倾向于年龄属性的聚类，“★”所代表的聚类不再有明显的界限，聚类结果更加科学。

通过图 2 和图 3 中 K-means 算法和加权的 K-means 算法聚类结果对比可知，对于相同的成绩样本数据对象，加权的 K-means 算法聚类比普通的 K-means 算法聚类效果更加明显且更加合理。

四、结果分析

根据前文结论可知，使用加权 K-means 算法能更加准确探索学生的年龄与期末考试的成绩是否存在关联。前文为方便直接观察两种聚类结果，文中只选取了学生的综合成绩与年龄二维样本数据进行聚类分析。为了获取学生的年龄、形考成绩、卷面成绩和综合成绩的内在关联，使用加权的 K-means 算法对样本数据进行聚类分析，样本数据包含学生年龄、形考成绩、卷面成绩、综合成绩四个维度，聚类分析后，结果如表 1～3 所示。

从表 2 最终聚类中心间的距离可以看出最小的间距为 18.986，聚类中心之间的距离较大，聚类结果较好。

表 1　最终聚类中心

	聚类						
	1	2	3	4	5	6	7
形考成绩	92	28	94	91	89	91	66
卷面成绩	97	76	97	79	82	94	93
综合成绩	95	62	96	83	84	93	85
年龄	25	28	31	25	32	41	27

表 2　最终聚类中心间的距离

聚类	1	2	3	4	5	6	7
1		75.828	18.986	21.076	27.650	49.016	28.472
2	75.828		78.337	67.477	67.003	83.231	48.151
3	18.986	78.337		29.017	19.591	30.526	33.019
4	21.076	67.477	29.017		21.047	51.789	28.730
5	27.650	67.003	19.591	21.047		31.542	29.479
6	49.016	83.231	30.526	51.789	31.542		50.504
7	28.472	48.151	33.019	28.730	29.479	50.504	

表 3　聚类案例数

聚类	1	2	3	4	5	6	7	有效数
案列数	63	12	59	27	33	32	14	240

对表 1 最终的聚类中心进行分析可以发现，聚类 1 和聚类 4 对比可以看出这两个聚类学生较年轻，形考成绩较高，但卷面成绩和综合成绩差距较大，说明这个年龄段的学生学习时间比较充裕，有较多的时间完成课后作业。但卷面成绩有高有低，说明他们对知识的掌握程度有好有坏，授课老师可以针对这个年龄段的学生加强知识点的巩固。

聚类 2 和聚类 7 中的学生年龄相仿，但形考成绩都不理想，说明这类学生忽视平时作业的完成，可能是因为这个年龄段的学生平时学习时间少，老师可以在平时作业方面多给予指导，提升他们的形考成绩。

聚类 3 和聚类 5 中学生的年龄相仿，三项成绩都很平均，说明这类学生学习有一定的主动性，老师可以从平时的作业中提前了解他们的学习情况。如果他们平时的作业完成度好，那么他们的期末考试卷面成绩也不会差；如果他们的平时成绩完成得不好，老师可以适当地对他们提出要求。

聚类 6 中的三项成绩都较高且较为平均，说明在这个年龄段的学生既有一定的时间完成课后作业，同时也注重知识点的掌握与巩固，可能是因为这个阶段的学生对知识更加渴望，老师在授课过程中可以适当地降低对他们的关注，将精力更多地放在其他学生身上。

五、结语

加权的 K-means 算法适合用于样本数据对象各个维度的数据范围相差较大或明确知道某个维度的数据更重要的场景。使用加权的 K-means 算法可以更好分析学生的成绩与年龄的相关性，获得更加准确的结果，为教师提供更多的学生数据信息。通过对聚类结果的分析，教师可以更加了解每个年龄段学生优势和弱点，有的放矢，因材施教，更加精准地指导每个学生的学习。在花费同等精力的同时，更加高效地提升学生的整体学习效果。

参考文献：

[1] 邓兴平，王金鑫.“互联网＋”背景下会计课程自主学习模式探索[J].新教育时代电子杂志(教师版)，2021(27)：80-81.

[2] 熊忠阳，陈若田，张玉芳.一种有效的 K-means 聚类中心初始化方法[J].计算机应用研究，2011，28(11)：4188-4190.

[3] 蔺小清.大数据时代 K-means 聚类算法应用于在线学习行为研究[J].电子设计工程，2021，29(18)：181-184.

[4] 杨俊闯，赵超.K-Means 聚类算法研究综述[J].计算机工程与应用，2019，55(23)：7-14，63.

[5] 王慧，申石磊.一种改进的特征加权 K-means 聚类算法[J].微电子学与计算机，2010，27(7)：161-163，167.

[6] 思绪无限.K-means 聚类算法详解[DB/OL].(2018-5-16)[2021-12-31].https://blog.csdn.net/qq_32892383/article/details/80107795.

[7] HAN Jiawei，KAMBER M.数据挖掘概念与技术[M].北京：机械工业出版社，2012：293.

[8] 李晓明.K-means 类型变量加权聚类算法的研究与实现[D].哈尔滨：哈尔滨工业大学，2006：18.

[9] 钟文精，焦中明，蔡乐.基于 K-means 算法的学生成绩聚类分析[J].教育信息技术，2021(5)：56-58.

[10] 王建仁,马鑫,段刚龙.改进的 K-means 聚类 k 值选择算法[J].计算机工程与应用,2019,55(8):27-33.

Student Performance Analysis Based on the Improved K-means Clustering Algorithm

ZHANG Yun

(Teaching Affairs Office, Anhui Open University, Hefei 230022, China)

Abstract: The scores of the final exams of undergraduate Computer Science and Technology classes in the spring semester of 2021 are selected as data samples from the educational administration system of Anhui Open University, and the improved K-means clustering algorithm is used to cluster and analyze the results in order to obtain potential relationships between student performance and age. The results show that the improved K-means clustering algorithm can better analyze learning data related to students of different age groups and provide teachers with more information about students' learning. Thus, teachers can take appropriate measures based on the results to improve students' learning effect.

Keywords: Open University; K-means algorithm; K value selection; performance analysis

[责任编辑 李潜生]

(上接第 75 页)

[6] 戴锦华.经典电影十八讲:镜与世俗神话[M].北京:中信出版社,2014:130.

[7] 戴锦华.电影批评[M].北京:北京大学出版社,2015:6.

[8] 巴赫金.陀思妥耶夫斯基诗学问题[M].白春仁,顾亚铃,译.北京:三联书店,1988:29.

[9] 吉尔·德勒兹.电影 2:时间-影像[M].谢强,译.长沙:湖南美术出版社,2004:121.

[10] 埃尔塞·瑟托马斯.非线性叙事的回归/转向:反事实历史和环形叙事[J].张振,译.当代电影,2020(4):26.

[11] CASETTI F. The Lumière Galaxy: 7 Key Words for the Cinema Today[M]. New York: Columbia University Press, 2015:213.

The Layered Structure and Polyphonic Narrative of *Little Women*

LIN Wei

(Foreign Language School, Guangzhou University of Chinese Medicine, Guangzhou 510006, China)

Abstract: The film *Little Women*, adapted and directed by Greta · Gerwig in 2019, reinterprets the original work under a non-linear narrative structure, thereby reshaping the audience's empathic identification experience with the characters in the work. Nonlinear narrative time challenges the linear time characterized by plot drive and causality, and shapes a spatialized narrative time experience. The narrative layer of the film constitutes a structure of Bakhtin-style "polyphonic dialogue", which not only points at the protagonist Jo's identity as a writer in the conception of the novel, but also implies the functioning of the "film machine" outside the film screen. Therefore, the film is endowed with multiple meanings of fiction, expressing the symbiotic relationship between literary consumption behavior and narrative economy in a media self-referential way, as well as the economic drive behind the screen narrative.

Keywords: *Little Women*; non-linear narrative; narrativelayer; polyphonic dialogue; media self-reflectivity

[责任编辑 夏强]