Computer Science and Application 计算机科学与应用, 2023, 13(6), 1179-1187
Published Online June 2023 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2023.136115

Hans汉斯

基于K-Means聚类模糊算法的学生特征聚类研究

周子安，薛欢欢，杨雨潇，李旭东

嘉兴南湖学院信息工程学院，浙江 嘉兴

收稿日期：2023年5月2日；录用日期：2023年5月31日；发布日期：2023年6月7日

摘　要

聚类分析是一种关于大量数据挖掘的重要技术。而对于学生存在的个性化差异，各自会表现出不同的特征。以不同学生所体现的特征来实现在教学过程中对学生因材施教，其特征类型的统计是一种大量数据处理的操作。所以，对于这种维度高、数据集大的海量数据，可以利用K-means算法结合数学模型给出聚类的学生特征，再引入模糊数学中的隶属度，来提取出更具解释性的聚类特征。首先，介绍了K-means算法的思想原理并分析其优缺点，并引入K-means++算法；其次，针对初始聚类中心点的选取和K值的确定；然后通过对大学生计算机专业的课程成绩为例，进行聚类分析，其实验结果表明，基于K-means聚类模糊算法的学生特征聚类相比基本的K-means或系统聚类，聚类结果上体现出更好的科学性和解释性；最后，对K-means算法技术关于学生学习成绩和效率进行展望。同时，本文利用蒙特卡罗和用频数来确定隶属度的想法，是团队首创，确保了创新性。

关键词

学生特征提取，K-Means算法，K-Means++，模糊聚类，数据处理

Research on Student Feature Clustering Based on K-Means Clustering Fuzzy Algorithm

Zi'an Zhou, Huanhuan Xue, Yuxiao Yang, Xudong Li

School of Information Engineering, Nanhu University, Jiaxing Zhejiang

Received: May 2nd, 2023; accepted: May 31st, 2023; published: Jun. 7th, 2023

文章引用: 周子安, 薛欢欢, 杨雨潇, 李旭东. 基于 K-Means 聚类模糊算法的学生特征聚类研究[J]. 计算机科学与应用, 2023, 13(6): 1179-1187. DOI: 10.12677/csa.2023.136115

Abstract

Cluster analysis is an important technique for large-scale data mining. And for the personalized differences that students have, they will exhibit different characteristics. Teaching students according to their aptitude in the teaching process based on the characteristics reflected by different students is a large-scale data processing operation. Therefore, for such high-dimensional and large datasets, K-means algorithm can be combined with mathematical models to provide clustered student features, and membership degree in fuzzy mathematics can be introduced to extract more explanatory clustering features. Firstly, the concept and principle of the K-means algorithm were introduced, and its advantages and disadvantages were analyzed, followed by the introduction of the K-means++ algorithm; secondly, regarding the selection of initial clustering center points and the determination of K values; then, taking the course grades of college students majoring in computer science as an example, clustering analysis was conducted. The experimental results showed that the student feature clustering based on K-means clustering fuzzy algorithm showed better scientific and interpretive results compared to basic K-means or system clustering; Finally, the prospects of K-means algorithm technology for students' academic performance and efficiency are presented. Also, the idea of using Monte Carlo algorithm and using frequency to determine affiliation in this paper is a first for the team and ensures innovation.

Keywords

Student Feature Extraction, K-Means Algorithm, K-Means++, Fuzzy Clustering, Data Processing

1. 引言

在高校教学中，为了更好地指导学生，我们需要根据学生的学习情况进行个性化的指导[1]。评价毕业生情况的一个重要指标是各门学科的成绩，这也是评价素质和效果的依据。十一名学员的考试成绩是衡量他们学习表现和教师教学质量的重要指标，可以通过学校的教务系统查询到他们的各项学习成绩，包括日常表现、次考试成绩以及综合素质评估结果。为了更准确地反映学生成绩的分布情况，我们使用K-means 聚类计算对学员的多门学科成果加以分类[2]。通过对比不同类别学员的业绩，我们可以给出相应的指导建议，以便更好地评估学生的学习情况。此外，这种方法还可以为教学质量的反馈提供理论支撑[3]。

然而，K-means 方法的最终聚合效果受到早期聚类中心的选择性的限制较大，而且很难准确地区分模糊类的属性。而对于模糊综合评价而言，我其中的隶属函数的赋予是个常见的痛点，而常用的指派法又太过于主观且无法充分利用已知样本数据[4]。对于上述二点，可以选择使用蒙特卡洛算法来大量尝试K-means 初始聚类中心的选择，我并通过纪录其各个属性被划分到不同聚类的频率，来模拟一个隶属度，以此达到处理模糊聚类关系的目的[5]。

2. K-Means 算法

K-means 技术是一种有效的聚类分配技术，它的主要思路是：在给定 K 值和最初始簇核心的情况下，

将各种数据内容分配到最近的簇核心所表示的类簇中，然后按照簇内的各种信息内容再次计算出该类簇的核心，并不断迭代，直至簇内核心的变异极小，甚至超过规定的迭代数量，从而实现有效的聚类分配(见图 1)。

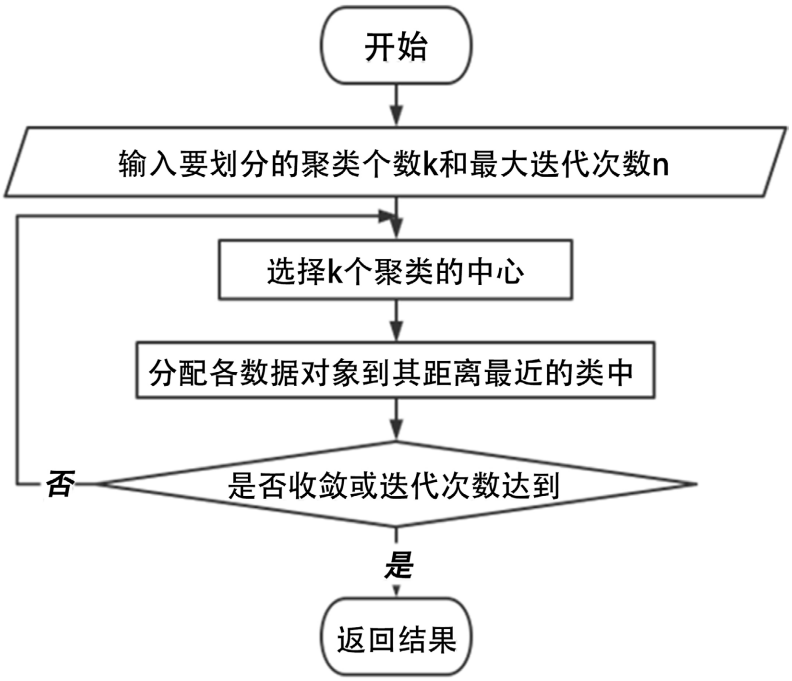

Figure 1. The flowchart of K-means algorithm
图 1. K-means 算法流程图

优点：

1) 该算法简单快速；

2) 对于处理大型数据集，该算法具有相对较高的性价比。

缺点：

1) 用户必须提前准备好生成的簇的数量 k，以确保最佳性能；

2) 其最终的分类结果，会因初始聚类中心的选择不同，而发生较大的变化；

3) 对于孤立点的数据极为敏感。

4) 聚类结果，非此即彼，对于某些很难明确划分的属性来说，缺少可解释性。

K-means++算法在选择初始聚类中心时，应尽可能保持它们之间的距离较远，以便更好地捕捉和分析特征[6]。

步骤 1：随机选择一个样本作为第一聚类中心；

步骤 2：首先，测量每个采样点与当前现有聚类中心之间的最短时间。该值越大，采样点被选择为聚类中心的概率就越大；其次，使用轮盘赌方法(基于概率进行提取)从中选择最近的聚类中心；

步骤 3：重复步骤 2，直到选择了 k 个聚类中心。如果选择了两个或多个聚类中心，则取先前确定的聚类中心的重心。在选择了起点之后，仍然使用标准的 K-均值算法。

3. 数据分析

数据源：学生成绩数据集.xlsx

(一) 算法配置：

算法：聚类分析(K-means)

变量：

{离散数学/4.0，高等数学 A2/4.0，高级语言程序设计实训/1.0，线性代数/3.0，专业认识实习(计算机类)/0.5，思想道德与法治/2.0，体育/1.0，大学生职业生涯规划与就业指导/0.5，创业基础/0.5，大学物理 A/4.0，大学英语 2/3.0，“互联网+”与大学生创新创业/1.0，大学生心理健康教育/1.5，大学英语 1/3.0，高等数学 A1/5.0，高级语言程序设计/4.0，计算机导论/1.0，军事理论/0.5，中国近现代史纲要/2.0，应用文写作/2.0，形势与政策/1.0，体育/1.0.1，选修课/2.0，加权平均分}

参数：聚类个数：{4}。

(二)步骤

1) 先将数据集带入系统聚类，得到不同聚类数下的误差平方和。再将误差平方和绘制成折线图，根据肘部定理，确定聚类数。

2) 将数据带入到 K-means 聚类模型，通过蒙特卡罗随机选取初始点来进行聚类，并将每次聚类后该点所属类别记录下，最后求得各个样本在每个类中出现的频率。

3) 把每个样本对于每个类别的频率作为该样本对于每个聚类的隶属度。

4) 对分析进行综述。

(三)分析结论：

聚类分析基于数据特征，将所有样本划分为几类：2 聚类结果共分为 4 类，聚类类别_1 的频数为 12，所占百分比为 42.857%；聚类类别_2 的频数为 11，所占百分比为 39.286%；聚类类别_3 的频数为 4，所占百分比为 14.286%；聚类类别_4 的频数为 1，所占百分比为 3.571%。

(四)详细结论分析：

1) 聚类数对比图(见图 2)

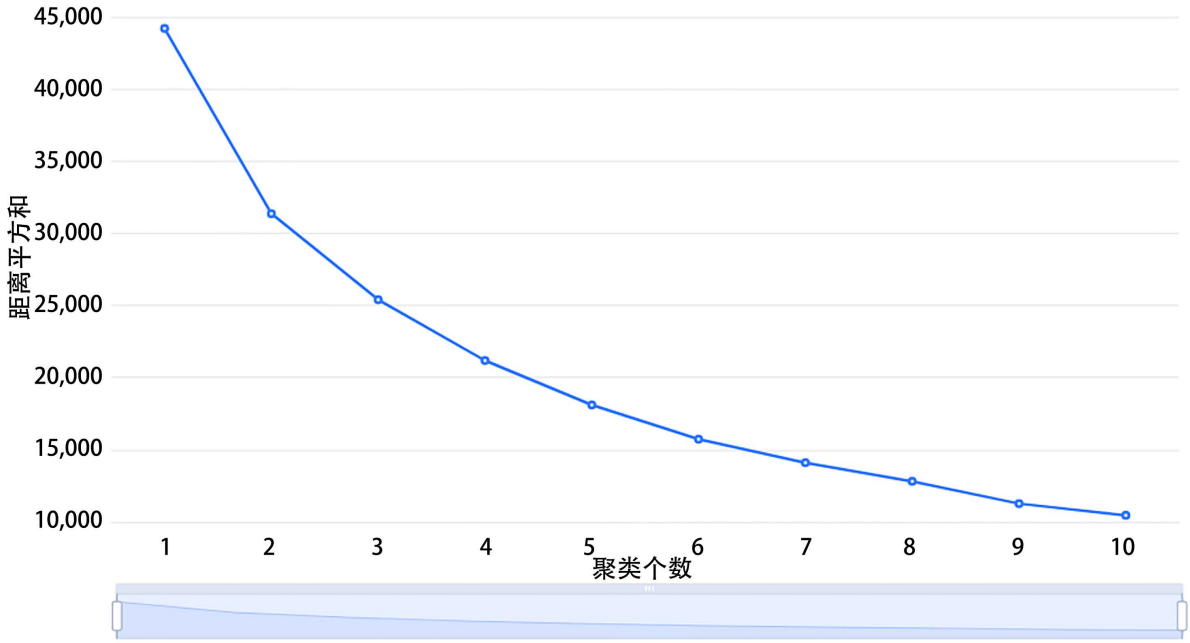

Figure 2. The result graph of elbow method
图 2. 肘部法则结果图

这张图用于确定最佳的聚类数量，横坐标表示聚集的数量，纵坐标表示 K 平均聚集的损失函数，它反映了每个数据到类别中心的时间平方和，也意味着误差平方和(越小表明聚集效果越好)。可以通过“坡度趋于平缓”找出最佳的类簇数量。

这里选择了 4 为聚类数，以同时满足数据严谨性与可解释性。

2) 字段差异分析(详见表 1)

Table 1. Field difference analysis display table
表 1. 字段差异分析展示表

	聚类类别(平均值 ± 标准差)				F	P
	类别 1 (n = 12)	类别 2 (n = 11)	类别 3 (n = 4)	类别 4 (n = 1)		
离散数学/4.0	72.333 ± 5.479	80.273 ± 6.78	57.375 ± 5.406	63.0 ± nan	15.167	0.000***
高等数学 A2/4.0	81.917 ± 5.022	93.227 ± 5.027	69.5 ± 4.435	80.0 ± nan	24.962	0.000***
高级语言程序设计实训/1.0	82.458 ± 3.928	88.864 ± 4.523	82.625 ± 1.75	87.5 ± nan	5.625	0.005***
线性代数/3.0	71.667 ± 5.844	85.682 ± 6.063	58.75 ± 8.292	61.0 ± nan	22.247	0.000***
专业认识实习(计算机类)/0.5	89.042 ± 5.887	89.864 ± 5.404	83.75 ± 7.5	80.0 ± nan	1.759	0.182
思想道德与法治/2.0	81.583 ± 5.329	82.773 ± 6.436	75.125 ± 3.25	70.0 ± nan	3.117	0.045**
体育/1.0	74.667 ± 8.669	75.636 ± 8.31	74.0 ± 5.715	0.0 ± nan	26.887	0.000***
大学生职业生涯规划与就业指导/0.5	89.667 ± 6.103	92.227 ± 4.86	87.5 ± 0.0	83.5 ± nan	1.518	0.235
创业基础/0.5	84.917 ± 2.961	87.409 ± 4.2	87.375 ± 5.422	80.0 ± nan	1.758	0.182
大学物理 A/4.0	68.542 ± 6.021	81.818 ± 5.001	60.25 ± 2.5	64.0 ± nan	21.754	0.000***
大学英语 2/3.0	70.792 ± 5.75	76.955 ± 8.487	59.125 ± 4.871	62.0 ± nan	7.162	0.001***
“互联网+”与大学生创新创业/1.0	91.208 ± 4.869	91.591 ± 3.917	85.5 ± 2.309	87.5 ± nan	2.379	0.095*
大学生心理健康教育/1.5	89.042 ± 4.943	89.864 ± 4.237	91.25 ± 4.33	83.5 ± nan	0.828	0.491
大学英语 1/3.0	66.125 ± 5.301	71.182 ± 8.483	56.0 ± 6.151	61.5 ± nan	4.964	0.008***
高等数学 A1/5.0	86.208 ± 5.695	92.409 ± 4.466	66.75 ± 10.874	77.0 ± nan	17.649	0.000***
高级语言程序设计/4.0	70.625 ± 9.088	84.182 ± 7.366	65.875 ± 7.099	75.5 ± nan	7.413	0.001***
计算机导论/1.0	83.917 ± 5.076	86.364 ± 6.017	77.125 ± 1.493	79.5 ± nan	3.301	0.037**
军事理论/0.5	92.667 ± 5.297	89.455 ± 15.029	94.75 ± 5.439	98.0 ± nan	0.429	0.734
中国近现代史纲要/2.0	83.292 ± 7.791	87.545 ± 7.885	75.0 ± 4.262	65.0 ± nan	4.809	0.009***
应用文写作/2.0	78.375 ± 4.987	79.864 ± 4.833	70.125 ± 2.658	62.0 ± nan	7.963	0.001***
形势与政策/1.0	89.0 ± 4.729	89.5 ± 5.705	84.75 ± 5.5	80.0 ± nan	1.703	0.193
体育/1.0.1	70.083 ± 7.537	71.091 ± 7.035	72.75 ± 7.399	61.0 ± nan	0.725	0.547
选修课/2.0	85.958 ± 6.017	94.045 ± 2.423	89.375 ± 3.75	87.5 ± nan	6.106	0.003***
加权平均分	77.725 ± 2.112	84.785 ± 2.609	68.988 ± 0.821	70.084 ± nan	58.292	0.000***

注：***、**、*分别代表 1%、5%、10%的显著性水平。

上表展示了定量字段差异性分析的结果，包括均值 ± 标准差的结果、F 检验结果、显著性 P 值。

- 分析每个分析项的 P 值是否显著(P < 0.05)。
- 如果它是显著的，并且原始假设被拒绝，则意味着两个数据集之间存在显著差异。可以使用平均值法±标准差来分析差异，反之亦然，这表明数据没有差异。

3) 聚类汇总(见表 2)

Table 2. Clustering summary
表 2. 聚类汇总

聚类类别	频数	百分比%
聚类类别_1	12	42.857
聚类类别_2	11	39.286
聚类类别_3	4	14.286
聚类类别_4	1	3.571
合计	28	100

上表展示了模型聚类的结果，包括频数、所占百分比。

4) 聚类汇总图(见图 3)

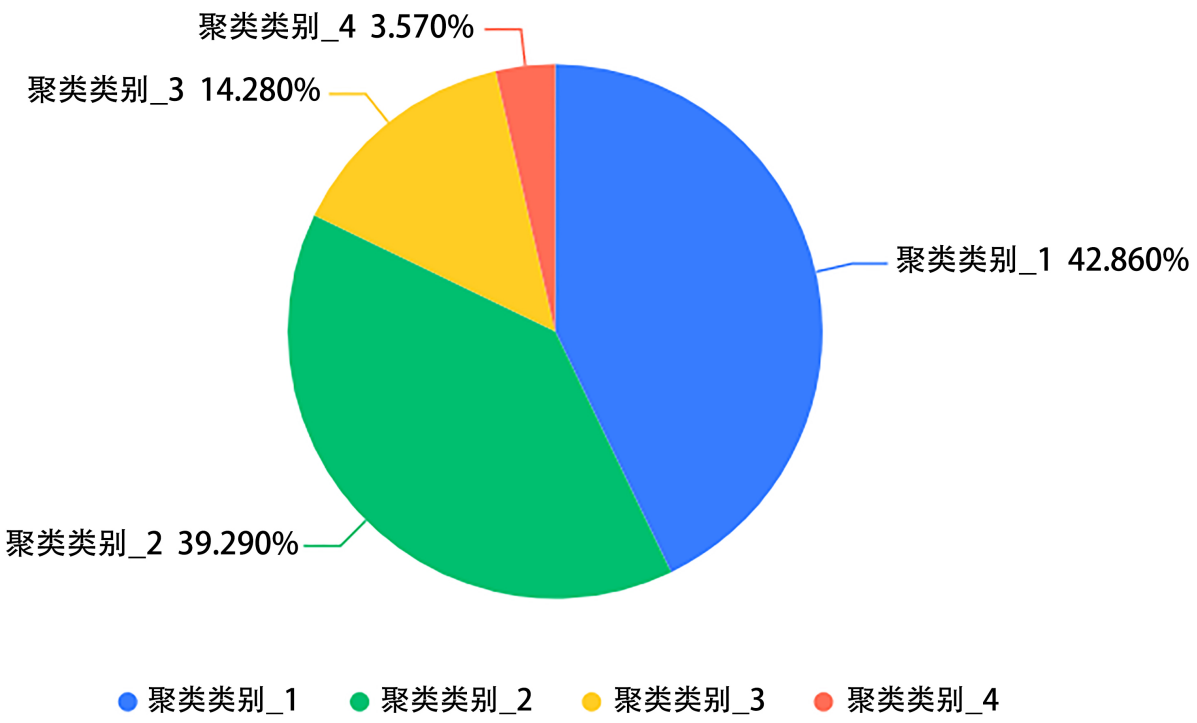

Figure 3. Clustering summary chart
图 3. 聚类汇总图

上图以可视化的形式展示了模型聚类的结果，包括频数、所占百分比。

5) 聚类中心点坐标(见表 3)

Table 3. Clustering centroid coordinates
表 3. 聚类中心点坐标

聚类种类	中心值_离散数学/4.0	中心值_高等数学A2/4.0	中心值_高级语言程序设计实训/1.0	中心值_线性代数/3.0	中心值_专业认识实习(计算机类)/0.5	中心值_思想道德与法治/2.0	中心值_体育/1.0	中心值_大学生职业生涯规划与就业指导/0.5	中心值_创业基础/0.5	中心值_大学物理A/4.0	中心值_大学英语2/3.0	中心值_“互联网+”与大学生创新创业/1.0	中心值_大学生心理健康教育/1.5	中心值_大学英语1/3.0
1	72.33333333	81.91666667	82.45833333	71.66666667	89.04166667	81.58333333	74.66666667	89.66666667	84.91666667	68.54166667	70.79166667	91.20833333	89.04166667	66.125
2	80.27272727	93.22727273	88.86363636	85.68181818	89.86363636	82.77272727	75.63636364	92.22727273	87.40909091	81.81818182	76.95454545	91.59090909	89.86363636	71.18181818
3	57.375	69.5	82.625	58.75	83.75	75.125	74	87.5	87.375	60.25	59.125	85.5	91.25	56
4	63	80	87.5	61	80	70	0	83.5	80	64	62	87.5	83.5	61.5

6) 聚类散点图(见图 4)

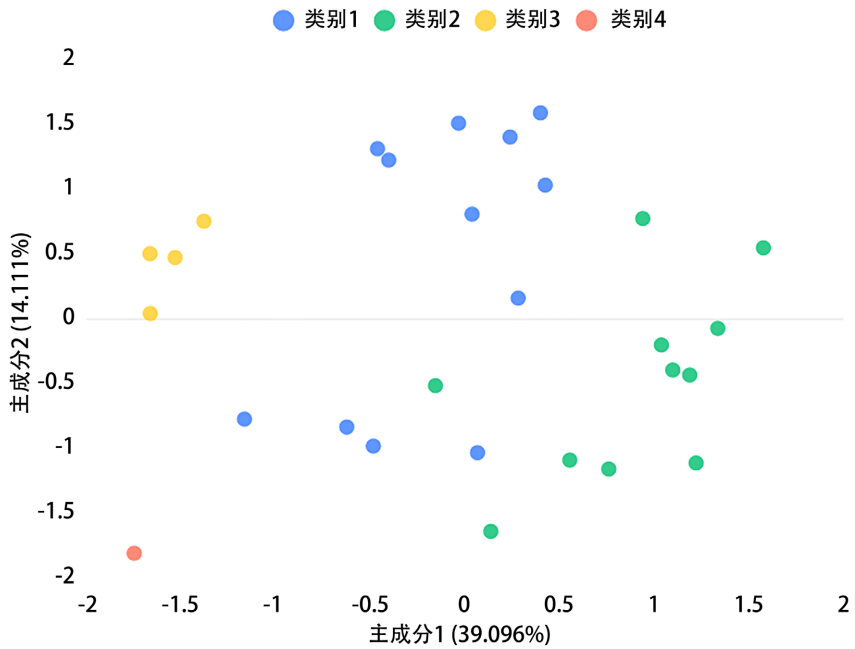

Figure 4. Clustered scatter plots
图 4. 聚类散点图

7) 评价指标

轮廓系数	DBI	CH
0.173	1.244	8.69

- 轮廓系数：这是一个重要的参数，用于测量样本集的样本，它可以反映出样本之间的相似性和差异性。轮廓系数的取值范围一般在[−1, 1]之间，越接近同一类别的样本，其轮廓系数就越大，从而提高聚类效果。
- DBI (Davies-boudin)：这个指标用来衡量两个簇之间的距离，它是一个比值，表示簇内距离与簇间距离之比该指标越小表示聚类效果越好。
- 啊 CH (Calinski-Harbasz Score)：透过统计类内各点与类中央的间距平方和，可以衡量出类内的紧密集程度(分母)，而透过统计类间的中心线与类别中心线的间距平方和，可以衡量出类别的分离度(分母)。CH 指标是由分离度与紧密度的差值来决定的，CH 越大，表明聚类效果越好。

8) 最终聚类模糊隶属度(见表 4)

Table 4. Final clustering fuzzy affiliation
表 4. 最终聚类模糊隶属度

No.	对于聚类 1 的隶属度	对于聚类 2 的隶属度	对于聚类 3 的隶属度	对于聚类 4 的隶属度
1	0.25769249	0.257814858	0.28553267	0.198959983
2	0.236975068	0.293994018	0.265583694	0.20344722
3	0.248396216	0.261694314	0.287690594	0.202218876
4	0.293970425	0.214650808	0.252263466	0.239115301
5	0.28644898	0.217622266	0.258563832	0.237364922
6	0.287514435	0.222523967	0.268911784	0.221049814
7	0.281133194	0.233773424	0.280223965	0.204869417
8	0.26393656	0.238065838	0.292379075	0.205618527
9	0.254470552	0.245942086	0.295959382	0.203627981
10	0.258946804	0.240510779	0.290941097	0.20960132
11	0.244370558	0.198780291	0.223515818	0.333333333
12	0.292946666	0.219050334	0.249631769	0.23837123
13	0.280947928	0.249229671	0.250497553	0.219324848
14	0.289525416	0.214993002	0.254016219	0.241465363
15	0.249381716	0.274218714	0.25057035	0.22582922
16	0.267589401	0.240500172	0.259247252	0.232663175
17	0.244033182	0.286753026	0.261783564	0.207430228
18	0.292100176	0.233674	0.255499687	0.218726136
19	0.254723707	0.274152959	0.265132144	0.205991189
20	0.276613307	0.252039895	0.255178638	0.216168159
21	0.293004385	0.217560427	0.250760627	0.238674561
22	0.260408919	0.257311484	0.287397148	0.194882449
23	0.239823391	0.298044027	0.263804553	0.198328028
24	0.237876087	0.285521563	0.268407547	0.208194803
25	0.242140802	0.289130421	0.264071381	0.204657396
26	0.286825326	0.239103968	0.26887611	0.205194596
27	0.283612019	0.244384127	0.252652729	0.219351124
28	0.264400384	0.277873118	0.263386675	0.194339823

4. 模型总结及评价

通过对于聚类后的四个类赋予基础解释性有：

类别 1：各科成绩在班级中游，有部分学科偏弱，需要加强学习薄弱学科便能更上一层楼。

类别 2：各学科成绩十分优秀，尤其是学分高的科目。这类学生稳定优秀，善于抓住主要矛盾

类别 3：学业成绩不理想，分数相对落后。这类学生需要老师引导兴趣或给予帮助。

类别 4：学习成绩比较偏科，部分学科成绩很差，但也有强势学科。老师应帮助引导，夯实弱项，发展强项。

上述这种由基础 K-means 引出的解释性分类难以考虑到极端情况，且单一的划分并不能做到很好的因材施教[7]。所以在此基础上我们通过引入模糊的隶属度概念，使一个学生可以同时隶属于多个类别，这样就产生的类别上更多的细分，同时在数据上教育者也更好的能分析该学生情况。

比如，某些学生因转专业或身体原因，部分课没有成绩，对于现有常用到 K-means 聚类，可能就误将其判为偏科。或者某个学生因竞赛或课程安排等原因，上学期的成绩与下学期的成绩很不一致，对于原有的 K-means 聚类也很难给出可信的解释。

或以上述第 6 号样本为例，在原 K-means 下被定义为第 4 类别，但实际上其对于第 4 类别的隶属度与对第 2 类别的隶属度几乎没有差异。且第 4 类别仅此一个样本，这就使得其原模型很难有好的解释性。

但增加了隶属度概念后，原本盖棺定论的“误判”会因隶属度的存在，稀释了特殊值或离群点对 K-means 的不利影响，进而有了可变性灵活性，使得聚类结果更具解释性。

5. 结语

K-means 算法是一个极其经典的聚类算法，自提出以来，以其思想简单、聚类速度快、结果良好而得到广泛应用。K-means 算法具有一些显著的优势，它能够有效地收敛域全局最优解，而且不会局限于一个局部最优解。但在这里，K-means 的“非此即彼”的属性，容易产生小因子的偏差。在此基础上，我们引入蒙特卡罗及模糊隶属度，使得聚类的结果更具解释性，且有更多细分。这样就能更好地通过数学模型实现“因材施教”的目的。

参考文献

[1] 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.

[2] 刘凤, 戴家佳, 胡杨. 基于局部密度离群点检测 K-means 算法[J]. 重庆工商大学学报(自然科学版), 2021, 38(4): 30-35.

[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] 王森, 刘琛, 邢帅杰. K-means 聚类算法研究综述[J]. 华东交通大学学报, 2022, 39(5): 119-126. https://doi.org/10.16749/j.cnki.jecjtu.20220914.001

[5] 蒋林岑, 樊晓唯, 刘向东. 对 K-means 聚类算法初始值的研究[J]. 电脑知识与技术, 2022, 18(11): 95-97. https://doi.org/10.14004/j.cnki.ckt.2022.0698

[6] 陶永辉, 王勇. 基于初始聚类中心选取的改进 K-means 算法[J]. 国外电子测量技术, 2022, 41(9): 54-59. https://doi.org/10.19652/j.cnki.femt.2203817

[7] 邵小青, 贾钰峰, 章蓬伟, 等. 基于 K-Means 聚类算法的数据分析[J]. 科学技术创新, 2021(23): 85-86.