文章编号:1001-5132（2010）04-0067-04

基于 K-means 算法的学生试卷成绩分析

张晓翊[1]，孟德欣[2]，余翠兰[3]

（1.江汉大学 商学院，湖北 武汉 430056; 2.宁波职业技术学院 计算机系，浙江 宁波 315800;
3.德宏师范高等专科学校 计算机系，云南 德宏 678400）

摘要： 目前在教学管理中，通常采用算术平均线性划分法对学生成绩进行分析评价，难以准确地反映学生真实的学习情况. 笔者运用 K-means 算法对上海市某高级中学某班的一次试卷成绩进行聚类，并对聚类结果进行了详细分析，为制定有效的教学及学习策略提供相关依据.

关键词： K-means 算法；聚类技术；学生成绩；成绩分析

中图分类号： TP311 **文献标识码：** A

学生考试成绩是学生学习成果的显性体现，是针对性制定教学策略的重要依据. 在日常教学中，产生了大量反映学生学习成效的数据——学习成绩. 在一般的教学管理中，通常采用算术平均线性划分法对学生成绩进行分析，即根据学生的原始成绩按照一定的取值尺度进行归类，若学生成绩存在多维属性，则通过简单算术相加转化为一维数据，然后再按照一维数据的划分标准给出等级评定[1]. 这种最常见的学生等级评定方法存在的不足是简单机械，需要有丰富教学经验的教师才能够得到比较合理和理想的学生等级细分，否则得到的分组可能无法充分反映学生的特点；主要表现在同一等级段的学生在某些特征方面并不相似，而不同学生细分段中的学生在某些特征方面存在相似性. 因此，这种成绩细分方法并没有真正起到区分不同成绩等级的作用.

聚类分析技术是数据挖掘及模式识别等研究方向的重要内容之一，在识别数据的内在结构方面具有极其重要的作用，已被广泛应用于语音识别、图像分割和机器视觉等领域[2]. 笔者将聚类分析技术应用于学生成绩分析和等级评定，希望从数据中发现某些规律，为制定有效的教学策略提供依据.

1 聚类分析技术与 K-means 算法

目前，被广泛采纳的关于聚类所下的定义为[3]：1 个类簇内的实体是相似的，不同类簇的实体是不相似的；1 个类簇是测试空间中点的会聚，同类簇的任意 2 个点间的距离小于不同类簇的任意 2 个点间的距离；类簇可以描述为 1 个包含密度相对较高点集的多维空间中的连通区域，它们借助包含密度相对较低点集的区域与其他区域(类簇)相分离. 并且，聚类确定了数据集中所有数据的归属[4].

聚类算法大致分成层次化聚类算法、划分式聚类算法、基于密度和网格的聚类算法和其他聚类算法. 其中，K-means 算法是一种基于划分的经典聚类算法，由 MacQueen 于 1967 年首次提出. 该算法

收稿日期：2010-04-26. 宁波大学学报（理工版）网址: http://3xb.nbu.edu.cn
第一作者：张晓翊（1969－），女，湖北武汉人，讲师，主要研究方向：信息管理及知识工程. E-mail: zhangxiaoyi2599@263.net

的核心思想是: 给定 1 个数据集合和需要聚类的数目 k (k 通常由用户指定), K-means 算法根据某个距离函数迭代运算, 将所有数据分入到 k 个聚类中. K-means 算法将给定的数据划分为 k 个聚类, 每个聚类有 1 个聚类中心(cluster centroid). 通常用聚类中心来表示 1 个类, 它即是这个聚类中所有数据的均值. K-means 算法描述如下所示.

Algorithm k-means(k, D)

Choose k data points as the initial centriods m_j, j=1, …,k;

Repeat

Initialize s_j: =0, j=1,⋯, k;

Initialize n_j: =0, j=1, ⋯,k;

For each data point x∈D do

j←argmin$_{i\in(1, 2, 3, \cdots, k)}$dist(x, m_i)

assigne x to the cluster j;

s_j: = s_j +x;

n_j: = n_j +x;

endfor

m_j←s_j/ n_j, j=1, ⋯, k;

until the stopping criterion is met

算法开始随机选取 k 个数据点作为初始聚类中心, 然后计算每个数据点与各个种子聚类中心间的距离, 将数据点分配给距离最近的聚类中心. 聚类中心以及分配给它的数据点就代表 1 个聚类. 一旦全部数据点都被分配了, 每个聚类的聚类中心会根据聚类中现有的数据点被重新计算. 这个过程会反复迭代, 直至满足某个终止条件为止.

K-means 算法能对大型数据集进行高效分类, 且适合于对数值型数据进行聚类, 其计算复杂性为 $O(tKmn)$, 其中, t 为迭代次数, K 为聚类数, m 为特征属性数, n 为待分类对象数, 通常, K, m, t<<n.

2 基于 K-means 聚类算法的聚类过程和结论分析

2.1 数据来源

笔者旨在对上海某高中某班级的全体学生的综合成绩进行分析, 原始文件为“某高级中学某班综合成绩表.xls”, 数据来源于该班某次全市调研考试成绩.

2.2 数据预处理

该班学生共 46 人, 全部参加考试, 且考试成绩真实有效. 本次考试测试生语文、英语、数学、物理、化学等 5 门课程, 所有成绩为百分制, 且最小记分单位为 1, 无数量级上的差别, 无需标准化. 为便于分析, 将 5 门课程按通行标准分为文、理 2 类, 其中文为语文和英语成绩之和, 理为数学、物

表 1 学生文理成绩表

学号	K1/分	K2/分	学号	K1/分	K2/分	学号	K1/分	K2/分	学号	K1/分	K2/分
01	145	211	13	133	207	25	142	253	37	145	171
02	96	136	14	153	272	26	153	267	38	146	169
03	147	245	15	142	207	27	137	215	39	129	228
04	143	253	16	128	201	28	143	273	40	144	231
05	137	193	17	171	229	29	128	271	41	150	184
06	152	222	18	135	192	30	149	249	42	112	155
07	149	240	19	152	248	31	114	192	43	91	76
08	140	195	20	130	202	32	146	278	44	118	181
09	120	242	21	136	140	33	112	165	45	124	171
10	144	223	22	135	247	34	114	149	46	162	281
11	152	199	23	132	189	35	130	215			
12	140	199	24	142	170	36	157	236			

理和化学成绩之和. 整合后的数据见表 1. 表中的 K1 表示文科成绩, K2 表示理科成绩.

2.3 聚类分析中“亲疏程度”的度量

聚类分析中, 个体之间的“亲疏程度”是极为重要的, 它将直接影响最终聚类结果[5]. 对个体间亲疏程度的测度一般有两个角度: (1)个体间的相似程度; (2)个体间的差异程度. 笔者在此以个体间的差异程度来进行聚类. 将案例中的个体看成是 $k=2$ 的二维空间上的点, 并以此定义个体间的距离, 计算出 46 个个体间的亲疏程度. 个体间的距离越小, 意味着它们越亲密, 越有可能聚成一类. 个体间的距离越大, 意味着他们越疏远, 越有可能分别属于不同的类. 研究中的距离采用欧氏距离, 公式如下: $EUCLID(x,y)=\sqrt{\sum_{i=1}^{K}(x_i-y_i)}$.

2.4 聚类分析

采用 weka3.6.1 版本. 根据教学实际需要, 拟将全部学生分为 4 类, 即 $k=4$, 采用欧式距离, 迭代的最大次数为 500, 初始种子由系统指定, 初始种子数为 4. 运行结果见表 2 和表 3. 聚类结果可视化表示如图 1 所示.

表 2　学生学习成绩聚类结果

Attribute	Full Data (46)	Cluster# 0 (7)	1 (20)	2 (2)	3 (17)
K1	136.7283	118.2857	137.6	93	148.4412
K2	210.2609	164.7143	200.6	106	252.6471

表 3　各类学生人数及比例

Cluster#	学生数/人	所占比例/%
0	7	15
1	20	43
2	2	4
3	17	37

2.5 聚类结果分析

(1) 从表 2 可以看出该班 47 位学生被聚为 4 类, 各类的中心为(93, 106)、(118.2857, 164.7143)、(137.6, 200.6)、(148.4412, 252.6471), 各类学生的数量分别为 2, 7, 20, 17(表 3). 如果所聚 4 个类分别对应优良中差 4 个等级, 则成绩优良学生(Cluster 1、Cluster 3)人数为 37 人, 占比 80%, 成绩中等学生(Cluster 0)人数为 7, 占比 15%, 成绩差等生人数为 2, 占比 4%. 以上数据说明在这次考试中, 学生成绩总体令人满意.

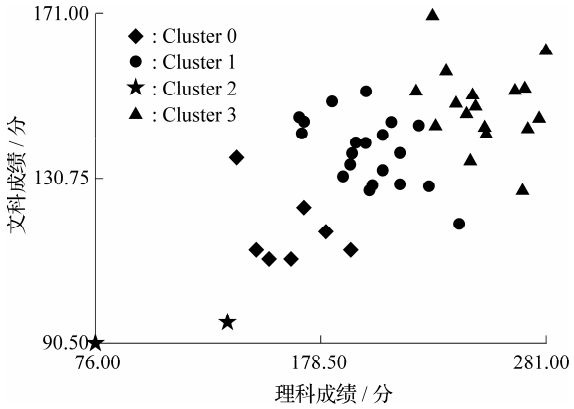

图 1　学生成绩聚类结果可视化图

(2) 聚类中心指出了聚类的中心所在的位置. 用聚类中心来表示每个聚类是使用最广泛的聚类表示方法[6]. 在本次聚类中, 形成了 4 个聚类中心, 分别对应优、良、中、差 4 个等级. 以欧氏距离计算类间距离, 相关数据见表 4.

表 4　4 个类间欧氏距离矩阵

Cluster	Euclidean Distance 1: 优	2: 良	3: 中	4: 差
1: 优	0	53.1642	92.95984	156.7772
2: 良	53.1642	0	40.75323	104.5864
3: 中	92.95984	40.75323	0	63.92758
4: 差	156.7772	104.5864	63.92758	0

从表 4 中可知: 等级为“中”的聚类中心距等级为“优”、“良”的聚类中心的距离分别是 92.95984 和 53.164, 说明成绩中等学生与成绩优良生存在较大的距离, 这类学生如果希望提高学习成绩, 进入学习优良生的行列, 需付出较大努力. 成绩差等生与成绩中等生之间的距离接近 64, 与成绩优秀生的差距更高达 156.7772, 说明他们在学习中存在巨大的学习困难, 成绩提高面临相当大的难度.

(3) 从图 1 中可以看出, 有些学生处于类的边缘, 如学生 10(K1: 144, K2: 223)处于 Cluster 1 的边

缘, 接近 Cluster 3; 学生 40 的文理成绩为(K1: 144, K2: 231)属于 Cluster 3. 两者 5 科成绩的差距仅 8 分. 对于学生 10 来说, 稍作努力即可进入成绩优秀学生行列. 而学生 6、9、40 则位于 Cluster 3 的边缘, 接近 Cluster 1, 如果他们希望继续保持在学习成绩上的优势地位, 则不能有丝毫懈怠.

(4) 从图 1 中还可以看出个别学生的文理成绩失衡. 如学生 17(K1: 171, K2: 229)和学生 29(K1: 128, K2: 271). 学生 17 文科成绩处于全班第 1, 而理科成绩位于全班第 17 位; 学生 29 理科成绩名列前茅, 但其文科成绩却低于全班平均水平.

3 结论

文中仅采用 K-means 算法对学生成绩进行了简单聚类. 在实际应用中, 还可以应用因子分析法分析各科成绩之间是否具有相关性, 进一步合并因子. 也可以根据教师丰富的教学经验, 指定各聚类中心, 形成聚类结果. 与传统成绩等级评定的刚性划分方法相比, 聚类分析通过比较各实例之间的差异性, 将具有相似特点的实例聚集成簇, 提供了一种新的分析方法和视角. 教师或学生可以根据聚类分析结果, 针对性地制定教学/学习策略, 提高教学/学习效果.

参考文献:

[1] 王立新, 许晖. 评定学生学习成绩等级的方法[J]. 延边大学学报, 2001, 27(4):304-307.

[2] 孙吉贵, 刘杰, 赵连宇. 聚类算法研究[J]. 软件学报, 2008(1):48-59.

[3] Jain A K, Dubes R C. Algorithms for clustering data[M]. Cambridge: Prentice Hall College Div, 1988.

[4] 朱扬勇, 熊赟. 数据学[M]. 上海: 复旦大学出版社, 2009.

[5] 薛薇. 基于 SPSS 的数据分析[M]. 北京: 中国人民大学出版社, 2006.

[6] Liu Bing. Web 数据挖掘[M]. 俞勇, 译. 北京: 清华大学出版社, 2009.

Examination Grade Analysis Based on K-means Methods

ZHANG Xiao-yi[1], MENG De-xin[2], YU Cui-lan[3]

(1.Commercial Institute, Jianghan University, Wuhan 430056, China; 2.Department of Computer, Ningbo Ploytechnic, Ningbo 315800, China; 3.Department of Computer, Dehong Teacher's College, Dehong 678400, China)

Abstract: The present teaching management adopts the arithmetical average method to analyze and evaluate students' performances, which is difficult to truly reflect the states of students learning process. In this paper, based on a fully completed score-sheet of academic performance sampled from a class in a senior high school in Shanghai, the K-means approach is adopted to cluster the examination score, and a detailed analysis is subsequently conducted on the clustering results. The introduced approach proves more realistic and convincing, which can be used for more effective teaching and learning planning.

Key words: K-means algorithm; clustering; the academic performance; the examination analysis

CLC number: TP311 **Document code:** A

（责任编辑 章践立）