文章编号：2095-2163(2019)01-0080-04 中图分类号：TP301.6 文献标志码：A

基于全局中心聚类算法的学生成绩评价研究

段桂芹
(广东松山职业技术学院 计算机系，广东 韶关 512126)

摘 要：针对课程难度差异大而带来的学生成绩评价难的问题，提出了一种基于全局中心聚类算法的学生成绩评价方法。首先，使用 min-max 归一化方法对样本集进行预处理；然后，采用全局中心聚类算法对学生的多科成绩进行聚类；最后，使用内部评价指标 CH 对多组聚类结果进行评价，得出最优聚类数和最优聚类划分。通过对某高校学生成绩的聚类分析结果表明：该方法能够有效地挖掘出学生多科成绩的分布情况，可为个性化教学的实施提供一种新的思路。

关键词：成绩评价；全局中心聚类；内部评价指标；数据标准化

Student performance evaluation based on global center clustering algorithm

DUAN Guiqin
(Depart. of Computer Science, Guangdong Songshan Polytechnic, Shaoguan Guangdong 512126, China)

【Abstract】 In view of the difficulty of students´ performance evaluation caused by the difference of curriculum difficulty, a new method of student performance evaluation based on global center clustering algorithm is proposed. Firstly, the sample set is pre-processed by min-max normalization method. Then, the global center algorithm is used to cluster the students´scores of many courses. Finally, the internal evaluation index CH is used to evaluate the results of multi-group clustering, and the optimal clustering number and optimal clustering partition are obtained. Through the cluster analysis of students´performance in a university, the results show that the method can effectively mine the distribution of students´ scores in many courses. It can provide a new idea for the implementation of personalized teaching.

【Key words】 performance evaluation; global central clustering; internal evaluation indicators; data standardization

0 引 言

高校在评价学生综合素质时，常采用平均分或总分作为衡量学生成绩的等级标准，在实际教学反馈中这种评价方式简单易行，但却忽略了由于试卷难易程度无法统一而导致评价结果的单一性和片面性。这种评价方式无法客观真实地反映学生间的成绩相对分布与学情分类情况[1]，例如：当 2 门课程的平均成绩分别为 80 分和 70 分，某学生 2 门课程同为 79 分，则该生第二门课程的评价等级要高于第一门课程，但是这种评价结果却无法通过原始卷面成绩直接反映出来，尤其当教师需要对学生进行分组教学并制定与学情相适应的个性化指导时，这种有缺陷的评价方式很难科学地对学生进行合理分组。因此，本文提出采用全局中心聚类算法对学生成绩进行聚类，将距离相对较近(高相似度)的学生聚为一类，通过分析比较各类学生成绩，给出相应的改进建议，为学生的成绩评价、个性化发展以及教师的差异化教学提供理论依据。

1 聚类算法

聚类分析作为一种探索性分析方法被广泛应用于模式识别、计算机视觉、数据挖掘等领域中，其目的是根据相似性原则将物理或抽象的对象集合分成若干个子集，并分析各子集中数据对象的内在联系、规律和特点[2]。K-means 聚类算法是应用最为广泛的划分方法之一，其实现简单、快速，能有效地处理大数据集，但该算法对初始聚类中心和异常数据较为敏感，且不能用于发现非凸形状的簇，因此聚类结果存在不稳定性。为了解决 K-means 算法的这些问题，研究人员围绕簇中心的选择与优化提出了新的计算方法[3-6]，提高了原算法的聚类质量，减少了聚类时间。

1.1 全局中心聚类算法

全局中心聚类算法由距离矩阵构建、初始聚类中心选择和簇中心更新 3 部分构成。首先，使用距离公式计算各数据对象间的距离；再从距离矩阵中选取 k 个首尾相连且距离乘积最大的数据对象作为

基金项目：韶关市科技计划项目(2017CX/K055)；广东松山职业技术学院重点科技项目(2018KJZD001)。

作者简介：段桂芹(1979-)，女，硕士，讲师，主要研究方向：数据挖掘。

收稿日期：2018-11-07

初始聚类中心集合 V；然后，根据集合 V 完成初次聚类，选取簇内距离之和最小的样本作为簇中心，生成临时簇中心集合 V'；最后，按最小距离将各样本划分到相应簇中，重复簇中心迭代过程，直至聚类误差平方和函数收敛，完成聚类。

1.2　相关定义与公式

全局中心聚类算法中的相关定义和公式如下所述。

设样本集 X 为含有 n 个学生样本的集合，$X=\{X_1,X_2,\cdots,X_i,\cdots,X_n\}$，每个学生样本由 p 门课程成绩组成，第 i 个样本对象可以表示为：$X_i=(X_{i1},X_{i2},\cdots,X_{ip})$。现将样本集划分为 k 个簇，每簇含学生样本数为 m，则样本集 $X=\{C_1,C_2,\cdots,C_k\}$，簇中心集合 $V=\{v_1,v_2,\cdots,v_k\}(k<n)$。

定义 1　min-max 标准化是对原始数据的线性变换，使结果落到[0,1]区间，转换函数如下：

$$x'=\frac{x-\min}{\max-\min} \tag{1}$$

其中，x 为某门课程的原始成绩；max 为该门课程的最高分；min 为该门课程的最低分。

定义 2　空间两点间的欧氏距离定义为：

$$d(X_i,X_j)=\sqrt{\sum(X_i^w-X_j^w)^2} \tag{2}$$

其中，$i=1,2,\cdots,n$；$j=1,2,\cdots,n$；$w=1,2,\cdots,p$。

定义 3　样本集 X 的空间距离矩阵 $\boldsymbol{X}'$

$$\boldsymbol{X}'=\begin{bmatrix} 0 & d(X_1,X_2) & \cdots & d(X_1,X_n) \\ d(X_2,X_1) & 0 & \cdots & d(X_2,X_n) \\ \cdots & \cdots & 0 & \cdots \\ d(X_n,X_1) & d(X_n,X_2) & \cdots & 0 \end{bmatrix} \tag{3}$$

定义 4　样本 X_i 的簇内距离定义为 X_i 与其所属同一簇的样本间的距离之和，即：

$$DistSum(X_i)=\sum_{i,j\in C_k}d(X_i,X_j) \tag{4}$$

其中，$i=1,2,\cdots,n$，$j=1,2,\cdots,n$

定义 5　第 k 簇的簇内距离之和矩阵定义为：

$$DistSum\_array(C_k)=\begin{bmatrix} DistSum(X_i) \\ DistSum(X_2) \\ \cdots \\ DistSum(X_m) \end{bmatrix} \tag{5}$$

定义 6　将第 k 簇的簇内距离之和最小的样本 X_i 作为中心，即：

$$V_k=find(\min(DistSum\_array(C_k))) \tag{6}$$

定义 7　聚类误差平方和 E 定义为：

$$E=\sum_{i=1}^{k}\sum_{j=1}^{m}|X_{ij}-V_i|^2 \tag{7}$$

其中，X_{ij} 是第 i 簇的第 j 个数据对象，V_i 是第 i 簇的中心。

定义 8　*CH* 指标(Calinski-Harabasz)[7]

$$CH(k)=\frac{\sum_{i=1}^{k}m_i d^2(v_i,c)/(k-1)}{\sum_{i=1}^{k}\sum_{x\in C_i}d^2(x,v_i)/(n-k)} \tag{8}$$

CH 指标将各簇中心点与样本集的均值中心的距离平方和作为数据集的分离度，将簇中各点与簇中心的距离平方和作为簇内的紧密度，将分离度与紧密度的比值视为 *CH* 的最终指标。该指标越大表示各簇之间分散程度越高，簇内越紧密，聚类结果越优。Milligan 在文献[8]中，对 *CH* 等评价指标的性能进行了深入探讨。实验结果表明，*CH* 指标在多数情况下，都要优于其它的指标。

2　学生成绩聚类

2.1　聚类流程

使用全局中心聚类算法对学生成绩进行聚类的整个流程分为 3 部分：数据预处理、多聚类结果比较和最优聚类结果输出。其中，多聚类结果比较环节中的聚类数 k 由内部评价指标 *CH* 确定，设计流程如图 1 所示。

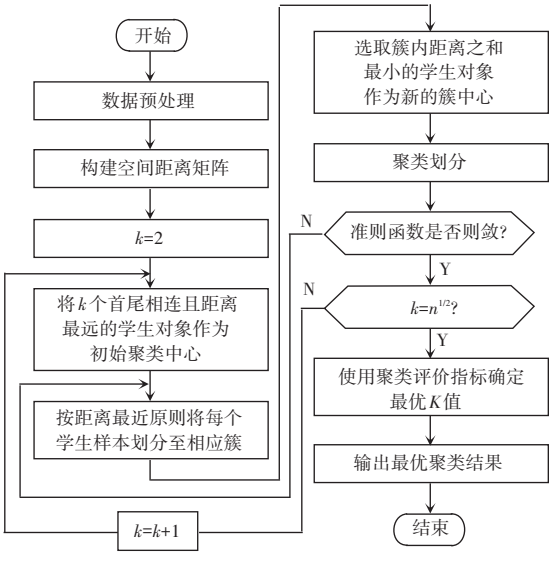

图 1　学生成绩聚类流程

Fig. 1　Student performance clustering flow chart

具体实施步骤如下：

(1)根据式(1)完成数据预处理；

(2)根据式(2)计算样本集 X 中各学生对象之

间的距离;

(3) 根据式(3) 构建全部学生的空间距离矩阵 $\boldsymbol{X}'$;

(4) 从 $\boldsymbol{X}'$ 中选择满足 k 个首尾相连且距离乘积最大的学生对象作为初始聚类中心,既将条件为 $\max(d(X_a',X_b')\times d(X_b',X_c')\times\cdots\times d(X_j',X_k'))$ 的 k 个数据对象加入簇中心集合 V 中,使得 $V=\{X_a',X_b',X_c',\cdots,X_k'\}$;

(5) 将非簇中心的学生对象按距离最近原则划分至相应簇中;

(6) 使用式(4)、(5) 得出簇内距离之和矩阵,再根据式(6) 从矩阵中筛选出簇内距离之和最小的学生对象作为新的簇中心存入集合 V' 中;

(7) 重复步骤(6),更新各簇的中心,直到 $|Z'|=k$,再用 V' 取代 V;

(8) 重复步骤(5);

(9) 根据式(7) 判断函数 E 是否收敛,如果收敛,则聚类算法结束,否则转到步骤(4) 继续执行;

(10) 使用式(8) 中的 CH 指标对 $k=\{2,3,\cdots,n^{1/2}\}$ 的聚类结果进行评价,将 CH 指标取最大值时的聚类划分作为最优聚类结果输出。

2.2 数据预处理

本文中的样本数据来源于某高校 2016 级 81 名同学第四学期 JAVA 、HTML5、C#数据访问技术 3 门课程的成绩。总成绩由平时成绩(20%) 和期末成绩(80%)构成,均采用百分制形式计算,原始数据见表 1,采用 min-max 标准化后的数据见表 2。

表 1 学生原始成绩

Tab. 1 Original student grade

学号	JAVA	C#数据访问技术	HTML5
1	68	68	60
2	50	60	78.8
3	84	77	75.2
4	82	64	39.2
5	66	64	64.4
6	72	70	70
7	60	52	60.4
8	73	67	62.8
9	96	93	62
10	81	83	60
…	…	…	…

2.3 K 值的确定与最优聚类结果

使用全局中心聚类算法对学生成绩聚类后,用 CH 指标对 $k=\{2,3,\cdots,9\}$ 的聚类结果进行对比,指标值与聚类数目的关系如图 2 所示。由 CH 指标的特征可知,该值最大时的 k 值即为最优聚类数,此时的聚类结果为最优划分。从图 2 可以看出,最优聚类数 $k_{opt}=3$,此时的各簇数据分布情况如图 3 所示,各簇中心点位置详见表 3。

表 2 预处理后的学生成绩

Tab. 2 Student grades after pretreatment

学号	JAVA	C#数据访问技术	HTML5
1	0.46	0.46	0.55
2	0.12	0.32	0.80
3	0.77	0.61	0.75
4	0.73	0.39	0.28
5	0.42	0.39	0.61
6	0.54	0.49	0.68
7	0.31	0.19	0.56
8	0.56	0.44	0.59
9	1.00	0.88	0.58
10	0.71	0.71	0.55
…	…	…	…

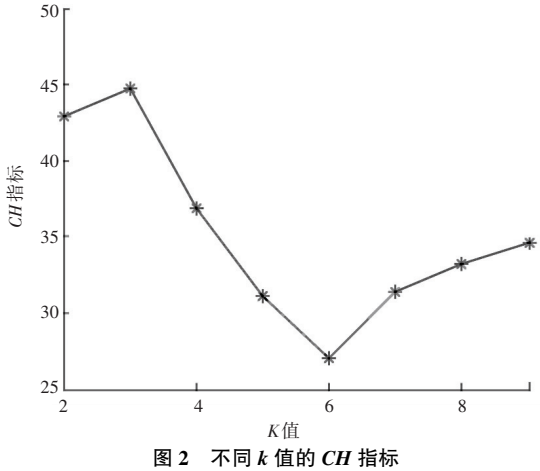

图 2 不同 k 值的 CH 指标

Fig. 2 CH indicator for different k values

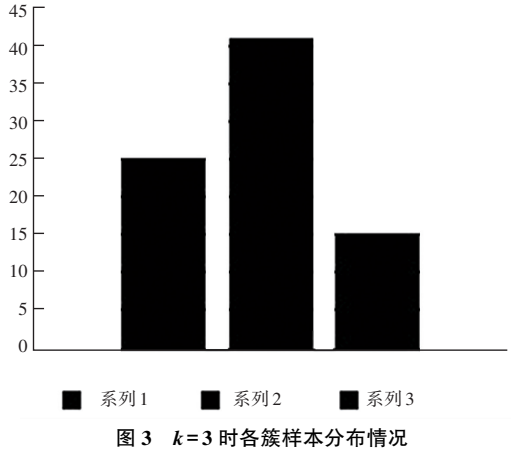

图 3 $k=3$ 时各簇样本分布情况

Fig. 3 Distribution of each cluster sample when $k=3$

表 3　标准化后各簇中心点分布

Tab. 3　The central point distribution of each cluster after standardization

	JAVA	C#数据访问技术	HTML5
I	0	0.24	0.55
II	0.96	0.92	0.98
III	0.83	0.32	0

3　聚类结果分析

从图 3 的最优聚类结果得出：学生可以划分为 3 类，各类人数分别为 25 人、41 人、15 人，所占比例分别为：30.86%、50.62%、18.52%，聚类划分基本符合正态分布，结合这 3 类同学的平时成绩，对本次聚类结果分析如下。

(1)相比于其它 2 类，第一类学生的综合素质较高，逻辑思维能力较强，程序设计能力较为突出，建议在现有水平的基础上，适当提高学习目标，深入学习更为前沿的知识技术；

(2)第二类学生成绩比较稳定，HTML5 课程的成绩有较大提升空间，建议尝试改进现有学习方法，提高应试技巧，加强主动学习意识；

(3)第三类学生的成绩低于及格线，说明这部分学生的学习态度消极或者学习方法不正确，需要教师、辅导员给予学生更多的关心，帮助学生树立正确的学习方法，鼓励学生在学习上投入更多的精力。

4　结束语

本文使用全局中心聚类算法结合 *CH* 评价指标对学生成绩进行聚类分析与评价。通过 min－max 标准化方法完成数据的归一化，相比传统的均值聚类算法，本文算法通过计算内部评价指标解决了无类标样本聚类数 k 难以确定的问题。所得到的最优聚类结果符合实际情况，有效地克服了因课程之间难度差异大而带来的评价不合理的问题，并针对各类学生的学习成绩给出了相应的改进建议。

参考文献

[1] 李巧君，李伟. 数据挖掘技术在学生成绩分析中的应用研究[J]. 微型电脑应用，2015，31(4)：35－36，40.

[2] 唐东凯，王红梅，胡明，等. 优化初始聚类中心的改进 K－means 算法[J]. 小型微型计算机系统，2018，39(8)：1819－1823.

[3] 邹臣嵩，杨宇. 基于最大距离积与最小距离和协同 K 聚类算法[J]. 计算机应用与软件，2018，35(5)：297－301，327.

[4] 段桂芹. 基于均值与最大距离乘积的初始聚类中心优化 K－means 算法[J]. 计算机与数字工程，2015，43(3)：379－382.

[5] 王彬宇，刘文芬，胡学先，等. 基于余弦距离选取初始簇中心的文本聚类研究[J]. 计算机工程与应用，2018，54(10)：11－18.

[6] 邹臣嵩，刘松. 基于谱聚类的全局中心快速更新聚类算法[J]. 计算机与现代化，2018 (10)：6－11.

[7] CALINSKI T，HARABASZ J. A dendrite method for cluster analysis[J]. Communications in Statistics，1974，3(1)：1－27.

[8] MILLIGAN G W，COOPER M C. An examination of procedures for determining the number of clusters in a data set [J]. Psychometrika，1985，50(2)：159－179.

(上接第 79 页)

4　结束语

本文基于卷积神经网络来识别婴儿的大便图像，从而对胆道闭锁进行早期的筛查。通过 GoogleNet 网络 InceptionV3 模型来训练数据，对模型的参数进行适当的修改，观察拟合带的程度，从而确定最佳的模型参数。相较之前对大便图像的识别分类，本文提出的方法获得了较高的准确率。卷积神经网络在图像的分类识别领域有很大的优势，利用卷积神经网络中的模型来对婴儿大便图像进行识别分类，有助于患儿早期胆道闭锁的诊断，提高了患病婴儿的存活率[5]。

参考文献

[1] 詹江华，陈扬，钟浩宇. 粪便比色卡在胆道闭锁早期筛查中的应用[J]. 临床小儿外科杂志，2017，16(2)：109－112.

[2] 熊复，徐静，朱书瑶，等. 应用大便比色卡进行胆道闭锁筛查的研究进展[J]. 中国儿童保健杂志，2016，24(11)：1172－1173.

[3] CHEN Shanming，CHANG M H，DU J C，et al. Screening for biliary atresia by infant stool color card in Taiwan [J]. Pediatrics，2006，117(4)：1147－1154.

[4] 李俊阳，雷鑫，宋宇，等. 基于卷积神经网络的脱机单个手写汉字识别[J]. 智能计算机与应用，2018，8(2)：92－95，99.

[5] 高震宇. 基于深度卷积神经网络的图像分类方法研究及应用[D]. 合肥：中国科学技术大学，2018.

[6] 刘欣. 基于卷积神经网络的联机手写汉字识别系统[D]. 哈尔滨：哈尔滨工业大学，2015.

[7] SZEGEDY C，VANHOUCKE V，IOFFE S，et al. Rethinking the inception architecture for computer vision [J]. arXiv preprint arXiv：1512.00567v3，2015.

[8] MACFAUL R. Screening for biliary atresia [J]. Lancet.，1993，342(8874)：811.

[9] 杨吉刚，马大庆，李春林. 胆道闭锁的临床及影像学诊断[J]. 实用儿科临床杂志，2006，21(23)：1668－1670.

[10] 孙梦熵，胡春红. 胆道闭锁的影像学诊断研究进展[J]. 国际生物医学工程杂志，2017，40(1)：42－45，57.