文章编号:1673-9469(2010)01-0096-03

基于MMD聚类算法及在高校成绩分析中的应用

顾洪博,赵万平
(大庆石油学院 计算机与信息技术学院,黑龙江 大庆市 163318)

摘要:介绍了在聚类算法中广泛使用的k均值算法。针对其受选择初始质心和聚类个数影响的缺点,给出了改进的k均值算法。使用最大最小距离法选择初始聚类中心,并确定聚类个数。进行了改进前后的对比实验。实验结果表明,改进后的算法比较稳定、准确。将改进后的算法应用到高校成绩分析中,达到较好的分类效果。

关键词:聚类分析;成绩分析;最大最小距离;
中图分类号:TP301.6 **文献标识码**:A

Cluestering algorithm based on Max-min Distance for students' score analysis in universities and applications

GU Hong-bo, ZHAO Wan-ping
(School of Computer & Information Technology, Daqing Petroleum Institute, Heilongjiang Daqing 163318, China)

Abstract: The classic algorithm of k-means is discussed, that is one of the most widespread methods in clustering, including both strongpoint's and shortages. Not only is it sensitive to the original clustering center, but also it may be affected by the k. Given these shortages, an improved algorithm is discussed, which makes improvements in k and selection of original clustering center. To select original clustering center based on the max-min distance. This paper presents the application which all show that the improved algorithm can lead to better and more stable solutions than k means algorithm. The experiment and application affection by the outliers is down to a much low figure. The improved algorithm was used to the students' score analysis in universities and had a good closer.

Key words: clustering algorithm; score analysis; max-min distance

聚类是按照某个特定标准将数据集划分成不同个簇或类的过程,同一组内的数据对象具有较高的相似度而不同组之间的数据对象相似度较低[1]。聚类应用广泛,在经济学、生物学、信息工程等领域都有十分重要的应用[2]。因此,对聚类的要求较高,提出准确且高效的聚类算法十分重要。划分方法是对于一个给定的数据集,通过把数据分成 k 个组,每个组为一个簇。最著名的划分聚类算法是 k 均值算法。

k 均值算法对大型数据集进行高效分类[3]。其不足是要确定聚类个数 k,不同的初始聚类中心会影响到最终的聚类结果[4]。目前,针对这些不足,许多研究者提出了改进的方法。如,文献[5]先把全部样本看成一个类,样本总均值点就是第1类的初始聚类中心;然后,由这一类的初始聚类中心和离它最远的一个样本作为两类的初始聚类中心;依此类推由($k-1$)类的代表点和离他们最远的一个样本点作为 k 类问题的初始聚类中心。文献[6]数据标准化后形成矩阵向量;依此作为输入求得聚类个数,利用遗传算法的全局搜索能力选取初始中心。最大最小距离法(Max-min Distance,最大最小距离)是模式识别领域中的一种方法[7]。本文提出一个基于MMD的 k 均值优化算法,将MMD和 k 均值结合,使用MMD来选择初始聚类中心,避免随机选择初始聚类中心对算法的影响。

收稿日期:2009-12-28
基金项目:黑龙江省自然科学基金(No.F200603)
作者简介:顾洪博(1976-),女,黑龙江宾县人,硕士,讲师,从事数据库应用及数据挖掘的教学与科研工作。

1 基本思想

1.1 k 均值

首先,从 n 个数据对象中随机选择 k 个对象作为初始簇的中心,对剩下的其它对象根据它们与这些簇中心的距离,分别将它们分配给与其最近簇中心所代表的簇。然后再计算每个新簇的中心(该簇中所有对象的平均值)。不断重复这一过程直到准则函数收敛为止[8]。

1.2 MMD 选择聚类中心

经典 k 均值算法随机选取 k 个点作为初始聚类中心进行操作。由于是随机选取,则变化较大,初始点选取不同,聚类结果也不同,并且聚类准确率相异。提出一种基于 MMD 选择聚类中心。

算法 1:(1) 在数据集 S 任取一个对象 X_1,把 X_1 作为第一个类的聚类中心,则有 $c_1 = X_1$。(2) 从集合 S 中找出到 c_1 距离最大的对象作为第二个类的聚类中心 c_2。(3)对 S 中剩余的每个对象 X_i,分别计算到 c_1 和 c_2 的距离 d_{i1} 和 d_{i2},令其中较小值为 $\min(d_{i1}, d_{i2})$,即找到剩余的每个对象到已有聚类中心的最近的距离。(4)计算 $\min(d_{i1}, d_{i2})$的最大值,记为:$\max(\min(d_{i1}, d_{i2}))$,对应的那个对象记为 X_j。(5)若 $\max(\min(d_{i1}, d_{i2})) > m \times | c_2 - c_1 |$,则取 X_j 为第三个聚类中心。其中 m 通常 $1/2 \leqslant m < 1$。(6)再比较剩余的其他点,用同样的方法找到 $\max(\min(d_{i1}, d_{i2}, d_{i3}))$的对象。(7)检验条件为 $\max(\min(d_{i1}, d_{i2}, d_{i3})) > m \times (\text{averge}(| c_2 - c_1 |, | c_3 - c_2 |))$。如满足条件,将该对象作为新的聚类中心,重复第(6)步,直到再找不到符合条件的新的聚类中心。

2 基于 MMD 聚类算法及实验分析

2.1 基于 MMD 聚类算法

假设要将数据对象集合 $U(n)$分为 k 类,$m = 0.5$。首先要用算法 1,来选择 k 初始聚类中心,最终有 $(c_1, c_2, \cdots, c_K)$。然后用经典 k 均值算法,来进行聚类分析,得到 $U = (A_1 \cup A_2 \cup \cdots\cdots \cup A_k)$。

2.2 基于 MMD 聚类算法的实验

实验采用知名的 Wine 数据集。该数据集 1991 年建立,共有 178 条记录,共有 3 个类,聚类结果可靠,适合做聚类分析的基准数据集。每个记录有 13 个属性,分别是 Alcohol、Malic acid、Ash、Alcalinity of ash、Magnesium、Total phenols、Flavanoids、Nonflavanoid phenols、Proanthocyanins、Color intensity、Hue、OD280/OD315 of diluted wine、Proline,如(1,14.23,1.71,2.43,15.6,127,2.8,3.06,0.28,2.29,5.64,1.04,3.92,1 065)。为了方便,只选取 Wine (Alcohol、Malic acid、Ash、Alcalinity of ash) 作为实验对象。实验环境:VC + + 6.0 编程语言。实验正确率是实验 10 次后得到的平均值。实验结果见表 1。

2.3 改进后的实验分析

从上面的实验中可以看出,k_1 到 k_2 的结果表明经典 k 均值算法可以比较准确的分类。Wine 数据集数据中随机选取初始聚类中心一定会影响分类效果。可以看出,当每次分类结果选择的初始聚类中心不同,分类结果就有差异。k_3 到 k_6 是是改进后的算法得到的结果。分类的准确率有所提高。这说明基于 MMD 聚类算法使用 MMD 来形成的初始聚类中心,并依此来寻找数据,因而产生的初始聚类中心较符合数据实际分布。实验表明,与传统随机选取初始聚类中心的方法相比,改进后的方法可得到更好的划分效果,寻找较为准确的 k 个聚类中心。

表 1 改进前后的聚类结果

Tab.1 The result of both the improved and classic clustering

算法	初始聚类中心	正确率
k_1	(1,14.23,1.71,2.43)(2,12.17,1.45,2.53)(3,12.7,3.55,2.36)	0.58
k_2	(1,14.2,1.76,2.45)(2,12.21,1.19,1.75)(3,12.6,2.46,2.2)	0.61
k_3	(1,14.83,1.64,2.17)(2,13.49,1.66,2.24)(3,12.53,5.51,2.64)	0.63
k_4	(1,14.12,1.48,2.32)(2,13.03,.9,1.71)(3,13.62,4.95,2.35)	0.65
k_5	(1,14.19,1.59,2.48)(2,12,.92,2,19)(3,13.88,5.04,2.23)	0.67
k_6	(1,13.05,2.05,3.22)(2,11.84,.89,2.58)(3,13.5,3.12,2.62)	0.68

3 在成绩管理中的应用分析

3.1 应用

学校主要是通过各种考试评价学生,判断学生在知识、能力上达到的水平。通常根据学生的考试分数,对学生作出不同的分等、分类。这种划分有利于研究学生的学习成绩,为后继的教学与学习过程提供反馈信息,具有重要意义。

采用聚类分析方法以后,在学生成绩评价中,每个簇就是一个成绩群,处于每个簇中心的数据就是该成绩群的中心成绩。不同的簇相应地对各个成绩群进行划分,也相应地给出了不同成绩群的中心成绩。这些中心成绩就是学生成绩的等级划分参考标准之一。因此,基于聚类分析的成绩划分是相对成绩的划分,对学生的成绩评价也更为准确。实验数据取计算机2008级1-8班的243名学生在08-09-12学期主要课程的考试分数。全部课程共6门,分别为:高等数学、线性代数、形式与政策、大学英语1级、计算机导论、大学体育。对学生成绩进行聚类时,数据集为243个数据,每个数据有6维属性。其算法结果比较见表2。

表2 算法结果比较

Tab.2 Comparison of the algorithms

比较条件	10次平均正确率/%		10次平均运行时间/ms	
	经典	改进后	经典	改进后
$k=3$	58.41	61.69	213.18	279.147
$k=4$	61.02	64.29	276.58	331.28
$k=5$	62.17	66.58	317.41	427.12

3.2 应用的分析

对数据集的数据进行聚类分析,无法事先判断243名学生能分成几类,故本次聚类个数从2变化到5。当 k 发生变化时,聚类结果也改变,最终看几类适合。改进算法使用MMD法或者经典算法选择初始聚类中心,每次都是任选一个数据对象作为第一聚类中心。因此会大大增强算法的随机性。所以每个算法都运行10次,取其平均结果。

从执行时间上看,使用经典k均值算法时用时最少;基于MMD寻找初始聚类中心,所以改进算法执行时间较长。当 $k=3$ 时用时最短。当 $k=5$ 时分类时间最长。其聚类结果见表3。

在算法运行效果上看,经典算法的分类结果低于基于MMD算法。当 $k=3$、4时,基于MMD算法比经典算法正确率提高了3.25%和3.17%。$k=5$,正确率提高了4.41%。这说明使用MMD后聚类结果有所提高。所以,改进算法的比较成功的。

从聚类中心看,A类为学习优秀类,此类学生成绩好,公共课平均成绩都≥80分,尤其专业课平均成绩都在90分以上,体育成绩平均为优秀。B类学生为学习良好类,公共课平均成绩≥72分,尤其专业课的平均成绩都80~85间。C类学生为学习中等偏上类,总平均成绩在74以上,专业课平均成绩都≥75分,体育课大多都在80分以上,但英语成绩大都在60分左右。D类为学习中等偏下类,其专业课的平均成绩都在70分左右,个别接近60分,体育成绩较好。E类学生为学习较差生类,其专业课、公共课的平均成绩都在75分以下,个别学生偏科现象严重,英语都不及格。

4 结束语

实验结果证实改进后的算法能够得到较高且稳定的准确率,更适用于对实际数据的聚类。把改进的算法应用于实际教学中,为以后改进教学与教育提供科学的信息,为因材施教提供科学的依据,从而提高教学质量,推进教学改革。

表3 当k=5时改进聚类结果

Tab.3 The result of improved clustering while k=5

算法	初始聚类中心	正确率
k_1	(92,88,91,94,93,86)(77,83,86,79,84,90)(84,73,79,68,71,78)(71,64,60,67,64,72)(64,83,69,51,68,60)	0.54
k_2	(91,92,83,95,90,91)(89,92,95,77,82,80)(78,89,81,69,91,82)(71,92,87,67,65,89)(79,94,45,76,81,65)	0.56
k_3	(89,91,84,95,89,93)(79,89,89,77,76,84)(76,91,84,64,81,92)(81,82,77,71,64,69)(78,84,43,64,87,66)	0.55
k_4	(93,94,86,91,93,90)(87,91,90,81,71,88)(78,87,80,67,90,87)(72,92,84,64,67,87)(77,89,52,66,73,64)	0.57

(下转第108页)

综合以上分析和计算结果可以看出,无论在收敛精度还是在计算效率和收敛性能方面,MRS-CGA 均好于常规的 GA。

4 结论

从提高进化效率的角度出发,提出了基于多保留策略的复合型遗传算法(简称 MRS-CGA),利用 Markov 链理论和仿真技术从各个层面分析了算法的收敛性能,理论分析和实例仿真比较表明,MRS-CGA 不仅具有良好的结构和可操作性,而且在计算效率和收敛稳定性方面均明显优于常规的遗传算法,为进一步构建复合优化方法奠定了基础,在一定程度上推广和丰富了现有的智能优化理论和方法。

参考文献:

[1] SRINIVAS M, PATNAIK M. Genetic algorithm: A survey [J]. IEEE Computer, 1994, 27(6): 17-26.

[2] FOGE D B. An introduction to simulated evolutionary optimization[J]. IEEE Trans. on SMC, 1999, 24(1): 3-14.

[3] ATMAR W. Noteson the simulation of evolution[J]. IEEE Trans. on SMC, 1994, 24(1): 130-147.

[4] HOLLAND J H. Adaptation in nature and artificial systems [M]. USA: Univ. of Michigan, 1975.

[5] 巩敦卫,孙小燕,郭西进. 一种新的优胜劣态遗传算法[J]. 控制与决策, 2002(6): 908-912.

[6] 韩万林. 遗传算法的改进[J]. 中国矿业大学学报, 2001(1): 102-105.

[7] 李京涛, 何丽丽, 高瑞贞, 等. 改进遗传算法在桁架拓扑优化中的应用[J]. 河北工程大学学报(自然科学版), 2009, 26(3): 19-22.

[8] 夏道行, 吴卓人, 严绍宗, 等. 实变函数与泛函分析[M]. 北京: 高等教育出版社, 1979.

[9] 方兆本, 缪柏其. 随机过程[M]. 合肥: 中国科学技术大学出版社, 1993.

[10] 盛 骤. 概率论与数理统计[M]. 北京: 高等教育出版社, 1989.

[11] 王小平, 曹立明. 遗传算法理论、应用与软件实现[M]. 西安: 西安交通大学出版社, 2002.

[12] 陈国良. 遗传算法及其应用[M]. 北京: 人民邮电出版社, 1996.

(责任编辑 马立)

(上接第 98 页)

参考文献:

[1] 杨小兵. 聚类分析中若干关键技术的研究[D]. 杭州: 浙江大学计算机学院, 2005.

[2] 袁 方, 周志勇, 宋 鑫. 初始聚类中心优化的 k-means 算法[J]. 计算机工程, 2007, 33(3): 65-66.

[3] HUANG Z. A fast clustering algorithm to cluster very large categorical data sets in data mining. In: Proc. of the SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery. Tucson, 1997. 146-151[EB/OL]. http://www.informatik.uni-trier.de/~ley/db/conf/sigmod/sigmod97.html.

[4] SAMBASIVAM S, THEODOSOPOULOS N. Advanced data clustering methods of mining Web documents[J]. Issues in Informing Science and Information Technology, 2006, (3): 563-579.

[5] 黎 敏. 数据挖掘算法研究与应用[D]. 大连: 大连理工大学, 2004.

[6] 孟 岩, 刘希玉, 刘艳丽. 一种基于蚁群算法的 K-means 算法—在公路运输枢纽宏观布局规划中的应用[J]. 计算机工程与应用, 2008, 44(1): 179-182.

[7] 周 涓, 熊忠阳, 张玉芳, 等. 基于最大最小距离法的多中心的聚类算法的研究[J]. 计算机应用, 2006, 26(6): 1425-1427.

[8] MARQUES J P, WRITTEN, WU Y F, TRANS. Pattern recognition concepts, methods and applications. 2nd ed [M]. Beijing: TsinghuaUniversity Press, 2002.

(责任编辑 刘存英)