一种改进的 K_means 算法在旅游客户细分中的应用

汪永旗
（浙江旅游职业学院 旅管系, 浙江 杭州 311231）

摘要: 针对传统 K_means 算法存在的问题, 提出一种基于密度的初始中心点选择方法, 并利用几何三角形三边关系理论简化了迭代中的计算次数, 以缩短大数据集聚类时间. 针对旅游电子商务的特点, 基于 RFM 模型设计了一种 RFMVCI 扩展模型. 新算法的有效性和扩展模型的合理性在实验和旅游客户细分实践中获得了验证.

关键词: K_means; 密度; RFM 扩展模型; 游客细分
中图分类号: TP39 文献标识码: A 文章编号: 1001-5132（2012）03-0058-04

近年来, 随着我国旅游业的快速发展, 旅游电子商务企业的数量和规模也在快速增大, 但是企业的客户服务质量的提高却相对滞后, 如何提高游客的服务质量已经成为大多数旅游电子商务企业发展的瓶颈问题, 而客户细分正是提高客户服务能力的重要手段.

聚类分析是数据挖掘领域的一个重要分支, 通过聚类可以将数据对象分成若干个簇. 常用的聚类算法有基于平面划分的算法、基于层次的算法、基于密度的算法、基于网格的算法等. K_means 算法是一种基于划分的聚类算法, 算法简单、快速, 但其聚类结果的好坏取决于聚类数、初始聚类中心选择、样本输入次序以及数据的几何特性等. 传统 K_means 算法的初始聚类中心是随机从数据集中产生的, 容易陷入最小局部最优解, 且聚类结果不稳定. 针对传统 K_means 算法缺点, 研究人员提出许多改进算法, 如 Dhillon 等[1]调整了迭代中聚类中心计算方法; Pelleg 等[2]提出了 X_means 算法以加快迭代过程; Sarafis 等[3]在构建目标函数中引入遗传算法; Alsabti 等[4]用 k-d 树结构改进 K_means 算法; 曹志宇等[5]提出了一种新的基于数据样本分布选取初始聚类中心的方法.

基于密度的聚类算法对噪声数据和数据输入顺序不敏感, 可以弥补 K_均值算法中随机选择初始中心点的缺点, 因此, 笔者结合两者提出了一种基于密度的 K_means 改进算法, 并在迭代中引入几何三角形三边关系理论来简化计算的次数.

1 K_means 算法

K_means 算法是划分聚类方法中的一种基于质心的技术, 它以 k 为参数, 将 n 个对象分为 k 个类, 以使类具有较高的相似度, 而类间的相似度则较低, 而相似度的计算则根据类中对象的平均值来进行. 设 k 是算法的输入参数, 代表输出的聚类数量, 数据集有 n 个对象组成, 初始化时, 根据输入参数 k 从数据对象 $\{i_1, i_2, \cdots, i_n\}$ 随机找出 k 个 $\{w_1, w_2, \cdots, w_k\}$ 作为簇的初始中心, 对剩余的每个对象根据其与各个簇中心的距离, 将它分配给最相似的簇, 然后计算每个簇的新均值; 不断重复这个过程, 直到准则函数(1)收敛, 相应公式为:

$$\sum_{j=1}^{k}\sum_{i_l \in c_j}\left|i_l - w_j\right|^2 . \tag{1}$$

K_means 算法描述如下: 输入: 包含 n 个对象的数据集及簇的数目 k; 输出: k 个簇的集合.

算法步骤如下:

(1) 初始化 k 个簇中心 $\{w_1, w_2, \cdots, w_k\}$, 其中 $w_j = i_l, j \in \{1,2,\cdots,k\}, l \in \{1,2,\cdots,n\}$;

(2) repeat;

收稿日期: 2012−02−07. 宁波大学学报（理工版）网址: http://3xb.nbu.edu.cn
基金项目: 浙江省旅游科学研究项目（2011ZC33）; 浙江旅游职业学院课题项目（2011XQZD02）.
作者简介: 汪永旗（1973−）, 男, 浙江衢州人, 硕士/系统分析师, 主要研究方向: 旅游电子商务和数据挖掘. E-mail: wyqi@tczj.com

(3) for 每个输入向量 i_l, 其中 $l \in \{1,2,\cdots,n\}$ do;

(4) 将 i_l 分配给最近的簇中心 w_j^* 所属的聚类 C_j^*, 即 $|i_l - w_j^*| \leqslant |i_l - w_j|, j \in \{1,2,\cdots,k\}$;

(5) for 每个聚类 C_j, 其中 $j \in \{1,2,\cdots,k\}$ do;

(6) 将簇中心更新为当前的 C_j 中所有样本的中心点;

(7) 计算准则函数 E;

(8) until E 不再明显地改变或成员不再变化.

2 K_means 算法的改进

虽然 K_means 算法应用广泛, 但它存在着几个缺点: (1)较难选择合适的 k 值, 它往往与应用领域密切相关; (2)随机中心点的选取会造成迭代次数和算法复杂度的不同, 聚类结果也可能不同; (3)因为 1 个极值可能很大程度上扭曲数据的分布, 所以 K_means 算法对噪声是敏感的, 且不能解决任意形状的数据聚类问题[6]; (4)在实际应用中, 如果数据量较大时, 算法的时间开销相当可观[7-9].

针对以上算法中的第(1)~(3)个缺点, 笔者引入了基于密度的方法来确定初始中心点, 从而弥补 K_means 算法只适合于解决凸状分布的数据类型问题; 针对第(4)个缺点, 笔者引入几何三角形三边关系理论来简化比较和计算的次数.

2.1 基于密度方法改进初始中心点选择

基于密度方法的基本思想是只要单个区域中的点密度大于某个阈值, 就将它加到与之相近的聚类中. 传统 K 均值算法用欧氏距离作为相似性度量的标准, 随机选择初始中心不能体现数据的分布情况, 而相互距离最远的 k 个对象更具有代表性. 因此为了避免低密度区中噪声点的干扰, 笔者从高密度区域中选择 k 个相距最远的点作为初始聚类中心.

定义 **1**　2 个对象的距离公式为:

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}, \tag{2}$$

其中, d_{ij} 为 2 个 p 维对象 i 和 j 的距离; x_{ik} 为对象 i 在 k 维的数据.

定义 **2**　ε 邻域为给定对象半径 ε 内的区域. 例如, 1 个点 p 的 ε 邻域为 $N_\varepsilon(p) = \{o \in D \mid dist(p, o) \leqslant \varepsilon\}$.

定义 **3**　核心对象: 如果 1 个对象 (p) 的 ε 邻域至少包含最小数目 MinPts 个对象, 则称该对象为核心对象.

改进后的基于密度的中心点初始化算法描述如下: 输入: 包含 n 个对象的数据集、簇的数目 k、邻域半径 ε 以及邻域包含对象的最小数目 MinPts; 输出: k 个初始中心点对象.

算法步骤如下:

(1) 计算所有 n 个对象的距离 d_{ij};

(2) 计算所有对象的 ε 邻域, 将所有 ε 邻域大于 MinPts 的对象加入到集合 T 中, 同时保存每个对象的 ε 邻域值;

(3) 在 T 中找出 ε 邻域最大的对象 c_1, 加入到初始中心集合 D 中, 并将 c_1 从 T 中删除;

(4) 计算 c_1 和集合 T 中所有对象的距离, 找出离 c_1 最远的数据对象 c_2, 将 c_2 加入初始中心集合 D 中, 并从集合 T 中删除 c_2 对象;

(5) 在集合 T 中找出离 c_1 和 c_2 最远的对象 c_3, 将它加入到 D 中, 并从集合 T 中删除 c_3 对象;

(6) 继续在 T 中查找对象 c_j, 直到找到第 k 个初始中心点对象.

2.2 基于三边关系理论简化迭代中的计算

在传统 K_means 算法中, 把每个对象归属到离它最近的中心所在类这个过程的时间复杂度为 $O(nkd)$, 其中, n 为对象的个数, k 为聚类数, d 为数据对象的维数. 在数据量较大和数据对象维数较多的实际应用中, 该算法的时间开销较为可观.

由于 K_means 算法中采用欧氏距离来衡量对象之间的相似性, 所以笔者考虑用三角形中两边之和大于第三边的关系理论来简化计算过程, 在一次迭代过程中的算法如下:

(1) 计算任意 2 个聚类中心的距离 $d(c_i, c_j)$, 其中, $i = 1,2,\cdots,k, j = 1,2,\cdots,k$;

(2) 计算对象 x_i 与本来所在类中心的距离 $d(x_i, C_m)$, 若 $d(C_m, C_n) \geqslant 2d(x_i, C_m)$, 则不成立, 就计算 $d(x_i, C_n)$; 若 $d(x_i, C_m) < d(x_i, C_n)$, 则暂时将 x_i 归到 C_m 类中;

(3) 继续第(2)步, 直到将 x_i 归属到最近的类.

该改进算法时间复杂度为 $O(nvd)$, 其中 v 为一次迭代过程中 1 个对象的平均计算次数, $1 \leqslant v \leqslant k$; 即在一次迭代中, 对象到中心点的计算次数最

好情况下是计算 1 次, 最坏情况下是计算 k 次, 如果样本集 (n) 较大时, 算法的效率提高是明显的.

3 实验分析

为了验证改进算法的可行性和有效性, 笔者将传统的 K_means 算法、本文改进算法以及 SOM 算法进行对比实验. 采用的测试数据集是 UCI 的 Iris、Pima-indians-diabetes 和 Wine 数据. UCI 数据库中数据都有确定分类, 因此可以用准确率来直观地表示聚类的质量. Iris 数据集包含 150 条记录, 4 个属性; Pima-indians-diabetes 数据集包含 768 条记录, 8 个属性; Wine 数据集包含 178 条记录, 13 个属性. 在实验中, 对于传统的 K 均值算法, 随机选取不同的聚类初始中心点进行 15 次实验; 对于传统的 SOM 算法, 训练次数为 300 次; 而本文的改进算法只需进行 1 次实验, 实验结果见表 1.

表 1 三种算法实验聚类精度比较 %

数据集	传统 K_means			SOM	本文算法
	min	max	avg		
Iris	65.33	87.33	72.31	78.67	86.67
Pima-indians-diabetes	54.95	88.41	71.88	68.75	83.46
Wine	57.30	71.35	64.57	66.29	60.67

从表 1 中可以看出, 对于传统算法而言, 由于初始中心点的选择是随机的, 并不考虑数据的实际分布情况, 因此初始中心点选择对最后聚类精度影响较大, 所以其最大值与最小值相差较大. 而改进算法一开始就找到了准确的初始中心, 所以能得到稳定的聚类精度. 除 Wine 数据集外, 改进算法的聚类准确率相比于传统算法和 SOM 算法都有较大提高; 但 Wine 数据集的聚类精度却不如传统算法, 其主要原因是其 13 个属性的取值范围差距较大, 如 Alcohol 的范围是 0.34~5.08, 而 Proline 的范围则是 278~1680, 这就造成了在新算法中计算距离时, 相互距离最远的点并不能很好地代表数据的实际分布情况, 也说明该算法有局限性.

另外, 为了验证有效性, 用传统 K_means 算法和本文算法进行实验, 算法的执行时间见表 2.

表 2 算法执行时间比较

数据集	K_means/ms	本文算法/ms
Iris	21	83
Pima-indians-diabetes	24	135
Wine	20	91

由表 2 可以看出, 虽然本文算法的执行时间比传统的 K_means 算法要长, 分别是 4 倍、6.1 倍和 4.8 倍, 但对于数据集中数据量不是太大的情况下, 改进算法的延长时间可以接受. 但也应看到, 对于 Wine 数据集而言, 聚类精度不如传统 K_means 算法, 并且执行时间更长, 说明改进算法有一定的局限性. 针对于该局限性, 在实际应用中的解决办法是对原始数据进行规一化, 保持各个属性的取值范围不要相差太大, 从而保证在新算法中的计算距离值与实际分布情况的一致性.

4 改进算法在旅游客户细分中的应用

RFM 分析是一种经典的客户细分方法, RFM 模型是衡量客户价值和客户创利能力的重要工具和手段, 该模型通过单个客户的近期购买行为 (R)、购买的总体频率(F)以及花费(M)这 3 项指标来描述该客户的价值状况.

4.1 RFM 模型扩展

旅游电子商务是旅游行为和电子商务技术相结合的产物, 客户不仅在网站上完成了除旅游活动之外所有交易过程, 还可以在完成旅游活动之后, 在网站上进行点评和推荐活动, 而点评和推荐在旅游交易中是非常重要的影响因素. 因此, 基于旅游电子商务的特点, 除了传统的 RFM 指标以外, 笔者另外设计了 3 个新的指标: 网站访问(V)、网上点评(C)和网上推荐行为(I). 网站访问指标是指客户访问电子商务网站的次数, 客户访问网站时可能消费也可能不消费; 网上点评指标是指客户对某个旅游产品或自己的旅游经历进行点评的次数; 网上推荐指标是指推荐他人在该网站上进行消费的行为次数和金额.

笔者从某知名大型旅游电子商务网站的客户记录中抽取了 1000 条用户行为数据作为样本, 并对 6 个指标进行归一化处理, 表 3 则列示了 4 个样本数据.

4.2 应用分析

将包含 R,F,M,V,C,I 6个指标的1000个样本归一化数据通过改进的 K_means 算法聚类后, 得到 9 个类别, 表 4 则列示了其中 4 个典型分类.

表 3　旅游电子商务网站中客户样本的 6 项指标

序号	*R*	*F*	*M*	*V*	*C*	*I*
1	0.19	0.52	0.47	0.64	0.19	0.66
2	0.34	0.31	0.63	0.46	0.67	0.00
3	0.79	0.37	0.81	0.10	0.17	0.65
4	0.08	0.12	0.18	0.46	0.12	0.10

表 4　改进的 K_means 算法聚类后的客户分类情况

聚类类别	客户/人	*R*	*F*	*M*	*V*	*C*	*I*
C2	105	0.203	0.311	0.705	0.370	0.551	0.520
C4	277	0.032	0.446	0.316	0.709	0.788	0.389
C7	92	0.701	0.388	0.783	0.309	0.090	0.064
C9	52	0.000	0.000	0.000	0.230	0.112	0.000

从表 4 中可以看出, 类别 C2 的客户是忠诚客户, 有消费能力并且有时间花在网上; 类别 C4 的客户目前消费力不强, 可能是经常进行网上消费的年轻人群体, 是潜在忠诚客户; 类别 C7 的客户消费力很强, 但不太点评, 可能是中产阶级群体; 类别 C9 的客户没有消费, 但是有访问和点评, 说明对网站有关注, 可能是未来的客户. 根据这些分类结果, 针对不同的群体, 旅游电子商务网站可以有针对性地制订不同的营销策略和客户服务策略.

5　结论

从优化初始中心出发, 提出一种基于密度算法的改进 K_means 方法. 同时, 为了减少传统算法中每次迭代的计算次数, 提出了一种基于三角形三边关系理论的改进方法, 并且用 UCI 数据集聚类实验证明了改进算法的有效性. 此外, 在数据集较大的旅游电子商务客户聚类实践中也验证了该算法的实用价值. 今后, 将进一步研究初始中心选择过程中相邻高密度区的合并问题, 使初始中心更具代表性, 从而进一步减少迭代中的计算量.

参考文献:

[1] Dhillon I, Guan Y, Kogan J. Refining clusters in high dimensional data[C]. The 2nd SIAM IC-DM, Workshop on Clustering High Dimensional Data, Arlington, 2002.

[2] Pelleg D, Moore A. X-means: Extending K_means with efficient estimation of the number of the clusters[C]. Proceedings of the 17th ICML, 2000.

[3] Sarafis I, Zalzala A M S, Trinder P W. A genetic rule-based data clustering toolkit[C]. Congress on Evolutionary Computation,Honolulu, 2002.

[4] Alsabti K, RANKA S, SINGH V. An efficient K_means cluste-ring algorithm[C]. IPPS/SPDP Workshop on High Performance Data Mining,Florida Orlando,1998.

[5] 曹志宇, 张忠林, 李元韬. 快速查找初始聚类中心的 K_means 算法[J]. 兰州交通大学学报, 2009, 28(6):15-18.

[6] 杨杰, 姚莉秀. 数据挖掘技术及其应用[J]. 上海交通大学学报: 工学版, 2011(1):173-178.

[7] 汪军, 王传玉, 周鸣争. 半监督的改进 K-均值聚类算法[J]. 计算机工程与应用, 2009, 45(28):137-139.

[8] Chaturvedi A. K_modes clustering[J]. Journal of Classification, 2008, 18(1):35-55.

[9] Huang Z. A note on K_modes clustering[J]. Journal of Classification, 2007, 20(2): 257-261.

An Improved K_means Algorithm and its Application to Tourists Classification

WANG Yong-qi

(Department of Tourism Management, Tourism College of Zhejiang, Hangzhou 311231, China)

Abstract: An improved density-based K_means algorithm is presented for the existing problems of traditional K_means clustering algorithm, in which selection of initial center pointer is optimized. Also, the triangular trilateral relation theorem is introduced to reduce calculation complexity. An expanded RMF model (RFMVCI) is presented in applications of tourism electronic business, and the validity of new algorithm and rationality of extended model are validated in practice of tourism customer classification.

Key words: K_means; density; extended RFM model; tourists classification

（责任编辑　章践立）