自适应特征权重的K-means聚类算法

李四海[1],满自斌[2]
(1.甘肃中医学院,甘肃 兰州 730000;2.兰州理工大学,甘肃 兰州 730050)

摘 要:为提高传统K-means聚类算法在医学数据聚类中的准确率和稳定性,提出了一种自适应特征权重的K-means聚类算法AFW-K-means。该算法首先通过计算属性的均方差选取初始聚类中心,然后根据当前的迭代结果,按照类内紧密、类间远离的原则调整属性在距离公式中的特征权重,以便更准确地反映数据点在欧氏空间中的真实距离,最后选取UCI上的BCW乳腺肿瘤等数据集对算法的有效性进行验证。结果表明:算法的准确率和稳定性均明显好于传统K-means算法。

关键词:K-means;医学数据聚类;自适应特征权重;聚类评价;混淆矩阵
中图分类号:TP181 **文献标识码**:A **文章编号**:1673-629X(2013)06-0098-04
doi:10.3969/j.issn.1673-629X.2013.06.025

K-means Clustering Algorithm Based on Adaptive Feature Weighted

LI Si-hai[1], MAN Zi-bin[2]
(1. Gansu College of Traditional Chinese Medicine, Lanzhou 730000, China;
2. Lanzhou University of Technology, Lanzhou 730050, China)

Abstract: In order to improve the accuracy and stability of traditional K-means algorithm on medical data clustering, proposed an adaptive feature weighted K-means clustering algorithm named AFW-K-means. Firstly, initial clustering center was chosen by calculating mean square deviation of feature attribute. Then, according to the results of each iteration, the feature attribute weight in distance formula is modified based on the principle of minimum-in-cluster-distance and maximum-between-cluster-distance, which can reflect the true distance among the data points in the Euclidean space. Finally, the validity of the proposed approach is demonstrated by the experiment of UCI data set such as Breast Cancer Wisconsin data set. The results showed that the algorithm has higher precision of prediction and better stability than traditional K-means algorithm.

Key words: K-means; medical data clustering; AFW; cluster evaluation; confusion matrix

0 引 言

聚类算法将给定的一个数据集按照特定的距离度量划分为多个类,使得同一类中的对象之间尽可能相似,不同类中的对象之间尽可能相异。聚类算法广泛应用于语音识别、图像分割、机器视觉、数据压缩、基因工程及信息检索等领域[1,2]。

在各种聚类算法中,基于划分的传统K-means算法由于简单、易于实现且能对大型数据集进行有效分类而得到广泛使用。但该算法的聚类结果对初始中心点非常敏感,不当的初始聚类中心将导致聚类结果不稳定。为了选择合适的初始聚类中心,国内外进行了大量研究[3~6]。事实上,该算法还对数据维度敏感,由于算法的相似性度量是欧几里得距离,且认为所有属性在计算欧氏距离时的重要性相同,这种对属性重要性不加区分的处理方法很可能导致数据点在欧氏空间中产生距离失真:如果空间中的两点在重要属性上距离很近,但由于其他无关属性对距离的放大作用,这两点在欧氏空间中很可能被度量为最远[7]。由此可见,通过对属性赋予不同的特征权值,能够更准确地反映对象之间的相似性并改善聚类性能。

目前,度量属性对聚类重要性的方法有多种:基于Fisher线性判别率[8,9]、基于属性信息熵[10]、基于小波低频能量熵等方法。由于K-means算法是迭代算法,使用固定的特征权值进行相似性度量还不能很好地发挥其对欧式空间中坐标轴的伸缩作用,使用可变的特征权值将能够进一步改善聚类性能。

收稿日期:2012-09-10 **修回日期**:2012-12-18 **网络出版时间**:2013-03-05
基金项目:国家自然科学基金资助项目(51069004)
作者简介:李四海(1972-),男,甘肃榆中人,硕士,讲师,研究方向为模式识别、小波分析。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20130305.0816.023.html

针对固定特征权重不够灵活的不足,引入类内距离之和与类间距离之和的比值作为度量属性重要性的依据,该值越小,属性越重要。根据每次迭代的结果,动态地为属性赋予不同的权重,随着迭代的进行,算法能够自动识别不同属性的重要程度,重要属性的权重会不断增大,而次要属性的权值会逐步减小,权值的调整方向真实反映了不同属性对聚类的重要性[11]。使用 UCI 数据集对算法的有效性进行验证,结果表明,算法在准确率和稳定性方面均高于固定权重的 K-means 算法。

1　传统 K-means 算法

输入:n 个对象的数据集,聚类数目 k;

输出:k 个类的划分,满足平方误差准则。

K - means 算法:

a) 从 n 个对象中任选 k 个作为初始聚类中心;

b) 根据每个聚类中对象的均值,按最小距离原则对所有对象进行划分;

c) 重新计算每个聚类的均值;

d) 重复执行 b) 和 c) 直到每个聚类不再变化。

传统 K - means 算法使用欧氏距离来度量对象之间的相似度,距离越小,对象越相似:

$$d(m,n)=\sqrt{\sum_{j=1}^{m}(x_{mj}-x_{nj})^2} \tag{1}$$

上式表明,所有属性对相似度的计算具有同等作用。

2　自适应特征权重 K-means 聚类算法

2.1　相关定义

1) 将 n 个 m 维待聚类对象表示为如下的矩阵形式:

$$\boldsymbol{X}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}$$

为使不同属性上的数据具有可比性,也为了方便计算属性贡献度,将上述矩阵按维度归一化至[0.01, 1]。

2) 设当前迭代后将 n 个对象划分为 K 个聚类,每个聚类中的对象个数分别为:$n_1,n_2,\cdots,n_k$,则所有 K 个聚类在第 j 维属性上的类内距离之和为:

$$d_n=\sum_{k=1}^{K}\sum_{i=1}^{n_k}(x_{ij}-m_{kj})^2 \tag{2}$$

m_{kj} 为聚类 k 在第 j 维属性上的均值。

3) 所有 K 个聚类在第 j 维属性上的类间距离之和为:

$$d_w=\sum_{k=1}^{K}(m_{kj}-m_j)^2 \tag{3}$$

m_j 为数据集在第 j 维属性上的均值。

4) 根据当前迭代结果,计算属性 j 对聚类的贡献度:

$$c_j=d_w/d_n \tag{4}$$

类内紧凑、类间远离通常用来度量聚类的整体性能。对于单个属性,如果聚类的结果在该属性上是类内紧凑且类间远离的,则该属性区分对象的能力强,对聚类的贡献大;反之,如果该属性上类内松散且类间近邻,则该属性区分对象的能力弱,对聚类的贡献小。

2.2　特征权重

第 j 维属性的特征权重为:

$$w_j=c_j/\sum_{j=1}^{m}c_j \quad w_j\in[0,1],\ \sum_{j=1}^{m}w_j=1 \tag{5}$$

使用上式修正欧氏距离公式(1),得到加权的欧氏距离公式:

$$d(m,n)=\sqrt{\sum_{j=1}^{m}w_j(x_{mj}-x_{nj})^2} \tag{6}$$

特征权重是根据各属性的贡献度计算得出的。特征权重越大,说明该属性对聚类越重要,该属性在欧氏空间中的坐标轴应该进行较大拉伸;特征权重越小,说明该属性对聚类作用不大,该属性在欧氏空间中的坐标轴应该进行较大缩减。w_j 的调节作用相当于对欧氏空间进行归约,在归约后的子空间中进行聚类,更能反映数据集在欧氏空间中的分布情况,从而提高聚类性能。

2.3　初始聚类中心确定

K-means 算法对初始聚类中心敏感,不同的聚类中心对应不同的聚类结果。传统的基于距离度量的初始聚类中心确定方法没有考虑各属性的重要程度,存在距离失真现象,选取的初始聚类中心并不能完全刻画数据集内部的类结构。为了在实验阶段准确评估文中提出的 K-means 聚类算法在不同数据集上的性能,提出如下确定初始中心点的方法:分别确定对象在各属性上的中心点,由各属性上的中心点得到整个数据集的初始聚类中心。具体做法为:首先计算对象在各属性上的均值和均方差,均方差反映了该属性上的值相对于均值的离散程度,然后以均方差和聚类个数 K 构造偏移因子,最后根据均值和偏移因子计算得到初始聚类中心。这种方法简单、易用,能更真实地反映初始聚类中心在各个属性上的分布情况,能够满足实际聚类的需要。

2.4　算法描述

AFW-K-means 算法:

输入:$n\times m$ 数据集(n 和 m 分别为对象个数和属性

个数),聚类个数 k;

输出:k 个聚类,使每个对象到其所在聚类中心的加权欧氏距离最小。

算法步骤:

(1) 计算对象在各属性上的均值 ***mean*** 和均方差 v,结果均为 1 ×m 的行向量。

(2) 构造初始聚类中心 C:

$$C=\begin{cases}\{\mathbf{mean}\pm\dfrac{2v}{k-1}\times j,j=1,\cdots,k/2\}\cup\{\mathbf{mean}\},k\text{ 为奇数}\\ \{\mathbf{mean}\pm\dfrac{2v}{k}\times j,j=1,2,\cdots,k/2\},k\text{ 为偶数}\end{cases}$$

以上两式中的 $2v/k$ 和 $2v/(k-1)$ 称为偏移因子。

(3) 特征权重初始化:

$w_j=1/m,j=1,2,\cdots,m$

(4)计算每个数据对象到聚类中心的加权欧氏距离,按照距离最小原则为每个对象分配聚类号。重新计算聚类中心。

(5)判断是否存在对象个数为 0 的聚类,如果是,说明数据对象分布非常密集,偏移因子过大。将偏移因子减半,重新选取聚类中心。

(6)根据迭代结果,按照属性类内紧密、类间远离的原则调整每个属性的特征权重。

(7)重复执行(4)和(6)直至达到预定的迭代次数或每个聚类不再变化为止。

3 实验结果及分析

3.1 有效性指标

有效性指标对于评价聚类算法对类结构的刻画程度十分重要。常用的指标有聚类准确率和聚类熵。聚类准确率指标简单、直观,是评价聚类性能的一个有效指标。但是,由于聚类熵指标不能对 K-means 算法的均匀效应做出有效评价,所以该指标对非平衡数据集并不适用。文中使用聚类准确率、VD 和 VI 指标。设聚类结果的混淆矩阵为:

$$\begin{bmatrix} & C_1 & C_2 & \cdots & C_k\\ P_1 & n_{11} & n_{12} & \cdots & n_{1k}\\ P_2 & n_{21} & n_{22} & \cdots & n_{2k}\\ \cdots & \cdots & \cdots & \cdots & \cdots\\ P_k & n_{k1} & n_{k2} & \cdots & n_{kk}\end{bmatrix}$$

其中,$P_1,P_2,\cdots,P_k$ 表示由聚类算法得到的 k 个划分,$C_1,C_2,\cdots,C_k$ 是数据集内在的类结构。则混淆矩阵反映了类对象在各个划分上的分布情况。VD 和 VI 指标定义如下[12]:

$$\mathrm{VD}=(2n-\sum_i\max_j n_{ij}-\sum_j\max_i n_{ij})/2n$$

$$\mathrm{VI}=-\sum_i p_i\log p_i-\sum_j p_j\log p_j-2\sum_i\sum_j p_{ij}\log p_{ij}/p_ip_j$$

在计算以上两个指标时同时扫描混淆矩阵的行和列,所以其结果能够对不同 K-means 聚类算法的均匀效应做出有效评价,当以上两个指标的值较小时,说明聚类算法性能较好。由于聚类熵指标只扫描混淆矩阵的行,所以 VI 是聚类熵指标的改进版本。

3.2 仿真实验及结果分析

为验证算法的有效性,在 Matlab R2010b 平台下,对 K-means,基于信息熵的固定权重 K-means[8] 及文中算法的有效性指标进行检验,比较不同聚类算法的性能。

首先选取 UCI 上的 Iris 数据集说明文中算法对权重的调整过程。该数据集共有 4 个属性,其中 petal length 和 petal width 两个属性对聚类结果影响较大。K-means 算法连续运行 10 次,其平均迭代次数为 7.3 次,基于信息熵的固定权重 K-means 算法迭代次数为 5 次,文中算法经过 4 次迭代后收敛,说明文中的算法能够显著减少迭代次数。文中算法对 Iris 数据集各属性特征权重的调整情况如图 1 所示:

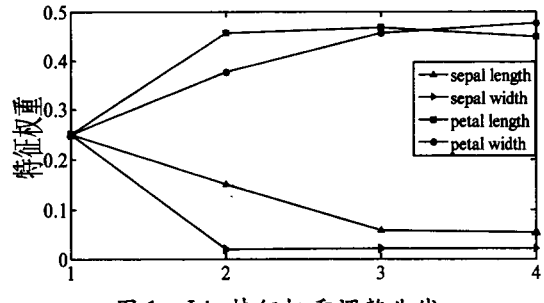

图 1 Iris 特征权重调整曲线

由图 1 可知,随着迭代的进行,算法能够自动识别属性的重要性,重要属性的特征权重逐步增大,次要属性的权重不断减小,最终 petal length 和 petal width 两个属性的权重由最初的 0.25 分别调整为 0.4484 和 0.4769。这种动态调整反映了各属性对类内紧密、类间远离聚类结果的重要程度,能够更真实地反映对象在欧氏空间中距离,减小距离失真程度,有利于提高聚类性能。

为验证文中算法在医学数据聚类中的有效性,选取美国威斯康辛州大学医学院的 Breast Cancer Wisconsin (BCW)数据集,该数据集包含细胞核图像的 10 个量化特征,分别为细胞核半径、质地、周长、面积、光滑性、紧密度、凹陷度、凹陷点数、对称度和断裂度,根据这些特征对患者诊断的结果为 B(benign)和 M(malignant)。三种算法中,K-means 算法平均准确率为 85.41%,文中算法经过 7 次迭代后收敛,聚类结果如

表 1 所示。在所有 357 个良性样本中,有 6 个被误诊为恶性,而在所有 212 个恶性样本中,有 32 个被误诊为良性,算法平均准确率为 93.32%。以上结果表明,文中算法用于乳腺肿瘤的医学辅助诊断是可行和有效的。

表 1　BCW 数据集聚类结果

	B	M
B	351	6
M	32	180

检验文中算法在 BCW 数据集的 VD 和 VI 指标,结果如表 2 所示:

表 2　BCW 数据集 VD 和 VI 比较

指标	K-means	固定权重 K-means	文中算法
VD	0.1459	0.0791	0.0668
VI	0.9264	0.7579	0.6516

VD 和 VI 的值较小时,说明聚类算法的有效性较好。由表 2 可知,文中算法的 VD、VI 值较小,说明对类结构的刻画程度要优于其他两种算法。事实上,当聚类算法的准确率都较高时,VD 指标和准确率指标等价,因为此时混淆矩阵每一行中的最大元素都出现在对角线上,所以有:准确率 = 1-VD。

以下选取更多的数据集,比较三种算法的迭代次数及聚类准确率。先选取 UCI 上的 4 个数据集,将三种算法各连续运行 10 次,比较其迭代次数,结果如图 2 所示,其中 K-means 算法为平均迭代次数。

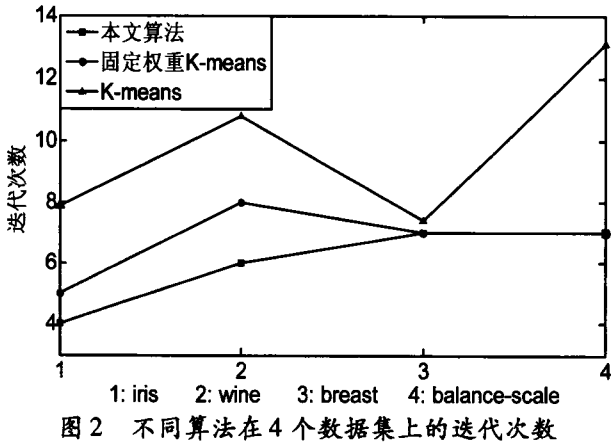

图 2　不同算法在 4 个数据集上的迭代次数

由图 2 可知,由于 K-means 聚类算法随机选取初始中心,算法稳定性不高,其平均迭代次数最多。文中算法的迭代次数略低于固定权重聚类算法,其收敛性能也优于固定权重聚类算法。通过与 K-means 算法迭代次数的比较,说明文中提出的初始聚类中心选取方法能够更好地反映原始数据集的内部结构,对于减少迭代次数是有效的。

为进一步验证文中算法的推广能力,选用 UCI 上的 10 个数据集,将三种算法各连续运行 10 次,聚类准确率的比较结果如表 3 所示,其中 K-means 算法的准确率是指 10 次运行的平均准确率。

表 3　不同 K-means 聚类算法的测试结果对比

数据集	K-means (%)	固定权重 K-means(%)	文中算法(%)
Iris	89.33	96	96
image	51.34	65.19	63.98
Wine	87.96	91.01	92.13
Breast	85.41	92.09	93.32
Balance_scale	48.00	47.04	54.88
lonosphere	71.23	70.66	72.36
Vehicle	36.29	37.59	44.21
Haberman	50.00	50.65	50.65
Waveform	49.76	50.08	50.22
Pima	67.06	67.06	67.89

分析测试结果发现,文中算法在 Breast、Haberman 和 Pima 等非平衡医学数据集上的准确率均明显高于传统 K-means 算法。当数据集各属性上的信息熵差别较大时,文中算法略优于固定权重 K-means,在 Balance_scale 和 Vehicle 等数据集上,各属性上的信息熵差别不大,固定权重 K-means 算法准确率较低,文中算法优势明显,这是由于此时固定权重 K-means 算法几乎退化为传统的 K-means 算法,如在 Balance_scale 上,各属性的信息熵均为 0.25,特征权值对对象相似度的计算不再有调节作用,此时该算法的准确率与 K-means 算法相当。文中提出的算法基于贪心策略,能够自动区分不同属性的重要性,根据每次的迭代结果,计算最好的特征权重,所以能保持相对较高的准确率且聚类结果稳定,当数据集有较为明显的无关属性时,效果更为显著。

4　结束语

提出了一种简单,易于使用的初始中心点选取方法,该方法能真实描述聚类中心在各属性上的分布情况,减少迭代次数,提高聚类算法的稳定性;在对象相似性度量中引入自适应特征权重,根据每次迭代的结果,在每个属性上计算类间距离之和与类内距离之和的比值,对欧氏空间进行一定程度的归约,消除无关属性对聚类的影响,以便更真实地反映对象之间的相似程度。实验结果表明,与传统的 K-means 算法相比,文中算法的聚类准确率更高,聚类结果更为稳定,能够有效提高非平衡医学数据集聚类性能。

参考文献:

[1]　孙吉贵, 刘　杰, 赵连宇. 聚类算法研究[J]. 软件学报,

(下转第 105 页)

码方式,利用领域搜索策略的特性,提出了能够有效地解决排课过程中的数学问题的优化算法 HQA。经过实验证明,HQA 符合实际要求,采用此算法可以排列出可靠、满意、稳定的班级课表和教师课表[13-15]。从排课过程中出现的数学问题出发,利用局部优化策略进行排列与组合,有效地减少了搜索次数,采取生优隔差的策略进行编码与解码,得到一组最优解。组合与优化是一种长期对资源分配与利用的过程,需要密切的配对与重生,组建成一个最优解。

因此,利用改进的混合量子算法提高管理水平,使教学的工作变得更加科学化和规范化,提高资源的利用率,保证课程的教学效果,加强学校的教学质量。

参考文献:

[1] 邢文训,谢金星.现代优化计算方法[M].第2版.北京:清华大学出版社,2005.

[2] 夏梦雨,叶春明,吴　勇.利用博弈演化算法求解置换 Flow Shop 调度问题[J].制造业自动化,2007,29(7):93-95.

[3] 何洋林,叶春明,朱露露.变批量车间调度文化进化算法研究及应用[J].机械科学与技术,2008,27(2):188-191.

[4] 王　萌,李蜀瑜.基于量子免疫进化算法的 Web 服务演化框架[J].计算机技术与发展,2012,22(3):39-43.

[5] 赵　吉,孙　俊,须文波.一种求解多峰函数优化问题的量子行为粒子群算法[J].计算机应用,2006,26(12):2956-2960.

[6] 周　殊,潘　炜,罗　斌.一种基于粒子群优化方法的改进量子遗传算法及应用[J].电子学报,2006,34(5):897-901.

[7] 傅家旗,叶春明.求解旅行商问题的混合量子算法[J].上海理工大学学报,2010,32(5):150-155.

[8] 叶春明,谢金华.改进混合量子算法在 Job Shop 调度中的研究[J].计算机工程与应用,2009,45(30):1077-1079.

[9] 李红婵,户　刚,朱颢东.基于群体优势遗传算法的高校排课问题研究[J].计算机工程与应用,2011,47(8):1235-1237.

[10] 金保华,李红婵.采用新型编码 GA 的高校排课问题仿真研究[J].计算机工程与应用,2011,47(13):1066-1069.

[11] 张立岩,张世民,秦　敏.基于改进粒子群算法排课问题研究[J].河北科技大学学报,2011,32(3):55-58.

[12] Tu Zhenguo, Lu Yong. A Robust Stochastic Genetic Algorithm (stGA) for Global Numerical Optimization[J]. IEEE Transactions on Evolutionary Computation, 2004, 8(5): 456-470.

[13] Talbi H, Draa A, Batouche M. A New Quantum-inspired Genetic Algorithm for Solving the Traveling Salesman Problem [C]//2004 IEEE International Conference on Industrial Technology. [s. l.]: [s. n.], 2004: 8-10.

[14] Yuan Bo, Gallagher M. On the Importance of Diversity Maintenance in Estimation of Distribution Algorithms[C]//Proceedings of the 2005 Genetic and Evolutionary Computation Conference. Washington, D C: ACM, 2005: 719-726.

[15] Li P C, Li S Y. Quantum-inspired evolutionary algorithm for continuous space optimization based on bloch coordinates of qubits[J]. Neuroeomputing, 2008, 72(1-3): 581-591.

(上接第 101 页)

2008,19(1):48-61.

[2] 金永波.动态聚类算法及其在医学数据上的应用[D].杭州:浙江大学,2011.

[3] 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.

[4] 周爱武,于亚飞.K-means 聚类算法的研究[J].计算机技术与发展,2011,21(2):62-65.

[5] Xu Junling, Xu Baowen, Zhang Weifeng. Stable Initialization Scheme for K-means Clustering[J]. Wuhan University Journal of Natural Sciences, 2009, 14(1): 24-28.

[6] Kang P, Cho S. K-means clustering seeds initialization based on centrality, sparsity, and isotropy[C]//Proceedings of the 10th International Conference on Intelligent Data Engineering and Automated Learning. Berlin: Springer, 2009: 109-117.

[7] 王熙熙,王亚东,湛　燕,等.学习特征权值对 K-均值聚类算法的优化[J].计算机研究与发展,2003,40(6):869-873.

[8] Modha D S, Spangler W S. Feature Weighting in K-means Clustering[J]. Machine Learning, 2003, 52(3): 217-237.

[9] 杨鹤标,薛艳峰,冯进兰,等.基于 Fisher 线性判别率的加权 K-means 聚类算法[J].计算机应用研究,2010,27(12):4439-4442.

[10] 原福永,张晓彩,罗思标.基于信息熵的精确属性赋权 K-means 聚类算法[J].计算机应用,2011,31(6):1675-1677.

[11] Tsai C Y, Chiu C C. Developing a feature weight self-adjustment mechanism for a K-means clustering algorithm[J]. Computational Statistics & Data Analysis, 2008, 52(10): 4658-4672.

[12] Wu Junjie, Xiong Hui, Chen Jian. Adapting the Right Measures for K-means Clustering[C]//Proceedings of The 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2009). Paris: [s. n.], 2009: 877-886.