结合初始中心优化和特征加权的K-Means聚类算法

王宏杰 师彦文
(西南石油大学计算机科学学院 成都 610500)

摘 要 为了提高传统K-Means聚类算法的聚类准确性,提出一种结合初始中心优化和特征加权的改进K-Means聚类算法。首先,根据样本特征对聚类的贡献程度获得初始特征权重,构建一种加权距离度量。其次,利用提出的初始聚类中心选择方法获得k个初始聚类中心,并结合初始特征权重进行初步聚类。然后,根据聚类精度来调整特征权重并再次执行聚类过程。重复执行上述过程直到聚类精度不再变化,获得最终的聚类结果。在UCI数据库上的实验结果表明,与现有相关K-Means聚类算法相比,该算法具有较高的聚类准确性。

关键词 K-Means聚类,贡献因子,特征加权,初始聚类中心优化
中图法分类号 TP393 **文献标识码** A

K-Means Clustering Algorithm Based on Initial Center Optimization and Feature Weighted

WANG Hong-jie SHI Yan-wen
(School of Computer Science, Southwest Petroleum University, Chengdu 610500, China)

Abstract In order to improve the clustering accuracy of traditional K-Means clustering algorithm, an improved K-Means clustering algorithm based on initial center optimization and feature weighted was proposed. Firstly, the initial feature weight is obtained based on the contribution factor of sample feature for clustering, and a weighted distance metric is constructed. Next, the k initial clustering centers are obtained by using the proposed initial clustering center selection method, and the initial clustering is performed with the initial feature weight. Then, the feature weights are adjusted according to the clustering accuracy and the clustering process is performed again. The above process is repeated until the clustering accuracy is no longer changed, resulting in the final clustering result. The experimental results on the UCI database show that the algorithm has high clustering accuracy compared with the existing K-Means clustering algorithm.

Keywords K-Means clustering, Contribution factor, Feature weighted, Initial clustering center optimization

1 引言

聚类技术分为监督学习和无监督学习两种。监督学习需要大量标记样本集来构建预测模型。在无监督学习中,基于集群相似性可将未标记样本分组到不同聚类中。每一对样本之间的近似度以欧氏距离为依据,且对于所有特征都是同等的[1]。K-Means聚类[2]是一种常见的聚类算法,通过将每个成员与聚类质心之间的距离最小化来实现聚类。然而,传统K-Means算法具有一些缺陷。例如其初始聚类中心是随机选择的,不同的聚类中心会获得不同的聚类结果[3],使聚类结果不稳定,且精度不高;另外,传统K-Means聚类过程中,所有样本特征的权重都是一样的。然而,实际上不同的特征与聚类的相关性不同,甚至一些特征与聚类不相关[4]。

目前,学者提出了一些改进型K-Means算法。例如典型的有Constrained K-Means算法[5],其从分类样本导出约束,使得同一类中的点具有最小的距离,而不同类中的点具有更大的距离。文献[6]将K-Means算法与模拟退火相结合,用来获得特征权重,进而使用加权欧氏距离度量来形成群集。然而,模拟退火算法需要大量迭代才能获得所需的权重,计算复杂度高。文献[7]提出了一种加权K-Means算法,其使用聚类中的大多数类样本之间的距离来修改权重并执行聚类。文献[8]在Constrained K-Means算法的基础上提出了一种成对Constrained K-Means(PC-KMeans)算法,其存在两种约束,即Must-Link约束和Cannot-Link约束,缓解了仅仅靠单一约束进行聚类的限制。

基于上述分析,针对传统K-Means算法的初始聚类中心选择和特征加权问题,提出了一种改进型K-Means算法。根据样本特征对聚类的贡献程度来获得初始特征权重;根据样本与样本之间的距离定义了一个聚类中心近似度量,以此获得k个初始聚类中心;基于每次聚类后的聚类精度来调整特征权重。重复执行上述过程直到获得最优的聚类结果。在UCI数据库进行了相关实验,结果证明了所提算法具有较高的聚类性能,平均聚类准确性达到了89%。

2 K-Means聚类算法

K-Means是一种流行的自聚类算法,通过最小化数据点与聚类中心之间的距离,将数据集分成k个聚类[9]。假设数据点集表示为$X=\{x_1,x_2,\cdots,x_N\}$,$x_i\in R^d$,则目标函数定义如下:

$$\Phi=\sum_{l=1}^{K}\sum_{X_i\in X_l}\|x_i-\mu_l\|^2 \tag{1}$$

王宏杰(1992-),男,硕士生,主要研究领域为机器学习、粒计算等;师彦文(1991-),男,硕士生,主要研究领域为机器学习、代价敏感等。

给定数据点集和目标函数，K-Means 算法将 X 分割为 K 个簇 $\{X_l\}_{l=1}^{K}$ 和质心集 $C=\{\mu_1,\mu_2,\cdots,\mu_K\}$。

使用欧氏距离度量来表示两个数据点 x_i 与 x_j 之间的距离：

$$d(x_i,x_j)=\sqrt{\sum_{m=1}^{p}(x_{im}-x_{jm})^2} \tag{2}$$

其中，$x_i=\{x_{i1},x_{i2},\cdots,x_{ip}\}$，$x_j=\{x_{j1},x_{j2},\cdots,x_{jp}\}$。

3 提出的改进 K-Means 聚类算法

3.1 加权欧氏距离度量

对于无监督学习，式(2)所示的距离度量中，所有特征在形成聚类时被认为同等重要。然而，数据集可以包含多个聚类解决方案，这取决于其使用目的，并且用于聚类的特征将决定如何形成聚类。因此，为了获得一个用于特定目标的最佳的聚类解决方案，只需要考虑相关的特征。一些特征可以被赋予权重 w_m 来显示其重要性，使其对距离函数具有不同的相对重要性，称为加权距离度量[10]。加权欧氏距离度量如下式所示：

$$\delta w(x_i,x_j)=\sqrt{\sum_{m=1}^{p}w_m(x_{im}-x_{jm})^2} \tag{3}$$

对于权重的设定，通常是先随机设置权重初始值，然后根据适应度函数迭代修改权重。这种方式所需的迭代次数较多，耗时较长。

本文使用多重判别函数来设定特征的初始权重。判别分析是一种降维技术，可用于查找可判别分组之间的预测因子。判别分析的线性方程表示为：

$$D=V_1X_1+V_2X_2+V_3X_3+\cdots+V_iX_i+a \tag{4}$$

其中，D 为判别函数，V 为该变量的判别系数或权重，X 为受访者对该变量的评分，a 为一个常数，i 为预测变量的数量。

3.2 初始权重的计算

在多类数据集中会得到多个判别函数，其中每个函数表示两个组之间的间隔。多个判别函数将给出不同特征的不同系数。为达到单一的权重，使其能够最好地描述在判别分组对象中特征的相关性，本文使用贡献因子这一概念。

贡献因子为一个在估计多于一个判别函数时，独立变量判别能力的综合衡量。贡献值包括变量对判别函数的贡献以及函数对整体解的相对贡献。贡献因子是所有重要判别函数的个体贡献值之和，它显示了变量的相对位置或等级。

贡献因子通过以下两步来计算。

步骤 1 计算每个变量对每个重要函数的贡献值(PV)：

$$PV_{ij}=DL_{ij}^2\times RE_j \tag{5}$$

其中，PV_{ij} 为变量 i 对函数 j 的贡献值。DL_{ij} 为判别负载，RE_j 为函数 j 的相关本征值。其中相关本征值 E_j 计算如下：

$$RE_j=\frac{E_j}{\sum_{j=1}^{N}E_j} \tag{6}$$

其中，E_j 表示判别函数 j 的本征值，$\sum_{j=1}^{N}E_j\sum_{i=1}^{N}E_i$ 表示所有重要判别函数的本征值之和。

步骤 2 计算变量对所有重要函数的综合贡献因子，其中，变量 i 的综合贡献因子 CPV_i 表示如下：

$$CPV_i=\sum_{j=1}^{N}PV_{ij} \tag{7}$$

贡献因子提供了特征的重要性等级，在包含 n 个特征的数据集中，第 i 个特征的权重计算如下：

$$W_i=CPV_i\times\frac{2}{n(n+1)} \tag{8}$$

3.3 初始聚类中心的选择

针对传统 K-Means 聚类中随机选择 k 个样本作为聚类中心，导致聚类结果不稳定[11]的问题，提出一种用于聚类中心选取的方法。其中，定义了一个聚类中心近似度量 Q_i，表示第 i 个样本 I_i 到其他任一样本的距离，表达式为：

$$Q_i=\sum_{j=1}^{K}e^{-\phi\|I_i-I_j\|^2} \tag{9}$$

其中，ϕ 为一个常数。提出的初始聚类中心选择算法的步骤如下：

1)计算每个样本的 Q_i 值，选择具有最大 Q_i 值的样本为第一个聚类中心 c_i。

2)计算样本与该聚类中心的隶属度值，根据隶属度值，将 c_i 和与 c_i 具有很高隶属度值的样本从下一次迭代中移除，防止产生伪聚类中心。

3)通过再一次计算 Q_i 值来确定下一个聚类中心，即 $Q_{i+1}=Q_i-Q_1^*e^{-\phi\|I_i-I_j^*\|^2}$。距离聚类中心越近的样本，其 Q_i 值就越小，成为下一个聚类中心的概率也就越低。

4)迭代执行上述过程，直到第 t 次迭代的 Q_t^* 满足 $Q_t^*<\varepsilon Q_1$ 为止。其中 ε 为一个很小的正数，这里取 $\varepsilon=10^{-3}$。

3.4 改进 K-Means 算法的流程

数据集包含标记和未标记的样本。标记样本用于获得判别函数，从而得到特征的初始权重。使用具有加权欧氏距离度量的 K-Means 算法进行聚类。根据簇中的大多数标记样本，为每个簇设定一个类别标记，这里簇的数量等于类别数量。

聚类精度被定义为正确分类到相应簇的样本百分比，表示如下：

$$\text{聚类精度}=\frac{\text{正确分类样本数}}{\text{总样本数}}\times 100\% \tag{10}$$

基于数据集中的标记样本计算聚类精度。然后，计算某个特征对聚类精度的贡献来修改权重值。即对于每个特征，测量当没有该特征时的聚类精度，用于获得该特征对聚类精度的贡献。如果新聚类精度小于初始聚类精度，则表明该特征较为重要，需要增加其权重；如果新聚类精度大于或等于初始聚类精度，则表明该特征不相关，需要去除该特征。

将修改后的权重进行归一化，使得权重之和为 1。基于新的权重，再次执行 K-Means 聚类，并计算聚类精度。如果聚类精度有所提高，则可以接受新权重。持续执行该过程，直到聚类精度不再明显变化。算法伪代码如算法 1 所示。

算法 1 提出的改进 K-Means 聚类

1. 对数据集进行判别分析，得到相关特征列表。
2. 计算每个特征的贡献因子，以此获得初始权重 A=(W_1x_1,W_2x_2,…,W_nx_n)。
3. 迭代执行提出的聚类中心选择方法，以此确定 k 个初始聚类中心。
4. 基于加权特征和初始聚类中心，执行 K-Means 以获得 K 个簇。
5. 计算初始聚类精度 C_{init}。
6. 对于 A 中的每个特征 i，执行没有该特征时的 K-Means 聚类，并计算聚类精度 C_i；若 $C_i<C_{init}$，则增加其权重 $W_{inew}=W_i(1+\frac{C_{init}-C_i}{C_{init}})$；否则移除该特征。
7. 归一化权重，基于新权重执行 K-Means 聚类，并计算聚类精度

C_{final}；若 $C_{final}>C_{init}$，则接受该新权重，并令 $C_{init}=C_{final}$；否则保持旧的权重不变。

8. 执行步骤 6，直到聚类精度没有明显改善。

4　实验及分析

4.1　数据集

采用了来自 UCI 数据库[12] 的 6 个基准数据集，分别为 Iris，Glass，Waveform，Vowel，Ionosphere，Wine。对于每个数据集，将其中 60%作为训练集，其余 40%作为测试集。这些数据集的属性描述如表 1 所列。

表 1　数据集的描述

序号	数据集	样本数	特征数	类别数
1	Iris	150	4	3
2	Glass	214	9	6
3	Waveform	5000	40	3
4	Vowel	990	13	11
5	Ionosphere	351	34	2
6	Wine	178	13	3

4.2　性能指标

为了更好地对聚类算法的性能进行评估，采用了聚类中的几种常用性能指标。

(1)准确性(Accuracy)

Accuracy 定义为正确聚类的数据比率[13]。为了有效表示准确性，使用了灵敏度和特异度两个度量。灵敏度(Sensitivity)表示正确识别的正值与所有正值的比值，特异性(Specificity)表示正确识别的负值与所有负值的比值。表达式分别如下：

$$sen=\frac{t_{pos}}{pos},spe=\frac{t_{neg}}{neg} \tag{11}$$

其中，t_{pos} 为正确聚类的正值的数量，pos 为正值的总数量，t_{neg} 为正确聚类的负值的数量，neg 为负值的总数量。则准确性可表示为：

$$Accuracy=(sen*\frac{pos}{pos+neg})+(spc*\frac{neg}{pos+neg}) \tag{12}$$

(2)调整兰德指数(Adjusted Rand Index，ARI)

假设 C 表示实际类别，K 表示聚类结果，a 表示在 C 与 K 中都是同类别的样本对数，b 表示在 C 与 K 中都是不同类别的样本对数。$C_2^{n_{samples}}$ 表示所有样本的可能组合数量。则兰德指数(Rand Index，RI)表示为：

$$RI=\frac{a+b}{C_2^{n_{samples}}} \tag{13}$$

RI 的取值范围为[0，1]，值越大表示聚类结果越准确。为了解决在随机产生聚类结果的情况下，RI 指标应该接近零的问题，形成了 ARI 指标，其取值范围为[−1，1]，表示为[14]：

$$ARI=\frac{RI-E[RI]}{\max(RI)-E[RI]} \tag{14}$$

(3)标准互信息(Normalized Mutual Information，NMI)

NMI 用来衡量两个数据分布的吻合程度，两个聚类的互信息表示为[15]：

$$I(X,Y)=\sum_{x,y}p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{15}$$

NMI 是基于熵将 MI 值调整到 0 与 1 之间，表达式如下：

$$NMI=\frac{I(X,Y)}{(H(X)+H(Y))/2} \tag{16}$$

4.3　性能比较

将提出的方法与传统 K-Means 聚类、半监督 Constrained K-Means 聚类和文献[8]提出的 PC-KMeans 方法，在 Iris，Glass，Wavrform，Vowel，Ionosphere，Wine 数据集上分别进行比较。其中，图 1 给出了 Accuracy 指标比较，图 2 给出了 ARI 指标比较，图 3 给出了 NMI 指标比较。

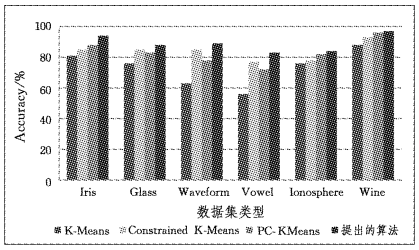

图 1　3 种聚类方法在 Accuracy 指标上的比较结果

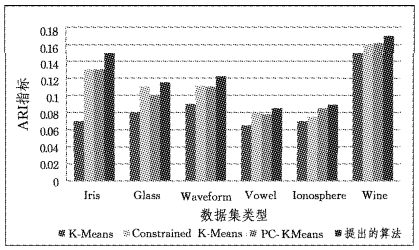

图 2　3 种聚类方法在 ARI 指标上的比较结果

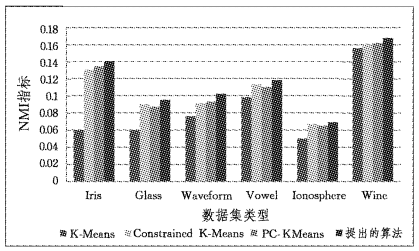

图 3　3 种聚类方法在 NMI 指标上的比较结果

可以看出，整体上而言，传统 K-Means 聚类算法的性能最差，Constrained K-Means 聚类和 PC-KMeans 聚类的性能相近，而本文提出的改进 K-Means 聚类性能最好。这是因为 Constrained K-Means 聚类是一种半监督聚类算法，在 K-Means 的基础上融入了由少数标签形成的 seed 集，并将其分为 K 个聚类，用来对算法进行初始化。Constrained K-Means 聚类明显提高了传统 K-Means 聚类算法的性能。然而，当样本中存在不规则簇时，Constrained K-Means 聚类的性能将会受到影响。PC-KMeans 聚类采用了两种约束进行聚类，但其仍然受到非球状分布数据的影响。而本文提出的聚类算法融入了初始聚类中心选择，有效解决了随机设置初始中心对聚类效果产生不利影响的问题。另外，根据特征对聚类精度的贡献因子对特征进行加权，提高了聚类的准确性。所提方法在 6 个数据集中的平均聚类准确度达到了 89.2%。

结束语　为了解决传统 K-Means 算法的初始聚类中心选择和特征加权问题，提出了一种结合初始中心优化和特征加权的改进 K-Means 聚类算法。根据特征贡献因子获得初始特征权重，利用一个聚类中心近似度量获得 k 个初始聚类中心。通过迭代聚类过程来优化权重并获得最优聚类。提出的算法有效解决了传统 K-Means 算法对初始聚类中心的敏感性和特征冗余的问题，有效提高了聚类性能。

(下转第 502 页)

表 4 全球 CAPE 型散货船舶轨迹(实验数据集 3)数据压缩效果对比

阈值/m	动态 D-P 压缩率/%	快速 D-P 压缩率/%	动态 D-P 误差/%	快速 D-P 误差/%	动态 D-P 压缩时间/s	快速 D-P 压缩时间/s
50	78.16	**80.64**	0.002411	**0.002154**	277	**152**
100	83.48	**83.67**	0.002521	**0.002174**	273	**147**
150	**86.26**	85.44	0.002632	**0.002200**	263	**145**
200	**88.10**	86.70	0.002735	**0.002233**	258	**139**

5.3 实验结论

综上所述,快速 D-P 算法相比于经典 D-P 算法和动态 D-P 算法在失真率更低的前提下,不仅提高了压缩率,而且在压缩速度上也最优。其可以有效地从船舶 AIS 数据找出可以体现原始轨迹特征的子集,舍弃了冗余、重复数据,从根本上解决了船舶 AIS 数据的存储难题。

结束语 本文提出的快速 Douglas-Peucker 算法具有更全面保留轨迹信息、压缩率更高、压缩速度更快的特点,对于解决全球船舶 AIS 轨迹的海量数据压缩具有重要的意义,有助于提升 AIS 大数据的加工和分析能力。该算法也可以推广应用到车辆、行人轨迹的压缩处理中,并对三维曲线和矢量图像压缩算法具有借鉴意义。同时,本文提出的压缩算法仍有进一步改善的空间:一方面,可以在本文算法的基础上提出对应的实时压缩算法;另一方面,可以考虑在本文算法中加入更多维的特征信息。AIS 数据中包含很多信息,本文压缩过程中仅仅利用了时间、经度、纬度信息,而船首向、船速等有效字段数据被忽略了。若可以将数据中的所有有效字段加以利用,将可以更清楚地体现船舶的航行特征,以达到进一步降低失真率的目的。

参 考 文 献

[1] DOUGLAS D H, PEUCKER T K. Alogrithms for the reduction of the number of points required to represent a digitized line or its caricature[J]. The Canadian Cartographer, 1973, 10(2): 112-122.

[2] 周岳明. 船舶自动识别系统的应用及其关键技术的研究[D]. 大连:大连海事大学,2002.

[3] 张宇. 通用船舶自动识别系统(AIS)及其关键技术研究[D]. 武汉:武汉理工大学,2004.

[4] 毕月琨. AIS 基站系统中的数据解析与压缩[D]. 杭州:浙江海洋学院,2014.

[5] 时学凯,王文珂,黄辉,等. 基于压缩域的脑成像大数据体可视化方法[J]. 计算机科学,2017,44(3):27-31.

[6] 冯飞,刘培学,李晓燕,等. 离散余弦变换在图像压缩算法中的研究[J]. 计算机科学,2016,43(11):240-241.

[7] 徐凯,李燕. 基于 Arduino 的 AIS 数据分布式采集研究[J]. 计算机测量与控制,2017,25(1):188-191.

[8] 李名,胡勤友,孟良. 基于 AIS 的船舶运动轨迹压缩技术研究[J]. 航海技术,2010(1):11-13.

[9] 王笑天,吕海洋. 基于第一特征点的道格拉斯-普克压缩算法[J]. 软件导刊,2016,15(11):68-70.

[10] 黄伟明,杨建宇,陈彦清,等. 基于扇形筛选法的矢量数据压缩方法[J]. 武汉大学学报(信息科学版),2016,41(4):487-491.

[11] 张树凯,刘正江,张显库,等. 基于 Douglas-Peucker 算法的船舶 AIS 航迹数据压缩[J]. 哈尔滨工程大学学报,2015(5):595-599.

[12] 王平利. 船舶自动识别系统应用关键技术研究[D]. 武汉:武汉理工大学,2007.

[13] ROSEN I. Real-time GPS track simplification algorithm for outdoor navigation of visually impaired[J]. Journal of Network and Computer Applications, 2012, 35(5): 1559-1567.

[14] SHI S Z, CHARLTON M. A new approach and procedure for generalizing vector-based maps of real-world features[J]. Gi-science & Remote Sensing, 2013, 50(4): 473-482.

[15] XU K, ZHEN H, LI Y, et al. Big Data Acquisition and Analysis Platform for Intermodal Transport[J]. International Journal of Database Theory and Application, 2016, 9(12): 67-78.

(上接第 459 页)

参 考 文 献

[1] 李孟杰,谢强,丁秋林. 基于正交非负矩阵分解的 K-means 聚类算法研究[J]. 计算机科学,2016,43(5):204-208.

[2] 胡海涛,朱建民. 基于 K-means 聚类的大学教学管理利益相关者分析[J]. 湘潭大学自然科学学报,2015,37(3):107-114.

[3] LIN K, LI X, ZHANG Z, et al. A K-means clustering with optimized initial center based on Hadoop platform[C]// International Conference on Computer Science & Education. IEEE, 2014: 263-266.

[4] YUNOH M F M, ABDULLAH S, NOPIAH Z M, et al. Fatigue Feature Extraction Analysis based on a K-Means Clustering Approach[J]. Journal of Mechanical Engineering and Sciences, 2015, 8(3): 1275-1282.

[5] CHEN Z, XUAN L, FAN Y. Constrained K-means with external information[C]// International Conference on Computer Science & Education. IEEE, 2013: 490-493.

[6] 何云斌,张晓瑞,万静,等. 基于改进遗传模拟退火 K-means 的心电波形的分类研究[J]. 计算机应用研究,2014,31(11):3328-3332.

[7] AMORIM R C D, MIRKIN B. Selecting the Minkowski Exponent for Intelligent K-Means with Feature Weighting[J]. Springer Optimization & Its Applications, 2014, 9(2): 103-117.

[8] COVOES T F, HRUSCHKA E R, GHOSH J. A study of K-Means-based algorithms for constrained clustering[J]. Intelligent Data Analysis, 2013, 17(3): 485-505.

[9] 王越,王泉,吕奇峰,等. 基于初始聚类中心优化和维间加权的改进 K-means 算法[J]. 重庆理工大学学报,2013,27(4):77-80.

[10] 王慧,申石磊. 一种改进的特征加权 K-means 聚类算法[J]. 微电子学与计算机,2010,27(7):161-163.

[11] AI H, LI W. K-means initial clustering center optimal algorithm based on estimating density and refining initial[C]// Information Science and Service Science and Data Mining. IEEE, 2012: 603-606.

[12] ASHOK P, NAWAZ G M K. Outlier Detection Method on UCI Repository Dataset by Entropy Based Rough K-means[J]. Defence Science Journal, 2016, 66(2): 113.

[13] 李翠霞,史苇杭,李占波. 一种基于密度的加权模糊均值聚类算法[J]. 计算机科学,2012,39(5):180-182.

[14] FEYEREISL J, AICKELIN U. Privileged information for data clustering[J]. Information Sciences, 2012, 194(5): 4-23.

[15] 侯勇,郑雪峰. 基于数据集特点的增强聚类集成算法[J]. 计算机应用,2013,33(8):2204-2207.