[文章编号] 1007－7405(2021)02－0139－07 DOI:10.19715/j.jmuzr.2021.02.08

基于逻辑回归函数的加权 K-means 聚类算法

林 丽[1]，薛 芳[2]

（1. 集美大学计算机学院，福建 厦门 361021；2. 集美大学信息化中心，福建 厦门 361021）

[摘要] 传统 K-means 聚类算法通过欧式距离计算样本的相似度，将数据所有的属性特征均平等对待，忽略每个属性特征的不同贡献，导致样本相似度计算的准确率不高。针对这个不足，提出一种特征加权的 K-means 算法进行优化。首先，运用 Softmax 和 Sigmoid 逻辑回归函数计算特征权重，使得加权的欧式距离更能准确地表示样本相似度；其次，优化初始聚类中心选择策略，选择距离较大的 K 个样本作为初始聚类中心，可有效避免样本的错误聚类及空簇问题。实验结果表明，在 UCI 标准数据集中采用加权 K-means 聚类算法可以有效减少迭代次数，提高聚类的准确率、精确率和召回率。

[关键词] 欧式距离；特征加权的 K-means 算法；逻辑回归函数；初始聚类中心

[中图分类号] TP 301.6

A Weighted K-means Clustering Algorithm Based on Logistic Regression Functions

LIN Li[1], XUE Fang[2]

(1. College of Computer Engineering, Jimei University, Xiamen 361021, China;
2. Informatization Center, Jimei University, Xiamen 361021, China)

Abstract: Traditional K-means clustering algorithms calculate the similarity of samples according to their Euclidean distance. All attributes of the data are treated equally and the potentially different contribution of each attribute is ignored. This can lead to a lack of accuracy in sample similarity calculations. To rectify this deficiency, a feature-weighted K-means algorithm is proposed. First of all, Softmax and Sigmoid logistic regression functions are used to calculate feature weights. The Euclidean distance after feature-weighting is able to represent the similarities and differences between samples more accurately. After this, the K samples with the largest distances between them are selected as clustering centers to optimize the strategy for selecting initial clustering centers. This can effectively avoid incorrect or empty sample clustering. Experimental results for application of the weighted K-means clustering algorithm to UCI standard datasets show that it is able to reduce the number of iterations and has better clustering accuracy, precision and recall rates than traditional K-means clustering approaches.

Keywords: Euclidean distance; feature-weighted K-means algorithm; logistic regression function; initial clustering center

[收稿日期] 2020－04－05
[基金项目] 国家自然科学基金项目（U1936114）；福建省自然科学基金项目（2020J01697）
[作者简介] 林丽（1983—），女，讲师，从事大数据技术、数据挖掘、机器学习方向研究。

0 引言

聚类分析在数据挖掘、文本摘要、图像识别领域有广泛应用，它是一种非常重要的机器学习算法。聚类算法能自动把数据对象划分成不同类别，每个类别中数据具有相似特征。通过聚类算法，可以在茫茫数据中挖出数据的规律。K-means 算法是一种基于划分的硬聚类算法，其运算速度快，尤其适用于高维数据、大规模及文本数据的聚类，是目前比较常用的一种聚类算法。

K-means 算法的做法是随机选择 K 个聚类中心，通过欧式距离计算各个样本和聚类中心的相似度。将样本分配给距离最近的类。K-means 算法存在 2 个问题：1）聚类结果不稳定；2）用欧式距离计算样本相似度。所有特征参与欧式距离计算且贡献度都一样，这往往带有大小不等的随机波动。针对问题 1），文献［1－3］提出 K-means + +算法，其采用概率来选择初始聚类中心，极大改善了传统 K-means 算法随机选择聚类中心的不确定性，提高聚类准确率。但是 K-means + +算法只考虑优化聚类中心，并没有考虑特征的不同贡献。关于问题 2）存在如下例子：如有二维样本（身高，体重），其中身高数值范围是 150 ~ 190，体重数值范围是 50 ~ 60，现有三个样本：$a(180,50)$，$b(190,50)$，$c(180,60)$。按照欧式距离计算样本相似度，$\mathrm{Dist}(a,b) = \mathrm{Dist}(a,c)$，那么身高 10 cm 真的等价于体重 10 kg 么？显然不是。而根据不同特征贡献加权的欧式距离去计算，样本的相似度才更准确。关于特征加权，文献［4－5］提出通过特征的总方差、均值来表示特征权重，但是方差并不能准确表示特征的差异情况。文献［6］提出的特征赋权主要基于同类内特征权重计算及学习。文献［7］提出采用 Pearso 相关系数来对数据对象间的距离进行加权，考虑样本和类中心的特征关联计算特征权重。文献［6］和文献［7］的特征赋权考虑是样本间的关联，但是并没有考虑特征本身的差异对特征权重的影响。

Softmax 函数可以凸显特征差异，Sigmoid 函数平滑极大极小特征差。本文提出一种特征赋权算法，它结合归一化指数函数 Softmax 和激活函数 Sigmoid 计算特征权重，特征差异低的特征属性贡献少则赋予较低权重，差异大的特征贡献大则赋予较大权重，以此尝试更好地解决 K-means 算法的两个问题。

1 相关知识

待聚类样本集 $X = \{x_i \mid x_i \in X, i = 1,2,\cdots,n\}$，每个样本都有 m 个特征 $x_i = \{x_{i1},x_{i2},\cdots,x_{im}\}$。

定义 1 样本 x_i 和样本 x_j 的欧几里得距离为：$\mathrm{Dist}(x_i,x_j) = \sqrt{\sum_{k=1}^{m}(x_{ik} - x_{jk})^2}$。其中 k 表示每个样本对应特征列。

定义 2 聚类的准确率 $r = (m/n) \times 100\%$。其中 m 为正确聚类的样本个数，n 为总样本个数。

定义 3 类误差平方和（SSE）$S = \sum_{i=1}^{n}\sum_{j=1}^{k}(x_i - c_j)^2$。其中 c_j 表示某个中心点，共有 k 个。

2 基于逻辑回归函数赋权的 LWK-means 算法

相对于传统 K-means 算法，本文提出的基于逻辑回归函数赋权的 LWK-means 算法改进了 2 个方面内容：一是聚类初始中心点优化，选择尽量远离样本作为初始聚类中心；二是设计特征权重计算方法，改进欧式距离计算公式。

2.1 初始聚类中心优化

聚类中心的优化策略为：先随机选一个样本做聚类中心，从第二类开始，计算每个数据对象 x 与已选聚类中心的最小距离 $\mathrm{minDist}(x)$。其中：在 $\mathrm{minDist}(x)$ 中选择最大值所在样本 j 作为下一个聚类中心；每一个聚类中心的确定都保证是和已有聚类中心的距离是较大的；可以让选择的聚类中心更贴近数据分布。实验表明，尽量远离聚类中心的做法，可以提高聚类算法的准确率。

InitCenter(X,k) 具体做法：

1）选择第一个样本作为第一个聚类中心 C_1，k'为当前类数量，$k'=1$。

2）根据定义 2 计算样本集 X 和已有类中心的距离 Dist(X,C)。

3）在 Dist(X,C) 中选择离现有类的最小距离作为样本和目前类 C_k 的距离值，

$$\min \text{Dist}(X) = \min(\text{Dist}(X,C_k))(k = 1,2,\cdots,k')。$$

4）min Dist(X) 中找出最大值所属的样本 j，作为下一个聚类中心 $C_{k'}$，$k' = k' + 1, C_{k'} = \arg\max(\min \text{Dist}(X))$。

5）重复步骤 2），直到选完 k 个聚类中心。

2.2　特征赋权

特征权重是改进算法中重要的步骤。特征权重计算主要是观察各个特征的差异情况。如果某个特征数据变化少，这个特征的差异度就少，该特征在分类中贡献度也少，赋低权重；如果某个特征数据变化大，该特征情况显著，对分类贡献大，赋高权重。特征差异度是在原始数据集上，即数据集未标准化前计算。保证特征差异度符合数据原始分布。

定义 4　特征差异度 $\boldsymbol{p} = \{p_1,p_2,\cdots,p_m\}$，表示每一列特征变化情况。是个一维向量。

特征差异度一般都用方差表示特征整体误差情况。由于各个特征取值范围不同，某些特征即使差异较大，若取值较小，也会导致方差较小。本文提出一个新的计算差异度公式衡量特征差异情况。令

$$p_i = (\max(x_i) - \min(x_i))/\text{avg}(x_i), \tag{1}$$

其中：x_i表示某一列的数据；$\max(x_i)$ 表示该列的最大值；$\min(x_i)$ 表示该列最小值；$\text{avg}(x_i)$ 表示该列平均值。极大极小差表示该列数据最大差异度，其与平均值比例可以从整体了解该列数据差异情况。$\boldsymbol{p}$ 值越大，特征差异度越大，该特征数据变化大。

定义 5　特征差异度的最大比值 $\max r = \max(p_i/p_j)$。p_i，p_j 分别表示第 i 列与第 j 列的特征差异度。主要通过这个参数了解是否存在极大极小特征差异度。

定义 6　特征权值 $w = \{w_1,w_2,\cdots,w_m\}$，表示 m 个特征在欧式距离计算的不同贡献度。使用逻辑回归函数 Softmax 计算，公式为：

$$w_i = e^{p_i}/\sum_{i=1}^{m} e^{p_i}, \tag{2}$$

其中 p_i 为每个特征差异度。p_i 越小，其贡献度越小；反之，则贡献度越大。

使用 Softmax 函数可以凸显最大值，抑制低于最大值的其他分量。但是 Softmax 函数会出现特征权重极大极小情况，当 $\max r$ 值超过 10，w_i 会完全偏向某一个特征。如某个数据集有 3 个特征权重分别为 [0.070,0.012,0.910]，第三个特征值比前 2 个大很多，特征权重计算会出现极大极小情况。由于特征值的取值范围不同，导致权重值完全失衡。Sigmoid 函数有很强的鲁棒性，各个特征权重值可以映射到 (0,1) 区间，平衡特征权重间的差异，故公式 (2) 也可以用 Sigmoid 函数设计：

$$w_i = (1 + e^{-p_i})^{-1}/\sum_{i=1}^{m}(1 + e^{-p_i})^{-1}。 \tag{3}$$

每个特征赋权后，欧式距离及类误差平方和 SSE 也改为加权欧式距离及加权的 SSE(S)。

$$\text{Dist}(x_i,x_j) = \sqrt{\sum_{k=1}^{m} w_k(x_{ik} - x_{jk})^2}, S = \sum_{i=1}^{n}\sum_{j=1}^{k} w\,(x_i - c_j)^2。 \tag{4}$$

2.3　LWK-means 算法

基于逻辑回归函数赋权的 LWK-means 算法流程为：

ⅰ）输入　样本数据集 X，聚类个数 K

　　输出　K 个聚类数

ⅱ）步骤

1）根据函数 InitCenter(X,k)，选择 k 个初始聚类中心 C_i。

2）根据公式（1）计算数据集的特征差异度 P_i。

3）根据定义 5 计算最大特征比值 max r。若 max $r>10$，说明存在极大极小特征权重问题，则选择公式（3）计算特征权重 w_i；若 max $r<10$，则选择公式（2）计算特征权重 w_i。

4）根据公式（4），计算样本和中心点的相似度 Dist(x,C)，取最小相似度作为样本归属类别。将样本分配至类别 L_i。

5）根据 L_i 划分的样本，计算同类样本在每一个特征的平均值，更新聚类中心 C_i。

6）加入特征权重计算误差平方和 SSE（公式（4）），重复步骤 2）→3）→4）→5），直至 SSE 不变或达到指定的迭代次数。

3 实验结果与分析

3.1 数据集

为了验证 LWK-means 算法的有效性及合理性。选用 UCI 数据库中的 6 个数据集作为仿真数据测试。表 1 为数据集说明。针对每个数据集，运行各算法 100 次。以算法的迭代次数、误差平方和、准确率作为有效性数据分析依据。并分别与 K-means、SWK-means、LWK-means 的聚类结果进行对比。其中：K-means 算法的聚类中心是随机选择，欧式距离不带权重；SWK-means 和 LWK-means 初始聚类中心的选择策略一样；SWK-means 算法的欧式距离权重值为各列的方差；LWK-means 算法的欧式距离权重值是基于特征差异度及逻辑回归函数计算的结果。数据预处理方面，采用 Z-Score 标准处理原始数据，保证不同维度数据的标准化。

表 1 UCI 数据集及说明

Tab. 1 UCI data set and description

数据集 Data set	样本数 The number of samples	特征数 The number of features	类别数 The number of clusters
Iris	150	4	3
Wine	178	13	3
Balance	625	4	3
Glass	214	9	6
Haberman	305	3	2
Seeds	210	7	3

3.2 数据集特征权重

特征权重是衡量各个特征贡献的指标。如 Iris 数据集，类别划分主要依据第 3、4 特征的数据，由于第 3、4 特征的数据差异度比第 1、2 特征大，故第 3、4 特征贡献大。使用 Softmax 函数凸显第 3、4 特征，根据公式（1）、（2）计算各特征权重后，对第 3、4 特征在欧式距离测量中赋予较高权重[0.29,0.45]（见表 2）。

表 2 数据集的特征权重

Tab. 2 Feature weights of data set

数据集 Data set	特征权重 Feature weights
Iris	[0.11,0.13,0.29,0.45]
Wine	[0.02,0.137,0.03,0.04,0.04,0.05,0.15,0.06,0.11,0.15 ,0.05,0.04,0.09]
Balance	[0.25,0.25,0.25,0.25]
Glass	[0.07,0.09,0.12,0.13,0.07,0.14,0.11,0.14,0.14]
Haberman	[0.30,0.24,0.44]
Seeds	[0.15,0.13,0.12,0.13,0.14,0.20,0.13]

从表 3 的实验结果可看到，Iris 数据经过特征赋权后，聚类正确率高达 95%。Haberman 和 Glass 数据集中特征差异度出现极大极小情况，使用 Sigmoid 回归函数平滑特征权重，即改善特征权重间的不平衡问题，也能按照特征的不同贡献对特征赋权（见表 2）。Balance 数据集每个特征都为离散型数据，列取值区间为［1,5］，各个特征变化情况一致，故各个特征权重一样（见表 2）。

3.3　实验结果及分析

运行 K-means、SWK-means、LWK-means 算法各 100 次，再取平均值。各算法在迭代次数、类误差平方和 SSE、类划分准确率的比较结果见表 3 和图 1 ~ 图 3。

表 3　迭代次数、类误差平方和 SSE、类划分准确率指标对比结果

Tab. 3　Comparison results of iterations, SSE, accuracy

数据集 Data set	迭代次数/次 Iterations/Times			SSE			准确率 Accuracy/%		
	K-means	SWK-means	LWK-means	K-means	SWK-means	LWK-means	K-means	SWK-means	LWK-means
Iris	6.50	4	4	149.27	25.0	24.24	78	86	95
Wine	7.54	8	7	1320.00	32.8	90.86	93	71	94
Balance	14.58	5	5	1753.05	442.0	449.79	46	65	65
Glass	8.44	5	6	856.77	76.0	97.72	41	41	51
Haberman	8.81	7	3	694.96	234.0	212.57	57	75	75
Seeds	9.06	8	7	430.85	52.2	64.62	92	90	92

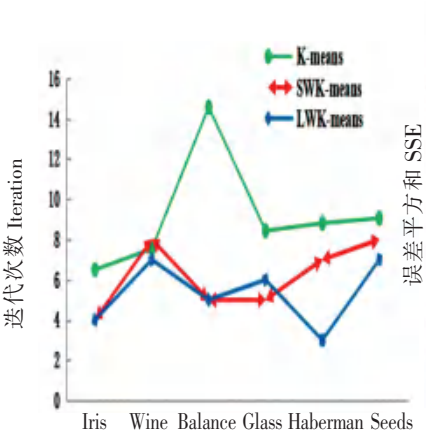

图 1　迭代次数比较结果

Fig.1　Comparison results of iterations

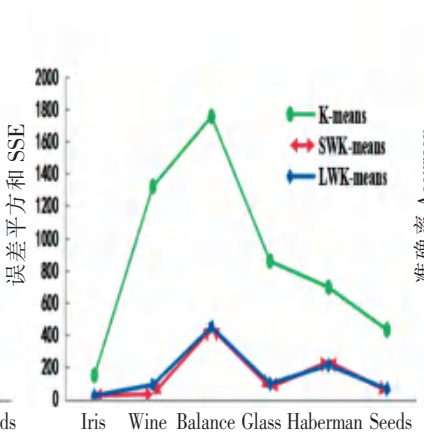

图 2　误差平方和比较结果

Fig.2　Comparison results of SSE

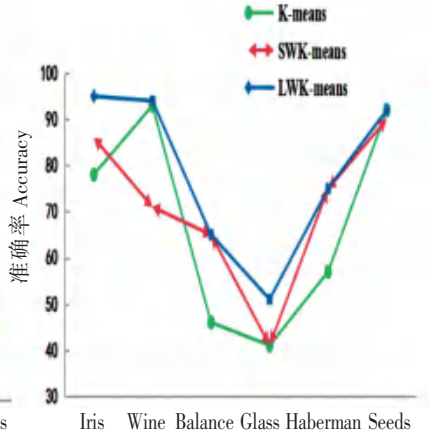

图 3　聚类正确率比较结果

Fig.3　Comparison results of accuracy

从表 3 及图 1 可以看到，在迭代次数方面，LWK-means 算法迭代次数均少于传统 K-means、SWK-means 算法。聚类中心优化后，LWK-means 算法的迭代次数在数据集 Haberman 中表现出更快的收敛，运行速度更优。类误差平方和 SSE 是聚类效果的一个重要衡量标准。SSE 越小，说明类内的误差越小，类越紧凑，聚类效果也越好。表 3 及图 2 显示加权后的 LWK-means 和 SWK-means 算法的 SSE 更小，加快了算法的收敛。Iris 数据集中，欧式距离经过特征加权后，LWK-mans 及 SWK-means 算法的准确率明显高于 K-means；其他数据集中，LWK-means 算法平均准确率也高于 K-means、SWK-means 算法。

准确率是一个很直观的评价指标，但是准确率高并没有反映出模型真正的能力，并不能代表某个算法就好。因此，引入另外两个指标：精确率（Precision）P 和召回率（Recall）R，P = 类别正确归类的样本数量/聚类后新类别的样本数量，R = 类别正确归类的样本数量/该类别原始的样本数量。精准确率用于检验聚类结果的有效性，召回率用于检查聚类结果的完整性。表 4 展示 2 个数据集中每个类别的精确率和召回率。图 4 和图 5 是 6 个数据集的精确率和召回率取均值后的柱状图。

由表 4、图 4、图 5 可以看出，LWK-means 在 6 个数据集中的精确率、召回率都优于其他 2 个算法；Balance、Haberman、Glass 数据集中，LWK-means 算法特征加权后的精确率和召回率都明显比未加权的 K-means 算法高。表 5 是对 6 个数据集的精确率、召回率、准确率取均值的比较结果：LWK-means 算法的精确率较未加入权重的 K-means 算法提高 7.6%，较方差加权的 SWK-means 算法提高 1.7%；其召回率较 K-means 算法提高 4.2%，较 SWK-means 算法提高 1.7%；其准确率较 K-means 算法提高 12.4%，较 SWK-mean 算法提高 3.3%。这 3 个指标再次证明 LWK-means 算法的有效性及稳定性。

K-means 算法采用欧式距离计算样本相似度，适用于球形数据分布，对非球形数据及离群点不敏感。从实验中看到，关于 Glass 和 Balance 等非球形数据集，3 种算法准确率较低。需要引入监督算法训练权重值。但从表 3 及图 3 可看到 LWK-means 算法平均准确率还是高于传统的 K-means 及 SWK-means 算法。

表 4 各算法精确率和召回率的比较结果

Tab. 4 Comparison results of precision and recall

数据集 Data set	类别 Cluster	精确率 Precision/%			召回率 Recall/%		
		K-means	SWK-means	LWK-means	K-means	SWK-means	LWK-means
Iris	类别 1	100	100	100	93	100	100
	类别 2	69	75	78	80	90	78
	类别 3	60	87	77	58	70	78
	平均值	76	87	85	77	87	85
Wine	类别 1	92	98	92	99	78	98
	类别 2	98	74	97	89	70	89
	类别 3	92	48	94	97	62	98
	平均值	94	73	94	95	70	95

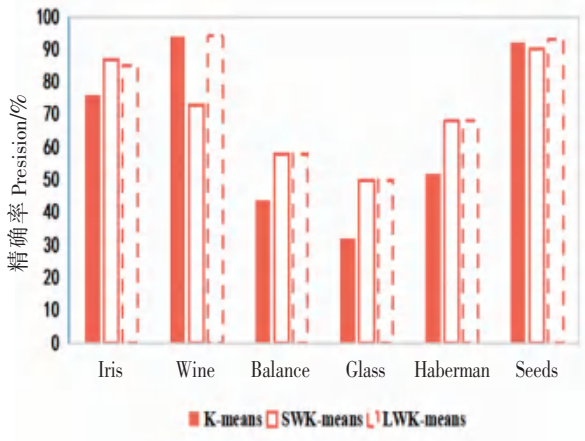

图 4 精确率指标对比结果

Fig.4 Comparison results of precision

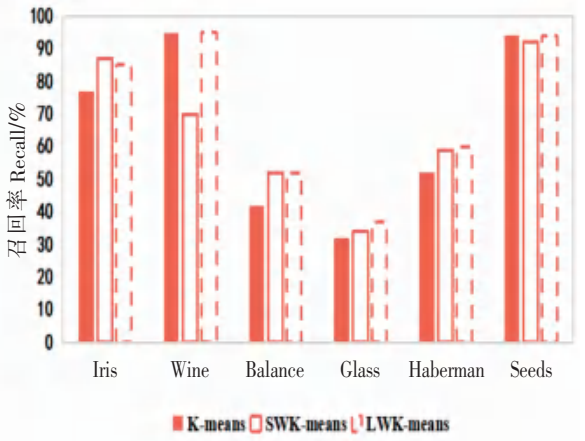

图 5 召回率指标对比结果

Fig.5 Comparison results of recall

表 5 各算法精确率、召回率、准确率指标对比结果

Tab. 5 Comparison results of precision, recall and accuracy

聚类算法 Clustering algorithm	精确率 Precision/%	召回率 Recall/%	准确率 Accuracy/%
K-means	65.3	65.1	66.5
SWK-means	71.2	67.6	75.6
LWK-means	72.9	69.3	78.9

4　结论

传统的 K-means 算法存在随机选择聚类中心、特征权重均等情况，导致聚类结果不稳定、样本相似度计算不准确的问题。本文针对 K-means 算法的不足，提出基于逻辑回归函数赋权的 LWK-means 算法。首先，优化聚类中心。选择距离较大的 K 个样本作为聚类中心，解决了传统 K-means 聚类中心不稳定问题，初始中心点尽量贴近类别本身。然后，根据每个特征差异度，使用 Softmax 和 Sigmoid 回归函数计算特征权重，为特征分配不同贡献度。经过加权的欧式距离计算方法可以提高样本相似度的计算准确率。实验选择了 UCI 数据库中的 6 个数据集，通过与 K-means、SWK-means 算法比较，LWK-mean 算法在迭代次数、误差平方和、准确率、精确率、召回率方面都表现更优性能。但是，在实验中也发现，欧式距离的计算方法并不适用所有数据集，后续研究工作中，可以引入监督学习方法改进样本相似度计算方法，提高聚类准确率。

[参 考 文 献]

[1] MUHAMMED MARUF ÖZTÜRK, UNAL CAVUSOGLU, AHMET ZENGIN. A novel defect prediction method for web pages using K-means + + [J]. Expert Systems with Applications, 2015, 42: 6496-6506.

[2] ARTHUR D, VASSILVITSKII S. K-means + +: the advantages of careful seeding [C] //Proceeding of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorihtms. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics. 2007: 1027-1035.

[3] 林涛，赵璨. 最近邻优化的 K-means 聚类算法 [J]. 计算机科学，2019，46(增刊 2)：216-219.

[4] 郭靖. 对 K-means 聚类算法欧氏距离加权系数的研究 [J]. 网络安全技术与应用，2016(10)：74-75.

[5] 冯荣耀，上官廷华，柳宏川. 一种基于均方差属性加权的 K-means 算法 [J]. 信息技术，2010，34(3)：55-57.

[6] 任江涛，施潇潇，孙婧昊，等. 一种改进的基于特征赋权的 K 均值聚类算法 [J]. 计算机科学，2006(7)：186-187.

[7] 刘建生，吴斌，章泽煜. 基于相关性加权的 K-means 算法 [J]. 江西理工大学学报，2018，39(1)：87-92.

[8] 李顺勇，张苗苗. 一种带权的混合数据聚类个数确定算法 [J]. 计算机应用与软件，2019，36(1)：284-290.

[9] 王巧玲，乔非，蒋友好. 基于聚合距离参数的改进 K-means 算法 [J]. 计算机应用，2019，39(9)：2586-2590.

[10] REDA M ELBASIONY, ELSAYED A SALLAM, TARE E ELTOBELY. A hybrid network intrusion detection framework based on random forests and weighted K-means [J]. Ain Shams Engineering Journal, 2013(4): 753-762.

[11] 张国锋，吴国文. 基于核函数的改进 K-means 文本聚类 [J]. 计算机应用与软件，2019，36(9)：281-284，301.

[12] 林志捷. 半监督与大规模数据聚类集成方法研究 [D]. 厦门：厦门大学，2018.

[13] ŠÁRKA BRODINOVÁ, PETER FILZMOSER, THOMAS ORTNER, et al. Robust and sparse K-means clustering for high-dimensional data [J]. Advances in Data Analysis and Classification, 2019, 13(4): 905-932.

[14] ARMANDO DI NARDO, MICHELE DI NATALE, CARLO GIUDICIANNI, et al. Weighted spectral clustering for water distribution network partitioning [J]. Applied Network Science, 2017, 2(1): 19-34.

(责任编辑　朱雪莲　英文审校　黄振坤)