Universe Scientific Publishing
Education and Learning,教育与学习（2）2020，12
ISSN:2705-0408(P);2705-0416(0)

簇方差加权 K-means 算法

赵挺祺　付学良

（内蒙古农业大学计算机与信息工程学院　内蒙古呼和浩特　010000）

【摘　要】随着信息技术的不断进步，如何从海量信息中有效地提取用户感兴趣的知识，已经成为当前数据挖掘的重要研究课题。而聚类作为数据挖掘的重要工具，通过将数据划分成多个类，使得类内数据尽可能相似，而类间数据的相似度尽可能小。从而挖掘类中的难以发现的隐含知识模式，而成为研究热点。聚类算法中 K-means 因为其简单、快速，常常被人们采用，但是 K-means 算法也存在对初始值敏感，容易被离群点影响聚类结果等缺点。本文提出了一种基于簇误差加权的 CVWK-means 算法，通过对误差大的簇进行加权处理提升聚类效果。实验结果表明，本文所提算法较原始 K-means 算法有更好的聚类效果。

【关键词】聚类；K-means；方差加权，聚类算法

DOI:10.18686/jyyxx.v2i12.39330

伴随着互联网时代的到来，用户可获取的数据信息海量增加，同时机器算力也呈摩尔倍数地上升，这为机器学习算法的应用提供了大量需求场景。机器学习算法包括分类算法和聚类算法，相比于分类算法，聚类算法因为其不依赖数据标注，更容易落地应用。K-means 算法因其简单易用，从众多聚类算法中脱颖而出。尽管其使用次数远超其他算法，但是并不能掩盖其固有的缺点。K-means 算法通过迭代最终得到簇中心和聚类结果，其运行结果对簇中心初始值选取和离群点非常敏感。

针对传统 K-means 算法存在的这些问题，学者们从不同角度尝试了许多改进方案。Alexandropoulos A，Plessas F，Birbas M 等人通过从样本整体中心从近及远采样的方法，选取初始聚类中心[1]。K-means＋＋是使用较广的一种方法，其思想是随机地选取相对较远的点作为初始聚类中心，因为直观上一个好的聚类应该是分散的[2]。实验证明，K-means＋＋可以保证较好的聚类效果和较快的收敛速度。不同于传统 K-means 的硬聚类算法，Arthur D，Vassilvitskii S 等人引入模糊聚类概念，使每个样本对所有簇都有一个隶属度。通过更新隶属度矩阵，得到各样本隶属度最大的簇作为其标签[3]。Bezdek J C，Ehrlich R 等人使用 GMM 算法进行聚类，认为每个样本符合不同的簇的高斯分布，使用样本数据拟合混合高斯分布[4]。Zhu X，Goldberg A B 等人使用 PCA 做数据预处理，然后应用高斯混合模型对处理后的数据聚类[5]。Liu J S，Zhang J L 等人认为不同的特征对聚类的贡献不同，提出一种基于特征加权的 K-means 算法，并应用 RliefF 算法自动计算特征权重[6]。

综上所述，学者们从簇中心初始值的选取、样本对应不同簇的隶属度、特征权重度量等等角度，对 K-means 算法尝试进行改进，并且取得了一定成果。基于对 K-means 的聚类目标函数的研究，本文发现 K-means 不同簇的方差在总体的目标函数中是和的关系，那么某一个聚类效果差的簇的方差，会被其他簇的方差平均。基于这个观察，本文提出一种簇方差加权目标函数，改进 K-means 算法本身。针对不同簇的方差权重，本文提出两种计算方法，自动地计算权重。通过在 UCI 数据集上的实验验证，本文算法在一些场景下明显提升 K-means 聚类效果，有效减少了部分簇聚类效果差的情况。

1　K-means 算法

K-means 算法的目的是要得到一个使得目标函数最小的簇划分，从而达到是得生成的簇尽可能紧凑和独立的划分结果。在描述 K-means 前定义一些符号。

聚类样本：$X(x^{(1)},x^{(2)},\cdots\cdots,x^{(n)}),x^{(i)}\in R^m$

聚类数目：K

簇中心：$C(c_1,c_2,\cdots\cdots,c_K),c_i\in sR^m$

对第 i 个样本 $x^{(i)}$，其归属于各个簇的概率为

$Z^{(i)}(z_1^{(i)},z_2^{(i)},\cdots\cdots,z_K^{(i)}),0\leqslant z_k^{(i)}\leqslant 1$

则 K-means 算法的目标函数为。

$$J=\sum_{i=1}^{N}\sum_{k=1}^{K}z_k^{(i)}\left||x^{(i)}-c_k\right||^2$$

K-means 算法本质上是一种 EM 算法，其优化步骤如下。

Step 1：初始化簇中心 C^0；

Step 2：根据每个样本距各个簇中心的距离，更新每个样本的类别标签。这一步对应 EM 算法的 E 步，更新样本归属于各个簇的概率，如果 $\left||x^{(i)}-c_k\right||^2=min_{1\leqslant j\leqslant k}\left||x^{(i)}-c_j\right||^2$，$P(z_k^{(i)}\mid x^{(i)},c)=1$；

否则，$P(z_k^{(i)}\mid x^{(i)},c)=0$。

Step 3：更新簇中心 C^t。这一步对应 EM 算法的 M 步，即最大化

$$\sum\nolimits_Z P(Z\mid X,c^{(i)})\,log\,(P(Z,X))$$

样本 Z 和簇标签 C 的联合概率 $P(Z,X)$ 计算方式为，如果 $\left||x^{(i)}-c_k\right||^2=min_{1\leqslant j\leqslant k}\left||x^{(i)}-c_j\right||^2$，

$$P(z_k^{(i)},x^{(i)}\mid c)=e^{-\left||x^{(i)}-c_k\right||^2}$$

否则，$P(z_k^{(i)},x^{(i)}\mid c)=0$

带入 Step 2 计算的 $P(Z\mid X,C)$，求极值得到使目标最大化的簇中心为：

$$c_k=\frac{\sum_{i=1}^{n}P(z_k^{(i)}\mid x^{(i)},c^{(i)})x^{(i)}}{\sum_{i=1}^{n}P(z_k^{(i)}\mid x^{(i)},c^{(i)})}$$

Step 4：反复 step 2 和 step 3，直到簇中心不再发生

变化。

2 簇方差加权 K-means 算法

观察 K-means 算法的目标函数 J，发现，如果有些簇聚类效果差，换言之簇内方差大，那它的方差会被方差小的簇平均，那么就得不到优化。基于这个想法，本文提出簇方差加权 K-means 算法（Cluster Variance Weighting K-means algorithm，CVWK-means）。通过对不同簇增加不同的权重，从而达到改良局部簇聚类质量的目的。CVWK-means 算法的目标函数为：

$$J_{cw}=\sum_{i=1}^{N}w_k^{\alpha}*\sum_{k=1}^{K}z_k^{(i)}\left\|x^{(i)}-c_k\right\|^2$$

其中 w_k 为簇权重系数，α 为调节簇权重系数影响的权重因子，要求满足，

$$\sum_{k}^{K}w_k=1,w_k>0$$

怎么计算簇权重系数 W 是簇方差加权 K-means 算法的关键，本文提出两种计算方法：

方法一，启发式方法，给予平均方差大的簇更高的惩罚，也就是更高的权重系数，则对第 j 个簇，其权重计算方式为：

$$h_j=\frac{\sum_{i=1}^{n}P(z_k^{(i)}\mid x^{(i)},c)\left\|x^{(i)}-c_j\right\|^2}{\sum_{i=1}^{n}P(z_k^{(i)}\mid x^{(i)},c)}$$

$$w_j=\frac{h_j}{\sum_{k=1}^{K}h_k}$$

方法二，自动学习，对簇方差大的簇施加一个大的惩罚系数，等价于如何求得一个使得 J_{cw} 值尽可能大的系数，这就转化为一个优化问题。

$$maxJ_{cw}=\sum_{i=1}^{N}w_k^{\alpha\sum_{i=1}^{K}z_k^{(i)}\left\|x^{(i)}-c_k\right\|^2}$$

满足条件，

$$\sum_{k}^{K}w_k=1,w_k>0$$

使用拉格朗日乘子法求解，可以得到权重的更新公式：

$$V_k=\sum_{i=1}^{N}z_k^{(i)}\left\|x^{(i)}-c_k\right\|^2$$

$$w_k=\frac{V_k^{\frac{1}{1-\alpha}}}{\sum_{j=1}^{K}V_j^{\frac{1}{1-\alpha}}}$$

此外为了使得收敛更平稳，权重更新采用移动指数加权平均方式。

$$w_k^{(t)}=\beta w_k^{(t-1)}+(1-\beta)\left(\frac{V_k^{\frac{1}{1-\alpha}}}{\sum_{j=1}^{K}V_j^{\frac{1}{1-\alpha}}}\right)$$

簇方差加权 K-means 算法具体步骤如下：

Step 1：初始化簇中心，初始化簇权重为 $w_i=1/K$；

Step 2：根据每个样本距各个簇中心的距离，更新每个样本的类别标签；

如果 $w_k\left\|x^{(i)}-c_k\right\|^2=min_{1\leqslant j\leqslant k}w_j\left\|x^{(i)}-c_j\right\|^2$

$$P(z_k^{(i)}\mid x^{(i)},c)=1,$$

否则，$P(z_k^{(i)}\mid x^{(i)},c)=0$

Step 3：更新簇中心 C^t；

$$c_k=\frac{\sum_{i=1}^{n}P(z_k^{(i)}\mid x^{(i)},c^{(i)})x^{(i)}}{\sum_{i=1}^{n}P(z_k^{(i)}\mid x^{(i)},c^{(i)})}$$

Step 4：更新簇权重 W，以启发式计算方式为例；

$$V_k=\sum_{i=1}^{N}z_k^{(i)}\left\|x^{(i)}-c_k\right\|^2$$

$$w_k^{(t)}=\beta w_k^{(t-1)}+(1-\beta)\left(\frac{V_k^{\frac{1}{1-\alpha}}}{\sum_{j=1}^{K}V_j^{\frac{1}{1-\alpha}}}\right)$$

Step 5：重复 Step 2-Step 5，直到簇中心不再变化。

3 实验及分析

实验的硬件环境为 Intel(R)Core(TM)i5-6500 3.20 GHz，8G 内存，软件环境为 Matlab2016b，Windows7 操作系统。实验数据集选择 UCI 数据集中的 Iris、Wine、Statlog-Australian credit approval、Yeast、Statlog-Image Segmentation 共计五个数据集，数据集的主要信息如表 1 所示。

表 1 UCI 数据集说明

Data set	样本数	特征数	类别数
Iris	150	4	3
Wine	178	13	3
Statlog-Australian credit approval	690	14	2
yeast	1484	8	10
Statlog-Image Segmentation	2310	19	7

实验参数及方法说明：

（1）簇权重更新方式采用第一种，权重调节因子、$\alpha = 0.4$；

（2）移动指数加权平均系数 $\beta = 0.7$；

（3）每个算法，在各数据集上，运行 10 次，取效果最好一次结果为最终结果；

（4）算法参数最大更新次数，max_iter = 100；

（5）另外，为了消除不同特征量纲差异对结果造成的影响，对数据进行了标准化预处理（使数据变为均值为 0，方差为 1 的分布）

算法性能度量指标说明见表 2。

表 2　聚类算法性能度量指标

purity	纯度，越接近 1 表示聚类结果越好；
NMI（Normalized Mutual Information）	归一化互信息，越大越好；
AMI(Adjusted Mutual Information)	调整后互信息，越大越好；
FMI（Fowlkes and Mallows Index）	FM 指数，越大越好。

聚类实验结果见表 3。

表 3　UCI 聚类算法效果

data set	algorithm	purity	NMI	AMI	FMI
Iris	K-Means	0.887	0.742	0.733	0.811
	K-Means++	0.887	0.742	0.733	0.811
	GMM	**0.967**	**0.9**	**0.897**	**0.936**
	CW-K-means	0.833	0.659	0.655	0.745
Statlog-Image Segmentation	K-Means	0.53	0.493	0.479	0.423
	K-Means++	0.558	0.533	0.506	0.483
	GMM	0.529	0.496	0.447	0.472
	CW-K-means	**0.575**	**0.681**	**0.562**	**0.62**
Wine	K-Means	0.702	0.429	0.423	0.584
	K-Means++	0.702	0.429	0.423	0.584
	GMM	0.697	0.518	0.477	0.672
	CW-K-means	**0.955**	**0.847**	**0.843**	**0.909**
Yeast	K-Means	0.524	0.256	0.216	0.282
	K-Means++	0.521	0.28	0.242	0.302
	GMM	0.448	0.184	0.159	0.271
	CW-K-means	**0.533**	**0.299**	**0.274**	**0.342**
Statlog-Australian credit approv	K-Means	0.562	0.034	0.007	0.707
	K-Means++	0.562	0.034	0.007	0.707
	GMM	0.562	0.034	0.007	0.707
	CW-K-means	**0.836**	**0.367**	**0.358**	**0.737**

可以看出，5 个数据集中，4 个均超过了常用算法，且在 Wine 和 Stalog-Australian credit approval 两个数据集上，大幅提升了性能。说明本文提出的簇方差加权 K-means 算法可以有效提升部分簇聚类效果差的问题。

在 Iris 数据集上，簇方差加权 K-means 算法效果不理想，经分析，造成这种差异的原因是不同算法适用于不同分布的数据聚类，比如 K-means 就不擅长非球类分布的数据聚类，而 GMM 在 Iris 数据集上性能远超其他算法。

分别进行实验来对比 K-means 算法聚类效果。通过测试这经典的 5 个数据集，将 K-means 算法、GMM 算法、K-means＋＋算法、CW-K-means 算法相比较。实验结果表明 CW-K-means 算法可以显著优化 K-means 部分簇聚类效果差的缺点，从而达到更好的聚类效果。

5　结语

本文提出了一种簇方差加权 K-means 算法，解决了

K-means算法聚类过程中，可能存在的部分簇聚类效果差的问题，本质上是针对潜在的簇分布不均匀问题进行了优化。实验结果表明：簇方差加权K-means算法在多个UCI公开数据集上获得了比其他聚类算法更优的效果。

K-means算法本身存在无法自动获取聚类簇数，初始值点敏感，异常值敏感等问题，尽管已有研究已经做出了一些改进，然而还没有一个更简单通用的算法取代K-means被大家接受，所以这些研究方向依然存在价值继续探索。

此外，聚类算法聚类的依据是什么，可能同样的数据存在不同层面的意义，比如给定一批任务信息数据，那聚类的结果那到底是以性别为依据，还是以身高为依据，聚类算法输出并不会给出这样的解释。所以后续聚类算法及其输出的解释，也是一个很有价值的研究点。

作者简介：赵挺祺(1993.9—)，女，研究生，研究方向：机器学习算法。

基金项目：典型湖泊水生态综合治理综合评价技术体系(2019YFC0409205)；大数据环境下乌梁素海湖泊湿地评价模型与保护机理研究（NO.2019MS06015）；大数据环境下寒旱区湖泊湿地智能监测与保护机理研究(61962047)。

【参考文献】

[1]Alexandropoulos A，Plessas F，Birbas M. A dynamic DFI-compatible strobe qualification system for Double Data Rate (DDR) physical interfaces[C]//2010 17th IEEE International Conference on Electronics，Circuits and Systems. IEEE，2010：277-280.

[2]Wong J A H A . Algorithm AS 136：A K-Means Clustering Algorithm[J]. Journal of the Royal Statistical Society，1979，28(1)：100-108.

[3]Arthur D，Vassilvitskii S . K-Means＋＋：The Advantages of Careful Seeding[C]// Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms，SODA 2007，New Orleans，Louisiana，USA，January 7-9，2007. ACM，2007.

[4]Bezdek J C，Ehrlich R，Full W . FCM：The fuzzy c-means clustering algorithm[J]. Computers &Geoences，1984，10(2-3)：191-203.

[5]Zhu X，Goldberg A B . Introduction to Semi-Supervised Learning[J]. Synthesis Lectures on Artificial Intelligence and Machine Learning，2009，3(1)：130.

[6]Liu J S，Zhang J L，Palumbo M J，et al. Bayesian clustering with variable and transformation selections[J]. Bayesian statistics，2003，7：249-275.

[7]徐艳，付学良，李宏慧.一种基于特征加权的K-Means算法研究[J].计算机科学与应用，2018，8(8)：8.