doi:10.3969/j.issn.1671-637X.2015.07.013

基于协同熵的K-均值算法

罗蜀君, 侯 飞, 毛 鑫
(中国航空工业集团公司洛阳电光设备研究所,河南 洛阳 471000)

摘 要:针对传统K-均值算法容易受到野点和噪声点的影响,缺乏鲁棒性的问题,提出了一种基于协同熵的K-均值算法。该方法利用协同熵作为一种局部的相似度度量手段,并依赖最大协同熵准则进行最优聚类中心的求解。采用迭代重加权的优化算法可以用来快速实现最优聚类中心的求解。对于残差较大的野点和噪声,它们在聚类中心更新的过程中将被赋予较小的权重。实验结果表明,基于协同熵的K-均值算法具有较好的鲁棒性,并获得较好的聚类效果。

关键词:K-均值算法; 协同熵; 聚类
中图分类号:O213.2 **文献标志码**:A **文章编号**:1671-637X(2015)07-066-04

K-Means Algorithm Based on Co-entropy

LUO Shu-jun, HOU Fei, MAO Xin
(Luoyang Institute of Electro-Optical Equipment, AVIC, Luoyang 471000, China)

Abstract: Considering the fact that conventional K-means algorithm is susceptible to the outliers and noise points, and lacking in robustness, a new K-means algorithm based on co-entropy is proposed. The proposed algorithm employs co-entropy as a means of local similarity measurement, and follows the co-entropy maximization principle to solve the optimal cluster centers. An iteratively reweighted optimization technique is employed to quickly find the optimal cluster centers. For outliers and noisy data points with larger residuals, they will be assigned smaller weights in updating the cluster centers. Experimental results demonstrate that the proposed co-entropy based K-means algorithm is robust, winning a better clustering effect.

Key words: K-means algorithm; co-entropy; cluster

0 引言

聚类分析是统计数据分析中一门很重要的技术,其目的是利用静态分类的方法将相似的对象分成不同的组别或者子集,使得在同一个子集中的对象具有相似的属性。近年来,涌现出很多聚类分析算法,它们被用来解决不同领域中的实际问题或者作为某些问题的有效预处理手段[1-8]。如在社交网络的研究中,可以利用聚类方法从大量用户中识别出社区;在图像分割中,聚类算法可以将数字图像分成不同的区域以进行边缘检测或者目标识别;推荐系统的作用是基于用户的偏好进行新项目的推荐,为了预测某一用户的偏好,通常利用聚类算法来检测出该用户所在的集群,进而利用集群中相似用户的偏好来完成预测。

收稿日期:2015-04-10 **修回日期**:2015-05-05

作者简介:罗蜀君(1979—),女,河南洛阳人,工程师,研究方向为信息处理与计算机技术。

国际数据挖掘会议(ICDM)在2006年评选出了数据挖掘领域的十大经典算法,K-均值算法是唯一入选的聚类分析算法。K-均值算法以空间中的k个点为中心,对最靠近它们的样本进行聚类。该算法的最大优势在于简洁和快速。但是因为野点和噪声点的存在,聚类中心的更新很容易发生较大的偏差,进而限制了K-均值算法的鲁棒性,影响了聚类效果。

一般认为,K-均值算法中采用的全局误差度量方法,即均方误差,使得算法的鲁棒性不够好。近年来,研究人员通过熵和粗糙熵的概念来提升算法的鲁棒性[9-11]。最新的研究表明协同熵提供了一种局部的度量方法来描述两组变量的相似性。为了进一步提升K-均值算法的鲁棒性,本文提出使用最大化协同熵的准则来寻找最优的聚类中心。对于每一个聚类,其聚类中心可以由一种迭代重加权的优化算法高效地求解得到。由于野点和噪声点通常具有较大的残差,在聚类中心的更新中,它们将被赋予较小的权重,进而保证了聚类中心更新的准确性和鲁棒性。通过在若干真实数据集上

与传统 K-均值算法的聚类结果的比较，新提出的基于协同熵的 K-均值算法具有明显的优势。

1　K-均值算法

K-均值算法作用在由 d 维向量组成的集合上，$\boldsymbol{D}=\{\boldsymbol{x}_i \mid i=1,\cdots,N\}$，其中，$\boldsymbol{x}_i \in \mathbf{R}^d$ 表示第 i 个样本。算法初始化阶段，选取 $\mathbf{R}^d$ 空间中的 k 个点作为初始的 k 个聚类中心，有很多方法可以用来选择这些初始点。例如从数据库中随机抽样，使用数据库的某个小子集的聚类结果，或者对数据库中所有样本的均值进行 k 次的扰动。在获得初始的聚类中心之后，K-均值算法将迭代以下两个步骤直至收敛。

步骤 1　样本的分配。每个样本点将被分配到距离它最近的聚类中心，如果样本与不同的聚类中心距离相等，则进行随机分配。该步骤将完成对样本的划分。

步骤 2　中心的调整。每个聚类的中心调整为被分配到该聚类的所有样本点的均值。该步骤的公式化描述为

$$c_i = \arg\min_{c} \sum_{x_j \in C_i} \| \boldsymbol{x}_j - c \|^2 \text{。} \tag{1}$$

当样本分配的结果（即 k 个聚类的中心 $\{c_1,\cdots,c_k\}$）不再变动时，K-均值算法达到收敛。注意到每次迭代的复杂度由 $N\times K$ 次距离比较决定，算法收敛所需要的迭代次数依赖于样本的个数 N。图 1 描述了 K-均值算法的执行过程。

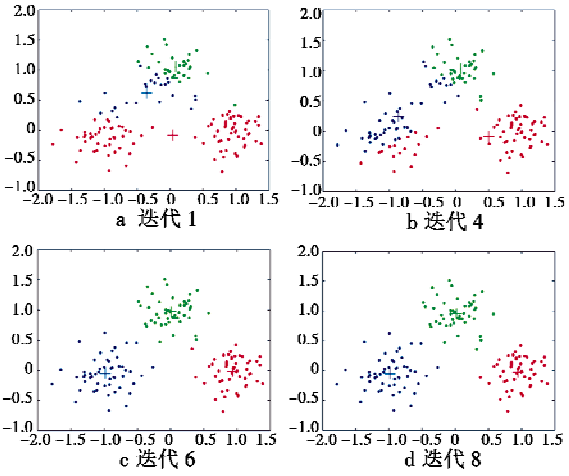

图 1　K-均值算法执行过程中聚类中心（由“+”表示）和样本分配（由颜色表示）的改变

Fig. 1　Changes in cluster centers (indicated by “+”) and sample allocation (indicated by color) during the execution of K-means algorithm

在步骤 2 中，新的聚类中心将由该聚类中所有样本点的均值决定。隐藏在该更新操作背后的一个很重要的假设是平等地看待不同的样本，即它们对于聚类中心的更新有着相同权重的影响。然而大量实验表明，这种更新的方法缺乏鲁棒性，很容易受到野点和噪声的影响。如果某个聚类中存在一个严重偏离真实中心的野点，聚类中心将被尽可能地调整到野点附近，以最大程度地减小野点带来的影响。

2　基于协同熵的 K-均值算法

传统 K-均值算法的步骤 2 通过最小化均方误差来求解最优的聚类中心。由于均方误差是一种全局的度量手段，它极易受到野点和噪声的影响，缺乏鲁棒性。基于协同熵的优良性质，本文提出通过最大化协同熵的方法来求解最优的聚类中心。

协同熵最初由文献[12]提出，并被用来处理非高斯的噪声和脉冲噪声，它与 2 阶 Renyi 熵有着紧密的联系。2 阶 Renyi 熵使用 Parzen 窗法来估计数据的分布。对于两个任意的随机变量 A 和 B，它们的局部相似度度量是

$$V_\sigma(A,B)=E[k_\sigma(A-B)] \tag{2}$$

式中：$k_\sigma(\cdot)$ 是一个核方程；$E[\cdot]$ 是期望操作符。协同熵利用核技巧将输入空间非线性地映射到一个高维的特征空间，与传统的核方法不同，它并不依赖于两两的样本对。以清晰的理论为基础，协同熵具有对称性、正性和有界性等良好性质。

在实际的计算中，A 和 B 的联合概率密度方程往往是未知的，并且仅能获得有限数量的样本 $\{(A_j,B_j)\}_{j=1}^m$。因此，协同熵估计算式为

$$\hat{V}_{m,\sigma}(A,B)=\frac{1}{m}\sum_{j=1}^{m} k_\sigma(A_j-B_j) \tag{3}$$

式中，$k_\sigma(\cdot)$ 是高斯核，其算式为

$$k_\sigma(x)=e^{-x^2/2\sigma^2} \text{。} \tag{4}$$

文献[12]进一步将基于采样的协同熵推广到任意两个离散向量的广义相似度度量，即协同熵诱导矩阵（CIM）。给定任意两个向量 $\boldsymbol{A}_e=(a_1,\cdots,a_m)^{\mathrm{T}}$ 和 $\boldsymbol{B}_e=(b_1,\cdots,b_m)^{\mathrm{T}}$，它们的协同熵诱导矩阵定义为

$$CIM(\boldsymbol{A}_e,\boldsymbol{B}_e)=\left(k_\sigma(0)-\frac{1}{m}\sum_{i=1}^{m} k_\sigma(a_j-b_j)\right)^{1/2} \tag{5}$$

式中，残差 e_j 的定义为 $e_j=a_j-b_j$。对于一个自适应系统，关于残差 e_j 的协同熵

$$\max_{\theta} \frac{1}{m}\sum_{i=1}^{m} k_\sigma(e_j) \tag{6}$$

被称作最大协同熵准则，其中，θ 是模型的参数。从概率上讲，最大协同熵准则意味着最大化在原点处的误差概率密度，和全局的均方误差不同，协同熵是一种局部的度量。图 2 描绘了均方误差和协同熵函数的图像，以及它们对应的导数的图像。从导数的曲线发现，随着残差的增大，残差对均方误差的影响将持续增大，而协同熵则对于残差的变化不再敏感。

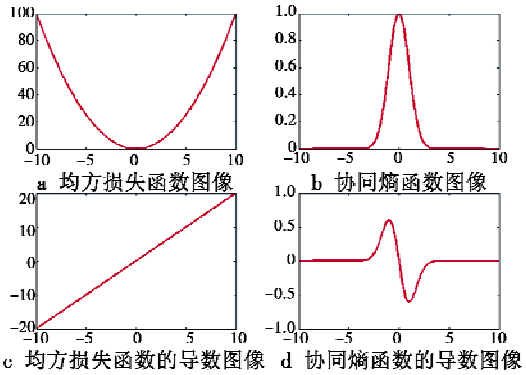

图 2 均方损失和协同熵的比较

Fig. 2 Comparison between squared error loss and co-entropy

对于包含 n_i 个样本的聚类 $C_i = \{x_1, \cdots, x_{n_i}\}$，其中，$x_j$ 指代 C_i 中的第 j 个样本，根据式(6)，可以给出该聚类的最优聚类中心 c 的求解方法为

$$\max_c J(c) = \frac{1}{n_i}\sum_{j=1}^{n_i} k_\sigma(x_j - c) \tag{7}$$

$J(c)$ 关于聚类中心 c 的导数为

$$\frac{\partial J(c)}{\partial c} = \frac{1}{n_i}\sum_{j=1}^{n_i} -\frac{x_j - c}{\sigma^2} k_\sigma(e_j) \tag{8}$$

式中，$e_j = x_j - c$，为第 j 个样本的残差。将 $\frac{\partial J(c)}{\partial c}$ 置为 0，并化简得到 $c = \sum_{j=1}^{n_i} x_j k_\sigma(e_j) \Big/ \sum_{j=1}^{n_i} k_\sigma(e_j)$。由于直到模型完成拟合才能得到真实的残差 $\{e_1^*, \cdots, e_{n_i}^*\}$，并且在残差未知时无法完成上式的计算，单独一个步骤不可能得到最优的聚类中心。因此，使用迭代重加权的方法来完成第 i 个聚类的最优聚类中心的求解，其具体过程如下所述。

1）随机初始化聚类中心 c，并计算每一个样本的残差 $\{e_1, \cdots, e_{n_i}\}$。

2）根据式(4)，计算样本的权重 $\{k_\sigma(e_1), \cdots, k_\sigma(e_{n_i})\}$。

3）更新聚类中心 c。

4）更新所有样本的残差 $\{e_1, \cdots, e_{n_i}\}$。

5）迭代执行 2）~4）直到聚类中心 c 不再变化。

与使用聚类中样本的均值作为新聚类中心的策略相比，通过最大化协同熵，不同样本将对聚类中心有着不同权重的影响。由于野点或者噪声有着较大的残差，在聚类中心的更新中，它们将被赋予较小的权重，从而避免聚类中心的估计朝着野点或者噪声样本偏移。在完成聚类中心的调整后，重复样本分配步骤以进行下一轮的迭代。考虑到协同熵的计算复杂度为 $O(N)$（这里 N 是样本个数）[12]，因此，基于协同熵的 K-均值算法的时间复杂度为 $O(Nkt)$，其中，N 是样本个数，k 是聚类数，t 是迭代次数。

3 实验结果

为了验证基于协同熵的 K-均值算法的有效性，本文选择了 UCI 机器学习数据库[12-13] 中的 Iris，Wine，Segmentation 和 Balance 作为测试数据集。各数据集的统计信息如表 1 所示。

表 1 数据集描述

Table 1 Data sets description

数据集	样本数	属性数	类别数
Iris	150	4	3
Wine	178	13	3
Segmentation	210	19	7
Balance	625	4	3

为评价聚类结果，采用了常用的纯度[13] 来衡量，其定义为

$$p(C,D) = \frac{1}{N}\sum_k \max_j |D_k \cap C_j| \tag{9}$$

式中，$C = \{C_1, \cdots, C_k\}$ 是聚类的集合，C_j 表示第 j 个聚类的集合，与聚类集合 D 的定义类似。

在实验中，分别运行随机选择初始聚类中心的传统 K-均值算法和基于协同熵的 K-均值算法，所得聚类结果由纯度表示，如表 2 所示。

表 2 传统 K-均值算法与改进的 K-均值算法聚类结果的纯度比较

Table 2 Clustering results comparison in purity between conventional K-means algorithm and co-entropy based K-means algorithm

数据集	K-均值算法	改进的 K-均值算法
Iris	0.92	0.96
Wine	0.95	0.98
Segmentation	0.70	0.79
Balance	0.83	0.86

从表 2 中可以看出，对于不同的数据集，使用基于协同熵的 K-均值算法可以获得比传统 K-均值算法更好的聚类结果。例如，在数据集 Segmentation 上，相比传统 K-均值算法，基于协同熵的 K-均值算法可以获得近 13% 的纯度的提升。这主要是由于传统 K-均值算法使用最小化均方误差的原则来更新聚类中心，不能对野点和噪声点进行区别对待。本文提出的基于协同熵的 K-均值算法使用最大化协同熵的准则来更新聚类中心，考虑了不同样本点的残差，降低了野点和噪声点对聚类中心选择的影响，进而保证聚类中心求解的准确性。

协同熵的计算中需要确定参数 σ。为了测试基于协同熵的 K-均值算法对于参数 σ 的敏感性，图 3 呈现了基于协同熵的 K-均值算法在 Segmentation 数据集上针对不同参数 σ 的聚类结果。

从图 3 中可以看出，对于不同的参数 σ，基于协同

熵的 K-均值算法的聚类结果略有波动，但是均优于传统的 K-均值算法的聚类结果。参数 σ 过小使得聚类中心的更新只由很少一部分残差较小的样本点决定；相反，过大的参数 σ 将降低模型辨别野点和噪声点的能力。因此，选择一个合适的参数 σ，将获得更为良好的聚类结果。

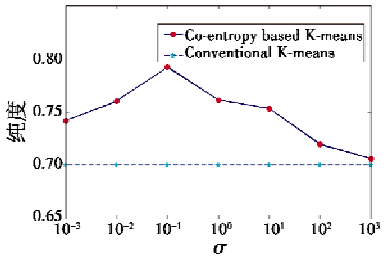

图 3　Segmentation 数据集上基于协同熵的 K-均值算法对参数 σ 的敏感性

Fig. 3 Sensitivity analysis of the co-entropy based K-means algorithm to parameter σ in the Segmentation dataset

4　结束语

K-均值算法是一种广泛应用的聚类分析算法，但是聚类中心更新所采用的最小均方误差原则忽略了野点和噪声点的影响，限制了算法的鲁棒性。本文提出的基于协同熵的 K-均值算法使用协同熵作为局部相似性的度量手段，并利用最大协同熵原则来求解最优的聚类中心。通过一种迭代重加权的优化方法，残差较大的野点和噪声点在聚类中心更新的过程中被赋予较小的权重，进而保证聚类中心调整的准确性和鲁棒性。标准数据集熵的实验结果证明了算法的有效性和可行性。由于本文算法具有很好的鲁棒性，未来考虑将算法应用于医学图像分割，图像和语音的压缩等领域。

参 考 文 献

[1] XU C, TAO D C, XU C. Large-margin multi-viewinformation bottleneck[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(8):1559-1572.

[2] XU C, TAO D C, XU C, et al. Large-margin Weakly supervised dimensionality reduction[C]//Proceedings of the 31st International Conference on Machine Learning, Beijing, 2014:865-873.

[3] XU C, TAO D C, LI Y X, et al. Large-margin multi-view Gaussian process for image classification[C]//Proceedings of the 15th International Conference on Internet Multimedia Computing and Service, ACM, 2013:7-12.

[4] XU C, TAO D C, XU C. A survey on multi-view learning [EB/OL]. [2015-04-10]. http://arxiv.org/abs/1304.5634.

[5] PETORS D, ALAN M F, RAVI K, et al. Clustering large graphs via the singular value decomposition[J]. Machine Learning, 2004, 56(1):9-33.

[6] TAO D P, LIANG L Y, JIN L W, et al. Similar handwritten Chinese character recognition by kernel discriminative locality alignment[J]. Pattern Recognition Letters, 2014, 35(1):186-194.

[7] TAO D P, JIN L W, WANG Y F, et al. Rank preserving discriminant analysis for human behavior recognition on wireless sensor networks[J]. IEEE Transactions on Industrial Informatics, 2014, 10(1):813-823.

[8] TAO D P, JIN L W, YANG Z, et al. Rank preserving sparse learning for kinect based scene classification[J]. IEEE Transactions on Cybernetics, 2013, 43(5):1406-1417.

[9] 桂云苗，朱金福. 一种用信息熵确定聚类权重的方法[J]. 统计与决策，2015(16):29-30. (GUI Y M, ZHU J F. Determining the weights of clustering method using information entropy[J]. Statistics & Decision, 2015(16):29-30.)

[10] 周漩，张凤鸣，惠晓滨，等. 基于信息熵的专家聚类赋权方法[J]. 控制与决策，2011，26(1):153-156. (ZHOU X, ZHANG F M, HUI X B, et al. Method for determining experts' weights based on entropy and cluster analysis[J]. Control and Decision, 2011, 26(1):153-156.)

[11] 徐怡，李龙澍，李学俊. 基于粗糙熵和 K-均值聚类算法的图像分割[J]. 华东理工大学学报：自然科学版，2007，33(2):255-258. (XU Y, LI L S, LI X J. Image segmentation based on rough entropy and K-means clustering algorithm[J]. Journal of East China University of Science and Technology Natural Science Edition, 2007, 33(2):255-258.)

[12] LIU W F, PUSKAL P P, JOSE C P. Correntropy: properties and applications in non-Gaussian signal processing [J]. IEEE Transactions on Signal Processing, 2007, 55(11):5286-5298.

[13] JAIN, ANIL K M, MURTY N M, et al. Data clustering: a review[J]. ACM Computing Surveys, 1999, 31(3):264-323.