

核模糊 C 均值算法的聚类有效性研究^{*}普运伟^{1,2,3} 金炜东¹ 朱明^{1,2} 胡来招²(西南交通大学信息科学与技术学院 成都 610031)¹(电子对抗国防科技重点实验室 成都 610036)² (昆明理工大学计算中心 昆明 650093)³

摘要 针对核模糊 C 均值聚类(Kernelized Fuzzy C-Means, KFCM)算法的有效性评价,以核非线性映射为工具,将原空间中的六个著名有效性指标推广到高维特征空间,得到其对应的核化形式,并通过数值比较实验考察这些核化指标的性能及其对高斯核宽度 β 和模糊指数 m 的敏感特性。结果表明,在所考察的指标中,著名的 Xie-Beni 指标 V_{XB} 及其改进指标 V_K 的核化版本具有最好的性能和可靠性,可优先作为 KFCM 聚类算法的有效性准则。

关键词 核聚类,核模糊 C 均值,聚类有效性,最佳聚类数

On Cluster Validity for Kernelized Fuzzy C-Mean Algorithm

PU Yun-Wei^{1,2,3} JIN Wei-Dong¹ ZHU Ming^{1,2} HU Lai-Zhao²(School of Information Science and Tech., Southwest Jiaotong University, Chengdu 610031)¹ (National Electronic Warfare Laboratory, Chengdu 610036)² (Computer Center, Kunming University of Science & Technology, Kunming 650093)³

Abstract For the sake of evaluating the quality of cluster results obtained by the KFCM (kernelized fuzzy c-means) algorithm, six noted validity indices for standard FCM are generalized into high-dimensional feature space for the purpose of acquiring their corresponding kernelized expressions. And then, the performances and the dependencies of these kernelized indices on Gauss kernel-width β and fuzzy exponent m are examined via some tests of numerical experiments. The results show that, the kernelized versions of the Xie-Beni index V_{XB} and its emendatory index V_K are the most effective and reliable of the indices considered, so can be given the priority as the cluster validity criteria for KFCM algorithm.

Keywords Kernel clustering, Fuzzy c-means, Cluster validity, Optimal number of clusters

1 引言

聚类是模式识别和数据挖掘中广为使用的数据分析手段。将经典的聚类算法推广到核 Hilbert 空间,是无监督学习领域近年来的研究热点之一^[1~6]。然而,和原空间中的聚类算法一样,核化的聚类算法也需要事先给定聚类数目 c ,这直接影响了这些算法的易用程度和实际性能。众所周知,聚类的质量和聚类数目密切相关,过大的 c 值将使聚类结果复杂化并难于解释和分析,而过小的 c 值将引起信息的损失并误导最终的决策^[7]。因此,如何得到最佳聚类数目 c 一直是聚类有效性研究的关键问题^[8]。针对经典的 FCM,人们已提出许多有效性评价指标。这些指标按计算过程中是否使用样本集的结构信息通常可分为两类:一类仅使用隶属度矩阵,如划分系数 V_{FC} ^[9],划分熵 V_{FE} ^[10],D.-W. Kim 的 V_{OS} ^[11]等;另一类除使用隶属度矩阵外,还使用数据集本身,如 Xie-Beni 的 V_{XB} ^[12],Fukuyama-Sugeno 的 V_{FS} ^[13],Kwon 的 V_K ^[14],Rezaee 的 V_{CWB} ^[15],Boudraa 的 V_B ^[16],以及 Kim 的 V_{SV} ^[17]等。当将 FCM 推广到特征空间时,由于样本集被执行了非线性映射,样本间的几何结构也相应发生了改变。因此,尽管上述第一类指标可不在作任何修改的前提下直接使用,而第二类指标却由于计算过程涉及数据集的结构信息,必须作适当的处理才能使用。为此,本文利用核函数与特征空间内积的对应关

系,将上述第二类有效性指标推广到核 Hilbert 空间,并通过比较实验研究所得核化指标的性能和稳健性。结果显示, V_{XB} 及其改进指标 V_K 的核化形式具有较好的性能及稳定性,可配合 KFCM 给出合理的划分。

2 核模糊 C 均值聚类

设原空间样本集为 $X = \{x_1, x_2, \dots, x_N\}$, $x_j \in R^d$, $j = 1, 2, \dots, N$,核非线性映射为 $\Phi: x \rightarrow \Phi(x)$,则 KFCM 聚类的目标函数为^[2, 5, 6]

$$J_m(X; U, V) = \sum_{i=1}^c \sum_{j=1}^N u_{ij}^m \|\Phi(x_j) - \Phi(v_i)\|^2 \\ = \sum_{i=1}^c \sum_{j=1}^N u_{ij}^m [\kappa(x_j, x_j) - 2\kappa(x_j, v_i) + \kappa(v_i, v_i)] \quad (1)$$

式中 v_i 为第 i 类的类中心, $\Phi(v_i)$ 为该中心在相应核空间中的像,且有

$$\Phi(v_i) = \sum_{k=1}^N u_{ik}^m \Phi(x_k) / \sum_{k=1}^N u_{ik}^m \quad (2)$$

为最小化(1)式,需计算 $\kappa(x_j, v_i)$ 和 $\kappa(v_i, v_i)$ 。注意到 $\langle x_i, x_j \rangle = \langle \Phi(x_i), \Phi(x_j) \rangle$,则可得

$$\kappa(x_j, v_i) = \langle \Phi(x_j), \Phi(v_i) \rangle = \sum_{k=1}^N u_{ik}^m K(x_k, x_j) / \sum_{k=1}^N u_{ik}^m \quad (3)$$

$$\kappa(v_i, v_i) = \langle \Phi(v_i), \Phi(v_i) \rangle = \sum_{k=1}^N \sum_{l=1}^N u_{ik}^m u_{il}^m K(x_k, x_l) / (\sum_{k=1}^N u_{ik}^m)^2$$

^{*} 国家自然科学基金(No. 60572143),国家电子对抗技术预研基金(NEWL51435QT220401)。普运伟 博士生,主要研究方向为模式识别,核方法,智能信号处理;金炜东 博士,教授,博导,主要研究方向为智能信息处理,优化理论与优化方法;朱明 博士生,主要研究方向为进化计算,神经网络,电子战信号处理;胡来招 博士,研究员,博导,主要研究方向为电子战信号处理,侦察接收机。

$$u_{ij}^m)^2 \quad (4)$$

因此,若设有效性准则为 $V(c)$,则 KFCM 算法可描述如下:

Step1 设置目标函数精度 ϵ ,模糊指数 m ,最大迭代次数 Tm ,最大聚类数 $cmax$,并令 $c=2$;

Step2 对样本集 x 作归一化处理,并选择核函数及合适的参数,计算核矩阵 K ;

Step3 随机初始化模糊划分矩阵 $U=[u_{ij}]_{c \times n}$;

Step4 for $t=1:Tm$

a) 计算目标函数 $J(t)$,并通过下式更新 U 阵;

$$u_{ij} = \frac{[1/d_k^2(\Phi(x_j), \Phi(v_i))]^{1/(m-1)}}{\sum_{k=1}^c [1/d_k^2(\Phi(x_j), \Phi(v_k))]^{1/(m-1)}}$$

b) 若 $|J(t) - J(t-1)| < \epsilon$ 或 $t \geq Tm$,则转 Step5; 否则, $t \leftarrow t+1$,转 a);

Step5 计算有效性指标值 $V(c)$;

Step6 若 $c < cmax$,则 $c \leftarrow c+1$,转 Step3; 否则,算法结束,并根据 $V(c)$ 取值确定最佳聚类数及其对应的划分。

3 模糊聚类有效性指标的核化表示

如前所述,对于计算中既涉及 U 阵又涉及数据集结构信息的指标,必须作适当处理才能用于 KFCM 算法。这类指标通常采用紧致性来评价类内的内聚程度,采用分离性来评价不同聚类间的隔离程度。该类指标在模糊聚类的有效性研究中最为普遍,数量也最多。本节选择较具代表性的 V_{XB} , V_{FS} , V_K , V_{CWB} , V_B , V_{SV} 六个指标进行推广,得到相应的核化表示。

3.1 Xie-Beni 指标 V_{XB} 的核化

在原空间中,著名的 Xie-Beni 指标 V_{XB} 定义如下^[12]:

$$V_{XB} = \frac{\sum_{i=1}^c \sum_{j=1}^n u_{ij}^m \|x_j - v_i\|^2}{n(\min_{j \neq k} \|v_j - v_k\|^2)} \quad (5)$$

式中, $\frac{1}{n} \sum_{i=1}^c \sum_{j=1}^n u_{ij}^m \|x_j - v_i\|^2$ 用于评价聚类的紧致性, $\min_{j \neq k} \|v_j - v_k\|^2$ 用于评价类间的分离性。由于好的聚类划分应使得样本在各自类中尽可能少分散而类间距离应尽可能大,因此 V_{XB} 为极小值时的划分为最佳划分。将其推广到核空间,显然对应式的核化指标 V_{KXB} 可表示为:

$$V_{KXB} = \frac{\sum_{i=1}^c \sum_{j=1}^n u_{ij}^m \|\Phi(x_j) - \Phi(v_i)\|^2}{n(\min_{j \neq k} \|\Phi(v_j) - \Phi(v_k)\|^2)} \quad (6)$$

观察(1)式,可知 V_{KXB} 的分子即为 KFCM 的目标函数 $J(Tm)$,而分母中的两聚类质心间的距离平方为:

$$\|\Phi(v_j) - \Phi(v_k)\|^2 = \kappa(v_j, v_j) - 2\kappa(v_j, v_k) + \kappa(v_k, v_k) \quad (7)$$

显然, $\kappa(v_j, v_j)$ 和 $\kappa(v_k, v_k)$ 可由(4)式计算,而

$$\kappa(v_j, v_k) = \langle \Phi(v_j), \Phi(v_k) \rangle = \frac{\sum_{i=1}^c \sum_{l=1}^n u_{il}^m u_{ik}^m K(x_l, x_l)}{\sum_{i=1}^c \sum_{l=1}^n u_{il}^m u_{ik}^m} \quad (8)$$

因此,根据(1),(4),(8)式便可得到 V_{KXB} 。文[4]曾提出一个有效性指标 V_{LL} ,对比(6)式不难发现, V_{LL} 和 V_{KXB} 在计算分离性时是一样的,两者的主要区别仅在于紧致性的计算: V_{KXB} 对每点到每个类的中心距离均考虑,并以 u_{ij}^m 加权;而 V_{LL} 仅考虑各点到其所属类的距离,并以 $1/u_{ij}^m$ 加权。

3.2 Fukuyama-Sugeno 指标 V_{FS} 的核化

V_{FS} 也采用类似 V_{XB} 的紧度和分离度评价方法,主要的不同之处在于 V_{FS} 的类间分离性是通过累加每个类到所有类的

中心平均值来实现的。原空间中的 V_{FS} 被定义为^[13]:

$$V_{FS} = \sum_{i=1}^c \sum_{j=1}^n u_{ij}^m \|x_j - v_i\|^2 - \sum_{i=1}^c \sum_{j=1}^n u_{ij}^m \|v_i - \bar{v}\|^2 \quad (9)$$

式中, $\bar{v}$ 为所有类中心的平均值。采用核化 V_{KXB} 同样的方法,可得

$$V_{KFS} = \sum_{i=1}^c \sum_{j=1}^n u_{ij}^m \|\Phi(x_j) - \Phi(v_i)\|^2 - \sum_{i=1}^c \sum_{j=1}^n u_{ij}^m \|\Phi(v_i) - \Phi(\bar{v})\|^2 \quad (10)$$

可见,第一部分和 V_{KXB} 的分子相同,为 $J(Tm)$ 。若将核空间中的所有类的中心平均值表示为:

$$\Phi(\bar{v}) = \frac{1}{c} \sum_{i=1}^c \Phi(v_i) \quad (11)$$

则第 i 类的中心到类平均中心的距离平方为:

$$\begin{aligned} \|\Phi(v_i) - \Phi(\bar{v})\|^2 &= \kappa(v_i, v_i) - 2\langle \Phi(v_i), \Phi(\bar{v}) \rangle + \langle \Phi(\bar{v}), \Phi(\bar{v}) \rangle \\ &= \kappa(v_i, v_i) - \frac{2}{c} \sum_{k=1}^c \kappa(v_i, v_k) \\ &\quad + \frac{1}{c^2} \sum_{l=1}^c \sum_{s=1}^c \kappa(v_l, v_s) \end{aligned} \quad (12)$$

利用(4)和(8)式即可计算上式的结果。

3.3 Kwon 指标 V_K 的核化

V_K 通过在 V_{XB} 的分子部分引入一惩罚项 $(\sum_{i=1}^c \|v_i - \bar{v}\|^2)/c$ 来避免 V_{XB} 在聚类数目趋于样本数目时单调下降的趋势^[14]。由于已得到 V_{KXB} 的各组成部分的计算式(1),(4),(7),(8),同时注意到该惩罚项和 V_{FS} 的第二项具有完全相同的形式,可通过(12)式计算。因此,不难计算 V_K 的核化表示 V_{KK} 。

3.4 Kim 指标 V_{SV} 的核化

V_{SV} 由欠划分项和过划分项组成,定义为^[17]:

$$\begin{aligned} V_{SV} &= V_{under}(V; X) + V_{over}(V) \\ &= \frac{1}{c} \sum_{i=1}^c \sum_{j=1}^n \|x_j - v_i\| / n_i + c / \min_{i \neq j} (\|v_i - v_j\|) \end{aligned} \quad (13)$$

显然,其核化指标 V_{KSV} 可表示为:

$$\begin{aligned} V_{KSV} &= \frac{1}{c} \sum_{i=1}^c \sum_{j=1}^n \|\Phi(x_j) - \Phi(v_i)\| / n_i + c / \min_{i \neq j} (\|\Phi(v_i) - \Phi(v_j)\|) \end{aligned} \quad (14)$$

利用(1)和(7)式可计算其 V_{KSV} 的值。

3.5 Rezaee 指标 V_{CWB} 的核化

从本质上讲, V_{CWB} 也是基于紧致性和分离性测度的指标,但该指标却使用方差来描述类内的分散程度,继而使用类内平均分散函数来描述模式集的平均紧致性。同时,采用一距离函数来描述类间分散特性。原空间中的 V_{CWB} 定义为^[15]:

$$\begin{aligned} V_{CWB} &= \alpha Scat(c) + Dist(c) \\ &= \alpha \frac{\sum_{i=1}^c \sum_{j=1}^n u_{ij} (x_j - v_i)^2}{cn \|\sigma(X)\|} + \frac{D_{max}}{D_{min}} \sum_{k=1}^c (\sum_{z=1}^c \|v_k - v_z\|)^{-1} \end{aligned} \quad (15)$$

式中, $\sigma(X)$ 为模式集的方差, D_{max} 和 D_{min} 分别是类间最大和最小距离, $\alpha = Dist(c_{max})$ 是调整两项取值范围平衡的加权因子。

(15)式在核空间可表示成如下形式:

$$\begin{aligned} V_{KCWB} &= \alpha \frac{\sum_{i=1}^c \sum_{j=1}^n u_{ij} (\Phi(x_j) - \Phi(v_i))^2}{cn \|\sigma(\Phi(X))\|} + \frac{D_{max}}{D_{min}} \sum_{k=1}^c (\sum_{z=1}^c \|\Phi(v_k) - \Phi(v_z)\|)^{-1} \end{aligned} \quad (16)$$

若采用欧式范数,则上述第一项的分子部分就类似 V_{KXB} 的分子;而 D_{max} , D_{min} 以及 $\|\Phi(v_k) - \Phi(v_z)\|$ 可由(7)式开方

后得到。因此,若再计算出 $\|\sigma(\Phi(X))\|$, 则 V_{KCB} 的值便可求得,而

$$\begin{aligned} \|\sigma(\Phi(X))\| &= \frac{1}{n} \sum_{j=1}^n \left[\Phi(x_j) - \frac{1}{n} \sum_{k=1}^n \Phi(x_k) \right]^2 \\ &= \frac{1}{n} \sum_{j=1}^n \left[\kappa(x_j, x_j) - \frac{2}{n} \sum_{k=1}^n \kappa(x_j, x_k) + \frac{1}{n^2} \sum_{l=1}^n \sum_{s=1}^n \kappa(x_l, x_s) \right] \\ &= \frac{1}{n} \sum_{j=1}^n \kappa(x_j, x_j) - \frac{1}{n^2} \sum_{j=1}^n \sum_{k=1}^n \kappa(x_j, x_k) \end{aligned} \quad (17)$$

3.6 Boudraa 指标 V_B 的核化

V_B 同样采用方差来描述类内聚集性,采用 $D_{\max}/D_{\min}$ 来描述数据集的几何性质。在原空间中, V_B 被定义为^[16]:

$$V_B = \alpha V_{w_i}(c) + G(c) = \alpha \frac{\sum_{k=1}^p \sum_{q=1}^c \text{var}_q(k)}{\sum_{q=1}^c \text{var}_q(q)} + \frac{D_{\max}}{D_{\min}} \quad (18)$$

式中, $\text{var}_q(k)$ 表示第 k 类样本的第 q 个特征的方差,而 $\text{var}_q(q)$ 表示整个数据集的第 q 个特征的方差, α 的作用同 V_{CWB} 中的 α ,但取值为 $G(c_{\max})/V_{w_i}(2)$ 。显然 $\sum_{q=1}^p \text{var}_q(q)$, 可通过 (17) 式计算,因此, V_B 的核化主要是计算 $\sum_{q=1}^p \sum_{k=1}^c \text{var}_q(k)$ 。在原空间中

$$\begin{aligned} \sum_{q=1}^p \sum_{k=1}^c \text{var}_q(k) &= \frac{1}{n_1} \sum_{j=1}^{n_1} \|x_j - v_1\|^2 + \frac{1}{n_2} \sum_{k=1}^{n_2} \|x_k - v_2\|^2 + \\ &\quad \dots + \frac{1}{n_c} \sum_{l=1}^{n_c} \|x_l - v_c\|^2 \\ &= \sum_{i=1}^c \left(\frac{1}{n_i} \sum_{j=1, f(x_j)=i}^n \|x_j - v_i\|^2 \right) \end{aligned} \quad (19)$$

其中, n_i 为第 i 类样本的个数, $f(x_j)=i$ 表示样本 x_j 被指派到第 i 类。采用 (2) 式描述的类中心,将 (19) 式推广到核空间有:

$$\begin{aligned} &\sum_{i=1}^c \left(\frac{1}{n_i} \sum_{j=1, f(x_j)=i}^n \|\Phi(x_j) - \Phi(v_i)\|^2 \right) \\ &= \sum_{i=1}^c \left(\frac{1}{n_i} \sum_{j=1, f(x_j)=i}^n [\kappa(x_j, x_j) - 2\kappa(x_j, v_i) + \kappa(v_i, v_i)] \right) \end{aligned} \quad (20)$$

可见,当利用 KFCM 算法对各数据项进行划分后,便可通过 (3), (4) 式计算上式的值。结合 (7), (17), (20) 式便可得到核化的 Boudraa 指标 V_{KB} 。

4 实验结果及讨论

本节通过 4 个数据集验证和比较上述核化指标的有效性 & 稳健性。实验测试的指标包括: $V_{PC}, V_{PE}, V_{CS}, V_{LL}$ 以及本文推广的 $V_{KXB}, V_{KFS}, V_{KK}, V_{KCB}, V_{KB}, V_{KSV}$ 共 10 个指标。 ϵ 设为 $1e-5$, T_m 为 100, 所有实验均采用高斯核。

4.1 Five Gaussian 数据集

该数据集由二维五高斯分布数据组成,常用于检验有效性指标的基本性能和稳定性,数据集信息可参见文[3]。实验中, $c_{\max}$ 取为 10。表 1 为核宽度 $\beta=1.0$, 模糊指数 $m=2$ 时的各有效性指标值。为了考察各指标随 β 的性能变化,实验测试了 $\beta=0.3 \sim 50$ 的范围。图 1 给出了 V_{KXB} 随 c, β 的变化情况。其余各指标具有类似的变化趋势,不再给出。另外,由于 KFCM 算法的性能与模糊指数 m 有关,加之部分有效性指标的定义本身含有 m , 因此实验也测试了各指标对 m 的敏感性(据文[18]选择 1.5~2.5 的范围),其中 β 为 5.0 时的实验结果列于表 2, V_{KXB} 随 c, m 的变化如图 2 所示。

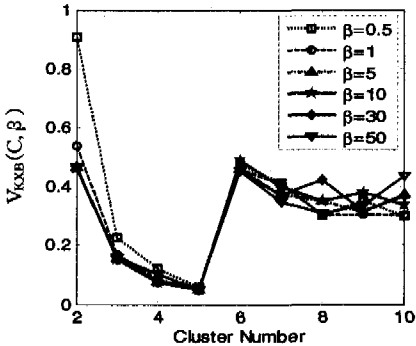


图 1 Five Gaussian 数据集 V_{KXB} 随 c, β 的变化情况, $m=2$

表 1 Five Gaussian 数据集的有效性指标值, $\beta=1.0, m=2.0$

类数 c	V_{PC}	V_{PE}	V_{CS}	V_{KXB}	V_{LL}	V_{KFS}	V_{KK}	V_{KCB}	V_{KB}	V_{KSV}
2	0.62	0.56	1.89	0.54	1.58	12.23	161.37	7.64	10.15	6.84
3	0.64	0.64	0.75	0.17	0.79	-3.60	50.86	7.44	6.36	8.54
4	0.67	0.66	0.36	0.08	0.86	-17.33	25.50	5.61	4.26	10.48
5	0.76	0.53	0.09	0.05	0.15	-23.95	16.64	4.75	3.08	15.09
6	0.68	0.68	0.12	0.46	1.91	-23.01	145.91	16.84	7.59	52.53
7	0.64	0.78	0.11	0.41	2.00	-22.48	129.59	18.51	8.28	61.86
8	0.61	0.87	0.11	0.30	1.69	-21.45	95.77	20.38	8.59	64.76
9	0.58	0.96	0.10	0.31	2.03	-20.55	98.16	27.09	9.54	78.10
10	0.56	0.99	0.09	0.30	2.50	-19.63	99.17	25.99	9.79	92.54

由表 1 和图 1 可知,在 $\beta=0.3 \sim 50$ 范围内, KFCM 均能正确聚类(本实验无错分样本),且除 V_{KSV} 错误地选择聚类数 2 外,所有核化指标包括只利用隶属度值的 3 个指标均正确识别出真实的聚类数 5。包括图 1 在内的 β 依赖关系实验进一步表明,各有效性指标的取值趋势和表 1 相同,对高斯核宽度的变化具有较强的稳定性。同时,表 2 的 m 依赖关系结果表明,不同指标对 m 的敏感程度不尽相同。 $V_{PC}, V_{KXB}, V_{LL}, V_{KK}, V_{KB}$ 五个指标在整个 m 取值范围内始终能正确识别出真实聚类数 5,表现出最好的可靠性。而 $V_{PE}, V_{CS}, V_{KFS}, V_{KCB}$ 则只在部分 m 范围内有效。 V_{KSV} 虽然对 m 的取值不敏感,但却始终给出错误的聚类数目 2。在其它核宽度 β 的测试中,结果类似表 2。

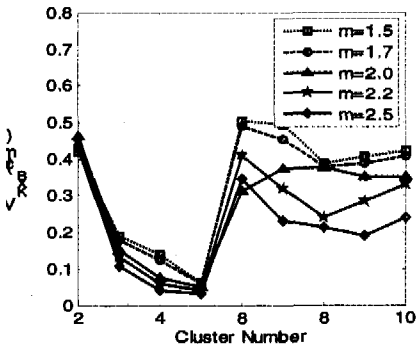


图 2 Five Gaussian 数据集 V_{KXB} 随 c, m 的变化, $\beta=5$

表2 Five Gaussian数据集各有效性指标所确定的最佳聚类数随 m 的变化关系, $\beta=5.0$

m	V_{FC}	V_{PE}	V_{CS}	V_{KXB}	V_{LL}	V_{KFS}	V_{KK}	V_{KCWB}	V_{KB}	V_{KSV}
1.5	5	5	5	5	5	9	5	3	5	2
1.6	5	5	5	5	5	10	5	3	5	2
1.7	5	5	5	5	5	7	5	3	5	2
1.8	5	5	5	5	5	5	5	4	5	2
1.9	5	5	5	5	5	5	5	3	5	2
2.0	5	5	5	5	5	5	5	5	5	2
2.1	5	2	5	5	5	5	5	5	5	2
2.2	5	2	10	5	5	5	5	5	5	2
2.3	5	2	10	5	5	5	5	5	5	2
2.4	5	2	8	5	5	5	5	5	5	2
2.5	5	2	10	5	5	5	5	5	5	2

4.2 Ring数据集

Ring数据集由两类组成,分别为半径为3的环形以及边长为1的正方形均匀分布数据,两类的类中心均为(0,0),数据点数为100。同时,两类数据分别被加上方差为0.1,0.3的高斯白噪声^[2,3]。图3给出了KFCM在 $\beta=0.2, m=2$ 时的聚类结果,此时各有效性指标的取值列于表3。

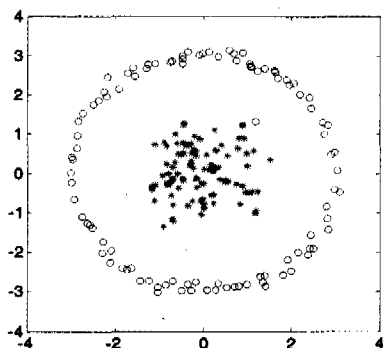


图3 Ring数据集聚类结果, $\beta=0.2, m=2$

从图3可知, KFCM对非球状分布的Ring数据具有较好

表3 Ring数据集的有效性指标值, $\beta=0.2, m=2$

类数 c	V_{FC}	V_{PE}	V_{CS}	V_{KXB}	V_{LL}	V_{KFS}	V_{KK}	V_{KCWB}	V_{KB}	V_{KSV}
2	0.54	0.65	1.50	2.24	10.87	71.07	448.29	0.69e+5	30777	6.62
3	0.38	1.03	0.84	1.43e+6	2.62e+7	44.32	2.83e+8	1.85e+5	58252	7124.3
4	0.29	1.31	0.59	1.22e+6	3.31e+7	32.20	2.45e+8	2.11e+5	50590	3.48e+6

表4 Iris数据集各有效性指标所确定的最佳及次佳聚类数随 β 的变化,括号内为次佳聚类数, $m=2$

β	V_{FC}	V_{PE}	V_{CS}	V_{KXB}	V_{LL}	V_{KFS}	V_{KK}	V_{KCWB}	V_{KB}	V_{KSV}
0.5	2(3)	2(3)	10(9)	2(3)	2(3)	10(9)	2(3)	2(3)	4(3)	2(3)
1.0	2(3)	2(3)	10(9)	2(3)	2(3)	7(5)	2(3)	2(3)	4(8)	2(3)
5.0	2(3)	2(3)	10(9)	2(3)	2(3)	7(8)	2(3)	2(3)	4(3)	2(3)
10	2(3)	2(3)	10(9)	2(3)	2(3)	7(8)	2(3)	2(3)	4(3)	2(3)

总结上述4个实验的结果可知, KFCM算法可以适应多种分布类型的数据集,是解决各种复杂数据集聚类问题的重要途径之一。多数情况下,核化后的有效性指标可以配合KFCM较好地工作,给出较为合理的最佳聚类数目和相应划分,且各指标的性能一般不随核宽度 β 的变化而变化,具有较好的稳定特性。但实验结果也同时表明,各核化指标也类似它们的原空间表示,对模糊指数 m 的依赖程度不尽相同^[11]。在所考察的有效性指标中, V_{KXB} , V_{KK} 显示出了最好的判断能力及稳健性,其次是 V_{LL} 和 V_{FC} ,而 V_{CS} , V_{KFS} 的表现却不尽人

意。实际上,注意到(1)式KFCM算法目标函数最小化样本到质心距离平方和的本质,可以预期建立在同一原则基础上的Xie-Beni指标及其改进形式将有相对较为理想的性能。因此,在实际应用KFCM时,可考虑优先选择 V_{KXB} 和 V_{KK} 指标作为聚类有效性测度。

4.3 Iris数据集

Iris是检验聚类算法和有效性指标性能最为著名的UCI真实数据集,其详细信息参见文[19]。对于该数据集,本文实验表明, KFCM在 $\beta \in [0.5, 50]$ 的较宽范围内均能稳定可靠地完成聚类。各有效性指标所确定的最佳聚类数随 β, m 的变化情况分别见表4和表5。由这些结果可知, V_{FC} , V_{PE} , V_{KXB} , V_{LL} , V_{KK} , V_{KCWB} , V_{KSV} 给出的最佳聚类数均为2,次佳聚类数多为3,且结果随 β, m 的变化相对较为稳定。而 V_{CS} 始终呈单调递减趋势, V_{KFS} , V_{KB} 表现较不稳定。这说明纯粹从聚类的角度看, Iris数据集似乎更应该该分为两类,这基本反映了Iris数据的几何结构和实际分布^[11]。

4.4 Wine数据集

Wine是UCI数据库中另一个著名的数据集。该数据集由3类共178个样本组成,每个样本13个属性^[19]。对于该高维数据集,实验结果显示,在 $m=2$ 且 $\beta \in [1, 50]$ 的范围内, KFCM的错聚样本数均为9,且在 $m=2.0$ 时,仅有 V_{KXB} , V_{KK} , V_{KCWB} , V_{KB} 四个指标给出真实的聚类数3,而 V_{FC} , V_{PE} , V_{LL} , V_{KSV} 得到的最佳聚类数为2, V_{KFS} 和 V_{CS} 则始终呈单调下降趋势。Wine数据集的 m 依赖性的实验结果见表6。由该表可见,除 V_{KCWB} 和 V_{KB} 表现出对 m 的敏感特性之外,其余指标在整个 $m \in [1.5, 2.5]$ 的范围内确定的最佳聚类数较为稳定。

表5 Iris数据集各有效性指标所确定的最佳及次佳聚类数随 m 的变化,括号内为次佳聚类数, $\beta=1.0$

m	V_{FC}	V_{PE}	V_{CS}	V_{KXB}	V_{LL}	V_{KFS}	V_{KK}	V_{KCWB}	V_{KB}	V_{KSV}
1.5	2(3)	2(3)	10(9)	2(3)	2(3)	10(6)	2(3)	2(3)	3(4)	2(3)
1.7	2(3)	2(3)	10(9)	2(3)	2(3)	8(10)	2(3)	2(3)	3(4)	2(3)
1.9	2(3)	2(3)	10(9)	2(3)	2(3)	6(5)	2(3)	2(3)	3(4)	2(3)
2.1	2(3)	2(3)	10(9)	2(3)	2(3)	4(5)	2(3)	2(3)	8(4)	2(3)
2.3	2(3)	2(3)	10(9)	2(3)	2(3)	4(3)	2(3)	2(3)	9(4)	2(3)
2.5	2(3)	2(3)	10(9)	2(3)	2(3)	3(4)	2(3)	2(3)	4(3)	2(3)

优于 HS 算法,但此算法的初始化步也会影响结果,对有些数据点以第一主成分做初始值不能得到理想主曲线。如图 3,对于在正弦上为加正态分布噪声的数据点,(1)是用第一主成分做初值提取的主曲线,(2)换做 s 形折线做初值,显然,后者提取的主曲线才能更好地拟合其本来结构。

对于多边形算法,当处理结构比较复杂的数据类型时,往往第一主成分线不能达到较好的效果,需要调整初始值,使其基本体现数据点的整体结构。

结论 通过对上述两种算法进行分析,我们可以得出这样的结论:局部最优未必得到全局最优。

由于主曲线与主成分有密切联系,是主成分分析的非线性推广,在提取主曲线的算法中往往以第一主成分线做初始值,这在理论上被认为是合理的选择。但是,我们通过实验,比较分析得出不同的结论。对于 HS 算法把初始值换作原点会得到比原算法更好的结果。而多边形算法要针对具体数据点的结构特点适当调整初始值。

参考文献

- 1 Hastie T. Principal curves and surfaces; [PhD thesis]. Stanford University, 1984
- 2 Hastie T, Stuetzle W. Principal curves. Journal of the American Statistical Association, 1989, 84: 502~516
- 3 Zhang Junping, Wang Jue. An Overview of Principal Curves. Chinese Journal of Computers, 2003, 26(2): 129~146

- 4 Kegl B. Principal curves: Learning, design, and applications; [PhD dissertation]. Concordia University, Canada, 1999
- 5 Delicado P. Another look at principal curves and surfaces. Journal of Multivariate Analysis, 2001, 77(1): 84~116
- 6 Duchamp T, Stuetzle W. Geometric properties of principal curves in the plane. Robust Statistics, Data Analysis, and Computer Intensive Methods. New York: Springer Press, 1996, 109: 135~152

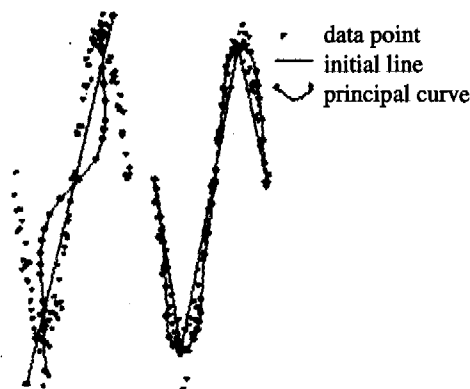


图3 对同一组数据点(灰色点)用多边形算法提取的主曲线(光滑曲线)。(1)中初始值为第一主成分(图中直线), (2)中初始值为图中折线。

(上接第 210 页)

表6 Wine 数据集各有效性指标所确定的最佳聚类数随 m 的变化, $\beta=10$

m	V_{PC}	V_{PE}	V_{OS}	V_{KXB}	V_{LL}	V_{KFS}	V_{KK}	V_{KCB}	V_{KB}	V_{KSV}
1.5	2	2	8	3	2	8	3	2	3	2
1.7	2	2	8	3	2	8	3	2	3	2
1.9	2	2	8	3	2	8	3	3	7	2
2.1	2	2	8	3	2	8	3	3	3	2
2.3	2	2	8	3	2	8	3	3	3	2
2.5	2	2	8	3	2	8	3	2	4	2

结论 本文针对核空间中聚类算法的有效性评价问题,在简要介绍 KFCM 算法的基础上,借助核函数与核空间中内积的计算关系,将六个著名的有效性指标 V_{XB} , V_{FS} , V_K , V_{SV} , V_{CB} , V_B 推广到核 Hilbert 空间,得到其核化表示。同时,通过数值实验测试,比较了这些核化指标随模糊指数 m 、高斯核宽度 β 的依赖关系。结果表明,在所考察的指标中, Xie-Beni 指标及其改进的 Kwon 指标的核化形式具有最好的性能和稳健性,可优先作为 KFCM 聚类算法的有效性指标。

参考文献

- 1 张莉,周伟达,焦李成. 核聚类算法. 计算机学报, 2002, 25(6): 587~590
- 2 伍忠东,高新波,谢维信. 基于核方法的模糊聚类算法. 西安电子科技大学学报(自然科学版), 2004, 31(4): 533~537
- 3 Girolami M. Mercer Kernel-Based Clustering in Feature Space. IEEE Trans. Neural Networks, 2002, 13(3): 780~784
- 4 Li K, Liu Y S. KFCSA: A Novel Clustering Algorithm for High-Dimension Data. FSKD 2005, LNAI 3613, Springer-Verlag, Berlin Heidelberg, 2005. 531~536
- 5 Kim D-W, Lee K Y, Lee D, et al. Evaluation of the Performance of Clustering Algorithms in Kernel-Induced Feature Space. Pattern Recognition, 2005, 38: 607~611

- 6 Chen S C, Zhang D Q. Robust Image Segmentation Using FCM with Spatial Constraints Based on New Kernel-Induced Distance Measure. IEEE Trans. SMC PART B: CYBERNETICS, 2004, 34(4): 1907~1916
- 7 Xu R, Wunsch II D C. Survey of Clustering Algorithms. IEEE Trans. Neural Networks, 2005, 16(3): 645~678
- 8 Halkidi M, Batistakis Y, Vazirgiannis M. Cluster Validity Methods; Part I. Available at: <http://citeseer.ist.psu.edu/534869.html>
- 9 Bezdek J C. Numerical Taxonomy with Fuzzy Sets. J. Math. Biol., 1974, 1: 57~71
- 10 Bezdek J C. Cluster Validity with Fuzzy Sets. J. of Cybernetics, 1974, 3: 58~72
- 11 Kim D-W, Lee K H, Lee D. On Cluster Validity Index for Estimation of the Optimal Number of Fuzzy Clusters. Pattern Recognition, 2004, 37: 2009~2025
- 12 Xie X L, Beni G. A Validity Method for Fuzzy Clustering. IEEE Trans. PAMI, 1991, 13(8): 841~847
- 13 Fukuyama Y, Sugeno M. A New Method of Choosing the Number of Clusters for the Fuzzy C-Means Method. In: Proc. of the fifth Fuzzy Systems Symposium, 1989. 247~250
- 14 Kwon S H. Cluster Validity Index for Fuzzy Clustering. Electronics Letters, 1998, 34(22): 2176~2177
- 15 Rezaee M R, Lelieveldt B P F, Reiber J H C. A New Cluster Validity for the Fuzzy C-Mean. Pattern Recognition Letters, 1998, 19: 237~246
- 16 Boudraa A-O. Dynamic Estimation of Number of Clusters in Data Sets. Electronics Letters, 1999, 35(19): 1606~1608
- 17 Kim D J, Park Y W, Park D J. A Novel Validity Index for Determination of the Optimal Number of Clusters. IEICE Trans. Inform. Syst. D-E, 2001, 84(2): 281~285
- 18 Pal N R, Bezdek J C. On Clustering for the Fuzzy C-Means Model. IEEE Trans. Fuzzy Systems, 1995, 3(3): 370~379
- 19 UCI Repository of Machine Learning Databases[DB/OL]. Available at: <ftp://ftp.ics.uci.edu/pub/machine-learning-database>