基于互包含度的数据分类效果评价研究*)

吴成茂　范九伦
(西安邮电学院信息与控制系　西安710061)

摘　要　针对模糊C-均值聚类算法对初始化分类参数的选择比较敏感而导致分类结果差异性较大的不足，提出了基于互包含度的有效性函数进行数据分类效果好坏的评价。实验结果表明，本文定义的分类效果评价方法是可行的。

关键词　模糊C-均值聚类，互包含度，分类效果

Study on Evaluating Data Classifying Quality Based on Mutual Subsethood

WU Cheng-Mao　FAN Jiu-Lun
(Department of Information and Control, Xi'an Institute of Post and Telecommunications, Xi'an 7100061)

Abstract　Based on the shortage of fuzzy c-means algorithm which initialized classification parameter is sensitivity to data classifying quality, and different initialized classification parameters generate classifying result with bigger otherness. A new evaluating criterion based on mutual subsethood puts forward to assess data classifying quality in this paper. Experimental results show that an evaluating criterion proposed in this paper is feasible.

Keywords　Fuzzy c-means algorithm, Mutual subsethood, Classifying quality

1　引言

模糊C-均值聚类(FCM)[1]算法是一种非监督模式识别方法，在模式识别、图像处理、模糊控制、计算机视觉、数据挖掘等许多科学领域有着极为广泛的应用。尽管FCM已得到广泛应用，但FCM算法的结果一般是局部最优而非全局最优，这就意味着在给定分类数 c 和 m 时，对不同的初始化分类参数(起始聚类中心位置或初始化分类隶属度矩阵)，FCM算法可能会产生不同的划分结果，错误划分会给解决实际问题带来难以预料的后果。如何对这些划分结果进行选择，得到"全局最优解"是一个需要探讨的问题。本文将提出基于互包含度的有效性函数并用于解决上述问题。

2　模糊聚类结果评价准则

模糊聚类问题可表示成下面的数学规划问题

$$\min J_m(U,V)=\sum_{i=1}^{n}\sum_{j=1}^{c}u_{ij}^m d_{ij}^2$$

使得 $\sum_{j=1}^{c}u_{ij}=1, 1\leqslant i\leqslant n; u_{ij}\geqslant 0, 1\leqslant i\leqslant n, 1\leqslant j\leqslant c; \sum_{i=1}^{n}u_{ij}>0, 1\leqslant j\leqslant c$，这里 $X=\{x_1,x_2,\cdots,x_n\}\subset R^s$ 是 s 维欧氏空间中的数据集，n 是数据集中元素的个数，c 是聚类中心数($1<c<n$)，m 是权重系数($m>1$)，$d_{ij}=\|x_i-V_j\|$ 是样本点 x_i 和聚类中心 V_j 的欧氏距离，$V_j\subset R^s(1\leqslant j\leqslant c)$，$u_{ij}$ 是第 i 个样本属于第 j 个类的隶属度，$U=[u_{ij}]$ 是一个 $n\times c$ 矩阵，$V=[V_1,V_2,\cdots,V_c]$ 是一个 $s\times c$ 矩阵。在文[1]中，Bezdek提出了解决上述数学规划问题的模糊C-均值聚类算法。

在应用模糊C-均值聚类算法时，必须给定数据的分类数。为了确定数据集的分类数，文[1,2]提出了选取样本集最佳分类数的许多函数。由于FCM算法对初始化分类参数的选择比较敏感，选取不同的初始化分类参数进行聚类得到的划分(依据最大隶属度原则)结果差异性较大，其中有的划分结果比较接近样本的实际分类情况，有的划分结果与样本的实际分类情况相差甚远。图1给出了分2类的二菱形数据[2]，图2和图3分别是二菱形数据在两种不同初始化聚类中心下得到的数据划分结果，显然图2的划分就较接近二菱形的实际分类情况，图3的划分就与二菱形数据的实际分类情况相差甚远。为了度量划分结果与样本的真实分类情况之间的一致性程度，本文我们采用了互包含度概念，并用来构造效果有效性函数作为聚类算法对选取不同初始化分类参数后样本划分效果好坏的评价标准，以便从不同初始化得到的分类结果中选取最优数据分类划分。

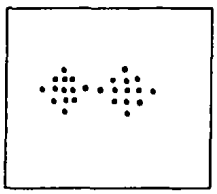

图1　二菱形数据

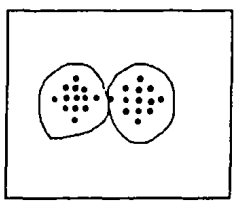

图2　二菱形数据的正确划分图

*)国家自然科学基金项目(批准号：69972041)。

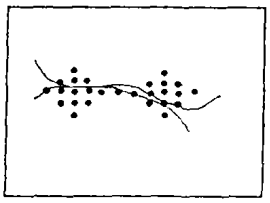

图3 二菱形数据的错误划分

3 基于互包含度的分类效果评价函数

用 X 表示论域，$F(X)$表示 X 上的所有模糊集之集。模糊集 A 在点 $x\in X$ 处的隶属度记作 $A(x)$，文[3]给出如下定义。

定义1 实函数 $c:F(X)\times F(X)\to[0,1]$叫 $F(X)$上的一个包含度，如果 c 满足如下性质：(1)如果 $A\subset B$，则 $c(A,B)=1$；(2)$c(X,0)=0$；(3)如果 $A\subset B\subset C$，则 $c(C,A)\leqslant c(B,A)$，$c(C,A)\leqslant c(C,B)$。

上述定义中的第一条表明如果 A 包含于 B，则 A 是 B 的子集的程度为1；第二条表明隶属度全为1的集合 X 包含于隶属度全为0的集合的程度为零；第三条是对包含度的一个约束。文[4]列举了构造了许多包含度公式，对于一般情况，很显然这些包含度公式的取值互相之间没有必然的联系，但是文[5]发现对于模糊 c-划分 M_{nc}，有些包含度公式之间是等价的。这里，假设论域 $X=\{x_1,x_2,\cdots,x_n\}$，记 $M(A)=\sum_{x\in X}A(x)$，最常用的包含度公式为：

$$c(A,B)=\begin{cases}1 & A=\Phi\\ \dfrac{M(A\cap B)}{M(A)} & A\neq\Phi\end{cases}\qquad(1)$$

文[6,7]讨论基于包含度来构造互包含度公式，其互包含度的定义为：

定义2 基于包含度的互包含度公式为：

$$mc(A,B)=\sqrt{c(A,B)\cdot c(B,A)}\qquad(2)$$

若互包含度定义中的包含度公式选用(1)式，则互包含度公式可以简化为：

$$mc(A,B)=\frac{M(A\cap B)}{\sqrt{M(A)\cdot M(B)}}=\frac{\sum_{i=1}^{n}\min(A(x_i),B(x_i))}{\sqrt{\sum_{i=1}^{n}A(x_i)\cdot\sum_{i=1}^{n}B(x_i)}}\qquad(3)$$

从互包含度公式的定义来看，其数学本质是一种贴近度。另外，文[8]给出许多互包含度定义的新方法[8]。

下面我们给出基于互包含度的数据分类效果好坏的评价，以便从不同初始化得到的分类结果中选取最优数据分类划分。

定义3 给定聚类数 c，基于互包含度公式(3)的数据分类效果评价函数为：

$$A(U)=\sum_{k=j+1}^{c}\sum_{j=0}^{c-1}mc(U_j,U_k)=\sum_{k=j+1}^{c}\sum_{j=1}^{c-1}\left(\sum_{i=1}^{n}\min(u_{ij},u_{ik})\Big/\sqrt{\sum_{i=1}^{n}u_{ij}\cdot\sum_{i=1}^{n}u_{ik}}\right)\qquad(4)$$

在固定权重系数不变的情况下，数据分类效果越好时，$A(U)$的值越小；数据分类效果越差时，$A(U)$的值越大。如果存在 U^* 满足 $A(U^*)=\min_{\Omega_U}A(U)$（$\Omega_U$ 表示数据的模糊 C-划分矩阵的集合），则 U^* 是最佳的数据划分，也即分类结果与样本的实际分类相吻合并使得数据误分率达到最小。

为了说明基于互包含度的数据分类效果评价函数的可行性，记基于包含度公式(1)的数据分类效果评价函数为：

$$A'(U)=\sum_{k=j+1}^{c}\sum_{j=0}^{c-1}c(U_j,U_k)=\sum_{k=j+1}^{c}\sum_{j=1}^{c-1}\left(\sum_{i=1}^{n}\min(u_{ij},u_{ik})\right)\Big/\sum_{i=1}^{n}u_{ij}$$

4 实验结果

在模糊 C-均值聚类算法中，m 的取值范围是一个值得重视的问题。针对已知样本集的分类数 c 的聚类有效性问题而言，常限制 m 的取值在[1.5,2.5]。这一点在文[9]中有较详细的说明，同时，文[10]认为模糊 C-均值聚类算法在实际应用中加权指数 m 取值为2.0最合适，并给出了 $m=2.0$时模糊 C-均值聚类的物理解释。本文仍然选取 m 的三个典型值1.5，2.0和2.5作为实验，其目的是更加充分说明互包含度可以作为分类效果评价准则，而包含度是不宜作为分类效果评价准则的。我们使用2个人造数据和著名的 IRIS 数据进行实验，测试 $A(U)$作为评价数据的分类效果的可行性。在具体实验过程中，我们采用随机初始化聚类中心方法对每个样本数据进行了100次随机初始化聚类中心的实验，然后统计数据划分的误分率、$A'(U)$和 $A(U)$值。由于篇幅所限，表1仅列出各个数据样本进行100次随机初始化的聚类结果中差别比较显著的误分率、$A'(U)$和 $A(U)$值的大小。

二菱形数据 该数据由27个样本构成，数据分为2类。图1给出了该数据的分布图。从图示来看，我们可分该数据为2类[2]。图2是该数据的理想分类效果。若该数据聚类时初始化聚类中心选取不当，可能造成误分率很大的2类，如图3所示。

三类数据 该数据包括49个样本构成，数据分为3类，第1类由6个样本构成，第2类由40个样本构成，第3类由3个样本构成[11]。图4给出该数据的分布图。从图示来看，我们可分该数据为3类。若该数据聚类时初始化聚类中心选取不当，可能造成误分率很大的3类，如图5所示。

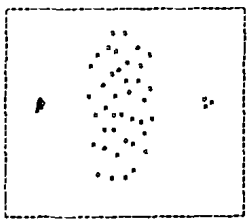

图4 三类数据的分布图

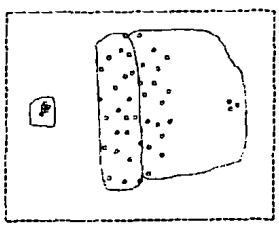

图5 误分率很大的分类结果

IRIS 数据 该数据包括150个4维数据样本，它表示为3个实际类且每类50个数据样本，第1类很好地从别的两个类分开，第2类与第3类之间有交叉。文[14]中指出 Anderson 版本

的Iris数据有错误，而Fisher版本的Iris数据是正确的。本文采用Fisher版本的Iris数据进行实验。若该数据聚类时初始化聚类中心选取不当，可能造成误分率很大的3类。如将第1类错分成2类，实际数据中的第2类和第3类合并分成了一个类。

表1 m 等于1.5所对应数据划分的误分率 E、$A'(U)$和 $A(U)$值

数据样本	误分率 E(%)	$A'(U)$	$A(U)$
二菱形	00%	0.051009	0.050959
	30%	0.993748	0.993748
三类数据	00%	0.087797	0.144019
	20%	0.3723760	0.262518
Iris数据	11%	0.141761	0.167488
	48%	0.402504	0.340861

表2 m 等于2.0所对应数据划分的误分率 E、$A'(U)$和 $A(U)$值

数据样本	误分率 E(%)	$A'(U)$	$A(U)$
二菱形	00%	0.151566	0.151659
	30%	0.998215	0.998215
三类数据	00%	0.992426	0.724163
	27%	0.910916	0.780326
Iris数据	11%	0.476770	0.513306
	48%	1.0990624	0.913319

表3 m 等于2.5所对应数据划分的误分率 E、$A'(U)$和 $A(U)$值

数据样本	误分率 E(%)	$A'(U)$	$A(U)$
二菱形	00%	0.279410	0.279600
	30%	0.998221	0.998221
三类数据	27%	1.310634	1.189948
	82%	1.493402	1.669985
Iris数据	9%	0.857706	0.901683
	11%	0.856864	0.902619

从表1、2和3的实验结果来看，若 $A(U)$值越小，对应样本数据划分的误分率越少，亦即数据分类效果越好，这说明函数 $A(U)$作为评价聚类算法对数据划分结果优劣的标准是合适的；然而，从三类数据 m 等于2.0和IRIS数据 m 等于2.5时所对应数据划分的误分率 E 和函数 $A'(U)$的值来看，$A'(U)$不宜作为评价聚类算法对数据划分结果优劣的标准。

结论 基于目标函数的模糊C-均值聚类算法及其推广形式采用数据的类内紧致性对数据进行划分，但如何评价数据划分的好坏，至今仍是一个待解决的问题。一般认为作为数据划分结果好坏的评价准则应与数据聚类的准则不应相同。为此，本文从模糊C-均值聚类的类间模糊集的互包含程度入手给出了评价数据划分效果好坏的评价标准。实验表明，本文给出的评价数据分类效果方法是可行的。

参考文献

1 Bezdek J C. Pattern Recognition with Fuzzy objective Function algorithms. New york, 1981. 95～107
2 范九伦. 模糊聚类新算法与聚类有效性问题研究. 西安电子科技大学, 1998. 53～54
3 Fan J L, Xie W X, Pei J H. Subsethood measures new definitions. Fuzzy Set and Systems, 1999, 106(1): 201～209
4 范九伦. 模糊熵理论. 西北大学出版社, 1999
5 范九伦, 吴成茂. 用于聚类有效性判定的包含度公式. 模糊系统与数学, 2002, 16(1): 80～86
6 贾克斌. 信息系统设计中聚类分析方法的研究. 北京工业大学学报, 1999, 25(3): 31～36
7 贾克斌, 周俊林. DB设计中聚类分析方法的研究. 计算机应用, 1998, 15(2): 55～58
8 范九伦. 若干新的贴近度公式. 西安邮电学院学报, 2002, 7(3): 69～71
9 PaL N R, Bezdek J C. on cluster validity for the fuzzy C-means model[J]. IEEE Trans. Fuzzy System, 1995, 3(3): 370～379
10 Bezdek J C. A physical interpretation of fuzzy ISODATA. IEEE Trans. Systems, Man and cybernetics, 1976, 6(5): 387～389
11 Bensaid A M, Hall L O, Bezdek J C, et al. Validity-Guided (Re) Clustering with Applications to Image Segmentation. IEEE Trans. Fuzzy system, 1996, 4(2): 112～122
12 Anderson E. The Irises of the Gaspe peninsula. Bull. Amer. IRIS Soc., 1939, 59: 2～5
13 Fisher R A. the use of multiple measurements in taxonomic problems. Ann. Eugen., 1936, 7(2): 179～188
14 Bezdek J C, Keller J M, Krishnapuram R, et al. Will the Real Iris Data Please Stand Up?. IEEE Trans. PAMI, 1999, 7(3): 368～369

(上接第69页)

GSPML不能指明行为规则的语义，该部分的语义由策略翻译程序来解释，策略翻译程序将行为规则的描述转换为有限自动机，称为策略引擎，在组会话期间指导组的安全行为。

结束语 SIMM安全体系结构中独立实现了一个策略服务器PS，作为公共服务设施，为多个多播组提供策略管理服务。SIMM PS的功能模块分为两部分，其中组策略编辑器、解析器、协商器和组策略发布位于一个集成环境中，而组策略翻译、鉴定、执行作为策略引擎嵌入到SIMM安全组服务中。策略服务器以XML文档方式发布组策略，组策略实例发布通过约定的信道(可能是保密的，也可能是广播等非保密的形式)，或是嵌入到SDP[10]协议中。

参考文献

1 McDaniel P, Harney H, Colegrove A, et al. Multicast Security Policy Requirements and Building Blocks. Internet Research Task Force, Secure Mutlicast Research Group (SMuG), Internet Engineering Task Force, November 2000. (dra ft-irtf-smug-polreq-00. txt)(Draft)
2 McDaniel P, Harney H, Dinsmore P, Prakash A. Multicast Security Policy. Internet Engineering Task Force, June 2000, (draft-irtf-smug-mcast-policy-00. txt) (draft)
3 Harney H, Colegrove A, Harder E, et al. Group Secure Association Key Management Protocol. Internet Engineering Task Force, May 2000, draft-harney-sparta-gsakmp-sec-01. txt (Draft)
4 Harney H, McDaniel P, Colgrove A, Dinsmore P. Group Security Policy Token. Internet Research Task Force, September 2001, (draft-ietf-msec-gspt-00. txt)(Draft)
5 Dinsmore B P, Heyman M, Kruus P, Scace C. Dynamic Cryptographic Context Management (DCCM) Report #4: Final Report: [NAI Report #0776]. April 6, 2000
6 McDaniel P, Prakash A. Antigone: Implement Policy in Secure Group Communication. http://www. eecs. umich. edu/～pdmcdan/docs/CSE-TR-426-00. pdf
7 McDaniel P, Prakash A. Ismene: Provisioning and Policy Reconciliation in Secure Group Communication. http://citeseer. nj. nec. com/384963. html, 2000
8 周伟, 尹青, 郭金庚. 多播安全体系结构的研究与实现. 计算机工程与应用, 2002, 5(9)
9 尹青, 周伟, 王清贤. 基于XML的组安全策略描述. 计算机科学, 2003(5)
10 Handley M, Jacobsen V. SDP: Session Description Protocol. RFC 2327, April 1998