文章编号：1001-0920(2009)12-1881-04

基于层次分析法的模糊分类优选模型

李春生[1,2]，王耀南[1]，陈光辉[2]，蒋宏锋[2]
(1. 湖南大学 电气与信息工程学院，长沙 410082；2. 广东商学院 数学与计算科学系，广州 510320)

摘 要：不同的模糊分类算法在同一个数据集合上常会产生不同的模糊分类. 究竟哪种方法最能揭示数据的真实结构，对此，以模糊分类有效性指标为评价指标，应用层次分析法对各模糊分类进行综合评价，建立了一个模糊分类优选模型. 大量实验表明，该优选模型所选出的最优模糊分类，其模式识别率高，能揭示数据的真实结构.

关键词：模糊分类优选模型；模糊分类；模糊分类有效性指标；层次分析法
中图分类号：TP39 **文献标识码**：A

A selection model for optimal fuzzy clustering based on hierarchical analytic process

LI Chun-sheng[1,2], *WANG Yao-nan*[1], *CHEN Guang-hui*[2], *JIANG Hong-feng*[2]
(1. College of Electrical and Informational Engineering, Hu'nan University, Changsha 410082, China; 2. Department of Mathematics and Computational Science, Guangdong University of Business Studies, Guangzhou 510320, China. Correspondent: LI Chun-sheng, E-mail: lcs5812084@sina.com)

Abstract: Different fuzzy clustering algorithms often generate different fuzzy clusterings over the same data set. Which algorithm can best discover the real structure of the data set is a difficult problem. A selection model for the optimal fuzzy clustering is proposed. This model employs hierarchical analytic process to comprehensively evaluate each alternative fuzzy clustering with multiple cluster validity indexes for fuzzy clusterings and select the optimal one from the alternative fuzzy clusterings. Many experiments show that the optimal fuzzy clustering selected from the alternatives is of the highest pattern recognition rate and perfectly can discover the real structure of the data set.

Key words: Selection model for optimal fuzzy clustering; Fuzzy clustering; Cluster validity index; Hierarchical analytic process

1 引 言

在数据的探索性分析中，最令人困惑的是，面对众多的模糊分类算法，使用者不知道选用哪一种，因为不同的模糊分类算法在同一个数据集合上常会产生不同的模糊分类，甚至同一个模糊分类算法，由于其参数值不同，在同一个数据集合上也会产生不同的模糊分类. 更糟糕的是，没有哪一种模糊分类算法能在任何情况下都产生好的模糊分类. 尽管有的学者对常用的一些模糊分类算法做过比较研究，得出了一些选用模糊分类算法的指导性意见，比如，FCM(Fuzzy c-means clustering algorithm)[1]适合于超球型、无噪音的数据；AFCM (Alternative FCM)[2]，PFCM (Possibilistic FCM)[3]与 PCA (Possibilistic clustering algorithm)[4]适合于有噪音的数据；G-K[5]适合于超椭球形、无噪音的数据；FCS(fuzzy c-shell)[6]适合于曲线形数据等. 但由于相关信息匮乏，使用者依然无法确定最适合数据结构特征的模糊分类算法.

解决上述问题的简单方法是，应用不同的模糊分类算法在同一个数据集合上产生多个模糊分类，然后从中选出一个最优的. 该方法的关键是如何准确地评价各个模糊分类. 虽然模糊分类有效性指标很多，但各有偏好，即不同的模糊分类有效性指标对同一个模糊分类的评价结果可能不一样，而且，没有哪一个模糊分类有效性指标总能做出准确的评价. 这就需要应用多个模糊分类有效性指标，对各个模糊分类进行综合评价，从中选出最优者. 常用且有效的综合评价模型当属层次分析法(AHP).

收稿日期：2008-12-23；**修回日期**：2009-04-22.
基金项目：国家 863 计划重点项目(2007AA04Z224，2008AA04Z214)；国家自然科学基金项目(60775047).
作者简介：李春生(1972—)，男，江西余江人，讲师，博士生，从事模式识别的研究；王耀南(1955—)，男，江西吉水人，教授，博士生导师，从事智能控制、图像处理等研究.

本文以模糊分类有效性指标为评价指标，各个模糊分类为备选对象，构造一个 3 层的层次分析模型，应用层次分析法[7]综合评价各个模糊分类，赋予它们以相应的优先级系数，从中选出优先级别最高的模糊分类为最优模糊分类.

2 模糊分类优选模型

假定数据集合 $X=\{x_1,x_2,\cdots,x_N\}$ 上共有 n 个模糊分类，每个模糊分类有 g 个类. 记 $U_k=(u_j^{(k)}(x_i))_{g\times N}$ 是 X 上的第 $k(k=1,2,\cdots,n)$ 个模糊分类，I_j 是第 j 个模糊分类有效性指标($j=1,2,\cdots,m$). 以模糊分类有效性指标为评价指标，以模糊分类为备选对象，构造一个层次分析模型，其结构如图 1 所示.

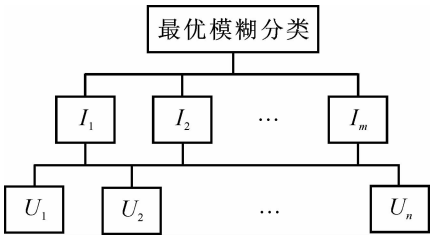

图 1 层次分析法的结构

应用层次分析法[7]，从 n 个模糊分类中选出一个最优的.

对于一个只有 3 层的层次分析，只需两步便可完成：Step1：用各评价指标独立地评价各备选对象，并赋予它们以相应的局部优先等级；Step2：综合所有评价指标全面地评价各备选对象，并赋予它们以相应的全局优先等级，优先等级最高者即为最优者. 下面详述这两步.

Step1 利用各评价指标独立地评价各备选对象，并赋予它们以相应的局部优先等级. 记 $\mathrm{IV}_j(U_k)$ 为模糊分类 U_k 的第 j 个指标 $I_j(j=1,2,\cdots,m)$ 的值，依据指标 I_j 的属性，将各模糊分类的指标值 $\mathrm{IV}_j(U_k)$ 转换为等级系数 $r_j(U_k)$. 若指标 I_j 的值越大，分类效果越好，则将 $\mathrm{IV}_j(U_k)(k=1,2,\cdots,n)$ 按照由大到小的顺序排列，$\mathrm{IV}_j(U_k)$ 对应的序号即为它的等级系数 $r_j(U_k)$. 例如 $(\mathrm{IV}_j(U_1),\mathrm{IV}_j(U_2),\mathrm{IV}_j(U_3))=(1.4501,0.7311,1.1068)$，将这些指标值按照由大到小的顺序排列后，各指标值对应的等级系数为

$$(r_j(U_1),r_j(U_2),r_j(U_3))=(1,3,2).$$

若指标 I_j 的值越小，分类效果越好，则将 $\mathrm{IV}_j(U_k)(k=1,2,\cdots,n)$ 按照由小到大的顺序排列，$\mathrm{IV}_j(U_k)$ 对应的序号就是它的等级系数 $r_j(U_k)$. 此时，上例中的各个模糊分类的等级系数为

$$(r_j(U_1),r_j(U_2),r_j(U_3))=(3,1,2).$$

应用等级系数，构造比较矩阵 $B^{(j)}=(b_{st}^{(j)})_{n\times n}$，其中

$$b_{st}^{(j)}=r_j(U_t)/r_j(U_s). \tag{1}$$

例如，当 $(r_j(U_1),r_j(U_2),r_j(U_3))=(3,1,2)$ 时

$$B^{(j)}=(b_{st}^{(j)})_{n\times n}=\begin{bmatrix}1 & 1/3 & 2/3\\ 3 & 1 & 2\\ 3/2 & 1/2 & 1\end{bmatrix};$$

而当 $(r_j(U_1),r_j(U_2),r_j(U_3))=(1,3,2)$ 时

$$B^{(j)}=(b_{st}^{(j)})_{n\times n}=\begin{bmatrix}1 & 3 & 2\\ 1/3 & 1 & 2/3\\ 1/2 & 3/2 & 1\end{bmatrix}.$$

最后，应用加权最小平方法(WLS)[8]求解各模糊分类的局部优先等级系数 $P_{kj}(k=1,2,\cdots,n)$，即

$$\min\sum_{s=1}^{n}\sum_{t=1}^{n}(P_{sj}-b_{st}^{(j)}P_{ij})^2,$$

$$\text{s.t.}\sum_{k=1}^{n}P_{kj}=1. \tag{2}$$

Step2 利用所有评价指标综合评价各备选对象，选出最优备选对象. 以加权和的方式将各模糊分类的局部优先等级系数 $P_{kj}(k=1,2,\cdots,n)$ 综合成全局优先等级系数 $\mathrm{GP}_k(k=1,2,\cdots,n)$，即

$$\mathrm{GP}_k=\sum_{j=1}^{m}w_jP_{kj}, \tag{3}$$

其中 w_j 是第 j 个指标 I_j 的加权系数，满足 $\forall j\in\{1,2,\cdots,n\}$，$w_j\geqslant 0$，$\sum_{j=1}^{m}w_j=1$.

下面给出基于层次分析法的模糊分类优选模型的虚拟代码，见表 1.

表 1 基于层次分析法的模糊分类优选模型的虚拟代码

Step1：输入备选的模糊分类 $U_k(k=1,2,\cdots,n)$.

Step2：利用各模糊分类有效性指标独自评价各备选模糊分类 $U_k(k=1,2,\cdots,n)$.

For $j=1$ to m

第 j 个模糊分类有效性指标 I_j 应用式(1)和(2)，赋予模糊分类 $U_k(k=1,2,\cdots,n)$ 以相应的局部优先等级系数 $P_{kj}(k=1,2,\cdots,n)$.

End

Step3：用所有评价指标综合评价各备选对象，赋予它们以相应的全局优先等级系数，并从中选出最优备选对象.

由式(3)计算模糊分类 $U_k(k=1,2,\cdots,n)$ 的全局优先等级系数 $\mathrm{GP}_k(k=1,2,\cdots,n)$；

选出最优模糊分类 U_k，其中 $k=\arg\max\limits_{1\leqslant i\leqslant n}\{\mathrm{GP}_i\}$.

3 仿真实验

为了测试本文提出的模糊分类优选模型，这里选用了 4 种模糊分类算法，即 FCM，AFCM，PFCM 与 PCA. 利用它们在同一个数据集合上产生 4 个不同的模糊分类，分别记为 U_{AFCM}，U_{FCM}，U_{PFCM} 与 U_{PCA}. 然后，选用 10 个模糊分类有效性指标，它们分别是 SCG[9]，BS[10]，F_Sil[11]，FS[12]，fpbm[13]，V_t[14]，SVI[15]，V_{gd}[16]，V_w[17]，V_{sc}[18]. 以它们为评价指标，设定

各评价指标的权重系数相等，应用层次分析法对上述 4 个模糊分类进行综合评价，从中选出最优模糊分类.

本文采用的计算协议为：每个模糊分类算法的模糊因子 $q=2$，收敛条件 $\varepsilon=0.001$，最大迭代次数为 100 次. PFCM 专有的参数值设定为 $\eta=1.5$，$a=1$，$b=3$. AFCM 的初始分类中心由分类中心初始化方法 CCIA(Cluster center initialization algorithm)[19] 提供，其余 3 个分类算法的初始中心或初始隶属度矩阵由 AFCM 在数据集合产生的分类结果提供. 在计算 U_{PCA} 的各模糊分类有效性指标值之前，先将各个数据属于各类的可能隶属度按照文献[4]的方式转化为模糊隶属度.

为测试本文提出的模糊分类优选模型的性能，从公用数据库[20,21] 中选择了 12 个真实数据——thyroid，iris，vehicle，satimage，zoo，segment，vowel，splice，sonar，monks_3，german，svmguide3. 有关它们的简要说明见表 2.

表 2　数据简要说明

数据名称	属性数	数据个数	分类数
thyroid[20]	5	215	3
iris[20]	4	150	3
vehicle[21]	18	846	4
satimage[21]	36	4435	6
zoo[20]	9	214	7
segment[21]	19	2310	7
vowel[21]	10	528	11
splice[21]	60	1000	2
sonar[21]	60	208	2
monks_3[20]	6	122	2
german[21]	24	1000	2
svmguide3[21]	22	1243	2

鉴于模式识别率是评价分类效果的标准指标，这里将 4 个模糊分类的模式识别率与层次分析法赋予它们的全局优先等级系数进行对比. 若全局优先等级系数最大的模糊分类，其模式识别率也最大，则本文提出的模糊分类优选模型成功地选出了最优模糊分类，即本文模糊分类优选模型选出的最优模糊分类一定是模式识别率最高的模糊分类. 全部实验结果见表 3. 该表显示，在 12 个数据集合上，本文模糊分类优选模型成功地选出了最优模糊分类，即选出的所有最优模糊分类的模式识别率都是最大的. 有趣的是，在 12 个数据集合上，U_{FCM} 有 6 次当选为最优模糊分类，U_{AFCM} 与 U_{PCA} 各有 3 次当选为最优模糊分类. 这表明 FCM 比其余 3 个模糊分类算法更具普适性. 表 3 同时还显示，4 个模糊分类的全局优先等级系数与它们的模式识别率并非完全一致，即模式识别率高的模糊分类，其全局优先等级系数不一定也大. 例如，在数据 thyroid 上，U_{PCA} 的模式识别率大于 U_{AFCM} 的，但其全局优先等级系数却小于 U_{AFCM} 的. 这一现象表明，本文提出的模糊分类优选模型虽然选出了最优模糊分类，但其对各个模糊分类的评价并非全部正确，其性能还有待于进一步改进.

表 3　模糊分类的全局优先等级系数 GP 与模式识别率 PR

数据名称		模糊分类			
		U_{AFCM}	U_{FCM}	U_{PFCM}	U_{PCA}
thyroid	GP	0.2622	0.4762	0.1120	0.1497
	PR	58.14	80.93	46.51	61.40
iris	GP	0.0717	0.5046	0.1938	0.2298
	PR	52.00	89.33	52.67	52.00
vehicle	GP	0.2507	0.2251	0.1975	0.3266
	PR	40.31	37.12	37.71	41.25
zoo	GP	0.1777	0.4316	0.0717	0.3191
	PR	71.29	79.21	47.52	71.29
vowel	GP	0.3143	0.1685	0.1136	0.4036
	PR	27.27	14.39	17.05	29.17
splice	GP	0.4036	0.1202	0.1820	0.2943
	PR	62.70	62.60	62.40	60.00
sonar	GP	0.5417	0.1368	0.2175	0.1040
	PR	56.73	55.29	53.85	55.77
monks_3	GP	0.4971	0.2859	0.1497	0.0674
	PR	72.95	69.67	69.67	54.92
german	GP	0.2531	0.5939	0.0889	0.0641
	PR	56.00	66.70	53.50	53.60
svmguide3	GP	0.2654	0.2593	0.0835	0.3917
	PR	78.20	57.36	50.36	78.20
segment	GP	0.1895	0.5654	0.0921	0.1529
	PR	43.42	67.49	40.39	45.80
satimage	GP	0.1605	0.3632	0.1658	0.3105
	PR	59.68	70.35	61.42	60.34

4　结　　论

本文提出了一个模糊分类优选模型，该模型以多个模糊分类有效性指标为评价指标，应用传统的层次分析法对同一数据集合上的各个模糊分类进行综合评价，从中选出一个最优的. 通过实验结果，可得出以下结论：

1) 不同的模糊分类算法在同一个数据集合上确实会产生不同的模糊分类，且没有哪一个模糊分

类算法能在任何情况下都产生最好的模糊分类.

2) 本文提出的模糊分类优选模型在不同规模、不同分类数的数据集合上,都能选择模式识别率最高的模糊分类为最优模糊分类,这表明该优选模型是有效的.

3) 统计结果显示,12 个数据集合上,FCM 产生的模糊分类有 6 次当选为最优模糊分类,名列前茅.这表明,FCM 相对于另外 3 个分类算法,更具普适性,这与 FCM 被广泛应用这一事实相吻合.

必须指出的是,本文提出的模糊分类优选模型对 4 个模糊分类的评价结果与模式识别率对 4 个模糊分类的评价结果并非总是一致的.其中原因值得探究,以便进一步改进本文的优选模型.

参考文献(References)

[1] Bezdek J C. Pattern recognition with fuzzy objective function algorithms[M]. New York: Plenum Press, 1981.

[2] Kuo-Lung Wu, Miin-Shen Yang. Alternative *C*-means clustering algorithm[J]. Pattern Recognition, 2002, 35(10): 2267-2278.

[3] Nikhil R Pal, Kuhu Pal, James M Keller, et al. A possibilstic fuzzy *c*-means clustering algorithm [J]. IEEE Trans on Fuzzy Systems, 2005, 13(4): 517-530.

[4] Miin-Shen Yang, Kuo-Lung Wu. Unsupervised possibilistic clustering[J]. Pattern Recognition, 2006, 39(1): 5-21.

[5] Gustafson D, Kessel W. Fuzzy clustering with a fuzzy covariance matrix[C]. Proc of IEEE CDC. San Diego, 1979: 761-766.

[6] Dave R N. Generalized fuzzy *C*-shell clustering and detection of circular and elliptical boundaries[J]. Pattern Recognition, 1992, 25(7): 639-641.

[7] Saaty T L. The analytic hierarchy process: Planning, priority setting, resource allocation[M]. New York: McGraw-Hill, 1980.

[8] Chu A R Kalaba, Springam K. A comparison of two methods for determining the weights of belonging to fuzzy sets [J]. J of Optimization Theory and Applications, 1979, 127(4): 531-541.

[9] Bouguessa M, Wang S R. A new efficient validity index for fuzzy clustering [C]. Proc of 3rd Int Conf on Machine Learning and Cybernetics. Shanghai, 2004: 26-29.

[10] Cho S B, Yoo S H. Fuzzy Bayesian validation for cluster analysis of yeast cell-cycle data[J]. Pattern Recognition, 2006, 39(12): 2405-2414.

[11] Campello R J G B, Hruschka E R. A fuzzy extension of the silhouette width criterion for cluster analysis[J]. Fuzzy Sets and Systems, 2006, 157(21): 2858-2875.

[12] Kwon S H. Cluster validity index for fuzzy clustering [J]. Electronic Letters, 1998, 34(22): 2176-2177.

[13] Malay K Pakhira, Sanghamitra Bandyopadhyay, Ujjwal Maulik. Validity index for crisp and fuzzy clusters[J]. Pattern Recognition, 2004, 37(3): 487-501.

[14] Tang Y G, Sun F C, Sun Z Q. Improved validation index for fuzzy clustering[C]. Proc of the American Control Conf. Oregon, 2005: 1120-1125.

[15] George E Tsekouras, Haralambos Sarimveis. A new approach for measuring the validity of the fuzzy *c*-means algorithm [J]. Advances in Engineering Software, 2004, 35(8/9): 567-575.

[16] Xie Y, Raghavan V V, Dhatric P, et al. A new fuzzy clustering algorithm for optimally finding granular prototypes[J]. Int J of Approximate Reasoning, 2005, 40(1/2): 109-124.

[17] Zhang Yunjie, Wang Weina, Zhang Xiaona, et al. A cluster validity index for fuzzy clustering [J]. Information Science, 2008, 178(4): 1205-1218.

[18] Zahid N, Limouri M, Essaid A. A new cluster-validity for fuzzy clustering[J]. Pattern Recognition, 1999, 32(17): 1089-1097.

[19] Shehroz S Khan, Amir Ahmad. Cluster center initialization algorithm for *K*-means clustering [J]. Pattern Recognition Letters, 2004, 25(11): 1293-1302.

[20] Blake C L, Merz C J. UCI repository of machine learning database[EB/OL]. (1998). http://www.ics.uci.edu/~mlearn/MLRepository.html.

[21] Chih-Chung Chang, Chih-Jen Lin. LIBSVM: A library for support vector machines [EB/OL]. (2001). http://www.csie.ntu.edu.tw/~cjlin/libsvm.