我国全要素生产力改进C－均值聚类

於世为，诸克军

（中国地质大学 研究生院，湖北 武汉 430074）

摘　要：我国每个省、直辖市由于受各种不同因素的影响，因而经济发展水平不一样。采用软计算方法中改进C-均值聚类，对我国31个省、直辖市的经济进行分类，将其分为了5类，并用BP网络对聚类结果进行仿真，检测聚类的效果。

关键词：全要素生产力；C-均值聚类；BP网络

中图分类号：F014.1　　**文献标识码**：A　　**文章编号**：1001-7348(2005)01-0149-03

0　前言

要计算教育对国民经济增长的贡献率，根据教育经济学的基本理论，采用神经—模糊方法研究劳动者受教育程度（即人力资本）到社会生产率(人均国民收入)的映射关系，并对由于劳动者受教育程度提高而引起的国民收入的提高进行软计算。怎么映射呢？具有同等人力资本的劳动者在不同的生产力环境下的经济贡献率是不同的，有时甚至是不可比较的。因此，我们必须将研究的目标系统按生产力的差别进行分类，用软计算的方法进行模糊软分类。然后，在子系统类内，实现人力资本到技术进步以及人力资本、投资、环境资源到经济增长的双重映射。

本文所论述的就是计算教育对国民经济增长的贡献率的第一步。具体是应用软计算方法中的改进C-均值聚类方法来确定整个经济模式的满意分类数，然后利用软分类规则确定每一个类别的样本，最后用BP神经网络来检测分类的效果。

1　经典模糊-C均值划分

经典模糊-C均值划分方法也称为软分类方法，分类的结果是利用"迭代自组织分析技术"(ISODATA)[2]确定出所分类别在各类中的位置，并使各自类别中的样品与其聚类中心的距离尽量最小。

设待分类的样品集为 $U=\{u_1,u_2,\cdots,u_n\}\subset R^n$，$R^n$ 表示实数 n 维向量空间，每个样品有 m 个指标 $X_1,X_2,\cdots,X_m$，样品 $u_i=(x_{i1},x_{i2},\cdots,x_{im})$，其中 $x_{ij}(j=1,2,\cdots,m)$ 是样品 $u_i(i=1,2,\cdots,n)$ 的第 j 个属性值。所谓 U 的一个模糊 c-划分是指：

$$\tilde{F}c=\left\{U_{cm}\in M_{cn}|\mu_{ij}\in[0,1],\forall i,k;\sum_{i=1}^{c}\mu_{ij}=1\ \forall j;0<\sum_{j=1}^{n}\mu_{ij}<n,\forall i\right\}$$

$$(i=1,2,\cdots,c;j=1,2,\cdots,n)\qquad(1)$$

其中，M_{cn} 是 $c\times n$ 阶矩阵的集合，μ_{ij} 表示 u_j 样本属于第 i 类的隶属度。

记 $V^T=(v_1,v_2,\cdots,v_c)(v_i\in R^n,i=1,2,\cdots,c)$ 为聚类中心向量，Bezdek 的 FCM 算法的关键在对于给定的 c，选择隶属度 $\mu_{ij}(i=1,2,\cdots,c;j=1,2,\cdots,n)$ 和 $v_i,i=1,2,\cdots,c$ 使得误差函数：

$$J_m(U,V,c)=\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{h}d_{ji}^{2}=\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{h}\|u_j-v_i\|^2,\quad 1\leqslant h\leqslant\infty\qquad(2)$$

最小化。这里，$d_{ji}^2=\|u_j-v_i\|^2=(u_j-v_i)^T(u_j-v_i)$，且

$$v_i=\frac{\sum_{j=1}^{n}(\mu_{ij})^h u_j}{\sum_{j=1}^{n}(\mu_{ij})^h}\qquad i=1,2,\cdots,c\qquad(3)$$

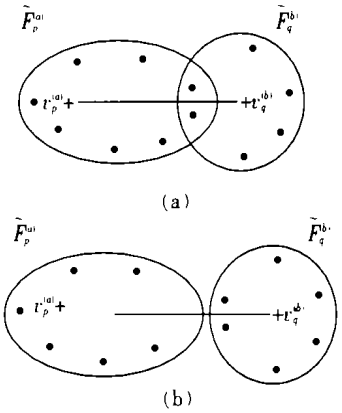
[Bezdek,1995]证明了当 $h>1$，可用(3)式及

$$\mu_{ij}=\left[\sum_{t=1}^{c}\left(\frac{d_{ji}}{d_{jt}}\right)^{\frac{2}{h-1}}\right]^{-1}\qquad(4)$$

进行迭代运算，该运算是收敛的[5]。

2　改进模糊-C划分[1]

对经典FCM，由于其类别数c必须事先给定，且仅仅考虑类别间的聚类中心的距离作为目标函数存在较多缺陷，例如附图所示，很显然两种分类的中心距离 $\|v_p-v_q\|$ 是相等的，但附图(b)的分法明显要比附图(a)的分法要好。



附图

为了将我国31个省、直辖市的技术进步情况进行比较客观合理的划分，我们采用

收稿日期：2004-06-28

基金项目：国家自然科学基金项目(70273044)

作者简介：於世为(1980-)，男，湖北黄冈人，中国地质大学管理科学与工程硕士(在读)，研究方向为信息管理与信息系统；诸克军(1953-)，男，湖北松滋县人，中国地质大学教授，研究方向为神经网络与不确定性的研究。

文献[1]中的改进 FCM 方法来进行聚类。

这种改进的聚类方法是基于类间的亲近度来衡量分类的好坏的。它将根据经典 FCM 分得的集合看成是一个模糊集，并且可以用 Zadeh[7]的表示方法来表示，

$$\tilde{F}_i=\frac{\sum_{j=1}^{n}\mu_{\tilde{F}_i}(x_j)}{x_j}=\frac{\mu_{\tilde{F}_i}(x_1)}{x_1}+\frac{\mu_{\tilde{F}_i}(x_2)}{x_2}+\cdots+\frac{\mu_{\tilde{F}_i}(x_n)}{x_n} \quad (5)$$

对给定的分类矩阵(U,V)，目标评价函数主要是能、通过计算类之间的相似度来获得类之间的亲近度。它既显示了不同类中对同一样本的重迭程度，也考虑了类之间的分离程度。小的亲近度意味着样本的重迭度低且类间的分离程度高，这种状态即是最佳分类。

这种目标评价函数的定义为：

假设$\tilde{F}_p$和$\tilde{F}_q$是属于模糊 C 划分类矩阵 U 的两个类别，$S(\tilde{F}_p,\tilde{F}_q)$是类$\tilde{F}_p$和类$\tilde{F}_q$之间的亲近度，则所谓最佳分类是使用下式值最小时的分类。

$$V_{proposed}(U,V;X)=\frac{2}{c(c-1)}\sum_{p=1}^{c}\sum_{q=p+1}^{c}S(\tilde{F}_p,\tilde{F}_q)=\frac{2}{c(c-1)}\sum_{p=1}^{c}\sum_{q=p+1}^{c}\left[\sum_{\mu}\sum_{j=1}^{n}\delta(x_j,\mu;\tilde{F}_p,\tilde{F}_q)w(j)\right],\quad(p\neq q) \quad (6)$$

式中$S(\tilde{F}_p,\tilde{F}_q)$是指在给定隶属度$\mu$下，两个类别$\tilde{F}_p$和$\tilde{F}_q$之间的总的亲近度函数

$$S(\tilde{F}_p,\tilde{F}_q)=\sum_{\mu}f(\mu;\tilde{F}_p,\tilde{F}_q)=\sum_{\mu}\sum_{j=1}^{n}\delta(x_j,\mu;\tilde{F}_p,\tilde{F}_q)w(j) \quad (7)$$

这里，

$$\delta(x_j,\mu;\tilde{F}_p,\tilde{F}_q)=\begin{cases}1.0 & \text{当}\mu_{\tilde{F}_p}(x_j)\geq\mu\text{且}\geq\mu_{\tilde{F}_q}(x_j)\geq\mu\text{时}\\ 0 & \text{其它}\end{cases} \quad (8)$$

$w(j)$是属于$[0,1]$的模糊数据权重，它由类之间分享数据 x_j 的程度所决定，数据越模糊 $w(j)$越大。

这种改进的 FCM 算法步骤为：其中，$V_{proposed}^{(mun)}$和 $c_{optimal}$ 分别表示 $V_{proposed}$ 最小值和最优聚类类别数。

(1)初始化与 FCM 相关的参数，$c=2$，$c_{optimal}=0$，$V_{proposed}=0$，$V_{proposed}^{(mun)}=\infty$。

(2)在给定的类别数(C)和权重系数 m 下，求 x_j 属于类别$\tilde{F}_i$的隶属度，使得$\sum_{j=1}^{n}\mu_{ij}=1$。

(3)使用(3)和(4)更新聚类中心 v_i 和隶属度向量 $U_i=[\mu_{ij}]$，$(i=1,2,\cdots,c)$。

(4)如果新的 $J_m(U,V)\leq\varepsilon$，转向第 5 步，否则返回第 3 步。

(5)由第 4 步得到的分类矩阵(U,V)，计算类间的亲近度值 $V_{proposed}$。

(6)如果 $V_{proposed}<V_{proposed}^{(min)}$，则 $V_{proposed}^{(mun)}\leftarrow V_{proposed}$，且 $c_{optimal}\leftarrow c$。

(7)如果 $c\leq c_{max}$，则 $c\leftarrow c+1$ 并且返回第 2 步，否则停止计算，将 $V_{proposed}^{(mun)}$和 $c_{optimal}$ 作为结果。

3 我国各地区生产力水平的改进 FCM 聚类

3.1 样本和指标的选取

若实际产出为，有个投入要素，则生产函数的一般形式为[5]：

$$Y=F(x_1,x_2,\cdots,x_n;t) \quad (9)$$

它代表产出与投入要素之间的某种依存关系。为了实现对我国 31 个省、市的生产力水平进行软分类，我们推广著名的 C—D 函数(柯布—道格拉斯)为如下形式：

$$Y=F(x_1,x_2,\cdots,x_n;t)=AK_i^{\alpha}L_i^{\beta}N_i^{\gamma} \quad (10)$$

式中 α、β、γ 是常数，并假设 $\alpha+\beta+\gamma=1$(即产出的规模效益不变)，这里，K，L，N 分别代表资本投入、劳动投入、土地(包括环境资源)投入。从(9)式容易得到：

$$A=YK_i^{-\alpha}L_i^{-\beta}N_i^{-\gamma} \quad (11)$$

(11)式表示一个地区的技术进步与该地区的产出、资本投入，劳动投入，土地(包括环境资源)投入具有一定的依存关系。正是基于这种依存关系对我国 31 个省、市的技术进步进行软划分。

全国各地区 2001 年的指标数据如表 1：GDP 单位(亿元/万人)，固定资产单位(亿元/万人)，单位面积交通(km/万 km²)，就业人数单位(万人)。

3.2 计算及结果

对 c_{max} 的选取，目前还没有统一的方法，我们采用文献[4]和[5]中的方法 $c_{max}\leq\sqrt{n}$，可见，在本文中 $c_{max}=\sqrt{31}\approx5$，$h=2$，$\varepsilon=10^{-6}$，给定的为$\mu=0.3$。初始$[\mu_{ij}]_{c\times n}$选为 5 类时的硬划分隶属度矩阵。

表 1 原始数据表

地区	GDP	固定资产	单位交通	就业人口
北京	2.0576	1.09423	8 955.13	629.5
湖北	0.7803	0.24879	5 122.40	2 452.5
天津	1.8328	0.70219	9 541.97	410.5
湖南	0.6039	0.17803	3 744.90	3 438.8
河北	0.8326	0.28549	3 583.43	3 379.6
广东	1.3681	0.4477	5 684.21	3 962.9
山西	0.5440	0.20281	3 861.89	1 412.9
广西	0.4660	0.13693	2 663.92	2 543.4
内蒙古	0.6503	0.21188	656.25	1 013.3
海南	0.6859	0.26799	6 279.48	339.7
辽宁	1.2001	0.33886	3 509.17	1 833.4
重庆	0.5650	0.22507	4 098.85	1 624.0
吉林	0.7553	0.26076	2 406.38	1 057.2
四川	0.5118	0.18721	2 091.47	4 414.6
黑龙江	0.9344	0.25284	1 620.16	1 631.0
贵州	0.2856	0.14109	2 179.82	2 068.2
上海	3.0674	1.24203	13 524.41	692.4
云南	0.4840	0.17221	4 256.94	2 322.5
江苏	1.2933	0.38385	8 154.55	3 565.4
西藏	0.5275	0.31658	2 842.96	124.6
浙江	1.4629	0.61436	5 465.01	2 772.0
陕西	0.5040	0.21138	2 391.98	1 784.6
安徽	0.5199	0.14118	5 369.67	3 389.7
甘肃	0.4165	0.17878	1 296.35	1 187.2
福建	1.2365	0.34096	4 728.01	1 677.8
青海	0.5754	0.37543	346.60	240.3
江西	0.5198	0.15094	4 090.25	1 933.1
宁夏	0.5300	0.3394	1 820.42	278.0
山东	1.0439	0.30845	4 865.05	4 671.6
新疆	0.7918	0.37633	718.64	685.4
河南	0.5903	0.1616	4 444.43	5 516.6

资料来源：中国统计年鉴 2002。

$$w(j)=\begin{cases}0.1,\mu_{\tilde{F}}(x_j)\geq0.8\\ 0.4,0.7\leq\mu_{\tilde{F}}(x_j)\leq0.8\\ 0.7,0.6\leq\mu_{\tilde{F}}(x_j)\leq0.7\\ 1.0,0\leq\mu_{\tilde{F}}(x_j)\leq0.6\end{cases} \quad (12)$$

运用上述改进 FCM 算法可得，当 $c=2$，3，4，5 时，$V_{proposed}$的值分别为 92.1，102.9，121.10，88.5。即当 $c=5$ 时，可得到最佳聚类结果，其分类隶属度矩阵如表 2。

根据最大隶属度原则，其分类结果如表 3所示：

4 BP 网络测试

为了检验分类结果是否正确，用神经网络 BP 网络来进行测试。我们将全国 31 个省市样本分为训练集和仿真集，如表 4。

所设计的 BP 网络结构为：输入层为 4 个神经元，8 个隐层，输出层 5 个神经元，传递函数都为 logsig，训练函数为 trainlm，训练目标误差为 10^{-6}。将样本的 4 个指标原始数据经 Matlab 中的 prestd()函数归一化后的值，作为训练集和仿真集的输入向量，属于

表2 最佳分类隶属度矩阵

地区	一	二	三	四	五
北京	0.8622	0.0256	0.0478	0.0334	0.031
湖北	0.0134	0.2222	0.1531	0.0932	0.5181
天津	0.4615	0.092	0.1842	0.1327	0.1296
湖南	0.0032	0.8808	0.035	0.0182	0.0628
河北	0.0076	0.7144	0.1219	0.0396	0.1166
广东	0.0086	0.0688	0.8771	0.0164	0.0291
山西	0.0042	0.0302	0.0201	0.1086	0.8370
广西	0.0091	0.2212	0.0599	0.1059	0.6038
内蒙古	0.0079	0.0403	0.0273	0.7470	0.1775
海南	0.0425	0.1029	0.1122	0.3702	0.3722
辽宁	0.0251	0.1404	0.1791	0.2497	0.4057
重庆	0.0024	0.0199	0.0131	0.0474	0.9172
吉林	0.0058	0.0288	0.0226	0.7246	0.2182
四川	0.012	0.7367	0.0956	0.0508	0.105
黑龙江	0.0128	0.0887	0.063	0.4080	0.4275
贵州	0.0097	0.1253	0.0485	0.1675	0.6490
上海	0.8290	0.0353	0.0587	0.0387	0.0383
云南	0.006	0.1135	0.0435	0.0630	0.7739
江苏	0.0205	0.104	0.7817	0.0327	0.0611
西藏	0.0115	0.0323	0.0288	0.7927	0.1346
浙江	0.0483	0.1162	0.6479	0.0790	0.1086
陕西	0.0046	0.047	0.0239	0.1344	0.7900
安徽	0.0102	0.6677	0.1078	0.0482	0.1661
甘肃	0.0096	0.0596	0.0353	0.5556	0.3399
福建	0.0326	0.1379	0.2244	0.2264	0.3786
青海	0.0117	0.0335	0.0283	0.8236	0.1028
江西	0.0026	0.0316	0.0156	0.0369	0.9133
宁夏	0.0052	0.0157	0.0134	0.9033	0.0624
山东	0.0187	0.4995	0.346	0.0458	0.0900
新疆	0.0076	0.0251	0.0218	0.8611	0.0843
河南	0.0268	0.5723	0.2045	0.0688	0.1276

第一类的目标值为(1 0 0 0 0),第二类为(0 1 0 0 0),依次类推,第五类推的目标值为(0 0 0 0 1)。在matlab6.5中,先用表4中的

表3 各省市(自治区)生产力水平分类表

类别	地区名称
一类	北京、天津、上海
二类	河北、山东、安徽、河南、湖南、四川
三类	江苏、浙江、广东
四类	内蒙古、吉林、西藏、甘肃、青海、宁夏、新疆
五类	山西、辽宁、黑龙江、福建、江西、湖北、广西、海南、重庆、贵州、云南、陕西

表4 训练和仿真集样本集

训练集	一类	北京、天津
	二类	河北、安徽、湖南、山东
	三类	江苏、广东
	四类	内蒙古、吉林、西藏、青海、新疆
	五类	山西、黑龙江、福建、湖北、海南、重庆、贵州、云南
仿真集	一类	上海
	二类	河南、四川
	三类	浙江
	四类	甘肃、宁夏
	五类	辽宁、江西、广西、陕西

训练集来训练网络，待网络训练完成后，用仿真集来进行仿真，以检验分类的结果是否正确。

我们进行了10次运算，其中有8次是与分类结果完全相同的，另外两次也只是个别样本不同，可见用改进的FCM对我国的生产力进行划分是比较正确的，实际上也是合理的。

参考文献：

[1]Dae-Won Kim,Kwang H.Lee,Doheon Lee.Fuzzy cluster validation index based on inter-cluster proximity[J].Patter Recognition Letters,24(2003):2561-2574.

[2]J.Bezdek.Pattern Recognition with Fuzzy Objective Function Algorithms[M].Plenum Press,New York,1981.

[3]Bezdk J C, Hath away R. Local convergence of the fuzzy c-means a births[J]. Pattern recognition 1986,19(6).

[4]Pal,N.R.,Bezdek,J.C.,On cluster validity for the fuzzy c-means model[J].IEEE Trans.Fuzzy Syst3(3),370-379,1995.

[5]Xie,X.L.,Beni,G.,A validity measure for fuzzy clustering[J].IEEE Trans.Pattern Anal.Mach.Intelligence,13(8),841-847,1991.

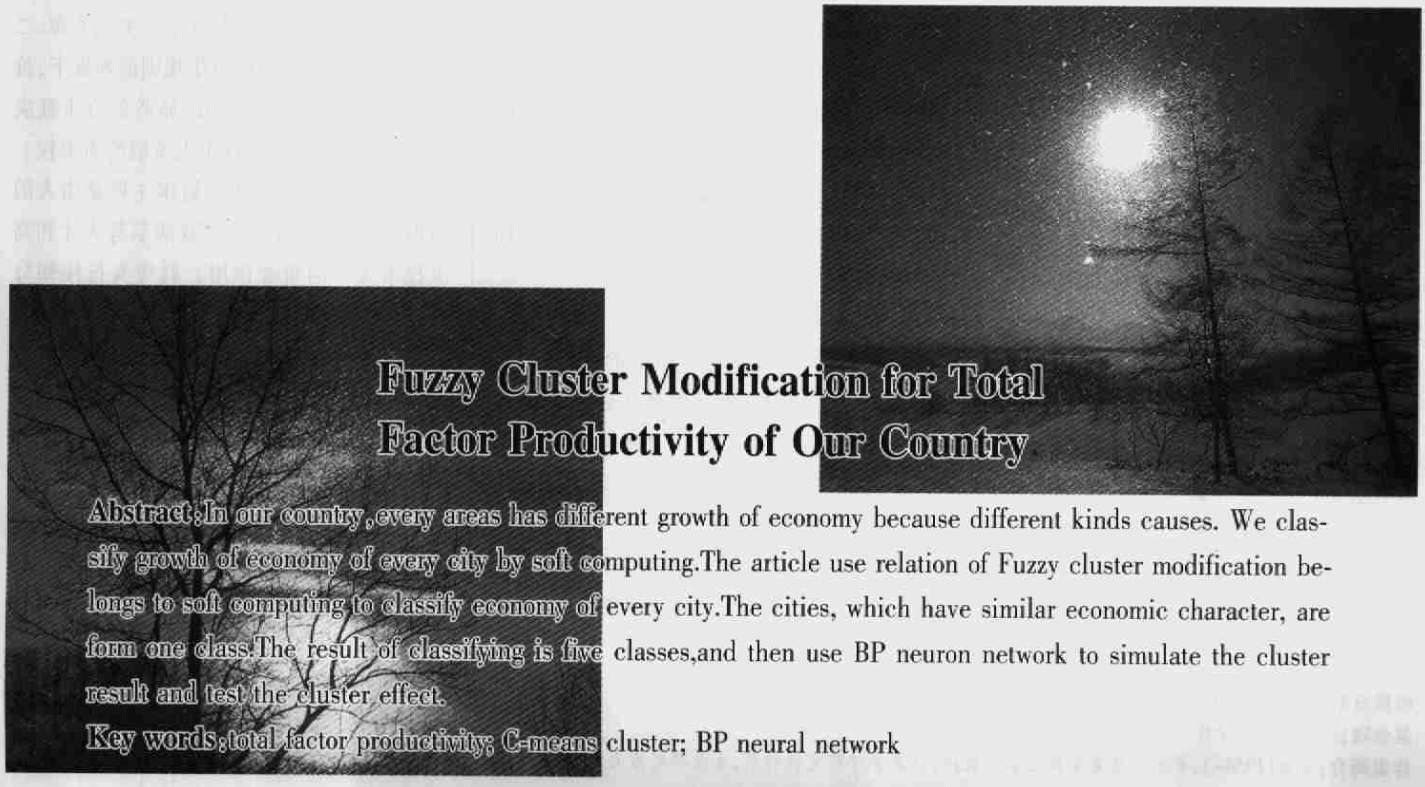
[6]L.O.Hall, B. Ozyurt, J.C. Bezdek. Clustering with a Genetically Optimized Approach[J].IEEE Trans. on Evolutionary Computation, Vol.3,No.2,1999. 103-112

[7]Zadeh L A.Fuzzy sets[J].Inform.and control, 1965,(8),338-353.

[8]郭嗣综,陈刚.信息科学中的软计算方法[M].沈阳:东北大学出版社,2001.68-72.

[9]许东,吴锋.基于MATLAB6.X的系统分析与设计[M].西安:西安电子科技大学出版社,2002.13-18.

(责任编辑:高建平)



Fuzzy Cluster Modification for Total Factor Productivity of Our Country

Abstract: In our country, every areas has different growth of economy because different kinds causes. We classify growth of economy of every city by soft computing.The article use relation of Fuzzy cluster modification belongs to soft computing to classify economy of every city.The cities, which have similar economic character, are form one class.The result of classifying is five classes,and then use BP neuron network to simulate the cluster result and test the cluster effect.

Key words: total factor productivity; C-means cluster; BP neural network