分级聚类算法及其在春季北半球环流形势客观分型和天气预报中的应用

朱盛明 姜翠宏
(江苏省气象科学研究所，南京 210008)

提 要

本文针对春季北半球环流形势宏观分型的物理问题性质，应用适合于类别样本呈超球形分布的两种聚类算法构成分级聚类算法，即先改进半模糊聚类方法，对 1979—1988 年 3—5 月北半球 500hPa 环流图作第一步分类，共得21类；再把 21 类的平均图作为种子样本，应用于自行设计适用于环流形势聚类的变 K-均值聚类方法，最终得 9 类。另外，我们根据预报员经验从原始资料集中挑选出 21 个种子样本，再用变 K-均值聚类方法重新分类，共得 11 类，比较两种聚类算法所得结果，分级聚类算法优越性主要有以下几点：

(1) 由于分级聚类法应用半模糊聚类方法选择种子样本用于变 K-均值方法，所以算法总迭代次数并不增加，而且也减少了用经验方法确定种子样本不当从而影响聚类结果。换言之，分级聚类结果更具客观性。另外，分级聚类算法第二步所采用的变 K-均值方法也克服了原 K-均值方法受初始类别数目影响的缺点。

(2) 分级聚类法所得的各类样本集中，平均天气图形势特征更清楚，异类区别也较明显，更具气象学意义和符合预报员经验。

(3) 分级聚类所得各类型分别与春季长江中下游连晴、过程性、连阴雨的天气有密切的关系，能作为这三类春季主要天气发生发展的背景场。

(4) 用两种聚类结果的分类样本所对应的高度场及其波谱分析资料中选取预报因子，分类建立判别方程，结果得出，无论在入选因子的置信水平还是业务预报准确率验证，都证明了分级聚类算法的优越性。

关键词：分级聚类算法；环流分型；天气预报。

一、引 言

由于聚类算法可以使主观定性的相似概念转化为具有客观性、可重复性和一致性的类别特征，它把大气环流形势划分为大型天气过程如连阴雨、连晴天气、强冷空气持续爆发南下、连续暴雨过程等的发生、发展、衰亡提供了背景场，并进而用分类建预报方程以提高准确率，在防灾、减灾中是一种有效的方法。在气象学中，过去所用的大部分聚类方法，包括后来引入模糊集等价关系再进行聚类的算法，都是单链算法，由于它是根据单个样本

1993 年 8 月 16 日收到，8 月 23 日收到修改稿。

间的亲疏远近构成谱系图来划分类别，因此并不适合于环流图类别样本呈超球形分布的情况，聚类结果的同类环流图象不太集中，不能很好地符合环流图分型的要求.

逐步聚类算法(或称动态聚类法)虽然改变了样本与样本间连系呈链状结构，而是根据样本和分类的中心亲疏远近进行分类，但聚类效果往往受预先选定的初始类别数目和种子样本(或称凝聚点)的影响，如 K-均值算法；也有的收敛性得不到保证，如 ISODATA 算法.

另一方面，我们的物理问题具有两面性. 环流图象的相似，特别是大范围（如半球范围)的图象相似具有模糊性质，某一张环流图不是确定地属于哪一类，而是在不同程度上隶属于所有类；但是，作为预报问题时，又希望根据高度场和槽脊位置求得一个可靠的背景场，以便建立预报工具，做出较正确的天气预报.

为此，本文用两种适应类别样本分布呈超球形的聚类方法作分级聚类，先改进半模糊聚类算法，使分类性能受初始的种子样本影响不大，用它来作第一步分类；然后，再把前面的分类结果用作种子样本，用以自行设计的适用于天气图分类的变 K-均值聚类算法，把所得分类结果与用单一的逐步聚类方法比较，总迭代次数不增加，但分类精度提高了.

二、半模糊聚类算法

半模糊聚类方法把每个样本的模糊性限制在所有种子样本的一个子集内. 在这样的数学模型里，种子样本的选取会对聚类结果有一定的影响，特别是对于环流图的种子样本，某一个本属 i 类的样本 S，在下面列出的迭代过程中，如果一开始就因离开种子样本较远而得到隶属函数 $\mu_{is}=0$，那末在以后的迭代中也不易增大. 为避免这一弊端，开始时可假定每个样本最多隶属 Q 类，再在迭代过程中给定减类阈值而逐步减少类别，再在给定的准则下结束聚类. 这样就可使分类结果受种子样本的影响不大. 具体算法步骤如下：

(1) 给定 Q，$2\leqslant Q\leqslant N$，N 为总样本数，并给定种子样本向量集 $H^{(0)}=\{h_i^{(0)}\}$，$1\leqslant i\leqslant Q$；

(2) 对环流图分类，我们对所有样本给定统一的减类阈值 IH 和 TH 每次减小的步长 ST，以及终止迭代的控制值 TOL；迭代次数记数 $IT=0$；

(3) $IT=IT+1$；

(4) 按下式[1]把 IT 步的隶属函数 $\mu^{(IT)}$ 更新为 μ'

$$\mu_{is}=\frac{1}{\sum_{q=1}^{Q}\left(\frac{d_{is}}{d_{qs}}\right)^{2}\Big/(m-1)} \qquad 1\leqslant S\leqslant N \qquad (1)$$

式中 $d_{is}=\|h_s-h_i\|$ 是第 S 个样本向量 h_s 与当步第 i 个类中心向量 h_i 差的模，m 是大于 1 的实数，由试验给定；

(5) 对样本 h_s，$1\leqslant S\leqslant N$；类中心 h_i，$1\leqslant i\leqslant Q$；选择隶属函数集合 $\{\mu'_{is}\}$ 中 Q 个较大的值，设为 $\mu'_{s1},\mu'_{s2},\cdots,\mu'_{sQ}$，并求 $\gamma'_s=\min\limits_{1\leqslant q\leqslant Q}\mu'_{sq}$，对于 $\{\mu'_{is}\}$ 中隶属函数

小于 γ_S 的值均置 0．并规格化求得第 $IT+1$ 步的隶属函数

$$\mu_{iS}^{(IT+1)}=\frac{\mu_{iS}'}{\sum_{q=1}^{Q}\mu_{qS}'} \tag{2}$$

（6）按下式[1]更新 Q 个类中心

$$h_i^{(IT+1)}=\frac{\sum_{S=1}^{N}(\mu_{iS})^m h_S}{\sum_{S=1}^{N}(\mu_{iS})^m} \qquad 1\leqslant i\leqslant Q \tag{3}$$

（7）若 $\max\limits_{\substack{1\leqslant i\leqslant Q\\ 1\leqslant g\leqslant G}}\{|h_{ig}^{(IT+1)}-h_{ig}^{(IT)}|\}>TH$，$G$ 为样本向量维数，则转第 3 步；否则转第 8 步；

（8）若 $\max\limits_{\substack{1\leqslant i\leqslant Q\\ 1\leqslant g\leqslant G}}\{|h_{ig}^{(IT+1)}-h_{ig}^{(IT)}|\}<TOL$，或 $Q=1$，则输出分类结果后停机；否则 $TH=TH-ST$，$Q=Q-1$，转第 3 步．

三、变 K-均值聚类算法

设有第 i 和第 S 两张环流图，定义相似指数

$$I_{iS}=\gamma_{iS}\left(1-\frac{E_{iS}}{n\sigma}\right) \tag{4}$$

式中

$$E_{iS}=\sqrt{\frac{1}{G}\sum_{g=1}^{G}[h_{ig}-h_{Sg}]^2} \tag{5}$$

h 为高度值，g 为图面格点序号，G 为格点数，即样本向量 h_i 和 h_S 维数．(5) 式是两张环流图的欧氏距离，表示高度场总的数值相似，而(4)式中

$$\gamma_{iS}=\sum_{g=1}^{G}\Delta h_{ig}\Delta h_{Sg}\Big/\sum_{g=1}^{G}\Delta h_{ig}^2\sum_{g=1}^{G}\Delta h_{Sg}^2 \tag{6}$$

Δh_{ig}、Δh_{Sg} 分别代表 i 日和 S 日上 G 个格点高度图面平均高度的偏差，γ_{iS} 称为两张环流图的相似系数．(6) 式直接与两张图面对应点的偏差值有关，反映图形的槽脊位置相似较好．相似指数把相似系数和欧氏距离结合起来就可以同时考虑两张天气图的高度场数值相似和槽脊位置的相似性．I 指数越大，两张图越相似．(4)式中

$$\sigma=\frac{1}{2}(\sigma_i+\sigma_S) \tag{7}$$

为两张天气图面高度场均方差的平均值，其中

$$\sigma_i=\sqrt{\frac{1}{G-1}\sum_{g=1}^{G}\Delta h_{ig}^2},\quad \sigma_S=\sqrt{\frac{1}{G-1}\sum_{g=1}^{G}\Delta h_{Sg}^2}$$

而 n 是正实数，用以消除欧氏距离的量纲，当调节 n 的大小使 E 的作用和 γ 相当，n 太小，

$\left(1-\frac{E}{n\sigma}\right)$ 会出现负值，I 就会反义，如果 n 取得过大，则 E 不起作用，I 值趋向 γ 值.

在相似指数的意义下，第 i 和第 S 张天气图的距离定义为

$$D_{iS}=1-L_S \tag{8}$$

K-均值聚类算法[2,3]是逐步聚类法，其中K指类数，均值即类中心，在环流图聚类中就是同类环流图的平均图. 通常的K-均值方法聚类效果受所选的种子样本数目的影响，本文设计的变K-均值算法，不受初始类别数目的影响，而是通过四个参数控制使得在迭代过程中类数K可增减以得到最佳的分类结果. 具体步骤如下:

(1) 确定四参数: 样本归类阈值 T，类中心合并阈值 V，以及终止迭代的两个参数——前后两次聚类后变类样本数上限 $TOL1$ 和不能归类的少数环流图的集合零类的样本数上限 $TOL2$; 变类记数 $NT=0$;

(2) 输入种子样本，作为类中心初值;

(3) 对所有样本求出与当步类中心的最近距离 D_M，若 $D_M<T$ 则归入最近类 LM 中;否则，1) 若 $D_M>T$，且总分类数 $K<20$ 时，如果上一次归类时属单样本数，则归入0类;否则 $K=K+1$，建新类; 2) 若 $K\geqslant 20$，亦归入0类;

(4) 对所有样本完成归类以后，和上次聚类结果比较，若改变属类则 $NT=NT+1$;

(5) 计算类中心及类中心间距离，若两类中心距离小于 V，则合并成一类重新计算类中心;

(6) 把所有只含一个样本的类并入0类，并整理类编号;

(7) 若 $NT<TOL1$，转入(8)，否则，转(3);

(8) 若0类样本数$>TOL2$，则改变 T，V 转(2);否则输出聚类结果停机.

四、分级聚类结果和检验

把分级聚类方法应用于春季北半球 500hPa 等压面上高度场形势图(简称环流图)分类，以求得春季长江中下游三类主要天气型: 连阴雨、连晴和晴雨交替的过程性天气背景形势场为目的. 资料取 1979 至 1988 年 3 至 5 月共 920 张图，每张图上取 15—75°N，10° 经度×5° 纬度的格点高度值，一张图共有 468 个网格点高度值.

首先，用半模糊聚类法进行分类，取 $N=920$，$G=468$，Q 初值为 30; 参考等高线间距为 8 位势什米，故定阈值 $TH=100$，$ST=8$; $TOL=8$; 取 $M=2.1$，经过 19 步迭代后收敛共得 21 类，对每个样本取最大隶属度所属的类别归类，并对 21 类环流图求出平均图.

用半模糊聚类法求得的 21 张平均图作为种子样本，应用变 K-均值聚类法进行第 2 级分类. 取归类阈值 $T=0.20$，并类阈值 $V=0.10$，容许相邻两次聚类的变类样本数 $TOL1=20$，0 类样本数上限 $TOL2=20$，都在总样本数的 2% 左右，符合业务预报的需要. 经 15 步迭代后收敛，共得 9 型(图略).

为与分级聚类法比较，我们用同样资料，根据预报员经验选择 21 张典型环流图，根据序号从参加聚类的 920 张天气图中调出作为种子样本，用变值K-均法取相同参数值，经

20 步迭代收敛，共得 11 型(图略)。

首先比较两种聚类的类平均图象。分级聚类法抓住了北半球环流形势主要特征，同类样本更为集中，异类间差异亦较明显，可见各类平均图在长江中下游有明显特征，第 I 型至第 III 型以西北气流为主；而第 IV 型至 VI 型多平直气流、多槽脊活动；而 VII 型至 IX 型则多西南气流。为进一步定量说明，表 1 给出 20 天中两种聚类方法所得各类天气型出现与否和前述三类天气出现与否的相关系数。由表可见分级聚类法所得环流分型和三类天气的关系比不分级方法密切得多。确切地说，I、II、III 型提供了连晴天气的环流背景，IV、V、VI 型提供了过程性天气的环流背景，而 VII、VIII，IX 型提供了连阴雨天气的环流背景。

表 1　不同聚类方法所得天气型分类与春季长江中下游连晴、连阴雨、过程性天气的相关系数

分型	连晴		过程性		连阴雨	
	分级	不分级	分级	不分级	分级	不分级
I	0.51	0.23	0.24	0.21	0.16	0.31
II	0.47	0.29	0.21	0.31	0.19	0.40
III	0.45	0.41	0.31	0.28	0.21	0.21
IV	0.31	0.38	0.48	0.32	0.30	0.19
V	0.28	0.40	0.46	0.31	0.28	0.18
VI	0.30	0.26	0.44	0.40	0.31	0.34
VII	0.21	0.25	0.30	0.41	0.51	0.38
VIII	0.14	0.24	0.29	0.29	0.46	0.33
IX	0.23	0.30	0.21	0.27	0.54	0.18
X		0.19		0.28		0.32
XI		0.21		0.26		0.35

背景并不等于某种分型下一定伴生某类天气过程。为应用于日常业务天气预报工作的需要，我们在分型条件下，应用 40 个初选因子，其中包括与三类天气过程发生与否关系较密切的 20 个高度场关键区因子和 20 个不同纬圈上谐波分析中选出的物理量：角动量输送；角动量输送的散度；经向、纬向和平均动能等的谱值，建立基于 Wilks Λ 准则的逐步判别方程。分级聚类分型下共有 $9 \times 3 = 27$ 个判别方程，变 K-均值法一步聚类分型下共 $11 \times 3 = 33$ 个判别方程。在引入和剔除变量时取 Wilks 统计量等价的 F 检验式，分级聚类下引入变量和剔除变量的 F 阈值 $F_1 = F_2 = 4.0$，入选 4 至 6 个因子；而一步聚类分型下 F_1 和 F_2 要降到 2.7 才可得 4 至 6 个因子。

表 2　1989,1990,1991 年 3—5 月平均准确率

项　目	分级聚类	一步聚类
未来 3 天	221/276=80%	185/276=67%
未来 4 天	218/276=79%	141/276=51%
未来 5 天	204/276=74%	138/276=50%

实际应用中未来的天气图形势用欧洲中期预报中心每日传送的 3—5 天预报图，按(8)式求它们与各类中心的距离，取距离最小的类别所对应的判别方程组，制作未来 3—5

天长江中下游连阴雨、连晴、过程性三类天气的预报. 经 1989、1990、1991 三年 3—5 月的检验,预报准确率如表 2 所示,分级聚类法明显优于一步聚类方法. 比较结果, 分级聚类法对应的方程组预报准确率达到或稍高于同期业务预报水平, 而一步聚类法对未来 4—5 天几乎没有预报能力. 实际业务预报的效益,再次证明了分级聚类法的优越性.

五、结 论

(1) 本文对半模糊聚类法作了改进,并自行设计了适用于天气图分类的变 K-均值聚类法,把两者结合起来实现了分级聚类方法, 保证了引入的初始类中心较为客观,减小了由于主观选择种子样本不当而对变 K-均值方法聚类结果产生的影响.

(2) 环流图的相似本身具有模糊性质,所以先用半模糊聚类方法是合适的;而为了求得一个较确切的背景形势场,用同时描述高度值和槽脊位置相似性的相似指数, 用变 K-均值方法做进一步聚类又是必要的. 所以,这样的分级聚类法符合气象问题的性质.

(3) 经与用变 K-均值法一步聚类结果比较,从图形的天气学意义, 它们与长江中下游春季连阴雨、连晴、过程性三类主要天气类型的相关性以及分类判别方程的检验效果看,分级聚类法是较好的聚类方法.

参 考 文 献

[1] Bezdek, J. C., 1981, Pattern receognation with fuzzy objective Function algorithms, Academic press.
[2] Hartigan, J. A., 1975, clustering algorithm, John Wiley and Sons.
[3] Anderberg, M. R., 1973, Clustering analysis for application, Academic Press.

Two Step Clustering Algorithm with Its Application to Objective Pattern Recognition of General Circulation over the Northern Hemisphere and Weather Forecasting

Zhu Shengming Jiang Cuihong
(Institute of Meteorology of Jiangsu Province, Nanjing 210008)

Abstract

In this paper a two step clustering algorithm is shown. First partial fuzzy clustering is improved and applied to pattern recognition of general circulation at 500 hPa over the Northern Hemisphere. And 21 patterns are obtained. These patterns are used as seed samples of a changeable K-means method to get g patterns finally.

The results indicate that the two step algorithm is useful to predictor selection and building of prediction model.

Key wrods: Two step clustering; Pattern recognition; Weather forecasting.