基于差异序列信息熵的降雨分区研究

管耀宗[1],张继国[2],朱永忠[1]

(1. 河海大学理学院,江苏 南京 210098; 2. 河海大学水利信息统计与管理研究所,江苏 常州 213022)

摘 要:充分挖掘降雨变量语法信息,利用基于差异序列信息熵测度理论的遗传算法聚类分析方法,对淮河流域蚌埠站以上区域99个雨量站进行了分区研究。根据各子区域内雨量站降雨序列差异测度得到最优分区,从而使得降雨信息在区域内具有较大的同质性。最后,以插值精度为标准,将不分区的情况为参照对象,对2、4、8三种分区结果进行校验,结果显示,相对于不分区,分区明显提升了降雨插值的精度。

关键字:降雨;差异序列信息熵;区域划分;遗传算法

中图分类号:P467 **文献标识码**:A **文章编号**:1000-0852(2015)05-0011-04

1 引言

水文模型是一个复杂的系统模型,降雨作为水文模型的一个重要输入量,其精度严重影响流域水循环,产汇流的模拟精度[1]。而降雨本身具有较强的时空分布不均匀性,因此,对这种不均匀性的考虑越充分,水文模型的精度就会越高。对于降雨的空间分布不均匀性,以已有站点为基础,将区域作分区化处理,减少不确定性因素的影响是一种切实有效的研究方法。因此在探讨降雨信息空间插值时,应首先将复杂的降雨测量站点系统划分成不同的子系统[2]。近些年来,不少学者对我国不同区域的降雨进行了分区研究,取得了具有较高学术价值的研究成果[3-5]。

传统的分区方法主要是依据站点雨量插值,即仅考虑了降雨信息的语义信息而忽视了其语法信息。本文根据差异信息序列的信息测度理论[6],利用差异信息熵反映降雨的语法信息,利用遗传算法聚类分析方法[7,8],将淮河流域蚌埠站以上的99个雨量站予以划分,使其各子区域内降雨信息差异最小。其研究结论将为流域内站网优化布局、降雨不均匀性分析、降雨空间插值提供重要的科学依据。

2 基本概念与方法

2.1 差异信息序列的测度理论

1948年,Shannon为了解决对信息的量化度量问题,提出了信息熵的概念,表明信息的信息量大小和它的不确定性之间有着直接的关系。基于信息熵理论,张岐山在灰色系统理论中指出差异就是信息,并基于差异信息理论给出差异信息序列的定义[9]。

设序列 $X=\{x(j)|j\in J\}$,指标集 J 为有限集或无限集,$X\geq 0$,称 X 为差异信息序列,当且仅当 X 具有下述内涵:(1)X 各分量值差异性越大,序列所含的信息量就越大;(2)X 各分量值无差异性,序列所含信息量为零。

Shannon信息熵是定义在概率空间上,所以无法直接用于非概率现实信息。于是,夏军在灰色系统的灰序列基础上提出了一种基于信息熵的非概率信息差异序列测度理论[6],并构造了差异信息序列的信息熵。

收稿日期:2014-09-03
基金项目:江苏省自然科学基金项目(BK20131135)
作者简介:管耀宗 (1990-),男,浙江桐乡人,硕士研究生,研究方向为水文不确定性分析。E-mail:413898309@qq.com

设 $X=\{x_1,x_2,\cdots,x_s\}\in A^s$，且至少 $\exists x_j\neq 0, j\in J, J=\{1,2,\cdots,s\}$ 为序列的指标集，如果 $x(i)$ 为非归一化序列，则总能找到差异信息序列的信息结构算子 f，使映射 $f: X\to Y=\{y_1,y_2,\cdots,y_s\}$，其中 $y_j=f\{x_j|x_i, i=1,2,\cdots,s\}$ 为归一化序列。如果选择信息结构算子为 $y_j=f(x_j)=\left(\frac{1}{1+x_j}\right)\Big/\left(\sum_{i=1}^{s}\frac{1}{1+x_i}\right)$，则称

$$I(x)=-K\sum_{j=1}^{s}y_j\ln(y_j)\quad j\in J, K=1/\ln 2 \tag{1}$$

为差异信息序列 X 的信息熵。

当且仅当各元素全部相同的时候，信息熵取得最大值为 $I_{\max}(X)=K\ln s$，即 $I_{\max}$ 对应的是无差异的特殊情况。在此，可定义

$$I_d(X)=I_{\max}(X)-I(X) \tag{2}$$

为差异信息序列的信息测度。

为了描述信息熵（潜在信息量）与差异信息量（现实信息量）之间的联系与转化关系，可定义差异信息序列相对测度。

由于信息序列的归一化约束，差异信息序列的信息熵一般取不到零值，但总有一个最小值 $I_{\min}(X)$。若定义最大差异信息量 $I_{d,\max}(X)$ 为

$$I_{d,\max}(X)=I_{\max}(X)-I_{\min}(X) \tag{3}$$

则差异信息量 $I_d(X)$ 与其最大差异信息量 $I_{d,\max}(X)$ 之比被称为差异信息的相对测度，记为 $I_a(X)$。

$$I_a(X)=\frac{I_d(X)}{I_{d,\max}(X)}\times 100\%=\frac{I_{\max}(X)-I(X)}{I_{\max}(X)-I_{\min}(X)}\times 100\% \tag{4}$$

实际应用中可以取各 $I(X)$ 中最小者为最小值 $I_{\min}(X)$。

2.2 遗传算法聚类分析

聚类就是按照某个特定标准把一个数据集分割成不同的类或簇，使得在同一个类内的数据对象有较大的同质性，而不同类之间有较大的异质性[10]。传统的 k-means 和 FCM 算法具有收敛速度快且易于扩展的优点，但是初始区域中心的选择、噪声数据的存在均会对聚类结果产生较大的影响[11]。遗传算法是一种宏观意义下的仿生算法，具有强大的全局搜索能力，基于遗传算法的聚类分析可以对大量数据进行聚类，而且有好的效果和效率[8]。

遗传算法主要步骤是将表现型映射到基因型即编码，每个编码对应一个问题解，称为染色体或个体。随机产生一组个体作为第一代种群，按照适者生存优胜劣汰的原则，逐代产生越来越好的近似解。在每一代，根据个体对应解的优劣作为适应度大小选择个体，并通过自然遗传学的遗传算子进行交叉变异产生新一代种群，最后将多次迭代后的结果作为问题的最优近似解。

3 研究实例

3.1 数据来源及处理

所研究的数据是来自淮河流域蚌埠站以上区域 99 个雨量站，站点基本情况可参见文献[10]。

对于原始降雨数据 $(X_1,X_2\cdots,X_N)$，其中 $X_i=\{x_i(1), x_i(2)\cdots,x_i(M)\}$ 代表每个站点的降雨时间序列，N 为站点数，M 为每个站点降雨量信息数。站点间差异信息序列定义为

$$X_{ij}=\{x_{ij}(1),x_{ij}(2)\cdots,x_{ij}(M)\}, 1\le i,j\le N \tag{5}$$

式中：

$$x_{ij}(k)=|x_i(k)-x_j(k)| \tag{6}$$

3.2 研究思路

（1）编码。由于遗传算法容易实现对二进制数的处理，因此本文首先将分区数简化为 $2^n(n=1,2,3\cdots)$ 个。假设以一个 n 位二进制数作为单元表示一个雨量站所属的分区，则可以用 $99\times n$ 位二进制数作为区域内雨量站分区情况的一种编码。

（2）算子选择。根据式（5）和式（6），计算得到 99 个站点间的信息差异序列矩阵 $(X_{\alpha\beta})_{99\times 99}, 1\le\alpha,\beta\le N$。再通过差异信息序列的信息测度理论求出两两雨量站间的差异信息相对测度，构成 99 个雨量站之间的差异信息矩阵 $(I_a(X_{\alpha\beta}))_{99\times 99}$。

设分区数为 k，第 j 个分区内雨量站数目为 n_j，定义同一分区内部信息差异 $I^{(j)}=\sum_{\alpha=1}^{n_j}\sum_{\beta=1}^{n_j}I_a(X_{\alpha\beta})$ 则分区后整体信息差异度记为 I_{dll}：

$$I_{dll}=\frac{1}{\sum_{j=1}^{k}n_j\times n_j}\sum_{j=1}^{k}I^{(j)} \tag{7}$$

I_{dll} 越小，区域内信息差异越小，相应的分区就越优。为此可将 I_{dll} 值的大小作为适应度大小进行选择。

（3）交叉和变异算子。运用一点交叉算子即在个体串中随机设定一个交叉点，在交叉时，该点前或后的个体部分进行互换，并生成两个新个体。

变异算子是以事先设定的变异概率 P_m 将个体上某些基因值取反。

（4）技术路线图。本实例的技术路线图如图 1 所示。

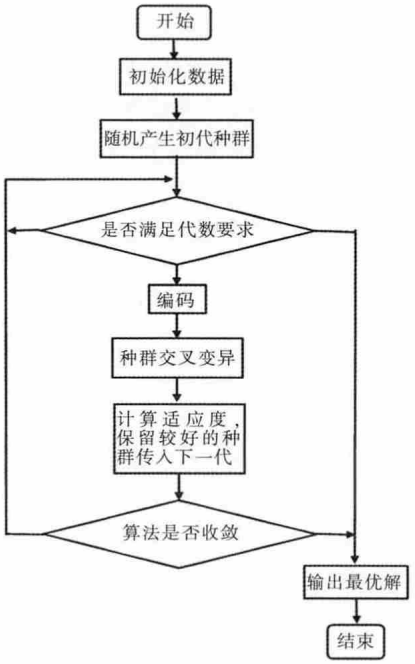


图 1　遗传算法聚类技术路线图
Fig.1　The technology roadmap of genetic algorithm

3.3　分区结果

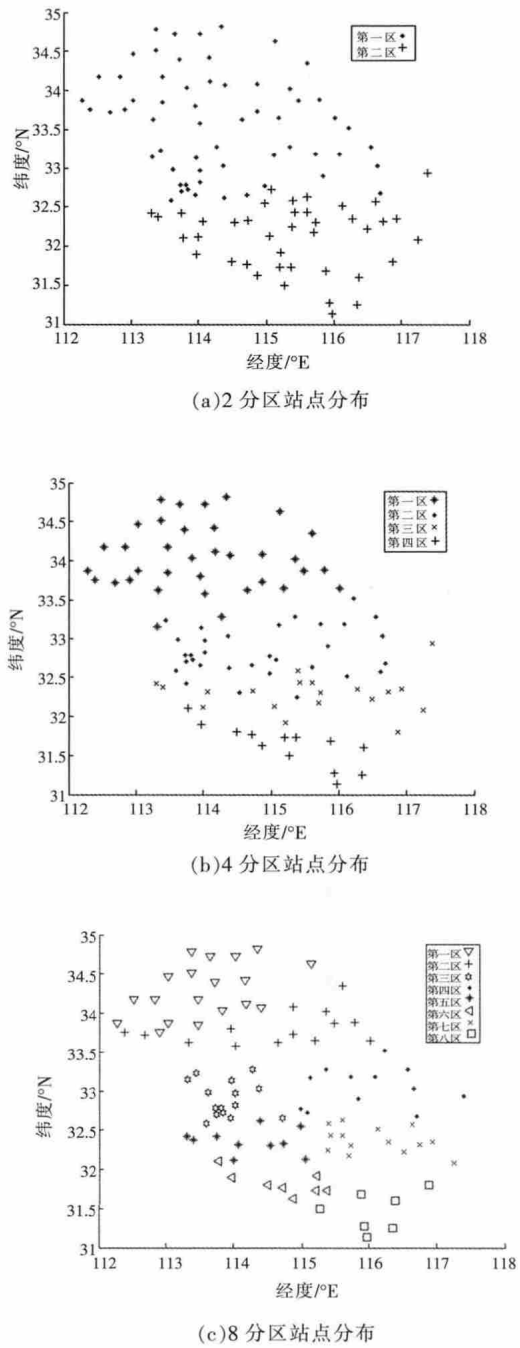
取 n=1,2,3 将分区数为 2、4、8 个的情况进行遗传算法聚类分析。分区结果如图 2 所示。

可以看出，算法算出的分区情况有明显的地理位置聚拢性，十分符合实际情况，即地理位置较近的地方,降雨量差异性较小。此外,分区多以纬度划分为主,说明降雨量在纬度变化中,差异变化大。

3.4　分区合理性验证

对雨量站分区一个主要作用是提高降雨量插值的精度,因此可以通过雨量插值精度反映分区的合理性。分别假设区域内某一个雨量站的雨量数据为未知,以反距离平方加权插值方法，对未知雨量站进行插值处理。得到分区情况下和不分区情况下两组插值雨量数据。以该雨量站真实测量数据为比较数据，根据公式(5)计算得到两种情况下的差异序列。对差异序列各分量求和得到差异值 $S_{分区}$,$S_{不分区}$。若差异值 S 越小,则表示插值结果与实际结果差异较小,反之,则表示插值结果与实际差异较大。因此可以以每个雨量站相对差异提升 $(S_{不分区}-S_{分区})/S_{不分区}$ 值来判别分区的合理性。相应的计算结果列于表 1。

找出平均站点相对差异提升值为正的站点，即分区后插值更加精确的站点。优化明显和优化不明显的站点见图 3。



(a)2 分区站点分布

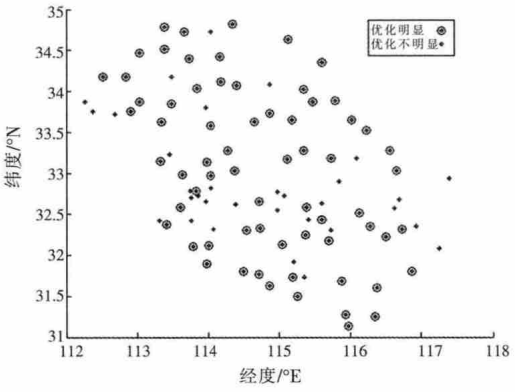
(b)4 分区站点分布

(c)8 分区站点分布

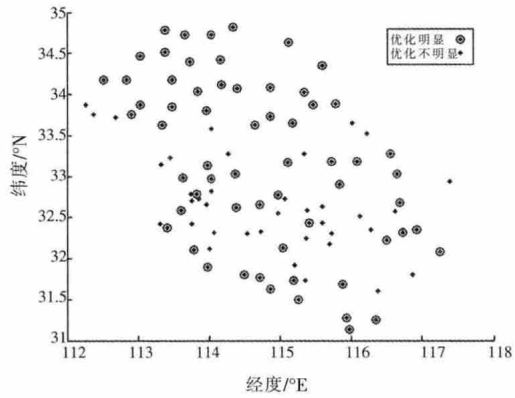
图 2　淮河流域蚌埠站以上 99 个站划分站点分区分布
Fig.2　The classification results of the 99 stations above the Bengbu station in the Huaihe River basin

表1　各分区情况插值平均提升值
Table1　The Average Increase of each Classification

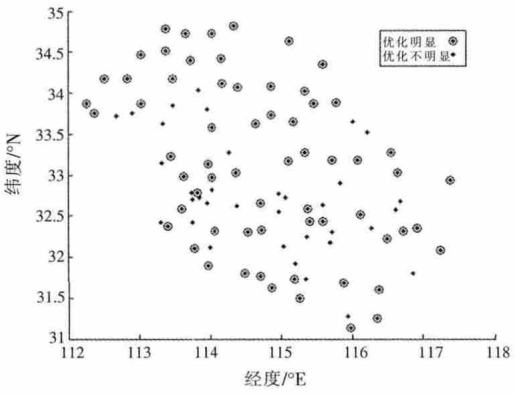
分区数	2 分区	4 分区	8 分区	不分区
S	16.91	16.69	16.33	17.76
平均站点相对差异提升值	4.81%	6.04%	8.06%	0



(a)2 分区插值优化站点分布



(b)4 分区插值优化站点分布



(c)8 分区插值优化站点分布

图 3 插值优化站点分布图

Fig.3 The interpolation optimization stations above the Bengbu station in the Huaihe River basin

可以看出,分区后大部分站点插值更加精确了,表明这种分区方法具有一定的优越性。然而,通过对比相应的分区图,发现处于各个子区域边界的雨量插值情况并没有得到太大的提升。

4 总结

通过分区化处理,可以将一个复杂的大系统问题化成多个简单子系统问题进行研究,有利于更加深入的探究系统内部关系,及其多种不确定因素。本文根据降雨量的差异信息熵,通过遗传算法,对淮河流域蚌埠站以上区域进行了区域划分,使得子区域内部站点差异信息熵最小。本文还通过插值精度验证了分区结果的合理性,并指出在区域边界处还可以做进一步研究。

参考文献:

[1] 姜红梅,任立良,袁飞. 降水空间不均匀性对径流过程模拟的影响[J]. 水文, 2004,24(2):1-6. (JIANG Hongmei, REN Liliang, YUAN Fei. Effect of spatial precipitation heterogeneity on runoff process [J]. Journal of China Hydrology, 2001,24(2):1-6. (in Chinese))

[2] 张继国,谢平,龚艳冰,等. 降雨信息空间插值研究评述与展望[J]. 水资源与水工程学报, 2012,23(1):6-9. (ZHANG Jiguo, GONG Yanbing, LIU Gaofeng. Review and perspectives of the research on spatial interpolation of rainfall information [J]. Journal of Water Resources and Water Engineering, 2012,23(1):6-9. (in Chinese))

[3] 郑永宏,林爱文,代侦勇. 湖北省降水分区研究[J].长江流域资源与环境, 2012,21(7):859-863. (ZHENG Yonghong, LIN Aiwen, DAI Zhenyong. Research on precipitation regionalization in Hubei province [J]. Resources and Environment in the Yangtze Basin, 2012,21(7):859-863. (in Chinese))

[4] 杨绚,李栋梁. 中国干旱气候分区及其降水量变化特征[J].干旱气象, 2008,26(2):17-24. (YANG Xuan, LI Dongliang. Precipitation variation characteristics and arid climate division in China [J]. Arid Meteorology, 2008,26(2):17-24.(in Chinese))

[5] 李生辰,徐亮,郭英香,等. 近34a青藏高原年降水变化及其分区[J]. 中国沙漠, 2007,27(2):307-314. (LI Shengchen, XU Liang, GUO Yingxiang, et al. Change of Annual Precipitation over Qinghai-Xizang plateau and sub-regions in recent 34 years [J]. Journal of Desert Research, 2007,27(2):307-314. (in Chinese))

[6] 夏军. 灰色系统水文学—理论、方法及应用[M]. 武汉:华中理工大学出版社,2000. (XIA Jun. Grey System Hydrology—Theory, Methods and Application [M]. Wuhan: Huazhong University of Science and Technology Press, 2000. (in Chinese))

[7] 陈锐,邹书蓉,张洪伟,等. 改进遗传算法及其在聚类分析上的应用[J]. 西南民族大学学报(自然科学版), 2009,35(6):1176-1179. (CHEN Rui, ZOU Shurong, ZHANG Hongwei, et al. Improvement of the genetic algorithm and its application on clustering [J]. Journal of Southwest University for Nationalities (Natural Science Edition), 2009,35(6):1176-1179. (in Chinese))

[8] 孙志胜, 曹爱增, 梁永涛.基于遗传算法的聚类分析及其应用[J].济南大学学报(自然科学版), 2004,18(2):127-129. (SUN Zhisheng, CAO Aizeng, LING Yongtao. Cluster analysis based on genetic algorithm and its application [J]. Journal of Shandong Institute of Building Materials, 2004,18(2):127-129. (in Chinese))

[9] 张岐山. 灰朦胧集的差异理论 [D]. 武汉: 华中科技大学,1996. (ZHANG Qishan. Disconfirmation Theory of Grey Hazy Set [D]. Wuhan: Huazhong University of Science and Technology, 1996. (in Chinese))

(下转第 29 页)

Changhui. Retrieving groundwater in Yellow River delta area using remote sensing [J]. Remote Sensing for Land&Resources, 2013,25(3):145-152. (in Chinese))

[16] 余涛,田国良. 热惯量法在监测土壤表层水分中的研究[J]. 遥感学报, 1997,1(1):24-31. (YU Tao, TIAN Guoliang. The application of thermal inertia method monitoring of soil moisture of north China plain based on Nova-Avhrr data [J]. Journal of Remote Sensing, 1997,1(1):24-31. (in Chinese))

[17] 徐涵秋. 利用改进的归一化差异水体指数(MNDWI)提取水体信息的研究[J]. 遥感学报,2005,9(5):589-595. (XU Hanqiu. A study on information extraction of water body with the modified normalized difference water index (MNDWI) [J]. Journal of Remote Sensing, 2005,9(5):589-595. (in Chinese))

[18] 吴德文,张远飞,朱谷昌. 遥感图像岩石信息提取的最优密度分割方法 [J]. 国土资源遥感, 2002,54(4):51-54. (WU Dewen, ZHANG Yuangfei, ZHU Guchang. The best density separation method for extracting rock information from remote sensing image [J]. Remote Sensing for Land&Resources, 2002,54(4):51-54. (in Chinese))

Study on Predicting Shallow Groundwater in Semi-arid Area Based on Soil Humidity Index of TM Data: Taking Chaoyang City as A Study Case

ZHENG Pu[1,2], DENG Zhengdong[1], WANG Daqing[1], XU Chunhua[1], DENG Feifan[1]

(1. *PLA University of Science & Technology*, *Nanjing* 210007, *China*; 2. *No.* 96528 *Troops of PLA*, *Beijing* 102202, *China*)

Abstract: In order to get the water information quickly in arid and semi-arid areas, Chaoyang City of Liaoning Province was taken as a study area. The soil humidity information was acquired by taking use of MNDWI and based on TM multi-spectral data. The soil humidity information classification map was made by classify the soil information with vegetation, rock and surface water being processed. At the same time, in order to compare the groundwater information of the area with higher soil humidity, investigation was executed to compare the the hydrological and geological data. The basic conclusion was consistent with the remote sensing, so as to prove that this method is feasibility and validity in arid and semi-arid areas.

Key words: remote MNDWI; shallow groundwater

(上接第 14 页)

[10] 杨小兵. 聚类分析中若干关键技术的研究[D]. 杭州: 浙江大学, 2005. (YANG Xiaobing. Research of Key Techniques in Cluster Analysis [D]. Hangzhou: Zhejiang University, 2005. (in Chinese))

[11] 王骏,王士同,邓赵红,等. 聚类分析研究中的若干问题[J]. 控制与决策, 2012,27 (3):321-328. (WANG Jun, WANG Shitong, DENG Zhaohong. Survey on challenges in clustering analysis research [J]. 2012,27(3):321-328. (in Chinese))

[12] 张继国. 降雨时空分布不均匀性信息熵研究[D]. 南京: 河海大学, 2004. (ZHANG Jiguo. Information Entropy Study on Rainfall Distribution in Time and Space [D]. Nanjing: Hohai University, 2004. (in Chinese))

Study on Precipitation Regionalization Base on Diversity Sequence Information Entropy

GUAN Yaozong[1], ZHANG Jiguo[2], ZHU Yongzhong[1]

(1. *College of Science, Hohai University, Nanjing* 210098, *China*;

2. *Institute of Hydraulic Information Statistics and Management, Hohai University, Changzhou* 213022, *China*)

Abstract: This paper fully dug the grammatical information of the rainfall as a variable, used the way of genetic algorithm clustering based on diversity sequence information entropy measure theory; and made regionalization research on the 99 precipitation stations above the Bengbu Station in the Huaihe River Basin. The optimal classification was obtained based on the measures of the rainfall sequences of the precipitation stations in the various sub-regions. Therefore, the research objective of high homogeneity among the rainfall information in the sub-regions was achieved. Finally, the 3 classifications of 2, 4 and 8 were checked by using the accuracy of interpolation, and taking un-categorization as reference object. The results show that the classification can improve the accuracy of rainfall interpolation.

Key words: rainfall; diversity sequence information entropy; regionalization; genetic algorithm