

丹顶鹤繁殖地气候数据特征的聚类分析*

彭丽潭, 晏寒冰, 唐旭清

(江南大学理学院, 江苏无锡214122)

摘要:以1981—1990年我国东北地区繁殖的候鸟野生丹顶鹤为例,调查其繁殖地分布区域内的月平均气温、月平均最高温度、月平均最低温度、月平均降水和月平均净辐射五个气候因子数据的基础上,采用分区间统计及基于归一化距离的分层聚类及聚类融合等理论和方法进行相关气候数据的特征分析,获得丹顶鹤在东北繁殖地的最适宜栖息时间为5~9月份。这一结果与东北地区丹顶鹤的生物学和生态学特征相吻合。与以往文献中用于物种分布预测的气候因子提取方法不同的是,该方法完全依赖于预测物种在调查分布区内的气候数据,通过数据挖掘与数据处理而获得,而不是通过预测物种的生物学与生态学特征及其在分布区内的气候因子相关性分析得到。这一研究方法具有一般性,将为未来气候变化对候鸟在东北地区栖息地的影响的研究提供基础。

关键词:丹顶鹤;气候因子;归一化距离;分层聚类;聚类融合

中图分类号:TP391;O29

文献标志码:A

文章编号:1001-3695(2014)03-0747-06

doi:10.3969/j.issn.1001-3695.2014.03.026

Clustering analysis on climate data of red-crowned crane breeding distribution area in Northeast China

PENG Li-tan, YAN Han-bing, TANG Xu-qing

(School of Science, Jiangnan University, Wuxi Jiangsu 214122, China)

Abstract: Based on the five climatic factors such as monthly mean temperature, monthly mean maximum temperature, monthly average lowest temperature, the average monthly net radiation and the monthly mean precipitation in the breeding distribution area of the typical migratory species such as red-crowned crane in Northeast China from 1981 to 1990, this paper studied the characteristics analysis on the related climate data by using the theory and methods of interval statistics, hierarchical clustering based on normalization distance and clustering fusion technology. The results show that the most appropriate time in breeding habitat was May to September. The conclusion coincided with the biological characteristics and ecological characteristics of red-crowned crane in Northeast China. Extracting climatic factors from climate data was critical to the predictive distribution of species. The used climatic factors in the past mostly depended on the biological and ecological characteristics of species and the correlation analysis for climatic factors in the distribution area of species. Instead, this method only depended on climate data in the distribution area of the predictive species entirely, and it obtained data mining and data processing. This set of research methods possess generality and will provide basis for the research about the impact of future climate change on the predictive distribution of the migratory species in Northeast China.

Key words: red-crowned crane; climatic factor; normalization distance; hierarchical clustering; clustering fusion

0 引言

全球变暖现已成了不争的事实。联合国政府气候变化专门委员会(IPCC)在第四次全球气候变化的评估报告中明确指出:气候系统变暖是毋庸置疑的,目前从全球平均气温和海温上升、大范围积雪和冰融化、全球平均海平面上升的观测中可以看出气候系统变暖是明显的^[1]。近100年(1906—2005年)地球表面气温上升了 $0.74 \pm 0.18^\circ\text{C}$,这一上升趋势大于第三次评估报告给出的 0.6 ± 0.2 的上升趋势(1901—2000年)^[2],到本世纪末,全球地球表面气温将上升 $3.75 \pm 2.65^\circ\text{C}$ 。除全球地球表面气温变化外,随着热带气旋度的增加,温带风暴路径向极地推移等现象的出现,全球气候变化还表现在降水、辐射和风向等气候因子上^[3]。这些气候因子的变化对全球生态

系统,特别是对陆地生态系统产生了深刻的影响^[4],进而影响到全球物种的物候与物种的分布。因此,未来气候变化对物种分布的影响与预测问题研究一直是人们关注的热点问题之一。

生态系统的结构和物种组成是系统稳定性的基础,生态系统的结构越复杂、物种越丰富,则系统表现出良好的稳定性,其抗干扰能力越强;反之,其结构简单、种类单调,则系统稳定性差,抗干扰能力相对较弱^[4,5]。千万年来,不同的物种为了适应不同的环境条件而形成了其各自独特的生理和生态特征,从而形成现有不同生态系统的结构和物种组成。由于原有系统中不同物种对气候因子变化的响应存在着很大的差别,因此气候变化将改变内陆生态系统的结构和物种组成。气候变化可通过以下途径使物种组成和结构发生改变^[4,6-9]:

a)温度胁迫。温度是物种分布的主要限制因子之一,高

收稿日期:2013-06-18; **修回日期:**2013-07-28 **基金项目:**国家自然科学基金资助项目(11271163);国家环保部环保公益性行业科研专项基金资助项目(200909070)

作者简介:彭丽潭(1988-),女,四川阆中人,硕士,主要研究方向为粒度计算、生态系统建模与仿真(pretty182@163.com);晏寒冰(1987-),男,硕士,主要研究方向为粒度计算、生态系统建模与仿真;唐旭清(1963-),男,教授,博士,主要研究方向为智能计算、生态系统建模与仿真、生物信息学。

温限制了北方物种分布的南界,而低温限制热带和亚热带物种向北分布。

b) 水分胁迫。随着全球气温升高,虽然现有大气环流模型(GCM)预测全球降雨量将有所增加,但降雨量随地区的不同和季节的不同而存在很大的差别。根据 IPCC 的研究表明:在 1900—2005 年期间,南美洲及北美洲东部等一些地区的降水呈显著增加趋势,但中纬度内陆地区尤其是在夏季其降雨会相对减少,在一些热带地区其干旱季节也将延长,如地中海地区、南非和南亚等地区呈干旱化趋势。

c) 日照和光强的变化。日照时数和光照强度的增加,将有利于阳性植物的生长和繁育,但对于耐阴性植物来说,其生长将受到严重的抑制,尤其是其后代的繁育和更新将受到强烈的影响。

d) 物候变化。冬季和早春温度的升高还会使春季提前到来,加上日照和光强的变化,从而影响到植物的物候,使其提前开花放叶,这将对那些在早春完成其生活史的林下植物产生不利的影响,甚至有可能使其无法完成生命周期而导致灭亡。

e) 有害物种的入侵。有害物种往往有较强的适应能力,它们更能适应强烈变化的环境条件而处于有利地位。

近年来国内外已广泛开展气候变化对物种分布影响与预测的研究。Parmesan 等人^[10]探讨了 1 700 多个物种在过去 20~140 年间分布区的变化,发现物种分布区的迁移与气候变化相关。Forsman 等人^[11,12]分析了气候变化对鸟类分布的影响,Erasmus 等人^[13]选择年及月均气温、年及月最高最低气温和繁殖期降水量,Forsman 等人^[14]选择最冷月气温和年均气温及繁殖期气温和降水量,Luoto 等人^[15]选择最冷月气温和降水量及积温大于 5℃ 的来分析气候变化对物种分布影响。张雷等人^[16]选用年平均气温、最冷月平均气温、最暖月平均气温、气温年较差、年平均降水量、平均夏季降水量及大于 5℃ 的积温 7 个气候变量和 10 个环境变量进行模拟,分析了我国主要树种的历史基准气候(1961—1990 年)和未来不同时期的潜在分布。这些研究都是通过比较物种在观测或基准情景下与未来单一年气候变化情景下分布差异,确定气候变化对物种分布的影响,同时表明:a) 物种的分布都与其分布区内的温度、降水量、蒸发量、净辐射等气候因子密切相关,其中都是通过比较物种在观测或基准情景下与未来单一年气候变化情景下分布差异,确定气候变化对物种分布的影响;b) 物种未来分布预测的气候因子的选取都依赖物种的生物学与生态学特征及其在分布区内的气候因子相关性分析。目前,不同长时间序列气候变化情景分析对物种分布与预测影响正日益受到关注^[3,17]。

丹顶鹤物种作为东北地区的典型候鸟物种,具有很高的生态价值和经济价值^[18-22]。对于不迁徙种群来说:高温限制了北方物种分布的南界,而低温则是热带和亚热带物种向北分布的限制因素。对于迁徙种群,丹顶鹤种群分布区域内各月平均最高温度、月平均最低温度等五个气候因子的数据自然就影响到丹顶鹤种群在该区域内逗留的时间段。由于每个物种的生物学特征是该物种自身长期进化以及与生态系统和气候环境相适应的结果,各物种对各气候因子的依赖具有相对稳定性。因此假定:丹顶鹤种群对各气候因子的依赖性稳定的,即不考虑丹顶鹤种群的生物学和生态学特征的变化,特别是随着全球气候变化而带来的丹顶鹤对各气候因子的适应性的改变。

文献[23~27]研究表明:1981—1990 年间丹顶鹤在繁殖

地的春季始见日和秋季终见日分别为每年的 3 月中下旬和 10 月下旬至 11 月中旬;丹顶鹤越冬地的春季终见日和秋季始见日分别为每年的 3 月中下旬和每年的 10 月中下旬;丹顶鹤的春季迁徙途径北戴河、黄河故道(河南)和山东的时间为每年的 3 月至 4 月,而秋季迁徙途径北戴河、黄河故道(河南)和山东的时间为每年的 10 月至 11 月上旬,即丹顶鹤的春季迁徙和秋季迁徙分别需要 1 个月左右的时间。这些信息表明:丹顶鹤在繁殖地的适宜栖息时间为 3 月中下旬至 11 月上旬;丹顶鹤的春季迁徙和秋季迁徙分别需要 1 个月左右的时间,即 3 月中下旬至 4 月上旬为丹顶鹤春季迁徙时间,而 10 月中下旬至 11 月中旬为丹顶鹤种群秋季迁徙时间,同时每年的 3~4 月份及 10~11 月份在繁殖地和越冬地均适宜丹顶鹤种群栖息,也就是说,每年的 5~9 月是丹顶鹤在繁殖地的最适宜栖息时间,相应地,每年的 11 中旬~12 月,及 1~3 月中旬是丹顶鹤在越冬地的最适宜栖息时间。因此,如果丹顶鹤种群的繁殖地和越冬地分布区域,以及迁徙路线和迁徙过程中停歇不变,且我国东北地区的丹顶鹤的春季迁徙和秋季迁徙所需要的时间不变,分别需要一个月左右的时间。因此,一旦确定了丹顶鹤种群每年在繁殖地逗留时间段,则一年中此时间段的之前和之后的一个月分别是丹顶鹤春季迁徙和秋季迁徙的时间段,从而丹顶鹤种群每年在越冬地逗留的时间段也随之确定。

本文将在 1981—1990 年丹顶鹤种群的繁殖地的月平均气温、月平均最高温度、月平均最低温度、月平均净辐射和月平均降水这五个气候因子数据的基础上,采用分层聚类方法,进行丹顶鹤种群的繁殖地气候数据特征分析,研究丹顶鹤种群的繁殖地区域内各气候因子数据的内在规律。

1 研究区域概况与数据来源

1.1 研究区域概况

研究区域包括丹顶鹤的繁殖分布区、越冬分布区和迁徙沿线。丹顶鹤种群繁殖地主要集中我国东北地区,包括黑龙江的齐齐哈尔、三江平原洪河、七星河流域、嫩江下游、兴凯湖、乌裕尔河下游和辉河流域,吉林的西部向海和莫莫格,辽宁的盘锦双台子河下游以及内蒙古达里诺尔湖等地,其中我国乌裕尔河下游的扎龙湿地自然保护区已是世界上最大的丹顶鹤繁殖地,约占现存丹顶鹤总数的 50%,地理坐标为东经 123°47′~124°37′,北纬 46°52′~47°32′,水禽理想的栖息地。保护区区内年均气温 3℃~5℃,最冷月 1 月的平均气温 -19.2℃ 左右,极端最低气温 -39.5℃,年均降雨量 420 mm 左右,年蒸发量 1 506 mm 左右,年均相对湿度 62% 左右。迁徙种群沿东、西两条迁徙路线分别在我国东南沿海各地与长江下游、朝鲜海湾等地越冬,即西线沿扎龙—双台河口—环渤海—黄河口—日照最后到达盐城,东线沿兴凯湖、长白山东缘,经图们、江口到达朝鲜半岛的三八线附近。丹顶鹤种群越冬地主要集中我国江苏盐城国家级珍禽自然保护区,该保护区属滨海湿地,地理位置位于东经 119°29′~121°16′,北纬 32°20′~34°37′,东濒黄海,海岸线长达 582 km,滩涂上散布着众多的河流和沼泽,主要入海河流均属淮河水系。主要湿地类型包括永久性浅海水域、滩涂、盐沼和人工湿地等。

1.2 数据来源

丹顶鹤种群繁殖地和越冬地分布区域数据来源东北林业

大学于1981—1990年间的调查数据及文献挖掘,应用地理信息系统GIS,可获得我国丹顶鹤种群在1981—1990年间繁殖地和越冬地分布区域位置的数据。应用这些地理信息系统GIS数据,通过MATLAB编程可得丹顶鹤种群在1981—1990年间繁殖地和越冬地分布区域如图1所示。图1的左下角对应的经纬度分别是东经115.602 955°和北纬28.856 615°;图1的每个网格跨度为0.008 421°,对应于1 km的实际距离,且所截的我国境内的陆地区域的行数为2 957,列数为2 321。在软件ArcGIS 9.3界面之下,提取在1981—1990年东北地区丹顶鹤种群的繁殖地和越冬地的渔网点(即图1中繁殖地和越冬地内行与列交叉的网格点个数)数据,且它们的个数分别是53 084和2 091个,记 $M_1=53\ 084$ 和 $M_2=2\ 091$ 。这些网格点数据集是获取相应年份气候数据的基础。

我国东北地区丹顶鹤种群的气候因子数据来源:由国家气候中心根据我国1951—2000年的气候数据,应用MIROC-RegCM模式模拟数据,可得气候的情景数据20C3M(1951—2000年)。取出我国东北局部地区的1981—1990年的气候平均数据,且以gis文本格式存放数据。在1981—1990年丹顶鹤种群的繁殖地的渔网点数据基础上,以及软件ArcGIS 9.3界面之下,可得到1981—1990年丹顶鹤种群在繁殖地分布区域内关于按月平均气温、按月平均最高温度、按月平均最低温度、按月平均降水和按月平均净辐射五项气候因子数据。这些气候数据将为以下进一步研究提供基础。

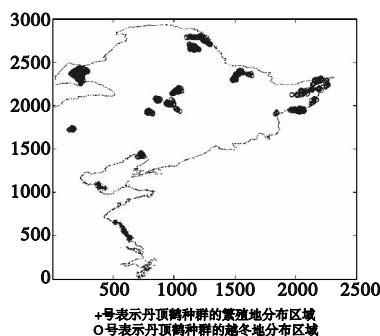


图1 1981—1990年我国丹顶鹤种群的繁殖地和越冬地分布区域

2 研究方法

研究方法是基于归一化距离的分层聚类和信息融合的技术。聚类技术,特别是模糊聚类技术在数据的一般变化趋势、数据间的相互联系与依存的规律发现及数据的结构信息的提取等方面起到了重要的作用^[28,29]。在过去的十多年中,随着信息的粒度化理论的发展^[30-33],有关信息的分层聚类和信息融合的理论与方法及其应用研究都得到了长足的发展,并形成了较为成熟的分层聚类(hierarchical clustering)和信息融合(information fusion)的技术。这里的信息融合是指:依据数据的每个属性的信息可以得到数据的一个聚类(或分类),综合数据的若干个属性的信息也可以获得数据的一个聚类(或分类),而获得这个数据的聚类(或分类)结果就称为信息的融合。以下介绍基于归一化距离的分层聚类和聚类融合理论。

2.1 归一化距离的分层聚类结构与算法

若距离 d 满足条件归一化条件(即 $\forall x, y \in X, 0 \leq d(x, y) \leq 1$),则 d 是 X 上的归一化距离。 d 诱导的粒度空间 $\mathcal{N}_{Td}(X)$ 是一个有序集,进一步粒度空间 $\mathcal{N}_{Td}(X)$ 就是归一化距离 d 的分层聚类结构,且这一聚类结构具有一致性^[33,34]。若 d 是有限

论域 $X = \{x_1, x_2, \dots, x_n\}$ 上的归一化距离, $D = \{d(x, y) \mid x, y \in X\} = \{d_0, d_1, \dots, d_m\}$,其中 $0 = d_0 < d_1 < \dots < d_m$,求 d 引导的分层聚类结构 $\mathcal{N}_{Td}(X)$ 的算法如下:

算法1^[34] 求解分层聚类结构

输入:基数为 n 的集合 X 上的一个归一化距离 d 。

输出:分层聚类结构 $\mathcal{N}_{Td}(X)$ 。

a) $i \leftarrow 0, X(d_i) = C = \{a_1, a_2, \dots, a_N\} (N \leq n)$;

b) Output $X(d_i) = C$;

c) $A \leftarrow C, i \leftarrow i + 1, C \leftarrow \emptyset$;

d) $B \leftarrow \emptyset$;

e) Taken $a_j \in A, B \leftarrow B \cup a_j, A \leftarrow A \setminus a_j$;

f) $\forall a_k \in A$, if there exists $x_j \in a_j, y_k \in a_k$ such that $d(x_j, y_k) \leq d_i$, then $B \leftarrow B \cup a_k, A \leftarrow A \setminus a_k$, otherwise goto g);

g) $C \leftarrow \{B\} \cup C$;

h) if $A \neq \emptyset$, goto d); otherwise, if $X(d_i) \neq X(d_{i-1})$, output $X(d_i) = C$;

i) if $i = m$ or $C = \{X\}$, goto j); otherwise goto c);

j) end.

2.2 基于分层聚类的归一化距离的最优聚类与最优聚类数

在有限论域 $X = \{x_1, x_2, \dots, x_n\}$ 上的归一化距离 d 的聚类结构 $\mathcal{N}_{Td}(X)$ 的基础上,引入 $d_i = (d(x_i, x_1), d(x_i, x_2), \dots, d(x_i, x_n))^T$ 表示 x_i 到 X 上所有元素的隶属函数值构成的 n 维向量($i = 1, 2, \dots, n$),记 $\bar{a} = \frac{1}{n} \sum_{i=1}^n d_i$ 表示这 n 个向量构成的形心。

对于给定的 $X(\lambda) = \{a_1, a_1, \dots, a_{C_\lambda}\} \in \mathcal{N}_{Td}(X)$,记 $k = 1, 2, \dots, C_\lambda$,则 $\sum_{i=1}^{C_\lambda} J_k = n$ 。记 $\bar{a}_k = \sum_{i=1}^{J_k} d_{k_i} / J_k, a_k = \{x_{k_1}, x_{k_2}, \dots, x_{k_{J_k}}\}$ 表示第 a_k 类的形心, $k = 1, 2, \dots, C_\lambda$ 。构造粒度 $X(\lambda)$ 的类间偏差 S_{between} 和类内偏差 S_{in} 如下:

$$\begin{aligned} S_{\text{between}} &= \sum_{i=1}^{C_\lambda} J_i \|\bar{a}_i - \bar{a}\|_2^2 / n \\ S_{\text{in}} &= \sum_{i=1}^{C_\lambda} \sum_{j=1}^{J_i} d_{ij} - \bar{a}_i \|^2 / J_i \end{aligned} \quad (1)$$

其中 $\|\cdot\|$ 表示2-范数,于是加权的总偏差为

$$\begin{aligned} S(X(\lambda)) &= \alpha S_{\text{between}} + (1 - \alpha) S_{\text{in}} = \\ &\alpha \sum_{i=1}^{C_\lambda} J_i \|\bar{a}_i - \bar{a}\|_2^2 / n + (1 - \alpha) \sum_{i=1}^{C_\lambda} \sum_{j=1}^{J_i} \|d_{ij} - \bar{a}_i\|_2^2 / J_i \end{aligned} \quad (2)$$

其中 $\alpha \in (0, 1)$ 为权系数。

从合理分类角度来说,任何一种合理的分类应体现其最大分类能力。因此基于结构聚类的优化原则就是在 $\mathcal{N}_{Td}(X)$ 中确定使 $S(X(\lambda), \alpha)$ 达到最大的聚类。按照这一原则可建立如下数学模型:

$$S(X(\lambda_0), \alpha) = \max_{X(\lambda) \in \mathcal{N}_{Td}(X)} \{S(X(\lambda), \alpha)\} \quad (3)$$

因此使 $S(X(\lambda), \alpha)$ 最大的聚类即为最优聚类,对应的聚类数即为最优聚类数。

2.3 基于归一化距离的分层聚类融合

以下通过一个引理给出基于归一化距离的分层聚类融合结论。

引理1^[34] 设 X_1 和 X_2 是 X 的两个粒度, d_1 和 d_2 分别是 X_1 和 X_2 上的归一化距离,记 $X_1 \cap X_2 = \{a \cap b \mid a \in X_1, b \in X_2\}$ 。定义模糊关系 d :

$$\forall a, b \in X_1 \cap X_2, d(a, b) = \max\{d_1(a_1, b_1), d_2(a_2, b_2)\} \quad (4)$$

其中 $a \subseteq a_i \in X_1, b \subseteq b_i \in X_2, i = 1, 2$ 。则 d 是 $X_1 \cap X_2$ 上的一个归一化距离。

在引理1中, d 也称为由 d_1 和 d_2 通过交运算而获得的,记为 $d = d_1 \cap d_2$ 。

在文献[34]中的进一步研究表明:引理 1 中通过交运算而获得的归一化距离 d 的分层聚类结构比 d_1 和 d_2 的分层聚类结构都要细。因此,从理论上解释了对于同一论域上不同的聚类(分类),通过引理 1 提供的交运算可获得更细化、更精确的聚类(分类)。

3 丹顶鹤繁殖地气候数据特征的聚类与结果分析

在 2.2 节提取的 1981—1990 年丹顶鹤繁殖地的气候因子数据集的基础上,结合归一化距离的分层聚类和信息融合理论和方法,本章将进行 1981—1990 年调查丹顶鹤在繁殖地气候数据的分层聚类研究。

在 1981—1990 年丹顶鹤繁殖地的气候因子数据集的基础上,用分区间统计的方法对气候因子数据集进行如下处理:

a)按繁殖地不同项的气候因子(即月平均气温、月平均最高温度、月平均最低温度、月平均降水和月平均净辐射五项气候因子)进行分类处理,即繁殖地的月平均气温数据(注:含 1~12 月份的全部月平均气温数据,即 $12 \times 53\ 084$ 个数据)、繁殖地的月平均最高温度数据等。记繁殖地关于气候因子 i 的分布区间为 $[a_i, b_i]$,其中 $i=1,2,3,4,5$ 分别表示气候因子月平均气温、月平均最高温度、月平均最低温度、月平均降水和月平均净辐射。经统计可得

$$\begin{aligned} [a_1, b_1] &= [-21, 25.68] \\ [a_2, b_2] &= [-9.6, 38.1] \\ [a_3, b_3] &= [-31.1, 18.2] \\ [a_4, b_4] &= [-2.7, 296.1] \\ [a_5, b_5] &= [3.4, 172.6] \end{aligned}$$

b)将繁殖地关于气候因子 i 的分布区间为 $[a_i, b_i]$ 分成等份,若记 $\Delta_i = \frac{(b_i - a_i)}{N}$,则 $[a_i, b_i]$ 的 N 等分区间分别记为 $[a_i, a_i + \Delta_i], [a_i, a_i + 2\Delta_i], \dots, [a_i, a_i + N\Delta_i]$ ($i=1,2,3,4,5$) 对繁殖地关于气候因子 i 在第 k 个月的数据进行繁殖地关于气候因子 i 的分布区间 $[a_i, b_i]$ 的 N 等分区间频数统计,可得频数向量 $(n_{ik1}, n_{ik2}, \dots, n_{ikN})$, $k=1,2, \dots, 12$, 且 $\sum_{s=1}^N n_{iks} = M_i$ 。于是繁殖地关于气候因子 i 在分布区间 $[a_i, b_i]$ 的 N 等分区间上第 k 个月的分布向量可记为

$$P_{ik} = (n_{ik1}, n_{ik2}, \dots, n_{ikN}) / M_i = (p_{ik1}, p_{ik2}, \dots, p_{ikN}) \quad (5)$$

其中 $i=1,2,3,4,5; k=1,2, \dots, 12$ 。

若取 $N=20$,对 1981—1990 年丹顶鹤的繁殖地气候因子数据集进行上述处理,可得繁殖地关于气候因子 i 在分布区间 $[a_i, b_i]$ 的 N 等分区间上第 k 个月的分布向量。例如繁殖地关于气候因子月平均温度在分布区间 $[a_1, b_1]$ 的 N 等分区间上的 12 个分布向量如下:

$$\begin{aligned} P_{11} &= (0.4333, 0.0610, 0.3223, 0.0.1835, 0.0, \\ &\quad 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0) \\ P_{12} &= (0.0, 0.3170, 0.1702, 0.3294, 0.0, 0.0, 0.0, 0.0, 0.1835, 0.0, \\ &\quad 0.0, 0.0, 0.0, 0.0, 0.0) \\ P_{13} &= (0.0, 0.0, 0.0, 0.0, 0.0761, 0.3540, 0.0571, 0.3294, 0.1835, \\ &\quad 0.0, 0.0, 0.0, 0.0, 0.0, 0.0) \\ P_{14} &= (0.0, 0.0, 0.0009, 0.4418, 0.0444, 0.3294, 0.0.1835, \\ &\quad 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0) \\ P_{15} &= (0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.3788, 0.0879, 0.0275, \\ &\quad 0.3233, 0.1835, 0.0, 0.0, 0.0, 0.0) \end{aligned}$$

$$\begin{aligned} P_{16} &= (0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0621, 0.3838, \\ &\quad 0.0428, 0.4214, 0.0843, 0.0, 0) \\ P_{17} &= (0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0542, \\ &\quad 0.4330, 0.3294, 0.1835, 0, 0) \\ P_{18} &= (0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0497, \\ &\quad 0.4268, 0.0178, 0.5058, 0.0, 0) \\ P_{19} &= (0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.2962, 0.1509, \\ &\quad 0.0400, 0.3294, 0.1835, 0.0, 0, 0, 0) \\ P_{110} &= (0.0, 0.0, 0.0, 0.0, 0.0009, 0.3921, 0.0909, 0.3326, \\ &\quad 0.0.1835, 0.0, 0.0, 0.0, 0.0, 0, 0) \\ P_{111} &= (0.0, 0.0, 0.0343, 0.4244, 0.0355, 0.3223, 0, \\ &\quad 0.1835, 0.0, 0.0, 0.0, 0.0, 0.0, 0, 0) \\ P_{112} &= (0.2329, 0.2507, 0.3315, 0.0015, 0.0038, 0.1796, \\ &\quad 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0, 0) \end{aligned}$$

类似地可计算其他分车区域。

c)在 b)的分布向量 P_{ik} 的基础上,通过内积运算获取繁殖地与气候因子 i 关于月份数据间的相关矩阵 $R_i = [r_{ikl}]_{n \times n}$, 其中:

$$r_{ikl} = \frac{(P_{ik}, P_{il})}{\sqrt{(P_{ik}, P_{ik})} \cdot \sqrt{(P_{il}, P_{il})}} = \frac{\sum_{s=1}^N p_{iks} \times p_{ils}}{\sqrt{\sum_{s=1}^N p_{iks}^2} \times \sqrt{\sum_{s=1}^N p_{ils}^2}} \quad (6)$$

由 R_i 得到 $d_i = E - R_i = [1 - r_{ikl}]_{n \times n}$ 。由于 d_i 是 12×12 维矩阵,可以通过验证 d_i 满足归一化距离的要求。于是可获得繁殖地与气候因子 i 关于月份数据间的距离矩阵 $d_i = (d_{ikl})_{12 \times 12}$, 其中 $i=1,2,3,4,5$ 。例如,利用上面的数据和式(6)可计算繁殖地的气候因子月平均温度关于 12 个月份的距离矩阵为

$$d_1 = \begin{bmatrix} 0 & & & & & & & & & & & \\ 0.3737 & 0 & & & & & & & & & & \\ 0.9838 & 0.9259 & 0 & & & & & & & & & \\ 1 & 0.9999 & 0.9625 & 0 & & & & & & & & \\ 1 & 1 & 1 & 0.9906 & 0 & & & & & & & \\ 1 & 1 & 1 & 1 & 0.9990 & 0 & & & & & & \\ 1 & 1 & 1 & 1 & 1 & 0.6740 & 0 & & & & & \\ 1 & 1 & 1 & 1 & 0.9997 & 0.0143 & 0.5226 & 0 & & & & \\ 1 & 1 & 1 & 0.9998 & 0.2359 & 0.9644 & 0.9999 & 0.9839 & 0 & & & \\ 1 & 0.9999 & 0.9589 & 0.0038 & 0.9901 & 1 & 1 & 1 & 0.9998 & 0 & & \\ 0.9598 & 0.6715 & 0.1936 & 0.9743 & 1 & 1 & 1 & 1 & 1 & 0.9728 & 0 & \\ 0.0615 & 0.1669 & 0.9795 & 0.9999 & 1 & 1 & 1 & 1 & 1 & 0.9999 & 0.9039 & 0 \end{bmatrix}$$

由于距离矩阵 d_i 都是对称矩阵,因此上面相关矩阵的表述略去了对角线上方的元素。类似地可计算其他气候因子。

在以上获得的关于月份数据间的相关矩阵的基础上,应用归一化距离的聚类结构提取的算法—算法 1,进行气候因子关于月份的分层聚类结构的提取与分析研究。以下以繁殖地的月平均温度的 12 个月份为例,进行分层聚类结构的提取与最优聚类数确定的研究。

记 $X = \{1, 2, \dots, 12\}$, $D_1 = \{d_{ikl} = d_1(k, l) \mid k, l \in X\}$ 。于是,应用算法 1 可得 d_1 引导的分层聚类结构 $N_{Td_1}(X)$ 如下:

$$\begin{aligned} X_1(0.0000) &= \{\{1\}, \{2\}, \dots, \{12\}\}; \\ X_1(0.0038) &= \{\{1\}, \{2\}, \{3\}, \{4, 10\}, \{5\}, \dots, \{11\}, \{12\}\}; \\ X_1(0.0143) &= \{\{1\}, \{2\}, \{3\}, \{4, 10\}, \{5\}, \{6, 8\}, \{7\}, \{9\}, \\ &\quad \{11\}, \{12\}\}; \\ X_1(0.0615) &= \{\{1, 12\}, \{2\}, \{3\}, \{4, 10\}, \{5\}, \{6, 8\}, \{7\}, \\ &\quad \{9\}, \{11\}\}; \\ X_1(0.1669) &= \{\{1, 2, 12\}, \{3\}, \{4, 10\}, \{5\}, \{6, 8\}, \{7\}, \{9\}, \\ &\quad \{11\}\}; \\ X_1(0.1936) &= \{\{1, 2, 12\}, \{3, 11\}, \{4, 10\}, \{5\}, \{6, 8\}, \{7\}, \\ &\quad \{9\}\}; \end{aligned}$$

$$\begin{aligned}
X_1(0.2359) &= \{ \{1,2,12\}, \{3,11\}, \{4,10\}, \{5,9\}, \{6,7,8\} \}; \\
X_1(0.5226) &= \{ \{1,2,12\}, \{3,11\}, \{4,10\}, \{5,9\}, \{6,7,8\} \}; \\
X_1(0.6715) &= \{ \{1,2,3,11,12\}, \{4,10\}, \{5,9\}, \{6,7,8\} \}; \\
X_1(0.9589) &= \{ \{1,2,3,4,10,11,12\}, \{5,9\}, \{6,7,8\} \}; \\
X_1(0.9644) &= \{ \{1,2,3,4,10,11,12\}, \{5,6,7,8,9\} \}; \\
X_1(0.9901) &= \{ \{1,2,\dots,12\} \} = X_0.
\end{aligned}$$

以下无特别说明的话, $\alpha = 0.5$ 。通过式(2),可得加权总偏差 $S(X_1(\lambda), \alpha)$ 与聚类数的对应值如表1所示。依据模型式(3),可得最优聚类数为2,且最优聚类为 $X_1(0.9644)$,即繁殖地的月平均温度关于1~12月可分为 $\{1,2,3,4,10,11,12\}$ 和 $\{5,6,7,8,9\}$ 两类。

表1 $S(X_1(\lambda), \alpha)$ 与聚类数的对应值

聚类数	$S(X_1(\lambda), \alpha)$	聚类数	$S(X_1(\lambda), \alpha)$
12	0.770	6	0.849
11	0.771	5	0.901
10	0.773	4	0.963
9	0.773	3	1.243
8	0.774	2	1.248
7	0.791	1	0.770

类似上面的讨论,可得到其他距离矩阵所对应的分层聚类结构 $\mathcal{N}_{m_i}(X)$ 及相应的最优聚类数。从所有相关矩阵所对应的最优聚类及相应的最优聚类数的结果来看:除繁殖地月平均最低温度和月平均降水量所对应的距离矩阵外,其他的距离矩阵所对应的最优聚类数都是2,且最优聚类都是 $\{ \{1,2,3,4,10,11,12\}, \{5,6,7,8,9\} \}$ 。同时这些不相同的结果分别是:

a) 繁殖地关于月平均最低温度的距离矩阵所对应的最优聚类数是2,且最优聚类是 $X_3(0.9379) = \{ \{1,2,3,4,5,9,10,11,12\}, \{6,7,8\} \}$ 。

b) 繁殖地关于月平均降水量的距离矩阵所对应的最优聚类数是3,最优聚类是 $X_4(0.4374) = \{ \{1,2,3,11,12\}, \{4,9,10\}, \{5,6,7,8\} \}$ 。

以下在引理1的基础上,引入基于归一化距离的聚类融合方法,对这两种具有不同聚类结果的情形分别处理。

关于月平均最低温度的情形。由于其最优聚类数为2,且聚类数为3的聚类结构 $X_3(0.7854) = \{ \{1,2,3,4,10,11,12\}, \{5,9\}, \{6,7,8\} \}$ 与一致的最优聚类结果较为接近,于是选取粒度 $X_3(0.7854)$ 作为一个研究对象。同时,由于繁殖地关于月平均温度的距离矩阵所对应的最优聚类数是2,且最优聚类是一致的最优聚类结果,因此可选取繁殖地的月平均温度气候因子所对应的聚类数为3的聚类结构 $X_1(0.9589) = \{ \{1,2,3,4,10,11,12\}, \{5,9\}, \{6,7,8\} \}$ 作为另一个研究对象。以下依据2.3节对聚类 $X_3(0.7854)$ 和 $X_1(0.9589)$ 进行融合。

记 $X^1 = X_3(0.7854) \cap X_1(0.9589) = \{ C_1^1, C_2^1, C_3^1 \}$, 其中 $C_1^1 = \{1,2,3,4,10,11,12\}$, $C_2^1 = \{5,9\}$, $C_3^1 = \{6,7,8\}$ 。于是, d_1 和 d_2 在粒度 X^1 上的归一化距离分别为^[33]

$$D_1^1 = \begin{bmatrix} 0 & & \\ 0.9379 & 0 & \\ 1 & 0.9592 & 0 \end{bmatrix}, D_2^1 = \begin{bmatrix} 0 & & \\ 0.9901 & 0 & \\ 1 & 0.9644 & 0 \end{bmatrix}$$

从而通过2.3节所给的交运算可得

$$D^1 = D_1^1 \cap D_2^1 = \begin{bmatrix} 0 & & \\ 0.9901 & 0 & \\ 1 & 0.9644 & 0 \end{bmatrix}$$

其中: D^1 是 X^1 上的一个归一化距离。而 D^1 在 X^1 上引导的分层聚类结构 $\mathcal{N}_{m^1}(X^1)$ 为

$$\begin{aligned}
X^1(0) &= \{ \{1,2,3,4,10,11,12\}, \{5,9\}, \{6,7,8\} \}; \\
X^1(0.0.9644) &= \{ \{1,2,3,4,10,11,12\}, \{5,6,7,8,9\} \}; \\
X^1(0.9901) &= \{ \{1,2,3,4,5,6,7,8,9,10,11,12\} \}.
\end{aligned}$$

取 $\alpha = 0.5$, 则加权总偏差 $S(X^1(\lambda), \alpha)$ 与聚类数的对应值如表2所示。

表2 $S(X^1(\lambda), \alpha)$ 与聚类数的对应表

聚类数	3	2	1
$S(X^1(\lambda), \alpha)$	0.324	0.401	0.324

依据模型式(3),可得最优聚类数为2,且最优聚类为 $X^1(0.4010)$,即融合繁殖地的气候因子月平均最低温度和月平均温度可将1~12月份分为 $\{1,2,3,4,10,11,12\}$ 和 $\{5,6,7,8,9\}$ 两类。类似地处理繁殖地月平均降雨量,可得最优聚类数为2,且最优聚类为 $X^2(0.9644) = \{ \{1,2,3,4,10,11,12\}, \{5,6,7,8,9\} \}$ 。

若同一物种的两个气候因子的最优聚类相同,则这两个分层聚类融合所获得的最优聚类结构不变。因此,融合我国东北地区的丹顶鹤繁殖地的月平均气温、月平均最高温度、月平均最低温度、月平均降水和月平均净辐射五个气候因子可获得它们一致的最优聚类为 $\{1,2,3,4,10,11,12\}$ 和 $\{5,6,7,8,9\}$ 。这一结果可解释为:丹顶鹤在东北繁殖地最适宜栖息停留的时间为5~9月份,4和10月份为迁徙所用月份,而1~3,11,12月份为丹顶鹤在越冬地适宜栖息的月份。这一结论与东北地区丹顶鹤的调查和研究的数据相吻合^[23~27]。

4 结束语

本文从数据挖掘的角度出发,在1981—1990年我国东北地区的典型候鸟物种丹顶鹤的调查分布区域内的月平均气温、月平均最高温度、月平均最低温度、月平均降水和月平均净辐射五个气候因子数据的基础上,采用分区间统计及基于归一化距离的分层聚类 and 聚类融合等理论和方法进行相关气候数据的特征分析,获得的结果与东北地区丹顶鹤的调查和研究的数据相吻合。与以往文献中用于物种分布预测中的气候因子提取方法不同的是,本文采用的方法完全依赖于预测物种在调查分布区内的气候数据,通过数据挖掘与数据处理而获得,而不是通过预测物种的生物学与生态学特征及其在分布区内的气候因子相关性分析得到。这一方法具有一般性。

接下来的研究工作,将在本文的研究基础上,通过东北地区候鸟物种丹顶鹤繁殖地与越冬地的气候数据分析,提取用于它们分布预测的气候因子,进行未来气候变化对东北地区候鸟物种分布预测影响的研究。

参考文献:

- [1] IPCC. Climate change 2007: the physical science basis-summary for policymakers of the working group I report [M]. Cambridge: Cambridge University Press, 2007.
- [2] IPCC. Climate change 2001: impacts, adaptation, and vulnerability-contribution of working group II to the forth assessment report of the intergovernmental panel on climate change [M]. Cambridge: Cambridge University Press, 2001.
- [3] IPCC. Climate change 2007: impacts, adaptation, and vulnerability-contribution of working group II to the forth assessment report of the

- intergovernmental panel on climate change [M]. Cambridge: Cambridge University Press, 2007.
- [4] 方精云. 全球生态学: 气候变化与生态响应[M]. 北京: Heidelberg: 高教出版社-Springer-Verlag 联合出版, 2000.
- [5] 徐汝梅. 生物入侵—数据集成、数量分析与预警[M]. 北京: 科学出版社, 2003.
- [6] 刘国华, 傅伯杰. 全球气候变化对森林生态系统的影响[J]. 自然资源学报, 2001, 16(1): 71-78.
- [7] WALKER B H, STEFFEN W. An overview of implications of global change for natural and managed terrestrial ecosystem [J]. *Conservation Ecology*, 1997, 1(2): art 2.
- [8] HOLMES T P, AUKEMA J E, Von HOLLE B, *et al.* Economic impacts of invasive species in forests: past, present and future[J]. *New York Academy of Sciences*, 2009, 1162: 18-38.
- [9] 吴军, 徐海根, 陈炼. 气候变化对物种影响研究综述[J]. 生态与农业环境学报, 2011, 27(4): 1-6.
- [10] PARMESAN C, YOHE G. A globally coherent fingerprint of climate change impacts across natural systems [J]. *Nature*, 2003, 421(6918): 37-42.
- [11] FORSMAN J T, MÖNKÖÖNEN M. The role of climate in limiting European resident bird populations [J]. *Journal of Biogeography*, 2003, 30(11): 55-70.
- [12] WESTPHAL M I, FIELD S A, TYRE A J, *et al.* Effects of landscape pattern on bird species distribution in the Mt. Lofty Ranges, South Australia [J]. *Landscape Ecology*, 2003, 18(4): 413-426.
- [13] ERASMUS B F N, VANJAARSVELD S A, CHOWN L S, *et al.* Vulnerability of South African animal taxa to climate change [J]. *Global Change Biology*, 2002, 8(7): 679-693.
- [14] FORSMAN J T, MÖNKÖÖNEN M. The role of climate in limiting European resident bird populations [J]. *Journal of Biogeography*, 2003, 30(1): 55-70.
- [15] LUOTO M, PÖYRY J, HEIKKINEN R K, *et al.* Uncertainty of bioclimatic envelope models based on the geographical distribution of species[J]. *Global Ecology and Biogeography*, 2005, 14(6): 575-584.
- [16] 张雷, 刘世荣, 孙鹏森, 等. 气候变化对物种分布影响模拟中的不确定性组分分割与制图—以油松为例[J]. 生态学报, 2011, 31(19): 5749-5761.
- [17] LOISELLE B A, JORGENSEN P M, CONSIGLIO T, *et al.* Predicting species distributions from herbarium collections: does climate bias in collection sampling influence model outcomes[J]. *Journal of Biogeography*, 2008, 35(1): 105-116.
- [18] 吴伟伟, 顾莎莎, 吴军, 等. 气候变化对我国丹顶鹤繁殖地分布的影响[J]. 生态与农业环境学报, 2012, 28(3): 243-248.
- [19] 徐海根, 王连龙, 包浩生. 我国丹顶鹤自然保护区网络设计[J]. 农村生态环境, 2003, 19(4): 5-9.
- [20] 王志强, 傅建春. 扎龙湿地丹顶鹤巢址空间分布变化及其对环境变化指征[J]. 生态环境学报, 2010, 19(3): 697-700.
- [21] 于文阁, 朱宝光. 洪河自然保护区丹顶鹤秋季觅食生境特征[J]. 东北林业大学学报, 2008, 36(10): 36-37.
- [22] 张曼胤. 江苏盐城滨海湿地景观变化及其对丹顶鹤生境的影响[D]. 长春: 东北师范大学, 2008.
- [23] 李金录, 程彩云. 丹顶鹤繁殖分布及研究[J]. 野生动物杂志, 1987, 8(2): 11-14.
- [24] 马逸清, 李晓民. 我国丹顶鹤资源的现状[J]. 国土与自然资源研究, 1990, 11(1): 62-64.
- [25] 严凤涛. 越冬区丹顶鹤动态分布[J]. 国土与自然资源研究, 1988, 9(1): 65-69.
- [26] 严凤涛. 盐城滩涂丹顶鹤越冬数量分布与生态研究[J]. 动物学杂志, 1991, 26(2): 34-36.
- [27] 冯科民, 李金录. 丹顶鹤的繁殖生态[J]. 东北林业大学学报, 1986, 14(3): 39-45.
- [28] FREY B J, DUECK D. Clustering by passing messages between data points[J]. *Science*, 2007, 315(5814): 972-976.
- [29] PEDRYCZ W. Collaborative fuzzy clustering[J]. *Pattern Recognition Letters*, 2002, 23(14): 1675-1686.
- [30] ZADEH L A. The roles of fuzzy logic and soft computing in the conception, design and deployment of intelligent systems [C]//Lecture Notes in Computer Science, vol 1198. 1997: 181-190.
- [31] ZHANG Ling, ZHANG Bo. The structure analysis of fuzzy sets[J]. *International Journal of Approximate Reasoning*, 2005, 40(1-2): 92-108.
- [32] 张铃, 张钊. 问题求解理论与方法—商空间粒度计算理论及应用[M]. 2版. 北京: 清华大学出版社, 2007.
- [33] 唐旭清, 朱平, 程家兴. 基于等腰归一化距离的模糊粒度空间研究[J]. 计算机科学, 2008, 35(4): 142-145.
- [34] TANG Xu-qing, ZHU Ping, CHENG Jia-xing. The structural clustering and analysis of metric based on granular space[J]. *Pattern Recognition*, 2010, 43(11): 3768-3786.

(上接第746页)

- [2] GLEISER P, DANON L. Community structure in jazz [J]. *Advances in Complex Systems*, 2003, 6(4): 565-573.
- [3] 汪小帆, 李翔, 陈关荣. 复杂网络理论及其应用[M]. 北京: 清华大学出版社, 2006.
- [4] FORTUNATO S. Community detection in graphs [J]. *Physics Reports*, 2010, 486(3-5): 75-174.
- [5] 袁超, 柴毅. 基于簇相似度的网络社团结构探测算法[J]. 物理学报, 2012, 61(21): 218901.
- [6] 罗家德. 社会网分析讲义[M]. 2版. 北京: 社会科学文献出版社, 2010.
- [7] 武澎, 王恒山. 突发事件信息传播超网络建模及重要节点判定[J]. 情报学报, 2012, 31(7): 722-729.
- [8] 邓智龙, 涂文燕. 复杂网络中的社团结构发现方法[J]. 计算机科学, 2012, 39(6): 103-108.
- [9] NEWMAN M E J, GIRVAN M. Finding and evaluating community structure in networks [J]. *Physical Review E*, 2004, 69(2): 026113.
- [10] DONGEN S. Graph clustering by flow simulation [D]. Dutch: University of Utrecht, 2000.
- [11] 刘臣. 组织内部知识网络的结构及知识共享博弈研究[D]. 哈尔滨: 哈尔滨工业大学, 2010.
- [12] 宋歌. SNA与MSA在揭示知识结构中的比较研究[J]. 图书情报工作, 2009, 53(8): 106-109, 121.
- [13] ZACHARY W. An information flow model for conflict and fission in small groups[J]. *Journal of Anthropological Research*, 1977, 33(4): 452-473.
- [14] 涂文燕, 赫南, 李德毅, 等. 一种基于拓扑势的网络社区发现方法[J]. 软件学报, 2009, 20(8): 2241-2254.
- [15] 杨博, 刘杰, 刘大有. 基于随机网络集成模型的广义网络社区挖掘算法[J]. 自动化学报, 2012, 38(5): 812-822.
- [16] 黄发良, 张师超, 朱晓峰. 基于多目标优化的网络社区发现方法[J]. 软件学报, 2013, 24(9): 2062-2077.