计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2017,26(9):10−15 [doi: 10.15888/j.cnki.csa.005961]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于互信息的生态位因子分析方法①

吕旭红 [1,2], 罗　泽 [2]

[1](中国科学院大学, 北京 100049)
[2](中国科学院 计算机网络信息中心, 北京 100190)

摘　要: 生态位因子分析方法是一种基于生态位概念的多变量分析方法, 然而该方法在计算相关性时所使用的协方差只考虑了变量间的线性关系, 而大部分变量间的关系是非线性相关的. 互信息可用于衡量两个变量间相互依赖的强弱程度, 且不局限于线性相关. 本文提出基于互信息的生态位因子分析方法, 采用互信息计算变量间的相关性, 分析斑头雁在青海湖地区的栖息地选择情况以及栖息地适宜性, 与传统生态位因子分析方法相比, 所提出的方法改变了特化向量, 提高了栖息地适宜性预测的准确率.

关键词: 互信息; 生态位因子分析方法; 栖息地选择; 栖息地适宜性

引用格式: 吕旭红,罗泽.基于互信息的生态位因子分析方法.计算机系统应用,2017,26(9):10–15. http://www.c-s-a.org.cn/1003-3254/5961.html

Ecological Niche Factor Analysis Based on Mutual Information

LV Xu-Hong[1,2], LUO Ze[2]

[1](University of Cinese Academy of Sciences, Beijing 100049, China)
[2](Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: Ecological-Niche Factor Analysis (ENFA) is a multivariable approach based on the concept of the ecological niche. But when computing the relevance between variables by covariance, it only handles linear dependencies, while most is nonlinear interaction. Mutual information measures the interdependence between variables and it's not limited to linear relations. ENFA based on mutual information (MIENFA) is presented which uses mutual information as the relevance. Through studies of Bar-headed Goose in Qinghai Lake, compared with the traditional ENFA, the proposed approach changes the specialization vector and improves the accurate rate of habitat suitability prediction.

Key words: mutual information; Ecological-Niche Factor Analysis (ENFA); habitat selection; habitat suitability

栖息地指的是能够为物种生存或繁殖使用的所有环境因素总和. 栖息地的好坏能影响到物种分布、种群密度、繁殖成功率及存活率, 因此, 动物对其栖息地具有一定的选择性. 栖息地选择指物种对不同栖息地产生不同反应的过程, 导致不成比例地使用栖息地, 从而影响物种或个体的生存和适合度[1]. 栖息地选择研究一直以来是动物生态学的研究热点, 与种群生态学、群落生态学等生态学分支有着密切的关系, 同时它也是开展珍稀濒危动物研究及生态多样性保护的基础, 并为动物保护措施的制定提供了重要的直接依据[2-4].

按照计算所需物种活动痕迹的不同, 可将栖息地选择模型分为三种: presence-absence 模型、enhanced presence-only 模型以及 simple presence-only 模型. 其中, simple presence-only 模型通过分析已知的物种出现地点的生态环境特点, 总结出统计规律, 主要包括 BIOCLIM、DOMAIN 等. 该模型计算过程简单且易于理解, 但预测准确度不高, 且只支持连续性数据输入; enhanced presence-only 模型只需要物种"出现"点的数据, 考虑环境变量及其相关性, 主要是基于生态位思想研究物种栖息地选择. 主要包括 ENFA、MADIFA、

① 基金项目: 中美软件合作研究项目(61361126011)
收稿时间: 2016-12-28; 采用时间: 2017-01-20

FANTER 等. 该模型支持连续和离散型数据, 能更好的描述物种分布, 但依赖于“出现”数据的可靠性; presence-absence 模型需要物种“出现”点和“非出现”点数据, 将问题转换为预测是否会出现, 主要包括 Logistic 回归、神经网络等. 该模型具有较高鲁棒性, 但依赖于“非出现”数据的可靠性[5,6].

Presence-absence 模型主要使用机器学习相关的算法, 具有较高鲁棒性, 但是缺乏“非出现”数据是生态学研究的主要问题, “非出现”数据通常难以精确获得. 误判的“非出现”数据的存在一定程度上会给分析带来偏差, 故通常考虑 presence-only 方法, 而生态位因子分析方法是比较常用[7]. 然而该方法在计算相关性时所使用的协方差只考虑了变量间的线性关系, 而大部分变量间的关系是非线性相关的.

针对该问题, 考虑采用互信息计算变量间的相关性, 提出基于互信息的生态位因子分析方法. 互信息在衡量变量间的相关性时不局限于线性相关. 通过分析斑头雁在青海湖地区的栖息地选择情况以及栖息地适宜性, 对所提出的方法和传统生态位因子分析方法进行比较.

1 生态位因子分析方法

生态位(Ecological Niche)的思想提供了一个只依赖“出现”数据的栖息地选择方法. 生态位指的是由 n 个环境变量构成的 n 维生态空间下的超体积, 在超体积中的点所构成的生态环境表示能够使物种无限生存的环境[7,8]. 如图 1(a)所示, 研究区域对应于图中的可利用空间(Available Space), 生物出现的区域对应于图中的利用空间(Used Space), 即生态位.

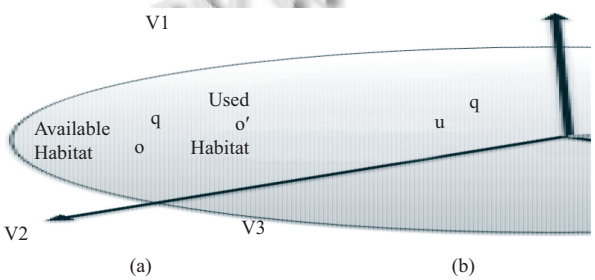

图 1 可视化显示生态位、边缘向量以及特化向量

生态位因子分析(Ecological Niche Factor Analysis, ENFA)方法是建立在 Hutchinson 生态位基础上的一种用于研究物种地理分布的多变量分析方法, 它假设物种在多种环境条件下不是随机分布的. 在多维空间下比较物种利用分布与可利用分布的差异性, 利用主成分分析方法提取出一套新的因子, 这些因子具有两层生态学含义, 一个是边缘性(Marginality), 指物种利用空间均值和可利用空间均值的差异, 值为正, 则说明该物种在该环境变量上偏好于该生态因子平均水平以上的环境; 为负则偏好于该生态位因子平均水平以下的环境, 绝对值越大, 偏好程度越高; 另一个是特化性(Specialization), 指在整个研究区域背景下, 物种生态位特化的程度, 值越大说明物种的生态位宽度越小, 越无法忍受该环境变量的变化[4,7-9].

2 基于互信息的生态位因子分析方法

2.1 互信息

互信息衡量两个变量间相互依赖的程度, 表示两个变量间共同拥有信息的含量[8]. 给定两个随机变量 X 和 Y, 若它们各自的边缘概率分布和联合概率分布分别为 $p(x)$, $p(y)$和 $p(x,y)$, 则它们之间的互信息 $I(X;Y)$定义为:

$$I(X;Y)=\sum_{x}\sum_{y}p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{1}$$

当变量 X 和 Y 完全无关或相互独立时, 互信息最小, 结果为 0.

这意味着两个变量之间不存在重叠的信息; 反之, 两者的相互依赖程度越高, 互信息的值越大, 所包含的相同信息也越多[10].

2.2 基于互信息的生态位因子分析方法

设 Z 为 N*V 的矩阵, 表示 V 个环境变量在 N 个栅格单元上的值, 定义了 V 维生态空间上的点集(可利用空间). S 为 N_S*V 的矩阵, 表示生物在 N_S 个栅格单元上出现所对应的 V 个环境变量的值, 定义了 V 维生态空间上的点集(利用空间). 为了消除变量之间由于量纲不同造成的差异, Z 和 S 已进行数据标准化预处理.

在 ENFA 方法中, 寻找一个单位向量μ, 使得可利用栖息分布投影到该向量后的方差($\mu^{\mathrm{T}}R_G\mu$)与利用栖息分布投影后的方差($\mu^{\mathrm{T}}R_s\mu$)比值最大, 同时该向量与边缘化向量 q 正交, 如图 1(b), μ即特化向量. 将问题公式化为:

$$f(\mu)=\arg\max_{\mu}\frac{\mu^{T}R_G\mu}{\mu^{T}R_S\mu} \tag{2}$$

$$\mu^T q = 0 \tag{3}$$

$$\mu^T \mu = 1 \tag{4}$$

$$R_G = \frac{1}{N} Z^T Z \tag{5}$$

$$R_S = \frac{1}{N_S - 1} S^T S \tag{6}$$

可以看出在计算可利用栖息地与利用栖息地方差时使用协方差反映变量间的相关性, 但它只能反映变量间的线性关系, 无法衡量变量间的非线性关系, 而互信息从信息论的角度出发, 评估变量间共有信息量, 不局限于线性关系, 与协方差相比有很大优势. 从公式(2-4)可以看出, 求解特化向量的过程相当于求解带约束的主成分分析的过程, 文献[10]中已给出了在主成分分析中将互信息替代协方差的可行性解释, 且基于互信息的主成分分析方法能够提高分类精度, 因此, 考虑在生态位因子分析中用互信息替代协方差, 提出一种基于互信息的生态位因子分析(ENFA based on mutual information, MIENFA)方法. 将式(5)、(6)分别改写为:

$$R_G = I(Z, Z) \tag{7}$$

$$R_S = I(S, S) \tag{8}$$

其中, 两个矩阵的对角线元素为变量的自信息, 非对角线元素为两个变量之间的互信息. 无论互信息或自信息均为实数, 当两个变量之间不相关时, 互信息为 0, 否则为正数, 因此矩阵为非负实数阵. 同时, 互信息满足 $I(X, Y)=I(Y, X)$, 可得矩阵为非负实数对称阵. 令:

$$H = (I_v - yy^T) W (I_v - yy^T) \tag{9}$$

$$y = \frac{z}{\sqrt{z^T z}}, z = R_G^{\frac{1}{2}} q, W = R_S^{-\frac{1}{2}} R_G R_S^{-\frac{1}{2}} \tag{10}$$

矩阵 R_G 和 R_S 为非负实数对称阵, 则式(9)有解且与矩阵 H 的特征向量v_i 相关. MIENFA 算法伪代码如表 1 所示.

3 实验

3.1 数据源及数据处理

3.1.1 斑头雁轨迹数据

使用 2007-2008 年斑头雁的轨迹数据, 原始数据 471774 条, 共 29 只斑头雁. 如表 2 所示, 数据记录主要包括以下几个字段. 字段 animal 表示被跟踪鸟类的唯一编号; record_id 表示数据获取的类型, LATEST ARGOS LOCS 表示使用 Argos 系统进行定位, LATEST GPS LOCS 表示使用 GPS 进行定位; latitude 和 longitude 分别表示经度和纬度; lc94 用来标记数据的卫星等级, 使用 GPS 进行定位的数据的级别为 LG, 使用 Argos 系统进行定位的数据等级分为 7 个, 按照准确度增大的顺序分别为 LZ、LB、LA、L0、L1、L2 和 L3; datetime 表示获取到数据的时间[11].

表 1 MIENFA 算法伪代码

MIENFA算法
Input: the number of cells in the research map N; The number of environmental variables V; The number of presence data Ns; The matrix of available habitat Z(N*V); The matrix of used habitat S(Ns*V);
Output: one marginal vector, (V) specialized vectors and eigenvalues;
Normalize Z and S
For j=1 to V do
Z[, j]=discretize(Z[, j])
End
Discretize S according to Z
Compute mutual information matrix of Z and S: R_G, R_S
#compute the marginality:
m=[]
For i=1 to Ns do
m[i]=0
for j=1 to V do
m[i]=m[i]+S[i, j]
End
m[i]=m[i]/Ns
End
Normalize m
compute the specialization:
Set $z=R_G^{\frac{1}{2}} q$, $y=\frac{z}{\sqrt{z^T z}}$, $W=R_S^{\frac{1}{2}} R_G R_S^{-\frac{1}{2}}$, $H=(I_v-yy^T)W(I_v-yy^T)$
Compute the eigenvalues $\lambda_1,\lambda_2,...,\lambda_V$ and eigenvectors $\{v_1, v_2, ..., v_V\}$ of H
For i=1 to V do
$\mu_i=\frac{R_S^{-1/2} v_i}{\sqrt{v_i^T R_S^{-1} v_i}}$
End
Return $(q,\{\mu_1,\mu_2,...,\mu_V\},\{\lambda_1,\lambda_2,...,\lambda_V\})$
End

根据文献[11], 一般使用 LG, L1-L3 这四种精度的数据进行分析, 过滤后数据量为 103486 条. 青海湖附近区域地理类型丰富, 且斑头雁的记录数据较多, 故实验选取青海湖区域(96°E, 101°E, 34°N, 38°N)作为研究区域, 该区域内的数据量为 61799 条.

表 2 斑头雁轨迹数据格式

animal	record_id	latitude	longitude	lc94	datetime
BH07_67580	LATEST ARGOS LOCS	36.726	99.82	L0	2007-06-19 03:52:44
BH07_67693	LATEST ARGOS LOCS	47.957	99.841	L1	2007-05-24 22:48:36
BH07_67582	LATEST ARGOS LOCS	37.09	99.666	L2	2008-05-11 18:45:03
BH07_67582	LATEST ARGOS LOCS	37.09	99.668	L3	2008-05-09 21:56:05
BH07_67580	LATEST ARGOS LOCS	36.744	99.842	LA	2007-06-23 06:50:29
BH07_67580	LATEST ARGOS LOCS	36.132	98.805	LB	2007-06-23 15:16:04
BH07_67580	LATEST GPS LOCS	36.73	99.804	LG	2007-06-20 23:00:00
BH07_67690	LATEST ARGOS LOCS	29.291	91.659	LZ	2008-01-04 23:31:39

3.1.2 环境变量

使用中国地区土地覆盖综合数据集, 网格数目为 4857×4045, 分辨率为 1 km, 不同环境变量对应不同的值, 值域为 1 到 22 的整数值, 如图 2 所示[12]. 表 3 为研究区域内的环境变量以及比重(X*i* 表示变量在土地覆盖中的值为 *i*). 其中, X1、X2、X10、X11、X14、X15、X17、X18、X21 所占的比重较小(<=5%), 考虑使用研究区域内各个像素与这些环境变量的最近距离替换, 以丰富这些变量的变化. 研究区域内环境变量的分布情况如图 3 所示.

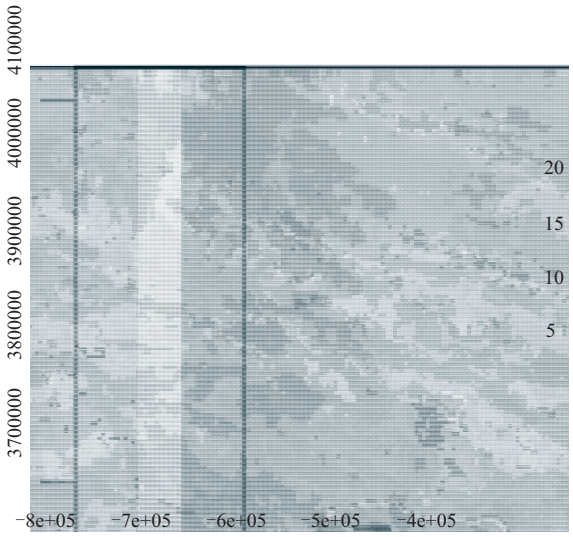

图 2 研究区域内土地覆盖数据

3.2 实验分析

首先, 进行栖息地选择, 得到边缘因子与特化因子, 从而定性分析出影响栖息地选择的主要环境变量. 为了定量比较两个算法的性能, 利用边缘因子和特化因子, 计算栖息地适宜性. 故主要进行两个实验, 首先, 分别使用 ENFA 和 MIENFA 分析斑头雁栖息地选择情况, 然后, 使用交叉检验方法分析和比较 ENFA、MIENFA 两种方法在栖息地适宜性的预测结果, 最后给出斑头雁栖息地适宜图.

表 3 研究区域内的环境变量及比重

变量	定义	中文	比重(%)
X1	needleleaved deciduoud forest	针叶阔叶林	0.16
X2	needleleaved evergreen forest	针叶常绿阔叶林	0.03
X8	alpine and sub_alpine meadow	高山和亚高山草甸	45.89
X10	plain grassland	平原草原	5.00
X11	desert grassland	荒漠草原	5.00
X12	meadow	草地; 牧场;	12.51
X14	river	河	0.0066
X15	lake	湖	2.47
X17	glacier	冰川	0.41
X18	bare rocks	裸露岩石	4.35
X19	gravels	砾石; 沙砾, 碎石	7.77
X20	desert	沙漠	6.67
X21	farmland	农田	0.36
X22	Alpine and sub-alpine plain grassland	高寒山地草原	9.3734

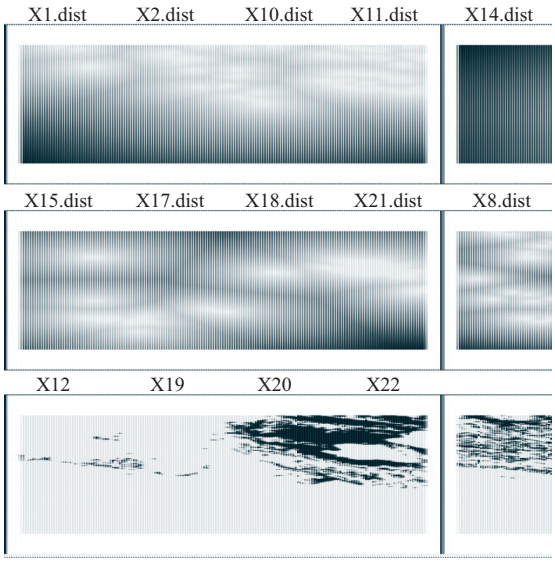

图 3 研究区域内环境变量的分布图

3.2.1 斑头雁栖息地选择

利用 R 中 adehabitatHS 包[13]的 enfa 方法以及自己编写的 mienfa 方法分析研究区域内斑头雁轨迹数据与环境变量的关系, 得到边缘向量和特化向量, 如表 4 所示, 两个方法的边缘向量相同, 特化向量不同.

表 4 ENFA 和 MIENFA 的边缘因子、第一个特化因子

variables	marginality	specialization 1	
		ENFA	MIENFA
X1.dist	-0.253	**0.301**	-0.00353
X2.dist	-0.125	-0.140	0.00345
X10.dist	-0.196	-0.123	-0.00348
X11.dist	-0.182	-0.139	0.000401
X14.dist	0.0572	0.0524	0.00518
X15.dist	**-0.493**	**0.183**	-0.0117
X17.dist	**-0.424**	0.0345	-0.000869
X18.dist	-0.206	0.0202	0.000165
X21.dist	-0.208	-0.159	-0.00445
X8	-0.113	-0.0171	-0.0378
X12	**0.539**	-0.00847	**0.0589**
X19	-0.118	0.123	**-0.449**
X20	-0.111	**-0.881**	**0.890**
X22	-0.0785	-0.00785	-0.0230

根据边缘向量对比斑头雁对不同环境变量的偏好, 可得出, 在研究区域内, 斑头雁主要选择在靠近湖、冰川以及牧场较多的地区栖息.

比较特化向量, ENFA 认为斑头雁无法忍受沙漠、与湖距离、与针叶阔叶林距离的变化, 而 MIENFA 认为斑头雁无法忍受沙漠、沙砾、牧场的变化.

3.2.2 斑头雁栖息地适宜图

利用 adehabitatHS 包[11]提供的 predict 方法计算得到栖息地适宜图(Habitat Suitability Mapping), 该方法先将多维空间中的点映射到边缘-特化低维空间上, 再计算每个点到中心的曼哈顿距离作为该点的栖息地适宜值, 距离越小, 则越适宜栖息. ENFA 和 MIENFA 得到的数据是连续的, 且范围不同. 为了方便比较, 进行数据标签化. 参照 FAO(Food and Agriculture Organization of United Nations)的土地适宜性分级标准, 采用 K-means 的离散化方法将栖息地适宜性分为 3 个等级: 适宜栖息、次适宜栖息和不适宜栖息[14,15].

采用交叉检验方法, 对两个模型进行检验: 先将斑头雁活动轨迹点均分成 10 份, 选择其中 9 份作为训练集用于生成栖息地适宜图, 剩下 1 份作为测试集用于计算模型精确度. 将“适宜栖息”作为正确分类点, 设为 1, 其他设为 0. 重复上述检验过程 10 次以保证每份活动轨迹点都参与模型精度计算. 采用分类混淆矩阵计算平均准确率($AVG_{Accuracy}$).

生态位因子分析方法最终生成一个边缘向量和多个特化向量, 特化向量的贡献率呈递减. 分别选择累积贡献率在 70-80%、80-90%、90% 以上的特化向量(表 5), 计算栖息地适宜值. 由图 4 可知, 在训练集上, 累积贡献率相近时, MIENFA 的准确率更高, 且随着累积贡献率的增加, 准确率增大. 相比于 ENFA, MIENFA 的准确率平均提高了 26.07%. 在测试集上, 两个方法的表现都没有训练集好, 但 MIENFA 还是优于 ENFA.

表 5 MIENFA 和 ENFA 方法累积贡献率

	累积贡献率(%)		
范围	70-80	80-90	90-100
mienfa	71.57	85.81	97.85
enfa	75.48	89.26	98.55

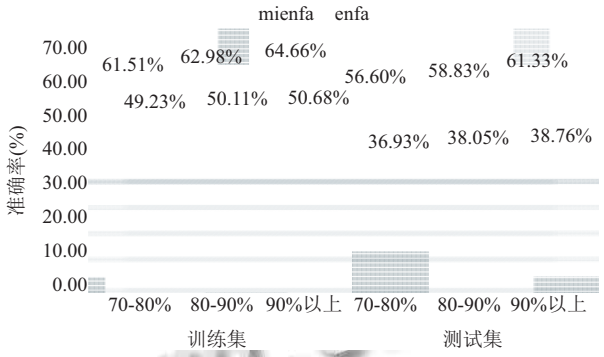

图 4 MIENFA 和 ENFA 方法在 90% 训练集、10% 测试集上的平均准确率

将数据划分为 50% 训练集和 50% 测试集, 分别使用 MIENFA 和 ENFA 进行栖息地适宜性分析, 如图 5, 可以看出训练集的准确度接近于测试集. 根据准确度计算公式以及实验可知, 图 4 中测试集准确率低于训练集可能是由于训练集的样本量太少导致的.

选择累积贡献率达 85.81% 的特化向量, 利用 MIENFA 对斑头雁数据进行分析, K-means 对栖息地适宜性进行分级, 得到不同等级的簇中心、适宜度范围以及所占比重, 表 6 即为分级评价标准. 图 5 为最终的栖息地适宜图.

将土地覆盖图与栖息地适宜图叠加, 得出适宜栖息地区主要是草地, 牧场, 高山和亚高山草甸地区; 次适宜栖息地区主要是高山和亚高山草甸、高寒山草原; 不适宜栖息地区主要是沙砾、沙漠.

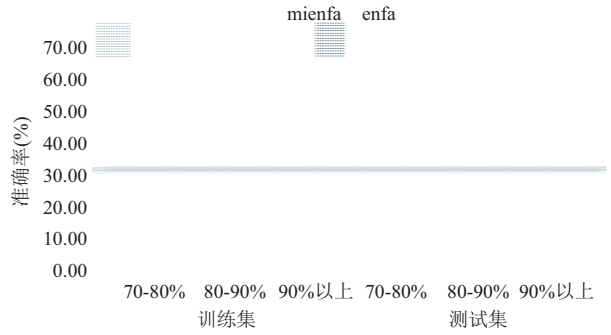

图 5 MIENFA 和 ENFA 方法在 50% 训练集、50% 测试集上的平均准确率

表 6 分级评价标准

适宜性等级	簇中心	适宜度范围	比重(%)
适宜栖息	0.59	[0, 0.955)	17.29
次适宜栖息	1.32	[0.955, 63.73)	41.48
不适宜栖息	126.14	[63.73, +∞)	41.23

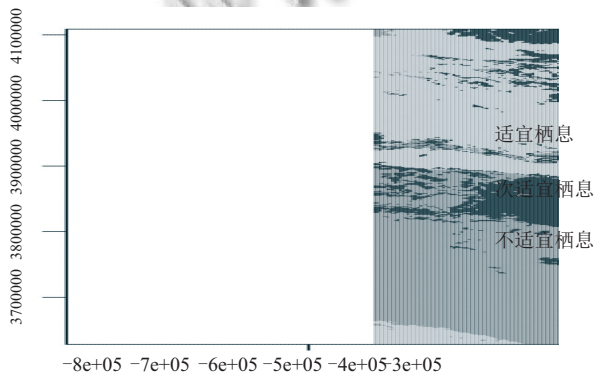

图 6 斑头雁栖息地适宜图

4 结语

针对传统生态位因子分析方法(ENFA)在求解过程中没有考虑环境变量间的非线性相关的问题, 本文在原有的 ENFA 方法基础上, 使用互信息来计算变量间的相关性, 提出基于互信息的生态位因子分析方法(MIENFA). 新的方法不再局限于描述环境变量间的线性相关, 能为数据建立更确切的描述模型.

在分析栖息地选择时, MIENFA 方法主要改变特化向量, 即生态位的宽度. 在计算栖息地适宜性时, MIENFA 在准确率上略高于 ENFA, 在一定程度上证明了该方法能更加确切的描述环境变量间的关系.

由于所选择的研究区域较大, 在计算每个栅格点的最近环境变量时所需时间较长, 后续可考虑并行处理.

参考文献

1 蒋爱伍, 周放, 覃玥, 等. 中国大陆鸟类栖息地选择研究十年. 生态学报, 2012, 32(18): 5918–5923.

2 孔维尧, 郑振河, 吴景才, 等. 莫莫格自然保护区白鹤秋季迁徙停歇期觅食生境选择. 动物学研究, 2013, 34(3): 166–173.

3 戴强, 顾海军, 王跃招. 栖息地选择的理论与模型. 动物学研究, 2007, 28(6): 681–688.

4 赵青山, 楼瑛强, 孙悦华. 动物栖息地选择评估的常用统计方法. 动物学杂志, 2013, 48(5): 732–741.

5 Senay SD, Worner SP, Ikeda T. Novel three-step pseudo-absence selection technique for improved species distribution modelling. PLoS One, 2013, 8(8): e71218. [doi: 10.1371/journal.pone.0071218]

6 陈辉荣. 基于多变量特征分析的栖息地选择分析算法研究及应用[硕士学位论文]. 北京: 中国科学院大学, 2014.

7 Hirzel AH, Hausser J, Chessel D, *et al*. Ecological-niche factor analysis: How to compute habitat-suitability maps without absence data? Ecology, 2002, 83(7): 2027–2036. [doi: 10.1890/0012-9658(2002)083[2027:ENFAHT]2.0.CO;2]

8 Basille M, Calenge C, Marboutin É, *et al*. Assessing habitat selection using multivariate statistics: Some refinements of the ecological-niche factor analysis. Ecological Modelling, 2008, 211(1-2): 233–240. [doi: 10.1016/j.ecolmodel.2007.09.006]

9 王学志, 徐卫华, 欧阳志云, 等. 生态位因子分析在大熊猫(Ailuropoda melanoleuca)生境评价中的应用. 生态学报, 2008, 28(2): 821–828.

10 范雪莉, 冯海泓, 原猛. 基于互信息的主成分分析特征选择算法. 控制与决策, 2013, 28(6): 915–919.

11 中国科学院计算机网络信息中心. 青海湖鸟类 GPS 跟踪数据库的详细信息. http://rsr.csdb.cn/rss01001Action.do?fromAction=rsl01005Action.do&sheetId1694&templatNameEn=DbMetadata09&owerSererName=%E9%9D%92%E6%B5%B7%E6%B9%96%E6%B5%81%E5%9F%9F%E5%9F%BA%E7%A1%80%E7%A7%91%E5%AD%A6%E6%95%B0%E6%8D%AE%E5%BA%93&conditin.inputTextValue=&conditin.selectStautsValue=2&conditin.selectTypeValue=&conditin.selectOwerValue=&conditin.orderbyItem=2&conditin.orderbyItemType=2&conditin.page=1&conditin.totalCount=2966. [2007-10-11].

12 冉有华, 李新, 卢玲. 中国地区土地覆盖综合数据集. 寒区旱区科学数据中心, 2010. [doi: 10.3972/westdc.007.2013.db]

13 Calenge C. Exploratory analysis of the habitat selection by the wildlife in R: The adehabitaths package. 2011. http://www2.uaem.mx/r-mirror/web/packages/adehabitatHS/vignettes/adehabitatHS.pdf.

14 FAO. Food and agriculture organization of the United Nations. Rome: FAO, 1997.

15 孔博, 张树清, 张柏, 等. 遥感和 GIS 技术的水禽栖息地适宜性评价中的应用. 遥感学报, 2008, 12(6): 1001–1009.