Секция «Экология растений»

Оптимизация моделей пространственного распространения растений (SDM): оценка вклада предикторов-климатических переменных

Научный руководитель – Сорокин Анатолий Александрович

Орлов М.А.[1], *Шелудков А.В.*[2]

1 - Московский государственный университет имени М.В.Ломоносова, Биологический факультет, Кафедра биофизики, Москва, Россия, *E-mail: orlovmikhailanat@gmail.com*; 2 - Институт географии РАН, Москва, Россия, *E-mail: a.v.sheludkov@gmail.com*

Модели пространственного распространения видов (Species distribution models, SDM) позволяют связать места фактического нахождения организмов и определяющие их ареал условия географической среды (например, климат, особенности рельефа, тип почвы и др.) Теоретической основой моделирования является концепция экологической ниши. Для вычислительной реализации используются алгоритмы машинного обучения с учителем (supervised machine learning), чаще всего алгоритм максимизации энтропии [1].

При построении моделей пространственного распространения важной задачей является надлежащий отбор независимых переменных-предикторов. Часто используемые биоклиматические переменные базы данных Worldclim [5] включают высоко скореллированные и являющиеся результатом математических преобразований других. Поскольку совместное использование таких переменных усиливает статистический шум и снижает показатели качества моделей, уместен их предварительных отбор. При этом возможны различные подходы: анализ информации о экологической нише исследуемого вида и ее соотнесение с экологическим значением отдельных переменных; их предварительный эксплораторный анализ (до этапа получения моделей) [2], тест пермутации с внесением случайных изменений отдельных переменных [3], собственные метрики машинного обучения. В данной работе использован последний подход, основанный на оценке вклада биоклиматических переменных в работу классификаторов SDM.

Регионами исследования являются полуостров Крым и остров Ванкувер. Выбор обусловлен изоляцией этих территорий морем, их сравнимой площадью, разнообразием условий рельефа и климата. При этом закономерности, определяющие климатическое районирование, значительно разнятся в каждом случае. Для названных территорий из базы данных GBIF [4] извлечены места фактического нахождения (локалитеты) представленного в обеих Crataegus monogyna (свыше 60 в каждой). Следует отметить, что данный вид нативен для п-ова Крым и является интродуцентом в случае о-ва Ванкувер. В качестве точек фона использованы по 120 взятых случайно местообитаний рассматриваемых территорий. Набор предикторов представлял собой 19 биоклиматических переменных базы данных Worldclim 2.0 [5]. Для построения моделей использованы следующие алгоритмы. Maxlike - аналог Maxent с небольшими техническими различиями при сравнимых результатах; bioclim.dismo - простой в использовании алгоритма типа "climate-envelope-model" и реже используемый для построения SDM алгоритм поддерживающих векторов svm. Их качество оценивалось по площади под ROC-кривой (AUC). Для каждого из двух регионов получены значения AUC в диапазоне 0.86-0.99. Для всех моделей на основе их внутренних метрик установлены отдельные биоклиматические переменные с минимальным вкладом в работу классификаторов. В случае Крыма это среднегодовая температура, средняя температура самого холодного кварталов и максимальная температура самого теплого месяца, в случае острова Ванкувер - среднемесячный разброс суточных температур, изотермальность, годовой разброс температур. Далее путем удаления названных переменных из исходных наборов данных получены редуцированные и построены новые модели трех

типов. Анализ площади под ROC-кривой показал очень слабые изменения (в третьем знаке), причем как в сторону увеличения, так и в сторону уменьшения.

Таким образом, получены модели пространственного распространения на основе различных алгоритмов машинного обучения. Показана возможность унифицированной оценки вклада предикторов в их работу за счет внутренних метрик. Такая оценка может оказаться ценной для анализа экологических особенностей расселения видов, населяющих регионы с различными условиями среды. Удаление малоинформативных предикторов не снижает показатели качества моделей, снижая уровень статистического шума. Вывод заключается в целесообразности совместного применения различных алгоритмов для получения SDM и оценки вклада предикторов с последующим отбором оптимальных.

Источники и литература

1) Дудов С.В. Моделирование распространения видов по данным рельефа и дистанционного зондирования на примере сосудистых растений нижнего горного пояса хр. Тукурингра (Зейский заповедник, Амурская область) // Журнал общей биологии. 2016. Т. 77, № 2. С. 122–134.

2) Orlov M. Climate data optimization for species distribution models using unsupervised machine learning // Abstracts of international scientific conference "Information Technologies in the Research of Biodiversity (BIT - 2018)". — 2018. — P. 38-39

3) Phillips S.J., Dudik M. Modeling of species distributions with MaxEnt: new extensions and a comprehensive evaluation // Ecography. 2008. V. 31. P. 161–175.

4) Global Biodiversity Information Facility: https://www.gbif.org/

5) WorldClim - Global Climate Data: http://www.worldclim.org/