Научный коллектив Института природно-технических систем в рамках выполнения проекта Российского научного фонда № 23-29-00558 разработал алгоритм обнаружения аномалий с помощью алгоритмов машинного обучения без учителя и прогнозных моделей для программного обеспечения автоматизированного комплекса биомониторинга водной среды, основанного на поведенческих реакциях двустворчатых моллюсков.

Грант № 23-29-00558 «Обнаружение аномалий в данных активности моллюсков алгоритмами машинного обучения для формирования сигнала тревоги в комплексах автоматизированного биомониторинга водной среды».

Исполнители – к.г.н., в.н.с. Вышкваркова Е.В. (руководитель), к.т.н., зам. руководителя центра Греков А.Н., к.б.н., вед. инженер-исследователь Трусевич В.В. и инженер Маврин А.С.

В качестве алгоритмов машинного обучения без учителя выбраны эллиптическая огибающая (elliptic envelope), изолирующий лес (iForest), одноклассовый метод опорных векторов (one-class SVM) и локальный уровень выбросов (LOF). Анализ данных проведен на языке программирования Python (V3.9.12) с использованием пакета машинного обучения scikit-learn (V 1.0.2). Для каждого алгоритма проведен выбор и настройка оптимальных гипермараметров, таких как уровень загрязнения (contamination rate), время осреднения данных, масштабирование, стандартизация и другие. Для трех аномалий лучшую скорость реакции на аномалию показал алгоритм машинного обучения IForest при усреднении данных за 15 мин, T = 50 и n, равном 70.

Для обнаружения аномалий в рядах активности двустворчатых моллюсков с использованием прогнозных моделей была проведена следующая работа: 1) Декомпозиция временных рядов. Позволяет определить размерность сезонности. Сезонный компонент можно наблюдать в виде закономерностей, которые повторяются через одинаковые промежутки времени. Анализируя эти закономерности, можно определить частоту или период сезонности данных; 2) Определение оптимальной модели, то есть определение гиперпараметров наших моделей; 3) Построение модели по оптимальным гиперпараметрам, определенным на шаге 2; 4) Прогнозирование данных. Использование построенной модели, чтобы предсказать данные временного ряда; 5) Оценка и анализ результатов. Проведенная декомпозиция данных показала наличие сезонности в данных, что подтвердило необходимость использования модели ARIMA с сезонной составляющей (SARIMA). Определение оптимальных параметров модели SARIMA осуществлялось двумя способами: первый - подбор оптимальных параметров путем минимизации ошибок MAPE и RMSE между прогнозируемыми и фактическими значениями, а второй – пошаговый алгоритм с оптимизацией по информационному критерию Акаике (AIC).

Сравнение результатов обнаружения аномалий моделью ARIMA с оценками обнаружения аномалий в этих же данных двустворчатых моллюсков с использованием четырёх алгоритмов машинного обучения без учителя на примере одной аномалии показало, что использование алгоритмов машинного обучения даёт небольшое преимущество в 10 минут по скорости обнаружения аномалии по сравнению с моделью SARIMA. Однако, при обнаружении аномалий алгоритмами машинного обучения без учителя в качестве входных данных для обучения моделей использовались данные 14 мидий с различными осреднениями за пятидневный интервал, что увеличивает вычислительную сложность по сравнению с моделью SARIMA на несколько порядков.

Результаты первого года проекта показали, что природные и технические аномалии в наборах данных об активности двустворчатых моллюсков можно обнаружить с помощью алгоритмов машинного обучения без учителя и прогнозными моделями.

Полученные результаты будут использованы в разработанном программном обеспечении комплекса для формирования сигнала тревоги в режиме реального времени для своевременного информирования заинтересованных лиц о неблагоприятных условиях (загрязнении) водной среды.