МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение высшего образования

«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»

Кафедра информатики и программирования

КЛАССИФИКАЦИЯ ТЕРРИТОРИИ ВОДОХРАНИЛИЩА ПО ДАННЫМ ДИСТАНЦИОННОГО ЗОНДИРОВАНИЯ НА РЕЧНЫЕ И ОЗЕРНЫЕ УЧАСТКИ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса, 441 группы

направления 02.03.03 – математическое обеспечение и администрирование информационных систем

факультета компьютерных наук и информационных технологий

Королькевича Ильи Дмитриевича

Научный руководитель:
к.ф.-м.н., доцент _____ Огнева М. В.

Заведующий кафедрой:
к.ф.-м.н., доцент _____ Огнева М. В.

Саратов 2025

ВВЕДЕНИЕ

Современные методы машинного обучения открывают новые возможности для анализа данных, в частности, дистанционного зондирования (ДЗЗ), особенно в задачах картографирования и мониторинга природных объектов. Одной из ключевых проблем в этой области остается точная классификация водных поверхностей [1-9], таких как водохранилища, где традиционные методы часто сталкиваются с трудностями из-за изменчивости спектральных характеристик, влияния погодных условий и антропогенных факторов.

Применение алгоритмов машинного обучения (деревья решений, случайные леса, метод опорных векторов, нейронные сети) позволяет автоматизировать обработку больших объемов данных ДЗЗ, выявлять сложные зависимости между параметрами водоемов и минимизировать субъективные ошибки ручного дешифрирования. Это особенно важно для задач оперативного мониторинга состояния водохранилищ, где точность классификации напрямую влияет на эффективность управления водными ресурсами.

Поэтому актуальной является тема данной работы – классификация территории водохранилища на речные и озерные участки по данным дистанционного зондирования Земли (ДЗЗ) с использованием признаков «цветения» водоемов, таких как индексов мутности, цвета, хлорофилла–А, альгоиндекса NDAI.

Выделение типов участков водохранилища на примере построения таких индексов позволяет эффективно анализировать различные типы акватории, выявлять их количественные и качественные характеристики. Индексный подход помогает обрабатывать дистанционно большие массивы данных, что существенно упрощает задачи мониторинга водных объектов.

Картографирование водных объектов с учетом их изменений в различные периоды времени позволяет создавать и накапливать большие массивы данных. Это, в свою очередь, создает информационную базу для научных исследований

и практических задач, связанных с прогнозом режима водного объекта, нахождению зависимостей показателей индексов от прочих гидрологических величин, особенностей территории и т.д.

Цель дипломной работы - выделить из водохранилища участки открытой воды («озерные») и участки закрытые («речные») – протоки, устья рек, межостровные пространства с использованием показателей «цветения» водоема.

Задачи:

1. Выполнить обзор источников, изучающих схожую тему анализа и классификации водоемов с использованием данных дистанционного зондирования.
2. Изучить теоретическую часть различных моделей машинного обучения, а также основные характеристики, используемые при классификации водоемов.
3. Изучить библиотеки для работы.
4. Подготовить и проанализировать данные, необходимые для обучения моделей
5. Обучить классификаторы, выполнить оценку качества
6. Провести анализ полученных результатов.

Методологические основы классификации водоемов перечислены в работах Т.И. Кутявиной и В.В. Рутмана [8], в которой для мониторинга эвтрофированных водоемов РФ исследователи предложили геоинформационную модель, интегрирующую космоснимки, пространственный аналих буферных зон и машинное обучение, а также M. Viso-Vázquez и X. Álvarez [7], которые провели пионерское исследование по детектированию цианобактериальных цветений в эвтрофных водохранилищах Испании с использованием данных Sentinel-2.

Теоретическая и практическая значимость бакалаврской работы состоит в том, что автоматическая классификация участков водных объектов с учетом их изменений в различные периоды времени позволяет создавать и накапливать большие массивы данных для научных исследований и практических задач,

связанных с мониторингом и прогнозом водоемов, нахождению зависимостей показателей индексов от прочих гидрологических величин и т. п.

Бакалаврская работа состоит из введения, 4 разделов, заключения, списка использованных источников и 9 приложений. Общий объем работы – 93 страницы, из них 57 страниц – основное содержание, включая 11 рисунков и 13 таблиц, цифровой носитель в качестве приложения, список использованных источников информации – 21 наименование.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Общие подходы» посвящен обзору исследовательских работ в области классификации водоемов на темы, схожие с темой бакалаврской работы, такие как мониторинг эвтрофикации, обнаружение водных объектов и оценка границ, классификации качества воды, прогнозирование параметров водохранилищ.

Повышение общего качества и количества спутниковых изображений предлагает новые возможности для задач автоматической обработки водных объектов, таких как обнаружение водных объектов и оценка границ [1] [2] [3], классификации качества воды [4], прогнозирование параметров водохранилищ [5] [6], мониторинг эвтрофикации [7] [8] [9].

Для мониторинга эвтрофированных водоемов РФ исследователи предложили геоинформационную, интегрирующую [8]:

- Космоснимки (Landsat-8)
- Пространственный анализ буферных зон
- Машинное обучение (SVM для классификации трофического статуса)

M. Viso-Vázquez и X. Álvarez провели пионерское исследование по детектированию цианобактериальных цветений в эвтрофных водохранилищах Испании с использованием Sentinel-2 [7]. Их модель на основе регрессионных деревьев предсказывала концентрацию хлорофилла-а (R^2=0.89) через связку спектральных каналов (B3, B5, B8).

Второй раздел «Используемые методы машинного обучения» описывает теоретические основы применяемых в работе алгоритмов машинного обучения. Были рассмотрены различные модели, которые изначально решают задачу многоклассовой классификации или могут быть обобщены на этот случай.

В данном разделе описываются следующие алгоритмы: метод ближайших соседей, метод опорных векторов, логистическая регрессия, дерево решений, случайный лес, многослойный персептрон, градиентный бустинг, а также его вариации: AdaBoost, LightGBM, CatBoost, XGBoost.

Третий раздел «Используемый инструментарий» посвящен обзору применяемых методов, алгоритмов и других инструментов для работы с данными, моделями машинного обучения, их визуализацией:

Geopandas — библиотека для работы с геопространственными данными, расширяющая возможности pandas. Позволяет анализировать и визуализировать географические данные. Интегрируется с matplotlib для создания тематических карт.

Matplotlib — основная библиотека для построения статических, анимированных и интерактивных графиков в Python. Используется для визуализации данных.

Numpy — инструмент для математических операций с массивами данных. Необходим для преобразования и нормализации числовых признаков перед обучением моделей.

Scikit-learn — это библиотека машинного обучения в Python, которая предоставляет инструменты для классификации, регрессии и кластеризации, в том числе следующие алгоритмы классификации: LogisticRegression, KNeighborsClassifier, LogisticRegression, LinearSVC, DecisionTreeClassifier, RandomForestClassifier.

AdaBoostClassifier, CatBoostClassifier, XGBClassifier, LGBMClassifier — библиотеки, предлагающие различные реализации методов градиентного бустинга на основании разных подходов к построению решающих деревьев.

Sequential — это линейный стек слоев нейронной сети, предоставляемый библиотекой Keras в рамках TensorFlow. Этот метод используется для простых архитектур, таких как многослойные перцептроны (MLP) или базовые сверточные сети (CNN).

Четвертый раздел «Практическая часть» посвящен подготовке данных для исследования, классификации и сравнению различных алгоритмов, и анализу полученных результатов.

В ходе работы был подготовлен набор данных, включающий в себя гидрологические показатели, используемые для классификации: индекс Хлорофилла-А, альгоиндекс, индекс мутности, индекс цвета и температура поверхности воды. Целевой признак – метка класса: коренной берег, суша в ложе водохранилища (острова), озерные участи и речные участки.

На основе этих данных было проведено тестирование различных методов машинного обучения, рассмотренных во втором разделе. Результаты показали, что модели оказались неспособны найти закономерности для наименее представленных в наборе классов, а именно классов берега и речных участков. Различные методы балансировки данных не смогли улучшить результат.

Для более подробного анализа проблемы задача была сведена к бинарной классификации двух ключевых классов: озерных и речных участков. Результаты показали, что модели имеют крайне похожие оценки F-меры для обоих классов, что навело на мысль о том, что модели ошибаются на одних и тех же объектах.

Для проверки этого предположения было проведено сравнение предсказаний, которое показало, что у каждой пары классификаторов более 70% ошибок – на одних и тех объектах.

Для дальнейшего анализа потребовалась возможность визуализации результатов, что потребовало пересобрать набор данных, включив в него столбец, отвечающий за координаты точки в пространстве. Бинарная классификация на новом наборе дала схожие результаты, однако визуализация показала, что у каждой модели основная часть ошибок сосредоточена в одних и тех же участках водоема. Проанализировав их, можно сказать, что это участки, которые являются:

- с одной стороны, достаточно крупными непрерывными акваториями без признаков речных участков внутри себя;
- с другой стороны, внешне они ограничены островами, и, по сути, все представляют собой крупные заливы островов, которые в основном относятся к участкам речным.

Другими словами, на этапе разметки у квалифицированного специалиста возникают сомнения в отнесении этих участков к классам "речные" или "озерные". Не удивительного, что и модель сомневается. Ее ошибки логичны и вполне оправданы. Есть вероятность, что модель в этих случаях более достоверна, чем исходная разметка данных.

ЗАКЛЮЧЕНИЕ

В ходе проделанной работы была изучена теоретическая часть методов машинного обучения, а также основные характеристики, используемые при классификации водоемов, протестированы различные модели классификаторов, таких как K-ближайших соседей, деревья решений, случайный лес, а также несколько различных реализаций градиентного бустинга, на нескольких датасетах, результаты тестирования были проанализированы.

Оценка классификаторов показала следующие результаты: общая точность предсказаний варьируется от 0.84 до 0.90. Наивысшую эффективность показали модели случайного леса, CatBoost и XGBoost.

Кроме того, анализ показал, что первостепенное значение для классификации водоемов имеет то, насколько верно были размечены данные, а достаточно мощные алгоритмы способны указывать на проблемные места в разметке данных.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Gharbia, R. Deep Learning for Automatic Extraction of Water Bodies Using Satellite Imagery / R. Gharbia // Journal of the Indian Society of Remote Sensing – 2023 – T. 51(7) - C. 1511–1521
2. Huang C. Mapping river and lake water bodies using Landsat imagery / C. Huang // International Journal of Remote Sensing - 2020
3. Mason, J., Pritchard, J. The use of remote sensing data for the classification of water bodies / J. Mason, J. Pritchard // Journal of Hydrology - 2018
4. Malek, N., Yaacob, W., Nasir, S., Shaadan, N., Prediction of Water Quality Classification using Machine Learning / N. Malek, W. Yaacob, S. Nasir, N. Shaadan // Water - 2022 - 14 - 1067
5. Lerios, J. L., Villarica, M. V., Pattern Extraction of Water Quality Prediction Using Machine Learning Algorithms of Water Reservoir / J. L. Lerios, M. V. Villarica // International Journal of Mechanical Engineering and Robotics Research – 2019 – Vol. 8, No. 6
6. Huang, J., Xu, X., Combining Satellite Imagery and a Deep Learning Algorithm to Retrieve the Water Levels of Small Reservoirs / J. Huang, X. Xu // Remote Sens – 2023 – №15
7. Viso-Vázquez, M., Álvarez, X. Remote Detection of Cyanobacterial Blooms and Chlorophyll-a Analysis in a Eutrophic Reservoir Using Sentinel-2 / M. Viso-Vázquez, X. Álvarez // Sustainability – 2021 – №13
8. Кутявина, Т.И., Рутман, В.В., Ашихмина, Т.Я., Применение пространственного геоинформационного анализа по материалам космоснимков для проведения мониторинга состояния эвтрофированных водоемов / Т.И. Кутявина, В.В. Рутман, Т.Я. Ашихмина // Международная конференция: Eco-TIRAS – 2021
9. Xie, Y., A review of remote sensing applications for water quality monitoring / Y. Xie // Sensors – 2021