DOI https://doi.org/10.46845/2541-8254-2024-5(47)-11-11
УДК 004.93

МЕТОДЫ ОБРАБОТКИ И ХРАНЕНИЯ ДАННЫХ ДЛЯ АВТОМАТИЗИРОВАННОЙ КЛАССИФИКАЦИИ СУБСТРАТОВ МОРСКОГО ДНА С ИСПОЛЬЗОВАНИЕМ НЕЙРОННЫХ СЕТЕЙ

Д. Е. Васильев, студент 2-го курса магистратуры
E-mail: danil.vasilev@digital-klgtu.ru
ФГБОУ ВО «Калининградский государственный технический университет»

Н. В. Шашин, студент 2-го курса магистратуры
E-mail: nikita.shashin@digital-klgtu.ru
ФГБОУ ВО «Калининградский государственный технический университет»

Ю. А. Ершов, студент 2-го курса магистратуры
E-mail: yurij.ershov@digital-klgtu.ru
ФГБОУ ВО «Калининградский государственный технический университет»

В статье рассматриваются методы обработки данных и современные подходы к организации хранения данных в целях их дальнейшего использования с применением технологий искусственного интеллекта и машинного обучения. Сформированы краткосрочный и долгосрочный планы разработки. Выделены факторы, повлиявшие на итоговый выбор.

Ключевые слова: *обработка данных, методы разметки, методы организации, хранилища данных.*

ВВЕДЕНИЕ

Изучение субстратов морского дна представляет собой значимую область исследований в контексте геоморфологии, физико-химических характеристик морского дна, распределения биоценозов и оценки потенциальных ресурсов. Традиционные методы классификации донных отложений, как правило, основываются на экспертной интерпретации данных многолучевого эхолокационного зондирования, что является процессом, требующим значительных временных и трудовых ресурсов, а также подверженным субъективному влиянию. В связи с этим разработка автоматизированных систем классификации становится насущной задачей, направленной на повышение точности и эффективности исследований.

Современные технологии способствуют улучшению определения характеристик морского дна, что имеет критическое значение для различных направлений морского исследования, гидрографических работы и других областей, где важным является точное картирование состава донного покрова. Одними из ключевых инструментов, облегчающих анализ данных, являются современные методы искусственного интеллекта. В частности, алгоритмы машинного обучения, такие как сверточные нейронные сети, демонстрируют высокую эффективность в решении задач автоматической сегментации и классификации изображений морского

дна, получаемых посредством гидроакустических систем. Применение этих алгоритмов способствует ускорению обработки больших объемов данных и обеспечивает более объективные и воспроизводимые результаты по сравнению с традиционными экспертными методами.

Кроме того, использование методов анализа больших данных и разработка интеллектуальных систем поддержки принятия решений создают перспективы для углубленного изучения закономерностей пространственно-временного распределения донных отложений. Это имеет большое значение для оптимизации управления морскими ресурсами и планирования хозяйственной деятельности.

ОБЪЕКТ ИССЛЕДОВАНИЯ

Объектом исследования являются алгоритмы и методы сегментации, распознавания образов, управления большими объемами неструктурированных данных, являющиеся неотъемлемой частью процесса решения задачи классификации типов субстратов морского дна.

ЦЕЛЬ И ЗАДАЧИ ИССЛЕДОВАНИЯ

Целью исследования является организация процесса обработки и хранения данных, касающихся типов субстратов морского дна, с дальнейшим их использованием для решения задачи классификации с применением технологий искусственного интеллекта и машинного обучения.

Задачами текущего этапа исследования являются:

- выбор наиболее оптимального подхода к обработке батиметрической съемки;
- принятие решения о способе организации хранения данных и доступа к ним.

МЕТОДЫ ИССЛЕДОВАНИЯ

Данные батиметрии представляют из себя огромные массивы данных, которые исчисляются гигабайтами. Подобные данные, без всякого сомнения, попадают под термин Big Data. Под Big Data понимаются большие объемы структурированных или неструктурированных данных, особенность которых заключается в их многообразии. Из-за особенностей многолучевого эхолота, окружающей среды, других независимых факторов сырые данные, очевидно, не подходят для работы. Для этого рассматриваются существующие методы обработки данных и организации их хранения.

Под обработкой данных понимается их нормализация, очистка, разметка. Непосредственно для обучения нейронной сети главную роль играют размеченные данные, которые будут частью обучающей выборки. Выбор оптимального метода обучения нейронной сети зависит от природы исследования, доступности размеченных данных. Сравнительный анализ различных подходов помогает выявить наиболее подходящий метод для задач классификации и сегментации изображений, например в исследовании донных отложений.

Методы разметки играют ключевую роль в предварительной обработке данных для обучения моделей. Перечислим наиболее распространенные методы.

Ограничивающие прямоугольники (Bounding Boxes). Это простейшая форма разметки, представляющая минимальные прямоугольники вокруг объектов на изображениях или в пространстве [1]. Они широко используются в компьютерном зрении и машинном обучении для задачи обнаружения объектов. Основное преимущество заключается в их простоте и скорости обработки, что делает их эффективными для большинства базовых задач. Однако такой метод может быть недостаточно точным для сложных объектов или объектов нестандартной формы.

Аннотирование многоугольниками. В отличие от ограничивающих прямоугольников, полигональная разметка предназначена для объектов с неправильными формами [2]. Этот метод используется для более точного определения границ объекта, что особо важно в медицине и требовательных приложениях, где детали играют критическую роль. Основной недостаток метода – трудоемкость и субъективность процесса ручной аннотации, а также сложность масштабирования на большие массивы данных.

Семантическая сегментация. Данный метод подразумевает присвоение меток каждому пикселю изображения, что обеспечивает более глубокую разметку сцены. Семантическая сегментация применяется в областях, требующих аналитики сложной визуальной информации, таких как медицинская диагностика и автоматизация транспортных средств. Она позволяет разработать более качественные и устойчивые модели. Тем не менее это требует высокого уровня вычислительной мощности и все еще остается менее точным по сравнению с методами точечного распознавания объектов [3–6].

Для выполнения сложных задач сегментации часто привлекаются модели глубокого обучения, такие как сверточные нейронные сети (CNN) [7]. Применение таких архитектур может обеспечить извлечение более точных признаков, необходимых для точной разметки. Например, модели типа DenseNet с более плотными связями между слоями позволяют эффективно комбинировать различные признаки, что особенно актуально для детальной семантической сегментации, применяемой в морских и экологических исследованиях.

Для дальнейшего использования обработанных данных рассматриваются существующие методы организации их хранения.

Реляционные базы данных (RDBMS) занимают значительное место в арсенале инструментов для управления данными, широко применяемых в современных информационных системах. Эти системы организуют данные в виде таблиц, создавая между ними связи при помощи ключей, что формирует структурированное представление информации. Системы управления реляционными базами данных (СУБД), такие как MySQL, PostgreSQL и Microsoft SQL Server, предлагают разнообразные функциональные возможности, включая поддержку транзакций, сложных запросов и обеспечения целостности данных [8, 9].

MySQL является популярной бесплатной RDBMS с открытым исходным кодом, востребованной благодаря своей простоте, высокой производительности и доступности для бесплатного использования. Чаще всего MySQL применяется в веб-разработке, где она активно поддерживает динамичные веб-приложения [10].

Microsoft SQL Server предлагает обширный спектр функциональных возможностей, включая поддержку хранимых процедур и репликацию данных. Благодаря своей масштабируемости и интеграции с другими продуктами Microsoft он становится предпочтительным выбором для корпоративных сред.

PostgreSQL выделяется как мощный и гибкий инструмент с открытым исходным кодом, предоставляющий такие функции, как поддержка JSON-структур данных и полнотекстовый поиск. Сильные стороны PostgreSQL делают ее идеальным выбором для задач, требующих высокой надежности и производительности [11].

Выбор конкретной СУБД часто зависит от характеристик и требований конкретного проекта, а также предпочтений разработчиков.

Современные тенденции формирования огромных объемов неструктурированных данных создают новые вызовы для традиционных RDBMS, которые могут не соответствовать динамичным и разнородным требованиям. NoSQL базы данных предлагают альтернативные подходы управления данными [12].

MongoDB демонстрирует высокую производительность при операциях записи, особенно в случае сложных обновляющих запросов. Она эффективно справляется с большими объемами данных при выполнении операций чтения и агрегирующих запросов [13].

Cassandra выделяется структурой, напоминающей SQL, что значительно облегчает переход от реляционных систем, обеспечивая баланс между производительностью чтения и записи и возможностью обработки масштабных нагрузок [14].

Redis превосходит конкурентов по скорости чтения в больших базах данных. Он особенно эффективен, когда требуется высокая производительность на малом количестве данных [15].

Выбор между этими решениями зависит от спецификации данных, требований к обработке и гибкости применения.

Объектные хранилища данных становятся неотъемлемой частью облачной инфраструктуры, предоставляя единый интерфейс для хранения и извлечения данных. Amazon S3, Yandex S3, как одни из известных представителей, обеспечивают надежное, масштабируемое и доступное решение для хранения данных в облаке.

S3 поддерживает события и триггеры, позволяя интегрировать вычислительные задачи для создания событийно-ориентированных приложений. Его интеграция с такими сервисами как AWS Lambda и Step Functions позволяет создать комплексные и устойчивые архитектуры.

С увеличением объемов данных и сложности работы с метаданными традиционные объектные хранилища сталкиваются с лимитами производительности. Hadoop Distributed File System (HDFS) предлагает решение, используя центральный NameNode для управления метаданными.

Чтобы улучшить масштабируемость и производительность HDFS, была разработана архитектура Dynamic Federated Metadata Management (DFMM). Она использует HDFS Federation для распределения метаданных между несколькими NameNodes, применяя техники шардирования. Это позволяет усовершенствовать масштабируемость и производительность, демонстрируя увеличение производительности более чем на 21% по сравнению с традиционным решением.

Рассматривая предложенные подходы, видно, что каждая методология обладает уникальными преимуществами и применяется в зависимости от требований проекта. Выбор между реляционными СУБД, NoSQL системами, объектными хранилищами или системами на основе Hadoop зависит от потребностей в обработке, хранения и организация данных. Использование каждого из решений привнесет ощутимые улучшения в управлении данными, обеспечивая значительное улучшение эффективности и позволяя преодолевать текущие и будущие технические вызовы в компьютерной науке.

S3 поддерживает события и триггеры, позволяя интегрировать вычислительные задачи для создания событийно-ориентированных приложений. Его интеграция с такими сервисами как AWS Lambda и Step Functions позволяет создать комплексные и устойчивые архитектуры [9, 16, 17].

С увеличением объемов данных и сложности работы с метаданными традиционные объектные хранилища сталкиваются с лимитами производительности. Hadoop Distributed File System (HDFS) предлагает решение, используя центральный NameNode для управления метаданными [18, 19].

Чтобы улучшить масштабируемость и производительность HDFS, была разработана архитектура Dynamic Federated Metadata Management (DFMM). Она использует HDFS Federation для распределения метаданных между несколькими NameNodes, применяя техники шардирования. Это позволяет усовершенствовать масштабируемость и производительность, демонстрируя увеличение производительности более чем на 21 % по сравнению с традиционным решением.

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

Разработка инструмента для классификации субстратов морского дна может быть структурирована в рамках краткосрочных и долгосрочных стратегических планов. В рамках краткосрочной перспективы предлагается перейти от задачи классификации типов субстратов к задаче определения непосредственно субстрата на карте морского дна. Этот подход обусловлен необходимостью постепенного и систематического решения поставленной задачи, чтобы обеспечить последовательное совершенствование методов и технологий.

В долгосрочной перспективе для решения задачи классификации типов донных субстратов предполагается использование методологии семантической сегментации. Этот метод обеспечивает более точное определение формы субстратов произвольной конфигурации, превосходя по точности альтернативные подходы, такие как традиционные методы классификации. К применению данной методики необходимо стремиться в рамках длительного процесса.

Учитывая сложность задачи, предполагается использование пошагового подхода. На первоначальном этапе разметка данных будет осуществляться с помощью ограничивающих прямоугольников. Этот метод был выбран из-за его относительной простоты и возможности более быстрой реализации, что позволяет эффективно закладывать основу для последующих, более сложных этапов обработки и анализа данных.

После изучения различных методов и решений, наиболее соответствующим потребностям проекта было признано использование объектного хранилища данных.

Основными факторами, определившими выбор объектного хранилища, стали:

- Масштабируемость и гибкость. Объектные хранилища обладают высокой масштабируемостью, позволяя наращивать объемы хранимых данных по мере необходимости. Кроме того, они предоставляют гибкие возможности по управлению данными, такие как группировка объектов, управление жизненным циклом и др.
- Надежность и безопасность. Объектные хранилища, как правило, обеспечивают высокий уровень защиты данных, включая механизмы шифрования и резервного копирования. Это критически важно для хранения чувствительных данных, получаемых с многолучевого эхолота.
- Интеграция. Выбранное объектное хранилище должно обеспечивать эффективную интеграцию с другими компонентами инфраструктуры, упрощая процессы анализа и обработки данных.
- Экономичность. Соотношение стоимости и функциональности объектного хранилища должно позволять эффективно управлять бюджетом на хранение и обработку данных.

ЗАКЛЮЧЕНИЕ

Резюмируя вышесказанное, для обработки информации будут использоваться методы с применением технологий искусственного интеллекта и машинного обучения.

На основе всестороннего анализа и оценки различных решений для хранения и управления большими данными, Yandex S3 Object Storage было признано наиболее подходящей платформой для организации распределенной системы обработки и хранения данных в рамках данного исследования.

СПИСОК ЛИТЕРАТУРЫ

1. Mikhaylov, A. A. Automatic data labeling for document image segmentation using deep neural networks / A. A. Mikhaylov // Proceedings of ISP RAS. – 2022. – Vol. 34, № 6. – P. 137–146.
2. Hu, Y. PolyBuilding: Polygon transformer for building extraction / Y. Hu [et al.] // ISPRS Journal of Photogrammetry and Remote Sensing. – 2023. – Vol. 199. – P. 15–27.
3. Shi, F. Supervised Semantic Image Annotation Using Region Relevance / F. Shi // Physics Procedia. – 2012.
4. Georg, G. Enhancing Underwater Image Segmentation: A Semantic Approach to Segment Objects in Challenging Aquatic Environment / G. Georg // Procedia Computer Science. – 2024. – Vol. 235. – P. 361–371.
5. Yang, C. Sonar image segmentation framework based on semi-supervised learning / C. Yang [et al.]. – 2023.
6. Wahyono. Region-based annotation data of fire images for intelligent surveillance system / Wahyono [et al.] // Data in Brief. – 2022. – Vol. 41. – P. 107925.
7. Wei, W. Automatic image annotation based on an improved nearest neighbor technique with tag semantic extension model / W. Wei [et al.] // Procedia Computer Science. – 2021. – Vol. 183. – P. 616–623.
8. Жук, М. М. Реляционные базы данных, язык SQL / М. М. Жук // StudNet. – Москва: ООО «Электронная наука», 2022. – № 6. – С. 5190–5196.
9. Sachdeva, S. Critical Analysis of Data Storage Approaches // Procedia Computer Science. 2020. – Vol. 173. – P. 264–271.

10. Григорьев, Ю. А. Сравнение времени выполнения запроса к хранилищу данных в среде MapReduce / Hadoop и СУБД MySQL / Ю. А. Григорьев, А. И. Устимов // Информатика и системы управления. – 2016. – № 49. – С. 3–12.

11. Татарникова, Т. М. Кластеризация данных на лету для СУБД PostgreSQL / Т. М. Татарникова // Программные продукты и системы. – 2023. – С. 196–201.

12. Rao, A. Insights into NoSQL databases using financial data: A comparative analysis / A. Rao // Procedia Computer Science. 2022. – Vol. 215. – P. 8–23.

13. Rameshwar, D. L. The MongoDB injection dataset: A comprehensive collection of MongoDB – NoSQL injection attempts and vulnerabilities / D. L. Rameshwar, H. V. Ankush // Data in Brief. – Elsevier, 2024. – Vol. 54. – P. 110289.

14. Suárez-Cabal M.-J. MDICA: Maintenance of data integrity in column-oriented database applications / M.-J. Suárez-Cabal, P. Suárez-Otero // Computer Standards & Interfaces. – North-Holland, 2023. – Vol. 83. – P. 103642.

15. Balis, B. Towards an operational database for real-time environmental monitoring and early warning systems / B. Balis, M. Bubak // Procedia Computer Science. Elsevier, 2017. – Vol. 108. – P. 2250–2259.

16. Ikhlasse, H. Recent implications towards sustainable and energy efficient AI and big data implementations in cloud-fog systems: A newsworthy inquiry // Computer and Information Sciences. – 2022.

17. Saadoon, M. Fault tolerance in big data storage and processing systems: A review on challenges and solutions / M. Saadoon // Ain Shams Engineering Journal. – 2022.

18. Guan, S. Hadoop-based secure storage solution for big data in cloud computing environment / S. Guan [et al.] // Digital Communications and Networks. 2024. – Vol. 10, № 1. – P. – 227–236.

19. Guan, Y. HDFS Optimization Strategy Based On Hierarchical Storage of Hot and Cold Data / Y. Guan, Z. Ma, L. Li // Procedia CIRP. – 2019. – Vol. 83. – P. 415–418.

METHODS OF DATA PROCESSING AND STORAGE FOR AUTOMATED CLASSIFICATION OF SEABED SUBSTRATES USING NEURAL NETWORKS

D. E. Vasilev, 2nd year master degree student
E-mail: danil.vasilev@digital-klgtu.ru
Kaliningrad State Technical University

N. V. Shashin, 2nd year master degree student
E-mail: nikita.shashin@digital-klgtu.ru
Kaliningrad State Technical University

Y. A. Ershov, 2nd year master degree student
E-mail: yurij.ershov@digital-klgtu.ru
Kaliningrad State Technical University

The article deals with the methods of data processing and modern approaches to the organization of data storage for further use together with artificial intelligence and machine learning technologies. Short-term and long-term development plans have been formed. The factors that influenced the final choice have been highlighted.

Keywords: data *processing, partitioning methods, organization methods, data warehouses.*