

Д. А. Гура^{1,2}, Р. А. Дьяченко^{1,3}, Т. А. Тихонов¹✉

Применение методик сбора пространственных данных для машинного обучения

¹Кубанский государственный технологический университет,
г. Краснодар, Российская Федерация

²Кубанский государственный аграрный университет, г. Краснодар, Российская Федерация

³Московский физико-технический институт (национальный исследовательский университет)
Москва, Российская Федерация
e-mail: timka2015@yandex.ru

Аннотация. В работе рассматриваются особенности формирования набора пространственных данных в виде облаков точек лазерного отражения для последующего применения в задачах машинного обучения при автоматизированном распознавании объектов недвижимости. Выполнено лазерное сканирование территории с использованием оборудования Leica BLK360. Полученные облака точек лазерного отражения были вручную классифицированы по семантическим признакам в программном обеспечении CloudCompare с последующей верификацией двумя независимыми операторами, что позволило сформировать эталонный набор данных для обучения нейросетевой модели. Анализ распределения классов выявил выраженную разбалансировку, обусловленную естественным неравенством площадей и объемов объектов, что негативно влияет на точность классификации редких классов обученной моделью. Приведены основные подходы, направленные на устранение указанной проблемы для повышения качества обучения моделей.

Ключевые слова: облака точек лазерного отражения, ТЛО, нейросетевая модель, классификация, лазерное сканирование

D. A. Gura^{1,2}, R. A. Dyachenko^{1,3}, T. A. Tikhonov¹✉

Applying Spatial Data Collection Techniques to Machine Learning

¹Kuban State Technological University, Krasnodar, Russian Federation

²Kuban State Agrarian University, Krasnodar, Russian Federation

³Moscow Institute of Physics and Technology (National Research University)
Moscow, Russian Federation
e-mail: timka2015@yandex.ru

Abstract. This paper examines the specifics of generating a spatial dataset in the form of laser reflectance point clouds for subsequent use in machine learning tasks for automated real estate recognition. A Leica BLK360 laser scan of the area was performed. The resulting laser reflectance point clouds were manually classified by semantic features using CloudCompare software and subsequently verified by two independent operators, enabling the formation of a reference dataset for training a neural network model. Analysis of the class distribution revealed a pronounced imbalance due to the natural inequality in the areas and volumes of objects, which negatively impacts the classification accuracy of rare classes by the trained model. The main approaches aimed at eliminating this problem and improving the quality of model training are presented.

Keywords: laser reflection point clouds, TLO, neural network model, classification, laser scanning

Введение

В условиях быстрого развития систем компьютерного зрения значительно возрастает потребность в пространственных данных, необходимых для обучения нейросетевых моделей [1,2]. Для задач автоматизированного распознавания объектов недвижимости под набором пространственных данных для эталонной разметки подразумеваются облака точек лазерного отражения (ТЛО), полученные методом лазерного сканирования, отличающиеся высокой плотностью, минимальной погрешностью измерения точек и достоверным отображением геометрии объектов. Собранные облака ТЛО используются для дальнейшей семантической классификации и формирования эталонной разметки, необходимой для обучения и валидации нейросетевых моделей [3-6].

Несмотря на активное развитие моделей машинного обучения, их эффективность при автоматизированной классификации объектов напрямую зависит от качества исходных данных, на которых они были обучены. Недостаточно плотные, шумные или нерепрезентативные данные приводят к ухудшению качества классификации и ошибкам в распознавании объектов. Поэтому особое значение приобретает разработка и применение стандартизированных методик получения облаков ТЛО, позволяющих формировать надежные эталонные выборки для обучения нейросетевых моделей [7-10].

Целью работы является разработка методик сбора данных в виде облаков ТЛО для последующего использования в задачах машинного обучения при автоматизированном распознавании объектов.

Для достижения поставленной цели в работе решаются следующие задачи:

- выполнить лазерное сканирование территории с соблюдением требований к нормативной точности, плотности точек и репрезентативности данных, необходимых для формирования эталонного набора облаков ТЛО;
- провести обработку и анализ размеченных облаков ТЛО;
- исследовать проблему разбалансировки классов при сборе облаков ТЛО и предложить подходы, направленные на решение этой проблемы.

Методы и материалы

Лазерное сканирование выполнялось с помощью прибора Leica BLK360 территории КубГТУ на ул. Московская, 2. В ходе работы были задействованы студенты направления «Землеустройство и кадастры» 21.03.02 и 21.04.02 в количестве 10 человек в рамках проведения учебных практик. Перед началом лазерного сканирования студенты прошли инструктаж по работе с оборудованием и ознакомление с программным обеспечением для предварительной обработки результатов лазерного сканирования (рис. 1).



Рис. 1. Выполнение этапа полевых работ

Результаты

В ходе проведенного лазерного сканирования был сформирован массив геопространственной информации. Общий объем данных составил около 200 ГБ, площадь съемки – 10 га, а общее число точек стояния сканера достигло 400.

После завершения лазерного сканирования и предварительной обработки, облако ТЛО было вручную классифицировано по семантическим признакам, что позволило сформировать эталонный набор данных для последующего обучения нейросетевой модели. Разметка данных осуществлялась в программном обеспечении CloudCompare. Особое внимание уделялось единообразию критериев классификации, поскольку точность и согласованность разметки напрямую влияют на качество формирования эталонного набора данных.

Каждый сформированный класс проходил процедуру проверки двумя независимыми операторами. В случаях обнаружения ошибки проводилось дополнительное согласование и корректировка классификационных точек. Верификация обеспечила повышение достоверности разметки, а также позволил сформировать

эталонный набор данных для обучения и последующей объективной оценки нейросетевой модели. В результате на основе разметки была подготовлена тренировочная выборка и выполнено предварительное обучение модели.

Анализ распределения классов в размеченном облаке ТЛО показал наличие выраженной разбалансировки. Это наглядно демонстрируется на гистограмме распределения классов (рис. 2), где видна значительная диспропорция между частыми и редкими объектами. Разбалансировка классов происходит из естественного неравенства площадей поверхности и объемов объектов.

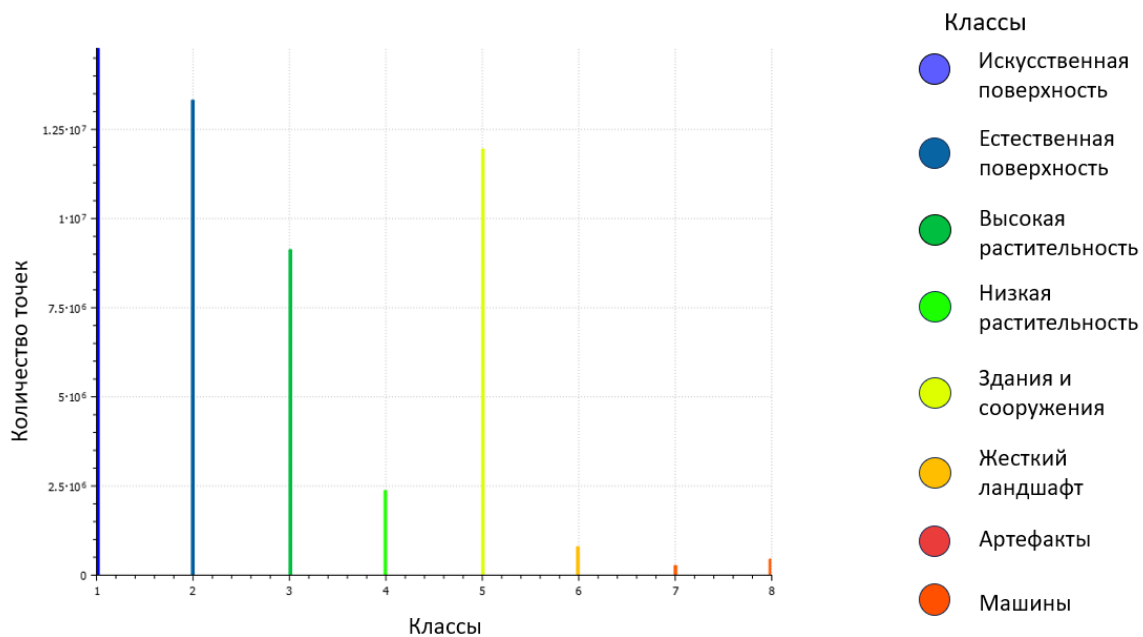


Рис. 2. Гистограмма классов в облаке ТЛО

Результаты предварительного обучения модели по эпохам (рис. 3) подтвердили влияние данной диспропорции: точность распознавания редких классов снижена и наблюдаются нестабильные значения метрик. Это связано с тем, что при сильной разбалансировке модель склонна «переучиваться» на наиболее распространенные классы и игнорировать малочисленные.

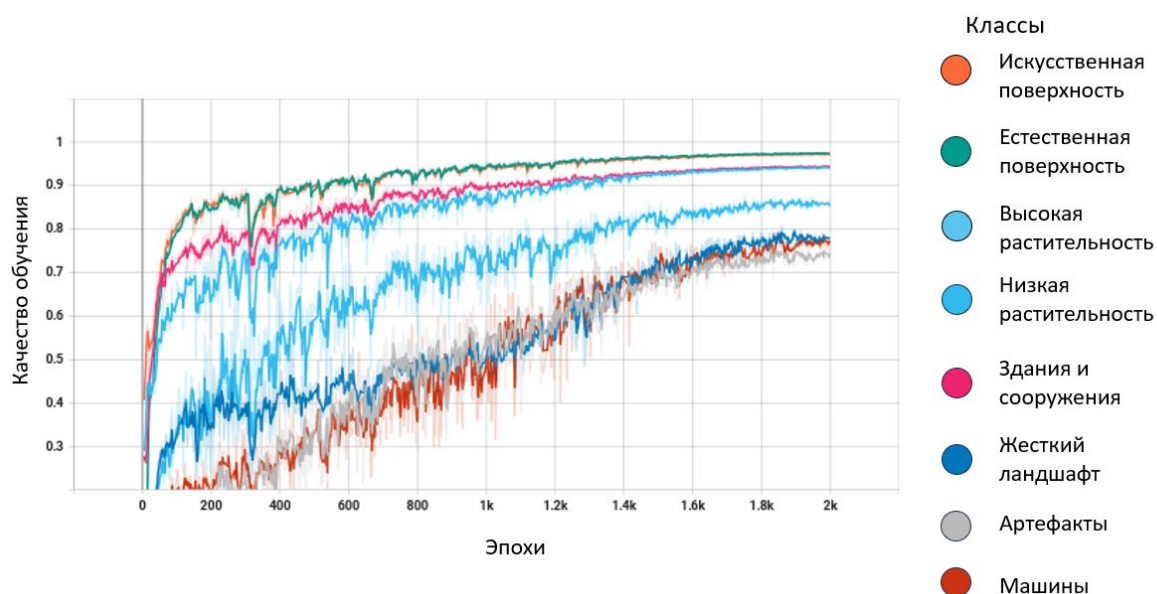


Рис. 3. График обучения 8 классов по эпохам

Таким образом, в результате проведенного исследования была выявлена необходимость применения специализированных методов устранения разбалансировки, направленных на более равномерное представление классов в обучающем наборе и повышение качества классификации редких объектов. Ниже приведены основные подходы, которые могут использоваться для решения данной проблемы в задачах классификации облаков точек (рис. 4).

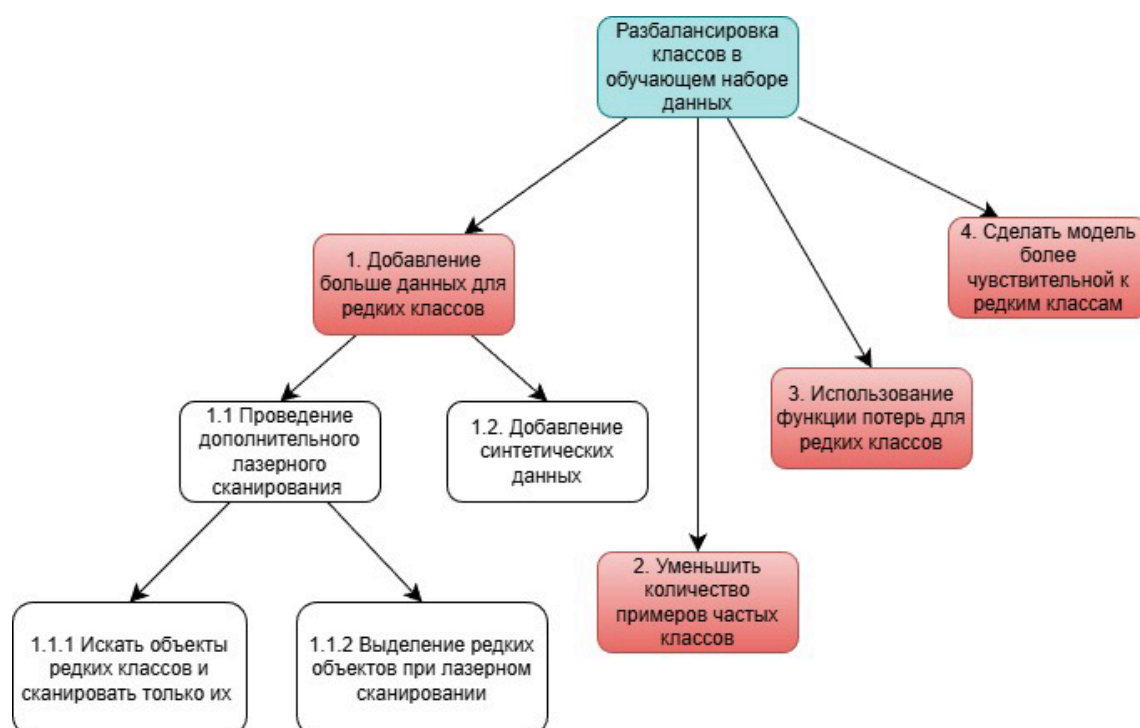


Рис. 4. Схема подходов к решению проблемы разбалансировки классов в обучающем наборе данных

Схема отражает совокупность подходов, направленных на устранение разбалансировки классов в обучающих данных для задач классификации облаков ТЛО. Основная идея заключается в том, чтобы обеспечить более равномерное представление используемых классов, поскольку недостаток примеров редких объектов неизбежно приводит к снижению качества их распознавания. Для этого используют методы расширения информации, такие как дополнительный сбор данных или генерация синтетических выборок. Также применяют подходы, направленные на уменьшение количества примеров наиболее распространенных классов, что помогает выровнять распределение объектов.

Обсуждение

Помимо манипуляций с данными, важную роль играют алгоритмические приемы, повышающие значимость редких классов в процессе обучения. К ним относятся функции потерь и адаптация модели таким образом, чтобы она более эффективно воспринимала и различала малочисленные классы. В совокупности эти решения позволяют компенсировать разбалансировку, вызванную неравномерной структурой выборки, и обеспечить более устойчивые результаты классификации, особенно в отношении редких и трудноразличимых объектов.

Заключение

Таким образом, проведенное исследование позволило подготовить эталонный набор данных для последующего обучения нейросетевой модели. Выполненные процедуры сканирования, предварительной обработки и семантической разметки обеспечили получение достоверного и репрезентативного материала.

Анализ результатов показал наличие выраженной разбалансировки классов, что снижает качество распознавания малочисленных объектов и требует применения специализированных методов решения проблемы разбалансировки классов. Полученные выводы подтверждают важность корректной подготовки облаков ТЛО и демонстрируют перспективность предложенной методики для дальнейшего использования в системах машинного обучения для целей геодезической и кадастровой деятельности.

Благодарности

Исследование выполнено при финансовой поддержке Кубанского научного фонда в рамках проекта № ЛАБ-24.1/2.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Гура Д. А., Дьяченко Р.А., Тихонов Т. А. Применение алгоритма CANUPO для классификации точек лазерного отражения // Интерэкспо Гео-Сибирь. – 2025. – Т. 4. – С. 210-216. – DOI 10.33764/2618-981X-2025-4-210-216.
2. Чжан Ч., Сахарова Е. К., Гришин И. А. Классификация пород деревьев с использованием моделей глубокого обучения на основе MLP // Искусственный интеллект в автоматизированных системах управления и обработки данных : Сборник статей II Всероссийской научной конференции: в 5 томах, Москва, 27–28 апреля 2023 года. – Москва: КДУ, Добросвет, 2023. – С. 377-383.

3. Сигаева Д. В., Логинов М. С. Влияние качества исходного набора данных для машинного обучения на точность диагноза // *Scientist (Russia)*. – 2022. – № 4(22). – С. 30.
4. Акимов А. А., Валитов Д. Р., Кубряк А. И. Предварительная обработка данных для машинного обучения // *Научное обозрение. Технические науки*. – 2022. – № 2. – С. 26-31. – DOI 10.17513/srts.1391.
5. Исмагилов А. В. Экспериментальное исследование влияния обучающих данных на производительность алгоритмов машинного обучения в задачах компьютерного зрения // *Инновации и инвестиции*. – 2024. – № 5. – С. 356-361.
6. Волошинов Д. В., Чернова А. А., Гвоздева А. В. Технология реконструкции и классификации трехмерных объектов с использованием облаков точек и искусственной нейронной сети // *Информатизация и связь*. – 2021. – № 8. – С. 66-71.
7. Джидид А. Д. Обзор методов сегментации и классификации облака точек архитектурных объектов // *Известия высших учебных заведений. Геодезия и аэрофотосъемка*. – 2019. – Т. 63, № 1. – С. 52-59. – DOI 10.30533/0536-101X-2019-63-1-52-59.
8. Федорова А. А., Нурлыева Д. Д., Желанкина А. С. Сравнение методов классификации облаков точек, полученных с помощью технологии lidar // *Искусственный интеллект в автоматизированных системах управления и обработки данных : Сборник статей Всероссийской научной конференции. В 2-х томах, Москва, 27–28 апреля 2022 года. Том 1*. – Москва: Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет), 2022. – С. 262-267.
9. Мсаллам М., Сырякин В. И. Повышение производительности классификации трехмерных облаков точек за счет увеличения данных // *Интеллектуальные системы 4-й промышленной революции : Сборник материалов V Международного форума*, Томск, 24–25 ноября 2022 года / Под редакцией В.И. Сырякина. – Томск: Общество с ограниченной ответственностью "СТТ", 2023. – С. 70-72.
10. Василенко Д. В. Разработка алгоритма классификации плотных облаков точек на примере городской застройки // *Вестник СГУГиТ (Сибирского государственного университета геосистем и технологий)*. – 2024. – Т. 29, № 6. – С. 44-52. – DOI 10.33764/2411-1759-2024-29-6-44-52.

© Д. А. Гура, Р. А. Дьяченко, Т. А. Тихонов, 2026