УДК 519.254

Очистка полуструктурированных и неструктурированных данных дистанционного зондирования Земли

Анатолий Дмитриевич Хомоненко[1, 2] — докт. техн. наук, профессор; профессор кафедры «Информационные и вычислительные системы»[1]; профессор кафедры математического и программного обеспечения[2]; область научных интересов: информационные системы, обработка больших данных, вероятностное моделирование геоинформационных систем, генетические алгоритмы, информационная безопасность. E-mail: khomon@mail.ru

Кириенко Андрей Борисович[2] — аспирант кафедры математического и программного обеспечения; область научных интересов: информационные системы, базы данных, обработка больших данных, вероятностное моделирование информационных систем, информационная безопасность. E-mail: anbokir@mail.ru

Злобин Сергей Евгеньевич[2] — адъюнкт кафедры математического и программного обеспечения; область научных интересов: информационные системы, базы данных, обработка больших данных, системы искусственного интеллекта; E-mail: zlobincergey15@gmail.com

Давыдова Даяна[1] — магистр; ассистент кафедры «Информационные и вычислительные системы»; область научных интересов: информационные системы, обработка больших данных. E-mail: dayana-0820@bk.ru

[1] Петербургский государственный университет путей сообщения Императора Александра I, Россия, 190031, Санкт-Петербург, Московский пр., 9

[2] Военно-космическая академия имени А. Ф. Можайского, Россия, 197198, Санкт-Петербург, ул. Ждановская, 13

Для цитирования: Хомоненко А. Д., Кириенко А. Б., Злобин С. Е., Давыдова Д. Очистка полуструктурированных и неструктурированных данных дистанционного зондирования Земли // Интеллектуальные технологии на транспорте. 2024. № 4 (40). С. 67–77. DOI: 10.20295/2413-2527-2024-440-67-77

Аннотация. *Затрагивается проблема обработки полуструктурированных и неструктурированных данных дистанционного зондирования Земли, полученных различными способами, включая спутники и беспилотные летательные аппараты.* ***Целью статьи*** *является исследование и реализация алгоритмов для эффективной очистки и предобработки полуструктурированных и неструктурированных данных дистанционного зондирования Земли. Подчеркивается важность очистки этих данных от шума, артефактов и ошибок для повышения их точности и значимости в прикладных научных исследованиях и практическом применении. Рассматриваются ключевые методики предобработки данных, включая удаление шума, коррекцию искажений, классификацию, сегментацию и стандартизацию данных, теоретические положения подкрепляются практическими примерами на Python с использованием таких библиотек, как GDAL, OpenCV и scikit-image. Приводятся примеры обнаружения воздушно-космических и транспортных объектов при помощи машинного обучения и глубокого обучения, подчеркивается значимость метрик точности, полноты и F1-Score при оценке качества очистки данных.* ***Практическая значимость*** *исследования заключается в оценке эффективности методов очистки данных, используемых для восстановления изображений при дистанционном зондировании Земли.*

Ключевые слова: *дистанционное зондирование Земли, предобработка данных, очистка данных, машинное обучение, глубокое обучение, фильтрация шума, коррекция искажений, классификация данных, сегментация изображений, воздушно-космические объекты, GDAL, OpenCV, scikit-image*

1.2.1 *— искусственный интеллект и машинное обучение (технические науки),* ***1.2.2*** *— математическое моделирование, численные методы и комплексы программ (технические науки),* ***2.9.8*** *— интеллектуальные транспортные системы (технические науки)*

Введение

Данные, полученные от авиации, космических аппаратов, беспилотных летательных аппаратов и других источников, различного типа датчиков, сенсоров и средств измерений, находящихся в космическом и воздушном пространстве, без их дополнительной предобработки и очистки зачастую непригодны для дальнейшего использования и анализа. Это вызвано различными факторами, такими как космическая радиация, технические неполадки и неисправности, ограничения оборудования, влияние погодных условий и различных радиоэлектронных средств, вызывающих помехи.

Очистка полуструктурированных и неструктурированных данных является важным этапом обработки, служащим для их исправления или удаления из набора некорректных, неполных, неформатированных или дублированных данных.

В контексте многоагентного управления космическими средствами корректно очищенные данные критически важны для точности и эффективности исполнения задач. Использование принципа многоагентных систем направлено на повышение эффективности, оперативности и боевой устойчивости космических систем за счет перераспределения функций управления между бортовым и наземным комплексами управления. Для получения точной и значимой информации для анализа и принятия решений в контексте многоагентного управления космическими средствами данные необходимо очищать от шума, артефактов и ошибок.

Ключевые аспекты и методики для очистки и предобработки данных

Целью восстановления изображения является восстановление скрытого чистого изображения x из его ухудшенного наблюдения:

$$y = x \otimes k \downarrow s + n, \qquad (1)$$

где $\otimes$ — двумерная свертка x с ядром размытия k; $\downarrow s$ — стандартная s-кратная понижающая дискретизация, то есть сохранение левого верхнего пикселя для каждого отдельного участка размером $s \times s$ и отбрасывание остальных; n — аддитивный белый гауссовый шум (additive white Gaussian noise, AWGN), определяемый стандартным отклонением (или уровнем шума) [1].

Формула (1) является моделью ухудшения и одновременно представляет общую модель сверхразрешения одиночного изображения, так как служит для описания процесса восстановления более качественного изображения на основе его низкоразрешенной версии.

Методы восстановления изображения (image restoration, IR) делятся на две основные категории, а именно: методы, основанные на моделях, и методы, основанные на обучении. Методы, основанные на моделях, направлены на замену поврежденных участков изображения реалистичными фрагментами. Методы, основанные на обучении, направлены на восстановление поврежденных фрагментов изображения с помощью нейронных сетей.

Методы, основанные на обучении, обычно моделируются как следующая задача двухуровневой оптимизации [2]:

$$\begin{cases} \min\limits_{\theta} \sum\limits_{i=1}^{n} \ell(\hat{x}_i, x_i), \\ \text{так, что } \hat{x}_i = f(y_i, \theta), \end{cases}$$

где θ — обучаемые параметры; $\ell(\hat{x}_i, x_i)$ измеряет потерю полученного чистого изображения $\hat{x}_i$ по отношению к истинному изображению x_i.

На рис. 1 показана иллюстрация связей между методами, основанными на глубоком обучении, методами, основанными на модели, методами глубокого PnP и глубокой развертки. Под методами глубокого PnP понимаются методы, объединяющие принципы глубокого обучения и концепцию Plug and Play (использования готовых моделей) [1].

Для методов *на основе моделей* обучение не требуется, обеспечиваются экономия времени, высокая гибкость, и имеет место явная интерпретация благодаря использованию предварительно заданных параметризованных моделей.

Для методов *на основе обучения* требуется обучение, обеспечиваются высокая скорость и эффективность, но имеет место ограниченная гибкость.

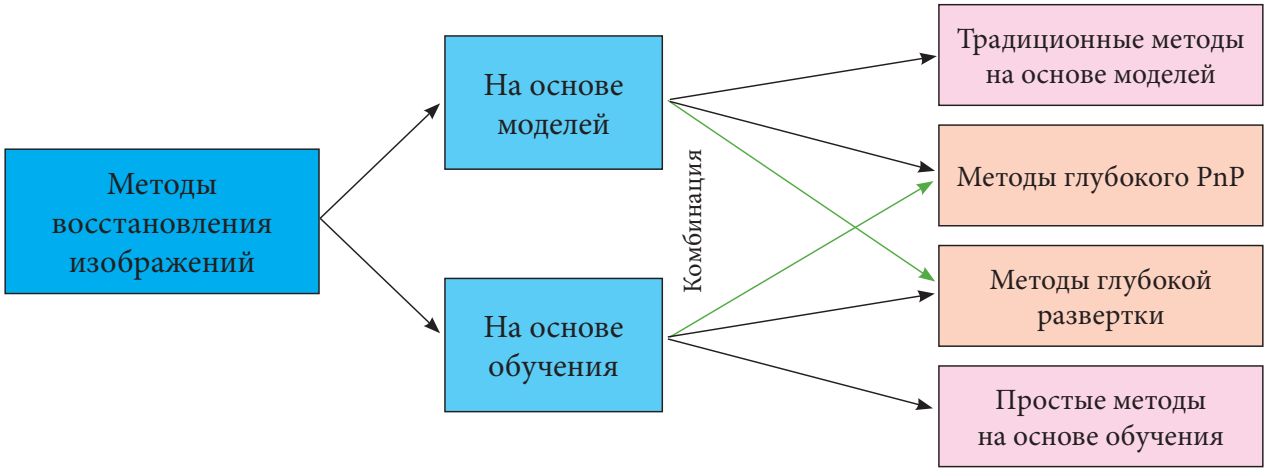

Рис. 1. Отношения между методами глубокого обучения, моделирования, глубокого PnP и глубокой развертки

Перечислим ключевые аспекты и методики, используемые для очистки и предобработки таких данных на основе моделей:

1. *Удаление шума и ошибок*. Применяются различные фильтры (например, медианные, гауссовые, ранжирующие, фильтры Винера, сглаживающие фильтры, спецификации сверточных нейронных сетей) для сглаживания изображений и устранения шума [3, 4]. Используются методы эрозии и дилатации для уточнения форм объектов на изображении.

2. *Коррекция искажений и размытий.* Геометрическая коррекция: преобразование данных для компенсации искажений, связанных с перспективой съемки, вращением Земли, помех, возникших в результате тряски беспилотного летательного аппарата, и другими факторами. Радиометрическая коррекция: коррекция различий в освещении, вызванных облачностью, углом солнца и другими факторами. Для повышения резкости изображений обычно применяются два метода: фильтрация изображений или деконволюция [5, 6].

3. *Классификация и сегментация.* Алгоритмы классификации изображений применяются для решения широкого круга проблем, таких как сегментация изображений и обнаружение изменений на изображениях, для мониторинга и контроля, анализа данных [7]. Применяются методы супервизионного и несупервизионного обучения: применение алгоритмов машинного обучения для классификации типов поверхности и других характеристик на основе обучающих данных или без них. Глубокое обучение: использование сверточных нейронных сетей для сегментации изображений и выделения интересующих объектов [8, 9].

4. *Удаление нерелевантной информации.* Маскирование: отсечение нерелевантных областей, таких как водные объекты, когда целью является анализ суши.

5. *Форматирование и стандартизация данных.* Приведение к единому масштабу: установка одинакового разрешения для всех наборов данных. Соответствие стандартам: обеспечение совместимости данных с общепринятыми форматами и протоколами для последующего анализа.

6. *Повышение точности геопространственной привязки.* Геопространственные данные от различных источников в виде спутниковых снимков, данных с беспилотных летательных аппаратов и результаты наземных съемок должны точно соответствовать друг другу. Некорректная привязка таких данных может привести к ошибкам в анализе, особенно когда геоданные накладываются друг на друга (например, границы земельных участков или маршруты инфраструктур). Повышение точности привязки помогает избежать таких ошибок и обеспечивает более точное позиционирование объектов.

Процесс очистки и предобработки данных дистанционного зондирования Земли (ДЗЗ) должен адаптироваться к специфике задачи и требованиям конкретного исследования. При правильном

подходе эти шаги значительно повышают качество и полезность собранных данных для анализа и принятия решений в областях экологии, городского планирования и многих других.

Для демонстрации приведем примеры кода на языке Python, которые используются для обработки и очистки данных ДЗЗ. Эти примеры включают использование библиотеки GDAL для работы с геопространственными данными и библиотеки OpenCV для обработки изображений. Некоторые из этих задач могут также включать применение библиотеки scikit-image для специализированных задач обработки изображений.

При проведении исследования использовалась ОС Ubuntu для установки библиотеки GDAL.

Установка библиотек:

```
sudo apt-get update
sudo apt-get install gdal-bin libgdal-dev
```

В дополнение может потребоваться установка специфической версии GDAL для соответствия версии библиотеки, установленной через apt. Чтобы найти версию GDAL, необходимо ввести команду «gdalinfo — version», а затем установить нужную версию Python-пакета, при этом важно, чтобы была установлена библиотека numpy и отключена изоляция сборки gdal:

```
pip install numpy
pip install -U setuptools wheel
pip install --no-build-isolation --no-cache-dir --force-reinstall gdal==<Версия>
pip install opencv-python scikit-image
```

Примеры обработки и очистки данных дистанционного зондирования Земли

Пример 1. Маскирование облачности.

Приведенный пример демонстрирует один из вариантов предобработки данных дистанционного зондирования Земли, а именно маскирование облачности, которое позволяет повысить точность и оперативность анализа полученных данных, а также оптимизировать работу алгоритмов машинного обучения.

Библиотека GDAL используется для чтения растрового изображения и OpenCV для создания маски облачности.

```
from osgeo import gdal
import cv2
import numpy as np
# Загрузка растрового изображения (указывается абсолютный либо относительный путь к файлу и необязательный аргумент, описывающий режим открытия — по умолчанию «только для чтения»):
dataset = gdal.Open('path_to_your_raster.tif')
band = dataset.GetRasterBand(1)
# Предполагается, что облака находятся в первом канале (модуль GDAL позволяет получить доступ как к любому каналу растра, так и ко всему растру сразу).
# Чтение данных в numpy массив:
image = band.ReadAsArray()
# Простая бинаризация для маскирования облачности:
_, cloud_mask = cv2.threshold (image, 200, 255, cv2.THRESH_BINARY)
# Предполагается, что облачность ярче (меняя параметр «200», можно установить минимальную яркость облаков).
# Применение маски к изображению (image):
masked_image = cv2.bitwise_and (image, image, mask=cloud_mask)
# Сохраняем результат:
cv2.imwrite ('masked_image.tif', masked_image)
```

Итог представлен на рис. 2.

Пример 2. Фильтрация шума с использованием медианного фильтра.

Приведенный пример описывает работу медианного фильтра. Преимущества медианного фильтра [10]:

• Эффективность в снижении шума: хорошо справляется с солью и перцем (шумом), сохраняя при этом края изображений. Это делает его более эффективным для некоторых типов шума по сравнению с гауссовыми фильтрами, которые могут размывать края.

а)

б)

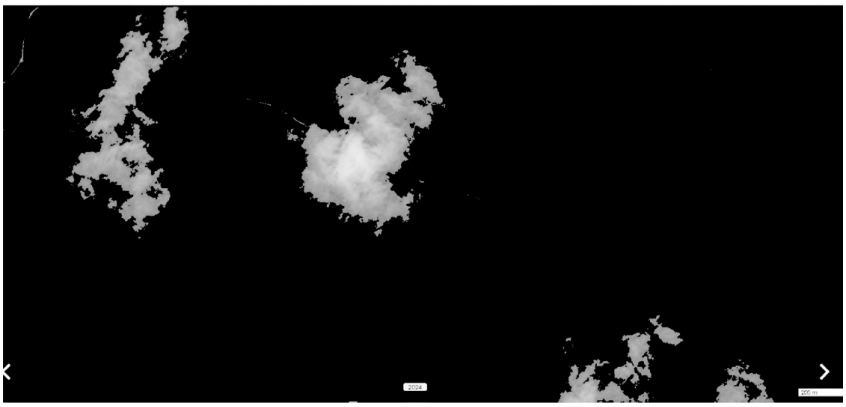

Рис. 2. а) исходное растровое изображение земной поверхности; б) результат

• Не меняет размеры: в отличие от некоторых других фильтров сохраняет размеры выходного изображения такими же, как и у входного, что удобно в обработке.

• Простота реализации: реализация в OpenCV достаточно проста и оптимизирована для работы с большими изображениями и в реальном времени.

• Не чувствителен к выбросам: не подвержен влиянию крайних значений (выбросов) в данных, что делает его более надежным для улучшения качества изображения.

```
import cv2
import numpy as np
# Чтение изображения (загружает изображение по указанному пути в градациях серого (черно-белом формате)):
image = cv2.imread (‘path_to_your_image.tif’, cv2.IMREAD_GRAYSCALE)
# Применение медианного фильтра к загруженному изображению с размером ядра 5, что помогает уменьшить шум в изображении и сохранить его края:
filtered_image = cv2.medianBlur (image, 5)
# Сохраняем результат:
cv2.imwrite (‘filtered_image.tif’, filtered_image)
```

Результат приведен на рис. 3.

Пример 3. Геометрическая коррекция (спецификация может зависеть от характера искажений).

Геометрическая коррекция является необходимым этапом обработки изображений ДЗЗ, обеспечивая надежность и точность полученных данных для последующего анализа и применения, например, приведения к единой системе координат, устранения искажений, повышения точности анализа, сравнения изображений во времени, сопоставления с векторными данными, упрощения проведения анализов.

Этот пример требует предварительного анализа для определения параметров коррекции, который обычно делается на основе контрольных точек, а также адаптации под конкретные задачи и данные.

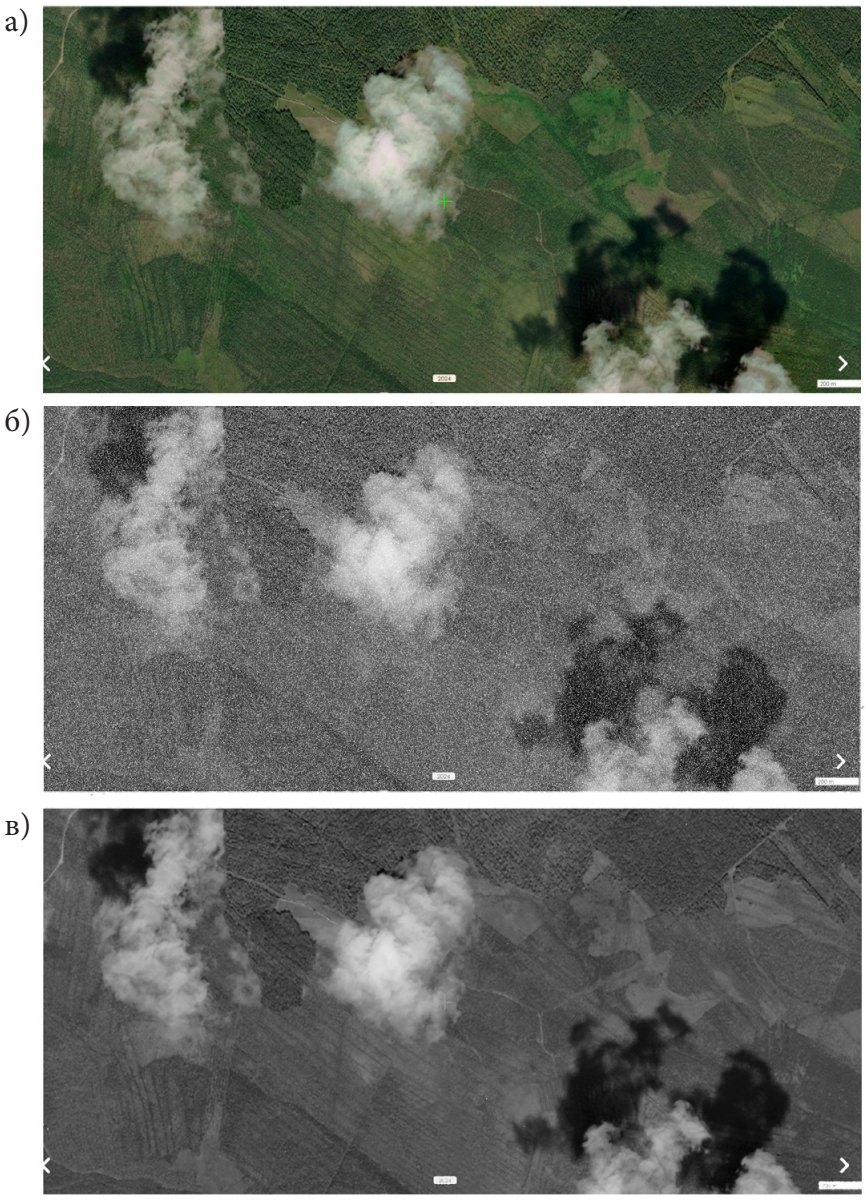

Рис. 3. а) исходное растровое изображение поверхности Земли;
б) изображение с наложением импульсного шума, в) очищенное от шума изображение

```
from osgeo import gdal, osr
# Загрузка исходного изображения по указанному пути:
dataset = gdal.Open ('path_to_your_distorted_image.tif')
# Определение новой геопривязки и системы координат (пример). Переменная geotransform задает параметры геопривязки изображения (положение, масштаб, угол наклона):
geotransform = (446720, 30, 0, 4611320, 0, -30)
srs = osr.SpatialReference()
# Пример использования EPSG кода WGS84 (создается объект srs (система координат) и импортируется система координат EPSG:4326 (WGS84)):
srs.ImportFromEPSG(4326)
# Создание нового файла (с использованием драйвера GTiff создается новый файл corrected_image.tif с теми же размерностями и типами данных, что и у исходного изображения; задаются геопривязка и проекция для нового изображения):
driver = gdal.GetDriverByName('GTiff')
out_ds = driver.Create('corrected_image.tif', dataset.RasterXSize, dataset.RasterYSize, dataset.RasterCount, dataset.GetRasterBand(1).DataType)
```

```
out_ds.SetGeoTransform(geotransform)
out_ds.SetProjection(srs.ExportToWkt())
# Копирование данных из исходного файла в корректированный (в цикле для каждого растра исходного изображения данные копируются в соответствующий растр нового файла):
for i in range(dataset.RasterCount):
    out_ds.GetRasterBand(i+1).WriteArray(dataset.GetRasterBand(i+1).ReadAsArray())
out_ds.FlushCache()  # Сброс в файл
```

Пример 4. Определение воздушно-космических объектов в данных, полученных с помощью ДЗЗ, с использованием библиотек машинного и глубокого обучения в Python.

Часто требуется распознавание и определение воздушно-космических объектов на полученных в результате ДЗЗ изображениях. Рассмотрим задачи обнаружения аэропортов и идентификации типов транспортных объектов на спутниковых снимках, приведем данные, которые можно использовать для этих целей, и предложим общую схему их решения с использованием библиотек машинного и глубокого обучения в Python.

4.1. Обнаружение аэропортов на спутниковых снимках.

Данные: использовались открытые спутниковые изображения высокого разрешения, например, предоставляемые сервисом Sentinel или Landsat. Google Earth API использовался для получения изображений аэропортов со всего мира для создания обучающего датасета.

Задача: разработать систему, способную автоматически обнаруживать и классифицировать аэропорты на спутниковых изображениях.

Решение: метод глубокого обучения с использованием сверточных нейронных сетей (CNN).

Необходимые классы для импорта:

- Sequential: модель, которая позволяет строить нейронные сети последовательным образом, добавляя слои один за другим.
- Conv2D: слой свертки, применяемый для обработки изображений.
- MaxPooling2D: слой подвыборки, который уменьшает размерность изображения.
- Flatten: преобразует многомерные данные в одномерный вектор.
- Dense: полносвязный слой, где каждый нейрон соединен со всеми нейронами предыдущего слоя.

Код с использованием библиотеки Keras:

```
#Импорт необходимых классов:
from tensorflow.keras.models import Sequential
#Импорт необходимых классов:
from tensorflow.keras.layers import Conv2D, MaxPooling2D, Flatten, Dense
# Создание модели CNN:
model = Sequential
([Conv2D (32, (3, 3), activation='relu', input_shape=(64, 64, 3)). # Создает первый слой свертки с 32 фильтрами размером 3 × 3. Использует активационную функцию ReLU (Rectified Linear Unit) для введения нелинейности. input_shape = (64, 64, 3) указывает, что входные изображения имеют размер 64 ×64 пикселя и 3 цветовых канала (RGB).
MaxPooling2D (2, 2). #Применяет подвыборку размером 2 × 2, что уменьшает размерность изображения в два раза и уменьшает вычислительные затраты.
Conv2D (64, (3, 3), activation='relu'). #Добавляет второй слой свертки с 64 фильтрами также размером 3 × 3 и функцией активации ReLU.
MaxPooling2D (2, 2). #Применяет второй слой подвыборки таким же образом.
Flatten (). # Преобразует выходные данные из 2D в 1D, подготавливая их для входа в полносвязный слой.
Dense (128, activation='relu'). # Создает полносвязный слой с 128 нейронами и активацией ReLU.
Dense (1, activation='sigmoid'). # На выходе создает слой с одним нейроном и функцией активации сигмоид для бинарной классификации, где выходное значение будет интерпретироваться как вероятность.
# Компиляция модели:
model.compile(optimizer='adam', loss='binary_crossentropy', metrics=['accuracy'])
# Модель готова к тренировке на данных изображений аэропортов. Параметры:
```

optimizer=’adam’ # Используется алгоритм оптимизации Adam, который адаптивно настраивает параметры обучения.

loss=’binary_crossentropy’ # Задается функция потерь для бинарной классификации, которая измеряет, насколько хорошо модель классифицирует входные данные.

metrics=[‘accuracy’] # Указывает, что модель будет отслеживать точность в процессе обучения.

4.2. Классификация типов состояния местности.

Данные: для анализа состояния местности использовались многополосные спутниковые снимки, например с Sentinel-2, которые включают инфракрасные каналы, полезные для дифференциации растительности.

Задача: идентификация и классификация местности на основе спектральных сигнатур.

Решение: метод машинного обучения с использованием классификаторов на основе случайного леса или градиентного бустинга.

Пример кода с использованием Scikit-learn:

```
from sklearn.ensemble import RandomForestClassifier # Импортирует класс RandomForestClassifier, который используется для создания модели случайного леса.
from sklearn.model_selection import train_test_split # Импортирует функцию для разделения данных на обучающую и тестовую выборки.
from sklearn.metrics import accuracy_score # Импортирует функцию для оценки точности модели.
import numpy as np
# Загрузка данных
# X — спектральные данные транспортных объектов, y – метки классов культур:
X, y = load_your_data()
# Разделение данных на обучающую и тестовую выборки:
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42);
# train_test_split() — функция, которая разделяет данные на две части: обучающую и тестовую;
# test_size=0.3 указывает, что 30 % данных будет использовано для тестирования, а 70 % — для обучения;
# random_state=42 задает фиксированное значение для генератора случайных чисел, обеспечивая воспроизводимость разделения данных.
# Обучение модели случайного леса:
clf = RandomForestClassifier(n_estimators=100, random_state=42) # Создается объект модели случайного леса.
#n_estimators=100 # Указывает, что модель будет состоять из 100 деревьев решений.
#random_state=42 # Обеспечивает воспроизводимость.
clf.fit(X_train, y_train) # Обучение модели (fit) на обучающих данных X_train и их метках y_train. Во время этой процедуры модель настраивает свои внутренние параметры на основе обучающих данных.
# Оценка модели:
y_pred = clf.predict(X_test)
print(f’Accuracy: {accuracy_score(y_test, y_pred)}’)
```

В обоих примерах кода центральным аспектом является подготовка данных: извлечение признаков, аннотация, разбиение на обучающую и тестовую выборки. Для обнаружения воздушно-космических объектов и классификации признаков на Земле данные ДЗЗ являются источником информации, который можно эффективно обрабатывать с помощью методов машинного и глубокого обучения.

Проведение ДЗЗ с помощью авиации предлагает уникальные возможности по сбору данных высокого разрешения. В отличие от спутникового ДЗЗ авиационный метод позволяет получать данные с более высокой частотой и меньшим влиянием атмосферы, однако также требует тщательной очистки данных от шумов, вызванных вибрацией, погодными условиями и другими факторами. Рассмотрим пример использования авиационного *ДЗЗ для мониторинга и анализа состояния местности*, в котором очистка данных может улучшить качество анализа.

Исходные данные: авиационные снимки местности, содержащие транспортные объекты.

Цели анализа: определить состояние местности в районе железнодорожных путей.

Параметры оценки:

• точность (Accuracy) — доля правильно классифицированных пикселей на всех снимках;

• полнота (Recall) — доля правильно определенных случаев состояния местности от количества фактических случаев;

• F1-Score — гармоническое среднее точности и полноты.

Проведение очистки

Допустим, что первоначальный анализ данных проводился без какой-либо предварительной очистки. Затем были применены следующие методы очистки данных:

• коррекция вибрации — удаление искажений, вызванных вибрацией камеры на борту;

• атмосферная коррекция — минимизация влияния атмосферных условий на качество снимков;

• фильтрация шума — применение спектральной фильтрации для устранения фонового шума.

Таблица 1

Результаты предварительной очистки данных мониторинга состояния местности

Показатели	До очистки	После очистки
Точность	70 %	85 %
Полнота	65 %	80 %
F1-Score	0,67	0,82

Вывод. Улучшение характеристик после очистки демонстрирует значительный выигрыш в качестве данных. Коррекция вибрации обеспечила более четкие изображения, благодаря чему улучшилась классификация состояний местности. Атмосферная коррекция снизила влияние тумана и других атмосферных эффектов, что подняло точность идентификации состояния местности. Фильтрация шума уменьшила количество ложных срабатываний, что положительно сказалось на полноте и F1-Score.

Пример иллюстрирует, как тщательная предобработка данных ДЗЗ может способствовать более точному и надежному анализу, особенно когда речь идет о критически важных применениях, таких как мониторинг условий эксплуатации железнодорожных объектов.

Заключение

Очистка и предобработка данных дистанционного зондирования Земли, полученных из космических и воздушных источников, представляет собой важный процесс, направленный на обеспечение их пригодности для дальнейшего анализа и принятия решений, в том числе в контексте многоагентного управления космическими средствами. Рассмотрены основные аспекты и методики, используемые для устранения шума, коррекции искажений, классификации и сегментации, а также удаления нерелевантной информации и стандартизации данных. В условиях современных вызовов и растущей сложности данных эффективные методологии очистки и предобработки становятся важнейшими инструментами, позволяющими обеспечить высокое качество анализа и оперативность обработки неструктурированных и слабоструктурированных данных в различных областях, в том числе позволяет более эффективно использовать возможности многоагентных систем. В дальнейшем углубление в развитие алгоритмов машинного обучения и автоматизации процессов предобработки будет способствовать более значительным изменениям в эффективности работы с данными.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Beyond a Gaussian denoiser: residual learning of deep CNN for image denoising / K. Zhang [et al.] // IEEE Transactions on Image Processing. 2017. P. 3142–3155.

2. Компьютерное зрение. Современные методы и перспективы развития / ред. Р. Дэвис, М. Терк; пер. с англ. В. С. Яценкова. М.: ДМК Пресс, 2022. 690 с.

3. Никитин Г. Ю. Повышение качества изображения на базе алгоритмов нейронных сетей // Компьютерные системы и сети: материалы 54-й научной конференции аспирантов, магистрантов и студентов (Минск, Беларусь, 23–27 апреля 2018 г.). Минск: Белорусский гос. ун-т информатики и радиоэлектроники, 2018. С. 242–244.

4. Булыга Ф. С., Курейчик В. М. Метод понижения шума на цифровых изображениях // Мировые научные исследования и разработки в эпоху цифровизации: материалы XV Международной научно-практической конференции (Ростов-на-Дону, Россия, 25 ноября 2021 г.): в 2 ч. Ч. 1. Ростов н/Д.: Изд-во Южного университета (ИУБиП), 2021. С. 143–147.

5. Кислянский Г. Н., Нестругина Е. С. Восстановление расфокусированных и смазанных изображений // Вестник Донецкого национального университета. Серия Г: Технические науки. 2020. № 4. С. 41–53.

6. Learning to Estimate and Remove Non-uniform Image Blur / F. Couzinié-Devy [et al.] // Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2013) (Portland, OR, USA, 23–28 June 2013). Institute of Electrical and Electronics Engineers, 2013. P. 1075–1082. DOI: 10.1109/CVPR.2013.143

7. Вершовский Е. А. Роевой алгоритм оптимизации в задаче кластеризации мультиспектрального снимка // Известия Южного федерального университета. Технические науки. 2010. № 5(106). С. 102–107.

8. Вершовский Е. А. Метод контроля качества кластеризации мультиспектрального изображения // Известия Южного федерального университета. Серия «Технические науки». 2010. № 7(108). С. 191–198.

9. Кузнецов А. А., Опарин А. Н., Шишкин В. А. Поиск и распознавание объектов на базисе нейросетевых алгоритмов и нейропроцессорных технологий // Прикладная физика. 2006. № 5. С. 97–100.

10. Сорокин С. В. Возможности и преимущества взвешенных медианных фильтров для удаления импульсного шума на изображении // Труды XIX Международного симпозиума «Надежность и качество» (Пенза, Россия, 26 мая — 01 июня 2014 г.). Т. 2. Пенза: Пензенский гос. ун-т, 2014. С. 203–204.

Дата поступления: 31.10.2024
Решение о публикации: 31.10.2024

Cleaning of Semi-Structured and Unstructured Earth Remote Sensing Data

Anatoly D. Khomonenko[1, 2] — Doctor of Technical Sciences, Professor; Professor of the Department of Information and Computing Systems[1]; Professor of the Department of Mathematics and Software[2]; research interests: information systems, big data processing, probabilistic modeling of geoinformation systems, genetic algorithms, information security. E-mail: khomon@mail.ru

Andrey B. Kirienko[2] — graduate student of the Department of Mathematics and Software; research interests: information systems, databases, big data processing, probabilistic modeling of information systems, information security. E-mail: anbokir@mail.ru

Sergey E. Zlobin[2] — Associate Professor of the Department of Mathematics and Software; research interests: information systems, databases, big data processing, artificial intelligence systems. E-mail: zlobincergey15@gmail.com

Dayana Davydova[1] — Master's Degree; Assistant of the Department of Information and Computing Systems; research interests: information systems, big data processing. E-mail: dayana-0820@bk.ru

[1] Emperor Alexander I St. Petersburg State Transport University, 9, Moskovsky pr., Saint Petersburg, 190031, Russia

[2] Mozhaisky Military Aerospace Academy, 13, Zhdanovskaya str., St. Petersburg, 197198, Russia

For citation: Khomonenko A. D., Kirienko A. B., Zlobin S. E., Davydova D. Cleaning of Semi-Structured and Unstructured Earth Remote Sensing Data // Intellectual Technologies on Transport. 2024. No. 4 (40). Pp. 67–77. DOI: 10.20295/2413-2527-2024-440-67-77 (In Russian)

Abstract. *The problem of processing semi-structured and unstructured Earth remote sensing (ERS) data obtained by various methods, including satellites and unmanned aerial vehicles, is touched upon.* ***The purpose*** *of the article is to study and implement algorithms for effective purification and preprocessing of semi-structured and unstructured Earth remote sensing data. The importance of cleaning these data from noise, artifacts and errors is emphasized in order to increase their accuracy and significance in applied scientific research and practical application. Key data preprocessing techniques are considered, including noise removal, distortion correction, classification, segmentation and standardization of data, reinforcing theoretical positions with practical examples in Python using libraries such as GDAL, OpenCV and scikit-image. Examples of detection of aerospace and transport objects using machine learning and deep learning are given, emphasizing the importance of accuracy, completeness and F1-Score metrics in assessing the quality of data purification.* ***The practical significance*** *of the study lies in evaluating the effectiveness of data purification methods used to restore images during remote sensing of the Earth.*

Keywords: *remote sensing of the Earth, data preprocessing, data purification, machine learning, deep learning, noise filtering, distortion correction, data classification, image segmentation, aerospace objects, GDAL, OpenCV, scikit-image*

REFERENCES

1. Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising / K. Zhang [et al.] // IEEE Transactions on Image Processing. 2017. P. 3142–3155.

2. Komp'yuternoe zrenie. Sovremennye metody i perspektivy razvitiya / red. R. Devis, M. Terk; per. s angl. V. S. Yacenkova. M.: DMK Press, 2022. 690 s. (In Russian)

3. Nikitin G. Yu. Povyshenie kachestva izobrazheniya na baze algoritmov nejronnyh setej // Komp'yuternye sistemy i seti: materialy 54-j nauchnoj konferencii aspirantov, magistrantov i studentov (Minsk, Belarus', 23–27 aprelya 2018 g.). Minsk: Belorusskij gos. un-t informatiki i radioelektroniki, 2018. S. 242–244. (In Russian)

4. Bulyga F. S., Kurejchik V. M. Metod ponizheniya shuma na cifrovyh izobrazheniyah // Mirovye nauchnye issledovaniya i razrabotki v epohu cifrovizacii: materialy XV Mezhdunarodnoj nauchno-prakticheskoj konferencii (Rostov-na-Donu, Rossiya, 25 noyabrya 2021 g.): v 2 ch. Ch. 1. Rostov n/D.: Izd-vo YUzhnogo universiteta (IUBiP), 2021. S. 143–147. (In Russian)

5. Kislyanskij G. N., Nestrugina E. S. Vosstanovlenie rasfokusirovannyh i smazannyh izobrazhenij // Vestnik Doneckogo nacional'nogo universiteta. Seriya G: Tekhnicheskie nauki. 2020. No. 4. S. 41–53. (In Russian)

6. Learning to Estimate and Remove Non-uniform Image Blur / F. Couzinié-Devy [et al.] // Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2013) (Portland, OR, USA, 23–28 June 2013). Institute of Electrical and Electronics Engineers, 2013. P. 1075–1082. DOI: 10.1109/CVPR.2013.143

7. Vershovskij E. A. Roevoj algoritm optimizacii v zadache klasterizacii mul'tispektral'nogo snimka // Izvestiya Yuzhnogo federal'nogo universiteta. Seriya “Tekhnicheskie nauki”. 2010. No. 5(106). S. 102–107. (In Russian)

8. Vershovskij E. A. Metod kontrolya kachestva klasterizacii mul'tispektral'nogo izobrazheniya // Izvestiya Yuzhnogo federal'nogo universiteta. Tekhnicheskie nauki. 2010. No. 7(108). S. 191–198. (In Russian)

9. Kuznecov A. A., Oparin A. N., SHishkin V. A. Poisk i raspoznavanie ob"ektov na bazise nejrosetevyh algoritmov i nejroprocessornyh tekhnologij // Prikladnaya fizika. 2006. No. 5. S. 97–100. (In Russian)

10. Sorokin S. V. Vozmozhnosti i preimushchestva vzveshennyh mediannyh fil'trov dlya udaleniya impul'snogo shuma na izobrazhenii // Trudy XIX Mezhdunarodnogo simpoziuma “Nadezhnost' i kachestvo” (Penza, Rossiya, 26 maya — 01 iyunya 2014 g.). T. 2. Penza: Penzenskij gos. un-t, 2014. S. 203–204. (In Russian)

Received: 31.10.2024
Accepted: 31.10.2024