

Методика формирования исходных данных для алгоритмов машинного обучения при идентификации элементов ландшафта

Виктор Мочалов,¹ Ольга Григорьева¹

¹Военно-космическая академия имени А.Ф.Можайского, г. Санкт-Петербург, 197198, Россия

Аннотация

Представлен типовой состав исходных данных для алгоритмов машинного обучения в части обучающих и контрольных выборок. Предложена методика обоснованного формирования исходных данных и пример её практической реализации. В качестве источника исходных данных приняты материалы отечественной мультиспектральной космической съёмки. Приведен пример автоматизированной идентификации элементов ландшафта и числовые значения показателей качества результатов обработки.

Ключевые слова

Мультиспектральная космическая съёмка, обучающая и контрольная выборки, семантическая сегментация, качество результатов обработки, цифровая карта

1. Введение

Развитие методов автоматизированной обработки данных дистанционного зондирования Земли является одним из основных направлений совершенствования новых информационных технологий, обеспечивающих оперативное решение актуальных практических задач. Часто возникает необходимость идентификации или семантической сегментации элементов ландшафта с последующим представлением результатов работ в виде цифровых топографических карт местности. Автоматизированная обработка материалов космической съёмки осуществляется, как правило, на основе алгоритмов машинного обучения, в том числе с использованием нейросетевых технологий. Успешная эксплуатация алгоритмов во многом зависит от исходных данных (ИД), в состав которых входят, в частности, обучающие и контрольные выборки. Вместе с тем ИД часто формируются из непроверенных источников. Кроме того, иногда считается, что расширение объёма ИД может привести к повышению качества результатов обработки материалов съёмки. Также не всегда осуществляется контроль достоверности данных, входящих как в обучающие, так и в контрольные выборки. Показатели качества результатов эксплуатации нейронных сетей часто определяются автоматически на основе заложенных в программные модули алгоритмов расчета обобщенных показателей. В данной работе предложена методика, опирающаяся как на проведённые теоретические исследования, так и на опыт решения практических задач обработки материалов мультиспектральной аэрокосмической съёмки. Представлена структура и основные этапы применения методики, апробированной при идентификации состояния лесных массивов. Обработывались материалы отечественной космической съёмки, выполненной с помощью космического аппарата «Канопус-В». Обеспечено формирование ИД, которые использовались для семантической сегментации элементов ландшафта. Результаты семантической сегментации представлены в виде фрагмента цифровой топографической карты.

7th International Workshop on Information, Computation, and Control Systems for Distributed Environments (ICCS-DE 2025), July 7–11, 2025, Irkutsk, Russia

EMAIL: vicavia@yandex.ru (A. 1); email2@mail.com (A. 2)

ORCID: 0009-0009-8553-1153 (A. 1); 0000-0001-6638-5204 (A. 2)



©2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

ICCS-DE 2025 Workshop Proceedings

DOI: 10.47350/ICCS-DE.2025.12

Проведено предварительное оценивание показателей качества результатов обработки и намечены пути их улучшения. Разработанная методика направлена на совершенствование технологий искусственного интеллекта, имеет важное значение при обработке пространственных данных и решении задач информационного обеспечения в различных областях практической деятельности.

2. Основная часть

2.1. Методический аппарат

Структурная схема методики формирования ИД для свёрточной нейронной сети (CNN) приведена на рисунке 1.

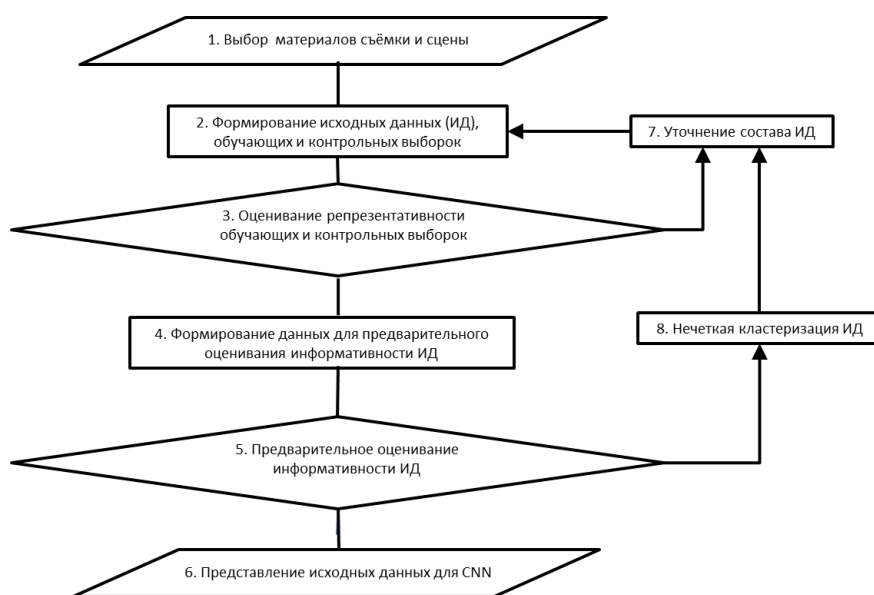


Рисунок 1: Структурная схема методики

На этапе 1 оператор осуществляет выбор сцены и её фрагмента в пределах условного кадра материалов съёмки с учётом содержания прикладной задачи. Фрагмент сцены содержит преимущественно идентифицируемые элементы ландшафта.

На этапе 2 оператор выбирает ИД. Состав ИД для двух элементов ландшафта приведен на рисунке 2. Исходные данные представляются в виде результатов измерения спектральных отражательных характеристик элементов ландшафта, имеющих строгую географическую привязку.

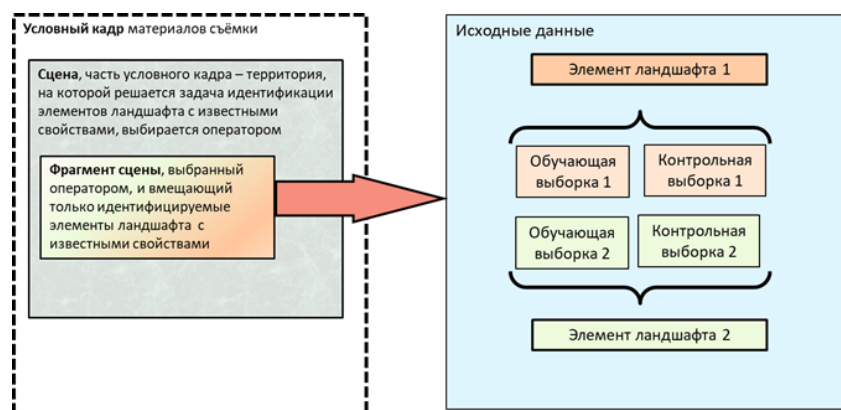


Рисунок 2: Состав исходных данных

На этапе 3 осуществляется оценивание репрезентативности ИД. Выборки считаются репрезентативными, если их статистические характеристики в части средних значений, дисперсии и среднеквадратичного отклонения отличаются от характеристик генеральной совокупности в пределах сцены не более чем на 5%. При невыполнении этого условия производится уточнение состава исходных данных.

На этапах 4 и 5 осуществляется формирование данных и предварительное оценивание показателей качества решения задачи идентификации элементов ландшафта. На основе результатов расчета информативных индексов или оцениваемых параметров для традиционных алгоритмов идентификации строятся гистограммы и с учётом планируемого размещения гиперпараметра осуществляется расчет базовых и обобщенных показателей качества решения задачи идентификации элементов ландшафта [1].

При невыполнении требований по информативности на этапе 8 выполняется операция корректировки ИД с помощью математического аппарата нечёткой кластеризации (FCM) [2].

2.2. Пример реализации методики

Методика апробирована на примере формирования ИД и идентификации трёх элементов ландшафта в пределах сцены, приведенной на рисунке 3. Рассматриваются материалы мультиспектральной космической съёмки, выполненной с помощью космического аппарата «Канопус-В» KV5_20109_19337_03_KANOPUS_20220812_081812_081849.SCN12.MS.L2.

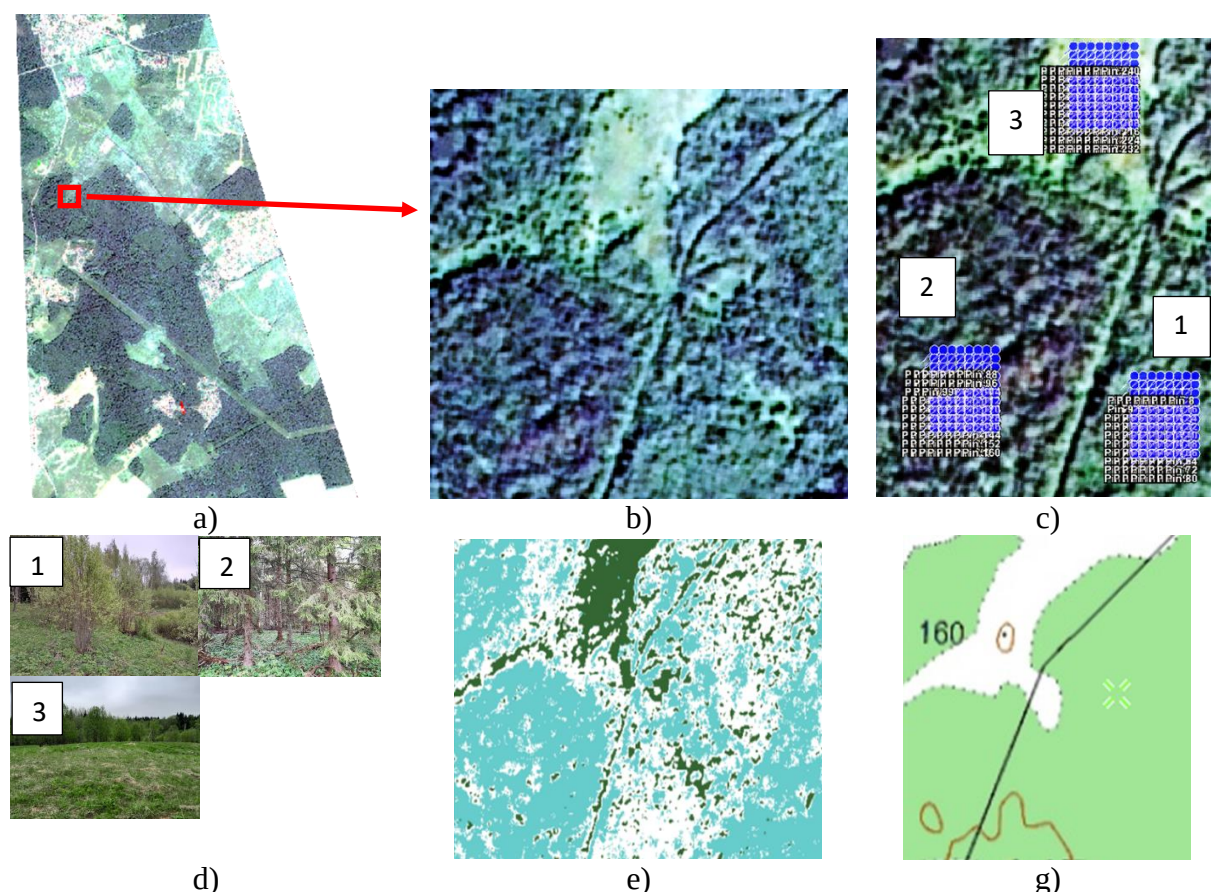


Рисунок 3 Пример формирования исходных данных и результаты идентификации элементов ландшафта: а) часть условного кадра материалов съёмки; б) сцена; в) исходные данные, сформированные оператором; д) результаты наземных обследований для семантического описания элементов ландшафта; е) результаты идентификации элементов ландшафта; г) фрагмент топографической карты сцены. Цифрами на рисунке обозначены

элементы ландшафта с условными наименованиями: 1 - редколесье; 2 - мощный лес; 3 - поляна.

На этапе 3 проведена проверка репрезентативности ИД, которая показала удовлетворительные результаты. При этом размер каждой выборки исходных данных составляет 80 единиц (пикселей).

На этапах 4 и 5 проведено предварительное оценивание качества результатов идентификации элементов ландшафта. Для этого построены гистограммы, представленные на рисунке 4, отражающие размещение пикселей из состава исходных в зависимости от спектрального индекса NDVI [3], характеризующего состояние растительности.

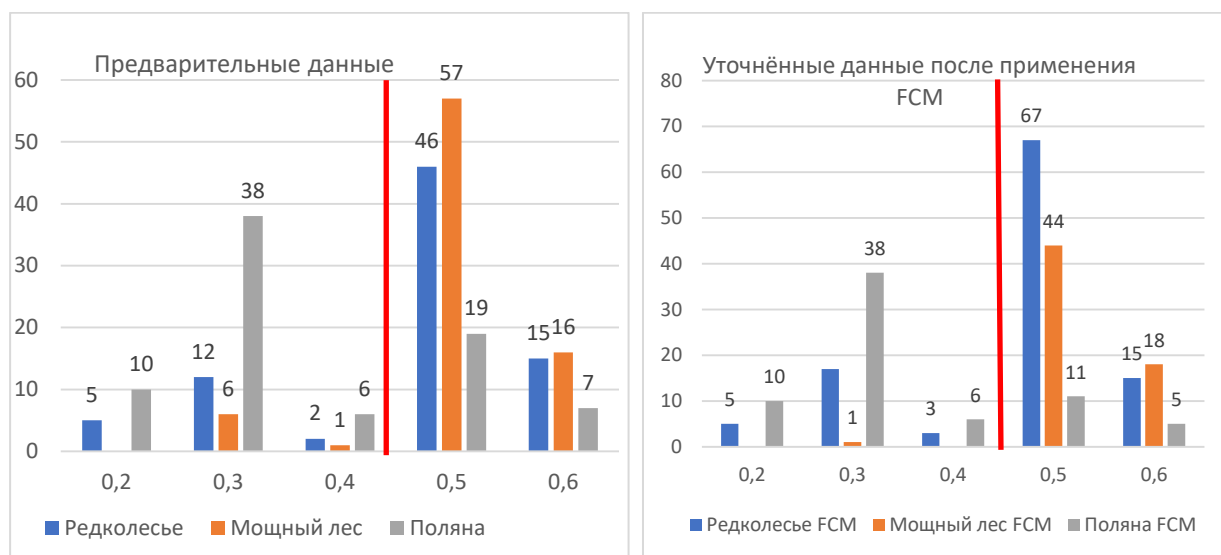


Рисунок 4 Гистограммы для расчета базовых показателей качества результатов идентификации элементов ландшафта: слева – на основе предварительного формирования исходных данных; справа – после применения математического аппарата нечёткой кластеризации (FCM). По оси абсцисс отложены значения индекса NDVI, по оси ординат – количество пикселей, характеризующихся соответствующим значением индекса. Красной линией обозначено значение гиперпараметра, назначаемое оператором.

В ходе анализа гистограмм получены предварительные оценки базовых и обобщенных показателей качества результатов идентификации. Значения показателей приведены в таблице 1. Анализ этих показателей позволил сделать вывод о необходимости повышения качества результатов идентификации элементов ландшафта путём уточнения состава исходных данных на основе применения математического аппарата нечёткой кластеризации. Результаты уточнения ИД отображены на рисунке 5.

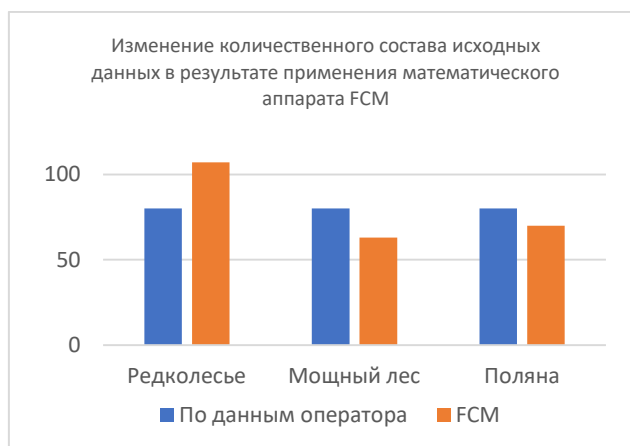


Рисунок 5 Результаты уточнения состава исходных данных после применения алгоритма нечёткой кластеризации (FCM).

В результате уточнения состава исходных данных улучшились предварительные значения обобщенных показателей качества результатов идентификации рассматриваемых элементов ландшафта.

Таблица 1

Показатели предварительного оценивания качества результатов идентификации элементов ландшафта с условными наименованиями «поляна» - «мощный лес» на основе исходных данных, сформированных оператором и после применения математического аппарата FCM

	TP	FP	TN	FN	OA
Оператор	54	7	73	26	0,79
FCM	54	1	62	16	0,87

В таблице приняты следующие обозначения:

TP (True Positive) - количество пикселей, представляющих действительно «поляну»;

FP (False Positive) - количество пикселей, представляющих ошибочно «поляну»;

TN (True Negative) - количество пикселей, представляющих действительно «мощный лес»;

FN (False Negative) - количество пикселей, представляющих ошибочно «мощный лес»;

OA (Overall Accuracy) – обобщенный показатель, качества результатов идентификации, рассчитываемый по формуле:

$$OA = \frac{(TP + TN)}{(TP + FP + TN + FN)}. \quad (1)$$

3. Заключение

Разработанная методика позволяет сформировать исходные данные для алгоритмов машинного обучения с обеспечением: репрезентативности объёма исходных данных; возможности предварительного оценивания качества результатов идентификации элементов ландшафта на основе сформированных выборок; возможности уточнения состава исходных данных.

4. Список литературы

- [1] Maxwell A.E., Warner T.A., Guillén L.A. Accuracy Assessment in Convolutional Neural Network-Based Deep Learning Remote Sensing Studies—Part 2: Recommendations and Best Practices. Remote Sens. 2021, 13, 2591. <https://doi.org/10.3390/rs13132591>
- [2] Mochalov, V., Grigorieva, O., Zhukov, D., Markov, A., Saidov, A. Remote Sensing Image Processing Based on Modified Fuzzy Algorithm Advances in Intelligent Systems and Computing, 2020, 1225 AISC, стр. 563–572 Conference Paper
- [3] <https://www.indexdatabase.de/>