ИЗВЕСТИЯ РАН. ТЕОРИЯ И СИСТЕМЫ УПРАВЛЕНИЯ, 2024, № 5, с. 126–137

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

УДК 004.932.2, 004.032.2

ВЫБОР МЕТОДОВ КЛАСТЕРИЗАЦИИ ПРИ МАШИННОМ ОБУЧЕНИИ ДЛЯ ИССЛЕДОВАНИЯ ЭКОЛОГИЧЕСКИХ ОБЪЕКТОВ ПО СПУТНИКОВЫМ ДАННЫМ

© 2024 г. В. Е. Воробьев[a, *], А. Б. Мурынин[a, b, **], А. А. Рихтер[a, *]**

[a]НИИ "АЭРОКОСМОС", Москва, Россия

[b]ФИЦ ИУ РАН, Москва, Россия

**e-mail: vvorobev.aero@yandex.ru*

***e-mail: amurynin@bk.ru*

****e-mail: urfin17@yandex.ru*

Поступила в редакцию 17.05.2024 г.

После доработки 02.08.2024 г.

Принята к публикации 16.10.2024 г.

Представлен способ подготовки данных для машинного обучения для семантической сегментации информативных классов на изображениях, основанный на кластеризации для решения задач космического мониторинга импактных районов. Приведена классификация методов кластеризации по различным критериям. Обоснован выбор иерархических методов кластеризации как наиболее эффективных для работы с кластерами произвольной структуры и формы. Приведена общая схема расчета модели кластеризации, включающая помимо самой кластеризации процедуры тайлирования данных, оценки оптимальных параметров кластеризации, регистрации объектов, оценку качества данных. Показана схема подготовки данных для машинного обучения, включающая построение эталонной разметки, расчет модели кластеризации, коррекцию разметки, тестирование моделей кластеризации для разных информативных классов на новых изображениях.

Ключевые слова: кластеризация, модель кластеризации, методы кластеризации, обучающая выборка, машинное обучение, семантическая сегментация, экологические объекты, импактные районы, изображения

DOI: 10.31857/S0002338824050088, **EDN:** TARMZN

CHOICE OF CLUSTERING METHODS IN MACHINE LEARNING FOR THE STUDY OF ECOLOGICAL OBJECTS BASED ON SATELLITE DATA

V. E. Vorobyov[a, *], A. B. Murynin[a, b, **], A. A. Richter[a, *]**

[a]ISR "AEROCOSMOS" Moscow, Russia

[b]FRC CSC RAS, Moscow, Russia

**e-mail: vvorobev.aero@yandex.ru*

***e-mail: amurynin@bk.ru*

****e-mail: urfin17@yandex.ru*

The paper presents a method for preparing data for machine learning for semantic segmentation of informative classes in images based on clustering for solving problems of space monitoring of impact areas. A classification of clustering methods by various criteria is given. The choice of hierarchical clustering methods as the most effective for working with clusters of arbitrary structure and shape is substantiated. A general scheme for calculating a clustering model is given, which includes, in addition to the clustering itself, procedures for data tiling, estimating the optimal clustering parameters, registering objects, and assessing the quality of the obtained data. A scheme for preparing data for machine learning is shown, including the construction of a reference markup, calculation of a clustering model, markup correction, and testing the obtained clustering models for different informative classes on new images.

Keywords: clustering, clustering model, clustering methods, training sample, machine learning, semantic segmentation, environmental objects, impact areas, images

Введение. Для выделения экологических объектов при космическом мониторинге импактных районов по мультиспектральным или гиперспектральным данным первостепенной задачей является сегментация изображений. Для решения этой задачи с помощью машинного обучения важно выбрать способ подготовки обучающих данных с учетом огромной площади импактных зон и разнообразия экологических классов [1–3]. Поэтому целесообразно разработать способы ускорения подготовки обучающей выборки [4]. Один из таких способов основан на кластеризации областей на изображениях при подготовке обучающих данных [5].

Кластерный анализ – многомерная статистическая процедура, включающая сбор информации о выборке объектов и затем упорядочивающая объекты в сравнительно однородные группы [6]. Кластеризуются пикселы полутонового, мультиспектрального или гиперспектрального изображения (выборки объектов). Задача кластеризации относится к классу задач обучения без учителя.

Результаты кластеризации могут быть использованы как данные для предобучения, которые впоследствии корректируются ручным или автоматизированным способом. Такой подход позволит ускорить построение обучающей выборки и умножить ее размеры многократно. Даже при наличии ошибок в предобученных данных объем этих данных может нивелировать наличие в них ошибок первого и второго рода.

1. Особенности объекта исследований. В качестве объекта исследования будем рассматривать экологические объекты, расположенные в труднодоступных импактных районах, относящиеся, в частности, к Арктическому региону. Импактные районы – это территории и акватории, подверженные интенсивному антропогенному воздействию и как следствие – сильному загрязнению окружающей среды. Поля загрязнений, обусловленные промышленными объектами, распространяются, захватывая все новые и новые пространства. В результате формируются измененные под воздействием загрязнений импактные экосистемы разного пространственно-временного масштаба, расположенные возле точечного источника эмиссии поллютантов и подверженные действию локальной токсической нагрузки от этого источника и сохраняющие устойчивость во времени [7–9].

По сравнению со многими районами нашей страны, считающимися относительно чистыми, есть «горячие точки», в которых масштабы деградации окружающей среды достигают опасных значений, а уровни загрязнения значительно превышают допустимые нормы. Особенно это заметно в Арктической зоне Российской Федерации, где выявлено более сотни таких точек [10, 11].

Горячие точки образуют крупные импактные районы с сильными техногенными нарушениями природной среды, пагубно сказывающимися не только на перспективах сохранения природно-ресурсного потенциала, но и на здоровье и благополучии населения России и в первую очередь – Арктического региона [7–9]. Импактные районы пролегают в местах, богатых природными ресурсами, т.е. где происходит их интенсивная добыча и переработка.

Ключевое отличие данного объекта от других (например, от участков с площадным загрязнением) заключается в градиентной природе загрязнения. С удалением от источника выбросов происходит постепенное (но не всегда гладкое) уменьшение поступления поллютантов. Соответственно экосистемы получают все меньшие дозы токсических нагрузок. Из-за этого импактные районы представляют собой специфическую пространственную структуру из концентрически расположенных зон с разной степенью загрязнения и, различным уровнем трансформации экосистем. Обычно выделяют три-четыре зоны трансформации: техногенную пустыню, импактную, буферную и фоновую. Пространственная граница импактного региона проходит там, где с помощью современных методов уже не удается отделить локальное действие токсической нагрузки от естественно-обусловленных изменений [7].

Постоянное развитие космических многоспутниковых систем, осуществляющих высокопериодическое зондирование поверхности Земли, позволяет выполнить как оперативную съемку, так и накопление обширных архивов данных дистанционного зондирования. Возросшая доступность и полнота данных дистанционного зондирования, а также современные вычислительные и информационные технологии открывают новые возможности для проведения мониторинговых исследований территорий. Это касается также создания геоинформационного обеспечения мониторинговых исследований, позволяющего осуществлять построение оценочных и прогностических пространственных моделей территорий на основе математического аппарата пространственного анализа данных и геопространственного моделирования.

В контексте анализа состояния импактных районов следует отметить, что полярноорбитальные средства дистанционного зондирования из космоса имеют большие потенциальные возможности для изучения природы высокоширотных импактных районов, включая и самые

труднодоступные. Большое значение в этой связи приобретает теоретическое обоснование, разработка и совершенствование методов сбора и обработки высокопериодической информации и больших массивов данных о состоянии и изменениях объектов в импактных районах, в том числе Арктического региона. Однако имеются некоторые принципиальные проблемы для космического дистанционного зондирования в зонах внутри пояса 75–90° с.ш. [12].

Нерегулярность освещенности северного полярного региона создает дополнительные трудности для оптического дистанционного зондирования, которое практически бесполезно в течение долгой полярной ночи. Кроме того, оптические сенсоры, особенно сканеры, перенасыщаются над арктическими территориями с высоким альбедо и часто изображают ледники со слишком малым контрастом. Интенсивные тени на космических изображениях, возникающие при сочетании низкого положения Солнца и горного рельефа, затрудняют их обработку. На широте 81° 20’ с.ш. ежедневные вариации в высоте Солнца над горизонтом, которые могли бы улучшить контраст, не превышают 17°, что не позволяет эффективно использовать их для получения изображений лучшей контрастности. Тепловые инфракрасные или тепловые микроволновые изображения свободны от этих недостатков, но обладают обычно более низким пространственным разрешением [13].

Главной проблемой для пассивного дистанционного зондирования импактных районов Арктики является, очевидно, своеобразный климат этого региона с ненадежными и неблагоприятными погодными условиями, частой сплошной облачностью и снегопадами, которые маскируют детали поверхности.

2. Выбор методов кластеризации. С учетом описанных особенностей объекта исследования проведем краткий анализ методов кластеризации для решения поставленной задачи (см. табл. 1) [14, 15].

Таблица 1. Классы методов кластеризации по разным критериям

Методы	Критерий	Класс методов	Пример метода
1	Способ формирования кластеров	Разделительные	k-means
2		Основанные на плотности	DBSCAN
3		Основанные на сетке	STING
4		Основанные на модели	Смеси Гауссианов
5		Основанные на графах	Спектральная кластеризация
6		Основанные на подпространствах	CLIQUE
7		Основанные на ансамбле	CSPA
8	Степень вложенности	Неиерархические (плоские)	ISODATA
9		Иерархические	HDBSCAN
10	Степень пересечения кластеров	Исключение	k-means
11		Перекрытие	MCOKE
12		Вложенность	HiSC
13		Нечеткость (стохастичность)	fuzzy k-means
14	Методический подход	Вероятностный подход	k-median
15		Дискреминантный анализ	На базе логистической регрессии
16		Генетический алгоритм	На базе фитнес-функции
17		Нейросетевой подход	Сеть Кохонена
18	Иерархическая кластеризация	Агломеративные методы	Single-link
19		Дивизионные методы	MST
20	Неиерархическая кластеризация	Четкие	Excpectation-Maximization
21		Нечеткие	fuzzy k-means

Методы 1 разделяют данные по минимизации внутрикластерного расстояния и максимизации межкластерного. Методы 2 группируют данные и отделяют группы высокой плотности от групп низкой. Методы 3 разделяют пространство на ячейки и анализируют плотность в каждой из них. В методах 4 предполагается пораждающая статистическая модель, на базе которой подбираются параметры кластеризации. Методы 5 базируются на теоретико-графовом представлении данных. В методах 6 данные кластеризуются в подпространствах признаков, найденных методами понижения размерности [16]. В методах 7 комбинируются (гибридизируются) различные методы кластеризации для получения надежного и стабильного разбиения.

В исключающих методах 10 элемент данных принадлежит строго одному кластеру, в перекрывающихся методах 11 – может принадлежать более, чем одному кластеру. В случае вложенности методов 12 каждый элемент относится как к определенному кластеру, так и ко всем его «предкам», в нечетких методах 13 – к разным кластерам с разной вероятностью.

В вероятностных методах 14 используются статистические распределения и эмпирические вероятности принадлежности точек кластерам. В методах 15, основанных на дискреминантном анализе, известны характеристики самих кластеров, к которым нужно отнести элементы данных. В методах 16 количество кластеризуемых объектов, образующих популяцию, соотносится с количеством компонентов вектора хромосом, каждая из которых отображает один из возможных вариантов кластеризации. В методах 17 результаты кластеризации, полученные другими методами, могут быть использованы для предобучения нейросетевой модели. Помимо перечисленных методических подходов 14–17 существуют и другие подходы.

Каждый метод кластеризации имеет свои достоинства и недостатки, задается набором параметров кластеризации. Также методы могут иметь разные модификации, меняющие шаги базового алгоритма, уменьшающие функцию потерь, применимые к разным типам данных и др.

Одним из наиболее популярных методов кластеризации является k-means.

Преимущества метода: 1) прост в реализации и понимании; 2) высокая скорость работы и точность на данных сферической формы; 3) наличие большого числа модификаций [14].

Недостатки метода: 1) задание числа кластеров / их начальных точек кластеров; 2) сходится к локальным максимумам, что дает несколько разные результаты кластеризации каждый раз при постоянном числе кластеров; 3) кластеризуются сфероподобные области (т.е. эллипсоили линейно подобные области кластеризуются значительно хуже); 4) в целом не учитывает плотность данных и неоднородность кластера [17].

Метод имеет модификации [14]: 1) вариации метода (lloyd's algorithm, elkan algorithm, minibatch k-means, k-medoids, k-modes); 2) способ выбора начальных центроидов (случайный выбор, k-means++, greedy k-means++ и др.); 3) способ подбора оптимального числа кластеров (метод локтя, метод силуэта, метод комплексной оценки); 4) «родственные» методы (c-means, k-median и др.).

Неиерархические методы 8 не учитывают иерархию кластеров (имеют один иерархический уровень). При этом неиерархические методы могут давать однозначную (методы 20) или стохастическую (методы 21) кластеризацию.

При кластеризации экологических объектов на аэрокосмических изображениях учет их структуры является одинм из наиболее важных условий выбора метода кластеризации. Большинство неиерархических методов либо ограничены в кластеризуемых структурных паттернах, либо имеют сложный алгоритм реализации. В частности, k-means работает лучше всего, когда кластеры: «круглые» или сферические; одинакового размера; одинаковой плотности; имеют самую высокую плотность в центре сферы; не загрязнены шумом/выбросами.

Иерархические методы 9 практически лишены данных недостатков. В основе иерархической кластеризации лежит построение кластерного дерева, представляемого обычно в виде дендрограммы. Данное дерево посредством связности точек, кластеров хранит практически полную информацию о структуре изображенных объектов. Однако с ростом размера кластеризуемых данных экспоненциально растет размер кластерного дерева, что не позволяет «напрямую» использовать данные методы при обработке данных больших размеров. В этом случае целесообразно подобрать способы тайлирования данных для их параллельной обработки. Иерархические методы 9 имеют более одного иерархического уровня. Они разделяются на алгомеративные 18 (новые кластеры создаются объединением более мелких в более крупные) и дивизионные 19 (новые кластеры создаются делением более крупных на более мелкие).

Разные методы кластеризации могут иметь иерархические расширения (в том числе k-means при иерархическом разложении данных на определенное количество кластеров). Основными параметрами иерархических методов являются метод вычисления связи (методы одиночной, полной, средней связи, метод Уорда) и метрика связи (дисперсия соединяемых / разделяемых кластеров, среднеарифметическое расстояний, максимальное / минимальное расстояние и др.) [14].

Преимущества иерархических методов: 1) способность обнаружения кластеров произвольной формы; 2) работоспособность с различными паттернами данных; 3) возможность формирования информативной иерархии кластеров для лучшего понимания структуры данных; 4) возможность получения оптимальной кластеризации; 5) производит значительно меньше шума.

Недостатки иерархических методов: 1) использование большого количества вычислительных ресурсов и памяти из-за работы со всей матрицей расстояний между объектами; 2) чувствительность к выбору критерия объединения кластеров и неустойчивость к шуму и выбросам, что может сильно искажать иерархию кластеров.

В табл. 2 рассмотрен пример выборки $w = 10$ точек для иерархической кластеризации.

Таблица 2. Выборка точек

n	Яркости			Координаты	
	q1	*q2*	*q3*	*x*	*y*
0	67	68	63	0	0
1	70	69	65	0	1
2	72	71	67	0	2
3	77	73	70	0	3
4	69	68	64	0	4
5	53	55	52	0	5
6	38	43	39	0	6
7	58	60	59	0	7
8	48	48	48	0	8
9	49	47	52	0	9

На рис. 1 показана дендрограмма (кластерное дерево) для данной выборки. Матрица связи для нее отражена в табл. 3 [18]. В качестве метрики расстояния взята евклидова метрика, тип связи – по ближайшей точке. Здесь n – номер точки; $L = w - 1$ – наибольший номер точки; *q1, q2, q3* – яркости на R-, G-, B- каналах соответственно; x и y – абсцисса и ордината точки на изображении; N – номер связи (иерархического кластера); *n1* и *n2* – номера связанных иерархических кластеров (точек); d – расстояние между иерархическими кластерами; m – число точек в составе иерархического кластера.

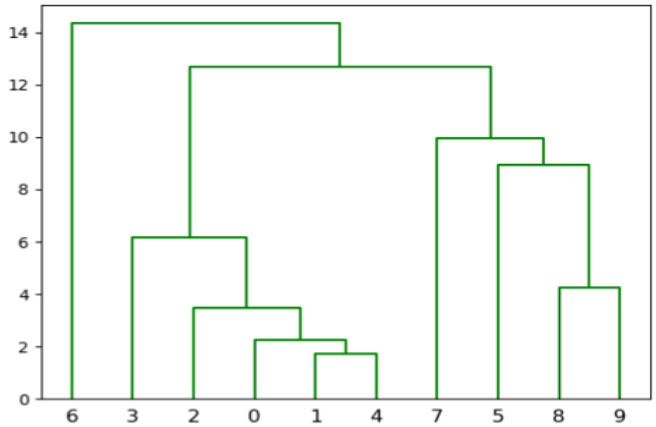

Рис. 1. Дендрограмма выборки точек.

Таблица 3. Матрица связи выборки точек

N	Матрица связи				Описание связи
	n1	*n2*	*d*	*m*	
1	1	4	1.73	2	Связь между точками $n = 1$ и $n = 4$
2	0	10	2.96	3	Связь точки $n = 0$ и кластера $N = 10 - L = 1$
3	8	9	4.24	2	Связь между точками $n = 8$ и $n = 9$
4	2	11	5.67	4	Связь точки $n = 2$ и кластера $N = 11 - L = 2$
5	3	13	8.95	5	Связь точки $n = 3$ и кластера $N = 13 - L = 4$
6	5	12	8.97	3	Связь точки $n = 5$ и кластера $N = 12 - L = 3$
7	7	15	13.9	4	Связь точки $n = 7$ и кластера $N = 15 - L = 6$
8	6	16	26.14	5	Связь точки $n = 6$ и кластера $N = 16 - L = 7$
9	14	17	40.76	10	Связь кластеров $N = 14 - L = 5$ и $N = 17 - L = 8$

На рис. 2, *а* показано входное изображение. Сопоставлены результаты кластеризации (предобученной разметки) методами k-means (рис. 2, *б*) и агломеративной кластеризации (рис. 2, *в*) для данного изображения. Как видно, иерархическая кластеризация образует значительно меньше шума, чем кластеризация методом k-means, что дает ощутимые преимущества при построении обучающей выборки.

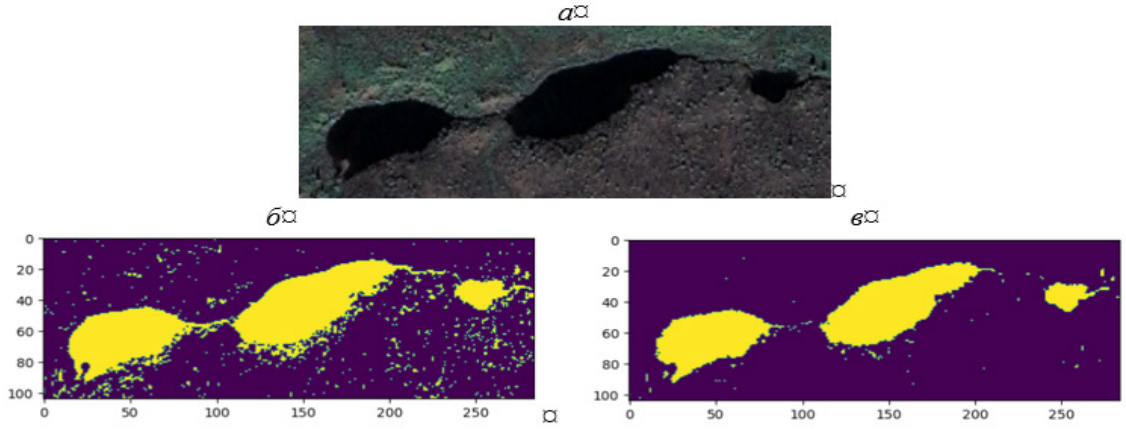

Рис. 2. Сопоставление для изображения (*а*) метода k-means (*б*) и метода агломеративной кластеризации (*в*).

3. Подготовка данных для обучения. На рис. 3 приведена общая схема кластеризации изображений.

Кластеризацию можно представить отображением $f(p, I) \to K$, где I – изображение для кластеризации, p – параметры кластеризации, f – метод и алгоритм кластеризации, K – оцененная маска кластеров (бинарная, пообъектная или маска полей энергии), размер которой равен размеру I. Итерируемые параметры $p' \subset p$ – параметры кластеризации, значения которых меняются для обеспечения оптимизации некоторой целевой функции. Основными параметрами кластеризации, которые также относятся к итерируемым параметрам, являются: количество m кластеров $\{k_i\}$, метрика схожести / разделимости кластеров. В качестве метрик могут быть метрики расстояний, угловые метрики, метрики на основе корреляций и др. [19, 20].

Для оптимизации может быть подана эталонная маска кластеров $K_э$ (эталоны объектов) изображения I. Оптимизация производится на базе целевой функции $F(p', \mu, K, K_э)$ или ее частного случая $F(p', \mu, K)$ при отсутствии эталонной кластеризации. Значение p', при ко-

Рис. 3. Схема получения модели кластеризации.

тором F достигает экстремума (минимального или максимального), считается оптимальным значением при всех неизменных других $p \setminus p'$ параметрах кластеризации.

Тайлирование (тайлинг) – разделение данных на части (тайлы), каждая из которых обрабатывается некоторым алгоритмом в отдельности. Как правило, необходимость в тайлировании обусловлена наличием больших размеров данных. При этом обработка алгоритмом данных целиком выражается через обработку алгоритмом тайлов данных в отдельности. Для обеспечения эквивалентности такой обработки необходимо провести обратную процедуру, т.е. слияние обработанных тайлов. Формально тайлирование можно записать в виде:

$$f(I) = \left\{ \bigcup_{j=1}^{l} f\left(I_j\right), \bar{f}\left(\left\{I_j\right\}\right) \right\}, I = \bigcup_{j=1}^{l} I_j,$$

где f – алгоритм обработки; I – данные целиком; $I_j, j = 1...l,$ – тайлы разбиения данных; l – количество тайлов разбиения; $\bar{f}$ – алгоритм слияния тайлов, порождаемый алгоритмом f. Над разбиением $\{I_j\}$ данных производится $\bar{f}$ такой, чтобы совместно с процедурами $f(X_1),..., f(I_j)$ результат обработки был равен или приближенно равен результату обработки процедуры $f(I)$.

На рис. 4 показаны некоторые способы тайлирования при кластеризации изображений. При слиянии в местах «сшития» тайлов I' и I'' производится соотнесение кластеров в соответствующих точках масок $K' = f(I')$ и $K'' = f(I'')$. Это необходимо для одинаковой интерпретации рассчитанных кластеров на разных тайлах.

Пообъектная регистрация – выделение связных компонент (объектов), их фильтрация по структуре и размеру растров, расчет для каждого отфильтрованного объекта его геометрических параметров, а также представление объекта в виде кортежа координат и яркостей его точек. На выходе образуется модель кластеризации, включающая: пообъектную регистрацию, маску кластеров, оптимальные параметры кластеризации, значения целевой функции, структуру входных данных (изображений) для обучения.

На рис. 5 – схема подготовки данных для машинного обучения с применением полученной модели кластеризации. Эталонная разметка (датасет) формируется для каждого информативного класса «ручным» и интерактивным способом, предобученная разметка – в результате

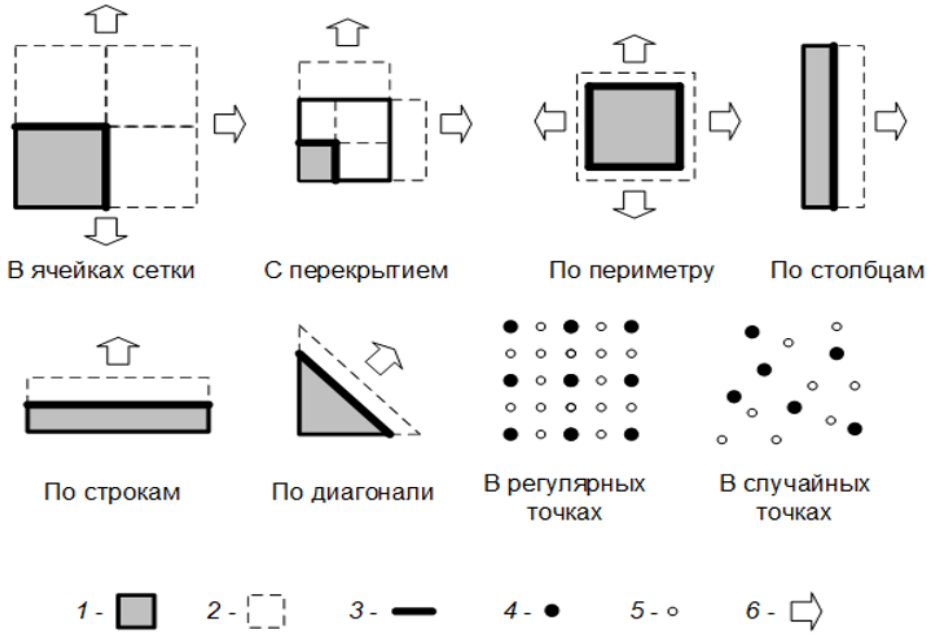

Рис. 4. Примеры способов тайлирования при кластеризации. Обозначения на рисунке: *1* – текущий тайл; *2* – следующий тайл; *3* – места сшития тайлов; *4* – элемент текущего таила; *5* – элемент следующего тайла; *6* – направление тайлирования.

Рис. 5. Схема подготовки данных для машинного обучения (ИНС – искусственная нейронная сеть).

расчета модели кластеризации, дообученная – по результатам коррекции и аугментации предобученной, обе выполняемые на кластеризованных областях.

Коррекция разметки состоит в изменении кластеров в рассчитанной оптимизированной модели кластеризации. В табл. 4 отражены основные эмпирические правила построения обучающей выборки, т.е. разметки объектов определенного информативного класса с установленными дешифровочными признаками [4].

Таблица 4. Типы правил построения обучающей выборки

Тип правила	Описание	Пример правила для информативного класса
Локализация	Обнаружение рабочей области поиска объектов	Свалки сосредотачиваются с большей вероятностью на периферии города, чем в его центре
Обнаружение	Обнаружение объектов в рабочей области на базе геометрических, текстурных, яркостных и других признаков	Одним из геометрических признаков окон на стене здания является регулярность их распределения
Выделение	Разграничение объекта от фона	Граница между мусорным захламлением и фоном лежит в местах максимальных перепадов зернистости текстуры захламления
Векторный тип	Особенности использования векторов при разметке	Ребро здания выделяется линейным вектором некоторой «толщины» (линейный вектор автоматически модифицируется в площадной с данной толщиной)
Предположение	Как размечать в условиях невидимости части объекта	В местах небольшой протяженности заслонения дороги деревьями или тенями предполагается такая же дорога
Разделимость	Как размечать в условиях наложения одних объектов на другие	Вагоны подвижного состава размечаются непересекающимися областями
Артефакты	Как размечать в условиях артефактов съемки	При наличии геометрических искажений крыши здания данная область исключается из обучающей выборки
Интерпретация объекта	Разметка в условиях неоднозначности или сложности в интерпретации	При схожести крыши здания с объектом из другого класса данная область исключается из обучающей выборки
Унификация объекта	Разметка в условиях типизации объектов	Если здание имеет одинаковую форму крыши, но с разными уровнями высоты, каждая часть здания с одним уровнем высоты размечается как отдельное здание
Исключение разметки	Как исключается область из обучения	Возможность построения исключаемых областей полигонами
Коррекция разметки	Как достраивать, удалять, перемещать разметку / часть разметки объекта	В местах наличия ложной разметки на части объекта (ошибки первого рода) построение стираемых областей полигонами

Аугментация обучаемых данных позволяет дополнительно «умножить» разметку с применением различных видов аугментации для размеченных областей: 1) изменение оттенка, насыщенности или яркости; 2) повороты, масштабирования или сдвиги; 3) преобразования перспективы; 4) копирование областей для стационарных объектов (имеющих постоянную локацию на земной поверхности) на изображения этих объектов в другие моменты времени с тем же ракурсом съемки.

В хранилище данных содержится: 1) изображения и их метаданные (для эталонной разметки, обучаемые, валидируемые и тестируемые при кластеризации); 2) модели кластеризации каждого информативного класса; 3) маски эталонной разметки и обучающей выборки информативных классов [21–23].

На вход модели нейронной сети подается обучающая выборка для эталонной разметки, а также разметки, рассчитанной посредством кластеризации и прошедшей процедуры коррекции и аугментации.

В качестве модели сети берется одна из ранее разработанных моделей сверточных сетей для сегментации экологических объектов [22–24]. Это многоклассовые и бинарные модели для сегментации различных экологических классов – мусорных свалок, разных типов жидкостей, объектов жидкостной и трубопроводной инфраструктуры, объектов автодорожной и железнодорожной инфраструктуры, зданий (в том числе производственных). Модели обучены на полутоновых и мультиспектральных изображениях импактных районов Арктики и Московского региона. Данные аугментированы с применением процедуры ROI Cover. В качестве функции потерь используется функция потерь Жаккарда (Joccard loss), а также данная функция в комбинации с бинарной перекрестной энтропией.

4. Результаты эксперимента. Покажем на примере г. Норильск результаты построения обучающей выборки объектов класса «водоемы» с применением кластеризации методом k-means.

На рис. 6, *а* приведен пример входного изображения. Средние координаты территории: широта – 69.417°, долгота – 87.907°; расположена в 12 км к северо-западу от жилой части города. Выбор данной территории обусловлен тем, что ее водные объекты являются объектами «непрямого» экологического загрязнения. Она расположена, с одной стороны, поблизости к новой очереди хвостохранилища Лебяжье, с другой стороны, «в стороне» от реки Амбарная – переносчика и распространителя экологического загрязнения от Норильского металлургического завода.

Для данного изображения произведена кластеризация со следующими основными параметрами кластеризации: количество кластеров – 5; количество запусков алгоритма с разными начальными центроидами – 10; максимальное количество итераций алгоритма для одиночного прогона – 300. На рис. 6, *б* представлена пообъектная маска кластеров. В маске выделен кластер, характеризующий водоемы (рис. 6, *в*). Произведена коррекция данного кластера (рис. 6, *г*) по установленным эмпирическим правилам для класса «водоемы». В частности, удалены шумовые составляющие – связные компоненты небольших размеров, описывающие в основном тени деревьев и резкие перепады поверхности. На рис. 6, *д* показано наложение скорректированного кластера на изображение. Водоемы в малых масштабах времени являются стационарными объектами. Поэтому можно произвести аугментацию копирования областей, т.е. взять размеченные области водоемов, но на других изображениях той же территории при условии, что они имеют тот же ракурс съемки. Например, ракурс изображения (на рис. 6, *е*) тот же, что и на рис. 6, *а*, но снято оно в других условиях погоды (появилась небольшая облачность) и времени (хронологическом и сезонном).

Как видно, обучающая выборка построена в полуавтоматическом режиме с применением кластеризации и коррекции, что значительно проще построения обучающей выборки в ручном режиме. Наблюдается хорошая точность полученной разметки по наложению областей на

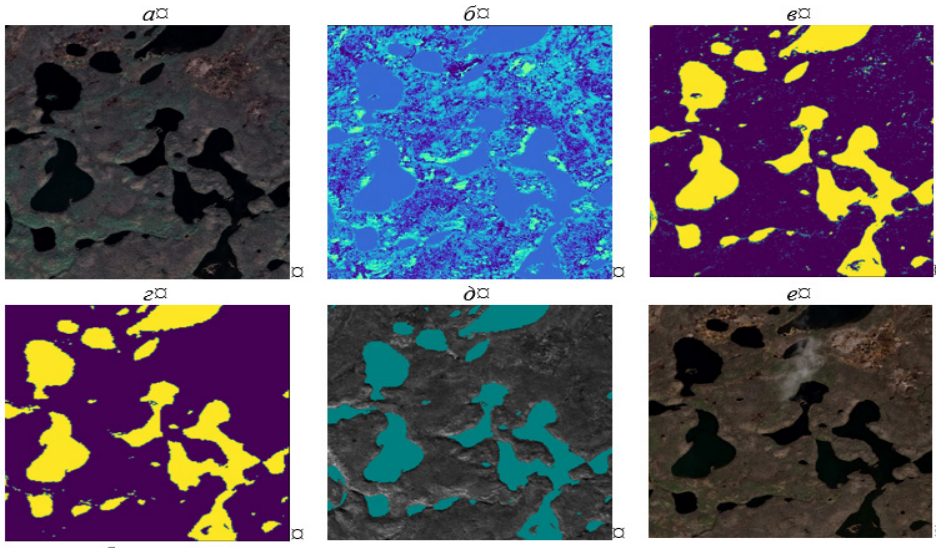

Рис. 6. Получение дообученной разметки с применением кластеризации: *а* – входное изображение; *б* – пообъектная маска кластеров; в-выделение кластера водоемов; *г* – коррекция кластера; *д* – наложение скорректированного кластера на изображение; *е* – другое изображение того же ракурса.

изображение. Автоматически выделены границы водоемов в местах максимальных перепадов яркости. При наличии в водоемах «островов» последние исключены из выборки. Значительно труднее сделать такое исключение в ручном режиме, так как «острова» относятся к трудноразмечаемым объектам. К другим трудноразмечаемым объектам принадлежат мелкие водоемы, поскольку их количество больше, чем крупных, и обнаружить их сложнее.

На примере изображения, приведенного на рис. 7, показан результат работы сверточной сети по сегментации водоемов, построенной на базе дообученной разметки. Дообучение проведено в окрестности фрагмента изображения, выделенного на рис. 7, *а*. В результате сегментации выделены крупные и мелкие водоемы и реки (слева – широкая река Амбарная, справа – узкая река Снежная). Разрывы, образованные в сегментации реки Амбарная, обусловлены в основном сужением водотока в излучине реки и пересыханием части прибрежной полосы, вероятно ввиду переноса поллютантов водным течением.

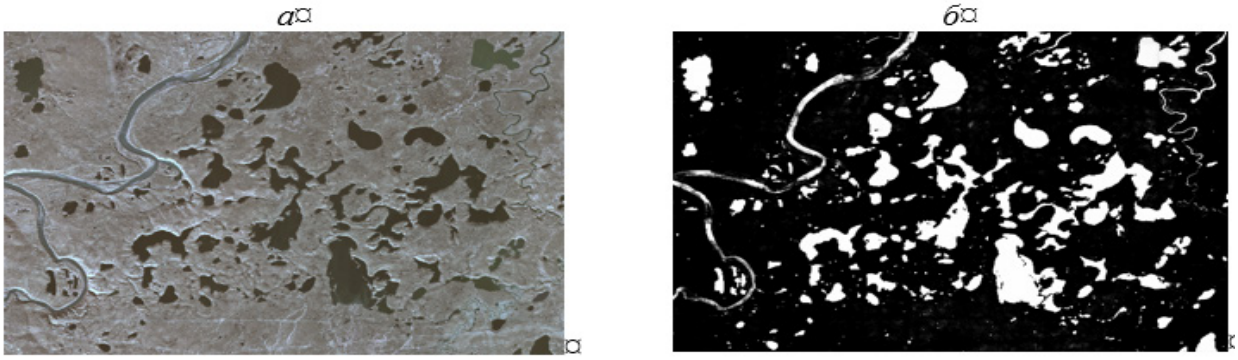

Рис. 7. Результат семантической сегментации водоемов (г. Норильск): *а* – входное изображение; *б* – выделение водоемов с помощью сверточной сети, дообученной на скорректированном кластере.

Заключение. Предложен подход к кластеризации для построения предобученной выборки для машинного обучения, которая «доучивается» оператором-разметчиком. Строится модель кластеризации, включающая пообъектную регистрацию, маски кластеров, полученные по набору обучаемых изображений, оптимальные параметры кластеризации. Данная процедура является составной частью процедуры подготовки данных для машинного обучения. В хранилище содержатся изображения и их метаданные, модели кластеризации каждого информативного класса, маски эталонной и обучающей разметки.

Достоинствами применения кластеризации для машинного обучения является автоматизация и ускорение построения разметки, а также допустимость ошибок первого и второго рода при наличии большого объема данных для обучения.

СПИСОК ЛИТЕРАТУРЫ

1. *Визильтер Ю.В., Выголов О.В., Желтов С.Ю., Рубис А.Ю.* Комплексирование многоспектральных изображений для систем улучшенного видения на основе методов диффузной морфологии // Изв. РАН. ТиСУ. 2016. № 4. С. 103–114.
2. *Желтов С.Ю., Себряков Г.Г., Татарников И.Б.* Компьютерные технологии создания геопространственных трехмерных сцен, использующих комплексирование географической информации и синтезированных пользовательских данных // Авиакосмическое приборостроение. 2003. № 8. С. 2–10.
3. *Ишутин А.А., Кикин И.С., Себряков Г.Г., Сошников В.Н.* Алгоритмы обнаружения, локализации и распознавания оптико-электронных изображений группы изолированных наземных объектов для инерциально-визирных систем навигации и наведения летательных аппаратов // Изв. РАН. ТиСУ. 2016. № 2. С. 85.
4. *Гвоздев О.Г., Козуб В.А., Кошелева Н.В., Мурынин А.Б., Рихтер А.А.* Построение трехмерных моделей ригидных объектов по спутниковым изображениям высокого пространственного разрешения с использованием сверточных нейронных сетей // Исследования Земли из космоса. 2020. № 5. С. 78–96.
5. *Мандель И.Д.* Кластерный анализ. М.: Финансы и статистика, 1988. 176 с. ISBN5–279–00050–7
6. *Shuyue G., Murray L.* An Internal Cluster Validity Index Using a Distance based Separability Measure // IEEE32nd Intern. Conf. on Tools with Artificial Intelligence (ICTAI)At: Baltimore, MD, USA, 2020. URL: https://arxiv.org/pdf/2009.01328

7. *Евсеев А.В., Красовская Т.М.* Закономерности формирования импактных зон в Арктике и Субарктике России // География и природные ресурсы. 1997. № 4.
8. *Евсеев А.В., Красовская Т.М.* «Горячие точки» Российской Арктики. Экологические проблемы российской Арктики // Вестн. МГУ. 2010. № 5.
9. *Душкова Д.О., Евсеев А.В.* Анализ техногенного воздействия на геосистемы Европейского Севера России // Арктика и Север. 2011. № 4. С. 1–34.
10. *Лукин Ю.Ф.* «Горячие точки» Российской Арктики //Арктика и Север. 2013. № 11. С.19, 20.
11. Программа ООН по окружающей среде. Диагностический анализ состояния окружающей среды арктической u1079 зоны Российской Федерации: Расширенное резюме. М.: Науч. мир, 2011.
12. *Бондур В.Г.* Основы аэрокосмического мониторинга окружающей среды. Курс лекций. М.: Московский государственный университет геодезии и картографии, 2008. 546 с.
13. *Савиных В.П. Соломатин В.А.* Оптико-электронные системы дистанционного зондирования. М.: Машиностроение, 2014. 431 с.
14. Хабр. Кластеризация в ML: от теоретических основ популярных алгоритмов к их реализации с нуля на Python. URL: https://habr.com/ru/articles/798331/#dbscan
15. Scikit-learn. Руководство пользователя URL: https://scikit-learn.ru/user_guide
16. *Рихтер А.А., Мурынин А.Б., Козуб В.А., Гвоздев О.Г.* Модели представления экологических объектов по данным гиперспектральной съемки // Матер. 21-й Всероссийск. конф. с междунар. участием: Математические методы распознавания образов (ММРО). М.: Российская академия наук, 2023.
17. *Гвоздев О.Г., Козуб В.А., Мурынин А.Б., Рихтер А.А.* Представление и обработка спектральных моделей по данным гиперспектральной съемки // Сб. тез. докл. 16-й Всероссийск. конф. «Современные проблемы дистанционного зондирования Земли из космоса». М.: ИКИ РАН, 2023. С. 19. URL: http:// http://conf.rse.geosmis.ru/files/books/2023/9992.htm
18. Scipy. Руководство пользователя. Метод linkage. URL: https://docs.scipy.org/doc/scipy/tutorial/index.html
19. *Shanmugam S., Srinivasaperumal P.* Spectral Matching Approaches in Hyperspectral Image Processing // Intern. J. Remote Sensing, 2014. V. 35. No. 24. P. 8217–8251. https://doi.org/10.1080/01431161.2014.980922. URL: https://www.researchgate. net/publication/270805406_Spectral_matching_approaches_in_hyperspectral_image_processing
20. *Jain A.K., Murty M.N., Flynn P.J.* Data clustering: a review // Association for Computing Machinery, 1999. URL: https://www.sci-hub.ru/10.1145/331499.331504?ysclid=lzwss1aw3q662345026
21. Ultralytics. Руководство пользователя. URL: https://docs.ultralytics.com/ru
22. *Гвоздев О.Г., Козуб В.А., Кошелева Н.В., Мурынин А.Б., Рихтер А.А.* Нейросетевой метод построения трехмерных моделей ригидных объектов по спутниковым изображениям // Мехатроника, автоматизация, управление. 2021. Т. 22. № 1. С. 48–55.
23. *Игнатьев В.Ю., Матвеев И.А., Мурынин А.Б., Усманова А.А., Цурков В.И.* Повышение пространственного разрешения панхроматических спутниковых изображений на основе генеративных нейросетей // Изв. РАН. ТиСУ. 2021. № 2. С.64–72.
https://doi.org/10.31857/S0002338821020074
24. *Гвоздев О.Г., Мурынин А.Б., Козуб В.А., Пуховский Д.Ю., Рихтер А.А.* Семантическая сегментация спутниковых изображений с использованием нейросетей для выявления антропогенных объектов в импактных районах Арктики // Матер. 20-й Междунар. конф. «Современные проблемы дистанционного зондирования Земли из космоса». М., 2022. С. 60.
https://doi.org/10.21046/20DZZconf-2022a