УДК 004.85, 528.85, 519.237.8

СРАВНЕНИЕ МЕТОДОВ КЛАСТЕРИЗАЦИИ K-MEANS И GAUSSIAN MIXTURE ДЛЯ АНАЛИЗА СПУТНИКОВЫХ СНИМКОВ

Белоусов Евгений Сергеевич[1], Добрынин Дмитрий Владимирович[2]

[1]Студент;
Государственный университет «Дубна»;
141980, Россия, Московская обл., г. Дубна, ул. Университетская, 19;
e-mail: bees.23@uni-dubna.ru.

[2]Старший преподаватель
Государственный университет «Дубна»;
141980, Россия, Московская обл., г. Дубна, ул. Университетская, 19;
e-mail: dobrynin.d.v@uni-dubna.ru.

В работе проведён сравнительный анализ двух алгоритмов кластеризации без учителя – K-Means и Gaussian Mixture Model (GMM) – применительно к задаче выделения природных объектов на спутниковых снимках. Признаковое пространство сформировано из попиксельных спектральных характеристик и локальных статистик (среднее и эксцесс яркости в окне 15×15) с последующей стандартизацией. Оптимальное число кластеров (k = 12) подобрано методом локтя. Для каждого алгоритма построены модели на спектральных признаках (S-варианты) и на расширенном пространстве (C-варианты), качество оценено по индексам Silhouette, Davies–Bouldin и Calinski–Harabasz и визуально. Дополнительно проверена устойчивость моделей при переносе на новое изображение. Показано, что выбор алгоритма должен определяться прикладной задачей: S-KMeans – быстрая базовая кластеризация, C-GMM – детекция границ объектов, S-GMM – масштабирование на новые снимки.

Ключевые слова: кластеризация, K-means, гауссовы смеси, дистанционное зондирование Земли, спутниковые снимки, признаковое пространство, обучение без учителя, сегментация изображений.

Для цитирования:

Белоусов Е. С., Добрынин Д. В. Сравнение методов кластеризации K-Means и Gaussian Mixture для анализа спутниковых снимков // Системный анализ в науке и образовании: сетевое научное издание. 2026. № 2. С. 29-35. EDN: PGGPNK. URL: https://sanse.ru/index.php/sanse/article/view/716.

COMPARISON OF K-MEANS AND GAUSSIAN MIXTURE CLUSTERING METHODS FOR SATELLITE IMAGE ANALYSIS

Belousov Evgeniy S.[1], Dobrynin Dmitry V.[2]

[1]Student;
Dubna State University;
19 Universitetskaya Str., Dubna, Moscow region, 141980, Russia;
e-mail: bees.23@uni-dubna.ru.

[2]Senior teacher;
Dubna State University;
19 Universitetskaya Str., Dubna, Moscow region, 141980, Russia;
e-mail: dobrynin.d.v@uni-dubna.ru.

This paper presents a comparative analysis of two unsupervised clustering algorithms – K-Means and the Gaussian Mixture Model (GMM) – applied to the task of extracting natural objects from satellite imagery. The feature space is built from pixel-wise spectral channels and local statistics (mean and kurtosis of brightness in a 15×15 window) followed by standardisation. The optimal number of clusters (k = 12) is selected via the elbow method. For each algorithm two models are trained – one on the spectral channels

only (S-variants) and one on the extended feature space (C-variants). Performance is evaluated using the Silhouette, Davies–Bouldin and Calinski–Harabasz indices together with visual inspection. The stability of the models under transfer to a new image is also tested. The results show that the choice of algorithm should be driven by the applied task: S-KMeans is best for fast baseline clustering, C-GMM for object-boundary detection, and S-GMM for transfer to new satellite images.

Keywords: clustering, K-means, Gaussian mixture model, Earth remote sensing, satellite imagery, feature engineering, unsupervised learning, image segmentation.

For citation:

Belousov E. S., Dobrynin D. V. Comparison of K-Means and Gaussian Mixture clustering methods for satellite image analysis. *System analysis in science and education*, 2026;(2):29-35 (in Russ). EDN: PGGPNK. Available from: https://sanse.ru/index.php/sanse/article/view/716.

Введение

Задача кластеризации спутниковых снимков востребована во многих прикладных областях – от сельского хозяйства, где она используется для оценки состояния посевов, до геологии, где по спектральным характеристикам выявляются перспективные участки месторождений. Развитие методов автоматической обработки данных дистанционного зондирования Земли (ДЗЗ) позволяет существенно ускорить интерпретацию большого объёма снимков, поступающих с современных космических аппаратов.

Однако применение классических подходов к сегментации и распознаванию объектов на спутниковых изображениях сталкивается с рядом ограничений. Во-первых, опора только на спектральные (цветовые) характеристики не позволяет различать объекты с близкими отражательными свойствами – например, лёд и облако. Во-вторых, природные объекты обладают разнообразной морфологией: от компактных озёр до вытянутых рек и дорог. В-третьих, размеченных наборов данных под конкретную задачу зачастую недостаточно, что делает актуальным применение алгоритмов обучения без учителя.

Цель настоящей работы – сравнительный анализ алгоритмов кластеризации для задачи распознавания природных объектов на космических снимках. Для достижения этой цели последовательно решены следующие задачи: сформировано расширенное признаковое пространство; обучены и настроены по гиперпараметрам модели *K-Means* и *Gaussian Mixture Model* (*GMM*); проведена количественная и визуальная оценка качества кластеризации; выбраны наиболее подходящие алгоритмы для различных классов прикладных задач.

1. Подготовка данных и формирование признакового пространства

В качестве исходных данных использовалось полноцветное (RGB) спутниковое изображение прибрежной арктической акватории разрешением 1356×1142 пикселя (рис. 1). Каждый пиксель рассматривался как объект выборки, описываемый тремя спектральными признаками – значениями яркости по красному, зелёному и синему каналам.

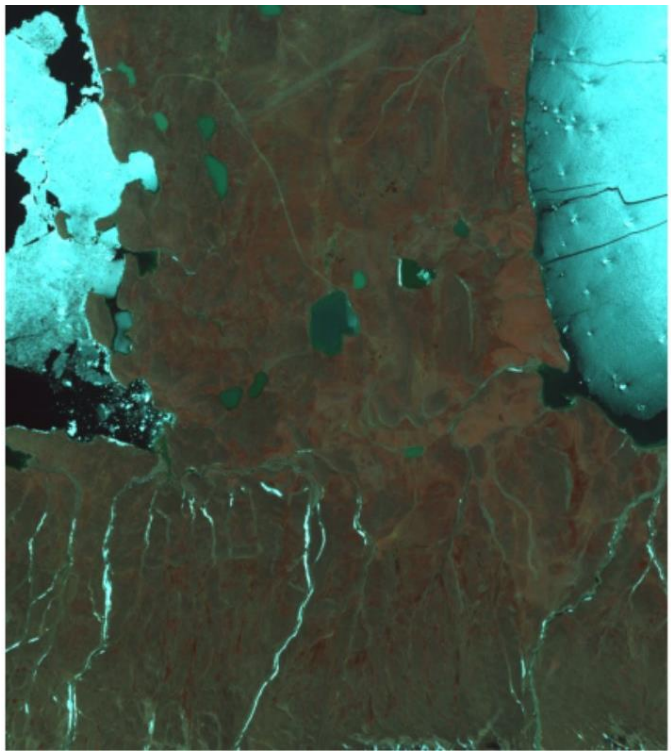

Рис. 1. Исходное спутниковое изображение прибрежной арктической акватории (1356×1142 пикселя)

Для разделения природных объектов, слабо различимых только по спектральным характеристикам, признаковое пространство было расширено локальными статистиками яркости. По каждому из трёх каналов для каждого пикселя в скользящем окне размера 15×15 пикселей вычислялись две характеристики: среднее значение яркости и эксцесс распределения яркостей. Эксцесс, как мера остроты распределения, был выбран в предположении, что он позволит отличать однородные текстуры от неоднородных. Размер окна 15×15 (соответствующий примерно 150 метрам на местности) был подобран экспериментальным путём.

Граничные пиксели обрабатывались методом отражения (*reflect*): значения вне исходного изображения дополнялись зеркальным отражением вдоль соответствующих осей. Итоговая матрица «объекты-признаки» содержала девять столбцов: *R, G, B, R_mean, G_mean, B_mean, R_kurtosis, G_kurtosis, B_kurtosis*. Признаки стандартизованы с помощью *StandardScaler* из библиотеки *scikit-learn*, чтобы привести их к единому масштабу и устранить доминирование переменных с большой амплитудой.

В дальнейшем рассматривались два варианта признакового описания: только спектральные каналы (далее – приставка *S-* к названию модели) и полное девятимерное пространство со статистиками (приставка *C-*).

2. Кластеризация методом K-Means

В качестве первого алгоритма был выбран *K-Means* как наиболее распространённый метод кластеризации. Оптимальное число кластеров определялось методом локтя по графику зависимости значения функции инерции от числа кластеров k. Как для модели, обученной только на спектральных каналах, так и для модели на полном признаковом пространстве, точка локтя соответствовала значению $k = 12$.

Модель *S-KMeans* ($k = 12$), обученная на спектральных каналах, показала приемлемое качество визуальной кластеризации: большинству природных объектов поставлены в соответствие разумные классы, плавно передан рельеф местности. Однако внутренние озёра отнесены к тому же классу, что

и тёмный лёд, что свидетельствует об ограниченности только спектральных признаков для различения близких по отражательным свойствам объектов.

Модель *C-KMeans* (*k* = 12), обученная на расширенном признаковом пространстве, визуально хуже передаёт плавность рельефа, однако значительно лучше выделяет границы объектов: айсбергов, береговых линий, контрастных участков рек. Кроме того, внутренние водоёмы в данном случае отнесены к классу океана, что также является шагом в сторону содержательной интерпретации.

Дополнительно была рассмотрена модель *C-KMeans* (*k* = 7), полученная максимизацией силуэтного коэффициента. Несмотря на формально лучшее значение метрики *Silhouette* (0,5911), такая кластеризация хуже передаёт рельеф и выделяет существенно меньше содержательных объектов, в связи с чем в дальнейшем анализе она не использовалась. Визуальные результаты кластеризации алгоритмом *K-Means* для обоих признаковых пространств приведены на рисунке 2, а сводка значений метрик для рассмотренных конфигураций – в таблице 1.

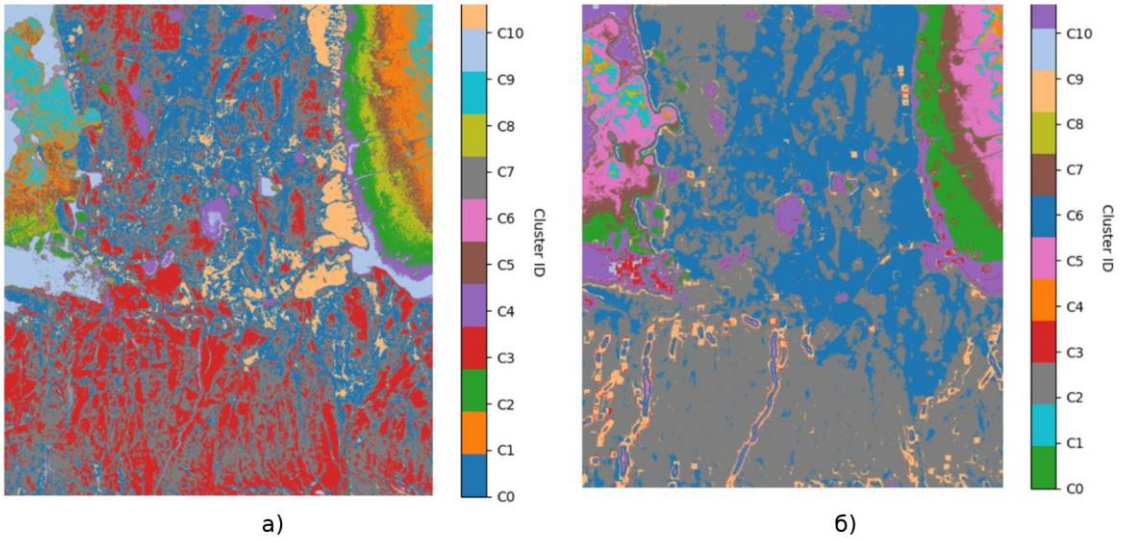

Рис. 2. Результаты алгоритма K-Means (k = 12): а) S-KMeans на спектральных каналах; б) C-KMeans на расширенном признаковом пространстве

Табл. 1. Метрики качества кластеризации алгоритмом K-Means

Метрика	**S-KMeans (k = 12)**	**C-KMeans (k = 12)**	**C-KMeans (k = 7)**
Silhouette Score (↑, макс. 1,0)	0,3649	0,3746	0,5911
Davies–Bouldin Index (↓, мин. 0,0)	0,9527	0,7609	0,8596
Calinski–Harabasz Index (↑)	935 055	3 985 908	1 015 742

По совокупности количественных метрик и качественной визуальной оценки модель *C-KMeans* (*k* = 12) превосходит *S-KMeans* (*k* = 12): индекс *Davies–Bouldin* снижается на 20 %, индекс *Calinski–Harabasz* возрастает более чем в четыре раза. Это подтверждает целесообразность включения локальных статистик в признаковое пространство для решения задач кластеризации ДЗЗ.

3. Кластеризация методом гауссовых смесей

В качестве второго подхода рассмотрен алгоритм *Gaussian Mixture Model* (*GMM*), использующий вероятностное описание данных в виде смеси многомерных нормальных распределений. Параметры модели (математические ожидания, ковариационные матрицы и веса компонент) оценивались алгоритмом максимизации правдоподобия *EM*, с начальным приближением методом k-means++. Оптимальное число компонент, как и для *K-Means*, было равно *k* = 12.

Модель *S-GMM* (k = 12), обученная только на спектральных каналах, визуально близка к *S-KMeans*, однако обладает большей детализацией: реки прослеживаются по всей своей длине, а не только на наиболее контрастных горных участках; на карте появляется дорога; частично выделяются айсберги. При этом сохраняется проблема отнесения внутренних озёр к классу тёмного льда.

Модель *C-GMM* (k = 12), обученная на расширенном признаковом пространстве, демонстрирует наиболее качественное выделение границ объектов среди всех рассмотренных моделей. По сравнению с *C-KMeans* (k = 12) увеличивается толщина границ айсбергов с одного до двух пикселей, что, с одной стороны, может усложнить последующую обработку, однако делает границы более устойчивыми и менее прерывистыми. Дополнительно выделяются трещины во льду, а также переходные зоны лёд–вода. Визуальные результаты кластеризации обеих моделей *GMM* приведены на рисунке 3, сводка значений метрик качества для всех четырёх основных моделей – в таблице 2.

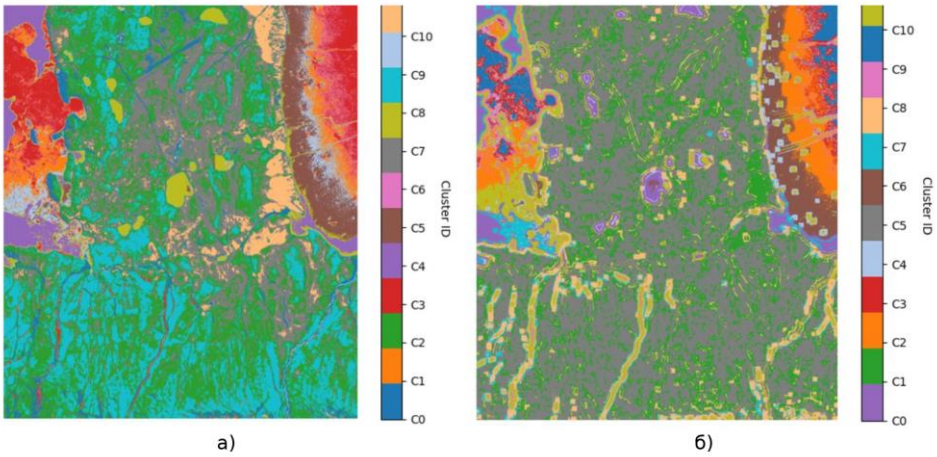

Рис. 3. Результаты алгоритма Gaussian Mixture (k = 12): а) S-GMM на спектральных каналах; б) C-GMM на расширенном признаковом пространстве. Хорошо видны выделенные границы рек, дороги в правой части снимка, айсберги и трещины во льду

Табл. 2. Сводные метрики качества кластеризации

Метрика	**S-KMeans (k = 12)**	**C-KMeans (k = 12)**	**S-GMM (k = 12)**	**C-GMM (k = 12)**
Silhouette Score (↑)	0,3649	0,3746	0,2900	0,0624
Davies–Bouldin Index (↓)	0,9527	0,7609	1,1451	2,4530
Calinski–Harabasz Index (↑)	935 055	3 985 908	1 857 846	333 540

По количественным метрикам алгоритмы *K-Means* существенно превосходят гауссовы смеси, особенно на расширенном признаковом пространстве. Однако визуальная оценка не позволяет сделать столь же однозначного вывода: модель *C-GMM*, имеющая наихудшие значения всех трёх индексов, тем не менее даёт наиболее информативное выделение границ объектов. Это указывает на ограниченную применимость рассмотренных метрик к задаче ДЗЗ: они склонны вознаграждать формирование плотных сферических кластеров, тогда как для выделения границ типичная конфигурация – это рассредоточенные кластеры, разделяемые лишь значениями локальных статистик. Поиск более адекватных метрик качества для подобных задач можно рассматривать как одно из направлений дальнейших исследований.

4. Проверка устойчивости при переносе на новые изображения

Важной практической характеристикой алгоритма кластеризации ДЗЗ является устойчивость классов при переносе на новые территории. Для оценки этого свойства была проведена кластеризация модифицированного снимка, на котором была наложена маска суши и удалены ледовые объекты.

Наиболее устойчивым подходом оказалось использование параметров распределения, полученных моделью *S-GMM* ($k = 12$) на исходном изображении, в качестве начального приближения для новой модели *GMM*, обучаемой на модифицированном снимке. В таком режиме «тёплого старта» (*warm start*) алгоритму потребовалось не более 50 итераций *EM* для сходимости, при этом классификация на новой поверхности оказалась содержательной. Алгоритм *C-GMM*, использующий расширенное признаковое пространство, отработал на этом сценарии хуже: после исчезновения ледовых объектов оптимальное число кластеров существенно меняется, и перенос параметров приводит к избыточной сегментации. Алгоритм *K-Means* в режиме тёплого старта устойчивых результатов не показал.

Таким образом, в задачах последовательной обработки серий снимков, охватывающих переходные зоны (например, прибрежные акватории, выходящие на материк), целесообразно использовать *S-GMM* с переносом параметров начального приближения.

Заключение

В работе проведён сравнительный анализ двух алгоритмов кластеризации без учителя – *K-Means* и *Gaussian Mixture Model* – применительно к задаче анализа спутниковых снимков. Рассмотрены конфигурации с двумя вариантами признакового пространства: только спектральные каналы (*S*-варианты) и расширенное пространство с локальными статистиками яркости в скользящем окне 15×15 (*C*-варианты).

На основании полученных результатов сформулированы следующие рекомендации по выбору алгоритма в зависимости от прикладной задачи (таблица 3): для быстрой базовой кластеризации одиночного снимка оптимален алгоритм *S-KMeans*, обеспечивающий приемлемое визуальное качество при минимальных вычислительных затратах; для задач детекции линейных и контрастных объектов (айсбергов, дорог, рек, трещин) наилучшие результаты даёт алгоритм *C-GMM*, обладающий максимальной детализацией границ; для задач, требующих переноса кластеризации на новые снимки той же сцены, рекомендуется *S-GMM*, устойчивый к смене признакового пространства и быстро сходящийся в режиме тёплого старта.

Табл. 3. Рекомендации по выбору алгоритма кластеризации

Задача	**Алгоритм**	**Обоснование**
Быстрая кластеризация одного снимка	S-KMeans	Минимальные вычислительные затраты, приемлемое визуальное качество.
Детекция линейных и контрастных объектов (айсберги, дороги, реки)	C-GMM	Наилучшая детализация границ объектов, выделение мелких структур.
Перенос кластеризации на новые снимки	S-GMM	Устойчивость к смене признакового пространства, быстрая сходимость (около 50 итераций).

Дополнительно показано, что классические метрики качества кластеризации (*Silhouette, Davies–Bouldin, Calinski–Harabasz*) плохо согласуются с визуальной оценкой результатов на задачах выделения границ объектов на ДЗЗ. Разработка специализированных метрик качества и расширение признакового пространства иными локальными дескрипторами (например, текстурными) представляют интерес для дальнейших исследований.

Список источников

1. Hartigan J. A., Wong M. A. Algorithm AS 136: A K-Means Clustering Algorithm // Journal of the Royal Statistical Society. Series C (Applied Statistics). – 1979. – Vol. 28. – №. 1. – Pp. 100–108.

2. Reynolds D. A. Gaussian Mixture Models // Encyclopedia of Biometrics / S. Z. Li, A. K. Jain (eds.). – Boston, MA: Springer, 2009. – Pp. 659–663.

3. Bishop C. M. Pattern Recognition and Machine Learning. – New York: Springer, 2006. – 738 p.

4. Scikit-learn: Machine Learning in Python / F. Pedregosa, G. Varoquaux, A. Gramfort [et al.] // Journal of Machine Learning Research. – 2011. – Vol. 12. – Pp. 2825–2830.

5. Rousseeuw P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis // Journal of Computational and Applied Mathematics. – 1987. – Vol. 20. – Pp. 53–65.

6. Davies D. L., Bouldin D. W. A Cluster Separation Measure // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 1979. – Vol. PAMI-1. – №. 2. – Pp. 224–227.

7. Caliński T., Harabasz J. A dendrite method for cluster analysis // Communications in Statistics. – 1974. – Vol. 3. – №. 1. – Pp. 1–27.

8. Lu D., Weng Q. A survey of image classification methods and techniques for improving classification performance // International Journal of Remote Sensing. – 2007. – Vol. 28. – №. 5. – Pp. 823–870.

9. Шовенгердт Р. А. Дистанционное зондирование. Модели и методы обработки изображений. – М.: Техносфера, 2010. – 560 с.

10. Кашкин В. Б. Дистанционное зондирование Земли из космоса. Цифровая обработка изображений / В. Б. Кашкин, А. И. Сухинин. – Москва: Логос, 2001. – 264 с.