

# ИЕРАРХИЧЕСКИЙ КЛАСТЕРНЫЙ АЛГОРИТМ ДЛЯ ТЕКСТУРНЫХ ДАННЫХ ПОВЕРХНОСТИ ЗЕМЛИ

В. С. Сидорова

*Институт вычислительной математики и математической геофизики СО РАН, 630090, Новосибирск*

УДК 528.852

Рассматривается автоматическая кластеризация и последующая сегментация аэроснимков по текстурным признакам. Использован дивизимный гистограммный иерархический алгоритм с поиском кластеров заданной делимости. Учтены особенности сегментации по статистическим текстурным признакам. Параметры модели изображения, известной как SAR, используются в качестве текстурных признаков для автоматической неконтролируемой классификации лесных ландшафтов на аэроснимках.

**Ключевые слова:** дистанционное зондирование, неконтролируемая классификация, многомерная гистограмма, делимость кластеров, иерархический алгоритм, текстура.

## Введение

В предлагаемом иерархическом алгоритме кластеризации текстурных изображений акцент сделан на оценку делимости кластеров. В [1] был предложен подобный алгоритм для мультиспектральных изображений. Однако статистические текстурные признаки имеют свои особенности, поэтому здесь дана модификация алгоритма. Внутри каждого кластера дальнейшая кластеризация осуществляется по алгоритму Нарендры [2]. В методе [2] гистограмма рассматривается как плотность вероятности векторов. Быстрый непараметрический алгоритм Нарендры разделяет векторное пространство признаков по унимодальным кластерам, модальные векторы которых соответствуют всем локальным максимумам гистограммы, границы кластеров проходят по долинам, областям низкой плотности векторного пространства. Этот жесткий алгоритм, используя графы, одновременно и быстро решает три основные задачи кластеризации: находит моды кластеров, устанавливает границы между кластерами и все вектора разносит по кластерам. Иерархический алгоритм позволяет автоматически выбирать различную детальность квантования в подобластях векторного пространства в зависимости от делимости кластеров в них. Что позволяет получить существенно меньше и хорошо разделенных кластеров. В этой работе рассчитываются статистические текстурные признаки в окрестности каждой точки по модели SAR [3]. В качестве примера рассматриваются изображения леса. На черно-белых полутонных аэроснимках масштаба 1:50000, текстура лесных сообществ формируется чередованием темных и светлых пятен, соответствующих группам деревьев различных пород. Визуальный анализ аэроснимков (дешифрирование) является составной частью инвентаризации и мониторинга в лесоводстве. Текстурных свойств изображений часто бывает достаточно, чтобы различить тип леса и его возраст [4].

## 1 Текстурные признаки

В качестве текстурных признаков используются приближенные значения параметров модели случайного некаузального поля SAR [3]. На дискретном двумерном фрагменте  $M \times M$  изображения определена матрица уровней серого тона  $g(x, y)$ ;  $x, y = 0, 1, \dots, M-1$ . Для приближенного вычисления параметров модели SAR в [3] предложен метод максимального правдоподобия. Эти параметры  $f_N = (\hat{\theta}_{(i,j)}); (i, j) \in N, (\hat{\rho}_N)$  могут интерпретироваться с точки зрения некоторых визуальных свойств текстуры. Величины  $\hat{\theta}$  могут характеризовать направленность. Параметр  $\hat{\rho}_N$  имеет прямую связь со степенью зернистости текстуры. Чем больше

значение  $\hat{\rho}_N$  образца, тем тоньше текстура. Рассматриваемое изображение предварительно эквализуется [5]. Вокруг каждой точки изображения рассматривается квадратное окно  $M \times M$ . Матрица  $g(x, y)$  модели SAR нормализуется вычитанием среднего из значения уровня серого в каждой точке окна и делением на стандартное отклонение. Найденные для каждой точки изображения параметры модели  $f_N$  растягиваются затем в пределы от 0 до 255 и используются как признаки для классификации, возможно в сочетании с другими характеристиками.

## 2 Меры качества

Параметром детальности является число уровней квантования  $N$  векторного пространства признаков. Мера изолированности для отдельного унимодального кластера  $M(N)$  (1), и мера качества распределения в целом  $m(N)$  по  $K(N)$  кластерам (2) были определены в [6]:

$$M^j(N) = \frac{1}{H^j(N)} \frac{1}{B^j(N)} \sum_{i=1}^{B^j(N)} h_i^j(N) \quad (1)$$

$$M(N) = \frac{1}{K(N)} \sum_{j=1}^{K(N)} M^j(N), \quad (2)$$

где  $h_i^j(N)$  — значение гистограммы в  $i$ -той точке границы кластера  $j$ ,  $B^j(N)$  — число точек границы кластера,  $H^j(N)$  — максимальное значение гистограммы. Чем меньше (1), тем лучше кластер. Минимумы (2) соответствуют лучшим классификациям. Границы кластера легко вычислить, используя список соседей всех векторов, построенный алгоритмом Нарендры. Всегда  $M^j(N) \leq 1$  и  $M(N) \leq 1$ . Можно показать, что мера (1) характеризует отношение объема точек вблизи границ кластера к его полному объему. Поэтому кластер характеризуется индивидуальной отделимостью. Мера (2) удовлетворяет условиям кластерной достоверности [7]. Она уменьшается с ростом компактности кластеров и с ростом расстояния между модальными векторами кластеров. На границах кластеров плотность вероятности не опускается до нуля, если кластеры тесно расположены. Такие кластеры типичны для данных дистанционного зондирования земной поверхности. Часто предлагается измерять компактность кластера среднеквадратичным отклонением признаков. Но среднеквадратичное отклонение признаков для таких кластеров зависит не только от их компактности, но и от их радиуса, поэтому при тесном расположении кластеров возникает неопределенность в нахождении разделимости. Мера (2) соотносит не расстояния, а плотности вероятности, что, кстати, намного дешевле.

## 3 Особенности сегментации по текстурным признакам

Статистические текстурные признаки являются пространственными характеристиками, а не точечными в отличие от спектральных, поэтому есть некоторые особенности кластеризации. Рассмотрены следующие особенности, требующие модификации кластерного алгоритма:

- а) статистика собирается по окрестности точки — окну, и требуется определить его размер;
- б) пиксели окна могут на границах объектов относиться к различным текстурам и создавать ложные кластеры, которые требуют специальной обработки.

Выбор размера окна. Текстурные признаки вычисляются в окрестности (окне) каждой точки изображения. Предложен автоматический выбор окна одного размера для всех точек изображения. От размера окна существенно зависят результаты кластеризации. Следующие соображения принимались во внимание. Размер должен быть достаточно велик, чтобы самой крупнозернистой текстуре была обеспечено достаточно статистики. Это можно проверить, постепенно увеличивая размер окна и оценивая значение текстурных признаков. Когда значение признака стабилизируется, то можно найти искомый размер. Если на изображении лишь одна текстура, то искомый размер должен быть не меньше найденного, но может быть как угодно большим. Однако, реальное изображение содержит различные текстуры, по-этому, чтобы точнее провести сегментацию, нужно использовать найденный, наименьший размер. Если изображение содержит достаточно протяженные объекты с одинаковой текстурой, а не одни границы, то для правильного размера окна все результаты кластеризации будут оставаться близкими (включая число полученных кластеров) для соседних значений размера окна. Особенно это справедливо, если на границах кластеров в векторном пространстве плотность векторов мала, а это свойство хорошо разделенных кластеров. Ранее сравнивались количества

полученных кластеров в оптимальных распределениях (соответствующих минимумам меры разделимости кластеров) при изменении детальности в алгоритме Нарендры. Теперь сравниваются количества полученных кластеров в распределениях, полученных для низких значений меры отделимости кластеров, заданной для нового иерархического гистограммного алгоритма кластеризации текстурных данных ДЗЗ.

Другая особенность текстуры в том, что на границах объектов с разной текстурой появляются ложные кластеры из-за того, что текстурные признаки рассчитываются по протяженному окну точек изображения, и в это окно попадают точки соседних различных текстур. Обычно эти ложные кластеры формируют ряд узких сегментов на границах текстур в плоскости изображения. После проведения гистограммной иерархической кластеризации текстурных признаков и глобальной предварительной сегментации изображения в соответствии с кластеризацией проводится обработка ложных кластеров. Для их индикации применяется пороговый метод, для ложных кластеров доля граничных точек по отношению к площади превышает заданный порог. Для объединения ложных кластеров и их присоединения к истинному предложена модификация алгоритма [8] (с анализом контекста изображения и кластерной разделимости в пространстве признаков). Не ложные кластеры назовем основными. Ложные кластеры объединяются с основными по следующему предложенному правилу.

Для ложного кластера сначала отыскиваются два его наиболее представительных соседа по изображению. Представительность определяется по длине границы. Для ее вычисления карта кластеров, полученная основным алгоритмом, единожды автоматически построчно просматривается, и суммируется число граничных точек сегментов каждого кластера по вертикали и горизонтали. Затем из этих двух представительных выбирается тот кластер, который хуже отделен от ложного в векторном пространстве. В формуле (1) для ложного кластера суммируются только векторы границы с представительным кластером.

Ложный кластер присоединяется к тому из двух претендентов, для которого значение меры (3) больше, и на границе с ним средняя плотность его векторов больше. Алгоритм для объединения последовательности ложных кластеров. На границах объектов с различной текстурой часто образуется последовательность ложных кластеров, чьи тонкие сегменты заключены между двумя соседними. Обычно они плохо разделены в векторном пространстве. Автоматически выбирается какое-то число наиболее крупных и компактных в плоскости изображения кластеров. Это число может быть вычислено, если задать другие ограничения: по минимальному объему кластера, или по значению максимальной разделимости кластера, или по порогу компактности  $d$ . Выбранные кластеры объявляются основными, и к ним присоединяются остальные (ложные) следующим образом. Рассмотрим последовательно список всех кластеров. Если кластер ложный, то исследуем двух его соседей для присоединения по описанному выше правилу, учитывая различный размер ячеек квантования векторного пространства в соседних кластерах, построенных иерархическим алгоритмом. Если кластер для присоединения окажется основным, то ложный присоединим к нему, иначе начнем выстраивать цепочку ложных кластеров, пока не попадется основной, и тогда всей цепочке присвоим его номер. Если же цепочка замкнется, то объявим ее новым кластером. В результате работы алгоритма соседние ложные кластеры будут объединены в областях высокой плотности векторов, граница же между объединенными кластерами будет проходить в областях низкой плотности, следуя принципу кластеризации метода Нарендры.

## 4 Пример

На рис. 1а представлено исходное эквализованное черно-белое изображение лесного ландшафта, полученное с самолета, масштаба 1:50 000. Эквализация оставила 30 уровней из 256. На рис. 1б лесоводами представлена карта, полученная с применением наземной таксации. Каждому выделу с номером на карте рис. 1б соотносится его описание по преобладающей породе. Здесь представлены кедровники и сосняки различных возрастных фаз, а также березняки II фазы развития, болота и немного (две 27 и 20) вырубков. Кедровники окрашены синими тонами, сосняки коричневыми и розовыми, березняки белыми, болота и вырубки черными. Сверху вниз протекает речка, сопровождаемая старицами (красные). Трудность при кластеризации состоит в том, что отдельные фазы развития кедровников и сосняков плохо различимы на исходном снимке. На рис. 1а представлено изображение лесного ландшафта ( $1781 \times 1157$ ), На рис. 1б картосхема наземной таксации.

Для вычисления текстурных признаков размер окна был определен  $16 \times 16$  (здесь одинаковый для всех областей). Заметим, что только для размеров окна при  $i = 3$  и  $i = 4$  (с близкими значениями числа кластеров)  $16 \times 16$  и  $18 \times 18$ , кластеры различных типов леса хорошо разделились с заданной отделимостью  $d = 0.15$ . Для кластеризации был выбран размер  $16 \times 16$ . Этот размер практически соответствует размеру области

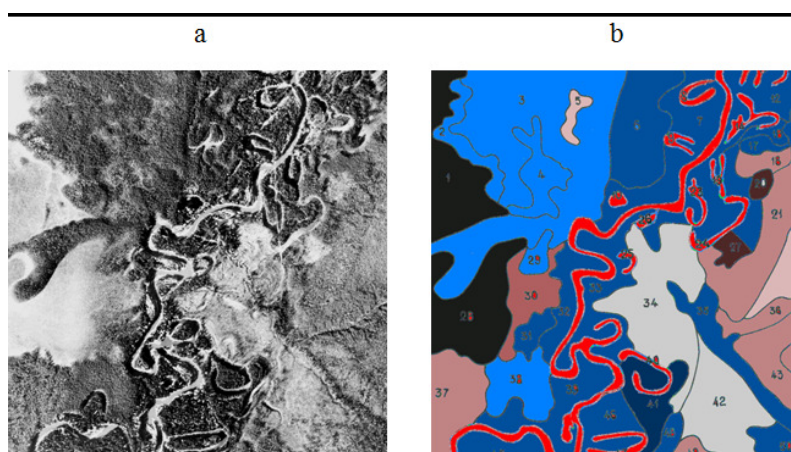


Рис. 1: а. Эквиализованное изображение лесного ландшафта на аэроснимке. В средней части по диагонали находится река. б. Картосхема наземной таксации, указаны номера выделов, их описание: 1, 28 — болота, 34, 42 — березовые леса, 2, 4, 6, 7, 12, 16, 17, 29, 31, 32, 33, 35, 38, 39, 41, 45, 46, 48, 50 — кедровые леса, сосновые леса: 18, 21, 22, 36, 43, 30, 37, 10, 12, 13, 27, 49

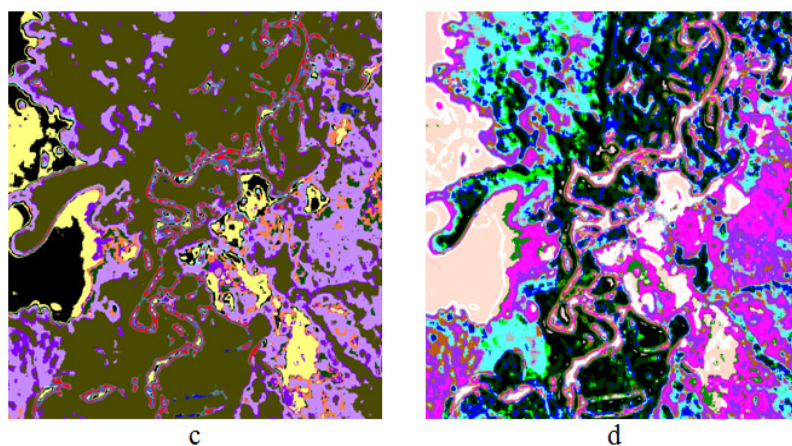


Рис. 2: с. Кластеризация для семи этапов иерархии. d. Кластеризация для четырнадцати этапов иерархии.

таксации, принятом в лесоводстве (1 га=100м × 100м). Так как текстура рассматриваемых лесов на изображении изотропна, то в качестве признаков модели SAR рассмотрим только признак ( $\hat{\rho}$ ), измеряющий степень зернистости. Таким образом, для кластеризации имеем три признака: средний уровень серого, стандартное отклонение и ( $\hat{\rho}$ ). отдельные фазы развития кедровников и сосняков плохо различимы на исходном снимке.

На рис. 2с при заданной отделимости кластера  $d = 0.15$  (всегда  $0 < d < 1$ ) представлена кластеризация семи этапов иерархии. Получено восемь кластеров. На этом этапе иерархии кластеры, относящиеся к соснякам и кедровникам, полностью разделились. Также выделились березники и болота. Каппа Коэна [9], характеризующая степень соответствия классификаций равна 0.91, что считается очень хорошим показателем согласованности. (Заметим, что выдел может содержать некоторый процент деревьев, отличных от основного насаждения). На рис. 2d полученная карта для 14 этапов иерархии. Здесь кластеры разделились по возрастам сообществ. Каппа Коэна равна 0.83.

## Заключение

Используя текстурные признаки аэроснимка, мы автоматически построили карту кластеров, которая соответствует карте лесных насаждений, полученной лесоводами с помощью наземной таксации. Иерархический подход с заданием максимальной отделимости кластера и дифференцированным выбором детальности в

подобластях пространства признаков позволил избавиться от плохо изолированных кластеров. Полученных кластеров немного. Они хорошо разделены, не смотря на то, что уровень квантования для отдельных областей высок, что позволило различить визуально близкие текстуры.

## Список литературы

- [1] V.S.Sidorova. Detecting Clusters of Specified Separability for Multispectral Data on Various Hierarchical Levels // Pattern Recognition and Image Analysis. 2014. V. 24, No. 1. P. 151–155.
- [2] Narendra P.M., Goldberg M. A Non-parametric Clustering Scheme for LANDSAT // Pattern Recognition. 1977. No. 9. P. 207–215.
- [3] Kashyap R.L., Chellapa R. Estimation and Choice of Neighbors in Spatial Interaction Models of Images. // IEEE Trans. Inform. Theory. 1983. vol.1. P. 60–72.
- [4] Sedykh V.H. // Shaping the Cedar Forests. Novosibirsk “Science” 1979.
- [5] R.M. Haralick, K. Shanmugam, I. Dinstein. Textural Features for Image Classification. IEEE Trans. Syst. Man. Cybern. 1973. Vol. SMS-3. P. 610–621
- [6] Sidorova V.S. Histogram Clustering Validation for Multispectral Image. Avtometria. 2007. Vol.43, No. 1. P. 37–43.
- [7] Halkidi M., Batistakis Y. and Vazirgiannis M. On clustering validation techniques. Journal of Intelligent Information Systems. 2001. 17(2-3). P. 107.
- [8] В.С.Сидорова. Алгоритм кластеризации текстурных данных дистанционного зондирования. Ж. Автометрии 2010 Т. 46, № 5, с. 43–52.
- [9] Грауэр Л.В., Архипова О.А. // Непараметрические критерии независимости. 2014. Санкт-Петербург -S Center, 1/30. [https://compscicenter.ru/media/slides/math\\_stat\\_2014\\_spring/2014\\_03\\_28\\_math\\_stat\\_2014\\_spring\\_1.pdf](https://compscicenter.ru/media/slides/math_stat_2014_spring/2014_03_28_math_stat_2014_spring_1.pdf)

*Валерия Сергеевна Сидорова — н.с. Института вычислительной  
математики и математической геофизики СО РАН;  
e-mail: svs@sscc.ru.*

*Дата поступления — 31 мая 2017 г.*