

Многомерная гистограмма и разделение векторного пространства признаков по унимодальным кластерам ¹

Валерия Сидорова
Институт Вычислительной Математики и Математической Геофизики СО РАН,
Новосибирск, Россия
svs@ooi.sccc.ru

Аннотация

Рассматривается быстрый непараметрический алгоритм неконтролируемой классификации многоспектральных данных, основанный на разделении многомерной гистограммы по унимодальным кластерам. Предлагается система хеширования для построения гистограммы. Алгоритм не использует никаких предположений о функциях распределения различных классов. Он также не требует никаких априорных данных (число кластеров, количество итераций и т.д.). Предлагается автоматическое управление детальностью кластеризации. Иллюстрируется работа алгоритма для трех, пяти и восьмиспектральных космических снимков поверхности Земли.

Ключевые слова: обработка изображений, кластерный анализ, многомерная гистограмма, дистанционное зондирование.

1. ВВЕДЕНИЕ

В дистанционном зондировании кластерный анализ применяется при анализе многоспектральных изображений, чтобы автоматически сгруппировать данные и сопоставить их классам земных покрытий. Особенностью космических снимков является громадный объем данных.

Предложено много кластерных алгоритмов. В их основе лежат различные подходы, например: кластеризация по K центрам, (широко известная ISODATA основана на этом методе), иерархическая, пороговая, и, наконец, кластеризация по многомерной гистограмме [1-10].

Здесь описана реализация метода разделения гистограммы по унимодальным кластерам на основе идеи построения графов [10]. Все значительные пики гистограммы объявляются кластерами. Это быстрый алгоритм, он не требует никаких априорных данных, не итеративный.

Если полученные кластеры будут очень хорошо разделены в пространстве признаков, то алгоритм можно непосредственно использовать для классификации объектов. Иначе его можно применять для предварительной группировки данных. Известно, насколько важно разделить пространство измерений на унимодальные кластеры для последующей классификации с применением теории распознавания образов [3].

Недостатком алгоритма является то, что он требует большой объем оперативной памяти компьютера (RAM) для хранения гистограммы. Существуют различные методы борьбы с этим недостатком. Можно определять ячейки эквивалентности в векторном пространстве многоспектральных измерений, т.е.

квантовать измерения по заданным размерам. Число реально присутствующих векторов на изображении обычно существенно меньше размерности пространства вследствие межканальной корреляции. Этот факт может использоваться. Например, Гонг преобразует пространство векторов в собственное пространство [7]. Мы выстраиваем присутствующие на изображении вектора в линейный список, как в [9]. Мы предлагаем систему хеширования для быстрого доступа к списку векторов при построении гистограммы.

2. ПОСТРОЕНИЕ ГИСТОГРАММЫ

Имеется N спектральных каналов изображения, и спектральные интенсивности каждого канала квантуются на 256 градаций. В рассматриваемом алгоритме используется тот факт, что число реально присутствующих на изображении N -мерных векторов существенно меньше всего объема пространства измерений, равного 256^N байтов. Гистограмма представляется таблицей векторов и соответствующих им частот. Гистограмма генерируется считыванием изображения и сравнением многоспектрального вектора в каждой точке с текущим списком векторов.

Для быстрого доступа к списку векторов используется система хеширования. Если позволяет оперативная память компьютера, то можно использовать простейшую элементарную хеш-функцию, которая рассчитывается как упаковка в целое слово значений интенсивности в N спектральных каналах (если $N \leq 3$). Эта хеш-функция взаимно однозначно связана со значением многоспектрального вектора, и, используя ее, гистограмму можно строить в лоб. Для этого нужно выделить для хранения гистограммы объем памяти, равный объему всего пространства измерений - $256^N * 4$ байтов.

Если предусматривается маскирование младших (шумовых) бит в каждом спектральном канале, то соответственно памяти нужно меньше. После построения гистограммы выбрасываются ячейки памяти, для которых значение гистограммы оказалось равным нулю. Должна получиться гистограмма для упорядоченных по возрастанию упаковок присутствующих векторов. Лишняя память затем освобождается.

Когда кластеры будут получены и нужно будет построить их карту, исходное изображение прочитаем снова, сожмем гистограмму и получим функцию новых номеров векторов от старых.

Однако уже при $N=4$ прямое построение гистограммы проблематично. В этом случае для вычисления XEIII-адреса ка-

¹ Работа выполнена частично при финансовой поддержке Российского фонда фундаментальных исследований (проект № 05-07-90057-В).

ждого вектора можно предложить, например, функцию вида[9]:

$$H = \sum_{i=1}^N C_i U_i,$$

где U_i - значение i -ой компоненты N спектрального вектора, C_i - постоянные для всего изображения компоненты.

Важно, чтобы значение адреса H было ограничено заданным числом HM , равным допустимому количеству присутствующих векторов.

Каждому значению ХЕШ-адреса теперь соответствует несколько различных векторов. Ясно, что чем большим можно задать максимальное H , тем меньше встретится неоднозначных векторов, и тем быстрее будет построена гистограмма. Для борьбы с неоднозначностью предлагается следующий подход. Создадим две таблицы. В одну (назовем ее ХЕШ-таблицей) будут заноситься вектора (в соответствии с их ХЕШ-номером) и их частоты, по мере просмотра изображения. Однако если место в таблице с полученным ХЕШ-номером окажется занятым, то перейдем к другой таблице (назовем ее SHELL-таблицей). В нее поместим вектора по принципу возрастания их значения. При этом новым вектором, только что считанным (и значением его частоты) мы заменим соответствующие данному ХЕШ-номеру значения в текущей ХЕШ-таблице.

Замещенные же значения переместим в SHELL-таблицу. Если там уже есть такой вектор, то заменим значение гистограммы на новое, а если нет, то добавим вектор. Предполагается, что, во-первых, большая часть векторов все-таки взаимно однозначно связана с ХЕШ-номерами, и таких перемещений по таблицам будет немного. Во-вторых, естественное изображение состоит из протяженных относительно однородных объектов. Поэтому отдаленные значения векторов, соответствующие одинаковым ХЕШ-номерам, будут соответствовать различным объектам, пространственно далеко расположенным. Т.е. предполагается, что ближайшие по считываемому изображению вектора не будут часто и резко меняться, к тому же еще таким образом, что их ХЕШ-номера окажутся одинаковыми. Наоборот, они долго должны оставаться в текущей ХЕШ-таблице с быстрым доступом. После однократного просмотра изображения, обе таблицы окажутся упорядоченными по возрастанию векторов. Их останется соединить в одну заполненную упорядоченную таблицу с помощью бинарной SHELL-сортировки. Для хранения гистограммы и списка векторов требуется $2(4+N)*HM$ байт. Выбор значения HM зависит от объема RAM компьютера.

Алгоритм кластеризации существенно использует тот факт, что вектора упорядочены по возрастанию. Сравнение векторов имеет здесь следующий смысл. Будем рассматривать два случая: 1) когда размерность пространства N больше трех, 2) меньше или равна трем. В первом случае каждый вектор хранится в памяти как поле однобайтовых векторов, где в каждом байте помещается значение яркости соответствующего спектрального канала. Позиция каждого канала в поле играет роль разряда числа, причем, чем старше разряд поля, тем левее он расположен. Сравнение векторов и необходимые арифметические операции над векторами осуществляется последовательно над байтами обычными логическими и арифметическими операциями, с учетом позиции байта в поле.

Во втором случае удобно запаковать яркости спектральных каналов в одно четырехбайтовое слово, т.е. в этом случае это та самая упаковка, которая является ХЕШ-номером вектора. Необходимые в дальнейшем арифметические и логические операции осуществляются над упаковками, т.е. над целыми словами. Это ускоряет работу алгоритма. Адекватность используемых арифметических действий рассматривается ниже подробно.

В любом случае в дальнейшем работаем с векторами таблицы, упорядоченными по возрастанию - это, безусловно, ускоряет и облегчает работу. Каждому вектору сопоставлено значение гистограммы в таблице.

3. КЛАСТЕРИЗАЦИЯ ГИСТОГРАММЫ

Суть алгоритма состоит в нахождении локальных максимумов многомерной гистограммы в дискретном векторном пространстве и отнесении векторов к соответствующим максимумам. Для каждого вектора составляется список его ближайших соседей (см. далее). В N -мерном пространстве их всего $3^N - 1$ для каждого вектора. Но в список заносятся

только присутствующие соседи. Для каждого вектора строится элементарный граф по направлению максимума положительного градиента плотности вероятности в списке соседей. Градиент рассчитывается как разность значения гистограммы соседа и самого вектора, деленную на расстояние

между ними. Такими расстояниями будут $1, 2, \dots, \sqrt{N}$. Если все соседи вектора имеют плотность меньшую, чем он сам, это значит, что гистограмме в этой точке векторного пространства соответствует локальный максимум и данный вектор является корнем соответствующего ориентированного графа.

Вектора связываются в деревья с помощью элементарных графов. Когда граф достигает локального максимума, то вся цепочка векторов относится к тому же кластеру, что и максимум (корень). Таким образом, формируются унимодальные кластеры. Вектора на границе кластеров относятся к кластерам по тому же принципу максимума положительного градиента. Некоторые трудности возникают, если гистограмма содержит плато. В каждой точке плато градиент равен нулю. Движение от вектора к вектору продолжается до встречи с помеченным кластером или до достижения максимума. В последнем случае всей цепочке векторов присваивается метка (положительное число) нового кластера, и начинается новая цепочка со следующего неклассифицированного вектора. Если же максимальное значение градиента по всем соседям данного вектора равно нулю, то это плато. В этом случае цепочка векторов помечается отрицательной меткой. Но поиск соседа продолжается, пока не попадется уже помеченный вектор. Можно остановиться, когда встретится положительный кластер, т.е. произойдет слияние плато с этим кластером. Или останавливаться только при замыкании цепочки с этим же плато. Хотя кластеры плато будут более раздроблены в последнем случае, но зато их легче анализировать в дальнейшем, т.к. они помечены отрицательной меткой. В обоих случаях не происходит закливания. Трассирование графов обеспечивает линейную зависимость количества операций от числа векторов. Более подробно в [11-13].

Для управления детальностью кластеризации мы задаем два режима: 1) убираем биты в каждом спектральном канале - грубый способ или 2)сглаживаем гистограмму заданное число раз по окрестности ближайших векторов. Ближайшими

считаются те, которые по любой из компонент отстоят от данного вектора не далее, чем на единицу. Это легко сделать, т.к. известны соседи каждого вектора. При сглаживании каждое значение гистограммы заменяется усредненным значением по точкам окрестности. Ориентиром является максимальное число кластеров, которое задает пользователь. Если после проведения кластеризации полученное число кластеров окажется больше заданного, то процедура кластеризации повторяется с новыми параметрами. Можно задать и автоматический режим. В этом режиме сначала происходит сглаживание до стабилизации, т.е. пока число кластеров не перестанет изменяться. Если число кластеров останется при этом больше заданного, то гистограмма возвращается в исходное состояние, и срезаются биты по спектральным каналам, и т.д. Следует заметить, что срез битов приводит к очень резкому изменению числа кластеров. Для плавного управления следует, видимо, вводить весовые множители, можно различные по спектральным каналам. Для управления детальностью можно также задать порог и рассматривать только те вектора, для которых значения гистограммы больше порога.

Алгоритм не гарантирует полного разделения кластеров в пространстве признаков (хотя срез битов задает определенное расстояние между всеми векторами) и могут в дальнейшем тестироваться по каким-либо критериям или анализироваться с помощью других кластерных алгоритмов. Но число кластеров значительно меньше входящих векторов, и все их характеристики занесены в полученные таблицы

4. СПИСОК БЛИЖАЙШИХ СОСЕДЕЙ

Окрестность упорядоченных по возрастанию упакованных векторов в дискретном пространстве содержит для каждого вектора $3^N - 1$ соседей. Любой из этих соседей отличается

от данного вектора некоторым сдвиговым вектором, состоящим из компонент $-1, 0, 1$. (Например, при $N=4$ сдвиговые вектора: $(0, 0, 0, 1)$, $(0, 0, 1, 0)$, $(0, 0, 1, 1)$ и т.д.). Пусть есть список M векторов. Для генерации списка соседей вектора каждый из M векторов должен быть проверен, попадает ли он в окрестность данного вектора. Дискретная природа данных позволяет просто и быстро найти соседей трассировкой линейного списка упорядоченных по возрастанию векторов. Эта схема требует не более $N \cdot M$ простых сравнений вместо $2M \cdot (M-1)$ вычислений нелинейной функции расстояния.

В алгоритме требуется производить сложение вектора последовательно со всеми сдвиговыми векторами. В случае $N > 3$, когда вектора представлены последовательностью байтов, удобно сначала проверить, может ли вообще существовать для данного сдвигового вектора сосед. Например, значение какой-либо компоненты вектора равно нулю, а значение соответствующей компоненты сдвигового вектора равно -1 . Для этого вектора в данном направлении соседа не будет. После проверки сложение векторов сводится к сложению компонент.

В случае $N \leq 3$, если вектора были упакованы с маскированием каждого байта хотя бы на единицу, то справедливо обычное арифметическое целочисленное сложение упакованного данного вектора и упакованного сдвигового вектора, т.к. обеспечивается правильное изменение компонент вектора, занимающих соответствующие байты в целом слове. Однако если ни один бит не отсекается, то необходима проверка каждого байта, является ли он заполненным или нуле-

вым, и для соответствующих направлений сразу определить отсутствие соседа.

Для каждого вектора в упорядоченном по возрастанию списке векторов ищется сосед по каждому сдвиговому вектору. Для найденного вектора соседом будет исходный вектор по противоположно направленному сдвиговому вектору с отрицательными компонентами. Его номер заносится в список соседей найденного вектора. Более подробно: для данного вектора определяется значение ближайшего соседа в данном направлении. Оно просто равно сумме данного вектора и соответствующего сдвигового вектора. Двигаясь от исходного вектора в сторону возрастания, сравниваем значения векторов списка с суммой. Если такой встретится, то его номер заносится в список соседей исходного вектора, и поиск в данном направлении прекращается. Поиск также прекращается, если встретится вектор, больший суммы, т.к. следующие вектора в списке могут быть только еще больше.

5. ЭКСПЕРИМЕНТЫ

Алгоритм обеспечивает разделение гистограммы на унимодальные кластеры. Их характеристики: значения максимального значения гистограммы и компонент соответствующего вектора, площади, граничные точки по всем спектральным каналам, ковариации заносятся в таблицы и могут быть показаны на экране и сохранены в файл. Карта кластеров получается в виде BMP-файла. Алгоритм реализован средствами VISUAL C++ в многооконном режиме, что позволяет работать с несколькими документами. Параметры кластеризации задаются в диалоговых окнах. Кластеры на полученной карте можно указывать курсором. Соответствующий кластер высвечивается инвертированным цветом и выводится его номер. Можно задать число самых больших по площади кластеров для вывода на экран или оконтурить их.

Раскраска кластеров на картах производится автоматически в соответствии со значением их модальных векторов. Для трехспектральных изображений цвет каждого кластера задавался равным цвету его модального вектора. Для многоспектральных значение каждой компоненты RGB на карте определялось как взвешенная линейная комбинация компонент векторов. Коэффициенты позволяют регулировать раскраску: получать контрастную с большим количеством оттенков или более мягкую и близкую к каким-то выбранным спектральным каналам.

Проиллюстрируем некоторые результаты.

Фрагмент космического трехспектрального снимка степного ландшафта представлен в верхней части рис.1. При заданных параметрах изображение содержит 42540 различных векторов. В нижней части рис.1 представлена полученная кластерная карта исходного изображения. Кластеризация проводилась в автоматическом режиме. Максимальное число кластеров задавалось равным 200. Было срезано по два бита в каждом спектральном канале. Гистограмма была сглажена дважды. Для этих параметров было получено 170 унимодальных кластеров. Три самых больших кластера оконтурены. Указанный кластер 6 выделен инвертированным цветом (желтый цвет этого кластера перешел в голубой). В правой части рис.1 показана полученная таблица характеристик кластеров.

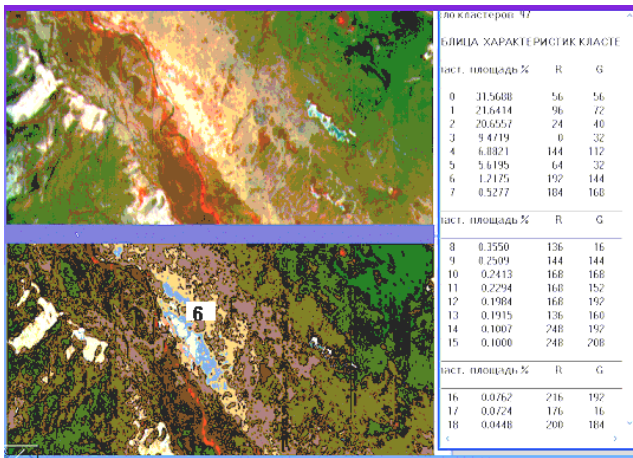


Рис.1. Вверху космический трехспектральный снимок степного ландшафта. По диагонали русло пересохшей реки. Большие светлые пятна в левой части снимка – озера. Внизу кластерная карта.

С помощью алгоритма исследовались изображения тающих снегов, полученные весной со спутника NOAA над территорией Западной Сибири. Изображения сканированы в пяти спектральных каналах: в видимой части спектра (в синем и зеленом диапазонах), в ближнем инфракрасном на длине волны 3.7 мкм и еще в двух инфракрасных 11 и 12 мкм.

Сначала кластерный алгоритм применялся для трехспектральных изображений. Были выбраны каналы: два в видимой области и в ближнем инфракрасном диапазоне. Изображение было записано как RGB файл, где B и G соответствовали синему и зеленому, а R инфракрасному диапазону.



Рис.2. Трехспектральный космический снимок тающих снегов.

При сопоставлении полученных цветных изображений с картами таяния снега, предоставленными метеорологами, видно, что выбранные спектральные каналы позволяют визуальнo различить стадии таяния снега: сплошной снег соответствует светло-голубому цвету, по мере таяния снега голубой темнеет и краснеет, это соответствует усилению инфракрасного излучения от теплеющих объектов поверхности Земли. Участки суши с растаявшим снегом имеют красный цвет на изображении.

Анализ полученных в результате работы алгоритма кластерных карт и характеристик кластеров показал, что пространство измерений хорошо разбивается на кластеры. То есть кластеров получается немного, они хорошо разделены в пространстве признаков и соответствуют объектам поверхности Земли. Различные фазы таяния снежного покрова относятся к различным кластерам при выборе достаточной детальности кластеризации [13]. На рис.2 представлен фрагмент космического снимка поверхности Земли. Изображение составляют 2120 различных трехспектральных векторов.

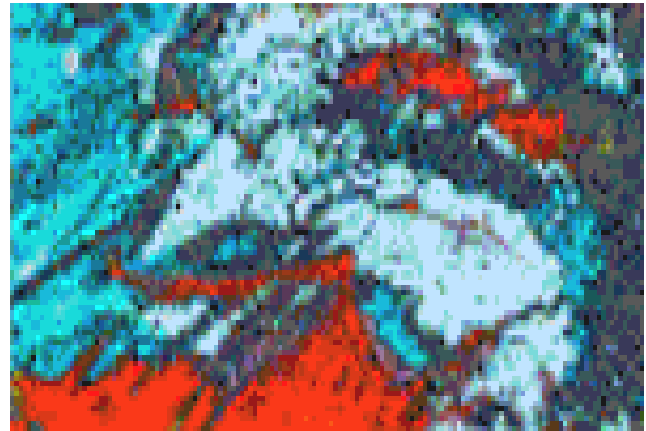


Рис.3. Кластерная карта изображения тающих снегов.

Проводилось сглаживание гистограммы. В результате работы алгоритма получено 25 унимодальных кластеров, причем только 12 имеют значительную площадь. Карта кластеров показана на рис.3.

Фрагмент трехспектрального космического изображения Западной Сибири 24 апреля показан на рис.4 слева. В левой части снег в основном растаял. На цветном изображении водные поверхности выделяются ярко-синими пятнами на красном фоне. В левом верхнем углу озеро Чаны, в нижнем левом озеро Кулундинское. Светлые оттенки на озерах - не растаявший лед. От середины немного вниз Обское водохранилище, оно еще под снегом. Справа вверху густое облако. Темные пятна в правой половине снимка – пробивающийся сквозь снег хвойный зеленеющий лес. Показаны результаты кластеризации с различными параметрами.

На рис.5 было срезано 2 бита и гистограмма сглажена два раза. Получено 15867 различных векторов и 158 кластеров. На рис.6 при срезе трех битов и сглаживании было получено 2962 вектора и 21 унимодальных кластера.

Даже при грубой детальности (отсечение трех битов в каждом спектральном канале) попадают в различные, хорошо разделенные кластеры следующие объекты: растаявшие участки суши (преобладание инфракрасного), снежный покров (отсутствие инфракрасного), растаявшие водные поверхности (исключительно синий диапазон), непрозрачные облака (белый цвет), хвойные насаждения (преобладание зеленого).

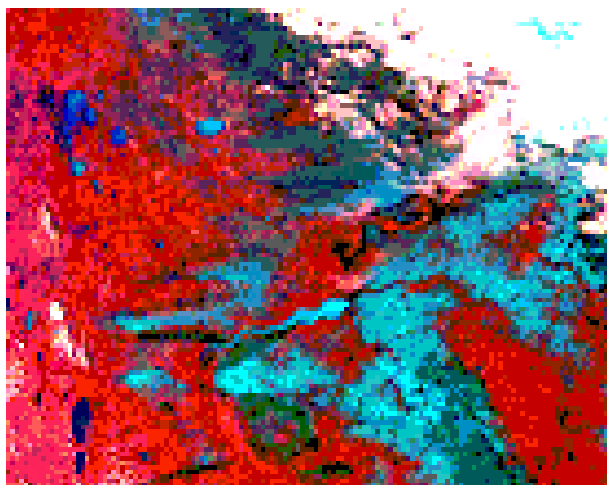


Рис.4. Трехспектральный космический снимок тающих снегов в Западной Сибири. 24 апреля.

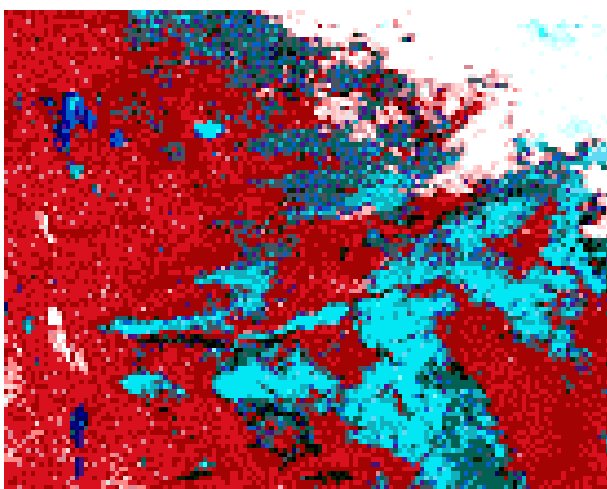


Рис.5. Кластерная карта изображения Западной Сибири. Здесь 158 кластеров.

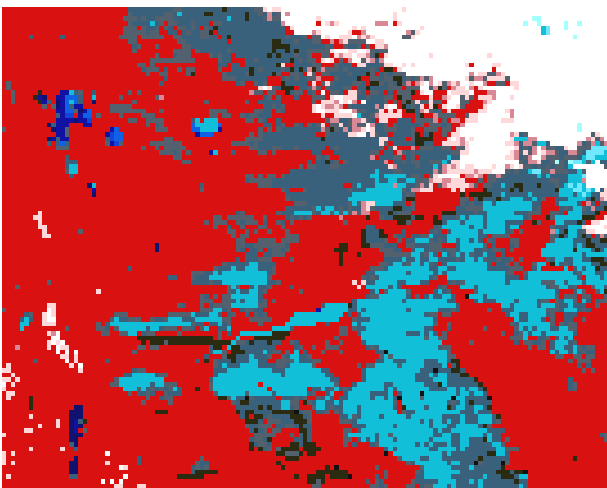


Рис.6. Кластерная карта Западной Сибири. Здесь 21 кластер.

Теперь проиллюстрируем результаты кластеризации пяти-спектрального изображения. На рис.7 представлено изображение трех первых каналов в цветосовмещенном формате

(RGB-файл). Здесь выбраны два видимых (синий и зеленый) и ближний инфракрасный диапазоны. А на рис.8 изображение в пятом инфракрасном канале на длине волны 12 мкм. Изображение в четвертом канале (11 мкм) визуальнo практически совпадает с пятым и здесь не приводится. Этот снимок также сделан над территорией Западной Сибири, но в более ранний период 17 апреля. Здесь большую часть занимают снега, сплошные и тающие. Только в левой части внизу оттаявшая поверхность Земли. В правом верхнем углу густое облако.

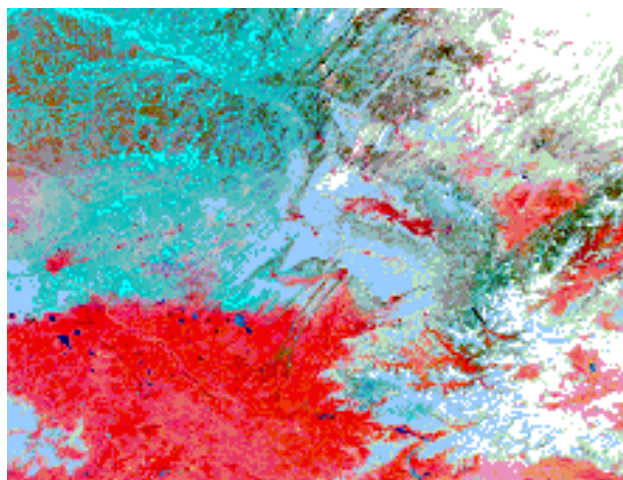


Рис.7. Изображение трех первых спектральных каналов в цветосовмещенном формате на космическом снимке Западной Сибири. 17 апреля.

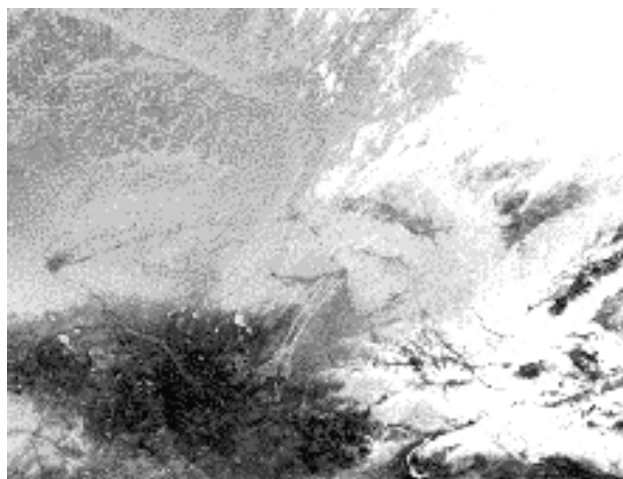


Рис.8. Пятый спектральный канал космического снимка Западной Сибири. 17 апреля.

Изображение имеет размер 1480*1124 и для пяти спектральных каналов его объем составляет около 8.3 Мегабайт. Кластеризация по пяти спектральным каналам дала следующие результаты. При срезе четырех бит получилось 16072 различных векторов и 44 унимодальных кластера (рис. 9). При срезе трех бит получилось 97575 векторов и 690 кластеров (рис. 10). Хотя на карту кластеров выводится 256 самых больших кластеров, остальных кластеров практически не видно на карте, т.к. они занимают чрезвычайно маленькую

площадь. На этих картах уже нет цветового соответствия с исходным изображением, т.к. раскраска карт является линейной комбинацией яркостей по пяти каналам.

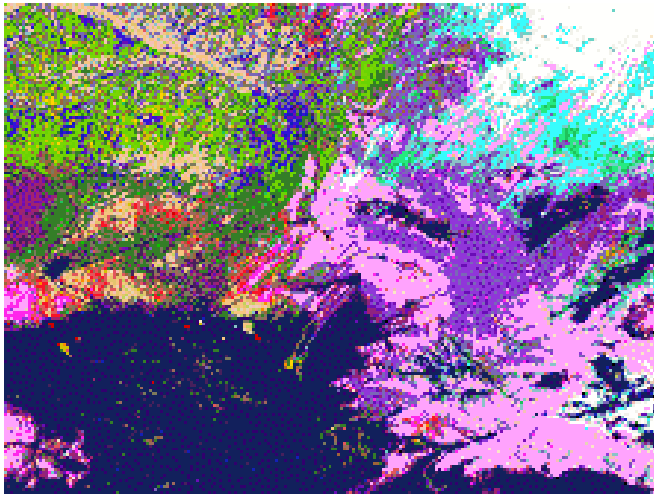


Рис.9. Кластерная карта пятиспектрального изображения Западной Сибири. Здесь 44 кластера.

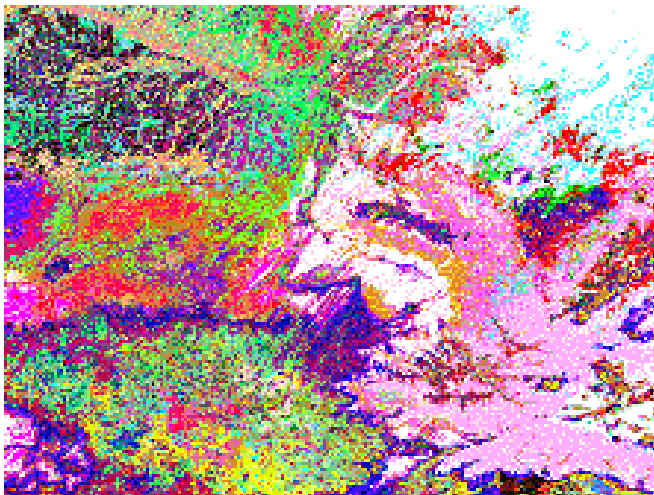


Рис.10. Кластерная карта пятиспектрального изображения Западной Сибири. Здесь 690 кластера.

Распределение яркостей для тающих снегов во всех каналах выглядит похоже, в то время как для оттаявшей суши существенно различается. Это, видимо, можно объяснить различной температурой участков и различным распределением излучения для диапазонов длин волн принимающих каналов. Была проведена кластеризация по фрагменту этого изображения, содержащего участок суши. На рис.11 слева изображение первых трех, справа 4 и 5 каналов, снизу полученная карта. На снимке в нижней части растаявшая поверхность Земли. По диагонали из левого верхнего угла просматривается русло Иртыша, в правой части изображения Ленточные боры.

Оконтурен третий кластер, соответствующий темному пятну внизу изображения для пятого и четвертого каналов, на RGB-изображении первых трех каналов не видно соответствующе-

го объекта. Не выделился подобный кластер и при кластеризации только по трем первым каналам. Таким образом, дополнительные два канала несут новую информацию, и это отразилось на карте кластеров. Из трех инфракрасных каналов, вклад в спектральную характеристику оконтуренного объекта вносит наиболее коротковолновый, соответствующий температуре. Т.е. объект, видимо, хорошо прогрет. Это участок поверхности Земли южнее Экибастуза.

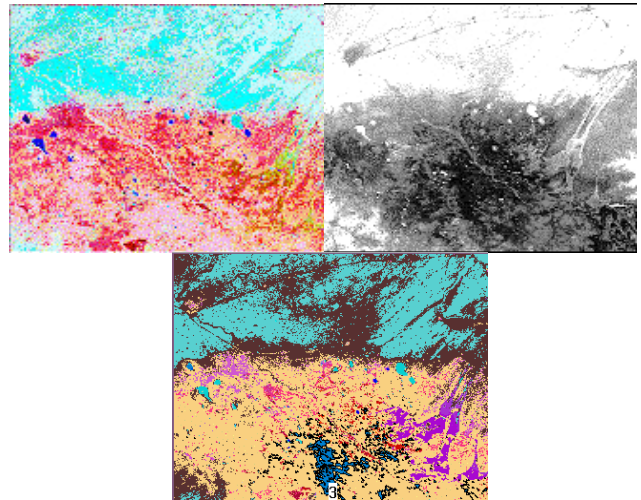


Рис.11. Вверху фрагмент пятиспектрального изображения.. Внизу кластерная карта.

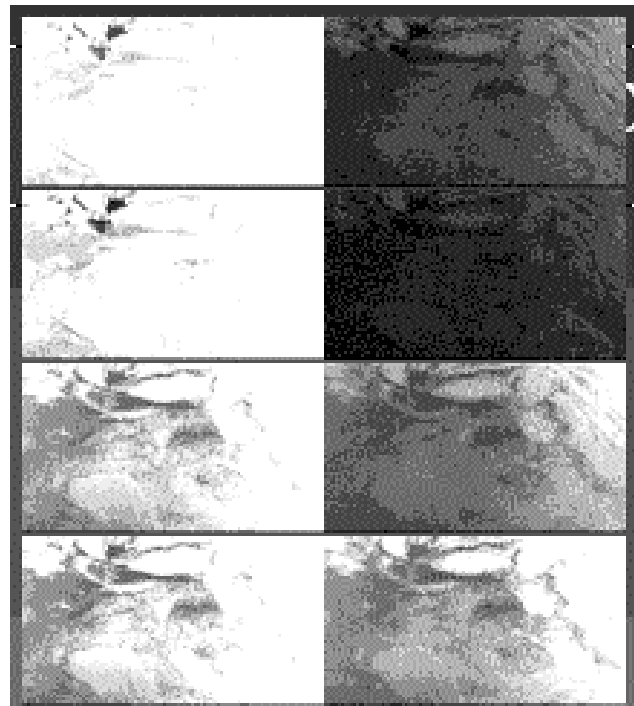


Рис.12. Изображения восьми каналов восьмиканального raw-файла.

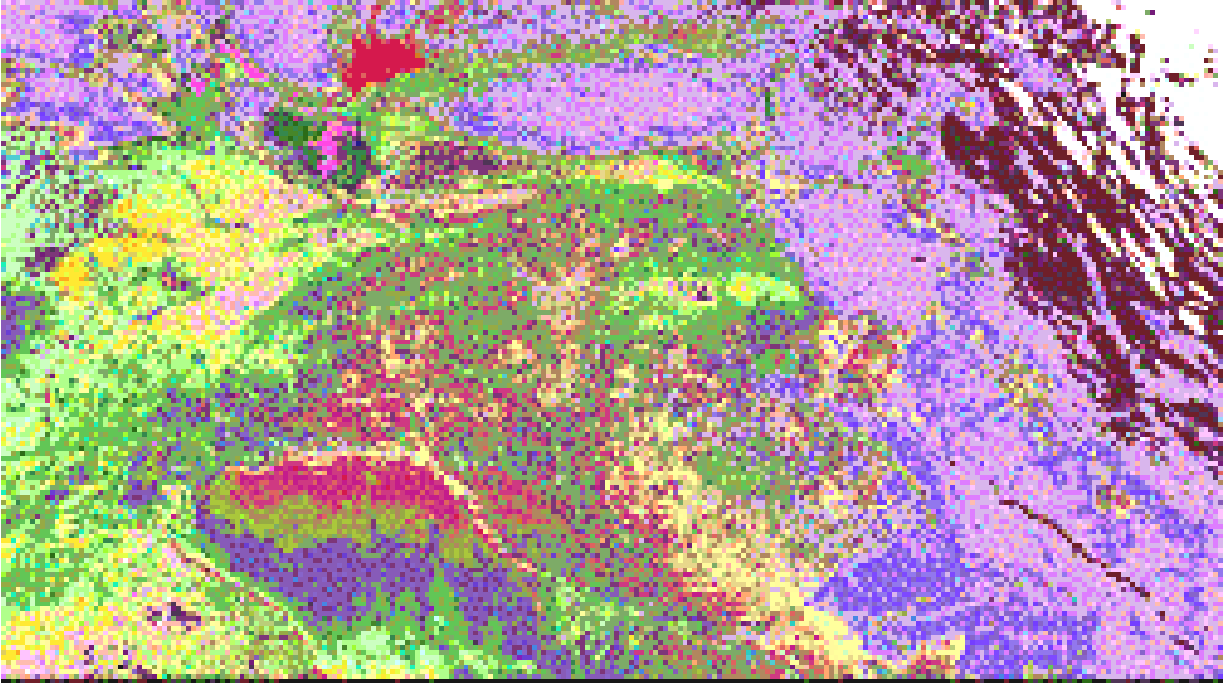


Рис.13. Кластерная карта восьмиканального изображения поверхности Земли. Получено 72 кластера.

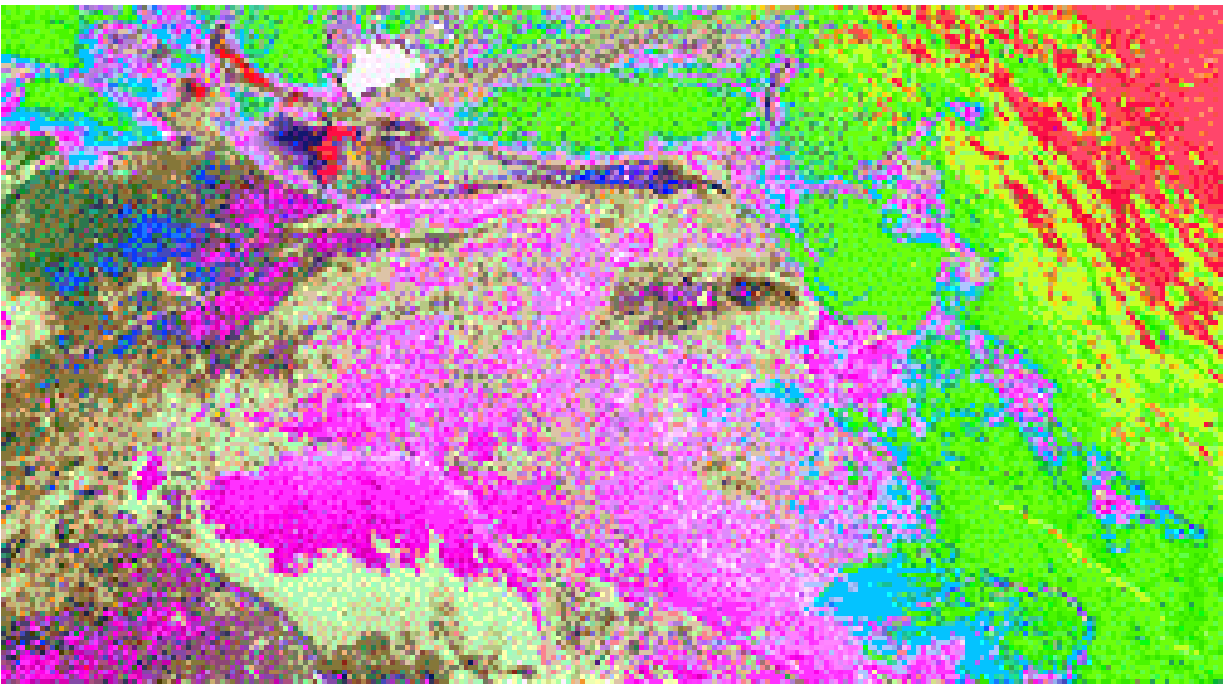


Рис.14. Кластерная карта восьмиканального изображения поверхности Земли. Получено 664 кластера.

Наконец, представим кластеризацию восьмиспектрального космического снимка поверхности Земли. На рис.12 изображения всех восьми каналов. 8-канальное изображение размером 2502*1410 пикселей записано в виде raw-файла и имеет объем около 28.2 мегабайт. Представлены кластерные карты, полученные для различных параметров детальности кластеризации. На рис.13 при срезе 3 бит, получено 9753 различных вектора и 72 кластера. На рис.14 при срезе 2 бит - 95659 векторов и 664 кластера. Это восьмиспектральное изображение обрабатывалось около часа на процессоре Celeron 2 GHz.. Оно было получено с другого спутника TERRA, поэтому соотношение между параметрами детальности кластеризации и полученным числом кластеров отличается от результатов для снимков со спутника NOAA.

6. ЗАКЛЮЧЕНИЕ

Рассматриваемый алгоритм разделяет пространство признаков многоспектральных изображений по унимодальным кластерам. Детальностью кластеризации можно управлять, убирая младшие разряды векторов яркости или сглаживая гистограмму. Поведение плотности вероятности в полученных кластерах может быть в дальнейшем исследовано и применены различные способы оценки разделенности кластеров, используя полученные таблицы характеристик и карты кластеров. Для контролируемой классификации унимодальные кластеры могут хорошо характеризовать выборки. Проиллюстрирована возможность применения алгоритма к анализу трех, пяти и восьмиспектральных изображений поверхности Земли.

7. ЛИТЕРАТУРА

- [1] Duda R.O. Hart P.E. Pattern, *Classification and Scene Analysis* (John Wiley and Sons, New York, 1973).
- [2] G.H.Ball & J.D.Hall, A clustering technique for summarising multivariate data, *Behav, Sci.12*, 1967, 153-155.
- [3] Philip H.Swain & Shirley M.Davis, *Remote Sensing.: the Quantitative Approach* (Edited by USA). McGraw-Hill, 1978).
- [4] Jack Bryant, On the Clustering of multidimensional pictorial Data, *Pattern Recognition, vol 11*, 1979,115-125.
- [5] V.I.Borisenko & Chesalin N.K., Algorithms of the Theme Decoding of the Multispectral Space Video Information. *Space research, Vol.XVI, Issue 3*, 1978, 3888-3893.
- [6] V.S.Sidorova, Age Recognition for the Forest Communities on Aeroimage Texture Features. *Proceedings of Pattern recognition for remote Sensing Workshop*, Andorra, 2000, 72-75.
- [7] Gong P. & P.J.Howarth, Frequency-based contextual classification and grey-level vector reduction for land-use identification, *Photogrametric Engineering and Remote Sensing, 58(4)*, 1992a, 421-437.
- [8] A.S.Alekseev & V.P.Pjatkin & V.S.Sidorova, *The Complex of Programs of Image Analysis (the Cluster Analysis and Segmenting)* (Preprint N 579) Novosibirsk, 1985.
- [9] В.А. Красиков, В.А.Шамис, Кластерная процедура на бае многомерной гистограммы распределения. Исследование Земли из космоса. N 2, сс.107-114.
- [10] W.Koontz, P.M.Narendra and K.Fukunaga, A graph theoretic approach to non-parametric cluster analysis, *IEEE Trans. Comput.C-23*, 936-944 (1967).

[11] А.С.Алексеев, В.П.Пяткин, В.С.Сидорова, Комплекс программ анализа изображений (кластерный анализ и сегментация). Препринт N 579. Новосибирск. 1985.

[12] В.С.Сидорова, Кластеризация многоспектральных изображений с помощью анализа многомерной гистограммы. Математические и технические проблемы обработки изображений. Вычислительный Центр. Сибирское отделение Академии Наук СССР. Новосибирск. 1986. сс.52-57.

[13] V.S.Sidorova, The Categorization of Multispectral Space Images of Earth Surface by Means of Separating Multimodal Histogram to Unimodal Clusters. *Proceedings of the International conference "Water Problems at the Beginning Initially Millenniums: hydrological, meteorological and ecological aspects"* Kazakhstan, 2004, 51-55.

Об авторе

Валерия Сидорова – научный сотрудник Института Вычислительной Математики и Математической Геофизики СО РАН, лаборатории Обработки изображений.

Адрес: Новосибирск, 630090, пр-т Лаврентьева, 6, ИВМиМГ. Телефон (3832)307-332. E-mail svs@ooi.sccc.ru.

Multidimensional Histogram and Separating of the Vector Feature Space on the Unimodal Clusters

Abstract

The fast nonparametric technique of uncontrollable classification of the multi spectral data, based on construction of the multidimensional histogram and it's separating on the unimodal clusters, is considered. The technique does not use any suppositions about functions of distributions of various classes. It also does not require any priori data (number of clusters, number of iterations etc.). The histogram is formed as the list for vectors which really present on the image. For a quick access to the list of vectors both the system of hashing and sorting of vectors on increasing are used. The essence of the technique is to find the maxima of the multidimensional histogram in a discrete vector space. We are deal with vectors that are ordered on increase. It is possible to manage the results of clustering using cutting bits of the vectors and smoothing the histogram. By means of the technique space images of melting springtime of snows on Earth surface were researched. There are five and eight spectral channels were considered.

Keywords: Image processing, clustering, multidimensional histogram, remote sensing.

About the author

Valeria Sidorova is a scientific researcher at Institute of Computational Mathematics and Mathematical Geophysics, Department of Image Processing. Her contact email is svs@ooi.sccc.ru.