Кластеризация почвенно-растительных объектов с помощью нейросетевого алгоритма Кохонена

Б.М. Балтер[1], Д.Б. Балтер[1], В.В. Егоров[1], А.А. Ильин[2], А.П. Калинин[3], А.Г. Орлов[2], И.Д. Родионов[4]

[1] Институт космических исследований РАН
117997, Москва, ул. Профсоюзная, 84/32
E-mails: balter@mail.ru, db.balter@gmale.com, victor_egorov@mail.ru
[2]Научно-технический центр "Реагент"
119991, Москва, ул. Косыгина, 4
E-mails: ilyinandrey@mail.ru, aorlov@reagent-rdc.ru
[3] Институт проблем механики РАН
119526, Москва, просп. Вернадского, 101/1
E-mail: kalinin@ipmnet.ru
[4] Институт химической физики им. Н.Н. Семенова РАН
119991, Москва, ул. Косыгина, 4
E-mail: irodionov@reagent-rdc.ru

В настоящей работе рассматривается задача кластеризации (классификации без обучения) почвенно-растительных объектов по данным гиперспектрального зондирования с борта авианосителя. Для этой цели используется отображение Кохонена из многомерного пространства спектральных данных на двумерную поверхность с последующей кластеризацией данных на этой поверхности методом ISODATA. Для кластеризации используются не весь объем данных, а данные с некоторых участков. Этими участками могут являться либо случайно выбранные точки зондируемой поверхности, либо интересующие нас объекты. На основании результатов кластеризации рассматривается проблема корректного выбора обучающих участков для применения методов классификации с обучением.

Введение

В настоящее время при дистанционном зондировании (ДЗ) Земли наряду с многоспектральными сканерами все более широкое применение находят бортовые гиперспектрометры [1-3], (http://www.satimagingcorp.com/gallery-quickbird.html; http://geo.arc.nasa.gov/sge/landsat/l7.html; http://eo1.gsfc.nasa.gov/Technology/Hyperion.html). Основной целью применения упомянутых сенсоров является классификация объектов зондирования и оценка их состояния [4] - предмет тематической обработки данных ДЗ. Однако методы тематической обработки гиперспектральных данных и оценки качества распознавания типов зондируемых объектов, несмотря на ряд работ [1, 2, 5], развиты недостаточно. Кроме того, не проводилось сравнение достоверности классификации объектов зондирования по данным гиперспектральной и многоспектральной съемки.

Поскольку растительный покров занимает значительную часть суши, задачи классификации типов растительности представляются весьма актуальными. Поэтому отработка методов тематической обработки и оценки достоверности классификации объектов по данным гиперспектральной съемки в настоящей работе решаются на примере исследования растительных сообществ европейской части России. Для этого используются результаты эксперимента по гиперспектральному зондированию и видеосъемки отдельных участков территории Пензенской области, занятых почвенно-растительным покровом.

В настоящей работе использовались данные вертолетной гиперспектральной съемки. Объекты были сгруппированы в 3 категории: целевой растительный объект, фоновые растительные объекты, участки без растительности (условно – почва).

В ходе работы рассматривалась задача оценки качества классификации без обучения с помощью нейросетевого алгоритма Кохонена. Этот метод можно использовать как самостоятельно, так и в качестве средства выбора обучающих участков для применения методов *классификации с обучением*, в частности, метода максимального правдоподобия [6, 7].

Процедура кластеризации

При описании способов кластеризации обычно используют понятие *точки* в многомерном пространстве R. В данном случае этой точкой x будет спектр, измеренный гиперспектрометром. Координаты точки x – это значения интенсивности отраженного сигнала на разных длинах волн. Размеренность пространства R - это число каналов гиперспектрометра.

Исходными данными для кластеризации является набор всех измерений гиперспектрометра – множество точек x_k. Эти измерения соответствуют разным классам зондируемых объектов Ω_i, $i = 1,\ldots,3$ (в нашем случае – почв и дорог, расположенных на зондируемой территории, и типов растительности). Принципиальным отличием кластеризации от процедуры классификации с обучением [6, 7] состоит в том, что в данном случае не используется информация о том, какое именно измерение из обучающих данных x_k какому классу Ω_i соответствует. Результатом кластеризации является разбиение всего множества точек x_k на подмножества, называемые кластерами. Разбиение производится таким образом, чтобы расстояния между точками одного кластера были бы меньше, чем расстояния между точками разных кластеров. Эту процедуру часто называют *классификацией без обучения*.

В качестве меры близости точек можно использовать евклидово расстояние или спектральный угол. В работе [7] описано, при каких предположениях о свойствах получаемых данных целесообразно использование такой метрики. Заметим, что в зависимости от математической модели измерительной системы (гиперспектрометра) возможно использование и других метрик.

Каждый кластер можно отнести к отдельному соответствующему классу Ω_i. Кластеры можно сгруппировать и сопоставить одному классу некоторый набор кластеров. Отнесение кластеров различным классам может проводиться компьютером полностью автономно на основании заранее заданных критериев или с некоторыми «подсказками» оператора в зависимости от реализации алгоритма. Количество классов может также определяться автоматически или задаваться оператором. Поскольку при использовании описываемого подхода отсутствует априорная информация о типах растительности, мы не можем сказать, какой тип растительности отнесен к некоторому классу Ω_i. В этом случае можно лишь сделать заключение, что фрагменты растительности, отнесенные к одному классу, в некотором смысле похожи между собой. При составлении карты зондируемой поверхности описываемым методом на практике фрагменты поверхности, отнесенные к одному классу, окрашивают одним цветом, выбор которого может быть случаен. Если при использовании метода кластеризации привлекается дополнительная информация о расположении тех или иных объектов на зондируемой поверхности, то кластеры могут быть осмысленно объединены в классы. В этом случае каждый класс уже будет содержать заданные объекты. Этот подход вплотную приближается к методам классификации с обучением.

После того как построены кластеры, можно произвести так называемую классификацию данных измерений без обучения. Для классификации некоторого измерения x_k достаточно отыскать ближайший к этой точке кластер и отнести это измерение к соответствующему классу.

Кластеризация с помощью нейронной сети Кохонена

Существует множество различных алгоритмов кластеризации. Одним из эффективных способов кластеризации точек в пространстве большой размерности является метод Кохонена на осно-

ве нейросети [8]. Его также называют отображением Кохонена или методом самообучающихся карт (Self Organizing Maps - SOM), который, вместе с тем, можно рассматривать и в качестве одного из способов проецирования входного многомерного пространства R в пространство с более низкой размерностью.

Искусственный нейрон (в частности, и сети Кохонена) имитирует в некотором приближении свойства биологического нейрона. На вход нейрона поступает некоторое множество сигналов, реагируя на которые нейрон формирует выходной сигнал или меняет свое состояние. Одним из отличительных свойств нейрона является его «автономность», то есть его поведение зависит только от входных сигналов и не зависит от поведения других нейронов. Поведение нейронной сети определяется совокупным поведением множества независимых нейронов. Далее опишем собственно нейронную сеть Кохонена.

Несмотря на то, что процедуру кластеризации часто называют «классификацией без обучения», при описании нейронной сети нельзя не использовать термин «обучение нейронной сети», - настолько этот термин стал общепринятым. «Обучающими данными» здесь является полный набор измерений x_k, без указания, какому классу каждое из них соответствует.

Как уже было упомянуто, нейронная сеть Кохонена проецирует исходное многомерное пространство в пространство более низкой размерности. Область пространства более низкой размерности (обычно одномерного или двумерного) равномерно заполняется нейронами и называется *слоем нейронов*. Каждый такой нейрон имеет заданные координаты в пространстве более низкой размерности и называется *нейроном из слоя нейронов* или *нейроном скрытого слоя*. Кроме того, каждому нейрону соответствует n-мерный вектор синаптических связей $w_k = (w_{k1}, w_{k2}, \ldots, w_{kn})$, где n - размерность исходного пространства R, а k - номер нейрона. Слово *синаптических* взято из биологии и используется в математике для обозначения связи нейронов. С помощью синаптической связи каждый нейрон слоя нейронов связан с *входными нейронами*, о которых будет рассказано далее. Количество нейронов слоя нейронов произвольно, а количество входных нейронов должно быть равно n - размерности исходного пространства R.

Некоторую точку x исходного пространства можно сравнить с точкой w_k, задаваемой вектором синаптической связи. Для этого вычисляется евклидово расстояние или спектральный угол (в зависимости от реализации алгоритма) между этими двумя точками. Сравнивая вычисленные расстояния для разных нейронов, определяют наиболее близкий нейрон, который называют «*нейрон-победитель*». Эта процедура осуществляется и при обучении нейронной сети, и при её использовании. В качестве точки x может быть выбрана любая точка исходного пространства R, но, как правило, это бывает некоторая точка обучающей выборки. Если задается некоторая точка x, то говорят, что на вход нейронной сети подается точка x. Входы, на которые подаются координаты точки исходного пространства, называют *входными нейронами*, и они связаны посредством синаптических связей со всеми нейронами слоя нейронов. В качестве примера на рис.1 изображена структура нейронной сети Кохонена, если пространство слоя нейронов является плоскостью. На этом рисунке жирными точками изображены входные нейроны. В виде шариков изображены нейроны из слоя нейронов, а линиями - синаптические связи между нейронами.

При помощи вектора синаптических связей w_k устанавливается соответствие между точкой исходного пространства (концом вектора w_k) и точкой слоя нейронов (пространства меньшей размерности), которая определяется координатами этого нейрона в слое. Таким образом, каждый нейрон описывает отображение между двумя точками разных пространств. Множество нейронов задает большое количество таких отображений, которые можно рассматривать как частные случаи преобразования исходного пространства в пространство слоя нейронов. Если необходимо спроецировать произвольную точку x в пространство слоя нейронов, то для простоты обычно поступают следующим образом: для точки x отыскивается нейрон-победитель, и его координаты

в пространстве слоя нейронов определяют проекцию точки x. Такой алгоритм используется для решения задач кластеризации.

Процесс обучения нейронной сети состоит в настраивании синаптических связей w_k. При этом ставится задача отобразить обучающее множество точек в пространство слоя нейронов с максимальным сохранением отношения соседства между точками. Процесс обучения нейронной сети Кохонена подробно описан, например, в работе [9].

После того, как построено отображение точек исходного пространства в пространство слоя нейронов, производится кластеризация нейронов. Для решения этой задачи можно использовать стандартные методы кластеризации. Таким стандартным методом является Isodata (Iterative Self-Organizing Data Analysis Techniques – итеративный самообучающийся метод анализа данных), описанный, например, в работе [10]. Эмпирически установлено, что кластеризация нейронов сети выгоднее, чем кластеризация исходных данных вышеупомянутым методом.

Далее, представим результаты обработки реальных измерений гиперспектрометра с использованием кластеризации методом нейронной сети Кохонена при наличии дополнительной информации об объектах на зондируемой поверхности (данных о разных видах растений, рассаженных на Мире).

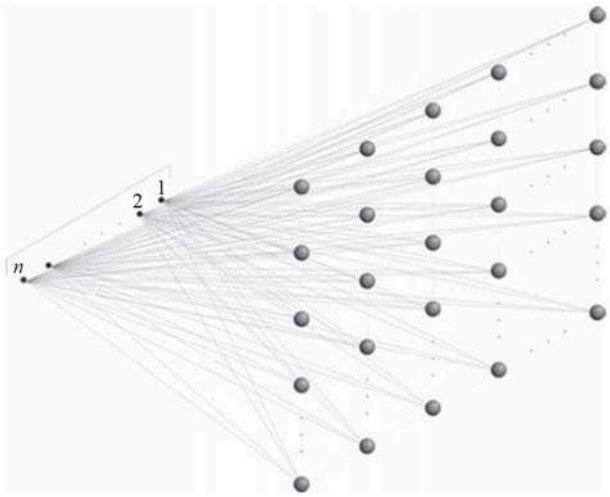


Рис. 1. Структура нейронной сети Кохонена

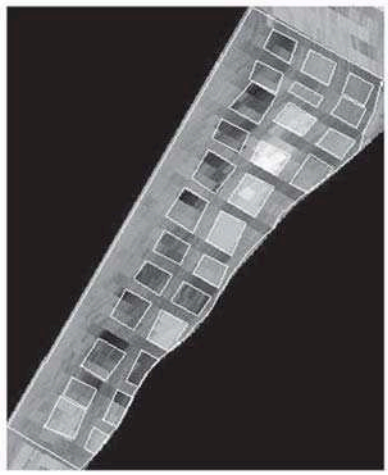


Рис. 2. Построение оператором обучающих участков на гиперспектральных данных

Исходные данные гиперспектральной съемки

При гиперспектральном мониторинге с борта вертолета основная рабочая часть трассы проходила через специально созданный участок с квадратными грядками из разных типов растительности (сокращенно – «Мира»). К сожалению, из-за трудностей наведения трасса не покрывает Миру полностью и проходит через нее наискось. Основная часть анализа данных гиперспектральной съемки проводилась в пределах Миры и ее ближайшей окрестности. Кроме того, классификация применялась ко всей трассе для визуальной оценки качества классификации на удаленных от Миры участках.

На каждом из «квадратов», составляющих обучающие участки (см. рис. 2) – свой вид растительности. Детальный видовой анализ растительности – это предмет дальнейших работ. Здесь мы сгруппировали все обучающие участки в 3 категории или 3 «макрообъекта»: целевой растительный объект, или «цель»; растительный фон, или просто «растительность»; участки без растительности, или «почва». На последних, конечно, есть какое-то количество растительности, но значительно меньше, чем на первых двух.

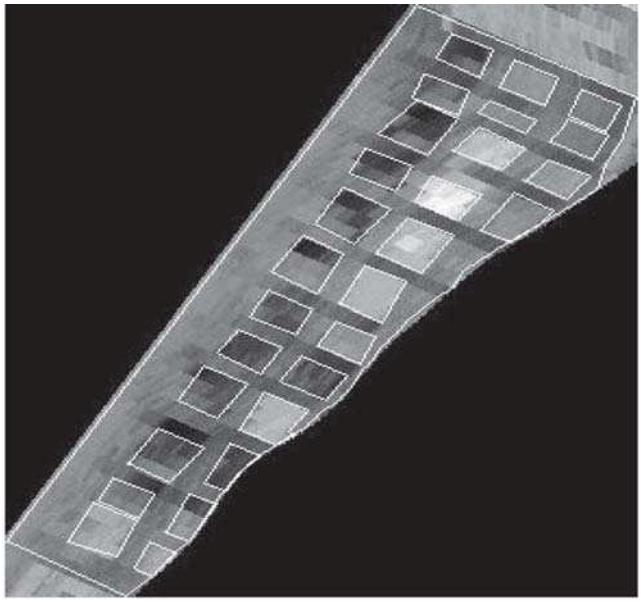


Рис. 2. Построение обучающих участков на гиперспектральных данных оператором

Кластерный анализ гиперспектральных измерений с использованием нейросетевого алгоритма Кохонена

Производилась кластеризация гиперспектральных измерений методом отображения Кохонена [11] из многомерного пространства спектральных данных на двумерную поверхность с последующей кластеризацией данных на этой поверхности методом ISODATA. При этом для сокращения времени счета использовались либо случайно выбранные точки вдоль трассы, либо только данные об интересующих нас объектах, т.е. обучающие участки Миры. В последнем случае мы игнорировали информацию о том, какой объект на каком участке, за исключением того случая, когда она используется для правильного приписывания объектов полученным кластерам.

На рис. 3 представлен пример отображения Кохонена. Двумерная плоскость разделена на 20 х 20 клеток, которые сгруппированы в кластеры, а те, в свою очередь, - в группы, соответствующие

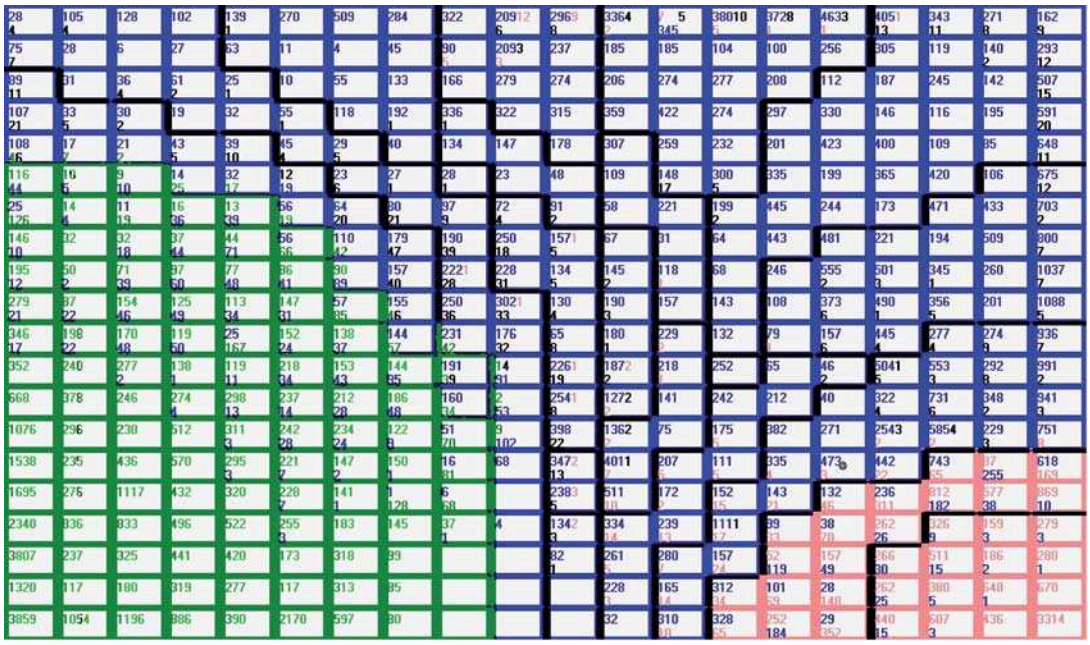
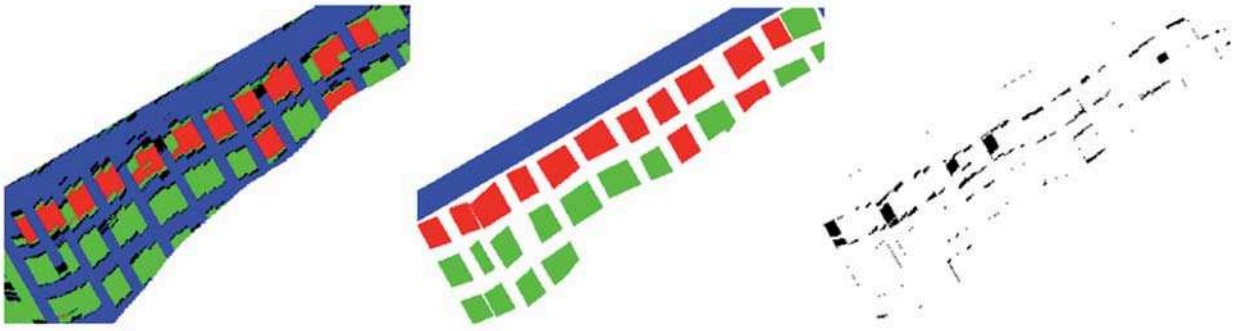


Рис. 3. Пример отображения Кохонена, построенного по обучающим участкам, с указанием числа точек каждого класса, попадающих в каждую клетку карты

3 классам объектов: цели, растительности и почве. Каждая клетка на рис. 3 – это один нейрон скрытого слоя. Объекты показаны на рисунке цветами: красный – цель, зеленый и черный – растительность, синий – почва. Показанная на рисунке группировка кластеров в классы произведена автоматически. Автоматическая группировка в классы производилась так: кластеру приписывается тот класс, у которого наибольшая доля точек попала в данный кластер. Возможна и ручная группировка кластеров в классы.

На рис. 4 показан результат классификации Миры методом Кохонена. Видно, что кластеризация дает более сложные формы участков, чем ручное выделение. Вопрос в том, насколько эти детали формы истинны. Поскольку грядки делались прямоугольными, можно было бы считать все извилистые формы ошибками, если бы не возможность того, что они связаны с реальными сложностями геометрии гиперспектральной съемки с борта вертолета.

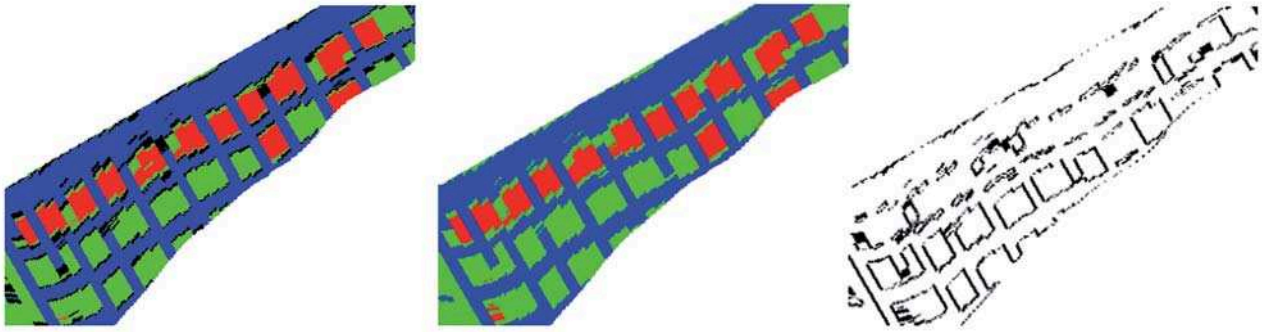


(а) *(б)* *(в)*

Рис. 4. Результат классификации Миры методом Кохонена: (а) – участки, полученные по результатам классификации на основе данных с «ядер» Миры (середин обучающих областей); (б) – участки, выделенные вручную; (в) – отличие приведенных обучающих участков (белый цвет – классы совпадают, черный – классы различны)

Влияние параметров кластеризации

Важно оценить устойчивость кластеризации к отбору данных, по которым строится отображение Кохонена. На рис. 5 показано два варианта классификации методом Кохонена: «ядра» (центральные области) обучающих участков Миры и случайно выбранные точки Миры. Представляется, что результаты достаточно устойчивы к отбору данных.



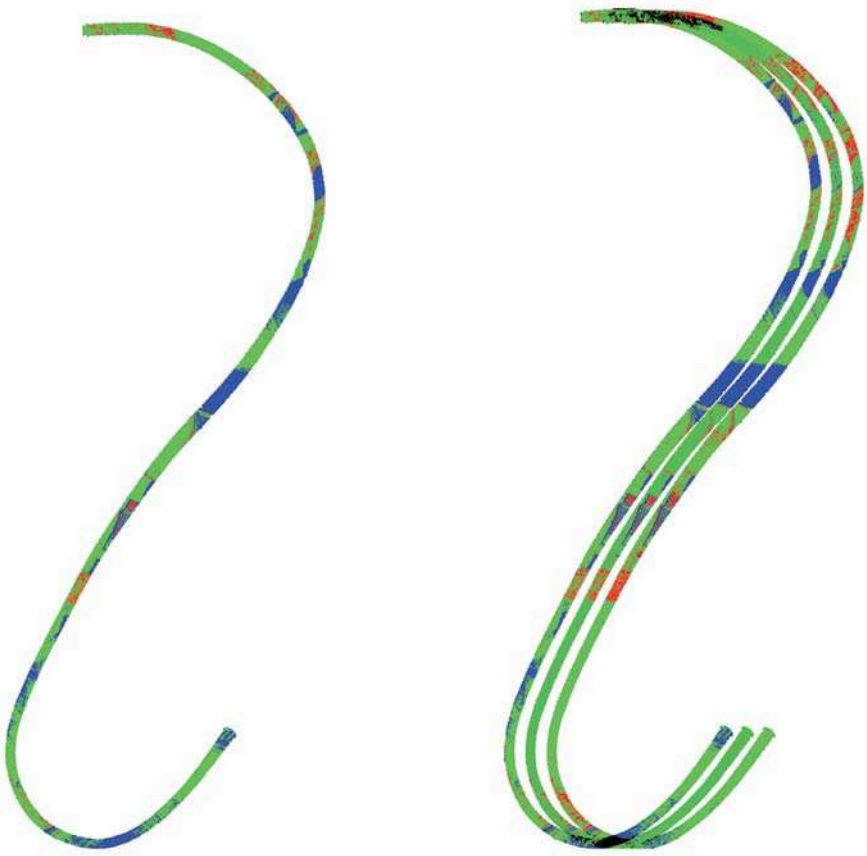
(а) *(б)* *(в)*

Рис. 5. Результаты классификации методом Кохонена: (а) - на основании данных с «ядер» (центральных областей обучающих участков), (б) - на основании случайно выбранного 1% точек Миры, (в) – отличие двух приведенных результатов (белый цвет – классы совпадают, черный – классы различны)

На рис.6 показан результат кластеризации для всей трассы в нескольких вариантах, различающихся исходными данными для построения отображения Кохонена. Видно, что результаты

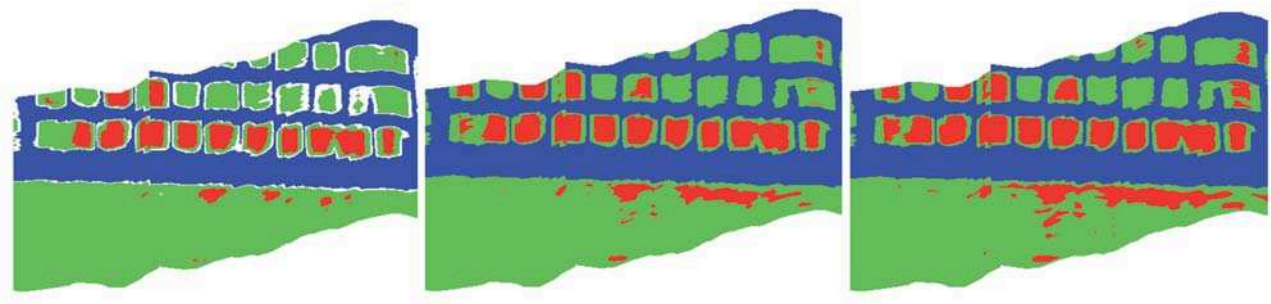
различаются в основном по количеству ложных тревог цели.

Естественно возникает мысль повторить кластеризацию, взяв в качестве исходных данных набор объектов, выделенный на первом шаге кластеризации. Продолжая этот процесс, можно надеяться, что он сойдется к «устойчивым» обучающим участкам, которые будут лучше, чем исходные. При этом возможны два варианта изменений от шага к шагу: изменение набора точек, взятых в качестве исходных данных, и изменение атрибуции этих точек категориям объектов (последнее имеет смысл, только если на этой основе производится атрибуция построенных кластеров объектам, как описано выше). Результат, полученный вторым способом, показан на рис. 7.



(а) *(б)*

Рис. 6. Результаты классификации методом Кохонена для всей трассы. Присвоение цветов 10 кластерам производилось оператором: (а) - Исходные данные с обучающих участков Миры; (б) - слева направо: исходные данные – с обучающих участков, случайная выборка 1:100 из Миры, случайная выборка 1:100 из всей трассы



(а) *(б)* *(в)*

Рис. 7. Последовательное уточнение обучающих участков повторением кластеризации по Кохонену: (а) – первая итерация, (б) – вторая итерация, (а) – третья итерация

Выводы

Применение кластеризации по Кохонену в чистом виде как средства классификации требует ручной агрегации кластеров в классы. И даже при этом условии ошибки больше, чем в методах с обучением. Поэтому мы не рассматриваем кластеризацию как самостоятельный метод. Ее можно использовать как средство для предварительной оценки территории.

Все методы классификации с обучением чувствительны к попаданию в обучающий участок «не своих» объектов. Это больше всех других факторов снижает точность классификации. Наилучшие перспективы у кластеризации – как у средства построения обучающих участков для методов классификации с обучением, в частности, метода максимума правдоподобия. При этом проявляется высокая устойчивость результатов кластеризации к выбору исходных данных. Точность метода максимума правдоподобия на таких участках по некоторым объектам заметно выше, чем на выбранных вручную участках.

Кластеризацию можно применять итеративно для улучшения обучающих участков. В этом варианте можно ожидать повышения точности сравнительно с ручным выбором обучающих участков, хотя бы потому, что участки, выбранные вручную, могут служить начальной точкой для итераций. Здесь требуются исследования сходимости этого процесса.

Литература

1. Б*алтер Б.М., Егоров В.В., Ильин А.А., Калинин А.П., Орлов А.Г., Останний А.Н., Родионова И.П., Родионов И.Д.* Оценка возможностей гиперспектральной съемки для дистанционного обнаружения заданного типа растительности // Препринт ИКИ РАН Пр-2134. 2007. 30 с.

2. *Непобедимый С.П., Родионов И.Д., Воронцов Д.В., Орлов А.Г., Калашников С.К., Калинин А.П., Овчинников М.Ю., Родионов А.И., Шилов И.Б., Любимов В.Н., Осипов А.Ф.* Гиперспектральное дистанционное зондирование Земли // Доклады Академии наук, 2004, Т 397, №1, С.45-48.

3. *Балтер Д.М., Белов А.А., Воронцов Д.В., Ведешин Л.А., Егоров В.В., Калинин А.П., Орлов А.Г., Родионов А.И., Родионова И.П., Федунин Е.Ю.* Проект спутникового гиперспектрометра, предназначенного для малого космического аппарата // Исслед. Земли из космоса, 2007, №2. С.43-55.

4. *Балтер Б.М., Егоров В.В.* Статистическая оценка состояния природных объектов по данным дистанционных измерений // Исслед. Земли из космоса, 1981, № 3, С.46-55.

5. *Воронцов Д.В., Егоров В.В., Калинин А.П., Орлов А.Г., Родионов И.Д., Родионова И.П.* Принципы обработки гиперспектральной информации и результаты летных испытаний прототипа авиационного гиперспектрометра // Вестник МГТУ им. Н.Э. Баумана, Сер. «Приборостороение», 2006, №4. С.27-37.

6. *Richards J.A.* Remote Sensing Digital Image Analysis. An Introduction // Springer-Verlag. Berlin, Heidenberg. 1993. 340 p.

7. *Балтер Б.М., Егоров В.В., Ильин А.А., Калинин А.П., Орлов А.Г., Останний А.Н., Родионова И.П., Родионов И.Д.* Оценка возможностей гиперспектральной съемки для дистанционного обнаружения заданного типа растительности // М.: Препринт ИКИ РАН, 2007, № 2134, 28 с.

8. *Kohonen T.* Self–organization and associative memory. Series in Information Sciences. V.8. // Berlin: Springer verlag, 1984.

9. *Круглов В.В., Дли М.И., Голунов П.Ю.* Нечеткая логика и искусственные нейронные сети. М.: Физматлит, 2001. 224 с.

10. *Ту Дж., Гонзалес Р.* Принципы распознавания образов. М.: Мир, 1978. 413 с.

11. *Балтер Б.М., Воронцов Д.В., Егоров В.В., Ильин А.А., Калинин А.П., Орлов А.Г., Останний А.Н., Родионов А.И., Родионов И.Д.* Распознавание типов растительности по данным авиационного гиперспектрометра и многоспектрального космического сканера Quickbird // М.: Препринт ИПМех РАН, 2007, № 834, 36 с.