|  |  |
| --- | --- |
| *voenmeh* | МИНОБРНАУКИ РОССИИ  федеральное государственное бюджетное образовательное учреждение  высшего образования  **«Балтийский государственный технический университет «ВОЕНМЕХ» им. Д.Ф. Устинова»**  **(БГТУ «ВОЕНМЕХ» им. Д.Ф. Устинова»)** |
| БГТУ.СМК-Ф-4.2-К5-01 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Факультет |  | И |  | Информационные и управляющие системы |
|  |  |  |  |  |
| Кафедра |  | И9 |  | Систем управления и компьютерных технологий |
|  |  |  |  |  |
| Направление |  | 09.04.04 Программная инженерия | | |

КУРСОВАЯ РАБОТА

на тему

|  |
| --- |
| Алгоритмы обработки данных дистанционного |
| зондирования Земли, решающие проблему |
| смешанных векторов |
|  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Выполнил студент группы | | | |  | | И9М33 |
| Харитонов А.С. | | | | | | |
| Фамилия И.О. | | | | | | |
| **РУКОВОДИТЕЛЬ** | | | | | | |
| Мишина О.А. | |  |  | | | |
| Фамилия И.О. Подпись | | | | | | |
| Оценка |  | | | |  | |
| «\_\_\_\_\_» |  | | | | 2018 г. | |

Содержание

[ВВЕДЕНИЕ 4](#_Toc497832512)

[1 ПРОБЛЕМА СМЕШАННЫХ ВЕКТОРОВ 6](#_Toc497832513)

[2 КЛАССИФИКАЦИЯ С ОБУЧЕНИЕМ 8](#_Toc497832514)

[3 НЕКОНТРОЛИРУЕМАЯ КЛАССИФИКАЦИЯ 11](#_Toc497832515)

[4 Метод K-средних 11](#_Toc497832516)

[5 Метод анализа мод многомерной гистограммы 13](#_Toc497832517)

[ЗАКЛЮЧЕНИЕ 16](#_Toc497832532)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 17](#_Toc497832533)

обозначения и сокращения

ДЗЗ – дистанционное зондирование Земли

ДДЗЗ – данные дистанционного зондирования Земли

Введение

Центральные вопросы тематической обработки (интерпретации) ДДЗЗ – вопросы повышения качества дешифрирования – непосредственно связаны с проблемой выбора адекватных алгоритмов распознавания. Возникающие при этом трудности обусловлены следующими причинами:

1. Структура реальных данных не соответствует модели данных, используемой в алгоритме. Например, невыполнение предположения о нормальном распределении векторов данных или невыполнение условия, что поле измерений является случайным. Опыт показывает, что такие ситуации возникают тогда, когда излучение от сканируемого объекта выходит за пределы динамического диапазона съемочной аппаратуры. В этих случаях приходится прибегать к приемам, повышающим дисперсию данных.
2. Непрезентативность обучающих последовательностей – недостаточное количество данных для восстановления параметров решающего правила.
3. Несоответствие обучающих данных и данных, предъявляемых на распознавание (загрязнение выборок смешанными векторами измерений, то есть векторами, которые образуются при попадании в элемент разрешения съемочной системы нескольких природных объектов).
4. Неточное соответствие обучающих данных, получаемых с помощью кластеризации, истинным тематическим классам.
5. Помехи аппаратуры, влияние атмосферных условий и т.п.

Таким образом, можно сказать, что современный опыт автоматизированного распознавания ДДЗЗ показывает: заранее практически невозможно установить, какой алгоритм будет лучше с точки зрения точности классификации. Поэтому в распознающую систему целесообразно закладывать несколько алгоритмов и выбор оптимального алгоритма проводить эмпирически [1].

Целью данной научно-исследовательской работы является изучение и анализ вариантов классификации ДДЗЗ, с помощью которых решается проблема определения коэффициента правдоподобности классов объектов и возможность их соотнесения друг с другом.

Отчет состоит из введения, четырех глав и заключения.

В первой главе описывается одна из главных проблем обработки ДДЗЗ – проблема смешанных векторов.

Во второй главе приводятся методы обработки ДДЗЗ при классификации с обучением.

В третьей главе приводятся методы обработки ДДЗЗ при автоматической классификации (без обучения).

1 ПРОБЛЕМА СМЕШАННЫХ ВЕКТОРОВ

Наличие смешанных векторов в ДДЗЗ является одним из наиболее серьезных источников ошибок при построении тематической карты классификации. Понятие тематической карты предполагает, что каждый участок земной поверхности можно пометить как принадлежащий к одному и только одному классу поверхности. В действительности же дистанционное зондирование реализует функцию получения векторов измерений с 2KN возможных значений для K уровней дискретизации (рисунок 1) и N спектральных полос. Когда это пространство измерений сжимается с помощью классификации до нескольких классов, мы пренебрегаем большим количеством информации для того, чтобы получить упрощенную тематическую карту [2].

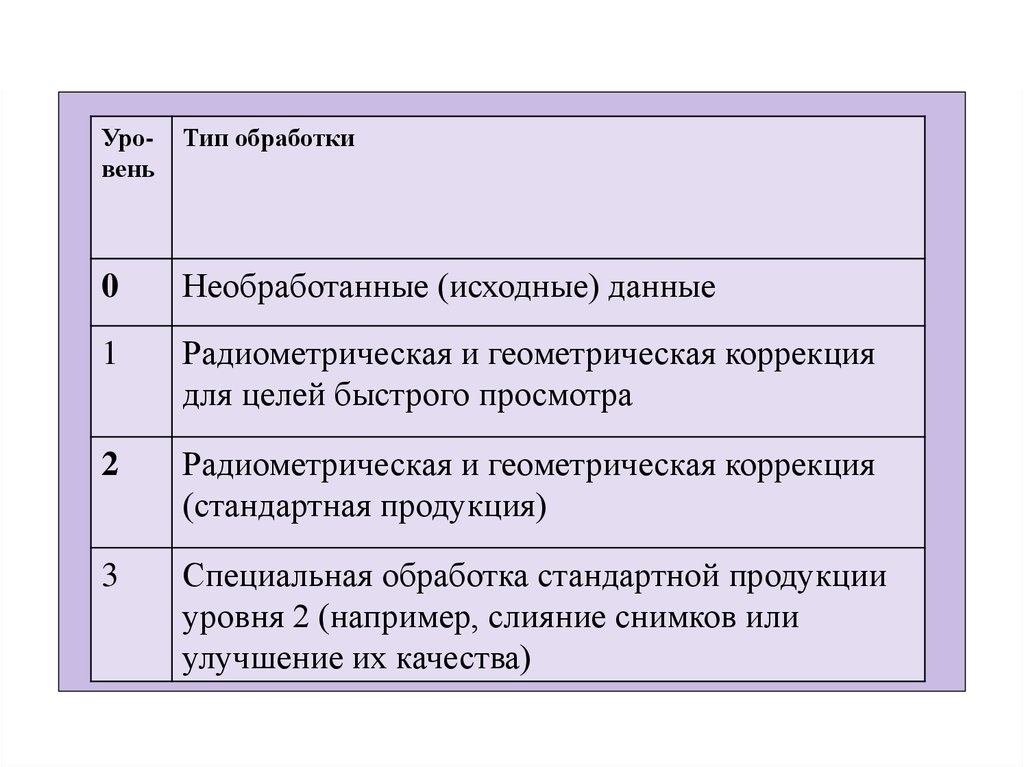


Рисунок 1. Основные уровни дискретизации ДДЗЗ

Считалось, что поскольку разрешение первых съемочных систем было невысоким, то проблема смешанных векторов исчезнет с увеличением разрешения. Однако оказалось, что это не так. При низком разрешении высокая вероятность вкладов от объектов разных классов в один и тот же смешанный вектор сочетается с относительно небольшим их количеством. При увеличении разрешения количество классов, представленных в каждом векторе, уменьшается, но увеличивается число самих смешанных векторов. Таким образом, смешанные векторы всегда являются источником ошибок при классификации ДДЗЗ. Большинство алгоритмов классификации для отнесения векторов признаков классам вычисляют для каждого вектора значения подходящей функции «правдоподобия». В случае зачисления вектора признаков в класс по максимальному значению функции правдоподобия получается так называемая *жесткая кластеризация*. Разделяющие границы в пространстве признаков для полученных при жесткой классификации классов являются четко определенными. Если же величины правдоподобия используются так, что допускается возможность существования множественных классов для каждого вектора, то получается *мягкая или нечеткая классификация*. Величины правдоподобия представляют в этом случае относительные пропорции каждого класса в пределах многоспектрального вектора признаков. В классификации ДДЗЗ чаще других используются методы, которые можно разбить на две группы: классификация с обучением (контролируемая классификация) и кластерный анализ (автоматическая классификация).

2 Классификация с обучением

Непосредственное использование традиционных алгоритмов для обработки данных гиперспектральных сканеров приводит к определенным трудностям, связанным с большой размерностью векторов измерений. Даже в тех местах, которые в обычном свете имеют одинаковый цвет, данные от этих датчиков могут отличаться на сотни и тысячи единиц [3]. С математической точки зрения это делает каждый пиксел изображения уникальным для обрабатывающего компьютера, эта проблема известна под названием эффекта Хьюза и это до последнего времени мешало широкому использованию гиперспектральных датчиков для расширения доступных людям знаний о нашей планете (рисунок 2).

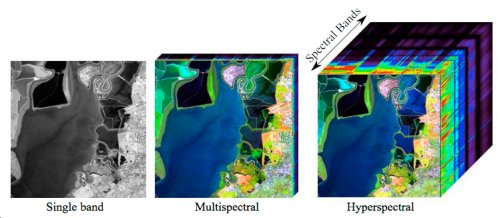


Рисунок 2. Сравнение примеров ДДЗЗ (с одним спектром, мультиспектрального и гиперспектрального)

Методы контролируемой классификации учитывают априорную информацию о типах объектов и вероятности их представления данными снимка, эталонных значениях спектральных характеристик этих объектов. В процессе проведения такой классификации значения яркости текущего пиксела сравниваются с эталонными и на основании этого сравнения пиксел относится к наиболее подходящему классу объектов. После проведения классификации по имеющимся наземным данным оценивается качество проведенной классификации. Обычно контролируемая классификация применяется, когда классы хорошо различаются на снимке и их число варьируется от 25 и выше [4].

При контролируемой классификации правила перехода от показателей спектральной яркости к классам объектов вырабатывают на «учебном» (тестовом, эталонном) участке, а затем автоматически применяют и на остальной части снимка. Контролируемую классификацию реализуют с помощью алгоритмов, использующих разные методы учета спектральных характеристик эталонных объектов:

* дискриминантных функций, которые в простейшем случае представляют собой линии, разделяющие объекты на классы, если они не пересекаются в пространстве признаков;
* минимальных расстояний, когда классифицируемые объекты относят к тому или иному классу, вычисляя евклидово расстояние в пространстве признаков между проверяемыми и эталонными пикселами и относя каждый пиксел к тому классу, до эталона которого это расстояние минимально;
* параллелепипеда, когда в признаковом пространстве выделяют замкнутые прямоугольные области, границы которых задают на основе выбора характерных диапазонов по гистограммам распределения яркостей эталонных объектов в двух, трех или более зонах спектра; каждый пиксел снимка относят к тому классу, в чей параллелепипед он попадает по своим значениям яркости;
* максимального правдоподобия, основанного на статистических оценках характеристик классов объектов. Этот алгоритм применяется наиболее часто. При его реализации для каждого дешифрируемого класса объектов нужно определить две функции условной вероятности, связанные формулой Байеса: Р(Х/Аi) — правдоподобие класса объектов Аi при данном векторе признаков X и P(Аi/X) — вероятность отнесения пикселов с их вектором признаков Х к классу А. Первая функция отражает различия в яркостях объектов разных классов, и ее параметры находят путем вычисления статистических характеристик распределения яркостей пикселов в пределах эталонных объектов на снимке — построения обучающих выборок. Процедура классификации относит каждый пиксел с измеренным Х к тому объекту, для которого правдоподобие максимально, т.е., согласно байесову решающему правилу, нахождение объекта в заданной области признакового пространства статистически наиболее правдоподобно (отсюда и название «метод максимального правдоподобия»).

Априори предполагается, что распределение спектральных яркостей пикселов объекта на многозональном снимке подчинено нормальному закону. Тогда их распределение в пределах объекта образует многомерные «эллипсы рассеяния» равной вероятности, определяемые условием Р(Х/Аi)=const, средними значениями яркости и стандартного отклонения. В большинстве ГИС-пакетов, используемых для выполнения классификации, процедуру обучения осуществляют в интерактивном режиме. Дешифровщик опознает на экране монитора эталонные объекты, относящиеся к выделяемым классам; в пределах каждого объекта строят многоугольник (полигон обучающей выборки), полностью вписывающийся в изображение объекта на экране. Границы этих полигонов вводятся в специальную программу, которая вычисляет для каждого класса средние значения Mi и Ki по значениям яркости ni , пикселов, попавших в полигон. Такая процедура позволяет обеспечить соответствие измерений обучающих (эталонных) и проверяемых (классифицируемых) объектов.

**3 НЕКОНТРОЛИРУЕМАЯ КЛАССИФИКАЦИЯ**

Алгоритмы неконтролируемой классификации, чаще называемые алгоритмами кластеризации, целесообразнее применять при отсутствии априорной информации об объекте съемки. Поскольку кластерный анализ относится к цифровым автоматизированным методам обработки космических изображений, то он позволяет выделять контура с неконтрастной по спектральной яркости структурой, например растительность, открытые почвы, воду, облака другие объекты. С использованием алгоритмов кластеризации возможно выполнить автоматическое разделение пикселов изображения на группы сходных по спектральным характеристикам пикселов – кластеры. При использовании алгоритмов неконтролируемой классификации необходимо иметь минимум исходной информации, например, число классов, длительность классификации и т.д [5].

После проведения неконтролируемой классификации полученная карта классификации более объективно отражает близкие по значениям дешифровочных признаков группы объектов, чем при контролируемой классификации, так как кластеры определяются автоматически. Однако полученная карта классификации требует дальнейшего объединения или разбиения классов, поскольку одни и те же объекты могут попасть в разные кластеры, например из-за условий освещения, а разные объекты – оказаться в одном кластере из-за одинаковой яркости. В первом случае необходимо объединить кластеры в единый класс, а во втором – привлечь дополнительные дешифровочные признаки для различения объектов [6].

Наиболее популярными среди алгоритмов неконтролируемой классификации (кластеризации) являются метод K-средних и метод анализа мод многомерной гистограммы.

**3.1. Метод K-средних**

Этот подход основан на итеративной процедуре отнесения векторов признаков классам по критерию минимума расстояния от вектора до центра класса. Оптимальным считается такое разбиение входных векторов на кластеры, при котором внутриклассовый разброс не может быть уменьшен при переносе какого-либо вектора из одного кластера в другой.

Алгоритм состоит в выполнении следующих шагов:

* На основе заданного соотношения α чистых и смешанных векторов производится разделение векторов на чистые и смешанные. Под смешанными мы понимаем векторы, компоненты которых либо формируются за счет попадания в поле зрения съемочной аппаратуры нескольких объектов, либо искажены влиянием фона. С этой целью вначале для исходного набора векторов измерений рассчитывается градиентное изображение и одновременно строится гистограмма градиентов. Исходя из заданного α, по гистограмме определяется порог, разделяющий векторы на смешанные и чистые [7].
* Объединение чистых векторов в связные компоненты. На этом этапе все чистые векторы объединяются в связные компоненты, которые последовательно нумеруются. Соответствующий алгоритм, близкий к алгоритму заполнения областей с произвольной границей по критерию связности, может выделять и нумеровать одновременно любое количество многосвязных областей без ограничений на их форму и ширину контуров. Для каждой связной компоненты вычисляется вектор средних.
* Итеративная кластеризация векторов средних. Начальные центры кластеров определяются по следующей схеме. В качестве первых двух центров берется пара векторов, наиболее далеких друг от друга. Затем вся выборка делится на кластеры по критерию близости к выбранным центрам. В каждом кластере отыскивается вектор, наиболее далекий от центра. Для всех таких векторов рассчитывается суммарное расстояние до всех центров. В качестве нового центра берется вектор, для которого суммарное расстояние максимально, и процедура распределения векторов по кластерам повторяется.
* Распределение связных компонент по кластерам. На этом этапе связные компоненты получают новые номера. Новый номер присваивается компоненте в соответствии с номером кластера, в который попал вектор средних этой компоненты.
* Кластеризация смешанных векторов. На завершающем этапе производится неитеративная кластеризация смешанных векторов по принципу минимума расстояния до центров кластеров *C1*, . . . , *Ck*. Смешанный вектор *z* будет отнесен к ближайшему кластеру *ωi*, если

**

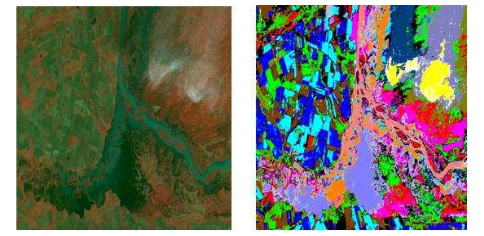


Рисунок 3. Результат работы алгоритма для следующих входных данных: количество выделяемых кластеров 15, соотношение количества смешанных и чистых векторов α= 0.35.

**3.2. Метод анализа мод многомерной гистограммы**

В основе второго подхода лежит предположение, что исходные данные являются хэш-кодирование. Первым шагом модального анализа является поиск ближайших соседей данного вектора списка среди других векторов списка [8,9]. По определению вектор *x* есть ближайший сосед вектора *y*, если  для *y=1,…, N.*

Каждый из возможных ближайших соседей данного вектора x может быть получен из него прибавлением вектора сдвига, компоненты которого принимают значения из множества{−1,0,1}. Алгоритмически *i*-ый вектор сдвига, *i= 1,2, . . . ,3N−1*, можно получить, уменьшив на 1 каждый из коэффициентов представления числа i в троичной системе счисления. Поскольку в реальной гистограмме присутствуют далеко не все ближайшие соседи, то для эффективного их поиска векторы предварительно упорядочиваются в многомерные бинарные деревья. В этом случае время поиска всех ближайших соседей данного вектора становится пропорциональным числу реально существующих соседей. При построении дерева векторы x рассматриваются как N-мерные ключи. Вначале рассчитываются дисперсии по всем координатам векторов и определяется координата j, имеющая максимальную дисперсию. Медианное значение выборки по этой координате используется в качестве ключа для разделения множества векторов на два подмножества: в одно подмножество помещаются векторы, значение которых по координате j меньше порогового значения, а в другое векторы, у которых значение координаты превосходит порог. Каждое из полученных подмножеств делится далее аналогичным образом.

Далее проводится локализация мод гистограммы. Вначале каждому вектору на основе анализа его ближайших соседей ставится в соответствие градиент. Вектору приписывается номер вектора с максимальным значением градиента. Если градиент меньше нуля, то это означает, что координаты вектора являются координатами локального максимума и вектору приписывается его собственный номер. В итоге каждой моде гистограммы сопоставляется выборкой из многомодового закона распределения, причем векторы, отвечающие отдельной моде, образуют кластер. Таким образом, задача сводится к анализу мод многомерных гистограмм [10].

Гистограмма генерируется последовательным просмотром векторов данных и сравнением каждого вектора с текущим списком векторов. При этом либо изменяется соответствующее значение частоты, либо вектор добавляется в список. Для вычисления адресов векторов в списке используется ориентированный граф, корень которого соответствует точке моды. Если количество получаемых кластеров (количество локальных максимумов гистограммы) больше заданного порога, то проводится сглаживание гистограммы. Сглаживание осуществляется либо путем замены частоты h(x) вектора x на среднее значение частот его ближайших соседей, либо путем уменьшения «разрешения» векторов данных, т. е. делением компонент векторов на 2. На завершающем этапе выполняется раскраска ориентированного графа одним цветом, т.е. всем вершинам графа присваивается значение, которое присвоено его корню. Результат кластеризации исходного изображения (на рисунке 3) описанным методом приведен на рисунке 4. Выделено 15 кластеров. Сравнивая результаты различных методов, можно заметить существенные различия: на рисунке 3 лучше разделены сельскохозяйственные угодья, а на рисунке 4 лучше «проработаны» водные поверхности (включая зоны паводкового затопления).

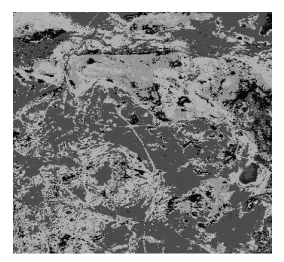


Рисунок 4. Результат обработки ДДЗЗ методом анализа мод многомерной гистограммы.

Эти различия говорят о том, что надежность результатов кластеризации часто можно оценить лишь сравнением нескольких вариантов обработки.

**ЗАКЛЮЧЕНИЕ**

В ходе выполнения курсовой работы за семестр был проведен краткий анализ основных алгоритмов классификации ДДЗЗ, а также приведены примеры их выполнения.

В настоящее время ДЗЗ является источником актуальной и оперативной пространственной информации и широко используется для решения задач в самых различных сферах жизни человека.

Список использованных источников

1. CyberLeninka. Жиляков Е.Г, Черноморец А.А. О дешифрировании изображений земной поверхности по космическим снимкам, 2010 г., с. 1-2 [Электронный ресурс] URL: https://cyberleninka.ru/article/n/o-deshifrirovanii-izobrazheniy-zemnoy-poverhnosti-po-kosmicheskim-snimkam (дата обращения: 13.6.2018)
2. CyberLeninka. Бучнев А.А., Пяткин В.П. Контролируемая и неконтролируемая классификация многоспектральных данных дистанционного зондирования Земли, 2012 г., с. 7-10 [Электронный ресурс] URL: https://cyberleninka.ru/article/n/kontroliruemaya-i-nekontroliruemaya-klassifikatsiya-mnogospektralnyh-dannyh-distantsionnogo-zondirovaniya-zemli (дата обращения: 13.6.2018)
3. CyberLeninka. Бучнев А.А., Пяткин В.П. Некоторые вопросы тематической обработки данных дистанционного зондирования Земли, 2007 г., с. 2-3 [Электронный ресурс] URL: https://cyberleninka.ru/article/n/ nekotorye-voprosy-tematicheskoy-obrabotki-dannyh-distantsionnogo-zondirovaniya-zemli (дата обращения: 14.6.2018).
4. СyberLeninka. Бучнев А.А., Пяткин В.П. Классификация с обучением гиперспектральных данных дистанционного зондирования, 2017 г., с. 9 [Электронный ресурс] URL: https://cyberleninka.ru/article/n/ klassifikatsiya-s-obucheniem-giperspektralnyh-dannyh-distantsionnogo-zondirovaniya-zemli (дата обращения: 14.6.2018)
5. Асмус В.В., Бучнев А.А., Пяткин В.П. Кластерный анализ и классификация с обучением многоспектральных данных дистанционного зондирования Земли / Journal of Siberian Federal University. Engineering & Technologies 1, 2009 г., с. 27-30.
6. О.С. Токарева. Обработка и интерпретация данных дистанционного зондирования: Учеб. пособие. Томск: издательство национального исследовательского томского политехнического университета, 2010. – 148 с.
7. А.А. Воробьева. Дистанционное зондирование Земли: Учеб.-методич. пособие. СПб: изд-во национального исследовательского университета информационных технологий, механики и оптики, 2012. – 168 с.
8. Зубков И.А., Скрипачев В.О. Применение алгоритмов неконтролируемой классификации при обработке данных ДЗЗ / ФГУП «Научный центр информационных космических систем и технологий наблюдения» - Москва – с. 56-57
9. Научный центр оперативного мониторинга Земли [Электронный ресурс], URL: http://www.ntsomz.ru/ks\_dzz (дата обращения: 14.6.2018)
10. Ecoruspace.me [Электронный ресурс], URL: http://ecoruspace.me (дата обращения: 14.6.2018).