Journal of Siberian Federal University. Engineering & Technologies 1 (2009 2), 23-31

УДК 004.4:528.9

Кластерный анализ и классификация с обучением многоспектральных данных дистанционного зондирования Земли

В.В. Асмус[a], **А.А. Бучнев**[b], **В.П. Пяткин**[b*]
[a] Научно-исследовательский центр "Планета", РОСКОМГИДРОМЕТ
123242 Россия, Москва, Большой Предтеченский пер., 7
[b] Институт вычислительной математики и математической геофизики СО РАН
630090 Россия, Новосибирск, пр. Лаврентьева, 6[1]

Получена 15.11.2008, окончательный вариант 10.12.2008, принята к печати 10.03.2009

Рассматриваются вопросы, связанные с проблемой выбора адекватных алгоритмов распознавания многоспектральных данных дистанционного зондирования Земли. Представлена система контролируемой классификации, основанная на стратегии максимального правдоподобия для нормально распределенных векторов признаков. Описывается система кластерного анализа, включающая алгоритм ***K****-средних и метод анализа мод многомерной гистограммы.*

Ключевые слова: дистанционное зондирование Земли, распознавание данных, контролируемая классификация, неконтролируемая классификация, кластерный анализ, решающее правило, обучение классификатора, метод К-средних, многомерная гистограмма.

1. Введение

Эффективность дистанционных исследований Земли из космоса во многом определяется используемыми методами тематической обработки данных дистанционного зондирования Земли (ДДЗЗ). При этом центральной при тематической обработке, безусловно, является система классификации. В модуль распознавания программного комплекса обработки ДДЗЗ, разработанного в ИВМиМГ СО РАН совместно с НИЦ "Планета" Роскомгидромета РФ, включены алгоритмы контролируемой и неконтролируемой классификации многоспектральных ДДЗЗ [1,2]. Центральный вопрос интерпретации ДДЗЗ (вопрос повышения качества дешифрирования) непосредственно связан с проблемой выбора адекватного алгоритма распознавания. Возникающие при этом трудности обусловлены следующими причинами:

1. Структура реальных данных не соответствует модели данных, используемой в алгоритме распознавания. Например, невыполнение предположения о нормальном распределении векторов данных или невыполнение условия, что поле измерений является случайным. Опыт показывает, что такие ситуации возникают при использовании данных в формате JPEG, а также тогда, когда излучение от сканируемого объекта выходит за пределы динамического диапазона съемочной аппаратуры. В этих случаях приходится либо вообще отказываться от методов, требующих обращения ковариационных матриц, либо прибегать к приемам, повышающим дисперсию данных (например, в описываемой ниже системе классификации к спектральным каналам с нулевой дисперсией возможно добавление гауссовского шума с нулевым средним и единичной дисперсией).

2. Нерепрезентативность обучающих последовательностей: недостаточное количество данных для восстановления параметров решающего правила; несоответствие обучающих

* Corresponding author E-mail address: pvp@ooi.sscc.ru

данных и данных, предъявляемых на распознавание ("загрязнение" выборок смешанными векторами измерений, т.е. векторами, которые образуются при попадании в элемент разрешения съемочной системы нескольких природных объектов, неточное соответствие обучающих данных, получаемых с помощью кластеризации, истинным тематическим классам, помехи аппаратуры, влияние атмосферных условий и т.п.) [2].

Таким образом, современный опыт автоматизированного распознавания ДДЗЗ показывает: заранее практически невозможно установить, какой алгоритм будет лучше с точки зрения соотношения точности классификации и стоимости. Поэтому в распознающую систему целесообразно закладывать несколько алгоритмов и выбор оптимального алгоритма проводить эмпирически на этапе обучения по результатам классификации тестовых данных. Выбранный алгоритм используется затем для распознавания всего набора векторов измерений.

2. Контролируемая классификация

Разработанная в ИВМиМГ СО РАН совместно с НИЦ "Планета" система *контролируемой классификации (классификации с обучением)* ДДЗЗ в программном комплексе состоит из семи классификаторов (один поэлементный классификатор и шесть объектных), основанных на использовании байесовской стратегии максимального правдоподобия, и двух объектных классификаторов, основанных на минимуме расстояния. Эта система является частью программного комплекса по обработке ДДЗЗ [3].

2.1. Поэлементная классификация. Под элементом здесь понимается N-мерный вектор измерений (признаков) $x = (x_1, \ldots, x_N)^T$, где N — число спектральных диапазонов. Предполагается, что векторы x имеют в классе ω_i нормальное распределение $N(m_i, B_i)$ со средним m_i и ковариационной матрицей B_i. В этом случае байесовская стратегия максимального правдоподобия для поэлементного классификатора формулируется следующим образом [4–6].

Пусть $\Omega = (\omega_1, \ldots, \omega_m)$ — конечное множество классов, $p(\omega_i)$ — априорная вероятность класса ω_i. Тогда дискриминантная функция класса ω_i имеет вид

$$g_i(x) = \ln(p(\omega_i)) - 0.5 \ln(|B_i|) - 0.5(x - m_i)^T B_i^{-1}(x - m_i). \tag{1}$$

Классическое решающее правило для классификатора принимает следующий вид: вектор x заносится в класс ω_i, если $g_i(x) > g_j(x)$ для всех $j \neq i$.

Для класса ω_i следующим образом определим параметр T_i:

$$T_i = \ln p(\omega_i) - 0.5A(N, Q) - 0.5 \ln |B_i|, \tag{2}$$

где $A(N, Q)$ — критическое значение уровня Q распределения χ^2. Пусть t_i — переменная, значение которой зависит от параметра классификатора thr:

$$t_i = \begin{cases} -\infty, & \text{если } thr = 1, \\ T_i, & \text{если } thr = 2, \\ \min\limits_{i=1}^{m} T_i, & \text{если } thr = 3, \\ \max\limits_{i=1}^{m} T_i, & \text{если } thr = 4, \\ \left(\sum\limits_{i=1}^{m} T_i \right) / m, & \text{если } thr = 5. \end{cases} \tag{3}$$

Тогда решающее правило для классификатора с учетом (3) принимает следующий вид [2]: вектор x заносится в класс ω_i, если $g_i(x) > g_j(x)$ для всех $j \neq i$ и $g_i(x) > t_i$ (при $t_i = T_i$ это означает следующее ограничение расстояния Махалонобиса до центра класса: $(x-m_i)^T B_i^{-1}(x-m_i) < A(N,Q)$). В противном случае вектор заносится в класс отклоненных векторов (класс с номером $m+1$).

Значения $A(N,q)$ для размерности вектора $N \leq 30$ находятся из статистических таблиц. Например, для $N = 4$ и уровня $Q = 0.05$ (т. е. 5% векторов может быть отклонено) $A = 9.488$. При $N > 30$ для нахождения значений $A(N,Q)$ используется аппроксимация

$$\chi^2_{1-Q}(N) \approx N\left(1 - \frac{2}{9N} + u_{1-Q}\sqrt{\frac{2}{9N}}\right)^3, \qquad (4)$$

где u_{1-Q} — значение стандартизованной нормальной величины для вероятности $1-Q$ (в частности, для уровня $Q = 0.05$ значение $u_{0.95} \approx 1.645$).

2.2. Объектная классификация. Под объектом мы понимаем блок смежных векторов квадратной или крестообразной формы. Поскольку физические размеры реально сканируемых пространственных объектов, как правило, больше разрешения съемочных систем, между векторами данных существуют взаимосвязи. Использование информации подобного рода дает возможность повысить точность классификации, если пытаться распознавать одновременно группу смежных векторов – объект в приведенном выше смысле. Рассмотрим вектор (объект) $X(x_1,\ldots,x_L)^T$, состоящий из смежных $N-$мерных векторов x_i, $i = 1,\ldots,L$ (например, в окрестности 3*3, 5*5,... элементов; мы работаем с объектами двух видов – квадратными либо крестообразными). Решение об отнесении центрального элемента объекта тому или иному классу принимается на основе результата классификации всего объекта.

Такой подход порождает целое семейство решающих правил. Во-первых, это использование принципа голосования, т.е. независимая классификация элементов объекта и отнесение центрального элемента к тому классу, которому было отнесено большинство элементов объекта. Во-вторых, это применение текстурных операторов (простейший пример – описание объекта через вектор средних составляющих его элементов) с последующим отнесением центрального элемента классу, к которому был отнесен параметр, характеризующий X. В-третьих, описание объекта случайным марковским полем, т. е. $p(X|\omega_i) = p(x_1|x_2,\ldots,x_L;\omega_i)\ldots p(x_L|\omega_i)$. В этом случае модель выглядит следующим образом. Пусть вектор x имеет в классе ω_i нормальное распределение $N(m_i,B_i)$ со средним m_i и ковариационной матрицей B_i. Тогда вектор также нормально распределен в классе ω_i со средним M_i размерности NL и ковариационной матрицей K_i размерности $NL\times NL$. Оценка этой матрицы при больших значениях NL (требуется очень большое количество обучающих данных), а также ее обращение на практике трудно реализуемо. Поэтому введем упрощающие структурные предположения. Если считать, что корреляция между элементами объекта во всех зонах съемки одинакова, то ковариационную матрицу K_i можно представить в виде прямого произведения матрицы пространственной корреляции R_i на ковариационную матрицу B_i. Если R_i является единичной, то $p(X|\omega_i) = \prod_{i=1}^{L} p(x_i|\omega_i)$ и мы имеем известное решающее правило при предположении, что элементы объекта независимы. Более адекватные модели возникают при других предположениях о структуре корреляционных связей. Например, вводя допущение о разделимости автокорреляционной функции элементов объекта по вертикали и горизонтали, получаем каузальную авторегрессионную модель первого либо третьего порядка (в зависимости от формы объекта).

Приведем алгоритмы работы некоторых объектных классификаторов. Предположим, что векторы внутри блока независимы. Будем рассматривать векторы, составляющие объект X, как один вектор размерности NL. Тогда дискриминантная функция класса ω_i имеет вид

$$g_i(X) = \ln(p(\omega_i)) - 0.5L\ln(|B_i|) - 0.5\sum_{l=1}^{L}(x^l - m_i)^T B_i^{-1}(x^l - m_i).$$

Решающее правило для данных классификаторов принимает следующий вид: центральный элемент объекта X заносится в класс ω_i, если $g_i(X) > g_j(X)$ для всех $j \neq i$ и $g_i(X) > t_i$. В противном случае центральный элемент объекта заносится в класс отклоненных векторов. Здесь t_i определяется по (2) и (3) с $A = A(LN, Q)$.

Снова считаем, что векторы внутри блока независимы. Классифицируется вектор x, равный среднему по всем векторам объекта X:

$$x = \frac{1}{L}\sum_{i=1}^{L} x^i. \tag{5}$$

Дискриминантная функция класса ω_i имеет вид

$$g_i(X) = \ln(p(\omega_i)) - 0.5L\ln(|B_i|) - 0.5L(x - m_i)^T B_i^{-1}(x - m_i).$$

Решающее правило для данных классификаторов таково: центральный элемент объекта X заносится в класс ω_i, если $g_i(X) > g_j(X)$ для всех $j \neq i$ и $g_i(X) > t_i$. В противном случае центральный элемент объекта заносится в класс отклоненных векторов. Здесь t_i определяется по (2) и (3).

Классифицируется средний вектор (5) блока в предположении, что векторы внутри блока независимы и ковариационные матрицы равны единичной. Фактически это объектные классификаторы, решающие правила которых основаны на минимуме евклидова расстояния до центра класса. Дискриминантная функция класса ω_i имеет вид

$$g_i(X) = \ln(p(\omega_i)) - 0.5(x - m_i)^T B_i^{-1}(x - m_i).$$

Решающее правило для данных классификаторов: центральный элемент объекта X заносится в класс ω_i, если $g_i(X) > g_j(X)$ для всех $j \neq i$ и $g_i(X) > t_i$. В противном случае центральный элемент объекта заносится в класс отклоненных векторов. Здесь $t_i = -A$, где $A > 0$ – число, задаваемое пользователем.

Система классификации содержит также объектные классификаторы, основанные на модели каузального марковского случайного поля первого и третьего порядка.

2.3. Обучение и работа классификаторов. Необходимые для построения дискриминантных функций классов оценки статистических характеристик – векторов средних, ковариационных матриц, коэффициентов пространственной корреляции между значениями координат соседних векторов в горизонтальном и вертикальном направлениях – определяются на основе векторов из обучающих выборок (полей). Кроме обучающих для каждого класса может быть задан набор контрольных полей.

Все классификаторы могут использоваться в двух режимах – тестовом и рабочем. По результатам работы классификаторов в тестовом режиме над векторами обучающих и контрольных полей рассчитываются матрица ошибок и оценки вероятностей правильной классификации. Известно (см., например, [4, 6]), что эти оценки для векторов из обучающих полей являются, в среднем, оптимистическими, а для векторов из контрольных полей также в

среднем пессимистическими. Анализируя эти данные, можно оценить (проконтролировать) качество обучения.

Как отмечалось выше, возможны ситуации, при которых нарушается условие о случайности поля измерений, следствием чего выступают нулевые дисперсии в некоторых каналах. Тогда формулы типа (1) становятся неприменимыми, т. к. ковариационные матрицы B вырожденные. Для исправления подобных ситуаций в системе классификации предусмотрена функция добавления к спектральным каналам с нулевой дисперсией гауссовского шума с нулевым средним и единичной дисперсией.

Результатом работы классификаторов в рабочем режиме служит одноканальное (байтовое) изображение, значениями пикселов которого являются номера классов. Это изображение окрашивается в предопределенные цвета, которые в интерактивном режиме могут быть заменены на цвета, определяемые пользователем. Кроме того, к этому изображению можно применить функцию редактирования, которая определяется как уточнение карты классификации на основе учета контекста без изменения перечня ранее выделенных классов. Эта функция может работать в двух режимах: в режиме **Vote**, при котором центральный пиксел окрестности 3*3 заменяется модой гистограммы окрестности, и в режиме **Allsame**, при котором центральный пиксел такой же окрестности меняется только тогда, когда все окружающие его пикселы имеют одинаковое значение.

Система контролируемой классификации имеет следующие характеристики: число обучающих образов – до 9, число классов – до 15, число обучающих и контрольных полей в классе – до 10, размер каждого поля – до 50*50 векторов, размер объекта – от 1*1 до 11*11, размерность векторов данных не ограничивается. На рис. 1 представлена тематическая карта ледовой обстановки в восточном секторе Арктики, полученная с использованием контролируемой классификации. В качестве распознаваемых изображений использовались мозаики радиолокационных и радиометрических снимков с ИСЗ “Океан-О1”. Для выбора тестовых участков использовалось цветосинтезированное изображение.

3. Неконтролируемая классификация

Неконтролируемая классификация (кластерный анализ) в программном комплексе представлена двумя алгоритмами – методом K-средних и методом анализа мод многомерной гистограммы [3].

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:

1. 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.

2. Объединение чистых векторов в связные компоненты. На этом этапе все чистые век-

торы объединяются в связные компоненты, которые последовательно нумеруются. Соответствующий алгоритм, идейно близкий к алгоритму заполнения областей с произвольной границей по критерию связности [3], может выделять и нумеровать одновременно любое количество многосвязных областей без ограничений на их форму и ширину контуров. Для каждой связной компоненты вычисляется вектор средних.

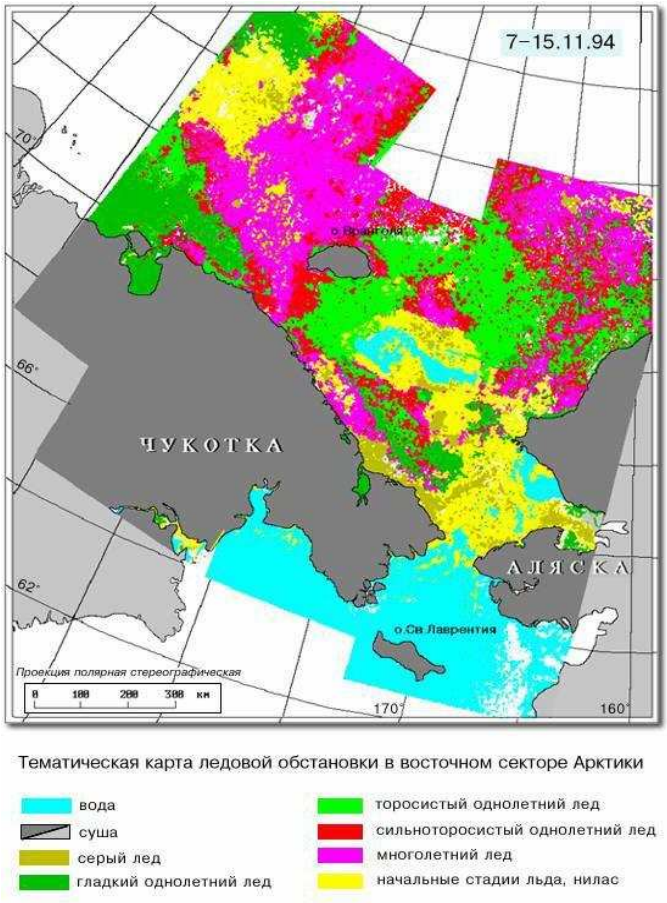

Рис. 1

3. 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.

4. 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.

5. Кластеризация смешанных векторов. На завершающем этапе производится неитеративная кластеризация смешанных векторов по принципу минимума расстояния до центров кластеров $C_1, \ldots, C_k$. Смешанный вектор z будет отнесен к ближайшему кластеру ω_i, если

$||C_i - z|| < 0.5 \max ||C_i - C_l||$, $i, l = 1, \ldots, k,\ k \neq l$.

Рис. 2

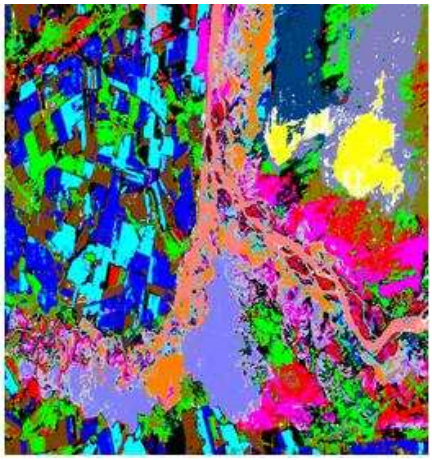

Рис. 3

Рис. 3 иллюстрирует результат работы кластеризации методом K-средних изображения, представленного на рис. 2. Рис. 3 демонстрирует результат работы алгоритма для следующих входных данных: количество выделяемых кластеров 15, соотношение количества смешанных и чистых векторов $\alpha = 0.35$.

3.2. Метод анализа мод многомерной гистограммы. В основе второго подхода лежит предположение, что исходные данные являются выборкой из многомодового закона распределения, причем векторы, отвечающие отдельной моде, образуют кластер. Таким образом, задача сводится к анализу мод многомерных гистограмм. В программный комплекс включена реализация метода, описание основных шагов которого приведено в [3].

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 хэш-кодирование. Первым шагом модального анализа является поиск ближайших соседей данного вектора списка среди других векторов списка. По определению вектор x есть ближайший сосед вектора y, если $|x_i - y_i| \leq 1$ для $i = 1, \ldots, N$. 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-й вектор сдвига, $i = 1, 2, \ldots, 3^N - 1$, 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 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.

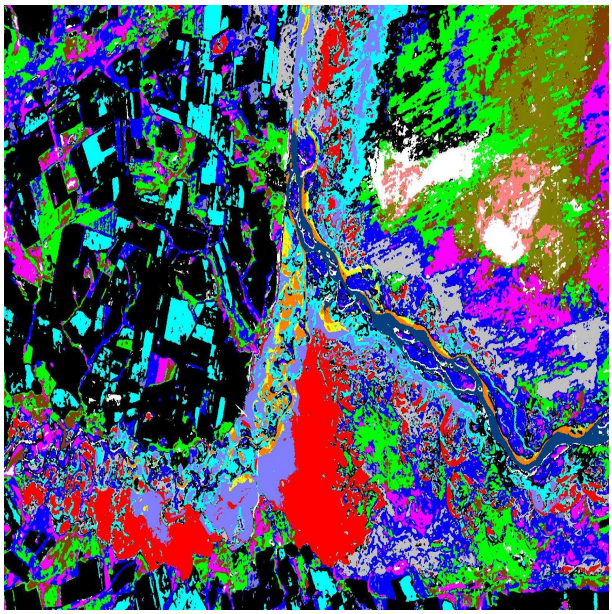

Рис. 4

На рис. 4 приведен результат кластеризации исходного изображения (рис. 2) описанным методом. Выделено 15 кластеров. Сравнивая рис. 3 и 4, можно заметить существенные различия: на рис. 3 лучше разделены сельскохозяйственные угодья, а на рис. 4 лучше "проработаны" водные поверхности (включая зоны паводкового затопления). Эти различия говорят о том, что надежность результатов кластеризации часто можно оценить лишь сравнением нескольких вариантов обработки.

4. Заключение

Практика решения конкретных прикладных задач дистанционного зондирования Земли с использованием предлагаемого подхода подтверждает его эффективность [2].

Работа выполнена частично при финансовой поддержке РФФИ (грант №07-07-00085).

Список литературы

[1] Remote Sensing: The Quantitative Approach // Edited by P.H. Swain and S.M. Davis. USA, McGraw-Hill, Inc., 1978. – 396 p.

[2] Асмус В. В. Программно-аппаратный комплекс обработки спутниковых данных и его применение для задач гидрометеорологии и мониторинга природной среды. Дис. д-ра физ.-математ. наук. На правах рукописи. – М., 2002. – 75 с.

[3] Асмус В. В., Бучнев А. А., Пяткин В. П. Программный комплекс для обработки данных дистанционного зондирования Земли. Труды XXXII Международной конференции "Информационные технологии в науке, образовании, телекоммуникации и бизнесе, IT+SE'2005", приложение к журналу "Открытое образование", 20-30 мая 2005, Украина, Крым, Ялта-Гурзуф. С. 229–232.

[4] Дж. Ту, Р. Гонсалес. Принципы распознавания образов. – М.: Мир, 1978. – 411 с.

[5] Гонсалес Р., Вудс Р. Цифровая обработка изображений. – М.: Техносфера, 2005, – 1072 с.

[6] Marques de Sa J.P. Pattern Recognition: Concepts, Methods and Applications. Springer-Verlag, Berlin, Heidelberg, 2001, – 318 p.

The Cluster Analysis and Classification with Training of Multispectral Data of Earth Remote Sensing

Vasily V. Asmus[a], Alexey A. Buchnev[b] and Valery P. Pyuatkin[b]

[a] Research center "Planeta"
7 Bolshoy Predtechensky, Moscow, 123242 Russia
[b] Institute of Computational Mathematics and Mathematical Geophysics SB RAS
6 pr. Ak. Lavrentjeva, Novosibirsk, 630090 Russia

It is obtained questions, connected with the problem of choosing appropriate algorithms of recognition of multispectral data of Earth remote sensing. It is submitted the system of supervised classification, based on a strategy of maximum probability for vectors of indications having the normal distribution. It is described the system of cluster analysis, including an algorithm for K-means method and analyzing method of mode of multidimensional histogram.

Key words: Earth remote sensing, data recognition, supervised classification, unsupervised classification, cluster analysis, decision rule, classifier training, K-means method, multidimensional histogram.