УДК 528.7

# УЛУЧШЕНИЕ ПРОЦЕССА ТЕМАТИЧЕСКОЙ ОБРАБОТКИ ГИПЕРСПЕКТРАЛЬНОЙ ИНФОРМАЦИИ

***Третьяков В.А.,*** *главный специалист ФГУП «Центральный научно-исследовательский институт машиностроения» (ФГУП ЦНИИмаш), e-mail: tretyakovva@tsniimash.ru;*
***Кротков А.Ю.,*** *главный специалист ФГУП «ЦНИИмаш», e-mail: krotkovay@tsniimash.ru;*
***Кривошеин В.В.,*** *ведущий инженер ФГУП «ЦНИИмаш», e-mail: krivosheinvv@tsniimash.ru;*
***Данилов Р.Ю.,*** *к.б.н., старший научный сотрудник ФГБНУ «Всероссийский научно-исследовательский институт биологической защиты растений», e-mail: daniloff.roman@yandex.ru.*

# IMPROVING THE THEMATIC PROCESSING OF HYPERSPECTRAL INFORMATION

***Tretiakov V.A., Krotkov A.Y., Krivoshein V.V., Ph.D. Biology Danilov R.Y.***

*Thematic processing of hyperspectral information is based on methods of statistical pattern recognition. In this paper, improving the process of the recognition of two classes of ground objects to reduce number and time of calculations is presented. For this purpose we consider the lower recognition error probability estimate limit by spectral features (reflectance). The main assumptions are: classification of the test sample, the equality of covariance matrices of two classes, the independence of spectral features.*
*Methods for determining the minimum required values of the recognition parameters of the selected object classes are developed under the given assumptions. Method based on one sample t-test is developed to determine the values of wavelengths corresponding to the spectral features, under the made assumptions.*
*Ground experimental hyperspectral measurements, carried out on the test site of Krasnodar, were done in order to verify the developed methods.*
*The minimum required volume of the sample, based on two sample Fisher's test is calculated. The necessary number of the spectral informative features is defined using given recognition error probability of two classes of ground objects.*

**Key words:** hyperspectral measurements, reflectance, spectral features, statistical pattern recognition, error probability, training data.

**Ключевые слова:** гиперспектральные измерения, коэффициент спектральной яркости, спектральные признаки, статистическое распознавание, вероятность ошибки, обучающая выборка.

*Предложен подход к улучшению процесса статистического распознавания двух классов наземных объектов с оценкой нижнего предела вероятности ошибки. Проведены расчеты с использованием экспериментально отработанных методик получения и обработки коэффициента спектральной яркости растительности на основе опыта многолетних наземных спектральных измерений, проведённых на Краснодарском тестовом участке для отработки разработанного метода. Определены объем выборки на основе двухвыборочного критерия Фишера и количество спектральных информативных признаков для распознавания объектов с заданной точностью, и оценкой нижней границы вероятности ошибки. С учетом сделанных допущений на основе одно выборочного t-критерия разработан метод определения значений длин волн, соответствующих спектральным признакам.*

## Введение

В настоящее время в нашей стране гиперспектральное (ГС) дистанционное зондирование Земли (ДЗЗ) является активно развивающимся направлением. Интерпретация данных ГС ДЗЗ и решение тематических природоресурсных задач требует разработки сложных алгоритмов, в основе которых лежат методы статистического распознавания объектов. Существует множество отечественных и зарубежных публикаций, в которых описываются подобные методы, а также результаты применения их в тематической обработке ГС данных. Однако, на сегодняшний день недостаточно проработан вопрос обоснования объема обучающей выборки в случае рассмотрения методов контролируемой классификации. В работе [1], посвященной классификации сельскохозяйственных культур с помощью гиперспектральных данных HYMAP, выбирается два набора данных, каждый из которых содержит 6 классов. Объем обучающей выборки из каждого класса выбирается равным 15 % от общего количества пикселей. Первый из двух наборов данных используется для обучения алгоритма классификации методом максимума правдоподобия. Второй набор данных необходим для оценки точности классификации данных. Рассматривался вопрос определения необходимого количества спектральных признаков и был предложен метод на основе вычисления максимального детерминанта ковариационной матрицы классов. Недостаток метода заключается в необходимости перебора каждого спектрального признака, пока добавление нового признака не перестанет увеличивать детерминант ковариационной матрицы.

В работе [2] классифицировались такие объекты, как реки, лес, поля с овощами, крыши домов, земли с за-

стройкой, на мультиспектральном изображении (6 каналов), полученном с оптического датчика TM космического аппарата Landsat различными методами классификации с использованием 12 наборов обучающих выборок объемами 20, 40, 60, 80, 100, 120, 140, 160, 180, 200, 220, 240.

Еще в одной работе [3] исследовалась точность классификации шести классов объектов на изображении высокого разрешения с оптического сенсора космического аппарата Quickbird методом опорных векторов, методом расстояния Махаланобиса, методом максимума правдоподобия и т.д. при различных объемах обучающей выборки 100, 200, 300, 400 для каждого класса. Наивысшая точность 80 % достигается при использовании метода опорных векторов при размере обучающей выборки 200. Метод максимума правдоподобия обеспечил точность распознавания 78,33 % при том же объеме обучающей выборки.

Во всех работах исследовалось влияние объема обучающих выборок на точность классификации хорошо различимых объектов (водные объекты, почвенные объекты, растительность, здания, дороги и т.д.). В нашей работе рассматривается процесс тематической классификации двух классов объектов с тонкими спектральными различиями, описанными в книге [4], и для его улучшения разрабатываются методы определения минимально необходимых параметров распознавания: объем выборки и количество информативных признаков на основе гиперспектральных измерений.

**Постановка задачи исследования**

Пусть на основе анализа процесса А в составе процесса Б, где А – процесс тематической классификации двух классов наземных объектов, Б – тематическая обработка ГС изображения, определены:

а) совокупность $X$ параметров элемента А, где $X$ – гиперспектральные данные;

б) показатель эффективности Pe элемента А, где Pe- вероятность ошибки классификации;

в) параметры $W$, оказывающие наибольшее влияние на выбранный показатель, где $W$ – совокупность таких параметров, как объем обучающей выборки ν, необходимое количество спектральных каналов $n$, значения длин волн $w$;

г) система основных ограничений и допущений $D$, принятая при проведении исследований элемента А, подробно описана ниже.

В ходе исследований требуется разработать математическую модель, которая бы позволила:

1) установить связь
$P_e = P_e\,(X, W, D)$
показателя $P_e$ с параметрами $X, (n, \nu, w) \in W$, при заданных значениях $D$;

2) определить $n$ при заданных значениях $P_e$ и $D$;

3) определить ν при заданных значениях $D$;

4) определить $w$ при заданных значениях $P_e$ и $D$.

**Методы решения задачи исследования**

В основе тематической обработки гиперспектральной информации лежат методы статистического распознавания объектов. Признаками распознавания являются коэффициенты спектральной яркости (КСЯ) наземных объектов, по которым строятся нормальные распределения рассматриваемых классов. Отдельный класс представляет собой набор КСЯ, каждый из которых свернут в одну точку [5]. Предположим, что два выбранных класса наземных объектов, являющиеся множеством КСЯ, подчиняются нормальному закону распределения. Для этого случая приведем систему условий D оценки нижней границы вероятности ошибки статистического распознавания двух классов объектов со своими многомерными нормальными распределениями яркости по спектральным признакам:

1) отсутствие корреляции спектральных признаков;

2) равенство ковариационных матриц двух рассматриваемых классов;

3) проведение классификации объектов по тестовой выборке.

Разберем каждое из приведенных условий.

1. В предыдущей работе [5] обсуждалось понятие n-мерного эллипсойда рассеяния. С его помощью можно описать расстояние Махаланобиса между центрами двух нормальных распределений классов объектов, являющееся их мерой разделимости. При отсутствии корреляции спектральных признаков достигается максимальное значение объема эллипсойда, что свидетельствует о наибольшей разделимости двух классов и, как следствие, о наименьшей вероятности ошибки распознавания. Объем данного эллипсойда пропорционален определителю диагональной ковариационной матрицы одного из классов.

2. Нижняя граница вероятности ошибки распознавания двух нормально распределенных классов достигается при распределениях с равными ковариационными матрицами, верхняя – при равных средних значениях при условии распознавания по одному спектральному признаку [6].

3. Обучающие выборки, использованные для построения классификатора, могут быть затем сами им классифицированы. На практике классификация обучающего набора дает оптимистические результаты, т.е. вероятность ошибки, оцененная по обучающим данным, оказывается ниже вероятности ошибки для полного множества данных. Таким образом, можно сказать, что в лучшем случае классификация обучающих выборок дает нижнюю границу или оптимистический нижний предел истинной вероятности ошибки [7].

***Метод определения объема обучающей выборки* ν**

Из условий независимости спектральных признаков и равенства диагональных ковариационных матриц, описывающих многомерные нормальные распределения яркостей двух классов, следует равенство дисперсий яркостей относительно средних значений каждого спектрального признака для двух классов. Таким образом, отношение несмещенных оценок данных дисперсий будет подчиняться закону Фишера-Снедекора [8]:

$$F = \frac{S^2_{\max}}{S^2_{\min}} \sim \frac{\chi^2_{\nu} / \nu}{\chi^2_f / f},$$

где $S^2_{\max}$ и $S^2_{\min}$ – несмещенные оценки дисперсий яркостей относительно средних значений одного признака для каждого класса, $\chi^2_v$ и $\chi^2_f$ – случайные величины, имеющие распределение $\chi^2$ с $v$ и $f$ – степенями свободы соответственно, $v = n_1\text{-}1$, $f = n_2\text{-}1$, где $n_1$ и $n_2$ – объем выборки для первого и второго класса соответственно.

По таблице распределения Фишера-Снедекора для заданного уровня значимости $\alpha$ определяется объем выборки.

***Метод определения количества спектральных признаков n***

В работе [5] описывался метод определения количества спектральных признаков для распознавания объектов с заданной точностью.

Для модельной ситуации распознавания двух классов с многомерными нормальными распределениями яркостных характеристик, равными ковариационными матрицами и равными априорными вероятностями появления данных классов справедливо выражение для ошибки распознавания в виде

$$P_e = \frac{1}{\sqrt{2\pi}} \int_{1/2\sqrt{J}}^{\infty} e^{-\frac{t^2}{2}} dt , \quad (1)$$

где $J$ – дивергенция, или мера разделимости двух классов по измеренным признакам в $m$ спектральных каналах. Спектральные каналы соответствуют определенным длинам волн, на которых вычисляются признаки распознавания.

Для $m$-канальной спектральной аппаратуры информационное расхождение можно определить по формуле

$$J = m\frac{(B_1 - B_2)^2}{\sigma_1^{\,2}}, \quad (2)$$

где $B_1$ и $B_2$ – средние значения яркостей распознаваемых объектов в одном спектральном канале для первого и второго класса соответственно, а $\sigma_1^2$ – дисперсия яркостей относительно средних значений в одном спектральном канале Приведенное соотношение справедливо при условиях: распределение яркостей ПО подчиняется нормальному закону; измеряемые признаки в разных каналах являются независимыми.

Задавшись вероятностью ошибки распознавания, можно определить дивергенцию двух классов объектов. Вычисляя затем среднее значение $\mu = \frac{B_1 - B_2}{\sigma_1}$, находят количество спектральных признаков $n$.

Заметим, что мы сокращаем количество спектральных каналов до $n$, поэтому количество спектральных признаков $n$ меньше первоначального количества спектральных каналов $m$.

После того как определено количество спектральных признаков $n$, необходимо ответить на вопрос о значениях длин волн $w$, соответствующих спектральным признакам распознавания.

***Метод определения значений длин волн w***

Находя в каждом из спектральных каналов отношение разности матожиданий к дисперсии для двух классов можно вычислить все возможные комбинации, среднее которых будет равно определенному ранее значению $\mu$. Воспользуемся $t$-критерием для одной выборки и проверим равенство матожиданий каждой из найденных комбинаций значению $\mu$.

$T$-критерий для одной выборки позволяет проверить гипотезу о равенстве выборочного среднего некоторому заданному числу. В одновыборочных $t$-критериях, наблюдаемое среднее $\overline{X}$ (вычисленное по реализации выборки отношений разности матожиданий к дисперсии для двух классов) сравнивается с ожидаемым (или эталонным) средним выборки $\mu$ (т.е. с некоторым теоретическим средним). Рассматриваются две гипотезы: $H_0 : \overline{X} = \mu$ и $H_1 : \overline{X} \neq \mu$.

Статистика критерия $t = \frac{\overline{X} - \mu}{s/\sqrt{n}}$ имеет $t$-распределение Стьюдента с ($n$-1) степенью свободы, (в нашем случае $n$-количество найденных спектральных признаков). Выборочное стандартное отклонение $s$ оценивается по наблюдаемой реализации выборки

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}} .$$

Вычисленное значение $t$ проверяют на предмет попадания в критическую область (критическое значение $t_{1-\frac{\alpha}{2}}(v)$ (где $v$ = $n$-1) можно определить по таблицам).

Иными словами, необходимо проверить условие

$$\left|\overline{X} - \mu\right| > \frac{t_{1-\frac{\alpha}{2}}(v)}{\sqrt{n}} s .$$

Если неравенство не выполняется, гипотезу $H_1$ отвергают в пользу $H_0$.

Приведенные известные положения математической статистики положены в основу разрабатываемого в настоящее время алгоритма классификации наземных объектов с тонкими спектральными различиями, в частности, здоровой пшеницы и пшеницы на ранних стадиях заражения бурой ржавчиной.

**Математическое моделирование**

Для математического моделирования использовались наземные ГС измерения двух классов объектов (здоровой и пораженной гербицидами пшеницы), полученные на Краснодарском тестовом участке в мае 2015 года с помощью спектрометра Ocean Optics MAYA 2000-Pro. Полевые измерения приводились к КСЯ, которые стали основой для статистической обработки спектральных характеристик пшеницы в программном пакете MSOffice Exel с помощью стандартных функций $f$-тест и $t$-тест, применение которых описано в работе. Последовательность измерений и предварительной обработки подробно описана в работе [9].

Для статистической обработки были выделены значения КСЯ в диапазоне от 400 до 1100 нм в 1637 спек-

тральных каналах. В электронных таблицах Exel были составлены матрицы КСЯ для первого и второго класса, и средствами пакета анализа данных был применен двух выборочный Фишер-тест в соответствии с приведенными выше формулами, описанными в методе A, для проверки условия равенства ковариационных матриц по объему выборки для каждого класса. Уровень значимости α был выбран по умолчанию 0,05. Объем выборки был выбран одинаковый для обоих классов, т.е. $n_1 = n_2 = \nu$. В процессе математического моделирования в Exel для разного объема выборок определялось наименьшее количество спектральных каналов, в которых не выполняется условие равенства дисперсий для двух классов. В результате была построена зависимость объема выборки от количества спектральных каналов, которая представлена на рис. 1.

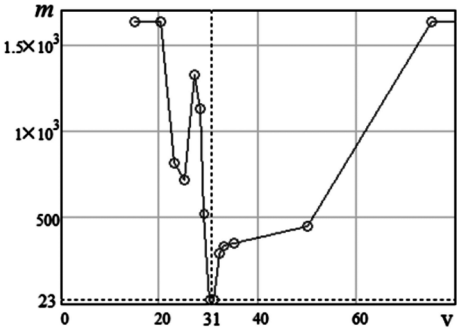


*Рис. 1. Зависимость количества спектральных каналов $m$ от объема выборки $\nu$ (количества наблюдений), в которых не выполняется условие равенства дисперсий в спектральном канале для двух классов*

Как видно из рисунка, наименьшее количество спектральных каналов, в которых не выполнено условие, равно 23. Такое значение соответствует количеству наблюдений 31.

Это количество наблюдений $\nu$, одинаковое для обоих классов, было взято за основу при создании двух обучающих выборок для проведения дальнейшего анализа и определения количества спектральных признаков $n$ с учетом сделанных допущений 1 и 2.

Далее строился график (рис. 2) зависимости вероятности ошибки распознавания двух классов от дивергенции $J$ в соответствии с соотношением (1). Задавшись вероятностью ошибки $P_e$, можно определить $J_{зад}$ по графику и переписать формулу (2) в виде:

$$J_{зад} = n\frac{(B_1 - B_2)^2}{\sigma_1^2}.$$

Отсюда определяется количество спектральных признаков $n$ для распознавания объектов с заданной вероятностью ошибки распознавания, предварительно вычисляя среднее отношение $\mu = \dfrac{B_1 - B_2}{\sigma_1}$ с учетом допущений 1, 2 и 3 для всех 1637 спектральных каналов. Для $P_e$ = 5 % количество спектральных признаков равно 6.

После найденного количества спектральных признаков $n$ и объема обучающей выборки для каждого класса $\nu$ анализировалась возможность определения длин волн $w$, соответствующих найденным спектральным признакам, среднее отношение разности матожиданий к дисперсии которых равно найденному значению $\mu = \dfrac{B_1 - B_2}{\sigma_1}$.

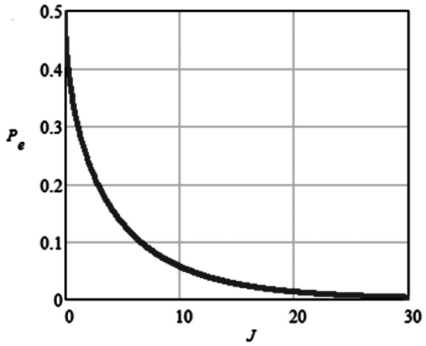


*Рис. 2. Зависимость вероятности ошибки распознавания от дивергенции*

Однако анализ показал, что количество комбинаций по 6 значений отношений разностей матожиданий для двух классов к дисперсии из 1637 составляет $C_{1637}^{6} = \dfrac{1637!}{6!(1637-6)!} = 2.65 \cdot 10^{16}$, поэтому проверка на равенство среднему $\mu$ в соответствии с $t$-критерием не проводилась.

**Обсуждение**

Разработанный метод определения количества спектральных признаков $n$ и соответствующих им значений длин $w$ не целесообразен в применении для обработки информации с оптической аппаратуры с большим количеством спектральных каналов (порядка 1000). Однако такой метод может найти применение на первых этапах обработки космических ГС изображений, получаемых в меньшем количестве спектральных каналов (до 300 для космических ГС приборов), в связи с уменьшением количества комбинаций спектральных каналов, дающих необходимый набор спектральных признаков для распознавания двух классов объектов с заданной вероятностью ошибки, при сделанных допущениях $D$. Этот вопрос требует дальнейшей проработки.

**Выводы и рекомендации**

Благодаря проведенным исследованиям был разработан метод определения необходимого объема обучающих выборок в условиях оценки нижней границы вероятности ошибки статистического распознавания двух классов объектов с тонкими спектральными различиями, позволяющий улучшить процесс тематической обработки космических ГС данных. Улучшение процесса состоит в найденном подходе к определению минимально необходимого объема обучающих выборок $\nu$ с учетом сделанных ограничений. На основе экспериментальных данных такой объем был определен и составил 31 для каждого класса объектов. Этот объем будет использован для получения уточненного объема выборок с учетом снятия наложенных условий, описанных в данной работе. Найденные выборки будут использованы в

дальнейшем для получения обучающих выборок при распознавании двух классов объектов на ГС изображениях, полученных из космоса. Поиску таких выборок и оценке вероятности ошибки распознавания двух классов объектов на основе найденных выборок будут посвящены дальнейшие исследования.

*Работа выполняется при финансовой поддержке РФФИ, проект №16-44-230264 р_а.*

**Литература**

1. Mader S., Vohland M., Jarmer T., Crop classification with hyperspectral data of the HyMAP sensor using different feature extraction techniques- Proceedings of the 2nd Workshop of the EARSeL SIG on Land Use and Land Cover. 28-30 September 2006.

2. Li C., Wang J., Wang L., Hu L., Gong P. Comparison of Classification Algorithms and training sample sizes in Urban Land Classification with Landsat Thematic Mapper Imagery –Remote Sensing. 2014. №6. P. 964-983.

3. Doma M.L., Gomaa M.S., Amer R.A. Sensitivity of pixel-based classifiers to training sample size in case of high resolution satellite imagery – Journal of Geomatics. 2015. V.9. № 1. P. 53-58.

4. Thenkabail Prasad S., Lyon G. John, Huete A. Hyperspectral Remote Sensing of Vegetation, CRC Press, USA, 2011, 782 p.

5. Третьяков В.А. Способ определения числа оптических каналов наблюдения природных объектов гиперспектральными средствами ДЗЗ для их классификации с заданной точностью, Тезисы докладов Второй международной научно-технической конференции «Актуальные проблемы создания космических систем дистанционного зондирования Земли». – М.:ОАО «Корпорация ВНИИЭМ», 2014г. С. 118-125.

6. Барабаш Ю.Л., Варский Б.В., Зиновьев В.Т., Кириченко В.С., Сапегин В.Ф. Вопросы статистической теории распознавания. – М.: Издательство «Советское радио», 1967. – 400 с.

7. Свейн Ф., Дейвис Ш. Дистанционное зондирование Земли: количественный подход. – М.: Издательство «Недра», 1983. – 401 с.

8. Андерсон Т. Введение в многомерный статистический анализ. – М.: Государственное издательство физико-математической литературы, 1963. – 500 с.

9. Акопов А.К., Баула Г.Г., Кривошеин В.В., Кротков А.Ю., Третьяков В.А. Разработка методики наземных валидационных измерений спектров сельскохозяйственных культур // Журнал «Космонавтика и ракетостроение». 2015. вып. № 6 (85). С. 45-50.

## *У в а ж а е м ы е  а в т о р ы !*

Редакция научно-технического журнала «Цифровая обработка сигналов» просит Вас соблюдать следующие требования к материалам, направляемым на публикацию:

**1) Требования к текстовым материалам и сопроводительным документам:**

- *Текст – текстовый редактор Microsoft Word.*
- *Таблицы и рисунки должны быть пронумерованы. На все рисунки, таблицы и библиографические данные указываются ссылки в тексте статьи.*
- *Объем статьи до 12 стр. (шрифт 12). Для заказных обзорных работ объем может быть увеличен до 20 стр.*
- *Название статьи на русском и английском языках.*
- *Рукопись статьи сопровождается: краткой аннотацией на русском и английском языках; номером УДК; сведениями об авторах (Ф.И.О., организация, должность, ученая степень, телефоны, электронная почта); ключевыми словами на русском и английском языках; актом экспертизы (при наличии в вашей организации экспертной комиссии).*

**2) Требования к иллюстрациям:**

Векторные (схемы, графики) – желательно использование графических редакторов Adobe Illustrator или Corel DRAW.

- *Растровые (фотографии, рисунки ) – М 1:1, разрешение не менее 300dpi, формат tiff.*