РОССИЙСКАЯ АКАДЕМИЯ НАУК

# ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ

(ОТДЕЛЬНЫЙ ОТТИСК)

МОСКВА

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 1998, том 38, № 10, с. 1788–1792

УДК 519.7

# ОБ ОПЫТЕ АВТОМАТИЧЕСКОГО СТАТИСТИЧЕСКОГО РАСПОЗНАВАНИЯ ОБЛАЧНОСТИ

© 1998 г. Н. Н. Апраушева*, И. А. Горлач**, А. А. Желнин**, С. В. Сорокин*

(*117967 Москва, ГСП-1, Вавилова, 40, ВЦ РАН;
**123376 Москва, Б. Предтеченский, 9-13, Гидрометцентр РФ)

Поступила в редакцию 22.01.98 г.

Приведены результаты распознавания многоспектральной информации, поступающей с искусственных спутников Земли (и. с. З.), по статистическому алгоритму автоматической классификации (с. а. а. к.), основанному на аппроксимации неизвестной функции плотности вероятности данного множества наблюдений смесью многомерных нормальных распределений. При заданном числе компонент смеси оптимальные оценки неизвестных параметров находятся по алгоритму Дэя–Шлезингера как одно из решений системы уравнений правдоподобия, максимизирующее функцию правдоподобия. Оптимальное число классов определяется методом последовательной проверки двух сложных статистических гипотез. Классификация данного множества наблюдений проводится по правилу Байеса. С целью уменьшения объема вычислений при использовании с. а. а. к. проведен предварительный анализ структуры исследуемого множества, позволяющий грубо оценить число классов и их основные характеристики. Дано описание результатов автоматической классификации – основных типов облачности и подстилающей поверхности.

## ВВЕДЕНИЕ

Необходимость применения данных об облачности и термических характеристиках атмосферы и поверхности Земли как в синоптической практике, так и в моделях анализа и прогноза погоды делает актуальной разработку автоматизированных методов распознавания различных типов облаков. Одним из наиболее перспективных источников информации об облачности являются данные, получаемые по измерениям радиометров высокого разрешения геостационарных спутников. Большие объемы поступающей со спутника информации, необходимость быстрой ее обработки выводят на первый план применение математических методов распознавания (дешифрирования) спутниковой информации.

Первые опыты по автоматическому дешифрированию данных и. с. З. с использованием накопленной априорной информации и. с. З. о различных типах облачности в разных географических условиях и по выбору ее в качестве эталона показали необходимость дальнейшего совершенствования методов обработки данных и. с. З. (см. [1]–[3]). Подход, основанный на исследованиях процессов переноса излучения в различных участках спектра через облака с разными свойствами и на выработке пороговой классификации облачности, также не привел к высокой успешности автоматизированного дешифрирования [4], [5]. Использование статистических алгоритмов автоматической классификации в решении этой задачи имеет ряд преимуществ и позволяет повысить точность распознавания до 75%–80% (см. [6], [7]). Поэтому при дешифрировании относительно небольших по площади участков изображения, занятых фронтальной облачностью, был выбран последний подход. Опробован статистический алгоритм автоматической классификации, основанный на аппроксимации неизвестной функции плотности вероятности данной совокупности наблюдений смесью многомерных нормальных распределений с различными векторами средних значений и с равными ковариационными матрицами.

## 1. ОПИСАНИЕ АЛГОРИТМА РАСПОЗНАВАНИЯ

Если имеется выборка $n$ $p$-мерных наблюдений ($p \geq 1$, $n > p$)

$$X^{(n)} = \{X_1, \ldots, X_n\}, \quad X_j = \{X_{j1}, \ldots, X_{jp}\}, \quad j = 1, 2, \ldots, n, \tag{1.1}$$

в которой значения каждого $i$-го числового признака ($i = 1, 2, \ldots, p$) принадлежат некоторому интервалу $[a_i, b_i]$, $X_{ji} \in [a_i, b_i]$, то ее можно рассматривать как реализацию некоторой непрерывной $p$-мерной случайной величины $\xi = (\xi_1, \ldots, \xi_p)$. Тогда неизвестную функцию плотности вероятно-

сти данной выборки $f(X, \Theta)$ можно аппроксимировать смесью $k$ нормальных распределений $f_i(\mu_i, \Sigma)$ (см. [8])

$$f(X, \Theta) = \sum_{i=1}^{k} \pi_i f_i(X, \mu_i, \Sigma), \tag{1.2}$$

где $\pi_i$ – априорная вероятность $i$-й компоненты смеси, $\mu_i$ – вектор среднего значения $i$-й компоненты смеси, $\Sigma$ – их ковариационная матрица:

$$\pi_i \geq 0, \quad \sum_{i=1}^{k} \pi_i = 1, \quad \Theta = (\pi_1, \ldots, \pi_{k-1}, \mu_1, \ldots, \mu_k, \Sigma).$$

В такой модели под классом подразумевается генеральная совокупность, описываемая унимодальной функцией плотности нормального распределения $f_i(\mu_i, \Sigma)$ ($i = 1, 2, \ldots, k$). Для классификации наблюдений (1.1) необходимо найти оптимальные оценки для неизвестных параметров распределения $k$ и $\Theta$. При известном значении $k$ оптимальной оценкой $\Theta - \Theta_{opt}$ является то решение системы уравнений правдоподобия (с. у. п.), которое обращает функцию правдоподобия $L(X^{(n)}, k, \Theta)$ в максимум:

$$L(X^{(n)}, k, \Theta) = \frac{|\Sigma|^{-n/2}}{(2\pi)^{pn/2}} \prod_{j=1}^{n} \left[ \sum_{i=1}^{k} \pi_i \exp\left( -\frac{1}{2}(X_j - \mu_i)\Sigma^{-1}(X - \mu_i)' \right) \right]. \tag{1.3}$$

При $k = 1$ с. у. п. имеет единственное решение [9], при $k \geq 2$ с. у. п. имеет несколько решений, получаемых по алгоритму Дэя–Шлезингера при различных начальных условиях (см. [10]–[13]).

Трудность использования алгоритма Дэя–Шлезингера состоит в том, что вероятность $P_{opt}$ получения оптимального начального значения параметра $\Theta$ при его случайном задании зависит от размерности выборочного пространства $p$, от расстояния Махаланобиса между классами $\rho_{st}$ ($s, t = 1, 2, \ldots, k$), от смещения оценок $\delta_i$ ($i = 1, 2, \ldots, r$, $r = k - 1 + pk + p(p + 1)/2$) неизвестных координат параметра $\Theta$, от направления больших осей эллипсоидов рассеяния, от числа классов $k$ (см. [13], [14]). Если значения $\delta_i$ ($i = 1, 2, \ldots, r$) велики, а значения $\rho_{st}$ ($s, t = 1, 2, \ldots, k$) малы, от величина $P_{opt}$ может стать сколь угодно близкой к нулю. Поэтому целесообразно не задавать начальные условия случайно, а грубо оценить начальные значения $\Theta_0$, $k_0$, проведя предварительный анализ (п. а.) структуры множества $X^{(n)}$ по подвыборкам, извлеченным из множества (1.1) методом случайного выбора без возвращения (м. с. в. б. в., см. [15]).

Вводя подходящее расстояние на множестве $X$, например евклидово $d$, вычисляем расстояния между всеми различными элементами полученной подвыборки $X'$, $\{d_{ij}\}$ ($i < j$, $i = 1, 2, \ldots, n_1 - 1$, $j = 2, 3, \ldots, n_1$, $n_1$ – объем подвыборки, $n_1 \ll n$). Упорядочивая множество $\{d_{ij}\}$ по возрастанию, имеем основной вариационный ряд (о. в. р.) множества $X'$ (см. [15]). Данные анализа о. в. р. дают оценку снизу $k_m$ для числа классов $k$ и оценку наибольшего диаметра классов $d_{max}$. Далее, применяя к полученной подвыборке $X'$ один из алгоритмов кластер-анализа [16], получаем очень грубые оценки для параметров смеси

$$k_0, \pi_{0i}, \mu_{0i}, \Sigma_0, \quad i = 1, 2, \ldots, k_0, \tag{1.4}$$

которые можно использовать как начальные приближения в алгоритме Дэя–Шлезингера [15]. Пороговое значение расстояния $d_п$ для алгоритма кластер-анализа задаем как часть $d_{max}$, например $d_п = d_{max}/2$. Варьируя $d_п$ с небольшим шагом, будем получать различные значения оценок параметров (1.4).

Оптимальная оценка параметра $k_{opt}$ определяется двумя способами. Один из них базируется на последовательной проверке двух сложных гипотез $H_k$ и $H_{k+1}$ ($k = k_m - 1, k_m, k_m + 1, \ldots, s$, $s \ll n$); гипотеза $H_k$ предполагает, что исследуемая выборка (1.1) состоит из $k$ классов [17]. Из испытуемых последовательных значений $k$ за оптимальное значение $k_1$ принимается то, при котором впервые гипотеза $H_k$ не отвергается по критерию отношения правдоподобия. В случае истинности гипотезы $H_k$ ($k \in \{k_m - 1, k_m, \ldots, k_s\}$, $s \ll n$) статистика

$$\lambda_{k, k+1} = -2\ln[L(X^{(n)}, k, \Theta_{opt}(k))/L(X^{(n)}, k + 1, \Theta_{opt}(k + 1))] \tag{1.5}$$

сходится по вероятности к $\chi^2$-распределению с числом степеней свободы $c$, $c = p + 1$, $p$ – размерность выборочного пространства; $L(X^{(n)}, k, \Theta_{opt}(k))$ – значение функции правдоподобия множества (1.1) при фиксированном значении $k$ при $\Theta = \Theta_{opt}$ (см. [17], [18]).

По второму способу за оптимальное значение $k$ принимается то его значение $k_2$, до которого последовательность значений асимптотических функций правдоподобия $\{L_{ac}(X^{(n)}, k, \Theta_{opt})\}$ ($k = k_m - 1, k_m, \ldots, l, l \ll n$) монотонно возрастает [19]:

$$L_{ac}(X^{(n)}, k, \Theta) = |\Sigma|^{-n/2}(2\pi)^{-pn/2}\prod_{s=1}^{k}\pi_s^{n_s}\prod_{x_{js}\in\omega_s}\exp[-(1/2)(X_{js}-\mu_s)\Sigma^{-1}(X_{js}-\mu_s)'], \quad (1.6)$$

$n_s$ – число элементов класса $\omega_s$. Если значения полученных двумя способами оценок $k$, $k_1$, и $k_2$ не равны, то за оптимальное значение можно принять любое за них или положить $k_{opt} = \min(k_1, k_2)$.

При получении $k_{opt}$, $\Theta_{opt}$ классификация наблюдений (1.1) проводится по правилу Байеса [9], [10]: элемент $X_j$ принадлежит такому классу $\omega_{i_0}$ ($i_0 \in \{1, 2, \ldots, k_{opt}\}$), для которого значение апостериорной вероятности

$$\mathbf{P}(X_j/\omega_i) = \pi_i\exp[-(X_j-\mu_i)\Sigma^{-1}(X_j-\mu_i)'/2]\left\{\sum_{s=1}^{k}\pi_s[\exp(X_j-\mu_s)\Sigma^{-1}(X_j-\mu_s)']/2\right\}^{-1} \quad (1.7)$$

максимально,

$$i_0 = \arg\max_i[\mathbf{P}(X_j/\omega_i)], \quad i = 1, 2, \ldots, k_{opt}. \quad (1.8)$$

В формулу (1.7) вместо истинных значений параметров смеси подставляются значения их оптимальных оценок.

## 2. РАСПОЗНАВАНИЕ ОБЪЕКТОВ ПО ДАННЫМ МЕТЕОРОЛОГИЧЕСКИХ И. С. З.

Для проверки пригодности описанного выше алгоритма для распознавания типов облачности и подстилающей поверхности по многоспектральной спутниковой информации были выбраны три района, наблюдавшиеся со спутников NOAA и METEOSAT. Результаты распознавания по двум районам описаны в [20]. Остановимся на результатах распознавания лишь в одном, наиболее сложном районе, расположенном в Северной Атлантике и наблюдавшемся 9 декабря 1991 г. со спутника METEOSAT.

Объем выборки в этом районе составил $100 \times 30$ пикселей (каждый пиксель – квадрат со стороной ~10 км). В качестве признаков были использованы данные в двух диапазонах спектра излучения: в инфракрасном и диапазоне излучения водяного пара. Таким образом, каждый пиксель был описан двумя слабо коррелированными признаками (их коэффициент корреляции составил 0.4):

$$X_j = (X_{j1}, X_{j2}), \quad j = 1, \ldots, 3000. \quad (2.1)$$

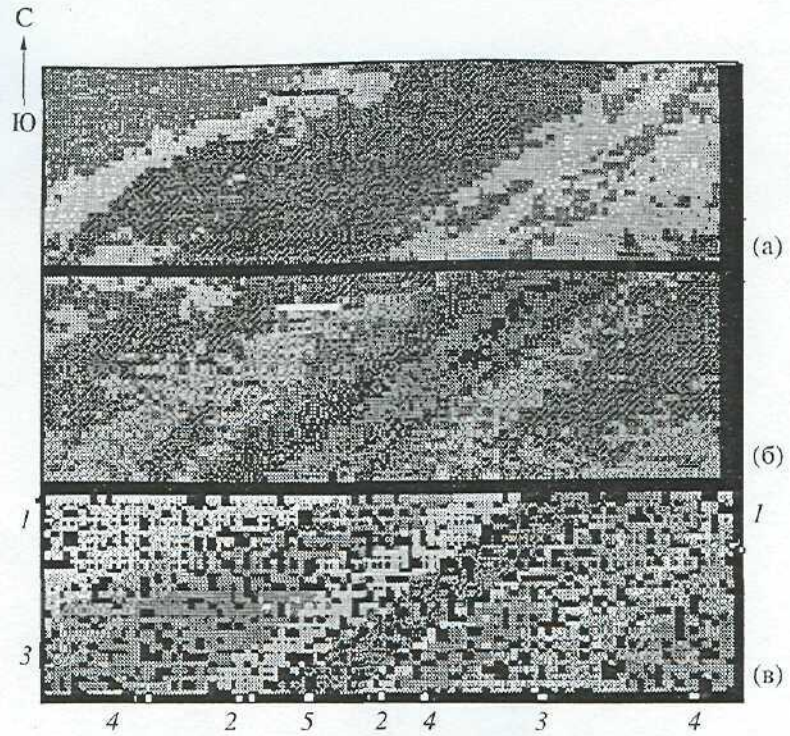
По данным п. а., который проводился по подвыборке объема $n_1 = 450$, получена оценка снизу для числа $k (k \geq 7)$ и значение наибольшего диаметра классов $d_{max} = 30$ (см. [15]). Начальные приближения $k_0$, $\Theta_0$ были получены классификацией этой подвыборки по алгоритму Маккина с использованием в качестве порогового значения для внутриклассовых расстояний $d_п = d_{max}/2$ (см. [15], [16]); ему соответствует оценка $k_0 = 7$. Варьируя значение $d_п$ ($d_п = 16, 15, 14, 12, 10$), по алгоритму Маккина получали различные значения оценок $k_0$ и $\Theta_0$ ($k_0 = 6, \ldots, 10$). При каждом из этих значений $k_0$ уточнялись оценки компонент параметра $\Theta$ по алгоритму Дэя–Шлезингера для подвыборок объемов $n_1 = 450$, $n_2 = 750$. Оптимальное значение $k$ из ряда значений ($k = 6, \ldots, 10$) определялось по значениям статистики $\lambda_{k,k+1}$ (см. (1.5)) для этих подвыборок, приведенным в табл. 1. Задав уровень значимости $\alpha = 0.02$, по таблице распределения $\chi^2$ с тремя степенями свободы имеем $\chi^2_{0.02}(3) = 9.8$ (см. [21]). Тогда по данным табл. 1 получаем

**Таблица 1**

| $n$ | $\lambda_{6.7}$ | $\lambda_{7.8}$ | $\lambda_{8.9}$ | $\lambda_{9.10}$ |
|---|---|---|---|---|
| **450** | 30 | 34 | 8 | 76 |
| **750** | 34 | 58 | 7 | –50 |

**Таблица 2**

| ln / $n$ | $L_{ac(6)}$ | $L_{ac(7)}$ | $L_{ac(8)}$ | $L_{ac(9)}$ | $L_{ac(10)}$ |
|---|---|---|---|---|---|
| **450** | –3169 | –3150 | –3126 | –3116 | –3175 |
| **750** | –6375 | –6344 | –6314 | –6317 | –6379 |




**Фигура.**

$\lambda_{6,7} > 9.8$, $\lambda_{7,8} > 9.8$, $\lambda_{8,9} < 9.8$ для двух подвыборок. Следовательно, $k_1 = 8$ (см. [17]).

В табл. 2 приведены значения логарифмов асимптотической функции правдоподобия (1.6) при $k = 6, \ldots, 10$ для тех же подвыборок.

По второму способу оценивания $k$ имеем $k_2 = 9$ для подвыборки объема **450**, $k_2 = 8$ – для подвыборки объема **750**. Следовательно, $k_{opt} = 8$ (см. [18]), и для алгоритма Дэя–Шлезингера в качестве начального значения вектора $\Theta$ берем тот вектор $\Theta_0$ в (1.4), полученный по алгоритму Маккина, который соответствует $k_0 = 8$. При $k_{opt} = 8$ по алгоритму Дэя–Шлезингера [10]–[12] проводится уточнение оценки векторного параметра $\Theta$ по всей выборке (2.1) с тем, чтобы использовать ее для классификации данных наблюдений по байесовскому правилу (1.8). Ввиду ограниченности оперативной памяти ЭВМ, на которой проводились расчеты, одновременно вводилась не вся выборка (2.1), а каждая из трех ее независимых подвыборок объема $n_i = 1100$ ($i = 1, 2, 3$), полученных м. с. в. б. в. (см. [15], [22]). При этом некоторые из данных наблюдений не попали ни в одну из подвыборок.

Фигура иллюстрирует изображения исследуемого района. В части (а) этой фигуры представлено изображение в инфракрасном диапазоне изучения электромагнитных волн – от 10.5 мкм до 12.5 мкм, в части (б) фигуры – изображение в диапазоне изучения водяного пара от 5.7 мкм до 7.1 мкм, в части (в) – изображение с результатами алгоритмической классификации, где хорошо видна структура интегрального представления наблюдаемой со спутника облачности и водной поверхности; наблюдения (2.1), не попавшие ни в одну из подвыборок, отмечены черным цветом.

Анализ приземных синоптических данных и карт барической топографии показывает, что выбранный район характеризуется поло-

**Таблица 3**

| Номер класса, $s$ | Средние значения | | Априорные вероятности, $\pi_s$ |
|---|---|---|---|
| | $\mu_{s1}$ | $\mu_{s2}$ | |
| 1 | 199 | 148 | 0.17 |
| 2 | 139 | 88 | 0.15 |
| 3 | 180 | 129 | 0.29 |
| 4 | 151 | 120 | 0.26 |
| 5 | 140 | 42 | 0.12 |
| 6 | 28 | 246 | 0.004 |
| 7 | 36 | 142 | 0.002 |
| 8 | 117 | 174 | 0.005 |

жением между двух хорошо выраженных фронтов с полосовой структурой типов облачности, ориентированных с юго-запада на северо-восток. Несмотря на отсутствие некоторых расчетных данных (изображенных черными квадратиками), алгоритм позволил выделить классы изображения, включающие четыре основных типа облачности: *1* – мощные слоисто-дождевые облака, *2* – перистообразная облачность, *3* – слоистообразная, *4* – слоисто-кучевая облачность и *5* – подстилающая поверхность (море). Кроме того, выделены совсем маленькие классы нанесенной привязки информации и бликов.

Оценки средних значений двух признаков по классам (в условных единицах), весов классов, полученные по алгоритму Дэя–Шлезингера для одной из подвыборок объема $n_i = 1100$, приведены в табл. 3. Значения оценок вторых моментов признаков для этой же подвыборки (дисперсий $\sigma_{11}$, $\sigma_{22}$, коэффициентов ковариации $\sigma_{12}$ и корреляции $r_{12}$) одинаковы для всех восьми классов: $\sigma_{11} = 44$, $\sigma_{22} = 81$, $\sigma_{12} = -14$, $r_{12} = -0.2$. Величина коэффициента $r_{12}$ сильно изменилась: с 0.4 перед классификацией до –0.2 после классификации.

Анализ всех полученных результатов автоматического распознавания спутниковой информации по трем выбранным районам позволяет надеяться, что возможно успешное распознавание облачных образований и подстилающей поверхности по описанному выше алгоритму в любом районе земного шара.

## СПИСОК ЛИТЕРАТУРЫ


1. *Соловьева И.С., Сонечкин Д.М., Харитонов В.Ф.* Обработка и анализ телевизионных изображений облачности на ЭВМ // Тр. Гидрометцентра СССР. Л.: Гидрометиздат, 1971. Вып. 73. С. 64–74.
2. *Бакст Л.А., Федорова Н.Н.* Исследование облачности по многоспектральным данным AVHRR ИСЗ NOAA для целей синоптического анализа // Иссл. Земли из Космоса. 1994. № 4. С. 3–8.
3. *Tokuno Masami, Tsuchija Kijoshi.* Classification of cloud types based on data of multiple Satellite sensors // Advances Space Res. 1994. V. 14. № 3. P. 199–206.
4. *Peak J.E., Tag P.M.* Segmentation of Satellite imagery using hierarchial thresholding and neural networks // J. Appl. Meteorology. 1994. V. 33. № 5. P. 605–616.
5. *Strabala K.I., Ackerman S.A., Menzel W.P.* Cloud properties inferared from 8–12 μm data // J. Appl. Meteorology. 1994. V. 33. № 2. P. 212–219.
6. *Bakst L., Fedorova N.* On some methods of synoptic analysis based on the study of the multyspectral satellite data variation // 9-th Meteosat Scient. User's Meeting. Locarno, Switzerland, 1992. P. 25–32.
7. *Bankertt R.L.* Cloud classification of AVHRR imagery in maritime regions using a probabilistic neural network // J. Appl. Meteorology. 1994. V. 33, № 8. P. 1023–1039.
8. *Волошин Г.Я., Бурлаков И.А., Косенкова С.Т.* Статистические методы решения задач распознавания, основанные на аппроксимационном подходе. Владивосток, 1992. Ч. 1.
9. *Андерсон Т.* Введение в многомерный статистический анализ. М.: Физматиз, 1963.
10. *Day N.E.* Estimating the components of a mixture of normal distributions // Biometrika. 1969. V. 56. № 3. P. 463–474.
11. *Шлезингер М.И.* Взаимосвязь обучения и самообучения в распознавании образов // Кибернетика. 1968. № 2. С. 81–88.
12. *Апраушева Н.Н.* Алгоритм расщепления смеси нормальных классов // Программы и алгоритма. М.: ЦЭМИ АН СССР, 1976. № 68.
13. *Апраушева Н.Н.* Исследование одного алгоритма расщепления смеси нормально распределенных классов // Многомерный статистич. анализ в социально-экономич. иссл. М.: Наука, 1974. С. 135–150.
14. *Апраушева Н.Н.* Преобразование признаков при статистическом решении одной задачи автоматической классификации // Изв. АН СССР. Сер. Техн. кибернетика. 1985. № 2. С. 167–174.
15. *Апраушева Н.Н.* Новый подход к обнаружению кластеров. М.: ВЦ РАН., 1993.
16. *Дюран Б., Оделл П.* Кластерный анализ. М.: Статистика, 1975.
17. *Апраушева Н.Н.* Определение числа классов в задачах классификации // Изв. АН СССР. Сер. Техн. кибернетика. 1981. № 3. С. 71–77.
18. *Уилкс С.* Математическая статистика. М.: Наука, 1973.
19. *Апраушева Н.Н.* Определение числа классов в задачах классификации // Изв. АН СССР. Сер. Техн. кибернетика. 1981. № 5. С. 153–160.
20. *Апраушева Н.Н., Бакст Л.А., Горлач И.А. и др.* О распознавании типов фронтальной облачности по данным ИСЗ. М.: ВЦ РАН, 1996.
21. *Крамер Г.* Математические методы статистики. М.: Мир, 1976.
22. *Феллер В.* Введение в теорию вероятностей и ее приложения. М.: Мир, 1984.