РОССИЙСКАЯ АКАДЕМИЯ НАУК
ВЫЧИСЛИТЕЛЬНЫЙ ЦЕНТР

СООБЩЕНИЯ ПО ПРИКЛАДНОЙ
МАТЕМАТИКЕ

Н. Н. АПРАУШЕВА, И. А. ГОРЛАЧ, Д. Е. ИВАНОВ

СТАТИСТИЧЕСКАЯ СИСТЕМА АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ

ВЫЧИСЛИТЕЛЬНЫЙ ЦЕНТР РАН
МОСКВА 1998

УДК 519

Ответственный редактор
доктор технических наук В. М. Солодов

В работе описана статистическая система автоматической классификации, предназначенная для персональных электронно-вычислительных машин. Система основана на математической модели смеси многомерных нормальных распределений. Приведен пример ее использования в решении задачи распознавания видов облачности по данным метеорологических искусственных спутников Земли.

Рецензенты: С. Н. Дулин,
Л. З. Яшин

Научное издание

Введение

Под классификацией (распознаванием образов) обычно понимается упорядочение объектов по их схожести. Под объектами подразумевают предметы, явления, события, процессы, ситуации и т.д. Класс (кластер, образ) можно определить как некоторое подмножество наиболее сходных между собой элементов данного множества [1-3].

Задачи классификации существуют и возникают во всех областях науки и практики. Их успешное решение во многом обусловлено применением математических методов и широким использованием ЭВМ, обладающих быстродей-

ствием, большой памятью, позволяющей анализировать большие массивы информации. В связи с этим возникает необходимость разработки и создания автоматизированных систем классификации, предназначенных для пользователей-специалистов самых различных областей науки и практики, возможно, весьма далеких от математики.

В данной работе представлена статистическая система автоматической классификации (ССАК) множества многомерных наблюдений любой природы, рассматриваемых как реализация некоторой многомерной случайной величины. Приведен пример использования этой системы на конкретных данных многоспектральной информации метеорологических искусственных спутников Земли (МИСЗ).

1. Постановка задачи классификации

Пусть имеется p-мерная выборка n наблюдений ($p \geq 1, n >> p$ **)**

$$X^{(n)} = \{X_1, X_2, \ldots, X_n\},\ X_i = (x_{i1}, x_{i2}, \ldots, x_{ip}), \qquad (1.1)$$

в которой значения каждого j-го числового признака x_{ij}, где $j = 1,\ldots,p$, принадлежит некоторому интервалу $\left[a_j, b_j\right]$, $x_{ij} \in \left[a_j, b_j\right]$, $\boldsymbol{i} = 1,\ldots,\boldsymbol{n}$. Тогда ее можно рассматривать как реализацию некоторой непрерывной p-мерной случайной величины $\xi = \left(\xi_1, \xi_2, \ldots, \xi_p\right)$, неизвестную функцию плотности вероятности $f(X, \Theta)$ которой аппроксимируем смесью k-нормальных распределений (СНР) $f_l(\mu_l, M_l)$ [4, 5],

$$f(X, \Theta) = \sum_{l=1}^{k} \pi_l f_l(X, \mu_l, \mathrm{M}_l), \qquad (1.2)$$

где Θ - r-мерный параметр, $r = k - 1 + kp + \frac{kp(p+1)}{2}$, $\Theta = \left(\theta_1, \theta_2, \ldots, \theta_r\right)$, или $\Theta = (\pi_1, \pi_2, \ldots, \pi_{k-1}, \mu_1, \mu_2, \ldots, \mu_k, \mathrm{M}_1, \mathrm{M}_2, \ldots, \mathrm{M}_k)$; π_l - априорная вероятность $\pi_l \geq 0$, $\sum_{l=1}^{k} \pi_l = 1$, μ_l - вектор среднего значения, $\mu_l = \left(\mu_{l1}, \mu_{l2}, \ldots, \mu_{lp}\right)$, M_l - ковариационная матрица l - й компоненты смеси, $M_l = \left(\sigma_{s,t}^{l}\right)$, $l = 1,2,\ldots,k$, $s,t = 1,2,\ldots,p$.

В такой модели под классом подразумевается генеральная совокупность, описываемая унимодальной функцией плотности нормального распределения $f_l(\mu_l, M_l)$, $l = 1,2,\ldots,k$.

Требуется найти оптимальные оценки для неизвестных параметров СНР $\Theta_{опт}$ и $k_{опт}$ для классификации наблюдений по правилу Байеса**.**

2. Алгоритм решения задачи классификации

Блок-схема алгоритма автоматической классификации множества наблюдений (1.1) представлена на рис. 1. Рассмотрим подробное описание каждого блока.


Начало
Блок предварительного анализа
Блок предварительной классификации
Блок уточнения параметров CHP
Блок определения числа классов
Блок заключительной классификации
Конец

Рис.1

Блок предварительного анализа. Сначала следует определить, состоит ли множество (1.1) из k отдельных классов или оно однородно.

Для ответа на поставленный вопрос зададим на множестве евклидову метрику [3],

$$r_{ig} = \left(\sum_{j=1}^{p} \left(x_{ij} - x_{gj} \right)^2 \right)^{\frac{1}{2}} , \; X_i, X_j \in X^{(n)}. \quad (2.1)$$

Вычислив по формуле (2.1) расстояния между всеми различными точками множества $X^{(n)}$ и упорядочив элементы полученного множества $R' = \left\{ r_{ig}, i = 1,\ldots,n-1, g = 2,\ldots,n, i < g \right\}$ по возрастанию, получим основной вариационный ряд (ОВР) множества $X^{(n)}$,

$$r_{(1)} \le r_{(2)} \le \ldots \le r_{(s)}, \quad s = \frac{n \cdot (n-1)}{2}. \quad (2.2)$$

Если множество $X^{(n)}$ состоит из k $(k \ge 2)$ далеко отстоящих друг от друга (в некотором смысле) кластеров (классов), то функция плотности вероятности расстояния между всеми его различными точками $f(r)$ имеет хотя бы один локальный минимум (ЛМИН). Число ЛМИН функции

$f(r)$ зависит не только от степени близости кластеров, но от их взаимного расположения.

Чтобы выяснить, имеет ли ЛМИН неизвестная функция $f(r)$, вводится понятие статистически значимого локального минимума гистограммы (СЗЛМИН) соответствующих наблюдений.

Пусть гистограмма наблюдений, построенная в соответствии с рекомендациями использования того или иного статистического критерия согласия [3], имеет ЛМИН на отрезке $\left[r_q, r_{q+1}\right]$, а на отрезках $\left[r_v, r_{v+1}\right]$, $\left[r_u, r_{u+1}\right]$ - локальные максимумы (ЛМАКС), ближайшие к этому минимуму (рис 2).

Для определенности положим $r_{v+1} \le r_q$, $r_{q+1} \le r_u$, $\widetilde{f}\left(r_u\right) \le \widetilde{f}\left(r_v\right)$, где $\widetilde{f}\left(r_u\right)$-значение гистограммы в полуинтервале $\left[r_u, r_{u+1}\right)$. По предположению в промежутке $\left[r_u, r_{u+1}\right)$ гистограмма имеет наименьший из двух рассматриваемых максимумов.

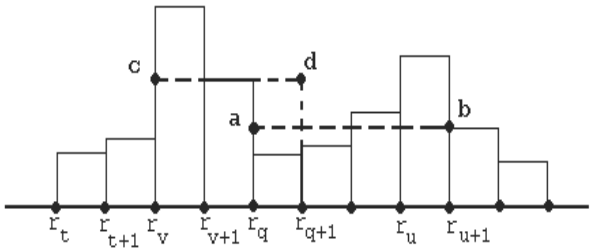


Рис. 2

Определение 2.1. ЛМИН, наблюдаемый в полуинтервале $\left[r_q\,,\,r_{q+1}\right)$ гистограммы, называется слабым или статистически незначимым, если на выбранном уровне значимости $\alpha\,(\alpha>0)$ в минимальном промежутке $\left[r_q\,,\,r_{u+1}\right)$, содержащем этот ЛМИН и наименьший из двух ближайших к нему ЛМАКС, не отвергается гипотеза H_0 о постоянстве функции $f(r)$,

$$H_0:\ f(r)=1\Big/\left(r_{u+1}-r_q\right),\quad r\in\left[r_q\,,\,r_{u+1}\right) \qquad (2.3)$$

при альтернативе H_1 о монотонности $f(r)$, $H_1 : f(r)$ - монотонно возрастающая (убывающая) функция при $r \in \left[r_q\ , r_{u+1}\right)$.

Определение 2.2. ЛМИН, наблюдаемый в полуинтервале $\left[r_q\ , r_{q+1}\ \right)$ гистограммы, называется статистически значимым (СЗЛМИН), если на выбранном уровне значимости $\alpha\ (\alpha > 0)$ в минимальном промежутке $\left[r_q\ , r_{u+1}\ \right)$, содержащем этот ЛМИН и наименьший из двух ближайших к нему ЛМАКС, отвергается гипотеза H_0 в пользу альтернативы H_1.

Очевидно, если в промежутке $\left[r_q , r_{q+1}\right]$ гистограмма имеет СЗЛМИН на выбранном уровне значимости $\alpha\ (\alpha > 0)$, то соответствующая ей неизвестная функция $f(r)$ в этом промежутке имеет ЛМИН с вероятностью $1 - \alpha$. На языке кластер – анализа имеем следующее утверждение: множество $X^{(n)}$ состоит из k $(k \geq 2)$ кластеров с вероятностью $1 - \alpha$, $\alpha > 0$, если гистограмма его ОВР имеет хотя бы один СЗЛМИН на выбранном уровне значимости α.

Если на гистограмме $f(r)$ обнаружен слабо выраженный ЛМИН, то неизвестно, имеет ли его соответствующая функция $f(r)$.

Возможна ситуация, когда функция $f(r)$ в некотором интервале имеет локальный экстремум, но он не обнаруживается на гистограмме используемыми статистическими критериями.

Для обнаружения СЗЛМИН гистограмм наблюдений могут использоваться такие статистические непараметрические критерии, как Колмогорова, ω^2, χ^2 и др. [6].

При использовании критерия χ^2 мера близости между распределениям и задается в виде

$$\chi^2 = \sum_{i=1}^{j_v} \frac{\left(v_i - v/j_v\right)^2}{v/j_v} \quad , \qquad (2.4)$$

где v - число наблюдений, попавших в промежуток $\left[r_q, r_{u+1}\right)$, j_v - число интервалов гистограммы, попавших в $\left[r_q, r_{u+1}\right)$, v_i - число наблюдений, содержащихся в каждом из этих интервалов.

Для проверки гипотезы H_0 при альтернативе H_1 задается уровень значимости α, например $\alpha = 0.05$. По таблице распределения χ^2 находим значение $\chi^2_{0.05}(t)$, t - число степеней свободы, $t = j_v - 1$ [3, 6]. Если значение χ^2, подсчитанное по формуле (2.4), больше или равно $\chi^2_{0.05}(t)$, то гипотеза H_0 отвергается и исследуемый локальный минимум гистограмм является статистически значимым [3].

Максимальное число ЛМИН функции плотности вероятности $f(r)$ равно: $m_{\max} = \frac{k(k-1)}{2}$. Если $\overline{m}$ - наблюдаемое число СЗЛМИН гистограммы наблюдений, то $\overline{m} \le \frac{k(k-1)}{2}$. Отсюда

$$k \ge E\left[\frac{\left(1+\left(1+8\overline{m}\right)^{1/2}\right)}{2} - \text{å}\right] + 1 \quad , \qquad (2.5)$$

где $E[y]$ - целая часть от y, å - малое положительное число.

Блок предварительной классификации предназначен для проведения предварительной классификации, целью

которой является нахождение грубых оценок для неизвестных параметров Θ и k с использованием оценок минимального числа классов $k_{мин}$ и максимального диаметра классов $D_{макс}$. Значение $D_{макс}$ вычисляется по гистограмме ОВР (рис.2):

$$D_{макс} = (r_q + r_{q+1})/2 . \qquad (2.6)$$

Для проведения предварительной классификации использован алгоритм Мак-Кина [7], который объединяет элементы множества (1.1) в классы по степени близости расстояния между ними. Перед началом работы алгоритма Мак-Кина необходимо задать каким-либо способом значение порогового расстояния $D_{пор}$, по которому будет осуществляться объединение объектов в классы. За начальное значение $D_{пор}$ принимается $D_{макс} / 2$. В один класс группируются все точки, евклидово расстояние каждой из которых до динамического центра тяжести μ_l не превышает $D_{пор}$. По результатам разбиения вычисляются грубые оценки значений априорных вероятностей π_l, векторов среднего значения μ_l и ковариационных матриц M_l классов по следующим формулам:

$$\pi_l = {}^{n_l}\!/_{n}\ , \qquad (2.7)$$

$$\mu_l = {}^{1}\!/_{n_l} \sum_{X_l \in \omega_l} X_l , \qquad (2.8)$$

$$\mathrm{M}_l = {}^{1}\!/_{n_l} \sum_{X_l \in \omega_l} \left(X_l - \mu_l\right)^T \times \left(X_l - \mu_l\right), \qquad (2.9)$$

где n_l - число наблюдений, попавших в l-й класс, n - общее количество точек, l - номер класса ($l = 1,2,\ldots,k_{грубб}$), X_l - наблюдение.

Блок уточнения параметров СНР предназначен для нахождения оптимальных оценок для параметров СНР $\Theta_{опт}$. Входными данными для этого алгоритма являются: предполагаемое число наблюдаемых классов $k_{грубб}$, а также грубые оценки параметров СНР $\Theta_{грубб}$, вычисляемые по формулам (2.7 - 2.9).

При известном значении числа классов оптимальной оценкой Θ - $\Theta_{опт}$ является то решение системы уравнений правдоподобия (СУП)

$$\frac{\partial \ln L\left(X^{(n)}, k, \Theta\right)}{\partial \theta_t} = 0\,,\ t = 1,2,\ldots,r\,,$$

$$r = k - 1 + kp + kp(p+1)/2\,, \qquad (2.10)$$

которое обращает функцию правдоподобия $L\left(X^{(n)}, k, \Theta\right)$ в максимум,

$$L\left(X^{(n)}, k, \Theta\right) = \frac{1}{(2\pi)^{pn/2}} \prod_{i=1}^{n} \left[\sum_{l=1}^{k} \frac{\pi_l}{|M_l|^{1/2}} \exp\left(-\frac{1}{2}\left(X_i - \mu_l\right) M_l^{-1} \left(X_i - \mu_l\right)^T \right) \right]. \qquad (2.11)$$

При $k = 1$ СУП имеет единственное решение [8], при $k \geq 2$ СУП имеет несколько решений, получаемых по алгоритму Дэя-Шлезингера при различных начальных значениях параметров дискриминантных поверхностей (ДП) [9 - 11].

Трудность использования алгоритма Дэя-Шлезингера состоит в том, что вероятность $P_{опт}$ получения оптимального значения параметра Θ при случайном задании параметров ДП зависит от расстояния Махаланобиса между классами $\rho_{s,t}$, $s,t = 1,2,\ldots,k$, от смещения оценок δ_t, $t = 1,2,...,r$, неизвестных параметров Θ_t, от числа классов k, от направления больших осей эллипсоидов рассеяния, от

размерности выборочного пространства [12-14]. Если значения δ_t велики, а значения $\rho_{s,t}$ малы, то величина P_{onm} может стать сколь угодно близкой к нулю. Поэтому целесообразно не задавать начальные условия случайно, а предварительно грубо оценить начальные значения Θ_0, k_0.

Алгоритм Дэя-Шлезингера позволяет оценить неизвестные параметры СНР при конкретном числе классов k методом последовательных итераций, который заключается во взаимном пересчете апостериорных вероятностей $P\left(l/X_i\right)$ относительно априорных вероятностей π_l, векторов средних значений μ_i и ковариационных матриц M_l. Система уравнений правдоподобия при $k \geq 2$ принимает вид

$$\boldsymbol{P}(\boldsymbol{l}/\boldsymbol{X_i}) = \frac{\dfrac{\pi_l}{\left|\boldsymbol{M_l}\right|^{1/2}} \exp\left[-\frac{1}{2}\cdot(\boldsymbol{X_i}-\mu_l)\cdot \boldsymbol{M_l}^{-1}\cdot(\boldsymbol{X_i}-\mu_l)^T\right]}{\sum_{s=1}^{k} \dfrac{\pi_s}{\left|\boldsymbol{M_s}\right|^{1/2}} \exp\left[-\frac{1}{2}\cdot(\boldsymbol{X_i}-\mu_s)\cdot \boldsymbol{M_i}^{-1}\cdot(\boldsymbol{X_i}-\mu_s)^T\right]}, \quad (2.12)$$

$$\pi_l = \frac{1}{n} \cdot \sum_{i=1}^{n} P\left(l/X_i\right), \qquad (2.13)$$

$$\mu_l = \frac{\sum_{i=1}^{n} X_i \cdot P\left(l/X_i\right)}{\sum_{i=1}^{n} P\left(l/X_i\right)}, \qquad (2.14)$$

$$M_l = \frac{\sum_{i=1}^{n} \left[\left(X_i - \mu_l\right)^T \left(X_i - \mu_l\right) P\left(l/X_i\right) \right]}{\sum_{i=1}^{n} P\left(l/X_i\right)}. \qquad (2.15)$$

Последовательность итераций $\grave{E}^{(s)} = (\pi_l^{(s)}, \mu_l^{(s)}, M_l^{(s)})$, $s = 0,\ldots,s_0$, получаемая по формулам (2.12 - 2.15) при заданном значении Θ^0 сходится к некоторому решению СУП $\widetilde{\Theta}$. Итерационный процесс прекращается, когда выполняется следующее неравенство:

$$\max_{1 \le l \le k} \left| \mu_l^{(t)} - \mu_l^{(t+1)} \right| \le \varepsilon, \qquad (2.16)$$

где t – шаг итерации, а ε задается, исходя из смысла данных наблюдений.

Блок определения числа классов предназначен для нахождения оптимальной оценки числа классов $k_{опт}$ множества (1.1). Алгоритм данного блока базируется на после-

довательной проверке двух сложных гипотез H_k и H_{k+1} [15, 16]. Гипотеза H_k предполагает, что исследуемое множество (1.1) состоит из k классов.

Задав уровень значимости $\alpha \geq 0$, например $\alpha = 0.05$, по таблице χ^2-распределения [6] находим $\chi^2_\alpha(c)$, $c = \frac{(p+1)(p+2)}{2}$. Если при каком-то значении $k \in \{k_{min} - 1, k_{min}, \ldots, s\}$, $s << n$, величина

$$\lambda_{k,k+1} = -2\ln\left(\frac{L\left(X^{(n)}, k, \theta_{opt}(k)\right)}{L\left(X^{(n)}, k+1, \theta_{opt}(k+1)\right)}\right) \quad (2.17)$$

удовлетворяет неравенству $\lambda_{k,k+1} \geq \chi^2_\alpha(c)$, то гипотеза H_k отвергается и проверяется гипотеза H_{k+1} при альтернативе H_{k+2}. Если же при проверке гипотезы H_k $\lambda_{k,k+1} < \chi^2_\alpha(c)$, то гипотеза H_k принимается.

Блок заключительной классификации предназначен для проведения окончательной классификации с использованием полученных оптимальных оценок числа классов $k_{опт}$ и параметра СНР $\theta_{опт}$. Классификация проводится по правилу Байеса: элемент X_i принадлежит такому классу l_0,

$l_0 \in \{1,2,\ldots,k_{опт}\}$, для которого значение апостериорной вероятности, вычисляемое по формуле (2.12), максимально, $l_0 = \arg\max_l \left(P\left(l \,/ X_i\right)\right)$.

3. Общая структурная схема статистической системы автоматической классификации

Схема реализованной автоматизированной системы автоматической классификации представлена на рис.3. Система представляет собой программный продукт, который может выполняться на любой ЭВМ, отвечающей следующим требованиям при наличии:

- оперативной памяти более 32 Мб (зависит от объема анализируемых данных);
- быстродействия более 200 MHz (анализ данных должен выполняться за минимальное количество времени);
- свободного места на жестком диске не менее 100 Мб для размещения самого программного продукта и анализируемых данных.

Представленная система реализует алгоритм, описанный в предыдущем разделе, работой которого управляет оператор (синоптик-дешифровщик).

Взаимодействие оператора с системой обеспечивается *блоком управления* (другими словами, он отвечает за организацию интерфейса между пользователем и системой) и осуществляется в двух направлениях:

1) с одной стороны, оператор задает начальные параметры, управляет ходом вычислительных процессов, вносит необходимые коррективы с помощью клавиатуры и мыши ЭВМ;
2) с другой стороны, система отображает полученные в ходе вычислительного процесса промежуточные и окончательные результаты на монитор либо на принтер и другие периферийные устройства.

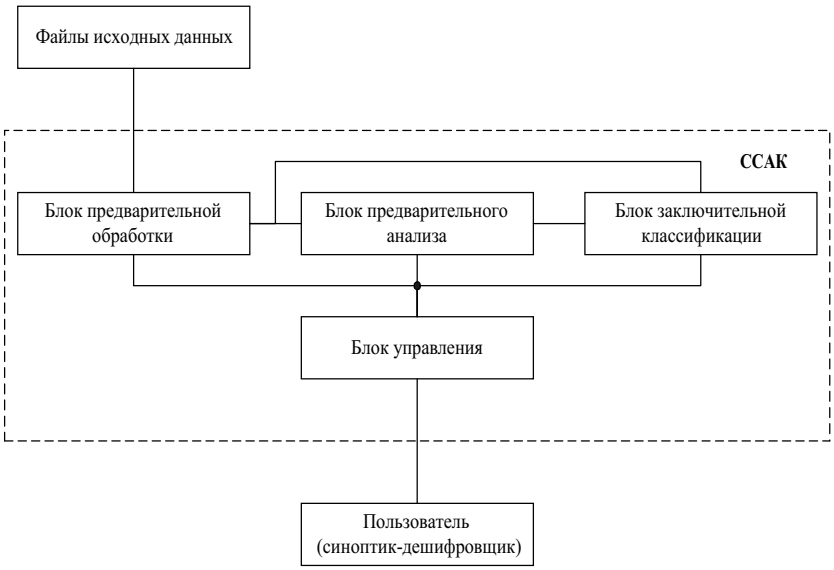


Рис.3

Блок предварительной обработки отвечает за выбор необходимых файлов исходных данных и их преобразование к формату системы (массива многомерных наблюдений). Кроме того, блок управляет предоставлением исходных данных другим блокам системы, осуществляющим их анализ, и дает возможность просмотра данных пользователем в удобном виде. При больших объемах исходных дан-

ных генерируется подвыборка для ускорения процесса предварительного анализа.

Блок предварительного анализа осуществляет предварительный анализ. Предварительный анализ подразумевает поиск оптимальных оценок числа классов и параметров СНР. Поиск предварительных оценок для числа классов, а также начальных значений параметров смеси для алгоритма Дэя-Шлезингера может проводиться по полной выборке при небольшом количестве наблюдений либо по предварительно сгенерированной подвыборке (при большом количестве наблюдений). Алгоритм Дэя-Шлезингера и проверка гипотез осуществляется по полной выборке.

Входные данные для *блока заключительной классификации* представляют собой исходное множество многомерных наблюдений, которое подлежит классификации, а выходные - оптимальные оценки числа классов и параметров СНР. Данный блок необходим, когда имеет место большой объем выборки, и предварительному анализу подлежит лишь малая часть. Результатом является соотнесение каждого наблюдения конкретному классу.

4. Распознавание информации, получаемой с метеорологических искусственных спутников Земли

Началом систематического обзора поверхности Земли из космоса можно считать запуск 1 апреля 1960 г. американского метеорологического спутника Tiros-1. Первый отечественный метеорологический спутник ″Космос-122″ был выведен на орбиту 25 июня 1966 г.

По мере совершенствования космической техники область применения данных дистанционного зондирования Земли (ДЗЗ) из космоса постоянно расширяется. Особая роль отводится использованию подобной информации в геоинформационных системах (ГИС), где результаты дистанционного зондирования поверхности Земли являются регулярно обновляемым источником данных для формирования природно-ресурсных кадастров и расчетов прогнозов погоды.

Наиболее ценная компонента наблюдений за состоянием атмосферы из космоса - возможность изучения структуры и динамики различных ансамблей облаков с помощью

МИСЗ. Во время наблюдений за облачностью необходимо выделять из общего потока измерений данные о различных типах облаков, например, перистых, высококучевых, слоисто-кучевых, кучевых и кучево-дождевых облаков; о структуре облачных систем средних размеров: полосовой, ячеистой и спиралевидной, а также облачных систем крупных размеров: фронтальной облачности, облачных спиралей, циклонов, облачности струйных течений.

Для повышения точности и достижения высокой скорости обработки данных о состоянии атмосферы, получаемых с метеорологических искусственных спутников Земли, используются различные математические методы. Так как объемы данных обычно слишком велики для применения этих методов ″вручную″, то назрела необходимость в создании автоматизированных систем, которые способны в реальное время решать поставленные задачи.

До недавнего времени создание автоматизированных систем по сбору, обработке и интерпретации информации, получаемой с МИСЗ, из-за больших объемов информации существенно зависело от наличия мощной вычислительной техники, способной в реальном масштабе времени решать

задачи по анализу и классификации получаемой информации. В настоящее время в связи с бурным развитием вычислительной техники решение подобных задач стало возможным даже на персональных ЭВМ. Уже существует множество систем по сбору, хранению и визуализации (на мониторе) количественных данных, получаемых по измерениям различных радиометров, установленных на МИСЗ. Однако задача по автоматизации процесса интерпретации этих данных не решена до сих пор ввиду сложности, множественности процессов, происходящих в атмосфере и на подстилающей поверхности, и неоднозначности перехода от данных измерений МИСЗ к физическим параметрам атмосферы.

Для сбора метеорологической информации используются как геостационарные, так и полярно-орбитальные спутники Земли. Источником информации, которая может быть получена с помощью аппаратуры, установленной на спутнике, является поле излучения земной поверхности и атмосферы.

Использование статистических алгоритмов автоматической классификации получаемой с МИСЗ информации, а также некоторых методов кластер-анализа, в решении

этой задачи имеет ряд преимуществ и позволяет значительно повысить точность распознавания.

Более высокая надежность статистических методов, возможность выявления классов кучевообразной и слоистообразной облачности атмосферных фронтов относительно небольших по площади участков изображения позволяет значительно повысить качество обработки данных. Исходя из сказанного выше, был выбран статистический подход к решению поставленной задачи с использованием методов кластер-анализа.

Результатом разработки явилась автоматизированная система распознавания типов облачности в виде приложения для платформ на основе интерфейса WIN32 API.

4.1. Функциональная схема взаимосвязи разрабатываемой автоматизированной системы распознавания типов облачности и автоматизированной системы сбора, обработки и хранения данных

Целью разработки автоматизированной системы (АС) распознавания типов облаков (РТО) является улучше-

ние качества анализа и интерпретации данных, получаемых с МИСЗ. Для полного и качественного функционирования АС необходимо обеспечить четкое взаимодействие между ней и АС СОХД, которая в при нашем подходе будет являться источником данных для АС РТО. Для обеспечения этих требований большое внимание уделялось вопросам совместимости двух рассматриваемых систем.

Функциональная схема взаимосвязи АС РТО и АС СОХД представлена на рис.4.

Рассмотрим основные элементы данной структуры.

- Метеорологический искусственный спутник Земли - геостационарный или полярно-орбитальный, на котором установлен спектрометр для измерения излучения отражающей поверхности в различных диапазонах спектра.
- Приемная станция данных представляет собой стационарный комплекс аппаратно-программных средств, входящих в подсистему сбора и обработки данных, для приема, обработки и хранения спутниковой информации. Подсистема визуализации данных приемной станции позволяет преобра-

зовывать спутниковую информацию в цифровой вид.

- Интерфейс передачи данных служит для передачи данных между приемной станцией и сервером.
- Сервер служит для хранения спутниковой информации в цифровом виде и обеспечения доступа к ней множества рабочих станций. Сервер выполнен на платформе Windows NT 3.5, поддерживающей интерфейс Win32 API.
- Интерфейс Win32 API - прикладной программный интерфейс (**A**pplication **P**rogramming **I**nterface). Win32 - это набор функций API операционной системы, которые прикладные программы могут использовать при работе.
- Рабочая станция представляет собой платформу Windows 95. Автоматизированная система обработки информации выполнена на рабочей станции в виде приложения. Исходная информация для этого приложения поступает на рабочую станцию с сервера в виде файлов специального формата.

- Синоптик-дешифровщик управляет работой системы АС и по полученным результатам осуществляет дешифровку результатов. Под дешифровкой результатов распознавания типов облачности понимается установка соответ-ствия между распознаваемыми классами и реальными типами облаков.

4.2 Входные и выходные данные АС РТО

Входными данными для разработанной автоматизированной системы являются оцифрованные данные ДЗЗ с любого метеорологического спутника Земли.

Аппаратура ДЗЗ, установленная на спутнике, позволяет получать данные в четырех основных диапазонах: ультрафиолетовом, видимом, инфракрасном и микроволновом, - только в этих областях спектра земная атмосфера прозрачна для электромагнитных волн.

Диапазоны спектра, по которым производится измерения, приведены в табл. 1.

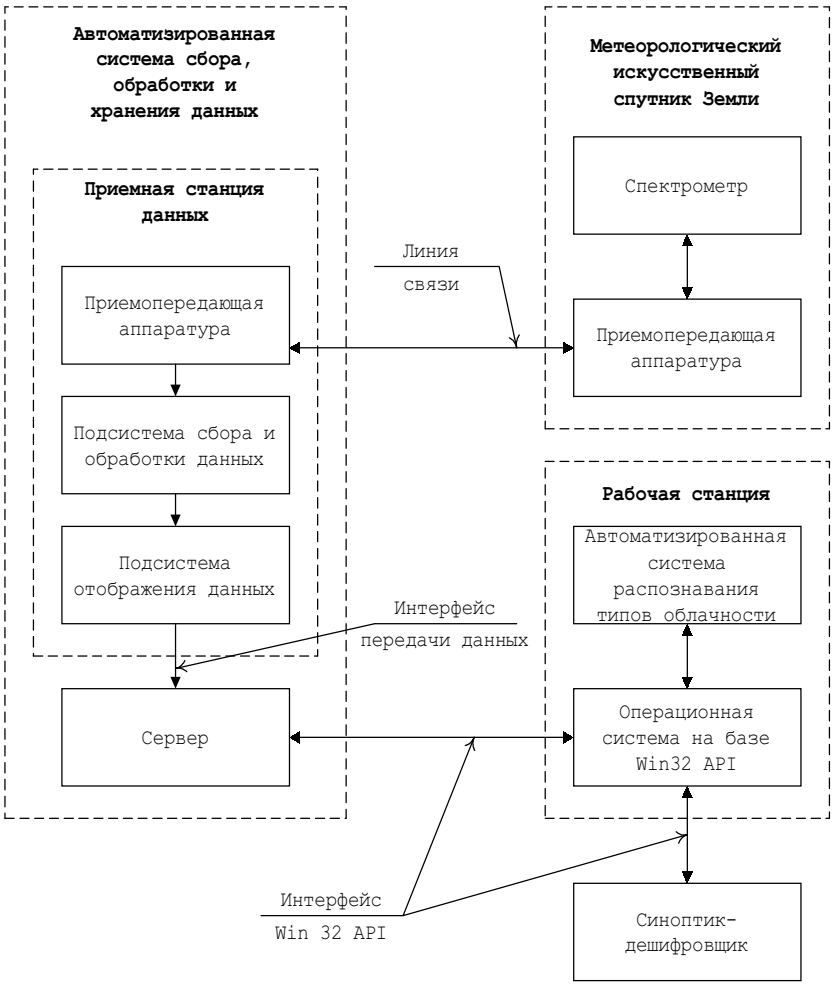

Автоматизированная система сбора, обработки и хранения данных
Приемная станция данных
Приемопередающая аппаратура
Подсистема сбора и обработки данных
Подсистема отображения данных
Сервер
Линия связи
Интерфейс передачи данных
Интерфейс Win 32 API
Метеорологический искусственный спутник Земли
Спектрометр
Приемопередающая аппаратура
Рабочая станция
Автоматизированная система распознавания типов облачности
Операционная система на базе Win32 API
Синоптик-дешифровщик

Рис. 4

Таблица 1

Номер канала	Название диапазона	Диапазон, мкм
1	видимый свет	0,58 - 0,68
2	ближний инфракрасный	0,725 - 1,1
3	средний инфракрасный	3,55 - 3,93
4	дальний инфракрасный	10,3 - 11,3
5	дальний инфракрасный	11,5 - 12,5

Цифровые данные, полученные на приемной станции и переданные по сети в виде двоичных файлов, пересылаются на сервер. Далее эти файлы становятся доступными для дальнейшей обработки на рабочей станции.

Данные в двоичном файле имеют условную размерность $n \times m$. Структура файлов исходных данных представлена на рис.5.

Выходными данными для разработанной автоматизированной системы является цветная карта классов, на которой разным цветом показывается расположение классов на снимке.

Кроме этого имеется возможность просмотреть параметры каждого класса, т.е. его априорную вероятность

(часть от общего числа точек, попавшую в данный класс), а также средние значения по каждому из признаков (измерений по каналам).

Сопоставление результатов классификации с данными наземных наблюдений производится, как показано на рис.6.

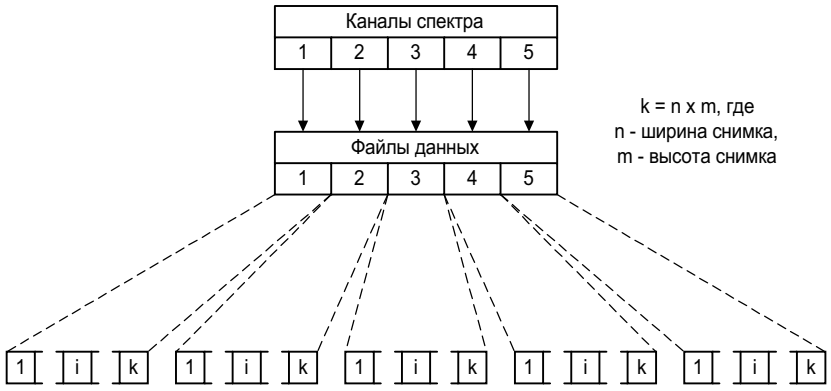


Рис.5

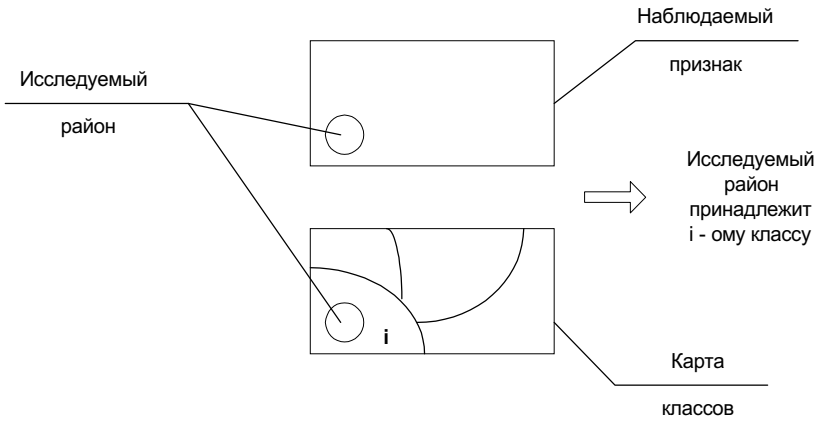


Рис.6

4.3. Функциональная схема автоматизированной системы распознавания типов облачности

Функциональная схема АС РТО приведена на рис.7.

Взаимодействие пользователя с АС РТО осуществляется в полуавтоматическом режиме, при котором оператор в определенных рамках сам управляет процессом классификации данных. Результаты предварительного анализа данных, а также окончательные результаты классификации выводятся на монитор ЭВМ.

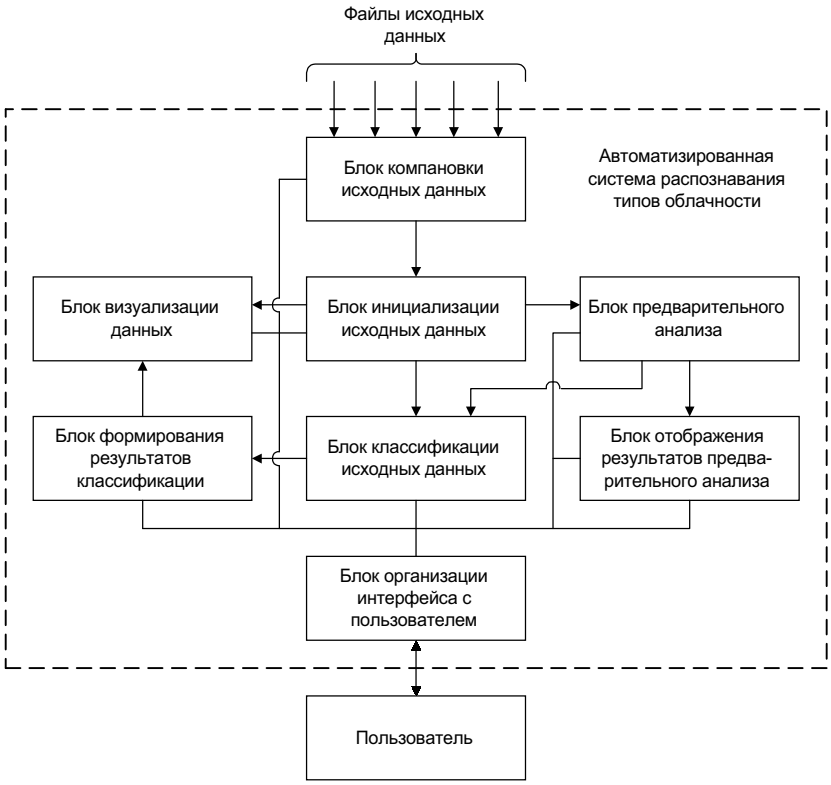


Рис. 7

Коротко рассмотрим назначение блоков структурной схемы.

Блок организации интерфейса с пользователем предназначен для организации взаимодействия между пользователем и АС РТО. С одной стороны, с помощью этого

блока пользователь осуществляет управление работой АС РТО, а с другой стороны, АС РТО осуществляется вывод пользователю результатов своей работы.

Блок компоновки исходных данных предназначен для объединения файлов исходных данных, полученных для одного района, но в различных участках спектра. Файлы объединяются в один проект, название которого определяет пользователь. Предусмотрен вариант ввода данных о размерности исходной информации.

Блок инициализации исходных данных осуществляет проверку соответствия введенной информации, считывает данные и формирует выборку, подлежащую классификации.

Блок предварительного анализа. Назначение и работа данного блока аналогична блоку предварительного анализа структурной схемы ССАК, представленной на рис.3.

Блок отображения результатов предварительного анализа предназначен для вывода пользователю промежуточных результатов, для анализа и корректировки процесса.

Блок классификации исходных данных предназначен для классификации всей выборки с использованием результатов блока предварительного анализа.

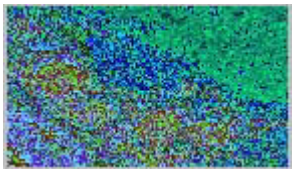
Блок формирования результатов анализа предназначен для преобразования результатов анализа в форму представления, удобную для анализа.

Блок визуализации данных предназначен для визуализации как исходных данных, так и результатов классификации, проведенной по всей выборке. После сопоставления всех данных пользователь проводит интерпретацию результатов классификации и делает соответствующие выводы о типах облачности, наблюдаемых над той или иной областью исследуемого района.

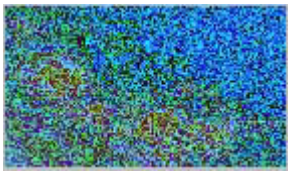
4.4. Экспериментальная часть

В данном разделе описан эксперимент, который проводился по реальным данным, полученным с МИСЗ NOAA. Размер района исследования составил 140x80 условных единиц (пикселей). При классификации исходных данных учитывались все пять признаков (файлов с данными по различным диапазонам спектра). Их изображение приведено на рис.8.

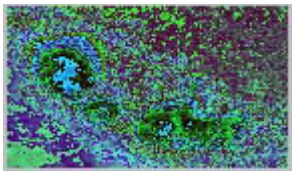
Для предварительного анализа была сгенерирована подвыборка из расчета 1% от общего объема данных.



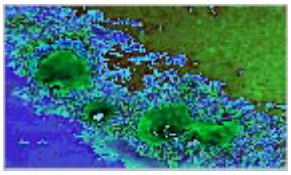
Признак № 1



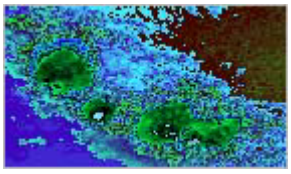
Признак № 2



Признак № 3



Признак № 4



Признак № 5

Рис.8

Гистограмма ОВР строилась по 25 диапазонам и имела вид, представленный на рис.9.

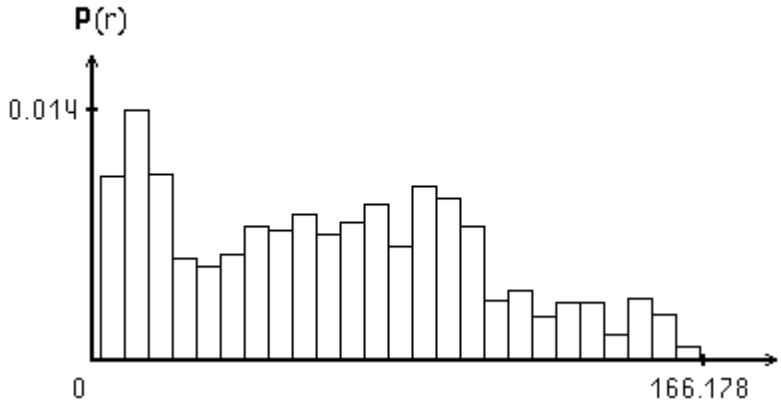


Рис.9

Оценка минимального числа наблюдаемых классов проводилась в автоматическом режиме. Получены минимальное число классов и оценка максимального диаметра класса. Далее производилась генерация гипотез с помощью алгоритма Мак-Кина. На первом и втором этапах расчетов получены разные максимальные диаметры класса. Поэтому потребовалось уточнение оценок параметров по алгоритму Дея-Шлезингера и затем снова проверка гипотез, в результате одна из гипотез была отвергнута. Принята гипотеза $H_{5,}$,

что означает выбор оптимального числа классов, равного 5. Уточненные значения параметров классов приведены в табл. 2. - 6. Следующим шагом распознавания была классификация полной выборки по алгоритму Мак-Кина и по правилу Байеса. На рис. 10, 11 соответственно показаны карты классов. Из рисунков хорошо видны похожие результаты расчетов, первоначальный оценочный и уточненный уже по полной выборке варианты.

Таблица 2

Номер класса	1				
Априорная вероятность	0,07521				
Вектор средних значений Каждого из 5 признаков	43,015	42,506	71,834	52,987	48,109
Корреляционная матрица	1	0,852	0,251	0,113	0,334
	0,852	1	0,648	0,568	0,661
	0,251	0,648	1	0,828	0,929
	0,113	0,568	0,828	1	0,907
	0,334	0,661	0,929	0,907	1

Таблица 3

Номер класса	2				
Априорная вероятность	0,43301				
Вектор средних значений каждого из 5 признаков	16,587	28,121	74,270	65,750	62,424
Корреляционная матрица	1	0,329	0,636	-0,785	-0,797
	0,329	1	0,177	-0,765	-0,262
	0,636	0,177	1	-0,291	-0,367
	-0,785	-0,765	-0,291	1	0,975
	-0,797	-0,262	-0,367	0,975	1

Таблица 4

Номер класса	3				
Априорная вероятность	0.21596				
Вектор средних значений каждого из 5 признаков	64,738	62,373	74,794	44,149	41,911
Корреляционная матрица	1	0,989	0,170	-0,122	-0,054
	0,989	1	0,247	-0,064	0,006
	0,170	0,247	1	0,927	0,949
	-0,122	-0,064	0,927	1	0,0997
	-0,054	0,006	0,949	0,997	1

Таблица 5

Номер класса	4				
Априорная вероятность	0,08491				
Вектор средних значений каждого из 5 признаков	96,667	93,333	38,778	15,778	9,722
Корреляционная матрица	1	0,999	-0,879	-0,851	-0,840
	0,999	1	-0,878	-0,850	-0,841
	-0,879	-0,878	1	0,982	0,973
	-0,851	-0,850	0,982	1	0,998
	-0,840	-0,841	0,973	0,998	1

Таблица 6.

Номер класса	5				
Априорная вероятность	0,19092				
Вектор средних значений Каждого из 5 признаков	98,666	88,577	77,196	50,160	48,505
Корреляционная матрица	1	0,979	-0,628	-0,594	-0,836
	0,979	1	-0,609	-0,898	-0,900
	-0,628	-0,609	1	0,780	0,772
	-0,594	-0,898	0,780	1	0,999
	-0,836	-0,900	0,772	0,999	1

В итоге после идентификации рассчитанных классов с метеорологическими данными выделены следующие классы: 1-й класс – распадающаяся конвективная облачность; 2-й класс – подстилающая поверхность (в данном случае это морская поверхность); 3-й класс – слоистообразная облачность; 4-й класс – распадающаяся конвективная облачность и 5 –й класс – полосы тумана.

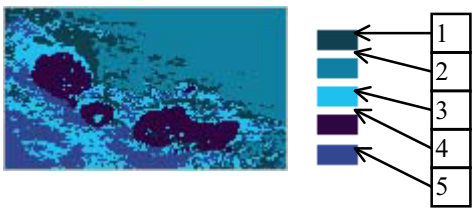


Рис.10

Цветовая шкала классов

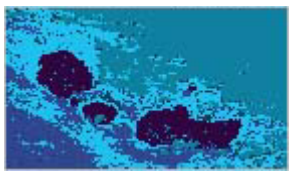


Рис.11

Таким образом, в результате использования системы автоматизированной обработки спутниковой информации получена реальная картина, отражающая распределение различных типов облачного покрова и подстилающей поверхности. Показано, что в случае недостатка времени или

отсутствия высоких требований к точности результатов распознавания типов облаков возможно применение алгоритма Мак-Кина без дополнительных уточнений.

Литература

1. Себестиан Г. С. Процессы принятия решений при распознавании образов. Киев: Техника, 1965.
2. Загоруйко Н. Г. Методы распознавания и их применение. М.: Сов. Радио, 1972.
3. Апраушева Н. Н. Новый подход к обнаружению кластеров. М.: ВЦ РАН, 1993.
4. Волошин Г. Я., Бурлаков. И. А., Косенкова С. Т. Статистические методы решения задач распознавания, основанные на аппромаксимационном подходе. Владивосток, 1992, ч.1.
5. Миленький А. В. Классификация сигналов в условиях неопределенности. М.: Сов. Радио, 1975.
6. Большев Л. Н., Смирнов Н. В. Таблицы математической статистики. М.: Наука, 1983.

7. Дюран Б., Оделл П. Кластерный анализ. М.: Статистика,1975.
8. Андерсон Т. Введение в многомерный статистический анализ. М.: Физматиз, 1963.
9. Day N. E. Estimating the components of a mixture of normal distributions// Biometrika, 1969. V. 56, № 3. P. 463 – 474.
10. Шлезингер М. И. Взаимосвязь обучения и самообучения в распознавании образов// Кибернетика. 1968. № 2, С. 81 - 88.
11. Апраушева Н. Н. Алгоритм расщепления смеси нормальных классов. Программы и алгоритмы. М.: ЦЭМИ АН СССР, 1976.
12. Апраушева Н. Н. Исследование одного алгоритма расщепления смеси нормально распределенных классов // Многомерный статистический анализ в социально-экономических исследованиях. М.: Наука, 1974. С. 135 – 149.
13. Апраушева Н. Н. Об использовании смесей нормальных распределений в распознавании образов. Диссертация на соискание ученой степени кандидата физ.-мат. наук. М., 1981.

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. Определение числа классов в задачах классификации// Изв. АН СССР. Сер. Техн. Кибернетика. 1981, №3. С. 71 – 77.

16.Уилкс. С. Математическая статистика. М.: Наука, 1973.

17. Кондратьев К. Я., Борисенков Е. П., Морозкин А. А. Практическое использование данных метеорологических спутников. Л.: Гидрометеоиздат, 1996.