

ПРИКЛАДНАЯ ТЕОРИЯ АВТОМАТОВ

DOI 10.17223/20710410/10/5

УДК 519.217

МНОГОПАРАМЕТРИЧЕСКАЯ КЛАССИФИКАЦИЯ
АВТОМАТНЫХ МАРКОВСКИХ МОДЕЛЕЙ НА ОСНОВЕ
ГЕНЕРИРУЕМЫХ ИМИ ПОСЛЕДОВАТЕЛЬНОСТЕЙ СОСТОЯНИЙ

А. Р. Нурутдинова, С. В. Шалагин

*Казанский государственный технический университет им. А. Н. Туполева,
г. Казань, Россия***E-mail:** sshalagin@mail.ru, Alsu124@mail.ru

Предложена методика многопараметрической классификации автоматных марковских моделей (АММ) на основе генерируемых ими последовательностей состояний при использовании метода дискриминантного анализа. При этом АММ заданы стохастическими матрицами, принадлежащими к априори заданным подклассам. Определено множество классифицирующих признаков (МКП), позволяющее разделить АММ, заданные матрицами из различных подклассов. Предложена методика, позволяющая определить минимальную длину последовательностей, требуемых для вычисления МКП с заданной доверительной вероятностью.

Ключевые слова: цепи Маркова, эргодические стохастические матрицы, идентификация, автономный вероятностный автомат, автоматная марковская модель, статистическая обработка данных, дискриминантный анализ, линейные дискриминантные функции, информативность.

Введение

Аппарат теории цепей Маркова (ЦМ) служит для моделирования поведения технических устройств, а также объектов из области экономики, социологии, образования и т. п. как в теоретическом, так и в прикладном аспектах [1–3]. В частности, автоматные марковские модели (АММ), определённые на основе аппарата теории ЦМ, позволяют описывать широкий класс устройств [4, 5].

Задача анализа ЦМ имеет важное прикладное значение, связанное с исследованием частотного распределения символов текстов по m -граммам, $m = 2, 3, \dots$ [6, 7].

Современные информационные технологии статистической обработки данных, такие, как пакет прикладных программ Statistica 6.0 [8], позволяют ставить и решать задачи анализа АММ [9, 10]. При этом особая роль отводится отработке методики решения поставленных задач, а также исследованию множества классифицирующих признаков.

В [9, 10] решены задачи классификации и идентификации эргодических стохастических матриц (ЭСМ) по функционалам, вычисленным на их базе. Что касается решения данных задач для реализаций ЦМ, то в [11] определены подходы, связанные с определением биграмм, на основе которых вычисляются определенные функционалы [11–13]. Подобная схема характеризуется тем, что биграммы вычисляются с некоторой погрешностью относительно ЭСМ [11]. Данное обстоятельство способствует снижению точности анализа на основе указанных функционалов. Выходом из создавшейся ситуации

служит создание интегрированных признаков, которые вычисляются непосредственно по реализации ЦМ и являются информативными в плане дискриминации марковских последовательностей, вычисленных на основе определенных подклассов ЭСМ.

В работе на основе результатов, представленных в [14–16], решена задача многопараметрического анализа последовательностей ЦМ, полученных на базе АММ, заданных ЭСМ. Для множества классифицирующих признаков (МКП) анализируется возможность выделения априори заданных подклассов АММ в зависимости от подклассов, к которым принадлежат ЭСМ, определяющие АММ. Анализ проведен при допущении, что имеет место сходимость по вероятности элементов МКП, вычисленных эмпирически на основе ЦМ конечной длины N , к соответствующим элементам МКП, вычисленным теоретически на базе ЭСМ. Задача решается методом дискриминантного анализа (ДА).

1. Объекты классификации

Определение 1. Простая конечная однородная ЦМ задана в виде [1]

$$(S, P, \pi_0), \quad (1)$$

где $S = \{s_0, s_1, \dots, s_{n-1}\}$ — множество состояний ЦМ; P — эргодическая [1] стохастическая матрица вида $P = (p_{ij})$, $i, j = \overline{0, n-1}$; π_0 — n -мерный вектор начального распределения вероятности появления каждого состояния ЦМ.

Определение 2. Автономным вероятностным автоматом будем называть систему [4]

$$(\bar{U}, S, P, \delta(u, s)), \quad (2)$$

где $\bar{U}$ — дискретная случайная величина [4], принимающая на входе значения $u_0, u_1, \dots, u_{l-1}$ с распределением $\bar{p} = (p_0, p_1, \dots, p_{l-1})$; l — размер имплицитующего вектора [4]; $\sum_{i=0}^{l-1} p_i = 1$, $0 \leq p_i \leq 1$; $\delta(u, s)$ — функция переходов, ставящая в соответствие

паре (u, s) однозначно новое состояние $s' \in S$. При этом $P = \sum_{i=0}^{l-1} p_i A_i$, где A_i — простая матрица (по терминологии [5]). Элементы A_i , обозначенные через $a_{kj}(i)$, $k, j = \overline{0, n-1}$, определены из следующего соотношения:

$$a_{kj}(i) = \begin{cases} 1, & \text{если } \delta(u_i, s_k) = s_j, \\ 0, & \text{если } \delta(u_i, s_k) \neq s_j. \end{cases}$$

Определение 3. Цепь Маркова (1) является эргодической, если она состоит из одного эргодического класса, т. е. можно из каждого состояния перейти в любое другое [1].

Определение 4. Стохастическая матрица P , задающая эргодическую цепь Маркова, называется стохастической эргодической матрицей.

Модель устройства, заданная согласно (1) и (2), определена как АММ. На АММ накладывается следующее ограничение: значения положительных элементов ЭСМ кратны определённой величине D , которую будем называть дискретностью элементов ЭСМ. АММ, заданную на основе ЭСМ P , обозначим АММ(P).

Исследуются марковские последовательности, которые являются реализациями простых конечных однородных ЦМ. Они определяются АММ, заданными на основе различных подклассов эргодических стохастических матриц: квазиправильных (Т) и

блочно-сообщающихся (Б). Матрицы класса Т подразделяются на верхние (T_v) и нижние (T_n). ЭСМ подкласса Б делятся на блочные правые (B_p) и блочные левые (B_l). Положительными элементами матриц заданных подклассов будут следующие:

- 1) для подкласса T_n — элементы p_{ij} , где $i = \overline{0, n-1}$, $j = \overline{0, i}$, и элемент $p_{0(n-1)}$;
- 2) для подкласса T_v — элементы p_{ij} , где $i = \overline{0, n-1}$, $j = \overline{i, n-1}$, и элемент $p_{(n-1)0}$;
- 3) для подкласса B_p — элементы p_{ij} , где $i, j = \overline{0, k-1}$, $k = [n/2]$, и p_{ij} , где $i, j = \overline{k, n-1}$, а также элементы $p_{(k-1)k}$ и $p_{k(k-1)}$;
- 4) для подкласса B_l — элементы p_{ij} , где $i = \overline{0, k-1}$, $j = \overline{k, n-1}$, $k = [n/2]$, и p_{ij} , где $i = \overline{k, n-1}$, $j = \overline{0, k-1}$.

«Угловые» элементы квазитреугольных матриц добавлены для сохранения у данных ЭСМ свойства эргодичности.

Возникает вопрос — последовательность состояний какой длины требуется сгенерировать с целью классификации АММ, определенной на основе ЭСМ, с заданной доверительной вероятностью?

2. Множество классифицирующих признаков

В работе решена задача распознавания АММ(P) в зависимости от того, к какому подклассу отнесены ЭСМ P , которые определяют порождаемые ими ЦМ.

Задача решена на основе заданного набора признаков $\tilde{f}_k$, $k = \overline{1-n, n-1}$, которые характеризуют частоту перехода ЦМ из состояния s_i в состояние s_j ($s_i, s_j \in S$, $i, j = \overline{0, n-1}$) для случая, когда $i - j = k$. Вычисление $\tilde{f}_k$ эквивалентно подсчету частот по биграмме, построенной на основе реализации ЦМ [7].

На основе реализации ЦМ длины N получим последовательность Y длины $(N-1)$, такую, что $y_z = k$ для $i - j = k$ ($y_z \in Y$, $z = \overline{1, N-1}$), если на z -м шаге ЦМ переходит из состояния s_i в состояние s_j .

Частота перехода ЦМ из состояния s_i в состояние s_j для случая, когда $i - j = k$, рассчитывается по формуле

$$\tilde{f}_k = \frac{1}{N-1} \sum_{z=1}^{N-1} \hat{y}_z, \quad \hat{y}_z = \begin{cases} 1, & \text{если } y_z = k, \\ 0, & \text{если } y_z \neq k, \end{cases} \quad k = \overline{1-n, n-1}. \quad (3)$$

Вероятности перехода ЦМ f_k ($k = \overline{1-n, n-1}$) из состояния s_i в состояние s_j для случая, когда $i - j = k$, являющиеся теоретическими оценками для величин $\tilde{f}_k$, определяются на базе ЭСМ(P), определяющих АММ(P), и рассчитываются по формуле

$$f_k = \frac{1}{n} \sum_{i=1}^{n-|k|} m(i, k), \quad (4)$$

где $m(i, k) = p_{i(i-k)}$ для $k = \overline{1-n, 0}$ и $m(i, k) = p_{(i+k)i}$ для $k = \overline{1, n-1}$. Величина (3) асимптотически стремится к (4) при бесконечном возрастании длины ЦМ. При решении практических задач величина N является конечной. Отклонение $\varepsilon_k = |\tilde{f}_k - f_k|$ ($k = \overline{1-n, n-1}$) рассчитывается исходя из того, что $\tilde{f}_k$ — частота появления события — перехода ЦМ из состояния s_i в состояние s_j для случая $i - j = k$, $i, j = \overline{0, n-1}$. В свою очередь, f_k , $k = \overline{1-n, n-1}$, есть вероятность данного события. Примем в качестве допущения, что значение ε_k распределено по закону Гаусса [17]. Данным допущением можно пользоваться уже при $N \geq 5$ [17]. При условии, что величины $\tilde{f}_k$ вычисляются на основе последовательности Y длины $(N-1)$, данное отклонение вы-

числяется по формуле [17]

$$\varepsilon_k = |f_k - \tilde{f}_k| = t_\alpha \sqrt{\frac{\tilde{f}_k(1 - \tilde{f}_k)}{N - 1}}, \quad k = \overline{1 - n, n - 1}, \quad (5)$$

где t_α — величина, определяющая для нормального закона число среднеквадратических отклонений, которое нужно отложить относительно центра рассеивания, чтобы вероятность попадания в полученный интервал была равна α . При этом для $\tilde{f}_k$ выполняется условие [17]

$$\tilde{f}_k \in \left[\frac{5}{N - 1}, \frac{N - 6}{N - 1} \right]. \quad (6)$$

Замечание 1. Условие (6) может быть записано, согласно (3), в виде $\sum_{z=1}^{N-1} \hat{y}_z \in [5, N - 6]$, $\hat{y}_z \in Y$, $z = \overline{1, N - 1}$.

В случае невыполнения условия (6) $\tilde{f}_k$ рассматривается как число появления заданного события в некотором числе опытов, которое распределено по биномиальному закону. В этом случае ε_k вычисляется согласно

$$\varepsilon_k = \max(|\tilde{f}_k - p_{1k}|, |p_{2k} - \tilde{f}_k|), \quad (7)$$

где p_{1k} и p_{2k} — верхняя и нижняя границы доверительного интервала, рассчитанные при заданном значении доверительной вероятности α :

$$\sum_{m=l}^{N-1} C_{N-1}^m p_{1k}^m (1 - p_{1k})^{N-m-1} = \frac{\alpha}{2}; \quad \sum_{m=0}^l C_{N-1}^m p_{2k}^m (1 - p_{2k})^{N-m-1} = \frac{\alpha}{2},$$

где l — количество переходов ЦМ из состояния s_i в состояние s_j для случая, когда $i - j = k$, $k = \overline{1 - n, n - 1}$ [16]. Перепишем (5) в виде

$$N = \frac{\tilde{f}_k(1 - \tilde{f}_k)}{|f - \tilde{f}_k|^2} t_\alpha^2 + 1 \leq \max_{\tilde{f}_k} \left(\frac{\tilde{f}_k(1 - \tilde{f}_k)}{\varepsilon^2} t_\alpha^2 \right) = \frac{1}{4\varepsilon^2} t_\alpha^2, \quad k = \overline{1 - n, n - 1}. \quad (8)$$

Тогда имеет место

Утверждение 1. Верхняя оценка длины последовательности N , на основе которой вычисляются значения признаков $\tilde{f}$, достаточной для идентификации с заданной доверительной вероятностью α АММ(P_1) и АММ(P_2), где P_1 и P_2 принадлежат различным подклассам ЭСМ, определяется согласно (8).

Согласно утверждению 1, определив признаки, имеющие максимальные из минимальных отклонений,

$$\varepsilon = \max_{k=\overline{1-n, n-1}} (\min f_k(P_1) - \min f_k(P_2)), \quad (9)$$

где $f_k(P_1)$, $f_k(P_2)$ — значения k -го признака, вычисленного для ЭСМ P_1 и P_2 соответственно, возможно определение N — минимальной длины ЦМ, достаточной для классификации АММ, определенной на основе ЭСМ, с заданной доверительной вероятностью α , согласно (8).

На основе утверждения 1 предложена методика дискриминации АММ(P) на подклассы, определяемые на основе ЭСМ P .

3. Методика решения задачи многопараметрической классификации (дискриминации)

Пусть элементы $p_{ij} > 0$ ЭСМ P кратны некоторому числу D и $p_{ij} \in [D, 1 - D]$, $i, j = \overline{0, n-1}$. На основе данного определения по формуле (4) вычисляется диапазон значений для величин f_k , $k = \overline{1-n, n-1}$.

Для многопараметрической дискриминации АММ(P), $P \in \{T_n, T_v, B_n, B_l\}$, на базе порождаемых ею последовательностей — реализаций ЦМ в зависимости от того, к какому подклассу принадлежит ЭСМ P , определяющая АММ(P), требуется решить следующие задачи.

Задача 1. Найти область определения каждого признака вида (4).

Задача 2. Выявить, какие признаки $\tilde{f}_k$ являются значимыми при выделении того или иного подкласса ЭСМ P , задающих АММ(P) (определить рабочее множество).

Задача 3. Для заданной ЦМ вычислить элементы рабочего множества $\tilde{f}_k$ вида (3) с априори указанной точностью. При этом определяется минимальная длина N последовательности, снимаемой с выхода АММ, которая требуется для достижения заданной точности ε_k , вычисленной согласно (5).

Задача 4. Разделить множество реализаций ЦМ, снимаемых с выхода АММ(P), на группы, определяемые подклассами T_n, T_v, B_n, B_l , используя метод дискриминантного анализа.

Последовательное решение задач 1 – 4 определяет методику многопараметрической дискриминации ЦМ на базе признаков вида (3).

Решение задачи 1 основано на следующем утверждении.

Утверждение 2. Для признаков f_k , $k = \overline{1-n, n-1}$, вычисляемых на основе множества V_k положительных элементов матрицы P согласно (4), их минимальные значения, $\min f_k$, достигаются при условии, что $h = D$ для всех $h \in V_k$, а максимальные значения, $\max f_k$, — при условии, что $h = 1 - D$ для всех $h \in V_k$.

Далее в формате $(\min f_k, \max f_k)$ приведены минимальные и максимальные значения признаков для рассматриваемых подклассов ЭСМ.

Минимальные и максимальные значения признаков для T_n равны

$$\begin{aligned} f_{1-n} &= (D, 1 - D), \quad f_k = (0, 0), \quad k = \overline{2-n, -1}, \\ f_0 &= (D, 1 - (2 + n(n-1))0,5Dn^{-1}), \\ f_k &= ((1 - kn^{-1})D, 1 - (n + k - 1)(n - k)0,5Dn^{-1}), \quad k = \overline{1, n-1}. \end{aligned} \quad (10)$$

Минимальные и максимальные значения признаков для T_v :

$$\begin{aligned} f_k &= ((1 - kn^{-1})D, 1 - (n + k - 1)(n - k)0,5Dn^{-1}), \quad k = \overline{1-n, -1}, \\ f_0 &= (D, 1 - (2 + n(n-1))0,5Dn^{-1}), \\ f_k &= (0, 0), \quad k = \overline{1, n-2}, \quad f_{n-1} = (D, 1 - D). \end{aligned} \quad (11)$$

Найдем минимальные и максимальные значения признаков для B_n , приняв $w = 0,5n$ и $v = [w] + 1$, где $n \geq 4$:

$$\begin{aligned} f_k &= (0, 0) \text{ при } k = \overline{1-n, -v} \text{ и } k = \overline{v, n-1}, \\ f_0 &= (D, 1 - 2Dn^{-1}((v-1)^2 - 1)). \end{aligned} \quad (12)$$

При нечетном n :

$$\begin{aligned} f_{-1} &= \left((n-1)Dn^{-1}, 1 - 2Dn^{-1} \left((v-1,25)^2 + \frac{23}{16} \right) \right), \\ f_1 &= ((n-1)Dn^{-1}, 1 - 2Dn^{-1}(v-1,5)^2 + 0,25), \\ f_k &= ((1+2k)Dn^{-1}, 1 - (v-1)(v+k) + D(v-2)(n+k-v)), \quad k = \overline{1-v, -2}. \end{aligned}$$

При четном n :

$$\begin{aligned} f_{-1} &= f_1 = ((n-1)Dn^{-1}, 1 - 2Dn^{-1}((w-1,5)^2 - 0,75)), \\ f_k &= ((1-2|k|)Dn^{-1}, 1 - Dn^{-1}(2w^2 - n(1+|k|) + 2|k| + 1)), \quad k = \overline{-w, -2}, \quad k = \overline{2, w}. \end{aligned}$$

Минимальные и максимальные значения признаков для B_n : $f_0 = (0, 0)$; при нечетном n :

$$\begin{aligned} f_k &= (kDn^{-1}, 1 - k[w]D), \quad k = \overline{1, [w]}, \\ f_k &= (-kDn^{-1}, 1 - k[w]D), \quad k = \overline{-[w], -1}, \\ f_k &= ([w]Dn^{-1}, 1 - 2[w]D), \quad k = [w] + 1, \\ f_k &= ((1-k)Dn^{-1}, 1 - (n-k)[w]D), \quad k = \overline{1-n, -[w]-2}, \quad k = \overline{[w]+2, n-1}; \end{aligned} \tag{13}$$

при четном n :

$$\begin{aligned} f_k &= (|k|Dn^{-1}, 1 + |k|(1-w)D), \quad k = \overline{-w+1, -1}, \quad k = \overline{1, w-1}, \\ f_k &= ((1-|k|)Dn^{-1}, 1 - (n-|k|)(w-1)D), \quad k = \overline{1-n, -w}, \quad k = \overline{w, n-1}. \end{aligned}$$

Результаты, полученные на основе (10)–(13), приведены в табл. 1

Решение задачи 2 требует введения следующих обозначений. Пусть Q — множество, обозначающее подкласс ЭСМ, принадлежащих одному и только одному из заданных подклассов — T_n , T_w , B_n или B_n ; J_Q — множество индексов признаков f_k , для которых $f_k(Q) = 0$.

Утверждение 3. Признаки f_k вида (4), объединенные во множество R_Q , являются значимыми для дискриминации АММ($P \in Q$) и включаются в рабочее множество, если диапазоны их значений не пересекаются: $f_k(Q) = 0$, а $\min f_k(\neg Q) > 0$, $k \in J_Q$.

Утверждение 4. Объединение множеств $R_{Q^{(1)}}$ и $R_{Q^{(2)}}$ признаков f_k вида (4), значимых для дискриминации АММ($P_1 \in Q^{(1)}$) и АММ($P_2 \in Q^{(2)}$), $Q^{(1)} \cap Q^{(2)} = \emptyset$, есть искомое рабочее множество.

Согласно (10)–(13), признаки f_k равны 0 при следующих условиях:

$$\begin{aligned} f_k(T_n) &= 0, \quad k \in J_{T_n} = \{2-n, 3-n, \dots, -1\}, \\ f_k(T_w) &= 0, \quad k \in J_{T_w} = \{1, 2, \dots, n-2\}, \\ f_k(B_n) &= 0, \quad k \in J_{B_n} = J_{B_{n1}} \cup J_{B_{n2}} = \{1-n, 2-n, \dots, -v\} \cup \{v, v+1, \dots, n-1\}, \\ f_0(B_n) &= 0. \end{aligned}$$

Т а б л и ц а 1

Характеристики признаков $f_k, k = \overline{1-n, n-1}$

Класс P	$\min f_k$	$\max f_k$
T_n	$D, k = 1 - n$	$1 - D, k = 1 - n$
	$D, k = 0$	$1 - (2 + n(n-1))0,5Dn^{-1},$ $k = 0$
	$(1 - kn^{-1})D,$ $k = \overline{1, n-1}$	$1 - (n + k - 1)(n - k)0,5Dn^{-1},$ $k = \overline{1, n-1}$
	$0, k = \overline{2-n, -1}$	$0, k = \overline{2-n, -1}$
T_b	$(1 - kn^{-1})D,$ $k = \overline{1-n, -1}$	$1 - (n + k - 1)(n - k)0,5Dn^{-1},$ $k = \overline{1-n, -1}$
	$D, k = 0$	$1 - (n + k - 1)(n - k)0,5Dn^{-1},$ $k = 0$
	$D, k = n - 1$	$1 - D, k = n - 1$
	$0, k = \overline{1, n-2}$	$0, k = \overline{1, n-2}$
	$n \bmod 2 = 1$	
	$\min f_k$	$\max f_k$
B_n	$kn^{-1}D, k = \overline{1, [w]}$	$1 - k[w]D,$ $k = \overline{1, [w]}, k = \overline{-[w], -1}$
	$[w]Dn^{-1},$ $k = -[w] - 1, k = [w] + 1$	$1 - 2[w]D,$ $k = -[w] - 1, k = [w] + 1$
B_n	$(1 - k)n^{-1}D,$ $k = \overline{1-n, -[w]-2},$ $k = \overline{[w]+2, n-1}$	$1 - (n - k)[w]D,$ $k = \overline{1-n, -[w]-2},$ $k = \overline{[w]+2, n-1}$
	$-kn^{-1}D, k = \overline{-[w], -1}$	
	$n \bmod 2 = 0$	
	$\min f_k $	$\max f_k $
B_n	$ k n^{-1}D,$ $k = \overline{-w+1, -1},$ $k = \overline{1, w-1}$	$1 - k (w-1)D,$ $k = \overline{-w+1, -1},$ $k = \overline{1, w-1}$
B_n	$(1 - k n^{-1}),$ $k = \overline{1-n, -w},$ $k = \overline{w, n-1}$	$1 - (n - k)(w-1),$ $k = \overline{1-n, -w},$ $k = \overline{w, n-1}$

На основе утверждений 3 и 4 предложена методика определения рабочего множества (РМ) для дискриминации АММ(P), $P \in \{T_n, T_b, B_n, B_n\}$, включающая три этапа (по количеству подклассов ЭСМ за вычетом единицы).

Этап 1. Выделяются признаки, которые разделяют последовательности, порождаемые $\text{АММ}(P \in T_n)$, от последовательностей, порождаемых $\text{АММ}(M)$, где $M \in \{T_v, B_n, B_l\}$.

Этап 2. Определяется множество признаков, которые разделяют последовательности, снимаемые с выхода $\text{АММ}(P \in T_v)$, от последовательностей, порождаемых $\text{АММ}(M \in B)$.

Этап 3. Вычисляются признаки, разделяющие реализации цепи Маркова, снимаемые с выхода $\text{АММ}(P \in B_n)$, от реализаций цепей Маркова, порождаемых $\text{АММ}(P \in B_l)$.

При выполнении этапа 1 определяются признаки, выделяющие T_n из множества $\{T_n, T_v, B_n, B_l\}$. Найдем максимальное отклонение для признаков, отделяющих T_n от T_v . Для этого определим следующие значения:

$$A_1 = \max_{k \in J_{T_n}} |\min f_k(T_n) - \min f_k(T_v)| = \max_{k \in J_{T_n}} (\min f_k(T_v)) = (n-1)Dn^{-1} \text{ для } k = -1,$$

$$B_1 = \max_{k \in J_{T_v}} |\min f_k(T_n) - \min f_k(T_v)| = \max_{k \in J_{T_v}} (\min f_k(T_n)) = (n-1)Dn^{-1} \text{ для } k = 1.$$

Максимальное отклонение для признаков, отделяющих B_n от T_n , найдем согласно формуле

$$C_1 = \max_{k \in (J_{T_n} \setminus J_{B_{n1}}) \cup J_{B_{n2}}} |\min f_k(T_n) - \min f_k(B_n)| =$$

$$= \max \left(\max_{k \in (J_{T_n} \setminus J_{B_{n1}})} (\min f_k(B_n)), \max_{k \in J_{B_{n2}}} (\min f_k(T_n)) \right). \quad (14)$$

При этом $J_{T_n} \setminus J_{B_{n1}} = \{-[w], -[w]+1, \dots, -1\}$. На этом интервале $\max_k (\min f_k(B_n)) = (n-1)Dn^{-1}$ при $k = -1$, а для $k \in J_{B_{n2}}$ получим $\max_k (\min f_k(T_n)) = (n-1)Dn^{-1}$ при $k = 1$. В результате $C_1 = (n-1)Dn^{-1}$.

Отклонения для признаков, отделяющих B_l от T_n , вычисляются, согласно (14), как $E_1 = \max_{k \in J_{T_n} \cup \{0\}} |\min f_k(T_n) - \min f_k(B_l)|$. В данном случае $J_{T_n} \cup \{0\} = \{2-n, 3-n, \dots, 0\}$, $\max_{k \in J_{T_n}} (\min f_k(B_l)) = [w]Dn^{-1}$, а $\min f_0(T_n) = D$. В результате $E_1 = D$.

Из этапа 1 следует вывод: для того чтобы выделить T_n из множества $\{T_n, T_v, B_n, B_l\}$, достаточно вычислить значения признаков с максимальным отклонением, равным $\min(A_1, B_1, C_1, E_1) = (n-1)Dn^{-1}$. В РМ войдут признаки $\tilde{f}_k$, для которых f_k удовлетворяет условию: $|\min f_k(T_n) - \min f_k(Q)| \geq (n-1)Dn^{-1}$, $Q \in \{T_n, T_v, B_n, B_l\}$. Определим данное РМ как R_{T_n} .

На этапе 2 выявлены признаки, выявляющие ЭСМ подкласса T_v из множества $B = \{B_n, B_l\}$.

Максимальное отклонение для признаков, отделяющих ЭСМ B_n от T_v , согласно (13), равно $A_2 = \max_{k \in J_{T_v} \setminus J_{B_{n2}}} |\min f_k(T_v) - \min f_k(B_n)| = (n-1)Dn^{-1}$, $k = 1$.

Максимальное отклонение для признаков, отделяющих B_l от T_v , равно, согласно (14), $B_2 = \max_{k \in \{0\} \cup J_{T_v}} |\min f_k(T_v) - \min f_k(B_l)| = D$. При этом $\{0\} \cup J_{T_v} = \{0, 1, \dots, n-2\}$, $\min f_0(T_v) = D$, а $\max_{k \in J_{T_v}} (\min f_k(B_l)) = [w]Dn^{-1}$.

Таким образом, на этапе 2, чтобы отделить T_v от B , достаточно выделить признаки с максимальным отклонением, равным $\min(A_2, B_2) = (n-1)Dn^{-1}$. В РМ, которое определим как R_{T_v} , войдут признаки $\tilde{f}_k$, для которых f_k удовлетворяет условию $|\min f_k(T_v) - \min f_k(B)| \geq (n-1)Dn^{-1}$.

На третьем этапе определены признаки, которые отделяют ЭСМ B_n от B_l . Максимальное отклонение для них, согласно (14), равно $A_3 = \max_{k \in \{0\} \cup J_{B_n}} |\min f_k(B_n) - \min f_k(B_l)|$.

При этом $\{0\} \cup J_{B_n} = \{1 - n, 2 - n, \dots, -v, 0, v, v + 1, \dots, n - 1\}$, $\min f_k(B_n) = D$ при $k = 0$, а $\max_{k \in J_{B_n}} (\min f_k(B_n)) = [w]Dn^{-1}$.

В РМ, которое определено как R_{B_n} , войдут признаки $\tilde{f}_k$, для которых f_k удовлетворяет условию: $|\min f_k(B_n) - \min f_k(B_n)| \geq D$.

Таким образом, согласно данным, полученных на этапах 1 – 3, имеет место

Замечание 2. РМ для дискриминации АММ(P), $P \in \{T_n, T_v, B_n, B_l\}$, на основе генерируемых ими ЦМ равно $R = R_{T_n} \cup R_{T_v} \cup R_{B_n}$. При этом максимально допустимое отклонение ε , вычисленное по формуле (9), равно $(n - 1)Dn^{-1}$.

Решение задачи 3. Согласно (3), вычислим элементы РМ R , которые позволяют дискриминировать на подклассы последовательности, порождаемые АММ(P), в зависимости от того, к какому подклассу принадлежит P . Минимальное количество испытаний, необходимых для выделения подклассов реализаций ЦМ, порождаемых АММ(P), определяется на основе (5) (при выполнении условия (6) для каждого из признаков РМ) согласно неравенству

$$N = \max_{f_k \in R} \left(\tilde{f}_k (1 - \tilde{f}_k) t_\alpha^2 \varepsilon_k^{-2} + 1 \right) \leq \frac{n^2 t_\alpha^2}{4(n - 1)^2 D^2} + 1, \quad (15)$$

где ε_k определяется как максимально допустимое отклонение $\tilde{f}_k \in R$ от f_k , определенное при решении второй задачи — $(n - 1)Dn^{-1}$. Чем выше ε_k , тем меньше минимальное число испытаний, определенное согласно (15). Примем $\alpha = 0,95$, откуда следует, что $t_\alpha = 1,96$ [17], $D = 5 \cdot 10^{-2}$. Таким образом, (15) будет представлено в виде: $N = \max_{f_k \in R} \left(1536,64 \cdot n^2 (n - 1)^{-2} \cdot \tilde{f}_k (1 - \tilde{f}_k) + 1 \right)$. В результате, согласно (15), при $\tilde{f}_k = 0,5$ получим $N \leq 384,16 \cdot n^2 (n - 1)^{-2} + 1$.

Замечание 3. В случае, когда (6) не выполняется для признаков R с множеством индексов J , минимальное число испытаний для $\tilde{f}_k$, $k \in J$, определяется на базе (7) путем решения системы уравнений относительно N_k :

$$\begin{cases} (n - 1)Dn^{-1} \geq \tilde{f}_k - p_{1k}, \\ (n - 1)Dn^{-1} \geq p_{2k} - \tilde{f}_k, \end{cases} \quad \text{или} \quad \begin{cases} p_{1k} \geq \tilde{f}_k - (n - 1)Dn^{-1}, \\ p_{2k} \leq (n - 1)Dn^{-1} + \tilde{f}_k, \end{cases} \quad (16)$$

где p_{1k} и p_{2k} определяются, согласно (7), при $\alpha = 0,95$:

$$\sum_{m=l}^{N_k-1} C_{N_k-1}^m p_{1k}^m (1 - p_{1k})^{N_k-m-1} = 0,475, \quad \sum_{m=0}^l C_{N_k-1}^m p_{2k}^m (1 - p_{2k})^{N_k-m-1} = 0,475.$$

Здесь l — количество переходов ЦМ из состояния s_i в состояние s_j для случая, когда $i - j = k$, $k = \overline{1 - n, n - 1}$.

Утверждение 5. Минимальное количество испытаний, необходимых для выделения при заданном α подклассов ЦМ, реализации которых порождаются АММ(P), определяется как $N = \max(N_k, N_z)$, где N_k , $k \in J$, вычислено на базе (16), а N_z находится, согласно (15), для $z = \overline{1 - n, n - 1}$.

Задачи 2 и 3 решены для ЦМ, заданных ЭСМ $P \in \{T_n, T_v, B_n, B_l\}$, при $n = 5$. Области определения для признаков f_k приведены в табл. 2; в ячейках заданы минимальные и максимальные значения выбранных признаков.

Таблица 2

Области определения для признаков для ЭСМ при $n = 5$

Признак	T_n	T_v	B_n	B_l
f_{-4}	$(D, 1 - D)/5$	$(D, 1 - 4D)/5$	$(0, 0)$	$(D, 1 - D)/5$
f_{-3}	$(0, 0)$	$(2D, 1 - 7D)/5$	$(0, 0)$	$(2D, 1 - 2D)/5$
f_{-2}	$(0, 0)$	$(3D, 1 - 9D)/5$	$(D, 1 - 2D)/5$	$(2D, 1 - 2D)/5$
f_{-1}	$(0, 0)$	$(4D, 1 - 10D)/5$	$(4D, 1 - 9D)/5$	$(D, 1 - D)/5$
f_0	$(5D, 1 - 11D)/5$	$(5D, 1 - 11D)/5$	$(5D, 1 - 10D)/5$	$(0, 0)$
f_1	$(4D, 1 - 10D)/5$	$(0, 0)$	$(4D, 1 - 8D)/5$	$(D, 1 - 2D)/5$
f_2	$(3D, 1 - 9D)/5$	$(0, 0)$	$(D, 1 - 3D)/5$	$(2D, 1 - 4D)/5$
f_3	$(2D, 1 - 7D)/5$	$(0, 0)$	$(0, 0)$	$(2D, 1 - 4D)/5$
f_4	$(D, 1 - 4D)/5$	$(D, 1 - D)/5$	$(0, 0)$	$(D, 1 - 3D)/5$

Результаты этапа 1 даны в табл. 3, где показано, с каким максимальным отклонением допускается вычислять значение того или иного признака. Для выделения T_n из множества подклассов $\{T_n, T_v, B_n, B_l\}$ достаточно вычислить признаки $\tilde{f}_k$, $k = \overline{-1, 1}$, с $\varepsilon_k \leq D$.

Таблица 3

Отклонения для признаков, выделяющих T_n из $\{T_n, T_v, B_n, B_l\}$

Подкласс	$\tilde{f}_{-4}$	$\tilde{f}_{-3}$	$\tilde{f}_{-2}$	$\tilde{f}_{-1}$	$\tilde{f}_0$	$\tilde{f}_1$	$\tilde{f}_2$	$\tilde{f}_3$	$\tilde{f}_4$	Макс. откл.
T_v	—	$0,4D$	$0,6D$	$0,8D$	—	$0,8D$	$0,6D$	$0,4D$	—	$0,8D$
B_n	$0,2D$	—	$0,2D$	$0,8D$	—	—	$0,4D$	$0,4D$	$0,2D$	$0,8D$
B_l	—	$0,4D$	$0,4D$	$0,2D$	D	—	—	—	—	D

Данные, полученные на этапе 2, сведены в табл. 4 — для выделения подкласса T_v из множества подклассов $\{T_v, B_n, B_l\}$ достаточно вычислить признаки $\tilde{f}_k$, $k = \overline{0, 1}$, с максимальным отклонением D .

Таблица 4

Отклонения для признаков, выделяющих T_v из $\{T_v, B_n, B_l\}$

Подкласс	$\tilde{f}_{-4}$	$\tilde{f}_{-3}$	$\tilde{f}_{-2}$	$\tilde{f}_{-1}$	$\tilde{f}_0$	$\tilde{f}_1$	$\tilde{f}_2$	$\tilde{f}_3$	$\tilde{f}_4$	Макс. откл.
B_n	$0,2D$	$0,4D$	$0,4D$	—	—	$0,8D$	$0,2D$	—	$0,2D$	$0,8D$
B_l	—	—	—	—	D	$0,2D$	$0,4D$	$0,4D$	—	D

В табл. 5 приведены результаты выполнения этапа 3, согласно которым достаточно вычислить признак $\tilde{f}_0$ с максимальным отклонением D .

Таблица 5

Отклонения для признаков, разделяющих подклассы B_n и B_l

Подкласс	$\tilde{f}_{-4}$	$\tilde{f}_{-3}$	$\tilde{f}_{-2}$	$\tilde{f}_{-1}$	$\tilde{f}_0$	$\tilde{f}_1$	$\tilde{f}_2$	$\tilde{f}_3$	$\tilde{f}_4$	Макс. откл.
B_l	$0,2D$	$0,4D$	—	—	D	—	—	$0,4D$	$0,2D$	D

На основе данных из табл. 2–4 определяем РМ для дискриминации реализаций ЦМ, порождаемых АММ(P), $P \in \{T_n, T_b, B_n, B_l\}$. При этом $\varepsilon_k = D$, $k = \overline{-1, 1}$.

Согласно формуле (3), вычислим элементы РМ $\tilde{f}_k$, $k = \overline{-1, 1}$, которые позволяют дискриминировать на подклассы последовательности, порождаемые АММ(P), в зависимости от того, к какому подклассу принадлежит P . Минимальное количество испытаний, необходимых для однозначного выделения подклассов ЦМ, порождаемых АММ(P), определяется на основе (5) и (14):

$$N = \max_{\tilde{f}_k \in R} \left(2401 \cdot \tilde{f}_k (1 - \tilde{f}_k) + 1 \right) \leq 602,$$

где ε_k определяется как максимально допустимое отклонение $\tilde{f}_k$ от f_k , $k = \overline{-1, 1}$, определенное при решении задачи 2 при $n = 5$. Чем выше данная величина, тем меньше минимальное число испытаний, определенное согласно (15); если $\alpha = 0,95$, то $t_\alpha = 1,96$; $D = 5 \cdot 10^{-2}$.

4. Получение и интерпретация результатов

Решение задачи 4 предполагает получение последовательностей, снимаемых с выхода АММ(P), где $P \in \{T_n, T_b, B_n, B_l\}$. Длина данных последовательностей — число испытаний ЦМ — определяется на основе (15) при выполнении условия (5) для всех признаков РМ, либо на базе (16) — в противном случае.

Указанные задачи решены для ЦМ, снимаемых с выхода АММ(P). При этом P принадлежит одному из подклассов T_n , T_b , B_n , B_l размерности $n = 5$. Объекты T_n , T_b , B_n , B_l получены с использованием генератора псевдослучайных чисел табличного процессора MS Excel. Дискретность представления положительных элементов ЭСМ заданных подклассов $D = 5 \cdot 10^{-2}$.

Проведено по 5 серий испытаний для АММ(P), каждая из которых определена на основе 20 ЭСМ каждого из подклассов T_n , T_b , B_n , B_l — всего 400 последовательностей. Верхняя граница количества испытаний в каждой серии определена на основе (15) и составляет 602, если для признаков из РМ, используемых при классификации полученных последовательностей, выполняется условие (5). Минимальное количество испытаний N , необходимых для выделения подклассов реализаций ЦМ, сгенерированных на основе АММ(P), не превышает 94 % от максимального значения $N_M = 602$.

Распределение значений величины N для последовательностей, генерируемых АММ, показано в табл. 6.

Таким образом, при проведении серий экспериментов для дискриминации реализаций ЦМ (при $\alpha = 0,95$), сгенерированных АММ(T), в более чем половине случаев требуется провести 509 испытаний, что соответствует 85 % от значения N_M . Для дискриминации с заданной точностью реализаций ЦМ, описываемых матрицами B , требуется в более чем половине случаев провести не менее 473 испытаний. Это составляет 79 % от N_M . Данные факты объяснимы более выраженной структурой ЭСМ B по сравнению с ЭСМ T . Для распознавания реализаций ЦМ, сгенерированных АММ(B),

требуется последовательность в среднем меньшей длины, чем для реализаций ЦМ, сгенерированных АММ(Т).

Т а б л и ц а 6

Распределение значений величины N для последовательностей, генерируемых АММ

Класс	Интервал изменения N	% от числа испытаний	% от значения N_M
T_H	380 – 564	87	63 – 94
T_H	86 – 233	13	14 – 39
T_B	380 – 564	99	63 – 94
B_Π	417 – 564	97	69 – 94
B_Δ	343 – 564	90	57 – 94

На основе признаков, входящих в РМ, проведено разделение полученных последовательностей на основе метода дискриминантного анализа. Результаты дискриминации представлены в табл. 7. Значение функции Фишера (F -критерия) для полученных результатов равно 221,8 при уровне значимости $p < 5 \cdot 10^{-5}$. Величина λ -статистики Уилкса равна $1,622 \cdot 10^{-2}$. Данные факты свидетельствуют о высоком качестве дискриминации.

Т а б л и ц а 7

Характеристики признаков РМ

Признак	λ -статистика Уилкса	F -критерий (3,291)	Довер. интервал
f_{-1}	0,073929	409,5078	0,718518
f_0	0,074529	413,8994	0,619251
f_1	0,094515	560,1085	0,668676

В табл. 7 приведены для каждой из переменных РМ значения следующих вычисленных параметров: λ -статистика Уилкса, F -статистика для исключения и доверительный интервал для его значения. Значение функции Фишера, вычисленное как соотношение усредненных межгрупповых и внутригрупповых дисперсий [11, 12], будет для данных признаков больше числа 409 [4, 7], а значение λ -статистики Уилкса не превышает 0,1. Это свидетельствует о высоком вкладе в дискриминацию каждой переменной РМ.

Доля верно дискриминированных последовательностей, достигаемая при использовании заданного РМ, составляет 100 %, что подтверждают приведенные в работе предположения.

5. Обсуждение полученных результатов

Экспериментальные данные подтверждают гипотезу относительно способности признаков, характеризующих частоту перехода ЦМ из состояния s_i в состояние s_j при $i - j = k$ ($s_i, s_j \in S$, $i, j = \overline{0, n-1}$), классифицировать последовательности, генерируемые АММ(P). При этом матрицы P ЦМ принадлежат к разным подклассам, которые различаются порядком расположения положительных элементов, а следовательно, и

характером реализаций ЦМ, генерируемых АММ(P). Определена длина последовательностей, достаточная для дискриминации порождающих их АММ(P) на априори заданные группы с заданной доверительной вероятностью.

Особую роль играет методика выбора элементов РМ — признаков, позволяющих дискриминировать генерируемые последовательности. Оценка их области определения при заданной дискретности элементов ЭСМ дает возможность выявить РМ.

Следует отметить, что решение задачи 3 для ЭСМ при $n = 5$ позволяет получить реализации ЦМ, длины которых могут в общем случае быть меньше, чем верхняя граница 602 за счет варьирования значений признаков из РМ и их отклонения от величины 0,5. Максимальное значение длины генерируемой реализации ЦМ для рассматриваемой задачи ограничено сверху, при этом увеличение размерности n ЭСМ способствует снижению данной оценки. Этот факт объясняется тем, что для ЭСМ большой размерности различие между их подклассами легче идентифицировать как за счет большего количества признаков — $(2n - 2)$, так и из-за лучшего проявления принадлежности ЭСМ к определенному подклассу. Напротив, при малых значениях n различия между подклассами ЭСМ стираются. Следствием этого является увеличение верхней оценки для длины генерируемой ЦМ.

Верхняя оценка N , вычисленная на базе (11), будет пропорциональна $\left(\max_{k \in J} \varepsilon_k\right)^{-1}$. Отклонение ε_k , в свою очередь, будет асимптотически стремиться к D при бесконечном возрастании размерности ЭСМ, а для матрицы размера 2×2 оно будет равно 0,5 D . Однако за счет особенностей структуры ЭСМ, выявляемых признаками из РМ, и за счет многомерности пространства признаков реальное значение длины ЦМ может быть уменьшено в среднем на 6–21 % относительно ее верхней оценки.

Заключение

Решена задача многопараметрической классификации АММ(P) на базе порождаемых ими последовательностей, где P принадлежит к априори заданным подклассам. Поставленная задача решается на основе вычисления частотных характеристик, а также выбора наиболее информативных из них.

Определена верхняя оценка для длины генерируемой последовательности, на базе которой с заданной точностью производится дискриминация порождающей ее АММ(P). Данная оценка может быть снижена в зависимости от величины отклонения вероятности перехода ЦМ из одного состояния в другое от значения 0,5 и от максимально допустимого отклонения данной вероятности относительно её теоретического значения. Допустимое отклонение вычисляется на основе областей определения частот перехода ЦМ из одного состояния в другое для ЭСМ, принадлежащих к априори заданным подклассам.

Предложена методика многопараметрической дискриминации реализаций ЦМ, заданных ЭСМ определённых подклассов, на базе заданного множества классифицирующих признаков.

ЛИТЕРАТУРА

1. Кемени Дж., Снелл Дж. Конечные цепи Маркова. М.: Наука, 1970. 272 с.
2. Романовский В. И. Дискретные цепи Маркова. М.; Л.: Гостехиздат, 1949. 436 с.
3. Раскин Л. Г. Анализ стохастических систем и элементы теории оптимального управления. М.: Сов. радио, 1976. 344 с.
4. Бухараев Р. Г. Основы теории вероятностных автоматов. М.: Наука, 1985. 287 с.

5. *Поспелов Д. А.* Вероятностные автоматы. М.: Энергия, 1970. 88 с.
6. *Алферов А. П., Zubov A. Ю., Кузьмин А. С., Черемушкин А. В.* Основы криптографии. М.: Гелиос АРВ, 2002. 480 с.
7. *Friedman W. F., Callimahos D.* Military crypto analyze. Part I. V.Z. Aegean Park Press, Laguna Hills CA, 1985. 356 p.
8. *Боровиков В. П.* Statistica: искусство анализа данных на компьютере. 2-е изд. СПб.: Питер, 2003. 700 с.
9. *Захаров В. М., Нурмеев Н. Н., Салимов Ф. И. и др.* К задаче дискриминантного анализа автоматных марковских моделей // Вестник КГТУ им. А. Н. Туполева. 2001. Т. 1. № 3. С. 37–39.
10. *Захаров В. М., Нурмеев Н. Н., Салимов Ф. И., Шалагин С. В.* Классификация стохастических эргодических матриц методами кластерного и дискриминантного анализа // Исследования по информатике. Казань: Отечество, 2000. С. 91–106.
11. *Ли И., Джадж Д., Зельнер А. М.* Оценивание параметров марковской модели по агрегированным временным рядам. М.: Статистика, 1977. 221 с.
12. *Ланкастер Л.* Теория матриц. М.: Наука, 1982. 272 с.
13. *Хинчин А. Я.* Понятие энтропии в теории вероятностей // Успехи матем. наук. 1953. № 3(55). С. 3–20.
14. *Сабитова А. Р., Шалагин С. В.* Дискриминантный анализ вероятностных моделей марковского типа // Наука. Технологии. Инновации: Материалы Всерос. конф. молодых ученых. Новосибирск: Изд-во НГТУ, 2007. С. 90–92.
15. *Сабитова А. Р., Шалагин С. В.* Многопараметрическая классификация марковских последовательностей // XV Туполевские чтения: Междунар. молодежная науч. конф. Казань: Изд-во КГТУ им. А. Н. Туполева, 2007. С. 78–79.
16. *Нурутдинова А. Р., Шалагин С. В.* Методика идентификации автоматных марковских моделей на основе порождаемых ими последовательностей // Вестник КГТУ им. А. Н. Туполева. 2010. № 1. С. 94–99.
17. *Вентцель Е. С.* Теория вероятностей. IV изд., стер. М.: Наука, 1969. 576 с.