СИСТЕМНЫЙ АНАЛИЗ

И.В. СЕРГИЕНКО, А.М. ГУПАЛ, А.В. ОСТРОВСКИЙ

УДК 519.217.2

ИСПОЛЬЗОВАНИЕ ЕМ-АЛГОРИТМА ДЛЯ КЛАССИФИКАЦИИ ГЕНОВ

Аннотация. Рассмотрены ЕМ-алгоритм для задачи разделения смесей распределений, описанных цепями Маркова, и связанная с ней проблема максимизации взвешенного правдоподобия. Предложены вспомогательные алгоритмы для выбора начального приближения и оптимального числа компонентов смеси, а также метод аппроксимации смеси распределений на основе известных данных с помощью метода опорных векторов. Полученные результаты применены к задаче классификации фрагментов генов.

Ключевые слова: цепь Маркова, классификация, ген, биоинформатика, нуклеотид, экзон, интрон, правдоподобие.

ВВЕДЕНИЕ

Проблема определения функциональных фрагментов генов с использованием методов машинного обучения в настоящее время остается одной из основных задач биоинформатики. Наиболее распространены методы ее решения на основе обобщенных моделей Маркова со скрытыми переменными [1], имеющие определенные допущения, затрудняющие применение моделей для других биологических видов. Как показано в [2], для нахождения фрагментов генов могут использоваться вероятностные модели, сочетающие обыкновенные скрытые модели Маркова и цепи Маркова высоких порядков. Для повышения качества классификации при рассмотрении геномов высших организмов эффективно применение композиций моделей с эксклюзивной компетентностью составляющих [3]. При этом разбиение множества генов на области компетентности проводится с помощью бинарных предикатов простого вида, зависящих от концентраций в генах отдельных нуклеотидов или их сочетаний; таким образом, генерируемое разбиение непосредственно не связано с выбранной вероятностной моделью.

В данной статье рассмотрен альтернативный подход: разбиение строится путем непосредственной максимизации правдоподобия для множества прецедентов, что позволяет точнее выявлять границы компетентности составляющих композиции. При этом применен итеративный алгоритм ожидания — максимизации (ЕМ-алгоритм), широко используемый в задачах машинного обучения, в частности, для кластеризации данных, распознавания образов и т.д. [4, 5].

ПОСТАНОВКА ЗАДАЧИ

Пусть задан определенный набор строк $X = \{x^i\}_{i=1}^n$ (прецеденты), которые имеют конечную длину и образованы из символов конечного алфавита Q:

$$\forall i = 1, \ldots, n \quad x^i \in Q^* \equiv \bigcup_{s=1}^{\infty} Q^s;$$

другими словами, $X \subset Q^*$.

Пусть плотность распределения вероятности на множестве R^* имеет вид смеси k распределений $M_1,\ M_2, \ldots, M_k$:

$$\forall x \in R^*\ \ P(x) = \sum_{j=1}^{k} w_j\, p(x \mid M_j),\ \ w_j \geq 0,\ \ \sum_{j=1}^{k} w_j = 1, \tag{1}$$

где $p(x \mid M_j)$ — функция правдоподобия для j-го компонента смеси, w_j — ее априорная вероятность. Функции правдоподобия для всех компонентов имеют вид, соответствующий цепям Маркова l-го порядка

$$p(x \mid M_j) = \varphi_j(|x|)\pi_j(x_1 \ldots x_l) \cdot \prod_{s=l+1}^{|x|} p_j(x_s \mid x_{s-l} \ldots x_{s-1}), \tag{2}$$

где приняты следующие обозначения:

- $\varphi_j(d) \equiv P\{|x| = d | M_j\}$ — распределение вероятности длины строк, генерируемых компонентом M_j;
- $\pi_j(u) \equiv P\{x_1 \ldots x_l = u \mid M_j\}$ — распределение начальных вероятностей;
- $p_j(v \mid u) \equiv P\{x_s = v \mid x_{s-l} \ldots x_{s-1} = u, M_j\}$ — распределение переходных вероятностей компонента M_j.

В качестве распределений φ рассмотрим функции, соответствующие усреднению эмпирических данных с шириной окна $2r+1$:

$$\varphi(d; F) = \frac{1}{2r+1} \sum_{s=d-r}^{d+r} F_s,\ \ r < d < d_{\max} - r,\ \ \sum_{s=0}^{d_{\max}} F_s = 1,$$

где граничная длина $d_{\max}$ выбрана так, чтобы вместить все (или почти все) эмпирические данные. При вычислении φ для длины, расположенной вблизи границ отрезка $[0, d_{\max}]$, количество членов, по которым проводится усреднение, уменьшается; таким образом, в наиболее общем случае формула имеет вид

$$\varphi(d; F) = \frac{1}{b-a+1} \sum_{s=a}^{b} F_s,\ \ a = \max\,(0, d-r),\ \ b = \min\,(d_{\max}, d+r). \tag{3}$$

Параметрами функции правдоподобия (3) являются $(d_{\max}+1)$ величин $\{F_d\}_{d=0}^{d_{\max}}$, $|Q|^l$ значений начальных вероятностей $\{\pi(u) \mid u \in Q^l\}$, а также $|Q|^{l+1}$ переходных вероятностей $\{p(v \mid u) \mid u \in Q^l, v \in Q\}$. Совокупность этих параметров для компонентов смеси M_j обозначим θ_j. Задача определения оптимальной смеси распределений с учетом введенных обозначений сводится к нахождению набора величин $\Theta \equiv \{w_1, w_2, \ldots, w_k; \theta_1, \theta_2, \ldots, \theta_k\}$, при котором согласно принципу максимума правдоподобия вероятность порождения множества прецедентов является максимальной

$$P(X) = \prod_{i=1}^{n} P(x^i) = \prod_{i=1}^{n} \sum_{j=1}^{k} w_j\, p(x^i \mid \theta_j) \to \max_{\Theta}. \tag{4}$$

После перехода к логарифму правдоподобия (4) сводится к задаче

$$Q(\Theta; X) = \sum_{i=1}^{n} \ln \sum_{j=1}^{k} w_j\, p(x^i \mid \theta_j) \to \max_{\Theta},\ \ \sum_{j=1}^{k} w_j = 1. \tag{5}$$

Функция Лагранжа для (5) имеет вид

$$L(\Theta; X) = \sum_{i=1}^{n} \ln \sum_{j=1}^{k} w_j\, p(x^i \mid \theta_j) - \lambda \left(\sum_{j=1}^{k} w_j - 1 \right).$$

Приравнивая частные производные этой функции по переменным w_j к нулю, получаем [6]

$$w_j = \frac{1}{n}\sum_{i=1}^{n} g_{ij}, \tag{6}$$

где g_{ij} — апостериорная вероятность генерации i-й строки выборки j-м компонентом смеси; согласно формуле Байеса для полной вероятности

$$g_{ij} = P\{\theta_j \mid x^i\} = \frac{w_j\, p(x^i \mid \theta_j)}{\sum\limits_{s=1}^{k} w_s p(x^i \mid \theta_s)}. \tag{7}$$

Кроме этого, равенства $\forall j = 1, \ldots, k \;\; \partial L / \partial \theta_j = 0$ эквивалентны k независимым задачам максимизации взвешенного правдоподобия

$$\sum_{i=1}^{n} g_{ij} \ln p(x^i \mid \theta_j) \to \max_{\theta_j}, \tag{8}$$

где $G_j = \{g_{ij}\}_{i=1}^{n}$. Таким образом, EM-алгоритм оптимизации выражения (5) сводится к выполнению следующих шагов.

Шаг 1. Определить начальное приближение параметров Θ.

Шаг 2 (E-шаг — ожидание). Вычислить апостериорные вероятности g_{ij} согласно формуле (7), пользуясь имеющимися в данный момент параметрами Θ.

Шаг 3 (M-шаг — максимизация). Решить задачи вида (8) для всех компонентов смеси, получив новые значения $\theta_1, \theta_2, \ldots, \theta_k$; вычислить веса компонентов w_j по формуле (6).

Шаг 4. Если выполнен критерий останова (например, достигнуто максимальное число итераций либо стабилизированы значения параметров), возвратить Θ; иначе перейти к шагу 2.

ЗАДАЧА МАКСИМИЗАЦИИ ВЗВЕШЕННОГО ПРАВДОПОДОБИЯ

Для выполнения шага максимизации EM-алгоритма необходимо построить алгоритм, решающий задачи вида (6)

$$WM(X, g) = \arg\max_{\theta} \sum_{i=1}^{n} g_i \ln p(x^i \mid \theta), \;\; X = (Q^*)^n, \;\; g \in R^n,$$

при ограничениях на вектор параметров $\theta \equiv (F, \pi, p)$ в виде равенств

$$\sum_{d=0}^{d_{\max}} F_d = 1, \;\; \sum_{|u|=l} \pi(u) = 1, \;\; \forall u \in Q^l \;\; \sum_{|v|=1} p(v \mid u) = 1.$$

Здесь и далее при суммировании по строкам полагается, что они состоят из символов алфавита Q.

Функция Лагранжа для поставленной задачи вычисляется следующим образом:

$$L_w(\theta; X, g) = \sum_{i=1}^{n} g_i \left[\ln \varphi(|x^i|) + \ln \pi(x_1^i \ldots x_l^i) + \sum_{s=l+1}^{|x^i|} \ln p(x_s^i \mid x_{s-l}^i \ldots x_{s-1}^i) \right] -$$

$$-\lambda_F \left(\sum_{s=0}^{d_{\max}} F_s - 1 \right) - \lambda_\pi \left(\sum_{|u|=l} \pi(u) - 1 \right) - \sum_{|u|=l} \lambda_u \left(\sum_{|v|=1} p(v \mid u) - 1 \right).$$

В силу сложности непосредственного нахождения параметров $\{F_d\}_{d=0}^{d_{\max}}$ из уравнений $\partial L_w / \partial F_d = 0$, $d = 0, \ldots, d_{\max}$, определим их, исходя из эвристических

соображений, как взвешенные эмпирические вероятности наблюдения строки соответствующей длины

$$F_d = \frac{\sum_{i=1}^{n} g_i [|x^i| = d]}{\sum_{i=1}^{n} g_i}. \tag{9}$$

При таком подходе в случае задачи с одинаковыми весами всех прецедентов, как и ожидается, величины F_d определяются эмпирическими вероятностями:

$$F_d = \frac{|\{|x^i| = d \mid i = 1, \ldots, n\}|}{n}.$$

Члены функции L_w, соответствующие начальным и переходным вероятностям, допускают перегруппировку, позволяющую избавиться от суммирования по прецедентам

$$L_w(\theta; X, g) = \sum_{|u|=l} N_{st}^w(u) \ln \pi(u) + \sum_{|u|=l} \sum_{|v|=1} N^w(uv) \ln p(v \mid u) -$$

$$-\lambda_\pi \left(\sum_{|u|=l} \pi(u) - 1 \right) - \sum_{|u|=l} \lambda_u \left(\sum_{|v|=1} p(v \mid u) - 1 \right) + \text{const}(\pi, p),$$

где введены следующие обозначения:

- $N_{st}^w(u) = \sum_{i=1}^{n} g_i [x_1^i \ldots x_l^i = u]$ — взвешенное количество строк множества X, начинающихся с определенной последовательности;
- $N^w(u) = \sum_{i=1}^{n} \sum_{j=1}^{|x^i|-|u|} g_i [x_j^i \ldots x_{j+|u|-1}^i = u]$ — взвешенное число вхождений последовательности u во все строки из X.

Из равенств

$$\forall u \in Q^l \quad \frac{\partial L_w}{\partial \pi(u)} = \frac{N_{st}^w(u)}{\pi(u)} - \lambda_\pi = 0 \Leftrightarrow N_{st}^w(u) = \lambda_\pi \pi(u)$$

после суммирования по всем строкам длины l следует

$$\lambda_\pi \sum_{|u|=l} \pi(u) = \lambda_\pi = \sum_{|u|=l} N_{st}^w(u) = \sum_{i=1}^{n} g_i,$$

откуда

$$\pi(u) = \frac{N_{st}^w(u)}{\sum_{i=1}^{n} g_i}. \tag{10}$$

Аналогично суммированием равенств

$$\frac{\partial L_w}{\partial p(v \mid u)} = \frac{N^w(uv)}{p(v \mid u)} - \lambda_u = 0$$

по строкам $v \in Q$ получаются оптимальные значения переходных вероятностей

$$p(v \mid u) = \frac{N^w(uv)}{\sum_{|z|=1} N^w(uz)} = \frac{N^w(uv)}{N^w(u)} \tag{11}$$

при условии, что $N^w(u) > 0$. Если $N^w(u) = 0$, что возможно тогда и только тогда, когда последовательность $u \in Q^l$ не входит ни в одну строку из множества X,

нельзя определить оптимальные значения переходных вероятностей вида $p(v \mid u)$. Это обстоятельство обусловливает ограничение порядка цепей Маркова l сверху.

ЭВРИСТИКИ И МОДИФИКАЦИИ

Вычисление вероятностей. Вследствие достаточно сложной структуры компонентов смеси при вычислении вероятностей вида $p(x^i \mid \theta_j)$ возникает проблема потери точности; в связи с этим эффективнее вычислять логарифмы вероятности

$$z_{ij} \equiv \ln\, p(x^i \mid \theta_j) = \ln \varphi_j(|x^i|) + \ln \pi_j(x_1^i \dots x_l^i) + \sum_{s=l+1}^{|x^i|} \ln\, p_j(x_s^i \mid x_{s-l}^i \dots x_{s-1}^i). \quad (12)$$

Апостериорные вероятности g_{ij} при таком подходе вычисляются как

$$g_{ij} = \frac{w_j \exp(z_{ij})}{\sum\limits_{s=1}^{k} w_s \exp(z_{is})} = \frac{w_j}{\sum\limits_{s=1}^{k} w_s \exp(z_{is} - z_{ij})};$$

критерий качества (5) преобразуется к виду

$$Q(\Theta; X) = \sum_{i=1}^{n} \ln \sum_{j=1}^{k} w_j \exp(z_{ij}) = \sum_{i=1}^{n} z_i^{\max} + \sum_{i=1}^{n} \ln\left(\sum_{j=1}^{k} w_j \exp(z_{ij} - z_i^{\max}) \right),$$

где $z_i^{\max} = \max\limits_{j=1,\dots,k} z_{ij}$.

Проблема локальных максимумов. Определенные сложности при вычислении апостериорных вероятностей возникают, когда по крайней мере для одного набора параметров θ_j выполняется равенство $p(x^i \mid \theta_j) = 0$, что означает невозможность порождения строки x^i соответствующей цепью Маркова. В таком случае $g_{ij} = 0$, т.е. прецедент x^i не учитывается при решении задачи максимизации взвешенного правдоподобия на М-шаге ЕМ-алгоритма. При некоторых обстоятельствах этого может быть достаточно, чтобы равенство $g_{ij} = 0$ выполнялось для всех последующих итераций (например, если x^i содержит подстроку длины $l+1$, не встречающуюся в других прецедентах), что приводит к «зацикливанию» алгоритма оптимизации вблизи точки локального максимума. Для решения этой проблемы при вычислении логарифмов правдоподобия (12) вместо функций φ_j, π_j, p_j, вычисляемых по формулам (3), (9)–(11), будут использоваться их приближения, ограниченные снизу:

$$\hat{\varphi}(d) = \max(\varphi(d), \varepsilon_\varphi);\ \hat{\pi}(u) = \max(\pi(u), \varepsilon_\pi);\ \hat{p}(v \mid u) = \max(p(v \mid u), \varepsilon_p),$$

где положительные величины ε_φ, ε_π и ε_p достаточно малы, чтобы не искажать результатов вычислений.

Аналогичная цель — «выбивание» алгоритма из точек локальных максимумов, и для модификации ЕМ-алгоритма, изложенной в [7], согласно которой векторы весов прецедентов G_j, используемые на шаге максимизации, определяются вероятностным образом: $\hat{G}_j = \{[\xi_{ij} \le g_{ij}] \mid i = 1, \dots, n\}$, где ξ_{ij} — независимые случайные величины, равномерно распределенные на единичном отрезке.

Выбор начального приближения и числа компонентов. Скорость сходимости ЕМ-алгоритма сильно зависит от выбора начального набора параметров Θ. В данной работе рассмотрены два хорошо зарекомендовавших себя способа выбора параметров и регулировки количества цепей Маркова в смеси:

- последовательное наращивание числа компонентов;
- последовательное удаление компонентов.

При использовании первого способа начальное приближение состоит из единственного компонента $\Theta = \{w_1; \theta_1\}$, где $w_1 = 1$, $\theta_1 = WM(X, \{1\}_{i=1}^n)$.

Алгоритм построения смеси заключается в циклическом выполнении следующих шагов вплоть до достижения желаемого числа компонентов.

Шаг 1. Найти строки из множества прецедентов X, которые плохо описываются смесью

$$X_b := \left\{ x \in X \mid P(x) \equiv \sum_{j=1}^{k} w_j p(x \mid \theta_j) < \delta \right\},$$

где k — текущее число цепей Маркова в смеси, δ — пороговая вероятность, определяемая как среднее значение вероятности строк выборки

$$\delta := \frac{1}{n} \sum_{j=1}^{n} P(x^i).$$

Шаг 2. Образовать из строк X_b новый компонент

$$w_{k+1} := \frac{|X_b|}{n};\ \theta_{k+1} := WM(X, \{[x^i \in X_b]\}_{i=1}^n);\ \Theta := \Theta \cup \{w_{k+1}; \theta_{k+1}\},$$

откорректировав веса остальных компонентов $\forall j = 1, \ldots, k\ \ w_j := w_j (1 - w_{k+1})$.

Шаг 3. Выполнить ЕМ-алгоритм, используя Θ в качестве начального набора параметров.

Для упрощения расчетов при выделении компонентов на шаге вместо вероятностей $P(x^i)$ можно использовать определенные ранее величины $z_i^{\max}$, которые приблизительно равны $\log P(x^i)$. При этом, как и в случае применения самих вероятностей, отдается предпочтение более длинным строкам; для более равномерного распределения строк по длинам в X_b величины $z_i^{\max}$ следует нормировать на длину $|x^i|$:

$$X_b := \left\{ x^i \mid i = 1, \ldots, n, \frac{z_i^{\max}}{|x^i|} < \frac{1}{n} \sum_{j=1}^{n} \frac{z_j^{\max}}{|x^j|} \right\}.$$

При последовательном удалении компонентов начальное приближение состоит из K цепей Маркова, где K заведомо превышает возможное число компонентов смеси. Параметры для каждой цепи получаются путем обучения на частях множества прецедентов, полученных разбиением X на приблизительно равные части случайным образом. Более формально

$$\forall j = 1, \ldots, K\ \ w_j := \frac{1}{n} \sum_{i=1}^{n} [\xi_i = j],\ \theta_j := WM(X, \{[\xi_i = j]\}_{i=1}^n),$$

где независимые случайные величины ξ_i равномерно распределены на множестве целых чисел $\{1, 2, \ldots, k\}$.

Алгоритм построения оптимальной композиции заключается в циклическом выполнении приведенных далее шагов.

Шаг 1. Удалить из смеси компонент с наименьшим весом, откорректировав веса остальных компонентов

$$t := \arg\min_{j=1,\ldots,k} w_j,\ \Theta := \Theta \setminus \{w_t; \theta_t\},\ \forall j \neq t\ \ w_j := \frac{w_j}{1 - w_t}.$$

Шаг 2. Выполнить ЕМ-алгоритм с начальным приближением Θ.

Критерием останова, как и в предыдущем случае, является достижение требуемого числа компонентов в смеси.

РАСПОЗНАВАНИЕ ФРАГМЕНТОВ ГЕНОВ

Для предсказания функциональной структуры генов можно применять описанный ранее подход. Гены всех живых организмов представляют собой последовательности, состоящие из четырех нуклеотидов: аденина A, цитозина C, гуанина G и тимина T [8]. Основными функциональными фрагментами генов являются чередующиеся один за другим экзоны — участки, кодирующие белки, и интроны — участки, не принимающие участия в синтезе белка. Таким образом, алфавит, позволяющий кодировать информацию о нуклеотидах генов и их распределении по функциональным участкам, состоит из восьми символов: $Q = \{A, C, G, T, a, c, g, t\}$, где заглавными буквами обозначены нуклеотиды, входящие в состав экзонов, строчными — интронные нуклеотиды. Множество прецедентов $X \subset Q^*$ совпадает с геномом — набором всех генов определенного биологического вида.

Задача нахождения структуры гена сводится к определению строки $x \in Q^*$ по известной последовательности нуклеотидов гена $s = \Pr_s(x) \in \{A, C, G, T\}^* \equiv O^*$, где функция $\Pr_s$ определена как $\Pr_s: (A, C, G, T, a, c, g, t) \to (A, C, G, T, A, C, G, T)$.

Для ее решения можно использовать вероятностные модели на основе цепей Маркова [2]

$$x = \arg\max_{z \in Q^*} P(z)[\Pr_s(z) = s], \tag{13}$$

где $P(z)$ рассчитывается по формуле (2). Этот подход эффективен для определения структуры генов простых организмов (например, растений и насекомых); вместе с тем для генов высших организмов (например, млекопитающих) он приводит к неудовлетворительному качеству классификации. Для повышения качества можно использовать композиции алгоритмов с эксклюзивной компетентностью компонентов [3], для которых функция плотности вероятности аналогична (1)

$$P(x) = \sum_{j=1}^{k} w_j P(x \mid \theta_j)[Pr_s(x) \in G_j], \tag{14}$$

где $\{G_j\}_{j=1}^{k}$ — покрытие множества O^*.

Итак, EM-алгоритм можно использовать непосредственно для построения оптимальных композиций, однако при этом возникают определенные затруднения:

- высокий порядок цепей Маркова, необходимый для удовлетворительного качества классификации, приводит к сильному переобучению при выделении компонентов;
- задача максимизации (13) при использовании плотности распределения вероятности (1) существенно сложнее, чем при использовании функции плотности (14), для решения которой, как и в случае одной цепи Маркова, можно использовать алгоритм Витерби.

Пусть $j^*(x)$ — номер наиболее вероятной цепи Маркова из композиции для строки x:

$$j^*(x) = \arg\max_{j=1,\ldots,k} w_j\, p(x \mid \theta_j) = \arg\max_{j=1,\ldots,k} P(\theta_j \mid x).$$

Если композиция эффективно разделяет множество прецедентов, выполняются неравенства $\forall j \neq j^*(x)\ P(\theta_j \mid x) << P(\theta_{j^*(x)} \mid x)$, откуда

$$P(x) \approx w_{j^*(x)} P(x \mid \theta_{j^*(x)}), \tag{15}$$

что совпадает с выражением (14). При этом решается проблема максимизации и вторая из упомянутых ранее проблем: значения $j^*(x)$, полученные при построении композиции с низким порядком цепей Маркова l, можно использо-

вать для разделения обучающей выборки на части $\{x \in X \mid j^*(x) = s\}_{s=1}^{k}$ и обучения на этих частях цепей более высокого порядка.

В реальных задачах, когда структура гена x неизвестна, для оценки $j^*(x)$ можно использовать какой-либо метод классификации, основанный на числовых признаках, полученных из известной нуклеотидной записи $s = \Pr_s(x)$. Один из возможных наборов признаков — эмпирические оценки переходных вероятностей m-го порядка для строки s:

$$\hat{p}_m(v \mid u) = \frac{\sum_i [s_i \ldots s_{i+m} = uv]}{\sum_i [s_i \ldots s_{i+m-1} = u]}, \quad u \in O^m, \ v \in O; \tag{16}$$

в этом случае каждой строке s соответствует вектор из 4^{m+1} чисел, расположенных на отрезке $[0,1]$. В качестве алгоритма классификации в настоящей работе рассмотрен метод опорных векторов (SVM) [9], адаптированный для задачи классификации с произвольным количеством классов путем построения отдельных классификаторов для каждой пары классов [10].

ВЫЧИСЛИТЕЛЬНЫЙ ЭКСПЕРИМЕНТ

Для проверки эффективности построения смесей вероятностных распределений с помощью EM-алгоритма рассматривались геномы шести видов, доступные в репозитории NCBI [11]: *Homo sapiens* (человек), *Gallus gallus* (курица), *Mus musculus* (мышь), *Rattus norvegicus* (крыса), *Papio anubis* (павиан) и *Sus scrofa* (свинья). Принимались во внимание гены, для которых полностью известна нуклеотидная запись; для генов человека и мыши для ускорения вычислений вводилось ограничение на длину: не более 40000 нуклеотидов.

Эффективность оптимизации EM-алгоритмом оценивалась с помощью функционала (5), для удобства нормализованного на количество строк во множестве прецедентов X:

$$Q_n(\Theta; X) = \frac{1}{|X|} Q(\Theta; X).$$

При работе алгоритма использовались следующие параметры:

- порядок цепей Маркова $l = 5$;
- полуширина окна усреднения для распределения строк по длинам $\varphi = 50$;
- максимальная длина строк в этом же распределении $d_{\max} = 20000$;
- нижние пороги вероятностей $\varepsilon_\varphi = 10^{-7}$, $\varepsilon_\pi = \varepsilon_p = 10^{-4}$;
- критерий остановки алгоритма — выполнение 20 итераций.

Результаты оптимизации при использовании последовательного добавления и удаления компонентов смеси отображены на рис. 1 и 2 соответственно. Здесь $\Delta Q = Q_n(k) - Q_n(1)$ и $\Delta Q = Q(k) - Q_n(5)$ — приращения функционала качества (5); k — число классов.

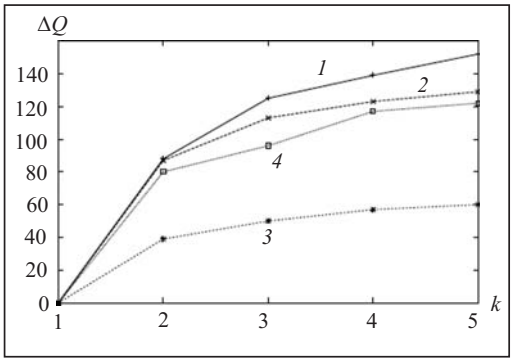

Рис. 1. Функционал качества для алгоритма последовательного добавления компонентов смеси для геномов: *Homo sapiens* — кривая *1*; *Gallus gallus* — *2*; *Mus musculus* — *3; Papio anubis* — *4*

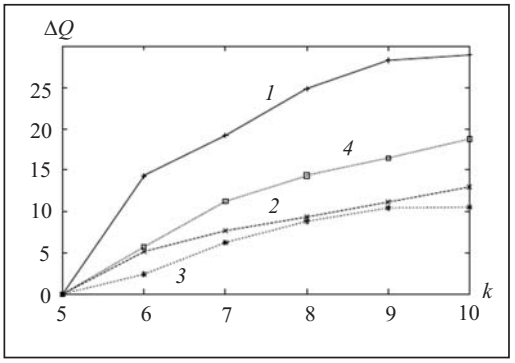

Рис. 2. Функционал качества для алгоритма последовательного удаления компонентов смеси для геномов: *Homo sapiens* — кривая *1*; *Gallus gallus* — *2*; *Mus musculus* — 3; *Sus scrofa* — *4*.

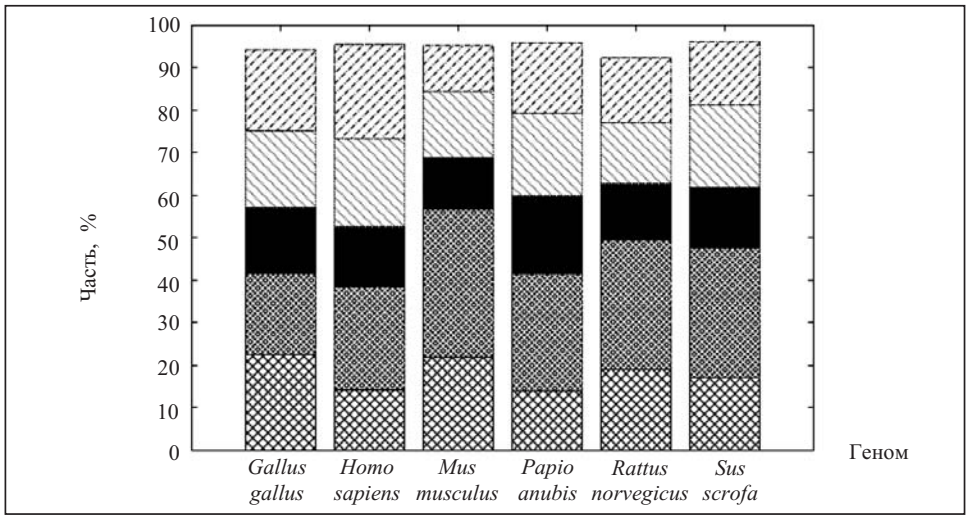

Рис. 3. Распределение генов исследуемых видов по пяти компонентам

Таблица 1. Функционал подобия (17) смесей из четырех компонентов для различных видов

Композиция *X*	**Значения функционала для пары *X* и *Y***					
	Композиция *Y*					
	Gallus gallus	*Homo sapiens*	*Mus musculus*	*Papio anubis*	*Sus scrofa*	*Rattus norvegicus*
Gallus gallus	0,0	409,2	176,4	256,5	349,8	148,9
Homo sapiens	359,1	0,0	132,6	9,5	356,7	128,1
Mus musculus	259,5	300,8	0,0	187,1	314,9	22,1
Papio anubis	305,1	14,4	123,2	0,0	330,0	115,6
Sus scrofa	319,5	375,7	169,3	239,9	0,0	141,8
Rattus norvegicus	250,3	323,0	24,3	198,1	305,4	0,0

На рис. 3 показан график распределения генов по наиболее вероятным компонентам $j^*(x)$ при ограничении значений апостериорной вероятности $\max_j P(x\,|\,\theta_j) > 0{,}95$. Как видно, гены распределены по составляющим смеси достаточно равномерно; количество генов, для которых ни одна апостериорная вероятность не превышает 0,95, пренебрежимо мало (около 5% для каждого вида), что обосновывает возможность перехода (15).

Несмотря на сложность смесей (1) по сравнению с деревьями разбиений, описанными в [3], их можно эффективно использовать для классификации других геномов. Подобие смесей двух геномов: X и Y, рассчитывалось по формуле

$$Q_n(\Theta(X); X) - Q_n(\Theta(X); Y), \qquad (17)$$

т.е. равнялось снижению функционала качества смеси, полученной с помощью ЕМ-алгоритма на генах X, при вычислении на другом геноме. Для двухкомпонентных смесей величины (17) для всех пар (X, Y) приблизительно равны нулю. Последнее означает, что они почти подобны. При увеличении числа компонентов различия между композициями становятся более явными, за исключением пар похожих организмов: мыши и крысы, человека и павиана (табл. 1).

Для аппроксимации значений $j^*(x)$ использовались переходные вероятности (16) четвертого порядка, дающие $4^5 = 1024$ признака на каждую строку. При-

Таблица 2. Точность аппроксимации разбиений геномов методом опорных векторов

Геном	Точность аппроксимации разбиений (%) для числа классов			
	2	3	4	5
Gallus gallus	94,73	91,49	86,09	81,61
Homo sapiens	96,25	92,78	89,79	87,30
Mus musculus	95,46	90,21	86,67	81,96
Papio anubis	95,45	88,99	87,68	84,10
Sus scrofa	96,17	91,44	86,01	83,85
Rattus norvegicus	95,09	89,63	84,08	80,74

Таблица 3. Качество классификации фрагментов генов при использовании смесей и их аппроксимаций

Геном	Число алгоритмов	Значения меры качества, %					
		NSp	*NSn*	*CC*	*ACP*	*ESp*	*ESn*
Gallus gallus	1	54,41	64,65	56,06	78,12	47,39	32,03
	3	69,11	65,51	64,97	82,50	52,72	37,58
	3.SVM	69,66	64,55	64,76	82,40	52,97	37,63
Homo sapiens	1	35,58	89,56	49,48	76,45	27,64	31,11
	5	57,04	86,75	66,39	83,73	44,95	48,91
	5.SVM	57,48	86,15	66,42	83,71	45,19	49,70
Mus musculus	1	59,97	85,32	67,20	83,97	42,22	40,39
	5	76,00	81,49	75,77	87,90	53,38	47,62
	5.SVM	76,28	81,58	75,96	87,99	53,47	47,87
Papio anubis	1	39,91	86,65	52,28	77,45	30,41	31,42
	5	68,52	81,59	71,51	85,86	50,62	49,17
	5.SVM	68,23	81,13	71,09	85,65	50,78	49,74
Sus scrofa	1	33,24	85,66	47,60	75,75	24,87	26,64
	5	57,92	78,85	64,31	82,46	41,05	41,01
	5.SVM	55,41	78,20	62,31	81,53	40,77	41,86
Rattus norvegicus	1	61,73	83,59	67,47	84,01	40,75	36,49
	4	76,45	78,20	74,10	87,05	48,55	40,28
	4.SVM	76,30	78,34	74,00	87,00	48,85	40,63

менялась реализация метода опорных векторов из библиотеки libsvm с радиально-базисными функциями и управляющим параметром $C = 1$. В результате пятикратной кросс-валидации было установлено, что SVM показывает достаточно высокое качество классификации, постепенно уменьшающееся при увеличении количества компонентов в смеси (табл. 2).

Вычислительный эксперимент завершился проверкой эффективности композиций вида (1) для классификации экзонов и интронов. Для измерения качества использовалась пятикратная кросс-валидация с шестью метриками, описанными в [1]:

- метрики *NSp* (нуклеотидная специфичность), *NSn* (нуклеотидная чувствительность), *CC* (коэффициент корреляции), *ACP* (средняя условная вероятность) оценивают качество классификации отдельных оснований;
- меры *ESp* (экзонная специфичность) и *ESn* (экзонная чувствительность) измеряют качество определения границ между экзонами и интронами.

В результате вычислений выяснено, что оптимальные смеси, которые строятся при кросс-валидации, порождаемые ими значения $j^*(x)$ и их аппроксимации с помощью метода опорных векторов мало отличаются от соответствующих им

величин, построенных для полных выборок. В соответствии с этим, как и в работе [3], для ускорения вычислений использовались смеси распределений, полученные на целых геномах. Результаты для композиций цепей Маркова седьмого порядка приведены в табл. 3, где под 3.SVM подразумевается использование композиции из трех алгоритмов с аппроксимацией разбиения с помощью SVM. Как видно, применение смесей позволяет существенно повысить качество классификации; при этом для метода SVM метрики практически не меняются.

ЗАКЛЮЧЕНИЕ

Рассмотрена общая постановка задачи разделения смеси распределений, представляющих собой цепи Маркова произвольного порядка. Для решения задачи предложен EM-алгоритм с вспомогательными алгоритмами, используемыми для определения выбора начального приближения и оптимального числа компонентов смеси. Показано, как приведенные рассуждения можно применить для решения задачи определения фрагментов генов (экзонов и интронов). Полученные композиции алгоритмов с эксклюзивной компетентностью позволяют ощутимо повысить качество классификации, что свидетельствует о возможности применения цепей Маркова высоких порядков для описания генов.

В качестве направлений для дальнейших исследований можно выделить:

- кластеризацию с помощью полученного математического аппарата интронов и экзонов;
- применение EM-алгоритма для повышения качества классификации в других задачах биоинформатики, в частности, задачи определения вторичной структуры белков.

СПИСОК ЛИТЕРАТУРЫ

1. Knapp K., Chen Y.-P.P. An evaluation of contemporary hidden Markov model genefinders with a predicted exon taxonomy // Nucleic Acids Research. — 2007. — **35**. — P. 317–324.
2. Сергиенко И.В., Гупал А.М., Островский А.В. Распознавание фрагментов генов в ДНК с применением моделей Маркова со скрытыми переменными // Кибернетика и системный анализ. — 2012. — № 3. — С. 58–67.
3. Гупал А.М., Островский А.В. Использование композиций моделей Маркова для определения функциональных участков генов // Там же. — 2013. — № 5. — С. 61–68.
4. Шлезингер М.И. О самопроизвольном распознавании образов // Читающие автоматы. — К.: Наук. думка, 1965. — С. 38–45.
5. Dempster A.P., Laird N.M., Rubin D.B. Maximum likelihood from incomplete data via the EM algorithm // J. the Royal Statistical Society. Ser. B. — 1977. — **34**. — P. 1–38.
6. Bishop C. Pattern recognition and machine learning. — Cambridge: Springer, 2006. — 749 p.
7. Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: классификация и снижение размерности. — М.: Финансы и статистика, 1989. — 607 с.
8. Ридли М. Геном: автобиография вида в 23 главах. — М.: Эксмо, 2008. — 432 с.
9. Cortes C., Vapnik V. Support vector machines // Machine Learning. — 1995. — **20**. — P. 273–293.
10. Knerr S., Pesonnaz L., Dreyfus G. Single-layer learning revisited: a stepwise procedure for building and training a neural network // Neurocomputing: Algorithms, Architectures and Applications / F.F. Soulie, J. Herald (Eds). — Berlin: Springer, 1990. — P. 41–50.
11. Национальный центр биотехнологической информации США. — http://ncbi.nlm.nih.gov/.

Поступила 08.04.2014