РАСПОЗНАВАНИЕ СТРУКТУРЫ ГЕНЕТИЧЕСКИХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ НА ОСНОВЕ МАРКОВСКИХ МОДЕЛЕЙ

Е. А. Медведь

ВВЕДЕНИЕ

Генетическая информация, заложенная в каждой клетке любого живого организма, играет определяющую роль в процессах его развития, жизнедеятельности и увядания. Данная информация, будучи правильно извлечена и проанализирована, представляет собой огромный интерес для специалистов в области генетики и медицины. Например, такая информация помогает разрабатывать новые лекарства, эффективные вакцины и средства диагностики заболеваний, выращивать принципиально новые сорта растений, штаммы микроорганизмов с заранее запрограммированными свойствами.

В настоящее время, благодаря стремительному развитию вычислительной техники, возрос интерес к ранее чересчур трудоёмким (в вычислительном плане) технологиям автоматического анализа ГП. Одно из перспективных направлений развития такого рода технологий – обнаружение и анализ марковской зависимости в ГП.

В предшествующей работе [1] была исследована возможность применения классической модели цепи Маркова (ЦМ) порядка s [2] в анализе ГП. Было показано, что из-за экспоненциального (по s) роста числа независимых параметров данной модели невозможно получить адекватные оценки параметров марковской зависимости достаточно высокого порядка (на практике – более 10) по существующим ГП. Всвязи с этим было предложено использовать следующие малопараметрические модели ЦМ.

ИСПОЛЬЗУЕМЫЕ МОДЕЛИ

1. Модель Джекобса-Льюиса. Подробное описание и строгое определение этой модели можно найти в [2]. Концепция её выражена соотношением:

$$x_t = \mu_t x_{t-\eta_t} + (1-\mu_t)\xi_t,\ t > s,$$

где μ_t – случайная величина (СВ) из распределения Бернулли с параметром ρ, η_t – дискретная СВ, задаваемая распределением вероятностей $\boldsymbol{P}\{\eta_t = j\} = \lambda_j,\ j \in \{1,\ldots,s\}$; ξ_t – дискретная СВ, задаваемая распределением $\boldsymbol{P}\{\xi_t = i\} = \pi_i,\ i \in S = \{1,\ldots,L\}$; S – пространство состояний модели

(алфавит уникальных символов последовательности); s – порядок модели. Параметры модели: ρ, λ_j, $j \in \{1,\ldots,s\}$, π_i, $i \in \{1,\ldots,L\}$.

2. MTD-модель Рафтери. Подробнее можно прочесть в [3]. Данная модель предполагает, что матрица вероятностей одношаговых переходов ЦМ представима в следующем «малопараметрическом» виде:

$$p_{i_1,\ldots,i_s,i_{s+1}} = \sum_{j=1}^{s} \lambda_j \boldsymbol{P}\{x_{t+1} = i_{s+1} | x_{t-s+j} = i_j\} = \sum_{j=1}^{s} \lambda_j q_{i_j,i_{s+1}}, i_1,\ldots,i_s,i_{s+1} \in S,$$

где $q_{ij} \geq 0$, $i, j \in S$, $\sum_{j=1}^{L} q_{ij} = 1$, $i \in S$; $\lambda_1 > 0$, $\lambda_j \geq 0$, $j \in \{2,\ldots,s\}$, $\sum_{j=1}^{s} \lambda_j = 1$.

Параметры модели: q_{ij}, $i, j \in S = \{1,\ldots,L\}$, λ_j, $j \in \{1,\ldots,s\}$.

3. Цепь Маркова порядка s с r частичными связями. Подробное описание может быть найдено в [4]. Эта модель предполагает представление матрицы вероятностей одношаговых переходов ЦМ в следующем виде:

$$p_{J_1^{s+1}} = p_{j_1,\ldots,j_s,j_{s+1}} = q_{j_{m_1^0},\ldots,j_{m_r^0}}, \quad J_1^{s+1} \in S^{s+1},$$

где $J_1^{s+1} = (j_1,\ldots,j_s,j_{s+1})$ – $(s+1)$-мерный вектор индексов; r – число связей, $1 \leq r \leq s$; $M_r^0 = (m_1^0,\ldots,m_r^0)$ – целочисленный вектор с r упорядоченными по возрастанию компонентами $1 \leq m_1^0 < m_2^0 < \ldots < m_r^0 \leq s$, называемый шаблоном связей; Q – стохастическая квадратная матрица размера $(r+1)$. Параметры модели: M_r^0, Q.

4. Цепь Маркова условного порядка. Данная модель впервые предложена и подробно описана в [5]. Сущность этой модели заключается в замене полносвязной ЦМ, где распределение вероятностей каждого последующего символа зависит от s предыдущих, последовательностью, в которой распределение последующего символа задаётся матрицей, определяемой так называемым базовым фрагментом памяти (БФП) – B_* предшествующими символами, и одним символом из s предшествующих, не входящим в БФП, порядковый номер которого также определяется БФП.

5. F-семейство условно-бернуллиевских временных рядов порядка s. Эта модель [6] разработана для анализа двоичных последовательностей и задаётся следующей формулой:

$$\boldsymbol{P}\{x_t = i \mid X_{t-s}^{t-1}\} = \begin{cases} \theta(X_{t-s}^{t-1}) = F\left(\sum_{j=1}^{m} a_j \psi_j\left(X_{t-s}^{t-1}\right)\right), i = 1, \\ 1 - \theta(X_{t-s}^{t-1}), i = 0, \end{cases}$$

где $\Psi\left(X_{t-s}^{t-1}\right)=\left(\psi_1\left(X_{t-s}^{t-1}\right),...,\psi_m\left(X_{t-s}^{t-1}\right)\right)^T$ – вектор-столбец m заданных базисных функций, $F(u):R^1\to[0,1]$ – произвольная известная функция распределения, $a=(a_j)\in R^m$ – вектор-столбец параметров модели.

Оценки перечисленных параметров представленных моделей строятся методом максимального правдоподобия. Каждая из моделей 1–5 представляет собой ЦМ с матрицей вероятностей одношаговых переходов специального, параметризованного вида, поэтому нахождение функций правдоподобия не составляет труда. Оценки максимального правдоподобия (ОМП) параметров моделей 3, 4 получены аналитически. Для моделей 1, 2, 5 ОМП получаются численной максимизацией соответствующих функций правдоподобия. Для модели 5 также удаётся получить состоятельную оценку вектора a на основе анализа вероятностно-статистических свойств s-фрагментов.

Для определения порядков моделей 1–5, а также числа связей в модели 3 и длины БФП в модели 4 используются информационные критерии (ИК) – статистические правила, позволяющие оценить качество построенной по известной выборке статистической модели без проверки статистических гипотез. Подробнее о сущности информационных критериев можно узнать в [7]. В [1] описан пример применения ИК для определения порядка классической модели ЦМ.

Построенные алгоритмы оценивания параметров моделей 1–5 реализованы в программном обеспечении (ПО) на языках C++, R и Python. С помощью созданного ПО были проанализированы смоделированные по соответствующим моделям последовательности (для проверки корректности алгоритмов оценивания), а также генетические последовательности из баз генетических данных [8]. Для моделей 1, 2, 5 оценивался порядок, параметры модели 3 оценивались при заданных максимальном порядке и числе связей, модели 4 – при заданных максимальном порядке и длине БФП. Описанные модели были применены при распознавании генетических последовательностей из [8]. Было отмечено, что визуализации избранных параметров моделей могут быть полезны при анализе последовательности специалистом.

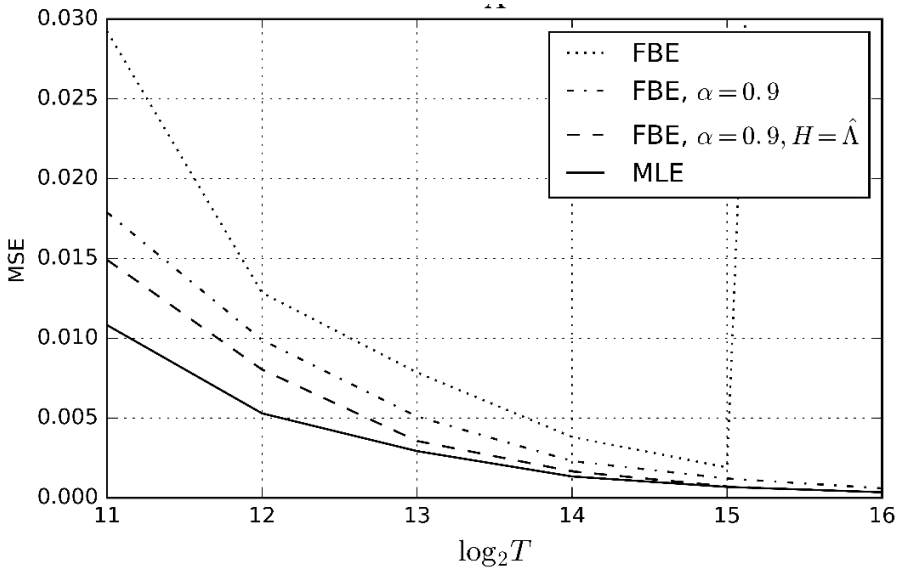

Рис. 1. Среднеквадратичная ошибка оценивания a

На рисунке 1 изображён график сходимости среднеквадратичной ошибки оценивания a (MSE) при оценивании численной максимизацией функции правдоподобия (MLE), а также с помощью различных модификаций (α – параметр регуляризации, H – матрица для учёта относительных частот s-грамм, подробно описаны в [6]) частотной оценки (FBE), призванных повысить её робастность и точность.

Литература

1. *Медведь, Е. А.* Статистическое обнаружение марковской зависимости в генетических последовательностях / Е. А. Медведь // XVI Республиканская научно-практическая конференция молодых учёных: сборник материалов. Брест, 15 мая 2014 г. В трёх частях. Часть 1 / под общ. ред. А.Е. Будько – Брест: БрГУ, 2014. – С. 87–89.
2. *Jacobs P. A., Lewis P. A. W.* Discrete time series generated by mixtures / P. A. Jacobs, P. A. W. Lewis // J. of the Royal Stat. Soc., 1978 – Vol. 40, No. 1–3.
3. *Raftery A., Tavare S.* Estimation and modelling repeated patterns in high order Markov chains with the Mixture Transition Distribution model / A. Raftery, S. Tavare // J. of Applied Statistics, 1994 – Vol. 43, No. 1, p. 179.
4. *Kharin Yu. S.* Long-memory discrete-valued time series: models and methods / Yu. S. Kharin // Computer Data Analysis and Modelling: Theoretical and Applied Stochastics, Minsk: Publ. Center of BSU, 2013 – Vol. 1, p. 64.
5. *Харин Ю. С.* Алгоритмы статистического анализа цепей Маркова с условной глубиной памяти / Ю. С. Харин, М. В. Мальцев // Информатика. 2011. № 1. С. 34.
6. *Kharin Yu. S.* Statistical estimation of parameters for binary conditionally nonlinear autoregressive time series / Kharin Yu. S., Voloshko V., Medved E. [presented for publication].
7. *Akaike H. A* New Look at the Statistical Model Identification / H. Akaike // IEEE Transactions on Automatic Control, 1974 – Vol. AC-19, No. 6, p. 716–723.
8. NCBI Nucleotide database [Электронный ресурс]. – Режим доступа http://www.ncbi.nlm.nih.gov/nuccore. – Дата доступа: 12.03.2017.