УДК 004.42

ПРОГРАММНЫЙ КОМПЛЕКС ДЛЯ СТАТИСТИЧЕСКОГО АНАЛИЗА ДАННЫХ НА ОСНОВЕ МАЛОПАРАМЕТРИЧЕСКИХ МОДЕЛЕЙ

Мальцев М.В., Палуха В.Ю.
НИИ прикладных проблем математики и информатики
E-mail: maltsew@bsu.by, palukha@bsu.by

Abstract. *The paper deals with parsimonious models of Markov chains. We present description of the developed software, which realizes methods and algorithms of statistical analysis based on these models.*

Малопараметрические марковские модели

Задачи статистического анализа данных требуется решать во многих областях: экономике, защите информации, генетике и др. Зачастую требуется определить, присутствуют ли в данных зависимости большой глубины и подобрать модель, которая наиболее адекватно их описывает. Существует универсальная математическая модель для описания зависимостей высоких порядков – цепь Маркова порядка s, $s < \infty$. Случайная последовательность $\{x_t: t \in \mathbf{N}\}$, принимающая значения из N-элементного множества $A = \{0, 1, ..., N–1\}$, называется однородной цепью Маркова порядка s (ЦМ(s)), если выполняется следующее марковское свойство [1]:

$$\mathrm{P}\{x_{t+1} = i_{t+1} \mid x_t = i_t, \ldots, x_1 = i_1\} = \mathrm{P}\{x_{t+1} = i_{t+1} \mid x_t = i_t, \ldots, x_{t-s+1} = i_{t-s+1}\},\ t > s,\ i_1, \ldots, i_{t+1} \in A,$$

т.е. распределение вероятностей процесса в момент времени $t + 1$ определяется не всей предысторией, а лишь s предыдущими состояниями.

Главным недостатком этой модели является экспоненциальный рост числа параметров модели D с ростом порядка s, которое определятся по формуле:

$$D = N_s\,(N - 1).$$

Таким образом, непосредственное использование ЦМ(s) на практике для обнаружения в данных зависимостей большой глубины затруднительно: для идентификации модели требуется иметь реализацию не всегда доступной длительности. В связи с этим разрабатываются так называемые малопараметрические марковские модели, которые представляют собой частные случаи ЦМ(s), число параметров которых зависит от порядка s полиномиально. Наиболее известными такими моделями являются:

1) цепь Маркова s-го порядка с r частичными связями (ЦМ(s, r)), $1 \le r \le s$ [2];
2) модель Рафтери (MTD-модель) [3];
3) модель Джекобса – Льюиса [4];
4) цепь Маркова условного порядка (ЦМУП) [5].

К примеру, для ЦМ(s, r) распределение вероятностей процесса в момент времени $t + 1$ определяется не s предыдущими состояниями, а r избранными состояниями.

Данные модели используются в разработанном программном комплексе для статистического анализа дискретных данных.

Программный комплекс «Markov Models»

Программный комплекс предназначен как для генерирования реализаций представленных выше моделей, так и для идентификации поступающих на обработку последовательностей. Кроме того, имеется возможность выбора наиболее адекватной модели для дискретной последовательности с помощью байесовского информационного критерия (BIC) [6].

Параметры модели ЦМ(s, r) оцениваются при помощи алгоритма из [2]. Для оценивания параметров модели Рафтери используется итерационный алгоритм из [3]. Параметры модели Джекобса-Льюиса оцениваются при помощи итерационного алгоритма из [4]. Оценки параметров модели ЦМУП строятся согласно [5].

Главное окно программного комплекса с результатами работы представлено на рисунке 1. В меню «Файл» имеется возможность выбора файла с входной последовательностью для оценивания параметров и файла для сохранения оценённых параметров либо файла для сохранения выходной последовательности в зависимости от выбранного режима. В меню «Опции» задаются параметры моделей для генерации и параметры алгоритмов оценивания. Мощность алфавита и длина генерируемой последовательности задаются в главном окне. Результаты работы выводятся в правой нижней части главного окна.

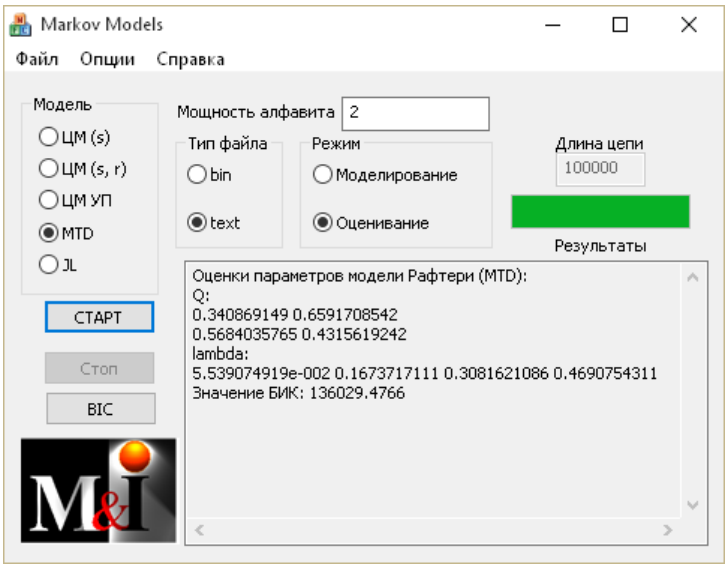

Рисунок 1 – Программный комплекс «Markov Models»

Список литературы

1. Doob, J. L. Stochastic processes / J. L. Doob. – NY : Wiley, 1953. – 654 p.

2. Харин, Ю. С. Цепь Маркова с частичными связями ЦМ(s, r) и статистические выводы о ее параметрах / Ю. С. Харин, А. И. Петлицкий // Дискретная математика. – 2007. – Т. 19, № 2. – С. 109 – 130.

3. Berchtold, A. Estimation of the Mixture Transition Distribution Model / A. Berchtold // Journal of Time Series Analysis. – 2001. – Vol. 22, № 4. – P. 379–397.

4. Харин, Ю.С. О статистическом анализе одной модели дискретной авторегрессии DAR(m) / Ю.С. Харин, А.Н. Ярмола // Вестник БГУ. Серия 1. – 2004. – № 3. – С. 65–69.

5. Kharin, Yu. S. Markov Chain of Conditional Order: Properties and Statistical Analysis Yu. S. Kharin, M. V. Maltsew // Austrian Journal of Statistics. – 2014. – Vol. 43, № 3–4. – P. 205–216.

6. Csiszar, I. Consistency of the BIC order estimator / I. Csiszar, P.C. Shields // Electronic research announcements of the American mathematical society. – 1999. – Vol. 5. – P. 123–127.