Р А Д И О Ф И З И К А

УДК 519.217.2

ПРИМЕНЕНИЕ ДИСКРИМИНИРУЮЩЕЙ ФУНКЦИИ В ЗАДАЧЕ ОЦЕНИВАНИЯ ПОРЯДКА ДВОИЧНОЙ МАРКОВСКОЙ ЦЕПИ

© 2008 г. ***С.А. Авдашов, Е.А. Коньков***

Нижегородский госуниверситет им. Н.И. Лобачевского

ekon@nifti.unn.ru

Поступила в редакцию 05.05.2008

Предложен метод оценивания порядка двоичной марковской цепи на основе дискриминирующей функции для биномиального распределения. Исследована эффективность метода в задаче оценивания порядка двоичной марковской цепи. Приведены результаты сравнения эффективности предложенного метода оценивания порядка двоичной марковской цепи и методов, основанных на информационных критериях Акаике и Байеса.

Ключевые слова: двоичная марковская цепь, дискриминирующая функция, биномиальное распределение.

Введение

Дискретные марковские модели – марковские цепи – находят применение в различных задачах обработки данных в качестве моделей анализируемых сигналов [1 – 4]. Многосвязные марковские процессы представляют собой обобщение простых марковских процессов [3]. Одной из важнейших характеристик многосвязной марковской цепи является её порядок, он в первую очередь подлежит оценке.

Рассмотрим последовательность отсчетов сигнала в дискретном времени, который представляет собой реализацию случайного процесса – однородной по времени марковской цепи некоторого неизвестного конечного порядка *M*, значения отсчетов этого процесса принадлежат конечному множеству *A*. Задача состоит в оценке порядка *M* двоичной марковской цепи по имеющейся реализации случайного процесса.

Существуют несколько методов оценки порядка марковской цепи. Условно их можно разделить на несколько классов: методы, основанные на анализе поведения энтропии [4–6], методы, основанные на информационных критериях Акаике [7, 8] и Байеса [9, 10]; кроме того, существует группа методов, в которых анализируются непосредственно значения переходных распределений вероятностей [5, 11].

В данной работе предлагается метод оценивания порядка двоичной марковской цепи на основе дискриминирующей функции для биномиального распределения.

Метод оценивания порядка двоичной марковской цепи

Введем обозначение для случайной последовательности из *n* бит в дискретном времени

$$x_1^n = \{x_1, x_2, x_3, ..., x_n\}, \qquad (1)$$

где x_i может принимать значение из множества $A = (0, 1)$. Обозначим конкретную реализацию случайной последовательности следующим образом:

$$a_1^n = \{a_1, a_2, a_3, ..., a_n\}, \ a_i \in A \forall i .$$

Количество повторений в выборке x_1^n последовательности бит a_1^k, $k < n$, можно выразить как

$$N_n(a_i^k) = \sum_{j=0}^{n-k} \delta\left(x_{j+1}^{j+k} = a_1^k\right), \qquad (2)$$

где

$$\delta\left(x_{j+1}^{j+k} = a_1^k\right) = \begin{cases} 1, & x_{j+1}^{j+k} = a_1^k , \\ 0, & x_{j+1}^{j+k} \neq a_1^k . \end{cases}$$

Используя данные обозначения, можно выразить эмпирические совместные вероятности появления последовательности бит a_1^{k+1} в выборке x_1^n

$$\hat{P}_n\left(a_1^{k+1}\right)=\frac{1}{n-k}N_n\left(a_1^{k+1}\right). \qquad (3)$$

Эмпирические переходные вероятности выражаются следующим образом

$$\hat{P}_n\left(a_{k+1}\left|a_1^k\right.\right)=\frac{N_n\left(a_1^{k+1}\right)}{N_{n-1}\left(a_1^k\right)}, \qquad (4)$$

причём $\hat{P}_n\left(a_{k+1}\left|a_1^k\right.\right)=0$, если $N_{n-1}\left(a_1^k\right)=0$.

Метод, предлагаемый в данной работе, основан непосредственно на анализе количеств повторений (2) различных комбинаций бит, входящих в выражение для переходных вероятностей (4). Он заключается в сравнении переходных вероятностей вида (4), у которых длина битовой последовательности в условии различается на 1 и остальные биты условной части совпадают между собой. Сравнение осуществляется с помощью дискриминирующей функции для биномиального распределения [12] следующим образом.

По заданной выборке последовательно оцениваются переходные вероятности (4) порядка k от 1 до k_0, где $k_0 > M$. Между соответствующими переходными вероятностями моделей соседних порядков рассчитывается значение дискриминирующей функции

$$d\left(m_1,n_1,m_2,n_2\right)=\frac{\left(\frac{m_1+1}{n_1+2}-\frac{m_2+1}{n_2+2}\right)^2}{\frac{(m_1+1)(n_1-m_1+1)}{(n_1+2)^2(n_1+3)}+\frac{(m_2+1)(n_2-m_2+1)}{(n_2+2)^2(n_2+3)}+\left(\frac{m_1+1}{n_1+2}-\frac{m_2+1}{n_2+2}\right)^2}, \qquad (5)$$

где $m_1=N_n\left(a_1^{k+1}\right)$, $n_1=N_{n-1}\left(a_1^k\right)$, $m_2=N_n\left(a_1^{(k+1)+1}\right)$, $n_2=N_{n-1}\left(a_1^{k+1}\right)$.

К достоинствам дискриминирующей функции (5) можно отнести те факты, что ее значение лежит в интервале от 0 до 1, в выражении (5) отсутствует операция логарифмирования, что исключает вычислительные трудности, возникающие при стремлении аргумента логарифма к нулю, характерные для алгоритмов на основе анализа поведения энтропии [4–6].

Для каждой пары моделей соседних порядков k и $k + 1$ получается всего 2^{k+1} значений дискриминирующей функции. Иллюстрация этой вычислительной процедуры в виде графа приведена на рис. 1. Граф представляет собой дерево, в узлах которого приведены значения переходных вероятностей и количества соответствующих комбинаций бит, через которые эти вероятности вычисляются по формуле (4). На ребрах, соединяющих узлы графа, приведены значения дискриминирующей функции (5), рассчитанной на основе величин, приведенных в узлах на концах ребра. Корневой узел дерева, который находится слева, соответствует последовательности независимых одинаково распределенных случайных бит. Узлы, находящиеся на одном уровне, соответствуют марковскому процессу, порядок которого равен номеру уровня. Номер уровня определяется количеством бит предыстории, стоящих в аргументе условной вероятности. Данная иллюстрация получена по выборке длиной $5\cdot10^5$ бит, представляющей собой реализацию марковской цепи третьего порядка ($M = 3$).

Рассмотрим переходную вероятность для некоторой комбинации бит в марковской модели текущего порядка k и соответствующую ей пару переходных вероятностей в марковской модели порядка $k + 1$. Если $k \geq M$, то учет еще одного бита предыстории приведет к тому, что переходные вероятности марковской модели порядка $k + 1$ будут (в идеальном случае) совпадать с соответствующими им переходными вероятностями марковской модели порядка k, а значение дискриминирующей функции (5) будет мало. Это можно наблюдать на рис. 1 между переходными вероятностями третьего и четвертого порядка. В противном случае, когда $0 \leq k < M$, учет дополнительного бита предыстории существенно изменит значения переходных вероятностей и значение дискриминирующей функции (5) будет велико.

В качестве критерия, по которому производится оценка марковского порядка исследуемого двоичного процесса, будем использовать среднее значение дискриминирующей функции (5), усредненное по всем 2^{k+1} ее значениям:

$$\bar{d}(k)=\frac{1}{2^{k+1}}\sum_{a_1^{k+1}\in A^{k+1}}d\left(N_n\left(a_1^{k+1}\right),N_{n-1}\left(a_1^k\right),\right.$$
$$\left.N_n\left(a_1^{(k+1)+1}\right)N_{n-1}\left(a_1^{k+1}\right)\right). \qquad (6)$$

Зависимость среднего значения дискриминирующей функции от предполагаемого порядка марковской цепи показана на рис. 2, где первые четыре точки отражают ситуацию, приведенную на рис. 1. Величина $\hat{M}$ является оценкой порядка марковской цепи, если выполняются следующие условия:

$$\begin{aligned}&\bar{d}(k)>\Delta, \quad k<\hat{M},\\ &\bar{d}(k)<\Delta, \quad k\geq\hat{M}.\end{aligned} \qquad (7)$$

Из анализа представленной на рис. 2 зависимости можно сделать вывод, что при $k < M = 3$

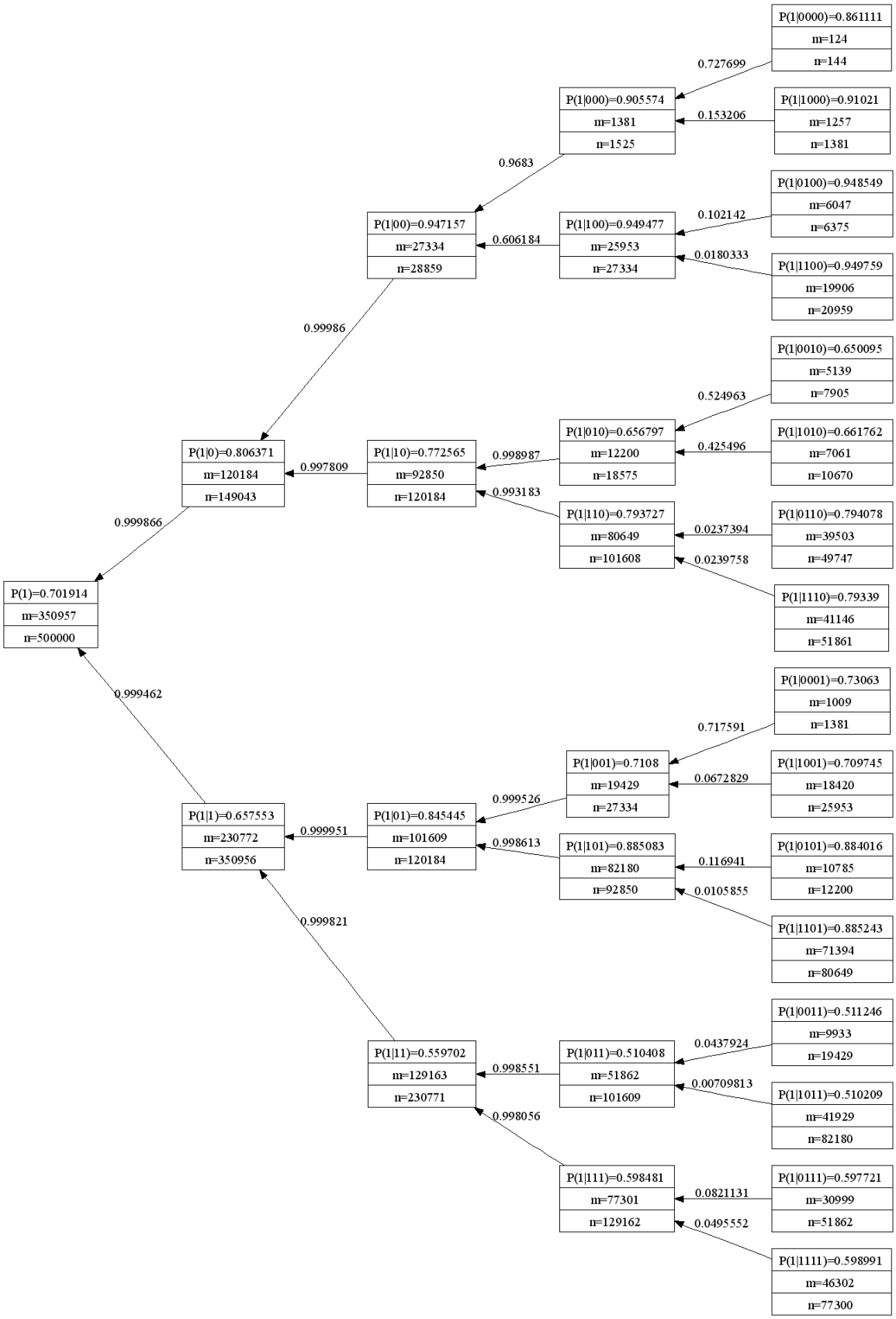

Рис. 1. Граф (дерево), иллюстрирующий вычисление дискриминирующей функции

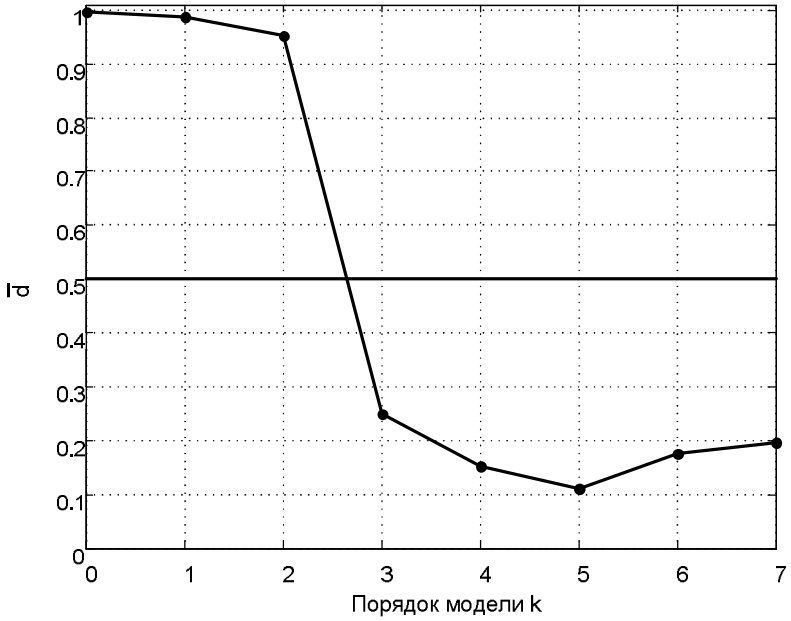

Рис. 2. Зависимость среднего значения дискриминирующей функции от предполагаемого порядка марковской цепи

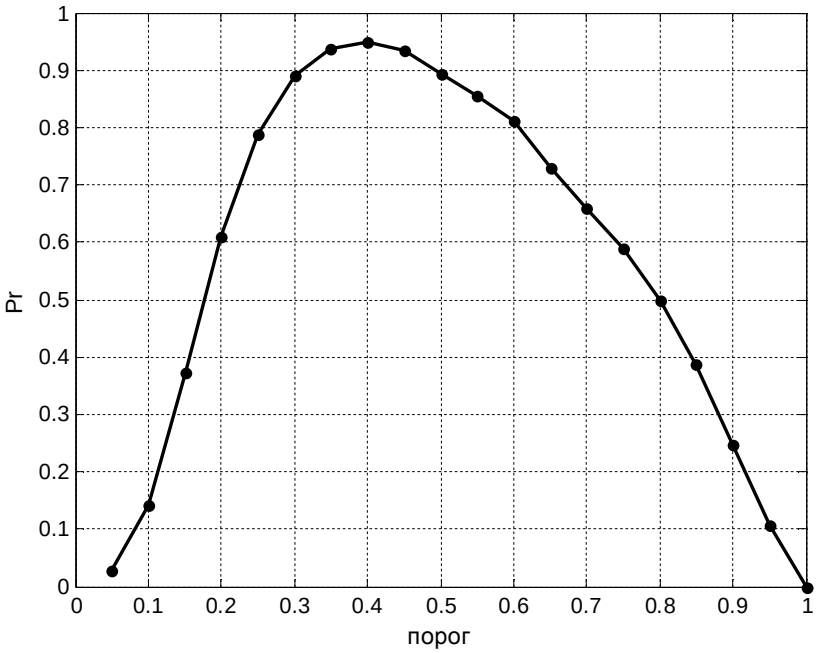

Рис. 3. Вероятность правильной оценки порядка марковской цепи предложенным методом в зависимости от порога для марковских цепей второго порядка длиной 1000 бит

среднее значение дискриминирующей функции больше порога $\Delta = 0.5$, а при $k \geq M$ среднее значение дискриминирующей функции меньше этого порога. Порог Δ является параметром предложенного метода и должен быть выбран таким, чтобы минимизировать ошибку определения порядка марковской модели.

Для случая когда дискриминирующая функция (5) используется в процедурах проверки гипотезы и принятия решения, зависимость свойств от величины порога изучена в [4]. В данной работе используется среднее значение дискриминирующей функции (6) и свойства предложенного алгоритма в зависимости от величины порога Δ определяются по результатам компьютерного моделирования.

Результаты и обсуждение

На рис. 3 представлена зависимость вероятности правильной оценки порядка марковской цепи Pr предложенным методом от величины

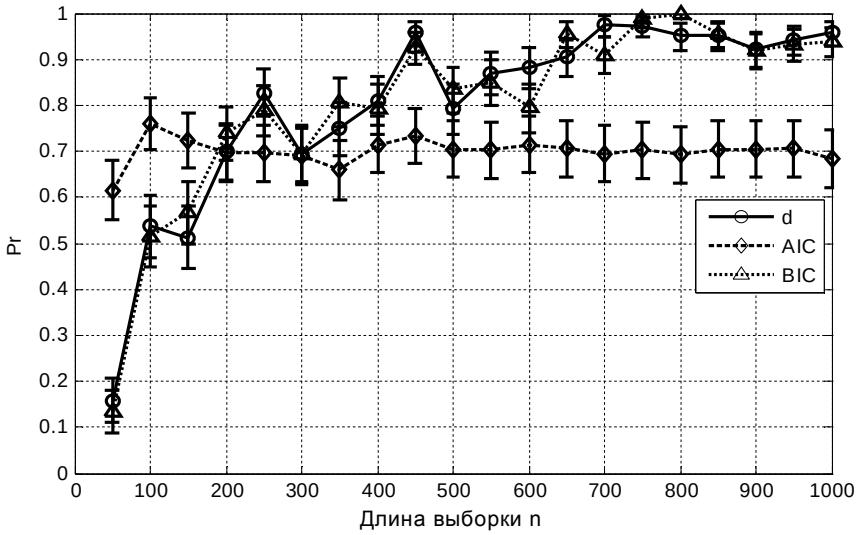

Рис. 4. Вероятность правильной оценки порядка марковской цепи в зависимости от длины выборки для марковских процессов третьего порядка

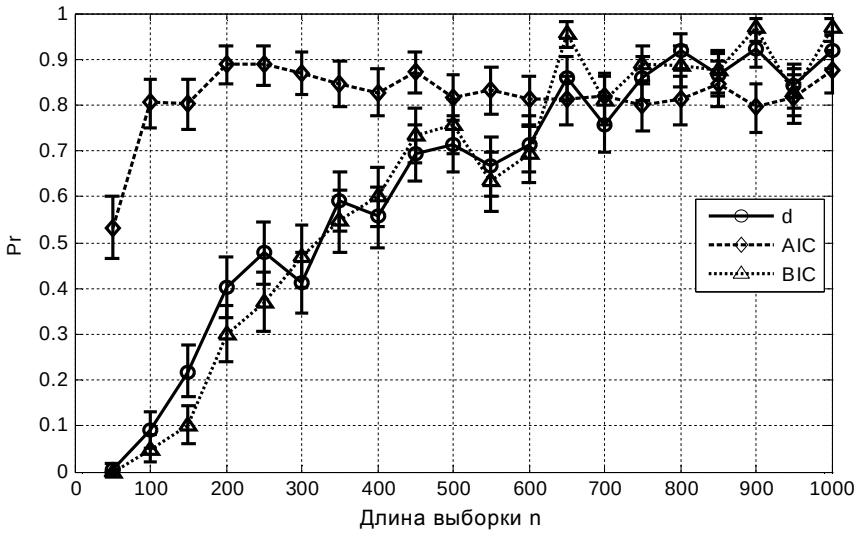

Рис. 5. Вероятность правильной оценки порядка марковской цепи в зависимости от длины выборки для марковских процессов четвертого порядка

порога Δ. Ошибки, как в сторону завышения, так и в сторону занижения на любую величину порядка марковской цепи, учитывались одинаково. Зависимость получена по реализациям различных марковских процессов второго порядка длиной 10^3 бит. Усреднение производилось по 100 выборкам для 50 распределений, каждая точка на графике получена усреднением по $5 \cdot 10^3$ выборкам. Значения порога варьировались от 0.05 до 1 с шагом 0.05. Анализ зависимости на рис. 3 показывает, что для процесса второго порядка при значении порога Δ в интервале от 0.3 до 0.5, с вероятностью 0.9 и выше оценка порядка марковской цепи будет верной. Результаты компьютерного моделирования для марковских цепей третьего порядка дают интервал оптимальных значений порога от 0.3 до 0.4, для четвёртого порядка — от 0.25 до 0.35. Таким образом, интервалы оптимальных значений порога Δ для исследованных процессов пересекаются и можно рекомендовать выбирать порог $\Delta = 0.35$.

На рис. 4 и 5 приведены зависимости вероятности правильной оценки порядка марковской цепи от её длины. По вертикальной оси отложена вероятность правильной оценки порядка марковской цепи, по горизонтальной оси – длина выборки (кривая *d* – предложенный метод, кривая AIC – метод на основе информационного критерия Акаике, кривая BIC – метод на основе информационного критерия Байеса).

Значения на графиках, получались следующим образом. Генерировалось по 20 случайных распределений каждого порядка. По каждому распределению генерировалось по 20 реализаций, которые подвергались анализу. Таким образом, каждая точка на графиках получена усреднением по 400 выборкам. Значение порога

было выбрано в соответствии с результатами проведенного исследования равным 0.35. На графиках указаны интервалы погрешностей, рассчитанные для доверительного уровня, равного 0.99.

Анализ рис. 5 показывает, что предложенный метод в данных условиях с учетом погрешностей по эффективности сравним с методом, основанным на информационном критерии Байеса, и несколько превосходит метод, основанный на информационном критерии Акаике, уже для длины выборки от 600 бит. Анализ рис. 5 показывает, что метод, основанный на информационном критерии Акаике, оказывается более предпочтительным при малых длинах выборок, а при длине выборок от 600 бит эффективность всех трех методов становится примерно одинаковой в пределах погрешностей. Вычисление критерия в предложенном методе не сопряжено с трудностями, характерными для методов, основанных на информационных критериях Акаике и Байеса, при стремлении аргумента логарифма к нулю.

Таким образом, в работе предложен метод оценивания порядка двоичной марковской цепи на основе дискриминирующей функции для биномиального распределения. Исследованы свойства метода в зависимости от величины порога принятия решения, установлен оптимальный диапазон значений порога. Проведено сравнение эффективности предложенного метода с методами на основе информационных критериев Акаике и Байеса.

Список литературы

1. Карлин С. Основы теории случайных процессов. М.: Мир, 1971.

2. Баруча-Рид А.Т. Элементы теории марковских процессов и их приложения. М.: Наука, главная редакция физико-математической литературы, 1969.

3. Романовский В.И. Дискретные цепи Маркова. М.–Л.: Государственное издательство технико-теоретической литературы, 1949.

4. Коньков Е.А., Солдатов Е.А., Морозов О.А. // Вестник Нижегородского государственного университета им. Н.И. Лобачевского. Сер. Радиофизика. Вып. 1 (3). Н. Новгород: Изд-во ННГУ, 2005. С. 148–155.

5. Peres Y., Shields P. // arXiv.org — arXiv:math.ST/0506080 v1.

6. Merhav N., Gutman M., Ziv J. // IEEE Transactions on Information Theory. 1989. V. 35, № 5. P. 1014–1019.

7. Akaike H. // IEEE Transactions on Automatic Control. 1974. V. AC-19, № 6. P. 716–723.

8. Tong H. // Journal of Applied Probability. 1975. V. 12. P. 488–497.

9. Csizar I., Shields P. // The Annals of Statistics. 2000. V. 28, № 6. P. 1601–1619.

10. Schwarz G. // The Annals of Statistics. 1978. V. 6. № 2. P. 461-464.

11. Finesso L., Liu Ch.-Ch., Narayan P. // IEEE Transactions on Information Theory. 1996. V. 42, № 5. P. 1488–1497.

12. Бурланков Д.Е., Коньков Е.А. // Труды РНТОРЭС им. А.С. Попова. Серия: Цифровая обработка сигналов и ее применение. Вып. IX-2. М.: РНТОРЭС, 2007. С. 449.

APPLICATION OF DISCRIMINANT FUNCTION IN THE ORDER ESTIMATION PROBLEM OF A BINARY MARKOV CHAIN

S.A. Avdashov, E.A. Konkov

A method of binary Markov chain order estimation based on a discriminant function for a binomial distribution has been proposed and its efficiency has been studied. The efficiency of the method has been compared with those based on Akaike and Bayesian information criteria. The results of this comparison are presented.