

СРАВНИТЕЛЬНЫЙ АНАЛИЗ ТОЧНОСТИ НЕЛИНЕЙНЫХ МОДЕЛЕЙ ПРИ ПРОГНОЗИРОВАНИИ СОСТОЯНИЯ СИСТЕМЫ НА ОСНОВЕ МАРКОВСКОЙ ЦЕПИ

© 2013 В.Н. Клячкин, Ю.С. Донцова

Ульяновский государственный технический университет

Поступила в редакцию 26.09.2013

В статье проводится описание и сравнительный анализ прогнозной точности моделей, используемых при нахождении будущих состояний марковского случайного процесса, протекающего в системе.

Ключевые слова: марковская цепь, прогнозирование состояния системы, нелинейная модель, логистическая регрессия.

Пусть некоторая система S может находиться в одном из состояний конечного (или счетного) множества возможных состояний S_1, S_2, \dots, S_n , а переход из одного состояния в другое возможен только в определенные дискретные моменты времени t_1, t_2, \dots, t_n , называемые шагами. Предположим, что вероятность перехода из любого состояния S_i в любое состояние S_j не зависит от того, как и когда система S попала в это состояние. В таком случае функционирование системы S описывается дискретной цепью Маркова.

В качестве x_1, x_2, \dots, x_m будем рассматривать внешние факторы, предположительно оказывающие влияние на систему S . В зависимости от области исследования данные переменные могут нести в себе различную смысловую информацию.

Требуется на основе имеющихся данных системы, включая воздействие внешних факторов, построить прогноз состояний системы на следующем шаге $(k+1)$ с помощью нелинейных моделей, а также сравнить точность прогнозирования.

Традиционно, задачи такого плана решаются с помощью нахождения матрицы переходных вероятностей P . Переходной вероятностью p_{ij} называют условную вероятность того, что из состояния i (в котором система оказалась в результате некоторого испытания, безразлично какого номера) в итоге следующего испытания система перейдет в состояние j . Таким образом, в обозначении p_{ij} первый индекс указывает номер предшествующего, а второй – номер последующего состояния.

Введем начальное распределение вероятностей марковской цепи:

$$P_1(0), P_2(0), \dots, P_i(0), \dots, P_n(0). \quad (1)$$

Клячкин Владимир Николаевич, доктор технических наук, профессор кафедры «Прикладная математика и информатика». E-mail: v_kl@mail.ru

Донцова Юлия Сергеевна, аспирант кафедры «Прикладная математика и информатика».

В частном случае, если первоначальное состояние системы S в точности известно $S(0) = S_i$, то начальная вероятность $P_j(0) = 1$, а все остальные равны нулю. Поскольку система может пребывать в одном из n состояний, то для каждого момента времени t необходимо задать n^2 вероятностей перехода p_{ij} , которое удобно представить в виде следующей матрицы:

$$\|p_{ij}\| = \begin{pmatrix} p_{11} & \dots & p_{1n} \\ \vdots & \ddots & \vdots \\ p_{n1} & \dots & p_{nn} \end{pmatrix}. \quad (2)$$

Так как в каждой строке матрицы помещены вероятности событий (перехода из одного и того же состояния i в любое возможное состояние j), которые образуют полную группу, то сумма вероятностей этих событий равна единице. Другими словами, сумма переходных вероятностей каждой строки матрицы перехода равна единице:

$$\sum_{i,j}^n p_{ij} = 1. \quad (3)$$

Такая матрица называется матрицей переходных вероятностей. Если матрица перехода на шаге $(k+1)$ будет известна, то прогноз распределения вероятностей состояний системы может быть получен, используя распределение вероятностей состояний системы за предыдущий шаг. То есть, распределение вероятностей состояний на следующем шаге вычисляется как:

$$P(k+1) = \pi(k)P(k, k+1), \quad (4)$$

где $p(k)$ – безусловное распределение вероятностей состояний системы.

Для нахождения переходных вероятностей будем использовать нелинейные модели из класса логистической регрессии [1].

Рассмотрим модель бинарной логистической регрессии, записываемую уравнением:

$$p(i, j) = \frac{1}{1 + e^{-\overline{ax}}}, \quad (5)$$

где $p(i, j)$ – переходная вероятность, \overline{ax} – линейная комбинация предикторов:

$$\overline{ax} = \alpha_{ij}^1 + \sum_{l=2}^{m+1} \alpha_{ij}^l \cdot x^{l-1}, \quad (6)$$

где α – неизвестные параметры модели, m – количество внешних факторов.

Перепишем модель следующим образом:

$$p(i, j) = \frac{\exp(\alpha_{ij}^1 + \sum_{l=2}^{m+1} \alpha_{ij}^l \cdot x^{l-1})}{1 + \exp(\alpha_{ij}^1 + \sum_{l=2}^{m+1} \alpha_{ij}^l \cdot x^{l-1})}. \quad (7)$$

Получили формулу для нахождения переходной вероятности, с учетом воздействия на систему внешних факторов. Однако при этом не учитываются прошлые переходы системы, которые также могут оказывать некоторое влияние. В связи с этим, добавим в уравнение дополнительный фактор, показывающий, был ли совершен рассматриваемый переход шаг назад.

Тогда с учетом данной поправки модель примет следующий вид:

$$p(i, j) = \frac{\exp(\alpha_{ij}^1 + \sum_{l=2}^{m+1} \alpha_{ij}^l \cdot x^{l-1} + \alpha_{ij}^{m+2} y_{ij-1})}{1 + \exp(\alpha_{ij}^1 + \sum_{l=2}^{m+1} \alpha_{ij}^l \cdot x^{l-1} + \alpha_{ij}^{m+2} y_{ij-1})}. \quad (8)$$

В общем случае введем историю предыдущих переходов:

$$H_{ij} = \{y_{ij}, k = 1, \dots, j - 1\}. \quad (9)$$

Следовательно, общая модель для нахождения переходной вероятности будет выглядеть так:

$$p(i, j) = \frac{\exp(\alpha_{ij}^1 + \sum_{l=2}^{m+1} \alpha_{ij}^l \cdot x^{l-1} + \sum_{r=1}^q \beta_{ij}^r \cdot y_{ij-r})}{1 + \exp(\alpha_{ij}^1 + \sum_{l=2}^{m+1} \alpha_{ij}^l \cdot x^{l-1} + \sum_{r=1}^q \beta_{ij}^r \cdot y_{ij-r})}, \quad (10)$$

где q – порядок модели, показывающий учет количества предыдущих переходов, β – неизвестные параметры.

На следующем шаге необходимо оценить все неизвестные параметры. Существует несколько способов, наиболее распространенными из которых являются метод наименьших квадратов и принцип максимального правдоподобия. Применим последний из них для модели общего вида.

Согласно принципу максимального правдоподобия за оценки параметров принимаются те

значения, которые дают максимум для функции правдоподобия:

$$F(\overline{f}) = \prod_{i=1}^v p(i, j)^{g_i} [1 - p(i, j)]^{1-g_i}, \quad (11)$$

где $\overline{\eta} = (\alpha^1, \dots, \alpha^{m+1}, \beta^1, \dots, \beta^q)$ – неизвестные параметры, v – объем выборки, g_i – зависимая переменная, принимающая следующие значения:

$$g = \begin{cases} 1, & \text{если был совершен переход для заданной пары состояний } (i \rightarrow j) \\ 0, & \text{иначе.} \end{cases}$$

Запишем логарифм функции правдоподобия и найдем максимум:

$$F(\overline{\eta}) = \sum_{i=1}^v g_i \ln p(i, j) + (1 - g_i) \ln(1 - p(i, j)). \quad (12)$$

$$F(\overline{\eta}) \rightarrow \max$$

В результате получим систему нелинейных уравнений относительно неизвестных параметров $\overline{\eta}$:

$$\left\{ \begin{aligned} w_0(\overline{\eta}) &= \sum p(i, j) - \sum_{\{i: g_i=1\}} 1 = 0, \\ w_1(\overline{\eta}) &= \sum p(i, j)x^1 - \sum_{\{i: g_i=1\}} x^1 = 0, \\ &\dots \\ w_m(\overline{\eta}) &= \sum p(i, j)x^m - \sum_{\{i: g_i=1\}} x^m = 0, \\ w_{m+1}(\overline{\eta}) &= \sum p(i, j)y_{ij-1} - \sum_{\{i: g_i=1\}} y_{ij-1} = 0, \\ &\dots \\ w_{m+q}(\overline{\eta}) &= \sum p(i, j)y_{ij-q} - \sum_{\{i: g_i=1\}} y_{ij-q} = 0. \end{aligned} \right. \quad (13)$$

Существует множество способов решения систем нелинейных уравнений, но одним из обеспечивающих наилучшую сходимость является метод Ньютона-Рафсона [2], в котором поиск решения осуществляется путём построения последовательных приближений и основан на принципах простой итерации.

Возьмем следующее начальное приближение для параметров модели:

$$\overline{\eta}^0 = (0, \dots, 0). \quad (14)$$

В качестве итерации запишем:

$$\overline{\eta}^{k+1} = \overline{\eta}^k - W(\overline{\eta}^k) J^{-1}(\overline{\eta}^k), \quad (15)$$

где

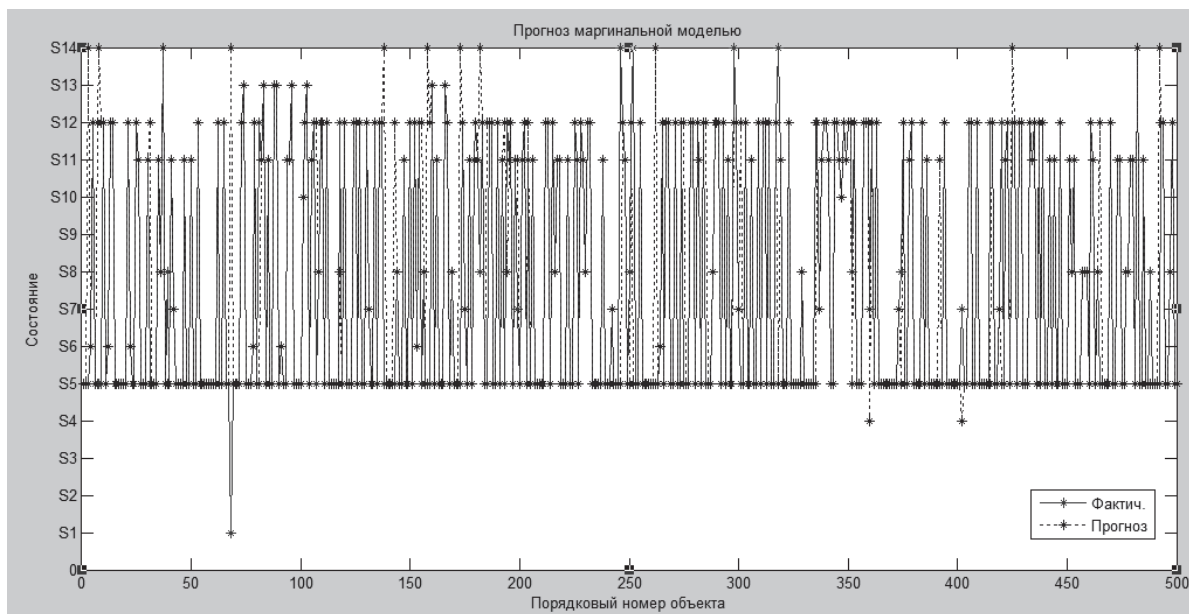


Рис. 1. Прогноз на базе маргинальной модели

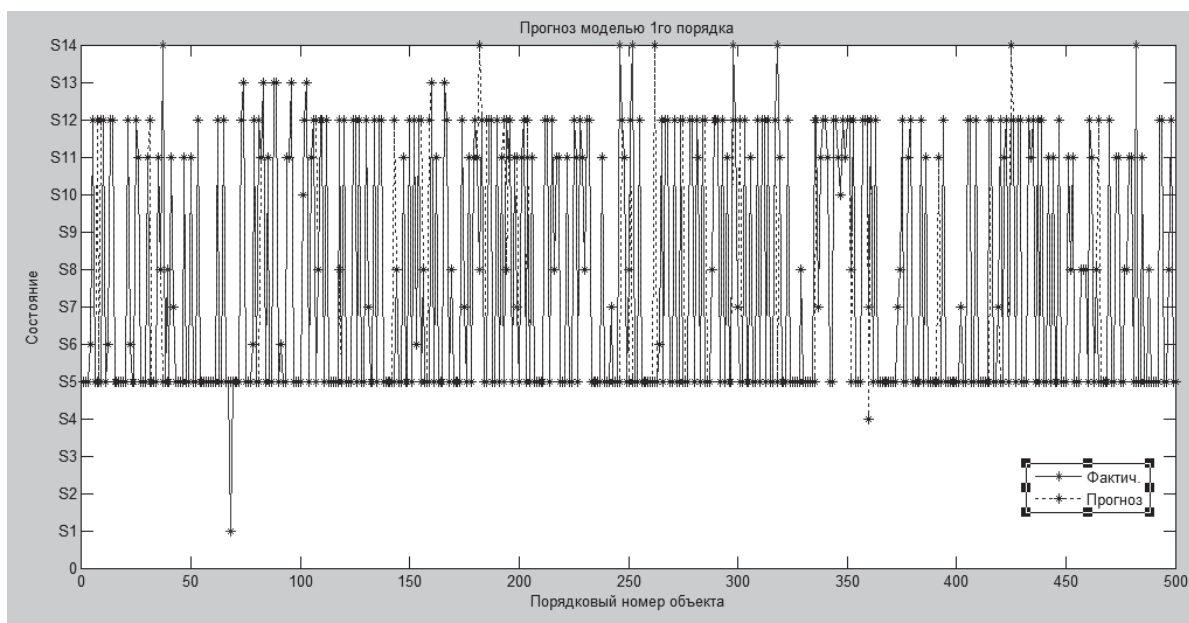


Рис. 2. Прогноз на базе модели 1-го порядка

$$W(\bar{\eta}^k) = (w_0(\bar{\eta}^k), w_1(\bar{\eta}^k), \dots, w_m(\bar{\eta}^k), w_{m+1}(\bar{\eta}^k), \dots, w_{m+q}(\bar{\eta}^k)),$$

– вектор частных производных, $J(\bar{\eta}^k)$ – матрица Якоби.

Для сравнительного анализа прогнозной точности описанных выше моделей рассмотрена выборка, состоящая из 5000 объектов, принимающих на каждом шаге одно из 14 возможных состояний. На основе данных, имеющихся за 17 шагов, строится прогноз состояний на следующем шаге и сравнивается с фактическими данными.

В результате применения нелинейных моделей погрешность прогноза составила менее 10%. При применении модели, содержащей факторы,

учитывающие предыдущие переходы более одного шага, результат ухудшается. Следовательно, в рассматриваемом примере имеет смысл использовать нелинейную логистическую модель, содержащую лишь факторы внешнего воздействия (маргинальная модель), либо логистическую модель с учетом одного дополнительного фактора (модель 1-го порядка), отвечающего за переход на предыдущем шаге. На рис. 1 и 2 показано сравнение опытных и прогнозируемых данных на основе первых 500 исследуемых объектов.

Исследование выполнено при поддержке Министерства образования и науки Российской Федерации, соглашение 14.В37.21.0672 (федеральная целевая программа “Научные и научно-педагогические кадры инновационной России”).

СПИСОК ЛИТЕРАТУРЫ

1. *Diggle P.J., Liang K.Y., Zeger S.L.* Analysis of Longitudinal Data. Oxford University Press: Oxford, 2002.
2. *Васильев Н.П., Егоров А.А.* Опыт расчета параметров логистической регрессии методом Ньютона-Рафсона для оценки зимостойкости растений // Математическая биология и биоинформатика. 2011. Т. 6. № 2. С.190-199.

**THE COMPARATIVE ANALYSIS OF ACCURACY OF NONLINEAR MODELS
AT FORECASTING OF THE CONDITION OF SYSTEM
ON THE BASIS OF THE MARKOVSKY CHAIN**

© 2013 V.N. Klyachkin, J.S. Dontsova

Ulyanovsk State Technical University

In article to be spent the description and the comparative analysis прогнозной accuracy of the models used at a finding of the future conditions марковского of casual process, proceeding in system.

Keywords: марковская a chain, forecasting of a condition of system, nonlinear model, logistical регрессия