

Применение теоретико-информационных методов прогнозирования к анализу геофизических и экономических данных

П. А. Приставка

В статье описывается и экспериментально исследуется метод прогнозирования стационарных и эргодических процессов, основанный на универсальной мере. На примере прогнозирования геофизических и экономических данных показывается, что точность полученных таким образом результатов может быть выше, чем у известных методов.

Ключевые слова: временные ряды, непараметрические методы, универсальная мера, универсальное кодирование, уровень моря, кросс-курсы.

1. Введение

Прогнозирование является важной задачей для многих областей. В настоящее время существует большое количество эффективных методов прогнозирования, которые основаны на различных математических моделях (см., например, [1]). Многие из подобных подходов основаны на математической статистике, теории вероятностей и некоторых других областях.

В этой статье мы описываем метод прогнозирования, который основан на так называемых универсальных мерах (или универсальных архиваторах) и экспериментально исследуем его эффективность.

По определению, универсальные коды, или универсальные методы сжатия данных без потерь, служат для «компрессии» текстов, т.е. кодирования их таким образом, чтобы длина кодовой последовательности текста была меньше длины исходного текста (при этом, конечно, должна оставаться возможность восстановить изначальный текст по его кодовой последовательности). Важно отметить, что априорные сведения о статистических характеристиках текста отсутствуют и именно поэтому подобные коды и называют универсальными. Фактически, универсальные коды в неявном виде используют оценки неизвестных статистических характеристик текста для его сжатия.

Универсальный код называется оптимальным, если он кодирует последовательность символов, порождённую дискретным источником, таким образом, что длина полученной кодовой последовательности асимптотически минимальна. Впервые универсальные коды, которые являются оптимальными для источников независимых и одинаково распределённых символов, были открыты Б. М. Фитингофом [2] и А. Н. Колмогоровым [3]. Универсальные коды для стационарных и эргодических дискретных источников были описаны в 1980-ые (см. [4]).

Очевидно, что такие универсальные коды могут рассматриваться как инструмент математической статистики, и вполне естественно попытаться применить их для решения традиционных проблем этой дисциплины (таких как проверка гипотез, оценка параметров и других). Впервые подобная попытка была предпринята в 1980-ые (см. [4 – 6]), и на сегодняшний день известно, что универсальные коды могут успешно применяться для проверки гипотез,

оценки параметров и прогнозирования временных рядов (см. [7, 8]). Однако в настоящее время имеются лишь предварительные данные о результатах применения данных методов прогнозирования к решению практических задач (см. [9]). Можно сказать, что имеющиеся результаты скорее служат для демонстрации самой возможности подобного применения, а не для предоставления информации о точности получаемых прогнозов.

Целью данной работы является разработка, реализация и экспериментальное исследование предложенного метода прогнозирования, основанного на универсальных кодах и так называемых универсальных мерах. В качестве объектов исследования были рассмотрены различные временные ряды, содержащие кросс-курсы некоторых валют и данные об уровне моря, т.е. процессы, имеющие важное практическое значение.

Полученные экспериментальные результаты показывают, что методы прогнозирования, основанные на универсальных кодах, обладают высокой точностью. В принципе, любой универсальный код может быть использован в качестве основы для метода прогнозирования, но в данной работе мы использовали универсальный код, предложенный Б. Я. Рябко в [4] и обладающий некоторыми дополнительными полезными свойствами (см. [7, 10]).

Статья имеет следующую структуру. Второй раздел содержит необходимые определения и информацию об универсальных кодах и мерах. Третий раздел посвящён результатам прогнозирования рассмотренных временных рядов. В четвёртом разделе дано краткое заключение.

2. Описание метода

Поясним сначала кратко задачу прогнозирования. Пусть есть некоторый стационарный и эргодический источник, который порождает последовательности $x_1 x_2 \dots$ элементов из некоторого множества A , которое может быть как конечным множеством, так и некоторым непрерывным интервалом. Предполагается, что распределение вероятностей этого источника $P(x_1 = a_{i_1}, x_2 = a_{i_2}, \dots, x_t = a_{i_t})$ (или плотность $p(x_1, x_2, \dots, x_t)$) неизвестно. Пусть источник порождает сообщение вида $x_1 \dots x_{t-1} x_t$, $x_i \in A$ для всех i , и следующий элемент требуется спрогнозировать.

Далее, приведём определение универсальной меры и поясним, как она связана с универсальными кодами. По определению, мера μ называется универсальной, если для любого стационарного и эргодического источника P верны следующие равенства:

$$\lim_{t \rightarrow \infty} \frac{1}{t} (-\log P(x_1 \dots x_t) - (-\log \mu(x_1 \dots x_t))) = 0$$

с вероятностью 1, и

$$\lim_{t \rightarrow \infty} \frac{1}{t} \sum_{u \in A^t} P(u) \log(P(u)/\mu(u)) = 0.$$

(Здесь и ниже $\log(x) = \log_2(x)$.) Данные равенства показывают, что, в некотором смысле, мера μ_U является непараметрической оценкой для (неизвестной) меры P . По этой причине универсальные меры могут быть использованы для оценки статистических характеристик процесса и прогнозирования.

В принципе, теперь мы могли бы перейти к описанию конкретной универсальной меры, которая была использована для прогнозирования в данной статье, и вообще не рассматривать универсальные коды. Однако мы хотим отметить, что универсальные меры имеют глубокую взаимосвязь с универсальными кодами, и, если есть универсальный код, то можно легко получить на его основе универсальную меру. Это важно для практических приложений, потому что многие универсальные коды доступны в виде компьютерных программ (так называе-

мых архиваторов), которые, следовательно, могут быть использованы для получения универсальной меры и непосредственно применяться для прогнозирования. По этой причине мы приведём здесь краткую информацию об универсальных кодах (их подробное описание можно найти, например, в [11]).

Можно сказать, что код отображает слова из множества A^t , $t \geq 1$, на множество слов в алфавите $\{0,1\}$. По определению, код U называется универсальным, если для любого стационарного и эргодического источника P верны следующие равенства:

$$\lim_{t \rightarrow \infty} |U(x_1 \dots x_t)|/t = H(P)$$

с вероятностью 1, и

$$\lim_{t \rightarrow \infty} |E_P(U(x_1 \dots x_t))|/t = H(P),$$

где $E_P(f)$ – среднее значение f по отношению к P , а $H(P)$ – энтропия P по Шеннону, т.е.

$$H(P) = \lim_{t \rightarrow \infty} -t^{-1} \sum_{u \in A^t} P(u) \log(P(u)).$$

Следующая теорема показывает, что любой универсальный код определяет универсальную меру.

Теорема 1. Пусть U – универсальный код и

$$\mu_U(\omega) = 2^{-|U(\omega)|} / \sum_{u \in A^{|\omega|}} \mu_U(u).$$

(Простое доказательство теоремы приводится в [11].) Итак, можно видеть, что, в некотором смысле, мера μ является подходящей (непараметрической) оценкой вероятностей для класса всех стационарных и эргодических источников на конечном алфавите.

Теперь опишем универсальную меру R , которая была использована в качестве основы для метода прогнозирования в данной статье. Для этих целей сначала опишем меру Кричевского K_m , которая является универсальной для множества марковских источников с памятью, или связностью, m , $m \geq 0$; если $m = 0$, то это источник независимых и одинаково распределённых символов. В некотором смысле эта мера является оптимальной для этого множества (см. [10]). По определению,

$$K_m(x_1 \dots x_t) = \begin{cases} \frac{1}{|A|^t}, & \text{если } t \leq m, \\ \frac{1}{|A|^m} \prod_{v \in A^m} \frac{\prod_{\alpha \in A} (\nu_x(v\alpha) + 1/2) / \Gamma(1/2)}{(\Gamma(\bar{\nu}_x(v) + |A|/2) / \Gamma(|A|/2))}, & \text{если } t > m; \end{cases} \quad (1)$$

где $\nu_x(v)$ – число последовательностей v , встречающихся в x , $\bar{\nu}_x(v) = \sum_{\alpha \in A} \nu_x(v\alpha)$, $x = x_1 \dots x_t$, а $\Gamma()$ – это гамма-функция.

Определим также распределение вероятностей $\{\omega = \omega_1, \omega_2, \dots\}$ для целых $\{1, 2, \dots\}$ как

$$\omega_i = 1/\log(i+1) - 1/\log(i+2), \quad (2)$$

$i=1, 2, \dots$. (В дальнейшем мы будем использовать это распределение, однако результаты будут справедливы для любого распределения ненулевых вероятностей.)

Мера R определяется как

$$R(x_1 \dots x_t) = \sum_{i=0}^{\infty} \omega_{i+1} K_i(x_1 \dots x_t) \quad (3)$$

Итак, значение меры R , вычисленное на основе формулы (3), может служить оценкой вероятности исхода процесса и использоваться для решения задачи прогнозирования.

При применении данного метода на практике может возникнуть вопрос о том, последовательность какой длины нужно использовать в качестве входных данных. Может показаться, что наилучший результат будет достигнут при использовании максимально возможного количества известных членов временного ряда. Однако в действительности это не всегда так, потому что статистические характеристики могут изменяться со временем, и, следовательно, «старые» данные не будут содержать информации о «новых» статистических характеристиках. Итак, если ответ на данный вопрос не очевиден, то следует использовать меру $\tilde{R}$, определение которой приводится далее. Пусть $x_1 \dots x_t$ – некоторая последовательность источника. Пусть N – множество различных натуральных чисел. Для каждого n из N рассмотрим выборку $x_{t-n+1} \dots x_t$. Другими словами, фактически каждая выборка состоит из n последних элементов исходной последовательности. Каждую отдельно взятую выборку данного множества назовем «окном», а соответствующее n – его размером. Предположим, $N = \{n_1, \dots, n_k\}$, $k \geq 1$, тогда

$$\tilde{R}(x_1, \dots, x_t) = \sum_{i=1}^k \tilde{\omega}_i R(x_{t-n_i+1} \dots x_t),$$

где $\tilde{\omega}_1, \tilde{\omega}_2, \dots, \tilde{\omega}_k$ – это распределение вероятностей на $N = \{n_1, \dots, n_k\}$. Аналогично величине R , $\tilde{R}$ может использоваться для оценки вероятности исхода процесса и использоваться для прогнозирования. Как видно из определения, при расчёте данной величины учитываются последовательности («окна») различной длины, и соответствующие R «склеиваются» с некоторыми вероятностями. Вычисления на основе $\tilde{R}$ должны автоматически обеспечивать оптимальные, с точки зрения простоты использования метода и точности прогноза, результаты (подробности и теоретическое доказательство см. в [11]). В дальнейшем подобный способ вычислений мы будем называть адаптивным режимом.

Подробное описание схемы прогнозирования и реализации алгоритма могут быть найдены в [12].

3. Экспериментальные результаты

Известно, что статистические методы прогнозирования часто используются вместе с предварительным преобразованием временного ряда. Сочетание данного подхода и исследуемого метода может повысить точность прогноза. В данном разделе была выполнена экспериментальная оценка метода прогнозирования на основе универсальной меры R , описанной в предыдущем разделе, в сочетании с предварительным дифференцированием временного ряда. Под операцией дифференцирования будем понимать преобразование исходной последовательности $x_1 \dots x_t$ в последовательность $\varphi_1 \dots \varphi_{t-1}$, где $\varphi_i = x_{i+1} - x_i$, $i = 1, 2, \dots, t-1$. Прогнозное значение для преобразованного временного ряда можно прибавить к последнему известному члену исходной последовательности и получить итоговый прогноз значения x_{t+1} .

3.1. Прогнозирование показателей уровня моря

В качестве объектов исследования были рассмотрены 15-минутные показатели уровня моря. Все наборы данных, использовавшиеся в экспериментах, могут быть найдены на Интернет-сайте Британского океанографического центра данных (далее – БОЦД), в разделе «UK Tide Gauge Network» [13]. Веб-интерфейс сайта позволяет любому зарегистрированному пользователю загрузить файл с историей данных о 15-минутных показателях уровня моря. В этом файле для каждой временной отметки некоторого периода времени приводится реально зафиксированное значение, а также величина, вычисленная как разность между прогнозом и зафиксированным значением (далее – «разность»). Согласно предоставленной службой поддержки БОЦД информации, прогнозные значения вычисляются в Национальном океанографическом центре Соединенного Королевства (далее – НОЦСК) на основе гармонического анализа приливов, который выполняется с помощью специальной компьютерной программы. Мы рассматривали данные, зафиксированные контрольно-измерительным прибором, установленным в Дувре, в 2005 году, так как в данных за этот период отсутствуют пропущенные и интерполированные значения. Суть каждого проводимого на данном этапе эксперимента заключалась в следующем. Имея в распоряжении только данные об t последовательных значениях временного ряда, требовалось спрогнозировать значение его $(t+1)$ -го члена. Для того чтобы иметь возможность сравнить результаты, полученные с помощью рассматриваемого метода, с соответствующими реально зафиксированными величинами процесса, как период основания, так и период упреждения прогноза были взяты из уже прошедшего временного интервала. В ходе вычислений по каждому выбранному процессу для анализа зависимости точности прогноза от длины последовательности известных значений использовались временные ряды различной продолжительности. Кроме того, с целью изучения возможности оптимизации вычислений, дополнительно проводились расчёты в адаптивном режиме. (Подробное описание данного метода см. в предыдущем разделе). Следует отметить, что в качестве размеров «окон» при вычислениях в адаптивном режиме для каждого процесса были взяты все длины, используемые для данного временного ряда при вычислениях в обычном режиме, а значения вероятностей при «склеивании» были одинаковыми. Для каждой длины временного ряда определённого процесса, а также при соответствующих вычислениях в адаптивном режиме, проводилось 30 экспериментов над независимыми наборами данных. Результаты вычислений содержатся в табл. 1.

Таблица 1. Результаты прогнозирования 15-минутных показателей уровня моря

Метод	500	1000	2000	5000	Адаптив.
R	0.034	0.038	0.0396	0.037	0.037
NOC	0.207	0.2133	0.09883	0.0779	н.д.

Первая колонка содержит названия методов, на основе которых строился прогноз. Здесь R – это исследуемый метод на основе универсальной меры, а NOC – это метод НОЦСК. Оставшиеся колонки содержат значения средней абсолютной погрешности, полученные либо при использовании соответствующей длины временного ряда, либо при вычислениях в адаптивном режиме. Следует отметить, что для прогнозов NOC данная величина вычислялась на основе «разностей», расположенных в файле истории для значений, которые прогнозировались исследуемым методом. Так, на основе таблицы можно сказать, что для прогнозов методом, основанным на универсальной мере, с использованием 5000 известных значений временного ряда величина средней абсолютной погрешности для 30 экспериментов составила 0.037. А величина средней абсолютной погрешности, вычисленной на основе «разностей», предоставленных НОЦСК для 30 спрогнозированных исследуемым методом значений, составляет 0.0779. Надпись «н.д.» в ячейке означает «нет данных».

3.2. Прогнозирование кросс-курсов валют

Простейшим способом оценки точности метода прогнозирования является его сравнение с так называемым инерционным методом, в котором последнее известное значение временного ряда считается прогнозным. В данной части работы мы проведем указанное сравнение на примере прогнозирования ежедневных кросс-курсов доллара США (USD) и британского фунта стерлинга (GBP) по отношению к евро (EUR). Кроме того, в данном подразделе было выполнено сравнение точности прогнозов, полученных исследуемым методом и программой прогнозирования Autobox 6.0 в автоматическом режиме. Данные о значениях кросс-курсов были взяты за период с 03 января 2001 по 17 января 2011. Все данные, использованные в вычислениях, были взяты с сайта fxhistoricaldata.com [14]. Здесь также рассматривалась задача краткосрочного прогнозирования на один шаг вперед. Вычисления проводились аналогичным способом из предыдущего подраздела образом. Для каждой длины временного ряда проводилось 10 экспериментов над независимыми наборами данных. Результаты сравнения содержатся в табл. 2 и табл. 3.

Таблица 2. Прогнозирование кросс-курса USD/EUR

Метод	500	1000
R	0.0030	0.0058
Инерционный	0.0030	0.0059
Autobox 6.0	0.0142	0.0191

Таблица 3. Прогнозирование кросс-курса GBP/EUR

Метод	500	1000
R	0.0019	0.0009
Инерционный	0.0025	0.0010
Autobox 6.0	0.0195	0.0012

Таблицы имеют аналогичную структуру. Первая колонка содержит название метода прогнозирования, а в оставшихся колонках расположены значения средней абсолютной погрешности, полученные при использовании соответствующей длины временного ряда. Надпись «н.д.» в ячейке означает «нет данных». На основе полученных данных можно сказать, что в целом метод прогнозирования на основе универсальной меры оказался более точным, чем инерционное прогнозирование и программа Autobox 6.0 в автоматическом режиме.

7. Заключение

В данной статье рассматривались описание и экспериментальная оценка метода прогнозирования, основанного на универсальной мере R . Анализ результатов исследования показал достаточно высокую точность полученных результатов. При прогнозировании показателей уровня моря были выявлены параметры, при которых наблюдается устойчивое преимущество рассмотренной реализации метода над программой прогнозирования НОЦСК. Кроме того, проведенное для временных рядов, содержащих данные о кросс-курсах некоторых валют, сравнение с инерционным методом и программой Autobox 6.0 показало более высокую точность исследуемого метода.

Литература

1. De Gooijer J.G., Hyndman R.J. 25 years of time series forecasting // International Journal of Forecasting. 2006. V. 22, № 3. P. 443-473
2. Фитингоф Б. М. Оптимальное кодирование при неизвестной и меняющейся статистике сообщений // Проблемы передачи информации. 1966. Т. 2, № 2. С. 3-11.
3. Колмогоров А. Н. Три подхода к определению понятия “количество информации” // Проблемы передачи информации. 1965. Т. 1, № 1, P. 3-11.
4. Рябко Б. Я. Дважды универсальное кодирование // Проблемы передачи информации. 1984. Т. 20, № 3. С. 24-28.
5. Rissanen J. Universal coding, information, prediction, and estimation // IEEE Transactions on Information Theory. 1984. V. 30, № 4. P. 629-636.
6. Рябко Б. Я. Прогноз случайных последовательностей и универсальное кодирование // Проблемы передачи информации. 1988. Т. 24, № 2. P. 3-14.
7. Ryabko B., Astola J. Universal codes as a basis for time series testing // Statistical Methodology. 2006. V. 3, № 4, P. 375-397.
8. Ryabko B. Compression-based methods for nonparametric prediction and estimation of some characteristics of time series // IEEE Transactions on Information Theory. 2009. V. 55, № 9. P. 4309-4315.
9. Ryabko B., Monarev V. Using information theory approach to randomness testing // Journal of Statistical Planning and Inference. 2005. V. 133, № 1. P. 95-110.
10. Krichevsky R. Universal Compression and Retrieval. Springer, 1 edition, 1994. 236 p.
11. Ryabko B. Applications of Universal Source Coding to Statistical Analysis of Time Series. In: Isaac Woungang et al. (Eds.), Selected Topics in Information and Coding Theory, World Scientific Publishing, 2010. P. 289 - 338.
12. Приставка П., Экспериментальное исследование метода прогнозирования, основанного на универсальных кодах // Вестник СибГУТИ. 2010. № 4. С. 26-35.
13. British Oceanographic Data Centre. UK National Tide Gauge Network. [Электронный ресурс]. URL : https://www.bodc.ac.uk/data/online_delivery/nts1f/ (дата обращения: 05.04.2011).
14. FXHISTORICALDATA.COM [Электронный ресурс]. URL: <http://www.fx-historicaldata.com/> (дата обращения: 05.04.2011).

*Статья поступила в редакцию 05.02.2011;
переработанный вариант — 18.05.2011.*

Приставка Павел Анатольевич

Ассистент кафедры прикладной математики и кибернетики СибГУТИ, e-mail: pra@ngs.ru

Application of information-theoretic forecasting methods to the analysis of geophysical and economic data

P. A. Pristavka

We describe and experimentally investigate a method to construct forecasting algorithms for stationary and ergodic processes based on universal measures (or so-called universal data compressors). Using some geophysical and economical time series as examples, we show that the precision of thus obtained predictions is higher than that of known methods.

Keywords: time series, nonparametric methods, universal measure, universal coding, sea level, cross-rates.