УДК 519.246.8
DOI 10.25205/1818-7900-2018-16-3-145-158

К. С. Чирихин [1, 2], **Б. Я. Рябко** [1, 3]

[1] *Новосибирский государственный университет*
ул. Пирогова, 1, Новосибирск, 630090, Россия

[2] *Сибирский государственный университет телекоммуникаций и информатики*
ул. Кирова, 86, Новосибирск, 630102, Россия

[3] *Институт вычислительных технологий СО РАН*
пр. Академика Лаврентьева, 6, Новосибирск, 630090, Россия

chirihin@gmail.com

ЭКСПЕРИМЕНТАЛЬНОЕ ИССЛЕДОВАНИЕ ТОЧНОСТИ МЕТОДОВ ПРОГНОЗА, БАЗИРУЮЩИХСЯ НА АРХИВАТОРАХ

В теории информации известно, что методы сжатия данных могут быть использованы для прогнозирования стационарных процессов. В данной работе предложен базирующийся на архиваторах алгоритм прогнозирования временных рядов и проведено экспериментальное исследование его эффективности. В процессе работы описанного алгоритма могут быть использованы произвольные методы сжатия данных, причем прогнозные значения от разных методов комбинируются, и наибольшее влияние на конечный результат оказывает метод, способный сильнее других сжать временной ряд. Данный алгоритм может быть использован для прогнозирования рядов с дискретными и непрерывными алфавитами. Для повышения точности прогноза возможно применение существующих методов предварительной обработки данных. Экспериментальное исследование эффективности предложенного алгоритма проводилось на временных рядах из M3 Competition и ряде Т-индекса, при этом были использованы хорошо известные архиваторы. Результаты вычислений показали, что полученный метод обладает сравнительно высокой точностью и быстродействием.

Ключевые слова: универсальное кодирование, прогнозирование временных рядов.

Введение

Задача прогнозирования привлекает внимание многих исследователей в силу ее большой практической значимости. Например, существует большое количество приложений в экономике (можно построить прогноз для уровня безработицы, объемов промышленного производства и т. д.), геофизике (прогнозирование числа солнечных пятен, уровня моря) и во многих других областях человеческой деятельности. Математически временной ряд может быть описан как случайный процесс с дискретным временем [1], значения которого, как правило, находятся на равном расстоянии друг от друга. При прогнозировании временного ряда требуется оценить значения процесса в нескольких будущих моментах времени на основании имеющейся в наличии его предыстории.

Для решения описанной задачи разработано большое количество разнообразных методов – как статистических, так и из области машинного обучения. К наиболее распространенным можно отнести экспоненциальное сглаживание, модель авторегрессии – скользящего

Чирихин К. С., *Рябко Б. Я.* Экспериментальное исследование точности методов прогноза, базирующихся на архиваторах // Вестн. НГУ. Серия: Информационные технологии. 2018. Т. 16, № 3. С. 145–158.

среднего и различные ее модификации, а также экспертные системы и нейронные сети [2; 3]. Тем не менее задача построения высокоточного прогноза еще далека от разрешения.

Поскольку между сжимаемостью последовательности и ее вероятностью существует тесная связь, одним из возможных подходов к решению данной задачи является использование методов сжатия данных. В статье [4] было показано, что любой архиватор может быть использован для прогнозирования. Важно отметить, что в архиваторах, помимо теоретических результатов, используются различные эвристики, повышающие их способность улавливать закономерности во встречающихся на практике данных и улучшающие степень сжатия.

В данной статье разрабатывается и исследуется метод прогнозирования временных рядов, основанный на распространенных архиваторах и библиотеках для сжатия данных (таких, как библиотека zlib [1], лежащая в основе архиватора gzip, библиотека ppmd [5], используемая среди прочих алгоритмов в 7-z, и др.). Для экспериментальной оценки эффективности метода были проведены расчеты для временных рядов из известного исследования M3-Competition [3] (далее M3C), основной целью которого было сравнение точности различных методов прогнозирования на реальных данных преимущественно из социально-экономической сферы. Проведено сравнение предлагаемого метода с методами, участвовавшими в этом исследовании. Еще одним рядом, результаты вычислений для которого приводятся в данной статье, является временной ряд T-индекса [2], тесно связанный с солнечными пятнами.

Результаты экспериментального исследования показывают, что описываемый метод обладает точностью, сравнимой с другими методами прогнозирования, и представляет практический интерес.

Прогнозирование с помощью методов сжатия данных для случая конечного алфавита

Сначала покажем, как связаны сжатие данных и прогнозирование. Пусть $X = x_1, x_2, ..., x_t$ – временной ряд (или в терминах теории информации – передаваемое сообщение), порожденный некоторым вероятностным источником. Все члены временного ряда x_i принадлежат конечному множеству A, называемому *алфавитом*. В теории информации хорошо известно, что для любого разделимого кода выполняется неравенство Крафта – Макмиллана [6]:

$$\sum_{X \in A^t} 2^{-|\phi(X)|} \le 1, \tag{1}$$

где A^t – множество всевозможных последовательностей длиной t над алфавитом A, $|\phi(X)|$ – длина кодового слова для сообщения X при кодировании по методу ϕ.

Величину $2^{-|\phi(X)|}$ далее для краткости будем называть *кодовой вероятностью* сообщения X.

В работе [4] было предложено использовать неравенство (1) для задания распределения вероятностей на множестве кодируемых сообщений:

$$P_\phi(X) = \frac{2^{-|\phi(X)|}}{\sum_{Y \in A^t} 2^{-|\phi(X)|}}.$$

Условная вероятность того, что следующие h символов $x_{t+1}, x_{t+2}, ..., x_{t+h}$ будут равны соответственно $a_1, a_2, ..., a_h$, $a_i \in A$, может быть найдена с использованием имеющейся предыстории по формуле

$$P_\phi\left(x_{t+1} = a_1, x_{t+2} = a_2, ..., x_{t+h} = a_h \middle| x_1, x_2, ..., x_t\right) =$$

[1] Zlib Home Site URL: https://zlib.net (дата обращения 23.03.2018).

[2] T Index FAQ. Australian Government/Bureau of Meteorology. URL: http://www.sws.bom.gov.au/Educational/5/2/1 (дата обращения 23.03.2018).

$$= \frac{2^{-|\phi(x_1,x_2,...,x_t,a_1,a_2,...,a_h)|}}{\sum_{Y\in A^h} 2^{-|\phi(x_1,x_2,...,x_t,y_1,y_2,...,y_h)|}}. \qquad (2)$$

Пример 1. Приведем пример вычислений по формуле (2). Построим прогноз на два шага вперед для последовательности 0110011001 над алфавитом $\{0,1\}$. Будем поочередно дописывать в конец данной последовательности различные комбинации длиной 2 из нулей и единиц и сжимать ее каким-либо архиватором. Например, воспользуемся библиотекой zlib версии 1.2.11, в частности, ее функцией compress2 с флагом Z_BEST_COMPRESSION. Получившиеся длины кодовых слов и дальнейшие вычисления приведены в табл. 1. Жирным шрифтом в столбце 1 выделены два «дописанных» символа. Последовательности в программе на языке C++ были представлены как массивы типа unsigned char, на каждый символ последовательностей отводился 1 байт (единица была представлена байтом со значением 1_{10}, а нуль – со значением 0_{10}).

Таблица 1

Пример построения распределения вероятностей для следующих двух элементов временного ряда

Последовательность	Размер сжатого представления, бит	Кодовая вероятность	Вероятность
011001100**100**	128	2.939E-39	1.520E-5
011001100**101**	128	2.939E-39	1.520E-5
011001100**110**	112	1.926E-34	9.961E-1
011001100**111**	120	7.523E-37	3.891E-3
Сумма	–	1.933E-34	1.000

Как видно из табл. 1, согласно данному методу, следующими двумя символами будут 10 с вероятностью 0.996. Таким образом, архиватор смог успешно выявить закономерность даже на короткой последовательности.

Более подробное изложение метода, описанного в данном и следующем разделах, а также обоснование приведенных в них формул можно найти в книге [7].

Прогнозирование временных рядов с непрерывным алфавитом

Описанный в предыдущем разделе метод применим только для временных рядов с конечным алфавитом, но на практике чаще всего встречаются ряды, у которых алфавитом является некоторый отрезок вещественной прямой. В таких случаях необходимо перейти от подобного отрезка к конечному алфавиту. Это можно сделать путем разбиения отрезка на k непересекающихся равных интервалов с номерами $A = \{0, 1, ..., k-1\}$ (обозначим интервал с номером i через q_i), последующего преобразования временного ряда в ряд номеров интервалов и прогнозирования номеров для следующих значений. Предположим, что требуется построить прогноз на h шагов вперед. Тогда вероятность того, что в момент времени i, $1 \le i \le h$, значение процесса попадет в интервал с номером $j \in A$, может быть получена из маргинального распределения вероятностей по совместному распределению номеров, вычисленному по формуле (2):

$$P_\phi\left(x_{t+i} \in q_j\right) = \sum_{a_1,...,a_{i-1},a_{i+1},...,a_h \in A} P_\phi\left(a_1,...,a_{i-1},j,a_{i+1},...,a_h\right). \qquad (3)$$

В качестве точечного прогноза можно использовать математическое ожидание случайной величины, значениями которой являются середины интервалов, и их вероятности задаются по формуле (3).

Рассмотрим далее вопрос о выборе количества интервалов *k*. Проблема заключается в том, что при малых *k* точность прогноза может оказаться низкой из-за грубого округления, а при больших – из-за слишком короткой предыстории процесса и присутствия шумов в данных. Кроме того, с увеличением *k* экспоненциально возрастает время вычислений. В данной работе был использован подход, описанный в статье [8]. Отрезок, из которого принимают значения члены временного ряда, разбивается последовательно на 2^i интервалов, $i = 1, 2, ..., n$, $k = 2^n$. Для каждого *i* прогноз строится независимо, и затем прогнозы взвешиваются. Пусть x_i – член исходного временного ряда, а $x_i^{[j]}$ – соответствующий ему номер интервала при разбиении исходного отрезка на 2^j интервалов. Взвешивание можно провести по следующей формуле:

$$P_\phi\left(x_1^{[n]}, x_2^{[n]}, ..., x_t^{[n]}\right) = \frac{\sum_{i=1}^{n} \omega_i 2^{-\left|\phi\left(x_1^{[n]}, x_2^{[n]}, ..., x_t^{[n]}\right)\right| + t(n-i)}}{\sum_{i=1}^{n} \sum_{Y \in N_i^t} \omega_i 2^{-|\phi(Y)| + t(n-i)}}, \quad (4)$$

где

$N_i = \left\{0, 1, ..., 2^i - 1\right\}$ – алфавит, состоящий из номеров интервалов;

$\omega_i \geq 0$, $\sum_{i=1}^{n} \omega_i = 1$ – весовые коэффициенты, в данной работе вычислялись по формуле

$$\omega_i = \begin{cases} \frac{1}{i} - \frac{1}{i+1}, i \neq n, \\ \frac{1}{n+1}, i = n. \end{cases}$$

Поясним, зачем прибавлять величину $t(n-1)$ к длине кодового слова в формуле (4). Без нее нельзя сравнивать длины кодовых слов для сообщений из разных алфавитов. Предположим, что область значений временного ряда поочередно разбивалась на 2 и 4 интервала (разбиения 1 и 2 соответственно). В данном случае $n = \log_2 4 = 2$. Заметим, что каждому интервалу разбиения 1 соответствует 2 интервала разбиения 2 (нулю соответствуют нуль и единица, единице – два и три). Для того чтобы сообщить, в какой из двух возможных интервалов разбиения 2 попадает элемент из ряда с разбиением 1, требуется 1 бит. Поскольку всего элементов *t*, требуется добавить $t(2-1) = t$ бит к ряду с разбиением 1.

Для того чтобы избежать больших отрицательных степеней при расчетах по формуле (4), после того, как длины всех кодовых слов будут известны, можно вычесть наименьшую из них из каждого кодового слова. Данная идея будет проиллюстрирована в примере 2.

Далее покажем, как можно взвесить прогнозы, полученные по различным архиваторам. Пусть имеется *n* методов сжатия данных (архиваторов) $\phi_1, \phi_2, ..., \phi_n$. Можно скомбинировать прогнозы, полученные от каждого архиватора в отдельности, в общий прогноз по следующей формуле:

$$P_{\phi_1, \phi_2, ..., \phi_n}\left(X \middle| Y\right) = \frac{\sum_{i=1}^{n} \gamma_i 2^{-|\phi_i(YX)|}}{\sum_{Z \in N^h} \sum_{i=1}^{n} \gamma_i 2^{-|\phi_i(YX)|}}, \quad (5)$$

где

$\gamma_i \geq 0$, $\sum_{i=1}^{n} \gamma_i = 1$ – весовые коэффициенты;

$Y = y_1, y_2, ..., y_t$ – данный временной ряд;

$X = x_1, x_2, ..., x_h$ – одно из возможных продолжений ряда;

YX – ряд, полученный путем дописывания ряда X в конец ряда Y.

В данной работе при вычислениях по формуле (5) были использованы равные веса: $\gamma_i = \frac{1}{n}$.

Отметим, что формулы (4) и (5) можно использовать совместно. В вычислениях, приводимых в данной статье, сначала оценки вероятностей, полученные при различных мощностях разбиений, взвешивались по формуле (4), и тем самым получалось единственное распределение интервалов (содержащее номера интервалов из наибольшего рассматриваемого разбиения) для каждого используемого архиватора. Затем распределения от разных архиваторов объединялись в одно по формуле (5).

Пример 2. Рассмотрим, как построить прогноз для временного ряда с непрерывным алфавитом при помощи двух архиваторов. Пусть дан следующий временной ряд:

3.4	0.1	3.9	4.8	1.5	1.8	2.0	4.9	5.1	2.1

Требуется построить по нему прогноз на два шага вперед. Для этого воспользуемся библиотекой zlib, реализацией алгоритма сжатия Re-Pair [9] и формулой (5). Сначала перейдем к конечному алфавиту. Разобьем интервал, в который попадают все значения временного ряда, на 2 и 4 части одинаковой длины, построим прогнозы независимо для этих разбиений, а затем скомбинируем их по формуле (5).

Наименьшее значение в рассматриваемом временном ряде $m = 0.1$, наибольшее $M = 5.1$. Хотя все значения ряда попадают в отрезок $[0.1; 5.1]$, отступим на некоторую величину от крайних значений (поскольку следующее значение может быть больше (меньше) предыдущего максимума (минимума)). При расчетах, результаты которых приводятся в данной статье далее, отступ составлял 10 % от ширины интервала, поэтому в данном примере поступим также. Получим $m' = m - 0.1(M - m) = -0.4$, $M' = M + 0.1(M - m) = 5.6$. При разбиении на два интервала одинаковой длины в случае, если значение временного ряда меньше чем 2.6, будем считать, что оно попало в интервал с номером 0, иначе в интервал с номером 1. Получим следующий временной ряд:

3.4	0.1	3.9	4.8	1.5	1.8	2.0	4.9	5.1	2.1
1	0	1	1	0	0	0	1	1	0

Будем последовательно дописывать к ряду всевозможные комбинации из символов $\{0,1\}$ и сжимать получающиеся сообщения. Размеры сжатых файлов и соответствующие вычисления вероятностей приведены в табл. 2. При «склеивании» кодовых вероятностей, полученных от zlib и Re-Pair, были использованы веса $\gamma_1 = \gamma_2 = 0.5$.

Проведем аналогичные вычисления для разбиения области значений ряда на 4 интервала. Получим следующий временной ряд:

3.4	0.1	3.9	4.8	1.5	1.8	2.0	4.9	5.1	2.1
2	0	2	3	1	1	1	3	3	1

Интервалу с номером 0 при разбиении на 2 интервала соответствуют интервалы 0, а при разбиении на 4 интервала – 1, соответственно интервалу с номером 1 – интервалы 2 и 3. Поэтому, например, прогнозному значению 13 при разбиении на 4 интервала соответствует прогнозное значение 01 при разбиении на два интервала.

К значениям в пятом и шестом столбцах табл. 2 было прибавлено 12 бит, поскольку длина t каждой из последовательностей равна 12, и $\log_2 4 - \log_2 2 = 1$. В данной таблице наименьшая длина сжатого сообщения составила 100 бит, для уменьшения риска переполнения при вычислениях с плавающей точкой эту величину можно вычесть из всех значений в столбцах 2, 3, 5, 6 табл. 2.

Таблица 2

Результаты вычислений при разбиении на два и четыре интервала

Сообщение (4 интервала)	Размер сжатого сообщения, бит		Сообщение (2 интервала)	Размер сжатого сообщения, бит	
	zlib	Re-Pair		zlib	Re-Pair
202311331**00**	160	112	101100011**00**	120 + 12	104 + 12
202311331**01**	160	112			
202311331**10**	144	120			
202311331**11**	136	104			
202311331**02**	160	120	101100011**01**	128 + 12	88 + 12
202311331**03**	160	112			
202311331**12**	144	120			
202311331**13**	144	120			
202311331**20**	160	120	101100011**10**	136 + 12	88 + 12
202311331**21**	160	112			
202311331**30**	160	112			
202311331**31**	160	112			
202311331**22**	160	112	101100011**11**	136 + 12	88 + 12
202311331**23**	160	120			
202311331**32**	160	112			
202311331**33**	144	112			

Далее нужно перейти к кодовым вероятностям, а затем взвесить и нормировать их. Окончание расчетов приведено в табл. 3.

Теперь построим прогноз на первый и второй шаги. Просуммировав вероятности в строках таблицы, в которых первый дописанный символ равен 0, получим вероятность того, что следующее значение временного ряда попадет в интервал с номером 0. Аналогичным образом вычисляются вероятности для интервалов с номерами 1, 2 и 3. Распределения вероятностей номеров интервалов, а также середины соответствующих интервалов, приведены в табл. 4. В качестве прогнозных значений используются математические ожидания, значения которых также приведены в таблице.

Уменьшение трудоемкости алгоритма

Обозначим через $|A|$ мощность алфавита источника (максимальная мощность разбиения в случае непрерывного алфавита). Если требуется построить прогноз на h шагов, то потребуется сжать $|A|^h$ последовательностей. В то же время в МЗС для ежемесячных данных требовалось вычислять прогноз на 18 шагов вперед, что приводит к необходимости сжатия $2^{18} = 262144$ последовательностей при минимально возможной мощности алфавита $|A| = 2$. Сократить время вычислений позволяет следующий подход. Предположим, что h является четным. Удалим из ряда все члены с четными номерами и построим прогноз на $h/2$ шагов вперед с нечетными номерами. Затем удалим из исходного ряда все элементы с нечетными номерами и построим прогноз для недостающих $h/2$ элементов с четными номерами. В результате вместо $|A|^h$ возможных продолжений ряда получим $2|A|^{h/2}$ вариантов. Для повышения точности можно построить прогноз на первые $h/2$ шагов по полному ряду. Данный метод легко можно обобщить. К примеру, если h кратно 6, то на первом этапе нужно оставить только члены ряда с номерами i, для которых $i \bmod 6 = 0,$ и вычислить прогноз на $h/6$ шагов вперед и т. д.

Таблица 3

Окончание расчетов из примера 2

Сообщение (4 интервала)	Кодовая вероятность (4 интервала)		Сообщение (2 интервала)	Кодовая вероятность (2 интервала)		Кодовая вероятность после «склеивания»	Вероятность после «склеивания»
	zlib	Re-Pair		zlib	Re-Pair		
202311133100	2.939E-39	2.441E-04	10110001100	2.328E-10	1.562E-05	6.485E-05	2.150E-05
202311133101	2.939E-39	2.441E-04				6.485E-05	2.150E-05
202311133110	1.926E-34	9.537E-07				4.053E-06	1.344E-06
202311133111	4.930E-32	0.063				0.016	0.005
202311133102	2.939E-39	9.537E-07	10110001101	9.095E-13	1.000	0.250	0.083
202311133103	2.939E-39	2.441E-04				0.250	0.083
202311133112	1.926E-34	9.537E-07				0.250	0.083
202311133113	1.926E-34	9.537E-07				0.250	0.083
202311133120	2.939E-39	9.537E-07	10110001110	3.553E-15	1.000	0.250	0.083
202311133121	2.939E-39	2.441E-04				0.250	0.083
202311133130	2.939E-39	2.441E-04				0.250	0.083
202311133131	2.939E-39	2.441E-04				0.250	0.083
202311133122	2.939E-39	2.441E-04	10110001111	3.553E-15	1.000	0.250	0.083
202311133123	2.939E-39	9.537E-07				0.250	0.083
202311133132	2.939E-39	2.441E-04				0.250	0.083
202311133133	1.926E-34	2.441E-04				0.250	0.083
Сумма	–		–	–		3.016	1.000

Таблица 4

Вычисление прогнозных значений на два шага вперед для ряда из примера 2

Номер интервала	Маргинальная вероятность		Середина интервала
	шаг 1	шаг 2	
0	0.166	0.166	0.35
1	0.171	0.171	1.85
2	0.332	0.332	3.35
3	0.332	0.332	4.85
Математическое ожидание	3.093	3.093	–

Используемые архиваторы

При проведении экспериментальных расчетов, результаты которых приведены в данной статье, были использованы библиотеки zlib, ppmd [3] и реализация алгоритма Re-Pair [9]. В их основе лежат разные идеи. В zlib реализована схема Deflate, в которой используется алгоритм LZ-77 совместно с кодом Хаффмана. Библиотека ppmd является реализацией алгоритма PPM (Prediction by Partial Matching, предсказание по частичному совпадению). Re-Pair – алгоритм сжатия данных, основанный на грамматической модели. В процессе работы этого алгоритма выполняется построение контекстно-свободной грамматики, задающей язык из единственной цепочки – сжимаемого файла. При этом сжатие достигается за счет замены повторяющихся фрагментов во входной последовательности на нетерминальные символы.

Экспериментальное исследование

В рамках данной работы была выполнена программная реализация описанного метода и проведены вычисления для набора временных рядов из исследования M3-Competition [3], а также временного ряда Т-индекса. Соревнование M3-Competition проводилось в 2000 г., и основной его целью являлось сравнение точности методов прогнозирования на реальных данных. В нем были представлены 3 003 временных ряда, длина которых составляла от 14 до 144 наблюдений. Присутствовали ежегодные, ежемесячные и ежеквартальные данные, а также небольшое количество рядов, не относящихся ни к одной из вышеперечисленных категорий (категория «другие»). Требовалось построить прогноз на 6 шагов для ежегодных данных, на 8 шагов для ежеквартальных данных и данных из категории «другие», на 18 шагов для ежемесячных данных. На веб-сайте Международного института прогнозистов [4] размещены временные ряды из M3C и реально зафиксированные значения прогнозируемых величин, поэтому имеется возможность провести аналогичные вычисления и оценить качество построенного прогноза.

В данной статье используется один из способов оценки точности в M3C – симметричная средняя абсолютная процентная ошибка (symmetric mean absolute percentage error, sMAPE), определяемая по следующей формуле:

$$\text{sMAPE}\left(\hat{X}, X\right) = \sum_{i=1}^{h} \frac{\left|\hat{x}_i - x_i\right|}{\left(\hat{x}_i + x_i\right)/2} \cdot 100, \tag{6}$$

где

$\hat{X} = \hat{x}_1, \hat{x}_2, ..., \hat{x}_h$ – ряд прогнозных значений;

$X = x_1, x_2, ..., x_h$ – ряд зафиксированных значений.

Поскольку ни один временной ряд из M3C не содержит отрицательных значений, величина, вычисляемая по формуле (6), не может быть отрицательной.

При вычислениях по описываемому в данной статье алгоритму использовалась предварительная обработка данных. В частности, для ежемесячных и ежеквартальных данных использовалось выделение сезонной компоненты временного ряда (была использована реализация [5] метода STL [10]). Затем выполнялось построение прогноза для тренда и случайной составляющей, а сезонная компонента принималась постоянной. Кроме того, для всех категорий временных рядов осуществлялся переход к первой разности (первой разностью ряда $x_1, x_2, ..., x_t$ называется ряд $x_2 - x_1, x_3 - x_2, ..., x_t - x_{t-1}$) и применялось сглаживание по формуле

[3] ppmd_sh URL: https://github.com/Shelwien/ppmd_sh (дата обращения 23.03.2018).

[4] M3-Competition/International Institute of Forecasters URL: https://forecasters.org/resources/time-series-data/m3-competition (дата обращения 23.03.2018).

[5] URL: http://stat.ethz.ch/R-manual/R-devel/library/stats/html/stl.html (дата обращения 23.03.2018).

$x_i = \left(2x_i + x_{i-1} + x_{i-2}\right)/4$. В процессе экспериментального исследования была предпринята попытка выбора такого порядка разности, при котором среднеквадратичное отклонение для ряда будет наименьшим. За исключением одного случая (ежемесячные данные) такой подход приводил к небольшому ухудшению точности прогноза по сравнению с постоянным выбором первой разности.

Поскольку при вычислениях возникает потребность в работе с очень маленькими числами с плавающей точкой, в программной реализации описанного метода была использована библиотека для арифметики с высокой точностью [6].

При построении всех прогнозов применялась ранее описанная процедура оптимизации вычислений при помощи прореживания ряда. При прогнозировании ежегодных, ежеквартальных и других данных оставлялся каждый второй элемент временного ряда, при прогнозировании ежемесячных – каждый шестой. Выбор таких значений был обусловлен соображениями трудоемкости: при прогнозировании ежегодных данных для каждого ряда дважды строился прогноз на 3 шага вперед, ежемесячных данных – шесть раз на 3 шага вперед и т. д.

Результаты вычислений приведены в табл. 5–8. Формат таблиц близок к таблицам из [3]. В таблицах приводятся данные, полученные по различным архиваторам и некоторым их комбинациям. В строках «Лучший М3С» и «Худший М3С» содержатся соответственно наименьшие и наибольшие ошибки на каждый шаг среди всех участвовавших в М3С архиваторов.

В двух из четырех таблиц метод прогнозирования на основе архиваторов показал точность, сравнимую с другими методами. На ежеквартальных и ежемесячных данных его точность оказалась сравнительно низкой, но, тем не менее, при прогнозировании на первые четыре шага сопоставимой с методами из М3С. При этом увеличение количества интервалов с 16 до 32 ни в одном случае не привело к существенному повышению точности прогноза.

Далее приведем результаты вычислений для ряда T-индекса, который рассчитывается метеорологическим бюро Австралии и тесно связан с солнечными пятнами. Временной ряд ежемесячных значений T-индекса можно найти на веб-сайте http://listserver.ips.gov.au/mailman/listinfo/ips-tindex-predictions. На этом же сайте размещается прогноз T-индекса (на данный момент доступен прогноз на 44 шага вперед), а также архивные данные. В рамках нашей работы для каждого из месяцев с апреля 2011 по апрель 2016 г. был построен прогноз на 18 шагов вперед (в этом временном интервале отсутствуют пропущенные значения), и проведено сравнение его точности с прогнозом метеорологического бюро.

В данном ряде, в отличие от рядов из М3С, встречаются отрицательные значения, поэтому точность прогноза оценивалась по средней абсолютной ошибке (mean absolute error, MAE), вычисляемой по следующей формуле:

$$\mathrm{MAE}\left(\hat{X}, X\right) = \frac{1}{h}\sum_{i=1}^{h}\left|\hat{x}_i - x_i\right|, \tag{7}$$

где

$\hat{X} = \hat{x}_1, \hat{x}_2, ..., \hat{x}_h$ – ряд прогнозных значений;

$X = x_1, x_2, ..., x_h$ – ряд зафиксированных значений.

Вычисления были проведены следующим образом. На основании данных из файла за некоторый месяц (например, январь 2014 г.) строился прогноз на 18 шагов вперед. Затем по более позднему файлу на 18 месяцев по формуле (7) рассчитывалась ошибка вычисленного прогноза и прогноза, содержащегося в исходном файле (за январь 2014 г. в данном примере, в каждом файле помимо зафиксированных значений приводятся прогнозные значения). Результаты расчетов приведены в табл. 9. Как видно из этой таблицы, при прогнозировании на 1 шаг метод на основе архиваторов оказался точнее метода метеорологической службы, в остальных случаях точность прогноза метеорологической службы оказалась выше.

[6] Bignum C++ library URL: https://www.ttmath.org/ (дата обращения 23.03.2018)

Таблица 5

Результаты прогнозирования ежегодных данных из M3C

Метод	Ошибка sMAPE для номера шага *h*						Среднее		Количество рядов
	1	2	3	4	5	6	1–4	1–6	
Лучший M3C	7.6	12.1	16.1	18.2	20.8	22.7	13.65	16.42	645
Худший M3C	10.7	15.2	20.8	24.1	28.1	31.2	17.57	21.59	645
zlib (16 интервалов)	9.9	14.9	20.8	22.9	28.0	28.2	17.13	20.79	645
ppmd (16 интервалов)	10.1	14.5	20.6	22.4	27.3	27.2	16.76	20.25	645
rp (16 интервалов)	10.5	14.9	19.5	21.9	26.2	27.4	16.70	20.06	645
zlib + ppmd + rp (16 интервалов)	10.4	14.5	19.4	21.8	26.5	27.3	16.51	19.97	645
zlib + ppmd + rp (32 интервала)	10.3	14.5	19.3	21.8	26.4	27.3	16.49	19.95	645

Таблица 6

Результаты прогнозирования ежеквартальных данных из M3C

Метод	Ошибка sMAPE для номера шага *h*							Среднее			Количество рядов
	1	2	3	4	5	6	8	1–4	1–6	1–8	
Лучший M3C	4.8	6.6	7.4	8.8	9.4	10.9	12	7	8.04	8.96	756
Худший M3C	7.7	8.9	9.1	10.7	11.8	13.7	15.7	8.86	9.6	10.96	756
zlib (16 интервалов)	5.8	7.6	9.0	11.5	14.0	14.4	17.0	8.47	10.39	12.02	756
ppmd (16 интервалов)	5.8	7.5	8.8	10.6	12.7	13.3	15.4	8.16	9.77	11.12	756
rp (16 интервалов)	6.5	9.0	10.0	12.0	13.5	13.6	15.2	9.35	10.75	11.96	756
zlib + ppmd + rp (16 интервалов)	5.8	7.5	8.8	10.5	13.0	13.2	15.0	8.16	9.80	11.11	756
zlib + ppmd + rp (32 интервала)	5.8	7.5	8.8	10.5	13.0	13.2	14.9	8.16	9.81	11.11	756

Таблица 7

Результаты прогнозирования ежемесячных данных из M3C

Метод	Ошибка sMAPE для номера шага *h*										Среднее			Количество рядов
	1	2	3	4	5	6	8	12	15	18	1–4	1–8	1–18	
Лучший M3C	11.2	10.7	11.7	12.4	11.8	12.2	12.6	13.2	16.2	18.2	11.54	12.06	13.85	1428
Худший M3C	15.3	13.8	15.7	17.0	15.3	15.6	17.4	17.5	22.2	24.3	15.39	15.89	18.4	1428
zlib (16 интервалов)	14.4	14.9	17.3	19.4	20.9	17.9	20.4	19.0	25.6	26.5	16.51	18.42	21.23	1428
ppmd (16 интервалов)	14.0	14.1	15.7	20.2	16.8	20.6	17.7	18.0	24.6	25.5	17.23	18.76	19.95	1428
rp (16 интервалов)	15.6	15.7	17.9	19.7	21.2	18.5	19.1	20.7	26.4	26.9	15.44	17.26	21.55	1428
zlib + ppmd + rp (16 интервалов)	14.0	14.1	15.8	19.6	21.2	18.3	19.1	20.6	25.9	26.7	15.86	18.02	21.13	1428
zlib + ppmd + rp (32 интервала)	14.0	14.1	15.8	19.6	21.2	18.4	19.1	20.6	26.0	26.7	15.86	18.02	21.13	1428
ppmd (16 интервалов, подбор порядка разности)	13.4	13.2	14.7	17.5	19.1	16.4	17.5	17.6	22.3	24.4	14.70	16.52	18.87	1428

Таблица 8

Результаты прогнозирования прочих (other) данных из М3С

Метод	Ошибка sMAPE для номера шага *h*							Среднее			Количество рядов
	1	2	3	4	5	6	8	1–4	1–6	1–8	
Лучший М3С	1.6	2.7	3.8	4.3	5.3	5.1	6.0	3.17	3.86	4.38	174
Худший М3С	2.7	3.8	5.4	6.3	7.8	7.6	9.2	4.38	5.49	6.3	174
zlib (16 интервалов)	2.5	3.4	4.9	6.3	8.3	7.7	8.9	4.25	5.49	6.33	174
ppmd (16 интервалов)	2.4	3.2	4.7	5.1	7.1	6.9	8.2	3.85	4.90	5.68	174
rp (16 интервалов)	2.7	3.9	5.8	6.9	8.0	8.5	10.3	4.84	5.97	6.87	174
zlib + ppmd + rp (16 интервалов)	2.4	3.2	4.7	5.1	7.3	6.8	8.1	3.85	4.91	5.70	174
zlib + ppmd + rp (32 интервала)	2.4	3.2	4.7	5.0	7.2	6.8	8.1	3.84	4.90	5.69	174

Таблица 9

Результаты расчетов для временного ряда Т-индекса

Метод	Ошибка MAE для номера шага *h*										Среднее			Количество рядов
	1	2	3	4	5	6	8	12	15	18	1–4	1–8	1–18	
Прогноз метеорологического бюро	13.0	14.2	14.8	15.7	16.5	18.0	21.2	23.5	25.6	27.0	14.43	16.69	21.09	61
zlib (16 интервалов)	12.3	16.0	18.3	20.4	19.1	21.5	24.9	24.5	28.8	26.5	16.75	19.70	23.68	61
ppmd (16 интервалов)	11.9	15.2	17.4	20.9	21.5	22.7	18.8	25.7	30.6	25.5	16.34	18.56	22.87	61
rp (16 интервалов)	12.5	18.1	18.6	18.3	24.9	24.5	22.6	33.4	38.3	41.9	16.86	21.10	28.2	61
zlib + ppmd + rp (16 интервалов)	11.9	15.2	17.4	20.8	21.5	22.0	18.8	24.9	29.4	24.2	16.31	18.46	22.57	61
zlib + ppmd + rp (32 интервала)	11.8	15.2	17.4	21.2	21.5	22.0	18.8	25.4	28.7	26.1	16.39	18.51	22.67	61

Выводы

В рамках данной работы был разработан и экспериментально исследован метод прогнозирования временных рядов, базирующийся на широко известных архиваторах. Показано, что точность данного метода во многих случаях сравнима с точностью других методов прогнозирования и он представляет практический интерес. Возможно совмещение предлагаемого метода с распространенными методами предварительной обработки данных для повышения точности прогноза.

Список литературы

1. *Kendall M. G.*, *A. Stuart.* The Advanced Theory of Statistics: Design and analysis, and time-series. The Advanced Theory of Statistics. Hafner, 1976.

2. *Hyndman R. J.*, *Athanasopoulos G.* Forecasting: principles and practice. OTexts, 2014.

3. *Makridakis S.*, *Hibon M.* The M3-Competition: results, conclusions and implications // International journal of forecasting. 2000. Vol. 16. No. 4. P. 451–476.

4. *Рябко Б. Я.* Прогноз случайных последовательностей и универсальное кодирование // Проблемы передачи информации. 1988. Т. 24, №. 2. С. 3–14.

5. *Shkarin D.* PPM: One step to practicality // Proc. Data Compression Conference. IEEE, 2002. P. 202–211.

6. *Cover T. M.*, *Thomas J. A.* Elements of information theory. John Wiley & Sons, 2012.

7. *Ryabko B.*, *Astola J.*, *Malyutov M.* Compression-based methods of statistical analysis and prediction of time series. Switzerland: Springer International Publishing, 2016.

8. *Ryabko B.* Compression-based methods for nonparametric prediction and estimation of some characteristics of time series // IEEE Transactions on Information Theory. 2009. Vol. 55. No. 9. P. 4309–4315.

9. *Bille P.*, *Gørtz I. L.*, *Prezza N.* Space-Efficient Re-Pair Compression // Data Compression Conference. IEEE, 2017. P. 171–180.

10. *Cleveland R. B.*, *Cleveland W. S.*, *Terpenning I.* STL: A seasonal-trend decomposition procedure based on loess // Journal of Official Statistics. 1990. Vol. 6. No. 1. P. 3.

Материал поступил в редколлегию 10.05.2018

K. S. Chirikhin [1, 2], **B. Ya. Ryabko** [1, 3]

[1] *Novosibirsk State University*
1 Pirogov Str., *Novosibirsk*, *630090*, *Russian Federation*

[2] *Siberian State University of Telecommunications and Information Sciences*
86 Kirov Str., *Novosibirsk*, *630102*, *Russian Federation*

[3] *Institute of Computational Technologies SB RAS*
6 Academician Lavrentiev Ave., *Novosibirsk*, *630090*, *Russian Federation*

chirihin@gmail.com

EXPERIMENTAL STUDY OF THE ACCURACY OF COMPRESSION-BASED FORECASTING METHODS

In information theory it is known that methods of data compression can be used for forecasting of stationary processes. In this paper an compression-based algorithm for time series forecasting was proposed and empirical study of its accuracy was carried out. The algorithm can operate with arbitrary methods of data compression. During the steps of the algorithm predicted values from different methods are combined, and the greatest impact on the end result is exerted by the method with the best compression ratio for the series. The algorithm can be used for forecasting of time series with discrete and continuous alphabets. To improve the accuracy of the forecast existing meth-

ods of time series preprocessing can be used. The empirical study of the efficiency of the proposed algorithm was conducted on time series from the M3 Competition and the T-index series. To generate forecasts well-known archivers were used. The results of the calculations showed that the obtained method has a relatively high accuracy and speed.

Keywords: universal coding, time series forecasting.

References

1. Kendall M. G., Stuart A. The Advanced Theory of Statistics: Design and analysis, and time-series. The Advanced Theory of Statistics. Hafner, 1976.

2. Hyndman R. J., Athanasopoulos G. Forecasting: principles and practice. OTexts, 2014.

3. Makridakis S., Hibon M. The M3-Competition: results, conclusions and implications. *International journal of forecasting*, 2000, vol. 16, no. 4, p. 451–476.

4. Ryabko B. Ya. Prediction of random sequences and universal coding. *Problems of information transmission*, 1988, vol. 24, no. 2, p. 87–96. (in Russ.)

5. Shkarin D. PPM: One step to practicality. *Proc. Data Compression Conference*. IEEE, 2002, p. 202–211.

6. Cover T. M., Thomas J. A. Elements of information theory. John Wiley & Sons, 2012.

7. Ryabko B., Astola J., Malyutov M. Compression-based methods of statistical analysis and prediction of time series. Switzerland, Springer International Publishing, 2016.

8. Ryabko B. Compression-based methods for nonparametric prediction and estimation of some characteristics of time series. *IEEE Transactions on Information Theory*, 2009, vol. 55, no. 9, p. 4309–4315.

9. Bille P., Gørtz I. L., Prezza N. Space-Efficient Re-Pair Compression. *Data Compression Conference*. IEEE, 2017, p. 171–180.

10. Cleveland R. B., Cleveland W. S., Terpenning I. STL: A seasonal-trend decomposition procedure based on loess. Journal of Official Statistics, 1990, vol. 6, no. 1, p. 3.

For citation:

Chirikhin K. S., Ryabko B. Ya. Experimental Study of the Accuracy of Compression-Based Forecasting Methods. *Vestnik NSU. Series*: *Information Technologies*, 2018, vol. 16, no. 3, p. 145–158. (in Russ.)

DOI 10.25205/1818-7900-2018-16-3-145-158