Использование суперкомпьютера для экспериментального исследования метода прогнозирования, основанного на универсальных кодах

П.А. Приставка
Аспирант кафедры прикладной математики и кибернетики СибГУТИ
e-mail: ppa@ngs.ru

Аннотация

В работе предлагается и экспериментально исследуется метод прогнозирования, основанный на универсальных кодах. На примерах прогнозирования солнечной активности показывается, что указанный метод обладает высокой точностью прогноза. Использование параллельных вычислений и ресурсов суперкомпьютера позволило существенно сократить время работы программной реализации метода.

1. Введение

На сегодняшний день задача прогнозирования является актуальной для множества областей человеческой деятельности. Следует отметить, что на практике очень часто априорные сведения о распределении вероятностей исходов прогнозируемого процесса отсутствуют, и это затрудняет решение поставленной задачи. В таком случае можно воспользоваться точными оценками указанных величин. В конце 1980-х было установлено (см. [1,2]), что универсальные коды представляют собой средство, позволяющее обнаруживать скрытые периодичности и зависимости в данных, и могут быть применены для получения таких оценок. Однако до сих пор результаты практического применения (см [3,4]) данного подхода остаются малоизученными.

Целью работы, которой посвящен доклад, являлось экспериментальное исследование предложенного в [5] метода прогнозирования, в основе которого лежит универсальный код R, описанный в [1]. Выбор именно этого кода объясняется тем, что он асимптотически оптимален. (Однако, важно отметить, что его практическая точность была неизвестна). Доклад представляет собой краткое описание основных результатов работы.

Возможность использования при расчетах ресурсов высокопроизводительного распределенного вычислительного комплекса позволила в значительной мере сократить время работы программной реализации указанного метода прогнозирования.

В качестве объекта исследования были взяты показатели солнечной активности (СА), и рассматривалась задача, когда по N известным значениям временного ряда необходимо спрогнозировать его следующее, (N+1)-ое, значение. Следует отметить, что задача прогнозирования СА представляет большой практический и теоретический интерес ввиду ее влияния на многие процессы, идущие на Земле. Так, например, в настоящее время широко исследуются и выявляются статистические связи погоды и климата с СА [6]. При этом, согласно [7], недостаточность знаний о механизмах, определяющих поведение СА, оставляют актуальным вопрос точности соответствующих прогнозов.

Полученные в ходе работы экспериментальные результаты показывают, что методы прогнозирования, основанные на применении универсальных кодов, обладают достаточно высокой точностью.

2. Описание метода и разработанного алгоритма

2.1. Описание метода

Сформулируем задачу прогнозирования, которая рассматривалась в проведенной работе. Пусть есть некоторый стационарный и эргодический источник, который порождает последовательности $x_1 x_2 ...$ элементов из некоторого множества A, которое может быть как конечным множеством, так и некоторым непрерывным интервалом. Предполагается, что распределение вероятностей этого источника неизвестно. Пусть источник порождает сообщение вида $x_1 ... x_{t-1} x_t$, $x_i \in \mathrm{A}$ для всех i, и следующий элемент требуется спрогнозировать.

Теперь перейдем к описанию метода прогнозирования.

Введем определение меры R, которая является оценкой вероятностей для класса всех стационарных и эргодических источников на конечном алфавите.

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

где $\{\omega = \omega_1, \omega_2, ...\}$ - это распределение вероятностей, каждый член которого находится по формуле:

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

где i – целое, из множества $\{1,2,..\}$, а $K_i(x_1 ... x_t)$ - мера Кричевского для источника с памятью i.

Важно заметить, что мера R, основывается на универсальном коде из [1] и может применяться при решении задач прогнозирования.

Рассмотрим прогнозирование для различных типов последовательностей, порождаемых источником.

1) *Источник порождает символы из конечного алфавита*

Пусть $x_1 ... x_t$ - некоторая известная последовательность, порожденная источником. Тогда для любого элемента $a \in \mathrm{A}$ можно вычислить вероятность того, что следующим символом на выходе источника будет он (т.е. условную вероятность этого символа), воспользовавшись формулой:

$$R(a \mid x_1 ... x_t) = \frac{R(x_1 ... x_t a)}{R(x_1 ... x_t)}$$

2*) Источник порождает значения из непрерывного интервала*

Пусть стохастический процесс принимает значения из некоторого непрерывного интервала *[A;B]*, $\{\Pi_n\}, n \geq 1$ - возрастающая последовательность конечных разбиений этого интервала и $x^{[k]}$ обозначает элемент Π_k, который содержит точку x.

Определим плотность r_U как

$$r_U(x_1...x_t) = \sum_{i=1}^{\infty} \omega_i R(x_1^{[i]}...x_t^{[i]}) \Big/ M_t(x_1^{[i]}...x_t^{[i]}) \qquad (2)$$

где ω_i - соответствующий член распределение из (1), а M - сигма-конечной мера.

В [5] показывается, что для оценки условной вероятности величины a можно воспользоваться формулой

$$r_U(a \mid x_1...x_m) = \frac{r_U(x_1...x_m a)}{r_U(x_1...x_m)} \qquad (3)$$

Иногда на практике может возникнуть вопрос о том, последовательность какой длины нужно взять в качестве входных данных, ведь наибольшая длина не всегда будет обеспечивать наибольшую точность . В случае если ответ на данный вопрос неочевиден, то в качестве вероятностной оценки можно использовать величину $\tilde{r}_U$, определение которой приводится ниже.

Пусть $x_1...x_t$ - некоторая последовательность источника. Рассмотрим множество различных натуральных чисел $\{N_i\}$ и множество выборок $\{x_{t-N_i+1}...x_t\}, 1 \le N \le t, 1 \le i \le t$. Каждую отдельно взятую выборку данного множества назовем «окном», а соответствующее N_i – его размером. Пусть *i=k*, тогда

$$\tilde{r}_U = \sum_{i=1}^{k} \tilde{\omega}_i r_U(x_{t-N_i+1}...x_t),$$

где $\{\tilde{\omega} = \tilde{\omega}_1, \tilde{\omega}_2, ..., \tilde{\omega}_k\}$ - распределение ненулевых вероятностей, $\sum_{i=1}^{k} \tilde{\omega}_i = 1$.

2.2 Реализация алгоритма

Рассмотрим некоторые аспекты практической реализации исследуемого метода. Итак, пусть есть некоторый источник, который порождает значения из непрерывного отрезка $[A;B]$, и известна последовательность членов временного ряда $x_1...x_t$ длины *t*. Требуется спрогнозировать x_{t+1}.

Шаг 1. Вычислим $r_U(x_1...x_t)$. Разобьем интервал [A; B] на 2 равные части и преобразуем $x_1...x_t$ в последовательность символов, каждый из которых равен номеру участка разбиения, который содержит соответствующую точку x_i. Далее, вычислим первое слагаемое суммы из формулы (2). В качестве M_n возьмем произведение длин всех участков разбиения, содержащих x_i, а меру *R* будем вычислять для последовательности, полученной на этом шаге на основе $x_1...x_t$. После этого снова разделим каждый из имеющихся участков (их будет два) на 2 равные части и для этого нового разбиения выполним аналогичные действия с целью нахождения второго слагаемого из правой части (2). Будем продолжать данный процесс до тех пор пока не получим разбиение, для которого все различные значения из временных рядов (в том числе и используемые на следующем шаге) не будут принадлежать различным участкам. Сложив все имеющиеся слагаемые, получим величину $r_U(x_1...x_t)$ из (2).

Шаг 2. Рассмотрим множество $\{a_i\}$, которое состоит из величин: $A, A+h, A+2h, ..., B$, где h – некоторый достаточно маленький шаг. В данной работе использовался шаг h=0.01. Подставляя по очереди каждый элемент этого множества в конец последовательности $x_1 ... x_t$, вычислим соответствующие величины $r_U(x_1 ... x_t a)$, где a –элемент из множества $\{a_i\}$. При вычислениях будем использовать те же способ и разбиение, что и на шаге 1.

Шаг 3. Теперь по формуле (3) для всех величин из множества $\{a_i\}$ вычислим оценки соответствующих условных вероятностей и на их основе определим прогнозное значение. В данной работе прогнозным значением считалась та величина, чья оценка условной вероятности была наибольшей среди всех остальных.

3. Экспериментальные результаты

В качестве объектов исследования были выбраны временные ряды, состоящие из следующих показателей: среднемесячное и сглаженное среднемесячное число солнечных пятен, абсолютные ежедневное и ежемесячное значение солнечного излучения. Данные, использованные в экспериментах, можно найти на сайте National Geophysical Data Center (NGDC) в разделе «Space Weather & Solar Events». Все исследуемые процессы принимали значения из некоторого непрерывного интервала. Проводимые в рамках данного раздела вычисления можно разделить на две независимые части. В первой рассматривалась задача прогнозирования значений процессов на один шаг вперед, а во второй – сравнение между собой точности краткосрочных прогнозов для сглаженного среднемесячного числа солнечных пятен: опубликованного на сайте NGDC и полученного помощью метода на основе кода *R*.

3.1 Прогнозирование значений временных рядов

Суть каждого проводимого на данном этапе эксперимента заключалась в следующем. Имея в распоряжении только данные об N последовательных значениях временного ряда, требовалось спрогнозировать значение его (N+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 на самом деле уже прошедшего временного интервала. Результаты экспериментальных вычислений приводятся в табл. 1.

Первая колонка таблицы содержит наименование исследуемого временного ряда, вторая – рассматриваемый диапазон принимаемых им значений. Остальные колонки содержат значения средней абсолютной погрешности за 25 экспериментов, полученные при вычислениях с использованием соответствующих входных данных. Следует отметить, что в качестве размеров «окон» для каждого процесса были взяты все длины, используемые для данного временного ряда при вычислениях без «окна», а значения вероятностей $\tilde{\omega}_i$ были одинаковыми. Надпись «н/д» («нет данных») в ячейке означает, что вычисление не проводилось по причине отсутствия требующихся данных.

Таблица 1. Результаты экспериментальных вычислений первой части

Временной ряд	Диапазон значений процесса	Исп. «окно»	Длина						
			500	700	1000	1200	2000	3000	4000
Среднемесячное число солнечных пятен	[0; 256]	23.97	6.54	2.56	9.58	15.85	21.7	19.63	н/д
Сглаженное среднемесячное число солнечных пятен	[0; 210]	2.34	1.5	1.1	1.99	0.77	3.36	2.56	н/д
Абсолютное ежедневное солнечное излучение	[50; 300]	1.78	1.17	1.17	2.71	5.52	8.35	1.72	1.45
Абсолютное ежемесячное солнечное излучение	[580; 2540]	45.29	211.29	45.88	н/д	н/д	н/д	н/д	н/д

3.2 Сравнение краткосрочных прогнозов

На этом этапе оценки экспериментальных результатов производилось сравнение точности краткосрочных прогнозов для сглаженного среднемесячного числа солнечных пятен: предоставляемого NGDC на основе улучшенного метода МакНиша-Линкольна (McNish-Lincoln) и составленного с помощью исследуемого метода.

В качестве прогноза, сформированного программой NGDC, был взят файл sunspot.predict от 06.08.2008, в котором для каждого месяца текущего солнечного цикла указано соответствующее прогнозное значение и его доверительный интервал. Здесь следует уточнить, что в качестве величины для сравнения будет использоваться непосредственно предоставленное прогнозное значение, без учёта доверительного интервала. На момент проведения вычислений этой работы в июне 2010 года имелось 21 уже известное значение исследуемого временного ряда. Вычисления исследуемым методом проводились по следующей схеме. Рассматривался такой временной ряд некоторой фиксированной в рамках эксперимента длины, что его последний член содержал данные за февраль 2008-го года. На основании содержимого данного ряда строился прогноз на один шаг вперед, т.е. март 2008-го года. После этого, на следующей итерации, производилось смещение границ временного ряда с известными значениями на один элемент вправо, при этом последний элемент (соответствующий реальному значению за март 2008) заменялся на полученное на предыдущем шаге прогнозное значение. На основании такого преобразованного ряда снова строился прогноз на один шаг вперед - апрель 2008-го года. И так далее. В общей сложности для каждой рассматриваемой длины временного ряда было проведено 21 вычисление. Результаты экспериментальных вычислений для исследуемого метода приводится в табл. 2. Первая строчка содержит данные о величине средней абсолютной погрешности при

использовании исследуемого метода и соответствующих входных данных. Вторая - среднюю абсолютную погрешность, вычисленную на основе прогноза программы NGDC.

Таблица 2. Результат прогнозирования временного ряда для сглаженных среднемесячных значений солнечных пятен

R	Длина			Исп. «окно»
	500	1000	2000	
	1.75	1.87	2.23	1.42
NGDC	16.39			

Для наглядности полученные в данной части экспериментальной оценки результаты представлены на рис. 1 ниже. На данном графике по оси N отложено сглаженное среднемесячное число солнечных пятен, а по оси t – условные единицы времени, каждая из которых соответствует некоторому месяцу солнечного цикла. Область графика, где $t \le 0$ – является периодом основания прогноза, а область, где $1 \le t \le 21$– периодом упреждения прогноза. Сплошная жирная линия отображает реальные значения временного ряда, пунктирные линии из коротких и длинных штрихов – прогнозные значения, полученные программой NGDC и вычисленные на основе кода R с использованием «окна» соответственно.

Можно отметить ярко выраженную тенденцию к росту ошибки прогноза улучшенного метода МакНиша-Линкольна с увеличением его порядкового номера.

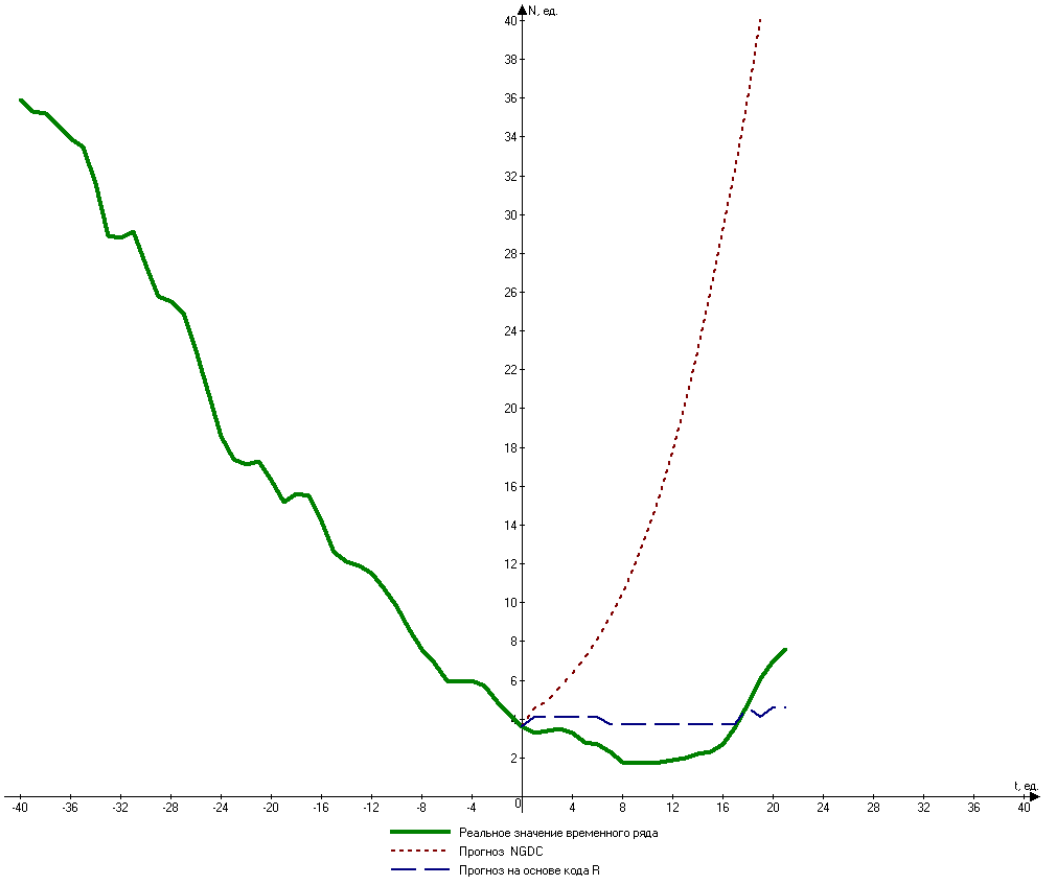

Рис. 1. Результаты сравнения краткосрочных прогнозов

Таким образом, на основании графика можно сказать, что в данном случае при построении краткосрочного прогноза исследуемый метод на основе универсального кода *R* оказался более эффективным. Данные на сервере NGDC представлены так, что провести другие эксперименты по сравнению точности невозможно.

4. Заключение

В ходе проделанной работы были рассмотрены реализация и экспериментальная оценка метода прогнозирования, основанного на универсальном коде *R*. Анализ итогов работы показал достаточно высокую точность полученных результатов. Таким образом можно сделать вывод о том, что универсальные коды являются эффективным инструментом при построения методов прогнозирования для решении практических задач.

Литература

1. B. Ya. Ryabko, Twice-universal coding // Problems of Information Transmission. 1984. V.20, № 3. P. 173–177.
2. B. Ya. Ryabko, Prediction of random sequences and universal coding // Problems of Information Transmission. 1988. V. 24, №. 2. P. 87–96.
3. A. Gruzin, B. Ryabko. Practical Application of Universal Codes to Time Series Forecasting // 2009 XII International Symposium on Problems of Redundancy in Information and Control Systems, Proceedings. P. 10 - 15.
4. B. Ryabko, V. Monarev. Experimental investigation of forecasting methods based on data compression algorithms // Problems of Information Transmission. 2005. V. 41, № 1. P. 65-69.
5. B. Ryabko. Compression-based methods for nonparametric on-line prediction, regression, classification and density estimation of time series // Festschrift in Honor of Jorma Rissanen on the Occasion of his 75th Birthday. Tampere, 2008. P. 271 – 288.
6. Солнечная активность // Википедия: свободная электронная энциклопедия: на русском языке (Последнее изменение: 10:46 21.06.2010) [Электронный ресурс] URL: `http://ru.wikipedia.org/wiki/Солнечная_активность` (дата обращения: 25.06.2010)
7. Котов Ю. Д. Солнечный спутниковый проект «Коронас-Фотон» // Земля и Вселенная. 2009. № 3. С. 3–19.
8. Space Weather & Solar Events [Электронный ресурс] // National Geophysical Data Center: [сайт] URL: `http://www.ngdc.noaa.gov/stp/spaceweather.html` (дата обращения: 04.06.2010)