УДК 519.246.85

А.Л. Кернога, Т.И. Бурак

Белорусский государственный университет информатики и радиоэлектроники,
г. Минск, Республика Беларусь

СРАВНЕНИЕ ПОДХОДОВ К ПРОГНОЗИРОВАНИЮ МЕТОДОМ БЛИЖАЙШИХ СОСЕДЕЙ

Приводится сравнение различных подходов к прогнозированию временных рядов с использованием метода ближайших соседей.

Ключевые слова: временные ряды; предсказание; метод ближайших соседей; авторегрессионный анализ.

A.L. Kernoga, T.I. Burak

Belorussian State University of Informatics and Radioelectronics,
Minsk, Republic of Belarus

COMPARISON OF FORECASTING APPROACHES FOR NEAREST NEIGHBORS METHOD

This article compares different approaches to time series forecasting using Nearest Neighbors method.

Keywords: time series; forecasting; Nearest Neighbors method; autoregression.

Анализ временных рядов – это совокупность математико-статистических методов анализа, предназначенных для выявления структуры временных рядов и их прогнозирования. В сфере моделирования на основании данных системы анализ позволяет получить представление об её структуре, а также спрогнозировать будущие состояния.

Одним из подходов к анализу временных рядов является использование метода ближайших соседей. Идея этого метода заключается в поиске наиболее близких участков данных согласно некоторой функции близости. Метод ближайших соседей относится к локальным методам анализа временных рядов, так как использует не всю историю ряда, а выбирает только наиболее существенные для прогнозирования данные.

Задача прогнозирования временного ряда Z состоит в том, чтобы по известному отрезку данных $\{Z_0, Z_1, \dots, Z_T\}$ предсказать следующие n значений $\{Z_{T+1}, Z_{T+2}, \dots, Z_{T+n}\}$. Алгоритм поиска ближайших соседей состоит из следующих шагов:

1. Обозначим Z_t^m вектор-историю длиной m, состоящую из следующих значений:

$$Z_t^m = \{Z_{t-m+1}, Z_{t-m+2}, \dots, Z_{t-1}, Z_t\}. \quad (1)$$

Разобьём ряд Z на множество векторов-историй Z_t^m, где $t = (m, \dots, T)$. В таком случае вектор Z_T^m будет последней доступной историей для ряда Z.

2. Среди векторов-историй Z_t^m выберем k векторов, являющихся ближайшими соседями вектора Z_T^m. Для определения близости векторов в данной работе будем использовать евклидово расстояние (2). Ближайшие соседи обозначим как nn_i, при этом коэффициент i показывает степень близости соседа к предыстории Z_T^m: чем меньше индекс, тем меньше расстояние до предыстории.

$$\rho_E(Z_T^m, Z_t^m) = \sqrt{\sum_{j=0}^{m-1}\left(Z_{T-j} - Z_{t-j}\right)^2}. \quad (2)$$

3. Полученные векторы $nn_i = \{Z_{t_i-m+1}, Z_{t_i-m+2}, \dots, Z_{t_i-1}, Z_{t_i}\}$, где $i = (1, \dots, k)$, используются для прогноза будущих значений ряда. Существует несколько подходов использования ближайших соседей для предсказания значений ряда:

а) в работе [1] предлагается использовать значения, следующие за ближайшими соседями nn_i, для получения значения Z_{T+1} следующим образом:

$$\hat{Z}_{T+1} = \frac{\sum_{i=1}^{k} Z_{t_{i+1}}}{k}. \quad (3)$$

В этом случае пункты 1–3 повторяются для предсказания каждого из значений $\{\hat{Z}_{T+1}, \hat{Z}_{T+2}, \dots, \hat{Z}_{T+n}\}$. Обозначим этот подход как SA (от англ. *simple average* – простое среднее);

б) другой подход, описанный в работах [1] и [2], основывается на авторегрессионном анализе ближайших соседей и последней предыстории ряда. Предполагается, что значения Z_{T+1} зависит от последней предыстории Z_T^m следующим образом:

$$\hat{Z}_{T+1} = \alpha_0 \cdot Z_T + \alpha_1 \cdot Z_{T-1} + \dots + \alpha_{m-1} \cdot Z_{T-m+1} + \alpha_m. \quad (4)$$

Для определения неизвестных коэффициентов $\{\alpha_0, \alpha_1, \dots, \alpha_m\}$ строится и решается система уравнений:

$$\begin{bmatrix} Z_{t_1+1} \\ Z_{t_2+1} \\ \vdots \\ Z_{t_k+1} \end{bmatrix} = \alpha_0 \cdot \begin{bmatrix} Z_{t_1} \\ Z_{t_2} \\ \vdots \\ Z_{t_k} \end{bmatrix} + \alpha_1 \cdot \begin{bmatrix} Z_{t_1-1} \\ Z_{t_2-1} \\ \vdots \\ Z_{t_k-1} \end{bmatrix} + \cdots + \alpha_{m-1} \cdot \begin{bmatrix} Z_{t_1-m+1} \\ Z_{t_2-m+1} \\ \vdots \\ Z_{t_k-m+1} \end{bmatrix} + \alpha_m. \quad (5)$$

Аналогично подходу SA пункты 1–3 повторяются для предсказания каждого из значений $\{\hat{Z}_{T+1}, \hat{Z}_{T+2}, \dots, \hat{Z}_{T+n}\}$. Обозначим этот подход как LAR (от англ. *local autoregression* – локальная авторегрессия);

в) в работе [3] описан подход долгосрочного предсказания. Для каждого ближайшего соседа $nn_i = \{Z_{t_i-m+1}, Z_{t_i-m+2}, \dots, Z_{t_i-1}, Z_{t_i}\}$ определим векторы-продолжения $p_i = \{Z_{t_i+1}, Z_{t_i+2}, \dots, Z_{t_i+n-1}, Z_{t_i+n}\}$ и вычислим весовые коэффициенты, характеризующие близость этого соседа к вектору Z_T^m:

$$W_i = \left(1 - \left(\frac{\rho_E(nn_i, Z_T^m)}{\rho_E(nn_{k+1}, Z_T^m)}\right)^2\right)^2, \quad (6)$$

$$w_i = \frac{W_i}{\sum_{j=0}^{k} W_j}. \quad (7)$$

Вектор предсказаний $\hat{Z}_{T+n}^n = \{\hat{Z}_{T+1}, \hat{Z}_{T+2}, \dots, \hat{Z}_{T+n}\}$ получается следующим образом:

$$\hat{Z}_{T+n}^n = \sum_{i=0}^{k} w_i \cdot nn_i. \quad (8)$$

Преимуществом данного подхода является то, что предсказания всех необходимых значений получаются за одну итерацию. Обозначим этот подход как LTP (от англ. *long-time prediction* – долгосрочное предсказание).

Проанализируем описанные выше подходы (SA, LAR, LTP) с точки зрения точности предсказания. Для этого протестируем методы на различных данных: на значениях среднемесячной температуры $f_1(t)$ (рис. 1), показателях содержания углекислого газа в атмосфере $f_2(t)$ (рис. 2) и показателях розничной торговли в ЕС $f_3(t)$ (рис. 3). Временной ряд $f_1(t)$ обладает строгой периодичностью, у ряда $f_2(t)$ присутствует тренд, ряд $f_3(t)$ не обладает видимым трендом и периодичностью.

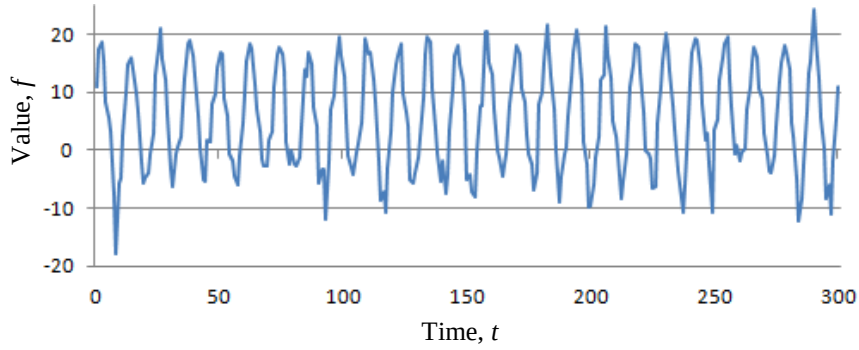

Рис. 1. Данные о среднемесячной температуре

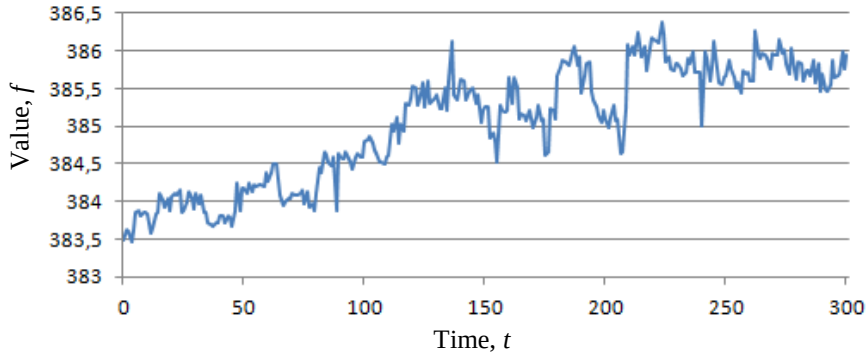

Рис. 2. Данные о содержании углекислого газа в атмосфере

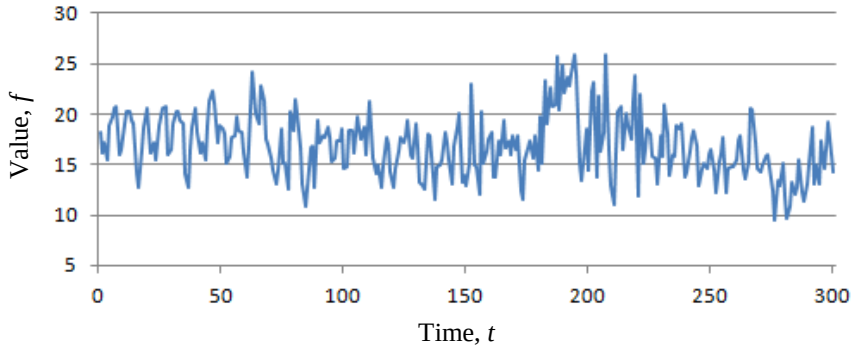

Рис. 3. Данные о показателях розничной торговли в ЕС

Тестируемые ряды содержат по $T = 300$ значений. Зафиксируем длину предыстории $m = 30$ и получим предсказания для последних $n = 40$ значений ряда. Оптимальное количество ближайших соседей k будем выбирать для каждого метода и ряда таким образом, чтобы оно минимизировало значение ошибки MAE (9).

$$\text{MAE} = \frac{1}{n} \cdot \sum_{i=1}^{n} \left| Z_{T+i} - \hat{Z}_{T+1} \right|,$$

$$\text{MAPE} = \frac{1}{n} \cdot \sum_{i=1}^{n} \left| \frac{Z_{T+i} - \hat{Z}_{T+1}}{Z_{T+i}} \right| \cdot 100\ \%, \qquad (9)$$

$$\text{SMAPE} = \frac{1}{n} \cdot \sum_{i=1}^{n} \left| \frac{Z_{T+i} - \hat{Z}_{T+1}}{(Z_{T+i} + \hat{Z}_{T+1})/2} \right| \cdot 100\ \%.$$

Результаты тестирования методов приведены в табл. 1–3.

Таблица 1

Результаты тестирования для ряда $f_1(t)$

	SA	LAR	LTP
k	4	60	14
MAE	1,832979	2,06220552	1,831438
MAPE, %	15,07885	14,0776146	10,39049
SMAPE, %	6,298938	10,1293612	6,048118

Таблица 2

Результаты тестирования для ряда $f_2(t)$

	SA	LAR	LTP
k	19	65	96
MAE	0,193325	0,16309	0,298683
MAPE, %	0,050095	0,042267	0,077396
SMAPE, %	0,050113	0,042274	0,077439

Таблица 3

Результаты тестирования для ряда $f_3(t)$

	SA	LAR	LTP
k	6	70	6
MAE	1,84180654	3,140744901	2,358595
MAPE, %	13,68217444	23,03360718	18,1913
SMAPE, %	12,6814555	20,68496811	15,79495

Для периодического ряда $f_1(t)$ лучшие результаты показали подходы LTP и SA (рис. 4). Функционала этих подходов достаточно для выявления закономерностей у периодических рядов, при этом они просты в реализации, а подход LTP имеет значительно меньшее время выполнения.

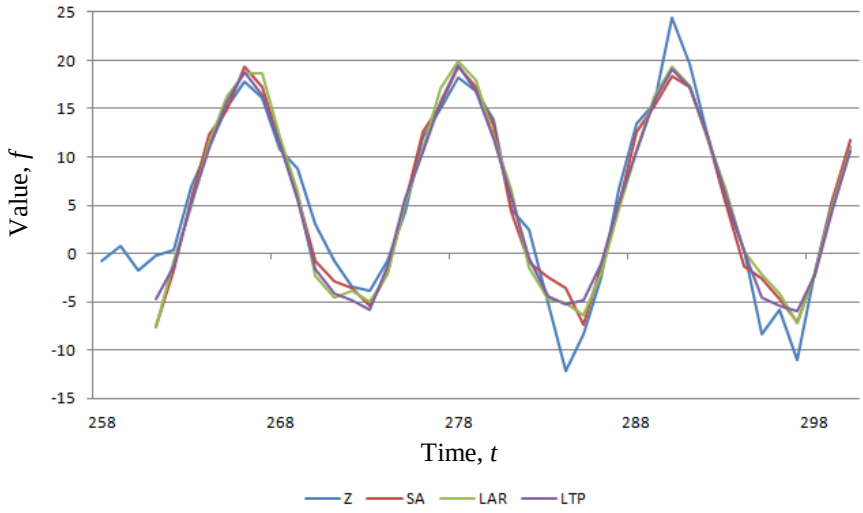

Рис. 4. Прогноз $f_1(t)$ для методов SA, LAR, LTP

Для ряда с трендом $f_2(t)$ лучшие результаты показал подход LAR (рис. 5). Используемый в нем авторегрессионный анализ позволяет определить тенденции значений ряда и построить более точный прогноз.

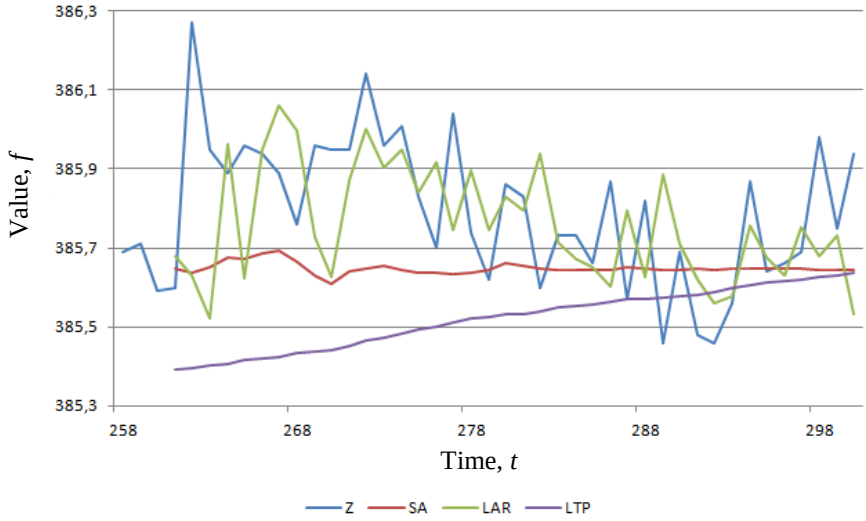

Рис. 5. Прогноз $f_2(t)$ для методов SA, LAR, LTP

Для ряда $f_3(t)$ лучшие результаты показал подход SA (рис. 6). Можно сделать вывод, что показатели подходов LAR и LTP значительно ухудшаются при наличии шума, в то время как подход SA относительно устойчив к шумам.

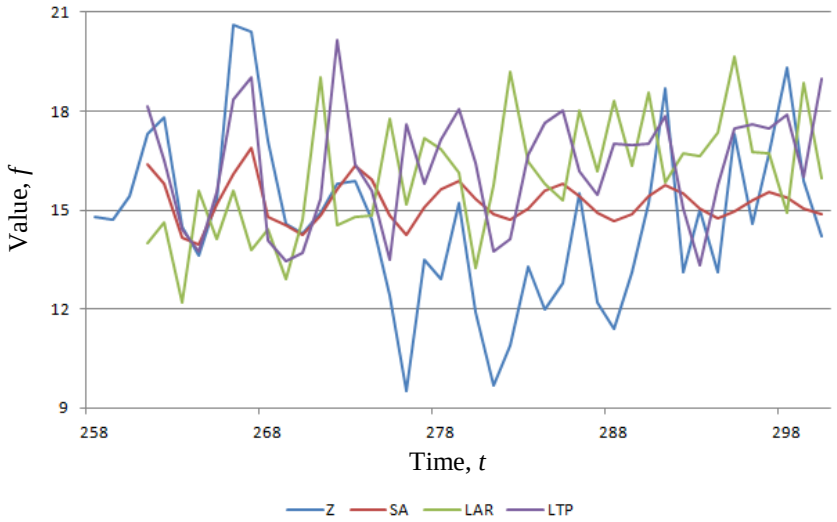

Рис. 6. Прогноз $f_3(t)$ для методов SA, LAR, LTP

Из результатов таблиц следует, что ни одна из метрик одновременно не дает на всех исследуемых рядах оптимальный результат. По полученным в работе данным можно оценить качество прогнозирования различных временных рядов при помощи метода ближайших соседей. Отметим, что в данном анализе длина предыстории m считалась фиксированной, что, однако, может повлиять на оптимальность использования того или иного алгоритма.

Библиографический список

1. Scherer Perlin M. Nearest neighbor method // Revista Electronica de Administracao. – 2007. – Vol. 13, № 2. – 15 p.

2. Fernandez-Rodriguez F., Sosvilla-Rivero S., Andrada-Felix J. Nearest-Neighbour Predictions in Foreign Exchange Markets // Fundacion de Estudios de Economia Aplicada. – 2002. – № 5. – 36 p.

3. Варфоломеева А.А. Локальные методы прогнозирования с выбором метрики // Машинное обучение и анализ данных. – 2012. – № 1(3). – P. 367–375.

References

1. Scherer Perlin M. Nearest neighbor method. *Revista Eletronica de Administracao*, 2007, vol. 13, no. 2. 15 p.

2. Fernandez-Rodriguez F., Sosvilla-Rivero S., Andrada-Felix J. Nearest-Neighbour Predictions in Foreign Exchange Markets. *Fundacion de Estudios de Economia Aplicada*, 2002, no. 5. 36 p.

3. Varfolomeeva A.A. Lokal'nye metody prognozirovaniia s vyborom metriki [Local forecasting methods with a metrics choice]. *Mashinnoe obuchenie i analiz dannykh*, 2012, no. 1(3), pp. 367-375.

Сведения об авторах

Кернога Анастасия Леонидовна (Минск, Республика Беларусь) – магистрантка кафедры электронных вычислительных машин Белорусского государственного университета информатики и радиоэлектроники (220013, Республика Беларусь, г. Минск, ул. Петруся Бровки, 6, e-mail: a.kernoga@gmail.com).

Бурак Тимофей Игоревич (Минск, Республика Беларусь) – магистрант кафедры электронных вычислительных машин Белорусского государственного университета информатики и радиоэлектроники (220013, Республика Беларусь, г. Минск, ул. Петруся Бровки, 6, e-mail: timburik@gmail.com).

About the authors

Kernoga Anastasia Leonidovna (Minsk, Republic of Belarus) – the master student of Electronic Computing Machines Department of the Belarusian State University of Informatics and Radioelectronics (220013, Republic of Belarus, Minsk, Piatrusia Broŭki St., 6, e-mail: a.kernoga@gmail.com).

Burak Timofey Igorevich (Minsk, Republic of Belarus) – the master student of Electronic Computing Machines Department of the Belarusian state university of Informatics and Radioelectronics (220013, Republic of Belarus, Minsk, Piatrusia Broŭki St., 6, e-mail: timburik@gmail.com).

Получено 20.02.2015