Локальные методы прогнозирования с выбором метрики*

А. А. Варфоломеева
annette92@mail.ru
Московский физико-технический институт, ФУПМ, каф. «Интеллектуальные системы»

В данной работе рассматривается локальный метод прогнозирования временных рядов. Исследуется вопрос выбора функции близости для нахождения похожих участков ряда. Проводится сравнение эффективности алгоритма построения прогноза при использовании различных метрик на модельных данных и временных рядах потребления электроэнергии и цен на сахар.

Ключевые слова: *локальное прогнозирование, функция близости, функционал качества, k ближайших соседей.*

Local forecasting with metrics selection*

A. A. Varfolomeeva
Moscow Institute of Physics and Technology

In this article the local method of time series prediction is considered. The method is based on the algorithm of k nearest neighbors. The author investigates the question of the choice of metrics in order to find similar parts of the series. A comparison of the effectiveness of the algorithm for constructing prediction using different metrics is illustrated on synthetic data and time series of electricity consumption and sugar prices.

Keywords: *local forecasting, k nearest neighbors, loss function, metrics.*

Введение

Методы построения прогноза временных рядов делятся на глобальные (использующие всю предысторию ряда) и локальные (используют только её часть). В общем случае прогнозирования рядов решается задача поиска оптимальных парамеров алгоритма: число ближайших соседей, длина предыстории и выбор функции близости. В данной работе рассматривается локальный метод прогнозирования, основанный на алгоритме поиска k ближайших соседей, который был описан в работах Дж. Макнеймса [1] и Ю.И. Журавлева [3]. В работе В.П. Федоровой [5] подробно рассмотрен вопрос оптимизации некоторых параметров метода прогнозирования, основанного на алгоритме "k ближайших соседей". В настоящей работе изучается проблема поиска метрики, дающей наибольшую эффективность данного алгоритма. Рассматривая набор определенных метрик и минимизируя функционал ошибки алгоритма, выбирается наиболее подходящая под данный временной ряд метрика.

Постановка задачи

Рассматриваются одномерные временные ряды, т.е. такие, в которых значением ряда в каждый момент времени является вещественное число. Задача прогнозирования временного ряда состоит в том, чтобы по известному отрезку временного ряда

$$(f_1, f_2, \ldots, f_n)$$

предсказать следующие t его значений:

$$(f_{n+1}, f_{n+2}, \ldots, f_{n+t}).$$

Научный руководитель В. В. Стрижов

Решается задача построения локального метода прогнозирования временных рядов, основанного на алгоритме "ближайших соседей". Для оценки степени близости объектов предлагается использовать различные метрики и сравнить качество прогноза при их использовании. В данной работе длина предыстории считается фиксированной и равной l. Алгоритм "ближайших соседей" состоит из следующих этапов:

1. Найти в предыстории среди всех векторов размерности l, составленных из отрезков временного ряда $(f_i, f_{i+1}, \ldots, f_{i+l-1})$, k векторов, наиболее похожих после линейных преобразований на вектор $(f_{n-l+1}, f_{n-l+2}, \ldots, f_n)$. При этом мера сходства определяется с помощью одной из рассматриваемых в работе метрик.
2. Пусть $\{(f_{i_1-l+1}, \ldots, f_{i_1}), \ldots, (f_{i_k-l+1}, \ldots, f_{i_k})\}$ — k ближайших соседей для предыстории $(f_{n-l+1}, \ldots, f_n)$. Прогноз $(\hat{f}_{n+1}, \hat{f}_{n+2}, \ldots, \hat{f}_{n+t})$ вычисляется как взвешенное среднее арифметическое k векторов:

$$\{(f_{i_1+1}, \ldots, f_{i_1+t}), \ldots, (f_{i_k+1}, \ldots, f_{i_k+t})\}.$$

Данная работа посвящена анализу различных метрик и зависимости качества прогноза от выбора одной из них. Рассматривается набор метрик P:

— Стандартная Евклидова метрика:

$$\rho_E(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})^T(\mathbf{x}-\mathbf{y})}. \tag{1}$$

— Диагонально взвешенная Евклидова метрика:

$$\rho_{wE}(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})^T\Lambda^2(\mathbf{x}-\mathbf{y})}, \text{ где } \Lambda = \mathrm{diag}(\lambda). \tag{2}$$

— Метрика Минковского L_p:

$$\rho_{L_p}(\mathbf{x}, \mathbf{y}) = (\sum_i |x_i - y_i|^p)^{1/p}, \text{ где } p \in \mathbb{N}. \tag{3}$$

Для оценки качества алгоритма используется функционал ошибки SMAPE(Symmetric Mean Absolute Percent):

$$\mathrm{SMAPE}(f, \hat{f}, n, t) = \frac{1}{t}\sum_{i=1}^{t} \frac{|\hat{f}_{n+i} - f_{n+i}|}{|\hat{f}_{n+i} + f_{n+i}|/2} * 100\%. \tag{4}$$

Этот функционал учитывает отклонение прогноза от точного значения относительно величины значения. Тогда задача прогнозирования состоит в поиске такого вектора $(\hat{f}_{n+1}, \hat{f}_{n+2}, \ldots, \hat{f}_{n+t})$, что

$$(\hat{f}_{n+1}, \hat{f}_{n+2}, \ldots, \hat{f}_{n+t}) = \arg\min_{(\hat{f})^\rho} SMAPE(f, (\hat{f})^\rho, n, t), \tag{5}$$

где $(\hat{f})^\rho = ((\hat{f}_{n+1})^\rho, (\hat{f}_{n+2})^\rho, \ldots, (\hat{f}_{n+t})^\rho)$ — прогноз, вычисленный с использованием одной из рассматриваемых метрик $\rho \in P$.

Описание алгоритма

Опишем более подробно используемый алгоритм. Для начала напомним основные обозначения:

— $\mathbf{S} = (f_1, f_2, \ldots, f_N)$ — известный временной ряд.
— l — длина предыстории, считается фиксированной.
— t — длина прогнозируемого отрезка.

— $\rho(\mathbf{x}, \mathbf{y}) \in P$ — функция близости между векторами $\mathbf{x}$ и $\mathbf{y}$.

Основная процедура. Пусть, для определенности, прогнозируются последние t значений временного ряда $\mathbf{S}^{T\times 1}$:

$$(f_{N-t+1}, f_{N-t+2}, \ldots, f_N).$$

Тогда за известный отрезок временного ряда принимается $(f_1, f_2, \ldots, f_n)$, где $n = N - t$. Для начала выделим предысторию прогнозируемого вектора $\mathbf{y} = (f_{n-l+1}, f_{n-l+2}, \ldots, f_n)$ и составим матрицу $\mathbf{X}^{l\times(n-t)}$, состоящую из всех векторов $\mathbf{x}_i = (f_i, f_{i+1}, \ldots, f_{i+l-1})$ размерности l, входящих в известный временной ряд $\mathbf{S}$, и имеющих после себя как минимум t известных значений: они рассматриваются как потенциальные соседи. Далее введем матрицу $\mathbf{D}^{4\times(n-t)}$, в которую для выбранной функции близости и каждого потенциального соседа $\mathbf{x}_i$ запишем его характеристики:

— индекс первого элемента вектора $\mathbf{x}_i$ i;
— расстояние до предыстории $\rho(\mathbf{x}_i, \mathbf{y})$, вычисленное с помощью определенной метрики;
— параметры линейного преобразования вектора при вычислении расстояния $\rho(\mathbf{x}_i, \mathbf{y})$.

Отсортируем матрицу $\mathbf{D}$ по величине расстояния $\rho(\mathbf{x}_i, \mathbf{y})$ и выделим первые k векторов $\mathbf{x}_i$, сооответствующие ближайшим соседям. Найдем для каждого ближайшего соседа $\mathbf{x}_i$ его продолжение $\mathbf{k}_i = (f_{i+l}, f_{i+l+1}, \ldots, f_{i+l+t-1})$ и запишем их в матрицу $\mathbf{K}^{t\times k}$. Также вычисляем веса W_i, с которыми каждый из векторов $\mathbf{k}_i$ учитывается при построении прогноза: они зависят от расстояния до предыстории $\mathbf{y}$ по формуле, предложенной в [4]:

$$W_i = \left(1 - \left(\frac{\rho(\mathbf{k}_i, \mathbf{y})}{\rho(\mathbf{k}_{k+1}, \mathbf{y})}\right)^2\right)^2, \tag{6}$$

где $\rho(\mathbf{k}_i, \mathbf{y})$ — расстояние до i-го ближайшего соседа. Для нормировки весов W_i, вычислим их общую сумму и поделим каждый из весов на нее:

$$w_i = \frac{W_i}{\sum_{j=1}^{t} W_j}.$$

Далее строится прогноз $\hat{\mathbf{x}}_{n+1} = (\hat{f}_{n+1}, \hat{f}_{n+2}, \ldots, \hat{f}_{n+t})$ как взвешенное среднее арифметическое найденных отрезков:

$$\hat{\mathbf{x}}_{n+1} = \sum_{i=1}^{t} w_i * \mathbf{k}_i, \tag{7}$$

и записывается в продолжение данного временного ряда:

$$\hat{S} = (f_1, f_2, \ldots, f_n, \hat{f}_{n+1}, \hat{f}_{n+2}, \ldots, \hat{f}_{n+t}). \tag{8}$$

Вычисление расстояния между отрезками. Во всех исследуемых метриках похожие отрезки ищутся с точностью до линейного преобразования: $\tilde{\mathbf{x}} = a * \mathbf{x} + b$, где $a, b \in \mathbb{R}$. Следовательно расстояние между векторами $\mathbf{x}$ и $\mathbf{y}$ определяется как:

$$\rho(\mathbf{x}, \mathbf{y}) = \min_{a,b} \rho(\tilde{\mathbf{x}}, \mathbf{y}), \tag{9}$$

при этом $\rho \in P$.

Для диагонально взвешенной Евклидовой метрики общим предположением является то, что конец предыстории для прогноза более важен, чем его начало, поэтому параметры

λ_{ii} увеличиваются с порядковым номером i. В работе предполагается, что последовательность весовых параметров λ_{ii} имеет степенной вид:

$$\lambda_{ii} = \lambda^{t-i+1}, 0 < \lambda_i \leqslant 1, i = 1, \ldots, t. \tag{10}$$

Следовательно, требуется задать, например, $\lambda_{11} = \lambda$.

Вычислительный эксперимент

Алгоритм протестирован на модельных и реальных данных. В качестве модельных данных взят ряд, образованный с помощью функции

$$f_1(t) = sin(t) * cos(0.01t) \text{ , где } t = 1, 2, \ldots, 1000.$$

Считаем известными первые 800 точек ряда и строим по ним прогноз длиной $t = 200$. Для исследуемых модельных данных $f_1(t)$ зафиксируем длину каждого соседа $l = 80$. В качестве реальных временных рядов использованы данные о почасовом потреблении электроэнергии $f_2(t)$ (рис.1) и данные о ценах на сахар $f_3(t)$ (рис.2). Первый ряд является строго периодичным и почти не содержит шумов, тогда как второй является зашумленным. В случае реальных данных $f_2(t)$ и $f_3(t)$ установим длину соседа $l = 60$ и будем прогнозировать последние $t = 400$ значений. На всех графиках, отражающих построение прогноза, зеленым цветом выделены линии, полученные с помощью используемого в работе алгоритма, синим — линии, построенные по известному временному ряду, а красные линии отражают поточечную разницу между прогнозом и известным значением ряда.

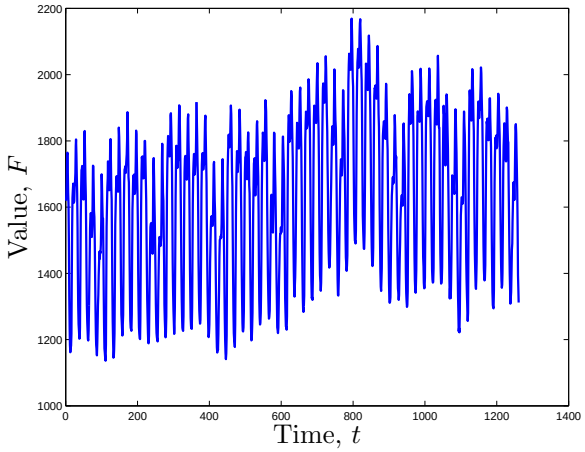

Рис. 1. Вид данных о потреблении электроэнергии.

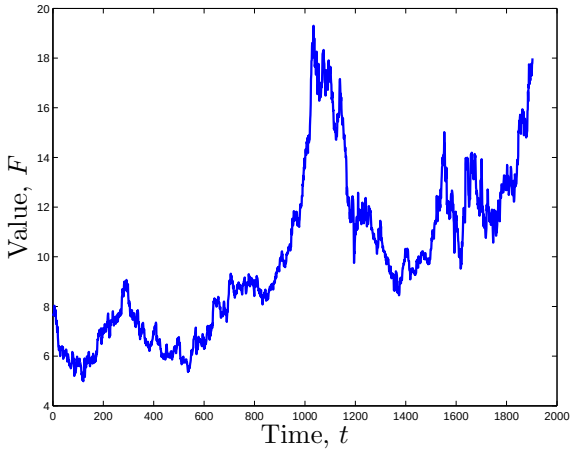

Рис. 2. Вид данных о ценах на сахар.

Стандартная Евклидова метрика.

Рассмотрим качество работы алгоритма, использующего стандартную Евклидову метрику (1). Исследуемым параметром для оптимизации в данном случае служит количество ближайших соседей k. Зависимость величины ошибки SMAPE (4) от количества ближайших соседей k и построение прогноза для рассматриваемых рядов отражены на рис.3, рис.4 и рис.5.

Диагонально взвешенная Евклидова метрика. При использовании метрики (2) необходимо оптимизировать также и весовые параметры метрики λ_{ii}, которые задаются как степенной ряд (10). Зависимость величины ошибки SMAPE (4) от этих двух параметров и построение прогноза для рассматриваемых рядов отражены на рис.6, рис.7 и рис.8.

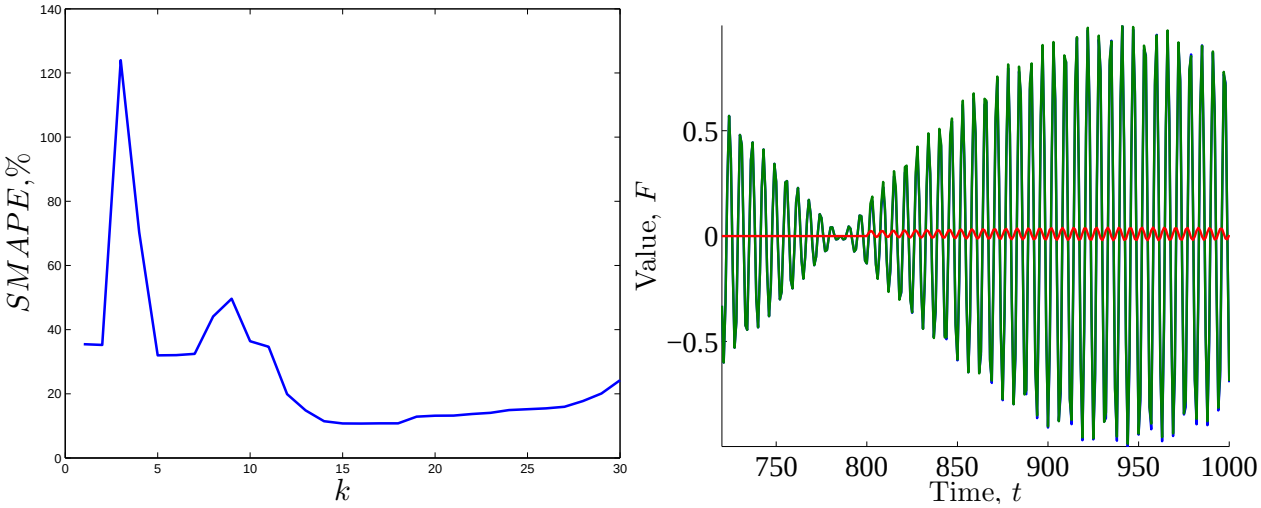

Рис. 3. Величина ошибки и построение прогноза для данных $f_1(t)$ ($k = 16$)

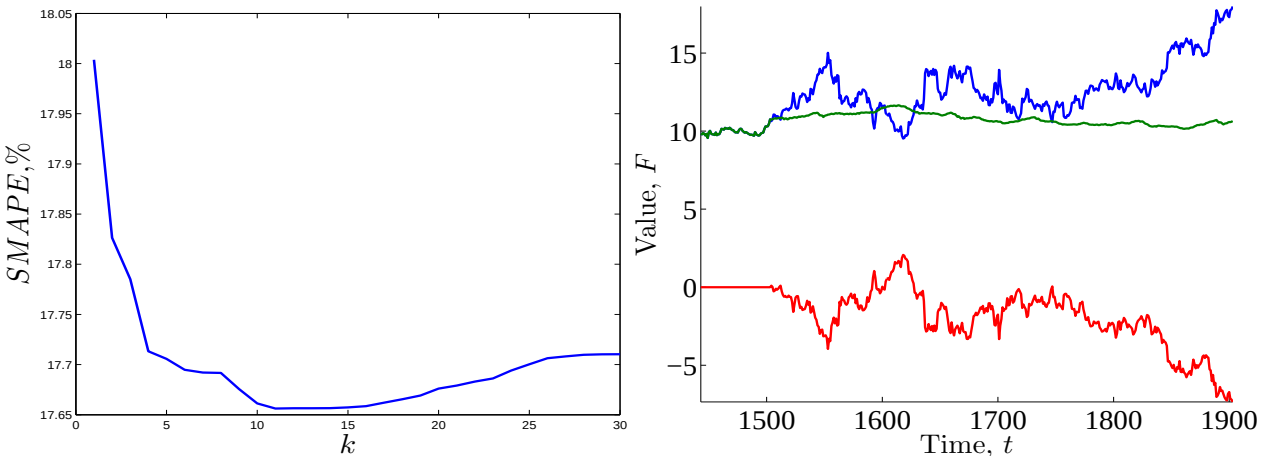

Рис. 4. Величина ошибки и построение прогноза для данных $f_2(t)$ ($k = 11$)

Метрика Минковского. Параметр метрики Минковского (3) p, очевидно, существенно влияет на качество прогноза. Также остается зависимость от количества ближайших соседей k. В данной работе рассматривается параметр метрики p в диапозоне от 1 до 10. Графики полученных зависимостей и построение прогноза приведены на рис.9, рис.10 и рис.11.

Сравнение результатов

Приведем в таблицах сравнительные результаты работы алгоритма на исследуемых данных при использовании различных метрик (1, 2, 3):

Из данных в таблицах следует, что ни одна из метрик одновременно не дает на всех исследуемых рядах оптимальный результат. Для строго периодичного модельного ряда $f_1(t)$ оптимальной является метрика Минковского: ошибка при её использовании менее 5%. Это объясняется тем, что чем больше параметр метрики p, тем меньше весовые параметры w_i тех ближайших соседей, порядковый номер которых близок к k. На зашумленных данных о ценах на сахар $f_2(t)$ наилучший результат, как и ожидалось, дает диагонально взвешенная Евклидова метрика: в этом случае алгоритм использует намного меньшее число

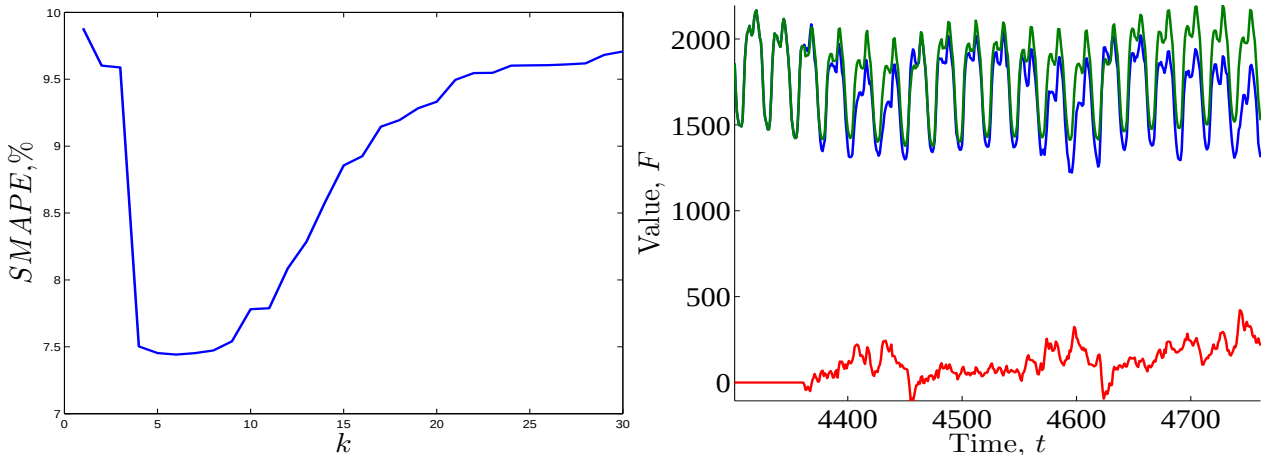

Рис. 5. Величина ошибки и построение прогноза для данных $f_3(t)$ ($k = 6$)

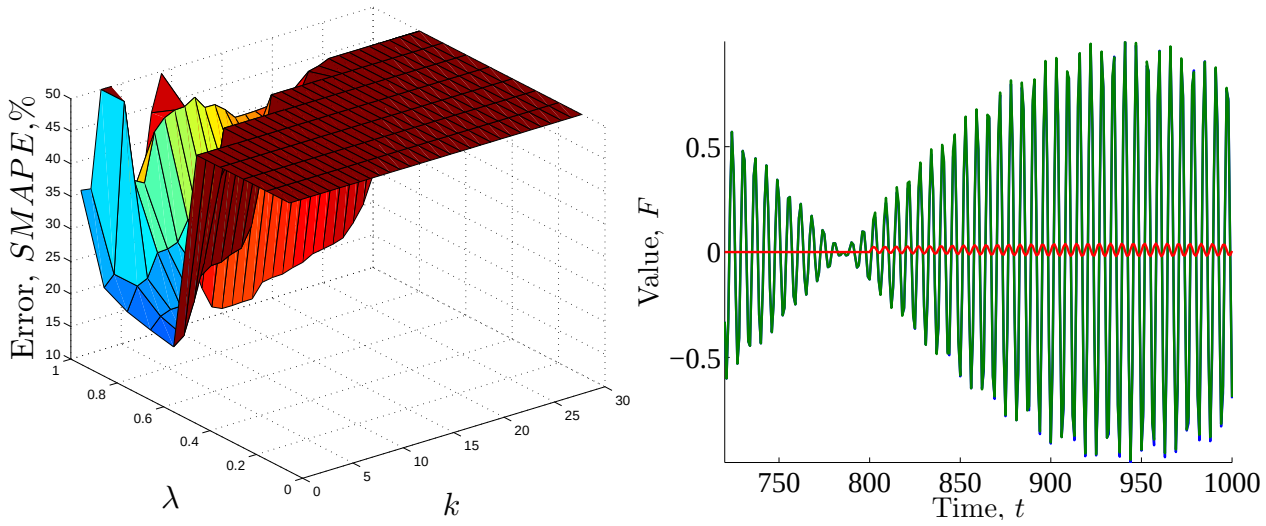

Рис. 6. Величина ошибки и построение прогноза для данных $f_1(t)$ $(k = 16, \lambda = 1)$

Metrics	(1)	(2)	(3)
best k	16	16	17
λ	1	1	1
p	2	2	4
SMAPE,%	10,71	10,71	4,91

Таблица 1. Сравнение результатов работы алгоритма на модельных данных $f_1(t)$.

ближайших соседей, но становится менее восприимчивым к шуму за счет весового параметра λ_{ii}. На этих данных использование метрики Минковского не дает улучшения по сравнению со стандартной Евклидовой метрикой: оптимальным параметром p является $p = 2$. Для данных о потреблении электроэнергии $f_3(t)$ наилучшей также оказалась диагонально взвешенная Евклидова метрика, но в этом случае она, напротив, использует намного большее число ближайших соседей и меньший весовой параметр λ_{ii}, что фактически означает, что начало предыстории сильно менее важно чем ее окончание. Однако,

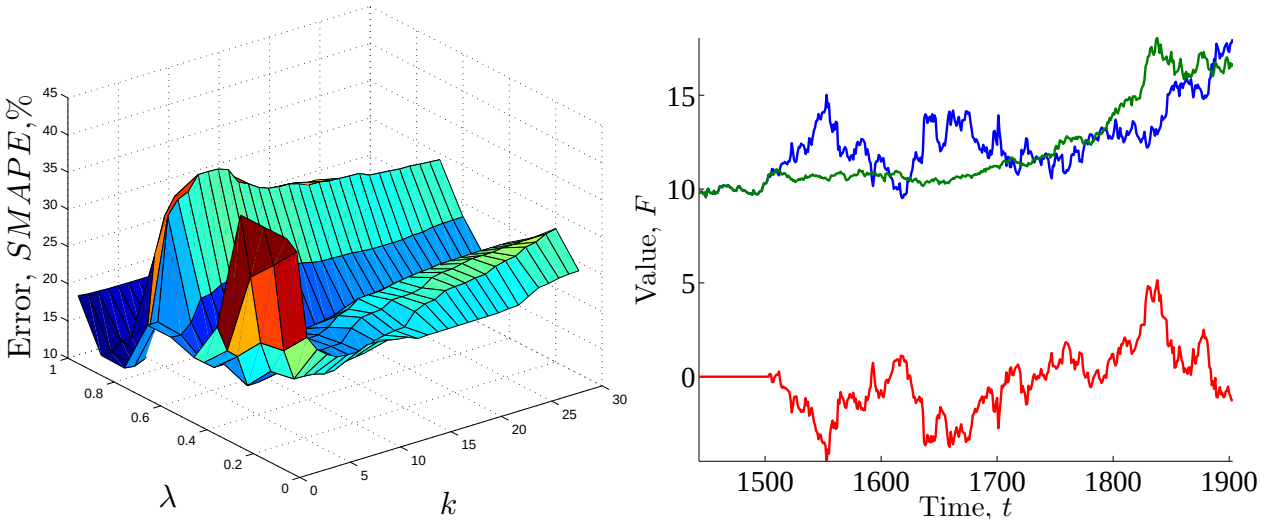

Рис. 7. Величина ошибки и построение прогноза для данных $f_2(t)$ $(k = 2, \lambda = 0.8)$

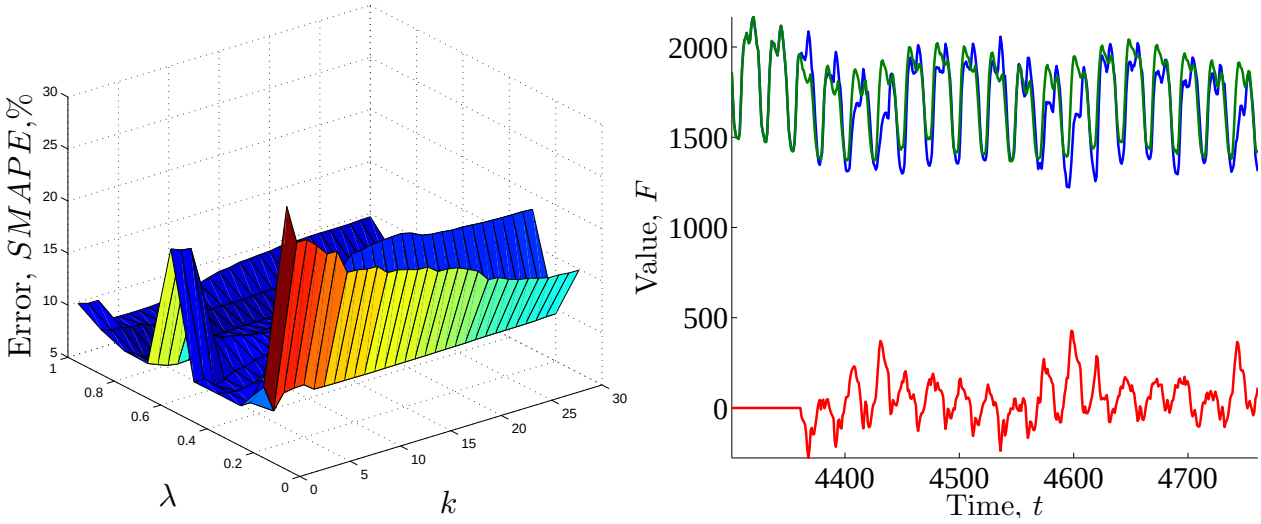

Рис. 8. Величина ошибки и построение прогноза для данных $f_3(t)$ $(k = 26, \lambda = 0,6)$

Metrics	(1)	(2)	(3)
best k	11	2	11
λ	1	0,8	1
p	2	2	2
SMAPE,%	17,66	11,3	17,66

Таблица 2. Сравнение результатов работы алгоритма на данных о ценах на сахар $f_2(t)$.

разница в величине ошибки небольшая, и поэтому судить об оптимальности применения данной метрики на рядах похожего типа не представляется возможным.

Важным фактором является то, что в данной работе длина предыстории l считалась фиксированной, что ограничивает возможность судить об оптимальности применения той

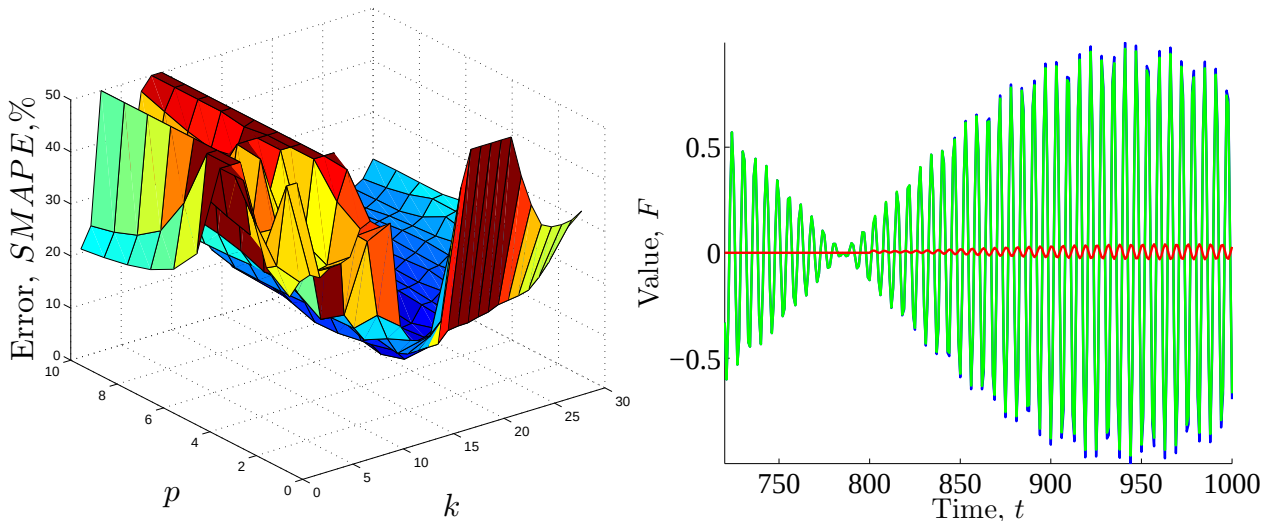

Рис. 9. Величина ошибки и построение прогноза для данных $f_1(t)$ $(k = 17, p = 4)$

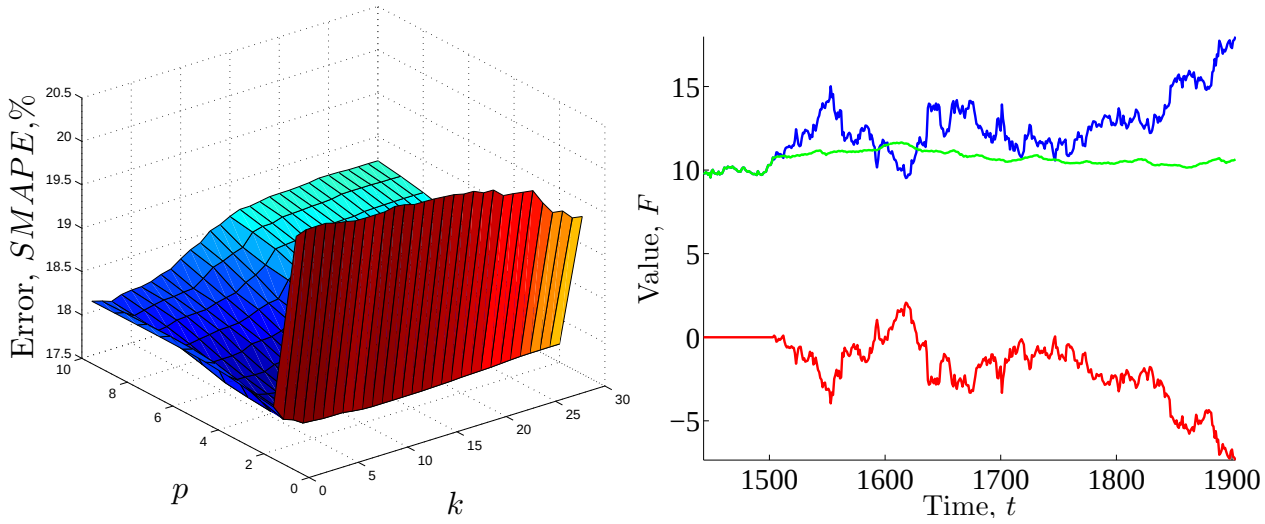

Рис. 10. Величина ошибки и построение прогноза для данных $f_2(t)(k = 11, p = 2)$

Metrics	(1)	(2)	(3)
best k	6	26	6
λ	1	0,6	1
p	2	2	5
SMAPE,%	7,44	5,74	7,37

Таблица 3. Сравнение результатов работы алгоритма на данных о потреблении электроэнергии $f_3(t)$.

или иной метрики: наилучшая длина l может значительно отличаться для различных метрик.

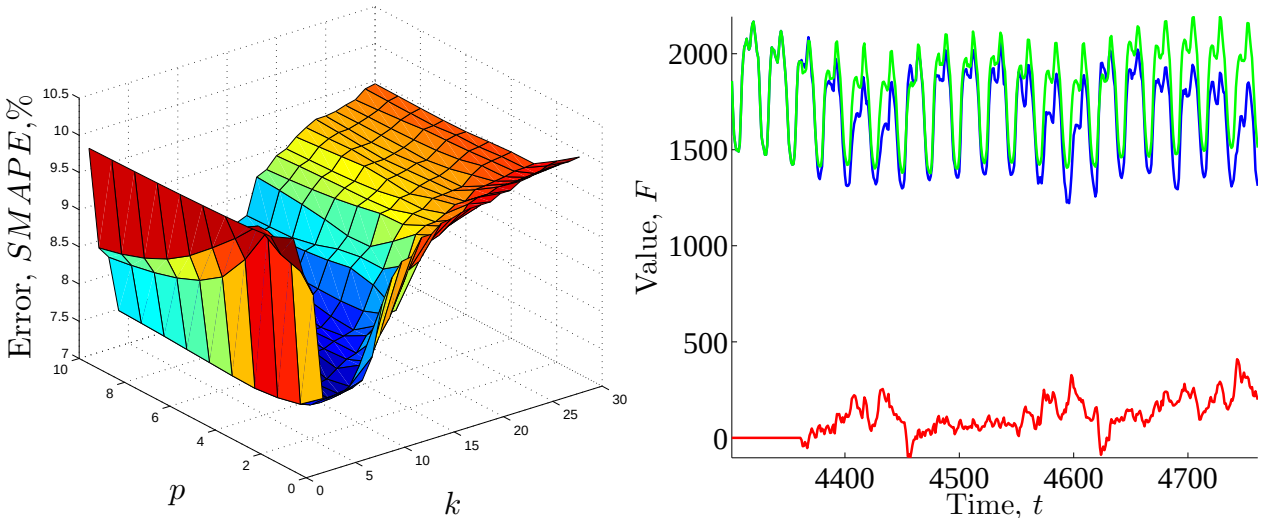

Рис. 11. Величина ошибки и построение прогноза для данных $f_3(t)$ $(k = 6, p = 5)$

Заключение

В данной работе рассмотрен локальный метод пронозирования временных рядов, основанный на алгоритме поиска "k ближайших соседей", исследована зависимость качества прогноза от используемой функции близости и от количества k ближайших соседей, проиллюстрированы результаты работы алгоритма на модельных рядах и реальных данных: о потреблении электроэнергии и о ценах на сахар, сравнительные результаты сведены в таблицы.

Литература

[1] McNames J.,*Innovations in local modeling for time series prediction* // Ph.D. Thesis, Stanford University, 1999.

[2] Воронцов К. В. Курс лекций *Математические методы обучения по прецедентам*

[3] Журавлев Ю. И., Рязанов В. В., и Сенько О. В. *Распознавание. Математические методы. Программная система. Практические применения.* // Фазис, Москва, 2005.

[4] Магнус Я. Р., Катышев П. К., Пересецкий А. А. *Эконометрика* // Дело, 2004, стр. 34-37

[5] Федорова В. П., *Локальные методы прогнозирования временных рядов* // Москва, 2009.

[6] Временной ряд (библиотека примеров) *http://www.machinelearning.ru/wiki/*