Розглянуто задачу прогнозування часових рядів великої довжини. Запропоновано метод побудови скорочених зважених рядів, що базується на сегментації початкового ряду та аналізі отриманих сегментів. Виконано розширення відомих методів прогнозування на зважені часові ряди

Ключові слова: часовий ряд, прогнозування даних, сегментація

Рассматривается задача прогнозирования временных рядов большого размера. Предложен метод построения сокращенных взвешенных рядов, основанный на сегментации начального ряда и анализе полученных сегментов. Выполнено расширение известных методов прогнозирования на взвешенные временные ряды

Ключевые слова: временной ряд, прогнозирование данных, сегментация

The problem of the large time series prediction is considered. The method of construction of the brief self-weighted time series, based on segmentation of initial row and analysis of the got segments is offered. Expansion of the known methods of prediction on the self-weighted time series is executed

Keywords: time series, data prediction, segmentation

УДК 004.67

МЕТОД ПОСТРОЕНИЯ ВЗВЕШЕННЫХ ВРЕМЕННЫХ РЯДОВ ДЛЯ РЕШЕНИЯ ЗАДАЧИ ПРОГНОЗИРОВАНИЯ

А.С. Миненко
Доктор физико-математических наук, доцент,
заведующий кафедрой
Кафедра системного анализа и моделирования*
Контактный тел.: (062) 304-92-58
E-mail: Minenko@suiai.edu.ua
Е.В. Волченко
Кандидат технических наук, доцент
Кафедра программного обеспечения интеллектуальных систем*
Контактный тел.: (062) 342-91-21, 050-932-45-72
E-mail: LM@mail.promtele.com
С.А. Шишкин*
Контактный тел.: (062) 342-91-21
E-mail: Shishkin.Sergey.ua@gmail.com
*Институт информатики и искусственного интеллекта
Донецкий национальный технический университет
пр. Б. Хмельницкого, 84, г. Донецк, Украина, 83050

1. Постановка проблемы и анализ литературы

Задача прогнозирования является одной из наиболее сложных и актуальных задач в области аналитической обработки данных [1]. В большинстве случаев приходится находить закономерности в небольших наборах данных, для которых на сегодняшний день разработано значительное количество эффективных методов прогнозирования, однако существует группа прикладных задач, в которых исследователь сталкивается с непрерывным увеличением объема информации. Нет сомнений, что необходимость построения прогнозов по большим базам данных усложняет и без того нетривиальную задачу анализа. Такая ситуация особенно характерна для бизнес - аналитики, где аккумулируется огромное количество информации, связанной с транзакциями: чеки, платежи, индексы и т.п. [2, 3]. На первый план выходит проблема построения группировок по исходным временным рядам, т.е. проблема сегментации ряда.

Понятие сегментации временных рядов заключается в следующей статистической проблеме: дан временной ряд t, необходимо найти разбиение этого ряда на m сегментов, которые являются внутренне однородными [4]. В зависимости от применения и цели разбиения необходимо найти стабильные периоды времени, точки изменения, или просто сжать исходный временной ряд в более компактное представление [3].

Для решения задач сегментации в последнее время все чаще прибегают к инструментарию теории динамических систем [5], использованию нечеткой логики [6], вейвлет-анализу исходного временного ряда [7]. При этом отсутствует возможность учета количества элементов в сегментах, удаленности сегментов от начала временного ряда, разброса элементов внутри сегмента.

Несмотря на значительную разнообразие существующих методов прогнозирования, основными являются ARIMA – регрессионные, авторегрессионные и скользящего среднего модели и классические эконометрические модели, основанные на восстановлении условных распределений будущих значений процесса относительно прошлых [8]. Данный класс методов достаточно широк, но не обладает универсальностью по отношению к входной информации. Так, если число воздействующих факторов велико и имеет место существенно нестационарное изменение данных, то применение моделей из этого класса приводит к большим ошибкам в прогнозных значениях. Также в настоящее время при прогнозировании довольно часто прибегают к нейросетевым моделям, показавшим свою эффективность при решении различных задач

(классификация и распознавание образов, кластеризация, управление). Однако в случае большого объема информации их использование оказывается невозможным из-за значительного времени, требуемого для предобработки данных и обучения [9, 10].

В работе [11, 12] для решения задачи сокращения объема обучающих выборок был предложен переход к взвешенным обучающим выборкам. Исходное множество объектов разбивается на подмножества, каждое из которых заменяется одним объектом. Для сохранения информации о количестве и взаимном расположении заменяемых объектов вводится дополнительный параметр – вес. Данная работа является продолжением исследований авторов в области решения задач на основе взвешенных выборок объектов и направлена на решение проблемы построения прогнозов по данным большого объема с использованием сокращенных взвешенных временных рядов.

Цель статьи – разработка метода построения взвешенных временных рядов и расширение существующих методов прогнозирования данных на взвешенные ряды.

2. Основные понятия. Постановка задачи

Рассматривается следующая постановка задачи. Дан временной ряд $t=\{y_i|1\le i\le n\}$ – конечное множество n отсчетов, отмеченных моментами времени $t_1...t_n$, $y=\{y_1,y_2,...,y_n\}$ – случайная наблюдаемая векторная функция. Пусть сегмент ряда t – это множество последовательных временных точек $S(a,b)=\{y_i\,|\,a\le i\le b\}$, т.е. $y_a,y_{a+1},...,y_b$. Необходимо выполнить сегментацию временного ряда t, т.е. найти разбиение t на m непересекающихся сегментов $S_t^m=\{S_i(a_i,b_i)|1\le i\le m\}$, таких что $a_1=1$, $b_m=n$ и $a_i=b_{i-1}+1$.

Введем для описания взвешенных сегментов дополнительный параметр w_i – весовой коэффициент сегмента.

Тогда каждый сегмент будем характеризовать парой параметров $s_i=\{(c_i,w_i)|1\le i\le m\}$, где c_i – значение центра сегмента. На выходе алгоритма будет получен взвешенный временной ряд $\{(c_1,w_1),(c_2,w_2),...,(c_m,w_m)\}$ с усредненными значениями элементов сегмента c_i и весом w_i.

3. Алгоритм сегментации временного ряда с заданным количеством уровней

Для решения задачи сегментации временного ряда предлагается алгоритм, основанный на дискретизации диапазона изменения данных прогнозной величины y. Исходный диапазон разбивается на заданное число уровней l, которое зависит от:

– величины разброса значений отсчетов ряда относительно их среднего значения;

– вида временного ряда (возрастающий, убывающий, периодический);

– длины ряда.

Выбор значения l осуществляется в зависимости от конкретной задачи, в общем случае число уровней предлагается рассчитывать по следующей формуле:

$$l=\lceil \log_{10} n\rceil+\left\lceil \frac{y_{max}-y_{min}}{y_{max}-\text{ср.знач.}}\right\rceil+\left\lceil \frac{y_{max}-y_{min}}{\text{ср.знач.}-y_{min}}\right\rceil,$$

где n – количество отсчетов в исходном временном ряду,

y_{max} – максимальное значение среди всех элементов сегмента,

y_{min} – минимальное значение среди всех элементов сегмента,

$\lceil\ \rceil$ – оператор округления до старшего целого.

Тогда формализованный алгоритм сегментации с заданным количеством уровней может быть описан следующим образом.

1. Вычисляются y_{max}, y_{min}, ср.знач по исходному временному ряду.

2. Рассчитывается количество уровней l и размер одного уровня по формуле:

$$size=\frac{y_{max}-y_{min}}{l}.$$

3. Определяется уровень, в котором находится первый отсчет временного ряда.

4. Создается первый сегмент, в который помещается первый отсчет временного ряда.

5. Для всех отсчетов временного ряда, начиная со второго, выполняются шаги 6 – 7.

6. Если текущий отсчет временного ряда находится в том же уровне, что и предыдущий – он помещается в текущий сегмент.

7. Если текущий отсчет временного ряда и предыдущий находятся в разных уровнях – создается новый сегмент, в который помещается текущий рассматриваемый отсчет.

Пример сегментации временного ряда на основе данного алгоритма представлен на рис. 1.

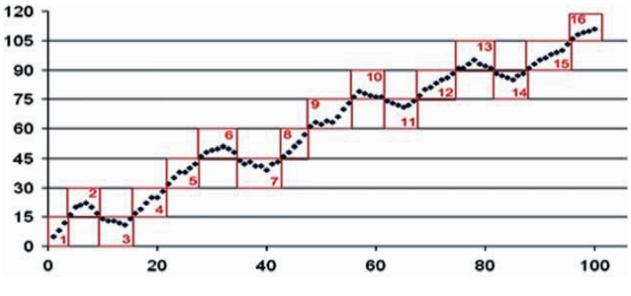

Рис. 1. Разделение временного ряда на сегменты

Для решения задачи сокращения ряда по каждому сегменту вычисляются центр и весовой коэффициент.

Значение центра вычисляется по формуле:

$$c_j=\frac{1}{n_j}\sum_{i=1}^{n_j} y_i,$$

где n_j – количество элементов в j-ом сегменте,

y_i – текущий элемент сегмента.

Использование взвешенных сегментов позволяет учитывать распределение точек по сегментам, а также степень влияния каждого сегмента на будущий прогноз. Весовые коэффициенты сегментов предлагается рассчитывать по одной из следующих формул.

1. Количество точек в сегменте:

$$w_j = n_j.$$

2. Площадь сегмента:

$$w_j = (y_{max\,j} - y_{min\,j}) \cdot (x_{max\,j} - x_{min\,j}).$$

3. Описание сегмента окружностью минимального радиуса:

$$w_j = \max[(y_{max\,j} - \text{ср.знач}_j), (\text{ср.знач}_j - y_{min\,j})] \cdot n_j.$$

4. Отклонение от среднего значения:

$$w_j = \sum_{i=1}^{n_j} |y_i - \text{ср.знач}_j|.$$

Поскольку во многих методах прогнозирования более поздним наблюдениям придается больший вес [2, 13], то все весовые коэффициенты предлагается пересчитать в соответствии с формулой:

$$w_j = w_j \cdot \frac{j}{m}.$$

Тогда нормализация весовых коэффициентов может быть осуществлена по следующей формуле:

$$w_j := \frac{w_j}{w_{max} + w_{max} \cdot k},$$

где w_{max} – максимальный весовой коэффициент, найденный по всем сегментам,

k – коэффициент нормализации.

4. Расширение алгоритмов прогнозирования данных на взвешенные временные ряды

Существующие методы прогнозирования позволяют строить прогнозы лишь по отсчетам исходного временного ряда, при этом нет возможности учета весовых коэффициентов. Для прогнозирования по взвешенным сегментам в настоящей работе предлагается ряд расширений наиболее известных и эффективных методов прогнозирования данных.

Алгоритм экстраполяция на основе среднего и скользящего среднего значений. Как и в оригинальном алгоритме, предполагается, что средний уровень ряда не имеет тенденции к изменению или это изменение незначительно. При этом в зависимости от весовых коэффициентов сегменты внесут разный вклад в среднее значение.

Прогноз на следующий период времени может быть найден по формуле:

$$\hat{Y}_{n+1} = \frac{\sum_{i=1}^{m} c_i \cdot w_i}{\sum_{i=1}^{n} w_i}.$$

Аналогично среднему значению находится и скользящее среднее, однако для усреднения используются значения только t последних сегментов. Таким образом, прогноз равен

$$\hat{Y}_{n+1} = \frac{\sum_{i=m-t}^{m} c_i \cdot w_i}{\sum_{i=m-t}^{m} w_i},$$

где t – количество сегментов, используемых для прогнозирования.

Метод среднего темпа. При прогнозировании средним темпом прогнозное значение может быть получено по формулам:

$$\hat{Y}_{n+1} = c_m + T,$$

$$\hat{Y}_{n+1} = c_m \cdot T,$$

где c_m – значение центра последнего сегмента;

T – средний темп роста.

Начальное значение темпа роста вычисляется по формулам:

$$T = c_2 - c_1,$$

$$T = c_2 / c_1.$$

В зависимости от выбранной модели, аддитивной или мультипликативной, средний темп роста взвешенных сегментов с учетом их весовых коэффициентов вычисляется по формулам:

$$T = T \cdot \left(1 - \frac{w_i - w_{i-1}}{2}\right) \cdot w_{cp} + (c_i - c_{i-1}) \cdot (1 - w_{cp}) \cdot \frac{w_i - w_{i-1}}{2},$$

$$T = T \cdot \left(1 - \frac{w_i - w_{i-1}}{2}\right) + \left(\frac{c_i}{c_{i-1}}\right) \cdot \frac{w_i - w_{i-1}}{2},$$

где w_{cp} – среднее значение весовых коэффициентов от первого до i -го сегментов.

Метод Хольта. В качестве расширения для данного алгоритма предлагается сглаженное значение уровня ряда строить по центрам сегментов, а при вычислении трендовой составляющей учитывать весовые коэффициенты.

Обобщенные формулы имеют вид:

$$\begin{cases} S_t = a \cdot c_t + (1-a) \cdot (S_{t-1} - b_{t-1}) \\ b_t = \beta \cdot (S_t - S_{t-1}) \cdot \left(1 - \frac{w_t + w_{t-1}}{2}\right) + (1-\beta) \cdot \frac{w_t + w_{t-1}}{2} \cdot b_{t-1} \end{cases},$$

где S_t – сглаженное значение уровня ряда;

b_t – трендовая составляющая;

α, β – коэффициенты сглаживания;

c_t – значение центра сегмента;

w_t – значение весового коэффициента сегмента.

Выбор начальных параметров S и b осуществляется следующим образом:

$$S_1 = c_1,$$

$$b_1 = c_2 - c_1.$$

Для прогнозирования следующего значения используется формула:

$$\hat{Y}_{t+1} = S_t + b_t.$$

Метод Хольта-Винтерса. Отличием метода Хольта-Винтерса от метода Хольта является попытка учесть сезонные составляющие в данных. В этом случае сглаженное значение уровня ряда и индексы сезонности предлагается вычислять по центрам сегментов, а в трендовой составляющей учитывать весовые коэффициенты. Система уравнений, описывающих данный метод, будет выглядеть следующим образом:

$$\begin{cases} S_t = a \cdot \dfrac{c_t}{I_{t-L}} + (1-a) \cdot (S_{t-1} - b_{t-1}) \\ b_t = \beta \cdot (S_t - S_{t-1}) \cdot (1 - \dfrac{w_t + w_{t-1}}{2}) + (1-\beta) \cdot \dfrac{w_t + w_{t-1}}{2} \cdot b_{t-1} \\ I_t = \gamma \dfrac{c_t}{S_t} + (1-\gamma) \cdot I_{t-L} \\ \hat{Y}_{t+i} = (S_t + i \cdot b_t) \cdot I_{t-L+i} \end{cases},$$

где α, β, γ – коэффициенты сглаживания;
S_t – сглаженное значение наблюдения;
b_t – коэффициент тенденции;
I_t – индекс сезонности;
$\hat{Y}_{t+i}$ – прогноз на i периодов вперед;
t – индекс текущего наблюдения;
c_t – значение центра сегмента;
w_t – значение весового коэффициента сегмента.

Коэффициенты α, β, γ подбираются таким образом, чтобы минимизировать среднеквадратическую ошибку.

Индексы сезонности рассчитываются следующим образом:

1) для каждого сезона рассчитывается среднее значение:

$$A_j = \frac{\sum_{i=1}^{L} c_{ji}}{L},$$

где j – изменяется от 1 до n;

2) для каждого периода рассчитывается индекс сезонности:

$$I_i = \frac{\sum_{j=1}^{n} \frac{c_{ji}}{A_j}}{n},$$

где c_{ji} – наблюдение, соответствующее i-му периоду j-го сезона;

i – изменяется от 1 до L.

Для оценки тенденции используется следующая формула:

$$b = \frac{1}{L} \cdot \left(\frac{c_{L+1} - c_1}{L} + \frac{c_{L+2} - c_2}{L} + \ldots + \frac{c_{L+L} - c_L}{L} \right).$$

Поскольку для всех рассмотренных методов прогнозирование осуществляется по сегментам, объединяющим в себе несколько отсчетов, предполагается, что результатом может являться как отдельное значение ряда (центр прогнозируемого сегмента), так и целая группа значений. Следовательно, полученный прогноз может быть продублирован для v значений:

$$\hat{Y}_{n+1} = \hat{Y}_{n+2} = \hat{Y}_{n+3} = \ldots = \hat{Y}_{n+v}$$

где v – среднее количество отсчетов в сегментах.

5. Анализ результатов экспериментальных исследований

Моделирование преследовало цель оценить качество предложенных метода построения взвешенных временных рядов и модификаций алгоритмов прогнозирования. Для оценки совокупной ошибки прогноза использовалась средняя относительная ошибка, которая определяется по формуле:

$$\delta_o = \frac{\sum_{t=1}^{n} \left(\frac{|y_t - \hat{y}_t|}{y_t} \right) \cdot 100\%}{n},$$

где y_t – значение временного ряда;
$\hat{y}_t$ – спрогнозированное значение;
n – количество прогнозируемых элементов.

Для экспериментальных исследований использовались временные ряды размерностью 100, 500 и 1000 отсчетов с показаниями температуры воздуха [14]. Исходные временные ряды и построенные по ним прогнозы представлены на рис. 2 – 4.

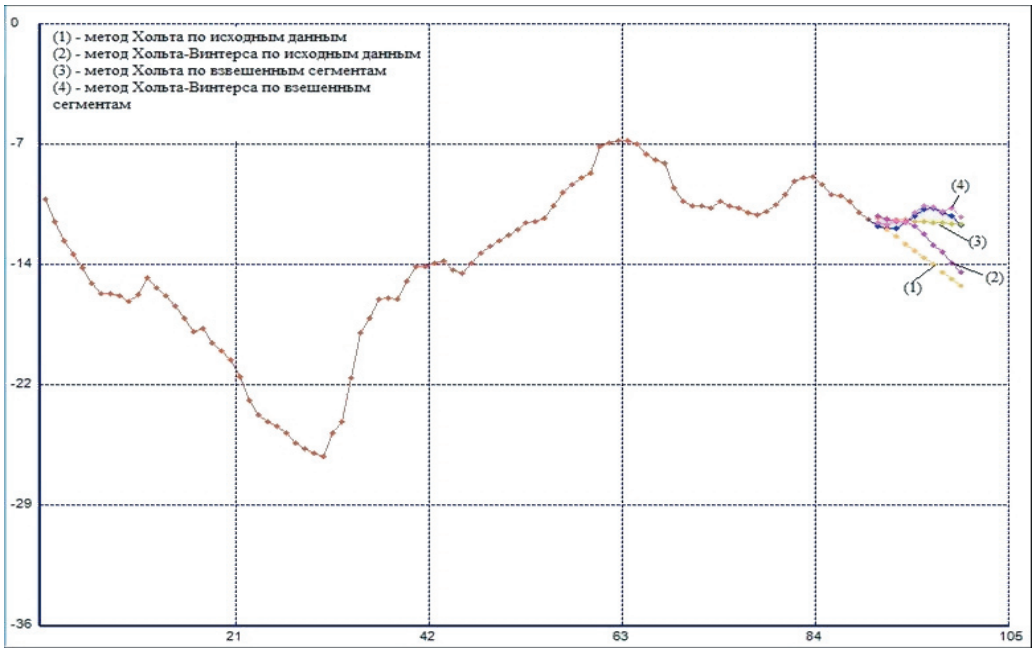

Рис. 2. Прогнозирование ряда размером 100 отсчетов

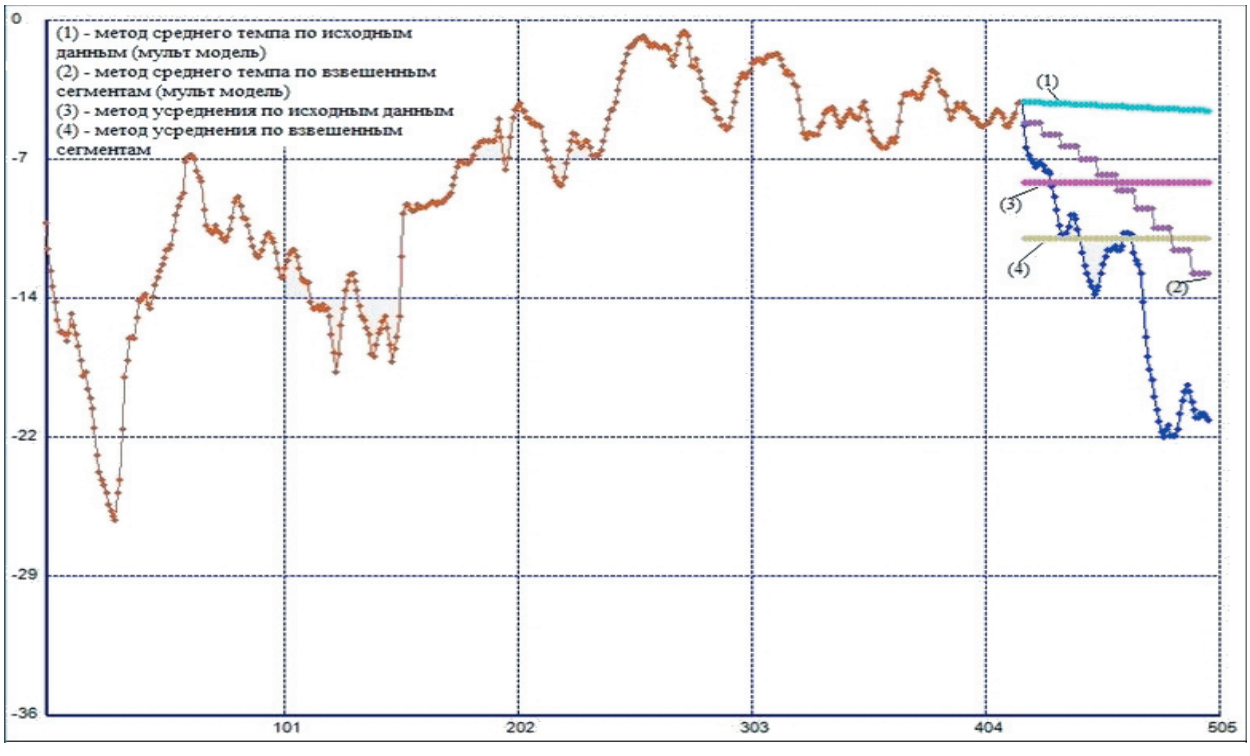

Рис. 3. Прогнозирование ряда размером 500 отсчетов

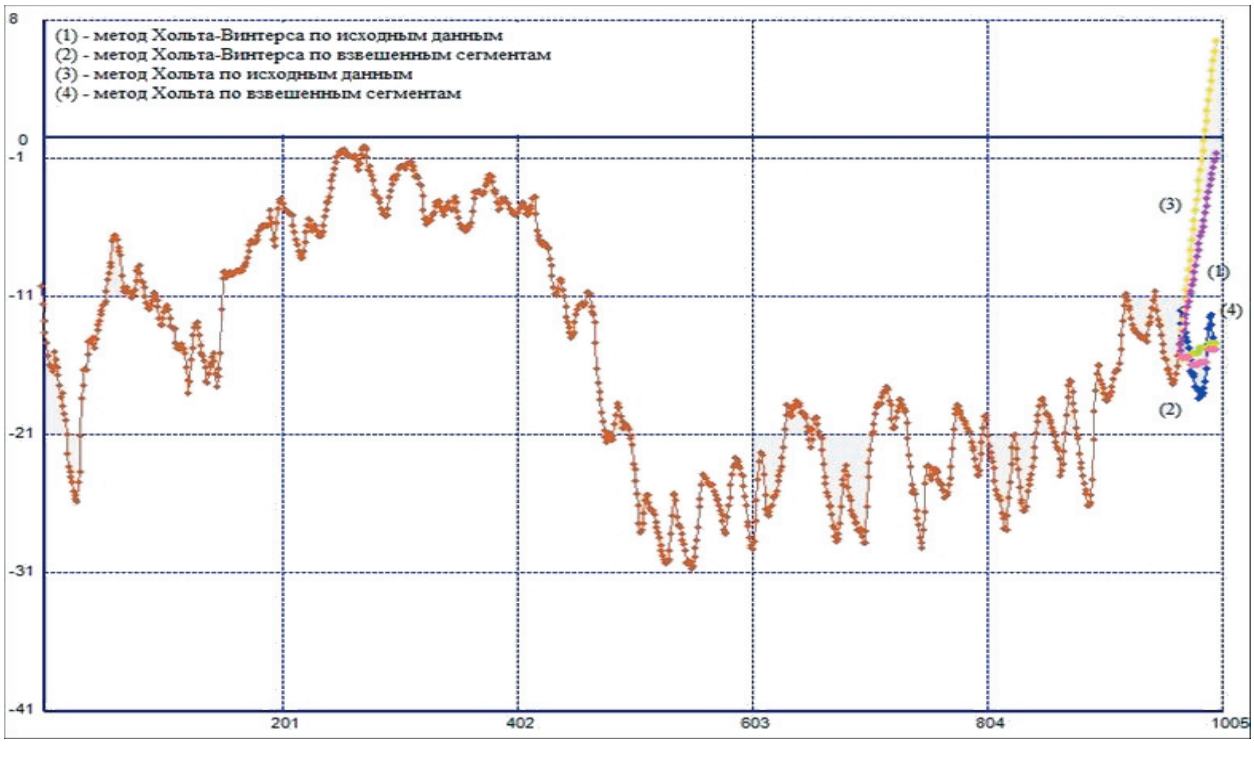

Рис. 4. Прогнозирование ряда размером 1000 отсчетов

На рис. 5 представлены результаты расчета ошибки для методов Хольта и Хольта-Винтерса, весовой коэффициент – отклонение от среднего значения, для метода Хольта-Винтерса размер сезона равен 11, для обучения и сегментации использовалось 90 отсчетов. На рис. 6 представлены результаты расчеты ошибки для методов среднего темпа и усреднения, весовой коэффициент – площадь сегмента, для обучения и сегментации использовался 421 отсчет.

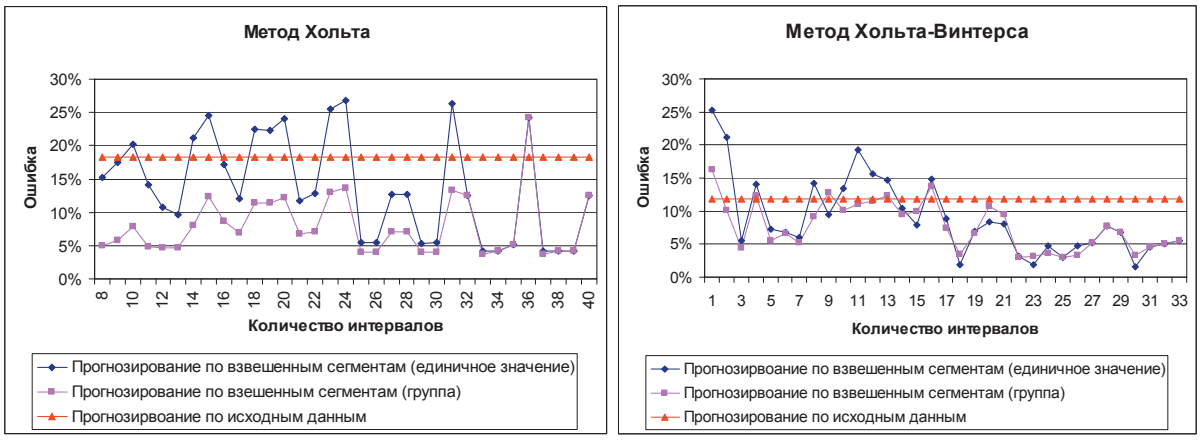

Рис. 5. Ошибка прогнозирования для 100 отсчетов

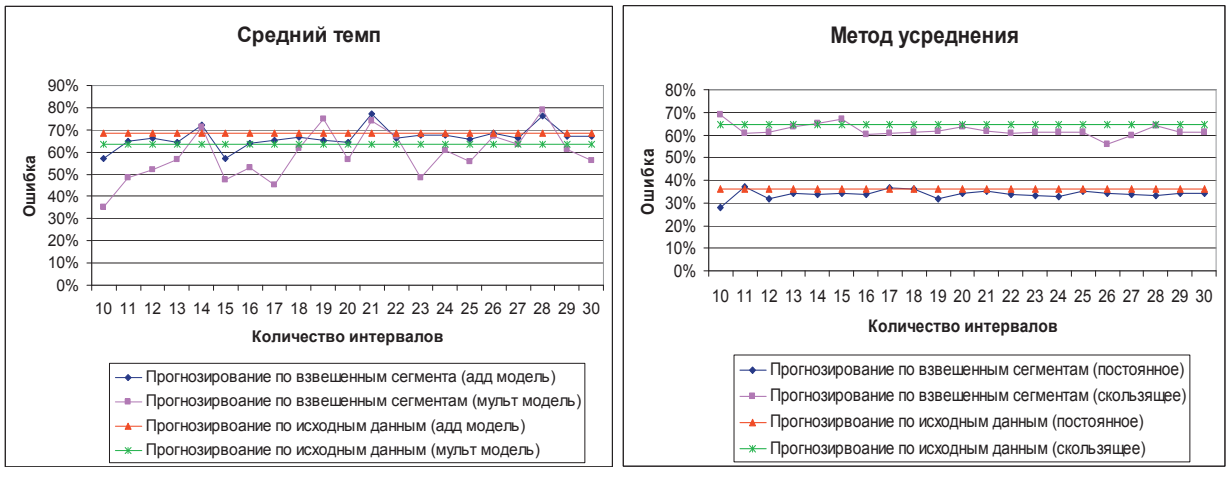

Рис. 6. Ошибка прогнозирования для 500 отсчетов

На рис. 7 представлены результаты расчета ошибки для методов Хольта и Хольта-Винтерса при прогнозировании группы значений, весовой коэффициент – площадь сегмента, для метода Хольта-Винтерса размер сезона равен 9, для обучения и сегментации использовалось 970 отсчетов.

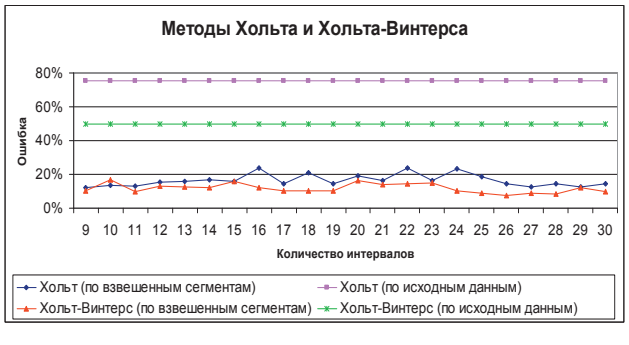

Рис. 7. Ошибка прогнозирования для 1000 отсчетов

Результаты моделирования продемонстрировали сокращение относительной ошибки прогноза по сравнению с оригинальными методами, а именно:

- при прогнозировании методом Хольта-Винтерса ряда размером 100 отсчетов относительная ошибка сократилась в среднем на 3% для единичных прогнозов и на 4% для группы, для ряда размером 1000 отсчетов ошибка сократилась в среднем на 37%;

- при прогнозировании методом Хольта ряда размером 100 отсчетов относительная ошибка сократилась в среднем на 5% для единичных прогнозов и на 10% для группы, для ряда размером 1000 отсчетов ошибка сократилась на 59%.

- для метода среднего темпа при прогнозировании ряда размером 500 отсчетов относительная ошибка сократилась в среднем на 2% для аддитивной модели и на 4% для мультипликативной;

- при прогнозировании методом усреднения ряда размером 500 отсчетов относительная ошибка сократилась в среднем на 2%.

Отметим, что для всех временных рядов размер взвешенного ряда не превышал 60% исходного ряда.

5. Выводы

В работе предложен новый подход к решению задачи прогнозирования данных по временным рядам большого размера, состоящий в построении сокращенных взвешенных временных рядов на основе выполнения сегментации. Предложенный метод сег-

ментации строит по отсчетам исходного временного ряда сокращенный ряд сегментов с усредненными значениями элементов, а использование весовых коэффициентов позволяет учитывать распределение точек в каждом сегменте.

Для построения прогнозов по взвешенным временным рядам предложены расширения существующих методов прогнозирования: алгоритмов экстраполяции на основе среднего и скользящего среднего значений, метода среднего темпа, метода Хольта и Хольта-Винтерса.

Результаты выполненных экспериментальных исследований на тестовых и реальных временных рядах показали, что использование взвешенных сегментов позволяет в среднем уменьшить относительную ошибку прогнозирования на 9% и сократить длину временного ряда на 40%.

Полученные результаты (сокращение размера временного ряда и сокращение относительной ошибки прогноза) свидетельствуют об эффективности предложенного подхода для решения актуальной задачи прогнозирования временных рядов большого размера.

Таким образом, на основе предлагаемого авторами подхода, заключающегося в построении взвешенных выборок данных, была успешно решена близкая к рассмотренным ранее задача прогнозирования временных рядов, что позволяет говорить об универсальности и эффективности использования взвешенных выборок для задач обработки больших массивов данных.

Литература

1. Larose D. T. Discovering knowledge in data: an introduction to data mining [Текст] / D. T. Larose. - New Jersey: John Wiley & Sons Inc., 2005. – 240 p.
2. Giudici P. Applied data mining: statistical methods for business and industry [Текст] / P. Giudici. - Chichester: John Wiley & Sons Inc., 2003. – 380 p.
3. Last M. Knowledge discovery in time series databases [Текст] / M. Last, Y. Klein, A. Kandel. – IEEE Transactions on Systems, man and cybernetics, 2000. – P. 60-69.
4. Vasko K. Estimating the number of segments in time series data using permutation tests [Текст] / K. Vasko, H. Toivonen. – IEEE International Conference on Data Mining, 2002. – P. 466 – 473.
5. Лоскутов А. Ю. Проблемы нелинейной динамики III: локальные методы прогнозирования временных рядов [Текст] / А. Ю. Лоскутов, О.Л. Котляров, И.А. Истомин, Д.И. Журавлев // Вестник Московского университета: Физика. Астрономия. – №6. – 2002. – С. 3 – 21.
6. Зайцев П. Н. Нечеткая сегментация временных рядов [Текст] / Зайцев П. Н. // Вестник ВГУ. Серия: Системный анализ и информационные технологии. – 2009. – №1. – С. 60-67.
7. Востров Г. М. Сегментация экономических временных рядов с использованием вейвлет-анализа [Текст] / Г. М. Востров, М. В. Поляков, В. В. Любченко // Труды Одесского политехнического университета. – 2003. – № 1(19). – С. 119 – 127.
8. Катулев А.Н. Стохастические модели прогнозирования цены [Текст] / А. Н. Катулев, Ан. Н. Сотников // Дискретный анализ и исследование операций. – 2002. – Т. 9. – №1. – С. 61 – 77.
9. Хайкин С. Нейронные сети: полный курс [Текст] / С. Хайкин.; [пер с англ.]. – 2-е издание – М.: Издательский дом «Вильямс», 2006. – 1104 с.
10. Волченко Е.В. Компактный генетический алгоритм выбора размера окон при нейросетевом прогнозировании временных рядов [Текст] / Е.В. Волченко // Проблеми інформатизації та управління. – 2011. – №1. – С. 42 – 48.
11. Розробка теоретичних засад і методів реалізації відкритих систем автоматичного розпізнавання, що навчаються: способи оптимізації навчаючих вибірок і методи побудови зважених вирішуючих правил класифікації [Текст] : звіт з НДР (заключний) : Тема GP/F32/130, Грант Президента України для підтримки наукових досліджень молодих учених на 2011 рік / керівник роботи О.В. Волченко. – Донецьк, ДВНЗ «ДонНТУ», 2011. – 67 с.
12. Волченко Е.В. Метод построения взвешенных обучающих выборок в открытых системах распознавания [Текст] / Е.В. Волченко // Доклады 14-й Всероссийской конференции «Математические методы распознавания образов (ММРО-14)», Суздаль, 2009. – М.: Макс-Пресс, 2009. – С. 100 – 104.
13. Сазонов В. Г. Прогнозирование и планирование в условиях рынка [Текст] / В.Г. Сазонов. – Владивосток: ДВГУ ТИДОТ, 2001. – 149 с.
14. Профессиональный информационно-аналитический ресурс, посвященный машинному обучению и интеллектуальному анализу данных [Электронный ресурс]. – Режим доступа: http://www.machinelearning.ru – Загл. с экрана.