И.Г. Сидоркина, Д.С. Шумков

КУСОЧНО-ЛИНЕЙНАЯ АППРОКСИМАЦИЯ ПРИ РЕШЕНИИ ЗАДАЧ ИЗВЛЕЧЕНИЯ ДАННЫХ

***Аннотация**. Прогнозирование является одним из основных вопросов, которые возникают при анализе временных рядов. При этом ставится задача определить будущее поведение временного ряда по его известным прошлым значениям. В данной работе предложен метод для прогнозирования временных рядов, который базируется на идеи выделения базовых паттернов (шаблонов) из исходных данных и позволяет установить внутренние закономерности исследуемого ряда.*
На сегодняшний момент одним из подходов, в котором ведутся исследования в области прогнозирования временных рядов, является системы Data Mining или "раскопка данных". Это связано с тем, что классические методы, основанные исключительно на линейных (ARIMA) и нелинейных (GARCH) моделях прогнозирования, не позволяют достичь необходимой точности прогноза. Используя методы, разработанные в рамках данной технологии, можно увеличить эффективность прогнозирования и выявить скрытые закономерности в исследуемых временных рядах.
***Ключевые слова:** временной ряд, аппроксимация, кусочно-линейная аппроксимация, прогнозирование, раскопки данных, базовые шаблоны, паттерны, локальные экстремумы.*

Основными этапами в предложенном методе прогнозирования временных рядов, основанного на выделении базовых паттернов являются следующие:

1. Построение кусочно-линейной аппроксимации временного ряда;
2. Выделение основных шаблонов (паттернов);
3. Построение таблицы переходов шаблонов из одного состояния в другое;
4. Сравнение текущего состояния с основными паттернами для прогнозирования будущего поведения значений временного ряда.

Рассмотрим каждый пункт метода более подробно. В качестве исходного временного ряда возьмем ряд, представленный на рисунке (рис 1).

Рис. 1

Исследуемый временной ряд

DOI: 10.7256/2305-6061.2013.2.7943

Он обладает достаточно сложной структурой и для того, чтобы сгладить его значения и выделить основные шаблоны, проведем кусочно-линейную аппроксимацию.

Известно достаточное количество алгоритмов для построения кусочно-линейной аппроксимации, но, несмотря на это, каждый из них может быть отнесен к одной из этих групп:

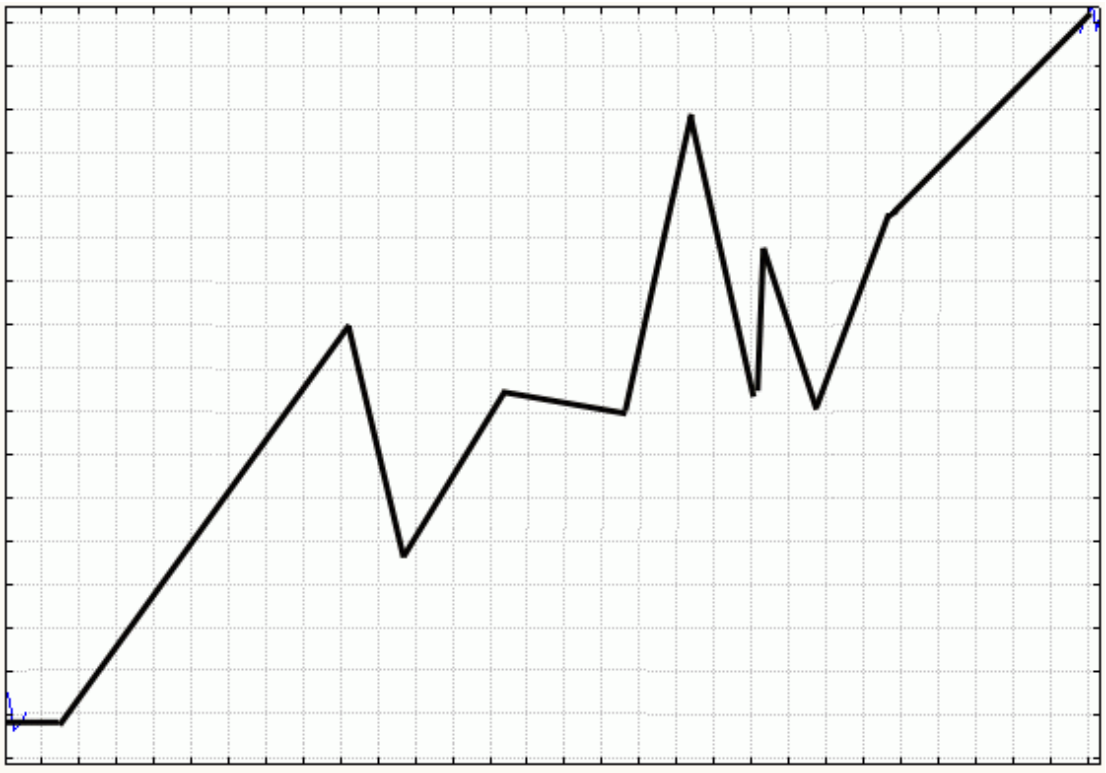
– Sliding Window Algorith (SW) или алгоритм скользящего окна;
– The Top-Down Algorithm (TD) или алгоритм спуска сверху вниз;
– The Bottom-Up Algorithm (BU) или алгоритм снизу-вверх.

В работе [1] была предложена методика сравнения эффективности каждого алгоритма кусочно-линейной аппроксимации путем расчета максимальной ошибки *max_error*. В результате экспериментов было выявлено, что наиболее «слабым» методом является SW алгоритм. При этом эффективность алгоритмов BU и TD является почти одинаковой, хотя в ряде случаев использование BU является более предпочтительным.

В связи с тем, что алгоритм BU дает лучшие результаты, он был принят как основной метод для кусочно-линейной аппроксимации в предложенном методе прогнозирования. Таким образом, исходный ряд был преобразован алгоритмом BU к следующему виду (рис 2).

Рис. 2

Аппроксимируемый временной ряд



Перейдем ко второму пункту предложенного метода прогнозирования, на котором главной задачей является выделение основных паттернов временного ряда. К двум наиболее известным методам, в основе которых лежит кусочно-линейная аппроксимация, относятся:

– Adaptive piecewise constant approximation (APCA) или адаптивная кусочно-постоянная аппроксимация [2];
– Landmark method или алгоритмы выделения базовых элементов (например, локальных экстремумов [3,6]) в числовой последовательности.

DOI: 10.7256/2305-6061.2013.2.7943

В данной работе представлен оригинальный алгоритм выделения основных шаблонов, в основе которого лежит модель инвариантная к следующим трансформациям:

– изменение масштаба времени;
– временному сдвигу.

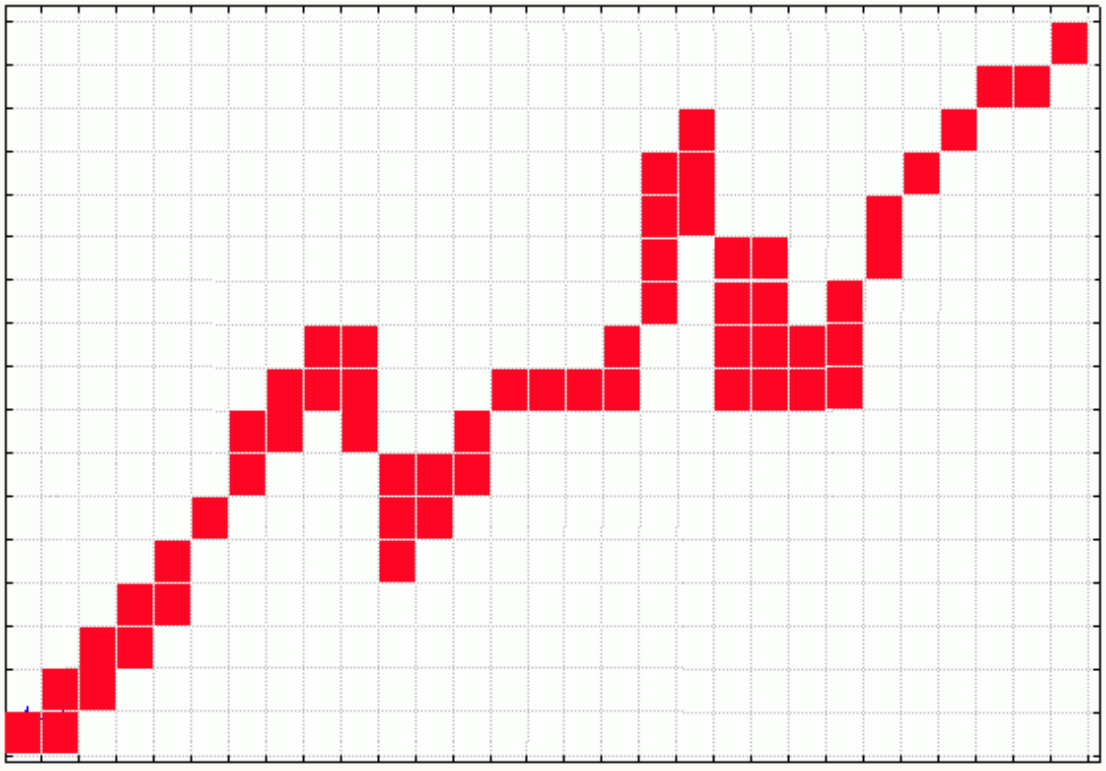
Он более гибок по сравнению с APCA и более точен по сравнению с методом локальных экстремумов.

Основными пунктами предложенного алгоритма выделения паттернов являются следующие шаги:

1. Строится таблица *M* размером *n* x *k*, которая накладывается на участок временного ряда, преобразованный методом кусочно-линейной аппроксимации;
2. Далее в ячейку $M_{i,j}$ ставится значение *1*, если хотя бы одно значение временного ряда лежит внутри данной ячейки, и *0* в обратном случае. На плоскости это можно представить следующим рисунком (рис 3), где ячейки матрицы с «*1*» представляют собой закрашенные «прямоугольники».

Рис. 3

Преобразованный временной ряд



Полученная матрица представляет собой образ исходного ряда. Ее особенностью является то, что изменяя количество исходных столбцов и строк можно менять точность отображения временного ряда.

В виду того, что матрица содержит значения только «*1*» и «*0*» и является бинарной, сравнение паттернов друг с другом для определения «схожести» выполняется очень быстро. В качестве оценки позволяющей определить насколько один шаблон отличается от другого, введем величину *pattern_error*. Данная величина рассчитывается следующим образом. Пусть имеется 2 паттерна, которые описываются своими бинарными матрицами M^1 и M^2. Количество значений $M^1_{i,j}$ и $M^2_{i,j}$, в которых $M^1_{i,j} <> M^2_{i,j}$ и будет равняться *pattern_error*.

Введем также величину *max_pattern_error*, которая будет описывать максимальную ошибку, при которой один шаблон отличается от другого. Варьируя данную величину, можно задавать насколько значения данного паттерна могут отличаться в рамках одного класса.

DOI: 10.7256/2305-6061.2013.2.7943

Итак, для того, чтобы выделить основные паттерны в исследуемом ряде, необходимо сравнить каждый полученный после аппроксимации шаблон друг с другом. Если *pattern_error > max_pattern_error*, тогда экземпляры необходимо определить в два различных класса. Проведя данную операцию над всеми паттернам в исходном временном ряду, получается набор базовых шаблонов для данного ряда.

На третьем шаге метода прогнозирования необходимо построить таблицу переходов одного базового паттерна в другой. Необходимо также учесть, что после одного и того же базового образа в зависимости от положения временного ряда переход может быть осуществлен в разные паттерны, поэтому требуется сохранить количество переходов в каждый их базовых шаблонов. Таким образом, будет получена таблица следующего вида:

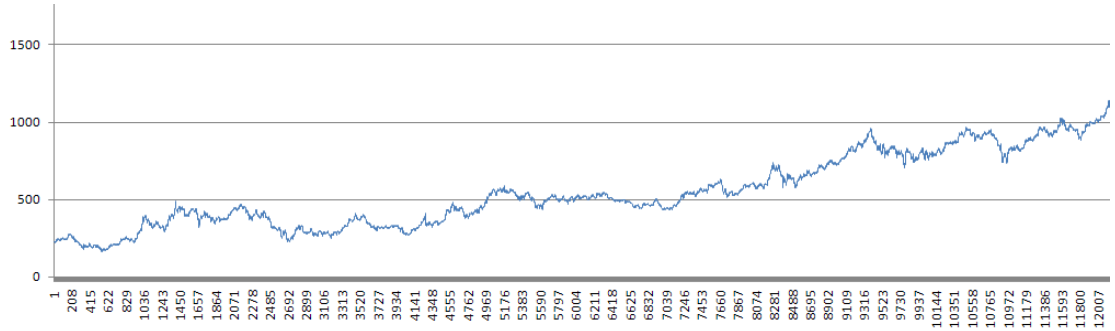
$P_1 \rightarrow P_4$: 4 $P_1 \rightarrow P_6$: 10 $P_2 \rightarrow P_3$: 1 $P_3 \rightarrow P_1$: 6 …… $P_i \rightarrow P_j$: m	где P_i… P_j это базовые паттерны, а цифры «4, 10, 1, 6, … , m» количество переходов из одного состояния в другое.

На четвертом шаге предложенного метода происходит прогнозирование временного ряда. Для этого вычисляется базовый паттерн для последних значений ряда, которые определяют текущий базовый шаблон. В таблице переходов выбираются те базовые элементы, из которых переходит текущий базовый паттерн. Они и определяют, как будет вести себя временной ряд дальше. При этом, чем больше значение соответствующее каждому переходу в одноименной таблице, тем больше вероятность того, что этот базовый паттерн даст более точный прогноз.

Для проверки эффективности предложенного метода прогнозирования ниже приводятся результаты его работы для следующего временного ряда, содержащего 12010 отсчетов (рис 4). Особенностью представленного процесса является его нестационарность.

Рис. 4

Исследуемый временной ряд



После первого и второго этапа было выделено 107 базовых паттернов и построена таблица переходов из одного базового паттерна в другой. В виду того, что каждый паттерн мог перейти в любой другой, для дальнейшего прогнозирования использовались только те шаблоны, у которых число переходов было максимальным. Другими словами, использовались наиболее вероятные базовые паттерны. В результате данного эксперимента была рассчитана средняя величина ошибки прогнозирования, составившая 17%. Одним из возможных решений, которые могут помочь снизить величину ошибки и улучшить прогноз, является метод, использующий для расчета не только наиболее вероятные базовые

DOI: 10.7256/2305-6061.2013.2.7943

паттерны, но и дополнительные шаблоны переходов. Этот вариант является предметом дальнейшего исследования.

Таким образом, предложен и исследован метод для прогнозирования временных рядов, основанный на выделении базовых паттернов из исходных данных. Одной из его особенностей является оригинальный способ выделения шаблонов с использованием кусочно-линейной аппроксимации. Достоинством данного метода является то, что он может применяться для временных рядов различной природы и структуры, используемых в системах *Data Mining или "раскопка данных"*. Одним из необходимых условий его корректной работы является достаточная длина исследуемого ряда.

Список литературы:

1. Last M., Kandle A, Bunke H. Data Mining in Time Series Databases. Series in Machine Perception and Artificial Intelligence. Vol. 57. — World Scientific, 2004. — 205 p.
2. Witten, I.H. (Ian H.) Data mining: practical machine learning tools and techniques / Ian H. Witten, Eibe Frank. — 2nd ed. — Elsevier, 2005. — 558 p.
3. Keogh, E.J., Chakrabarti, K., Mehrotra, S., and Pazzani, M.J. (2001a). Locally Adaptive Dimensionality Reduction for Indexing Large Time Series Databases. Proc. 2001 ACM SIGMOD Conf. on Management of Data. Pp. 151–162.
4. Kim, S., Park, S., and Chu, W.W. (2001). An Index-Based Approach for Similarity Search Supporting Time Warping in Large Sequence Databases. Proc. 17th Int. Conf. on Data Engineering (ICDE). Pp. 607–614.
5. Дюк В. Data Mining: учебный курс (+ CD). — СПб: Питер, 2001. — 386 с.
6. Лебедев Е.К. Вычисление вероятностей переходов для цепей Маркова, аппроксимирующих сигналы в фазовых системах / Е.К. Лебедев, Н.А. Галанина, Н.Н. Иванова // Вестник Чувашского университета. — 2001. — № 3. — С. 89-100.

References (transliteration):

1. Last M., Kandle A, Bunke H. Data Mining in Time Series Databases. Series in Machine Perception and Artificial Intelligence. Vol. 57. — World Scientific, 2004. — 205 p.
2. Witten, I.H. (Ian H.) Data mining: practical machine learning tools and techniques / Ian H. Witten, Eibe Frank. — 2nd ed. — Elsevier, 2005. — 558 p.
3. Keogh, E.J., Chakrabarti, K., Mehrotra, S., and Pazzani, M.J. (2001a). Locally Adaptive Dimensionality Reduction for Indexing Large Time Series Databases. Proc. 2001 ACM SIGMOD Conf. on Management of Data. — Pp. 151-162.
4. Kim, S., Park, S., and Chu, W.W. (2001). An Index-Based Approach for Similarity Search Supporting Time Warping in Large Sequence Databases. Proc. 17th Int. Conf. on Data Engineering (ICDE). Pp. 607–614.
5. Dyuk V. Data Mining: uchebnyi kurs (+ CD). — SPb: Piter, 2001. — 386 s.
6. Lebedev E.K. Vychislenie veroyatnostei perehodov dlya cepei Markova, approksimiruyushih signaly v fazovyh sistemah / E.K. Lebedev, N.A. Galanina, N.N. Ivanova // Vestnik Chuvashskogo universiteta. — 2001. — №3. — S. 89–100.

DOI: 10.7256/2305-6061.2013.2.7943