УДК 330.47:004.021+519.728.2

Минц А.Ю., к.э.н., доцент,
доцент кафедры финансов и банковского дела
Приазовский государственный технический университет

МЕТОД УПРОЩЕНИЯ ДИНАМИЧЕСКИХ РЯДОВ С ИСПОЛЬЗОВАНИЕМ ГЕНЕТИЧЕСКИХ АЛГОРИТМОВ

Минц А.Ю. Метод упрощения динамических рядов с использованием генетических алгоритмов. В статье описан метод упрощения динамических рядов путем выделения значимых точек перегиба. При этом обеспечивается высокая степень сжатия данных и сохранение информации о пиковых уровнях и закономерностях в них. Для выделения точек перегиба с заданной значимостью использован аппарат генетических алгоритмов.

Ключевые слова: динамические ряды, квантование, упрощение, сжатие, фильтрация, генетические алгоритмы, Matlab.

Мінц О.Ю. Метод спрощення динамічних рядів із використанням генетичних алгоритмів. У статті описано метод спрощення динамічних рядів шляхом виділення значущих точок перегину. При цьому забезпечується висока ступінь стиснення даних і збереження інформації про пікові рівні та закономірності в них. Для виділення точок перегину із заданою значимістю використано апарат генетичних алгоритмів.

Ключові слова: динамічні ряди, квантування, спрощення, стиснення, фільтрація, генетичні алгоритми, Matlab.

Mints O.Y. The method of simplifying the time series using genetic algorithms. This article describes a method to simplify the time series by allocating significant inflection points. In compare to existing methods, it provides a high compression level and save the data patterns and peak levels. To highlight inflection points with desired importance the genetic algorithms are used.

Keywords: time series, quantization, simplification, compression, filtering, genetic algorithms, Matlab.

Постановка проблемы. Исследования динамики появления и накопления информации в современном обществе объективно показывают, что оно находится в состоянии, которое получило название «информационный взрыв». Массовое внедрение компьютерных технологий привело к тому, что количество информации, которая нуждается в обработке, увеличивается в геометрической прогрессии [1, с. 63] примерно на 30% ежегодно. В этих условиях увеличивается актуальность методов, позволяющих упростить восприятие данных, а также их хранение и передачу.

Анализ последних исследований и публикаций. Проблема упрощения данных, представленных в виде динамических рядов, до сих пор исследовалась в основном применительно к теории передачи информации и цифровой обработке аналоговых сигналов (Л. Бриллюэн, Л.Р. Рабинер, М.В. Гашников, Н.Б. Паклин и др.). Кроме того, похожие методы существуют в рамках экономической статистики и некоторых других дисциплинах. Однако все они разрабатывались в условиях дефицита вычислительных ресурсов и поэтому реализуют лишь простые приемы, не позволяя одновременно получить большую степень сжатия данных и сохранить их адекватность.

Постановка задания. Целью исследования является разработка эффективного метода упрощения динамических рядов. Для достижения этой цели проанализированы преимущества и недостатки существующих методов, на основании чего сформулированы принципы, позволяющие избежать этих недостатков. Построена математическая модель поиска оптимального представления данных с различной степенью детализации. Данная модель адаптирована к оптимизации с использованием генетических алгоритмов и реализована в системе Matlab. Проанализированы результаты моделирования.

Изложение основных результатов. Темпы роста объемов информации опережают скорость развития средств её обработки. Вследствие этого на конкурентоспособность экономических субъектов и социальных групп существенно влияют как неравенство доступа к средствам информационных технологий, получившее название «первый цифровой разрыв» (англ. digital divide), так и неравенство в знаниях об использовании таких технологий (второй цифровой разрыв) [2 с. 5]. Причем если страны – лидеры технологического прогресса в настоящее время работают над решением проблем второго цифрового разрыва, то в Украине всё еще актуальна проблема первого разрыва.

Именно с проблемой обработки больших объемов информации связано бурное развитие технологий обработки больших объемов данных (Big Data), позволяющих находить закономерности в базах из

миллионов записей. В частности, очередной пик популярности переживает такой инструмент интеллектуальной обработки информации, как нейронные сети, возможности которых по распознаванию графических и звуковых объектов в настоящее время достигли человеческих, а в некоторых случаях даже превосходят их.

По сути, основная задача, которую решают эти и другие подобные методы, сводится к снижению размерности данных до уровня, достаточного для дальнейшей обработки традиционными методами анализа или для непосредственного восприятия человеком. То же предназначение имеет и метод упрощения динамических рядов, рассматриваемый в данной статье.

Динамическим рядом называется ряд чисел или ряд однородных статистических величин, показывающих изменения размеров какого-либо явления или признака во времени [3, с. 85, 53]. В зависимости от составляющих величин различают несколько типов динамических рядов.

Базовыми являются динамические ряды, построенные из абсолютных величин (курс валюты, объемы реализации товара и так далее). Остальные типы получаются путем обработки абсолютных показателей, а именно:

– производные динамические ряды, представленные относительными величинами и демонстрирующие изменения каких-либо коэффициентов (изменение цен, изменение объемов реализации и так далее);

– динамические ряды, состоящие из средних величин, например показателей средней реализации товаров за период времени, среднего дохода на душу населения и так далее.

Каждый динамический ряд состоит из двух элементов: отрезков времени (периодов), в рамках которых был зафиксирован определенный показатель и показателей, характеризующих объект исследования (уровни ряда). Если состояние объекта описывается m-показателями, то массив данных A будет иметь вид:

$$A=\{t_i, a_{i1}, a_{i2}, ..., a_{ij}\} \quad i=1..m,\ j=1..n.$$

В некоторых случаях, если отсчет показателей ведется с одинаковыми интервалами времени, элемент t_i в массиве данных может быть пропущен, поскольку легко восстанавливается из t_0, i и Δt, которые обычно известны.

Современные информационные технологии дают возможность получать данные с высокой частотой, что позволяет лучше отслеживать изменения в состоянии объектов, но усложняет восприятие полученных данных и их последующую обработку, так как они содержат большое количество избыточной информации – шума. Такие данные могут быть подвергнуты обработке с целью удаления небольших изменений показателей и выделения значимых тенденций. Поскольку в результате количество точек отсчета (m) уменьшается, можно говорить об упрощении данных.

Среди существующих методов решения этой задачи можно выделить как общие, так и специальные. К общим относится, например, метод квантования по уровню – разбивка диапазона значений непрерывной или дискретной величины на конечное число интервалов. Если в течение некоторого периода времени значение величин динамического ряда не выходило за пределы одного интервала, то в результате квантования все эти величины будут заменены одним значением [4, с. 184].

Достоинствами алгоритма квантования являются простота реализации и минимальные требования к вычислительным ресурсам, что обусловило его распространение в методах цифровой обработки сигналов. Недостаток метода – сильные искажения при большой степени сжатия.

Специальные методы применяются в узких предметных областях и максимально учитывают их особенности. Так, для сжатого представления биржевых данных используется метод квантования по времени, модифицированный так, что каждый дискретный временной интервал описывается не одним, а четырьмя значениями (цена открытия и закрытия, а также максимальная и минимальная цены за период). Существует несколько стандартных временных интервалов, например 1 час, 4 часа, 1 день. Этот метод удобен тем, что независимо от степени сжатия сохраняются важнейшие с точки зрения биржевых операций характеристики временного ряда. Кроме того, результаты обработки могут быть наглядно представлены в виде «японских свечей» (рис. 2) или графика с засечками [5].

Вместе с тем данный метод является узкоспециализированным, поскольку рассчитан на обработку только одного показателя – цены, а кроме того, требует от человека определенной подготовки для восприятия результатов.

Общим недостатком рассмотренных методов является наличие фиксированного шага квантования, что не позволяет получить высокую степень сжатия данных при сохранении достаточного уровня аппроксимации.

Также можно отметить ряд методов механического сглаживания динамических рядов, привнесенных из статистики. Например, метод укрупнения интервалов, аналитического выравнивания, сглаживания рядов при помощи скользящих средних и подобные. Все они также не лишены недостатков, основными из которых являются сглаживание пиковых уровней и потеря закономерностей в данных [6]. Кроме того, в результате сглаживания не всегда происходит сокращение точек отсчета значений ряда.

Большинства из перечисленных недостатков позволяет избежать предлагаемый в данной статье метод, основанный на квантовании по времени с переменным шагом. Рассмотрим его.

В общем виде решаемую задачу сформулируем следующим образом.

Пусть имеется некоторая последовательность данных, образующих динамический ряд, который задан массивом $A=\{t_i, a_i\}$, $i=1..m$.

Показатель m, соответствующий количеству точек отсчета для данного ряда, назовем *сложностью* ряда.

Упрощением динамического ряда назовем такое преобразование:

$$A \to \overline{A}, \quad (1)$$

где

$$A=\{t_i, a_i\}, \quad i=1..m,$$

при котором

$$\overline{m} < m, \quad (2)$$

а мера соответствия рядов A и $\overline{A}$

$$\Psi(A, \overline{A}) \to \max. \quad (3)$$

Однако для практического использования условий (1–3) недостаточно, поскольку очевидным решением будет ряд, отличающийся от исходного только на один элемент. Поэтому необходимо ввести дополнительное условие минимизации точек отсчета:

$$\Theta(\overline{m}) \to \min, \quad (4)$$

где Θ – мера сложности ряда.

Выражения (1–4) составляют законченную модель упрощения ряда, но определение функций Ψ и Θ требует дополнительных пояснений.

Функция Ψ имеет экономический смысл премии за соответствие рядов A и $\overline{A}$. При этом соответствие выражается в совпадении трендов в рядах и зависит от силы тренда.

Функция Θ имеет экономический смысл штрафа за каждую точку отсчета. Варьируя величину штрафа можно управлять относительной сложностью ряда $\overline{A}$.

Рассмотрим практические примеры таких функций.

Пусть A – динамический ряд, содержащий m пар элементов $\{t_i, a_i\}$, причем

$$a_i \in [0;1], \ \text{ti=i}. \quad (5)$$

Если значения элементов a_i не соответствуют условиям (5), то необходимо провести нормирование, а если данные в ряде располагаются не через равные промежутки времени, – то интерполяцию.

Поскольку решение задачи сводится к определению точек перегиба ряда, введем в модель вектор решений системы, $S=\{s_j\}$, $j=1..m$, где s_j – индексный элемент, численно равный элементу t_i исходного ряда A, то есть месту расположения точки перегиба.

Функция соответствия Ψ тогда запишется следующим образом:

$$\Psi = \sum_{j=2}^{\overline{m}} \left| a_{t_j} - a_{t_{j-1}} \right|, \quad (6)$$

$$t_1 = 1$$

Функцию штрафов можно записать так:

$$\Theta = l \cdot \overline{m}, \quad (7)$$

где l – штраф за каждую точку отсчета.

В общем виде целевая функция модели имеет вид:

$$\max z = \Psi - \Theta. \quad (8)$$

Очевидно, что наибольшим соответствием в общем случае обладает ряд, отсчитывающийся по всем тем же точкам, что и исходный. Но, с другой стороны, для него наибольшей будет и функция штрафов. При этом чем больше разница между значениями результирующего ряда в соседних точках отсчета, тем больше значение целевой функции. Таким образом, решение представляет собой компромисс между точностью соответствия исходным данным и количеством точек отсчета, с другой стороны.

Процесс решения задачи сводится к нахождению оптимального размера и значений вектора S. Единственным путем нахождения глобального оптимума такой задачи является полный перебор всех возможных размеров и значений вектора S. Поскольку вычислительная сложность такого решения возрастает экспоненциально с увеличением количества показателей m, оно относится к классу NP и за приемлемое время может быть найдено с использованием алгоритмов многофакторной оптимизации, которые позволяют за приемлемое время найти достаточно хорошее решение. К ним относятся генетические алгоритмы, метод имитации отжига, метод муравьиных колоний и другие. В данном случае рассматривается использование генетических алгоритмов (ГА).

Модель (6–8) была реализована автором в системе Matlab с использованием функций пакета расширений Genetic Algorithm and Direct Search Toolbox. Входными параметрами программы являются сам динамический ряд, значение штрафа за точку отсчета и некоторые параметры самого генетического алгоритма (размер популяции, условия остановки, вероятность мутации).

Результаты моделирования представляются как в графическом виде, так и в виде значения функции приспособленности лучшей особи в популяции. Ввиду особенностей данного пакета расширения задачи оптимизации могут решаться только на минимум. Поэтому при адаптации модели к программной среде Matlab целевая функция была переписана в виде

$$\min z = \Theta - \Psi,$$

а также сделаны некоторые другие не принципиальные упрощения.

В качестве входных данных, на которых тестировался алгоритм, выбраны данные об изменении курса евро по отношению к доллару США за январь 2016 г. (120 точек отсчета) и данные по изменению

дневной температуры в г. Мариуполь с января по сентябрь 2016 г. (259 точек отсчета). Для облегчения визуального анализа качества работы алгоритма на графиках присутствуют как исходный, так и результирующий ряды.

На рис. 1 и рис. 2 показаны графики изменения показателя дневной температуры, обработанные алгоритмом с различной степенью сжатия.

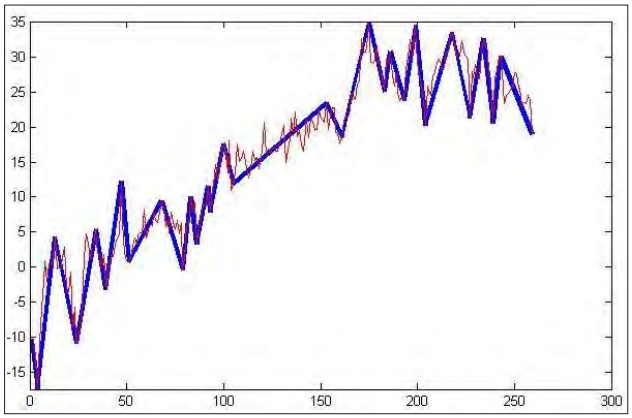

Рис. 1. Исходный и упрощенный графики изменения показателя дневной температуры в январе-сентябре 2016 г. при степени сжатия *l*=0.1

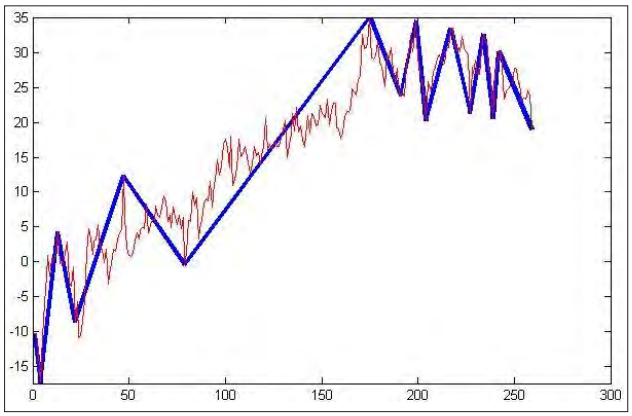

Рис. 2. Исходный и упрощенный графики изменения показателя дневной температуры в январе-сентябре 2016 г. при степени сжатия *l*=0.2

Как видим, даже при небольшой степени сжатия при сохранении практически всех основных тенденций, наблюдаемых на графике, количество точек отсчета снижается с 259 до 30, то есть более чем в восемь раз.

Последующее повышение уровня сжатия позволяет снизить количество точек отсчета до 16 (рис. 2). При еще большем увеличении сжатия количество точек отсчета доходит до трех (начало ряда, точка максимума и конец ряда). Но даже в этом случае получившегося графика достаточно для отслеживания основных тенденций изменения исходного показателя.

Аналогичные результаты были получены и при обработке данных о колебаниях валютного курса (рис. 3).

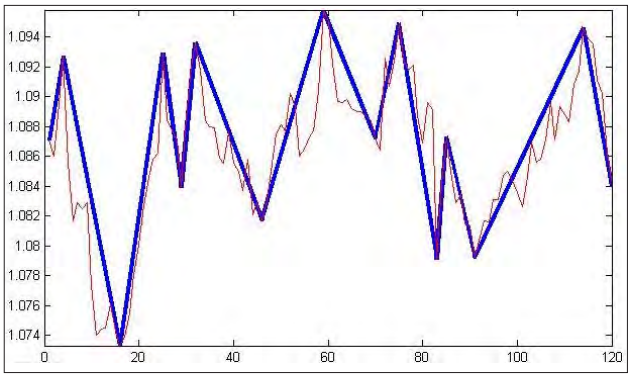

Рис. 3. Исходный и упрощенный графики колебаний курса EUR/USD в январе 2016 г. при степени сжатия *l*=0.2 и приспособленности 4.4756

При l=0.2 алгоритмом были выделены все основные точки перегиба, общее количество которых сократилось со 120 до 15 (в восемь раз). При этом время работы программы на компьютере автора (довольно устаревшей конфигурации с процессором Athlon 64 4400+) составило около пяти секунд.

Следует отметить, что решения, найденные при различных запусках алгоритма, могут несколько отличаться друг от друга и, соответственно, иметь различные значения приспособленности. Для проверки того, насколько сильно различаются найденные решения, было сделано 20 тестовых запусков алгоритма по обработке колебаний валютного курса. Гистограмма распределений значений функции приспособленности показана на рис. 4.

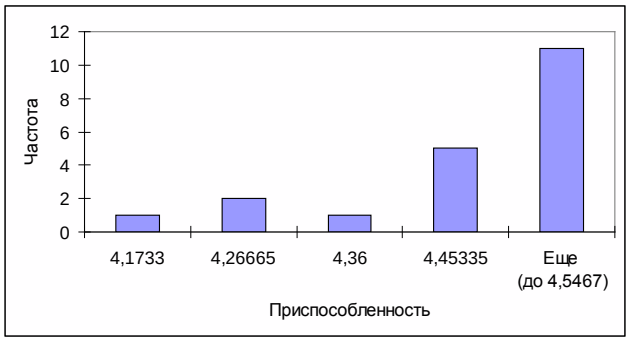

Рис. 4. Распределение значений функции приспособленности при 20 запусках генетического алгоритма

Как видно из анализа данных рис. 4, в 11 случаях из 20 решения практически не отличались по качеству от лучшего. Еще в пяти случаях значение приспособленности отличалось от лучшего из найденных не более чем на 0.1. И только в четырех случаях были зафиксированы более значительные отклонения. Причем даже худшее из найденных решений адекватно отображает основные тенденции ряда и содержит только две небольших ошибки в определении позиции точек перегиба.

Выводы. Таким образом, предложенный в статье метод упрощения динамических рядов позволяет быстро и эффективно сокращать количество точек отсчета ряда с любой степенью сжатия данных, которая регулируется изменением всего лишь одного показателя. По сравнению с существующими данный метод обеспечивает сохранение пиковых значений ряда, а также более эффективное сжатие данных, особенно если значения показателей ряда меняются сравнительно медленно.

Метод упрощения динамических рядов с использованием генетических алгоритмов может использоваться в экономической статистике и анализе, в процессе подготовки презентаций, а также во всех других случаях, когда требуется выделить основные тенденции в большой последовательности данных. Отдельно следует отметить возможность применения принципа квантования по времени с переменным шагом в системах передачи информации с ограниченной пропускной способностью.

Дальнейшие исследования в этом аспекте могут быть направлены на развитие изложенных принципов. Так, следует проверить эффективность реализации модели (6–8) с использованием других методов многокритериальной оптимизации, в частности – метода имитации отжига. Также следует рассмотреть возможность использования других функций штрафов и премий в реализации модели (1–4).

Список литературы:

1. Hilbert Martin, López Priscila. The World's Technological Capacity to Store, Communicate, and Compute Information. / Science 01 Apr 2011: Vol. 332, Issue 6025. – Pp. 60–65.

2. OECD. Understanding the Digital Divide. – Paris. Online-Quelle (Zugriff am 08.11.2002) [Электронный ресурс]. – Режим доступа : http://www.oecd.org/dataoecd/38/57/1888451.pdf.

3. Лопатников Л.И. Экономико-математический словарь: словарь современной экономической науки / Л.И. Лопатников. – М. : Дело, 2003. – 520 с.

4. Дьяконов В. Математические пакеты расширения Matlab. Специальный справочник / В. Дьяконов, В. Круглов. – СПб. : Питер, 2001. – 480 с.

5. Интерактивный график EUR/USD [Электронный ресурс]. – Режим доступа : http://ru.investing.com/currencies/eur-usd-advanced-chart.

6. Батракова Л.Г. Теория статистики / Л.Г. Батракова. – М. : КноРус, 2013. – 528 с.