Методы прогнозирования временных рядов с большим алфавитом на основе универсальной меры и деревьев принятия решений

А. С. Лысяк[1], Б. Я. Рябко[2,3]
[1] *Новосибирский государственный университет, Россия*
[2] *Сибирский государственный университет телекоммуникаций и информатики, Новосибирск, Россия*
[3] *Институт вычислительных технологий СО РАН, Новосибирск, Россия*
e-mail: accemt@gmail.com, boris@ryabko.net

Лысяк А.С., Рябко Б.Я. Методы прогнозирования временных рядов с большим алфавитом на основе универсальной меры и деревьев принятия решений // Вычисл. технологии. 2014. Т. 19, № 2. С. 76–93.

Предлагаются новые подходы к непараметрическому прогнозированию временных рядов, основанные на теории универсальной меры и на деревьях принятия решений. Описаны классический подход к прогнозированию на основе универсального кодирования, методы его оптимизации, а также метод прогнозирования на основе универсальной меры. Кроме того, предложен новый подход к работе алгоритмов прогнозирования — метод разделения алфавита. Данный метод может быть внедрён в произвольный алгоритм прогнозирования и позволяет существенно сократить его трудоёмкость, а также повысить точность предсказания. Показано, как метод разделения алфавита можно внедрить в произвольный алгоритм прогнозирования.

Ключевые слова: универсальная мера, универсальные коды, деревья принятия решений, решающие деревья, прогнозирование, R-мера, временные ряды, метод разделения алфавита.

Lysyak A.S., Ryabko B.Ya. Forecasting methods for time series with large alphabets based on universal measure and decision trees // Comput. Technologies. 2014. Vol. 19, No. 2. P. 76–93.

We suggest and experimentally investigate methods to construct forecasting algorithms based on universal measure and decision trees. The classical method for forecasting is based on the universal coding along with methods for its optimization, and the method for forecasting on the basis of the universal measure. We propose a new approach to forecasting algorithms — the method of separation of the alphabet. This method can be implemented in an arbitrary prediction algorithm and it can significantly reduce the complexity and increase the accuracy of predictions. We show how to implement the method of separation of the alphabet into random prediction algorithm.

Key words: universal measure, decision trees, forecasting, prediction, R-method, time series.

Введение

В настоящее время методы прогнозирования представляют большой практический интерес и позволяют решать широкий спектр задач в науке, технике и экономике. К их числу можно отнести анализ социальных, экономических, геофизических процессов, предсказание природных явлений, экономических событий и др.

Методы прогнозирования служат для исследования системных связей и закономерностей функционирования и развития объектов и процессов с использованием современных методов обработки информации и представляют собой важное средство в анализе сложных прикладных систем, в работе с информацией, в целенаправленном воздействии человека на объекты исследования с целью повышения эффективности их функционирования.

Наиболее распространённой постановкой задачи прогнозирования является задача прогнозирования временных рядов, т. е. функции, определённой на оси времени. В последние два десятилетия были разработаны много методов прогнозирования, показавших свою достаточно высокую эффективность. В частности, в [1] описаны модели машинного обучения, представляющие серьёзную конкуренцию классическим статистическим моделям [2–4]. В работах [5–8] был предложен и развит метод прогнозирования на основе методов универсального кодирования или "сжатия данных", т. е. применения определённых способов кодирования информации, уменьшающих её конечный битовый размер. Преимущество данных методов состоит в выявлении скрытых закономерностей произвольного рода, что позволяет применять их в достаточно широких диапазонах.

На сегодня существуют достаточно много эффективных методов прогнозирования, связанных с мощным математическим аппаратом. К таковым, в частности, относятся прогнозирование на основе билинейной модели [9], авторегресионный анализ различных типов [10–13], прогнозирование на основе методов Монте-Карло [14], методы на основе построения экспертных оценок (рекурсивные стратегии, описание которых можно найти в [4, 5]) и мн. др. Несмотря на наличие приведённого спектра методов и алгоритмов, многие проблемы в задах прогнозирования ещё далеки от своего разрешения. Одна из важнейших в ряду таких проблем — повышение качества прогнозирования характеристик систем, описываемых временными рядами. В числе прочих решению данной проблемы и посвящена настоящая работа. В [15] методом на основе универсальной меры уже были получены некоторые экспериментальные данные по точности прогнозирования природных явлений, превосходящие результаты всех ранее существующих аналогичных методов. В настоящей работе получены новые экспериментальные данные по прогнозированию экономических временных рядов, показана точность прогнозов, которая во многих случаях превышает точность многих современных методов для данных видов рядов. Другой важной проблемой рассматриваемой области является отсутствие значимых (т. е. с приемлемым качеством прогноза и с достаточным для оценки качества прогноза количеством экспериментальных данных) и достаточно многочисленных результатов и методов прогнозирования на несколько шагов вперёд, несмотря на то, что данный класс задач весьма актуален. Причина небольшого числа существующих подходов состоит в существенных сложностях и в нерешённых проблемах, возникающих при их разработке. В частности, сюда относятся эффект накапливания ошибок, снижение качества прогноза и увеличение неопределённости с ростом числа прогнозируемых шагов. К существующим методам, решающим проблему накапливающихся ошибок, относятся методы, основанные на билинейной модели и сжатии данных [16],

однако точность их невысока. Предложенные в настоящей работе методы имеют высокую точность прогноза.

Важной проблемой многих существующих методов прогнозирования является также их высокая трудоёмкость. Так, методы, базирующиеся на универсальной мере и на решающих деревьях, для достижения желаемого качества прогноза требуют достаточно высоких вычислительных ресурсов (во многих случаях — суперкомпьютеров). Предложенный в статье метод разделения алфавита представляет собой не самостоятельный метод прогнозирования, а некоторую модификацию, которая может быть внедрена практически в любой метод прогнозирования, где присутствует оценка вероятности прогнозных событий (элементов).

Важным аспектом настоящей работы является и то, что она посвящена прогнозированию временных рядов с большими алфавитами, что ранее часто было невозможно из-за ограниченности вычислительных ресурсов. Предложенные методы прогнозирования временных рядов с большим алфавитом могут применяться и в случае стандартных временных рядов с небольшими алфавитами, при этом точность их прогноза, как показали полученные экспериментальные результаты, не меняется.

В статье предложены новые подходы к прогнозированию одномерных временных рядов, основанные на определённых моделях теории информации и методах когнитивного анализа данных. В частности, в качестве базовых методов решения задачи прогнозирования используются методы, построенные на основе универсальных кодов, а также основанные на деревьях принятия решений. Кроме того, предложен метод ускорения любых математических методов прогнозирования без потери точности прогноза. Описаны результаты экспериментальных исследований всех предложенных методов и модификаций, а также способы оптимизации рассмотренных методов прогнозирования.

Новизна предлагаемой работы состоит в разработке модификаций методов прогнозирования, позволяющих сократить трудоёмкость произвольных методов прогнозирования, а также нового метода прогнозирования, основанного на решающих деревьях.

1. Постановка задачи прогнозирования

В общем виде задача прогнозирования временных рядов может быть сформулирована следующим образом. Пусть имеется некоторый источник, порождающий последовательность элементов $x_1, x_2, \ldots$ из некоторого множества A, называемого алфавитом. Алфавит может быть как конечным, так и бесконечным (т. е. представлять собой некоторый ограниченный непрерывный интервал). Пусть при этом на момент времени t мы имеем конечную порождённую источником последовательность $x_1, x_2, \ldots, x_t$. Задача прогнозирования сводится к предсказанию элемента, следующего в момент времени $(t+1)$, т. е. элемента x_{t+1}. Когда алфавит A является дискретным и конечным, любой алгоритм прогнозирования может быть применён к данному случаю естественным образом, так как будет оперировать с конечным множеством алфавита A и с конечной выборкой $x_1, x_2, \ldots, x_t$.

Если алфавит A представляет собой непрерывный конечный интервал, то поступим следующим образом. Разобьём заданный интервал на фиксированное количество непересекающихся подмножеств (в общем случае подмножества могут быть произвольного неравного размера), сопоставим им целочисленные номера в соответствии с их порядком в исходном интервале. Количество возможных номеров будет совпадать с числом интервалов. При этом множество всех номеров будет представлять собой уже

новый конечный дискретный алфавит A'. Далее преобразуем исходный временной ряд из терминов в алфавите A в ряд, записанный в терминах нового алфавита A'. Таким образом, получим некоторую конечную выборку (ряд) из конечного алфавита и будем работать с ним, как с конечным дискретным алфавитом. При этом после прогнозирования очередного значения такого ряда ему сопоставляются соответствующий его номеру непрерывный интервал или точка из него (например центр интервала).

Количество букв алфавита обозначим через N. Предполагается, что процесс, или источник информации, является стационарным и эргодическим, т. е. неформально распределение вероятностей символов этого источника не изменяется со временем и не зависит от конкретной реализации процесса. Данное предположение связано с тем, что в работе [7] математически было доказано, что метод на основе универсальной меры выявляет закономерности именно для таких видов рядов. Метод на основе решающих деревьев сходен по видам выявляемых закономерностей и некоторым принципам действия с методом на основе универсальной меры. Как ведут себя эти методы в случае других видов рядов — неясно. Многие реальные временные ряды могут не являться одновременно стационарными и эргодическими, однако в задачу настоящей работы входит экспериментальное исследование применимости предложенных в ней методов к реальным временным рядам.

Пусть источник порождает сообщение $x_1, \ldots, x_{t-1}, x_t,\ x_i \in A,\ i = 1, 2, \ldots, t$, и требуется произвести прогнозирование n следующих элементов (в простейшем случае — одного элемента). Ошибкой прогноза называется (апостериорная) величина отклонения прогноза от действительного состояния объекта (т. е. величина $|x_i - x_i^*|$, где x_i^* — прогнозное значение, x_i — реальное значение). Здесь и далее под ошибкой прогнозирования n элементов будем понимать среднюю ошибку прогноза каждого из n элементов в отдельности. Напомним, что ошибка прогноза характеризует качество прогнозирования.

Очевидно, если распределение вероятностей исходов процесса известно заранее, то задача прогнозирования следующих значений решается достаточно просто (строится прогнозная функция в соответствии с известной закономерностью либо прогнозируются значения исходя из удовлетворения плотности распределения вероятностей ряда, полученного после вставки прогнозных элементов). Однако в большинстве практических задач описанные априорные данные отсутствуют, да и не всегда заданное распределение явно существует. В настоящей работе будет рассматриваться именно такой случай. В данной ситуации для решения задачи прогнозирования можно воспользоваться точными оценками указанных величин, полученными с помощью статистических методов, построенных на основе анализа взаимосвязи последовательных исходов процесса и выявления закономерностей.

В более общей постановке задачи прогнозирования элементы x_i могут быть не только конкретными числами (целыми или вещественными), но и векторами размерности k, где первый элемент вектора — значение прогнозируемой характеристики ряда, а оставшиеся $(k-1)$ атрибутов — какие-либо характеристики процесса или величины, коррелирующие со значениями ряда, известными для всех элементов ряда. Приведём пример. Пусть имеется ряд значений внутреннего валового продукта (ВВП) страны с интервалом в один месяц, величину которого требуется спрогнозировать. Как известно, на ВВП влияют такие параметры как уровень инфляции, индекс потребительских цен, объёмы промышленного производства, дефицит платёжного баланса и мн. др. Значения этих характеристик так же, как и значение ВВП, могут быть известны на каждый месяц прогнозируемого ряда ВВП. Таким образом, можно составить многомерный ряд и про-

гнозировать уже не одно значение временного ряда, а значения всего вектора. При этом интересовать нас будет только один — первый элемент прогнозного вектора.

В итоге задача прогнозирования может быть как одномерной, так и многомерной.

2. Метод прогнозирования, базирующийся на универсальной мере

В работах [7, 16] в качестве подхода для решения задачи прогнозирования временных рядов предлагается использовать методы, основанные на универсальной мере. Приведём определение универсальной меры и поясним связь между указанным и описанным в предыдущем разделе подходами. Мера μ называется универсальной, если для любого стационарного и эргодического источника P верны равенства

$$\lim_{t\to\infty}\frac{1}{t}\left(-\log_2 P\left(x_1,\ldots,x_t\right)-\log_2\mu\left(x_1,\ldots,x_t\right)\right)=0$$

с вероятностью 1 и

$$\lim_{t\to\infty}\frac{1}{t}\sum_{u\in A^t}P(u)\log_2(P(u)/\mu(u))=0.$$

Данные равенства показывают, что в некотором смысле мера μ является непараметрической оценкой для (неизвестной) меры P. По этой причине универсальные меры могут быть использованы для оценки статистических характеристик процесса и прогнозирования.

Понятие универсальной меры тесно связано с понятием универсального кода. Опишем эту взаимосвязь. Код U называется универсальным, если для любого стационарного и эргодического источника P верны равенства

$$\lim_{t\to\infty}|U(x_1,\ldots,x_t)|/t\ =H\left(P\right)$$

с вероятностью 1 и

$$\lim_{t\to\infty}E_P(|U\left(x_1,\ldots,x_t\right)|)/t=H\left(P\right),$$

где $E_P(f)$ — среднее значение f по отношению к P, а $H(P)$ — энтропия P по Шеннону, т. е.

$$H\left(P\right)=\lim_{t\to\infty}-t^{-1}\sum_{u\in A^t}P(u)\log P(u).$$

Таким образом, если есть универсальный код, то на его основе легко получить универсальную меру.

Следующая теорема, приведённая с доказательством в [15], говорит о том, что на базе любого универсального кода можно построить универсальную меру.

Теорема 1. *Пусть U — универсальный код и*

$$\mu_U\left(\omega\right)=2^{-|U(\omega)|}\Big/\sum_{u\in A^{|\omega|}}2^{-|U(u)|}.$$

Тогда μ — универсальная мера.

Теперь опишем универсальную меру R, которая была использована в качестве основы для метода прогнозирования в настоящей работе. Выбор именно этой меры связан с тем, что она построена на основе асимптотически оптимального универсального кода R, что доказано в [8].

В общем случае в качестве универсальной меры была взята мера Кричевского $K_m \geq 0$, являющаяся универсальной для множества марковских источников с памятью, или связностью, m, $m \geq 0$; если $m = 0$, то имеем источник независимых и одинаково распределённых символов. В некотором смысле эта мера является оптимальной для данного множества. По определению,

$$K_m(x_1, \ldots, x_t) = \begin{cases} \dfrac{1}{|A|^t}, & t \leq m, \\ \dfrac{1}{|A|^m} \displaystyle\prod_{v \in A^m} \dfrac{\displaystyle\prod_{a \in A} (\Gamma(\nu_x(va) + 1/2)/\Gamma(1/2))}{(\Gamma(\overline{\nu}_x(v) + |A|/2)/\Gamma(|A|/2))}, & t > m, \end{cases} \tag{1}$$

где $\nu_x(v)$ — число последовательностей v, встречающихся в x, $\overline{\nu}_x(v) = \sum\limits_{a \in A} \nu_x(va)$, $x = x_1, \ldots, x_t$, $\Gamma(\,)$ — гамма-функция.

Определим также распределение вероятностей $\{\omega = \omega_1, \omega_2, \ldots\}$ для целых {1,2,...} как

$$\omega_i = 1/\log(i+1) - 1/\log(i+2). \tag{2}$$

Далее будем использовать именно это распределение.

Мера R определяется как

$$R(x_1, \ldots, x_t) = \sum_{i=0}^{v} \omega_{i+1} K_i(x_1, \ldots, x_t). \tag{3}$$

Слагаемые ω_i играют в данном случае роль весовых коэффициентов, которые с ростом индекса i (т. е. порядка меры Кричевского) становятся всё меньше. Понятно, что слишком большие порядки в мере Кричевского должны иметь меньший вес и меньше влиять на прогноз. В результате в качестве весовых коэффициентов было выбрано распределение (2). В общем случае весовые коэффициенты представляют собой варьируемый параметр метода и могут меняться в зависимости от ряда и метода.

Посчитать бесконечную сумму (3) в реальных алгоритмах не удастся, поэтому все расчёты будут проводиться в соответствии с тем, что каждое следующее слагаемое в сумме (3) вносит всё меньший вклад в итоговое значение R. Данный факт можно без труда доказать теоретически и проверить эмпирическим путём. В силу этого свойства для расчёта меры R будем использовать первые m слагаемых суммы, а число m назовём глубиной анализа метода.

Итак, значение меры R, вычисленное на основе формулы (3), может служить оценкой вероятности исхода процесса и использоваться для решения задачи прогнозирования.

Перейдём далее к схеме прогнозирования на основе универсальной меры для источников как на дискретном, так и на непрерывном алфавите.

Вначале рассмотрим источник, порождающий значения из конечного алфавита. В данном случае схема вычисления меры R достаточна проста. Пусть $x_1, \ldots, x_t$ — име-

ющаяся временная последовательность. Для каждого $a \in A$ построим последовательность $x_1, \ldots, x_t a$ и вычислим условную вероятность на основе меры R:

$$R(a|x_1, \ldots, x_t) = R(x_1, \ldots, x_t a)/R(x_1, \ldots, x_t).$$

Полученные таким образом для каждого $a \in A$ величины можно использовать в качестве оценок соответствующих неизвестных вероятностей $P(x_1, \ldots, x_t a)$. Величина a, имеющая максимальную оценку вероятности, и будет прогнозным значением.

Рассмотрим теперь схему прогнозирования для источника из непрерывного интервала. Пусть имеется стохастический процесс, генерирующий последовательность X_t, каждый элемент которой принимает значения из стандартного борелевого пространства Ω, представляющего в нашем случае непрерывный интервал $[A, B]$. Пусть также $\{\Pi_n\}$, $n \geq 1$ — возрастающая последовательность конечных разбиений интервала $[A, B]$ на n частей (назовём этот процесс квантизацией). В нашем случае разбиение интервалов проводилось равномерно на равные подынтервалы, т. е. размер каждого подынтервала определялся как $h = (B - A)/n$. Обоснование выбора именно такого метода будет дано ниже. Определим также $x^{[k]}$ как элемент Π_k, содержащий точку x.

Оценку плотности вероятностей r выразим в виде

$$r\left(x_1, \ldots, x_t\right) = \sum_{s=1}^{v} \omega_s R\left({x_1}^{[s]}, \ldots, {x_t}^{[s]}\right). \tag{4}$$

Как показано в [8, 9], плотность $r\left(x_1, \ldots, x_t\right)$ является оценкой неизвестной плотности $p\left(x_1, \ldots, x_t\right)$, а соответствующая условная плотность

$$r(a|x_1, \ldots, x_t) = r(x_1, \ldots, x_t a)/r\left(x_1, \ldots, x_t\right) \tag{5}$$

является подходящей оценкой вероятности $p\left(a|x_1, \ldots, x_t\right)$. Количество слагаемых в сумме (4) при реализации описанного далее алгоритма, как и в дискретном случае, равно глубине анализа m.

3. Метод прогнозирования на основе решающих деревьев

В общем виде постановка задачи для решающих деревьев выглядит следующим образом. Пусть дано множество объектов A (всего в A находится N объектов, составляющих так называемую обучающую выборку), обладающих определёнными независимыми характеристиками (атрибутами с конечным множеством значений; всего имеется $(M + 1)$ атрибутов). Множество первых M атрибутов обозначим как Q. Для заданного множества A все $(M + 1)$ атрибутов известны. Для других (новых) элементов по известным первым M атрибутам требуется найти целевой $(M + 1)$-й атрибут. При этом на вход подаются число элементов в обучающей выборке N, число M и параметр $m \leq M$.

Как правило, данный метод применяется для задач классификации и кластеризации. В настоящей работе предложен подход, который показывает способ использования данных деревьев в прогнозировании временных рядов. Дерево принятия решений строится по описанному ниже алгоритму.

Введём некоторые важные определения.

Определение 1. *Энтропия* $H\left(A, S\right) = -\sum\limits_{i=1}^{Sn} \dfrac{|A_i|}{|A|} \log_2 \dfrac{|A_i|}{|A|}$, *S — целевой атрибут, A_i — элементы из A, у которых атрибут S равен i (а $|A| = N$).*

Определение 2. *Прирост информации определяется для каждого атрибута из Q по отношению к целевому атрибуту S и показывает, какой из атрибутов Q даёт максимальный прирост информации относительно значения атрибута S (т. е. относительно класса элемента). Прирост информации определяется по формуле*

$$\mathrm{Gain}\,(A, Q) = H(A, S) - \sum_{i=1}^{Qn} \frac{|A_i|}{N} H(A_i, S).$$

Далее опишем один из наиболее эффективных алгоритмов построения решающего дерева — ID3, зависящий от множества A, целевого атрибута S и множества атрибутов Q.

1. Создать корень дерева.
2. Если S равно какому-то q на всех элементах из A, поставить в корень метку q и выйти.
3. Если $Q = \{\emptyset\}$, то из множества значений S выбрать такое q, которому равно наибольшее число элементов из A, поставить q в корень и выйти.
4. Выбрать $q \in Q$, для которого $\mathrm{Gain}(A, q)$ максимален.
5. Поставить в корень дерева метку q.
6. Для каждого значения q_i атрибута q
   а) добавить нового потомка и пометить исходящее ребро меткой q_i;
   б) если в A нет элементов, для которых значение q равно q_i, то поступить в соответствии с п. 3;
   в) иначе — запустить $\mathrm{ID3}(Aq_i, S, Q\backslash\{q\})$ и добавить его результат как поддерево с корнем в этом потомке.

Дерево строится до окончания обучающего множества или до пустоты множества Q. В предлагаемой реализации данного алгоритма глубину дерева можно ограничивать искусственно — отдельным параметром. После достижения заданной глубины дерева выполняется пункт 3 алгоритма ID3.

Рассмотрим пример построения решающего дерева для прогнозирования игры в футбол заданной команды.

Пусть имеются следующие характеристики игры: позиция соперника в турнирной таблице (выше или ниже заданной команды), место игры (дома или в гостях), лидеры команды (на месте или нет), погода (будет дождь или нет). Требуется спрогнозировать результат игры при следующих известных характеристиках игры:

Позиция соперника в турнирной таблице	Место игры	Лидеры команды	Дождь	Победа
Выше заданной команды	Дома	На месте	Да	Нет
Выше заданной команды	Дома	На месте	Нет	Да
Выше заданной команды	Дома	Нет	Нет	Да
Ниже заданной команды	Дома	Нет	Нет	Да
Ниже заданной команды	В гостях	Нет	Нет	Нет
Ниже заданной команды	Дома	Нет	Да	Да
Выше заданной команды	В гостях	На месте	Да	Нет
Выше заданной команды	В гостях	На месте	Нет	?

Вычислим значения энтропии относительного целевого признака “Победа”:

$$H\,(A,\ \text{Победа}) = -\frac{4}{7}\log\frac{4}{7} - \frac{3}{7}\log\frac{3}{7} = 0.9852.$$

Теперь вычислим прирост информации для каждого из нецелевых признаков:

1) $\mathrm{Gain}\,(A,\ \text{Лидеры}) =$
$= H(A,\ \text{Победа}) - \frac{3}{7} H\,(A_{\text{на месте}},\ \text{Победа}) - \frac{4}{7} H\,(A_{\text{нет}},\ \text{Победа}) = 0.1281;$
2) $\mathrm{Gain}\,(A,\ \text{Играем}) =$
$= H(A,\ \text{Победа}) - \frac{5}{7} H\,(A_{\text{дома}},\ \text{Победа}) - \frac{2}{7} H\,(A_{\text{в гостях}},\ \text{Победа}) = 0.4696;$

и т. д. для всех четырёх свойств.

Следуя описанному алгоритму ID3, строим дерево, выбирая на каждом этапе признак с максимальным приростом информации (рис. 1). Для применения данного дерева в случае одномерного прогнозирования временного ряда возьмём в качестве признаков предыдущие значения ряда. Сделаем это следующим образом. Зададим параметр метода m, имеющий смысл, аналогичный глубине анализа в R-мере. Параметр m будет определять число признаков в дереве (и соответственно его максимальную глубину). Далее составим множество A по правилу: в качестве первого и целевого признака возьмём какое-то i-е значение ряда, а в качестве его $(m-1)$ атрибутов примем $(m-1)$ значений, стоящих перед i-м значением во временном ряде. В итоге получим множество A, состоящее из $(N-m)$ элементов, на основе которых строим дерево в соответствии с алгоритмом ID3, и далее, следуя по дереву и используя последние $(m-1)$ элементов ряда, получим прогнозное значение.

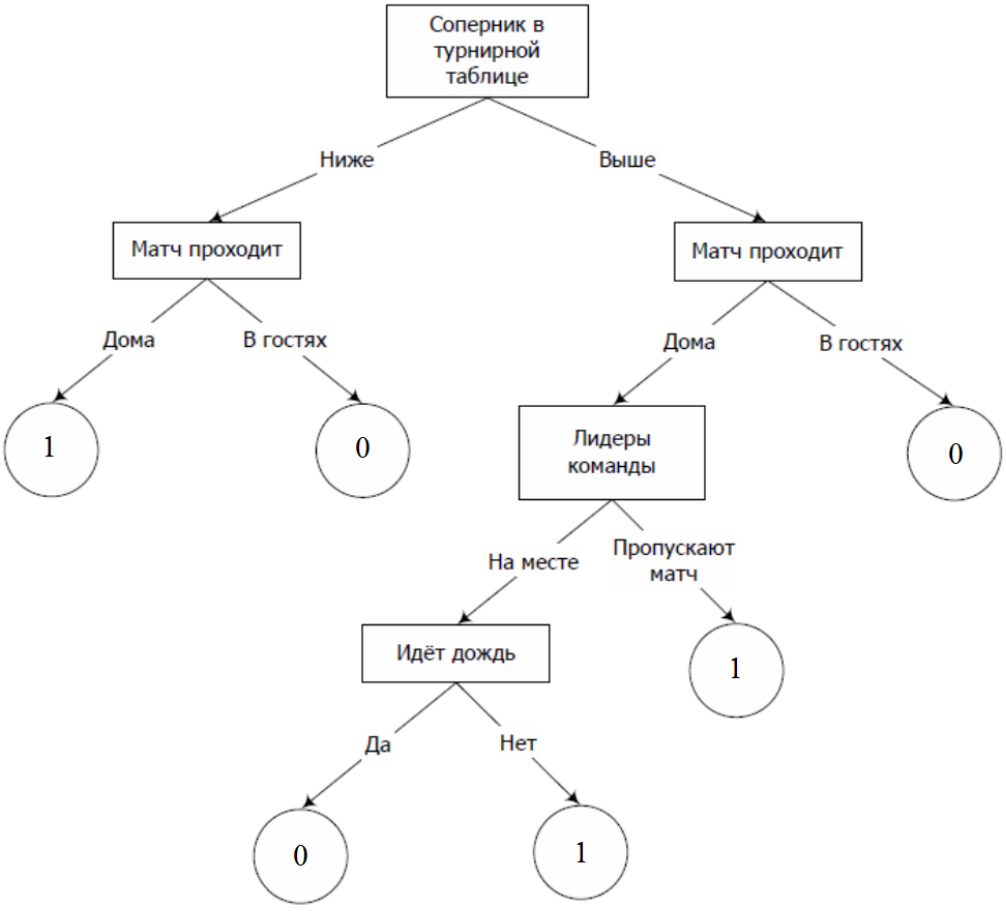

Рис. 1. Дерево, построенное по алгоритму ID3

В силу того что при большой глубине анализа и большом алфавите дерево будет слишком сильно разветвляться и трудоёмкость алгоритма соответственно возрастёт экспоненциально относительно значения параметра m, введём следующую модификацию алгоритма: зададим другой параметр m', показывающий максимальную глубину дерева, до которой работает алгоритм ID3. При достижении заданной в m' максимальной глубины следуем пункту 3 алгоритма ID3.

4. Результаты прогнозирования курсов валют методами R и ID3

Методы были реализованы на суперкомпьютере [19] и протестированы на прогнозах реальных данных. Необходимость работы на суперкомпьютере обусловлена высокой трудоёмкостью методов. Для обоих методов оценки сложности схожи; для стандартной реализации метода R оценка трудоёмкости на один прогнозный элемент равна $T = O(m \cdot n^{m+1} \cdot N^2)$ и для рассмотренного случая $m = 5$, $N = 500$, $n = 50$ составляет $1.8 \cdot 10^{16}$ операций на один прогнозный элемент. Результаты данных прогнозов приведены в табл. 1. Следует заметить, что при реализации алгоритма ID3 было учтено ограничение на максимальную глубину дерева m' (см. раздел 3). В таблице это ограничение (параметр m') и параметр m показаны в столбце 3 в виде m'/m.

Исследования проводилась в двух режимах: первый режим (on-line) означает прогнозирование значений временного ряда на один шаг вперёд, второй режим — на 10 шагов вперёд. Для прогнозирования на много шагов вперёд выбрано именно 10 шагов, так как это значение является достаточно большим, чтобы данные прогнозы отличались от метода on-line-прогнозирования, и вместе с тем не очень большим, чтобы не снизить точность прогноза (в силу нестационарности источников). Прогнозирование на 10 шагов вперёд проводилось следующим образом: вначале прогнозировалось очередное значение ряда, после чего выборка пополнялась прогнозным значением, затем

Т а б л и ц а 1. Данные прогноза курса валют евро/доллар с временными интервалами один день и один час в периоды соответственно 20.06.2012 – 08.11.2013 и 20.06.2012 – 11.07.2012

Размер выборки L	Разбиение n	m'/m^*	Метод ID3 (on-line)	Метод R (on-line)	Метод ID3 (10 шагов)	Метод R (10 шагов)
Временной интервал один день						
500	10	2 / 2	0.0079	0.0084	0.0103	0.0299
		5 / 5	0.0095	0.0084	0.0151	0.0299
	20	2 / 2	0.0088	0.0083	0.0105	0.0159
		5 / 5	0.0084	0.0083	0.0105	0.0159
	50	2 / 2	0.0089	0.0083	0.0119	0.0187
Временной интервал один час						
500	10	2 / 2	0.00114	0.00114	0.00131	0.00131
		5 / 5	0.00132	0.00114	0.00144	0.00131
	20	2 / 2	0.00106	0.00103	0.00131	0.00131
		5 / 5	0.00147		0.00110	0.00131
	50	2 / 5	0.00103	0.00104	0.00238	0.00141

*m'/m — максимальная глубина дерева m'/величина ошибки прогноза каждого из методов m (то же в табл. 2–4)

прогнозирование велось ещё на один шаг, но с использованием уже пополненного ряда и далее продолжалось до десятого прогнозного элемента, после чего считалась ошибка прогноза. Прогнозирование выполнялось на десяти разных выборках с последующим усреднением ошибки.

Важно отметить, что прогноз проводился не для абсолютных величин выборки, а для разницы между соседними элементами с последующей прибавкой спрогнозированной разницы к последнему элементу ряда (в результате получалось значение следующего за последним элемента ряда). Такой подход позволяет существенно снизить необходимый размер непрерывного интервала, в котором находятся прогнозные значения, и выявлять линейные и квазилинейные тренды и периоды на них, что было невозмож-

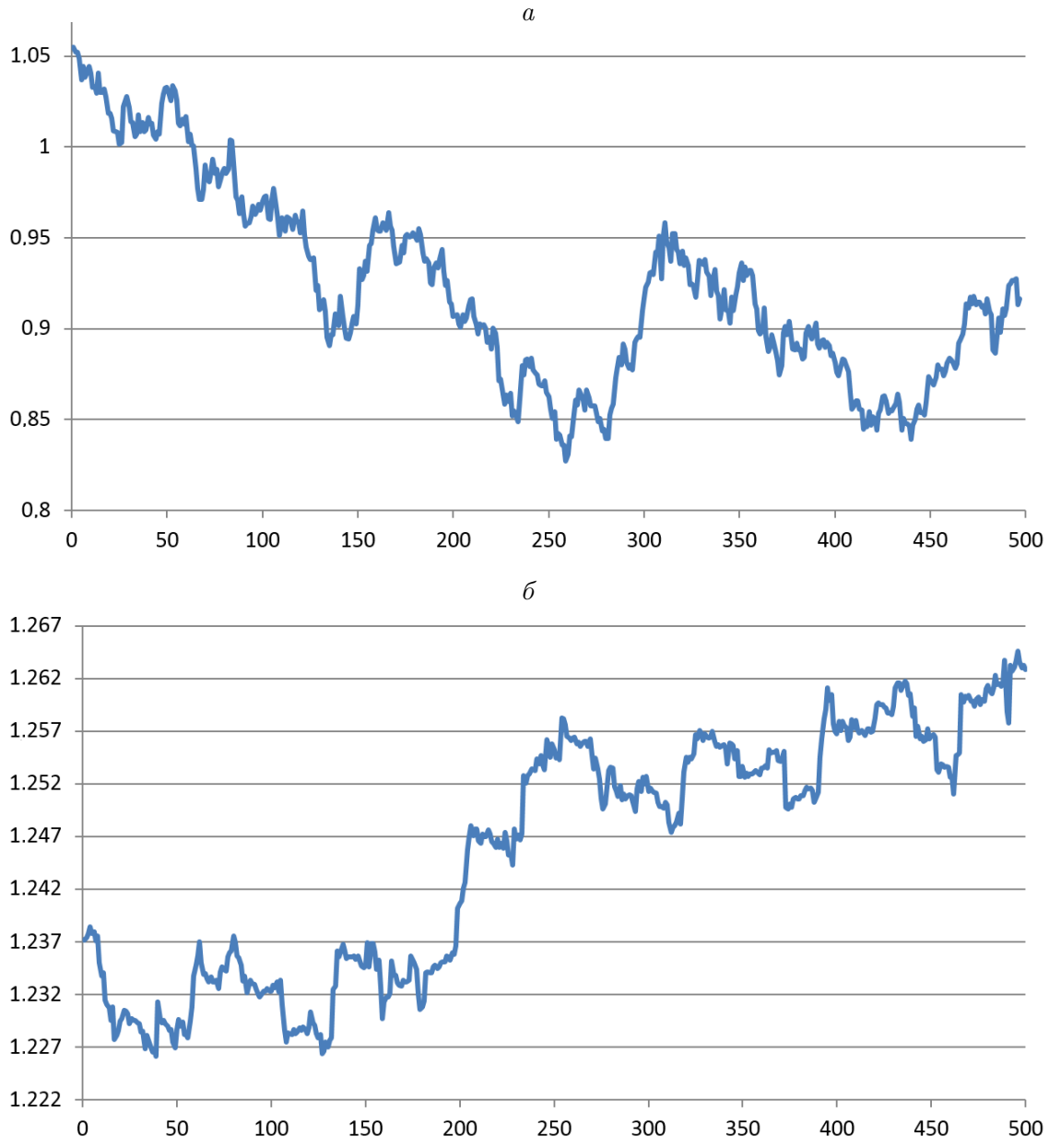

Рис. 2. Графики курса евро/доллар: *а* — временной интервал один день, *б* — временной интервал один час

но при прогнозировании абсолютных величин временного ряда. Определение границ интервала осуществлялось следующим образом: вычислялась величина максимального и минимального (с учётом знака) отклонения текущего и предыдущего элементов, полученные значения использовались в качестве левой и правой границ интервала, который далее разбивался на количество частей, равное мощности алфавита.

В табл. 1 приведены данные прогноза курса валют евро/доллар при следующих параметрах: размер выборки L, количество частей разбиения непрерывного интервала n, максимальная глубина дерева m' и величина ошибки прогноза каждого из методов m (m'/m). Размер выборки в табл. 1 (и далее в табл. 2) равен 500, что является относительно небольшой величиной, однако большее количество значений ряда приводит к существенному росту операционной сложности методов. Кроме того, все полученные ошибки сравнивались с максимальным изменением между соседними элементами ряда, делённым на число n. В большинстве полученных результатов для $n = 10$ и $n = 20$ все прогнозные элементы оказались в пределах одного элемента разбиения, что говорит о достаточности данного размера выборки для небольших n. Графики курса евро/доллар приведены на рис. 2.

Как видно из табл. 1, существует определённый предел размера алфавита (в нашем случае — разбиения непрерывного интервала), после которого точность методов не растёт. Этот факт справедлив для обоих рассматриваемых методов. Из полученных данных также следует, что глубина анализа после значения $m = 2$ улучшает точность прогноза достаточно несущественно и после какого-то заданного m точность, как и в случае с размером алфавита, уже не меняется. Это говорит о том, что для получения оптимальных прогнозов за приемлемое время достаточно подобрать такие минимальные значения размера алфавита и глубины анализа обоих методов, которые будут давать оптимальные (приближённые к границе точности) значения ошибок.

Наличие описанных границ точности методов объясняется достаточно просто: в случае прогнозирования сложных рядов, в которых нет видимых или относительно простых закономерностей, оба метода не находят их, усредняют тренд (разницу между соседними элементами) и это усреднённое значение используют в качестве прогноза. При наличии же в ряду каких-либо закономерностей алгоритмам потребуется бóльшая глубина анализа (при m меньше длины периода закономерности алгоритмы просто не выявят её).

5. Ускорение алгоритмов с использованием метода разделения алфавита

Специфика работы алгоритмов прогнозирования R и ID3 состоит в том, что для получения достаточно точных данных прогноза и учёта больших длин выборок и алфавитов зачастую нужны большие значения параметров глубины выборки и размера алфавита, что значительно повышает трудоёмкость работы этих алгоритмов, экспоненциально зависящей от глубины анализа и, как минимум, квадратично — от размера алфавита. В результате был разработан универсальный метод, позволяющий уменьшить время работы в общем случае любого алгоритма прогнозирования. Этот метод впервые предложен Б.Я. Рябко в [18], после чего был модифицирован и реализован для применения в алгоритмах прогнозирования R и ID3. Опишем данный алгоритм под названием “метод разделения алфавита”.

Пусть дан алфавит A элементов временного ряда какой-либо большой мощности N. Выберем некоторые рекуррентные подразбиения заданного алфавита A по следующему алгоритму:

- выберем набор непересекающихся подмножеств множества A: $B_1, B_2, \ldots, B_{N1}$, где $N1 << N$ и каждый B_i содержит один или несколько элементов из A;
- полученные элементы алфавита A, содержащиеся в каждом B_i, разобьём ещё на $N2$ частей $B_{i,N2}$, также содержащих соответствующие элементы множества из A;
- продолжим данный рекуррентный процесс до получения заданного числа подразбиений.

Далее продолжим процесс по следующему алгоритму:

- записываем исходный временной ряд в терминах алфавита B (т. е. все элементы исходного ряда преобразуем в соответствующие символы из алфавита B) и прогнозируем элемент из алфавита B. Обозначим его как B_i;
- фильтруем исходный ряд по прогнозному значению B_i: оставляем в нём только те элементы из A, которые принадлежат спрогнозированному B_i;
- записываем ряд в терминах алфавита B_i (т. е. ряд второго уровня);
- прогнозируем элемент алфавита второго уровня и продолжаем процедуру до последнего уровня разбиения.

Приведём пример работы предложенного алгоритма. Пусть даны алфавит $A = \{i\}$, где $i = \overline{1, \ldots, 12}$, и временной ряд $X(A)$: 1, 3, 5, 5, 6, 7, 8, 1, 3, 5, 5, 6, 7, 8, 1, 3, 5. Требуется предсказать следующий элемент (который, очевидно, должен быть равен 5). Разобьём исходный алфавит на четыре равных части, определив тем самым новый алфавит B: $\{B_i\}$, $i = \overline{1, \ldots, 4}$. Сделаем разбиение равномерным. В итоге получим следующие значения B_i: $B_1 = \{1, 2, 3\}$, $B_2 = \{4, 5, 6\}$, $B_3 = \{7, 8, 9\}$, $B_4 = \{10, 11, 12\}$.

В каждой из частей B_i содержится три элемента из множества A, которые и будут образовывать вторичный алфавит $B_{i,j}$. В итоге каждому A_i будет однозначно соответствовать элемент $B_{i,j}$.

Теперь перепишем исходный ряд в терминах алфавита B_i: $X(B)$: 1, 1, 2, 2, 2, 3, 3, 1, 1, 2, 2, 2, 3, 3, 1, 1, 2. Затем к полученному ряду применим какой-либо метод прогнозирования и найдём прогнозное значение в терминах алфавита B. В данном случае прогнозным значением будет 2.

Далее осуществим фильтрацию исходной последовательности по принципу: если элемент $X_i(A)$ находится во множестве B_2, то оставляем его в ряду, иначе — удаляем. В результате получим следующий ряд: 5, 5, 6, 5, 5, 6, 5. В нём присутствуют только два символа алфавита из 12 (так как мощность множества B_i равна 3, а число 4 в исходной последовательности не встречается ни разу). Поэтому можно переписать заданный ряд в терминах нового алфавита из трёх элементов (4 переходит в 1, 5 — в 2, 6 — в 3): 2, 2, 3, 2, 2, 3, 2. Затем в полученном ряде определяем прогнозное значение и приводим его к терминам исходного алфавита. Прогнозное значение равно числу 2, соответствующему числу 5 в исходном алфавите. Символ исходного алфавита 5 и будет являться результатом прогнозирования.

Количество подразбиений алфавита в общем случае не ограничено, однако для использования данного алгоритма в наших конкретных методах прогнозирования возьмём одно разбиение исходного алфавита. Результаты, приведённые ниже, были получены на обычном компьютере с теми же временными затратами, что ранее наблюдались только при работе на суперкомпьютере.

6. Результаты, полученные с использованием метода разделения алфавита

Рассмотрим случаи, описанные в разделе 4, с применением метода разделения алфавита для прогнозирования временных рядов с большим алфавитом. Предполагается, что алфавит является большим, если разбиение n составляет больше 160 элементов (что при максимальной разнице между соседними элементами ряда, равной 1.0, позволяет иметь точность прогноза в 2–3 знака после запятой). Полученные результаты приведены в табл. 2. При этом разбиение алфавита A осуществлялось на равномощные непересекающиеся подмножества. Размеры подмножеств указаны во второй колонке таблицы вместе с величиной разбиения n (в скобках — произведение двух чисел, первое из которых — число подмножеств множества A, второе — мощность каждого такого подмножества).

Как видно из табл. 2, точность рассматриваемых методов при использовании метода разделения алфавита не уменьшается, однако скорость работы алгоритма существенно увеличивается. Ошибка прогноза методов зависит только от размера исходного алфавита и несущественно — от выбранной глубины анализа. Прогнозирование на 10 шагов вперёд даёт ошибку прогноза, сравнимую в порядке с ошибкой прогноза on-line-режима, что говорит о хороших результатах применения методов R и ID3 при прогнозировании на множество шагов вперёд. При этом различие между обоих методов несущественно.

Т а б л и ц а 2. Точность, полученная с использованием метода разделения алфавита, с временными интервалами один день и один час

Размер выборки L	Разбиение n	m'/m	Метод ID3 (on-line)	Метод R (on-line)	Метод ID3 (10 шагов)	Метод R (10 шагов)
Временной интервал один день						
500	10 $(10\cdot 1)$	2 / 2	0.0077	0.0084	0.0076	0.0299
		5 / 5	0.0112	0.0084	0.0112	0.0299
	20 $(20\cdot 1)$	2 / 2	0.0068	0.0080	0.0068	0.0159
		5 / 5	—	0.0080	—	0.0159
	20 $(10\cdot 2)$	2 / 2	0.00721	0.0100	0.00716	0.0345
	20 $(5\cdot 4)$	2 / 2		0.0080		0.0159
	50 $(10\cdot 5)$	2 / 2	0.00735	0.0099	0.00693	0.0173
		2 / 5	0.00756	0.0080	0.00751	0.0119
	50 $(25\cdot 2)$	2 / 2	—	0.0083	—	0.0107
	50 $(50\cdot 1)$	2 / 2	—	0.0083	—	0.0187
Временной интервал один час						
500	10 $(10\cdot 1)$	2 / 2	0.00114	0.00114	0.00131	0.00131
		5 / 5	0.00132	0.00114	0.00144	0.00131
	20 $(20\cdot 1)$	2 / 2	0.00106	0.00103	0.00131	0.00131
		5 / 5	0.00147		0.00110	0.00131
	20 $(10\cdot 2)$	2 / 2	0.00103	0.00103	0.00185	0.00185
	20 $(5\cdot 4)$	2 / 2	0.00103	0.00103	0.00185	0.00185
	50 $(10\cdot 5)$	2 / 2	0.00094	0.00104	0.00137	0.00141
		2 / 5	0.00094		0.00137	0.00141
	50 $(25\cdot 2)$	2 / 5	0.00124	0.00104	0.00349	0.00141
	50 $(50\cdot 1)$	2 / 5	0.00103	0.00104	0.00238	0.00141

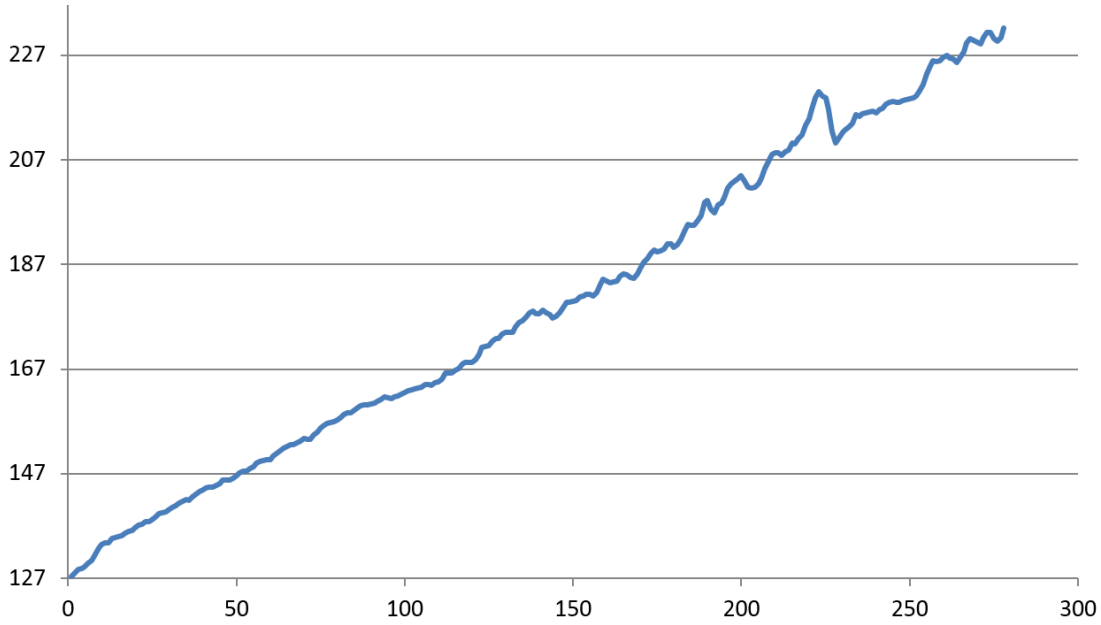

Рис. 3. Индекс потребительских цен для периода прогноза 01.1990 – 02.2013

Т а б л и ц а 3. Индекс потребительских цен для периодов прогноза 02.2012 – 02.2013 и 02.2002 – 02.2003

Размер выборки L	Разбиение n	m'/m	Метод ID3 (on-line)	Метод R (on-line)	Метод ID3 (10 шагов)	Метод R (10 шагов)
Период прогноза 02.2012 – 02.2013						
277	5 (5 · 1)	2 / 2	0.0966154	0.112769	0.256923	0.257077
	10 (10 · 1)	2 / 2	0.0926154	0.092615	0.088	0.088154
		2 / 5	0.1269231		0.264154	
	20 (20 · 1)	2 / 2	0.0847692	0.107846	0.110462	0.136
		2 / 5	0.0936923		0.165385	
	20 (10 · 2)	2 / 2	0.0926154	0.097692	0.136	0.136
	20 (5 · 4)	2 / 2	0.0976923	0.107846	0.136	0.136
	100 (10 · 10)	2 / 2	0.0946154	0.100769	0.115846	0.230308
		2 / 5	0.1249231		0.285846	
	100 (20 · 5)	2 / 2	0.0987692	0.111846	0.133538	0.230308
	100 (5 · 20)	2 / 5	0.1333846	0.111846	0.107692	0.230308
Период прогноза 02.2002 – 02.2003						
240	5 (5 · 1)	2 / 2	0.0761538	0.056	0.105231	0.105231
	10 (10 · 1)	2 / 2	0.0606154	0.050462	0.193077	0.193077
		2 / 5	0.0606154		0.193077	
	20 (20 · 1)	2 / 2	0.0627692	0.047692	0.065231	0.065231
		2 / 5	0.062		0.080308	
	20 (10 · 2)	2 / 2	0.0576923	0.047692	0.065231	0.065231
	20 (5 · 4)	2 / 2	0.0576923	0.047692	0.065231	0.095385
	100 (10 · 10)	2 / 2	0.0627692	0.047692	0.11	0.11
		2 / 5	0.0627692		0.11	
	100 (20 · 5)	2 / 2	0.0618462	0.047692	0.11	0.11
	100 (5 · 20)	2 / 5	0.0507692	0.047692	0.130154	0.11
	240 (24 · 10)	2 / 5	0.0518462	0.045692	0.061077	0.131231

Рассмотрим результаты прогнозирования некоторых экономических показателей, в частности, индексов потребительских и промышленных цен. В табл. 3 отражены данные прогноза общего суммарного индекса потребительских цен по территории США в период 01.1990 – 02.2013 с временным интервалом один месяц. Значения индекса потребительских цен прогнозировались для периодов 02.2012 – 02.2013 и 02.2002 – 02.2003. График для данного временного ряда приведён на рис. 3.

В верхней части табл. 3 представлены результаты прогноза последней части ряда, в которой дисперсия выборки явно больше, чем в середине (что хорошо видно из рис. 3). Там же приведены результаты прогноза центральной части графика, которая выглядит более гладкой. Таким образом, результаты прогноза ряда с более низкой дисперсией, несмотря на меньший размер выборки, ощутимо лучше, что вполне закономерно.

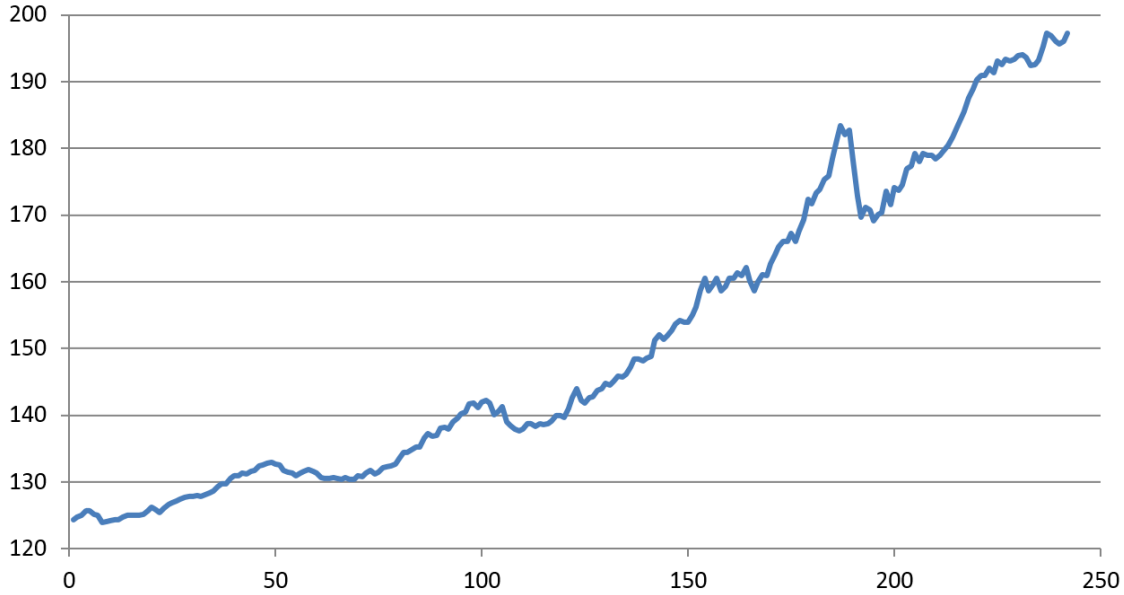

Рис. 4. Индекс промышленных цен для периода прогноза 01.1990 – 02.2013

Т а б л и ц а 4. Индекс промышленных цен для периода прогноза 02.2012 – 02.2013

Размер выборки L	Разбиение n	m'/m	Метод ID3 (on-line)	Метод R (on-line)	Метод ID3 (10 шагов)	Метод R (10 шагов)
277	5 $(5 \cdot 1)$	2 / 2	0.1109302	0.11093	0.217442	0.217442
	10 $(10 \cdot 1)$	2 / 2	0.115814	0.105814	0.158488	0.158488
		2 / 5	0.1537209		0.208488	
	20 $(20 \cdot 1)$	2 / 2	0.1244186	0.10593	0.108721	0.125116
		2 / 5	0.130814		0.125116	
	20 $(10 \cdot 2)$	2 / 2	0.1109302	0.10593	0.125116	0.125116
	20 $(5 \cdot 4)$	2 / 2	0.110814	0.10593	0.125116	0.125116
	100 $(10 \cdot 10)$	2 / 2	0.1177907	0.105814	0.14	0.117791
		2 / 5	0.1506977		0.173023	
	100 $(20 \cdot 5)$	2 / 2	0.1294186	0.11186	0.119767	0.118721
	100 $(5 \cdot 20)$	2 / 5	0.1169767	0.105814	0.276977	0.117791
	240 $(24 \cdot 10)$	2 / 5	0.1009302	0.105814	0.166163	0.11814

Результаты прогнозирования индекса промышленных цен и соответствующий график для периода 02.2012 –02.2013 представлены в табл. 4 и на рис. 4.

Полученные данные подтверждают сделанные выше заключения. Что же касается сравнения методов R и деревьев принятия решений ID3, то исходя из описанных результатов можно сделать следующие выводы:

- Прогнозирование периодических функций метод ID3 осуществляет лучше, чем метод R, как в режиме on-line, так и на несколько шагов вперёд.
- Прогнозирование курсов валют и индексов потребительских и промышленных цен в режиме на несколько шагов вперёд на малом разбиении лучше осуществляет метод решающих деревьев; в режиме on-line рассмотренные методы показывают примерно одинаковые результаты.
- Прогнозирование реальных данных на большом разбиении в режиме on-line несколько лучше осуществляет метод R, в режиме на несколько шагов вперёд оба метода примерно равноценны.
- Метод разделения алфавита существенно повышает скорость работы алгоритмов для фиксированного разбиения, не ухудшая при этом результаты прогноза.

Заключение

Как видно из полученных результатов, метод на основе универсальной меры и метод на основе деревьев принятия решений показывают достаточно высокий уровень точности при прогнозировании на реальных данных. Кроме того, была создана более эффективная реализация метода R, позволяющая сократить компьютерное время перебора различных вариантов прогнозных значений из алфавита. Трудоёмкость уменьшилась до константной относительно длины алфавита (была линейной). Кроме того, благодаря использованию метода разделения алфавита существенно снизились ограничения на аппаратные ресурсы и параметры, на которых будут работать данные методы,что позволило достичь ту же скорость работы, что ранее достигалась только на суперкомпьютере.

Метод, основанный на универсальной мере, даёт примерно ту же точность, что и метод решающих деревьев, но использует другие принципы анализа. Преимущество этих методов в модификации с методом разделения алфавита состоит в том, что в такой комбинации их можно легко обобщить на многомерные ряды, где, кроме значения, у каждого элемента имеются ещё другие атрибуты-свойства, коррелирующие с данным рядом.

Список литературы

[1] Bontempi G. Local Learning Techniques for Modeling, Prediction and Control. Ph.d., IRIDIA-Universit de Libre de Bruxelles, BELGIUM, 1999.

[2] Ahmed N. An empirical comparison of machine learning models for time series forecasting // Econometric Rev. 2010. Vol. 29, iss. 5-6. P. 594–621.

[3] Palit A.K., Popovic D. Computational Intelligence in Time Series Forecasting: Theory and Engineering Applications (Advances in Industrial Control). New York: Springer-Verlag, 2005.

[4] Zhang G., Patuwo B.E., Michael Y.H. Forecasting with articial neural networks: The state of the art // Intern. J. of Forecasting. 1998. Vol. 14, iss. 1. P. 35–62.

[5] Ryabko B. Compression-based methods for nonparametric prediction and estimation of some characteristics of time series // IEEE Trans. on Informat. Theory 2009. Vol. 55, No. 9. P. 4309–4315.

[6] Рябко Б.Я. Дважды универсальное кодирование // Проблемы передачи информации. 1984. Т. 20, № 3. С. 24–28.

[7] Рябко Б., Монарёв В. Экспериментальное исследование методов прогнозирования, основанных на алгоритмах сжатия данных // Там же. 2005. Т. 41, № 1. С. 65–69.

[8] Nevill-Manning C.G., Witten I.H., Paynter G.W. Lexically-generated subject hierarchies for browsing large collections // Intern. J. of Digital Libraries. 1999. Vol. 2, iss. 3. P. 111–123.

[9] Poskitt D.S., Tremayne A.R. The selection and use of linear and bilinear time series models // Intern. J. of Forecasting. 1986. Vol. 2, iss. 1. P. 101–114.

[10] Tong H. Non-linear Time Series: A Dynamical System Approach. Oxford Univ. Press, 1990.

[11] Tong H. Threshold Models in Nonlinear Time Series Analysis. Berlin: Springer-Verlag, 1983.

[12] Tong H., Lim K.S. Threshold autoregression, limit cycles and cyclical data // J. of the Royal Statistical. Ser. B (Methodological). 1980. Vol. 42, iss. 3. P. 245–292.

[13] Engle R. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom // Econometrica. 1982. Vol. 50, iss. 4. P. 987–1007.

[14] Clements M.P., Franses P.H., Swanson N.R. Forecasting economic and financial time-series with non-linear models // Intern. J. of Forecasting. 2004. Vol. 20, iss. 2. P. 169–183.

[15] Cheng H., Tan P-N., Gao J., Scripps J. Multistep-ahead time series prediction // Lecture Notes in Computer Sci. 2006. Vol. 3918. P. 765–774.

[16] Приставка П.А. Экспериментальное исследование метода прогнозирования, основанного на универсальных кодах // Вестник СибГУТИ. 2010. № 4. С. 26–35.

[17] Nevill-Manning C.G., Witten I.H. Identifying hierarchical structure in sequences: A linear-time algorithm // J. of Artificial Intelligence Res. 1997. Vol. 7. P. 67–82.

Поступила в редакцию 8 октября 2013 г.,
с доработки — 13 января 2014 г.