УДК 123.456.789

Прогнозирование временны́х рядов на основе универсальной меры и деревьев принятия решений

А.С. Лысяк, Б.Я. Рябко

В данной работе предложены и развиты два метода прогнозирования временных рядов базирующихся на методах сжатия информации. Показано теоретическое обоснование описываемых методов, а также способы применения данных методов прогнозирования к случаю прогнозирования вещественных рядов. Кроме того, приведены результаты экспериментальных исследований двух рассматриваемых методов на примерах прогнозирования реальных экономических рядов, таких как индексы промышленных и потребительских цен и курсов валют, а также проведено исследование эффективности данных методов и способы выбора эффективных параметров работы данных методов.

Ключевые слова: прогнозирование, временные ряды, R-метод, деревья принятия решений, решающие деревья, ID3.

1. Введение

Методы прогнозирования временны́х рядов представляют большой практический интерес и позволяют решать широкий спектр задач в науке, технике и экономике. К числу таких задач можно отнести анализ социальных, экономических, геофизических процессов, предсказание природных явлений и экономических событий.

Различные методы прогнозирования служат для исследования закономерностей и системных связей в функционировании и развитии различных объектов и процессов. Они являются важным средством в работе с информацией, в анализе работы сложных прикладных систем, а также целенаправленном воздействии человека на объекты исследования с целью повышения эффективности их функционирования.

Наиболее распространённой постановкой задачи прогнозирования является задача прогнозирования временны́х рядов, т. е. функции, определённой на оси времени. Данный тип прогнозирования является очень значимым в силу того, что соответствующий ему класс задач широко связан со многими проблемами экономики, геофизики и других областей нашей жизни. Эти методы играют ключевую роль в повышения эффективности, надёжности и качества технических, экономических, биологических, медицинских и социальных систем.

В последние два десятилетия появилось множество методов прогнозирования, показавших свою достаточно высокую эффективность. В частности, к таковым относятся модели машинного обучения [6], которые стали представлять собой серьёзную конкуренцию классическим статистическим моделям и методам прогнозирования [1, 2, 3].

В настоящее время существует достаточно много эффективных и разнообразных методов прогнозирования, связанных с мощным математическим аппаратом. К ним, в частности, относится прогнозирование на основе билинейной модели [12], авторегресионный анализ различных типов [13–16], прогнозирование на основе методов Монте-Карло [17], методы на основе построения экспертных оценок (так называемые рекурсивные стратегии, описание которых можно найти в [4, 5]) и многое другое. Несмотря на наличие описанного спектра методов и алгоритмов, многие проблемы в задах прогнозирования ещё далеки от своего разрешения. Одна из важнейших таких проблем – повышение качества прогнозирования характеристик систем, описываемых временны́ми рядами. Другой важной проблемой является от-

сутствие значимых или многочисленных результатов и методов прогнозирования на несколько шагов вперёд, несмотря на то, что данный класс задач также является очень важным и актуальным. Такое небольшое количество подходов к этой задаче связано с наличием большого количества сложностей и нерешённых проблем. В частности, к ним относится эффект накапливания ошибок, снижение качества прогноза и увеличение неопределённости с ростом числа прогнозируемых шагов.

В [18] был впервые предложен метод прогнозирования на основе так называемых методов универсального кодирования или «сжатия данных». Данный класс методов включает в себя применение определённых способов кодирования информации, уменьшающих её конечный битовый размер. Преимущество данных методов состоит в выявлении скрытых закономерностей произвольного рода, что позволяет применять метод в достаточно широких диапазонах. Чем меньше битовый размер выходной последовательности относительно входной, тем больше закономерностей имеется в ряду. Данные методы позволяют естественным образом оценивать вероятность всей последовательности при добавлении в конец новых (прогнозных) символов из алфавита и таким образом прогнозировать новые значения ряда.

В ранее опубликованных работах присутствуют преимущественно теоретические результаты без их практической реализации. В данной работе предлагается широкий набор экспериментальных результатов и проведённого на их основе исследования особенностей реализации рассматриваемых методов. В ранних исследованиях в области прогнозирования внимание уделялось преимущественно методам прогнозирования временны́х рядов с конечным алфавитом. В текущей работе предложен подход, позволяющий осуществлять прогнозирование процессов, которые принимают значения из непрерывного интервала.

В данной работе предложены новые подходы к прогнозированию одномерных временны́х рядов, основанные на определённых моделях теории информации и сжатии данных. В качестве базовых методов решения задачи прогнозирования используются методы, построенные на основе универсальных кодов и решающих деревьев. Предложены способы оптимизации описываемых методов и результаты экспериментальных исследований.

2. Постановка задачи прогнозирования

В общем виде задача прогнозирования временны́х рядов может быть сформулирована следующим образом. Пусть имеется некоторый источник, порождающий последовательность элементов x_1x_2 ... 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 $x_1x_2 \dots, x_t$. Задача прогнозирования сводится к предсказанию элемента, следующего в момент времени $(t+1)$, т.е. элемента x_{t+1}. В случае, когда алфавит A является дискретным и конечным, любой алгоритм прогнозирования может быть применён к данному случаю естественным образом, т.к. будет оперировать с конечным множеством алфавита A и с конечной выборкой $x_1x_2 \dots, x_t$.

В случае, если алфавит A представляет собой непрерывный конечный интервал, то поступим следующим образом. Разделим 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. Множество всех номеров при этом будет представлять собой новый, уже конечный дискретный алфавит A'. Далее, преобразуем исходный временной ряд из терминов в алфавите A в ряд, записанный в терминах нашего нового алфавита A'. Таким способом получим некоторую конечную выборку (ряд) из уже конечного алфавита и будем работать с ним, как со случаем конечного дискретного алфавита. При этом после прогнозирования очередного значе-

ния такого ряда ему сопоставляется 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. Пусть источник порождает сообщение $x_1 \dots x_{t-1}x_t, x_i \in A, \quad i = 1,2, \dots, t,$ и требуется прогнозировать n следующих элементов (в простейшем случае – 1 элемент). Ошибкой прогноза называется (апостериорная) величина отклонения прогноза от действительного состояния объекта. Здесь и далее под ошибкой прогнозирования n элементов будем понимать среднюю ошибку прогноза каждого из n элементов в отдельности. Понятно, что ошибка прогноза характеризует качество прогнозирования.

Очевидно, что если распределение вероятностей исходов процесса известно заранее, то задача прогнозирования следующих значений решается достаточно просто. Однако в большинстве практических задач описанные априорные данные отсутствуют, да и не всегда заданное распределение явно существует. В этой работе мы будем рассматривать 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.

3. Схема прогнозирования на базе универсальной меры

3.1. Предсказатель Лапласа и его свойства

Одним из первых исследователей, кто предложил решить задачу прогнозирования с использованием математического аппарата (в частности, теории вероятностей) был Лаплас. Он предложил предсказатель, который принимал на вход весь временной ряд, который требуется спрогнозировать, а также предполагаемое прогнозное значение из алфавита ряда. На выходе предсказатель Лапласа давал некоторую оценку вероятности предполагаемого символа в качестве следующего элемента в заданном ряду.

Предсказатель Лапласа выглядит следующим образом:

$$L_0(a|x_1 \dots x_t) = (\nu_{x_1 \dots x_t}(a) + 1)/(t + |A|),$$

где $\nu_{x_1 \dots x_t}(a)$ обозначает число символов a, встречающихся в слове $x_1 \dots x_t$, a – предполагаемый (прогнозный) элемент, а $x_1 \dots x_t$ – рассматриваемый временной ряд. Исходя из вида предсказателя, ясно, что спрогнозированная вероятность не может равняться нулю.

Лаплас рассмотрел применение данного предсказателя (назовём его предсказатель) для решения проблемы оценки вероятности того, что Солнце взойдет на следующий день (т.е. завтра). При этом мы знаем, что все предыдущие дни, сколько бы мы ни взяли, оно восходило. В данном случае алфавит A состоит из двух элементов: 0 («Солнце взойдет») и 1 («Солнце не взойдет»), t – это количество дней, которые мы рассматриваем в ряду. В итоге, получим ряд: $x_1 \dots x_{t-1}x_t = 0 \dots 00$. Легко видеть, что с ростом t вероятность того, что Солнце взойдёт, будет стремиться к 1, а того, что оно не взойдёт, – наоборот, к нулю.

Важно отметить, что предсказатель Лапласа даёт оценку неизвестных условных вероятностей рассматриваемого стохастического процесса, что довольно естественно. В [8] показано, что средняя ошибка предсказателя Лапласа при оценке вероятностей появления символов на выходе источника независимых и одинаково распределённых символов асимптотически стремится к 0 (т.е. при $t \to \infty$).

3.2. Универсальная мера и её свойства

В работах [4, 7] в качестве подхода для решения задачи прогнозирования временны́х рядов предлагается использовать методы, основанные на универсальной мере.

Приведём определение универсальной меры, а также поясним связь между данным и описанным в предыдущем пункте подходами. По определению, мера μ называется универсальной, если для любого стационарного и эргодического источника P верны следующие равенства:

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

с вероятностью 1, и

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

Данные равенства показывают, что, в определённом смысле, мера μ является непараметрической оценкой для меры P (в общем случае неизвестной). По этой причине универсальные меры могут быть использованы для оценки статистических характеристик процесса, а также для оценки вероятностей последовательностей, генерируемых любыми стационарными и эргодическими источниками на конечном алфавите.

Универсальная мера тесно связана с понятием универсального кода. Опишем взаимосвязь этих понятий. Код U называется универсальным, если для любого стационарного и эргодического источника P верны следующие равенства:

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

с вероятностью 1, и

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

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

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

Универсальные меры имеют глубокую взаимосвязь с универсальными кодами, и, если есть универсальный код, то можно легко получить на его основе универсальную меру.

Следующая теорема говорит о том, что на базе любого универсального кода можно построить универсальную меру.

Теорема 1. Пусть U – универсальный код и

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

тогда μ – это универсальная мера.

Универсальный код называется оптимальным, если он кодирует последовательность символов, порожденную дискретным источником, таким образом, что длина полученной кодовой последовательности асимптотически минимальна. Универсальные коды для стационарных и эргодических дискретных источников были описаны в 1980-ые [11].

3.3. Схема прогнозирования для источников из конечного алфавита

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

После представления Лапласом своего предсказателя возникла задача нахождения такого предсказателя, для которого средняя ошибка при оценке вероятностей для источника независимых и одинаково распределённых символов будет минимальной. Эту проблему решил Кричевский, предложив следующий предсказатель:

$$K_0(a|x_1 \dots x_t) = (\nu_{x_1 \dots x_t}(a) + 1/2)/(t + |A|/2), \qquad (1)$$

где $\nu_{x_1 \dots x_t}(a)$ обозначает число символов a, встречающихся в слове $x_1 \dots x_t$. Легко увидеть схожесть предсказателей Кричевского и Лапласа.

Затем Кричевский обобщил свой предсказатель на меру, которая является универсальной для множества марковских источников с памятью, или связностью $m, m \geq 0$; если $m = 0$. В некотором смысле эта мера является оптимальной для этого множества, что показано в [19].

Итак, мера Кричевского выглядит следующим образом:

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

где $\nu_x(\vartheta)$ – число последовательностей ϑ, встречающихся в x, $\overline{\nu}_x(\vartheta) = \sum_{a \in A} \nu_x(\vartheta a)$, $x = x_1 \dots x_t$, а $\Gamma()$ – это гамма-функция. Параметр m называется порядком меры Кричевского и фактически определяет глубину (количество элементов) в рассматриваемом процессе, на которой данная мера будет находить какие-либо закономерности.

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

$$\omega_i = 1/\log(i+1) - 1/\log(i+2). \quad (3)$$

В дальнейшем будем использовать именно это распределение.

Мера *R* определяется как

$$R(x_1 \dots x_t) = \sum_{i=0}^{\infty} \omega_{i+1} K_i(x_1 \dots x_t). \quad (4)$$

Слагаемые ω_i играют в данном случае роль весовых коэффициентов, которые с ростом индекса i (т.е. порядка меры Кричевского) становятся всё меньше и меньше. Понятно, что слишком большие порядки в мере Кричевского должны иметь меньший вес и меньше влиять на прогноз. В результате в качестве весовых коэффициентов и было выбрано распределение (3). В общем случае весовые коэффициенты представляют собой варьируемый параметр метода и могут меняться в зависимости от конкретного ряда и метода.

Таким образом, мы построили меру *R*, которая является оценкой вероятностей для класса всех стационарных и эргодических источников на конечном алфавите и может непосредственно применяться для прогнозирования.

Итак, значение меры *R*, вычисленное на основе формулы (4), может служить оценкой вероятности исхода стационарного и эргодического процесса и использоваться для решения задачи прогнозирования.

Рассмотрим схему прогнозирования на основе универсальной меры как для источников на дискретном, так и на непрерывном алфавите.

Для осуществления прогноза будем использовать меру *R* (4). Вычисление меры R будет состоять из вычисления суммы (4) до элемента $i = t$, где t – это длина ряда, и суммы (3) после этого элемента. Во второй части суммы все слагаемые будут одинаковы и равны $\frac{1}{|A|^t}$, что позволяет вычислить слагаемые меры R после элемента *t* следующим образом:

$$\sum_{i=t}^{\infty} \omega_{i+1} K_i(x_1 \dots x_t), = \sum_{i=t}^{\infty} (1/\log(i+1) - 1/\log(i+2)) \cdot \frac{1}{|A|^t} = \frac{1}{|A|^t \log(t+1)}.$$

Видно, что с ростом длины ряда рассматриваемая вторая часть суммы (4) стремится к 0. Таким образом, существенное влияние как на значение ряда (4), так и на сложность его вычисления оказывает только первая часть суммы. Исходя из определения весовых коэффициентов (3), видно, что с ростом *i* значение коэффициента w_i стремится к нулю и является небольшим при больших *i* (при $i > 5$ $w_{i+1} < 0.05$). Одновременно значение меры Кричевского при больших порядках также уменьшается, что видно исходя из вида меры (2), а также будет показано далее на примере практических результатов. Соответственно, вклад слагаемого $\omega_{i+1} K_i(x_1 \dots x_t)$ с ростом i будет небольшим и в целях уменьшениях сложности вычислений

его можно ограничить каким-либо параметром m, где $m = 1, \dots, t$. Назовём этот параметр глубиной анализа метода.

Вначале рассмотрим источник, порождающий значения из конечного алфавита. В данном случае схема вычисления меры R достаточна проста. Пусть $x_1 \dots x_t$ – имеющаяся временна́я последовательность. Для каждого $a \in A$ построим последовательность $x_1 \dots x_t a$ и вычислим условную вероятность на основе меры R:

$$R(a|x_1 \dots x_t) = R(x_1 \dots x_t a)/R(x_1 \dots x_t)$$

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

3.4. Схема прогнозирования для источников из непрерывного интервала

Очень актуальным стало прогнозирование элементов временно́го ряда для источника, принимающего значения из непрерывного интервала. Описанные ранее результаты в этой области имеют преимущественно теоретический характер, а полученные экспериментальные данные относятся к случаю конечного дискретного алфавита. В частности, данные результаты описаны в работе [7]. Практические экспериментальные результаты прогнозирования временны́х рядов существуют только для случая источника, принимающего значения из дискретного алфавита.

Рассмотрим схему прогнозирования для источника, принимающего значения 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 $[A, B]$. И 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 производилось равномерно на равные подынтервалы, т.е. размер каждого подынтервала определяется, как $h = \frac{B-A}{n}$. Обоснование выбора именно такого метода будет дано далее. Определим также $x^{[k]}$, как элемент Π_k, содержащий точку x.

Определим совместное распределение P_n для $(X_1, X_2, \dots, X_n)$, как функцию плотности вероятности $p(x_1 x_2 \dots x_n)$ по отношению к сигма-конечной мере L. В качестве L может выступать мера Лебега или какая-либо счётная мера.

Для целых s и n определим оценку плотности вероятности $p(x_1 x_2 \dots x_n)$:

$$p^s(x_1, \dots, x_n) = P({x_1}^{[s]}, \dots, {x_n}^{[s]})/L\left({x_1}^{[s]} \dots {x_n}^{[s]}\right). \quad (6)$$

Определим теперь оценку плотности вероятностей r следующим образом:

$$r(x_1 \dots x_t) = \sum_{s=1}^{\infty} \omega_s R({x_1}^{[s]} \dots {x_t}^{[s]}) \, / L({x_1}^{[s]} \dots {x_t}^{[s]}). \quad (7)$$

Коэффициенты ω_s определяются схемой (3) и несут роль весовых коэффициентов для случая каждого разбиения из Π_k. При этом происходит нормировка каждого слагаемого (представляющего собой некоторую оценку вероятности последовательности при заданном разбиении) по сигма-конечной мере L, которая представляет собой некоторую оценку вероятностей величин подынтервалов. Таким образом, мы соединяем между собой оценки плотности вероятностей для случая различных возрастающих разбиений, что избавляет нас от зависимости результатов прогноза от конкретного разбиения. Этот процесс соединения оценок плотностей вероятностей с нормировкой по L и с весовыми коэффициентами ω_s называется «склейкой». В итоге, мы можем использовать для прогнозирования произвольные последовательности конечных разбиений (также произвольных).

Как показано в [4, 8], величина $r(x_1 \dots x_t)$ является оценкой неизвестной плотности $p(x_1 \dots x_t)$, а соответствующая условная плотность

$$r(a|x_1 \dots x_t) = r(x_1 \dots x_t a)/r(x_1 \dots x_t) \quad (8)$$

является подходящей оценкой вероятности $p(a|x_1 \dots x_t)$. Количество слагаемых в сумме (7) при реализации описанного далее алгоритма, как и в дискретном случае, равно глубине анализа m.

4. 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 лежит 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 $(M+1)$-ый атрибут. При этом на вход подаётся число N (элементов в обучающей выборке), число M, параметр $m \leq M$.

Как правило, данный метод применяется для задач классификации и кластеризации. В данной работе предложен подход, который показывает способ применения данных деревьев к прогнозированию временны́х рядов. Дерево принятия решений строится по описанному ниже алгоритму.

Введём вначале некоторые важные определения.

Определение 1. Энтропия $H(A,S) = -\sum_{i=1}^{Sn} \frac{|A_i|}{|A|} \log_2 \frac{|A_i|}{|A|}$, 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 следующей формуле:

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

Далее, опишем непосредственно один из наиболее эффективных алгоритмов построения решающего дерева, названный 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 выбрать такое q из множества значений S, которому равно наибольшее число элементов из A, и поставить q в корень и выйти.
4. Выбрать $q \in Q$, для которого $\text{Gain}(A, q)$ максимален.
5. Поставить в корень дерева метку q.
6. Для каждого значения q_i атрибута q:
   a. Добавить нового потомка и пометить исходящее ребро меткой q_i.
   b. Если в A нет элементов, для которых значение q равно q_i, то поступить в соответствии с п.3.
   c. Иначе запустить $\text{ID3}(A_{q_i}, S, Q\backslash\{q\})$ и добавить его результат, как поддерево с корнем в этом потомке.

Дерево строится до исчерпания обучающего множества или до пустоты множества Q. В предлагаемой реализации данного алгоритма можно ограничивать глубину дерева искусственно – отдельным параметром. После достижения глубины дерева заданной глубины 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). Спрогнозировать требуется результат игры при известных характеристиках игры. При этом известны данные приведены в нижеследующей таблице.

Таблица 1. Начальные данные для решающего дерева

Соперник	**Играем**	**Лидеры**	**Дождь**	**Победа**
Выше	Дома	На месте	Да	Нет
Выше	Дома	На месте	Нет	Да
Выше	Дома	Пропускают	Нет	Да
Ниже	Дома	Пропускают	Нет	Да
Ниже	В гостях	Пропускают	Нет	Нет
Ниже	Дома	Пропускают	Да	Да
Выше	В гостях	На месте	Да	Нет
Выше	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. $Gain(A, \text{Лидеры}) =$
   $= H(A, \text{Победа}) - \frac{3}{7} H(A_{\text{на месте}}, \text{Победа}) - \frac{4}{7} H\left(A_{\text{пропускают}}, \text{Победа}\right) = 0.1281;$
2. $Gain(A, \text{Играем}) =$
   $H(A, \text{Победа}) - \frac{5}{7} H\left(A_{\text{дома}}, \text{Победа}\right) - \frac{2}{7} H(A_{\text{гостях}}, \text{Победа}) = 0.4696$

и т.д. для всех 4 свойств…

Следуя описанному алгоритму ID3, строим дерево, выбирая на каждом этапе признак с максимальным приростом информации. В итоге получим следующее дерево:

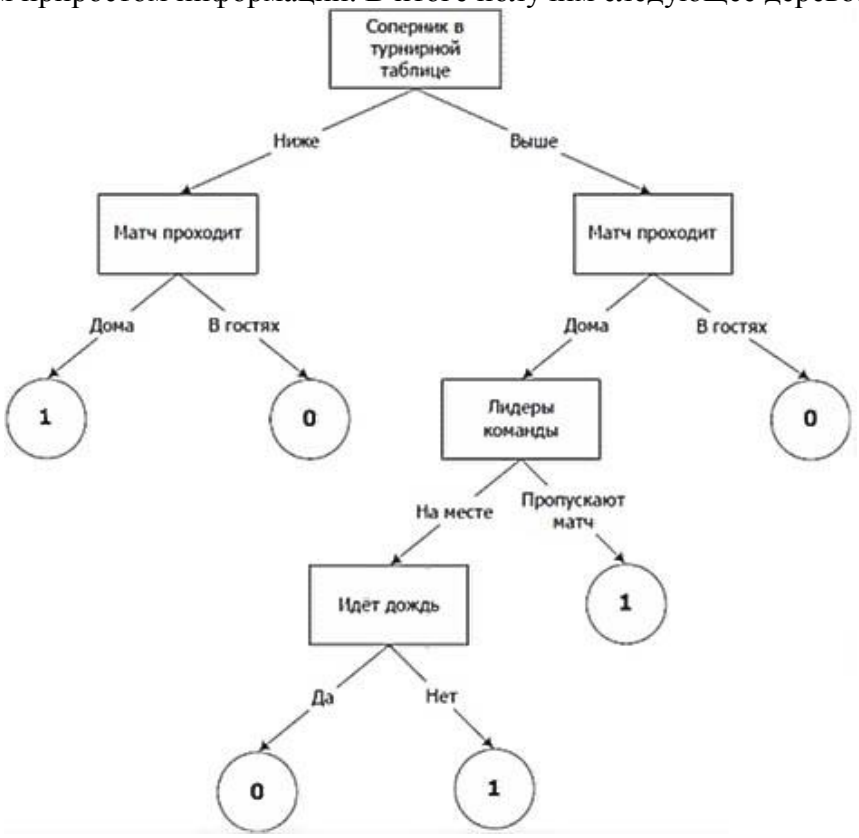

Рис. 1. Решающее дерево для прогнозирования игры в футбол

Для применения данного дерева в случае одномерного прогнозирования временно́го ряда возьмём в качестве признаков предыдущие значения ряда. Сделаем это следующим образом. Зададим параметр метода m, имеющий смысл аналогично 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-ое значение ряда, а в качестве его $m-1$ атрибутов примем $m-1$ значений, стоящих перед i-ым во временно́м 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 беря последние $m-1$ элементов ряда, получим прогнозное значение.

В силу того, что при большой глубине анализа и большом алфавите дерево будет слишком сильно разветвляться и трудоёмкость алгоритма будет расти 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.

5. Экспериментальные результаты прогнозирования методами R и ID3

Оба описанных метода реализованы на суперкомпьютере и протестированы на прогнозах реальных данных. Результаты данных прогнозов приведены в нижеследующих таблицах.

Все исследования проводилась в двух режимах. Первый режим– on-line – означает прогнозирование значений временно́го ряда на 1 шаг вперёд. Второй режим – на 10 шагов вперёд – прогнозирование значений ряда на 10 шагов вперёд. При этом прогнозирование на 10 шагов вперёд считалось следующим образом: прогнозировалось очередное значение ряда, после чего выборка пополнялась прогнозным значением, далее прогнозирование велось ещё на 1 шаг, но с использованием уже пополненного ряда и так далее продолжаем до 10-го прогнозного элемента, после чего считаем ошибку прогноза. Прогнозирование выполнялось на 10 разных выборках с последующим усреднением ошибки.

Важно отметить, что выполнялось прогнозирование не абсолютных величин выборки, а разницы между соседними элементами с последующей прибавкой спрогнозированной разницы к последнему элементу ряда (в результате получим значение следующего за последним элемента ряда); такой подход позволяет существенно снизить необходимый размер непрерывного интервала, в котором лежат прогнозные значения, а также позволяет выявлять линейные и квазилинейные тренды и периоды на них. Это было невозможно при прогнозировании абсолютных величин временно́го ряда. Определение границ интервала осуществлялось естественным образом: считалась величина максимального и минимального (с учётом знака) отклонения текущего и предыдущего элементов; полученные значения и брались в качестве левой и правой границе интервала, которые далее и разбивался на количество частей, равное мощности алфавита.

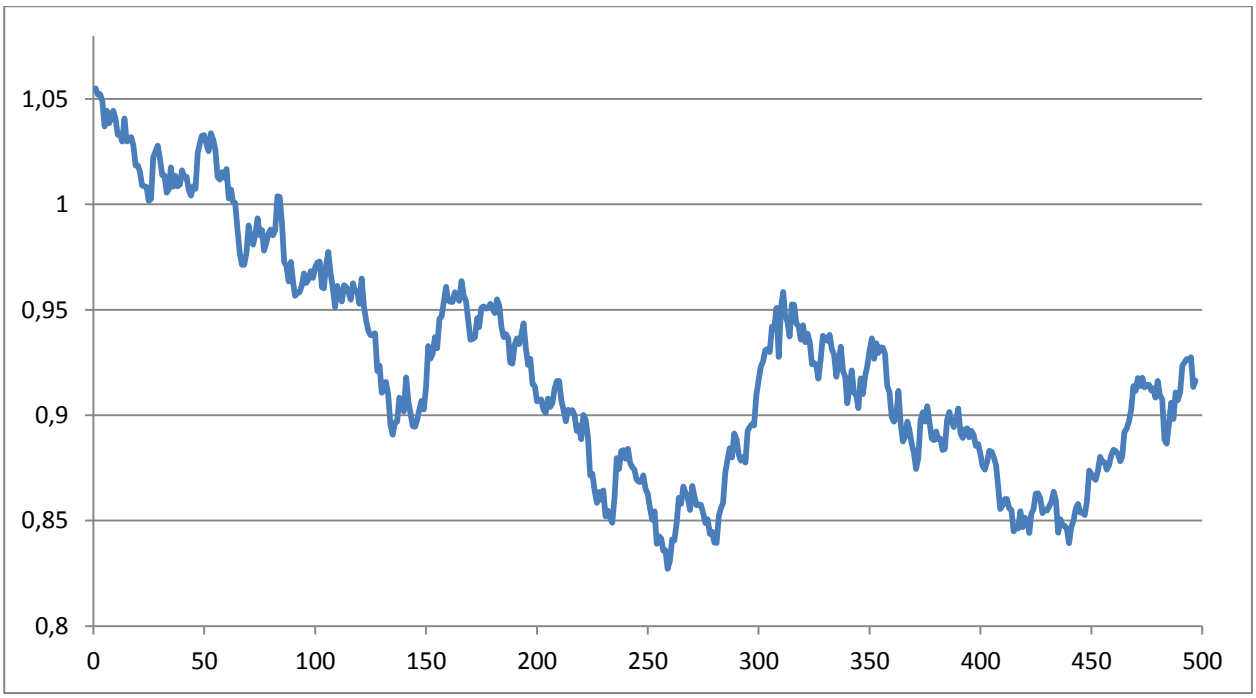

Рис. 2. График курса евро/доллар (период: 1 день)

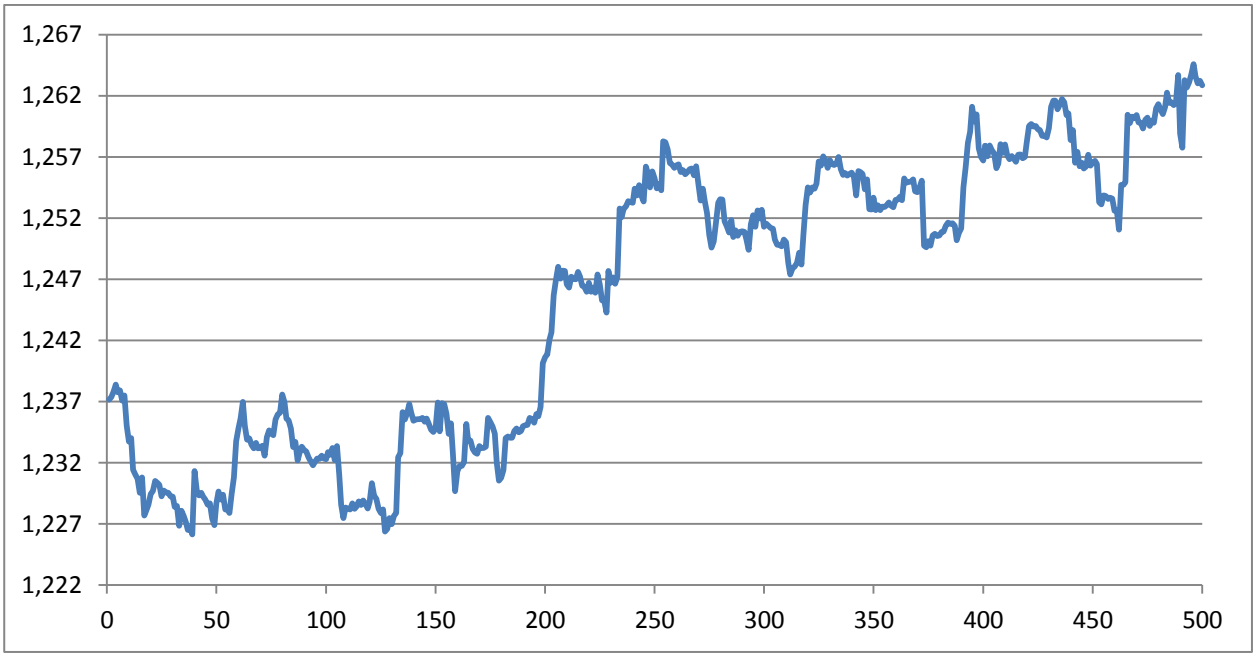

Рис. 3. График курса евро/доллар (период: 1 час)

В каждой из следующих ниже таблиц приведены параметры: размер выборки (n), количество частей разбиения непрерывного интервала, параметры m' (максимальная глубина дерева) и m (разделены знаком «/») и величина ошибки прогноза каждого из методов.

В табл. 2 приведены данные прогноза курса валют евро/доллар с временны́м интервалом (timeframe) один день (D1) в период с 20.06.2012. График ряда приведён на рис. 2.

Таблица 2. Прогнозирование курса евро/доллар (период: 1 день)

Размер выборки L	Разбиение *n*	Макс. глубина д-ва / m	Решающие деревья On-line	R-measure On-line	Решающие деревья 10 шагов	R-measure 10 шагов
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

В табл. 3 приведены результаты прогнозирования того же курса евро/доллар, но уже наряду с периодом (timeframe) 1 час. График ряда приведён на рис. 3.

Таблица 3. Прогнозирование курса евро/доллар (период: 1 час).

Размер выборки L	Разбиение *n*	Глубина анализа *m*	Решающие деревья On-line	R-measure On-line	Решающие деревья 10 шагов	R-measure 10 шагов
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

Из приведённых выше результатов видно, что после определённого предела размера алфавита (разбиения непрерывного интервала) ошибка прогноза методов перестаёт падать. Это справедливо как для метода R, так и для метода решающих деревьев. Кроме того, из полученных данных видно, что глубина анализа после значения $m = 2$ улучшает точность прогноза весьма несущественно и после какого-то заданного m точность, как и в случае с размером алфавита, уже не меняется. Фактически, 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 и просто усредняют значение тренда (разницу между соседними элементами) и используют в качестве прогноза. При наличии же каких-либо закономерностей в ряду, алгоритмам потребуется бóльшая глубина анализа (при m размером меньше, чем длина периода закономерности, алгоритмы просто её не выявят).

Рассмотрим другой экономический временнóй ряд: индекс потребительских цен США в период с 02.1992 по 02.2013 с интервалом между значениями в 1 месяц. График данного ряда приведён на рис. 4. Исходя из его вида, хорошо прослеживается достаточно понятная закономерность в виде повышающего тренда. Разницы между соседними элементами, которые мы будем прогнозировать, практически не меняются и близки константному значению. При этом более гладким выглядит начальная часть графика, что позволяет предварительно говорить о лучшем качестве прогноза обоих методов на нём. Ближе к концу ряда дисперсия ряда сильно возрастает, и качество прогнозов может ухудшиться. Проведём экспериментальные исследования.

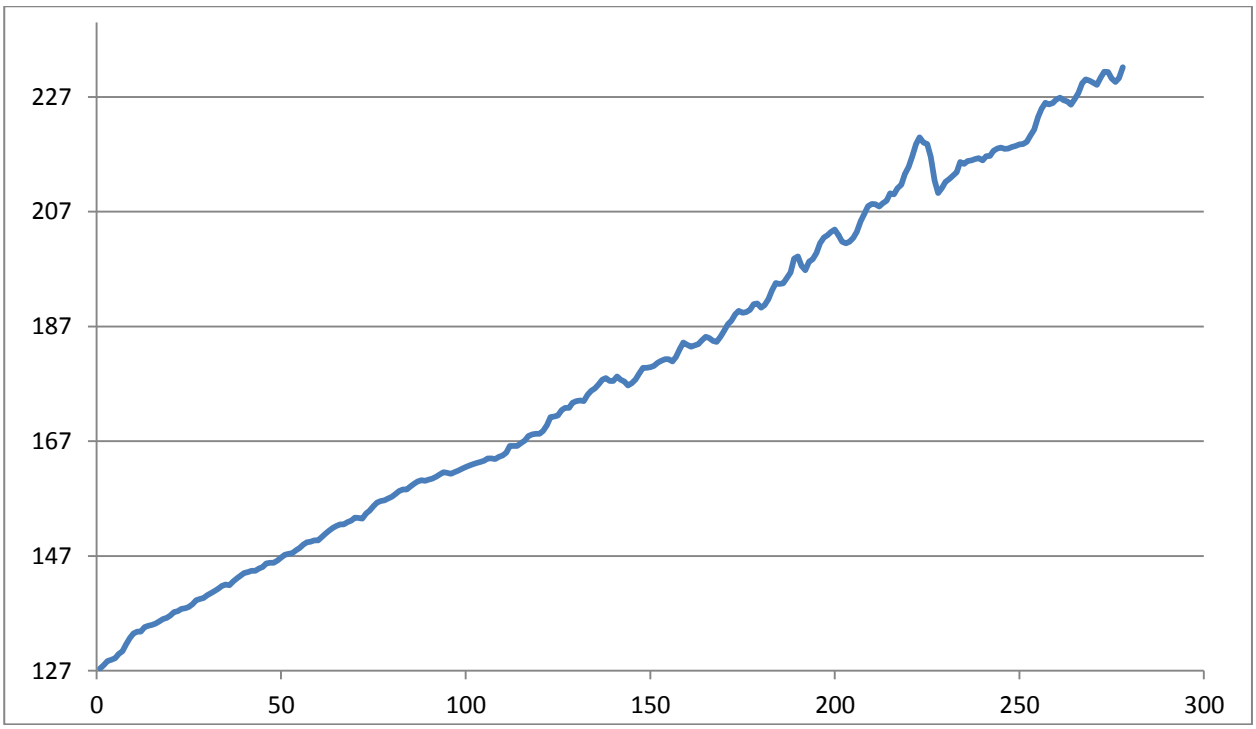

Рис. 4. Индекс потребительских цен

В табл. 4 отражены результаты прогноза последней части ряда, в которой дисперсия выборки явно больше, чем в середине (что хорошо видно из рис. 4). В табл. 5 показаны результаты прогноза начальной части временно́го ряда, которая выглядит более гладкой.

Таблица 4. Прогнозирование индекса потребительских цен. Период: 03.2012 – 02.2013

Размер выборки L	Разбиение *n*	Глубина анализа *m*	Решающие деревья On-line	R-measure On-line	Решающие деревья 10 шагов	R-measure 10 шагов
277	5	**2** / 2	0.0966154	0.112769	0.256923	0.257077
	10	**2** / 2	0.0926154	0.092615	0.088	0.088154
		2 / 5	0.1269231		0.264154	
	20	**2** / 2	0.0847692	0.107846	0.110462	0.136
		2 / 5	0.0936923		0.165385	
	100	**2** / 2	0.0946154	0.100769	0.115846	0.230308
		2 / 5	0.1249231		0.285846	

Таблица 5.Прогнозирование индекса потребительских цен. Период: 03.2002 – 02.2003

Размер выборки L	Разбиение *n*	Глубина анализа *m*	Решающие деревья On-line	R-measure On-line	Решающие деревья 10 шагов	R-measure 10 шагов
240	5	**2** / 2	0.0761538	0.056	0.105231	0.105231
	10	**2** / 2	0.0606154	0.050462	0.193077	0.193077
		2 / 5	0.0606154		0.193077	
	20	**2** / 2	0.0627692	0.047692	0.065231	0.065231
		2 / 5	0.062		0.080308	
	100	**2** / 2	0.0627692	0.047692	0.11	0.11
		2 / 5	0.0627692		0.11	
	240	**2** / 5	0.0518462	0.045692	0.061077	0.131231

Как видно по вышеприведённым результатам, во второй таблице, отражающей значения на меньшей выборке и с меньшей дисперсией, ошибка прогноза значительно меньше, чем в первой таблице, соответствующей выборке с большей длиной и бóльшим разбросом значений.

Рассмотрим далее индекс промышленных цен за тот же временной интервал, что и ряд индекса промышленных цен. График данного ряда приведён на рис. 5.

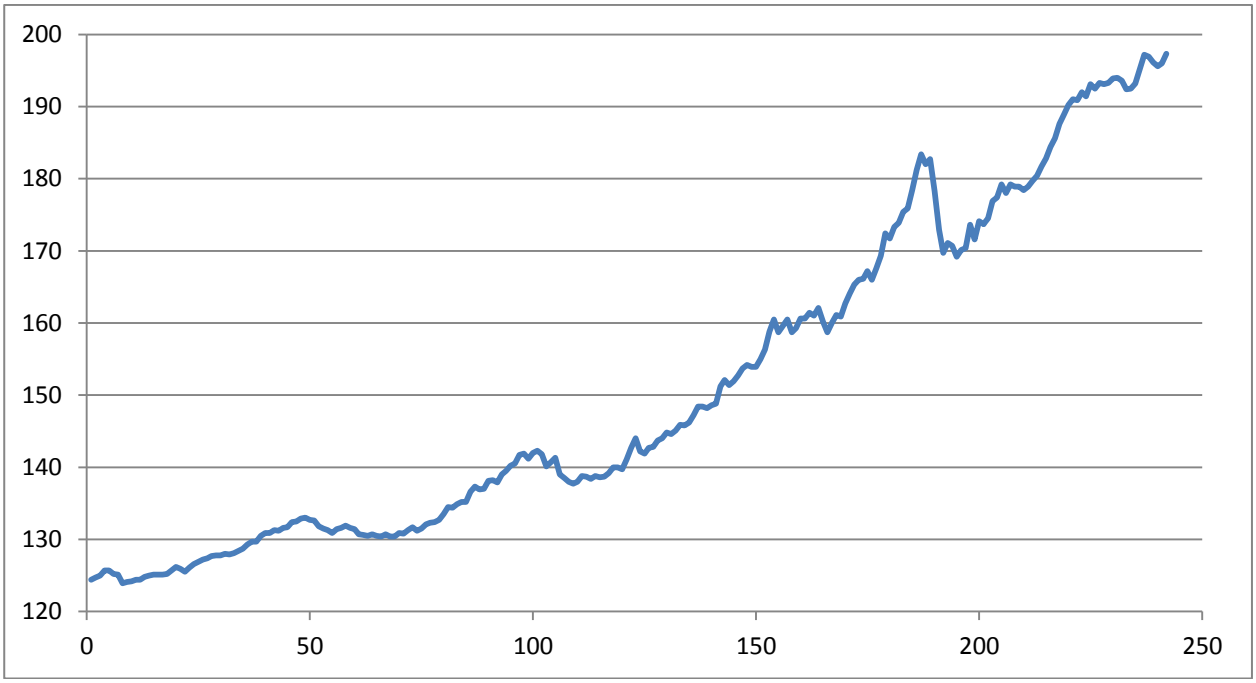

Рис. 5. Индекс промышленных цен

Далее рассмотрим прогнозирование индекса промышленных цен. Результаты приведены в табл. 6.

Таблица 6. Прогнозирование индекса промышленных цен.

Размер выборки L	Разбиение *n*	Глубина анализа *m*	Решающие деревья On-line	R-measure On-line	Решающие деревья 10 шагов	R-measure 10 шагов
277	5	**2** / 2	0.1109302	0.11093	0.217442	0.217442
	10	**2** / 2	0.115814	0.105814	0.158488	0.158488
		2 / 5	0.1537209		0.208488	
	20	**2** / 2	0.1244186	0.10593	0.108721	0.125116
		2 / 5	0.130814		0.125116	
	100	**2** / 2	0.1177907	0.105814	0.14	0.117791
		2 / 5	0.1506977		0.173023	
	240	**2** / 5	0.1009302	0.105814	0.166163	0.11814

Данные результаты подтверждают сделанные ранее выводы и показывают достаточно высокую эффективность обоих методов. Оба метода показывают примерно равную точность получаемых прогнозов. При этом деревья принятия решений, как правило, лучше справляются с короткими временны́ми рядами, выявляя на них закономерности быстрее, чем метод R.

6. Заключение

Как видно из приведённых выше результатов, метод на основе универсальной меры и метод на основе деревьев принятия решений показывают достаточно высокий уровень точности при прогнозировании на реальных данных. Это верно и по сравнению с другими методами прогнозирования, что подробно показано в [4]. Кроме того, была создана более эффективная реализация метода R, позволяющая сократить трудоёмкость перебора различных вариантов прогнозных значений из алфавита. Трудоёмкость снизилась до константной относительно длины алфавита (была линейной).

Важно отметить, что преимущество предложенных методов прогнозирования состоит в том, что их можно легко обобщить на случай прогнозирования многомерных рядов, где кроме значения у каждого элемента имеются ещё другие атрибуты-свойства, коррелирующие друг с другом.

Литература

1. *Ahmed N*. An empirical comparison of machine learning models for time series forecasting // Econometric Reviews. 2010,Vol. 29, Issue 5-6. P. 594-621.
2. *Palit A. K., Popovic D*. Computational Intelligence in Time Series Forecasting: Theory and Engineering Applications (Advances in Industrial Control). Springer-Verlag New York: Secaucus, NJ, USA, 2005.
3. *Zhang G., Patuwo B. E., Michael Y. H*. Forecasting with articial neural networks:: The state of the art // International Journal of Forecasting. 1998. Vol. 14, Issue 1. P. 35--62.
4. *Приставка П.А.* Экспериментальное исследование метода прогнозирования, основанного на универсальных кодах // Вестник СибГУТИ, 2010. №4, С. 26--35.
5. *Cheng H. et al.* Multistep-ahead time series prediction // Lecture Notes in Computer Science. 2006. V. 3918. P. 765--774.
6. *Bontempi G*. Local Learning Techniques for Modeling, Prediction and Control. Ph.d., IRIDIA-Universit de Libre de Bruxelles, BELGIUM, 1999.
7. *B. Ryabko*. Compression-Based Methods for Nonparametric Prediction and Estimation of Some Characteristics of Time Series. // IEEE Transactions on Information Theory, Vol. 55, No. 9, 2009.P. 4309--4315.
8. *Рябко Б. Я.* Дважды универсальное кодирование // Проблемы передачи информации. 1984. Т. 20, № 3. С. 24--28
9. *РябкоБ., МонарёвВ*. Экспериментальное исследование методов прогнозирования, основанных на алгоритмах сжатия данных // Проблемы передачи информации. 2005. С. 65--69.
10. *Nevill-Manning C.G., Witten I.H., Paynter G.W*. Lexically-GeneratedSubjectHierarchiesforBrowsingLargeCollections // International Journalof Digital Libraries. 1999. Vol. 2, Issue 3. P. 111--123.
11. *Nevill-Manning C.G., Witten I.H.*Identifying Hierarchical Structure in Sequences: A linear-time algorithm // Journal of Artificial Intelligence Research. 1997. Vol. 7. P. 67--82.
12. *Poskitt D.S., Tremayne A.R*. The selection and use of linear and bilinear time series models // International Journal of Forecasting. 1986. Vol. 2, Issue 1. P. 101--114
13. *Tong H.* Non-linear Time Series: A Dynamical System Approach. Oxford University Press, 1990.
14. *Tong H*. Threshold models in Nonlinear Time Series Analysis. Springer Verlag, Berlin, 1983.
15. *Tong H., Lim K. S*. Threshold autoregression, limit cycles and cyclical data //. Journal of the Royal Statistical. Series B (Methodological). 1980. Vol. 42, Issue 3. P. 245--292.
16. *Engle R*. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom // Econometrica. 1982. Vol. 50. Issue 4. P. 987--1007.

17. *Clements M.P. et al*. Forecasting economic and financial time-series with non-linear models // International Journal of Forecasting. 2004. Vol. 20.Issue 2. P. 169--183
18. *Рябко Б.Я.* Прогнозирование случайных поседовательностей и универсальное кодирование. // Проблемы передачи информации. 1988. №24. С.3--14.
19. *Krichevsky R*. Universal Compression and Retrival. KluverAcademicPublishers, 1993.

Статья поступила в редакцию 16.03.2014;
переработанный вариант — 16.05.2014

Лысяк Александр Сергеевич

Инженер лаборатории информационных технологий в вычислительных системах СибГУТИ, ассистент кафедры систем информатики НГУ (630090, Новосибирск, ул. Пирогова, 2), тел. +79130110883; e-mail: accemt@gmail.com.

Рябко Борис Яковлевич

д.т.н., ректор СибГУТИ, профессор кафедры прикладной математики и кибернетики СибГУТИ (630102, Новосибирск, ул. Кирова, 86), тел. (383) 2-698-272,

e-mail: boris@ryabko.net.

Prediction of time series based on a universal measure and decision trees

A.S. Lysiak, B.J. Ryabko

In this paper, we propose and develop two methods for time series forecasting based on the methods of data compression. Theoretical justification of the methods described is presented, as well as the ways of using the prediction methods to the real series forecasting. In addition, the results of experimental studies of two methods are considered illustrated by real economic series forecasting such as indexes of industrial and consumer prices and exchange rates. The effectiveness of these techniques and methods for selecting effective parameters of these methods are also investigated.

Keywords: forecasting, time series, R-method, decision trees, ID3.