Тема: разработка и исследование теоретико-информационных методов прогнозирования временных рядов, базирующихся на теории универсальной меры и решающих деревьях.

Состав коллектива:

- А.С. Лысяк (ст. преподаватель каф. КС ФИТ НГУ, аспирант ФИТ НГУ (очная форма, спец. 05.13.18), срок окончания: 01.10.2015 (на текущий момент: защита кандидатской диссертации и поступление в докторантуру) e-mail: accemt@gmail.com).
- П.А. Приставка (док. кафедры ПМиК СибГУТИ, зам. декана, к.т.н; e-mail: ppa@ngs.ru).
- Б.Я. Рябко (зав. лабораторией защиты информации ИВТ СО РАН, проф. каф. КС ФИТ НГУ, д.т.н., проф., e-mail: boris@ryabko.net).

Отчёт о проделанной работе и необходимость продления доступа:

1. Постановка задачи.

Пусть имеется некоторый источник, порождающий последовательность элементов $x_1, x_2, \dots, x_t, x_{t+1} \in A$ из некоторого множества A,называемого алфавитом. Задача прогнозирования состоит в определении распределения вероятностей для случайной величины $x_{t+1} \in A$, т.е. в определении для конечного дискретного алфавита условных вероятностей вида: $p(x_{t+1} = a \in A | x_1, x_2, \dots, x_t)$, а для случая, когда алфавит представляет собой ограниченный непрерывный интервал, условной плотности вероятности. В данном разделе описываются оба случая: когда алфавит A является конечным и когда представляет собой некоторый ограниченный непрерывный интервал. Ошибка прогноза при этом определяется следующим образом: $E_i = |x_i - x_i^*|$, где x_i^* – прогнозное значение (полученное из распределения вероятностей), а $x_i -$ истинное значение процесса в момент времени i.

В более общей постановке задачи прогнозирования элементы x_i могут быть не только конкретными числами (целыми или вещественными), а векторами размерности k, где первый элемент вектора – значение прогнозируемой характеристики ряда, а оставшиеся $(k-1)$ атрибутов – какие-либо характеристики рассматриваемого процесса или величины, коррелирующие со значениями ряда и известные для всех элементов ряда.

Для решения данной задачи будем рассматривать теоретико-информационный подход к определению плотности распределения x_{t+1}. В случае, если алфавит A является конечным, алгоритмы прогнозирования на основе вероятностного подхода, учитывающие распределение вероятностей символов источника и их конечных цепочек, могут применяться для прогнозирования таких источников естественным образом.

В случае, если алфавит A представляет собой непрерывный ограниченный интервал, то требуется оценить плотность вероятности распределения величины $x_{t+1} \in A$.

2. Необходимость использования кластера ИВТЦ

Разработанный и описанный ниже метод прогнозирования был реализован на кластере, что позволило сократить время вычислений в среднем в 20-30 раз. Опишем ниже схему распараллеливания данного алгоритма.

Пусть у нас имеется разбиение, состоящее из N частей, а количество элементов ряда равно $n = t$. Тогда вычислим первоначально $r(x_1 \dots x_t)$, который будем далее использовать для ускоренного вычисления значений $r(x_1 \dots x_t a_i), i = \overline{1,N}$, где $a_i \in [A,B]$ – произвольные точки из различных частей разбиения. Без ограничения общности будем считать, что a_i – это середины интервалов разбиения. Далее, имея $r(x_1 \dots x_t)$, вычисляем независимо и параллельно значения $r(x_1 \dots x_t, a_i)$. При этом каждое данное значение вычисляется отдельным подпроцессом. Для этого используем N параллельных процессов. В силу того, что разбиение в процессе практических экспериментов не превышало 100, число используемых процессов также было меньше или равно 100. На выходе данные процессы отправляют в управляющий процесс свои значения, на основе которых уже считаются условные вероятности (данное вычисление представляет собой одну математическую операцию и в распараллеливании не нуждается). Также, управляющий процесс вычисляет прогнозное значение $x_{\text{пр.}}$, которое совместно с посчитанной плотность вероятности отправляется на выход программы.

Общая схема параллельных вычислений представлена на рисунке 1а.

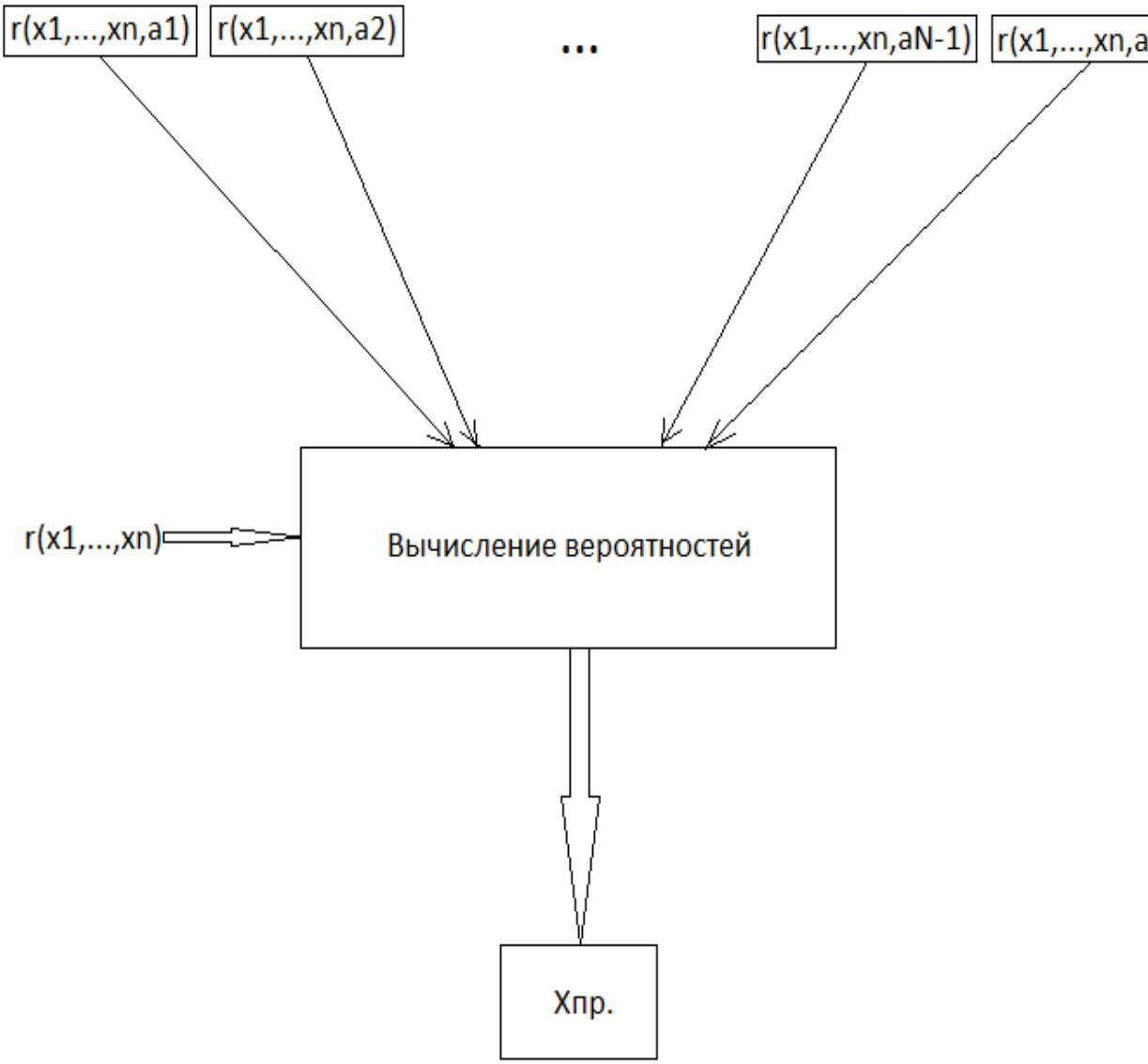

Рисунок 1а – Распараллеливание процесса вычисления прогнозного значения.

Реализация представленного алгоритма и схемы его работы была осуществлена на языке программирования C++. Для распараллеливания был использован интерфейс MPI (Message passing interface) в реализации библиотеки OpenMP.

3. Подробное описание работы, включая используемые алгоритмы:

3.1. Универсальная мера и её свойства

В 1988 году был предложен метод прогнозирования на основе использования сжатия данных. Точнее, было предложено использовать универсальную меру, базирующуюся на универсальных кодах.

Приведём определение универсальной меры, а также поясним связь между данным и описанным в предыдущем пункте подходами. Рассмотрим определение универсальной меры. Пусть дан стационарный и эргодический источник P. Тогда код U называется универсальным, если для любого такого 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.

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

Мера μ называется универсальной, если для любого описанного выше источника P верны следующие равенства:

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

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 определённом смысле, мера μ является непараметрической оценкой для неизвестного распределения источника P. Таким образом, универсальная мера может быть использована для оценки статистических характеристик процесса, а также для оценки вероятностей последовательностей, генерируемых любыми стационарными и эргодическими источниками на конечном алфавите.

Универсальные меры имеют глубокую взаимосвязь с универсальными кодами, и, если есть универсальный код, то можно легко получить на его основе универсальную меру и наоборот: на основе универсальной меры можно построить универсальный код. Следующее простое утверждение 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, порождённую источником P, таким образом, что средняя длина полученной кодовой последовательности асимптотически минимальна. Фактически оптимальный универсальный код максимально сжимает информацию, заключённую во временном ряде. Оптимальные универсальные коды для стационарных и эргодических дискретных источников были описаны в 1980-ых.

Рассмотрим универсальную меру R, которая использовалась для прогнозирования описанных в данной работе временных рядов. Выбор именно этой меры связан с тем, что она построена на основе асимптотически оптимального универсального кода.

В 1968 году после открытия универсального кодирования найден предсказатель, для которого погрешность асимптотически минимальна. Данный предсказатель

предложил Кричевский. Он описал следующий предиктор, позволяющий вычислить условные вероятности для следующего элемента ряда:

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

где $\nu_{x_1\dots x_t}(a)$ – число элементов a, встречающихся в слове $x_1,\dots,x_t$. Важно отметить, что для этого предсказателя погрешность асимптотически в два раза ниже, чем аналогичная ошибка для предсказателя Лапласа.

Для Марковских источников аналогичная (обобщённая) мера выглядит следующим образом:

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

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 Γ –гамма-функция. Данная мера является универсальной для множества Марковских источников связности m.

Мера R, универсальная для множества всех стационарных и эргодических источников, определяется следующим образом:

$$R(x_1,\dots,x_t) = \sum_{i=0}^{\infty} \omega_{i+1} K_i(x_1,\dots,x_t), \quad (3)$$

где множители ω_i являются некоторыми положительными весовыми коэффициентами, сумма которых равна 1:

$$\sum_{i=1}^{\infty} \omega_i = 1$$

Понятно, что слишком большие порядки в мере Кричевского должны иметь меньший вес и меньше влиять на прогноз, т.к. обнаружение длинных закономерностей с большей вероятностью окажется посторонним шумом. В результате, в качестве весовых коэффициентов было выбрано распределение (4). В общем случае весовые коэффициенты представляют собой варьируемый параметр метода и могут меняться, в зависимости от ряда и метода. В данной работе в качестве весовых коэффициентов было взято распределение вероятностей $\{\omega_i\}$, определяемое следующим образом:

$$\omega_i = 1/\log(i+1) - 1/\log(i+2) \quad (4)$$

В дальнейшем будем использовать именно это распределение.

Отметим, что мера R даёт оценку вероятностей для класса всех стационарных и эргодических источников на конечном алфавите, и будет использоваться для прогнозирования временных рядов, порождённых данным процессом.

3.2. Схема прогнозирования для источников из конечного алфавита

Рассмотрим схему прогнозирования на основе универсальной меры для источников, порождающих значения из конечного алфавита на примере меры R.

Вычисление меры R будет состоять из вычисления суммы (3) 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 (2) стремится к 0. Таким образом, существенное влияние как на значение ряда (2), так и на сложность его вычисления оказывает только первая часть суммы.

Рассмотрим стационарный и эргодический источник, порождающий значения из конечного дискретного алфавита A. И пусть также имеется порождённый данным источником временной ряд $x_1 \dots x_t$. Все значения $x_i \in A$, где A – некий конечный алфавит. Схема вычисления меры R достаточно проста. Для каждого $a \in A$ построим последовательность $x_1, \dots, x_t a$ и вычислим оценку условной вероятности на основе меры R следующим образом:

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

Полученные таким образом для каждого $a \in A$ величины и будут являться оценками соответствующих неизвестных условных вероятностей $P(x_{t+1} = a|x_1 \dots x_t)$.

3.3. Схема прогнозирования для источников из непрерывного интервала

На практике часто встречаются ряды, элементы которых являются числами из некоторого интервала. Таким образом, возникает естественная необходимость модификации описанного выше подхода для источника, принимающего значения из конечного непрерывного интервала. Описанные ранее результаты в этом направлении имеют преимущественно теоретический характер, а полученные экспериментальные данные относятся к случаю конечного алфавита.

Рассмотрим схему прогнозирования с использованием меры R для источника, принимающего значения из конечного непрерывного интервала. Пусть имеется стохастический процесс, генерирующий последовательность X_t, каждый элемент которой принимает значения из стандартного Борелевого пространства Ω, представляющего в нашем случае ограниченный 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 предлагаемом подходе разбиение интервалов производилось равномерно, т.е. на равные подинтервалы. Размер каждого подинтервала определяется, как $h = \frac{B-A}{n}$. Обоснование выбора именно такого метода будет дано ниже. В общем случае величины подинтервалов могут быть произвольными. 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 сигма-конечной мере L. В качестве L может выступать мера Лебега или какая-либо другая (в том числе счётная) мера.

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

$$p^s(x_1, \dots, x_n) = p({x_1}^{[s]}, \dots, {x_n}^{[s]})/L({x_1}^{[s]} \dots {x_n}^{[s]}) \quad (5)$$

Определим теперь оценку плотности вероятностей r следующим образом:

$$r(x_1 \dots x_t) = \sum_{s=1}^{\infty} \omega_s R({x_1}^{[s]} \dots {x_t}^{[s]}) / L({x_1}^{[s]} \dots {x_t}^{[s]}) \quad (6)$$

Коэффициенты ω_s определяются формулой (4) и несут роль весовых коэффициентов для случая каждого разбиения из Π_k. Как видно из формулы (6), в процессе вычисления меры r происходит нормировка каждого слагаемого (представляющего собой некоторую оценку вероятности последовательности при заданном разбиении с учётом фиксированного порядка меры) по сигма-конечной мере L. Таким образом, мы соединяем между собой оценки плотности вероятностей для случая различных возрастающих разбиений, что избавляет нас от зависимости результатов прогноза от конкретного разбиения. Описанный процесс соединения оценок плотностей вероятностей с нормировкой по L и умножением на весовые коэффициенты ω_s называется «склейкой».

Можно использовать для прогнозирования произвольные последовательности конечных разбиений. При этом какая-либо неравномерность разбиения не должна приводить к ухудшению прогнозов. Экспериментальным путём выяснили, что

равномерная квантизация даёт, как правило, наилучшую точность прогноза на реальных действительных рядах, потому она и была выбрана в качестве основной.

Величина $r(x_1, \dots, x_t)$ является оценкой неизвестной плотности вероятности $p(x_1, \dots, x_t)$, а соответствующая условная плотность

$$r(a|x_1, \dots, x_t) = r(x_1, \dots, x_t a)/r(x_1, \dots, x_t) \quad (7)$$

является оценкой плотности $p(a|x_1 \dots x_t)$. Количество слагаемых в сумме (6) при реализации описанного далее алгоритма, как и в случае источника с конечным алфавитом, определяется длиной выборки $t + 1$ (первые t слагаемых из первой части суммы и одно слагаемое из второй части суммы).

3.4. Оптимизация алгоритма вычисления R-меры

Оценим трудоёмкость описанного подхода к определению следующего прогнозного значения. Под трудоёмкостью (или сложностью) работы алгоритма будем понимать максимально возможное число элементарных операций в процессе его работы, оценивая это число асимптотически (используя O-символику). Трудоёмкость определения прогнозного элемента состоит из оценки трудоёмкости вычисления меры R, умноженной на параметр разбиения $(n + 1)$, где n – число подинтервалов в разбиении рассматриваемого непрерывного интервала для случая источника, порождающего значения из непрерывного интервала, и мощность алфавита для случая источников, порождающих значения из конечного алфавита. В свою очередь, вычисление меры R состоит из произведения параметра длины ряда t, умноженного на трудоёмкость вычисления K_m. Исходя из определения весовых коэффициентов (4), видно, что с ростом i значение коэффициента w_i стремится к нулю и является небольшим при больших i (при $i > 5w_{i+1} < 0.05$). Вклад слагаемого $\omega_{i+1}K_i(x_1 \dots x_t)$ с ростом i будет небольшим. Соответственно, в целях уменьшениях трудоёмкости вычислений, количество слагаемых в сумме (4) можно ограничить каким-либо параметром m, где $m = 1, \dots, t$. Назовём этот параметр глубиной анализа метода.

Таким образом, нам требуется оценить сложность вычисления выражения (2) при фиксированном параметре m. Его сложность равна следующей величине:

$$T(K_m) = 4 \cdot m \cdot t \cdot n^{m+1} = O(m \cdot n^{m+1})$$

Исходя из этого, сложность вычисления прогнозного значения всей вероятности будет равна:

$$T(R) = (n + 1) \cdot \sum_{k=1}^{t} 4 \cdot k \cdot t \cdot n^{k+1} = O(t \cdot m^2 \cdot n^{t+2})$$

Предположим, что мы вычислили меру R для последовательности $x_1, \dots, x_t$ на первом этапе и запомнили при этом все частоты $\nu_x(\vartheta)$ для каждого набора ϑ и каждого порядка m. Заметим, что $K_m(x_1, \dots, x_t a)$ отличается от $K_m(x_1, \dots, x_t)$ лишь тем, что к одному множителю во внутреннем произведении формулы (3) к частоте ϑa прибавится 1 (т.к. добавится проверка вхождения ϑa ещё в одной последовательности: самой последней, состоящей из $t + 1$ элемента). В знаменателе произойдут ровно те же изменения: к одному члену суммы $\sum_{a \in A} \nu_x(\vartheta a)$ добавится единица. А в силу свойства гамма-функций: $\Gamma(k + 1) = k \cdot \Gamma(k)$. Соответственно, вычисление оператора $K_m(x_1, \dots, x_t a)$ может быть записано следующим образом:

$$K_m(x_1, \dots, x_t a) = \frac{1}{|A|^m} \prod_{\vartheta \in A^m} \frac{\prod_{a \in A}(\Gamma(\nu_x(\vartheta a)+1/2)/\Gamma(1/2)))}{(\Gamma(\overline{\nu}_x(\vartheta)+|A|/2)/\Gamma(|A|/2))} = K_m(x_1 \dots x_t) \cdot \frac{\nu_x(\vartheta a)+1/2}{\sum_{a \in A} \nu_x(\vartheta a)} \quad (9)$$

Так как мы при первом вычислении меры R запомнили все частоты $\nu_x(\vartheta)$, то вычисление меры $R(x_1, \dots, x_t a)$ будет происходить за время $O(n)$.

В итоге, трудоёмкость подсчёта прогнозного элемента сократится до следующей величины:

$$T(R) = \sum_{k=1}^{m} 2 \cdot k \cdot t \cdot n^{k+1} + n \cdot m = O(t \cdot m^2 \cdot n^{t+1})$$

Исходя из данного соотношения, видно, что сложность вычислений уменьшилась в $(2 \cdot n)$ раз, что при достаточно больших n (т.е. при большом разбиении интервала) будет существенно влиять на время вычислений.

3.5. 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 временных 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. 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 для признака q определяется по следующей формуле:

$$Gain(A,q) \; = \; H(A,S) \; - \sum_{i=1}^{q_n} \frac{|A_{q_i}|}{N} H(A_{q_i}, S).$$

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-либо a на всех элементах из A, поставить в корень метку a и выйти.
   Вероятность каждого символа a будет при этом определяться следующим образом: $P(x_{t+1} = a | x_1, x_2, \dots, x_t) = 1.0;\ P(x_{t+1} = b \neq a | x_1, x_2, \dots, x_t) = 0.0$
3. Если $Q = \{\emptyset\}$, то выбрать такое a из множества значений S, которому равно наибольшее число элементов из A, поставить a в корень и выйти.
   Вероятность каждого прогнозного символа a будет определяться следующим образом: $P(x_{t+1} = a | x_1, x_2, \dots, x_t) = A_a / |A|$
4. Выбрать $q \in Q$, для которого $Gain(A,q)$ 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 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. Также, в предлагаемой реализации данного алгоритма можно ограничивать глубину

дерева искусственно – отдельным параметром. После достижения глубины дерева заданной глубины, выполняется пункт 3 алгоритма ID3.

Опишем разработанную методику применения данного подхода для случая прогнозирования элементов временного ряда. Пусть дан временной ряд $x_1, \dots, x_t$, где $x_i \in A$, A – конечный алфавит возможных значений элементов ряда. Требуется спрогнозировать значение элемента $x_{t+1} \in A$. Пусть также есть параметр метода m, определяющий максимальную глубину дерева. Параметр m будет определять число признаков у каждого элемента (и соответственно, максимальную глубину дерева). Далее, определим множество A по правилу: в качестве последнего– целевого – признака возьмём какое-то i-ое значение ряда x_i, а в качестве его $(m-1)$ атрибутов примем $(m-1)$ значений, стоящих в ряду перед i-ым, т.е. элементы $x_{i-m+1}, x_{i-m+2}, x_{i-m+2}, \dots, x_{i-1}$. 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 m 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.

3.6. Метод усреднения алфавита

Рассмотрим проблему выбора прогнозного значения. Пусть имеется ряд $x_1, x_2, \dots, x_t$ и какая-то оценка распределения вероятностей p для элемента x_{t+1}. В случае, если выбирать в качестве прогнозного элемента середину интервала, который имеет максимальную вероятность, возникает следующая проблема. Пусть имеется 2 соседних подинтервала с очень близкими и высокими вероятностями. Тогда возникает задача выбора одного из них. Поэтому для уменьшения величины ошибки прогноза предлагается выбирать в качестве прогнозного значения не середину подинтервала, имеющего максимальную вероятность, а считать математическое ожидание от всех подинтервалов. Таким образом, вычисление прогнозного значения будет сводиться к вычислению следующего соотношения:

$$x_{t+1} = \sum_{i=1}^{n} p_i \cdot k_i,$$

где p_i – вероятность i-го подинтервала, n – величина разбиения (число подинтервалов), k_i – середина i-го подинтервала. Таким образом, мы будем учитывать не только одну максимальную вероятность из всех подинтервалов, а вероятности всех подинтервалов, выбирая вероятностное среднее. Прогнозные элементы теперь не будут ограничены дискретным набором действительных величин, а смогут принимать любое значение из интервала прогнозирования.

3.7. Метод группировки алфавита

При прогнозировании временных рядов, порождённых источником, принимающем значения из непрерывного интервала, возникает проблема подбора оптимальных значений параметров разбиения (число n). Понятно, что чем меньше будет разбиение, тем ниже будет точность получаемых прогнозов, т.к. мы определяем только интервал, которому принадлежит x_{t+1}, и даже при существовании и выявлении методом определённой закономерности, точность будет оставаться в пределах величины длины подинтервалов. Соответственно, для увеличения точности надо уменьшать подинтервалы и увеличивать разбиение, но тогда возникает другая проблема: частота каждого подинтервала (символа алфавита) становится очень маленькой (в особенности при небольших размерах ряда),

многие символы могут иметь нулевую частоту. В разделе 3.4 была описана вычислительная сложность описываемого метода. Даже в оптимизированном варианте она остаётся достаточно высокой и растёт в зависимости от величины разбиения, как полином высокой степени $(m+1)$. Для решения двух описанных проблем в данной работе предлагается использовать так называемый метод группировки алфавита.

Опишем суть работы разработанного метода группировки алфавита.

Пусть дан алфавит A элементов временного ряда $x_1, \dots, x_N$. Пусть также есть некоторое разбиение алфавита A на $N1$ непересекающихся подмножеств: $B_1, B_2, \dots, B_{N1}$. Тогда под фильтрацией ряда $\mathrm{x}_1, \dots, \mathrm{x}_\mathrm{N}$ по прогнозному значению B_j назовём процесс выбора элементов данного ряда по следующему правилу. Идём от начала ряда до его последнего элемента и на каждом шаге определяем, оставить текущий элемент или же удалить его из ряда: если очередной $x_i \in B_j$, то элемент $\mathrm{x_i}$ оставляем, иначе – удаляем его из ряда. Выберем некоторые рекуррентные подразбиения заданного алфавита A по следующему алгоритму:

- Разобьём множество A на $N1$ непересекающихся подмножеств: $B_1, B_2, \dots, B_{N1}$, где $N1 << N$ и каждый B_i содержит один или несколько элементов из A.
- Каждое полученное подмножество B_i разобьём ещё на $N2$ частей и получим в общем итоге $N1 * N2$ подмножеств, содержащих элементы множества A.
- Продолжаем данный рекуррентный процесс до получения заданного числа подразбиений. В дальнейшем мы будем рассматривать случай только одного разбиения алфавита A, т.е. имеем алфавит B из $N1$ подмножеств множества A.
- Записываем исходный временной ряд в терминах алфавита B (т.е. все элементы исходного ряда преобразуются в соответствующие символы из алфавита B) и прогнозируем соответственно элемент из алфавита B. Обозначим его, как B_i.
- Фильтруем исходный ряд (в терминах алфавита A) по прогнозному значению B_i, т.е. оставляем в нём только те элементы из A, которые принадлежат множеству B_i.
- Далее, если подразбиение было не одно, то записываем ряд в терминах алфавита B_i (т.е. алфавита уже второго уровня). И продолжаем вышеописанный процесс рекуррентно до достижения последнего уровня подразбиения.
- Прогнозируем новый (отфильтрованный и уменьшенный) ряд в обычном режиме (в терминах исходного алфавита A).

Приведём пример работы алгоритма. Пусть дан алфавит $A = \{\mathrm{i}\}, \text{где } \mathrm{i} = \overline{1, \dots, 12}$. И пусть дан временной ряд: $\mathrm{X}(A)$: 1, 3, 5, 5, 6, 7, 8, 1, 3, 5, 5, 6, 7, 8, 1, 3, 5. Требуется предсказать следующий элемент. Разобьём исходный алфавит на 4 равных части, определив тем самым новый алфавит B: $\{B_i\}, \mathrm{i} = \overline{1, \dots, 4}$. 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 3 элемента из множества A, которые и будут образовывать сгруппированный алфавит $\mathrm{B_{i,j}}$. В итоге, каждому A_i будет однозначно соответствовать элемент $\mathrm{B_{i,j}}$.

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

Далее осуществим обработку исходной последовательности по следующему правилу: если элемент $\mathrm{X_i}(A)$ лежит во множестве B_2, то оставляем его в ряду, иначе – удаляем. Получим следующий ряд: 5, 5, 6, 5, 5, 6, 5. В нём присутствует только 2 символа алфавита из 12 (т.к. 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 3 элементов (4 переходит в 1, 5 – в 2, 6 – в 3): 2, 2, 3, 2, 2, 3,

2. Далее просто определяем прогнозное значение в полученном ряде и приводим его к исходному алфавиту. Очевидно, что прогнозное значение равно 2, которое соответствует числу 5 в исходном алфавите. Символ исходного алфавита 5 и будет являться результатом прогнозирования.

3.8. Склейка методов прогнозирования

В современное время, как уже было сказано, существует достаточно большое множество методов прогнозирования, эффективность которых весьма разнится в зависимости от конкретной ситуации и конкретного процесса, который требуется спрогнозировать. В определённых областях науки, техники и экономики часто существуют хорошо работающие методы прогнозирования, которые разрабатываются для прогнозирования рассматриваемых процессов с учётом их особенностей и существующих в данной сфере закономерностей. В этом случае возникает задача выбора и применимости различных методов: какие-то лучше работают на примере одних процессов, какие-то лучше прогнозируют другие типы процессов. Предлагаемые подходы на основе универсальной меры и решающих деревьев достаточно универсальны и не привязаны к конкретной реализации процесса. Однако их эффективность и применимость также ограничены, и проблема получения высокого качества прогнозов в зависимости и типа ряда и природы процесса существует. Предлагается решение данной задачи посредством использования так называемой «склейки методов». Поясним суть данного подхода.

Пусть имеется временной ряд $x_1, \dots, x_t$, где $x_i \in A$, A – конечный алфавит возможных значений элементов ряда. Требуется спрогнозировать одно или несколько значений после элемента $x_t \in A$. Для определённости будем рассматривать случай прогнозирования одной величины $x_{t+1} \in A$. И пусть также имеется несколько методов прогнозирования, определяющих вероятностное распределение вероятностей величины x_{t+1}. Обозначим их следующим образом:

$$M_i(x_{t+1} = a | x_1, \dots, x_t),$$

где a – предполагаемое прогнозное значение, а значение функции M_i– соответствующая условная вероятность. Всего будем считать, что у нас имеется n методов. Для соединения данных методов можно воспользоваться следующим соотношением:

$$P(x_{t+1} = a | x_1, \dots, x_t) = \sum_{i=1}^{n} k_i \cdot M_i(x_{t+1} = a | x_1, \dots, x_t), \quad (8)$$

где k_i – весовые коэффициенты, представляющие собой степень значимости i-го метода. Для k_i в целях сохранения корректности метода должно выполняться следующее соотношение: $\sum_{i=1}^{n} k_i = 1$. Важно отметить, что после получения итогового распределения вероятностей для вычисления прогнозного значения можно применить метод усреднения (т.е. взятие мат. ожидания). Это позволит учесть в прогнозе всё распределение и от всех методов (т.е. больше информации), а не только один элемент с максимальной вероятностью.

Таким образом, мы можем использовать при прогнозировании различные методы и, варьируя параметры k_i, можем определять большую значимость для тех методов, которые на основе предыдущей статистики лучше работают на данном конкретном процессе.

3.9. Прогнозирование поведений

При прогнозировании реальных процессов часто возникает проблема отсутствия в данных рядах каких-либо явных закономерностей и даже распределений в явном виде. Зачастую данные ряды не являются стационарными или эргодическими, и применимость к ним большинства существующих методов довольно сильно ограничена. В таких случаях можно пытаться спрогнозировать не конкретные вещественные значения ряда, а их поведение, т.е. направление движения значений процесса (т.е. его тренда): вверх, вниз или

останется на текущем уровне. При этом фактически мы будем прогнозировать дискретный временной ряд с алфавитом, состоящим из всего трёх элементов. В случае прогнозирования сложных временных рядов, данный подход зачастую себя оправдывает.

Рассмотрим реализацию данного подхода. Пусть имеется вещественный временной ряд $x_1, \dots, x_t$, и требуется спрогнозировать следующий элемент x_{t+1}. Определим два значения A, B, которые определяют нижнюю и верхнюю границу для элемента ряда, при нахождении в которых он будет подходить под определение «остаётся на текущем уровне».

Алгоритм прогнозирования значений в этом случае будет следующим:

1. Преобразуем исходный ряд $x_1, \dots, x_t$ в ряд $x'_1, \dots, x'_t$, где $x'_i = f(x_i)$, функция f определяется следующим образом:
$$f(x_i) = \begin{cases} 0, \text{если } x_i < A \\ 1, \text{если } x_i \in [A, B] \\ 2, \text{если } x_i > B \end{cases}$$
2. Прогнозируем каким-либо методом значение x'_{t+1} ряда $x'_1, \dots, x'_t$.
3. Выполняем обратное преобразование спрогнозированного значения x'_{t+1} по следующему правилу: выбираем какое-либо (в общем случае произвольное) значение из соответствующего интервала: $[A', A], [A, B], [B, B']$, где A' и B' – минимальное и максимальное значение из всех элементов временного ряда $x_1, \dots, x_t$.

Таким образом, мы получили некоторый инструмент для прогнозирования сложных временных рядов, в которых нет явных закономерностей.

3.10. Подход на основе многомерного прогнозирования

Достаточно очевидным является факт взаимосвязи различных реальных процессов, происходящих в мире. К примеру, внутренний валовый продукт оказывает влияние на курс валюты рассматриваемой страны, а показатель уровня жизни – на индекс потребительских цен. Все эти показатели и процессы зачастую представляют собой отдельные временные ряды, которые нам также известны. Если бы можно было учесть корреляции хотя бы некоторого ограниченного набора временных рядов, то мы смогли бы существенно повысить точность и эффективность получаемых прогнозов. Пример наличия простой взаимосвязи между рядами: при увеличении значений одного временного ряда всегда происходит увеличение значений другого временного ряда. Конечно, такие влияния могут быть «запоздалыми» или наоборот «спешащими», но существующая корреляция между разными временными рядами позволяет получить дополнительную информацию о том временном ряде, который мы хотим спрогнозировать.

Таким образом, качество прогнозирования временных рядов может быть существенно увеличено с использованием так называемого многомерного подхода, при котором в прогнозе учитываются другие временные ряды. Ранее, методов, учитывающих сразу несколько различных коррелирующих временных рядов, не было.

В данной работе предлагается подход, который позволяет учесть при прогнозировании одного временного ряда другой временной ряд, который коррелирует с первым. Важно отметить, что данный подход не зависит от используемого метода (алгоритма) прогнозирования. В качестве основы мы можем использовать любой математический метод прогнозирования стационарных и эргодических источников.

Пусть имеется K временных рядов, коррелирующих каким-то образом между собой:

$$x_1^1, x_2^1, x_3^1, \dots, x_t^1$$
$$x_1^2, x_2^2, x_3^2, \dots, x_t^2$$
$$\dots$$
$$x_1^K, x_2^K, x_3^K, \dots, x_t^K$$

При этом мы предполагаем, что все K временных рядов определены на одной и той же оси времени (с едиными начальными и конечными точками времени) и записаны в квантизованном виде (т.е. в виде номеров подинтервалов). Также у них одинаковая квантизация (разбиение). Нам требуется спрогнозировать следующий элемент первого ряда, т.е. элемент x_{t+1}^1. Построим временной ряд $(K+1)$ на основе первых K по правилу:

$$x_i = x_{l+i}^1 + x_i^2 \cdot N + x_i^3 \cdot N^2 + \cdots + x_i^K \cdot N^{K-1}$$

где N – мощность алфавита (разбиения), а l – сдвиг первого ряда назад относительно оставшихся $(K-1)$ рядов, $i = 1, \dots, t-l$. Сдвиг нужен для учёта отстающей по времени корреляции целевого ряда относительно других. Вышеприведённая формула является полиномиальных хешем от рассматриваемых K временных рядов (с учётом сдвига первого ряда). Далее осуществляем прогноз $(K+1)$-го ряда каким-либо классическим (в общем случае произвольным) методом прогнозирования с учётом суженного диапазона возможных значений (алфавита) элемента x_{t-l+1}. Суженный диапазон значений представляет собой целочисленное множество $A' = \{a | (x_{t+1}^2 \cdot N + x_{t+1}^3 \cdot N^2 + \cdots + x_{t+1}^K \cdot N^{K-1}) \le a \le (x_{t+1}^2 \cdot N + x_{t+1}^3 \cdot N^2 + \cdots + x_{t+1}^K \cdot N^{K-1} + N - 1)\}$. Далее, по полученной плотности вероятности элемента x_{t+1-l} восстановим плотность вероятности элемента x_{t+1}^1 по правилу:

$$p(x_{t+1}^1 = a \in A | x_1, x_2, \dots, x_t) = C \cdot p(x_{t-l+1} = b \in A' | x_1, x_2, \dots, x_t),$$

где $a = b \bmod N$, а C – нормирующий коэффициент. Оценка функции плотность вероятности строится по оценке условных вероятностей в виде соответствующей ступенчатой функции.

В случае, когда сдвиг l равен нулю, значения всех K временных рядов в момент времени $(t+1)$ неизвестны, поэтому целочисленное множество A' будет представлять собой множество всех возможных значений ряда $x_1, \dots, x_t$: $A' = \{a | 0 \le a \le N^K - 1\}$.

В простейшем случае можно соединить всего два коррелирующих между собой временных ряда. При этом вопрос поиска коррелирующих между собой рядов и определение оптимального сдвига l остаётся задачей исследователя. Однако при этом важно отметить, что приведённые далее экспериментальные данные показали, что при выборе не коррелирующих между собой рядов, точность получаемого прогноза остаётся на том же уровне, что и случае классического (не многомерного) прогнозирования одного ряда. Также важно отметить, что существуют и другие способы соединения (слияния) временных рядов в один ряд. Например, можно соединять два временных ряда по принципу чередования значений одного и другого. Однако, как показали приведённые ниже экспериментальные результаты, описанная выше методика соединения рядов является наиболее эффективной с точки зрения точности получаемых прогнозов.

Таким образом, для увеличения точности прогноза нам надо найти такие временные ряды, которые имеют ненулевую корреляцию между собой. Но этого мало. Известно, что некоторые процессы, не являющиеся абсолютно случайными, могут влиять на другие ряды с некоторым запозданием или наоборот опережением. Для нас важно найти такой временной ряд, который бы влиял на первый с опережением, т.е. такой, у которого факты увеличения / уменьшения значений, периоды, какие-либо ещё закономерности происходят раньше, чем у другого временного ряда. Только при соблюдении описанных условий мы сможем получить существенное увеличение эффективности работы выбранного метода прогнозирования.

Важно также отметить, что в данную схему можно ввести параметр сдвига временных рядов относительного первого. Данный параметр будет определяться в зависимости от предполагаемого среднего уровня опережения рассматриваемого временного ряда относительно первого временного ряда.

3.11. Методика экспериментальных исследований

Метод на основе R-меры был реализован с учётом оптимизации, предложенной в разделе 3.4, и был протестирован на прогнозах реальных данных. Все прогнозы осуществлялись с применением метода группировки алфавита.

Все эксперименты проводились в двух режимах. Первый режим – on-line – означает прогнозирование значений временного ряда на 1 шаг вперёд (т.е. нахождение элемента x_{t+1}). Второй режим – на 10 или 20 шагов вперёд – обозначает прогнозирование значений ряда на 10 последовательных шагов вперёд. При этом для уменьшения трудоёмкости работа во втором режиме осуществлялась по следующему алгоритму. Вначале осуществляем прогноз на 1 элемент текущего ряда и запоминаем 5 элементов с наиболее высокими вероятностями. Далее, добавляем к каждому из этих 5 элементов все возможные элементы алфавита, после чего прогнозируем комбинацию сразу из 2 элементов ряда. Всего на данном этапе получим $(n + 5 * n)$ прогнозов, где n – число разбиения ряда. Далее, осуществляем прогноз заданных пар значений и запоминаем 3 пары с наивысшими вероятностями, после чего добавляем к заданной паре ещё n возможных элементов ряда. В общем итоге, получаем $(n + 5 \cdot n + 3 \cdot n) = 9 \cdot n$ прогнозов и сразу 3 прогнозных элемента. Далее осуществляем прогнозирование ещё на 3 шага вперёд по описанному выше методу и так далее продолжаем до получения заданного числа шагов (10-20 в нашем случае), после чего считаем погрешность прогноза. Все прогнозы выполнялись на 10 выборках одного ряда с различным сдвигом по временной оси, в итоговую таблицу вносилась усреднённая ошибка.

Важно отметить, что иногда выполнялась предварительная обработка ряда, суть которой заключается в следующем. Из исходного ряда $x_1, \dots, x_t$ мы получаем ряд $y_1, \dots, y_{t-1}$ по принципу: $y_i = x_{i+1} - x_i$, т.е. ряд разниц между соседними элементами. Прогнозы осуществлялись именно на обработанном ряде, и в результате работы алгоритма получались разницы между прогнозируемым и последним элементом ряда. Прогнозное значение x_{t+1} получалось посредством прибавки спрогнозированной разницы y_t к последнему элементу ряда x_t. 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, что было невозможно при прогнозировании абсолютных величин временного ряда. Определение границ интервала осуществляем естественным образом: считаем величину максимальной и минимальной (с учётом знака) разницы между соседними элементами; берём полученные значения в качестве левой и правой границ интервала, который далее и разбивался на n частей. При этом величиной $delta$ будем называть максимальную разницу между двумя соседними элементами, т.е. фактически разницу между верхней и нижней границей интервала.

В описанных ниже результатах используются следующие параметры: размер выборки (L), количество частей разбиения непрерывного интервала (n), параметр глубины анализа (m) и величина ошибки прогноза для двух режимов работы метода. При этом таймфреймом будем называть временной интервал между двумя соседними значениями ряда (т.е. время между измерениями).

4. Полученные результаты:

4.1.Прогнозирование периодических функций

Рассмотрим на первом этапе прогнозирование самых простых временных рядов – рядов значений периодических функций. Возьмём функцию $f(x) = \frac{\sin(x)+\cos(3x)}{2}$. При этом будем прогнозировать разницы между соседними элементами. Разбиение N будет являться задаваемым параметром, одновременно размер алфавита будет задаваться отдельным независимым параметром n. Величина $delta$ определяется величиной разбиения и потому

содержится в таблице с результатами. Размер выборки будет равен 2.5 периода рассматриваемой функции. Результаты приведены ниже в таблице 1.

Таблица 1. Прогнозирование периодической функции.

Разбиение N	Алфавит n	Глубина анализа m	$delta$	R-метод on-line	R-метод 10 шагов
30	10	2	0,197	0,0111	0,1201
80	10	2	0,0739	0,0021	0,0541
30	20	2	0,197	0,0075	0,1158
120	20	2	0,049	0,0006	0,0172
60	40	2	0,099	0,0017	0,0093
120	40	2	0,049	0,0007	0,0038
240	80	2	0,025	0,0002	0,00081
400	80	2	0,014	0,00007	0,00135

В случае верного определения прогнозного элемента каким-либо методом ошибка прогноза не должна превышать размера одного подинтервала, который фактически определяется, как $delta/n$. Как видно из приведённых в таблице 1 результатов, в случае, когда разбиение N (т.е. длина ряда) достаточно большие (N> 60), ошибка прогноза никогда не превосходит величину $delta/n$, что говорит о точном выявлении существующей в рассматриваемом временном ряду закономерности. В случае, когда длина ряда небольшая, и разбиение сопоставимо по величине с размером алфавита, метод даёт чуть худшие результаты, что видно по ошибкам прогноза метода в режиме на 10 шагов вперёд. Отсюда можно сделать вывод о том, что метод R хорошо выявляет существующие в ряду периодические закономерности в случае, когда длина ряда достаточно велика (имеется 3-4 или более периодов). Также видно, что для устойчивого определения периода нужно брать разбиение, не меньшее, чем мощность алфавита (параметр n метода). В дальнейшем будем придерживаться данного правила.

Рассмотрим результаты прогнозирования решающих деревьев на этом же примере, но для различных значений длины ряда в периодах, и сравним два метода. Данные результаты прогнозирования для режимов on-line и на 10 шагов вперёд приведены в таблице 2 и 3, соответственно.

Таблица 2. Прогнозирование периодической функции методами R и решающих деревьев. Режим on-line.

Размер выборки L	Глубина анализа m	Разбиение n	Решающие деревья	R-метод
1 период	2 / 2	10	0,0074	0,0074
1 период	2 / 5	10	0,0091	
1 период	2 / 5	20	0,0064	0,0037
2 периода	2 / 2	10	0,0074	0,0074
2 периода	5 / 5	10	0,0074	0,0074
2 периода	2 / 5	20	0,0037	0,0037
2 периода	2 / 2	100	0,01141	0,01986

Таблица 3. Прогнозирование периодической функции методами R и решающих деревьев. Режим на 10 шагов вперёд.

Размер выборки L	Глубина анализа m	Разбиение n	Решающие деревья	R-метод

1 период	2 / 2	10	0,03856	0,03856
1 период	2 / 5	10	0,07665	
1 период	2 / 5	20	0,06116	0,00373
2 периода	2 / 2	10	0,03856	0,03856
2 периода	5 / 5	10	0,00742	0,03856
2 периода	2 / 5	20	0,00373	0,00373
2 периода	2 / 2	100	0,03959	0,06779

Исходя из приведённых результатов, можно увидеть, что на коротких выборках (длиной в 1 период) R-метод в среднем даёт результаты лучше, чем решающие деревья. В случае же длиной выборки в 2 периода, решающие деревья дают примерно те же результаты, в редких случаях – лучше, чем деревья. Объясняется данное наблюдение тем, что R-метод лучше, чем деревья, выявляет периодические закономерности и в случае существования периодов в явном виде сразу их определяет. Деревьям же требуется больше времени для определения периода, но тем не менее они его тоже успешно выявляют. Данные выводы справедливы как для случая работы методов в on-line режиме, так и для случая прогнозирования на несколько шагов вперёд.

Некоторое преимущество решающих деревьев состоит в выявлении в ряду более сложных, нежели периодические, закономерностей. Приведём пример. Пусть имеется ряд:
0 1 0 1 1 0 1 2 1 1 3 0 1 0 3 2 1 0 1 1 1 1 1 2 1 3 1 1 3 1 0 1 1 2 1 2 1 1 3 3 3 0 3 1 3 1 1 3 2 1 0

Суть закономерности в данном ряду заключается в следующем. Если в последовательности из пяти элементов на 2-ой и 4-ой позиции стоят единицы, то 5-ая цифра равна 1; если в последовательности из 4 элементов первая цифра 3, то 4-ая цифра равна 0. Решающие деревья успешно выявляет оба типа закономерностей; метод R в данном случае будет давать переменные результаты.

4.2. Прогнозирование цен на энергоносители в США

Рассмотрим результаты прогнозирования цен на топливо в США. В таблице 2 приведены результаты прогнозирования данного ряда с таймфреймом 1 неделя в период с 01.01.2002 по 01.10.2013. Длина ряда равна 615 элементам. Значение $delta$ для ряда цен на энергоносители равняется 0.832. График ряда приведён на рисунке 1.

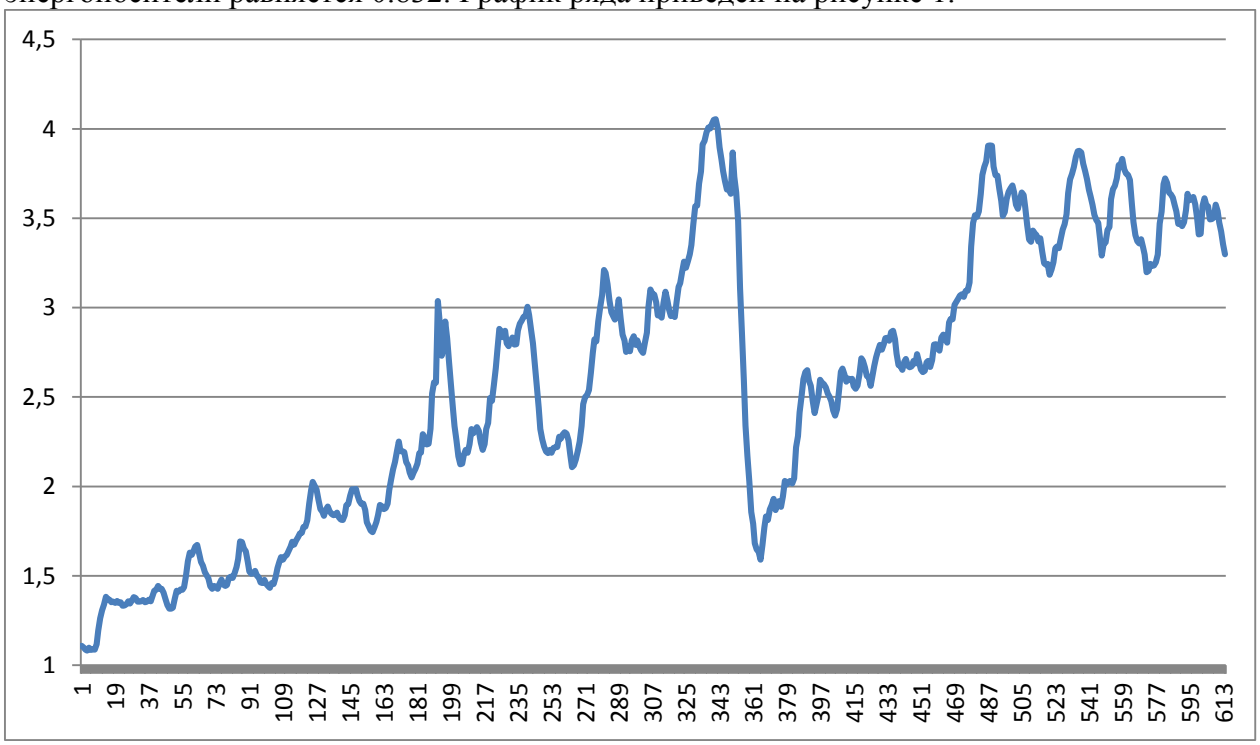

Рис. 1. График цен на энергоносители (таймфрейм: 1 неделя)

В таблице 2 приведены результаты прогнозирования данного ряда в 3 режимах: простой режим on-line, режим on-line с применением метода усреднения алфавита, а также режим на 20 шагов вперёд с применением метода усреднения алфавита. В таблице 3 приведены аналогичные данные прогноза для решающих деревьев в двух режимах. Также, во всех прогнозах применялся метод группировки алфавита. В процессе группировки разбиение осуществлялось на глубину алфавита 1, т.е. алфавит разбивался на определённое количество групп всего 1 раз. При этом здесь и далее разбиение по группам осуществлялось равномерно, т.е. на группы равного размера, обозначения разбиения при этом вводились следующие: $n\ (a * b)$, где n – стандартная величина разбиения интервала (равная мощности алфавита), a – количество групп в группировке алфавита. Параметр b обозначает количество элементов в каждой группе. Нетрудно понять, что всегда будет верно равенство $n = a * b$ (значения параметра a подбирались так, чтобы это было верно). Также, в ячейках ошибки прогноза приводилось 2 значения: до черты «/» указана обычная ошибка прогноза, после черты – ошибка прогноза, разделённая на величину $delta$.

Таблица 2. Прогнозирование цен на энергоносители США. R-метод.

Разбиение n	Глубина анализа m	R-метод. Усреднение. on-line	R-метод on-line	R-метод. Усреднение. 20 шагов
5	5	0.05211 / 0.0626	0.07260	0.44139 / 0.5305
10	2	0.04790 / 0.0575	0.04658	0.57510 / 0.6912
	5	0.04790 / 0.0575	0.04658	0.57510 / 0.6912
20	2	0.04971 / 0.0597	0.05202	0.29762 / 0.3577
	5	0.04971 / 0.0597	0.05202	0.29762 / 0.3577
50	2	0.04609 / 0.0554	0.03915	0.12638 / 0.1519
	5	0.04609 / 0.0554	0.03915	0.12638 / 0.1519

Таблица 3. Прогнозирование цен на энергоносители США. Решающие деревья.

Разбиение n	Глубина анализа m / макс. глубина дерева	Решающие деревья on-line	Решающие деревья 20 шагов
5	5	0.07192 / 0.0864	0.5457 / 0.6558
10	2	0.04658 / 0.0559	0.1089 / 0.1309
	5	0.04728 / 0.0568	0.1089 / 0.1309
20	2	0.05618 / 0.0675	0.3689 / 0.4434
	5	0.06476 / 0.0778	0.6393 / 0.7683
50	2	0.052996 / 0.0637	0.22954 / 0.2759
	5 / 2	0.049312 / 0.0593	0.05918 / 0.0711

Из приведённых результатов видно, что ошибка прогноза данных методов находится в пределах 1/25 от величины $delta$, что говорит о достаточно высокой точности предлагаемого подхода. При этом она слабо зависит от используемого разбиения при разбиении n от 10 элементов. Зависимости результатов от глубины анализа не наблюдается никакой. Также, можно заметить, что метод усреднения даёт ощутимое преимущество только при разбиении $n = 5$: получаемая ошибка получается сравнимой с разбиением $n = 10$. Решающие деревья дают чуть лучшие, чем R-метод, результаты при малых разбиениях (10-20 элементов) и чуть лучшие на более высоких разбиениях. В целом же методы дают сравнимую друг с другом точность прогнозов. Рассмотрим прогнозирование других временных рядов и проверим сделанные выводы.

4.3. Прогнозирование цен на энергоносители с использованием склейки методов

Рассмотрим склейку R-метода и решающих деревьев для случая прогнозирования ряда, рассмотренного в предыдущем разделе. Для этого введём дополнительный параметр

w, являющийся коэффициентом значимости первого метода, т.е. k_0 в формуле (9). При этом под первым методом будем иметь в виду R-метод. Соответственно, итоговая вероятность каждого символа a будет определяться следующим соотношением:

$$P(x_{t+1} = a|x_1, \dots, x_t) = w \cdot R(a|x_1, \dots, x_t) + (1 - w) \cdot DT(a|x_1, \dots, x_t),$$

где $DT(a|x_1, \dots, x_t)$ – вероятность $p(x_{t+1} = a|x_1, \dots, x_t)$ для решающего дерева. Таким образом, мы вычисляли распределение вероятностей для каждого из методов, потом склеивали данные вероятности и затем вычисляли мат. ожидание от склеенных вероятностей. В нижеследующей таблице 4 приведены данные прогнозирования ряда цен на энергоносители США с использованием метода склейки при различных параметрах коэффициента w. Прогнозирование осуществлялось в режиме on-line. При этом в конце работы метода использовалось усреднение (в качестве прогнозного значения бралось мат. ожидание от итогового распределения). Полученные результаты приведены в таблице 4.

Таблица 4. Прогнозирование цен на энергоносители США. Склейка.

Разбиение *n*	Глубина анализа *m*	Решающие деревья On-line	R-метод On-line	Параметр *w*	Склейка методов
5	5	0.08022	0.07260	0,2	0.07268
				0,5	0.06246
				0,7	0.05832
10	2	0.04658	0.04658	0,2	0.04684
				0,5	0.04724
				0,7	0.04751
20	2	0.05618	0.05202	0,2	0.5475
				0,5	0.5261
				0,7	0.5134

Из приведённых в таблице 4 результатов видно, что склейка двух методов во всех случаях даёт результаты не хуже, чем худший из двух методов. Во многих случаях она превосходит по точности прогноза оба метода, взятые в отдельности, что говорит о высокой эффективности представленной модификации.

4.4. Прогнозирование объёма промышленного производства США

Рассмотрим результаты прогнозирования объёмов промышленного производства в США. Для этого рассмотрим ряд данного процесса с таймфреймом 1 месяц в период с 01.1970 по 09.2013. Длина ряда равна 525 элементам. Значение *delta* для ряда цен на энергоносители равняется 7.2945. График данного ряда приведён на рисунке 2.

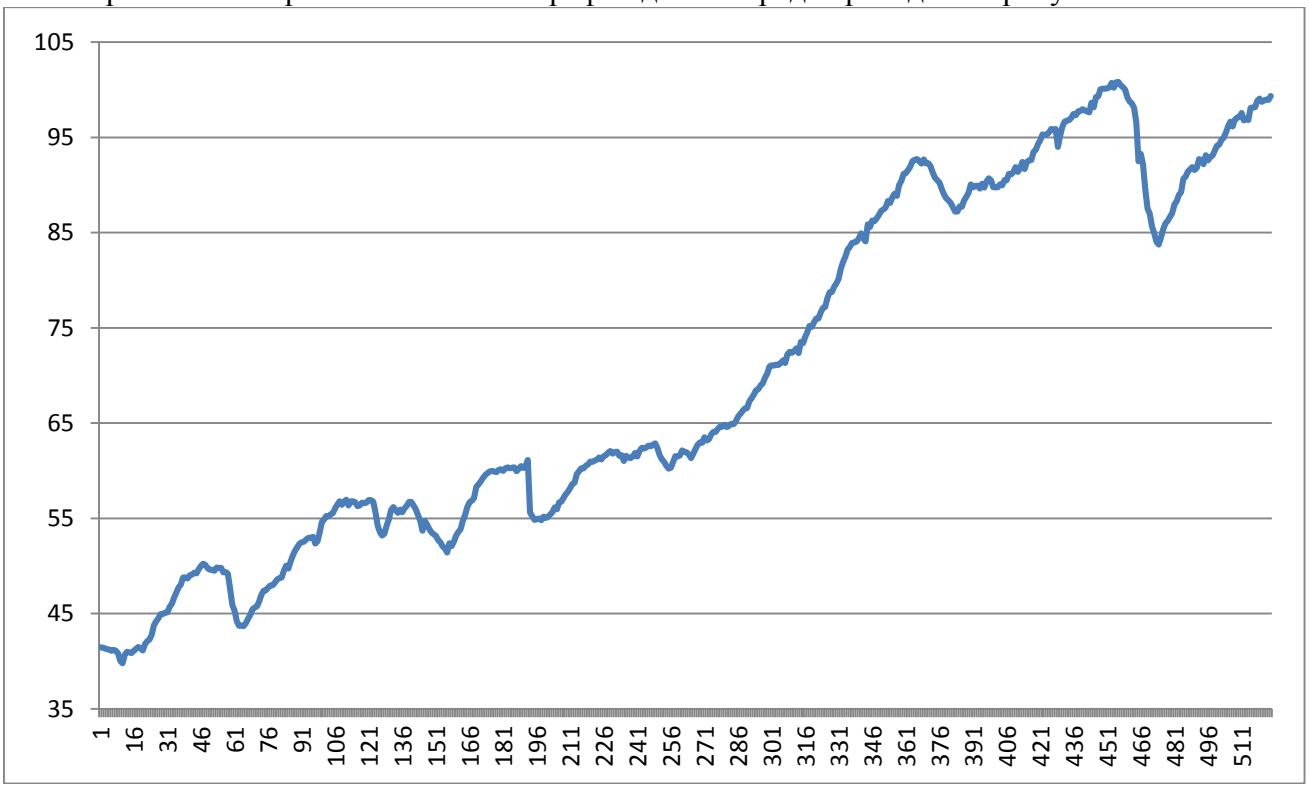

Рис. 2. График объёма промышленного производства (таймфрейм: 1 месяц)

Все прогнозы осуществлялись в тех же самых режимах, что и в случае прогнозирования временного ряда цен на энергоносители в США. Обозначения те же. Полученные результаты приведены в таблице 5.

Таблица 5. Прогнозирование объёма промышленного производства США.

Разбиение *n*	Глубина анализа	R-метод. Усреднение.	R-метод on-line	R-метод. Усреднение.

	m	on-line		20 шагов
5	5	0.35916 / 0.0492	0.64542 / 0.0884	2.21250 / 0.3033
10	2	0.34561 / 0.0474	0.39077 / 0.0536	1.42499 / 0.1954
	5	0.34561 / 0.0474	0.39077 / 0.0536	1.42499 / 0.1954
20	2	0.34808 / 0.0477	0.33722 / 0.0462	1.55850 / 0.2136
	5	0.34808 / 0.0477	0.33722 / 0.0462	1.55850 / 0.2136
50	2	0.36178 / 0.0496	0.34590 / 0.0474	2.12041 / 0.2907
	5	0.36178 / 0.0496	0.34590 / 0.0474	2.12041 / 0.2907

Из приведённых в таблице 5 результатов видно, что после определённого предела размера алфавита (разбиения непрерывного интервала) ошибки прогноза стабилизируются и меняются слабо. Также видно, что точность метода, как в случае прогнозирования предыдущего ряда, остаётся в пределах 1/25 от значения $delta$. Метод усреднения показывает свою эффективность только на малых разбиениях, что также совпадает с результатами, полученными в предыдущем примере. Фактически, это говорит о том, что для получения оптимальных прогнозов за приемлемое время достаточно использовать метод усреднения и подобрать такие минимальные значения разбиения n и глубины анализа m, которые будут давать оптимальные (приближенные к границе точности) значения ошибок прогнозов.

Наличие описанных границ точности предлагаемого метода и его модификаций объясняется достаточно просто: в случае прогнозирования сложных рядов, в которых нет видимых или относительно простых закономерностей, метод не находит их и просто усредняет значение тренда (разницу между соседними элементами), используя это значение в качестве прогнозного элемента. Если же какие-либо закономерности в ряду имеются, при достаточной длине ряда и глубине анализа m алгоритм ихвыявит, и итоговые ошибки прогноза будут меньше. Важно отметить, что в силу использования при вычислении прогнозного значения всего полученного распределения вероятностей (для всех элементов разбиения), теоретическая и практическая точность методов, как уже было показано, достаточно высокая.

4.5. Прогнозирование временных рядов IIF

Рассмотрим применение метода усреднения, описанного в разделе 4.1, на примере прогнозирования экономических временных рядов США, по которым известны результаты прогнозирования методами международного института прогнозистов (International institute of forecasters (IIF)). Было взято 4 временных ряда с сайта forecasters.org [21]: industry (индекс промышленного производства США), finance (1) и finance (2) (показатели финансовой активности США) и demographic (демографические показатели США). Данные временные ряды брались в следующих временных периодах. Ряд Industry в период с 01.1982 по 01.1994; Finance (1) в период с 01.1962 по 01.1974; ряд Finance (2) в период с 01.1965 по 01.1976; ряд Demographic в период с 01.1983 по 01.1994. Графики данных временных рядов приведены на рисунках 3-6.

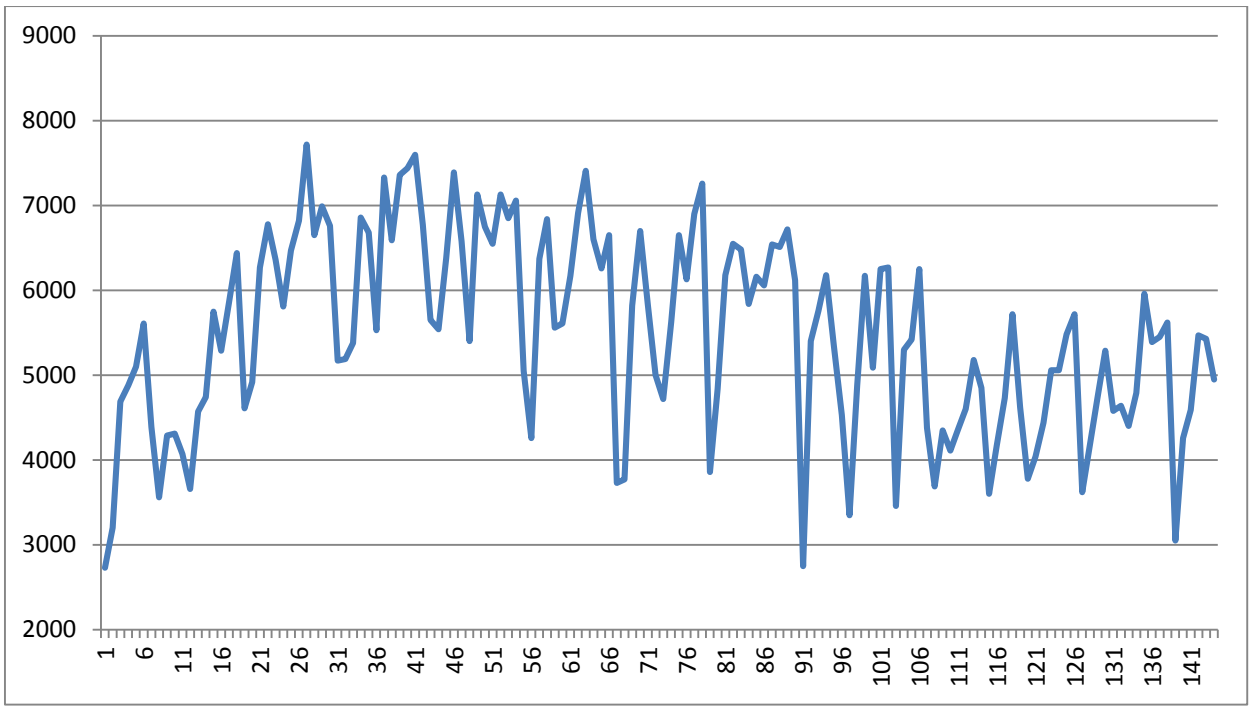

Рис. 3.Ряд Industry.

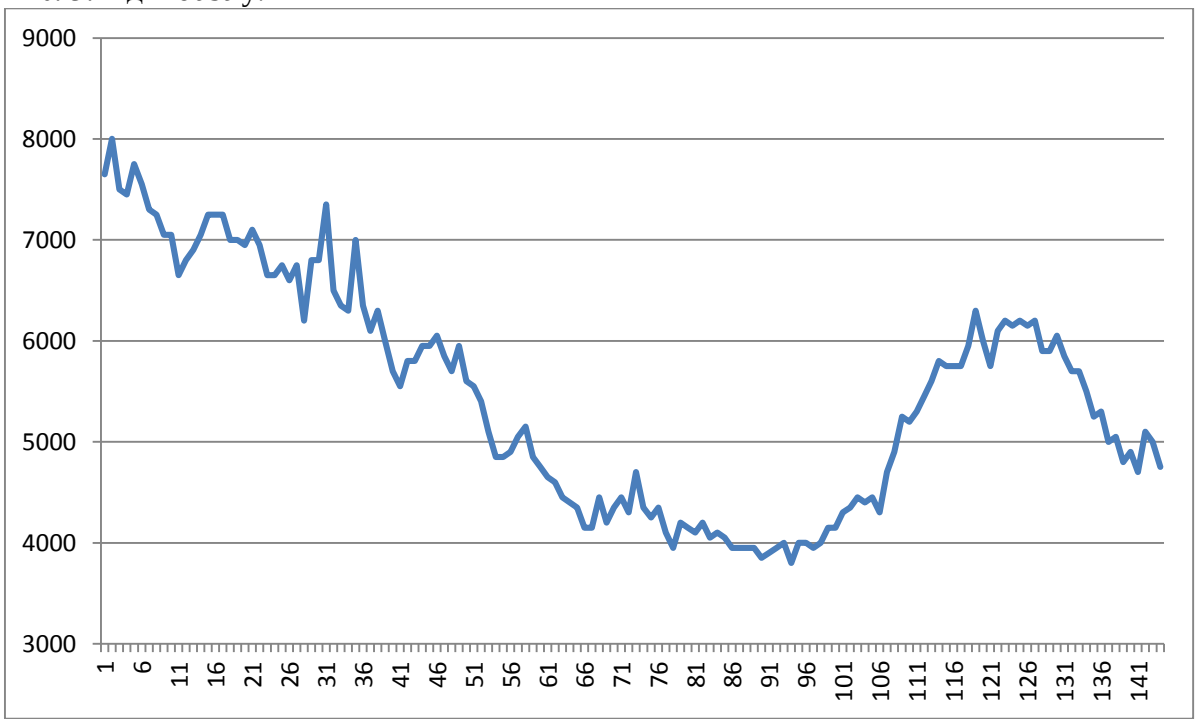

Рис. 4. Ряд Finance (1).

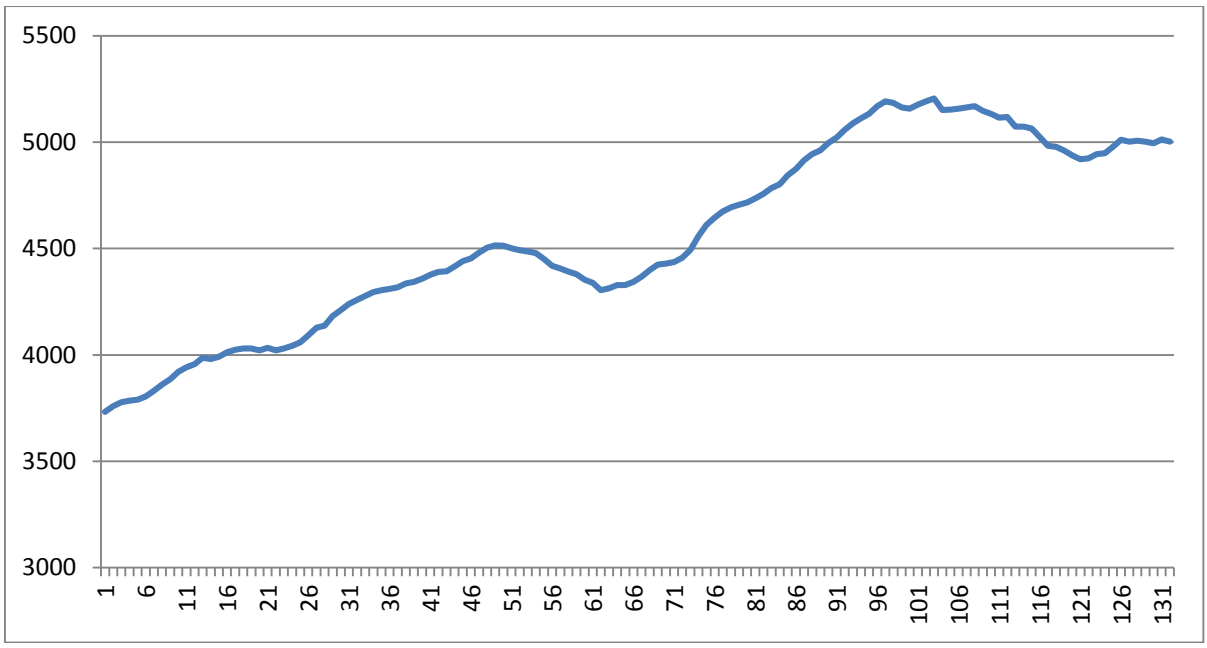

Рис. 5. Ряд Finance (2).

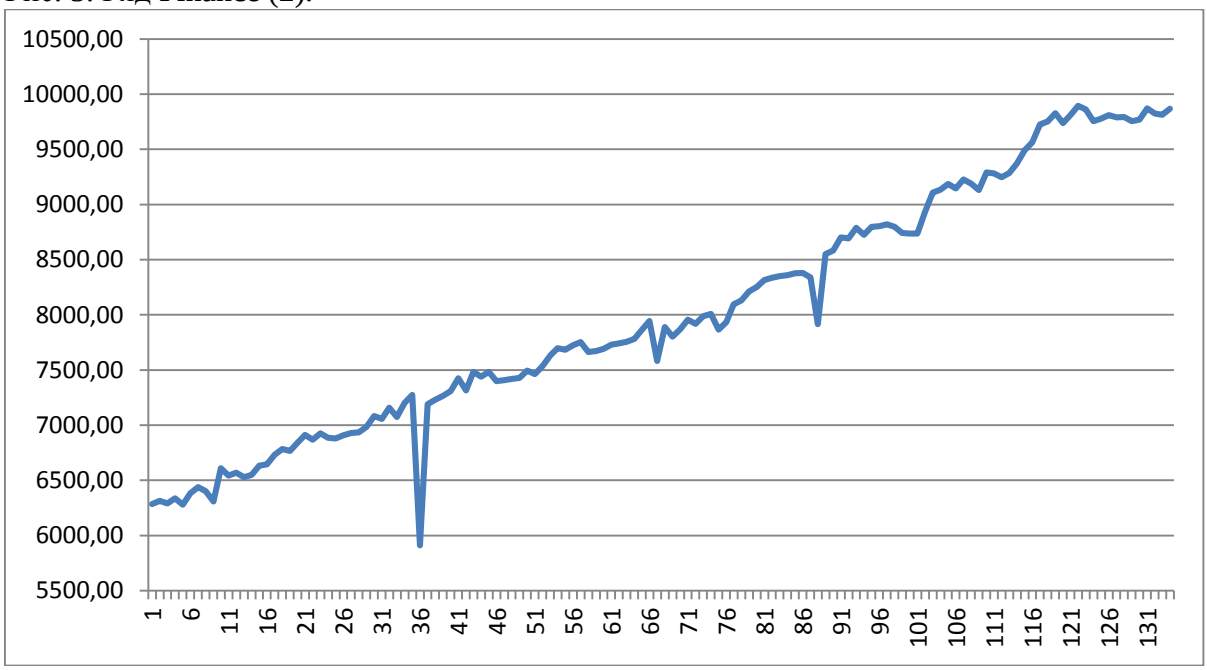

Рис. 6. Ряд Demographic.

Все прогнозы велись в режиме on-line. Эксперименты проводились на прогнозировании 18 различных элементов данных временных рядов. В качестве конкурентов методу R были выбраны следующие 3 наиболее известных метода, результаты по которым представлены на сайте IIF: AutoBox, ForecastPro, PP-Autocast. Глубина вычислений для всех рассмотренных случаев бралась равной трём. Эксперименты проводились на примере двух разбиений, т.к., большие величины разбиений, что было показано на практике, не дают существенного прироста точности, а в некоторых случаях могут её ухудшить. Результаты представлены в нижеследующих таблицах 6 и 7.

Таблица 6. Прогнозирование экономических временных рядов США.

Временной ряд	Размер выборки L	Разбиение n	$delta$	R-метод on-line	R-метод усреднение	Autobox
Industry	144	10	6050	768.06	703.53	340.72

Finance (1)	144		1550	167.5	165.44	680.49
Finance (2)	132		118	21.74	19.59	76.12
Demographic	134		2642	82.19	62.58	122.08
Industry	144	20	6050	718.61	706.52	340.72
Finance (1)	144		1550	162.64	164.48	680.49
Finance (2)	132		118	26.44	21.07	76.12
Demographic	134		2642	53.56	53.46	122.08

Таблица 7. Прогнозирование экономических временных рядов США.

Временной ряд	Размер выборки L	Разбиение n	$delta$	ForecastPro	PP-Autocast
Industry	144	10	6050	301.86	303.64
Finance (1)	144		1550	794.42	793.03
Finance (2)	132		118	71.98	41.40
Demographic	134		2642	152.71	286.19
Industry	144	20	6050	301.86	303.64
Finance (1)	144		1550	794.42	793.03
Finance (2)	132		118	71.98	41.40
Demographic	134		2642	152.71	286.19

Из приведённых таблиц 6 и 7видно, что R-метод в случае рядов Finance (1), Finance (2) и Demograpic даёт существенно меньшую ошибку прогноза по сравнению с другими известными методами. В среднем ошибка прогноза у метода R в 2 раза ниже, чем у других приведённых методов, что говорит о его высокой сравнительной эффективности. При этом внедрение усреднения алфавита в R-метод улучшает результат его работы.

4.6. Прогнозирование курсов валют

Рассмотрим многомерное прогнозирование некоторых экономических временных рядов, которые коррелируют между собой. Многомерный подход даёт возможность использовать в прогнозировании анализируемого ряда другие ряды, что даёт дополнительную информацию об исходном процессе и во многих случаях улучшает качество получаемых прогнозов. Проверим данное предположение на практике. Для этого рассмотрим прогнозирование индексов потребительских и промышленных цен США в период с 09.1983 по 03.2013. Их графики приведены на рисунках 8 и 9, соответственно. В качестве базового метода для реализации многомерного прогнозирования был выбран R-метод. Коэффициент сдвига целевого ряда l для всех случаев был выбран 1.

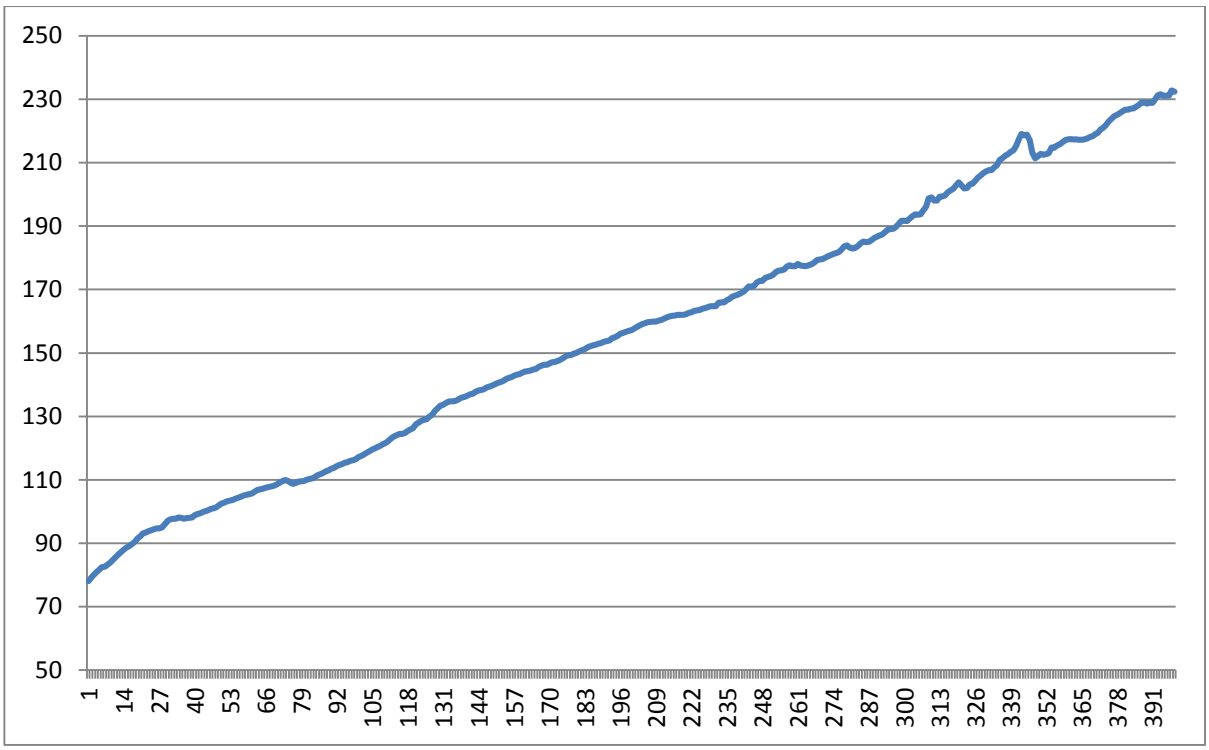

Рисунок 8 – Индекс потребительских цен США (CPI). 1983 – 2013гг.

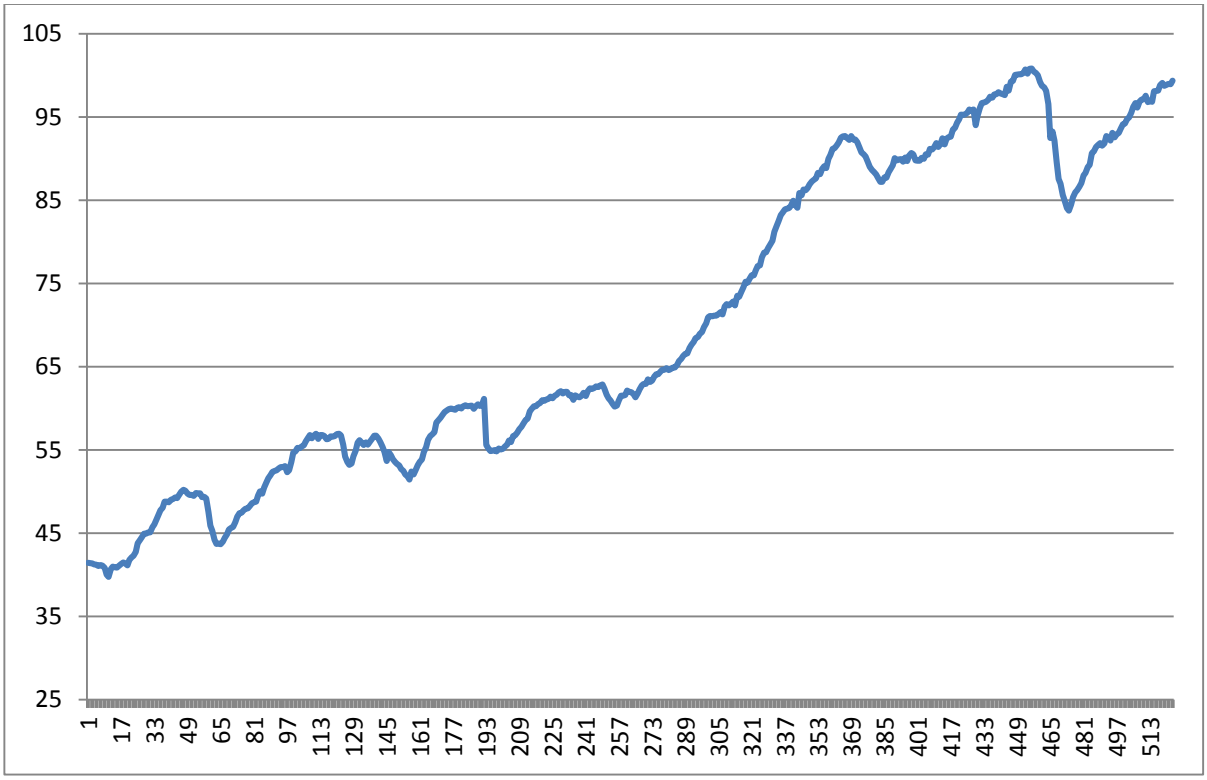

Рисунок 9 – Индекс промышленных цен США (PPI).

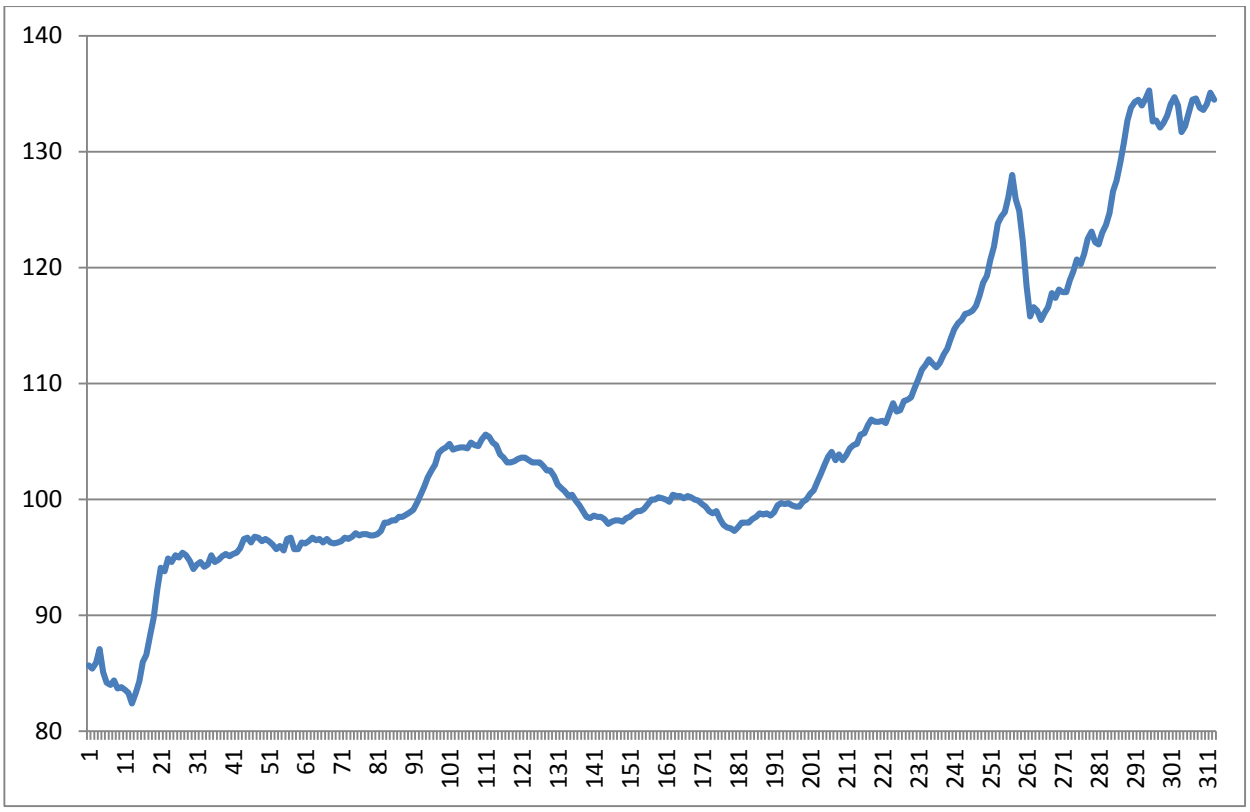

Рисунок 10 – Уровень экспорта США.

Размеры обоих рядов составляют 360 элементов, период между измерениями (таймфрейм) равен 1 месяцу. Для осуществления многомерного прогнозирования к заданным двум рядам присоединялись другие экономические ряды с теми же временными характеристиками (период; период между измерениями; длина) со сдвигом l равным 1. В частности, мы использовали дополнительно следующие ряды: уровень экспорта США (его график приведён на рисунке 10), курсы валют американский доллар / британский фунт стерлингов (USDGBP) и американский доллар / канадский доллар (USDCAD). Данные ряды были выбраны из соображений взаимосвязанности друг с другом (практического наличия корреляции) и вполне могут иметь выраженные корреляции.

Результаты прогнозирования индексов потребительских (CPI) и промышленных (PPI) цен приведены в таблицах 7 и 9, соответственно. При этом в первой строке идут результаты одномерного прогнозирования временного ряда (без присоединения к нему других рядов), а далее идёт двумерное прогнозирование с обозначением вида A + B. Данное обозначение говорит о том, что прогнозируются значения ряда A с присоединением к нему ряда B. Глубина анализа m = 3.

Таблица 7. Прогнозирование индекса потребительских цен США.

№	Временной ряд	R-метод on-line	R-метод 10 шагов
1	CPI	0.3922	0.5670
2	CPI + PPI	0.4308	0.5537
3	CPI + USDGBP	0.4533	0.5703

4	CPI + USDCAD	0.4533	0.5703
5	CPI + Уровень экспорта	0.7468	1.4239

Таблица 8. Прогнозирование индекса промышленных цен США.

№	Временной ряд	R-метод on-line	R-метод 10 шагов
1	PPI	1.3450	3.2575
2	PPI + CPI	1.0650	2.6825
3	PPI + USDGBP	1.2107	1.3571
4	PPI + USDCAD	1.2107	1.3571
5	PPI + Уровень экспорта	1.1393	2.9750

По графикам видно, что взятые дополнительные ряды не слишком сильно коррелируют друг с другом. В результате, в части прогнозирования индекса CPI получились результаты в среднем хуже или сравнимыми с одномерным случаем. В случае же прогнозирования индекса PPI результаты получились лучше, чем для одномерного ряда PPI, что говорит о нахождении предложенным подходом определённых корреляций.

5. Иллюстрации, визуализации результатов: приведены в разделе 4.

Эффект от использования кластера:

В процессе описанной работы активно использовались вычислительные ресурсы кластера, благодаря чему удалось получить столь значительные объёмы практических экспериментальных результатов и показать эффективность разработанных методов.

Перечень публикаций, содержащих результаты работы:

1. Lysyak, A.S. Gradient statistical attack at block cipher RC6 / A.S. Lysyak // Applied methods of statistical analysis. Simulations and statistical inference. – 2011. – P. 285–294.
2. Лысяк, А.С. Градиентная статистическая атака на блоковые шифры RC6, Blowfish / А.С. Лысяк // Материалы 50-й юбилейной международной научной студенческой конференции. – Новосибирск, 2012. – С. 18–23.
3. Lysyak, A.S. Analysis of gradient statistical attack at block ciphers RC6, MARS, CAST-128. / A.S. Lysyak // Proc. of XIII International Symposium on Problems of Redundancy in Information and Control Systems. – SpB., 2012. – C. 44-47.

4. Лысяк, А.С. Анализ эффективности градиентной статистической атаки на блоковые шифры RC6, MARS, CAST-128, IDEA, Blowfish в системах защиты информации. / А.С. Лысяк, А.Н. Фионов, Б.Я. Рябко // Вестник СибГУТИ. – 2013. – №1. – С. 85–109.
5. Lysyak, A. Universal coding and decision trees for nonparametric prediction of time series with large alphabets. A. Lysyak, B. Ryabko // Applied methods of statistical analysis. Simulations and statistical inference. – 2013. – P. 154–162.
6. Лысяк, А.С. Методы прогнозирования временных рядов с большим алфавитом на основе универсальной меры. / А.С. Лысяк, Б.Я. Рябко // Индустриальные информационные системы. – 2013. – С. 125–142.
7. Лысяк, А.С. Методы прогнозирования временных рядов с большим алфавитом на основе универсальной меры и деревьев принятия решений. / А.С. Лысяк, Б.Я. Рябко // Вычислительные технологии. – 2014. – Т. 19, №2. – С. 75–92.
8. Лысяк, А.С. Прогнозирование временных рядов на основе универсальной меры и деревьев принятия решений. / А.С. Лысяк, Б.Я. Рябко // Вестник СибГУТИ. – 2014. – №2. – С. 57–71.
9. Лысяк, А.С. Прогнозирование многомерных временных рядов. / А.С. Лысяк, Б.Я. Рябко // Вестник СибГУТИ. – 2014. – №4. – С.75–88.
10. Лысяк, А.С. Теоретико-информационные методы прогнозирования временных рядов. / А.С. Лысяк. – LAP Lambert Academic Publishing, 2014, ISBN 978-3-659-59737-4. – 72 с.

Итоги работы:

1. Подготовлена к защите кандидатская диссертация. Диссертация представлена в дисс. Совет ИВТ СО РАН и принята им к защите.
2. Намерен продолжать научно-исследовательскую работу по данной теме в рамках докторской диссертации у того же научного руководителя, а также в рамках своей научно-преподавательской деятельности в НГУ (как штатного сотрудника).

Необходимость продления доступа:

Требуется продление доступа до 01.10.2018 в связи с выполнением научно-исследовательской работы в рамках написания докторской диссертации по аналогичной теме, а также расширение используемых методов на область криптоанализа блоковых шифров.

Необходимость использования кластера ИВТЦ описана в разделе 2.

Необходимые аппаратные ресурсы остаются в рамках тех, что имеются на текущий момент.