На правах рукописи

Лысяк Александр Сергеевич

РАЗРАБОТКА И ИССЛЕДОВАНИЕ ТЕОРЕТИКО-ИНФОРМАЦИОННЫХ МЕТОДОВ ПРОГНОЗИРОВАНИЯ

Специальность 05.13.18 – «Математическое моделирование, численные методы и комплексы программ»

Автореферат

диссертации на соискание ученой степени

кандидата технических наук

Новосибирск – 2015

Работа выполнена в федеральном государственном автономном образовательном учреждении высшего образования «Новосибирский национальный исследовательский государственный университет», г. Новосибирск.

Научный руководитель: доктор технических наук, профессор, Рябко Борис Яковлевич

Официальные оппоненты: Дьячков Аркадий Георгиевич, доктор физико-математических наук, МГУ им. М.В. Ломоносова, г. Москва, профессор кафедры теории вероятностей.

Лемешко Борис Юрьевич, доктор технических наук, ФГБОУ ВО НГТУ, г. Новосибирск профессор кафедры теоретической и прикладной информатики.

Ведущая организация: Федеральное государственное бюджетное учреждение науки Институт математики им. С. Л. Соболева Сибирского отделения Российской академии наук, г. Новосибирск.

Защита состоится «4» декабря в 16:00 на заседании диссертационного совета ДМ003.046.01 на базе Федерального государственного бюджетного учреждения науки Института вычислительных технологий Сибирского отделения Российской академии наук по адресу 630090, г. Новосибирск, пр-т Академика Лаврентьева, 6, конференц-зал ИВТ СО РАН.

С диссертацией можно ознакомиться в библиотеке и на сайте Федерального государственного бюджетного учреждения науки Института вычислительных технологий Сибирского отделения Российской академии наук http://www.ict.sbras.ru/ru/Structure/disCouncil/lysyak2015

Автореферат разослан «___»_____2015 г.

Ученый секретарь диссертационного совета
кандидат физико-математических наук, доцент  Лебедев А.С.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность исследования.

Представленная работа посвящена исследованию теоретико-информационных методов прогнозирования временных рядов, описывающих прикладные процессы и реальные явления.

В настоящее время задача прогнозирования является актуальной при решении широкого спектра проблем в науке, экономике и технике. К их числу можно отнести анализ экономических, социальных, геофизических процессов, предсказание природных явлений, экономических событий и других прикладных областей. Кроме того, задача прогнозирования возникает при создании систем автоматического управления и систем поддержки принятия решений.

Методы прогнозирования служат для моделирования (исследования системных связей и закономерностей функционирования и развития) объектов и процессов с использованием современных методов обработки информации и являются важным средством в анализе сложных прикладных систем, моделировании их работы, работе с информацией, целенаправленном воздействии человека на объекты исследования с целью повышения эффективности их функционирования.

Наиболее распространённой постановкой задачи прогнозирования является задача прогнозирования временных рядов, т. е. прогнозирование функции какого-либо процесса, определённой на оси времени. В последние два десятилетия появилось множество методов прогнозирования, показавших достаточно высокую эффективность. Для решения этой задачи применяются модели машинного обучения, которые стали представлять собой серьёзную конкуренцию классическим статистическим моделям в сообществе специалистов по прогнозированию. Б.Я. Рябко был предложен и развит метод прогнозирования на основе универсального кодирования, представляющего способ кодирования информации, уменьшающий её конечный битовый размер[1]. Преимущество данного подхода состоит в выявлении скрытых закономерностей произвольного рода, что позволяет применять метод в достаточно широких диапазонах. Кроме того, в данной работе показано, что для решения задачи прогнозирования можно использовать методы из различных математических областей. В частности, в области интеллектуального анализа данных (data mining) имеется ряд методов, решающих задачу кластеризации. На базе данных методов, как показано в представленной работе, возможно проектирование новых алгоритмов прогнозирования.

Несмотря на разнообразие существующих методов прогнозирования, многие проблемы и задачи ещё далеки от своего решения. Количество публикаций, связанных с методами прогнозирования прикладных процессов,

[1] Рябко Б.Я. Прогнозирование случайных последовательностей и универсальное кодирование. // Проблемы передачи информации. 1988. №24, с.3–14.

постоянно растёт, что подтверждает важность выбранной области исследований.

Цель работы.

Разработка эффективных алгоритмов прогнозирования временных рядов, обладающих более высокой, чем у ранее известных методов, точностью, полиномиальной сложностью и учитывающих взаимные корреляции процессов, для решения задачи прогнозирования и криптоанализа блоковых шифров и генераторов псевдослучайных чисел (далее – ГПСЧ).

Задачи работы.

1. Разработка эффективных алгоритмов прогнозирования (обладающих невысокой вычислительной сложностью и использующих относительно небольшую память) на основе методов универсального кодирования и решающих деревьев.
2. Экспериментальное исследование разработанных методов при прогнозировании реальных экономических и социальных процессов.
3. Применение разработанных методов прогнозирования для анализа надёжности генераторов псевдослучайных чисел и блоковых шифров, а также для реализации градиентной статистической атаки на современные блоковые шифры.

Основные результаты, выносимые на защиту:

1. Предложены эффективные («быстрые» и не требующие большого объёма памяти) методы прогнозирования, базирующиеся на теории универсальной меры и на решающих деревьях.
2. Показано, что предлагаемые методы прогнозирования применимы для анализа надёжности генераторов случайных и псевдослучайных чисел, а также блоковых шифров.
3. Разработан универсальный метод [2] группировки алфавита, существенно уменьшающий вычислительную сложность и улучшающий качество получаемых прогнозов.
4. Разработан универсальный метод[2] многомерного прогнозирования, улучшающий точность получаемых прогнозов, благодаря учёту в прогнозе коррелирующих между собой временных рядов.
5. Показано, что качество работы предложенных методов при прогнозировании сложных экономических и социальных процессов выше, чем у ранее известных алгоритмов прогнозирования.

Достоверность результатов обеспечивается корректным применением методов интеллектуального анализа данных, математической статистики и теории вероятностей, а также сравнением полученных результатов с результатами, полученными другими авторами. Все экспериментальные результаты прогнозирования также сравнивались с реальными данными: получаемая точность методов оказалась высокой.

[2] Универсальный метод (универсальная модификация) – способ модификации какого-либо алгоритма прогнозирования, применимый к произвольному методу прогнозирования временных рядов.

Научная новизна.

Результаты экспериментальных исследований ряда теоретико-информационных методов прогнозирования важны как с точки зрения исследования сравнительной эффективности используемых математических подходов, так и с позиции дальнейшего развития методов прогнозирования, применимых при прогнозировании ряда прикладных процессов.

Предложена и экспериментально исследована модификация, применимая к произвольным методам прогнозирования, названная методом группировки алфавита, позволяющая существенно сократить трудоёмкость работы произвольного метода прогнозирования.

Разработаны два новых метода прогнозирования, основанных на применении и адаптации к этой задаче решающих деревьев.

Впервые предложены и экспериментально исследованы следующие универсальные модификации: методика создания гибридных методов прогнозирования, основанных на соединении нескольких различных алгоритмов; метод усреднения алфавита; моделирование поведений; метод многомерного прогнозирования, применимый к любому вероятностному алгоритму прогнозирования.

Впервые предложено приложение методов прогнозирования временных рядов к задачам криптоанализа блоковых шифров.

Практическая ценность работы. Разработанные методы прогнозирования и их реализации позволяют повысить эффективность работы автоматизированных систем, работающих со сложными прикладными процессами. Кроме того, предложенные в диссертационной работе методы являются эффективным средством поддержки принятия решений при управлении (как ручном, так и автоматизированном) сложными системами и процессами. Разработанные реализации предложенных методов прогнозирования превосходят по эффективности ранее известные методы.

Представление работы.

Основные положения и результаты диссертации докладывались и обсуждались на следующих конференциях:

- Applied methods of statistical analysis. Simulations and statistical inference (Россия, Новосибирск, 2011).
- Proc. of XIII International Symposium on Problems of Redundancy in Information and Control Systems (Россия, Санкт-Петербург, 2012).
- Applied methods of statistical analysis. Simulations and statistical inference (Россия, Новосибирск, 2013).
- Индустриальные информационные системы (Россия, Новосибирск, 2013).

Реализация и внедрение результатов работы.

Результаты представленной работы использовались при выполнении следующих проектов и государственных программ:

- Проект ООО ПКФ «Техпром»: «Моделирование спроса и предложения по отраслям в коммерческой организации».
- Проект ООО «РТИ-Югра»: «Разработка экспертных систем автоматической торговли на валютной бирже Forex».
- Проект федеральной целевой программы Минобрнауки РФ «Разработка теоретико-информационных методов оценки и повышения производительности компьютерных систем и сетей передачи данных». Государственный контракт №8239 от 17 августа 2012 года.
- Проект федеральной целевой программы Минобрнауки РФ «Эффективные методы построения защищённых высокоскоростных каналов передачи цифровых данных для предоставления доступа к широкополостным мультимедийным услугам». Государственный контракт №8229 от 6 августа 2012 года.
- Внедрение в учебный процесс кафедры Компьютерных систем ФАОУ ВПО НГУ по магистерским программам.
- Внедрение в учебный процесс кафедры Прикладной математики и кибернетики ФГОБУ ВПО СибГУТИ по магистерским программам.

Результаты диссертационной работы внедрены в учебный процесс на кафедре «Компьютерных систем» Новосибирского государственного университета в программе курса «Информационная безопасность» по направлению подготовки 230100 «Информатика и вычислительная техника».

Публикации. По материалам диссертации опубликовано 10 печатных работ, в том числе 4 работы в изданиях, внесённых в перечень журналов и изданий, утвержденных ВАК. Результаты работы отражены в отчетах по грантам и НИР, в рамках которых выполнялось исследование.

Личный вклад автора состоит в исследовании и программной реализации методов, построенных на основе универсальной меры; теоретической разработке и программной реализации ряда модификаций произвольных методов прогнозирования; экспериментальном исследовании всех рассмотренных методов и модификаций; выработке рекомендаций по подбору оптимальных параметров алгоритмов прогнозирования; оптимизации вычислительной сложности реализованных методов; теоретической разработке и реализации приложения методов прогнозирования к анализу надёжности генераторов случайных и псевдослучайных чисел, а также к градиентной статистической атаке на блоковые шифры.

Структура и объём работы. Представленная диссертационная работа состоит из 144 страниц текста и включает введение, пять глав, заключение, список литературы и приложение. Диссертация содержит 31 рисунок, 27 таблиц. Список литературы состоит из 38 источников.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** обсуждается актуальность темы исследований, формулируются основные цели и задачи, приводится обзор современных

тенденций в прогнозировании, а также описываются основные проблемы в изучаемой области.

В **главе 1** описываются теоретические основы прогнозирования, а также приводится обзор существующих методов прогнозирования.

В **разделе 1.1** описана общая постановка задачи прогнозирования временного ряда, которая может быть кратко сформулирована в следующем виде. Пусть имеется некоторый источник, порождающий последовательность элементов $x_1, x_2, \dots, x_t, x_{t+1} \in \mathrm{A}$ из некоторого множества A, называемого алфавитом. Задача прогнозирования состоит в определении распределения вероятностей для случайной величины $x_{t+1} \in \mathrm{A}$, 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 **разделе 1.2** сделан обзор современных тенденций и распространённых на практике методов прогнозирования, с которыми в дальнейшем будет производиться сравнение исследуемых в данной работе алгоритмов и подходов. В данном разделе проведён также анализ существующих методов по имеющимся проблемам: точность получаемых прогнозов, вычислительная сложность, учёт взаимных корреляций различных процессов. Показано, что существующие методы прогнозирования имеют существенные недостатки, которые исправлены в предложенной работе.

В **разделе 1.3** описан распространённый в настоящее время подход к прогнозированию временных рядов на основе методов сжатия данных.

В **главе 2** описаны исследуемые в представленной работе методы прогнозирования на основе универсального кодирования в классе стационарных и эргодических источников.

В **разделе 2.1** описывается предсказатель Лапласа, являющийся предшественником используемой в разработанных методах универсальной меры.

В **разделе 2.2** введено понятие универсального кода и универсальной меры. Рассмотрим их подробнее. 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.

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

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

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 неизвестного распределения источника P.

Таким образом, универсальная мера может быть использована для оценки вероятностей последовательностей, генерируемых любыми стационарными и эргодическими источниками.

В 1981 году Р.Е. Кричевский и В.К. Трофимов [3] предложили предсказатель для Марковских источников с фиксированной памятью $m \geq 0$, погрешность которого для источников независимых и одинаково распределённых элементов асимптотически минимальна:

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

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 с памятью (или связностью) m.

В 1988 году на базе предсказателя Кричевского Б.Я. Рябко была разработана мера R, универсальная для множества всех стационарных и эргодических источников[1]:

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

где множители ω_i являются некоторыми положительными весовыми коэффициентами, сумма которых равна 1. В качестве весовых коэффициентов было взято следующее распределение вероятностей $\{\omega_i\}$:

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

В **разделе 2.3** описан разработанный автором алгоритм прогнозирования на базе универсальной меры R для источников, порождающих значения из конечного алфавита. Оценка плотности вероятности происходит следующим образом:

$$p^*(x_{t+1}=a|x_1\dots x_t)=R(a|x_1\dots x_t)=R(x_1\dots x_t a)/R(x_1\dots x_t).$$

[3] Krichevsky, R.E., Trofimov V.K. The Performance of Universal Encoding, IEEE Trans. Information Theory, 1981, Vol. IT-27, № 2, pp. 199–207.

В **разделе 2.4** приведены разработанная автором схема прогнозирования с использованием меры R для источников, порождающих значения из непрерывного ограниченного интервала. Основная идея данного подхода состоит в построении возрастающей последовательности конечных разбиений $\{\Pi_n\}, n \geq 1$, интервала, в котором лежат значения временного ряда (источника). Разбиение в общем случае может быть произвольным. В процессе экспериментальных исследований применялись различные способы разбиения и сравнивались между собой. Лучшую эффективность среди всех испробованных показало равномерное разбиение, т.е. разбиение интервала на равные подинтервалы.

Оценка плотности вероятностей r при этом определяется следующим образом:

$$r(x_1 \dots x_t) = \sum_{s=1}^{\infty} \frac{\omega_s \mathrm{R}\left(x_1^{[s]} \dots x_t^{[s]}\right)}{\mathrm{L}(x_1^{[s]} \dots x_t^{[s]})},$$

где $x^{[k]}$ – элемент Π_k, содержащий точку x, $\mathrm{L}\left(x_1^{[s]} \dots x_t^{[s]}\right)$ – некоторая сигма-конечная мера (для нашей задачи была использована мера Лебега).

Соответствующая условная вероятность определяется так:

$$p^*(x_{t+1} = a | x_1 \dots x_t) = r(a|x_1, \dots, x_t) = r(x_1, \dots, x_t a)/r(x_1, \dots, x_t).$$

В **разделе 2.5** описан подход на базе адаптивной меры R. В случае прогнозирования нестационарных временных рядов, подход на базе адаптивной меры R придаёт последним участкам ряда больший вес (значимость), нежели участкам, находящимся ближе к началу ряда. Данное свойство достигнуто путём использования в мере R не всей длины последовательности, а некоторого «окна» фиксированного размера. При этом для подсчета «окна» используется подход «мнимое скользящее окно»[4], а используемая для прогнозирования универсальная мера определяется, как весовая сумма универсальных мер R всех «окон».

В **разделе 2.6** исследуется задача оптимизации вычислительной сложности меры R. Исходная сложность прогнозирования одного элемента равна следующей величине: $T(R) = O(t^3 \cdot n^{t+2})$.

В данном разделе предложен алгоритм вычисления меры R, при котором сложность её вычисления падает до величины: $T(R) = O(t \cdot m^2 \cdot n^{m+1})$, где n – величина разбиения (равная в большинстве оптимальных соотношений ~20), m – глубина анализа – параметр, который ограничивает число слагаемых в сумме (1). Видно, что сложность вычислений прогнозного значения в предложенном алгоритме линейно зависит от длины входной последовательности. Фактическое время вычислений на суперкомпьютере при длине ряда (t) до 1500 сократилось в $\left(2 \cdot n \cdot \left(\frac{t}{m}\right)^2\right)$, где $m \ll t$ раз, т.е. в 1000 – 1200 раз.

В **разделе 2.7** описаны технические особенности реализации оптимизированного вычисления меры R на суперкомпьютере. Приведена схема

[4] Рябко, Б.Я. Сжатие данных с помощью мнимого скользящего окна. Проблемы передачи информации, 1996, Т. 32, с. 22 – 30.

распараллеливания вычислений, а также рассмотренные варианты квантизации (разбиения) интервала.

В **главе 3** предложен подход к прогнозированию временных рядов, основанный на деревьях принятия решений (или решающих деревьях), применяющихся в задачах кластеризации и таксономии.

В **разделе 3.1** описан классический алгоритм построения решающих деревьев ID3, а также предложена его модификация, позволяющая применять данный подход в построении распределения вероятностей при работе с временным рядом. Кроме того, в данном разделе предложена модификация данного алгоритма, идея которой заключается в «обрезании» ветвей дерева на фиксированной глубине, что позволяет сократить максимальное число элементов дерева и как следствие – его сложность. Максимальная глубина дерева при этом, как и размер памяти m в мере R, будет равняться связности источника.

Кроме того, в данном разделе приведена оценка вычислительной сложности алгоритма, равная $T_{all\_tree} = O(t^2 \cdot n^m)$, где n – величина разбиения, m – глубина анализа (максимальная глубина дерева). Кроме того, предложена оптимизация метода прогнозирования с использованием решающих деревьев. Основная идея оптимизации состоит в построении вместо всего дерева только тех ветвей, которые требуются для построения искомого прогноза. В итоге сложность вычислений при прогнозировании на k шагов вперёд сократится до следующей величины: $T_{tree\_adaptive} = O(k \cdot t^2)$.

В **разделе 3.2** предложен ряд модификаций алгоритма на базе решающих деревьев, главной из которых является новый критерий ветвления, позволяющий снижать значимость уникальных признаков (таких как дата или номер), выбираемых в процессе построения дерева. В классическом критерии ветвления на основе прироста информации используется энтропия, значение которой с ростом уникальности значений признака растёт, что отрицательно сказывается на качестве прогноза. Предложенный критерий ветвления позволяет решить данную проблему и повысить качество прогноза.

В **разделе 3.3** описан метод прогнозирования, основанный на алгоритме кластеризации «Случайный лес» (Random forest), построенный на базе решающих деревьев. Основная идея алгоритма «случайного леса» состоит в построении множества решающих деревьев, которые используют случайный набор элементов из обучающей выборки с повторениями. Затем итоговая условная вероятность определяется как равномерная «склейка» (математическое среднее) оценок условных вероятностей, полученных в результате работы каждого дерева.

В данной главе предложен способ модификации описанного алгоритма для решения задачи построения распределения вероятностей временного ряда. Результатом работы алгоритма является распределение вероятностей.

В качестве основы для построения «случайного леса» были использованы адаптивные решающие деревья с модифицированным критерием ветвления, предложенные в разделе 3.2.

Кроме того, в данном разделе показана использованная схема распараллеливания вычислений при реализации предложенного алгоритма на суперкомпьютере. Описанная схема вычислений и хранения данных позволяет сократить время вычислений для адаптивного случайного леса в $3 * NTrees$ раз по сравнению со стандартной реализацией, где $NTrees$ – количество деревьев в лесу.

В **главе 4** описаны различные универсальные модификации, применимые для произвольных методов прогнозирования, оценивающих вероятность распределения источника. Данные модификации позволяют, как показано экспериментально, улучшить используемые методы прогнозирования в части точности получаемых прогнозов, а также вычислительной сложности. Все предложенные модификации исследовались экспериментально и сравнивались с уже известными методами прогнозирования (эффективность той или иной модификации подробно описана в главе 5).

В **разделе 4.1** описан метод усреднения, который применяется в алгоритме прогнозирования при работе с источником, порождающим значения из непрерывного интервала. Суть метода усреднения заключается во взятии в качестве прогнозного значения ряда не середины интервала с максимальной вероятностью, а математического ожидания на основе полученной плотности вероятности и середин подинтервалов.

В **разделе 4.2** предлагается метод группировки алфавита, предназначенного для снижения трудоёмкости используемого алгоритма, а также для улучшения точности получаемых прогнозов. Описан как сам алгоритм прогнозирования с использованием группировки алфавита, так и обоснована его применимость к произвольным алгоритмам прогнозирования, улучшающая точность их работы.

В **разделе 4.3** описана модификация склейки методов, позволяющая использовать в прогнозировании временных рядов несколько алгоритмов прогнозирования. При этом результирующая плотность вероятности получается путём весовой суммы плотностей каждого из участвующих методов.

В **разделе 4.4** описаны принципы моделирования поведений сложных прикладных систем. Смысл работы модуля моделирования заключается в построении прогнозов не значений ряда, а направления движения тренда. При этом выполняется прогнозирование ряда разниц $y_1, \dots, y_{t-1}$, построенного по исходному ряду $x_1, \dots, x_t$ следующим образом: $y_i = x_{i+1} - x_i$. Если количество направлений движения тренда равно k, выполняется прогнозирование ряда разниц с его разбиением на k частей. Результатом прогноза является номер полученного направления движения тренда.

В **разделе 4.5** предложен подход под названием «многомерное прогнозирование». Смысл данного подхода состоит в построении алгоритмов прогнозирования, учитывающих не только один источник (и соответственно, один временной ряд), а несколько коррелирующих между собой источников. К примеру, из макроэкономики известно, что внутренний валовый продукт

оказывает влияние на курс валюты изучаемой страны, а показатель уровня жизни влияет на индекс потребительских цен. Корреляций такого рода достаточно много и методы прогнозирования, которые могут учесть подобные взаимные влияния источников, способны дать намного лучший результат. Данный факт неоднократно подтверждён и показан в процессе экспериментальных исследований. В данном разделе описан метод, позволяющий строить многомерные прогнозы любого количества временных рядов, основанные на произвольном алгоритме прогнозирования.

Опишем подробнее суть многомерного подхода. Пусть имеется K временных рядов, коррелирующих каким-то образом между собой:

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

$$\dots$$

$$x_1^K, x_2^K, x_3^K, \dots, x_t^K.$$

При этом мы предполагаем, что все K временных рядов определены на одной и той же оси времени с едиными начальными и конечными точками. Также у них одинаковая квантизация (разбиение с алфавитом A), и сами ряды записаны в квантизованном виде. 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:

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

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 $A' = \{a | (x_{t+1}^2 \cdot N + x_{t+1}^3 \cdot N^2 + \dots + x_{t+1}^K \cdot N^{K-1}) \le a \le (x_{t+1}^2 \cdot N + x_{t+1}^3 \cdot N^2 + \dots + x_{t+1}^K \cdot N^{K-1} + N - 1)\}$. 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 $a = b \bmod N$, C – нормирующий коэффициент.

В **главе 5** описаны экспериментальные результаты всех предложенных методов и их модификаций на примере прогнозирования реальных экономических и социальных процессов.

В **разделе 5.1** описана общая схема проведения экспериментальных исследований. Опишем основные моменты экспериментальных исследований. В процессе прогнозирования рассмотренных далее временных рядов прогнозировались не абсолютные значения, а разницы между соседними

элементами рассматриваемого временного ряда. Это существенно увеличивает точность работы любого метода, а также снижает волатильность рядов. Таким образом, на базе заданного ряда, строился другой – ряд разниц.

Основные параметры исследования включают следующие: размер (длина) временного ряда (L), алфавит (разбиение) (n), параметр глубины анализа метода (m). Считалась величина ошибки прогноза для двух режимов работы метода: прогнозирование на 1 шаг вперёд (режим on-line) и прогнозирование сразу на 10-20 шагов вперёд. Определяется величина Δ, обозначающая максимальную разницу между двумя соседними элементами рассматриваемого временного ряда, т.е. фактически ширину интервала, из которого выбираются прогнозные значения. В этом случае значение Δ определяет максимально возможную ошибку прогноза.

Зная величину Δ для исследуемого ряда, а также разбиение ряда n, можно определить статистическую границу точности прогноза следующим образом:

$$Err_{max} = \frac{\Delta}{2 \cdot n}.$$

Данная формула показывает среднюю ошибку изучаемого метода прогнозирования, работающего при разбиении n, и не совершающего ошибок в прогнозировании номера подинтервала.

В **разделе 5.2** описаны экспериментальные результаты прогнозирования периодических функций методом R и методом решающих деревьев. Оба метода при длине последовательности больше 2 периодов функции показывают ошибку прогноза для обоих режимов работы, близкую к границе точности для любого разбиения. Данный результат вполне естественен, т.к. оба метода должны свободно выявлять явные периодические закономерности ряда. На более коротких последовательностях (1-2 периода) метод R даёт чуть лучшую точность.

В **разделе 5.3** приведены экспериментальные результаты прогнозирования ценовых экономических индексов США (индексы потребительских и промышленных цен) методами R и решающими деревьями. При этом в процессе экспериментов использовались различные варианты группировки алфавита и вариант без группировки. Наиболее эффективным из всех вариантов для обоих методов является группировка с равномерным разбиением на группы, т.е. вида n $(\sqrt{n} \cdot \sqrt{n})$, т.к. она приводит к существенному снижению временной сложности метода при сохранении или улучшении точки прогноза по сравнению с вариантом без группировки. Точность работы обоих методов сравнима и приближается во всех вариантах разбиения к пределу точности для разбиения $n = 5$.

В **разделе 5.4** описаны результаты прогнозирования цен на энергоносители в США методами R и решающими деревьями. При этом на примере R-метода была протестирована модификация усреднения. Точность получаемых прогнозов, в которых использовался метод усреднения, немного выше, чем у R-метода, не использующего усреднение. Точность R-метода и решающих деревьев при любых значениях параметров методов приближается к

границе точности для разбиения $n = 10$. Это говорит о том, что большие величины разбиений не дают существенного прироста точности, а в некоторых случаях могут её ухудшить. Решающие деревья на малых разбиениях дают чуть более высокую точность.

В **разделе 5.5** описаны результаты прогнозирования цен на энергоносители в США, но уже с использованием склейки методов R и решающих деревьев при различных весовых коэффициентах. Склейка показала во всех случаях неплохие результаты, которые всегда лучше худшего из двух склеиваемых методов. При этом в ряде случаев она показа точность лучшую, чем оба метода по отдельности.

В **разделе 5.6** показаны результаты прогнозирования объёмов промышленного производства в США методами R и решающими деревьями. Данный ряд обладает сравнительно невысокой волатильностью. При этом также использовался вариант методов с усреднением. Метод усреднения показывает лучшую по сравнению с классическим подходом эффективность только на малых разбиениях, что совпадает с результатами, полученными при прогнозировании ряда цен на энергоносители. Оба метода показывают точность, приближенную к границе точности при $n = 15$ при любых разбиениях (больших 15) и значениях глубины анализа. При том на маленьких разбиениях ($n = 5; 10$) точность методов с использованием усреднения, в отличие от стандартных вариантов, также приближена к границе точности $n = 15$. 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.

В **разделе 5.7** приведены результаты прогнозирования экономических и социальных временных рядов США, по которым известны результаты прогнозирования методами международного института прогнозистов (International institute of forecasters (IIF)). Было использовано 4 временных ряда, взятых из сайта forecasters.org: industry (индекс промышленного производства США), finance (1) и finance (2) (показатели финансовой активности США) и demographic (демографические показатели США). Все прогнозы велись в режиме on-line. Эксперименты проводились на прогнозировании 18 различных элементов данных временных рядов при помощи метода R с усреднением при разбиении $n = 20$ и глубине анализа $m = 5$. В качестве конкурентов методу R были выбраны следующие 4 наиболее известных и эффективных метода, результаты по которым представлены на сайте IIF: AutoBox, ForecastPro, PP-Autocast и Statistica. Результаты работы данных методов на представленных рядах взяты с сайта IIF[5]. Результаты прогнозирования и сравнения с первыми 3 методами приведены в таблице 1.

[5] Web-сайт Institute journal of forecasters: http://forecasters.org/resources/time-series-data.

Таблица 1. Результаты прогнозирования экономических рядов методами R, Autobox, ForecastPro и PP-Autocast.

Ряд	Размер ряда L	Δ	R-метод усреднение	Autobox	ForecastPro	PP-Autocast
Industry	144	6050	706.52	340.72	301.86	303.64
Finance (1)	144	1550	164.48	680.49	794.42	793.03
Finance (2)	132	118	21.07	76.12	71.98	41.40
Demographic	134	2642	53.46	122.08	152.71	286.19

Приведённые результаты говорят о том, 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.

В дальнейшем все экспериментальные результаты сравнивались с приведёнными методами прогнозирования. В подавляющем большинстве случаев предложенные автором алгоритмы показывают большую точность прогноза.

В **разделе 5.8** описаны экспериментальные результаты прогнозирования курсов валют адаптивным методом R, решающими деревьями и методом на основе случайного леса. Для прогнозирования использовались временные ряды курсов евро / доллар США и английский фунт стерлингов / доллар США при различных периодах между измерениями (называемых в дальнейшем таймфреймами): 1 час и 1 сутки. Данные ряды обладают очень высокой волатильностью и практически не имеют явных закономерностей.

В режиме on-line при таймфрейме в 1 сутки все три метода показывают сравнимые результаты, приближенные к границе точности при $n = 10$. При таймфрейме в 1 час ошибка приближается к границе точности при $n = 5$. При росте разбиения точность работы методов ухудшается. Описанные границы точности объясняются большей волатильностью ряда, получаемого при большем разбиении или большем таймфрейме, что ведёт к учёту исследуемым методом появляющихся в ряду шумов и негативно сказывается на точности прогнозов. Данная проблема легко решается при помощи использования группировки алфавита. В режиме прогнозирования на 10 шагов вперёд решающие деревья показывают в среднем в 1.5 раза меньшую ошибку прогноза. Метод на основе случайного леса показывает в среднем чуть лучшие, чем у двух других, результаты (в особенности при больших разбиениях).

Также важно отметить, что 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.

В **разделе 5.9** приведены результаты прогнозирования расхода электроэнергии в США методом R, решающими деревьями и методом на основе случайного леса. Данный ряд обладает некоторыми необычными закономерностями и относительно высокой волатильностью. Результаты прогнозирования данного ряда приведены в таблице 2.

Таблица 2. Прогнозирование расхода электроэнергии в США.

Тип прогнозирования	Δ	R-метод	Решающие деревья	Случайный лес
On-line	68.9	4.0813	3.9373	2.6727
10 шагов		38.3997	8.1417	2.2133

Из представленных результатов видно, что решающие деревья и метод на основе случайного леса дают существенно лучшие результаты в сравнении с R-методом. Данное свойство показывает лучшую работу обоих методов на основе решающих деревьев на последовательностях с высокой волатильностью (они выявляют определённые закономерности в то время, как работа R-метода сводится только к усреднению тренда). Получаемая погрешность методов на основе решающих деревьев близка к границе точности при разбиении $n = 10$.

В **разделе 5.10** показаны результаты по многомерному прогнозированию некоторых экономических временных рядов, имеющих взаимные корреляции. В частности, рассматривались следующие ряды: индекс потребительских цен США (CPI), индекс промышленных цен США (PPI), уровень экспорта США (Export), курс валют USD/GBP и USD/CAD, уровень безработицы США (Unemployment), обращения по безработице в США (Claims of unemployment), внутренний валовый продукт США (GDP), индекс промышленного производства США (IPI), цены на топливо (Gasoline). Прогнозирование велось на 1 шаг вперёд (on-line) и на 10 шагов вперёд. Базовый алгоритм для реализации многомерного подхода – адаптивный R-метод с усреднением, показавший свою высокую эффективность в предыдущих экспериментальных исследованиях.

Результаты многомерного прогнозирования уровня безработицы за период с 01.1970 по 08.2012 приведены в таблице 3. 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. Размер интервала (Δ) равен 1.65.

Таблица 3. Многомерное прогнозирование уровня безработицы.

№	Временной ряд	R-метод on-line	R-метод 10 шагов
1	Unemployment	0.106	0.425
2	Unemployment + Claims of unemployment	0.098	0.136
3	Unemployment + IPI	0.106	0.460
4	Unemployment + GDP	0.116	0.370

В таблице 4 приведены данные по многомерному прогнозированию уровня ВВП США за период с 01.1970 по 08.2012. Значение Δ равно 154.15.

Таблица 4. Многомерное прогнозирование уровня ВВП США.

№	Временной ряд	R-метод on-line	R-метод 10 шагов
1	GDP	9.363	79.132
2	GDP + Unemployment	16.988	50.475
3	GDP + IPI	9.209	78.133
4	GDP + Claims of unemployment	13.306	69.878
5	GDP + CPI	8.207	57.702

В таблице 5 представлено многомерное прогнозирование цен на топливо в США за период с 01.1992 по 08.2013. Размер интервала (Δ) равен 1.324.

Таблица 5. Многомерное прогнозирование цен на топливо.

№	Временной ряд	R-метод on-line	R-метод 10 шагов
1	Gasoline	0.162	0.211
2	Gasoline + Retail sales	0.122	0.134
3	Gasoline + Claims of unemployment	0.162	0.211
4	Gasoline + IPI	0.149	0.185
5	Gasoline + CPI	0.136	0.160

По представленным в таблицах 3, 4 и 5 данным видно, что добавление ряда, который коррелирует с основным, в большинстве случае существенно увеличивает точность получаемых прогнозов, что говорит о высокой эффективности предложенного многомерного подхода по сравнению с классическим (т.е. не учитывающим взаимные корреляции).

В **разделе 5.11** описано приложение методов прогнозирования в задачах проверки надёжности блоковых шифров и генераторов случайных и псевдослучайных чисел. Для выходной последовательности генератора случайных чисел верно

$$p(x_{t+1} = 0|x_1, x_2, \dots, x_t) = p(x_{t+1} = 1|x_1, x_2, \dots, x_t) = 1/2,$$

где p – оценка условной вероятности равенства $(t+1)$-ого бита выходной битовой последовательности $x_1 \dots x_t$ нулю или единице. Если некоторый метод прогнозирования предсказывает отклонения от этих вероятностей, то такой генератор случайных чисел не может использоваться в задачах криптографии.

То же правило справедливо и для выходной последовательности идеального блокового шифра.

Предложенные методы прогнозирования применялись для анализа надёжности и реализации градиентной статистической атаки на современные блоковые шифры. В частности, удалось получить новые результаты по отношению к следующему ряду блоковых шифров: RC6, MARS, CAST-128, IDEA, Blowfish. Под новыми результатами понимается следующее: количество раундов, на которых удалось найти отклонения от случайности и провести градиентную статистическую атаку, больше, чем было известно ранее по данному шифру.

Полученные результаты опубликованы в работах автора [1-4].

В **заключении** приводятся основные выводы по теме диссертационной работы:

- Разработаны эффективные (имеющие полиномиальную сложность и не требующие большого объёма памяти) алгоритмы для методов прогнозирования, базирующихся на универсальной мере и решающих деревьях.
- Показано, что точность работы предложенных методов прогнозирования, базирующихся на универсальной мере и решающих деревьях, выше, чем у ранее известных подходов (что показано на примере прогнозирования реальных экономических и социальных процессов).
- Разработан универсальный метод[2] группировки алфавита, существенно уменьшающий вычислительную сложность и улучшающий качество получаемых прогнозов.
- Разработан универсальный метод[2] многомерного прогнозирования временных рядов, улучшающий точность получаемых прогнозов, благодаря учёту в прогнозе коррелирующих между собой процессов.
- Показано, что предлагаемые методы прогнозирования можно применять для анализа надёжности генераторов случайных и псевдослучайных чисел и блоковых шифров.

ПУБЛИКАЦИИ В ИЗДАНИЯХ, РЕКОМЕНДОВАННЫХ ВАК

1. 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.

2. 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.
3. 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.
4. Лысяк, А.С. Прогнозирование многомерных временных рядов. / А.С. Лысяк, Б.Я. Рябко // Вестник СибГУТИ. – 2014. – №4. – С.75–88.

ПРОЧИЕ ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

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. – С. 44-47.
4. 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.
5. 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.
6. 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.