ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«НОВОСИБИРСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

На правах рукописи

Лысяк Александр Сергеевич

РАЗРАБОТКА И ИССЛЕДОВАНИЕ ТЕОРЕТИКО-ИНФОРМАЦИОННЫХ МЕТОДОВ ПРОГНОЗИРОВАНИЯ

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

Диссертация
на соискание ученой степени
кандидата технических наук

Научный руководитель:
д.т.н. Борис Яковлевич Рябко

Новосибирск – 2015

Оглавление

Введение

Актуальность исследования.

Представленная диссертационная 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 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, 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 свою достаточно высокую эффективность. В частности, в работе [1] описаны модели машинного обучения, которые стали представлять собой серьёзную конкуренцию классическим статистическим моделям в сообществе специалистов по прогнозированию [2,3,4]. В [5-8] был предложен и развит метод прогнозирования на основе универсального кодирования или «сжатия данных», т. е. применения определённых способов кодирования информации,

уменьшающих её конечный битовый размер. Преимущество данных методов состоит в выявлении скрытых закономерностей произвольного рода, что позволяет применять метод в достаточно широких диапазонах. Кроме того, в решении задачи прогнозирования могут использоваться методы из различных математических областей. В частности, в области интеллектуального анализа данных (data mining) имеется ряд методов, посвящённых решению задачи 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 разрешения. Количество публикаций, связанных с методами прогнозирования прикладных процессов, постоянно растёт, что подтверждает важность выбранной области исследований.

Объект исследования.

Объектом исследования представленной работы являются алгоритмы и методы прогнозирования зависящих от времени характеристик сложных прикладных процессов в различных экономических, социально-политических и естественнонаучных областях.

Цели и задачи исследования.

Цель данной диссертационной работы состоит в разработке эффективных алгоритмов прогнозирования временных рядов, обладающих высокой точностью, полиномиальной сложностью и учитывающих взаимные корреляции процессов, для решения задачи прогнозирования и криптоанализа блоковых шифров и генераторов псевдослучайных чисел (далее – ГПСЧ).

В данной работе предложены новые подходы к прогнозированию временных рядов, основанные на определённых моделях теории информации и методах когнитивного анализа данных. В частности, в качестве базовых методов решения задачи прогнозирования используются методы, построенные

на основе универсальных кодов, а также методы, основанные на решающих деревьях. Также, предложен ряд универсальных (т.е. применимых к любым методам прогнозирования) модификаций, существенно повышающих эффективность используемых алгоритмов. Описаны результаты экспериментальных исследований всех предложенных методов и модификаций, а также способы оптимизации обоих методов прогнозирования.

В области прогнозирования временных рядов существует большое количество актуальных задач и проблем. Некоторые из них были описаны и решены в представленной работе.

Одной из таких проблем является проблема высокой вычислительной сложности методов. В частности, методы, базирующиеся на универсальной мере, а также на решающих деревьях, требуют достаточно больших вычислительных ресурсов для достижения желаемого качества прогноза. Во многих случаях, для получения прогнозов желаемой точности необходимо использование суперкомпьютеров. Предложенный в данной работе метод группировки алфавита может быть внедрён практически в любой метод прогнозирования, где присутствует оценка вероятностей прогнозных событий (элементов).

Важным аспектом данной работы, вытекающим из выше упомянутой модификации, является также и то, что она посвящена методам прогнозирования временных рядов с большими алфавитами, прогнозирование которых ранее часто было невозможно из-за ограниченности в вычислительных ресурсах. Ранее, высокоэффективных методов прогнозирования для временных рядов с большим алфавитом не существовало. Предложенные в текущей работе методы прогнозирования временных рядов с большим алфавитом могут применяться и к стандартным временным рядам с небольшими алфавитами, при этом точность их прогноза при внедрении предлагаемых не меняется, а в некоторых случаях становится даже лучше, что показано ниже в экспериментальных результатах.

Помимо вышеописанных проблем, существует также и проблема получения высокого качества прогнозов в специализированных областях. В частности, рассматривая конкретные процессы в сферах техники и экономики, нередко существуют хорошо работающие методы прогнозирования, которые разрабатываются специализированно для применения в прогнозировании конкретно заданных процессов. В этом случае возникает задача выбора и применимости различных методов: какие-то подходы лучше работают на примере одних процессов, какие-то лучше прогнозируют другие типы процессов. В данной работе показано решение задачи создания метода, соединяющего в себе несколько методов прогнозирования с различным влиянием на результат. Данный подход называется склейкой методов.

Также, имеет место быть проблема прогнозирования временных рядов не на один, а на несколько шагов вперёд. В данной работе содержится описание возможных подходов к решению данной задачи, а также приведены реальные практические результаты её решения.

Кроме того, актуальной является задача повышения точности прогнозов за счёт учёта в процессе оценки распределения вероятностей будущего значения процесса других процессов (т.е. других временных рядов), которые каким-то образом коррелируют с целевым рядом. Подобного рода корреляции имеют место быть в большинстве реальных областей. К примеру, на уровень ВВП России с некоторым запозданием оказывают влияние индексы уровня безработицы, а также уровень цен на нефть. Таким образом, во многих ситуациях, зная в текущий момент времени значения характеристик коррелирующих с данным процессов, мы можем достаточно точно спрогнозировать будущее значение целевого процесса. Большинство из существующих методов прогнозирования данные взаимосвязи различных процессов не учитывают.

В соответствии с описанными выше проблемами, основные задачи проведённого исследования включают в себя следующие:

1. Исследование существующих методов и алгоритмов прогнозирования временных рядов.
2. Разработка эффективных алгоритмов прогнозирования (обладающих полиномиальной вычислительной сложностью и использующих относительно небольшую память) на основе методов универсального кодирования и решающих деревьев.
3. Оптимизация исследуемых методов и алгоритмов с целью снижения их трудоёмкости.
4. Разработка универсальных модификаций исследуемых методов с целью повышения эффективности их работы.
5. Программная реализация всех рассматриваемых методов прогнозирования и их практическое (экспериментальное) исследование на основе прогнозирования временных характеристик реальных экономических и социальных процессов.
6. Разработка универсального подхода в прогнозировании, позволяющего повысить точность прогнозов за счёт учёта взаимосвязей между различными коррелирующими процессами.
7. Применение разработанных методов прогнозирования для анализа надёжности генераторов псевдослучайных чисел и блоковых шифров, а также для реализации градиентной статистической атаки на современные блоковые шифры.

Научная новизна.

Научная новизна работы состоит в следующем:

1. Разработана и исследована экспериментально модификация, применимая к произвольным методам прогнозирования, названная методом группировки алфавита. Данная модификация основана на оценке распределения вероятностей, которая позволяют сократить трудоёмкость работы метода (ранее подобных методов опубликовано

не было и трудоёмкость многих эффективных методов требовала применение суперкомпьютеров).

2. Разработаны два новых метода прогнозирования, основанные на применении и адаптации к этой задаче решающих деревьев.
3. Разработана и проверена экспериментально методика создания гибридных методов прогнозирования, основанных на соединении нескольких различных методов.
4. Разработана и исследована экспериментально модификация, применимая к произвольным методам прогнозирования, основанная на усреднении алфавита и моделировании поведений.
5. Разработан метод многомерного прогнозирования, применимый к любому вероятностному алгоритму прогнозирования.
6. Предложены приложения методов прогнозирования временных рядов к задачам криптоанализа блоковых шифров.

Результаты, выносимые на защиту.

1. Разработаны эффективные («быстрые» и не требующие большого объёма памяти) алгоритмы для методов прогнозирования, базирующихся на универсальной мере и решающих деревьях.
2. Показано, что методы прогнозирования, базирующиеся на универсальной мере и решающих деревьях, обладают высокой точностью.
3. Показано, что предлагаемые методы прогнозирования применимы для анализа надёжности генераторов случайных и псевдослучайных чисел, а также блоковых шифров.
4. Разработан универсальный (применимый к произвольным алгоритмам прогнозирования временных рядов) метод группировки алфавита, существенно уменьшающий вычислительную сложность и улучшающий качество получаемых прогнозов.

5. Разработан универсальный метод многомерного прогнозирования, улучшающий точность получаемых прогнозов, благодаря учёту в прогнозе коррелирующих между собой временных рядов.
6. Показано, что качество работы предложенных методов при прогнозировании сложных экономических и социальных процессов выше, чем у ранее известных алгоритмов прогнозирования.

Практическая ценность полученных результатов.

Разработанная реализация предложенных методов и алгоритмов прогнозирования позволяет повысить эффективность работы некоторых 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 автоматическом) сложными системами и процессами. Разработанные реализации предложенных методов прогнозирования показывают результаты, превосходящие по своей эффективности ранее известные методы.

Внедрение результатов исследований.

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.

Апробация работы.

Основные результаты представленной диссертационной работы были представлены 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).

Публикации.

По теме диссертации опубликовано 10 печатных работ, в том числе 4 работы в научных журналах и изданиях, внесённых в перечень журналов и изданий, утвержденных ВАК, и 1 монография. Результаты работы отражены в отчетах по грантам и НИР, в рамках которых выполнялось исследование.

Структура диссертации.

Представленная диссертационная работа состоит из 144 страниц текста и включает введение, пять глав, заключение, список литературы и приложения. Диссертация содержит 31 рисунок, 28 таблиц. Список литературы состоит из 38 источников.

Глава 1 содержит обзор современных актуальных методов прогнозирования, а также общий анализ современных тенденций в области прогнозирования временных рядов. Также, в главе 1 описана общая постановка задачи прогнозирования и применяемая в дальнейшей части работы терминология.

Глава 2 посвящена описанию теоретических основ методов прогнозирования временных рядов, основанных на универсальной мере, а также способ оптимизации работы алгоритма данного метода с целью снижения его операционной сложности. В данной части работы описаны схемы прогнозирования для источников, порождающих значения из конечного множества и для источников, порождающих значения из непрерывного интервала. Кроме того, в данной главе предложен адаптивный метод, основанный на универсальной мере.

Глава 3 содержит описание методов прогнозирования, основанных на решающих деревьях. Предлагаются модификации методов кластеризации на основе решающих деревьев и на основе алгоритма кластеризации «Случайный лес». Предложенные модификации позволяют применять данные подходы из когнитивного анализа данных в задачах прогнозирования. Также, в данной главе предложены различные модификации представленных алгоритмов.

В главе 4 описаны различные модификации, применимые для произвольных методов прогнозирования, основанных на оценке вероятностного распределения: метод группировки алфавита, метод усреднения алфавита, склейка методов, моделирование поведений, а также многомерный подход в прогнозировании. Предложенный в данной части работы

многомерный подход позволяет модифицировать произвольные вероятностные методы прогнозирования таким образом, чтобы учитывать корреляции, существующие между различными прикладными процессами и системами, временные характеристики которых нам известны. Благодаря данному подходу стало возможно существенное повышение эффективности работы применяемых методов прогнозирования.

В главе 5 описаны экспериментальные результаты всех предложенных методов и их модификаций на примере прогнозирования реальных экономических и социальных процессов. Также, в главе 5 описаны приложения предложенных методов прогнозирования к задачам криптоанализа блоковых шифров.

В заключении описаны основные выводы, полученные в ходе диссертационного исследования.

Глава 1. Описание методов прогнозирования

1.1. Постановка задачи прогнозирования

В общем виде задача прогнозирования временных рядов может быть сформулирована следующим образом. Пусть имеется некоторый источник, порождающий последовательность элементов $x_1, x_2, \dots, x_t$ из некоторого множества A, называемого алфавитом. Алфавит может быть как конечным, так и бесконечным (т.е. представлять собой некоторый ограниченный непрерывный интервал). Пусть при этом в момент времени t мы имеем порождённую источником конечную последовательность $x_1, x_2, \dots, x_t$. Задача прогнозирования на 1 шаг вперёд состоит в определении распределения вероятностей для случайной величины $x_{t+1} \in$ A.

Для решения данной задачи будем рассматривать вероятностный подход к определению следующего элемента x_{t+1}. Для этого, в случае дискретного конечного алфавита, мы будем оценивать условные вероятности следующего вида: $p(x_{t+1} = a \in A | x_1, x_2, \dots, x_t)$, т.е. определять условную вероятность того, что следующий элемент ряда x_{t+1} равен элементу a ∈ A при условии, что известны предыдущие t элементов ряда. В случае прогнозирования ряда со значениями из непрерывного ограниченного интервала, мы будем оценивать плотность вероятности следующего вида: $p(x_{t+1} | x_1, x_2, \dots, x_t)$, где $x_{t+1} \in$ A – независимая переменная.

Таким образом, имея оценку плотности вероятности, мы можем вычислить оценки других статистических характеристик процесса: мат. ожидание, дисперсию, квантили, медиану и т.д,, т.е. мы будем иметь всю информацию об имеющемся процессе. Ясно, что чем точнее оценка плотности вероятности, тем точнее будут все вычисленные по ней характеристики. На практике наиболее востребованной является оценка мат. ожидания.

Кроме того, практический интерес представляет задача прогнозирования на несколько шагов вперёд. В этом случае для дискретного алфавита задача

прогнозирования состоит в оценке многомерного распределения вероятностей следующего вида:

$$p\left(x_{t+1} = a_{i_1}; x_{t+2} = a_{i_2}; \dots; x_{t+n} = a_{i_n} \middle| x_1, x_2, \dots, x_t\right)$$

где $a_{i_k} \in \mathrm{A}$, $k = 1, \dots, n$. В случае непрерывного алфавита задача будет состоять из оценки соответствующей многомерной плотности вероятности.

Если мы имеем источник, генерирующий временной ряд из дискретного конечного алфавита A, для решения задачи прогнозирования естественным образом может применяться множество математических алгоритмов прогнозирования, основывающихся на оценке условных вероятностей. Впервые подобный подход к прогнозированию элементов источников, порождающих значения из дискретного конечного алфавита, был описан в [9].

В случае, если алфавит A представляет собой непрерывный ограниченный интервал, нам требуется оценить плотность вероятности распределения величины $x_{t+1} \in \mathrm{A}$. Для этого применим следующий подход. Первоначально определим интервал [A, B], как минимальный непрерывный вещественный интервал, включающий в себя все элементы временного ряда $x_1, \dots, x_t$. Далее построим для интервала [A, B] возрастающую последовательность конечных разбиений $\{\Pi_k\}, k \geq 1$, где каждое Π_k – множество непересекающихся подмножеств каждого элемента из Π_{k-1}. В общем случае, разбиения Π_k могут быть произвольными (неравномерными). Далее сопоставим временному ряду $x_1, \dots, x_t$ целочисленные номера в соответствии разбиениями Π_i и получим множество рядов $x_1^{[s]}, \dots, x_t^{[s]}$, где $x_i^{[s]}$ – элемент Π_s, содержащий точку x_i. Фактически, мы получим множество временных рядов с элементами из конечных дискретных алфавитов Π_s и можем работать с данными рядами так же, как и в случае рядов из конечного дискретного алфавита, учитывая кроме того в алгоритме прогнозирования особенности разбиений Π_s. В результате работы алгоритма, получим множество условных вероятностей, на основе которых в дальнейшем мы можем построить оценку соответствующей функции

плотности. На предложенном подходе основаны разрабатываемые и исследуемые в данной диссертационной работе методы прогнозирования вещественных временных рядов.

Количество букв алфавита для случая конечного алфавита и максимальную мощность разбиения для случая непрерывного интервала обозначим через N. Предполагается, что процесс, или источник информации, является стационарным и эргодическим, т. е. неформально, распределение вероятностей символов этого источника не изменяется со временем и не зависит от конкретной реализации процесса. Данное предположение связано с тем, что в работе [7] математически доказано, что используемый в данной работе метод на основе универсальной меры выявляет закономерности именно для таких видов рядов. Другой используемый метод, основанный на решающих деревьях, сходен по видам выявляемых закономерностей и некоторым принципам действия с методом на основе универсальной меры. Важно отметить, что многие реальные временные ряды могут не являться стационарными и эргодическими, однако в задачи данной работы входит экспериментальное исследование применимости предложенных методов в предположении, что они таковыми являются, а также экспериментальное изучение поведения предложенных методов на произвольных реальных данных, свойства которых в общем случае неизвестны.

Пусть имеется источник, порождающий сообщение $x_1, \dots, x_{t-1}, x_t, x_i \in \mathrm{A}$, $\mathrm{i} = 1,2, \dots, t$, и требуется спрогнозировать n следующих элементов (в простейшем случае 1 элемент). При этом ошибкой прогноза E_i на i-ом шаге назовём апостериорную величину отклонения прогнозного значения x_i^* (полученного каким-либо образом из распределения вероятностей) от истинного значения процесса x_i в рассматриваемый i-ый момент времени, т.е. ошибкой прогноза является следующая величина:

$$E_i = |x_i - x_i^*|,$$

где x_i^* – прогнозное значение, а $x_i -$ истинное значение процесса. В простейшем случае прогнозное значение получается из плотности вероятности, как элемент (середина интервала в разбиении), имеющий максимальную вероятность. Под ошибкой прогноза на n шагов вперёд будем понимать среднюю ошибку прогноза каждого из n элементов в отдельности. Понятно, что ошибка прогноза характеризует качество прогнозирования и является основным критерием определения эффективности работы выбранного метода прогнозирования.

Очевидно, что если распределение вероятностей исходов процесса известно заранее, то задача прогнозирования следующих значений решается достаточно просто: в соответствии с известным распределением вероятностей (т.е. фактически с известной закономерностью) строится прогнозная функция, которая определяет все последующие элементы ряда. Либо же прогнозные значения выбираются, исходя из условия удовлетворения этих значений плотности распределения вероятностей ряда, полученного после вставки прогнозных элементов. Однако в большинстве практических задач описанные априорные данные отсутствуют, да и не всегда заданное распределение явно существует. В данной работе мы будем рассматривать именно такой случай. В такой ситуации для решения задачи прогнозирования можно воспользоваться точными оценками указанных величин, полученными с помощью статистических методов, построенных на основе анализа взаимосвязи последовательных исходов процесса и выявления закономерностей.

В более общей постановке задачи прогнозирования элементы x_i могут быть не только конкретными числами (целыми или вещественными), 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. Рассмотрим пример. Пусть имеется ряд значений ВВП страны с интервалом в

один месяц, значения которого требуется спрогнозировать. Как известно, на ВВП влияют такие параметры, как уровень инфляции, индекс потребительских цен, объёмы промышленного производства, дефицит платёжного баланса и многое другое. Значения всех таких характеристик так же, как и значение ВВП, могут быть известны на каждый месяц прогнозируемого ряда ВВП. Таким образом, мы можем составить многомерный ряд и прогнозировать уже не одно значение временного ряда, а значения всего вектора. При этом интересовать нас будет лишь один – первый – элемент прогнозного вектора. В итоге, задача прогнозирования может быть, как одномерной, так и многомерной.

1.2. Обзор современных тенденций в сфере прогнозирования

В настоящее время существует достаточно много эффективных и разнообразных методов прогнозирования, связанных с мощным математическим аппаратом. К наиболее широко используемым, в частности, относятся методы прогнозирования на основе билинейной модели [1-3], авторегрессионный анализ различных типов [5], спектральный анализ [6, 8], прогнозирование на основе методов Монте-Карло [6], методы на основе машинного обучения и экспертных оценок (рекурсивные стратегии [4,8], нейронные сети [7]), фрактальные стратегии, методы на основе многомерной регрессии (в том числе с использованием непараметрических оценок плотности распределения) [4] и многие другое. Данные методы в современное время являются одним из наиболее известных и широко распространённых подходов в прогнозировании. Важно также заметить, что для автоматизации процесса прогнозирования и соответствующего операционного контроля операций, часто используются специализированные программы, такие как Statistica и Autobox.

Несмотря на наличие описанного спектра методов и алгоритмов, многие проблемы в задах прогнозирования ещё далеки от своего разрешения. Среди таких проблем можно выделить следующие:

- Низкая точность прогнозов для процессов, принимающих вещественные значения.
- Высокая (часто экпоненциальная) вычислительная сложность существующих методов.
- Отсутствие качественных методов учёта взаимных корреляций различных процессов, что не позволяет учитывать их при прогнозе.

В результате проведённого анализа методов прогнозирования, перечисленных выше, было показано, что все рассмотренные методы обладают одним или несколькими из указанных проблем.

Решению данных проблем в числе прочих и посвящена данная диссертационная работа и в частности, разработанные в ней методы. В [10] уже были показаны некоторые экспериментальные результаты для метода на основе универсальной меры, в которых отражены результаты прогнозирования природных явлений (данные результаты превзошли результаты всех ранее существовавших методов прогнозирования природных явлений). В экспериментальных результатах текущей работы показаны новые экспериментальные результаты, посвящённые прогнозированию экономических временных рядов, показана точность получаемых прогнозов, которая во многих случаях превосходит точность многих современных методов для данных видов рядов.

На ряду с перечисленными важной проблемой является также отсутствие значимых (т.е. с приемлемым качеством прогноза и с достаточным для оценки качества прогноза количеством экспериментальных данных) и многочисленных результатов и методов прогнозирования на несколько шагов вперёд, несмотря на то, что данный класс задач также является очень важным и актуальным. Такое небольшое количество подходов к этой задаче связано с наличием большого количества сложностей и нерешённых проблем. В частности, к ним относятся эффект накапливания ошибок, снижение качества прогноза и увеличение неопределённости с ростом числа прогнозируемых шагов. К

существующим методам, решающих проблемы накапливающихся ошибок, относятся методы на основе билинейной модели и основанные на сжатии данных [1,17]. При этом методы на основе сжатия данных являются одними из наиболее распространённых современных подходов. Общая схема прогнозирования на основе данных методов описана в разделе 1.3. Тем не менее, точность работы указанных методов остаётся невысокой. Предложенные в текущей работе алгоритмы имеют большую точность прогноза.

1.3. Прогнозирование на основе сжатия данных и статистических тестов

В современное время очень распространён подход к прогнозированию на основе сжатия данных. Данный метод связан с возможностью выявления скрытых закономерностей во временном ряду путём применения методов сжатия (архивации) анализируемого ряда. Чем выше получилась степень сжатия, тем менее случаен (имеет больше закономерностей) рассматриваемый процесс. Таким образом, общая схема прогнозирования на базе методов сжатия выглядит следующим образом.

Пусть имеется процесс, порождающий ряд $x_1, \dots, x_{t-1}, x_t, x_i \in \mathrm{A}$, $i = 1,2, \dots, t$. И пусть требуется спрогнозировать 1 следующий элемент. Поступим следующим образом. Определим функцию $Compress(x)$, где x – последовательность данных, как функцию вычисления размера сжатой последовательности каким-либо фиксированным алгоритмом сжатия. Вычислим данную функцию по отношению ко всем последовательностям вида $x_1 \dots x_t a$, где $a \in \mathrm{A}$. Теперь условные вероятности соответствующих прогнозных элементов будут определяться следующим образом:

$$p(x_{t+1} = a \in A | x_1, x_2, \dots, x_t) = 1 - \frac{Compress(x_1 \dots x_t a)}{\sum_{p \in A} Compress(x_1 \dots x_t p)}$$

Данное определение плотности вероятности является корректным и может быть получено с использованием произвольного метода сжатия (архивации) данных.

Описанный выше общий подход может быть применён с использованием произвольных статистических критериев, основанных на каких-либо статистических тестах. В этом случае при выявлении каких-либо закономерностей в заданной последовательности, значение статистики критерия будет возрастать. Соответственно, схема вычисления условных вероятностей будет выглядеть следующим образом:

$$p(x_{t+1} = a \in A | x_1, x_2, \dots, x_t) = \frac{Stat(x_1 \dots x_t a)}{\sum_{p \in A} Stat(x_1 \dots x_t p)}, \quad (1)$$

где функция $Stat$ вычисляет значение статистики выбранного критерия.

Глава 2. Схема прогнозирования на основе универсальной меры

2.1. Предсказатель Лапласа и его свойства

Одним из первых исследователей, предложившим решать задачу прогнозирования с использованием условных вероятностей был Лаплас. Он предложил следующий вероятностный метод.

Пусть дан временной ряд $x_1, \dots, x_t$, где $x_i \in A$, A – конечный алфавит возможных значений элементов ряда. Требуется оценить неизвестную условную вероятность $p(x_{t+1} = a \in A | x_1 x_2, \dots, x_t)$. Предсказатель Лапласа представляет собой именно такую оценку неизвестных условный вероятностей стохастического процесса. Он принимает на вход предполагаемое следующее прогнозное значение ряда из алфавита при условии, что мы знаем все ранние значения ряда. На выходе предиктор Лапласа даёт некоторую числовую оценку условной вероятности предполагаемого прогнозного значения.

Предсказатель Лапласа вычисляется по следующей формуле:

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

Рассмотрим пример. Пусть алфавит $A = \{0,1\}$ и пусть дана временная последовательность $x_1 \dots x_7 = 1010101$. Тогда предикторы Лапласа будут определяться следующим способом:

$$L_0(x_8 = 0 | x_1 \dots x_7 = 1010101) = (3 + 1)/(7 + 2) = 4/9,$$

$$L_0(x_8 = 1 | x_1 \dots x_7 = 1010101) = (4 + 1)/(7 + 2) = 5/9.$$

Полученные величины $L_0(x_8 = 0 | x_1 \dots x_7 = 1010101)$ и $L_0(x_8 = 1 | x_1 \dots x_7 = 1010101)$ являются оценками условных вероятностей $P(x_8 = 0 | x_1 \dots x_7 = 1010101)$ и $P(x_8 = 1 | x_1 \dots x_7 = 1010101)$, соответственно.

Лаплас рассмотрел применение данного метода для решения проблемы оценки вероятности того, что Солнце взойдёт завтра с учётом знаний того, что оно всходило во все дни с момента сотворения Земли. При этом мы знаем, что все предыдущие дни, сколько бы мы ни взяли, оно восходило. В данном случае

алфавит A состоит из двух элементов: 0 («Солнце взойдёт») и 1 («Солнце не взойдёт»), t – это количество дней, которые мы рассматриваем в ряду. В итоге, получим ряд: $x_1, \dots, x_{t-1}, x_t = 0 \dots 00$. Предложенный Лапласом подход к получению множества оценок условных вероятностей, является достаточно естественным и эффективным, так как учитывает всю априорную информацию о процессе.

Предиктор γ, оценивающий условные вероятности заданного временного ряда $x_1, \dots, x_{t-1}, x_t$, называется универсальным для множества источников Ω, если его ошибка с ростом длины ряда t стремится к нулю. При этом под ошибкой предсказателя γ будем понимать следующую величину:

$$\rho(P||\gamma) = \sum_{x_1 \dots x_{t-1} x_t \epsilon A^t} P(x_1 \dots x_{t-1} x_t) \sum_{a \in A} P(a|x_1 \dots x_{t-1} x_t) \log \frac{P(a|x_1 \dots x_{t-1} x_t)}{\gamma(a|x_1 \dots x_{t-1} x_t)} \quad (2)$$

где P – истинное распределение вероятностей исследуемого процесса. В [9] показано, что ошибка предсказателя Лапласа при оценке вероятностей появления символов на выходе источника независимых и одинаково распределённых символов асимптотически (при $t \to \infty$) стремится к нулю. Соответственно, предсказатель Лапласа для указанного множества источников является универсальным.

2.2. Универсальная мера и её свойства

В 1988 году был предложен метод прогнозирования на основе использования сжатия данных [9]. 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 [9].

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, что доказано в [12].

В 1968 году после открытия универсального кодирования был найден предсказатель, для которого погрешность (2) для источников независимых и одинаково распределённых элементов асимптотически минимальна [13, 14]. Данный предсказатель предложил Кричевский. Он описал следующий предиктор, позволяющий вычислить условные вероятности для следующего элемента ряда:

$$K_0(a|x_1,\dots,x_t) = (\nu_{x_1\dots x_t}(a) + 1/2)/(t + |A|/2), \qquad (3)$$

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 [13].

Рассмотрим тот же пример, что и для предсказателя Лапласа. Пусть $A = \{0,1\}$, $x_1 \dots x_7 = 1010101$. Тогда

$$K_0(x_8 = 0|1010101) = (3 + 1/2)/(7 + 1) = 7/16,$$

$$K_0(x_8 = 1|1010101) = (4 + 1/2)/(7 + 1) = 9/16,$$

Рассмотрим далее обобщение предсказателя Кричевского на случай Марковских источников с какой-то фиксированной памятью $m \geq 0$. Для Марковских источников аналогичная (обобщённая) мера выглядит следующим образом [14]:

$$K_m(x_1, \dots, x_t) = \begin{cases} \dfrac{1}{|A|^t}, & t \leq m, \\ \dfrac{1}{|A|^m} \prod_{\vartheta \in A^m} \dfrac{\prod_{a \in A} (\Gamma(\nu_x(\vartheta a) + 1/2)/\Gamma(1/2)))}{(\Gamma(\overline{\nu}_x(\vartheta) + |A|/2)/\Gamma(|A|/2))}, & t > m; \end{cases} \quad (4)$$

где $\nu_x(\vartheta)$ – число последовательностей ϑ, встречающихся в x, $\overline{\nu}_x(\vartheta) = \sum_{a \in A} \nu_x(\vartheta a)$, $x = x_1 \dots x_t$, а $\Gamma -$ гамма-функция. Данная мера является универсальной для множества Марковских источников связности m. При этом для $m = 0$ мы получаем предсказатель Кричевского (3) и случай независимых и одинаково распределённых элементов.

В качестве примера рассмотрим аналогичную последовательность из предыдущего случая, приняв связность m равную 2:

$$K_2(1010101) = 2^{-2} \frac{1}{1} \frac{3}{2} \frac{1}{2} \frac{1}{2} \frac{5}{2} \frac{3}{2} \frac{1}{2} \frac{1}{6} \frac{1}{1} = \frac{45}{288} = \frac{5}{32}.$$

В 1988 году на базе предсказателя Кричевского была разработана мера R [9], универсальная для множества всех стационарных и эргодических источников, определяется следующим образом:

$$R(x_1, \dots, x_t) = \sum_{i=0}^{\infty} \omega_{i+1} K_i(x_1, \dots, x_t), \quad (5)$$

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

(6). 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:

$$\omega_i = 1/\log(i+1) - 1/\log(i+2) \quad (6)$$

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.

2.3. 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 (5) до элемента $i = t$, где t – это длина ряда, и суммы (5) после этого элемента. Во второй части суммы все слагаемые будут одинаковы и равны $\frac{1}{|A|^t}$, что позволяет вычислить слагаемые меры R после элемента t следующим образом:

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

Видно, что с ростом длины ряда рассматриваемая вторая часть суммы (4) стремится к 0. Таким образом, существенное влияние как на значение ряда (4), так и на сложность его вычисления оказывает только первая часть суммы.

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

2.4. Схема прогнозирования для источников из непрерывного интервала

На практике часто встречаются ряды, элементы которых являются числами из некоторого ограниченного интервала. Таким образом, возникает естественная необходимость модификации описанного выше подхода для источника, принимающего значения из непрерывного ограниченного интервала. Описанные ранее результаты в этом направлении имеют преимущественно теоретический характер [11], а полученные экспериментальные данные относятся к случаю конечного алфавита.

Рассмотрим схему прогнозирования с использованием меры R для источника, принимающего значения из непрерывного ограниченного 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 $\{\Pi_k\}, k \geq 1$ – возрастающая последовательность конечных разбиений интервала [A, B] (назовём этот процесс квантизацией). В предлагаемом подходе разбиение интервалов производилось равномерно, т.е. на равные подинтервалы. Размер каждого подинтервала определяется, как $h = \frac{B-A}{n}$, где n – количество частей разбиения. Обоснование выбора именно такого метода квантизации будет дано

ниже. В общем случае разбиение (величины подинтервалов) может быть произвольным. 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:

$$p^s(x_1, \dots, x_n) = p({x_1}^{[s]}, \dots, {x_n}^{[s]})/\mathrm{L}({x_1}^{[s]} \dots {x_n}^{[s]}) \quad (7)$$

Определим теперь оценку плотности вероятностей r следующим образом:

$$r(x_1 \dots x_t) = \sum_{s=1}^{\infty} \omega_s \mathrm{R}({x_1}^{[s]} \dots {x_t}^{[s]}) \,/\mathrm{L}({x_1}^{[s]} \dots {x_t}^{[s]}) \quad (8)$$

Коэффициенты ω_s определяются формулой (6) и несут роль весовых коэффициентов для случая каждого разбиения из Π_k. Как видно из формулы (8), 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 мы выяснили, что равномерная квантизация даёт наилучшую точность прогноза на реальных вещественных временных рядах, потому она и была выбрана в качестве основной.

Как показано в [11], величина $r(x_1, \dots, x_t)$ является оценкой неизвестной плотности вероятности $p(x_1, \dots, x_t)$, а соответствующая условная плотность

$$r(a|x_1, \dots, x_t) = r(x_1, \dots, x_t a)/r(x_1, \dots, x_t) \quad (9)$$

является оценкой плотности $p(a|x_1 \dots x_t)$. Количество слагаемых в сумме (8) 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).

2.5. Адаптивный метод прогнозирования на базе универсальной меры R

При использования схемы прогнозирования на базе универсальной меры на реальных данных возникает следующая проблема. Пусть имеется какой-либо стохастический процесс, и пусть дан временной ряд, описывающий этот процесс на каком-то фиксированном временном интервале. Тогда при использовании метода на базе меры R мы будем рассчитывать неизвестные плотности, основываясь на всей длине ряда, т.е. учитывая все закономерности, которые может выявить наша мера в рассматриваемой последовательности по всей её длине. Однако реальные процессы (носящие социально-экономический характер или описывающие природные явления) зачастую не обладают свойствами стационарности и эргодичности на протяжении всего времени своего существования, а обладают указанными свойствами только на каких-то локальных участках. Иными словами, со временем закономерности в каких-либо реальных процессах могут меняться и носить только временный и локальный характер. Соответственно, нет смысла учитывать историю какого-либо процесса на протяжении всей его длины равнозначно. В большей мере нас интересует современное время и современные (находящиеся в последней части ряда) закономерности, закономерности, находящиеся в более «далёкой» части нас интересует меньше. Отсюда возникает вопрос выбора необходимой длины

входной последовательности. Таким образом, если мы знаем, что закономерности могут изменяться, то поступим следующим образом.

Пусть дан некоторый временной ряд $x_1 \dots x_t$. И пусть задано число n, определяющее размер текущей выборки, для которой мы будем применять рассматриваемый метод R. Назовём это число n длиной окна. Рассмотрим множество последовательностей следующего вида:

$$x_{t-k\cdot n+1} \dots x_t,$$

где k – натуральное и удовлетворяет неравенству: $0 < k < \left\lfloor \frac{t+1}{n} \right\rfloor$. Другими словами, разделим нашу исходную выборку на множество вложенных подпоследовательностей длины $(k \cdot n)$. Каждую данную подпоследовательность назовем окном.

Определим меру R' для данного подхода следующим образом:

$$R'(x_1 \dots x_t) = \sum_{k=1}^{\left\lfloor \frac{t+1}{n} \right\rfloor} \omega'_k R(x_{t-k\cdot n+1} \dots x_t),$$

где ω'_k – это весовые коэффициенты, представляющее собой некоторое распределение, удовлетворяющее свойству $\sum_{i=1}^{\left\lfloor \frac{t+1}{n} \right\rfloor} \omega'_i = 1$. Суть меры R' состоит в том, что мы вычисляем значение меры R для каждого окна и потом склеиваем полученные значения с некоторыми весами. При этом больший вес даём ближнему к концу ряда окну.

Для источников из непрерывного интервала будем использовать аналогичную схему с величинами r:

$$r'(x_1 \dots x_t) = \sum_{k=1}^{\left\lfloor \frac{t+1}{n} \right\rfloor} \omega'_k r(x_{t-k\cdot n+1} \dots x_t)$$

Данный подход автоматически обеспечивает оптимальность результатов прогнозирования [26], а также существенно уменьшает трудоёмкость самого алгоритма. В дальнейшем будем называть данный подход адаптивным.

Отдельно стоит отметить, что данный адаптивный подход можно очевидным образом обобщить на любой статистический (т.е. дающий оценки плотности вероятностей) метод прогнозирования.

2.6. Оптимизация алгоритма вычисления меры R

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 -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 (6), видно, что с ростом i значение коэффициента w_i стремится к нулю и является небольшим при больших i (при $i > 5: w_{i+1} < 0.05$). 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 (5) 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 (4) 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 последовательности будет равна следующей величине:

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

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 (4) 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:

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

$$= K_m(x_1 \dots x_t) \cdot \frac{\nu_x(\vartheta a) + \frac{1}{2}}{\sum_{a \in A} \nu_x(\vartheta a)} \qquad (10)$$

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.

2.7. Практическая реализация алгоритма прогнозирования на базе меры R

2.7.1. Постановка задачи

Рассмотрим вопросы, связанные с использованием и реализацией описанных алгоритмов прогнозирования. Случай источника, порождающего значения из конечного интервала, является достаточно тривиальным и в отдельных пояснениях не нуждается. К тому же, в реальной практике намного чаще встречаются вещественные ряды. В силу означенных причин, здесь и далее будем рассматривать случай источника, порождающего элементы из непрерывного интервала.

Пусть имеется источник, порождающий элементы временного ряда из некоторого вещественного интервала $[A, B]$. Сам интервал может быть либо известен изначально, либо может быть определён следующим образом. В качестве левой точки возьмём минимальное значение из всех, имеющихся в последовательности $x_1 \dots x_t$, а в качестве правой точки – максимальное.

Пусть также дан сам временной ряд $x_1 \dots x_t$. Наша задача состоит в определении плотности вероятностей элемента x_{t+1}.

2.7.2. Реализация алгоритма на базе меры R

Рассмотрим схему вычисления плотности вероятностей элемента x_{t+1}, т.е. меры r, определяемую на основании (9).

Этап 1. Инициализация разбиения и вычисление значения $r(x_1 \dots x_t)$. Первоначально нам нужно получить последовательность разбиений Π_k нашего интервала. На первом шаге мы проведём одно разбиение (варианты различных методов разбиения будут даны далее), после чего вычислим значение первого слагаемого в сумме (8). Далее, каждый подинтервал разобьём ещё на некоторое количество частей, получив тем самым новое разбиение, позволяющее вычислить уже второе слагаемое из (8). Будем продолжать данный процесс до тех пор, пока не получим ситуацию, когда все элементы ряда находятся в

различных подинтервалах. Таким образом, после вычисления нужного числа слагаемых и после их сложения, мы получим знаменатель в выражении (9).

Теперь остановимся на выборе метода разбиения (или возрастающей последовательности конечных разбиений). В общем случае, алгоритм квантизации может быть любым. Однако в контексте решаемой задачи прогнозирования имеет смысл рассмотреть 2 варианта получения разбиений:

1. Метод равномерной квантизации интервала.

   Суть данного метода состоит в том, что на каждом этапе разбиения мы делим имеющиеся интервалы на равные подинтервалы. Таким образом, если на каждом этапе делить интервалы на n равных частей, то при квантизации -го порядка (т.е. на k-ом шаге) мы получим разбиение из n^k элементов размером по $\frac{1}{n^k}$ каждый. Вид такого разбиения для глубины подразбиений $k = 3$ приведён на рисунке 1.

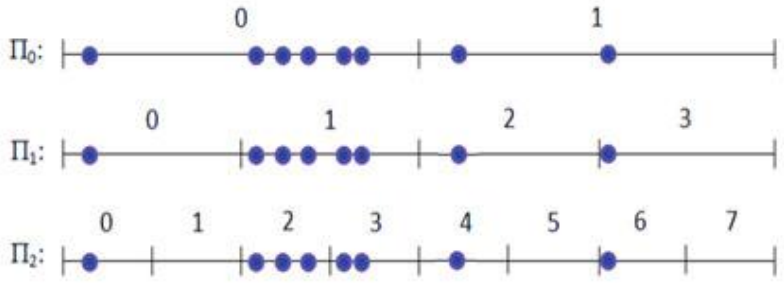

Рисунок 1 – Метод равномерной квантизации интервала.

   Числа над интервалами обозначают их номер $[x_i]^k$, на который и будет заменяться каждое вещественное число x_i.

2. Метод квантизации при условии равномерного распределения элементов.

   В данном подходе мы делим интервалы на каждом шаге таким образом, чтобы каждый интервал разбивался на части, содержащие примерно равное число элементов.

   Данное разбиение наглядно показано на рисунке 2.

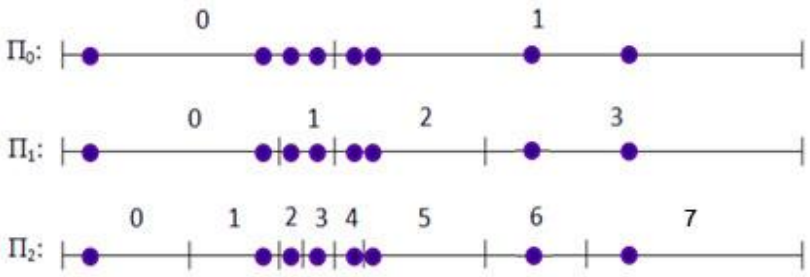

Рисунок 2 – Метод квантизации при условии равномерного распределения элементов.

Этап 2. Вычисление значений плотности вероятностей.

Будем использовать на данном этапе тот же метод квантизации интервала (и то же разбиение), что и на предыдущем этапе. Далее, будем в качестве элемента a подставлять точки из различных частей выбранного разбиения, получая в итоге последовательности вида $x_1 \dots x_t a$, где $a \in \Pi_2$. Точку из части Π_2 выбираем произвольным образом, но в соответствии с правилами: из одной части одну точку. Количество данных последовательностей будет равняться мощности выбранного разбиения. Далее, мы будем вычислять значения $r(x_1 \dots x_t a)$ для каждого такого a. В результате, мы получим по формуле (9) все n условных вероятностей, представляющих собой искомую плотность. Далее, по данному множеству вероятностей можно посчитать прогнозное значение. В простейшем случае в качестве прогнозного значения мы выбирали то, которое имеет максимальную вероятность. Можно выбирать прогнозное значение и другими способами, о которых более подробно будет описано в главе 4.

Следует отметить важный момент, связанный с вычислениями значений $r(x_1 \dots x_t a)$: все данные значения и соответствующие им вероятности независимы и соответственно, могут вычисляться параллельно. Возможность внедрения параллелизма в данный вычислительный алгоритм позволяет существенно сократить время вычислений, которое в случае работы с объёмными временными рядами будет весьма существенным. Предложенный метод прогнозирования был реализован на суперкомпьютере, 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 3.

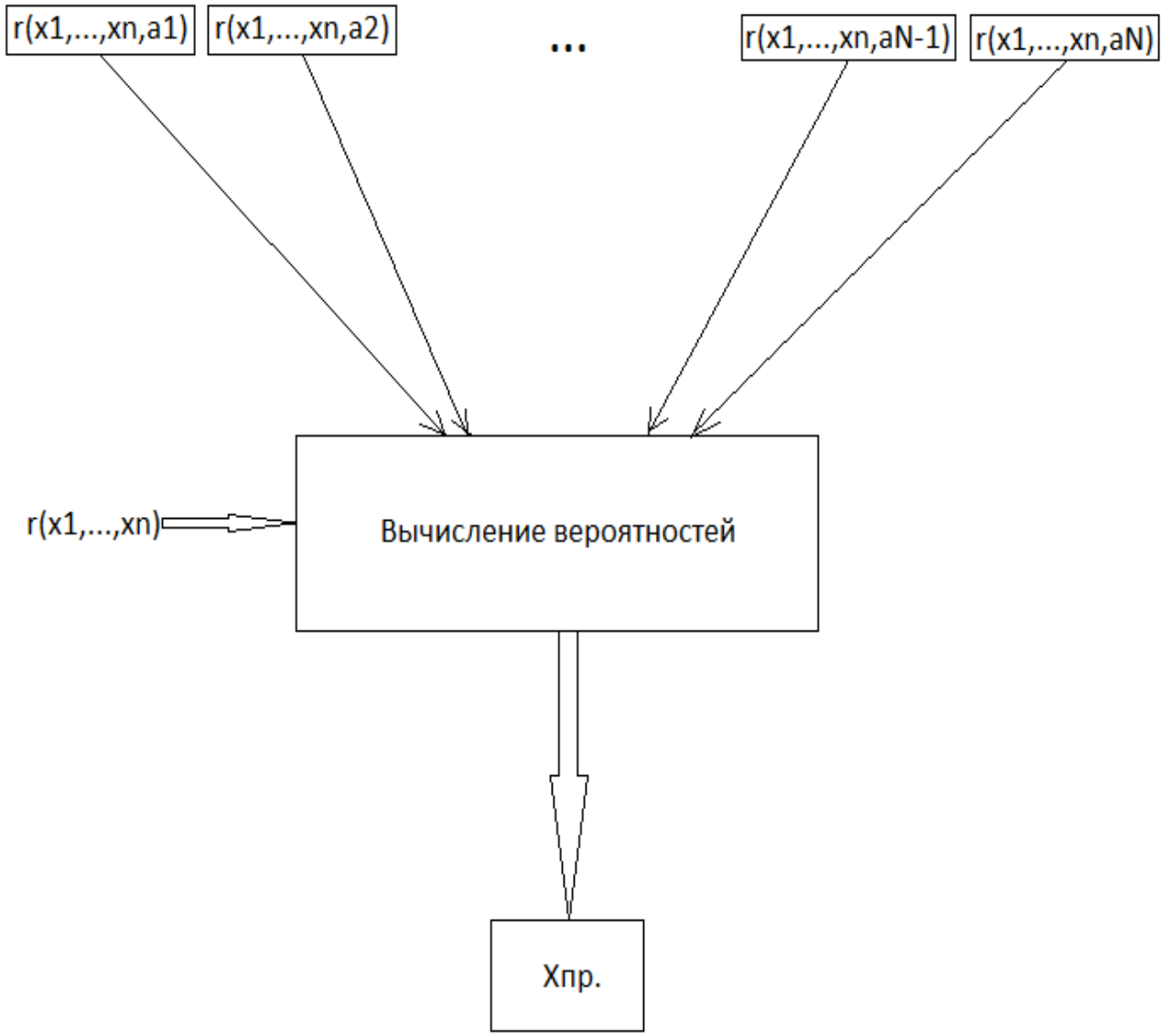

Рисунок 3 – Распараллеливание процесса вычисления прогнозного значения.

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. Проведение вычислений осуществлялось на высокопроизводительном кластере НГУ.

Глава 3. Методы прогнозирования на основе решающих деревьев

3.1. Описание метода 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 (обозначаемый далее как S). При этом на вход подаётся число N (количество элементов в обучающей выборке), число M и некоторый параметр $m \leq M$, называемый глубиной анализа.

Как правило, данный метод применяется для задач классификации и кластеризации [28]. 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 $H(A,S) = -\sum_{i=1}^{n} \frac{|A_i|}{|A|} \log_2 \frac{|A_i|}{|A|}$, 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:

$$Gain(A,q) = H(A,S) - \sum_{i=1}^{q_n} \frac{|A_{q_i}|}{|A|} H\left(A_{q_i}, S\right) \qquad (11)$$

Далее, опишем непосредственно алгоритм построения прогнозного дерева. Данный метод основан на одном из наиболее эффективных алгоритмов

построения деревьев принятия решений, который называется ID3. Цель ID3 состоит в решении задачи кластеризации. Предложенный подход имеет ряд модификаций для его применения к задаче прогнозирования. Алгоритм зависит 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 того, что атрибут S равен значению a, будет при этом определяться следующим образом: $P(S = a|PrevAttr) = 1.0$; $P(x_{t+1} = b \neq a|PrevAttr) = 0.0$, где $PrevAttr$ означает, что значения всех атрибутов, находящихся по дереву выше заданного узла, равны значениям соответствующей ветви.
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 того, что атрибут S равен значению a, будет определяться следующим образом: $P(x_{t+1} = a|PrevAttr) = A_a/|A|$, где $PrevAttr$ определяется таким же образом, как и в пункте 2 алгоритма.
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. Одно из отличий предложенного алгоритма от оригинального ID3 состоит в том, что предложенный вариант позволяет определять условные вероятности для всех возможных вариантов, а не только одно единственное

значение, как это происходит в любом оригинальном алгоритме кластеризации на основе решающих деревьев. Вероятности задаются в пунктах алгоритма 2 и 3 вполне естественным образом.

Также, в предлагаемой реализации данного алгоритма можно ограничивать глубину дерева искусственно – отдельным параметром m_{max}. После достижения заданной в m_{max} глубины дерева выбираем такое a из множества значений целевого признака S, которому равно наибольшее число элементов из A, ставим a в корень дерева и выходим, т.е. фактически выполняем пункт 3 алгоритма. Условные вероятности при этом определяется тем же способом: $P(x_{t+1} = a|PrevAttr) = A_a/|A|$.

Рассмотрим пример построения решающего дерева для более простого, нежели временной ряд, случая прогнозирования – для прогнозирования результатов игры в футбол заданной команды.

Пусть имеются следующие характеристики игры: позиция соперника в турнирной таблице (выше или ниже заданной команды), место игры (дома или в гостях), лидеры команды (на месте или нет), погода (будет дождь или нет). Спрогнозировать требуется результат игры при известных характеристиках игры. Известные данные приведены в нижеследующей таблице.

Таблица 1. Начальные данные для решающего дерева

Соперник	**Играем**	**Лидеры**	**Дождь**	**Победа**
Выше	Дома	На месте	Да	Нет
Выше	Дома	На месте	Нет	Да
Выше	Дома	Пропускают	Нет	Да
Ниже	Дома	Пропускают	Нет	Да
Ниже	В гостях	Пропускают	Нет	Нет
Ниже	Дома	Пропускают	Да	Да
Выше	В гостях	На месте	Да	Нет
Выше	В гостях	На месте	Нет	?

Вычислим значения энтропии относительного целевого признака «Победа»:

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

Теперь вычислим прирост информации для каждого из нецелевых признаков:

1. $Gain(A, \text{Лидеры}) =$

   $= H(A, \text{Победа}) - \frac{3}{7} H(A_{\text{на месте}}, \text{Победа}) - \frac{4}{7} H\left(A_{\text{пропускают}}, \text{Победа}\right) =$

   $= 0.1281$

2. $Gain(A, \text{Играем}) = H(A, \text{Победа}) - \frac{5}{7} H\left(A_{\text{дома}}, \text{Победа}\right) -$

   $- \frac{2}{7} H(A_{\text{гостях}}, \text{Победа}) = 0.4696$

и т.д. для всех 4 свойств.

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

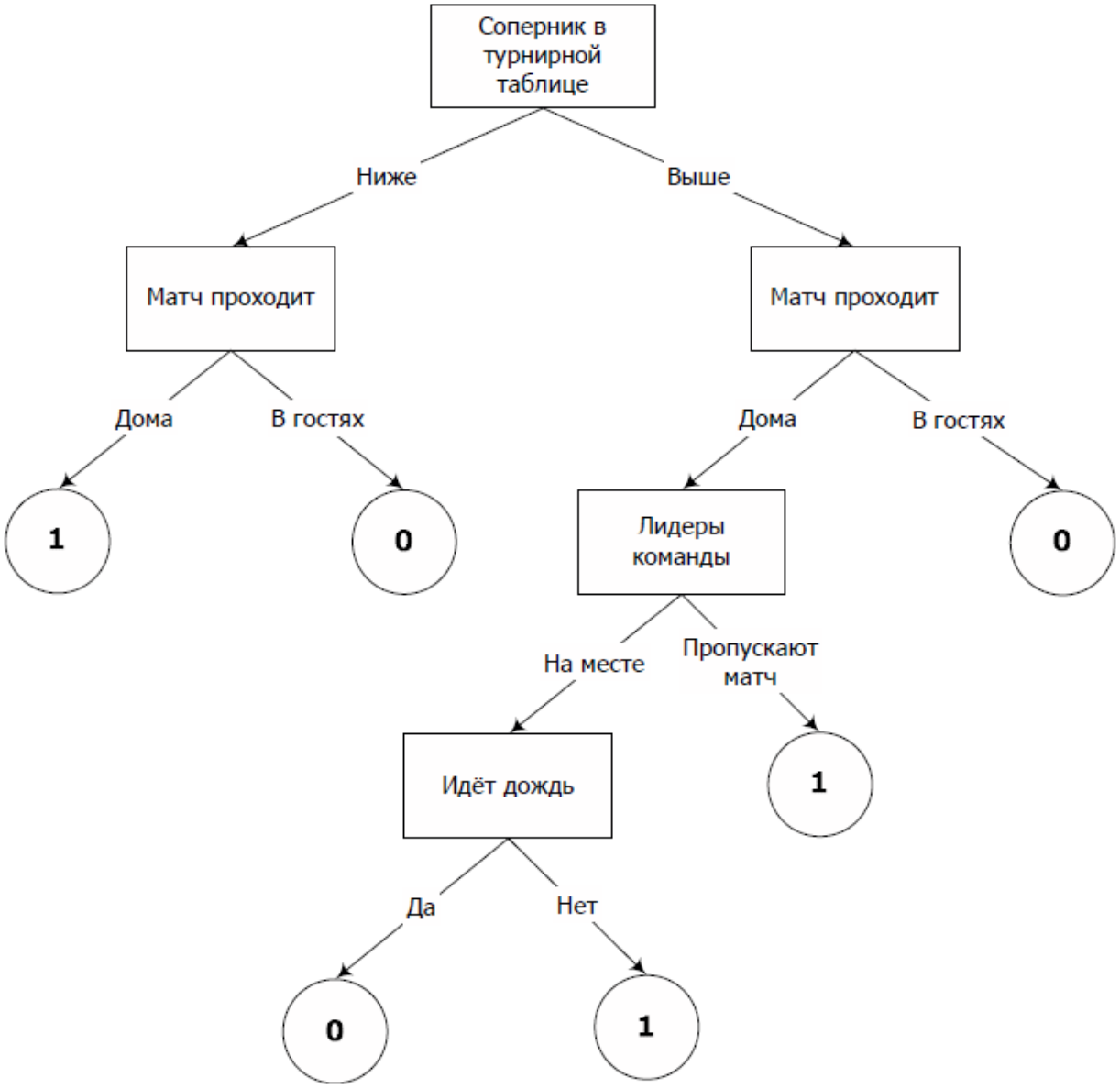

Рисунок 4 – Решающее дерево для прогнозирования игры в футбол.

Опишем разработанную методику применения данного уже модифицированного алгоритма 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. В общем случае, m может быть меньше или равно M. Без ограничения общности будем полагать, что параметр M, определяющий число признаков у каждого элемента выборки, равен m. На первом этапе мы должны произвести разбиение интервала имеющегося в

общем случае вещественного временного ряда. Для этого воспользуемся равномерным методом квантизации, предложенном в разделе 2.7.2. Выбор именно равномерной квантизации связан с особенностью работы решающих деревьев, в которых мы не можем ввести параметр учёта размера подинтервала, как это было сделано для прогнозирования посредством универсальной меры R (посредством введения сигма-конечной меры L, подробнее см. раздел 2.4). Далее, конвертируем имеющийся вещественный ряд в дискретный. После этого определим множество A по правилу: в качестве последнего – целевого – признака возьмём некоторое i-ое значение ряда x_i, а в качестве его $(m-1)$ атрибутов примем $(m-1)$ значений, стоящих в ряду перед i-ым (порядок при этом важен), т.е. элементы $x_{i-m+1}, x_{i-m+2}, x_{i-m+3}, \dots, x_i, i = m, \dots, t-1$. В итоге, получим множество A, состоящее из $(t-m+1)$ элементов, каждый из которых представляет собой вектор, состоящий из m атрибутов. В качестве значения параметра N в данном случае будет выступать число $(t-m)$. На основе полученного обучающего множества A строим дерево в соответствии с предложенным выше алгоритмом. При этом условные вероятности будут определяться следующим образом. Для пункта 2 алгоритма: $P(x_{t+1} = a|x_1, x_2, \dots, x_t) = 1.0$; $P(x_{t+1} = b \neq a|x_1, x_2, \dots, x_t) = 0.0$. Для пункта 3: $P(x_{t+1} = a|x_1, x_2, \dots, x_t) = A_a/|A|$.

Далее берём последний элемент ряда – вектор $(x_{t-m+1,}x_{t-m,+2}, \dots, x_t)$ и подставляем его в построенное дерево. Следуя по дереву с использованием заданного вектора, получаем нужное распределение вероятностей прогнозное значение.

Рассмотрим случаи прогнозирования элементов ряда, принимающего значения из конечного и вещественного множеств. Определим параметр разбиения n. Если элементы ряда принимают значения из некоторого конечного множества, то прогнозирование осуществляется естественным образом. Если же нам дан вещественный временной ряд, значения которого лежат в некотором непрерывном конечном интервале [A, B], то все

вышеуказанные рассуждения справедливы после процедуры квантизации заданного интервала, т.е. определения возрастающей последовательности конечных разбиений интервала [A, B] на N частей $\{\Pi_N\}, N \geq 1$. Условные вероятности при этом определяются по формуле (7). Схема прогнозирования в этом случае становится полностью аналогичной схеме прогнозирования вещественного ряда с использованием универсальной меры R, описанной в разделе 2.4.

Стоит также отметить, что для метода прогнозирования на основе решающих деревьев можно применять адаптивный подход, аналогичный описанному в разделе 2.5 для метода R. При этом будет анализироваться не только весь ряд, но и только конечные его части, обладающие большим весом и соответственно, оказывающие большее влияние на получаемый результат.

3.1.1. Трудоёмкость алгоритма на основе решающих деревьев

При практической реализации предложенной модификации алгоритма ID3 возникает ряд проблем, связанный с трудоёмкостью получающегося алгоритма. Оценим сложность использованного алгоритма в его оригинальном виде. Под трудоёмкостью (или сложностью) в данном случае будем понимать так же, как и в случае с мерой R, асимптотическую оценку максимального количества элементарных операций в процессе работы алгоритма.

Пусть дан некоторый временной ряд $x_1, x_2, \dots, x_t$, для которого требуется определить плотность вероятности $p(x_{t+1} = a | x_1, x_2, \dots, x_t)$. Примем, что $m << t$. Пусть также задан параметр разбиения n. Тогда трудоёмкость вычисления энтропии $H(A, S)$ будет определяться, как произведение мощности множества A (которое мы перебираем) и числа возможных значений целевого атрибута n:

$$T\big(H(A,S)\big) = O(|A| \cdot n)$$

Далее, сложность вычисления прироста информации для каждого узла дерева и каждого атрибута q будет состоять из следующей суммы, которая определяется на основании формулы (11):

$$T\big(Gain(A,q)\big) = T\big(H(A,S)\big) + O\left(|A| \cdot n \cdot T\left(H\big(A_q,S\big)\right)\right) = O(n^2 \cdot |A|^2)$$

Теперь, зная сложность подсчёта прироста информации, мы можем без труда определить трудоёмкость всего алгоритма построения дерева. Она будет равна некоторой сумме, состоящей из вычисления прироста информации вначале для всех m признаков и 1 узла, потом – для $(m-1)$ признаков и n узлов и т.д. до 2 признаков и n^{m-2} узлов. В итоге, сложность работы алгоритма будет равна следующему:

$$T_{all\_tree} = O\Big(|A| + m \cdot T\big(Gain(A,q)\big) + n(m-1) \cdot T\big(Gain(A,q)\big) + \cdots$$
$$+ 2 \cdot n^{m-2} T\big(Gain(A,q)\big)\Big) = O(n^{m-1} \cdot T\big(Gain(A,q)\big) =$$
$$= O(n^{m+1} \cdot |A|^2)$$

Осталось решить одну проблему: в посчитанной сложности участвует параметр $|A|$, который зависит от номера узла в дереве и его глубины. При этом вначале $|A|$ равно t, на глубине 1 оно будет равно в среднем t/n, на глубине 2 – t/n^2 и т.д.. В силу того, что n – фиксированная константа, которая много меньше t, в рамках O-символики можно принять $|A|$ равное t. В итоге, общая сложность будет равна следующему асимптотическому выражению:

$$T_{all\_tree} = O(t^2 \cdot n^m)$$

Данные расчёты полностью подтвердились на практике во время проведения экспериментальных исследований предложенного метода.

Видно, что сложность алгоритма построения дерева растёт квадратично относительно длины ряда, что делает невозможным применение данного метода для достаточно длинных рядов. Однако представленный алгоритм имеет существенное преимущество: при построенном дереве трудоёмкость прогнозирование произвольного числа элементов крайне низкая и равна глубине дерева, т.е. $O(m)$.

3.1.2. Адаптивный метод прогнозирования на основе решающих деревьев

Модифицируем применение предложенного алгоритма с целью уменьшить его сложность. Легко видеть, что число ветвей и листьев построенного дерева будет равно n^m, что при даже небольших значениях параметров n и m представляет собой большое число. При разбиении $n = 10$ (среднее эффективное значение при практических экспериментах) и глубине дерева $m = 5$ число ветвей будет равно 100000. Однако для прогнозирования значений временного ряда на небольшое количество шагов нам вовсе не требуются все имеющиеся ветви. При прогнозировании на k шагов вперёд (как правило, для сохранения высокой точности прогноза k не превышает 20-30) нам требуется не более чем k ветвей. Соответственно, можно изначально строить не все ветви дерева, а только те, которые нам потребуются при прогнозировании рассматриваемого временного ряда. Для этого изменим пункт 6 предложенного в разделе 3.1 алгоритма следующим образом. Помимо обучающей выборки A будем рассматривать элемент X, значение атрибута S которого требуется спрогнозировать. При этом мы знаем значения всех прочих атрибутов данного элемента. Тогда перепишем пункт 6 алгоритма следующим образом: «для атрибута q выберем значение, равное соответствующему значению атрибута q элемента X». Далее, в подпунктах a-c пункта 6 алгоритма в качестве q_i будем понимать значение выбранного атрибута q у прогнозного элемента X. Таким образом, мы построим ровно одну ветку дерева (вместо n^m), что уменьшит сложность прогноза на 1 шаг вперёд в n^m раз. Для прогнозирования следующего элемента проделаем ровно ту же процедуру. При этом в случае прогнозирования более чем 1 элемента, на каждом этапе требуется проверка: не существует ли уже требуемой ветви в дереве. Если она существует, вычислять прирост информации для множества текущих атрибутов не требуется и сложность будет ниже простого произведения сложности прогнозирования одного элемента на число прогнозных элементов.

В общем случае, для предложенной модификации, трудоёмкость будет равняться следующей величине:

$$T_{all\_tree\_adaptive} = O(k \cdot t^2),$$

где k – число элементов, которые требуется спрогнозировать.

3.2. Проблемы и модификации алгоритма решающих деревьев

В силу того, что при большой глубине анализа m и большом алфавите дерево будет слишком сильно разветвляться (в случае построения всего дерева) и трудоёмкость алгоритма, как было показано в разделе 3.1.1, будет расти экспоненциально относительно значения параметра m, введём следующую модификацию алгоритма: зададим параметр m', показывающий максимальную глубину дерева, до которой происходит построение дерева. При достижении заданной в m' максимальной глубины будем следовать пункту 3 алгоритма.

Важно отметить, что критерий ветвления (пункт 4 алгоритма) также является параметром метода и может быть в общем случае произвольным. В частности, вместо критерия прироста информации мы можем использовать так называемый критерий $Gini$. Для его задания нужно определить вначале так называемый индекс $Gini$.

Определение 3. Индекс $Gini(A,S) = 1 - \sum_{i=1}^{Sn} \frac{|A_i|}{|A|}$, где S – целевой атрибут; A_i – элементы из A, у которых атрибут S, имеющий s_n значений, равен i.

Теперь можно определить критерий $GainGini$.

Определение 4. $GainGini(A,q) = Gini(A,S) - \sum_{j=1}^{q_n} \frac{|A_j|}{|A|} Gini(A_j,S)$, где q_n – число возможных значений атрибута q.

В [27] показано, что критерий $Gini$ и критерий прироста информации эквиваленты, потому выбор из них в большинстве случае не приведёт к заметным изменениям эффективности работы приведённого метода.

Трудоёмкость алгоритма построения дерева с использованием критерия $Gini$ асимптотически остаётся той же, что и для критерия прироста информации.

В изначально предложенном критерии прироста информации имеется одна существенная проблема. Он склонен выбирать в качестве основных те атрибуты, которые имеют максимальную волатильность, т.е. наибольшее количество возможных значений. Покажем суть данной проблемы на уже рассмотренном выше примере прогнозирования результата игры футбольной команды.

Добавим к имеющимся 5 признакам (один из которых – целевой) шестой – дату проведения матча. Очевидно, что дата сама по себе не оказывает никакого явного влияния на исход матча и не должна учитываться при прогнозе. По крайней мере, она не должна фигурировать среди первых признаков, по которым происходит ветвление дерева. Посчитаем прирост информации для признака «Дата матча»:

$$Gain(A, \text{Дата}) = H(A, \text{Победа}) - \sum_{j=1}^{7} \frac{\left|A_{\text{Дата}=j}\right|}{7} H\left(A_{\text{Дата}=j}, \text{Победа}\right) =$$

$$= H(A, \text{Победа}) - \sum_{j=1}^{7} \frac{1}{7}(1 \cdot \log(1)) = H(A, \text{Победа})$$

Фактически, мы получили результат, свидетельствующий о том, что прирост информации с использованием признака «Дата» равен изначальной энтропии признака «Победа», т.е. Дата обладает максимально возможным приростом информации, что, конечно же, не так. Да и если бы даже было так – если бы дата оказывала какое-либо влияние на исход матча, – то это бы всё равно не играло положительной роли, т.к. даты всё равно никогда не повторяются. Указанная проблема особо актуальна в случае применения решающих деревьев к задаче прогнозирования, т.к. классический критерий прироста информации присваивает «шумным» признакам неоправданно большой вес. В результате, дерево может не найти многие даже простые

закономерности. Особенно актуальна данная проблема при больших разбиениях (малой сетке) на ряде, обладающем шумами: дерево будет учитывать шумы (записывать их частоты в отдельный элемент разбиения) и в результате существенно ухудшится выявление закономерностей, которыми обладает процесс. При слишком грубом разбиении проблема высокошумных рядов исчезает, но появляется проблема грубой сетки и соответствующей низкой точности прогнозов. В итоге, говоря потенциально о произвольном методе прогнозирования возникает дилемма выбора оптимального разбиения.

Для решения данной проблемы, а также указанной проблемной особенности критерия прироста информации (делающий работу деревьев на «шумных» рядах неэффективной) предлагается другой критерий ветвления. Назовём его $GainRatio$ и определим следующим образом.

Определение 5. Критерий $GainRatio(A,q) = \frac{Gain(A,q)}{SplitInfo(A,q)}$, где $SplitInfo(A,q) = -\sum_{i=1}^{q_n} \frac{|A_{q_i}|}{|A|} \log_2 \frac{|A_{q_i}|}{|A|}$.

Как видно из определения 5, критерий $GainRatio$ получается методом добавления делителя $SplitInfo(A,q)$ к стандартному критерию прироста информации $Gain$. Смысл $SplitInfo$ состоит в уменьшении значения критерия при росте количества возможных значений (вариантов) атрибута q. При росте числа уникальных значений выбранного атрибута (т.е. его «разнообразия») $SplitInfo(A,q)$ увеличивается, уменьшая тем самым значение $GainRatio(A,q)$. Таким образом, проблема «учёта шумов» деревьями решается.

Трудоёмкость алгоритма при введении данной модификации не изменится, так как основная операция при вычислении $SplitInfo(A,q)$ – это операция подсчёта $|A_{q_i}|$, который считается при вычислении $Gain(A,q)$, и соответственно, может быть использован без дополнительных трудозатрат.

3.3 Метод прогнозирования на основе случайного леса

Рассмотрим некоторую модификацию метода прогнозирования на основе решающих деревьев, связанную с учётом принципа статистической случайности. При строительстве одного дерева мы всегда используем всю имеющуюся информацию и выбираем на очередном шаге признак (номер позиции) для ветвления по строго заданному детерминированному алгоритму (по критерию наибольшего прироста информации или критерию $Gini$), от которого и будет зависеть фактически всё дерево и прогноз. Однако, если выбранный критерий ветвления на некотором наборе примеров не производит наилучший выбор признака (номера позиции), то дерево не будет давать сколько-нибудь эффективных результатов. Примером тому является описанная в разделе 3.2 проблема влияния волатильности выбираемого признака на значение критерия прироста информации. Следуя предположению, что данная проблема является не единственной, попробуем рассмотреть методы повышения эффективности работы алгоритмов на основе деревьев принятия решений.

Наиболее эффективный метод улучшения любых алгоритмов на основе решающих деревьев предложил в 2001 году Лео Брейман и Адель Катлер [29-30]. Они предложили алгоритм под названием Random forest (случайный лес). Основная его идея состоит в использовании вместо одного дерева целого ансамбля решающих деревьев, построенных по несколько модифицированному алгоритму. При этом суть борьбы с проблемой неэффективного выбора признаков заключалась в использовании в процессе построения дерева некоторых случайных выборок, что убирает детерминированность построения дерева и делает этот процесс стохастическим. Перейдём непосредственно к описанию алгоритма.

Опишем постановку задачи в общем случае. Пусть нам дано такое же множество объектов A мощности N, составляющих обучающую выборку и обладающих множеством из $(M + 1)$ атрибутов. Множество первых M

атрибутов обозначим за 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 параметр N, число $M \geq 1$, параметры $m \leq M$, $m_q \leq M$, некоторый параметр r $(0 < r \leq 1)$ и параметр числа деревьев в ансамбле $NTrees \geq 1$.

Обобщённый алгоритм прогнозирования на основе случайного леса выглядит следующим образом:

1. На основе исходного обучающего множества A генерируем случайную выборку с повторениями размером $r \cdot N$.
2. На основе сгенерированной выборки строим дерево решений по определённому алгоритму (в общем случае, для данной задачи может использоваться любой алгоритм построения решающих деревьев). При чём, в ходе построения очередного узла дерева из M имеющихся признаков, на основе которых можно разделить дерево, выберем m_q случайных. Решение о разбиении принимается на основе лучшего из $m_q \leq M$ выбранных признаков (т.е. на основе применения критерия ветвления к m_q признакам).
3. Процедура повторяется $NTrees$ раз. Полученные в результате $NTrees$ случайных деревьев объединяются в ансамбль.
4. Для очередного нового элемента определить классы (значение его $(M+1)$-го параметра), используя все $NTrees$ построенных деревьев и выбрать в качестве результирующего класса тот, за который «проголосовало» больше всего деревьев.

Эффективность описанного алгоритма основана на двух основных принципах: методе бэггинга Бреймана [28] и методе случайных подпространств Тим Кан Хо [29].

В общем виде, процесс прогнозирования одного конкретного элемента будет выглядеть так, как показано на рисунке 5.

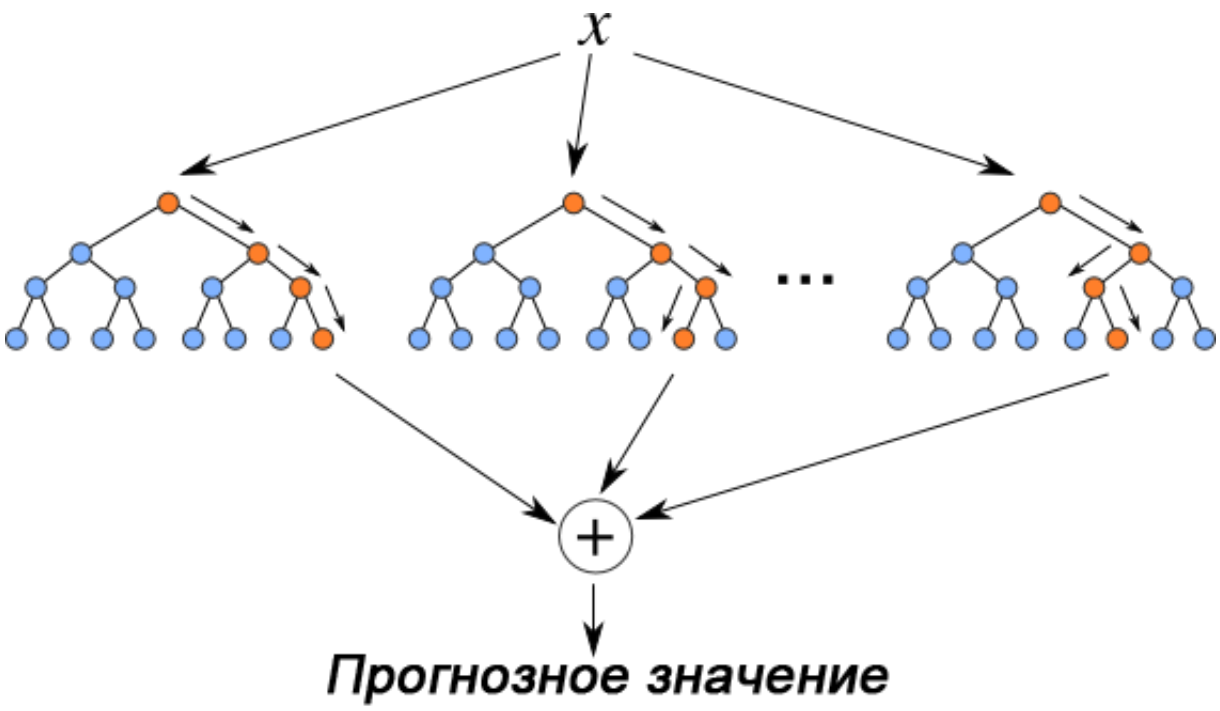

Рисунок 5 – Общая схема работы алгоритма «Случайный лес».

Приведём пример работы описанного алгоритма в общем случае. Пусть имеется обучающее множество A, состоящее из 6 элементов $X_1, X_2, \dots, X_6$. Каждый вектор X_i состоит из 6 атрибутов: 5 атрибутов p_i (входящих во множество признаков Q) и 1 последнего целевого признака S, могущего принимать 3 значения: A, B, C. Атрибуты p_i могут принимать различные множества значений. Примером решающего дерева, сгенерированного алгоритмом «Случайный лес», может быть, например, представленное на нижеследующем рисунке 6.

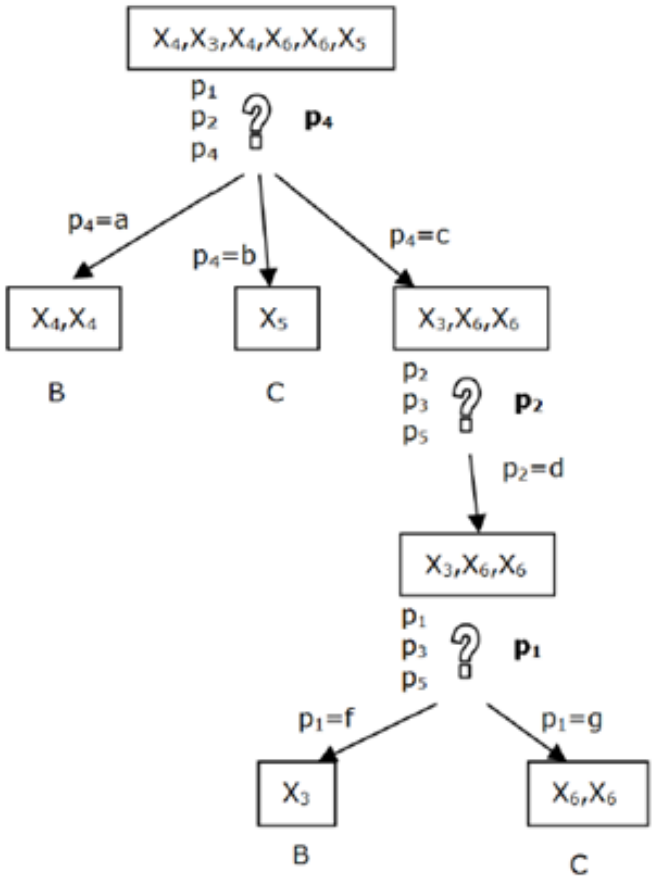

Рисунок 6 – Пример дерева, сгенерированного алгоритмом «Случайный лес».

Параметр $r = 1$, общее число признаков $M = 5$, а число выбираемых признаков $m = 3$. На рисунке 6 мы видим, что на первом этапе в соответствии с шагом 1 описанного алгоритма случайного леса была выбрана случайная выборка с повторениями из 6-элементного множества A: $X_4, X_3, X_4, X_6, X_6, X_5$. Далее, в соответствии с шагом 2 алгоритма, были выбраны 3 случайных атрибута p_1, p_2, p_4, из которых в соответствии уже с выбранным критерием ветвления выбирается один атрибут. В нашем случае выбрали атрибут p_4, по всем возможным значениям которого (а таковых три) дерево разветвляется (в соответствии с алгоритмом построения решающего дерева). Далее ставим в соответствие каждому получившемуся листу соответствующие элементы из исходной вершины, после чего снова происходит выбор случайного набора атрибутов (из оставшихся) и процедура построения дерева продолжается. Так

продолжаем до исчерпания множества атрибутов или до другого признака окончания построения дерева.

Для случая прогнозирования временных рядов описанный выше алгоритм был модифицирован следующим образом. В качестве алгоритма построения решающих деревьев был использован алгоритм, предложенный в разделе 3.1. При этом на листьях каждого дерева получаем не один элемент (значение целевого $(M+1)$ атрибута), а множество условных вероятностей (т.е. плотность вероятности). В качестве элементов обучающего множества A по аналогии с описанием в разделе 3.1 будут выступать вектора вида $(x_{i-m+1}, x_{i-m+2}, x_{i-m+2}, \dots, x_{i+1}), i = m, \dots, t-1$. Параметр N , равный мощности множества A, будет в данном случае равен $(t-m+1)$. Пункт 4 приведённого выше алгоритма изменится следующим образом. Для определения плотности вероятности всего ансамбля будем использовать следующую схему:

$$P(x_{t+1} = a | x_1 \dots x_t) = \frac{1}{NTrees} \sum_{i=1}^{NTrees} P_i(x_{t+1} = a | x_1 \dots x_t), \qquad (12)$$

где $P_i(x_{t+1} = a | x_1 \dots x_t)$ – соответствующая условная вероятность i-го дерева в ансамбле. Как видно из приведённой формулы (12), плотность вероятности определяем, как среднее арифметическое всех плотностей. Предложенный подход не нарушает свойства плотности вероятности и является корректным и эффективным, что было проверено в процессе экспериментальных исследований.

Схема работы алгоритма прогнозирования представлена на нижеследующем рисунке 7.

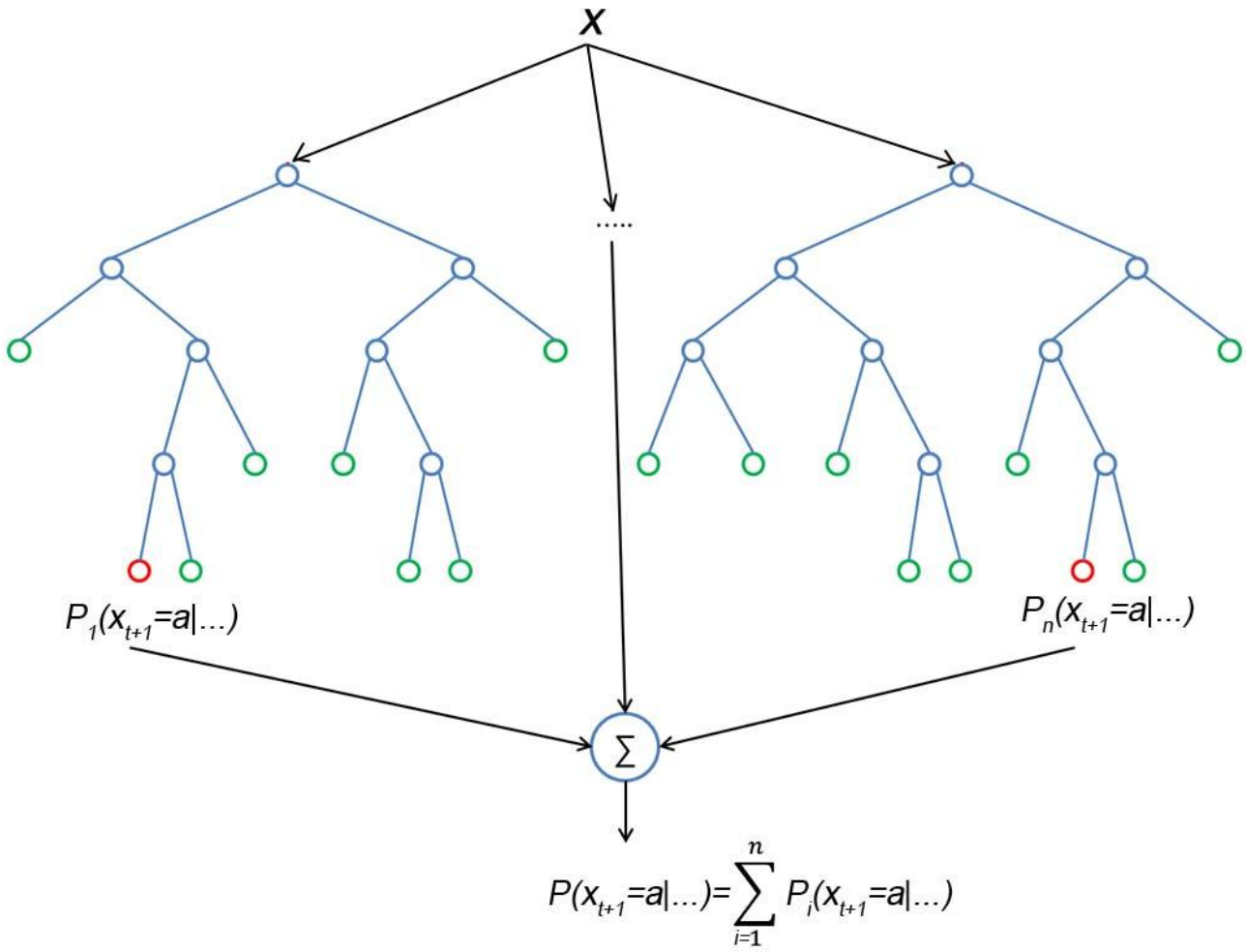

Рисунок 7 – Схема прогнозирования алгоритма «Случайный лес» в применении к временным рядам.

Прочие параметры определялись в процессе экспериментальных исследований экспериментальным путём. Методика их оптимального подбора будет описана в главе с практическими исследованиями предложенного метода.

3.3.1. Трудоёмкость алгоритма на основе случайного леса

Рассмотрим сложность работы предложенного алгоритма на основе случайного леса. В силу того, что рассматриваемый алгоритм использует в своей работе алгоритм решающих деревьев, предложенный в разделе 3.1, будем использовать оценки сложности для алгоритма построения решающего дерева, описанные в подразделе 3.1.1.

Будем рассматривать вариант работы алгоритма при следующих параметрах $r = 1, m_q = M$. При данных значениях параметров вычислительная сложность алгоритма будет наибольшей. Тогда сложность построения одного дерева будет совпадать со сложностью построения оригинального решающего дерева без модификаций. Так как в экспериментальных исследованиях в качестве алгоритма построения решающих деревьев был использован предложенный в разделе 3.1 алгоритм на основе ID3, возьмём в качестве базовой его трудоёмкость. В этом случае сложность работы всего алгоритма построения случайного леса будет определяться следующим образом:

$$T_{rf} = O(T_{all\_tree} \cdot NTrees) = O(NTrees \cdot t^2 \cdot n^m)$$

Данную сложность можно уменьшить, если не строить весь лес сразу, а заниматься построением только тех ветвей у каждого дерева из леса, которые нужны для прогнозирования данного конкретного элемента. Данная идея построения дерева в процессе получения прогнозных значений (плотности вероятности) описана в подразделе 3.1.2. Беря за основу предложенный адаптивный алгоритм построения деревьев в случайном лесу, мы асимптотически уменьшим сложность до следующей величины:

$$T_{rf} = O\left(T_{all\_tree\_adaptive} \cdot NTrees\right) = O(NTrees \cdot k \cdot t^2), \qquad (13)$$

где k – число элементов, которые требуется спрогнозировать. Назовём полученную модификацию алгоритма «Случайный лес» в применении к задаче прогнозирования временного ряда адаптивным алгоритмом «Случайный лес».

3.3.2. Схема вычислений алгоритма на основе случайного леса

Предложенный выше адаптивный алгоритм «Случайный лес» имеет множество независимых друг от друга шагов, который можно вычислять параллельно. Внедрение параллелизма в предложенный алгоритм позволит существенно сократить время вычислений плотности вероятностей и прогнозных значений. Схема распараллеливания приведена на нижеследующем рисунке 8.

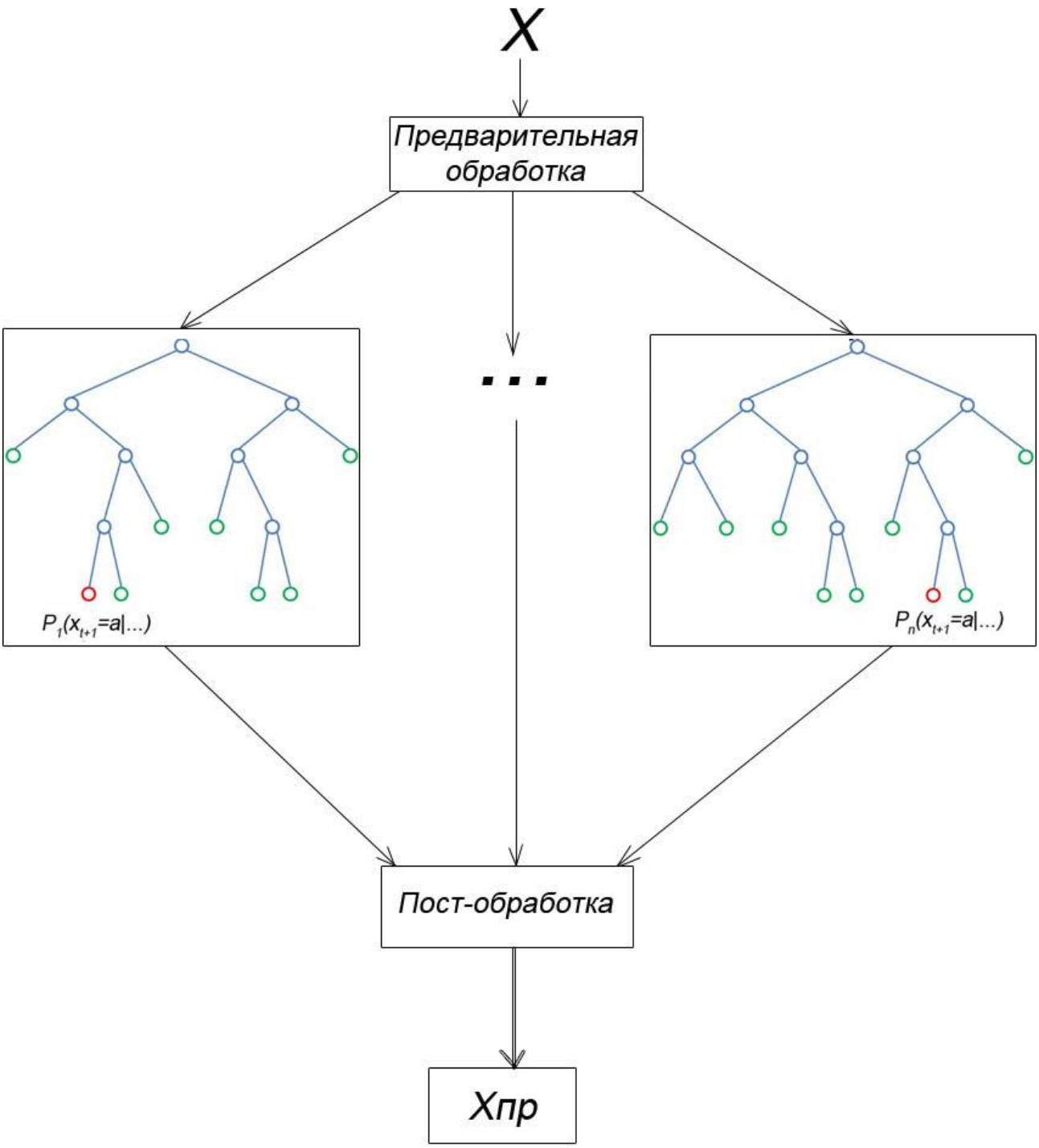

Рисунок 8 – Схема распараллеливания алгоритма «Случайный лес».

На представленном рисунке 8 элемент **X** – это входная последовательность, т.е. имеющийся изначально временной ряд. В общем случае, временной ряд будет вещественным. На этапе «Предварительная обработка» происходит определение интервала значений имеющегося временного ряда, его равномерной разбиение (квантизация) и последующая

дискретизация, т.е. ровно те же действия, что и на этапе предварительной обработки алгоритма построения решающего дерева, описанного в разделе 3.1. Далее на предварительном этапе проводится создание $NTrees$ обучающих множеств A мощности $r \cdot (t - m + 1)$ на основе случайного выбора с повторениями элементов-векторов вида $(x_{i-m+1}, x_{i-m+2}, x_{i-m+3}, \dots, x_{i+1}), i = m, \dots, t-1$. На этом шаг предварительной обработки завершается, все $NTrees$ обучающих множеств, а также атрибуты элемента, который требуется спрогнозировать, направляются в различные подпроцессы. Каждый из подпроцессов отвечает за построение одной ветви (соответствующей прогнозируемому элементу) одного дерева. Всего требуется $NTrees$ подпроцессов. На этапе пост-обработки подпроцессы отправляют посчитанные плотности вероятностей в исходный процесс, который считает итоговую плотность вероятности элемента x_{t+1} по формуле (12). На выходе представленной схемы получаем плотность вероятности и итоговое прогнозное значение.

Реализация предложенного алгоритма и схемы его работы была осуществлена на языке программирования C++. Для распараллеливания был использован интерфейс MPI (Message passing interface) в реализации библиотеки OpenMP. Проведение вычислений осуществлялось на высокопроизводительном кластере НГУ.

Оценим общую получившуюся сложность представленной версии параллельного адаптивного алгоритма «Случайный лес». В стандартном варианте мы использовали значение параметра $r = 1$. Как будет показано в разделе экспериментальных исследований, данное значение является наиболее оптимальным. Среднее число различных значений в исходной выборке при этом будет равно $\sim N/3$, где N – число всех возможных различных элементов, равное $(t - m + 1)$. Данный факт легко доказать в математической статистике, а также проверить экспериментально. В реализации алгоритма использовалась такая схема хранения элементов обучающего множества A, при которой

достаточно было хранить только различные элементы с указанием в отдельном поле объекта-вектора количества их повторений. Предложенная схема даёт существенное преимущество, как в пространственной сложности алгоритма, так и в операционной сложности построения одного дерева, по сравнению с хранением всех выбранных объектов. Так как теперь среднее число элементов обучающего множества в среднем равно $\sim N/3$, а все деревья строятся параллельно, то, несмотря на то, что асимптотическая сложность алгоритма остаётся равной (13), реальное время вычислений по сравнению со стандартной версией адаптивного алгоритма случайного леса сократится в $3 * NTrees$ раз, что является существенным улучшением.

Глава 4. Модификации произвольных методов прогнозирования

4.1. 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.

4.2. Метод группировки алфавита

При прогнозировании временных рядов, порождённых источником, принимающем значения из непрерывного интервала, возникает проблема подбора оптимальных значений параметров разбиения (число n). Выбор оптимального набора параметров показан на практических примерах в главе 5.

Опишем существующую проблему. При больших разбиениях (малой сетке) на примере ряда, обладающего «шумами», выбранный метод прогнозирования будет учитывать «шумы» (т.к. он будет записывать частоты даже немного отклонившихся («шумных») элементов в отдельный интервал разбиения; кроме того, будет множество пустых или низкочастотных интервалов разбиения), и в результате существенно ухудшится выявление закономерностей, которыми обладает анализируемый процесс. При слишком грубом разбиении проблема высокошумных рядов пропадает, однако появляется проблема грубой сетки и соответствующей низкой точности прогнозов, т.к. в большинстве методов мы определяем только интервал, которому принадлежит x_{t+1} и минимальная ошибка прогноза метода остаётся в пределах размера данного интервала, который при малом разбиении будет большим. В итоге, говоря потенциально о произвольном методе прогнозирования, возникает проблема выбора оптимальной величины разбиения.

Кроме того, в случае существенного роста разбиения возникает проблема пустых подинтервалов: при большом разбиении на маленькие подинтервалы во многих подинтервалах будет либо очень мало, либо не будет ни одного элемента исходного ряда. Приведём пример отрицательного влияния больших значений разбиений на результат прогнозирования.

Пусть дан вещественный ряд: 1, 10, 18, 14, 2, 9, 17, 15, 0, 11, 19, 14, 2, 9. Разобьём интервал ряда (от 1 до 19) на 5 частей, каждая из которых имеет длину 4: 0..3, 4..7, 8..11, 12..15, 16..19. В этом случае исходный ряд превратится в целочисленную последовательность: 1, 3, 5, 4, 1, 3, 5, 4, 1, 3. Хорошо виден период, который будет выявлен методом R, и следующим элементом станет подинтервал с номером 5. Точку можно определять, как середину подинтервала, т.е. 17.5. Максимально возможная ошибка при этом не будет превышать половины размера подинтервала, т.е. 2.5. Теперь возьмём разбиение не на 5, а на 10 частей. Длина подинтервала будет равна 2. Получим следующий

ряд: 1, 6, 9, 8, 2, 5, 9, 8, 1, 6, 10, 8, 2, 10. Никаких явных периодических закономерностей в данном ряду не видно. Соответственно, и прогноз в общем случае будет существенно хуже. Графический пример ряда с шумами, но явной закономерностью дан на рисунке 9.

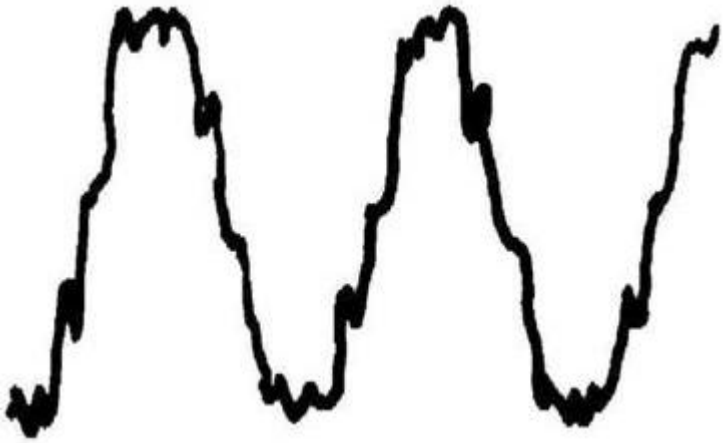

Рисунок 9 – Временной ряд, обладающий закономерностью с шумами.

Для решения описанных проблем в данной работе предлагается использовать так называемый метод группировки алфавита.

Идея метода группировки алфавита впервые была предложена в [16] для решения задачи построения адаптивных кодов, после чего на его основе автором был построен метод для применения его в алгоритме прогнозирования на основе R-меры. Опишем суть работы метода.

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 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 $A = \{\mathrm{i}\}$, где $\mathrm{i} = \overline{1, \dots, 12}$. И пусть дан временной ряд: X(A): 1, 3, 5, 5, 6, 7, 8, 1, 3, 5, 5, 6, 7, 8, 1, 3, 5. Требуется предсказать следующий элемент. 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.

Данный подход, как уже было сказано, решает две проблемы классических методов прогнозирования: проблему высокой сложности метода, а также проблему пустых подинтервалов. Поясним суть второй проблемы: в случае слишком высокого разбиения и, как следствие, маленьких размеров подинтервалов, эти подинтервалы нередко становятся пустыми или мало заполненными, что приводит к влиянию и учёту шумов метода, а также к неэффективности выявления закономерностей в рядах. Модификация группировки алфавита должна приводить к сохранению или улучшению точности работы метода, к которому она применялась, т.к. на первом этапе фактически происходит грубое разбиение (на малое число шагов). Очевидно, что если ряд со случайными шумами имеет какие-либо периодические

закономерности, то существует максимальное разбиение, на котором данная закономерность будет выявлена. При увеличении разбиения метод будет учитывать случайные шумы, присутствующие в любых прикладных процессах, что приводит к ухудшению качества работы метода. Данный факт будет многократно показан в процессе экспериментальных исследований рассматриваемых методов. При этом, если на разбиении n_1 закономерность выявлялась, то на разбиении $n_2 < n_1$ она гарантированно будет выявляться. Соответственно, на малом разбиении метод будет выявлять все закономерности, что выявлял на большом разбиении, и качество его работы будет ограничено лишь размером интервала разбиения. При фильтрации ряда и дальнейшем применении метода в случае наличия дополнительных закономерностей, они будут выявлены и использованы при построении прогноза. Одновременно, если при меньшем разбиении останутся лишь шумы, прогнозное значение будет выбрано в пределах интервала большого разбиения и качество прогноза не уменьшится. Приведённое заключение подтверждается экспериментальными результатами, приведёнными в главе 5.

4.3. 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 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:

$$P(x_{t+1} = a | x_1, \dots, x_t) = \sum_{i=1}^{n} \omega_i \cdot M_i(x_{t+1} = a | x_1, \dots, x_t), \quad (14)$$

где ω_i – весовые коэффициенты, представляющие собой степень значимости i-го метода. Для $\{\omega_i\}$ в целях сохранения корректности метода должно выполняться следующее соотношение: $\sum_{i=1}^{n} \omega_i = 1$. 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 ω_i, можем определять большую значимость для

тех методов, которые на основе предыдущей статистики лучше работают на данном конкретном процессе.

4.4. Моделирование 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 оценивать не плотность распределения и не конкретные вещественные значения ряда, а их поведение, т.е. направление движения значений процесса (т.е. его тренда). К примеру, процесс будет иметь динамику движения вверх, вниз или останется на текущем уровне. При этом фактически мы будем прогнозировать дискретный временной ряд с алфавитом, состоящим всего из трёх элементов. В случае прогнозирования сложных временных рядов, данный подход себя оправдывает.

Рассмотрим реализацию предложенного метода. Пусть имеется вещественный временной ряд $x_1, \dots, x_t$, и требуется спрогнозировать изменение тренда данного ряда (к примеру, текущий тренд будет слабеть, останется на прежнем уровне или будет усиливаться). Построим на основе данного ряда ряд разниц $y_i = x_{i+1} - x_i, i = 1, \dots, t-1$. Далее выберем для интервала [A, B] ряда y_i некоторое разбиение П, соответствующее выбранным направлениям. При этом интервал [A, B] – это минимальный по размеру непрерывный интервал, покрывающий всё множество значений ряда $y_1, \dots, y_{t-1}$. Далее запишем ряд $y_1, \dots, y_{t-1}$ в квантизованном виде, получив тем самым ряд $y_1', \dots, y_{t-1}'$, где y_i' - элемент П, содержащий y_i . Применим выбранными нами метод прогнозирования к ряду $y_1', \dots, y_{t-1}'$ и получим следующее множество условных вероятностей:

$$p(napr = a | y_1 \dots y_{t-1}) = p'(y_t' = a | y_1' \dots y_{t-1}') \frac{L(y_1' \dots y_{t-1}')}{L(y_1' \dots y_{t-1}' a)}$$

где y_i' - элемент П, содержащий y_i , L – мера Лебега, $a \in П$ – одно из направлений (к примеру: 0, 1 или 2). Множитель $\frac{L(y_1' \ldots y_{t-1}')}{L(y_1' \ldots y_{t-1}' a)}$ играет в данном соотношении роль коэффициента, нормирующего вероятности относительно размера интервала направления, что делает получаемые условные вероятности независимыми от размера выбранных интервалов для каждого направления. Направление динамики тренда, имеющее максимальную вероятность и будет искомым решением задачи.

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

4.5. Многомерное прогнозирование

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:

$$x_i = x_{l+i}^1 + x_i^2 \cdot N + x_i^3 \cdot N^2 + \dots + x_i^K \cdot N^{K-1} \qquad (15)$$

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 $A' = \{a|(x_{t+1}^2 \cdot N + x_{t+1}^3 \cdot N^2 + \cdots + x_{t+1}^K \cdot N^{K-1}) \leq a \leq (x_{t+1}^2 \cdot N + x_{t+1}^3 \cdot N^2 + \cdots + x_{t+1}^K \cdot N^{K-1} + N - 1)\}$. 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 $x_1, \dots, x_t$: $A' = \{a|0 \leq a \leq N^K - 1\}$.

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.

Глава 5. Экспериментальные результаты прогнозирования

5.1. Методика экспериментальных исследований

Метод на основе R-меры был реализован с учётом оптимизации, предложенной в разделе 2.5, 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 / 20 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. Разбиение использовалось равномерное. 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 во всех случаях 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 получался ряд разниц между соседними элементами исходного 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 Δ будем называть максимальную разницу между двумя соседними элементами, т.е. фактически разницу между верхней и нижней границей интервала, из которого выбираются прогнозные значения. В этом случае значение Δ определяет максимально возможную ошибку прогноза.

В описанных ниже результатах используются следующие параметры: размер (длина) временного ряда (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).

Важно отметить, что зная величину Δ для исследуемого ряда, а также разбиение ряда n, можно определить статистическую границу точности прогноза следующим образом:

$$Err_{max} = \frac{\Delta}{2 \cdot n} \tag{16}$$

Данная формула справедлива в силу того, что при точном определении подинтервала максимально возможная ошибка будет равна размеру этого

подинтервала, т.е. $\frac{\Delta}{n}$, а средняя ошибка будет равна, соответственно, $\frac{\Delta}{2 \cdot n}$. В дальнейшем под термином «граница точности» будет пониматься величина (16).

5.2. 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 Δ задаёт максимально возможную ошибку и определяется величиной разбиения, поэтому она содержится в таблице с результатами. Размер выборки будет равен 2.5 периодам рассматриваемой функции. Результаты приведены ниже в таблице 1.

Таблица 1. Прогнозирование периодической функции методом R.

Разбиение N	Алфавит n	Глубина анализа m	Δ	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 равен Δ/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 Δ/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. При этом будем рассматривать работу методов на различных длинах ряда. Длину ряда будем измерять в количестве периодов. Данные результаты прогнозирования для режимов on-line и на 10 шагов вперёд приведены в таблице 2 и 3, соответственно.

Таблица 2. Прогнозирование периодической функции методом R и методом 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 методом R и методом 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.

5.3. Прогнозирование ценовых индексов

Применим исследуемые методы прогнозирования к экономическим процессам, связанным с индексами промышленных и потребительских цен. Для этого рассмотрим первоначально временной ряд индекса потребительских цен в период с 01.1990 по 02.2013. Период между измерениями (ТФ) составляет 1 месяц. График данного ряда изображён на рисунке 10. Длина ряда составляет 280 элементов. Прогноз осуществлялся в 2 режимах: on-line и на 10 шагов вперёд. Для случая прогноза на 10 шагов выбирался интервал от 03.2012 до 02.2013. Значение параметра Δ равно 6,541.

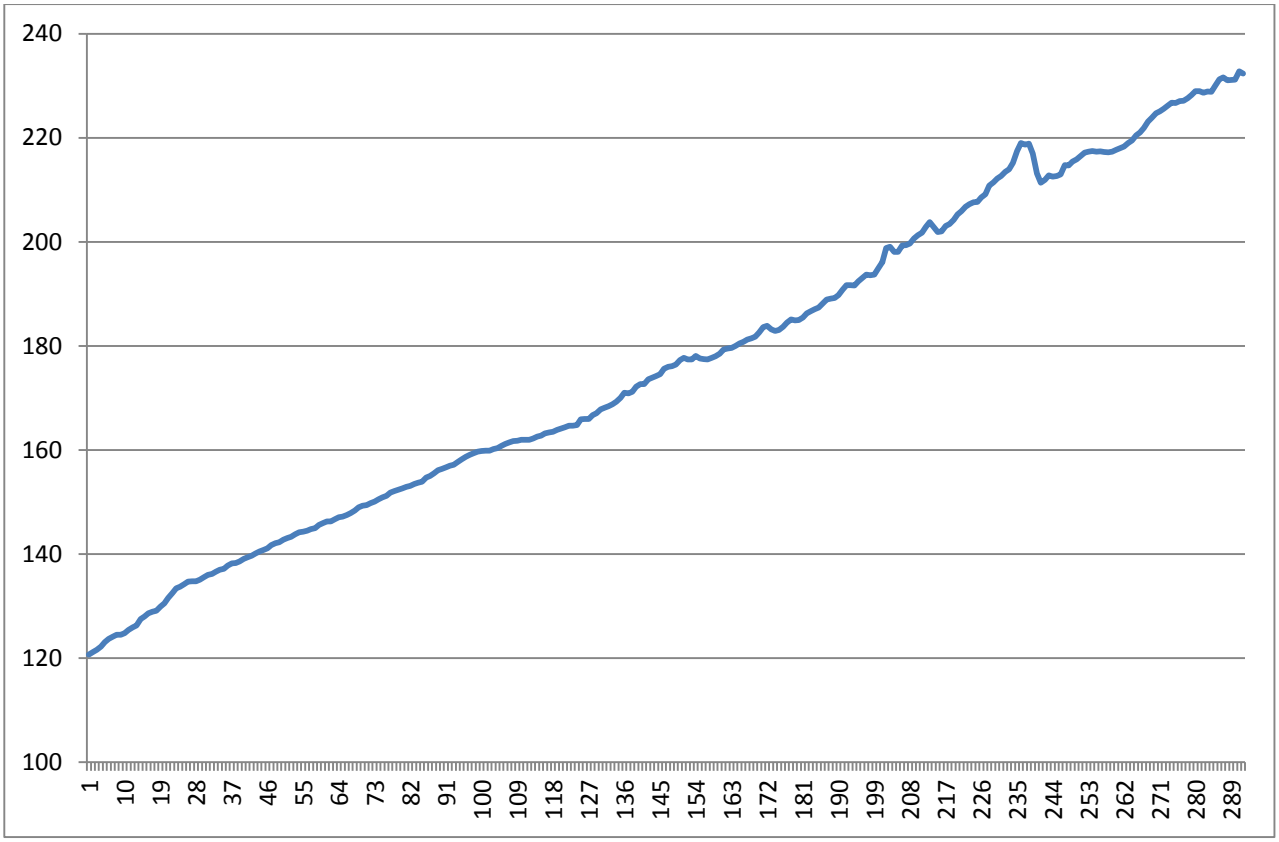

Рисунок 10 – Индекса потребительских цен (CPI). 1990 – 2013гг.

Результаты прогнозирования методом R, а также методом решающих деревьев приведены в таблице 4. В столбце глубины анализа приводилось 2 значения: до черты «/» указана глубина анализа m (применимая как к решающим деревья, так и к R-методу), после черты – максимальная глубина дерева (после которой ветвь обрезается, данный параметр никак не влияет на работу R-метода).

Важно отметить, что во всех случаях экспериментальных исследований применялся метод группировки алфавита, описанный в разделе 4.2. В процессе группировки разбиение прогнозного интервала 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. все группы были равномощны. Для демонстрации эффективности метода группировки в данной серии экспериментальных исследований в результатах указывались параметры группировки. 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 (т.е. мощность групп). Значения параметров a и b при этом подбирались так, чтобы было верно равенство $n = a * b$.

Таблица 4. Прогнозирование индекса потребительских цен (CPI) при разных группировках. 03.2012 – 02.2013.

Разбиение n	Глубина анализа m	Решающие деревья On-line	R-метод On-line	Решающие деревья 10 шагов	R-метод 10 шагов
5 (5*1)	2 / 2	0,628	0,733	1,670	1,671
10 (10*1)	2 / 2	0,602	0,602	0,572	0,573
	2 / 5	0,825		1,717	
20 (20*1)	2 / 2	0,551	0,701	0,718	0,884
	2 / 5	0,609		1,075	
20 (10*2)	2 / 2	0,602	0,635	0,884	0,884
20 (5*4)	2 / 2	0,635	0,701	0,884	0,884
100 (100*1)	2 / 2	0,615	0,655	0,753	1,497
	2 / 5	0,812		1,858	
100 (20*5)	2 / 2	0,642	0,727	0,868	1,497
100 (10*10)	2 / 5	0,867	0,727	0,700	1,497

Из представленных данных видно, что оба метода дают сравнимую друг с другом точность, что говорит о выявлении обоими методами одних и тех же закономерностей, которые использовались при построении прогноза. Однако решающие деревья в некоторых случаях дают лучшую точность, что говорит о их лучшей эффективности при работе на коротких рядах.

Точность методов в зависимости от разбиения практически не меняется и соответственно, от разбиения в данном случае не зависит. Это объясняется тем, что явно линейный тренд данного графика выявляется на любом разбиении, а ошибка возникает в результате присутствия случайных шумов на тренде, которые не учитываются методами ни на одном из разбиений. Точность приближена к границе точности для разбиения $n = 5$.

Важно отметить, что в случае, когда анализируемый ряд не имеет никаких явных или скрытых закономерностей, и методы, соответственно, их не выявляют, смысл их работы заключается во взятии математического среднего значения ряда. В силу того, что мы во всех случаях прогнозируем разницы, то в случае отсутствия выявляемых методами закономерностей, а также в точках бифуркации поведение методов на основе решающих деревьев и на основе универсальной меры сводится к взятию в качестве прогнозных значений основного тренда анализируемого ряда. Данное наблюдение справедливо и было замечено на всех последующих экспериментальных результатах, в которых точность была невысокой. Такое поведение методов в указанных ситуациях является приемлемым с точки зрения качества получаемых прогнозов и позволяет получать сравнительно неплохую точность даже для очень «шумных» рядов с высокой волатильностью. А в случае использования адаптивного R-метода и адаптивного метода решающих деревьев мы для случая рядов без выявляемых закономерностей получим прогнозирование на основе скользящей средней.

Рассмотрим прогнозирование данного ряда для интервала 05.2002 – 02.2003. До данного интервала график является линейным и практически не

имеет каких-либо возмущений, что должно привести к лучшей точности прогноза, нежели на конечной части графика, в которой шумов присутствует больше, чем в центральной. Длина ряда при этом будет составлять 160 элементов. Результаты прогноза приведены в таблице 5.

Таблица 5. Прогнозирование индекса потребительских цен (CPI) при разных группировках. 03.2002 – 02.2003.

Разбиение n	Глубина анализа m	Решающие деревья On-line	R-метод On-line	Решающие деревья 10 шагов	R-метод 10 шагов
5 (5*1)	2 / 2	0,495	0,364	0,684	0,684
10 (10*1)	2 / 2	0,394	0,328	1,255	1,255
	2 / 5	0,394		1,255	
20 (20*1)	2 / 2	0,408	0,310	0,424	0,424
	2 / 5	0,403		0,522	
20 (10*2)	2 / 2	0,375	0,310	0,424	0,424
20 (5*4)	2 / 2	0,375	0,310	0,424	0,620
100 (100*1)	2 / 2	0,408	0,310	0,715	0,715
	2 / 5	0,408		0,715	
100 (20*5)	2 / 2	0,402	0,310	0,715	0,715
100 (10*10)	2 / 5	0,330	0,310	0,846	0,715

Как мы видим из приведённых результатов, точность получаемых прогнозов для обоих методов заметно выше, чем для случая прогноза конечной части рассматриваемого ряда.

Что касается зависимости точности методов от применения модификации группировки алфавита, то из результатов обоих таблиц можно сделать следующие выводы. С применением модификации группировки алфавита методы дают точность, сравнимую со стандартным подходом (при отказе от применения данной модификации). Кроме того, в большинстве случаев модификация группировки алфавита даёт лучшую, нежели стандартный подход, точность, что полностью соответствует теоретическим исследованиям, показанным в разделе 4.2. Наиболее эффективным можно считать группировку вида $n\ (\sqrt{n} \cdot \sqrt{n})$, т.к. она приводит к наибольшему снижению временной сложности метода при сохранении высокой точности прогноза (сравнимой с другими видами группировки или отсутствия оной). Именно такая группировка использовалась во всех дальнейших экспериментальных исследованиях, в которых значение параметра разбиения n было больше или равно 20 (при меньших значениях разбиения проблем, которые бы решала группировка, нет).

Рассмотрим теперь прогнозирование временного ряда индекса промышленных цен США в период с 01.1990 по 02.2013. Период между измерениями (ТФ) составляет 1 месяц. График данного ряда изображён на рисунке 11. Видно, что данный ряд существенно сложнее, чем предыдущий. Длина ряда составляет 280 элементов. Прогноз осуществлялся в 2 режимах: on-line и на 10 шагов вперёд. Для случая прогноза на 10 шагов выбирался интервал от 03.2012 до 02.2013. Значение параметра Δ равно 8,614.

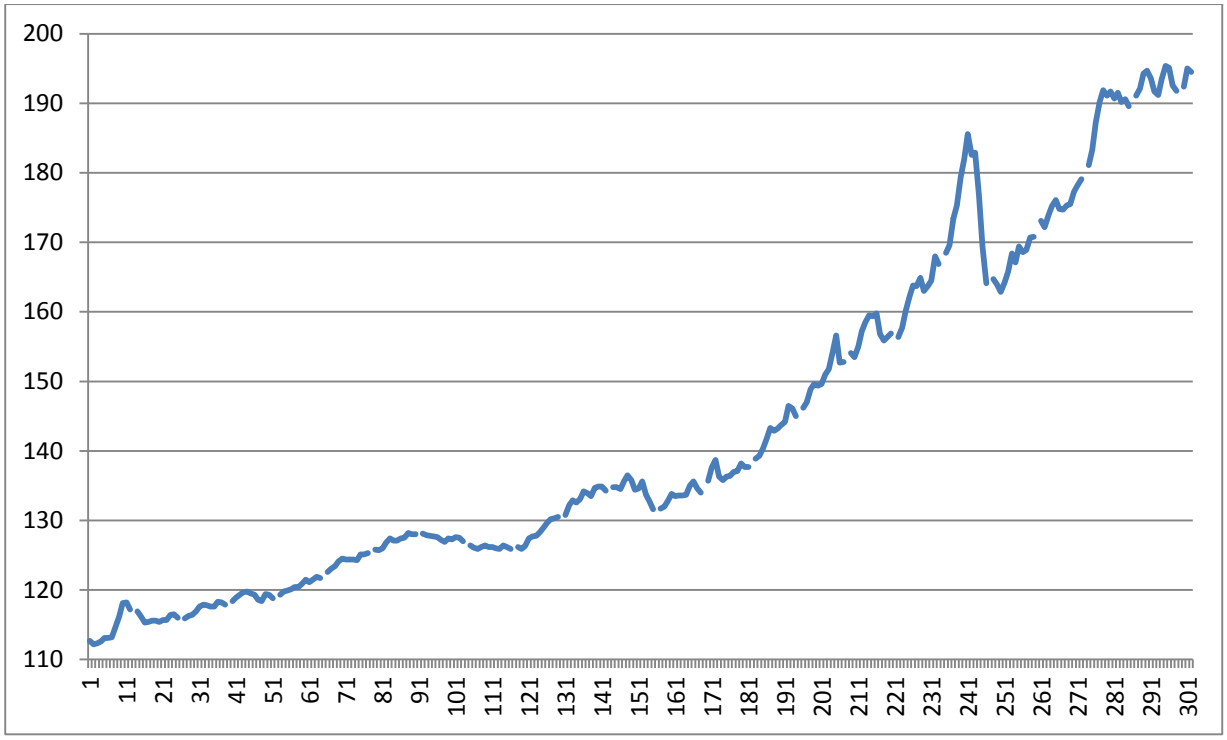

Рисунок 11 – Индекса промышленных цен (PPI). 1990 – 2013гг.

Результаты работы метода на основе решающих деревьев, а также R-метода приведены в таблице 6.

Таблица 6. Прогнозирование индекса промышленных цен (PPI) при разных группировках. 03.2012 – 02.2013.

Разбиение n	Глубина анализа m	Решающие деревья On-line	R-метод On-line	Решающие деревья 10 шагов	R-метод 10 шагов
5 (5*1)	2 / 2	0,954	0,978	1,87	3,276
10 (10*1)	2 / 2	0,996	0,978	1,363	3,276
	2 / 5	1,322		1,793	
20 (20*1)	2 / 2	1,07	0,910	0,935	1,363
	2 / 5	1,125		1,076	

20 (10*2)	2 / 2	0,954	0,907	1,076	1,311
20 (5*4)	2 / 2	0,953	0,894	1,076	1,324
100 (100*1)	2 / 2	1,013	0,910	1,204	1,122
	2 / 5	1,296		1,488	
100 (20*5)	2 / 2	1,113	0,911	1,030	1,055
100 (10*10)	2 / 5	1,006	0,904	2,382	1,074

Полученные результаты сравнимы с теми, что были получены для индекса CPI. В этом примере точность работы решающих деревьев в режиме прогнозирования на 10 шагов вперёд выше, чем для R-метода. Это происходит также из-за небольшой длины ряда, учитывая, что будущие прогнозы длинных рядов показывают примерно сходную точность работы обоих методов.

Точность работы методов практически не зависит от разбиения и приближается к пределу точности для разбиения $n = 5$. Зависимость погрешности работы методов от вида группировки алфавита полностью соответствует сделанным выше выводам.

5.4. Прогнозирование цен на энергоносители в США

Рассмотрим результаты прогнозирования цен на топливо в США. В таблице 7 приведены результаты прогнозирования данного ряда с периодом 1 неделя в интервале с 01.01.2002 по 01.10.2013. Длина ряда равна 615 элементам. Значение Δ для ряда цен на энергоносители равняется 0.832. График ряда приведён на рисунке 12.

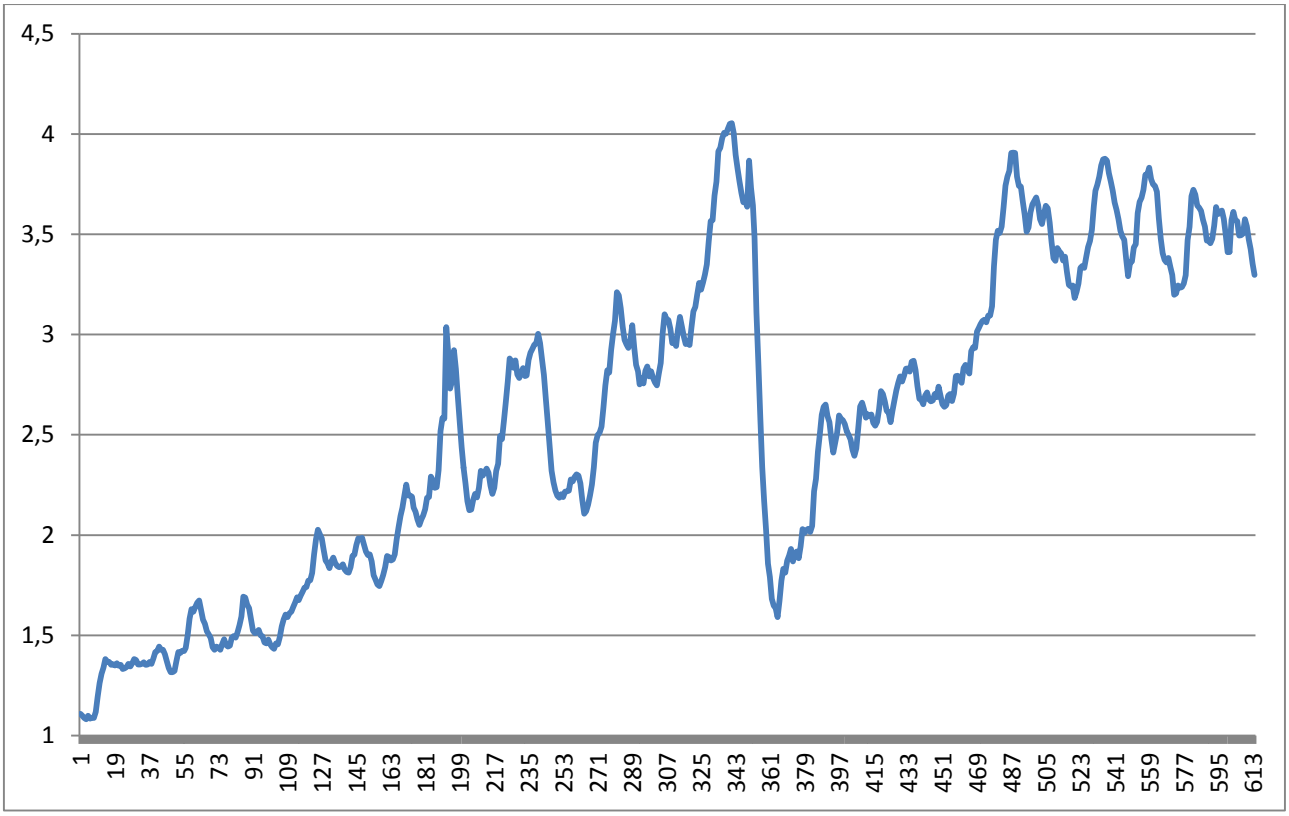

Рисунок 12 – График цен на энергоносители (ТФ: 1 неделя).

В таблице 7 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 8 приведены аналогичные данные прогноза для решающих деревьев в двух режимах.

Таблица 7. 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 8. 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 Δ, что говорит о достаточно высокой точности предлагаемого подхода, т.к. средняя ошибка при выборе случайного значения из интервала будет равна $\Delta/2$. При этом точность метода слабо зависит от используемого разбиения, если число разбиения n больше 10. 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.

5.5. 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 (14). 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 9.

Таблица 9. 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 9 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.

5.6. Прогнозирование объёмов промышленного производства в 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 Δ для ряда цен на энергоносители равняется 7.2945. График данного ряда приведён на рисунке 13.

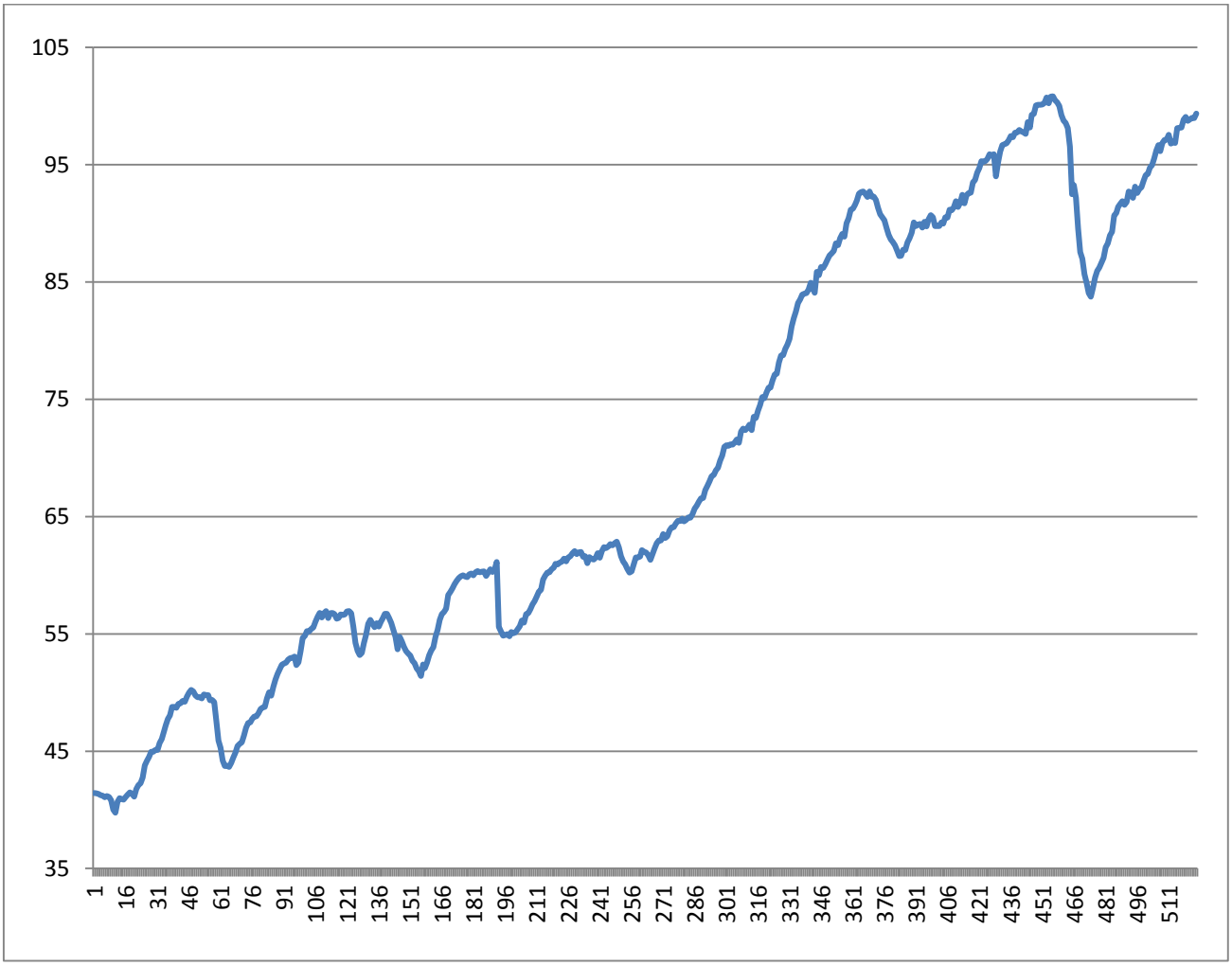

Рисунок 13 – График объёма промышленного производства (ТФ: 1 месяц).

Рассмотрим прогнозирование данного ряда посредством метода R. Прогнозы осуществлялись в следующих 3 режимах: on-line, on-line с использованием усреднения алфавита, на 20 шагов вперёд с использованием усреднения алфавита. Осуществлялось измерение абсолютной (до косой черты) и относительной (после косой черты) погрешности методов. Полученные результаты приведены в таблице 10.

Таблица 10. Прогнозирование объёма промышленного производства США. R-метод.

Разбиение n	Глубина анализа m	R-метод. Усреднение. on-line	R-метод on-line	R-метод. Усреднение. 20 шагов
5	5	0.35916 / 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 10 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 Δ. Метод усреднения показывает свою эффективность только на малых разбиениях, что совпадает с результатами, полученными при прогнозировании ряда цен на энергоносители. 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 их выявит, и итоговые ошибки прогноза будут меньше.

Рассмотрим прогнозирование рассматриваемого ряда с использованием решающих деревьев в стандартном варианте (без модификаций). Результаты приведены в нижеследующей таблице 11.

Таблица 11. Прогнозирование объёма промышленного производства США. Решающие деревья.

Разбиение n	Глубина анализа m / максимальная глубина дерева	Решающие деревья on-line	Решающие деревья 20 шагов
5	5 / 5	0.63855 / 0.0875	6.49907 / 0.8909
10	2 / 2	0.39077 / 0.0536	2.66946 / 0.3659
	5 / 5	0.32371 / 0.0444	1.34473 / 0.1843
20	2 / 2	0.40225 / 0.0551	3.10434 / 0.4256
	5 / 5	0.45208 / 0.0619	2.46412 / 0.3378
	5 / 2	0.46487 / 0.0637	0.50068 / 0.0686
50	2 / 2	0.42505 / 0.0583	1.198828 / 0.2725
	5 / 2	0.39845 / 0.0546	0.913448 / 0.1252

Из полученных результатов видно, что решающие деревья в режиме on-line дают чуть худшие результаты, чем R-метод, хотя результаты сравнимы. В режиме прогнозирования на 20 шагов вперёд решающие деревья выигрывают при разбиении на 50 частей. В остальных случаях также дают несущественно худшие результаты.

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 и приближена к границе точности прогноза для разбиения $n = 10$.

5.7. 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 14-17.

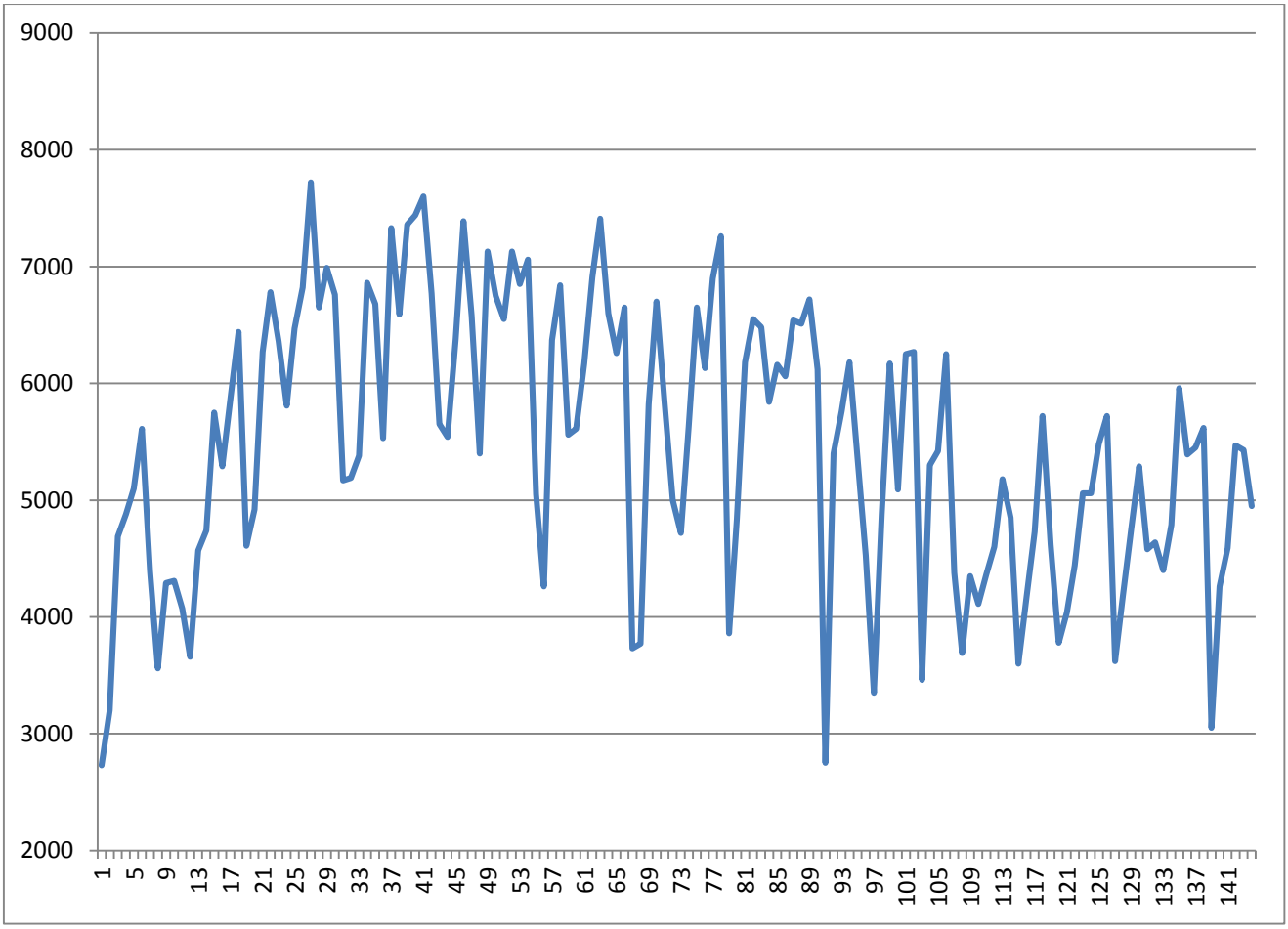

Рисунок 14 – Ряд Industry.

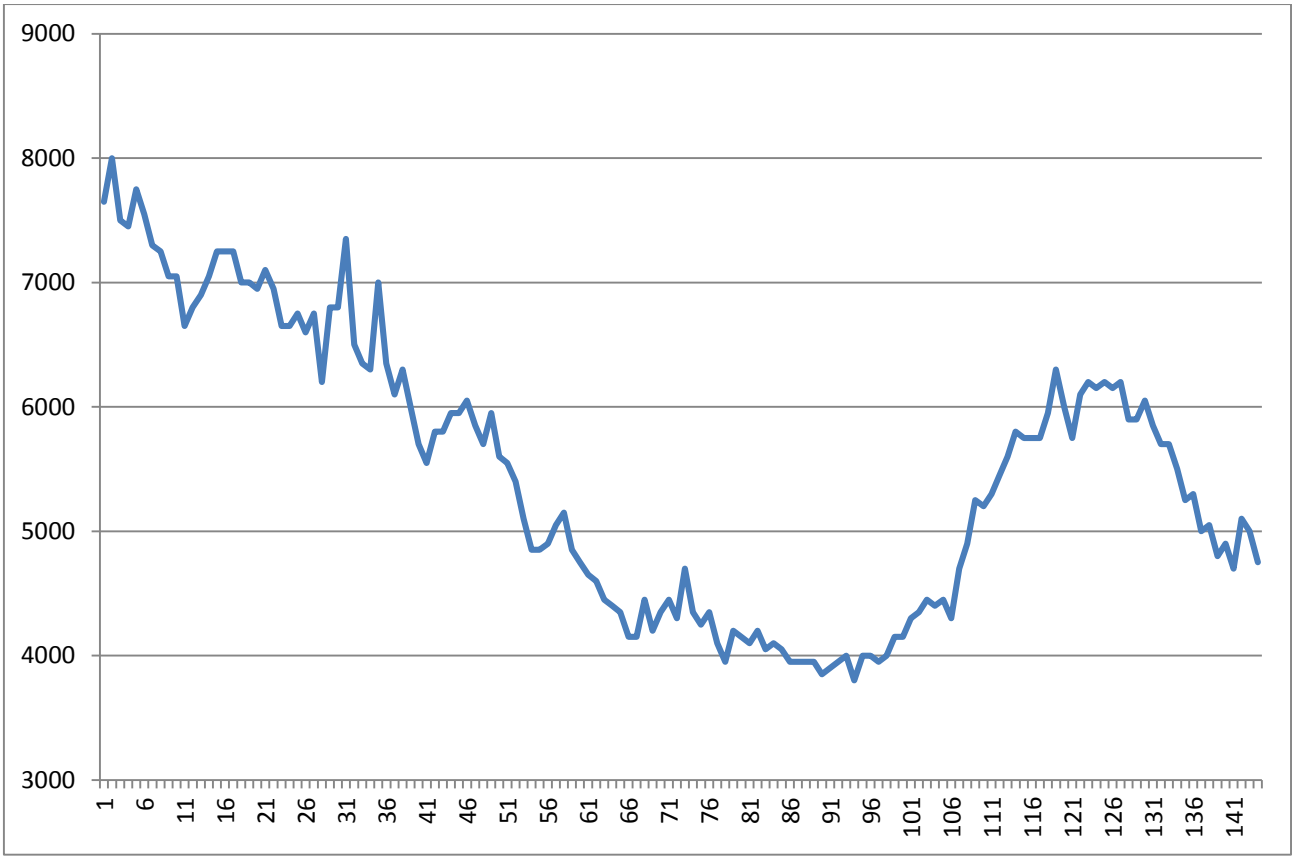

Рисунок 15 – Ряд Finance (1).

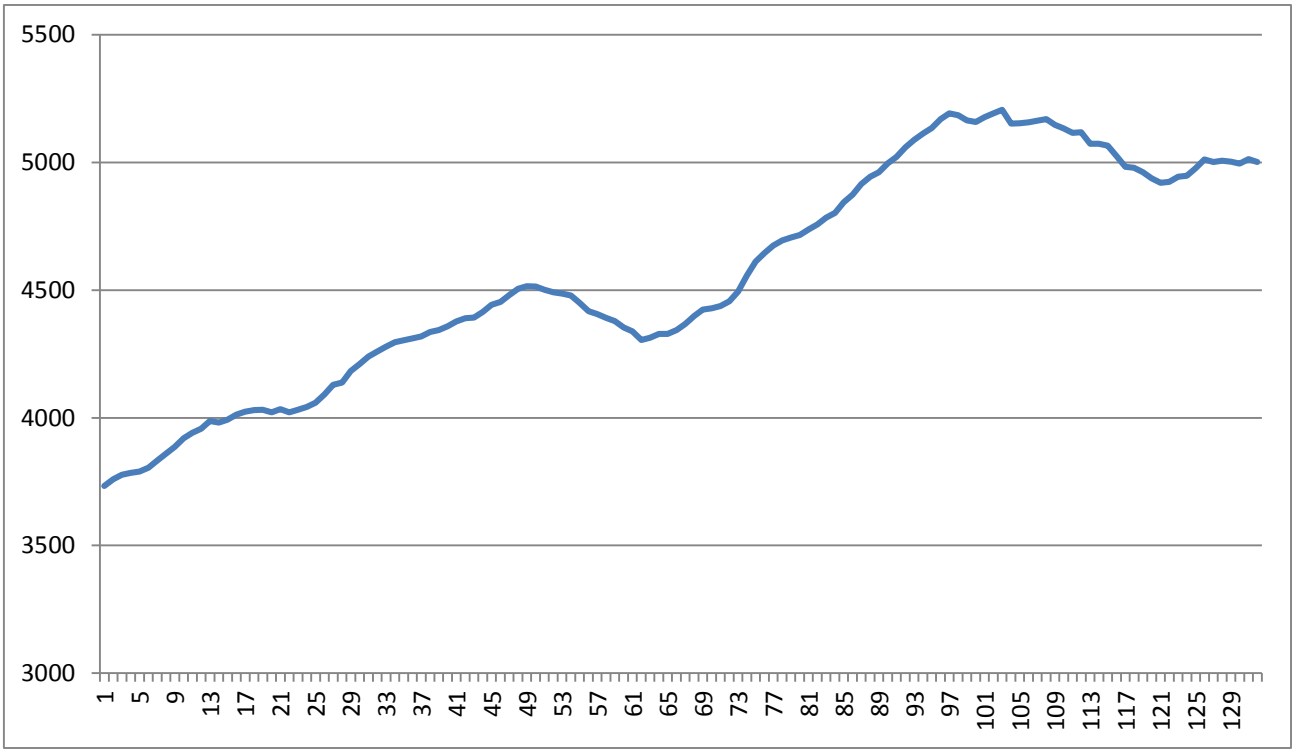

Рисунок 16 – Ряд Finance (2).

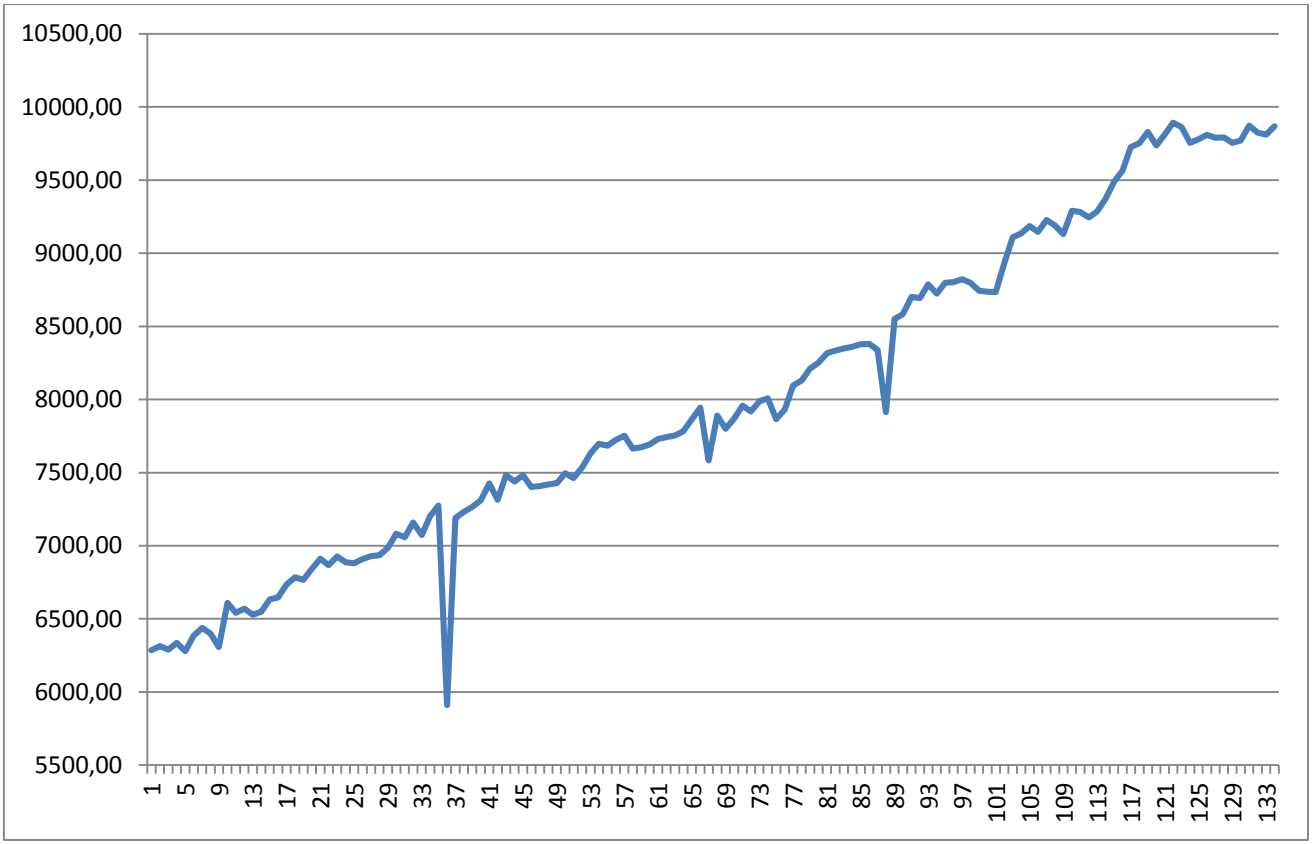

Рисунок 17 – Ряд Demographic.

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 анализа 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 12 – 14.

Таблица 12. Прогнозирование экономических временных рядов США. R-метод и R-метод усреднение.

Временной ряд	Размер выборки L	Разбиение n	Δ	R-метод on-line	R-метод усреднение
Industry	144	10	6050	768.06	703.53
Finance (1)	144		1550	167.5	165.44
Finance (2)	132		118	21.74	19.59
Demographic	134		2642	82.19	62.58
Industry	144	20	6050	718.61	706.52
Finance (1)	144		1550	162.64	164.48
Finance (2)	132		118	26.44	21.07
Demographic	134		2642	53.56	53.46

Таблица 13. Прогнозирование экономических временных рядов США. R-метод усреднение и Autobox.

Временной ряд	Размер выборки L	Разбиение n	Δ	R-метод усреднение	Autobox
Industry	144	10	6050	703.53	340.72
Finance (1)	144		1550	165.44	680.49

Finance (2)	132		118	19.59	76.12
Demographic	134		2642	62.58	122.08
Industry	144	20	6050	706.52	340.72
Finance (1)	144		1550	164.48	680.49
Finance (2)	132		118	21.07	76.12
Demographic	134		2642	53.46	122.08

Таблица 14. Прогнозирование экономических временных рядов США. Методы ForecastPro и PP-Autocast.

Временной ряд	Размер выборки L	Разбиение n	Δ	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 12, 13 и 14 видно, 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.

5.8. Прогнозирование курсов валют

5.8.1. Прогнозирование стандартных курсов валют

Рассмотрим прогнозирование курсов валют методом решающих деревьев. В качестве исследуемого ряда возьмём временной ряд валютной пары английский фунт стерлингов / доллар США (GBP/USD) в интервале с 01.03.2014 по 12.05.2014 с периодом (ТФ) в 1 час. График данного ряда приведён на рисунке 18. Длина ряда составляет 1424 элемента. Значение параметра Δ равно 0.01020.

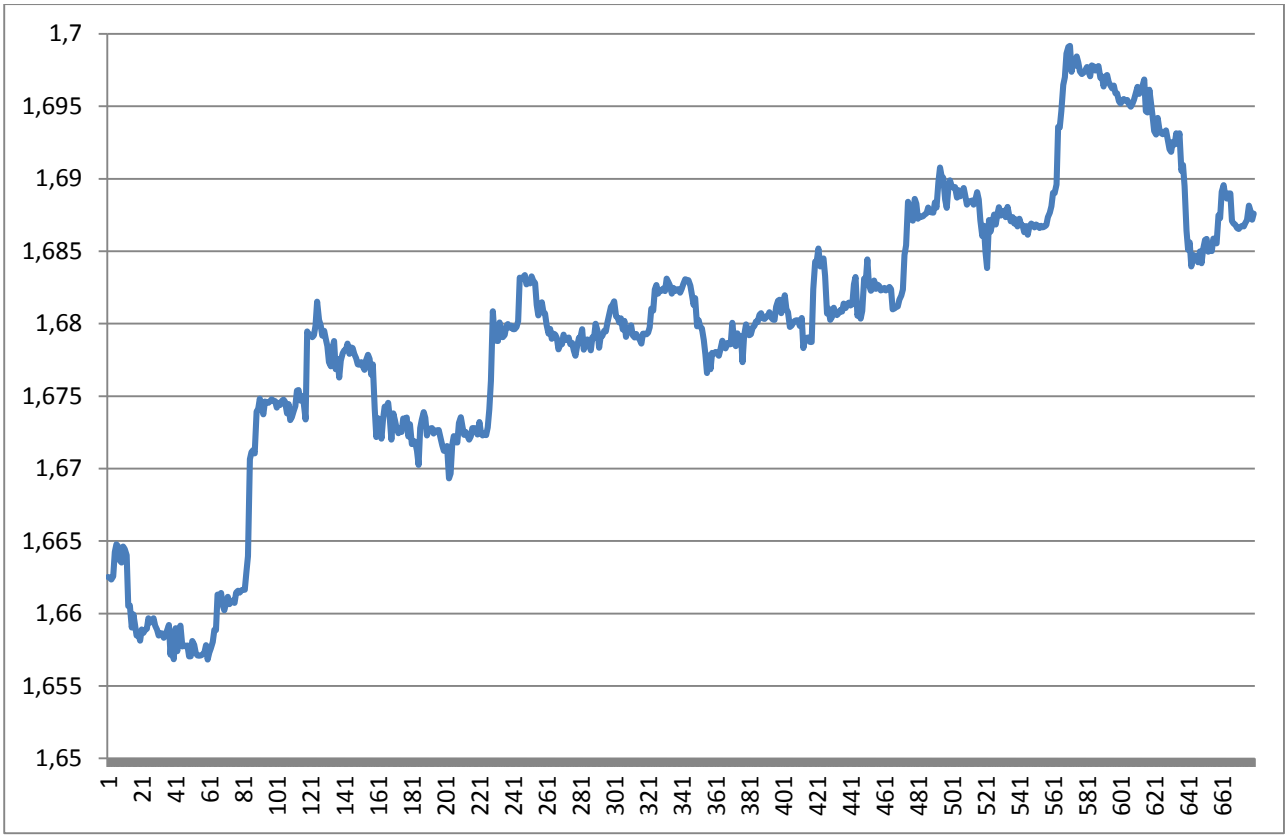

Рисунок 18 – График курса фунт стерлингов/доллар США (ТФ: 1 час).

Рассмотрим прогнозирование данного ряда в режиме on-line и в режиме на 20 шагов вперёд. Результаты прогнозирования данного ряда приведены в таблице 15. При этом в результатах, кроме абсолютной ошибки прогноза, содержится также и относительная ошибка, которая определяется следующим образом: отношение абсолютной ошибка к значению параметра Δ. Таким образом, по относительной ошибке мы можем определить, находит ли метод какие-либо закономерности в ряду и если да, то какова эффективность.

Таблица 15. Прогнозирование курса евро/доллар (ТФ: 1 час)

Разбиение n	Глубина анализа m / максимальная глубина дерева	Решающие деревья on-line	Решающие деревья 20 шагов
5	5 / 5	0.0006465 / 0.0634	0.004980 / 0.4882

10	2 / 2	0.0005105 / 0.0500	0.001261 / 0.1236
	5 / 5	0.0005615 / 0.0550	0.000937 / 0.0918
20	2 / 2	0.0005490 / 0.0538	0.001312 / 0.1286
	5 / 5	0.0006725 / 0.0659	0.001312 / 0.1286
50	2 / 2	0.0004479 / 0.0439	0.001261 / 0.1236
	5 / 2	0.0005399 / 0.0529	0.004935 / 0.4838

Из приведённых результатов видно, что решающие деревья эффективно находят закономерность в имеющемся ряду при разбиении 10 элементов. При разбиении 10 элементов значение относительной ошибки равно ~0.05, что в соответствии с формулой (16) является статистическим пределом максимально возможной точности прогноза. При большем, чем на 10 частей, разбиении точность не увеличивается, а в некоторых случаях даже ухудшается. Это объясняется большей волатильностью получаемого ряда и учёту исследуемым методом появляющихся в ряду шумов.

Рассмотрим прогнозирование той же валютной пары, но на периоде в 1 сутки. Интервал при этом возьмём с 29.06.2012 по 01.03.2014. Длина ряда составит, соответственно, 614 элементов. График описанного ряда показан на рисунке 19. Значение параметра Δ равно 0.04505.

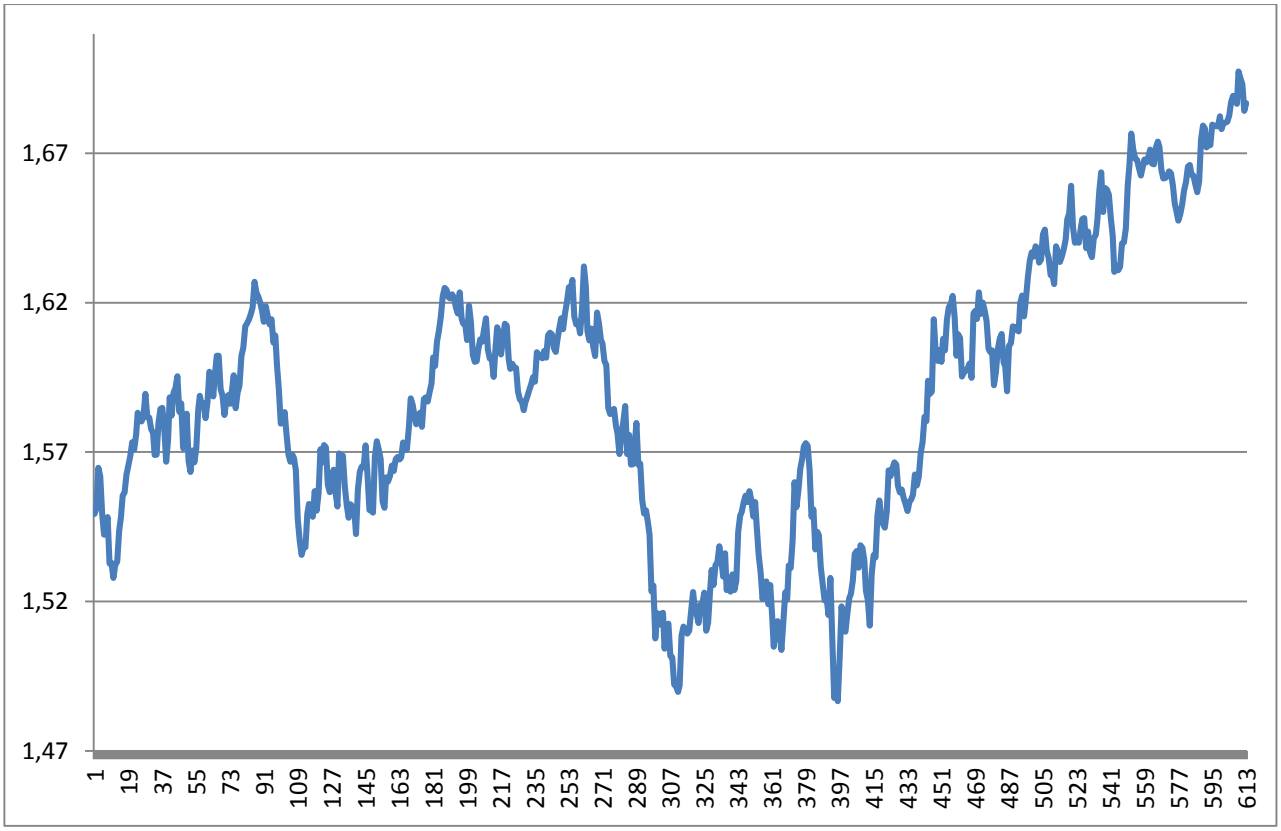

Рисунок 19 – График курса фунт стерлингов/доллар США (ТФ: 1 день).

Прогнозирование выполнялось по тем же правилам, что и для случая периода в 1 час. Результаты представлены в нижеследующей таблице 16.

Таблица 16. Прогнозирование курса евро/доллар (ТФ: 1 день)

Разбиение n	Глубина анализа m / максимальная глубина дерева	Решающие деревья on-line	Решающие деревья 20 шагов
5	5 / 5	0.003055 / 0.0678	0.006691 / 0.1485
10	2 / 2	0.003003 / 0.0666	0.001746 / 0.0388
	5 / 5	0.004090 / 0.0907	0.017460 / 0.3875
20	2 / 2	0.004466 / 0.0991	0.039861 / 0.6851
	5 / 5	0.005251 / 0.1165	0.027585 / 0.6123

50	2 / 2	0.005073 / 0.1126	0.033228 / 0.7375
	5 / 2	0.005832 / 0.1295	0.045211 / 1.0036

Как видим из результатов, в этом случае сохраняются те же тенденции эффективной работы метода на основе решающих деревьев. Наибольшая точность достигается при разбиении, равном 10. При большем разбиении точность заметно падает (растёт ошибка прогноза). При этом, как и в случае с меньшим периодом, работа решающих деревьев в режиме прогнозирования на 20 шагов вперёд даёт некоторое ухудшение точности, но тем не менее, она остаётся на высоком уровне. Отдельно следует отметить, что в данном примере работа решающих деревьев при глубине анализа 5 даёт результаты, которые во всех случаях хуже, чем при глубине 2, чем не было в случае прогноза ряда с периодом 1 час. Это можно объяснить спецификой данного конкретного ряда, в котором слишком велика доля шумов (больше, чем в предыдущем примере) и при большей глубине анализа шумы негативно влияют на выявление анализируемым методом закономерностей в ряду.

Теперь рассмотрим прогнозирование курса валютной пары евро/доллар адаптивным методом R, а также решающими деревьями, после чего сравним оба представленных метода. В таблице 17 приведены данные прогноза курса пары евро/доллар США с временным интервалом (ТФ) один день (D1) в период с 20.06.2012 по 12.05.2014. Длина ряда составляет 500 элементов. График ряда приведён на рисунке 20.

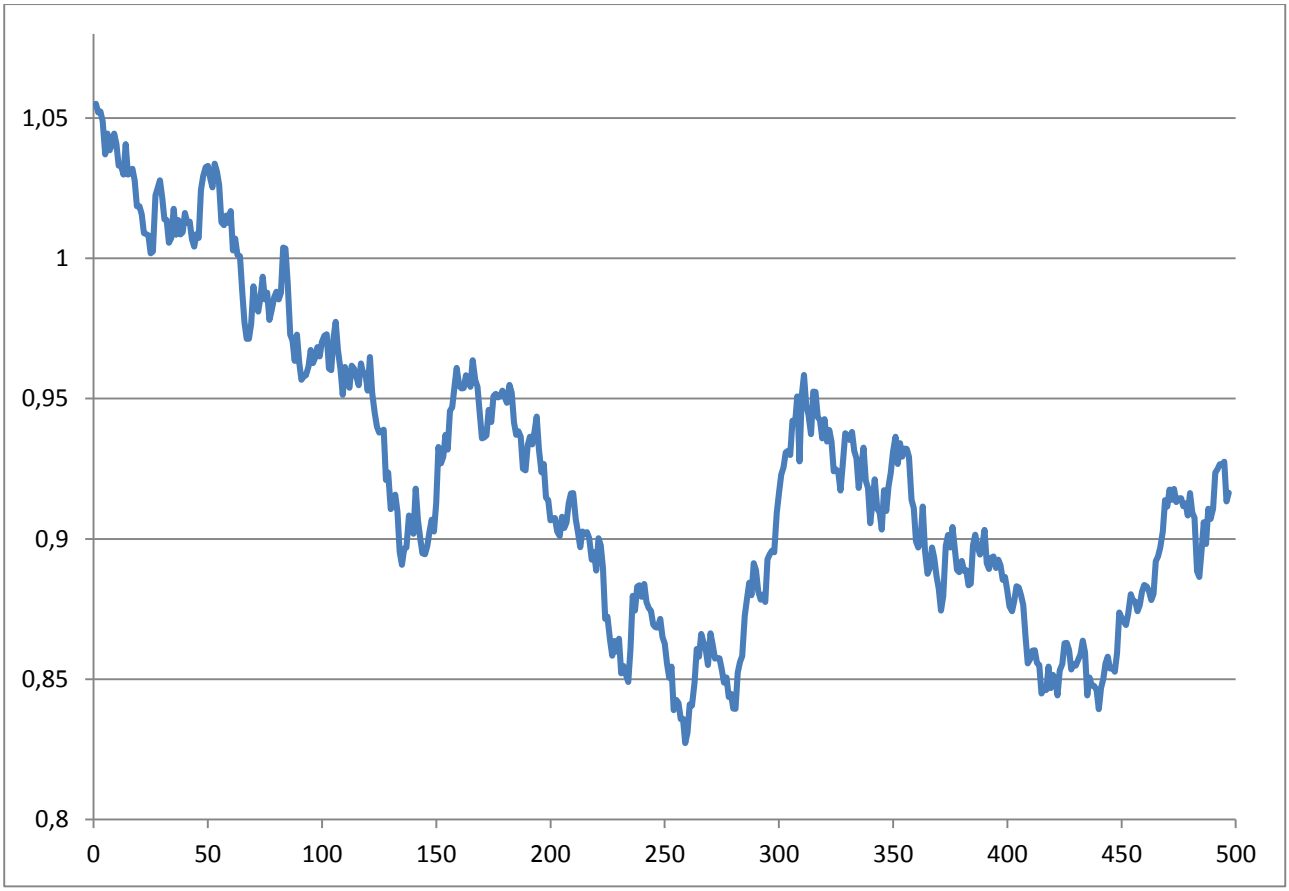

Рисунок 20 – График курса евро/доллар (ТФ: 1 день).

Результаты прогнозирования рассматриваемого ряда приведены в нижеследующей таблице 17.

Таблица 17. Прогнозирование курса евро/доллар (ТФ: 1 сутки).

Разбиение *n*	Глубина анализа *m*	Решающие деревья On-line	R-метод On-line	Решающие деревья 10 шагов	R-метод 10 шагов
10	2	0.0079	0.0084	0.0103	0.0299
	5	0.0095	0.0084	0.0151	0.0299
20	2	0.0088	0.0083	0.0105	0.0159
	5	0.0084	0.0083	0.0105	0.0159

50	2	0.0089	0.0083	0.0119	0.0187

В таблице 18 приведены результаты прогнозирования того же курса евро/доллар, но уже на ряду с периодом (таймфрейм) 1 час. График ряда приведён на рисунке 21.

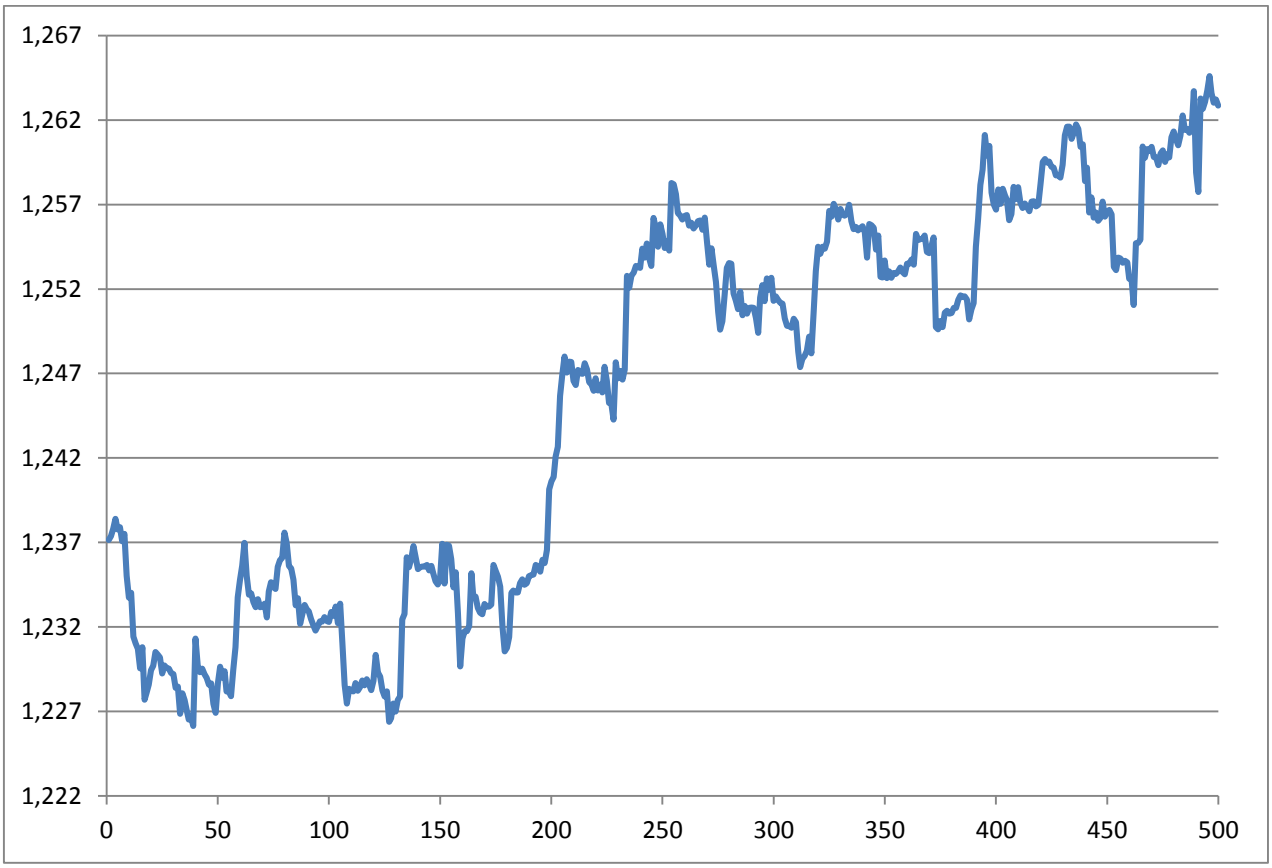

Рисунок 21 – График курса евро/доллар (ТФ: 1 час)

Таблица 18. Прогнозирование курса евро/доллар (ТФ: 1 час).

Разбиение *n*	Глубина анализа *m*	Решающие деревья On-line	R-метод On-line	Решающие деревья 10 шагов	R-метод 10 шагов
10	2	0.00114	0.00114	0.00131	0.00131
	5	0.00132	0.00114	0.00144	0.00131

20	2	0.00106	0.00103	0.00131	0.00131
	5	0.00147		0.00110	0.00131
50	2	0.00103	0.00104	0.00238	0.00141

Из приведённых выше результатов видно, что после определённого предела размера алфавита (разбиения непрерывного интервала) ошибка прогноза методов перестаёт уменьшаться. Это справедливо как для метода R, так и для метода решающих деревьев. Кроме того, из полученных данных видно, что глубина анализа после значения $m = 2$ улучшает точность прогноза достаточно несущественно и после какого-то заданного m точность, как и в случае с размером алфавита, уже не меняется. Фактически, это говорит о том, что для получения оптимальных прогнозов за приемлемое время достаточно подобрать такие минимальные значения размера алфавита и глубины анализа обоих методов, которые будут давать оптимальные (приближенные к границе точности) значения ошибок.

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

Исследуем эффективность работы метода на основе случайного леса, описанного в разделе 3.3, на примере прогнозирования курса электронной криптографической валюты Bitcoin. Курс данной валюты измеряется в долларах США за 1 Bitcoin. Рассмотрим курс описанной валюты за период с 01.04.2013 по 01.04.2014 с интервалом между измерениями (таймфреймом) 1

сутки. Размер полученного ряда 365 элементов. Значение величины Δ равно 300.05. График описанного курса представлен на рисунке 22.

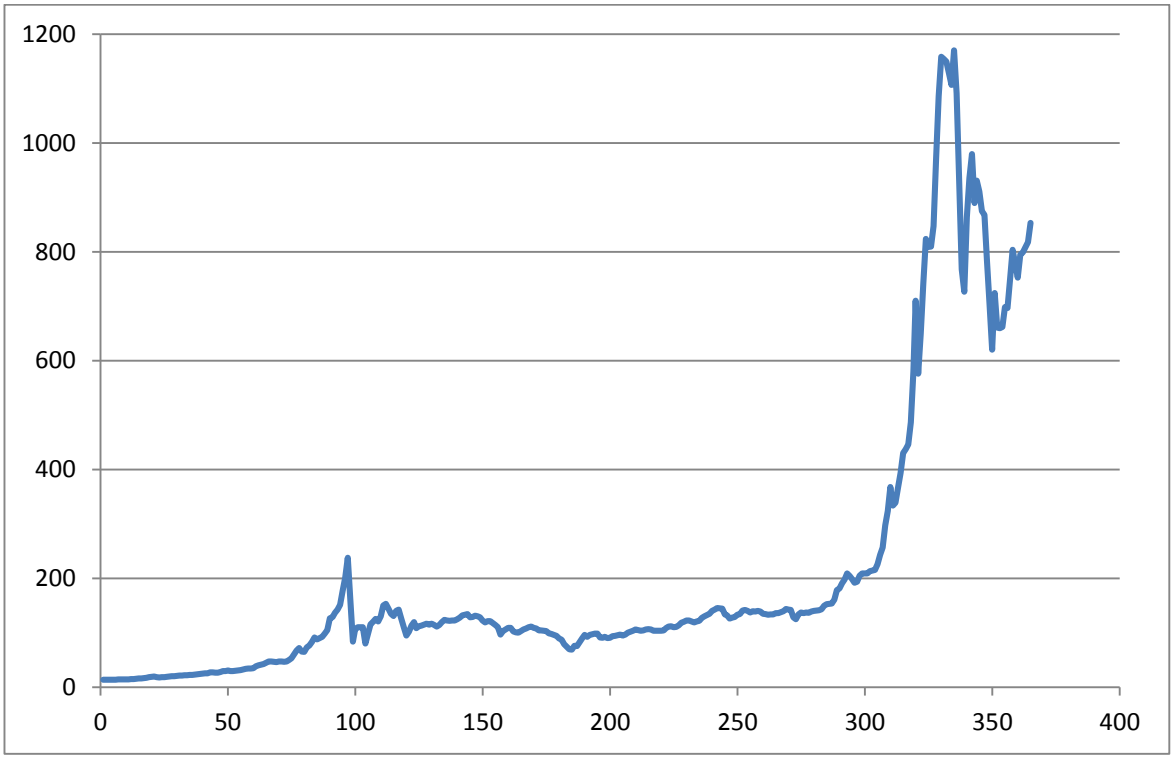

Рисунок 22 – Курс Bitcoin.

Для прогнозирования данного курса применялся метод R, решающие деревья и метод на основе случайного леса. Использованная во всех методах глубина анализа $m = 3$. Метод на основе случайного леса был протестирован на множестве специфичных именно для него параметров. В итоге, были выбраны оптимальные значения, равные следующим: параметр $r = 0.7$; число деревьев $NTrees = 40$. Методы были проверены на различных значениях разбиений n.

Полученные в процессе прогнозирования результаты для интервала прогнозирования с 25.12.2013 по 04.01.2014 в обоих режимах работы представлены в таблице 19.

Таблица 19. Прогнозирование курса Bitcoin. 25.12.2013 – 04.01.2014.

№	Разбиение n	R-метод on-line	Решающие деревья	Случайный лес
Режим on-line				
1	10	30,81	30,14	29,87
2	20	31,44	31,98	30,25
3	40	26,73	26,73	26,73
4	80	25,98	26,41	25,08
5	120	25,73	24,56	23,01
Режим на 10 шагов вперёд				
1	10	91,68	90,41	88,44
2	20	36,16	35,26	34,17
3	40	89,79	90,38	90,73
4	80	85,85	87,13	86,34
5	120	82,56	82,56	82,56

Исходя из представленных результатов видно, что точность прогноза близка к границе точности при $n = 6$. Такая невысокая точность работы объясняется высокой волатильностью конечной части ряда, по которой мы осуществляли прогноз. Данный отрезок не соответствует никаким трендам и закономерностям основной (начальной и центральной) части рассматриваемого ряда, что хорошо видно на графике (рис. 22). С ростом разбиения погрешность всех методов строго убывает. Все 3 метода показывают сравнимую друг с

другом точность работы. При этом метод на основе случайного леса даёт немного меньшую ошибку.

Рассмотрим прогнозирование рассмотренного ряда за период с 20.10.2013 по 01.11.2013, в который волатильность ряда заметно ниже. Все параметры работы методов при этом остаются прежними. Результаты работы алгоритмов приведены в таблице 20.

Таблица 20. Прогнозирование курса Bitcoin. 20.10.2013 – 01.11.2013.

№	Разбиение n	R-метод on-line	Решающие деревья	Случайный лес
Режим on-line				
1	10	23,64	23,64	23,64
2	20	14,46	14,10	13,81
3	40	15,24	16,23	16,73
4	80	15,83	15,41	15,29
5	120	15,24	14,94	14,94
Режим на 10 шагов вперёд				
1	10	51,12	51,03	50,54
2	20	36,38	36,38	35,71
3	40	43,17	44,65	45,87
4	80	47,10	46,38	45,07
5	120	43,17	42,68	42,21

Точность прогноза за указанный интервал у всех методов заметно повысилась, что полностью соответствует вышеприведённому предположению о влиянии волатильности на точность получаемых прогнозов. Прогнозирование при разбиении $n \geq 20$ ошибка прогноза уже равна границе точности при $n = 10$, что говорит о выявлении методами имеющихся в ряду закономерностей (большей точности в данном случае достичь не представляется возможным никаким методом прогнозирования из-за наличия в ряду шумов, которые спрогнозировать невозможно). Что касается соотношения работы методов, то тенденция осталась той же, что и для конечной части ряда: методы показывают сравнимую эффективность работы.

Рассмотрим результаты прогнозирования курса рубля к доллару также всеми тремя изучаемыми методами. Прогнозирование указанного ряда осуществлялось в период с 01.01.2013 по 01.12.2014 с периодом между измерениями (таймфреймом) 1 сутки. Длина ряда при этом составила 700 элементов. График ряда показан на рисунке 23. Параметр разбиения для всех методов $n = 15$. Значение величины Δ равно 4.6768. Метод на основе случайного леса был протестирован на множестве параметров. Параметры метода на основе случайного леса выбраны те же: параметр $r = 0.7$; число деревьев $NTrees = 40$.

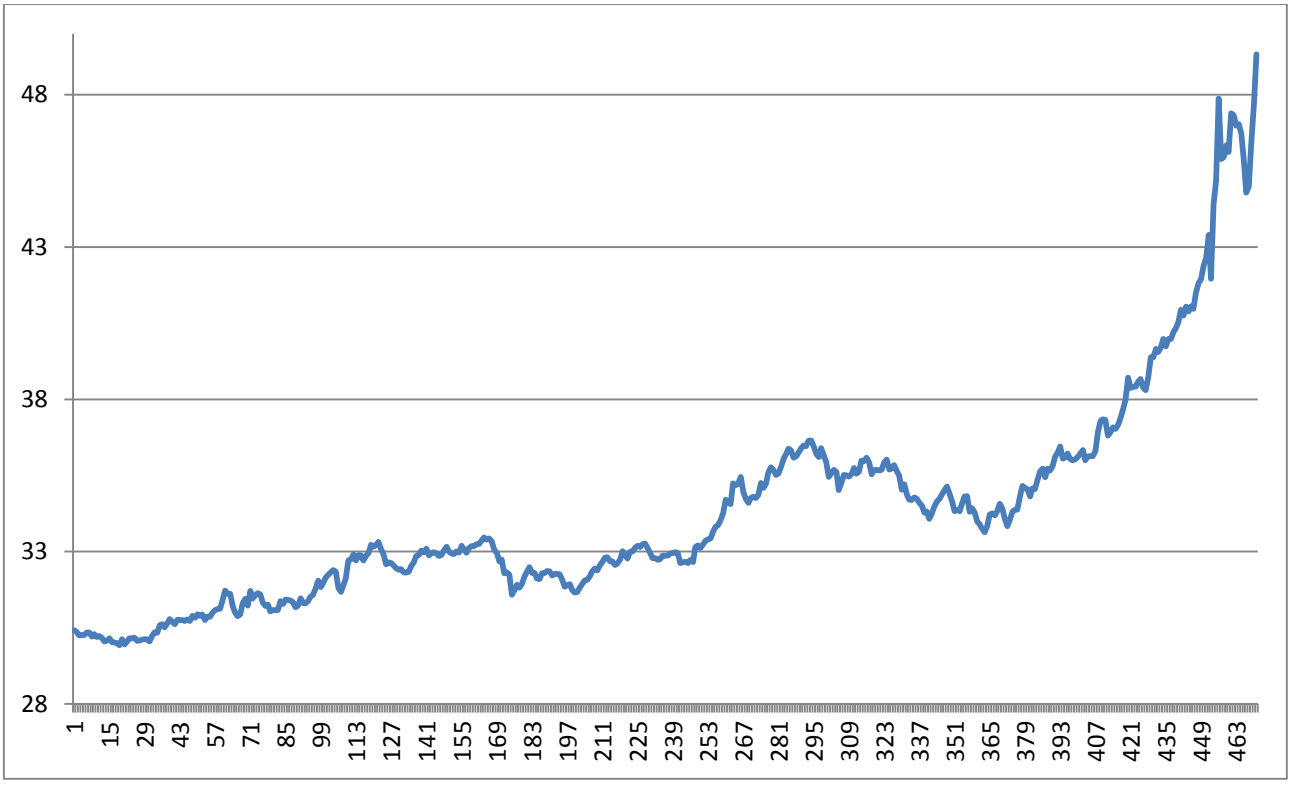

Рисунок 23 – Курс рубля к доллару. 01.01.2013 – 01.12.2014.

Результаты работы всех 3 алгоритмов представлены в таблице 19. В столбце глубина анализа m указано 2 значения: до черты «/» указана глубина анализа m (применимая к решающим деревьям, случайному лесу и R-методу), после черты – максимальная глубина дерева (после которой ветвь обрезается, данный параметр применим только к методу на основе решающих деревьев; деревья в методе на основе Случайного леса не обрезаются). При этом в таблице 21 представлено два варианта прогнозирования ряда: самой последней его части (с 20.11.2014 по 01.12.2014) и со сдвигом назад (с 01.11.2014 по 11.11.2014), обозначенные номерами 1 и 2, соответственно.

Таблица 21. Прогнозирование курса рубля к доллару.

№	Тип прогнозирования	Глубина анализа m / число признаков	R-метод	Решающие деревья	Случайный лес
1	On-line	5 / 3	0.7243	0.7243	0.7243

		5 / 4			0.7243
	10 шагов	5 / 4	1.2234	1.2234	1.2234
		5 / 3			1.2234
2	On-line	5 / 3	0.9439	1.0686	1.0063
		5 / 4			1.0686
	10 шагов	5 / 3	2.1150	2.1150	2.1150
		5 / 4			2.1150

Из полученных результатов видно, что методы дают очень близкие по точности прогнозные значения. Сделать однозначных выводов о том, какой из них лучше по своей эффективности нельзя. Общие значения погрешностей у всех методов не слишком высокие (далеки от границы точности прогноза для заданного разбиения) в силу существенного изменения экономической ситуации в ноябре-декабре 2014 года и резкого изменения тенденций развития курса рубль/доллар.

5.8.2. Автоматическая торговля на валютной бирже

Рассмотрим приложение методов прогнозирования к одной из наиболее актуальных в современное время областей – к торговле на бирже (фондовой, валютной, фьючерсной и т.д.). Одной из наиболее актуальных задач в этом направлении является задача создания автоматических торговых систем, называемых экспертными системами, которые приносили бы прибыль своим владельцам путём покупки-продажи определённых товаров на бирже в нужное время и в нужном объёме. В качестве базового примера возьмём торговлю на валютном рынке Forex по нескольким валютным парам.

Был разработан следующий алгоритм работы системы автоматической торговли на бирже:

1. Применяем какой-либо метод прогнозирования на 1 шаг вперёд при заданном разбиении, получая плотность вероятности следующего элемента. При этом прогнозируем разницы между элементами.
2. Смотрим значение точки (подинтервала) с максимальной вероятностью: если его модуль меньше некоторого порога, то сделки не открываем (т.к. маленький модуль прогнозного значения говорит об отсутствии тренда, т.е. о слабых намерениях рынка к изменению курса и высокой вероятности колебаний; прибыль будет либо низкая, либо отрицательная) и переходим к следующему прогнозному элементу (алгоритм начинается с шага 1). Иначе переходим на следующий шаг алгоритма.
3. Рассматриваем некоторую окрестность вокруг точки (подинтервала) с максимальной вероятностью и считаем суммарную вероятность попадания прогнозного значения в этот интервал (т.е. сумму вероятностей всех подинтервалов в выбранной окрестности).
4. Если полученная вероятность окрестности меньше некоторого порогового значения, то сделки не открываем (т.к. вероятность движения рынка по заданному направлению слишком низкая) и переходим к следующему прогнозному элементу (алгоритм начинается с шага 1). Иначе переходим на следующий шаг алгоритма.
5. Открываем сделку в соответствии со значением центра окрестности. Если значение положительно (т.е. курс будет расти, т.к. мы прогнозируем разности), то открываем сделку покупки с размером лота (вкладываемых в сделку средств), равным произведению некоторого параметра (размера лота), прогнозного значения и вероятности окрестности. Данный способ вычисления вкладываемых средств позволяет совершать сделки с учётом возможных рисков.

При достижении времени существования сделки порогового значения (параметр метода), сделка автоматически закрывается. При переходе к следующему прогнозному элементу алгоритм повторяется с шага 1. Как видно из алгоритма, метод имеет следующий набор параметров: величина окрестности вокруг точки максимума, пороговое значение минимума центра окрестности (точки максимума), пороговое значение вероятности окрестности, размер лота.

Рассмотрим экспериментальные результаты работы данного алгоритма на валютном рынке Forex. В качестве метода прогнозирования, применяемого в данной системе, был использован R-метод. В качестве рядов для проведения экспериментальных исследований были выбраны валютные пары евро / доллар США (EUR/USD), британский фунт стерлингов / доллар США (GBP/USD), доллар США / швейцарский франк (USD/CHF), Bitcoin / доллар США. Размер всех рядов равен 700 элементам. Разбиение $n = 20$. Число шагов работы алгоритма (количество временных точек) равно 15. Глубина анализа $m = 3$. Интервал между измерениями (таймфрейм) равен 1 часу (H1) и 1 суткам (D1): измерения проводились для обоих вариантов. Период всех рядов для ТФ 1 сутки: с 02.01.2012 по 12.02.2014. Период всех рядов для ТФ 1 час: с 01.10.2014 по 12.02.2014. Величина окрестности равна 10% от размера всего интервала возможных значений. Начальный объём средств равен 1000 единиц. Пороговые значение минимума центра окрестности и вероятности окрестности взяли для проведения начальных тестовых испытаний равными 0.5 и 0.3, соответственно.

Результаты проведения испытаний представлены в нижеследующей таблице 22. В таблице указано значение параметра Δ для каждой валютной пары, таймфрейм, количество шагов, на которые производилось открытие сделки (было взято 2 варианта: на 1 и на 3 шага), а также показана вероятность окрестности, точность получаемых в процессе работы алгоритма прогнозов и конечный итог (получившаяся в итоге сумма).

Таблица 22. Результаты работы системы автоматической торговли на валютной бирже.

Валютная пара	Δ	ТФ	Количество шагов сделки	Вероятность на окрестности	Точность прогноза	Итог
EUR/USD	0.0129	H1	1	0.7682	0.00075	1000
			3	0.7723	0.00111	1004
	0.0237	D1	1	0.3856	0.00404	982
			3	0.3877	0.00765	1010
GBP/USD	0.0212	H1	1	0.7851	0.00129	1008
			3	0.7869	0.00145	1010
	0.1142	D1	1	0.3847	0.00544	1002
			3	0.3865	0.00866	978
USD/CHF	0.0105	H1	1	0.7850	0.0006	998
			3	0.7855	0.0011	995
	0.0914	D1	1	0.5405	0.0096	997
			3	0.5383	0.0113	1065
Bitcoin/USD	300.06	D1	1	0.81	56.68	1083
			3	0.84	137.32	1156

Из представленных результатов видно, что в среднем погрешность применяемого метода прогнозирования сравнима с границей точности для разбиения $n = 6$, что является неплохим результатом для временных рядов валютных пар (т.к. данные ряды не имеют строгих закономерностей и обычно обладают высокой дисперсией). Суть работы R-метода для рядов EUR/USD, GBP/USD и USD/CHF сводилась к выявлению тренда и прогнозированию значений по этому тренду. Данное поведение метода было описано в разделе 5.3. Что касается ряда курса Bitcoin, то здесь прогнозные значения (центр

окрестности) существенно изменялись на разных шагах, методом выявлялись определённые закономерности и система приносила существенную прибыль.

Что касается результатов работы системы автоматической торговли, то для валютных пар EUR/USD, GBP/USD и Bitcoin/USD результаты получились неплохими: система отработала с прибылью практически для всех случаев. Для пары USD/CHF в среднем получился убыток. Для реального практического применения предложенного подхода требуется проведение множества тестов, в ходе которых был бы разработан алгоритм отбора оптимальных параметров метода. Однако проведённые эксперименты показывают возможность практической применимости разработанной системы и её потенциальную эффективность.

5.9. Прогнозирование расхода электроэнергии

Проведём прогнозирование расхода электроэнергии в США R-методом, решающими деревьями и методом на основе случайного леса. В качестве временного ряда для прогноза взяли ряд ежесуточных значений расхода электроэнергии в США за период с 01.10.2011 по 01.12.2013. Длина ряда при этом составила 745 элементов. Его график показан на рисунке 24. Значение Δ для представленного ряда равно 68.9. Прогноз вёлся с разбиением $n = 15$, в двух режимах: on-line и на 10 шагов вперёд. Значение параметра r метода на основе случайного леса равно 0.7. Число деревьев $NTrees$ равно 40.

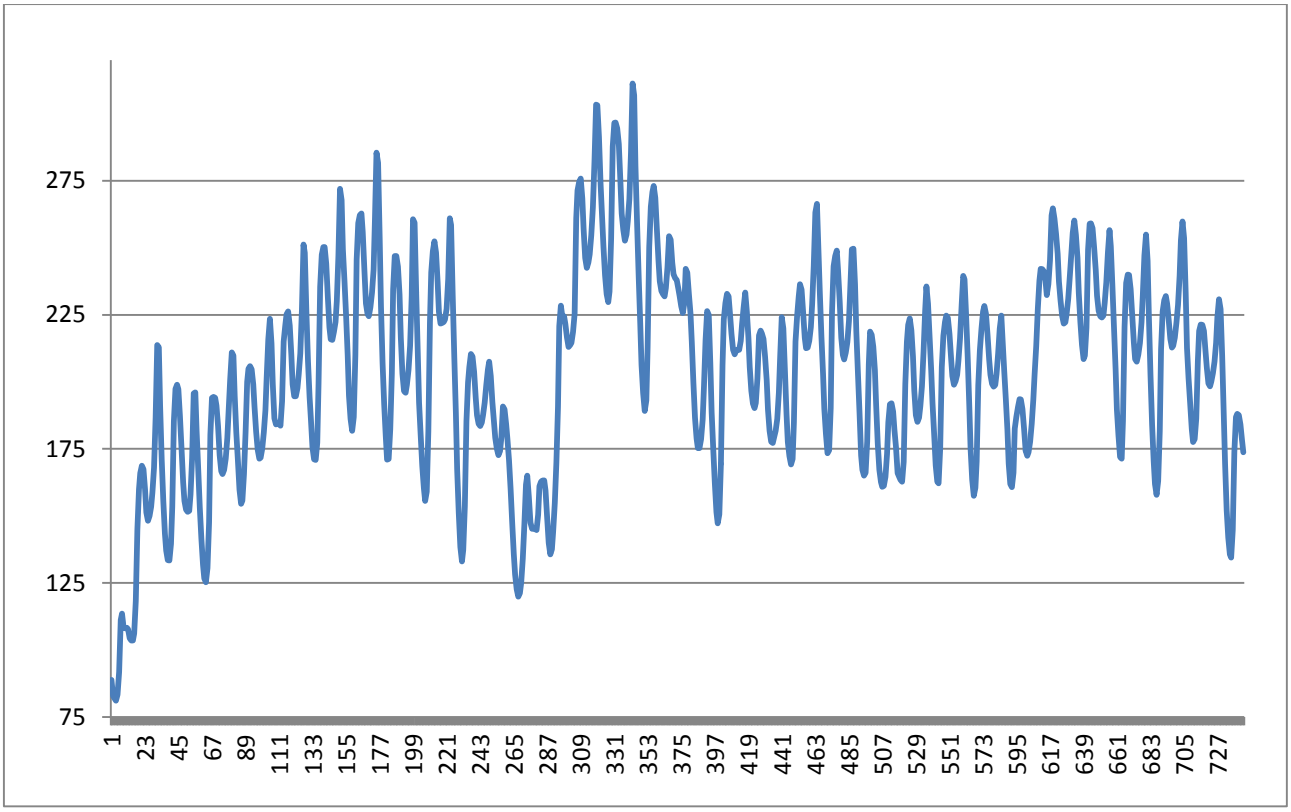

Рисунок 24 – Расход электроэнергии США.

Результаты прогнозирования приведены в таблице 23. В столбце с глубиной анализа после черты «/», как и в предыдущих примерах приведено значение максимальной глубины дерева для метода на основе решающих деревьев, после которой ветви дерева обрезаются. Для R-метода и метода на основе случайного леса данный параметр смысл не имеет.

Таблица 23. Прогнозирование расхода электроэнергии США.

№	Тип прогнозирования	Глубина анализа m	R-метод	Решающие деревья	Случайный лес
1	On-line	5 / 3	10.032	3.3453	5.5987
		5 / 4			3.3453
	10 шагов	5 / 4	30.0390	12.5843	23.6083

		5 / 3			24.5270
2	On-line	5 / 3	4.0813	3.9373	2.6727
		5 / 4			3.5800
	10 шагов	5 / 3	78.3997	8.1417	2.2133
		5 / 4			17.1417

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

5.10. Многомерное прогнозирование экономических процессов

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 25 и 26, соответственно. В качестве базового

метода для реализации многомерного прогнозирования был выбран R-метод. Коэффициент сдвига целевого ряда l для всех случаев был выбран 1.

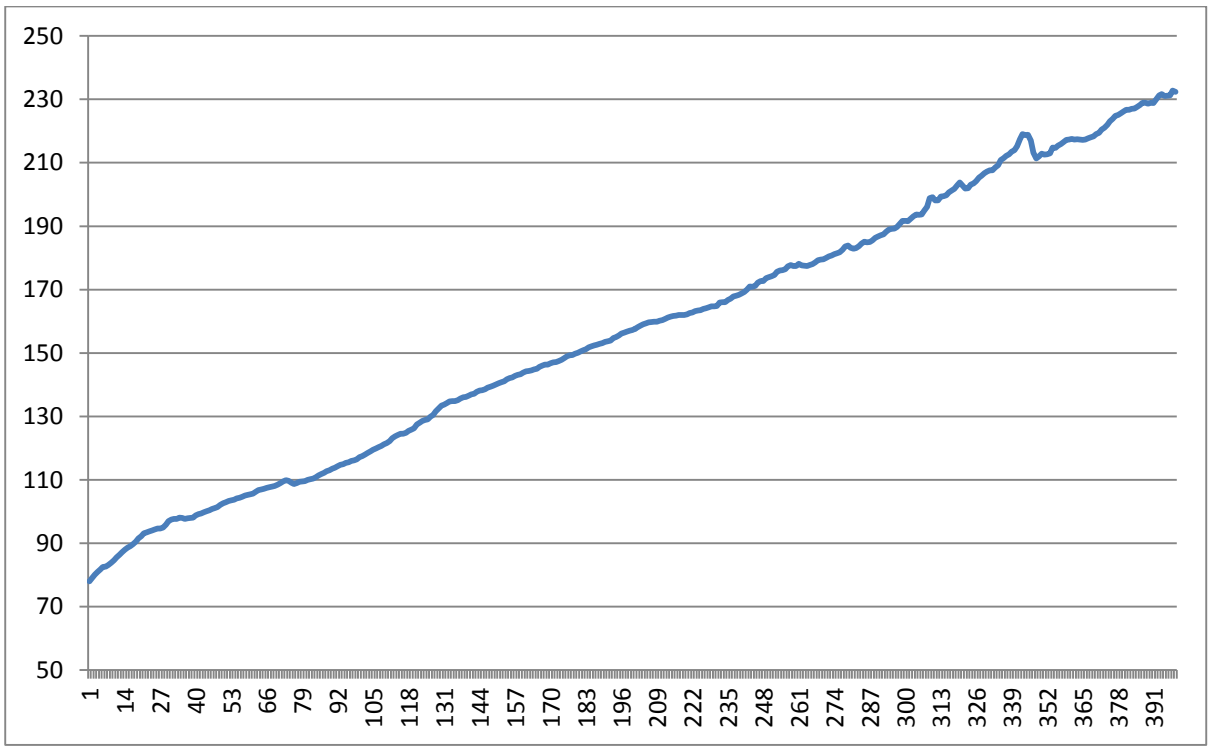

Рисунок 25 – Индекс потребительских цен США (CPI). 1983 – 2013гг.

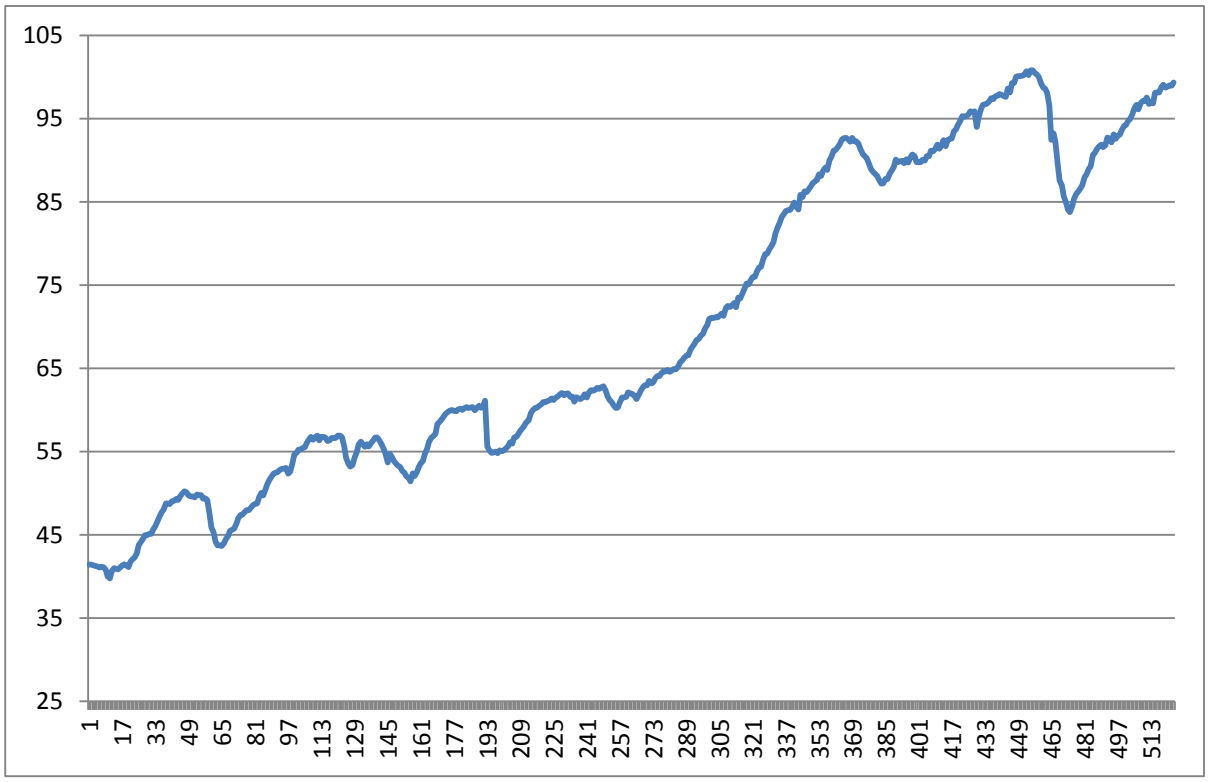

Рисунок 26 – Индекс промышленных цен США (PPI).

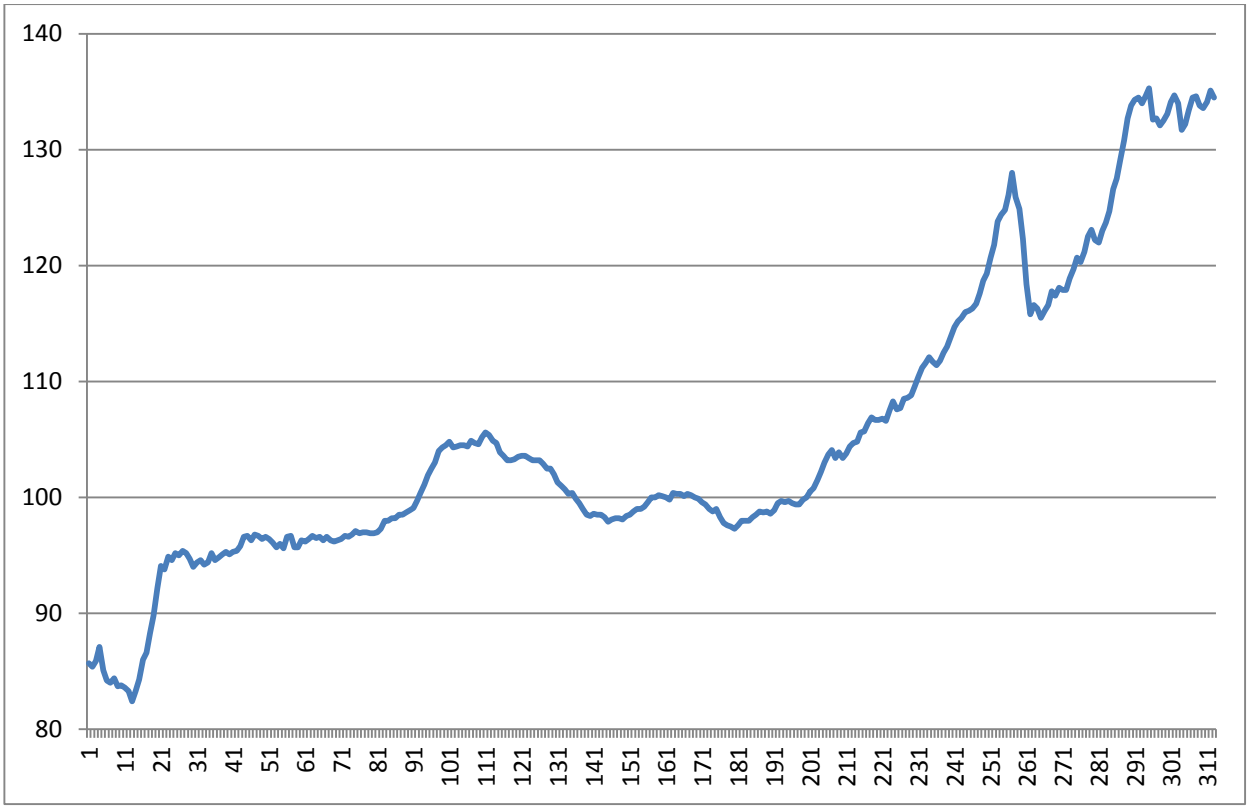

Рисунок 27 – Уровень экспорта США.

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 27), 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 24 и 25, 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 24. 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 25. 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.

Рассмотрим прогнозирование уровня безработицы в США. Его график представлен на рисунке 28. Для этого в качестве дополнительных рядов возьмём явно коррелирующий с уровнем безработицы ряд количество обращений по безработице, график которого показан на рисунке 29. Внешне хорошо видны корреляции и закономерности, однако ряды всё же отличаются и абсолютные значения и дисперсия (волатильность) у них совершенно различны. Для большей объективности возьмём также следующие ряды: уровень ВВП США (рисунок 30), индекс промышленных цен (рисунок 25) и индекс промышленного производства в США (рисунок 31). Периоды данных рядов составляют с 01.1970 по 08.2012. Размер рядов: 512 элементов. Глубина анализа $m = 3$. Величина Δ для ряда, представляющего уровень безработицы, равна 1.65.

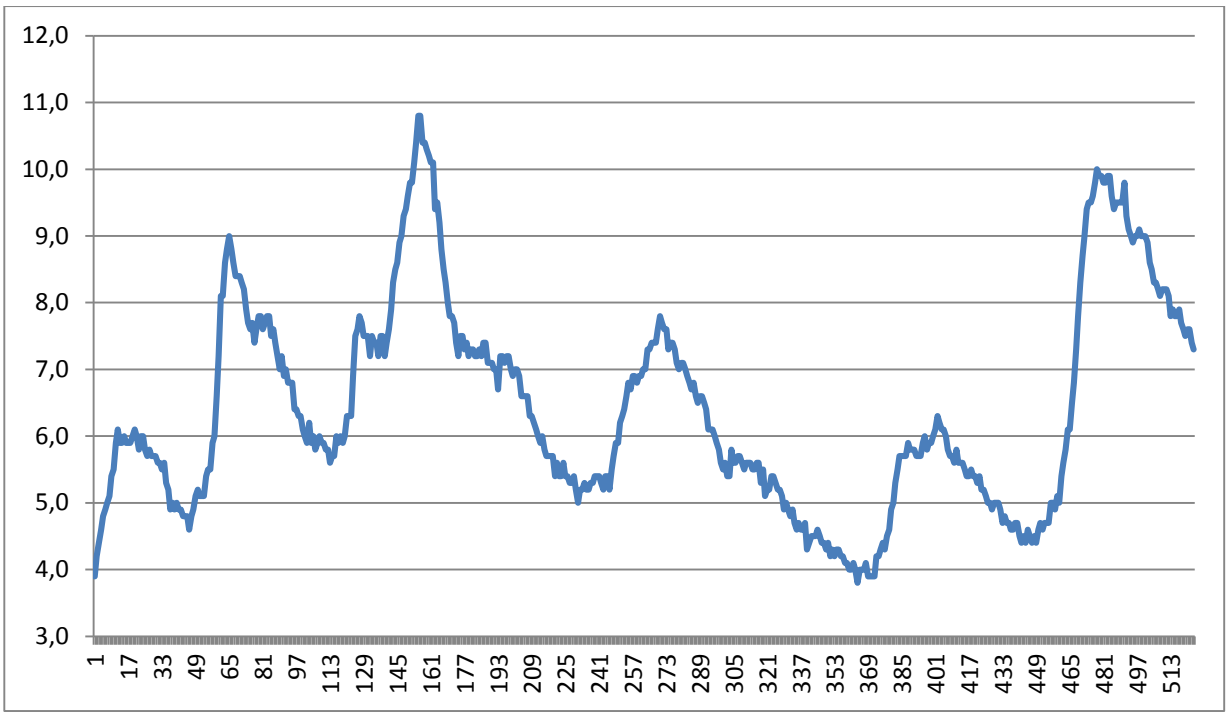

Рисунок 28 – Уровень безработицы в США.

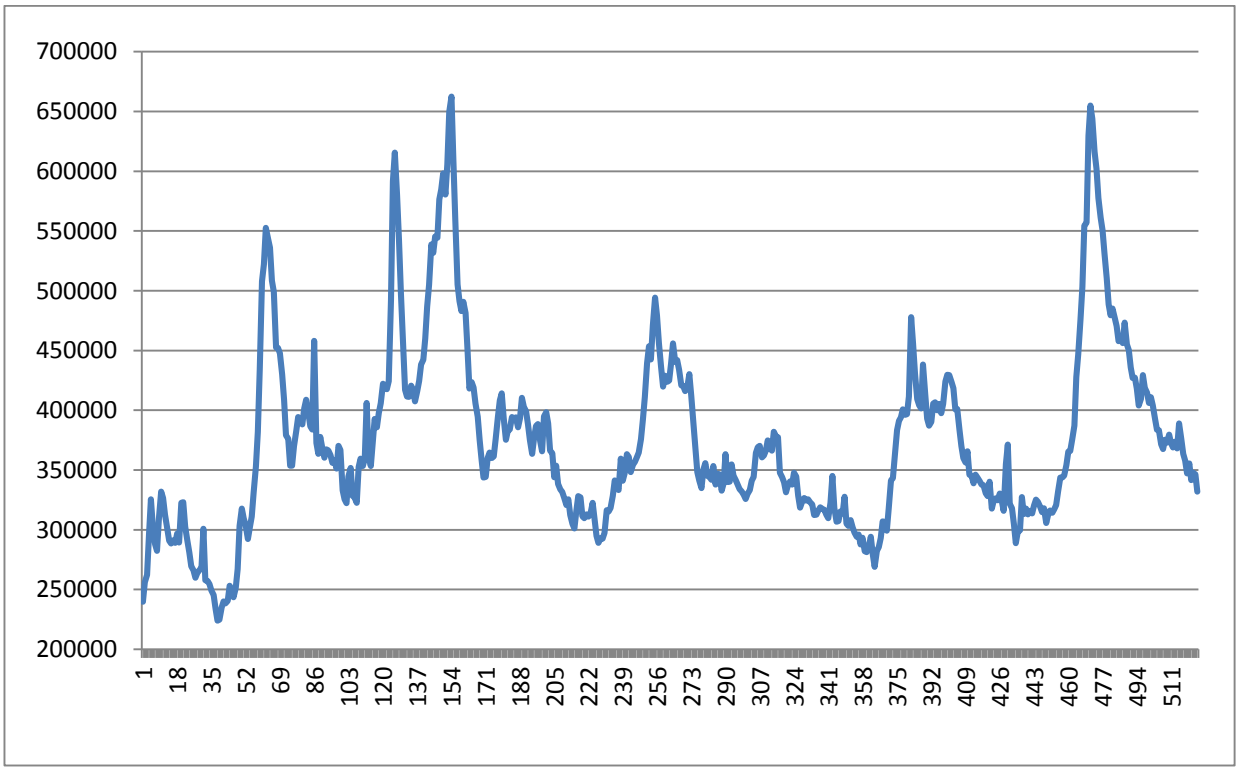

Рисунок 29 – Обращения по безработице в США.

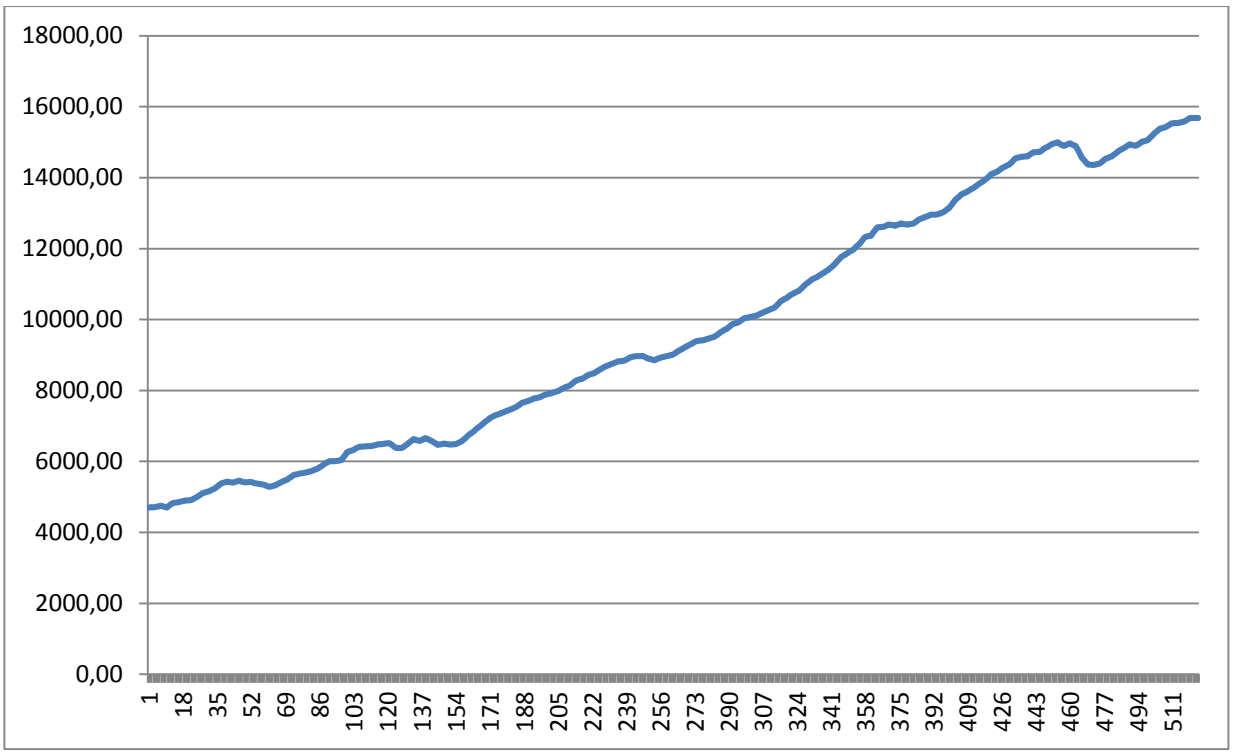

Рисунок 30 – Внутренний валовый продукт США (GDP).

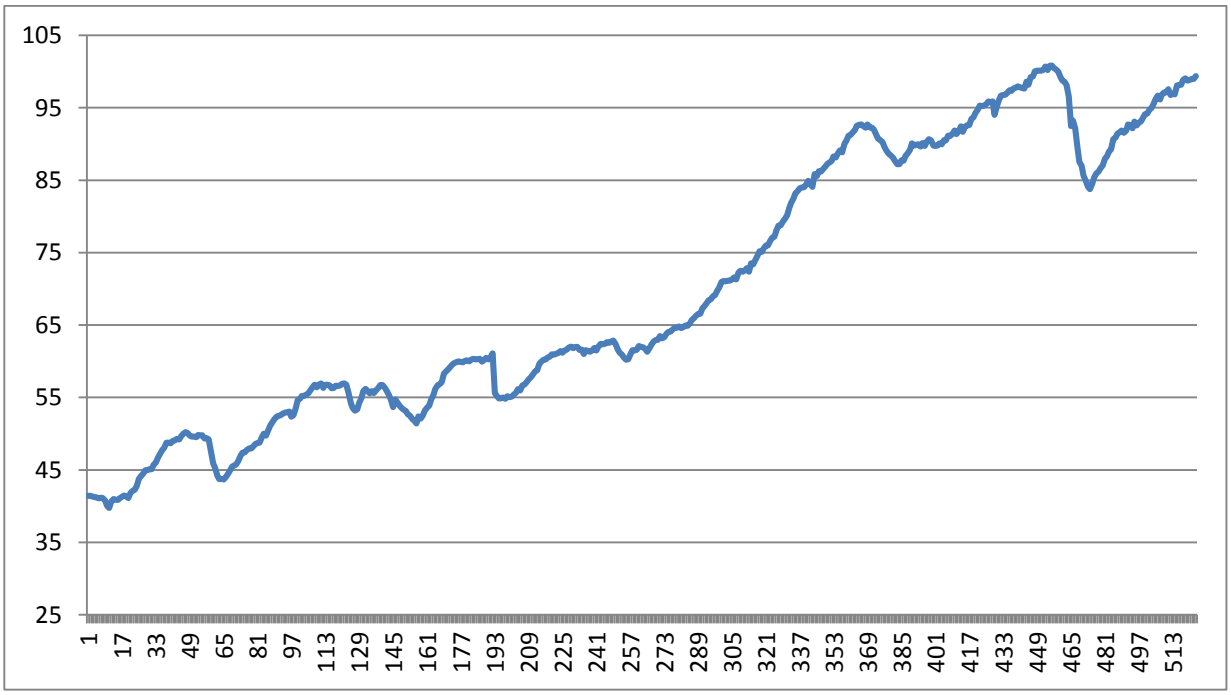

Рисунок 31 – Индекс промышленного производства США (IPI).

Результаты прогнозирования ряда уровня безработицы в США приведены в таблице 26.

Таблица 26. Прогнозирование уровня безработицы в США.

№	Временной ряд	R-метод on-line	R-метод 10 шагов
1	Безработица	0.106	0.425
2	Безработица + Жалобы по безработице	0.098	0.136
3	Безработица + Индекс промышленного производства	0.106	0.460
4	Безработица + ВВП	0.116	0.370
5	Безработица + Индекс потребительских цен	0.134	0.7155

По представленным в таблице 26 данным хорошо видно, что добавление ряда обращений по безработице существенно увеличивает точность получаемых прогнозов. В особенности, для случая прогнозирования на несколько шагов вперёд. Три других ряда прироста точности не дают, оставляя ошибку прогноза на уровне одномерного подхода, что вполне закономерно в виду отсутствия явных корреляций между рядами.

Рассмотрим прогнозирование уровня ВВП США с использованием в качестве дополнительных рядов уровень безработицы США, индекс промышленного производства США, а также индексы CPI и PPI. Их временные параметры аналогичны рассмотренным выше. Исходя из графиков данных временных рядов, можно увидеть, что ВВП сильно коррелирует с индексом потребительских цен США, а также – немного с индексом промышленного производства США. Исходя из этого, мы должны наблюдать прирост точности метода при соединении ряда GDP с двумя вышеназванными. Результаты прогнозирования приведены в таблице 27.

Таблица 27. Прогнозирование уровня ВВП США.

№	Временной ряд	R-метод on-line	R-метод 10 шагов
1	ВВП	9.363	79.132
2	ВВП + Безработица	16.988	50.475
3	ВВП + Индекс промышленного производства	9.209	78.133
4	ВВП + Жалобы по безработице	13.306	69.878
5	ВВП + Индекс потребительских цен	8.207	57.702
6	GDP + Индекс промышленных цен	15.146	88.720

Как видно из вышеприведённых результатов, заметное увеличение точности по сравнению с одномерным подходом наблюдается в случае добавления рядов индексов промышленного производства и потребительских цен, что вполне закономерно.

Рассмотрим прогнозирование цен на энергоносители США (Gasoline) в многомерном режиме. В качестве дополнительных коррелирующих рядов будем использовать объём розничных продаж США, уровень импорта США, обращения по безработице США, индекс промышленного производства США, индекс потребительских цен США, внутренний валовый продукт США. Прогноз выполнялся в период с 01.1992 по 08.2013. Таймфрейм равен 1 месяцу. Периоды и таймфреймы коррелирующих рядов совпадают с исходным рядом. Разбиение (параметр n) было равно 10. Глубина анализа $m = 3$. Величина Δ для исходного ряда равна цен на энергоносители равна 1.324.

Результаты прогнозирования приведены в таблице 28.

Таблица 28. Многомерное прогнозирование цен на энергоносители.

№	Временной ряд	R-метод on-line	R-метод 10 шагов
1	Цены на топливо	0,1224	0,1594
2	Цены на топливо + Уровень розничных продаж	0,0921	0,1012
3	Цены на топливо + Уровень безработицы	0,1224	0,1594
4	Цены на топливо + Индекс промышленного производства	0,1125	0,1397
5	Цены на топливо + Индекс потребительских цен	0,1027	0,1208
6	Цены на топливо + ВВП	0.1221	0.1672
7	Цены на топливо + Индекс потребительских цен	0.1364	0.1601

Из приведённых данных хорошо виден положительный результат работы методов. В случае использования рядом объёма розничных продаж, объёма промышленного производства и уровня потребительских цен точность получаемого прогноза становится выше, что говорит о существующей корреляции между указанными рядами, а также об эффективности многомерного подхода. Кроме того, в случае соединения исходного ряда с рядами объёма розничных продаж и индексом CPI метод R показывает погрешность, равную границе точности работы метода при заданном разбиении.

Все вышеприведённые результаты показывают высокую эффективность работы модификации метода R на основе многомерного подхода. Благодаря использованию предложенного подхода можно расширить возможности методов прогнозирования по выявлению закономерностей, не ограничиваясь лишь простыми закономерностями (например, периодическими).

5.11. Приложение методов прогнозирования к задаче криптоанализа блоковых шифров

Методы прогнозирования временных рядов, основанные на вычислении плотности вероятности, могут применяться и в задачах криптографического анализа. Рассмотрим задачу криптоанализа блоковых шифров. Как известно, основное требование ко всем современным блоковым шифрам, заключается в статистической неотличимости зашифрованной последовательности от абсолютно случайной никаким статистическим тестом. При этом данное требование должно быть выполнено для произвольной исходной (незашифрованной) последовательности. Данное требование гарантирует отсутствие в выходной последовательности шифра каких-либо обнаруживаемых закономерностей, которые могли бы помочь в подборе ключей алгоритма (в т.ч. раундовых ключей).

Рассмотрим основную идею реализации градиентной статистической атаки на блоковые шифры, связанные с отсутствием статистической случайности их выходной последовательности.

Пусть имеется некоторый блоковый шифр с ключом K. Его работа по шифрованию исходного открытого текста состоит из некоторого числа этапов, называемых раундами. На каждом раунде используются так называемые раундовые ключи K_i, формируемые на основе исходного K. При этом длина раундовых ключей, как правило, много меньше длины исходного ключа шифра, т.е. $|K_i| \ll |K|$. Пусть дана исходная последовательность x_0.

Тогда схема шифрования будет выглядеть следующим образом. $\mathrm{x}_1 = \mathrm{E}(\mathrm{x}_0, \mathrm{K}_1), \mathrm{x}_2 = \mathrm{E}(\mathrm{x}_1, \mathrm{K}_2), \dots, \mathrm{x}_\mathrm{r} = \mathrm{E}(\mathrm{x}_{\mathrm{r}-1}, \mathrm{K}_\mathrm{r})$, где x_0 – исходный блок данных, который необходимо зашифровать, E – операция (функция) шифрования на i-ом раунде, $\mathrm{x_i}$ – блок данных, являющийся «выходом» i-ого раунда и «входом» $(\mathrm{i}+1)$-ого, всего имеется r раундов. Дешифрование происходит по схеме, обратной к схеме шифрования. При этом используется некоторая функция $\mathrm{D}(\mathrm{x}, \mathrm{K})$. Пусть также имеется некоторый статистический тест $Stat(x)$,

показывающий значение «случайности» анализируемой последовательности. В итоге, мы перебираем раундовые ключи, начиная с последнего, по схеме. Если $Stat(\mathrm{x}_{i-1}) > Stat(\mathrm{D}(\mathrm{x}_i, \mathrm{K}_i))$, то мы нашли верный ключ раунда i и можем переходить к раунду $(i-1)$, иначе ключ неверный и нужно продолжить перебор ключей K_i. Так продолжаем до нахождения всех раундовых ключей, что равносильно нахождению исходного ключа шифра K.

Смысл статистического теста $Stat$ заключается в вычислении степени отклонения анализируемой последовательности от абсолютно случайной. В качестве такого теста может выступать произвольный метод прогнозирования, оценивающий условные вероятности. Как уже было сказано, для идеального шифра выходная последовательность не должна иметь отклонений от статистической случайности, что означает выполнение для его бинарной выходной последовательности следующего правила:

$$P(x_i = 1 | x_1 \dots x_{i-1}) = P(x_i = 0 | x_1 \dots x_{i-1}) = \frac{1}{2} \qquad (17)$$

где P – оценка условной вероятности равенства i-ого символа 0 последовательности $x_1 \dots x_i$ нулю или единице. Данная оценка должна быть справедлива для любого $i = 1, \dots, N$, где N – длина входной / выходной последовательности. На основе правила (17) и статистического критерия χ^2 статистическая функция $Stat$ определяется следующим образом:

$$Stat(x_1 \dots x_N) =$$

$$= \sum_{i=1}^{N} \frac{(i \cdot P(x_i = 1 | x_1 \dots x_{i-1}) - i \cdot 0.5)^2}{i \cdot 0.5} + \frac{(i \cdot P(x_i = 0 | x_1 \dots x_{i-1}) - i \cdot 0.5)^2}{i \cdot 0.5} \qquad (18)$$

В качестве функции оценки условных вероятностей P допустимо использовать произвольный метод прогнозирования. При этом, если глубина анализа выбранного метода равняется длине всей последовательности, то в формуле (18) допустимо учитывать только последнее слагаемое суммы, т.к. глубина анализа метода учитывает все другие подпоследовательтности. В

итоге, для методов с глубиной анализа, равной длине последовательности, допустимо использовать нижеследующую формулу (19):

$$Stat(x_1 \dots x_N) =$$
$$= 2 \cdot N \cdot ((P(x_N = 1|x_1 \dots x_{N-1}) - 0.5)^2 + (P(x_N = 0|x_1 \dots x_{N-1}) - 0.5)^2) \quad (19)$$

Предложенная схема была использована для реализации криптографической атаки на блоковые шифры RC6, MARS, IDEA, CAST-128 и Blowfish [32-34]. При этом по ряду шифров были получены результаты, превосходящие все ранее известные. Результаты градиентной статистической атаки на указанные шифры описаны в работах автора [32-35].

Кроме того, предложенная выше схема позволяет проверять надёжность произвольных блоковых шифров с использованием любого метода прогнозирования временных рядов на основе оценки плотности вероятности. Предложенные в данной работе алгоритмы прогнозирования применялись для анализа надёжности блоковых шифров [35].

Заключение

В рамках представленной работы были осуществлены как теоретические, так и экспериментальные исследования теоретико-информационных методов прогнозирования временных рядов, основанных на сжатии данных и на интеллектуальном анализе данных. Были разработаны новые методы и универсальные модификации, применимые к произвольным методам прогнозирования и повышающих их эффективность. Проведённые экспериментальные результаты показывают высокую эффективность предложенных в данном исследовании методов и модификаций с точки зрения точности получаемых прогнозов реальных экономических процессов. Свою высокую эффективность показали как методы, основанные на универсальной мере, так и методы на основе решающих деревьев (решающие деревья и метод на основе случайного леса). Модификации, связанные с методом группировки алфавита и методом усреднения, также показывают свою высокую эффективность, существенно уменьшая ошибку прогноза для всех предложенных методов и их модификаций. Модификация группировки алфавита, помимо снижения ошибки прогноза, существенно уменьшает трудоёмкость описанных алгоритмов, что позволяет использовать данную модификацию на настольных компьютерах и получать результаты, которые ранее были достижимы только с привлечением супервычислительной техники. Разработаны методы оптимизации работы предложенных алгоритмов, а также способы подбора оптимальных параметров работы методов.

Кроме того, эффективность работы предложенных методов была показана на примере внедрения результатов в маркетинговые исследования ООО ПКФ «Техпром», где разработанные подходы использовались для решения задачи ситуационного моделирования развития многоотраслевой коммерческой компании, а также в задачах автоматического управления работой на валютном рынке Forex.

Предложенный способ применения вероятностных методов прогнозирования к задаче криптоанализа блоковых шифров также показал свою высокую эффективность.

Разработанный и впервые предложенный в данной работе метод многомерного прогнозирования, как показывают практические результаты, оказался эффективнее традиционных методов в случае подбора коррелирующих между собой рядов и позволяет существенно уменьшить ошибку прогноза. Одновременно, в случае отсутствия каких-либо корреляций между соединяемыми рядами, данный метод показывает точность, сравнимую с классическим одномерным подходом на основе того же алгоритма.

Кроме того, описанный многомерный подход позволяет учитывать в прогнозировании не только 2 или более дополнительных рядов, но и какие-либо дополнительные атрибуты или свойства рассматриваемого процесса. Важно отметить, что рассмотренный подход может использоваться, как модификация по отношению к любым вероятностным методам прогнозирования (т.е. тем, выходом которых является распределение вероятностей).

Высокие результаты точности прогнозов подтверждаются и в сравнении с другими известными методами прогнозирования, что показано в данной работе экспериментально.

ЛИТЕРАТУРА

1. Poskitt, D.S. The selection and use of linear and bilinear time series models / D.S. Poskitt, A.R. Tremayne // International Journal of Forecasting. – 1986. – Vol. 2, Issue 1. – P. 101–114.
2. Tong, H. Non-linear Time Series: A Dynamical System Approach. / H. Tong – Oxford University Press, 1990.
3. Tong, H. Threshold models in Nonlinear Time Series Analysis. / H. Tong // Springer Verlag Inc. – Berlin, 1983.
4. Guerard, J. Introduction to Financial Forecasting in Investment Analysis / J. Guerard. – Hardcover, ISBN: 978-1-4614-5238-6, 2013. – 236 p.
5. Engle, R. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom / R. Engle // Econometrica. – 1982. – Vol. 50, Issue 4. – P. 987–1007.
6. Bontempi, G. Local Learning Techniques for Modeling / G. Bontempi // Prediction and Control. – BELGIUM, IRIDIA-Universit de Libre de Bruxelles, 1999.
7. Zhang, G. Forecasting with artificial neural networks: The state of the art / G. Zhang, B. E. Patuwo, Y. H. Michael // International Journal of Forecasting. – 1998. – Vol. 14, Issue 1. – P. 35–62.
8. Cheng, H. et al. Multistep-ahead time series prediction / H. Cheng et al. // Lecture Notes in Computer Science. – 2006. – Vol. 3918. – P. 765–774.
9. Рябко, Б.Я. Прогнозирование случайных последовательностей и универсальное кодирование. / Б.Я. Рябко // Проблемы передачи информации. – 1988. – №24. – С. 3-14.
10. Рябко, Б.Я. Экспериментальное исследование методов прогнозирования, основанных на алгоритмах сжатия данных / Б.Я. Рябко, В.А. Монарёв // Проблемы передачи информации. – 2005. – №41. – С. 74–78.
11. Ryabko, B. Compression-Based Methods for Nonparametric Prediction and Estimation of Some Characteristics of Time Series. / B. Ryabko // IEEE Transactions on Information Theory. – 2009. – Vol. 55, Issue 9. – P. 4309–4315.
12. Рябко, Б.Я. Дважды универсальное кодирование / Б.Я. Рябко // Проблемы передачи информации. – 1984. – Т. 20, № 3. – С. 24–28.

13. Кричевский, Р. Связь между избыточностью кодирования и достоверностью сведений об источнике. / Р. Кричевский // Проблемы передачи информации. – 1968. – №4. – С. 48–57.
14. Krichevsky, R. Universal Compression and Retrival. / R. Krichevsky. – Kluver Academic Publishers, 1993.
15. Приставка, П.А. Экспериментальное исследование метода прогнозирования, основанного на универсальных кодах / П.А. Приставка // Вестник СибГУТИ. – 2010. – №4. – С. 26–35.
16. Ryabko, B. Adaptive Coding and prediction of sources with large and infinite alphabets / B. Ryabko, J. Astola, A. Gammerman // IEEE transactions on information theory. – 2008. – Vol. 54, Issue 8. – P. 3808–3813.
17. Palit, A.K. Computational Intelligence in Time Series Forecasting: Theory and Engineering Applications (Advances in Industrial Control). / A.K. Palit, D. Popovic – New York, NJ, USA: Springer Verlag Inc, 2005.
18. Nevill–Manning, C.G. Lexically-Generated Subject Hierarchies for Browsing Large Collections / C.G. Nevill–Manning, I.H. Witten, G.W. Paynter // International Journal of Digital Libraries. – 1999. – Vol. 2, Issue 3. – P. 111–123.
19. Nevill-Manning, C.G. Identifying Hierarchical Structure in Sequences: A linear-time algorithm / C.G. Nevill-Manning, I.H. Witten // Journal of Artificial Intelligence Research. – 1997. – Vol. 7. – P. 67–82.
20. Clements, M.P. et al. Forecasting economic and financial time-series with non-linear models / M.P. Clements et al. // International Journal of Forecasting. – 2004. – Vol. 20, Issue 2. – P. 169—183
21. Web-сайт Institute journal of forecasters: http://forecasters.org/resources/time-series-data.
22. Web-сайт Independent statistics and Analysis. U.S. Energy information administration: http://www.eia.gov/petroleum/gasdiesel/.
23. 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.
24. 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.

25. 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.
26. Ryabko, B. Applications of Universal Source Coding to Statistical Analysis of Time Series. / B. Ryabko // Selected Topics in Information and Coding Theory. – 2010. – World Scientific Publishing. – P. 289–338.
27. Donskoy, V.I. Splitting criteria, binary decision tree synthesis, and algorithm LISTBB / V.I. Donskoy // Intellectual Archive. – 2013. – №1058. – 25 p.
28. Breiman, Leo. Classification and regression trees. / Leo Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone. – Monterey, CA: Wadsworth & Brooks, 1984.
29. Breiman, Leo. Random Forests. / Leo Breiman // Machine Learning. – 2001. – T. 45, №1. – P. 5–32.
30. Ho, Tin Kam. Random Decision Forest. / Tin Kam Ho // Proceedings of the 3rd International Conference on Document Analysis and Recognition. – Montreal, QC, 1995. – P. 278–282.
31. Breiman, L. Bagging Predictors. / L. Breiman // Machine Learning. – 1996. – P. 123–140.
32. 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.
33. 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.
34. 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. – P. 44–47.
35. 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.

36. Brockwell, P.J. Introduction to Time Series and Forecasting. / P. J. Brockwell, R. A. Davis. – Springer Publication, 2nd edition, 2003.
37. Makridakis, S. Forecasting: Methods and Applications / S. Makridakis, S. Wheelwright, R. J. Hyndman. – New York: John Wiley & Sons, 3rd edition, 1998.
38. Fildes, R. The Most Influential Articles in Forecasting / R. Fildes, P. Geoff Allen, eds. – SAGE Publications Ltd: Forecasting, 2011. – 2104 p.

Работы автора, в которых изложены основные результаты 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. 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.

ПРИЛОЖЕНИЕ А

Акты о внедрении результатов работы

Список организаций, в которых проводилось внедрение работы:

- ООО ПКФ «Техпром»;
- ООО «РТИ-Югра»;
- ФГБОУ ВО «Новосибирский национальный исследовательский государственный университет»;
- ФГБОУ ВО «Сибирский государственный университет телекоммуникаций и информатики».

ПКФ «ТЕХПРОМ»

Общество с Ограниченной Ответственностью
ПКФ«ТЕХПРОМ»

628400 Российская федерация,
Тюменская область,
Ханты-Мансийский Автономный округ
г. Сургут, ул. Производственная 2/1,оф. 11
Тел./факс: (3462) 45-30-30, 45-30-31

р/с 40702810600000014320
в ОАО «УРАЛТРАНСБАНК» г. Екатеринбург,
к/с 30101810200000000767
ИНН 8602009848 КПП 860201001
БИК 046551767

17.04.2015

В диссертационный совет ДМ 003.046.01

Институт вычислительных технологий СО РАН

630090, г. Новосибирск, пр. Ак. М.А. Лаврентьева, 6

АКТ О ВНЕДРЕНИИ

результатов диссертационной работы Лысяка А.С.

«Разработка и исследование теоретико-информационных методов прогнозирования временных рядов»

Настоящим подтверждается, что результаты диссертационного исследования Лысяка А.С. на тему: «Разработка и исследование теоретико-информационных методов прогнозирования временных рядов» обладают актуальностью и представляют большой практический интерес при моделировании поведения маркетинговых показателей работы коммерческих фирм. Предложенные в диссертационном исследовании Лысяка А.С. методы моделирования поведений были использованы в организации ООО ПКФ "Техпром" при моделировании показателей спроса и предложения по отраслям, что существенно повысило эффективность маркетинговой политики компании.

Генеральный директор
ООО ПКФ "Техпром"

Кучерявенко А.В.

Российская Федерация
Ханты-Мансийский автономный округ-Югра
Общество с ограниченной ответственностью
« РТИ-ЮГРА »

Юридический адрес: 628402, РФ, ХМАО-Югра, г. Сургут, ул. Нагорная, д. 9 кв. 38
Почтовый адрес: 628402, РФ, ХМАО-Югра, г. Сургут, ул. Нагорная, д. 9 кв. 38
ОГРН 1148602001470 ИНН 8602213226 КПП 860201001
☎ 8 (3462) 70-20-80 🖷 8 (3462) 25-86-72

Исх. №01-15 от "04" апреля 2015 г.

В диссертационный совет ДМ 003.046.01
Институт вычислительных технологий СО РАН
630090, г. Новосибирск, пр. Ак. М.А. Лаврентьева, 6

АКТ О ВНЕДРЕНИИ

результатов диссертационной работы Лысяка А.С.
«Разработка и исследование теоретико-информационных
методов прогнозирования временных рядов»

Настоящим подтверждается, что результаты диссертационного исследования Лысяка А.С. на тему: «Разработка и исследование теоретико-информационных методов прогнозирования временных рядов» обладают высокой актуальностью и представляют практический интерес при построении систем автоматической торговли на валютных биржах. Предложенные в диссертационном исследовании Лысяка А.С. методы прогнозирования были использованы в организации ООО «РТИ-Югра» при создании экспертных систем автоматической торговли на валютной бирже Forex, где показали свою высокую эффективность.

Генеральный директор **А.Б.Богатырев**

Исполнитель

"УТВЕРЖДАЮ"
ректор ФАОУ ВПО НГУ
д.ф.-м.н., проф. М. П. Федорук
«04» апреля 2015 г.

АКТ ВНЕДРЕНИЯ

результатов диссертационной работы Лысяка А.С.
«Разработка и исследование теоретико-информационных методов прогнозирования временных рядов»

Комиссия в составе: председатель комиссии Лаврентьев М.М., проф, и.о. декана ФИТ и члены комиссии Пищик Б.Н., к.т.н., заведующий кафедрой КС ФИТ, Романенко А.А., к.т.н., зам. декана по научной работе ФИТ констатирует, что результаты диссертационной работы ассистента кафедры Компьютерных систем федерального автономного образовательного учреждения высшего образования «Новосибирский национальный исследовательский государственный университет» (НГУ) Лысяка А.С. внедрены в учебный процесс на кафедре Компьютерных систем в курсе «Защита информации» (бакалавриат) и «Современные проблемы информатики» (магистратура) по направлению подготовки 230100 «Информатика и вычислительная техника».

Председатель комиссии /Лаврентьев М.М./

Члены комиссии /Пищик Б.Н./

/Романенко А.А./

"04" апреля 2015 г.

"УТВЕРЖДАЮ"
И.о. ректора
ФГОБУ ВПО СибГУТИ
В.Г. Беленький
_____
«07» 04 2015 г.

АКТ ВНЕДРЕНИЯ

результатов диссертационной работы Лысяка А.С.
«Разработка и исследование теоретико-информационных методов прогнозирования временных рядов»

Комиссия в составе:

Декан факультета ИВТ, д.т.н., проф.	Трофимов В.К. (председатель)
Зав. каф. ПМиК, д.т.н., проф	Фионов А.Н.
Доцент кафедры ПМиК, к.т.н.	Ситняковская Е. И.

констатирует, что результаты диссертационной работы ассистента кафедры Компьютерных систем федерального автономного образовательного учреждения высшего образования «Новосибирский национальный исследовательский государственный университет» (НГУ) Лысяка А.С. внедрены в учебный процесс на кафедре ПМиК в программе курса «Современные проблемы информатики» (магистратура) по направлению подготовки 09.04.01 «Информатика и вычислительная техника».

Председатель комиссии /Трофимов В.К./

Члены комиссии /Фионов А.Н./

/Ситняковская Е. И./

"УТВЕРЖДАЮ"
И.о. ректора
ФГОБУ ВПО СибГУТИ
В.Г. Беленький
_____
«07» 07 2015 г.

АКТ ВНЕДРЕНИЯ

результатов диссертационной работы Лысяка А.С.
«Разработка и исследование теоретико-информационных методов прогнозирования временных рядов»

Комиссия в составе:

Декан факультета ИВТ, д.т.н., проф.	Трофимов В.К. (председатель)
Зав. каф. ПМиК, д.т.н., проф	Фионов А.Н.
Доцент кафедры ПМиК, к.т.н.	Ситняковская Е. И.

Констатирует, что результаты диссертационной работы ассистента кафедры Компьютерных систем федерального автономного образовательного учреждения высшего образования «Новосибирский национальный исследовательский государственный университет» (НГУ) Лысяка А.С. были использованы при выполнении работ по Федеральной Целевой Программе Миноборнауки РФ «Разработка теоретико-информационных методов оценки и повышения производительности компьютерных систем и сетей передачи данных » (соглашение 8229 от 6.08.2012).

Председатель комиссии /Трофимов В.К./

Члены комиссии /Фионов А.Н./

/Ситняковская Е. И./

"УТВЕРЖДАЮ"
И.о. ректора
ФГОБУ ВПО СибГУТИ
В.Г. Беленький
_____
«07» 04 2015 г.

АКТ ВНЕДРЕНИЯ

результатов диссертационной работы Лысяка А.С.
«Разработка и исследование теоретико-информационных методов прогнозирования временных рядов»

Комиссия в составе:

Декан факультета ИВТ, д.т.н., проф.	Трофимов В.К. (председатель)
Зав. каф. ПМиК, д.т.н., проф	Фионов А.Н.
Доцент кафедры ПМиК, к.т.н.	Ситняковская Е. И.

Констатирует, что результаты диссертационной работы ассистента кафедры Компьютерных систем федерального автономного образовательного учреждения высшего образования «Новосибирский национальный исследовательский государственный университет» (НГУ) Лысяка А.С. были использованы при выполнении работ по Федеральной Целевой Программе Миноборнауки РФ «Эффективные методы построения защищённых высокоскоростных каналов передачи цифровых данных для предоставления доступа к широкополосным мультимедийным услугам» (соглашение 8329 от 6.08.2012).

Председатель комиссии /Трофимов В.К./

Члены комиссии /Фионов А.Н./

/Ситняковская Е. И./