На правах рукописи

Чирихин Константин Сергеевич

Использование методов теории информации и искусственного интеллекта для разработки и исследования высокоточных методов прогнозирования временных рядов

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

Автореферат
диссертации на соискание учёной степени
кандидата технических наук

Новосибирск — 2022

Работа выполнена в Федеральном государственном автономном образовательном учреждении высшего образования «Новосибирский национальный исследовательский государственный университет».

Научный руководитель: доктор технических наук, профессор
Рябко Борис Яковлевич

Официальные оппоненты: **Лемешко Борис Юрьевич,**
доктор технических наук, профессор, Новосибирский государственный технический университет, г. Новосибирск, профессор кафедры теоретической и прикладной информатики

Трифонов Петр Владимирович,
доктор технических наук, доцент, Университет ИТМО, г. Санкт-Петербург, профессор факультета безопасности информационных технологий

Ведущая организация: Федеральное государственное бюджетное образовательное учреждение высшего образования Томский государственный университет систем управления и радиоэлектроники

Защита состоится 29.09.2022 в 10:00 на заседании диссертационного совета Д 999.141.03 на базе Федерального государственного бюджетного учреждения науки Института динамики систем и теории управления им. В. М. Матросова Сибирского отделения Российской академии наук, Федерального государственного бюджетного научного учреждения «Федеральный исследовательский центр информационных и вычислительных технологий», федерального государственного бюджетного образовательного учреждения высшего образования «Сибирский государственный университет телекоммуникаций и информатики» по адресу: 630090, г. Новосибирск, пр. Академика Лаврентьева, 6, конференц-зал ФИЦ ИВТ.

С диссертацией можно ознакомиться в библиотеке и на сайте Федерального государственного бюджетного научного учреждения «Федеральный исследовательский центр информационных и вычислительных технологий»: http://www.ict.nsc.ru/ru/structure/discouncil/chirikhin.

Автореферат разослан «__» _____ 2022 г.

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

Общая характеристика работы

Актуальность исследования. Задача прогнозирования временных рядов обладает большой практической значимостью и привлекает внимание сотен исследователей во всём мире. Она возникает в самых разных областях человеческой деятельности. Например, методы прогнозирования временных рядов используются в экономике[1], здравоохранении[2], экологии[3], социальных науках[4] и др.

В настоящее время существует много подходов к решению данной задачи и уже получен ряд эффективных методов в рамках математической статистики, а в последние годы и с применением алгоритмов искусственного интеллекта (ИИ). Но несмотря на полученные методы, ряд важных проблем прогнозирования временных рядов остаётся нерешённым. Так, не решена задача выявления скрытых закономерностей в данных для использования их при прогнозировании. Существование подобных закономерностей возможно, например, в рядах экономических показателей (таких как биржевые цены), поскольку участники соответствующих процессов могут оказывать влияние на значения этих показателей, используя для этого сложные алгоритмы. В качестве другого примера можно привести временных ряды, связанные с космическими объектами. Вероятно, простейшим примером последовательности с закономерностью, которую известные методы не способны обнаружить, может считаться 01001000100001 . . . Поэтому задача разработки методов прогнозирования, способных выявлять подобные закономерности и использовать их для повышения точности прогнозов, является актуальной.

Объект исследования — алгоритмы и методы прогнозирования временных рядов социальных, экономических, физических и других показателей.

Предмет исследования — разработка и применение алгоритмов и методов прогнозирования стационарных и нестационарных временных рядов.

Цель работы — разработка, программная реализация и экспериментальное исследование методов прогнозирования временных рядов, способных обнаруживать не только периоды и простейшие тренды, но и «сложные» скрытые

[1] *Ghysels E.*, *Marcellino M.* Applied economic forecasting using time series methods. New York : Oxford University Press, 2018. 616 p.; *Franses P. H.*, *Dijk D.*, *Opschoor A.* Time Series Models for Business and Economic Forecasting (2nd ed.) Cambridge : Cambridge University Press, 2014. 311 p.; *Kim K.* Financial time series forecasting using support vector machines // Neurocomputing. 2003. Vol. 55, no. 1. P. 307–319.

[2] *Chimmula V. K. R.*, *Zhang L.* Time series forecasting of COVID-19 transmission in Canada using LSTM networks // Chaos, Solitons & Fractals. 2020. Vol. 135. P. 1–6.

[3] *Baum C. F.*, *Hurn S.* Environmental Econometrics Using Stata. College Station, Texas : Stata Press, 2021. 416 p.

[4] Time Series Analysis for the Social Sciences / J. M. Box-Steffensmeier [et al.]. Cambridge : Cambridge University Press, 2014. 297 p.

нестационарные закономерности в данных и использовать их для повышения точности прогнозов.

Задачи исследования:

1. Построение методов прогнозирования временных рядов, использующих алгоритмы искусственного интеллекта для обнаружения скрытых закономерностей в данных с целью повышения точности прогнозов;
2. Разработка метода преобразования алгоритмов прогнозирования к методам сжатия данных, что позволит: 1) использовать алгоритмы ИИ, применяемые в методах сжатия; 2) объединять различные методы прогнозирования в один комплекс;
3. Решение задачи выбора лучшего метода прогнозирования из имеющихся без их «полного» перебора с небольшими затратами времени.

Научная новизна:

1. Впервые построены методы прогнозирования, способные для повышения точности прогнозов находить в данных сложные закономерности за счёт использования многоголовочных автоматов и контекстно-свободных (КС) грамматик;
2. Впервые предложен и реализован теоретико-информационный метод интеграции нескольких методов прогнозирования, позволяющий «автоматически» выбирать наиболее подходящий метод для прогнозирования ряда;
3. Построены новые методы прогнозирования временных рядов, представляющие практический интерес (например, они были использованы для прогнозирования показателей Новосибирской области).

Теоретическая и практическая значимость работы. Предлагаемые в данной работе методы могут быть использованы для прогнозирования временных рядов из различных областей человеческой деятельности, но в первую очередь они разработаны для экономических рядов и рядов, связанных с космическими объектами. Например, результаты экспериментального исследования показывают, что методы обладают высокой точностью при прогнозировании среднемесячных чисел солнечных пятен и временного ряда Т-индекса — точность получаемых прогнозов на один шаг на исторических данных оказалась выше, чем точность прогнозов обсерваторий.

Результаты диссертации внедрены в учебный процесс на кафедре Компьютерных систем ФИТ НГУ при обучении аспирантов по направлению подготовки 09.06.01 Информатика и вычислительная техника, а также в учебный процесс на кафедре Прикладной математики и кибернетики СибГУТИ при обучении бакалавров по направлению подготовки 09.03.01 Информатика и вычислительная техника (акты о внедрении приведены в Приложении А диссертации).

Основная часть работы выполнена в рамках следующих проектов:

1. Проект РФФИ № 19-37-90009 «Методы прогнозирования временных рядов, базирующиеся на алгоритмах сжатия данных и искусственного интеллекта» (конкурс «Аспиранты»);
2. Проект РФФИ № 19-47-540001 «Разработка когнитивных методов прогнозирования и их применение для предсказания социально-экономических процессов в Новосибирской области». Отметим, что представленные в диссертации результаты прогнозирования показателей Новосибирской области были выполнены в рамках этого проекта, что является практическим внедрением разработанных методов (акт о внедрении в ФИЦ ИВТ приведён в Приложении А диссертации);
3. Проект «Социально-экономическое развитие Азиатской России на основе синергии транспортной доступности, системных знаний о природно-ресурсном потенциале, расширяющегося пространства межрегиональных взаимодействий», проводимый при финансовой поддержке Российской Федерации в лице Министерства науки и высшего образования России, Соглашение № 075-15-2020-804 от 02 октября 2020 г. (грант № 13.1902.21.0016).

Методология и методы исследования. При разработке алгоритмов использовались методы из теории информации, дискретной математики, искусственного интеллекта, теории сложности вычислений. Для исследования точности полученных методов выполнялось построение прогнозов для уже известных значений.

Основные положения, выносимые на защиту:

1. Разработаны методы прогнозирования временных рядов, способные находить сложные закономерности новых классов в данных за счёт использования многоголовочных автоматов и КС-грамматик;
2. Разработаны эффективные (с точки зрения точности и трудоёмкости вычислений) методы объединения алгоритмов прогнозирования временных рядов в один, имеющий наилучшую точность (из объединяемых);
3. Использование сжатия данных при прогнозировании позволяет применять реализованные в алгоритмах сжатия методы ИИ, а также объединять различные методы прогнозирования в один;
4. Предлагаемые методы обладают высокой точностью, что подтверждается результатами вычислительных экспериментов, проводимых с реальными данными.

Соответствие паспорту специальности. В работе получены результаты, соответствующие трём пунктам паспорта специальности 05.13.18 — «Математическое моделирование, численные методы и комплексы программ» по техническим наукам:

1. Разработка новых математических методов моделирования объектов и явлений;

2. Разработка, обоснование и тестирование эффективных вычислительных методов с применением современных компьютерных технологий;
3. Комплексные исследования научных и технических проблем с применением современной технологии математического моделирования и вычислительного эксперимента.

Представление работы. Основные результаты работы докладывались на:

– XXI Всероссийская конференция молодых учёных по математическому моделированию и информационным технологиям (Россия, Новосибирск, 2020);
– The 34th annual European Simulation and Modelling Conference (France, Toulouse, 2020);
– Российская научно-техническая конференция «Обработка информации и математическое моделирование» (Россия, Новосибирск, 2020);
– Международная научно-практическая конференция «Распределенные информационно-вычислительные ресурсы: цифровые двойники и большие данные» (DICR 2019) (Россия, Новосибирск, 2019);
– International Workshop «Applied Methods of Statistical Analysis. Statistical Computation and Simulation — AMSA'2019» (Russia, Novosibirsk, 2019);
– The 39th International Symposium on Forecasting (Greece, Thessaloniki, 2019).

Публикации. Основные результаты по теме диссертации изложены в 9 печатных изданиях: 1 статья в журнале, индексируемом в WoS и Scopus, 1 статья в журнале, индексируемом в Scopus и входящим в перечень ВАК, 1 статья в журнале из перечня ВАК и 6 публикаций, включённых в сборники трудов конференций. Получено свидетельство о государственной регистрации программы для ЭВМ.

Содержание работы

Во **введении** обосновывается актуальность исследований, проводимых в рамках данной диссертационной работы, формулируется цель, ставятся задачи исследования, излагается научная новизна и практическая значимость представляемой работы.

В **главе 1** развивается подход к прогнозированию временных рядов, основанный на использовании методов сжатия данных.

В параграфе 1.1 приводится постановка рассматриваемой задачи прогнозирования, заключающаяся в следующем. Пусть дана последовательность целых или вещественных чисел $x_1, x_2, \ldots, x_t$, порождённая некоторым вероятностным источником P, где x_i принадлежит множеству $\mathcal{A}$, называемому алфавитом. Предполагается, что $\mathcal{A}$ является либо конечным множеством целых чисел, либо ограниченным интервалом. Если $\mathcal{A}$ является конечным множеством

целых чисел, то требуется найти оценку для условного распределения вероятностей $P(x_{t+1} = a_{i_1}, x_{t+2} = a_{i_2}, \ldots, x_{t+h} = a_{i_h} | x_1, x_2, \ldots, x_t)$, где $a_{i_j} \in \mathcal{A}$, h — положительное целое. Если $\mathcal{A}$ является ограниченным интервалом, то требуется найти оценку для многомерной условной плотности вероятности $P(x_{t+1}, x_{t+2}, \ldots, x_{t+h} | x_1, x_2, \ldots, x_t)$. Точечный прогноз $\hat{x}_{t+i|t}$ для значения x_{t+i}, построенный с использованием первых t значений процесса, вычисляется как математическое ожидание по маргинальному условному распределению вероятностей (условной многомерной плотности вероятности).

В параграфе 1.2 содержится краткое изложение истории развития некоторых широко используемых в настоящее время методов прогнозирования временных рядов. Он начинается с рассмотрения методов на основе экспоненциального сглаживания, далее рассматриваются модели авторегрессии-скользящего среднего, затем модели авторегрессионной условной гетероскедастичности. Наконец, рассматриваются методы на основе моделей искусственного интеллекта и сжатия данных.

В начале параграфа 1.3 приводятся основные сведения о теоретико-информационном подходе к прогнозированию временных рядов. Пусть задана последовательность $X = x_1, x_2, \ldots, x_t$, в которой x_i принадлежит некоторому конечному $\mathcal{A}$, и требуется вычислить прогноз на h шагов вперёд. Для этого строится оценка для неизвестного условного распределения вероятностей следующих h значений X с помощью алгоритма сжатия данных без потерь φ (неформально φ иногда будем называть архиватором):

$$P_\varphi(x_{t+1} = a_{i_1}, \ldots, x_{t+h} = a_{i_h} | x_1, \ldots, x_t) = \frac{2^{-|\varphi(x_1,\ldots,x_t,a_{i_1},\ldots,a_{i_h})|}}{\sum\limits_{(b_{j_1},\ldots,b_{j_h}) \in \mathcal{A}^h} 2^{-|\varphi(x_1,\ldots,x_t,b_{j_1},\ldots,b_{j_h})|}}, \quad (1)$$

где $a_{i_k} \in \mathcal{A}$, $|\varphi(\alpha)|$ — длина в битах кодового слова для α.

Известно[5], что если φ является универсальным кодом, а вероятностная мера P стационарной и эргодической, то P_φ является в определённом смысле непараметрической оценкой для P.

Если требуется дать точечные прогнозы на h шагов, то в качестве прогнозов можно использовать математические ожидания, вычисленные по соответствующим маргинальным распределениям вероятностей.

Далее в параграфе рассматривается прогнозирование вещественных временных рядов. Отмечается, что для прогнозирования с использованием методов сжатия данных такие ряды необходимо предварительно преобразовывать к рядам с конечными алфавитами с помощью процедуры квантования. В простейшем варианте эта процедура заключается в разбиении отрезка, содержащего все значения ряда, на некоторое конечное число n равных пронумерованных интервалов и

[5] *Ryabko B.* Applications of Kolmogorov complexity and universal codes to nonparametric estimation of characteristics of time series // Fundamenta Informaticae. 2008. Vol. 83, no. 1/2. P. 177–196.

последующей замене каждого элемента ряда номером интервала, в который этот элемент попадает. Для «автоматического» выбора значения n, обеспечивающего хорошую точность, предлагается использовать следующий подход. Пусть $n = 2^k$, где k — некоторое положительное целое число. При построении прогноза будем использовать временные ряды, получающиеся при квантовании с разбиениями области возможных значений исходного ряда Y на 2^i, $i = 1, 2, \ldots, k$, интервалов. Если временной ряд $X = x_1, x_2, \ldots, x_t$, $x_j \in \mathcal{A}_i = \{0, 1, \ldots, 2^i - 1\}$, получен из ряда $Y = y_1, y_2, \ldots, y_t$, $y_j \in \mathbb{R}$, с помощью процедуры квантования с использованием 2^i интервалов, то будем обозначать X как $X^{[i]} = x_1^{[i]}, x_2^{[i]}, \ldots, x_t^{[i]}$, $x_j^{[i]} \in \mathcal{A}_i$. Будем вычислять $P_\varphi(x_1^{[k]}, x_2^{[k]}, \ldots, x_t^{[k]})$ как

$$P_\varphi(x_1^{[k]}, x_2^{[k]}, \ldots, x_t^{[k]}) = \frac{\sum\limits_{i=1}^{k} \omega_i 2^{-|\varphi(x_1^{[i]}, x_2^{[i]}, \ldots, x_t^{[i]})| + t(k-i)}}{\sum\limits_{i=1}^{k} \sum\limits_{(z_1, z_2, \ldots, z_t) \in \mathcal{A}_i^t} \omega_i 2^{-|\varphi(z_1, z_2, \ldots, z_t)| + t(k-i)}}, \tag{2}$$

где ω_i — неотрицательные весовые коэффициенты, причём $\sum\limits_{i=1}^{k} \omega_i = 1$. Варьируя значения этих коэффициентов, можно отдавать предпочтение разбиениям с большим или меньшим количеством интервалов.

В параграфе 1.4 предлагается способ объединения нескольких методов сжатия в один метод прогнозирования, заключающийся в следующем. Предположим, что задано некоторое конечное множество методов сжатия $F = \{\varphi_1, \varphi_2, \ldots, \varphi_k\}$. Для построения прогноза на h шагов для последовательности $X = x_1, x_2, \ldots, x_t$ с использованием всех методов из F предлагается следующая формула:

$$P_F(x_{t+1} = a_{i_1}, \ldots, x_{t+h} = a_{i_h} | x_1, \ldots, x_t) = \frac{\sum\limits_{i=1}^{k} \gamma_i 2^{-|\varphi_i(x_1, \ldots, x_t, a_{i_1}, \ldots, a_{i_h})|}}{\sum\limits_{(b_{j_1}, \ldots, b_{j_h}) \in \mathcal{A}^h} \sum\limits_{i=1}^{k} \gamma_i 2^{-|\varphi_i(x_1, \ldots, x_t, b_{j_1}, \ldots, b_{j_h})|}}, \tag{3}$$

где γ_i — неотрицательные весовые коэффициенты, причём $\sum\limits_{i=1}^{k} \gamma_i = 1$.

Отмечается, что наибольшее влияние на распределение вероятностей для $x_{t+1}, \ldots, x_{t+h}$, получаемое по формуле (3), оказывает метод, обеспечивающий наилучшую степень сжатия для X с некоторым возможным продолжением $b_{j_1}, \ldots, b_{j_h}$ на конце. Поэтому можно сказать, что в процессе вычислений по (3) наилучший для прогнозирования X метод «автоматически выбирается» из F.

В параграфе 1.5 рассматривается прогнозирование многомерных временных рядов с помощью методов сжатия данных. Пусть даны k рядов

$X_1, X_2, \ldots, X_k$, $X_j = x_{1j}, x_{2j}, \ldots, x_{tj}$, $j \in \{1, 2, \ldots, k\}$, и их элементы принимают значения из алфавита $\mathcal{A} = \{0, 1, \ldots, n-1\}$. Требуется построить прогноз на h шагов для каждого ряда. Построим новый ряд $X' = x'_1, x'_2, \ldots, x'_t$ с алфавитом $\mathcal{A}' = \{0, 1, \ldots, n^k - 1\}$ по следующему правилу:

$$x'_i = \sum_{j=1}^{k} x_{ij} |\mathcal{A}|^{j-1}. \tag{4}$$

Ясно, что по X можно построить прогнозные значения $x'_{t+1}, x'_{t+2}, \ldots, x'_{t+h}$ с помощью описанных ранее методов. Разложив x'_{t+j} обратно по формуле (4), получаем прогнозные значения для $X_1, X_2, \ldots, X_k$.

В параграфе 1.6 приводится описание подхода к прогнозированию временных рядов, базирующегося на использовании формальных грамматик. Как известно[6], универсальные коды могут быть основаны на грамматических моделях. В таких кодах для сжатия данных осуществляется построение компактной формальной грамматики, из которой однозначно выводятся сжимаемые данные. Затем вместо сжатия и передачи исходных данных сжимается и передаётся построенная грамматика, получив которую декодер может однозначно восстановить исходное сообщение. Известно, что если поиск компактной грамматики производится в классе контекстно-свободных (КС) грамматик, то задача является NP-трудной[7]. В связи с этим в алгоритмах сжатия на основе КС-грамматик применяются приближённые алгоритмы. В настоящее время существуют и находятся в открытом доступе несколько реализаций кодов, основанных на грамматиках. Две из них используются в данной работе путём объединения их с другими методами сжатия по формуле (3).

Вторая глава посвящена описанию метода прогнозирования временных рядов, основанного на многоголовочных автоматах. Рассматривается только прогнозирование целочисленных одномерных временных рядов с алфавитами вида $\mathcal{A} = \{0, 1, \ldots, n-1\}$, поскольку вещественные и многомерные временные ряды могут быть приведены к такому виду.

В параграфе 2.1 приводится мотивация разработки данного метода. Отмечается, что существующие алгоритмы прогнозирования временных рядов могут корректно обнаруживать периодические закономерности в данных, а также распознавать «запрещённые» комбинации символов. Например, в последовательности $X = 102333201231101$ после 1 никогда не встречается 3, поэтому при построении совместного распределения вероятностей для нескольких будущих значений X с помощью методов сжатия вероятности комбинаций, начинающихся с 3, будут низкими. Также уже разработаны методы для прогнозирования рядов с трендами.

[6] *Kieffer J.*, *Yang E.* Grammar-based codes: A new class of universal lossless source codes // IEEE Transactions on Information Theory. 2000. Vol. 46, no. 3. P. 737–754.

[7] The smallest grammar problem / M. Charikar [et al.] // IEEE Transactions on Information Theory. 2005. Vol. 51, no. 7. P. 2554–2576.

Однако существуют классы закономерностей, которые не могут быть обнаружены известными методами прогнозирования временных рядов. В качестве примера простой последовательности, содержащей закономерность подобного класса, приводится $010010001\ldots$

В параграфе 2.2 содержится описание алгоритма, способного обнаруживать подобные закономерности во временных рядах. Он основан на алгоритме[8] для прогнозирования полилинейных (multilinear) слов[9] и является его модификацией. Бесконечное слово называется полилинейным, если оно может быть записано в виде

$$q \prod_{n \geq 0} r_1^{a_1 n + b_1} r_2^{a_2 n + b_2} \ldots r_m^{a_m n + b_m},$$

где $\prod$ обозначает конкатенацию, q — некоторое конечное слово, m — положительное целое, для любого $i \in \{1, 2, \ldots, m\}$ r_i не пусто, а a_i и b_i — неотрицательные целые числа такие, что $a_i + b_i > 0$. Ясно, что слово $010010001\ldots$ является полилинейным, поскольку его можно записать как $\prod\limits_{n \geq 0} 0^{n+1}1$.

Было показано[8], что существует интеллектуальный (sensing) многоголовочный детерминированный конечный автомат (ИМДКА), способный правильно прогнозировать любое полилинейное слово над алфавитом $\mathcal{A}$, и приведён алгоритм его работы. Для функционирования ИМДКА входное бесконечное слово α записывается на ленту автомата. За один переход каждая из k головок ИМДКА либо остаётся на месте, либо передвигается на одну позицию вправо на ленте, автомат меняет своё состояние, а также выдаёт прогноз для следующего символа, формируя таким образом последовательность прогнозных значений $M(\alpha)$. Обозначим i-й символ α как $\alpha[i]$. Если записанное на ленту слово является полилинейным и автомат выполняет указанный алгоритм, то найдётся i_0 такое, что $\alpha[i] = M(\alpha)[i]$ для всех $i \geq i_0$. Для целей диссертационной работы нужно, чтобы этот алгоритм работал с конечными словами и результатом его выполнения была не последовательность одношаговых прогнозов автомата $M(\alpha)$, а оценка вероятности $P_M(X)$ появления входного конечного слова $X = x_1, x_2, \ldots, x_n$, $x_i \in \mathcal{A}$, на выходе неизвестного источника P. Это позволит получить длину кодового слова $|\varphi_M(X)|$, нужную для включения автомата в формулу (3) вместе с методами сжатия данных, как $-\log_2 P_M(X)$.

Для получения искомой оценки вероятности предлагается следующий подход. Вероятность появления слова вычисляется как произведение условных вероятностей появления отдельных его букв:

[8] *Smith T.* Prediction of infinite words with automata // Theory of Computing Systems. 2018. Vol. 62, no. 3. P. 653–681.

[9] *Smith T.* On infinite words determined by stack automata // IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS 2013). IIT Guwahati, India, 2013. P. 413–424.

$$P_M(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P_M(x_i|x_1, x_2, \ldots, x_{i-1}), \tag{5}$$

причём для первого символа $P_M(x_1|\ldots) = P_M(x_1) = 1/|\mathcal{A}|$.

В алгоритме[8] на некоторых шагах прогнозное значение, которое выдаёт автомат, строго указано (случай 1), в других местах в его качестве может быть выбран произвольный символ (случай 2). Оценки $P_M(x_i|x_1, x_2, \ldots, x_{i-1})$ предлагается по-разному вычислять для случаев 1 и 2.

Пусть уже просмотрено i букв X и оценивается вероятность $P_M(x_{i+1}|x_1, \ldots, x_i)$. Если согласно прогнозу автомата следующим символом будет $a \in \mathcal{A}$ и автомат «уверен» в своём выборе, т.е. имеем дело со случаем 1, то a приписывается вероятность

$$P_M(a|x_1,x_2, \ldots, x_i) = \frac{C + 1/2}{C + |\mathcal{A}|/2}, \tag{6}$$

где C — количество безошибочных предсказаний автомата, в которых он был «уверен», с момента последней ошибки (или с начала слова, если ошибок ещё не было). Всем остальным символам $\mathcal{A}$ приписываются одинаковые вероятности:

$$P_M(b|x_1,x_2, \ldots, x_i) = \frac{1/2}{C + |\mathcal{A}|/2}, \tag{7}$$

где $b \in \mathcal{A} \setminus \{a\}$.

Как только автомат обучится прогнозированию входного слова, C будет строго возрастать с увеличением i и вычисляемые по (6) вероятности будут стремиться к 1, а по (7) — к нулю.

Если автомат «не знает», какой символ будет следующим (случай 2), то для каждого символа $a \in \mathcal{A}$ его условная вероятность оценивается по контекстной модели нулевого порядка:

$$P_M(a|x_1, \ldots, x_i) = \frac{\nu_{x_1,\ldots,x_i}(a) + 1/2}{i + |\mathcal{A}|/2}, \tag{8}$$

где $\nu_{x_1,\ldots,x_i}(a)$ — количество вхождений буквы a в слово $x_1,x_2, \ldots, x_i$.

Далее в параграфе приведены несколько примеров прогнозирования полилинейных слов с помощью модифицированного алгоритма для автомата, методов сжатия, а также их комбинации по формуле (3). При прогнозировании полилинейных слов, не являющихся периодическими, методы сжатия данных ошибаются, в то время как автомат достаточно быстро начинает выдавать безошибочные прогнозы. Показано, что комбинация методов сжатия с автоматом также перестаёт ошибаться, но иногда для этого необходима более длинная предыстория процесса.

Третья глава посвящена описанию адаптивного метода прогнозирования, а также разработке способа преобразования произвольных методов прогнозирования к методам сжатия данных.

В параграфе 3.1 приводится мотивация разработки адаптивного метода. Отмечается, что вычисления по формуле (3) отличаются высокой трудоёмкостью, поскольку все последовательности, возникающие при прогнозировании, должны быть сжаты всеми используемыми архиваторами. Отсюда возникает потребность в методе сокращения трудоёмкости вычислений.

В параграфе 3.2 содержится описание адаптивного метода. Он основан на универсальных по времени кодах[10]. Предположим, что задано некоторое конечное множество методов сжатия $F = \{\varphi_1, \varphi_2, \ldots, \varphi_k\}$ и требуется сжать последовательность $X = x_1,x_2,\ldots,x_t$ символов из конечного алфавита $\mathcal{A}$ с помощью $\varphi_s \in F$, обеспечивающего наилучшее сжатие. Можно сжать небольшой фрагмент X, например, префикс $x_1, x_2, \ldots, x_r$, $r << t$, всеми архиваторами из F для того, чтобы выбрать лучший на этом фрагменте метод $\hat{\varphi}_s$ и затем сжать X целиком только с его помощью. При этом кроме сжатого представления X необходимо хранить номер s выбранного архиватора, иначе X не сможет быть декодирована. В представленной диссертационной работе этот подход применяется для целей прогнозирования. Заметим, что указанный фрагмент будет одинаковым у всех $|\mathcal{A}|^h$ временных рядов, получаемых при прогнозировании, поэтому его можно сжать только один раз, что позволяет сократить время вычислений даже при $r = t$. Добавим, что в адаптивном методе прогнозирования можно выбрать не один близкий к оптимальному метод сжатия, а несколько, и затем объединить их в один метод прогнозирования по формуле (3).

Раздел 3.3 посвящён описанию способа преобразования методов прогнозирования к методам сжатия данных. Это позволяет объединять методы сжатия с произвольными методами прогнозирования с возможным использованием адаптивного метода из предыдущей главы.

Пусть дан некоторый метод прогнозирования π. По последовательности чисел $Y = y_1, y_2, \ldots, y_t$ метод π должен быть способен выдавать прогнозное значение $\hat{y}_{t+1|t}$ для следующего члена последовательности: $\hat{y}_{t+1|t} = \pi(y_1, y_2, \ldots, y_t)$. Если Y — вещественный временной ряд, то к нему необходимо применить процедуру квантования из параграфа 1.3 для получения целочисленного ряда X, элементы которого принадлежат алфавиту $\mathcal{A} = \{0, 1, \ldots, n-1\}$, где n — количество интервалов при квантовании. Если же Y — целочисленный ряд, то какие-либо преобразования выполнять необязательно, т.е. $X = Y$, а $\mathcal{A}$ — множество всех значений X. Как и в случае с автоматом, требуется получить оценку вероятности $P_\pi(x_1, x_2, \ldots, x_t)$ появления X на выходе источника с помощью π, поскольку затем эту вероятность можно будет конвертировать в длину кодового слова как $-\log_2 P_\pi(X)$. Как и ранее, вероятность появления слова X вычисляется через произведение условных вероятностей его букв:

[10] *Ryabko B.* Time-Universal data compression // Algorithms. 2019. Vol. 12, no. 6. P. 1–10.

$$P_\pi(x_1, x_2, \ldots, x_t) = \prod_{i=1}^{t} P_\pi(x_i | x_1, x_2, \ldots, x_{i-1}), \tag{9}$$

причём для первого символа $P_\pi(x_1|\ldots) = P_\pi(x_1) = 1/|\mathcal{A}|$.

Условные вероятности с помощью π вычисляются следующим образом. Если $\pi(x_1, x_2, \ldots, x_{i-1})$, округлённое до ближайшего целого числа из $\mathcal{A}$, совпадает с x_i, то

$$P_\pi(x_i | x_1, x_2, \ldots, x_{i-1}) = \frac{(i-1) + 1/2}{(i-1) + |\mathcal{A}|/2}.$$

В противном случае,

$$P_\pi(x_i | x_1, x_2, \ldots, x_{i-1}) = \frac{1/2}{(i-1) + |\mathcal{A}|/2}.$$

Если π использует $k > 1$ предыдущих значений при вычислении прогноза для следующего символа, то k первых условных вероятностей оцениваются с помощью предсказателя Кричевского[11].

Отмечается, что в диссертации подобное преобразование было реализовано для модели Хольта-Уинтерса (Holt-Winters)[12].

В **четвёртой главе** представлены результаты экспериментального исследования разработанных методов.

Параграф 4.1 посвящён описанию методологии, использованной в вычислениях. Отмечается, что при построении прогнозов применялись 7 методов сжатия данных, основанных на разных подходах: zlib, bzip2, ppmd, rp, lcacomp, zstd и zpaq, а также автомат из главы 2 (rp и lcacomp используют КС-грамматики). При объединении различных алгоритмов по формуле (3), а также прогнозов, полученных с использованием разного числа интервалов при квантовании в формуле (2), были использованы равные веса, т.е. априорно не отдавалось предпочтение какому-либо алгоритму или разбиению. Для представления одного элемента временного ряда в памяти использовался 1 байт.

Для удаления присутствующих в прогнозируемых рядах трендов применялась процедура взятия n-ой разности. Также для уменьшения влияния выбросов использовалось сглаживание по следующей формуле:

$$\bar{x}_i = (2x_i + x_{i-1} + x_{i-2})/4. \tag{10}$$

Далее в параграфе описывается используемый подход к получению доверительных интервалов для прогнозируемых значений, а также отмечается, что для оценки точности построенных прогнозов чаще всего использовались средние абсолютные ошибки (mean absolute error, MAE) для каждого шага, реже использовались средние относительные ошибки.

[11] *Кричевский Р. Е.* Сжатие и поиск информации. М. : Радио и связь, 1989. 168 с.

[12] *Hyndman R.*, *Athanasopoulos G.* Forecasting: principles and practice. OTexts, 2018. 382 p.

Поскольку в формулах (1)–(3) количество сжимаемых последовательностей экспоненциально возрастает с числом шагов, на которое строится прогноз, при прогнозировании на большое количество шагов (больше 4 в данной работе) с целью сокращения трудоёмкости вычислений использовался следующий подход. Предположим для простоты, что количество элементов t в прогнозируемом временном ряде и число шагов h, на которое строится прогноз, являются чётными числами. Разобъём временной ряд X на два временных ряда $X_{\text{odd}} = x_1, x_3, \ldots, x_{t-1}$ и $X_{\text{even}} = x_2, x_4, \ldots, x_t$. Затем будем отдельно прогнозировать $x_{t+1}, x_{t+3}, \ldots, x_{t+h-1}$ по X_{odd} и $x_{t+2}, x_{t+4}, \ldots, x_{t+h}$ по X_{even}. Ясно, что аналогично можно разбить X на 3, 4 и т.д. временных ряда, а также воспользоваться данным подходом, если t и/или h являются нечётными числами.

В параграфе 4.2 приводятся результаты прогнозирования временных рядов из соревнования M3 Competition (M3C)[13] с помощью разработанных в рамках диссертации методов. Это соревнование было проведено в 2000 году и основной его целью являлось сравнение точности различных методов прогнозирования на реальных данных. На веб-сайте Международного института прогнозистов (`https://forecasters.org/resources/time-series-data/m3-competition/`, дата обращения 15.12.2017) размещены временные ряды из M3C вместе с их значениями за периоды времени, на которые требовалось построить прогнозы в рамках соревнования. Эти значения были недоступны участникам, но использовались организаторами для определения ошибок прогнозов. Поэтому сейчас имеется возможность провести аналогичные вычисления и выполнить сравнение точности полученных прогнозов с точностью прогнозов методов из M3C.

Для построения прогнозов в данном параграфе использовались три метода сжатия: zlib, ppmd и rp, а также их комбинация. При этом применялись метод STL[14] для выделения сезонной составляющей и процедура сглаживания по формуле (10). Кроме того, для всех категорий временных рядов осуществлялся переход к первой разности. Для сокращения трудоёмкости вычислений использовалась описанная в параграфе 4.1 процедура декомпозиции исходного временного ряда на несколько отдельных рядов: на два при прогнозировании ежегодных, ежеквартальных и «других» рядов и на 6 при прогнозировании ежемесячных рядов.

Для ежегодных данных и данных из категории «другие» метод прогнозирования на основе архиваторов показал точность, сравнимую с точностью методов, участвовавших в соревновании. Для ежемесячных и ежеквартальных данных в среднем по шагам 1–18 точность прогнозов оказалась более низкой, но тем не менее при прогнозировании на первые четыре шага сопоставимой с методами из M3 Competition.

[13] *Makridakis S.*, *Hibon M.* The M3-Competition: results, conclusions and implications // International Journal of Forecasting. 2000. Vol. 16, no. 4. P. 451–476.

[14] STL: A seasonal-trend decomposition / R. B. Cleveland [et al.] // Journal of Official Statistics. 1990. Vol. 6, no. 1. P. 3–73.

В параграфе 4.3 приводятся результаты прогнозирования физических данных. Параграф начинается с рассмотрения временного ряда среднемесячных чисел солнечных пятен. Отмечено, что Служба космической погоды (The Space Weather Services, SWS) Австралийского метеорологического бюро каждый месяц публикует свой прогноз для этого ряда. Кроме того, доступны для загрузки прогнозы, сделанные SWS в прошлые месяцы. Это позволяет сравнить точность прогнозов, получаемых с помощью предлагаемых методов, с точностью прогнозов бюро. Следуя схеме вычислений, изложенной в параграфе 4.1, были построены прогнозные значения на 4 шага для каждого месяца, начиная с февраля 2016, при этом использовались все 7 ранее перечисленных архиваторов, а также автомат. В вычислениях применялся адаптивный метод из параграфа 3.1, но для сравнения точности использовалось и простое объединение всех 8 методов по формуле (3) (этот метод назван комбинированным).

В адаптивном методе необходимо выбрать размер префикса ряда, который будет сжат всеми архиваторами для выбора близкого к оптимальному. В целях эксперимента, вычисления были проведены с 10 различными вариантами выбора размера этого префикса — сжимались от 10% до 100% значений ряда с шагом 10%. Максимальное рассматриваемое количество интервалов при квантовании было выбрано равным 16. Также от ряда бралась первая разность. В результате при использовании от 10% до 40% значений ряда в качестве лучшего метода выбирался ppmd, а для 50%–100% значений ряда лучшим методом становился zstd. Средние абсолютные ошибки для ppmd, zstd, комбинированного метода, а также прогнозов SWS приведены в таблице 1. Видно, что точность комбинированного метода при прогнозировании на 1 шаг вперёд в среднем оказалась выше, чем точность метода SWS. ppmd и zstd, выбираемые при использовании адаптивного подхода, обеспечивают точность, не уступающую точности комбинированного метода. Отмечается, что для эксперимента были построены прогнозы и с более ранней даты (май 1982), в этом случае zstd оказался точнее чем ppmd для шагов 1–3. При использовании 50% значений ряда в адаптивном методе ускорение вычислений по сравнению с комбинированным методом составило более чем 17 раз при построении одного прогноза на 4 шага вперёд.

Таблица 1 — Средние абсолютные ошибки, полученные при прогнозировании среднемесячных чисел солнечных пятен с использованием до 16 интервалов при квантовании

Метод	Номер шага			
	1	2	3	4
ppmd	7.2	9.2	10.1	10.2
zstd	8.1	10.3	11.8	13.3
ppmd+zstd	8.1	10.3	11.8	13.3
Комбинированный метод	8.1	10.3	11.8	13.3
SWS	8.3	9.1	9.7	9.8

Далее рассматривается прогнозирование временного ряда планетарного K-индекса (Planetary K-index, Kp-index). В работе используются значения ряда за период с 16.11.2020 по 07.12.2020, прогнозы на 4 шага вперёд строятся для второй половины значений этого ряда с помощью архиваторов zlib, rp и метода Хольта-Уинтерса. Средние абсолютные ошибки для прогнозов отдельных методов, а также для их комбинации с равными весами приведены в таблице 2. Как видно из этой таблицы, при прогнозировании данного ряда точнее оказывается модель Хольта-Уинтерса, при этом комбинированный метод практически не уступает ей в точности.

Таблица 2 — Средние абсолютные ошибки, полученные при прогнозировании второй половины временного ряда планетарного K-индекса

Метод	Номер шага			
	1	2	3	4
zlib	0.80	0.92	1.0	1.08
rp	1.06	1.34	1.08	1.46
holt-winters	0.58	0.71	0.76	0.72
zlib+rp+holt-winters	0.59	0.73	0.73	0.75

В конце раздела 4.3 рассматривается прогнозирование временного ряда T-индекса. В рамках диссертационной работы для каждого месяца с апреля 2011 года по апрель 2016 года был построен прогноз на 18 значений (в этом временном интервале отсутствуют пропущенные значения) и проведено сравнение точности этого прогноза с точностью прогноза метеорологического бюро. При этом использовалась та же методология, что и при прогнозировании данных М3С. При прогнозировании на 1 шаг предлагаемый метод оказался точнее метода метеорологической службы, в остальных случаях точность прогноза службы оказалась выше. Полученные результаты хорошо согласуются с результатами прогнозирования среднемесячных чисел солнечных пятен.

В параграфе 4.4 приводятся результаты прогнозирования социально-экономических показателей Новосибирской области (НСО). Отмечено, что рассматриваемые временные ряды могут быть найдены на официальном сайте Федеральной службы государственной статистики по НСО (`https://novosibstat.gks.ru`, дата обращения 15.03.2020). Для построения прогнозов совместно использовались архиваторы zlib, bzip2, ppmd, rp и автомат. Предварительно осуществлялся переход к первой или второй разности, а также применялось сглаживание по формуле (10). Для всех прогнозов строились доверительные интервалы.

Прогнозы были построены для следующих показателей НСО: прожиточный минимум (руб.), валовый региональный продукт (млн. руб.), количества браков и разводов, средний возраст матери, кредиторская и дебиторская задолженности организаций, средние цены на рынке жилья (руб. за 1 кв.м.), средняя ожидаемая продолжительности жизни, количество умерших, среднегодовая численность и

естественный прирост населения. Для некоторых показателей прогнозы были получены с помощью многомерного прогнозирования.

Отмечается, что приведённые в данном параграфе прогнозы были опубликованы в [1], и на момент их вычисления (15.03.2020) последние зафиксированные значения были доступны за 2018 год. По этой причине прогнозы не могут учитывать влияние пандемии COVID-19 — на исторических данных в период 2000–2018 гг. подобные явления в НСО не происходили. На текущий момент некоторые новые зафиксированные значения для всех прогнозируемых показателей стали известны, поэтому теперь возможно проанализировать точность построенных прогнозов. В диссертации приводится таблица, в ячейках которой содержатся зафиксированные и прогнозные значения для всех рядов. Анализируя эту таблицу, можно сделать вывод, что наиболее точными оказались прогнозы для среднегодовой численности населения (относительные ошибки прогнозов для 2019 и 2020 годов составили 0.035% и 0.671% соответственно), величины прожиточного минимума (относительные ошибки 3.715% для 2020 года, 4.331% для 2021-го и 5.299% для 2022-го), а также для количеств браков и разводов. Для естественного прироста населения и ожидаемой продолжительности жизни заметны большие ошибки прогнозов на 2020 год.

Пятая глава посвящена описанию разработанного программного комплекса, в котором реализованы предлагаемые методы прогнозирования.

В параграфе 5.1 перечисляются требования к программному комплексу, которые были учтены при его разработке. Особое внимание было уделено расширяемости и производительности, обеспечению кроссплатформенности, а также возможности интеграции комплекса с существующими системами анализа данных. В результате разработанная программа получилась состоящей из двух компонент: библиотеки itp_core, реализованной на языке C++, и основанного на ней пакета itp для Python.

Раздел 5.2 посвящён описанию библиотеки itp_core. Отмечается, что в ней реализованы: вычисление прогнозных значений для временного ряда по формулам (1)–(3), квантование (параграф 1.3), адаптивный метод прогнозирования из параграфа 3.2, процедура декомпозиции временного ряда на k более коротких рядов для их прогнозирования по отдельности (параграф 4.1), взятие n-ых разностей, а также базовые классы для работы с произвольными методами прогнозирования с модификацией, описанной в параграфе 3.3.

Раздел 5.3 посвящён описанию пакета itp. Использование этого пакета позволяет работать со стандартными средствами обработки и визуализации данных Python. Пакет обладает следующими возможностями: 1. Предоставляет доступ ко всей функциональности itp_core из Python; 2. Позволяет автоматически строить прогнозы для исторических значений рядов с целью оценки точности моделей и построения доверительных интервалов; 3. Поддерживает возможность добавления в itp_core новых методов прогнозирования, изначально не основанных

на сжатии, из Python; 4. Предоставляет возможность параллельного построения нескольких прогнозов с использованием библиотеки MPI (Message Passing Interface); 5. Предоставляет средства для визуализации исторических данных, прогнозов и доверительных интервалов, в том числе при работе в параллельном режиме; 6. Реализует модификацию из параграфа 3.3 для метода Хольта-Уинтерса.

В **заключении** приведены основные результаты работы, которые заключаются в следующем:

1. Разработан метод прогнозирования временных рядов, основанный на использовании сжатия данных с целью объединения нескольких методов прогнозирования в один, а также для использования алгоритмов искусственного интеллекта, реализованных в методах сжатия (таких, как построение компактной контекстно-свободной грамматики, из которой однозначно выводятся сжимаемые данные). В предлагаемом методе произвольные алгоритмы прогнозирования временных рядов преобразуются к методам сжатия и могут быть использованы совместно с универсальными кодами для вычисления прогнозов;
2. Разработан алгоритм прогнозирования временных рядов на основе многоголовочных конечных автоматов. Данный алгоритм способен правильно прогнозировать полилинейные слова, другие методы прогнозирования временных рядов этого делать не способны. За счёт объединения этого алгоритма с универсальными кодами, в том числе основанными на грамматических моделях, получен метод прогнозирования, способный находить «сложные» скрытые нестационарные закономерности в данных и использовать их для повышения точности прогнозов;
3. Построен адаптивный метод прогнозирования, базирующийся на универсальных по времени кодах и позволяющий существенно сократить время вычислений при объединении нескольких методов прогнозирования в один с помощью теоретико-информационного подхода. Использование данного метода при прогнозировании ряда солнечных пятен позволило сократить время вычислений более чем в 17 раз без потери точности прогнозов;
4. Выполнена программная реализация всех предложенных методов. Она включает в себя библиотеку itp_core, в которой поддержана работа с 7 алгоритмами сжатия данных, основанными на разных принципах, а также с предложенным в рамках настоящей работы алгоритмом на основе многоголовочных автоматов. На базе itp_core разработан пакет itp для Python, в котором реализовано преобразование для модели Хольта-Уинтерса, позволяющее рассматривать её как метод сжатия и использовать в itp_core, поддержана возможность организации параллельных вычислений, а также содержатся методы анализа и визуализации прогнозов;

5. С использованием пакета itp построены прогнозы для социально-экономических показателей Новосибирской области, временного ряда солнечных пятен, временных рядов T-индекса и планетарного K-индекса, а также рядов из M3 Competition. Проведено сравнение точности прогнозов, полученных с использованием предлагаемых методов, с точностью прогнозов для тех же данных, построенных сторонними исследователями и организациями, а также выполнено сравнение прогнозных значений с реальными значениями, неизвестными на момент прогнозирования. Результаты экспериментального исследования показывают, что предлагаемые методы обладают высокой точностью и могут быть использованы для прогнозирования временных рядов экономических, социальных, физических и других показателей.

В дальнейших исследованиях можно рассмотреть более сложные подходы к выбору близкого к оптимальному алгоритма в адаптивном методе. Например, использовать для этой цели методы многомерной оптимизации.

Публикации по теме диссертации

Статьи, опубликованные в рецензируемых научных журналах, входящих в перечень ВАК:

1. Чирихин К. С., Рябко Б. Я. Применение методов искусственного интеллекта и сжатия данных для прогнозирования социальных, экономических и демографических показателей Новосибирской области // Вычислительные технологии. — 2020. — Т. 25, № 5. — С. 80–90.
2. Чирихин К. С., Рябко Б. Я. Экспериментальное исследование точности методов прогноза, базирующихся на архиваторах // Вестник Новосибирского государственного университета. Серия: Информационные технологии. — 2018. — Т. 16, № 3. — С. 145–158.

Статьи, опубликованные в рецензируемых научных журналах, индексируемых в Scopus и Web of Science:

3. Chirikhin K., Ryabko B. Compression-Based Methods of Time Series Forecasting // Mathematics. — 2021. — Vol. 9, no. 3. — P. 1–11.

Свидетельство о государственной регистрации программы для ЭВМ:

4. Чирихин К. С. Пакет для Python «Information-theoretic predictor» — Свидетельство о государственной регистрации программы для ЭВМ № 2022613783 от 15 марта 2022 г.

Публикации в трудах конференций:

5. Чирихин К. С. Теоретико-информационный метод интеграции различных алгоритмов прогнозирования временных рядов // Тезисы XXI Всероссийской конференции молодых учёных по математическому моделированию и информационным технологиям. — Новосибирск, 2020. — С. 44.

6. Chirikhin K. Application of time-universal codes to time series forecasting // 34th Annual European Simulation and Modelling Conference, ESM 2020. — Toulouse, France, 2020. — P. 60–64.
7. Чирихин К. С. Реализация алгоритма прогнозирования временных рядов на основе методов сжатия данных и искусственного интеллекта // Тезисы Российской научно-технической конференции «Обработка информации и математическое моделирование». — Новосибирск, 2020. — С. 195–199.
8. Чирихин К. С. Применение методов сжатия данных и искусственного интеллекта для прогнозирования демографических и экономических показателей Новосибирской области // Распределенные информационно-вычислительные ресурсы. Цифровые двойники и большие данные (DICR-2019). — Новосибирск, 2019. — С. 238–243.
9. Chirikhin K. S., Ryabko B. Y. Application of artificial intelligence and data compression methods to time series forecasting // Applied Methods of Statistical Analysis. Statistical Computation and Simulation — AMSA'2019: Proceedings of the International Workshop. — Novosibirsk, 2019. — P. 553–560.
10. Chirikhin K. S., Ryabko B. Y. Application of data compression techniques to time series forecasting // The 39th International Symposium on Forecasting. — Thessaloniki, Greece, 2019. — P. 12.