Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ ИУ «ИНФОРМАТИКА И СИСТЕМЫ УПРАВЛЕНИЯ»

КАФЕДРА ИУ7 «ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ ЭВМ И ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ»

РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

НА ТЕМУ:

Методы хранения и прогнозирования временных рядов

Студент **ИУ7-52Б** _____ **Д.А. Тузов**

Руководитель _____ **А.С. Кострицкий**

2024 г.

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

УТВЕРЖДАЮ

Заведующий кафедрой ИУ7
(Индекс)
_____ И. В. Рудаков
(И.О.Фамилия)
«20» сентября 2024 г.

З А Д А Н И Е
на выполнение научно-исследовательской работы

по теме

Методы хранения и прогнозирования временных рядов

Студент группы **ИУ7-52Б**

Тузов Даниил Александрович

Направленность НИР (учебная, исследовательская, др.): **учебная.**

Источник тематики (кафедра, предприятие, НИР): **кафедра.**

График выполнения НИР: 25% к 4 нед., 50% к 9 нед., 75% к 13 нед., 100% к 15 нед.

Техническое задание

Проанализировать предметную область: ввести основные определения, обозначить основные вехи развития. Формализовать задачу хранения и прогнозирования временных рядов. Перечислить методы или группы методов решения, сформулировать критерии сравнения. Сравнить перечисленные методы по сформулированным критериям.

Оформление научно-исследовательской работы:

1. Расчетно-пояснительная записка на **12-15** листах формата А4.
2. Перечень графического (иллюстративного) материала (плакаты, слайды и т. п.): Презентация на **3** слайдах. В презентации должны быть отражены формализованная постановка задачи и результаты сравнения методов решения.

Дата выдачи задания «20» сентября 2024 г.

Руководитель НИР _____ **А. С. Кострицкий**
(Подпись, дата) (И.О.Фамилия)

Студент _____ _____
(Подпись, дата) (И.О.Фамилия)

Временная метка генерации: 24.11.24 г., 22:06:52

РЕФЕРАТ

Расчетно-пояснительная записка 15 с., 2 рис., 11 ист., 2 табл., 1 прил.

Ключевые слова: методы хранения временных рядов, методжы прогнозирования временных рядов, методы оптимизации, модели машинного обучения, случайный лес, регрессия, решающие деревья, гауссовский процесс.

Объект исследования: методы хранения и прогнозирования временных рядов.

Цель работы: исследование методов хранения и прогнозирование временных рядов.

В работе выполняется формализация задачи прогнозирования временных рядов, рассмотрение методов решения задачи и сравнение методов.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Авторегрессионные модели (AR) являются фундаментальной концепцией в анализе и прогнозировании временных рядов. Их часто применяют в различных областях, включая финансы, экономику, климатологию и многое другое.

Научные работы о прогнозировании временных рядов появились еще в 1970 году в труде Бокса [8]. Бокс рассматривает линейные стационарные модели и некоторые системы, сводящиеся к ним. В частности, рассматривается модель ARIMA и ее производные, учитывающие сезонность и внешние факторы.

Значительный вклад в развитие прогнозирования временных рядов внесла работа о LSTM [10] в 2012 году, в которой рассматривалась модификация рекурентной нейронной сети для задач распознавания речи. Появляется множество статей, в которых модель LSTM с некоторыми модификациями превосходит в точности все ранее известные модели [9].

В это же время исследуется вопрос построения ансамблей моделей, когда на основе множества «простых» моделей обучается другой алгоритм регрессии, минимизирующий общую ошибку.

Целью данной работы является исследование методов хранения и прогнозирования временных рядов.

Для достижения поставленной цели необходимо выполнить следующие задачи:

— формализовать задачу прогнозирования временных рядов;
— проанализировать основные известные методы решения;
— сравнить методы решения по заранее сформулированным критериям.

1 Формализация задачи

Схема формализации задачи в виде диаграммы IDEF0 представлена на рисунке 1.1.

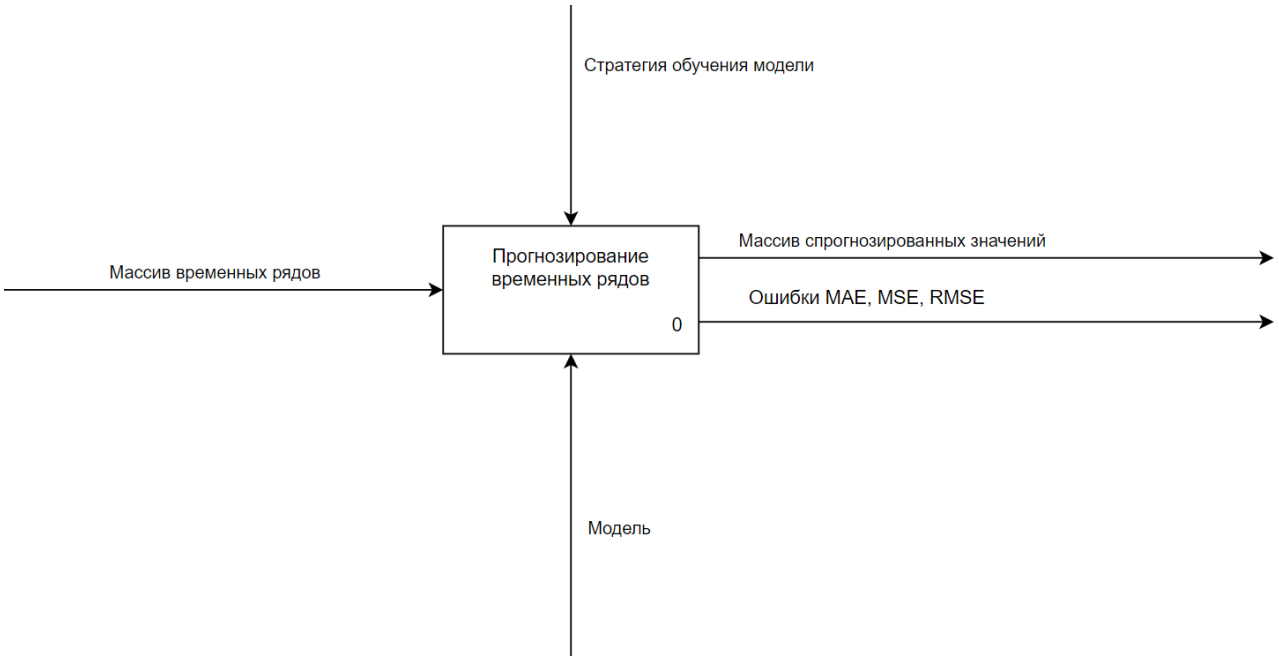

Рисунок 1.1 — IDEF0 диаграмма

1.1 Математическая формулировка

Даны временные ряды

$$X_i = \{x_{i,t}, t = \overline{1,k}\}, i = \overline{1,m} \tag{1.1}$$

где m — количество временных рядов, а k — размер временного ряда [5].

Задача состоит в том, чтобы построить такую модель $A(X_i)$, что

$$A(X_i) = x_{i,k+n} \tag{1.2}$$

где n количество значений, которые необходимо спрогнозировать [5].

При этом дополнительным условием является минимизация одной из ошибок MAE (Mean Abcolute Error), представленную формулой 1.3 или MSE (Mean Square Error), представленную формулой 1.4 для каждого временного

ряда.

$$MAE = \sum_{i=k+1}^{k+n} \frac{|y_i - x_i|}{n} \tag{1.3}$$

$$MSE = \sum_{i=k+1}^{k+n} \frac{(y_i - x_i)^2}{n} \tag{1.4}$$

В этих формулах y_i — истинное значение временного ряда, а x_i — спрогнозированное [4].

2 Анализ методов прогнозирования временных рядов

Временной ряд — это ряд точек данных, индексированных во временном порядке [11]. Чаще всего временной ряд — это последовательность, взятая в упорядоченных равноотстоящих точках времени. Таким образом, это последовательность дискретных временных данных.

Значения временного ряда является суммой его четырех основных компонент [1]:

— тренд — плавно меняющаяся компонента, описывающая чистое влияние долговременных факторов, т. е. длительную тенденцию изменения признака (например, рост населения);

— сезонность — компонента, отражающая повторяемость экономических процессов в течение не очень длительного периода (например, объем продаж товаров или перевозок пассажиров в различные времена года);

— цикличность — компонента, отражающая повторяемость экономических процессов в течение длительных периодов (например, влияние волн экономической активности Кондратьева);

— шум — случайная компонента, отражающая влияние не поддающихся учету и регистрации случайных факторов.

Чаще всего временные ряды хранят в виде массива значений, индексированного по времени, но иногда предпочитают хранить отдельно все компоненты ряда.

В этом разделе рассматриваются некоторые методы прогнозирования временных рядов.

2.1 ARIMA

ARIMA (Autoregressive Integrated Moving Average) — это модель авторегрессии с интегрированной скользящей средней. Определяется тремя параметрами: (p, d, q).

— авторегрессия AR(p) — регрессионная модель, которая использует зависимую связь между текущим наблюдением наблюдениями за предыдущий период;

— интеграция I(d) — использует дифференциацию наблюдений, чтобы

сделать временной ряд стационарным. Дифференциация включает вычитание текущих значений ряда из его предыдущих значений d раз;

— скользящее среднее MA(q) — модель, которая использует зависимость между наблюдением и остаточной ошибкой из модели скользящего среднего, применяемой к запаздывающим наблюдениям. Порядок q представляет собой количество членов, которые должны быть включены в модель.

Значения временного ряда считаются по следующей формуле 2.1.

$$X_i = c + \varepsilon_i + \sum_{k=1}^{p} \alpha_k X_{i-k} + \sum_{k=1}^{q} \beta_k \varepsilon_{i-k} \tag{2.1}$$

где c — некоторая константа, ε_i — значение шума, α_k — коэффициенты авторегрессии, β_k — коэффициенты скользящего среднего [1].

Недостатком этой модели является то, что она плохо справляется с данными, в которых ярко выражена компонента сезонности [1].

2.2 SARIMA

SARIMA (Seasonal Autoregressive Integrated Moving Average) — это расширение несезонной модели ARIMA, разработанное для обработки данных с сезонными закономерностями. Определяется четырьмя параметрами: (p, d, q, s).

Параметры (p, d, q) аналогичны параметрам модели ARIMA. Параметр s представляет сезонность, которая относится к повторяющимся закономерностям в данных [1].

Математическое представление модели выглядит следующим образом 2.2.

$$(1 - \phi_1 B)(1 - \Phi_1 B^s)(1 - B)(1 - B^s) y_t = (1 + \theta_1 B)(1 + \Theta_1 B^s)\varepsilon_t \tag{2.2}$$

где y_t — это наблюдаемый временной ряд в момент времени t, B — оператор обратного сдвига, представляющий оператор задержки (то есть $By_t = y_{t-1}$), ϕ_1 — коэффициент несезонной авторегрессии; Φ_1 — коэффициент сезонной авторегрессии, θ_1 — несезонный скользящий средний коэффициент, Θ_1 — сезонный скользящий средний коэффициент, s — сезонный период, ε_t — значение шума

в момент времени t [1].

2.3 VAR

Полулярной моделью связи между временными рядами является векторная авторегрессия (VAR – vector autoregression). Определяется одним параметром p, который задает параметры авторегрессии [7].

Математическое представление модели выглядит следующим образом 2.3.

$$y_t^i = a_0^i + \sum_{j=1}^{k} a_{1j}^i y_{t-1}^i + \sum_{j=1}^{k} a_{2j}^i y_{t-2}^i + ... + \sum_{j=1}^{k} a_{pj}^i y_{t-p}^i + \varepsilon_t^i \tag{2.3}$$

где X_t^i, $i = \overline{1,k}$ — i-ый временной ряд, а a_i^j — коэффициенты авторегрессии.

Равенство из формулы 2.3 можно переписать в векторной формуле 2.4.

$$\overrightarrow{y_t} = \overrightarrow{a_0} + \sum_{m=1}^{p} A_m \overrightarrow{y_{t-m}} + \overrightarrow{\varepsilon_t} \tag{2.4}$$

где A_m — матрица коэффициентов авторегрессии.

2.4 Решающие деревья

Решающие деревья (РД) представляют собой направленный иерархический граф. В его узлах стоят признаки, по которым идет разделение выборки, а в листьях — части выборки [6].

Построение решающих деревьев идет путем разделения выборки на части по вводимым признакам. Признаки и порог их значения, по которым делится выборка, нужно подбирать так, чтобы в листьях дерева оставались объекты одного класса.

Пусть в вершине X_m объектов. Выбираем порог t по критерию ошибки Q для признака j, минимизируя критерий ошибки 2.5.

$$Q(X_m, j, t) \to min \tag{2.5}$$

При использовании в качестве функционала ошибки в задачах регрессии значения среднеквадратичной ошибки, прогнозом будет среднее значение в

листе a_m, вычисленное по формуле 2.6.

$$a_m = \frac{1}{|X_m|} \sum_{i \in X_m} y_i \tag{2.6}$$

Пример решающего дерева представлен на рисунке 2.1.

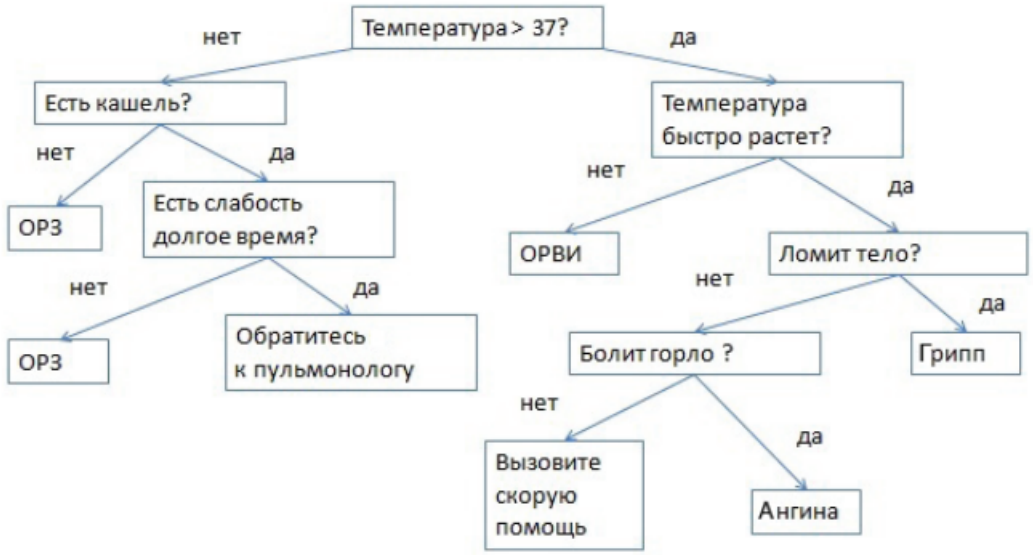

Рисунок 2.1 — Пример решающего дерева

2.5 Случайный лес

В методе случайного леса (СЛ) обучают каждый алгоритм из композиции, а ответом является усредненный результат по всем алгоритмам, входящим в композицию.

В случае регрессии ответ a(x) находится по формуле 2.7.

$$a(x) = \frac{1}{N} \sum_{i=1}^{N} b_n(x) \tag{2.7}$$

где N — количество деревьев в лесу, $b_i(x)$ — результат i-ого алгоритма [6].

2.6 Регрессия Гауссовского процесса

Gaussian Process Regression (GPR) — это гибкая непараметрическая техника регрессии. Она особенно полезна при работе с непрерывными данными, где связь между входными переменными и выходом не известна явно [3].

3 Сравнение методов

3.1 Критерии сравнения

В таблице 3.1 приведены критерии для сравнения методов прогнозирования временных рядов и их описание.

Таблица 3.1 — Критерии для сравнения методов прогнозирования временных рядов

Критерий	Описание
Параметры	Какие параметры, влияющие на прогноз, есть у метода
Дискретность	Поддерживает ли метод временные ряды, заданные дискретно
Непрерывность	Поддерживает ли метод временные ряды, заданные непрерывно
Сезонность	Поддерживает ли метод прогнозирование на данных с ярко выраженной сезонностью
Многомерность	Поддерживает ли метод прогнозирование многомерных данных
Композиция	Является ли метод композицией других методов

3.2 Сравнение

В таблице 3.2 представлены результаты сравнения методов прогнозирования временных рядов по заданным выше критериям.

Таблица 3.2 — Результаты сравнения методов прогнозирования временных рядов

Метод	ARIMA	SARIMA	VAR	РД	СЛ	GPR
Параметры	p, d, q	p, d, q, s	p	X_m, j, t	X_m, j, t	—
Дискретность	Да	Да	Да	Да	Да	Нет
Непрерывность	Нет	Нет	Нет	Да	Да	Да
Сезонность	Нет	Да	Да	Да	Да	Нет
Многомерность	Нет	Нет	Да	Да	Да	Нет
Композиция	Нет	Нет	Нет	Нет	Да	Нет

ЗАКЛЮЧЕНИЕ

Цель данной работы была достигнута: исследованы методы хранения и прогнозирования временных рядов. Были решены все задачи:

— формализована задача прогнозирования временных рядов;

— проанализированы основные методы решения задачи;

— сформулированы критерии для сравнения методов решения задачи;

— по сформулированным критериям проведено сравнение методов решения задачи.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. К.О. Кизбикенов. Прогнозирование и временные ряды. — Город: Барнаул, АлтГПУ, 2017. — 115 с.

2. В.Н. Афанасьев. Анализ временных рядов и прогнозирование. — Город: Саратов, Оренбург, Ай Пи Эр Медиа, 2020, 286 с.

3. С.В. Каштаева. Методы оптимизация. — Город: Пермь, ИПЦ «Прокростъ», 2020, 84 с.

4. А.А. Хватов, Н.О. Никитин, А.В. Калюжная. Современные методы оптимизации с примерами на Python. — Город: Санкт-Петербург, Редакционно-издательский отдел Университета ИТМО, 2023, 53 с.

5. Н. Л. Майорова, Д. В. Глазков. Методы оптимизации. — Город: Ярославль, ЯрГУ, 2015, 112 с.

6. О. В. Лимановская, Т. И. Алферьева. Основы машинного обучения. — Город: Екатеринбург, Издательство Уральского университета, 2020, 92 с.

7. В.П.Носко. Эконометрика. Введение в регрессионный анализ временных рядов, — Город: Москва, 2002, 254 с.

8. G.E.P. Box, G.M. Jenkins и Wisconsiv Univ Madison. «Dept. of statistics. Time Series Analysis: Forecasting and Control. Holden-Day series in time series analysis and digital processing», Holden-Day, 1970

9. Jian Cao, Zhi Li и Jian Li. «Financial time series forecasting model based on CEEMDAN and LSTM». В: Physica A: Statistical Mechanics and its Applications 519 (2019), с. 127—139.

10. Martin Sundermeyer, Ralf Shluter и Hermann Ney, «LSTM Neural Networks for Language Modeling», 2012

11. Suman Kalyan Adari Sridhar Alla. Beginning Anomaly Detection Using Python-Based Deep Learning, 2019, 538 с.

ПРИЛОЖЕНИЕ А

Презентация к научно-исследовательской работе

Презентация к научно-исследовательской работе содержит 3 слайда.